

Data Mining

**Arta si Știința de a obține
Cunoștințe din Date**

Explozia Informatională

- paradox -

- ❖ Informația a devenit o marfă care se produce și se vinde.
- ❖ Problema care se ridică este că se produce mai mult decât se poate consuma.
- ❖ Acest aspect ridică întrebarea CE este important să reținem în cursul procesului de cunoaștere din informația aflată la dispoziție

-Definiii- -Informatia-

Într-o definire - pe cât de sumară tot pe atât de informală și, deci, de inexactă - se poate spune că **informația se constituie într-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).

Informația **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 instrucțiune, 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 Climatic 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.

Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

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 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
 $1,000,000,000,000,000,000,000 \text{ bytes} = 1000^7 = 10^{21}$
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 societății de tip occidental sunt supuși unui adevărat bombardament informațional: conform unui studiu american recent, fiecare primește, zilnic, o cantitate de informație echivalentă cu cea cuprinsă în 147 de ziare!

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

Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

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 informatii,
- se trimit 50 milioane SMS per zi
- Aproape 73 produse sunt comandate pe Amazon in fiecare secunda
- Zilnic, o persoană produce și transmite altora, în medie, informație 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 și 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 servere

Este posibil ca

Google să dețină aproape un milion de servere.

Yahoo are aproximativ 700 000 cu 13 000 de angajați.

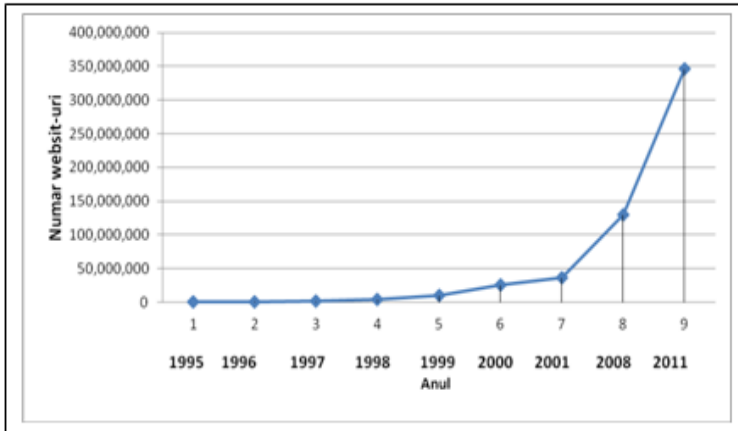
Wikipedia are 679 de servere și 95 de angajați

Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

Explozia Informatională

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 decursul 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.

Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

Explozia Informationala 2010 / date digitale

How big is **BIG DATA?**

In the last two decades, Internet speed and access has sharply increased across the globe causing an explosion in the amount of data being produced and collected. Top companies are scrambling to make sense of such huge data sets and all agree that the struggles and opportunities of BIG DATA surround the "3 Vs": volume (the amount of data), velocity (speed of data in and out), and variety (range of data types and sources). But exactly how much data are we talking about?

2,500,000,000,000,000,000

In 2012, we created 2.5 quintillion bytes of data **every day**.

90% of the world's data was created in the last two years alone. As a society, we're producing and capturing more data each day than was seen by everyone since the beginning of time.

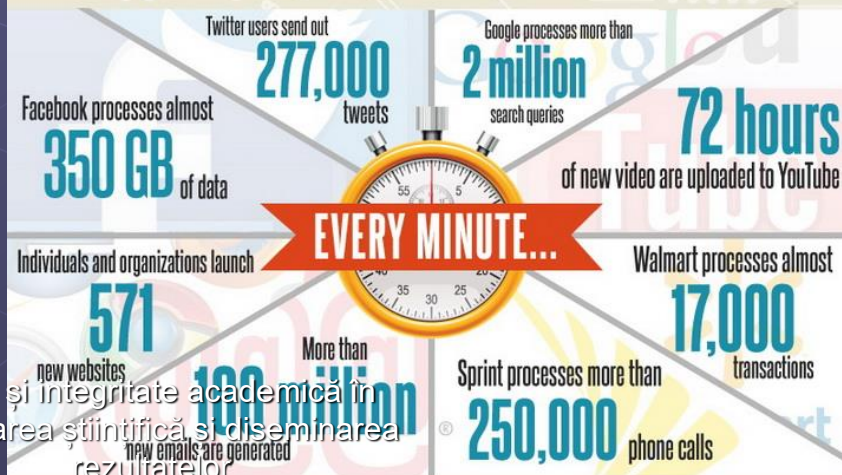


The text of all the books in the Library of Congress would fit comfortably on a stack of DVDs the height of a single-story house.



In 2010, the world created enough digital data to fill a stack of DVDs that would stretch from Earth to the moon...and back.

