

Curs 1:

# Introducere în Data Mining

# Preliminarii

Cum ați traduce **Data Mining**?

1. Analiza datelor
2. Explorarea datelor
3. Exploatarea datelor
4. Extragerea de cunoștințe din date
5. Mineritul datelor

# Preliminarii

- De ce Data Mining? – o scurtă motivație
- Ce este Data Mining? – concepte de bază
- Ce nu este Data Mining? – tematici corelate
- Categoriile de date
- Principalele tipuri de prelucrări
- Organizarea cursului și criteriile de evaluare

# De ce Data Mining?

La ora actuală se colectează și devine **accesibil** un **volum foarte mare de date** de diferite tipuri și provenind din diferite surse:

- tranzacții comerciale (ex: hipermarket-uri)
- tranzacții financiare (ex: bancomate)
- utilizarea unor resurse web (ex: comerț electronic, alte servicii web)
- interacțiuni sociale(ex: rețele sociale)
- date satelitare (ex: date privind Pământul și atmosfera colectate folosind senzori plasați pe sateliți)
- date genomice (ex: date referitoare la nivelul de exprimare a genelor colectate folosind dispozitive de tip microarrays)
- date medicale(ex: înregistrări medicale în format electronic)
- documente în format electronic (ex: documente scanate – în biblioteci, arhive electronice etc.)

...

# De ce Data Mining?

- Toate aceste date **încorporează o mulțime de cunoștințe** care ar trebui extrase în diferite scopuri:
  - Pentru a se genera **recomandări** (ex: pentru a ghida activitatea de marketing, pentru a sugera produse clienților)
  - Detectarea **comportamentului anormal** (ex: acces fraudulos la un cont bancar)
  - **Predicție** (ex: în meteorologie, evoluția pieței/ prețurilor)
  - Identificarea de **tipare** (ex: identificarea rolului unei gene)
  - Asistarea deciziei medicale (ex: furnizarea unor sugestii privind diagnosticul potențial)

# De ce Data Mining?

## **Exemplu 1: Date referitoare la fertilizarea in vitro**

[Witten, Frank, Hall – Data Mining. Practical Machine Learning Tools and Techniques - <http://www.cs.waikato.ac.nz/ml/weka/book.html>]

**Se pornește de la:** embrioni descriși prin 60 de caracteristici

**Problema:** selectarea acelor embrioni care au șanse de supraviețuire

**Date:** înregistrări istorice cu caracteristici ale embrionilor și informații privind viabilitatea lor (fertilizare cu succes sau fără succes)

## **Exemplu 2: Procesarea aplicațiilor pentru împrumut**

**Se pornește de la:** chestionar cu informații financiare și personale (ex: vârsta, date privind locul de muncă, starea de sănătate, starea financiară etc)

**Problema:** decizia dacă se acordă împrumutul sau nu

**Date:** înregistrări istorice conținând informații personale și financiare precum și privind rambursarea acestuia (dacă a fost rambursat la timp sau au existat probleme)

# De ce Data Mining?

**Exemplu 3: Predicția încărcării unei rețele de distribuție a energiei electrice** (estimarea cererii viitoare de putere electrică – util pentru companiile distribuitoare)

**Se cunoaște:** un model de încărcare a rețelei în cazul unor condiții climatice normale

**Problema:** predicția încărcării minime/ maxime la anumite momente (de exemplu din oră în oră)

**Date:** înregistrări istorice privind condițiile meteo (temperatura, umiditatea, viteza vântului, gradul de acoperire a cerului) și gradul de încărcare a rețelei

**Exemplu 4: Analiza coșului de cumpărături**

**Date:** bază de date cu tranzacții (o tranzacție conține informații despre produsele cumpărate de către fiecare client)

