



# Curs 1

Despre Data Mining si Data Science



# Despre mine



# Educație

- **Computational Biomedicine Laboratory**, University of Houston  
Bursa de cercetare post-doctorala  
Houston, Texas, SUA  
2005 – 2006
- **Department of Mathematics**, University of Iowa  
Doctorat Matematica, Media 4.0/4.0  
Iowa City, Iowa, SUA  
2005
- **Facultatea de Matematica**, Universitatea Transilvania  
Master Probabilitati, Statistica si Fiabilitatea Sistemelor, Media 9.8/10  
Brasov, Romania  
Mai 1998
- **Facultatea de Matematica**, Universitatea Transilvania  
Licentiat in Matematica si Informatica, Media 9.7/10  
Mai 1997

# Publicații

Bîldea, T. Ș., Gorin, T. "Towards capturing ancillary revenue via unbundling and cross-selling". *J Revenue Pricing Manag* **17**, 102–114 (2018)

Santamaria-Pang, A., Bîldea, T. Ș., Tan, S., & Kakadiaris, I. A. "Denoising for 3-d photon-limited imaging data using nonseparable filterbanks". *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, **17(12)**, 2312–2323 (2008).

Bîldea, T. Ș. "The Laplacian subalgebra of  $\mathcal{L}(\mathbb{F}_N)^{\overline{\otimes} k}$  is a strongly singular MASA." *Proceedings of the American Mathematical Society* **135**, no. 3 :823-31, (2007)

Santamaria-Pang, A., Bîldea, T. Ș., I. Konstantinidis and I. A. Kakadiaris, "Adaptive Frames-Based Denoising of Confocal Microscopy Data," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, pp. II-II. (2006)

Bîldea, T. Ș.. "Generalized amalgamated free product of C\*-algebras" - *J. Operator Theory* **56:1**, 143–165 (2006).

Bîldea, T. Ș. "On the Laplacian subalgebra of some tensors of group von Neumann algebras". *Contemp. Math.* **414**, 265–271(2006)

Bîldea, T. Ș., D. E. Dutkay and G. Picioroaga, "MRA super-wavelets", *New York J. Math.*, **11**, 1–19 (2005)



# Cursuri predate la Univ. Transilvania

- Tehnici și Procese de Data Mining 2016-2017, 2017-2018
- Știința Explorării și Explotării Datelor 2019-2020
- Aplicații ale Inteligenței Artificiale în Economie 2019-2020



Universitatea  
Transilvania  
din Brașov

# De la matematică la data science



Universitatea  
Transilvania  
din Brașov  
FACULTATEA DE MATEMATICĂ  
ȘI INFORMATICĂ



Licență, Matematică-Informatică-Analiză (1997)  
Master, Probabilități și Statistică (1998)



Doctorat, Analiză Funcțională (2005)



Computational Biomedicine Lab

Changing the Way People Look at Computers  
Computers People

Postdoctorat, Aplicații Wavelets (2006)



Optimizare de prețuri, cercetător



Științe actuariale- asurări,  
analist senior



Bioinformatică, cercetător senior



Cercetător principal, EMEA (2011)



Fondator, Data science software (2017)



Universitatea  
Transilvania  
din Brașov



Cadru didactic asociat, cursuri aplicative de data science și AI  
(2016-2020)

“Computers have promised us a fountain of wisdom but delivered a  
flood of data”