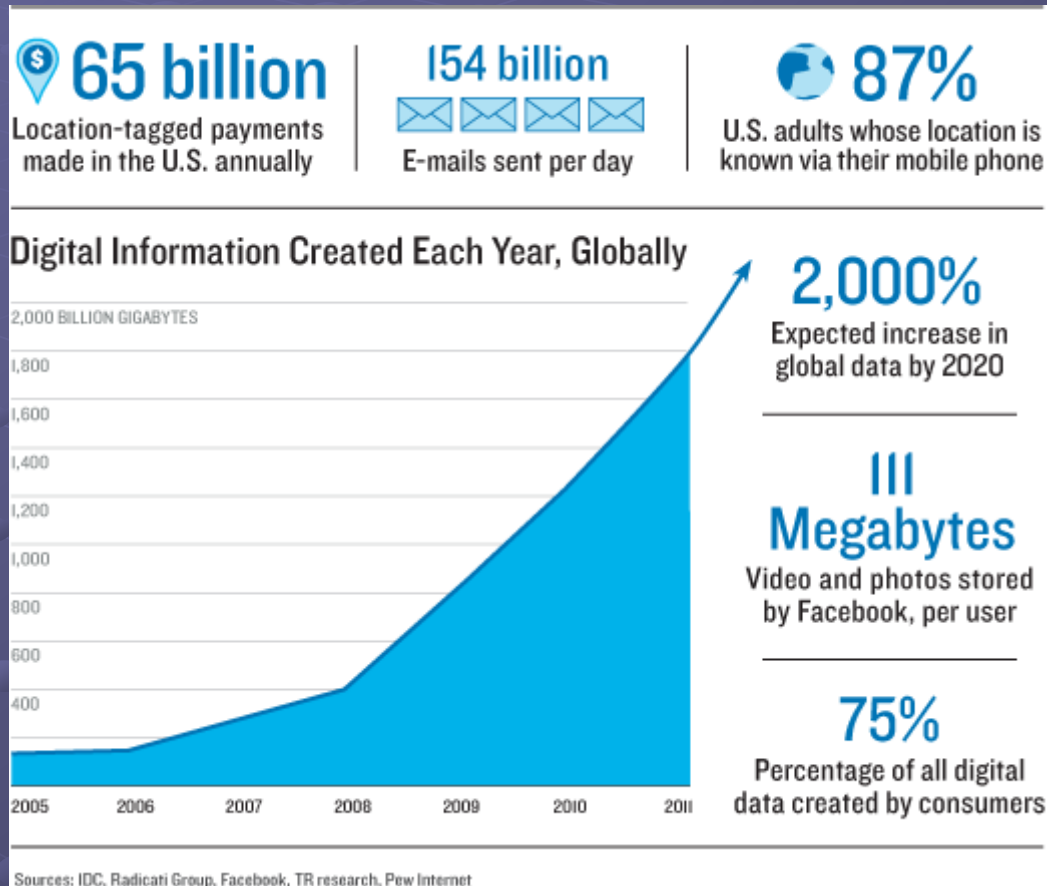
WHERE IS DATA COMING FROM?



Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

Explozia Informatională

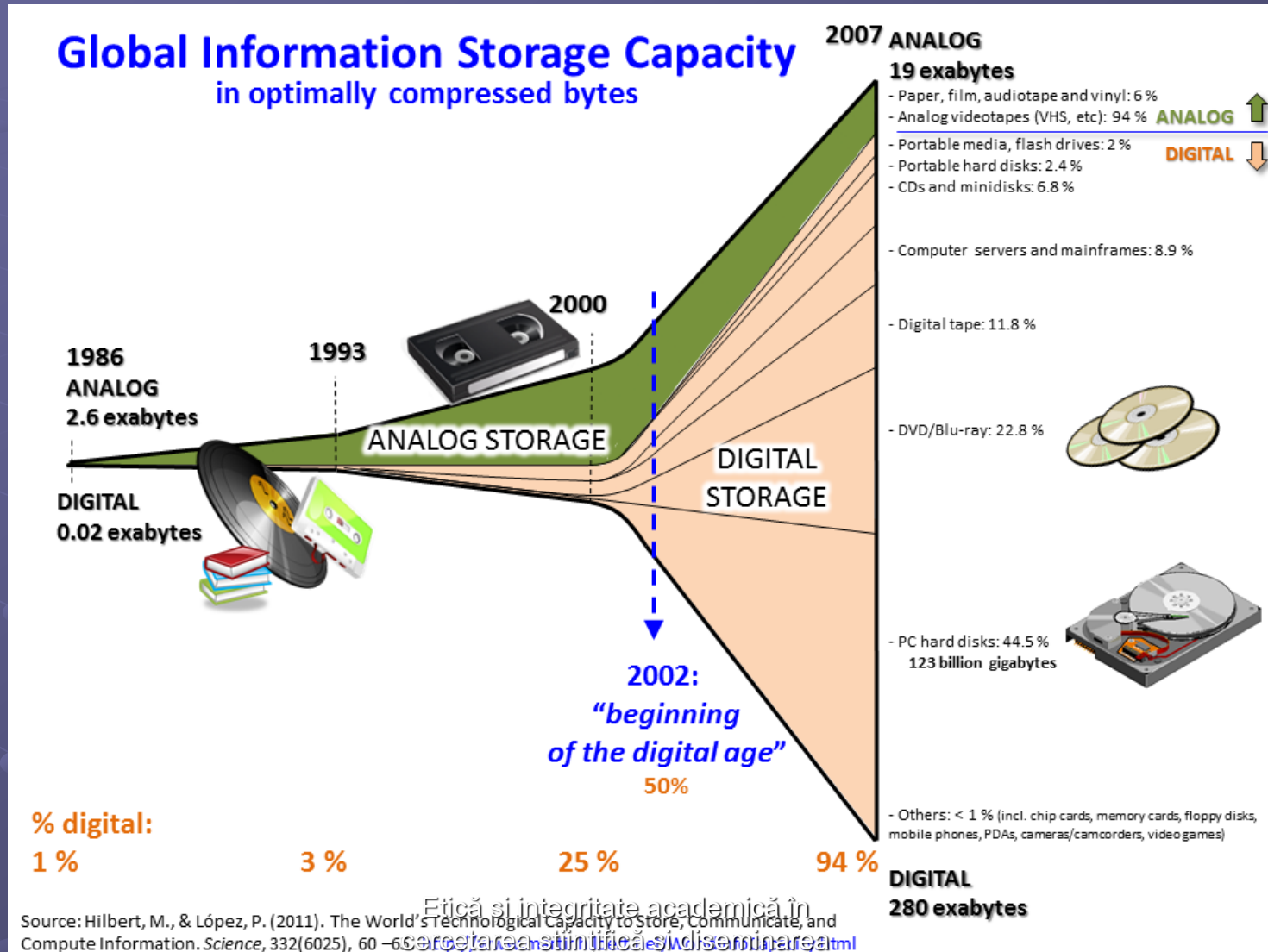
2010 / date digitale



Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

Explozia Informatională

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"

Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

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.

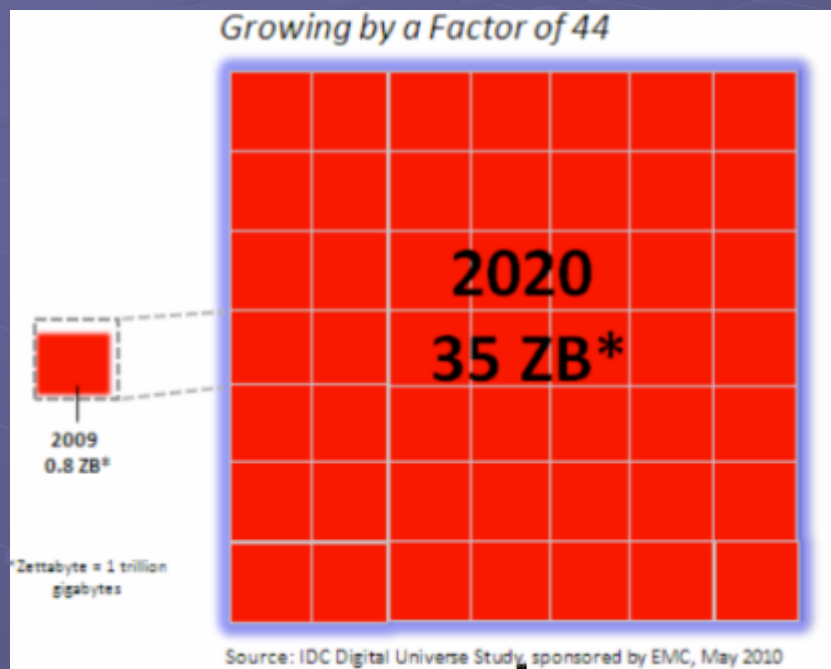
Creșterea surselor de date a alimentat și creșterea tipurilor de date. De fapt, 80% din datele generate în lume sunt date nestructurate.

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 Management). 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
 - in 2020 s-au produs 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 masinile de 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

Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

Explozia Datelor (cont.)

- Foarte puține date pot fi analizate și integrate de operatorul uman.
- Datele se colectează ușor, analiza lor este costisitoare.
- Există suspiciunea că în masivele de date pot exista cunoștințe ascunse.



