Data Mining

Arta si Ştiinţa de a obţine Cunoştinţe din Date

Prof. univ. dr. ing. Ștefan HOLBAN

Explozia Informationala - paradox -

- Informatia a devenit o marfa care se produce si se vinde.
- Problema care se ridica este ca se produce mai mult decat se poate consuma.
- Acest aspect ridica intrebarea CE este important sa retin in cursul procesului de cunoastere din informatia aflata la dispozit**ie**

-Definitii-
-Informatia-

Într-o definire - pe cât de sumară tot pe atât de informală și, deci, de inexactă - se poate spune că informația se constituie intr-o reprezentare a realității, dar și a reflecției și proiecției - care sunt operații tipice intelectului uman - prin intermediul unui set bine precizat și structurat de simboluri - de regulă accesibile simțurilor și rațiunii umane, dar și unora dintre dispozitive, precum cele de calcul automat (calculatoare).

Informatia nu este nici conținut (dar stările unui sistem pot fi asimilate cu acesta), nici agent (dar semnalele transmise printr-un canal pot fi asimilate cu acesta), nici proprietate, nici instructiune, nici proces și nici metoda. Informația se constituie într-o categorie de sine stătătoare, având o existență abstractă și subtilă - adică nematerială - categorie care este reflectată de stări, semnale etc. și constituie un element esențial în procesul cunoașterii.

În ultimele decenii ale sec. XX, creșterea gradului de informatizare a proceselor industriale precum și a creșterii gradului de folosire a informațiilor în rezolvarea problemelor a făcut ca informația să fie considerată ca o resursă economică, întrucâtva egală cu alte resurse cum ar fi munca, materia primă și capitalul.

Cât de mare este un Exabyte

pana in 2009 in ordine de marime

Kilobyte (KB)	1,000 bytes OR 10^3 bytes 2 Kilobytes: A Typewritten page. 100 Kilobytes: A low-resolution photograph.
Megabyte (MB)	1,000,000 bytes OR 10^6 bytes 1 Megabyte: A small novel OR a 3.5 inch floppy disk. 2 Megabytes: A high-resolution photograph. 5 Megabytes: The complete works of Shakespeare. Megabytes: A minute of high-fidelity sound. 100 Megabytes: 1 meter of shelved books. 500 Megabytes: A D-ROM.
Gigabyte (GB)	1,000,000,000 bytes OR 10^9 bytes 1 Gigabyte: a pickup truck filled with books. 20 A good collection of the works of Beethoven. 100 Gigabytes: A library floor of academic journals.
Terabyte (TB)	1,000,000,000,000 bytes OR 10^{12} bytes 1 Terabyte: 50000 trees made into paper and printed. 2 Terabytes: An academic research library. 10 Terabytes: The print collections of the U.S. Library of Congress. 400 Terabytes: National Climactic Data Center (NOAA) database.
Petabyte (PB)	1,000,000,000,000,000 bytes OR 10^{15} bytes 1 Petabyte: 3 years of EOS data (2001). 2 Petabytes: All U.S. academic research libraries. 20 Petabytes: Production of hard-disk drives in 1995. 200 Petabytes: All printed material.
Exabyte (EB)	1,000,000,000,000,000,000 bytes OR 10^{18} bytes 2 Exabytes: Total volume of information generated in 1999. 5 Exabytes: All words ever spoken by human beings.

Explozia Informationala

pana in 2009

Cresterea anuala a cantitatii de informatie stocata este estimata
la un procent anual de aproximativ ~30% ea dublându-se practic la 20 luni!

Studiile efectuate au aratat ca:

- **de la inceputul aparitiei omului si pana in 1999 au fost generate 12 terabyte de date.**
- **In lume cantitatea de date a crescut de la 5 exabytes in 2003**
  la 161 exabytes in 2006
- **In 2008 cantitatea cantitatea de date a crescut la 255 exabytes**
- **In 2010 s-au produs 988 exabytes.**
- **In 2013 cantitatea a crescut la 5 zettabytes (1 zettabytes = 1000 exabytes)**

Cantitatea totala de date produsa in lume (tiparit, film, optic, magnetic) in 2009
cere 1.5 miliarde de Gb de spatiu de stocare
Acesta este echivalent cu 250 MB de date pentru fiecare locuitor al acestei planete

Explozia Informationala

incepand cu 2010
ordine de marime

Multiples of bytes					v·d·e
SI decimal prefixes		Binary usage	IEC binary prefixes		
Name (Symbol)	Value		Name (Symbol)	Value	
kilobyte (kB)	10^3	2^{10}	kibibyte (KiB)	2^{10}	
megabyte (MB)	10^6	2^{20}	mebibyte (MiB)	2^{20}	
gigabyte (GB)	10^9	2^{30}	gibibyte (GiB)	2^{30}	
terabyte (TB)	10^{12}	2^{40}	tebibyte (TiB)	2^{40}	
petabyte (PB)	10^{15}	2^{50}	pebibyte (PiB)	2^{50}	
exabyte (EB)	10^{18}	2^{60}	exbibyte (EiB)	2^{60}	
zettabyte (ZB)	10^{21}	2^{70}	zebibyte (ZiB)	2^{70}	
yottabyte (YB)	10^{24}	2^{80}	yobibyte (YiB)	2^{80}	
See also: Multiples of bits · Orders of magnitude of data					

Un zettabyte este o unitate egala cu sextilion de bytes
1,000,000,000,000,000,000,000 bytes = $1000^7 = 10^{21}$
Un zettabyte este 1 miliard de terabytes

Explozia Informationala

incepand cu 2010

ordine de marime

Cum se utilizeaza aceasta informatie

•***Studiile facute au aratat ca in medie un cetatean SUA***

- ***vorbeste la telefon*** **16.17 ore pe luna**
- **asculta la radio 90 ore pe luna,**
- **priveste la TV 131 ore pe luna**

•Aproximativ 53% din populatia USA utilizeaza internetul intr-o luna:

- **25 ore si 25 minute acasa**
- **74 ore si 26 minute la lucru**

in total 13% din timpul disponibil / luna

•Membrii societăţii de tip occidental sunt supuşi unui adevărat bombardament informaţional: conform unui studiu american recent, fiecare primeşte, zilnic, o cantitate de informaţie echivalentă cu cea cuprinsă în 147 de ziare!