**Problema:** identificarea grupurilor de produse care apar frecvent împreună în aceeași tranzacție (ex: pâine și lapte)

# De ce Data Mining?

## Exemplu 5: Detectarea anomaliilor

**Date:** date privind tranzacții financiare

**Problema:** identificarea unei schimbări în comportamentul utilizatorilor

## Exemplu 6: Identificarea profilelor utilizator

**Date:** fișiere cu date de conectare la un server web (log files)

**Problema:** identificarea unor profile de utilizatori (grupuri de utilizatori caracterizați prin comportament similar)

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0"  
200 2326
```



# Ce este Data Mining?

Există diferite definiții:

Data mining = “colectarea, curățirea, procesarea, analiza datelor și extragerea de informații sau cunoștințe utile din ele” [C.Aggarwal – Data Mining. The Textbook, 2015]

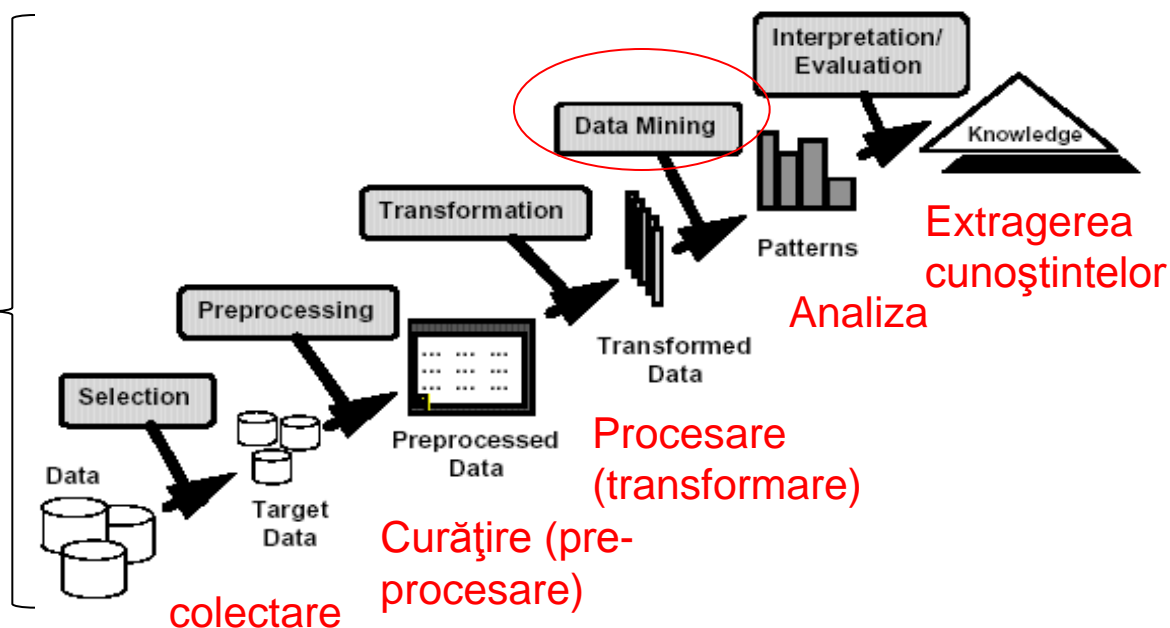
- **Colectare:** există diferite surse de date (senzori, documente scrise, servere web, dispozitive de tip microarray etc)
- **Curățire:** eliminarea zgomotului (a inconsistențelor sau a datelor eronate) și tratarea valorilor absente
- **(pre)Procesare:** transformarea datelor într-un format standardizat
- **Analiza:** identificarea tiparelor, a regularităților, a asocierilor sau a relațiilor existente în date
- **Extragere cunoștințe:** formularea unor reguli concise și aplicabile (care ar putea fi folosite de către utilizatori)

# Ce este Data Mining?

Există diferite definiții:

Data mining = “**extragerea** din date a cunoștințelor implicite, anterior necunoscute și potențial utile” [<http://www.cs.waikato.ac.nz/ml/weka/book.html>] sau “**explorarea** și **analiza**, prin mijloace automate sau semi-automate, a unei cantități mari de date cu scopul de a identifica tipare utile/ relevante” [Tan, Steinbach, Kumar – Introduction to Data Mining, 2004]