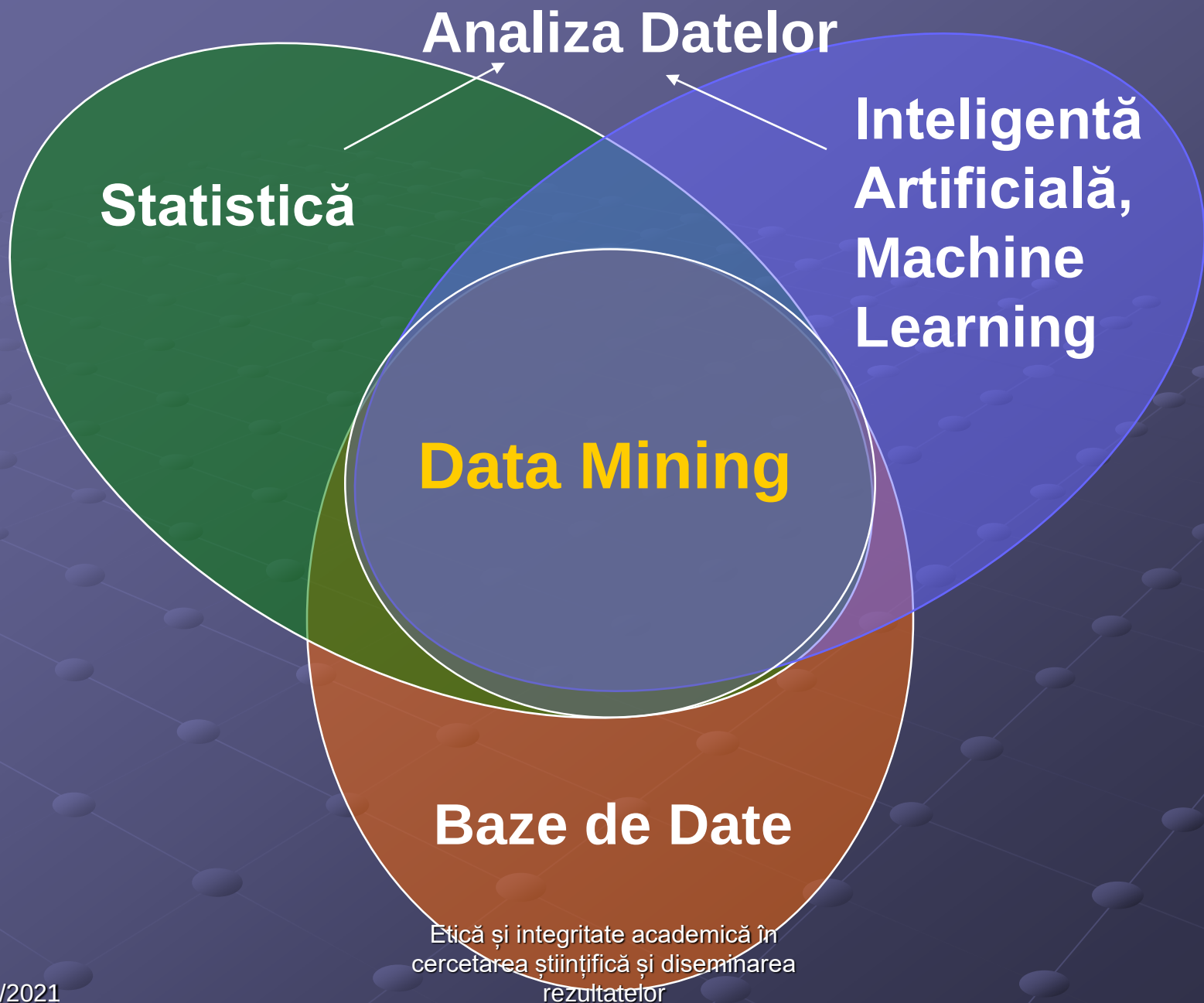
- **Descoperirea Cunoștințelor este NECESARA** pentru a da sens utilizării datelor.

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

Ce este Data Mining?

“Procesul de analiza a unor cantități mari de date în scopul determinării de relații care apar între elementele prezente în bazele de date și a determinării de machete (potențial 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)