William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus

# DATA NEVER SLEEPS 6.0

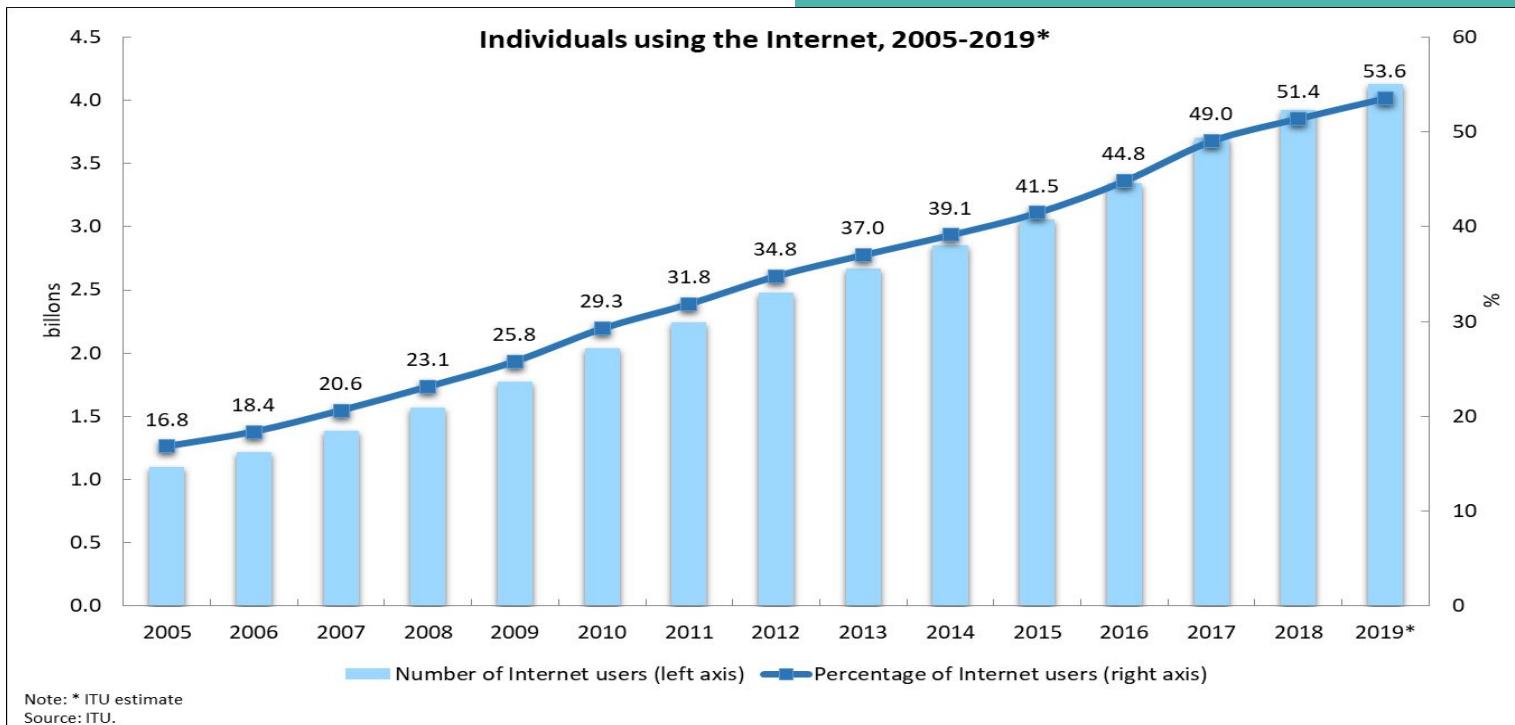
---

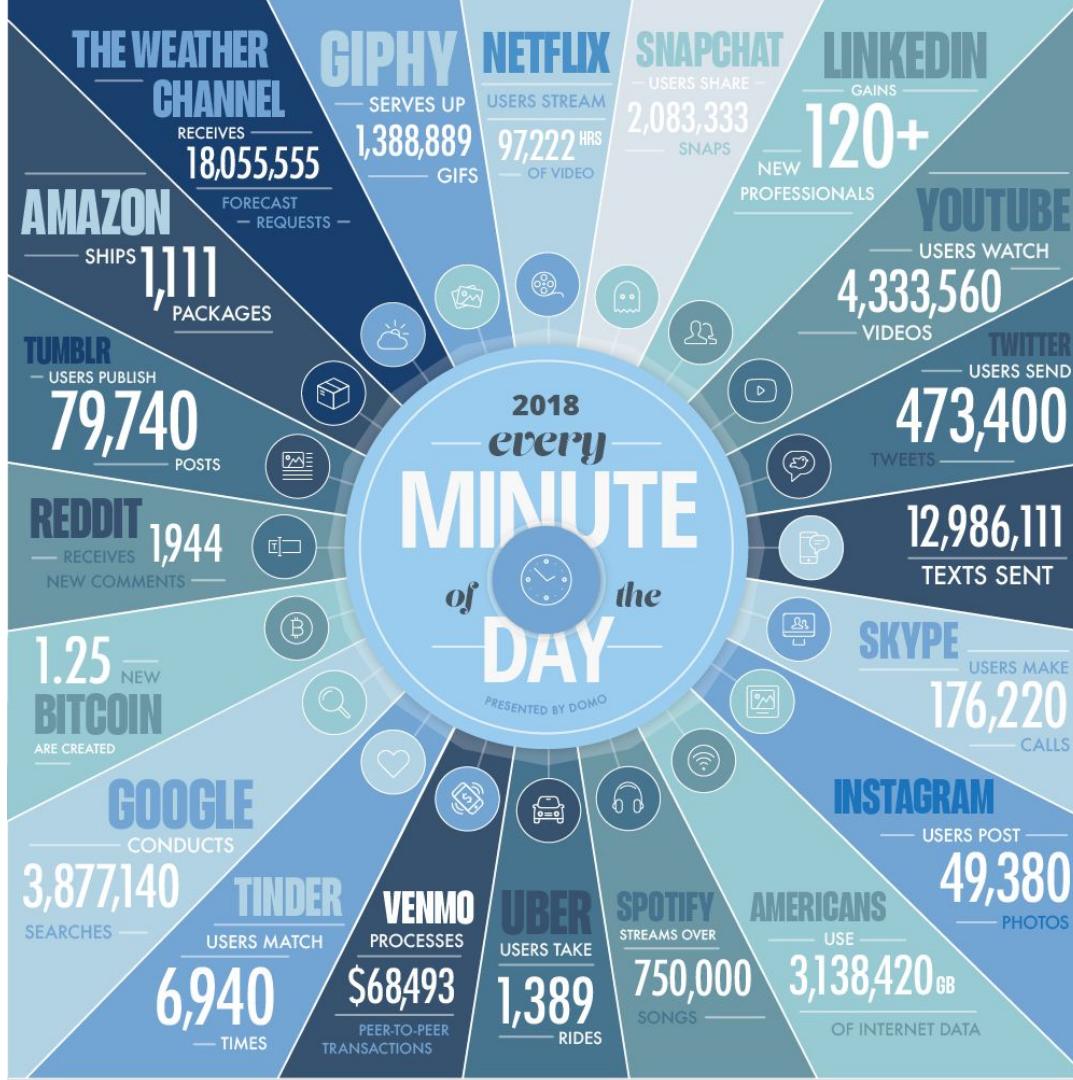
How much data is generated *every minute*?