•Dezvoltarea internetului, programele de televiziune disponibile 24 de ore din 24, precum şi răspândirea telefoanelor mobile au făcut ca, în ziua de azi, o persoană să primească, în fiecare zi, de 5 ori mai multă informaţie decât primea în 1986.

Explozia Informationala

incepand cu 2010

ordine de marime

Cum se utilizeaza aceasta informatie

- **Se trimit aproximativ 3 milioane emails / secunda,**
- **20 ore video sunt uploaded in YouTube in 60 secunde,**
- **Google proceseaza 24 petabytes de informatiie,**
- **se trimit 50 milioane SMS per zi**

•Aproape 73 produse sunt comandate pe Amazon in fiecare secunda

•Zilnic, o persoană produce şi transmite altora, în medie, informaţie într-o cantitate echivalentă cu cea cuprinsă în 6 ziare - de 200 ori mai mult decât în urmă cu 24 de ani, când fiecare "genera" doar două pagini şi jumătate.

•studiu se arata ca in 2008 sau consumat pana la 3.6 zettabytes sau 10,845 trillion de cuvinte , respectiv 34 gigabytes de persoana pe an

•DACA se stocheza datele digitale existente pana la sfrrsitul anului 2010 pe DVD se poate forma o stiva care sa acopere distanta de la luna si inapoi

Explozia Informationala

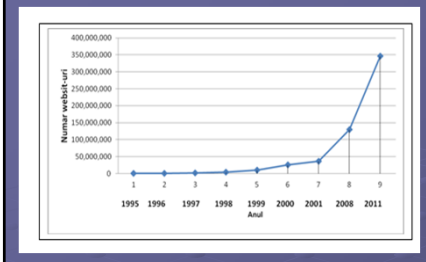
incepand cu 2010

Cine are cele mai multe servere Web?

OVH : 100.000 servere (firma , iulie, 2011)
SoftLayer : 100.000 servere (firma, decembrie 2011)
Akamai Technologies : 95,000 servere (firma, decembrie 2011)
Rackspace: 78717 de servere (companie 30 septembrie 2011)
Intel: 75,000 servere (firma , august, 2011)
1 & 1 Internet : 70000 servere (companie , februarie 2010)
Facebook: 60.000 servere (estimare, octombrie 2009)
LeaseWeb: 36,000 servere (firma, februarie 2011)
Intergenia: (PlusServer/Server4You), 30.000 de servere (companie , 2011)
SBC Communications: 29,193 servere (Netcraft)
Verizon : 25,788 servere (Netcraft)
Time Warner Cable : 24,817 servere (Netcraft)
HostEurope: 24.000 servere (Compania)
AT & T : 20,268 servere (Netcraft)

In lume exista aproximativ 44 milioane de severe

Este posibil ca
Google să dețină aproape un milion de servere.
Yahoo are aproximativ 700 000 cu 13 000 de angajați.
Wikipedia are 679 de servere și 95 de angajați



Explozia Informationala

incepand cu 2010

Cata informatie exista in spatiul Web?

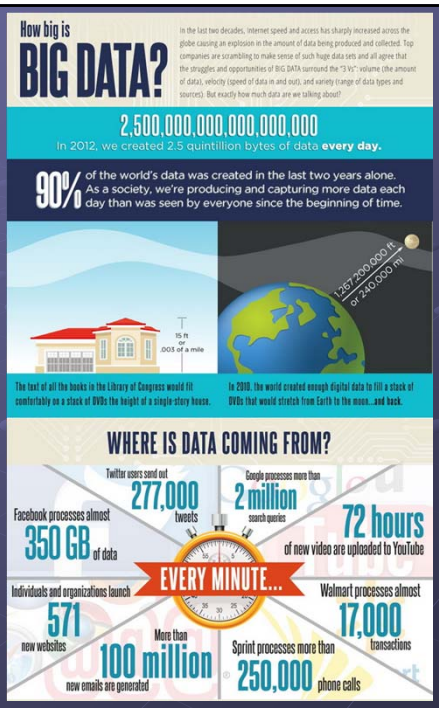
Spatiul Web este format din doua componente:
-"*Suprafata*" Web formata din situri publice cunoscute ca Web
-Spatiul Web de "*adancime*" format din situri specializate mai mare de 400 – 500 ori decat "suprafata"

Spatiul Web 2011 de tip "suprafata"
- ✓ **Suprafata Web a variat in decursului anului intre 25 pana la 50 terabytes**
- ✓ **existau la inceputul anului 2.5 miliarde documente**
- ✓ **In fiecare zi se adauga 7,300 000 noi pagini, ceea ce insemna 0.1 terabyte noi pe zi**

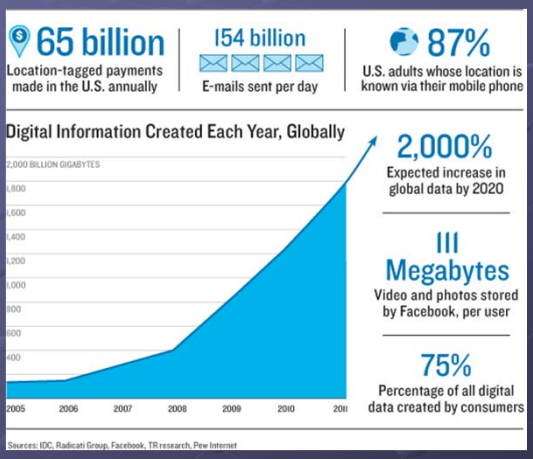
Spatiul Web 2011 de tip "adancime"
- ✓ **Adancimea Web are 7,500 terabytes de date**
- ✓ **Aproximativ 4,200 terabytes sunt date stiintifice**
- ✓ **Exista 550 miliarde de documente interconectate, 95% din aceasta informatie este accesibila publicului**

Email & Mailing Lists
- ✓ **Au fost trimise intre 900 – 1100 miliarde de email-uri in acest an**
- ✓ **O persoana primeste in medie 40 email-uri pe zi din care arhiveaza aproximativ 17 email-uri**
- ✓ **Cantitatea de informatii aferenta email-urilor trimise se ridica la gigantica cantitate de 11,285 pana 20,350 terabytes.**