Uneori acest proces este denumit **descoperirea cunoștințelor** iar termenul data mining referă doar o etapă a acestui proces



# Ce nu este Data Mining?

**Exemplu:** se consideră o bază de date ce conține informații despre clienții unei bănci:

- Căutarea tuturor clienților ce locuiesc într-un oraș specificat **nu este o prelucrare specifică pentru data mining**
- Determinarea numărului de clienți care au în cont o sumă mai mică sau mai mare decât o valoare specificată **nu este o prelucrare specifică pentru data mining**

... astfel de probleme se rezolvă prin **interogări simple ale bazei de date**

Pe de altă parte:

- Identificarea clienților cărora li se poate acorda un împrumut
- Identificarea operațiunilor anormale într-un cont

... sunt probleme care necesită **expertiză umană și/sau instrumente de data mining**

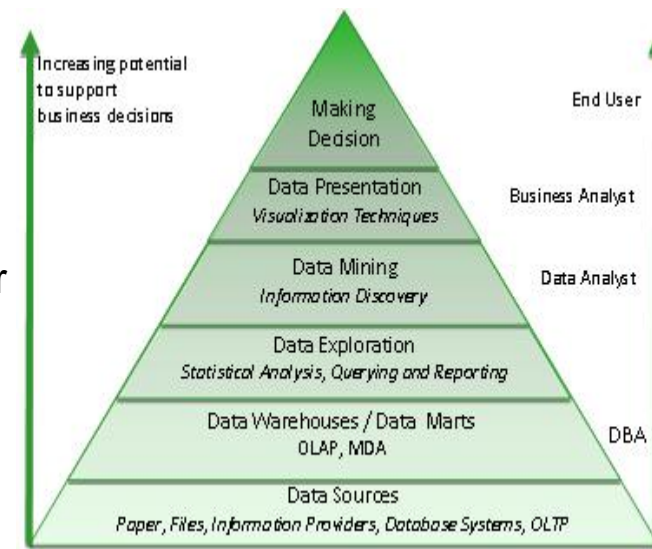
# Domenii înrudite

Data mining este un domeniu înrudit cu:

- **Statistica** – unele tehnici din data mining au rădăcini și se bazează pe metode statistice
- **Învățare automată** = extragerea de modele din date printr-un proces de învățare – cele mai multe modele din data mining se bazează pe metode de învățare
- **Baze de date** – cele mai multe date sunt stocate în baze de date
- Alte domenii:
  - **Vizualizare**: instrumente pentru vizualizarea datelor
  - **Optimizare**: multe procese de extragere a modelelor din date se bazează pe optimizarea unor criterii
  - **Algebră liniară**: datele sunt frecvent organizate în matrici, a.i. sunt frecvent folosite prelucrări asupra matricilor

## Altă definiție

Data mining =  
Aplicarea metodelor de învățare automată pentru extragerea de cunoștințe din date