There's no way around it: big data just keeps getting bigger. The numbers are staggering, but they're not slowing down. By 2020, it's estimated that for every person on earth, 1.7 MB of data will be created every second. In our 6th edition of Data Never Sleeps, we once again take a look at how much data is being created all around us every single minute of the day—and we have a feeling things are just getting started.

“Computers have promised us a fountain of wisdom but delivered a flood of data”

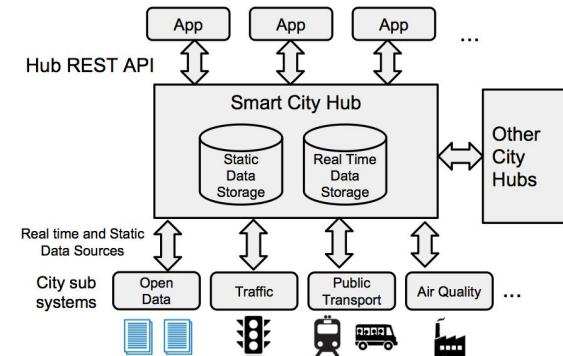
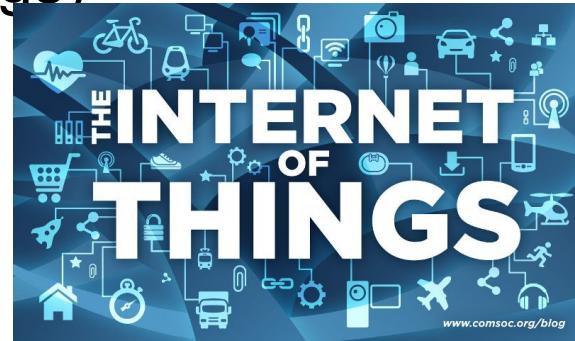
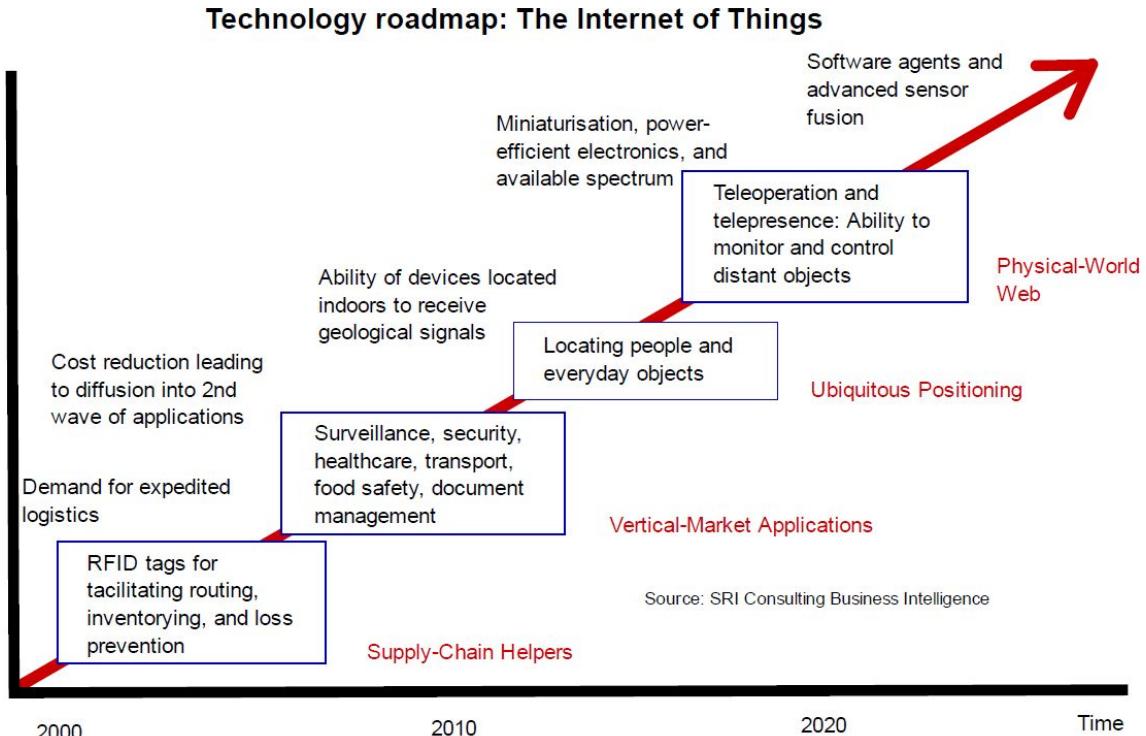
William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus





# Ce urmeaza ? Mai multe date via IoT (Internet of Things)

Technology Reach



# Cateva curiozitati

din experienta mea in modelare si data mining

Computational Biomedicine Laboratory, University of Houston:

Abia prin 2005 s-a observat ca neuronul excitat transmite semnal nu numai inainte, catre neuronii conectati la dendrite, dar si inapoi, catre neuronii care l-au facut sa emita.

Nationwide Insurance – locul 8 in SUA in 2015, cu 3% din piata:

Asiguratorii stiu de prin 2000, in urma unui studiu extins, ca scorul de credit este foarte puternic corelat cu despagubirile – cu alte cuvinte modul de cheltuire a banilor si de plata a creditelor indica/prezice stilul de sofat. Este insa ilegală utilizarea directa a acestei informatii confidentiale (credit score).

Pioneer, o companie DuPont (concurrent cu Monsanto)

Produc porumb experimental in 50 de locatii pe glob intr-o cantitate egala in 2006 cu cantitatea de porumb produsa de Australia

Intre 2006-2008 costurile de marcarea a ADN-ului s-au redus cu 50-75%. Numarul de markeri genetici a crescut de la sute la zeci de mii, la fel au crescut datele colectate si folosite in modele predictive bazate pe regresie la markeri = variabile independente

PROS, Inc. “The Revenue and Profit Realization Company“

In 2019, 57% din piata de bilete de avion foloseste (diverse versiunie ale unui) soft PROS de generare a preturilor optime (Lufthansa 25+ ani)

Pricing: soft de optimizare de preturi : HP, BP, Shell, Siemens, BASF, Disney;

- Optimizare de venit: Lufthansa, Iberia, Swiss, Etihad, Turkish Airlines, Southwest, etc. (!50 companii); Caribbean, Deutsche Bahn, Hertz, Avis;

# Agenda curs1

Ce este Data Mining?

Ce este un 'Data Scientist'?

Trecere in revista a tehnicilor de data mining

# Ce este Data Mining? \*

- Termen creat recent pentru o confluență de idei de la statistica și informatică (machine learning și metode de baze de date) aplicate la baze de date mari în știință, inginerie și afaceri.
- Definiri multiple, dispute legate de ce este și ce nu este data mining. Terminologia aferentă nu este standardizată încă (e.g. bias, clasificare, predicție)

# Definitii generice si punctuale

Definitiile generice includ metode statistice traditionale, cele punctuale scot in evidenta automatizari si metode euristice.

Exemple de nomenclatura:

- Data mining, data dredging (dragare de date), fishing expeditions
- Knowledge Discovery in Databases (KDD)

# Gartner Group

(<http://www.gartner.com/it-glossary/?s=Data+mining> + google translate )

“Data mining este procesul de a descoperi corelații, modele și tendințe noi, semnificative prin analizarea unor cantități mari de date stocate în arhive, folosind tehnologiile de recunoaștere a modelelor (pattern recognition) precum și tehnici statistice și matematice”

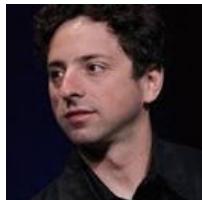
# Alte definitii si comentarii



Daryl Pregibon

[1] "KDD = Statistics at Scale and Speed"