Explozia Informationala
2010 / date digitale

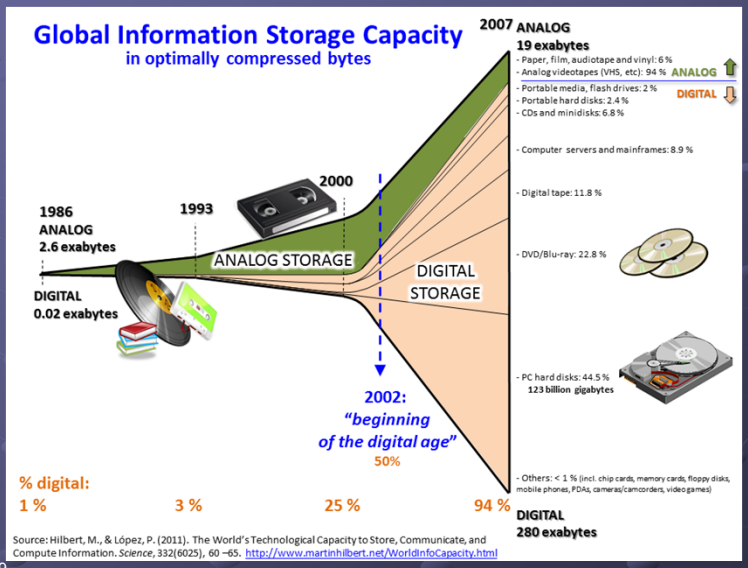


Explozia Informationala
2010 / date digitale



Explozia Informationala

2010 / date digitale



Explozia Informationala

2010 / BIG DATA

BIG DATA se referă la Datele păstrate și prelucrate în cantități imense, datorită unor medii de stocare mai ieftine, unor metode de procesare mai rapide și unor algoritmi mai performanți"

Explozia Informationala

2010 / BIG DATA

BIG DATA are 4 caracteristici principale:

1. **Prima caracteristică este VOLUMUL.**
   Volumul de date este în creștere exponențială. Experții prezic că volumul de date din lume, va crește la 35 de Zettabytes în 2020. Numărul de surse de date este de asemenea în creștere.

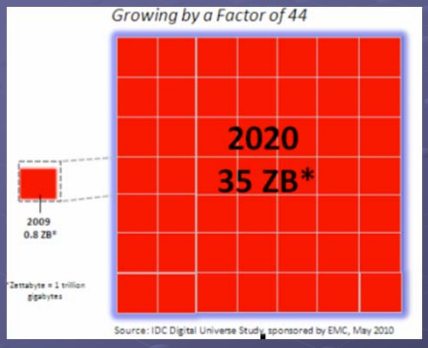
2. **A doua caracteristică este VITEZA.**
   Datele se creează la viteze din ce în ce mai mari.

3. **A treia caracteristică este VARIETATEA datelor.**
   Creșterea surselor de date a alimentat și creșterea tipurilor de date. De fapt, 80% din datele generate în lume sunt date nestructurate.

4. **A patra caracteristică este VERIDICITATEA datelor.**
   Datele pot veni de la sisteme tradiționale - sisteme de facturare, sisteme ERP (Enterprise Resource Planning) , sisteme CRM (Customer Relationship Managament). De asemenea, vin de la oameni - site-ul web, social media, etc. Acest lucru face foarte dificilă analiza datelor sociale - extragerea ideilor de conținut în mare parte sub formă de text într-un timp foarte scurt.

Explozia Informationala

perspective



- cantitatea de informatie digitala produsa a fost de :
  0.8 zettabytes in 2009
  5 zettabytes in 2013
  daca cresterea se mentine in 2020 se vor produce 35 ZB

In loc de concluzii

Intreaga istorie a omenirii din punct de vedere a cantitatii totale de informatie produsa pana in anul 1999 reprezinta aproximativ a miliarda parte din informatia generata in anul 2010. Exista cateva aspecte care merita sa fie relevate.

Pana in 1999	Din 2000
Informatia prezenta permite extragerea de cunostinte utile si consistente	Informatia nu mai permite extragerea de informatii utile. Sunt necesare unelte specializate de extragere a acesteia (vezi masinilede cautare de tip Google etc). In prezent cunostintele extrase au un grad scazut de credibilitate.
Favorizeaza insusirea si intelegerea aproape in totalitate a ceea ce insemna cunostinte specifice unui domeniu sau meserii.	Favorizeaza superficialitatea datorita imposibilitatii de a discerne ce este esential sau nu in procesul de filtrare a informatiilor.
Este favorizata aparitia unor personalitati enciclopedice cu o viziune de ansamblu asupra dezvoltarii societatii umane	Apar specializari extrem de inguste . Apare fenomenul de tip semidoctism
Se facea raportarea la o traditie intr-un domeniu	Nu mai exista traditie

Cantitatea mare de informatie generata in prezent nu mai favorizeaza procesul de cunoastere

Explozia Datelor (cont.)

• Foarte puţine date pot fi analizate si integrate de operatorul uman.
• Datele se colectează uşor, analiza lor este costisitoare.
• Există suspiciunea că in masivele de date pot exista cunoştinţe ascunse.

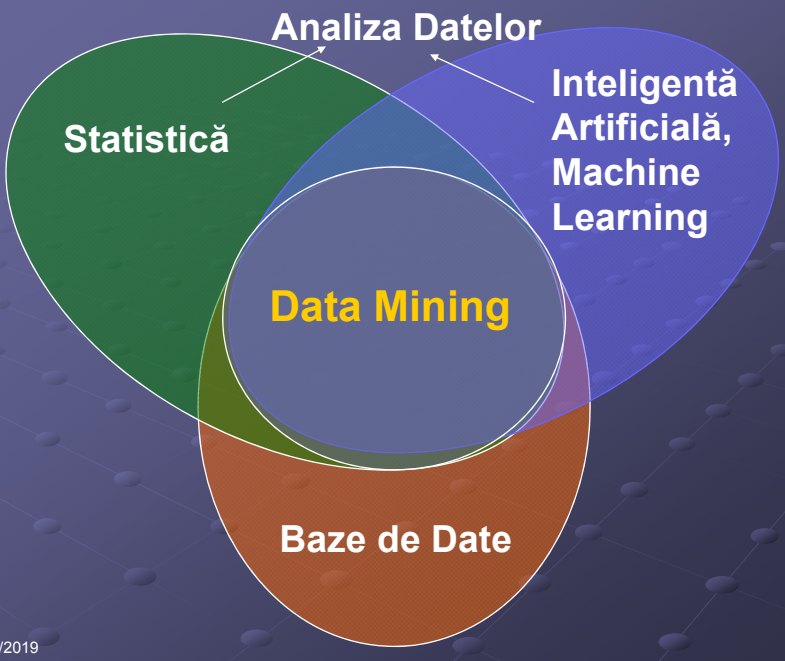
• Descoperirea Cunoştinţelor este **NECESARA** pentru a da sens utilizării datelor.