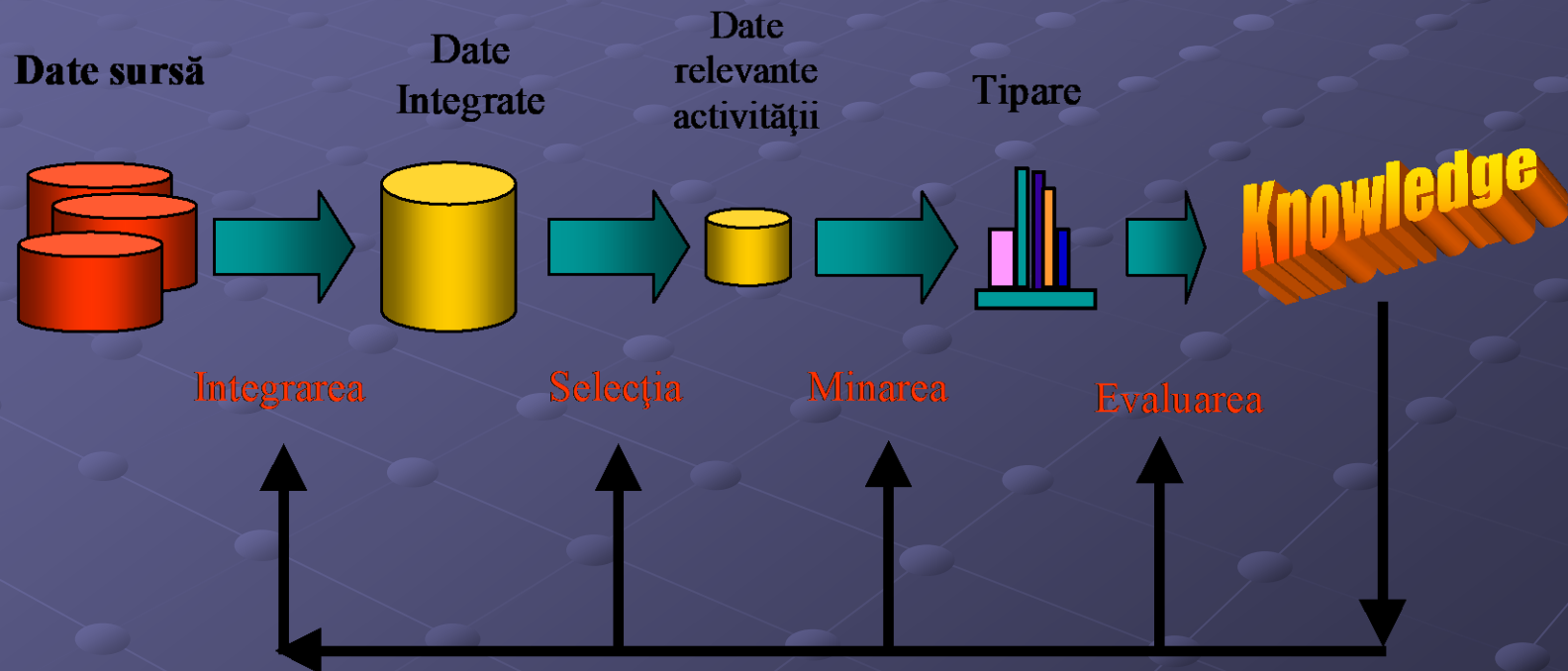


Definirea procesului de descoperire a cunoștințelor

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

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

● Procesul de descoperire de cunoștințe (etape)



Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

Data Mining: Tehnici

- ❖ Clasificare
- ❖ Corelatii
- ❖ Grupare
- ❖ Asociații

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
- ...

Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

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 știință și inginerie
- Provocări si oportunități

Definirea domeniului Data Mining

- Explozia datelor
- Introducere în data mining
- Provocari si oportunitati

Provocări și oportunități

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

(Google a fost creat de Sergey Brin și Larry Page în perioada când erau studenți la Stanford în urma cercetărilor acestora în baze de date și data mining din 1998)

- Aflat la granița dintre 3 domenii, prezintă o mare diversitate de tehnici și algoritmi care înglobează concepte ce asigură o flexibilitate care nu se întâlnește în alte domenii tehnologice
- Include tehnici de prelucrare paralelă și distribuită

Data Mining Software

Address <http://www.kdnuggets.com>

**Data Mining, Knowledge Discovery, Genomic Mining, Web Mining**
[Data Mining Consulting](#) | [Data Mining Jobs](#) | [Advertising](#) | [Site Map](#)

CLEMENTINE 7.0 = POWER, PREDICTION, PRODUCTIVITY
[SPSS Clementine 7.0 - The next generation of Data Mining](#)

[Free Webinar: Why Use Predictive Analytics?](#)

KDnuggets News, the Data Mining & Knowledge Discovery newsletter: data mining news, jobs, software, courses, ...
[2003 issues](#) | [Schedule](#) | [Archive](#) | [Submit](#) | [Subscribe!](#)

Current Issue: [NEW! 03:19, Oct 14, 2003: Data preparation; NSF deadline; ICDM-2003, Nov 19-22 ... \(29 items\)](#)

Match in: [help](#)

Software:
[Classification](#), [Suites](#), [Text](#)

Solutions:
[Bioinformatics](#), [CRM](#), [Web](#)

Companies:

Jobs:
[Industry](#), [Academic](#)

Courses: [Oct](#), [Nov](#), [Dec](#)
[Education](#)

Meetings:

Insightful Miner
Easy to Use & Extensible Data Mining

- Build predictive models easily
- Modern visual interface
- Advanced analytic methods
- Scalable capabilities

Free Webcast & Whitepaper!

[Insightful Miner](#)

Easy to Use & Extensible Data Mining

Poll

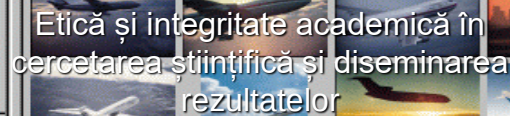
How frequently do you do a separate feature selection in classification (rather than have a learning algorithm do selection)

☐ Always
☐ Most of the time
☐ Frequently
☐ Rarely
☐ Never

[View Results](#)