Alți termeni corelați: **data science**, **big data**

# Categorii de date

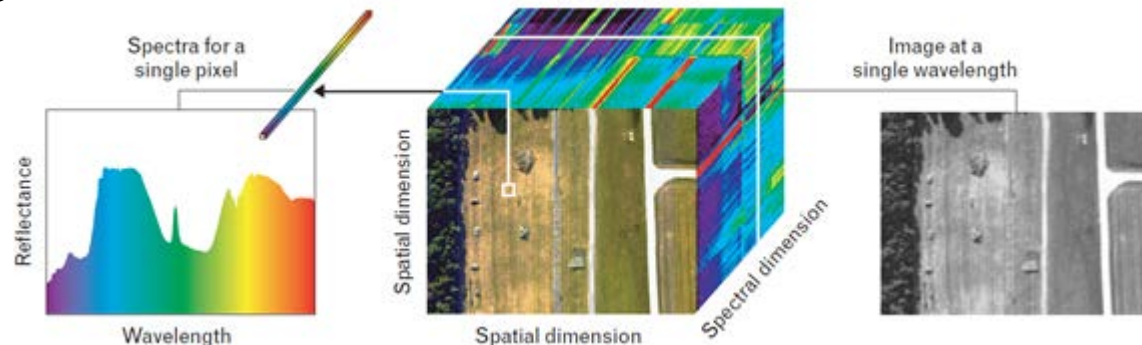
**Date structurate** = set de înregistrări/**instanțe**/articole conținând un număr fix de câmpuri/ **atribute**/caracteristici

Obs:

- Fiecare instanță corespunde unui obiect/entitate de analizat (ex: client, pacient, tranzacție, zi etc.)
- Fiecare atribut corespunde unei caracteristici măsurabile a obiectului (ex: vârsta, greutate, venit, temperatură etc.)

**Exemple:**

- **Tablouri bi-dimensionale (i.e. Matrice de date)**
  - Baze de date relaționale
  - Foi de calcul
- **Tablouri multi-dimensionale**
  - Imagini multi-spectrale



# Categorii de date

**Date structurate** = set de înregistrări/**instanțe**/articole conținând un număr fix de câmpuri/ **attribute**/caracteristici

**Exemplu:** Car Evaluation Database [<http://archive.ics.uci.edu/ml/datasets.html>]

1728 instanțe 6 attribute

**Scop:** clasificarea unei mașini în una din patru categorii: inacceptabilă, acceptabilă, bună, foarte bună

## Atribute

Instanța	Preț cumpărare	Preț întreținere	Nr uși	Capacitate	Dim. bagaj	Siguranța	Clasa
1	Very high	Very high	2	2	small	low	inaccept.
2	Very high	high	4	4	big	medium	inaccept.
3	Very high	medium	5more	4	big	medium	accept
4	low	low	5more	4	big	medium	bună

# Categorii de date

**Date semi - structurate** = date care nu au o structură standard (i.e. nu toate instanțele au aceleași attribute); există totuși unele elemente (e.g. tags) care ajută la identificarea unei structuri în date

**Exemplu:** fișier XML al unui CV [<http://www.eife-l.org>]

**Scop:** prelucrarea automată a CV-urilor cu scopul identificării expertizei (sarcină tipică pentru departamentele HR)