[2] "un amestec de statistica, inteligenta artificiala, si cercetare in baze de date [mari]"



Sergey Brin – cofondator Google, PhD in DM

"Google's mission is to organize the world's information, making it universally accessible and useful"

[1] Daryl Pregibon. 2001: A statistical odyssey. KDD Conference '99, 1999

[2] Daryl Pregibon 1997: Data Mining. Statistical Computing and Graphics

# De ce Data Science?

- Piata: tranzitia de la focus pe produs/serviciu la focus pe consumator
- IT: tranzitie de la inregistrari actualizate la zi catre cautare de sabloane/modele in inregistrari - Data Warehouses - OLAP

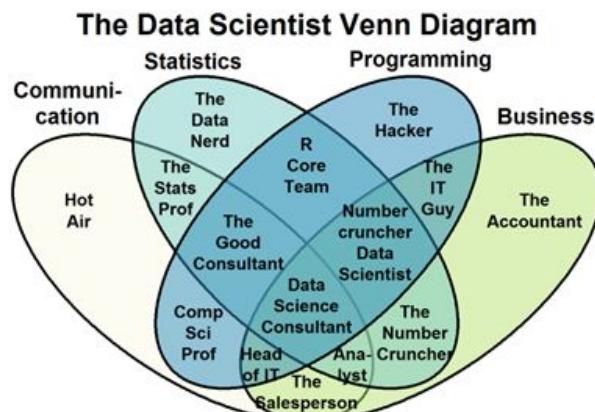
- Scadere dramatica a costurilor de stocare a datelor

Baze de date uriasă – exemplu: Walmart: 20 milioane tranzactii/zi, baza de date de date de 10 terabyte (2003) Google, Facebook, era video

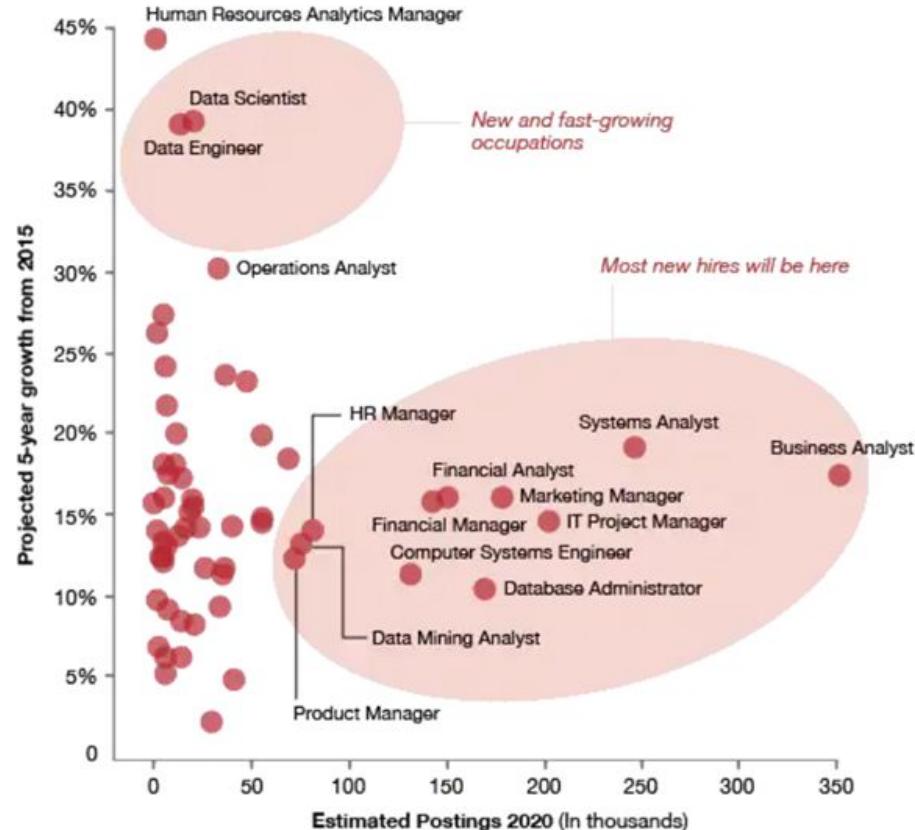
- Captare automatizata a datelor din tranzactii – e.g. Cod bare, POS, clicuri Mouse, localizari (GPS, tel inteligente)
- Internet: interactiuni personalizate (Facebook, Google), date longitudinale
- Accelerarea producerii de date in era mobile

# De ce data science?

- „Datele sunt noul petrol” - [Clive Humby](#), 2006
- „IA este noul curent electric” - [Andrew Ng](#), 2017
- „Software 2.0” - [Andrej Karpathy](#), Director of AI, Tesla, 2017



Data Scientist Venn Diagram by philippe therond Last updated about 3 years ago



Note: Each dot represents an occupation in the US jobs market where data science and analytics skills are required.  
Source: PwC analysis based on Burning Glass Technologies data, January 2017.