Din acest motiv mulţi cercetători au considerat extragerea cunoştinţelor din baze de date ca un domeniu semnificativ de investigat

Ce este Data Mining?

"Procesul de analiza a unor cantităţi mari de date în scopul determinării de relaţii care apar intre elementele prezente in bazele de date si a determinării de machete (potenţial utile) care pot caracteriza global bazele de date."

(din Advances in Knowledge Discovery and Data Mining, Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

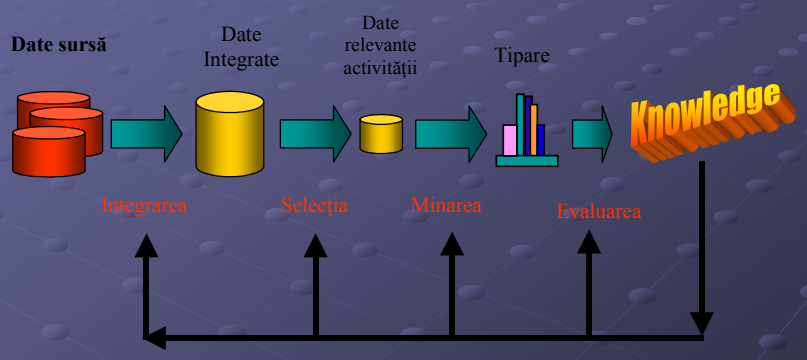


Definirea procesului de descoperire a cunoştinţelor

Procesul de descoperire de informaţii din baze de date mari cuprinde mai multe etape

1. ***definirea scopului urmărit***
2. ***interogarea surselor de date*** *si definirea structurii datelor supuse prelucrării,*
3. ***preprocesarea datelor*** *(selectarea, curăţarea, transformarea acestora),*
4. ***minarea datelor*** *pentru extragerea de tipare şi de modele apropiate,*
5. ***evaluarea şi interpretarea tiparelor*** *extrase pentru a decide ce constituie "cunoştinţă" (knowledge),*
6. ***consolidarea cunoştinţelor*** *şi rezolvarea conflictelor dintre cunoştinţele extrase anterior, oferirea cunoştinţelor spre utilizare.*

Procesul de descoperire de cunoştinţe (etape)



Data Mining: Tehnici

- Clasificare
- Corelatii
- Grupare
- Asociatii

Data Mining: Tehnici

Clasificare

- Linear Discriminant Analysis
- Naïve Bayes / Bayesian Network
- 1R
- Neural Networks
- Decision Tree (ID3, C4.5, …)
- K-Nearest Neighbors
- Support Vector Machines
- …

Corelare

- Multiple Linear Regression
- Principal Components Regression
- Partial Least Square
- Neural Networks
- Regression Tree (CART, MARS, …)
- K-Nearest Neighbors
- Support Vector Machines
- …

Grupare

- K-Mean Clustering
- Self Organizing Map
- Bayesian Clustering
- …

Asociere

- A Priori
- Markov Chain
- Hidden Markov Models
- …

Etape de construire a unui model în Data Mining

1. Definirea problemei
2. Construirea bazei de date de tip data mining
3. Explorarea datelor
4. Pregatirea datelor pentru modelare
5. Construirea modelului
6. Evaluarea modelului
7. Utilizarea modelului

Definirea domeniului Data Mining

- Explozia datelor
- Introducere in data mining
- Exemple de data mining in ştiinţă şi inginerie
- Provocări si oportunităţi

Exemple de data mining in inginerie

1. Data mining in inginerie Biomedicala

   "Controlul unui brat robotic utilizand Tehnici Data Mining"

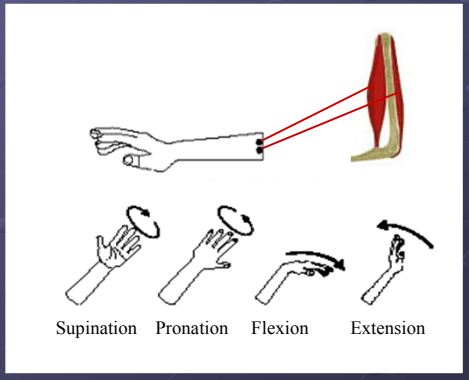
2. Data mining in inginerie Chimica

   "Data Mining pentru Monitorizarea imagini din procesul de extrudere mase plastice"

1. Definirea problemei

"Controlul unui brat robotic prin intermediul semnalelor EMG culese de pe muschii biceps si triceps."

Contractia muschiulara	Biceps	Triceps
Supination	H	H
Pronation	L	L
Flexion	H	L
Extension	L	H