```
....  
<Address type="Residence">  
  <oa:AddressLine sequence="1">myaddress</oa:AddressLine>  
  <oa:CityName>mycity</oa:CityName>  
  <CountryCode>FR</CountryCode>  
  <oa:PostalCode>29630</oa:PostalCode>  
  <UserArea> <europass:CountryLabel  
                xml:lang="fr">France</europass:CountryLabel>  
  </UserArea>  
</Address>
```

...

**Obs:** datele semi-structurate sunt de regulă transformate în date structurate înainte de a aplica tehnici de data mining – relativ ușor de parsat

# Categorii de date

**Date nestructurate** = nu sunt organizate într-o manieră predefinită (nu există un model al datelor) – sunt de obicei texte în format liber și scopul urmărit este extragerea de informații din text.

**Exemplu:** documente text

**Prelucrări:**

- sumarizarea documentelor (extragere cuvinte cheie, idei principale)
- Identificarea entităților cu nume (ex: nume de persoane, nume de instituții, locuri geografice etc)

**Dificultăți:**

- Datele pot fi ambigue (ex: Numele unei persoane poate apare în diferite variante: Ioan Popescu, I. Popescu, Popescu Ioan)
- Prelucrarea datelor de tip text necesită metode specifice prelucrării limbajului natural (ex: etichetarea părților de vorbire –substantive, verbe, adjective ...)



# Tipuri de prelucrări

## Prelucrări predictive

**Scop:** predicția unor valori necunoscute sau viitoare ale unor atribute pe baza valorilor celorlalte atribute

### Variante:

- **Clasificare** = identificarea clasei (categoriei) căreia ar trebui să îi aparțină o anumită instanță (pe baza valorilor atributelor ei)

**Exemple:** datele referitoare la fertilizarea in vitro, la evaluarea cererilor de împrumut bancar

- **Regresie** = estimarea valorii unui atribut pe baza valorilor altor atribute

**Exemplu:** predicția încărcării rețelei de distribuție a energiei electrice

# Tipuri de prelucrări

## Prelucrări descriptive

**Scop:** identificarea unor tipare interpretabile care permit descrierea sau explicarea datelor

### Variante:

- **Clustering (grupare)** = identificarea unor grupuri naturale în date  
**Exemplu:** identificarea profilelor utilizator
- **Asociere** = descoperirea unor reguli de asociere între atribute  
**Exemplu:** analiza coșului de cumpărături
- **Excepții sau anomalii** = identificarea entităților (instanțe) care par anormale într-un anumit sens (de obicei semnificativ diferite de celelalte)  
**Exemplu:** detecția activității frauduloase

# Clasificare

## Ce se cunoaște?

- O colecție de instanțe (înregistrări) pentru care se cunoaște clasa căreia îi aparțin (**set de antrenare**)
- Fiecare dintre instanțe conține un set de **atribute**, iar unul dintre aceste atribute este eticheta **clasei**

## Ce se dorește?

- un **model** care captează legătura dintre atributul clasă și celelalte atribute (modelul este extras pornind de la **setul de antrenare** printr-un proces numit **învățare supervizată**)

## Care este scopul final?

- Să se folosească modelul extras din date pentru a identifica clasa căreia îi aparține o instanță nouă (care nu face parte din setul de antrenare)

**Observație:** un model util trebuie să fie caracterizat printr-o bună capacitate de predicție (acuratețe); acuratețea modelului poate fi estimată utilizând date pentru care se cunoaște clasa căreia îi aparțin dar care nu au fost utilizate în extragerea modelului (**set de testare**)

# Clasificare

Exemplu:

- **Diagnoza medicală** = asociază unei înregistrări medicale o clasă (prezența sau absența unei boli)

Exemplu subset dintr-un set de date (breast-cancer-wisconsin - format arff—  
vezi Lab 1)