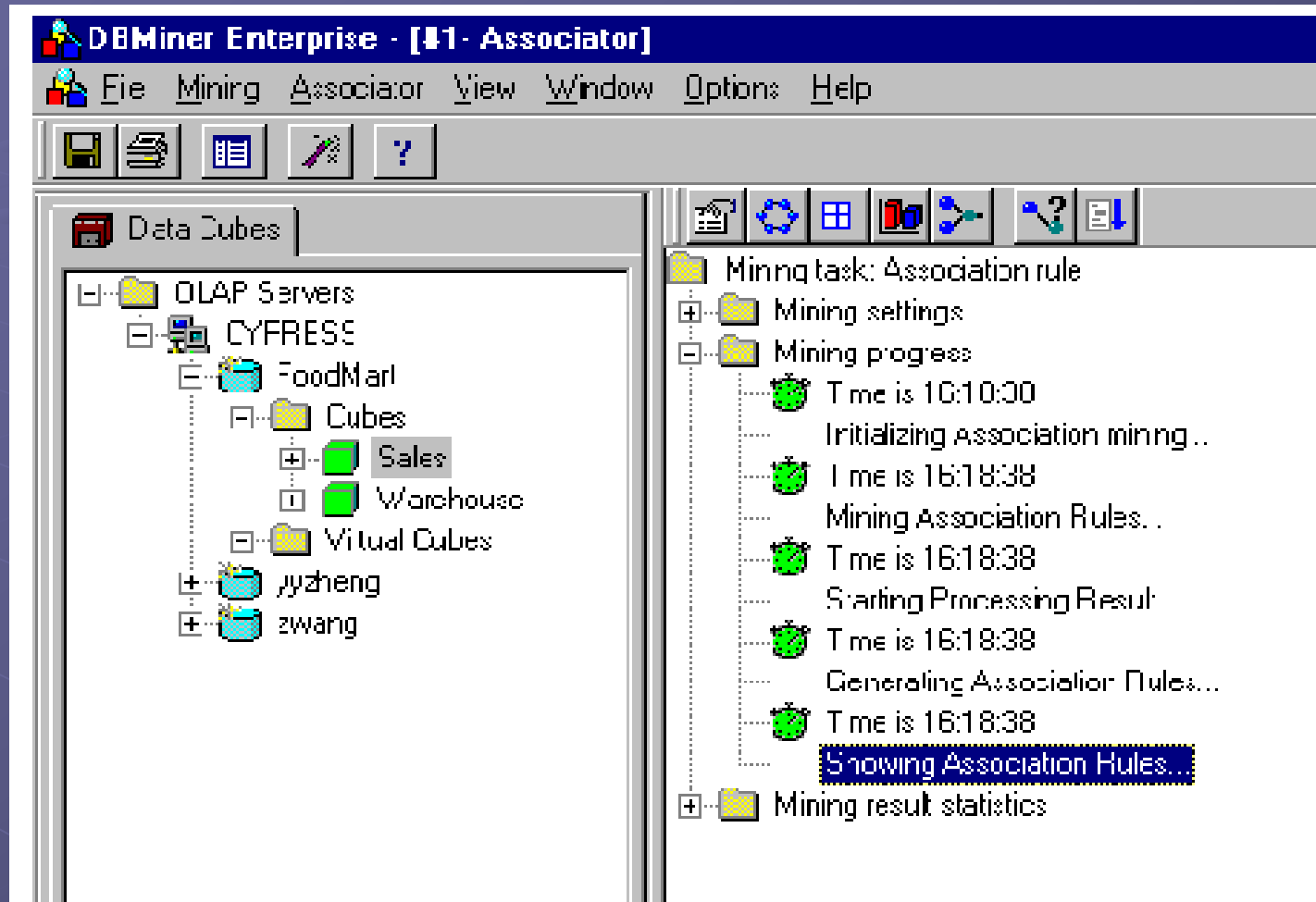
Mining Multimedia Databases in **MultiMediaMiner**

MultiMediaMiner



Data Mining Software (cont.)