<https://www.pwc.com/us/en/library/data-science-and-analytics.html>

## Nevoie acută de specialiști data science

Numărul companiilor care se bazează pe servicii data science de exploatare a datelor este în continuă creștere, generând o cerere continuă de specialiști în domeniu



# Tehnici esentiale in DM

- Statistica (adaptata pentru marimea datelor si viteza de procesare din secolul 21)

Exemple:

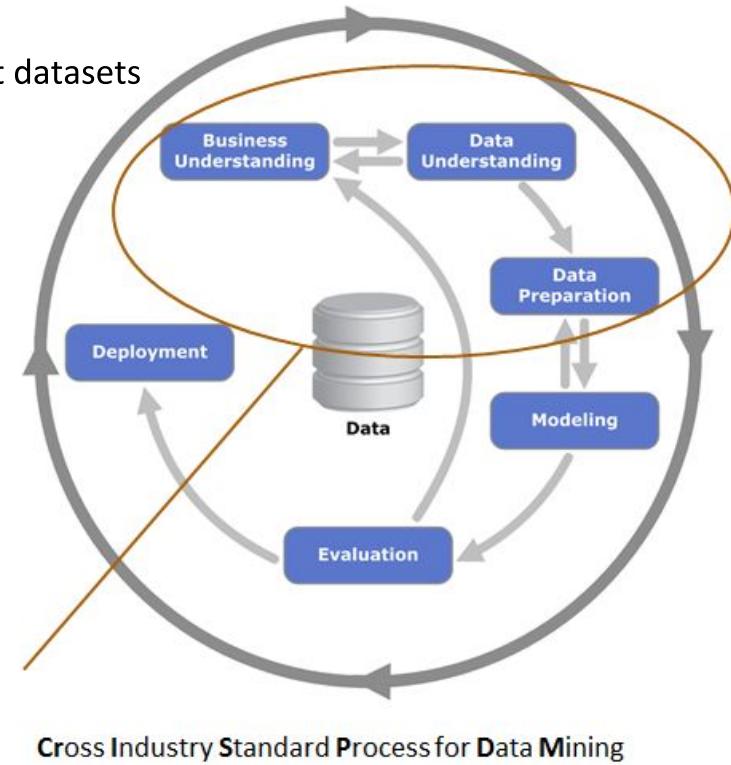
- Statistica Descriptiva (EDA)
- Vizualizari de date
- Modelare: Regresie, Analiza Cluster, Clasificare
- Machine Learning, Inteligenta Artificiala: e.g. Retele Neuronale
- Lucrul cu baze de date: e.g. Reguli de asociere (de exemplu pt recomandari de produse)
- Procesare paralela: e.g. metode cu arbori, k Nearest Neighbors, OLAP-EDA si mai nou cu baze de date non-SQL, Hadoop, algoritmi de tip map-reduce

# Pasii unui proiect de exploatare a datelor

1. Intelegerea scopului si a datelor
2. Asamblarea setului de date pentru studiu
3. Curatirea, asanarea si preprocesarea datelor
4. Reducerea/simplificarea/agregarea datelor
5. Alegerea tipului de tehnica DM
6. Alegerea algoritmului DM potrivit
7. Rularea algoritmului
8. Interpretarea rezultatului si iteratii ale pasilor 4-7 daca e necesar
9. Integrarea procesului intr-un sistem operational