```
@relation wisconsin-breast-cancer
@attribute Clump_Thickness integer [1,10]
@attribute Cell_Size_Uniformity integer [1,10]
@attribute Cell_Shape_Uniformity integer [1,10]
@attribute Marginal_Adhesion integer [1,10]
@attribute Single_Epi_Cell_Size integer [1,10]
@attribute Bare_Nuclei integer [1,10]
@attribute Bland_Chromatin integer [1,10]
@attribute Normal_Nucleoli integer [1,10]
@attribute Mitoses integer [1,10]
@attribute Class { benign, malignant}
@data
5,1,1,1,2,1,3,1,1,benign
5,4,4,5,7,10,3,2,1,benign
3,1,1,1,2,2,3,1,1,benign
8,10,10,8,7,10,9,7,1,malignant
1,1,1,1,2,10,3,1,1,benign
```

# Clasificare

## Exemplu:

- **Filtru anti-spam** = identificarea clasei (ilegitim=spam/legitim=ham) unui mesaj (e-mail sau SMS)

## Exemplu subset date (SMS spam collection dataset din UCI Machine Learning Repository)

ham What you doing?how are you?

ham Ok lar... Joking wif u oni...

ham dun say so early hor... U c already then say...

ham MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H\*

ham Siva is in hostel aha:-.

spam FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! unsubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop

spam Sunshine Quiz! Win a super Sony DVD recorder if you can name the capital of Australia? Text MQUIZ to 82277. B

spam URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

....

# Regresie

## Ce se cunoaște?

- O colecție de instanțe caracterizate prin atribute numerice (**set de antrenare**)

## Ce se urmărește?

- Un model al dependenței între unul dintre atribute (**atributul răspuns**) și celelalte atribute (**atribute predictor**)

## Care este scopul final?

- Să se prezică valoarea atributului răspuns pe baza valorilor cunoscute ale celorlalte atribute.

## Observație

- se poate presupune de la început că modelul de regresie satisface anumite proprietăți (este liniar sau neliniar); modelul poate fi fixat, ca în regresia statistică, sau poate fi flexibil (ca în cazul rețelelor neuronale sau a altor modele din inteligența computațională)

# Regresie

## Exemple:

- Predicția volumului de vânzări a unui produs nou în funcție de cheltuielile pentru publicitate.
- Predicția vitezei vântului în funcție de temperatură, umiditate, presiunea aerului etc.
- Predicția evoluției în timp a indicilor bursieri.

Set de date: predicția consumului de combustibil în funcție de caracteristicile mașinii (UCI Machine Learning Repository)

@relation autoMpg

@attribute cylinders { 8, 4, 6, 3, 5}      @attribute displacement real

@attribute horsepower real      @attribute weight real      @attribute acceleration real

@attribute model { 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82}

@attribute origin { 1, 3, 2}

@attribute MilesPerGallon real

@data

8,307,130,3504,12,70,1,18

8,350,165,3693,11.5,70,1,15

8,318,150,3436,11,70,1,18

8,304,150,3433,12,70,1,16

....

# Clustering

## Ce se cunoaște?

- Un **set de date** (nu neapărat structurate)
- O **măsură de similaritate/disimilaritate** între date (este specifică problemei)

## Ce se urmărește?

- Identificarea unui **model** care descrie modul în care pot fi grupate datele în clustere astfel încât datele aparținând aceluiași cluster sunt mai similare între ele decât datele aparținând unor clustere diferite

## Care este scopul final?

- Să se poată verifica dacă două date aparțin aceluiași cluster sau nu
- Să se identifice clusterul adecvat unei date
- Să se identifice/vizualizeze modul în care se grupează datele