DBMiner Enterprise

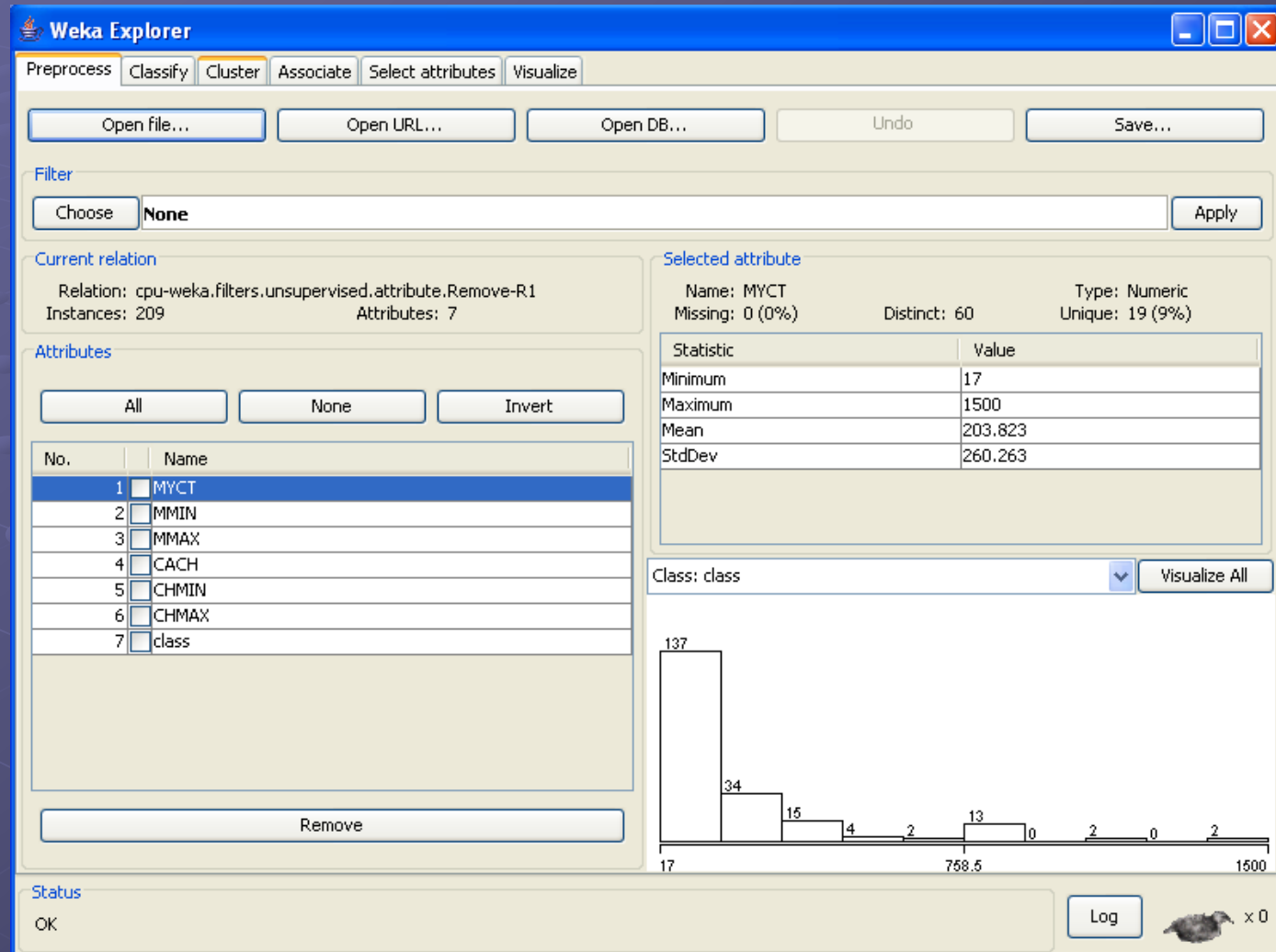
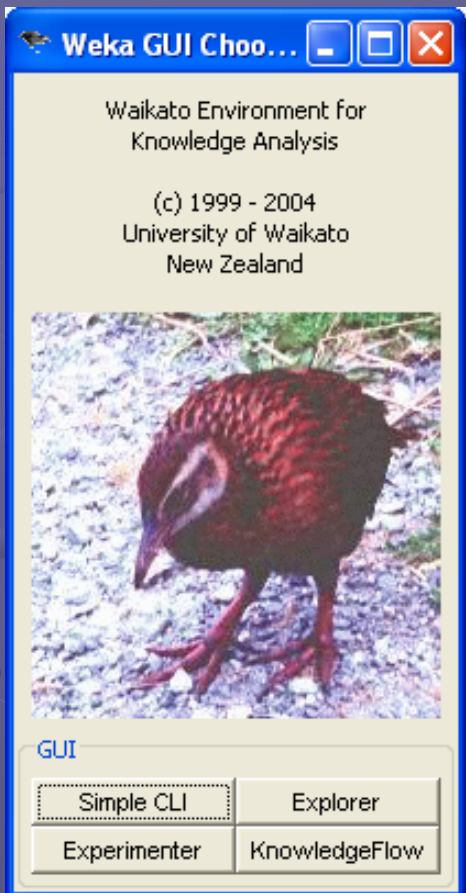


Este destinat obtinerii de cunostinte din date din lumea afacerilor

Etica și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

Data Mining Software (cont.)

Weka



The image shows the 'Weka Explorer' window. It has a title bar 'Weka Explorer' with standard window controls. The main content area has a light yellow background and contains the following sections:

- Preprocess** tab: 'Classify', 'Cluster', 'Associate', 'Select attributes', 'Visualize'.
- Buttons**: 'Open file...', 'Open URL...', 'Open DB...', 'Undo', 'Save...'.
- Filter**: 'Choose' button, 'None' dropdown, 'Apply' button.
- Current relation**: 'Relation: cpu-weka.filters.unsupervised.attribute.Remove-R1', 'Instances: 209', 'Attributes: 7'.
- Attributes**: 'All', 'None', 'Invert' buttons. A table with 7 rows: 'No.', 'Name'. The first row is selected: '1 MYCT'. The other rows are '2 MMIN', '3 MMAX', '4 CACH', '5 CHMIN', '6 CHMAX', '7 class'. A 'Remove' button is at the bottom.
- Selected attribute**: 'Name: MYCT', 'Missing: 0 (0%)', 'Distinct: 60', 'Type: Numeric', 'Unique: 19 (9%)'. A table with 2 columns: 'Statistic', 'Value'. The rows are: 'Minimum 17', 'Maximum 1500', 'Mean 203.823', 'StdDev 260.263'.
- Class: class**: A dropdown menu with a 'Visualize All' button.
- Visualize**: A histogram showing the distribution of the 'class' attribute. The x-axis ranges from 17 to 1500. The y-axis shows counts: 137, 34, 15, 4, 2, 13, 0, 2, 0, 2.
- Status**: 'OK' button, 'Log' button, and a small kiwi bird icon.

Data Mining Software (cont.)