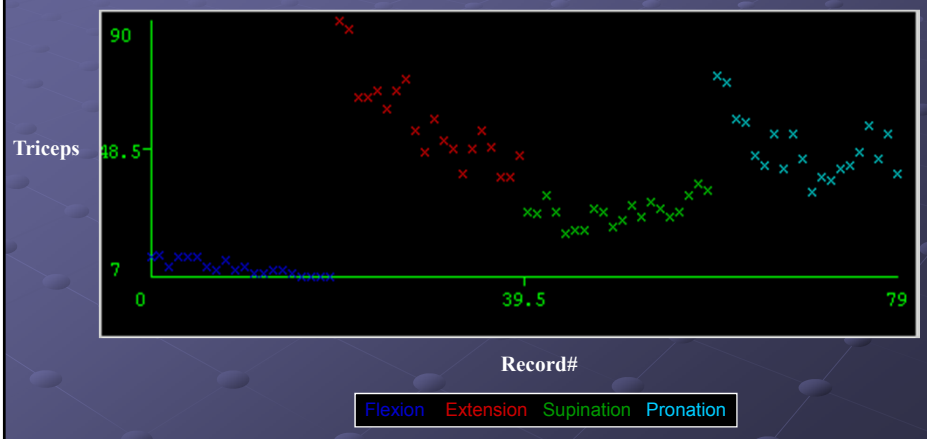


2. Construirea bazei de date de tip data mining

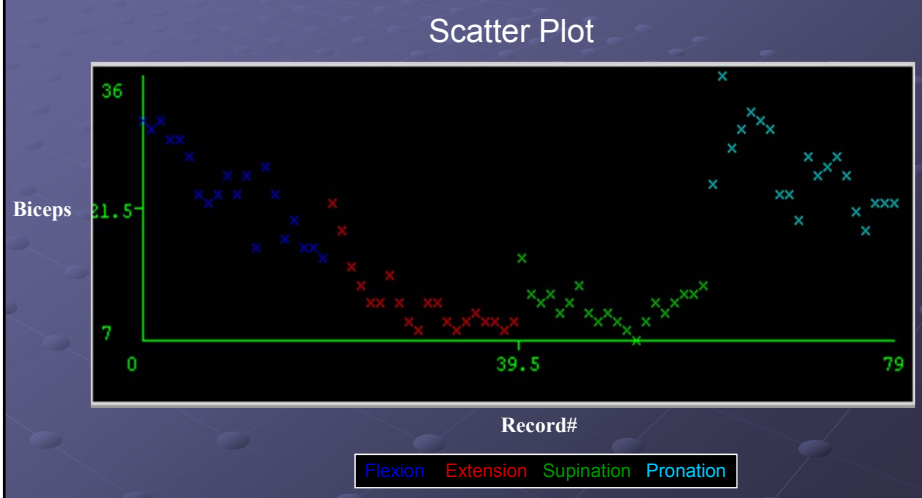
- Setul de date are un numar de 80 înregistrari.
- Există două variabile de intrare: semnalul de la biceps si semnalul de la triceps.
- Există o variabilă de ieşire cu patru posibile valori: supination, pronation, flexion si extension.

3. Explorarea datelor

Scatter Plot



3. Explorarea datelor(cont.)



4. Pregatirea datelor pentru modelare

- Translatarea setului de date in format ARFF:

```
@relation EMG

@attribute Triceps real
@attribute Biceps real
@attribute Move {Flexion,Extension,Pronation,Supination}

@data
13,31,Flexion
14,30,Flexion
10,31,Flexion
13,29,Flexion
.....
```

5. Construirea modelului

- **Clasificare**
  - 1R
  - Decision Tree
  - Naïve Bayesian
  - K-Nearest Neighbors
  - Neural Networks
  - Linear Discriminant Analysis
  - Support Vector Machines
  - ...

6. Evaluarea modelului

- Validarea modelului utilizand setul de testare

Rezultate validare

1R	76%
Decision Tree	90%
Naïve Bayesian	98%
1-Nearest Neighbors	100%
Neural Networks	100%

7. Utilizarea modelului

❖S-a implementat modelul de tip reţea neuronală intr-un brat robotic.

Exemple de data mining in inginerie

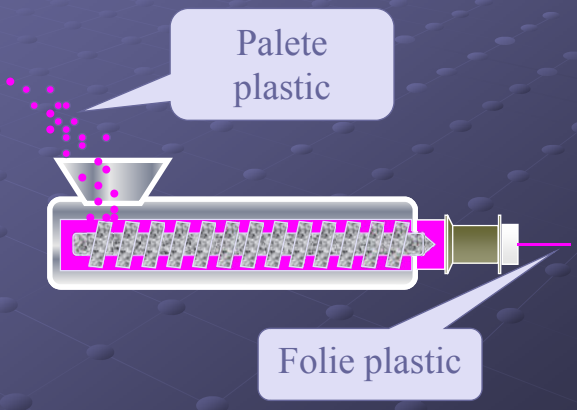
1. Data mining in inginerie Biomedicala

   "*Controlul unui braţ robotic utilizând Tehnici Data Mining*"

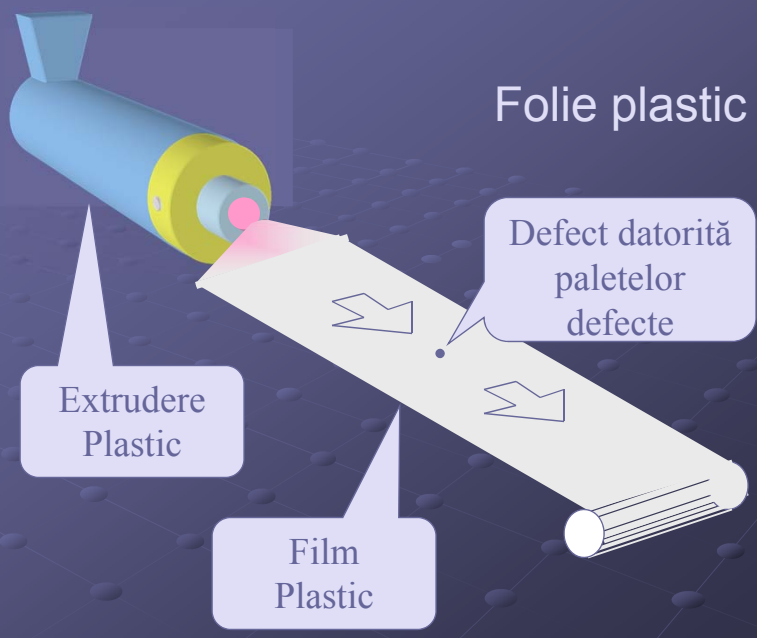
2. Data mining in inginerie Chimică

   "*Data Mining pentru Monitorizarea imagini din procesul de extrudere mase plastice*" K.Torabi, L D. Ing, S. Sayad, and S.T. Balke