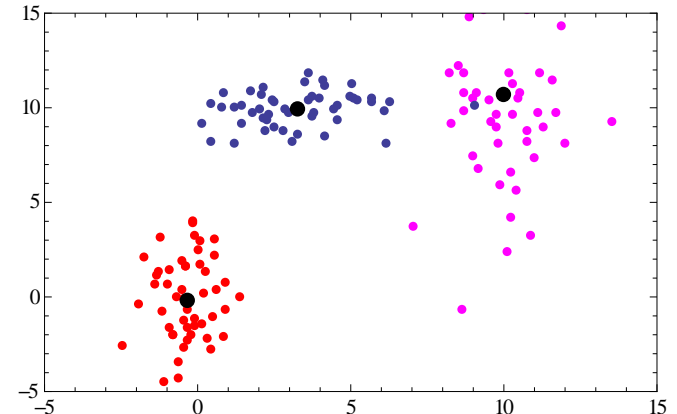
**Observație:** pentru unele metode de grupare nu e necesar să se cunoască datele primare fiind suficient să se cunoască matricea de (di)similaritate



# Clustering

## Exemple:

- **Gruparea clienților** = identificarea de grupuri de clienți cu obiceiuri similare de cumpărare
- **Sumarizarea datelor / gruparea documentelor** = identificarea de grupuri de documente pe baza conținutului
- **Extragerea profilelor de utilizatori** = identificarea grupurilor de utilizatori ai unui serviciu web caracterizați prin comportament similar
- **Segmentarea imaginilor** = identificarea de regiuni omogene în imagini



# Analiza excepțiilor

## Ce se cunoaște?

- Un **set de data** (nu neapărat structurate)
- O **măsură de similaritate/disimilaritate** între date (este specifică problemei)

## Ce se urmărește ?

- Identificarea unui model care corespunde comportamentului normal

## Care este scopul final?

- Identificarea excepțiilor, adică a datelor care se abat semnificativ de la model (valori atipice)

**Observație:** este oarecum complementară grupării datelor

# Analiza excepțiilor

## Exemple:

- **Sisteme de detecție a intrușilor**
  - Apeluri sistem anormale sau trafic anormal în rețea pot sugera prezența unei activități malițioase
- **Fraudă bancară**
  - Un comportament neobișnuit în utilizarea unei cărți de credit (e.g.utilizarea cardului din locații geografice neobișnuite sau la ore neobișnuite) poate sugera o posibilă activitate frauduloasă
- **Diagnoza medicală**
  - Structuri anormale observate pe imagini MRI(magnetic resonance imaging), PET (positron emission tomography) sau secvențe EKG pot indica prezența unor patologii

# Reguli de asociere

Ce se cunoaște?

- Un set de înregistrări, fiecare conținând obiecte (entități) dintr-o colecție

Ce se urmărește?

- Să se găsească un model care să permită estimarea prezenței unui obiect în ipoteza prezenței altor obiecte

Care este scopul final?

- Identificarea unor tipare de asociere între obiecte

# Reguli de asociere

**Exemplu:** analiza coșului de cumpărături (fiecare instanță corespunde unei tranzacții = listă de produse cumpărate)

T1: {lapte, pâine, carne, apă}

T2: {pâine, apă}

T3: {pâine, unt, carne, apă}

T4: {apă}

## Rezultate:

- **Itemset frecvent:** {pâine, apă} - suport 75% (perechea de produse apare în 3 din 4 tranzacții) – se poate spune că “pâinea și apa sunt cumpărate frecvent împreună”
- **Regulă de asociere:** pâine->apă (100% nivel de încredere: în toate cazurile atunci când este cumpărată pâine este cumpărată și apă)

# Structura cursului

## Tematici

1. Introducere (acest curs)
2. Pre-procesarea datelor
3. Tehnici de clasificare
4. Tehnici de grupare
5. Reguli de asociere
6. Regresie și analiza seriilor temporale
7. Analiza excepțiilor
8. Meta-modele și tehnici de tip ansamblu
9. Tehnici specifice (text mining, web mining, network analysis)

Materiale: <http://staff.fmi.uvt.ro/~daniela.zaharie/dm2018/RO>

- slide-uri curs
- exerciții laborator

# Structura laboratorului

1. Seturi și colecții de date. Introducere în Rattle, Weka
2. Pre-procesarea datelor (curățare, transformare, reducere dimensiune)
3. Clasificarea datelor (clasificatori bazați pe instanțe, arbori și reguli de decizie)
4. Clasificarea datelor (modele probabiliste, rețele neuronale, vectori suport)
5. Gruparea datelor (algoritmi partiționali, ierarhici, bazați pe densitate)
6. Reguli de asociere. Modele de regresie.