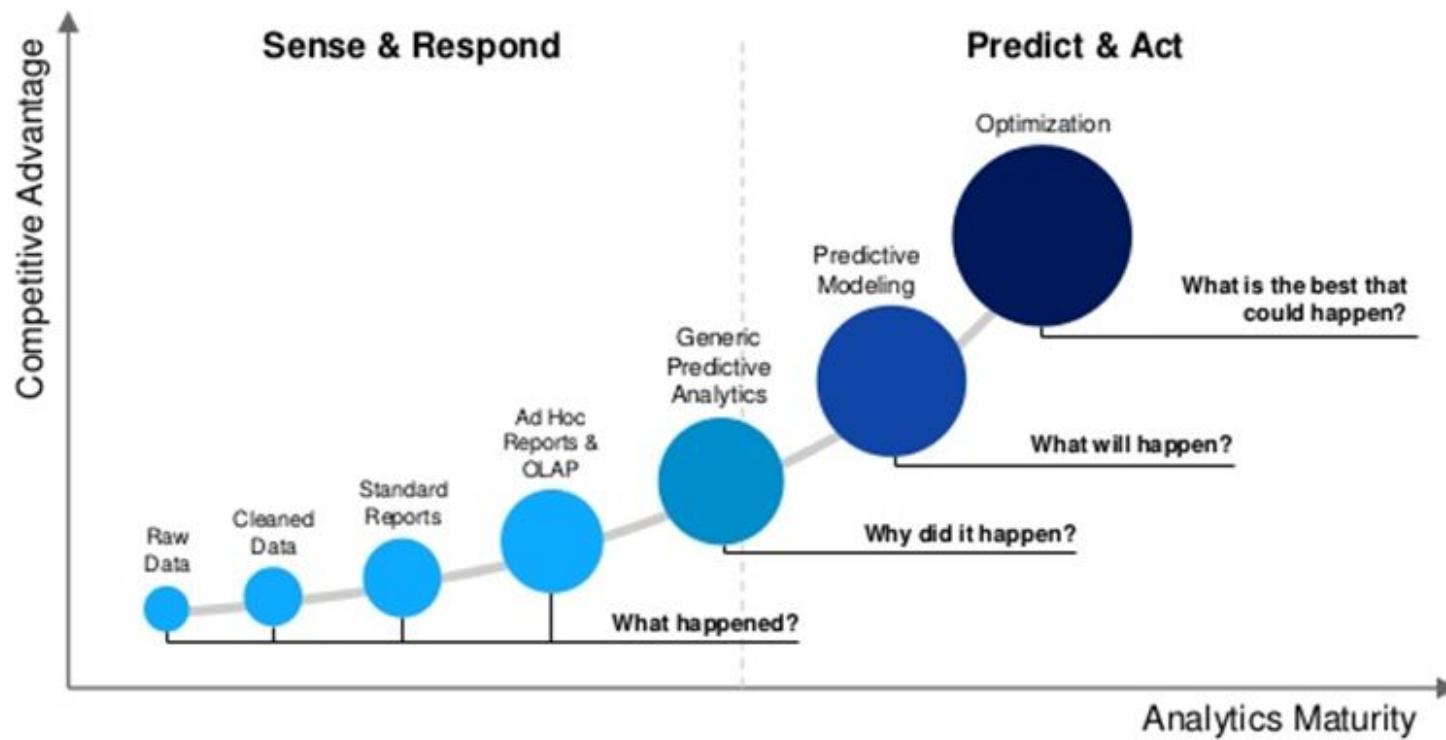
# Metodologii: SEMMA (SAS) vs. CRISP-DM

- Sample from data sets, Partition into Training, Validation and Test datasets
- Explore data set statistically and graphically
- Modify: Transform variables, Impute missing values
- Model: fit models e.g. regression, classification tree, neural net
- Assess: Compare models using Partition, Test datasets