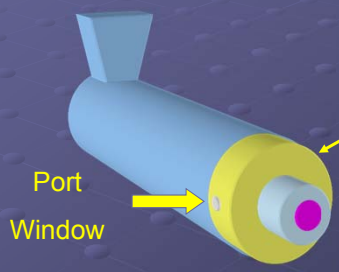
Extrudere masă plastică



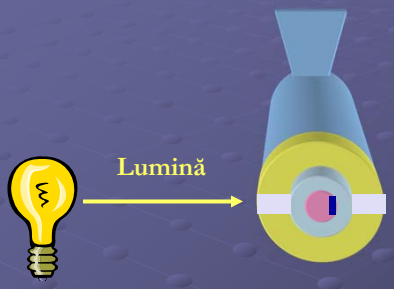
Folie plastic



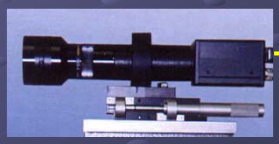
Monitorizare In-Line



Monitorizare In-Line



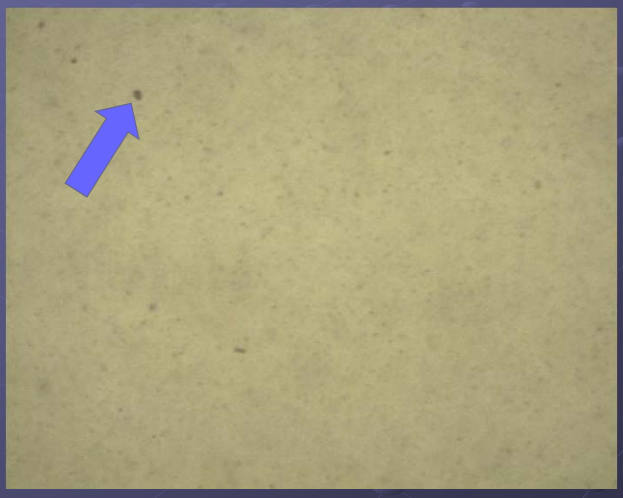
Ansamblu Optic



Imagine
Calculator

Film plastic fara defecte (FD)
- fara particule contaminante -

Film plastic cu defecte (CD)
- fara particule contaminante -



1. Definirea problemei

Se clasifica imaginile in doua clase corespunzatoare cazurilor film fara defecte (FD) si film cu defecte (CD).

FD

CD

2. Construirea bazei de date de tip data mining

- 2000 Imagini
- 54 variabile toate numerice
- O variabila de iesire cu doua posibile valori
  - cu defecte (cu particule CD) si
  - fără defecte (fara particule FD)

3. Explorarea datelor

Etapa nu este necesara

4. Pregatirea datelor pentru modelare

- Prelucrarea imaginilor pentru eliminarea zgomotelor
- Set 1 de date cu imagini curate: 1350 imagini care includ 1257 fara particule si 91 cu particule
- Set 2 de date cu imagini curate si cu zgomot : 2000 care includ 1909 fără particule si imagini cu zgomot si 91 cu particule
- 54 Variabile de intrare toate numerice
- O variabilă de ieşire, cu două valori posibile (CD si FD)

5. Construirea modelului

Clasificare:

- 1R
- Decision Tree
- 3-Nearest Neighbors
- Naïve Bayesian

6. Evaluarea modelului

Rezultate validare

Set Date	*Atrib.*	*Clase*	*1R*	*C4.5*	*3.N.N*	*Bayes*
Imagini curate	54	2	99.9	99.8	99.8	95.8
Imagini curate + zgomot	54	2	98.5	97.8	97.8	93.3
Imagini curate + zgomot	54	3	87	87	84	79

If densitatea de pixeli Max < 142 then CD

7. Utilizarea modelului

❖ Un program in Visual Basic s-a utilizat pentru implementarea modelului.

Exemple de data mining in ştiinţă

1. Data mining in Astronomie
   1. "*Detectarea de noi obiecte astronomice*"
   2. *"Clasificarea galaxiilor"*

2. Data mining in Relatii Internationale
   Sistem de cautare a relatiilor intre evenimente
3. Data mining in Meteorologie
   Detectarea cicloanelor tropicale:
   Estimarea vitezei maxime a vantului

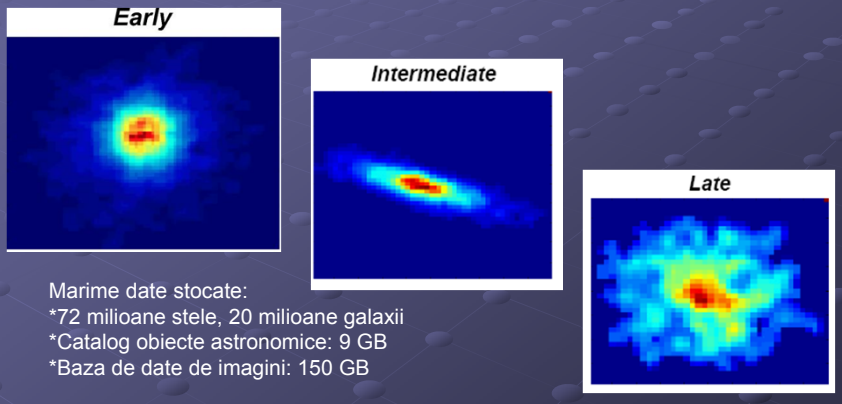
Detectarea de noi obiecte astronomice

- **Scop:** Definirea tipului de obiect astronomic (stea sau galaxie), prezent in imaginile achizitionate de la Observatorul astronomic Palomar
  –3000 imagini cu 23,040 x 23,040 pixels / imagine.
- **Mod de abordare:**
  – Segmentarea imaginii
  – Crearea unui numar de 40 caracteristici (atribute)
  – Construirea unui model de grupare
- **Rezultat**: Gasirea unui numar de 16 quasari!

Clasificarea galaxiilor

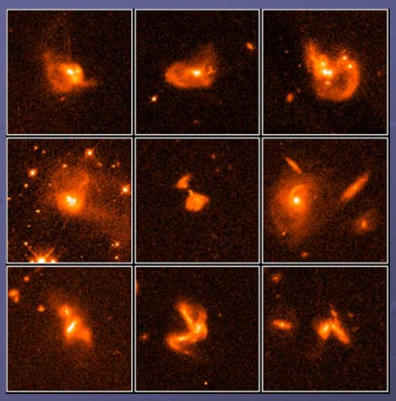
Clasa: Etapa de formare

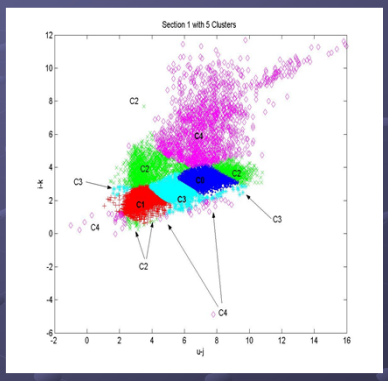
Atribute: Caracteristici imagine, Caracteristici lungime de unda primita, etc.