7. Analiza seriilor temporale. Metode de tip ansamblu. Text mining.

# Bibliografie

- C.C. Aggarwal, *Data Mining – The Text Book*, Springer, 2015
- M. H. Dunham. *Data Mining. Introductory and Advanced Topics*, Pearson Education 2003
- F. Gorunescu, *Data Mining. Concepts, Models and Techniques*, Springer, 2011
- C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.
- I.H. Witte, E. Frank, M.A. Hall. *Data Mining – Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, 2011



# Evaluare

- Examen cu acces la materiale bibliografice (20%) – 20 întrebări/ 90 minute
- Proiect (60%):
  - Raport (6-12 pagini)
  - Aplicație (în R, Python, Weka sau alt limbaj de programare)
  - Slide-uri pt prezentarea de la examen (cca 10 minute)
- Activitate laborator (20%)
  - participare
  - teme

# Exemple de aplicații

(Store Product Placement) A merchant has a set of  $d$  products together with previous transactions from the customers containing baskets of items bought together. The merchant would like to know how to place the product on the shelves to increase the likelihood that items that are frequently bought together are placed on adjacent shelves.

Ce prelucrare este adecvată?

(Product Recommendations) A merchant has an  $n \times d$  binary matrix  $D$  representing the buying behavior of  $n$  customers across  $d$  items. It is assumed that the matrix is sparse, and therefore each customer may have bought only a few items. It is desirable to use the product associations to make recommendations to customers.

Ce prelucrare este adecvată?

# Exemple de aplicații

(Store Product Placement) A merchant has a set of  $d$  products together with previous transactions from the customers containing baskets of items bought together. The merchant would like to know how to place the product on the shelves to increase the likelihood that items that are frequently bought together are placed on adjacent shelves.

Reguli de asociere

(Product Recommendations) A merchant has an  $n \times d$  binary matrix  $D$  representing the buying behavior of  $n$  customers across  $d$  items. It is assumed that the matrix is sparse, and therefore each customer may have bought only a few items. It is desirable to use the product associations to make recommendations to customers.

Clustering

# Exemple de aplicații

(Medical ECG Diagnosis) Consider a set of ECG time series that are collected from different patients. It is desirable to determine the anomalous series from this set.

Ce prelucrare este adecvată?

(Web Log Anomalies) A set of Web logs is available. It is desired to determine the anomalous sequences from the Web logs.

Ce prelucrare este adecvată?

# Exemple de aplicații

(Medical ECG Diagnosis) Consider a set of ECG time series that are collected from different patients. It is desirable to determine the anomalous series from this set.

(Web Log Anomalies) A set of Web logs is available. It is desired to determine the anomalous sequences from the Web logs.

Detecție anomalii  
Clasificare

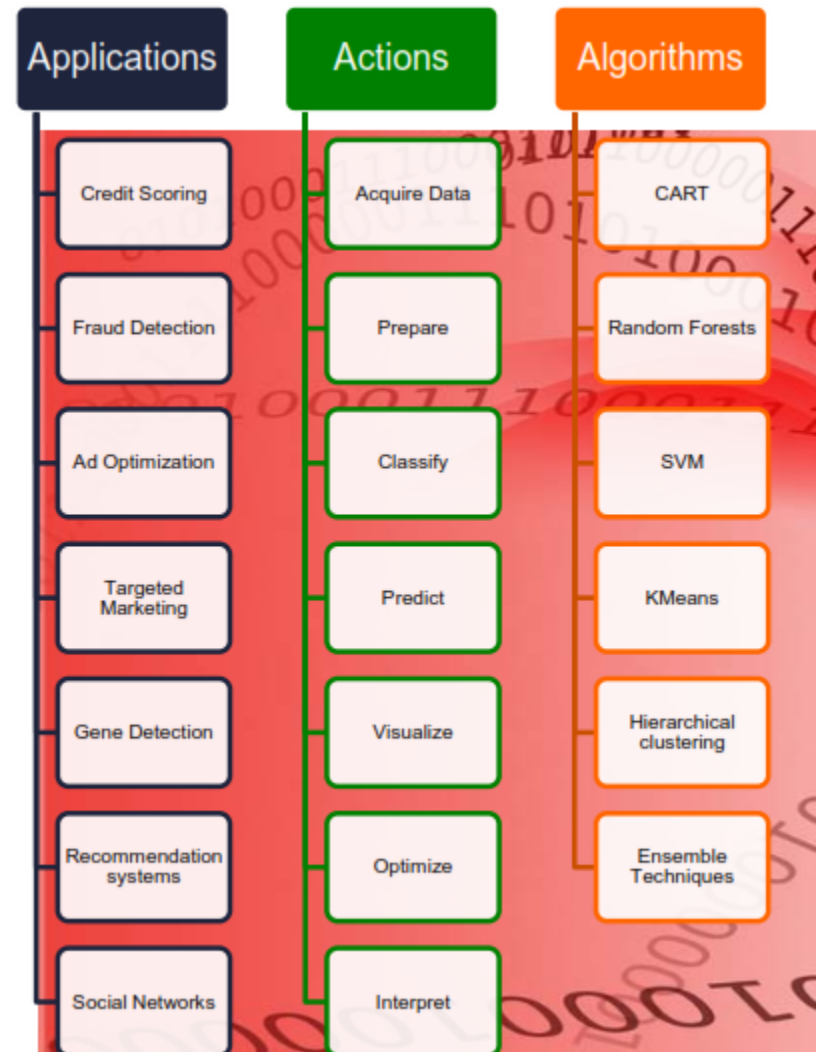
# Sumar

## Data mining:

- Aplicație
- Task (acțiune)
- Metoda (algoritm)

## Elemente cheie:

- Stabilirea întrebării adecvate
- Identificarea datelor adecvate
- Pregătirea datelor
- Selectarea algoritmilor adecvați
- Interpretarea rezultatelor



# Sumar: a roadmap for a data scientist

"A data scientist is a person who is better at statistics than any software engineer and better at software engineering than any statistician."