DataFit

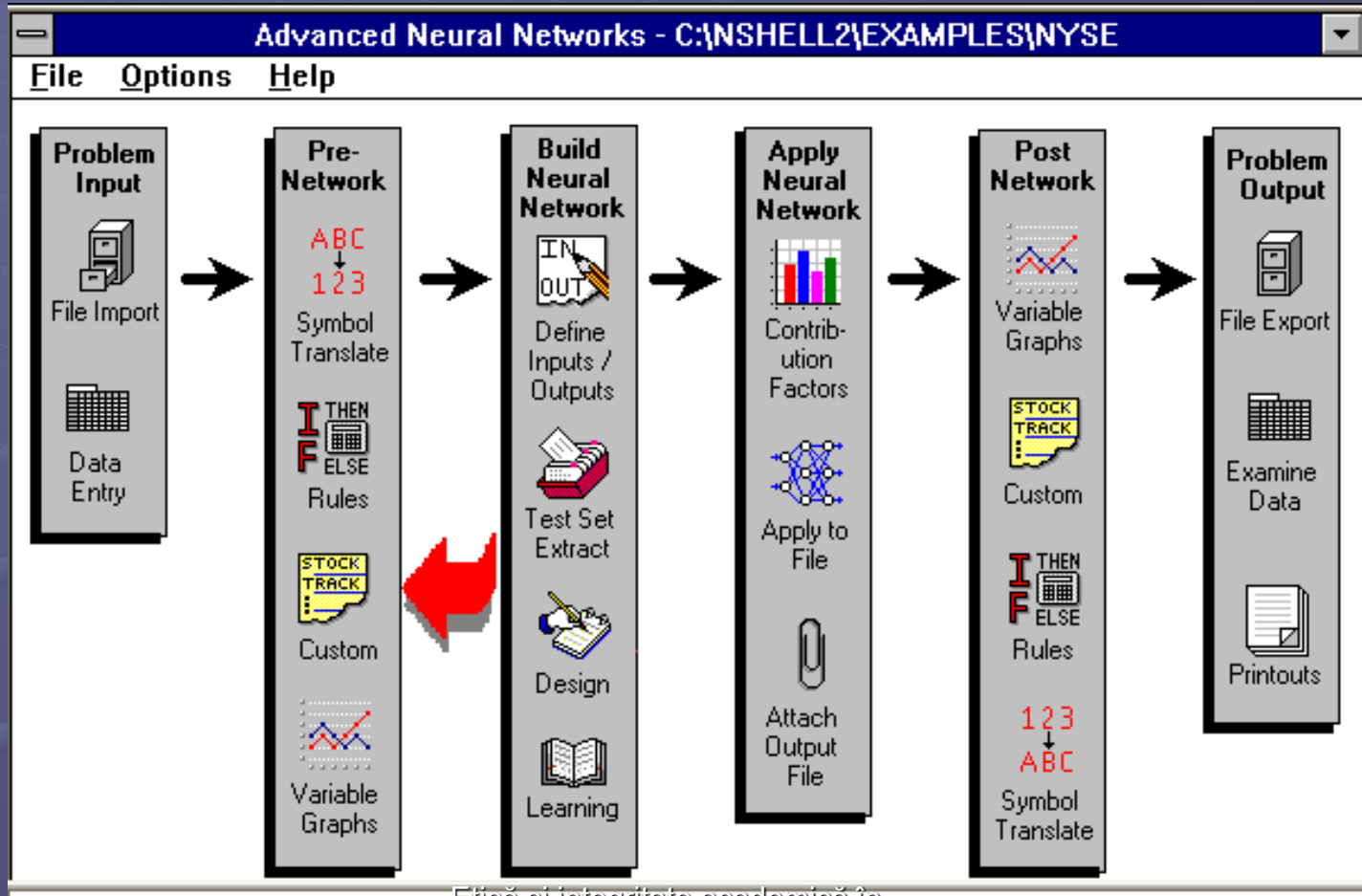
The screenshot shows the DataFit software window. The title bar reads "DataFit". The menu bar includes "File", "Edit", "Format", "Solve", "Results", "Export", "Plot", "Window", and "Help". The toolbar contains various icons for file operations, editing, and solving. The main window displays a data table with columns X1, X2, and X3. The first row has a value of 25 in the X1 column. The table lists 12 rows of data. To the right of the table, there is a section titled "Available Solutions Sorted By RSS" with two radio buttons: "Regression Models" (selected) and "Interpolation Models". Below these buttons, two regression models are listed: $a*x1+b*x2+c*x3+d*x4+e*x5+f*x6+g*x7+h$ and $a*x1+b*x2+c*x3+d*x4+e*x5+f*x6+g*x7$.

	X1	X2	X3
	25		
1	25	1	3
2	31	1	3
3	31	1	1
4	31	1	2
5	31	1	3
6	71	1	3
7	71	1	1
8	71	1	2
9	71	1	3
10	11	15	3
11	11	15	1
12	11	15	2

Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

Data Mining Software (cont.)

NeuroShell



Etică și integritate academică în
cercetarea științifică și diseminarea
rezultatelor

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
și interesant prin aceea că are abilitatea
de a rezolva un mare număr de
probleme științifice complexe.

MULȚUMESC!