Marime date stocate:
*72 milioane stele, 20 milioane galaxii
*Catalog obiecte astronomice: 9 GB
*Baza de date de imagini: 150 GB

Clasificarea galaxiilor

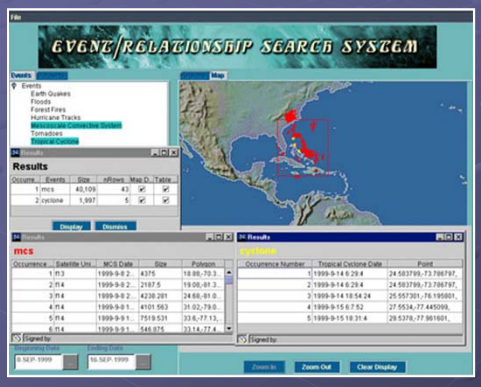




Galaxii care se formeaza :
-Prin fuziune
-Prin splitare

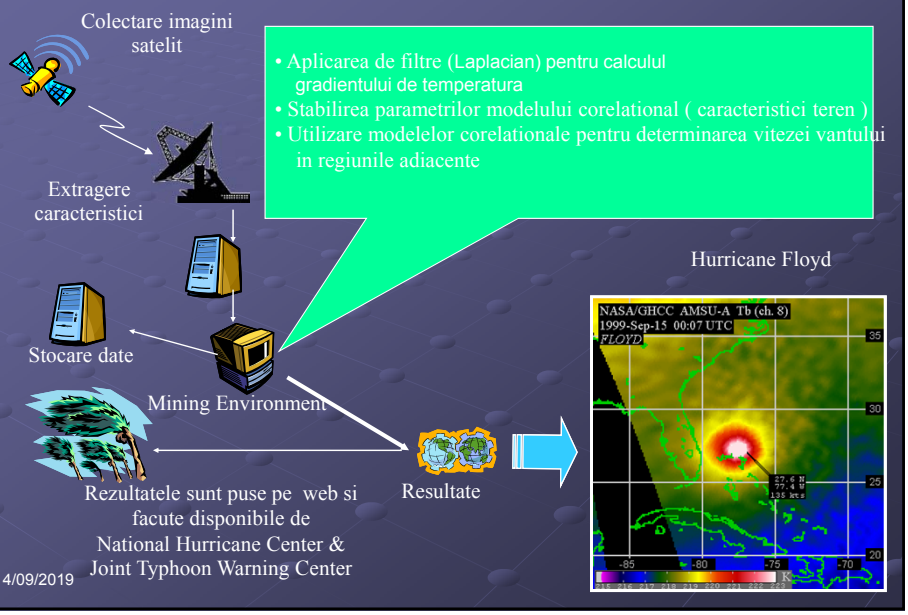
Utikizarea tehnicilor de Grupare si Clasificare pentru a le distige de o galaxie normala

Sistem de cautare a relatiilor intre evenimente



- Permite utilizatorului sa gaseasca corelatii intre evenimente. In ce masura un eveniment este cauza sau efect a unui alt eveniment
- Atributele cuprind informatii geografice, politice, configurationale care se intind pe perioade determinate de timp

Detectarea cicloanelor tropicale: Estimarea vitezei maxime a vantului



Definirea domeniului Data Mining

- Explozia datelor
- Introducere în data mining
- Exemple de data mining in stiinta si inginerie
- Provocari si oportunitati

Provocări şi oportunităţi

- Data mining este în topul primelor 10 tehnologii care sunt dezvoltate in prezent

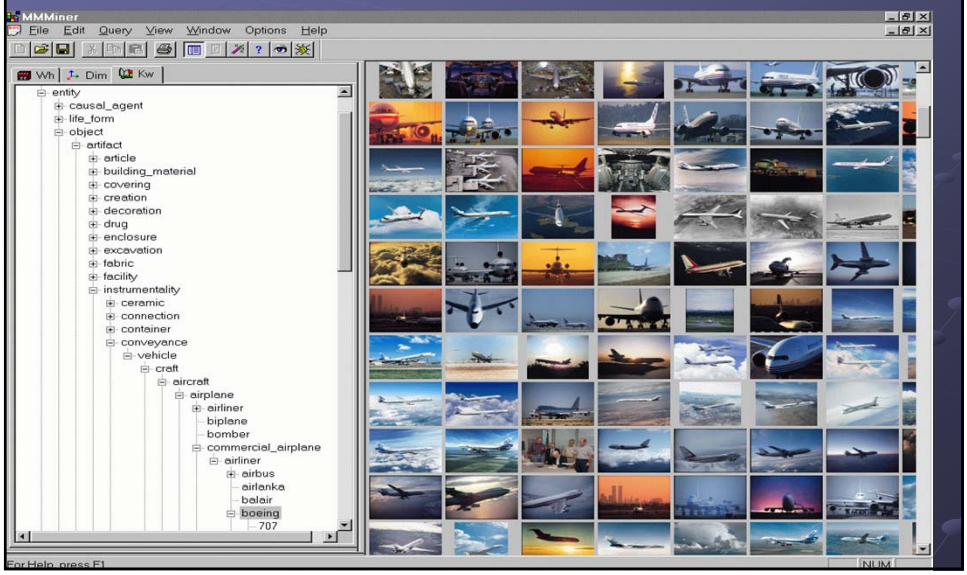
(Google a fost creat de Sergey Brin si Larry Page
in perioada cand erau studenti la Stanford in urma cercetarilor acestora in baze de date si data mining din 1998)

- Aflat la granita dintre 3 domenii, prezintă o mare diversitate de tehnici si algoritmi care inglobează concepte ce asigură o flexibilitate care nu se întâlneşte in alte domenii tehnologice
- Include tehnici de prelucrare paralelă si distribuită

Data Mining Software

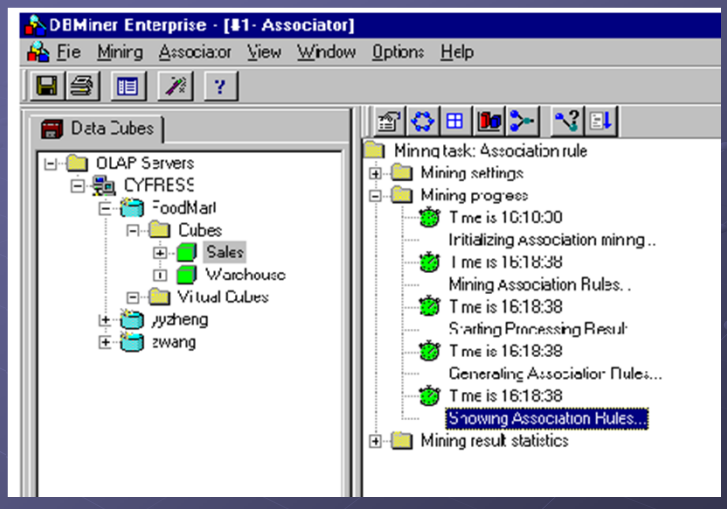
Data Mining Software (cont.)

Mining Multimedia Databases in **MultiMediaMiner**



Data Mining Software (cont.)

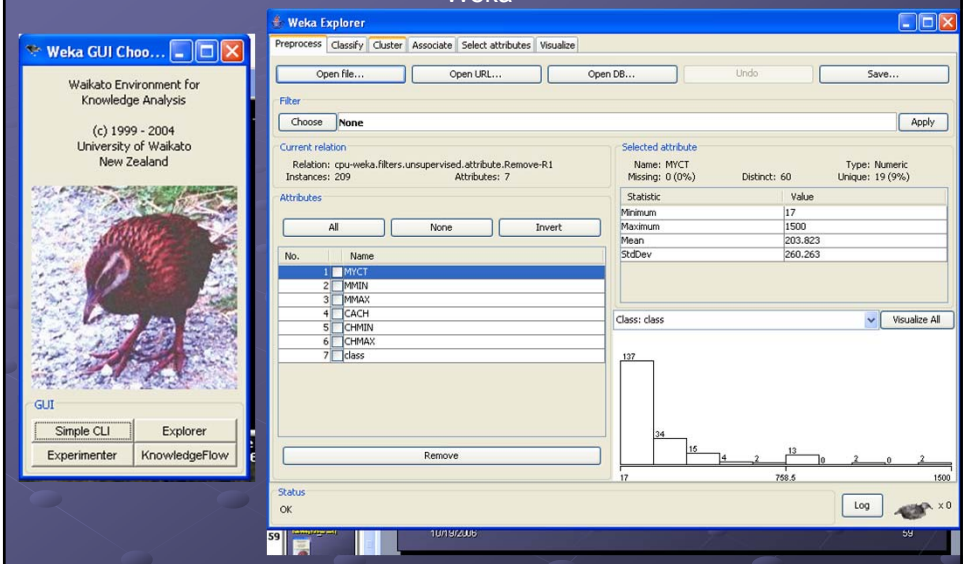
DBMiner Enterprise



Este destinat obtinerii de cunostinte din date din lumea afacerilor

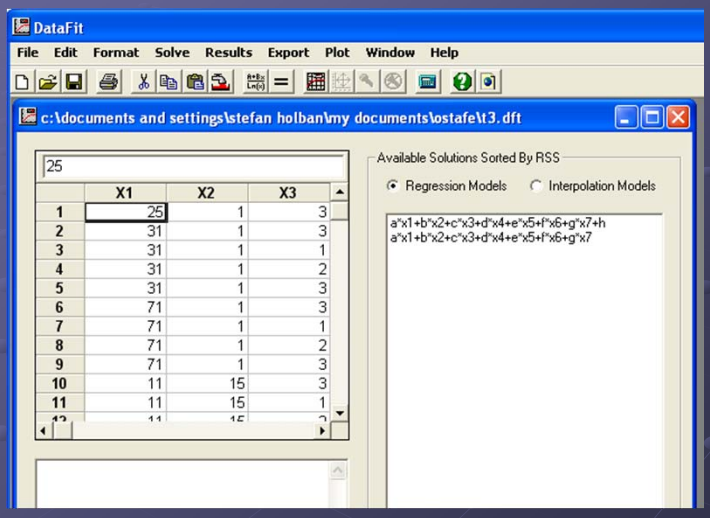
Data Mining Software (cont.)

Weka



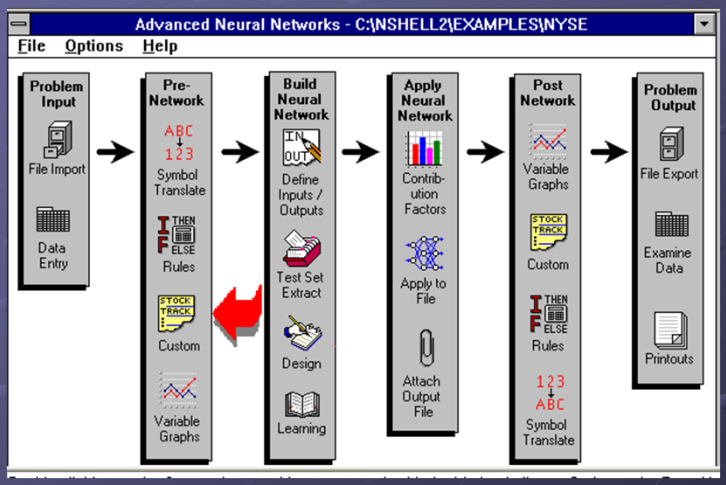
Data Mining Software (cont.)

DataFit



Data Mining Software (cont.)

NeuroShell



Data Mining Software (cont.)

- mining software cu licenta
  - SAS Enterprise Miner, SPSS Clementine, Statistica Data Miner, MS SQL Server, Polyanalyst, KnowledgeSTUDIO, …
  - lista adrese http://www.kdnuggets.com/software/suites.html
- mining software fara licenta
  - WEKA (Waikato Environment for Knowledge Analysis)
    - Free (GPLed) Java package with GUI
    - adresa www.cs.waikato.ac.nz/ml/weka
    - Witten and Frank, 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.*

Data mining reprezintă un domeniu vast şi interesant prin aceea ca are abilitatea de a rezolva un mare număr de probleme stiinţifice complexe.

MULŢUMESC!