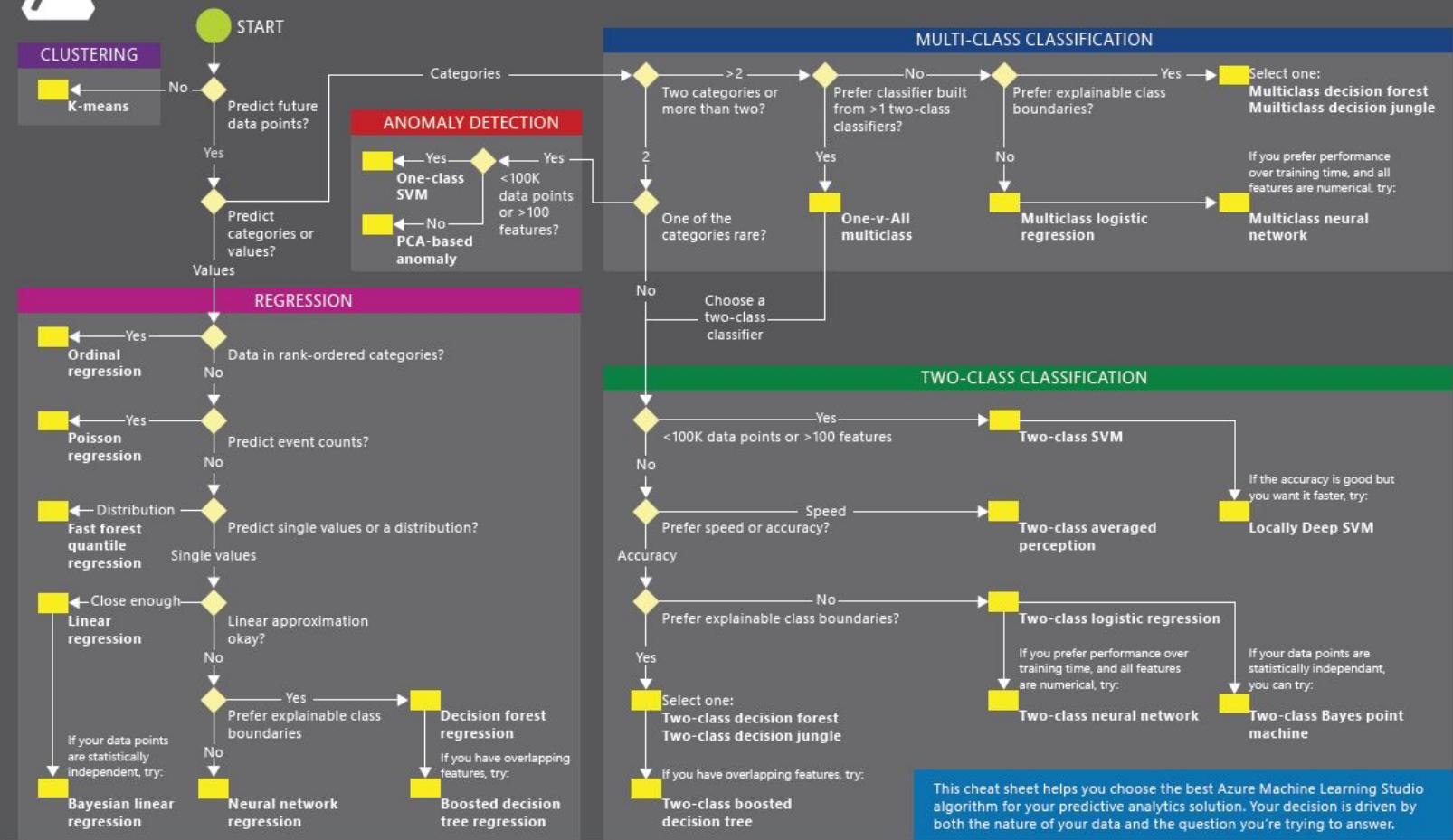
Esential, uneori 90% din proiecte  
repetitive.

# Descriptiv, Predictiv, Prescriptiv<sup>[1]</sup>





# Microsoft Azure Machine Learning: Algorithm Cheat Sheet



# Exemple de aplicatii <sup>[1], [2]</sup>

## Servicii de telefonie si utilitati

Problema:

- Clienti traseisti –CHURN

Solutie:

- Segmentarea clientilor pe baza datelor de facturare, a interactiunilor directe, a vizitelor pe site, etc.
- Construirea unui scor probabilistic de pierdere a clientului

Actiune:

- Oferte personalizate si stimulente pentru fidelizarea celor cu risk mare

# Exemple de aplicatii <sup>[1], [2]</sup>

## France Telecom

Odata cu implementare unei solutii de data warehouse, pe partea BI s-a introdus si un sistem CHURN de profilare a clientilor (Customer Profile System – CPS)

- Profilarea preventiva a dus la retentia sporita a clientilor

# Exemple de aplicatii [\[1\]](#), [\[2\]](#)

## Retail – comerțul cu amanuntul

Problema:

- Rezultate neclare din promocii

Solutie:

- Optimizarea promociilor prin segmentarea clientilor pe baza tiparului de cumpărături (purchase pattern)

Actiune:

- Clienti care cumpara des, putin, revin in timp, si au revenit de curand vor primi carduri de loialitate, si vor fi stimulati pentru upsell (catre branduri mai scumpe) si cross-sell (propuneri pentru extinderea cosului de cumpărături)
- Clienti care cumpara mult si foarte rar pot primi o oferta personalizata de tip 'win-back customer'

# Exemple de aplicatii [1], [2]

## Prevenirea fraudelor -Automobile Insurance Bureau of Massachusetts

Problema:

- Fraudele in asigurari maresc costurile si micsoreaza veniturile

Solutie:

- Folosind regresia logistica si retele neuronale s-au identificat caracteristicile cazurilor frauduloase pornind de la: date client, detalii ale accidentelor, detalii ale daunelor
- Prin segmentarea clientilor in grupe de risc se previn fraude ulterioare

Actiune:

- Clienti profilati ca problematici vor fi refuzati sau vor primi prime de asigurare mai mari astfel marindu-se venitul

# Exemple de aplicatii [1], [2]

## Agentii de Inteligenta, Criminalistica

Problema:

1. Detectarea zonelor cu grad sporit de risc
2. Detectarea persoanelor cu grad sporit de risc la granite
3. Profilarea informatiilor agentiilor de securitate

Solutie:

1. Segmentarea geografica pe baza unui scor din statisticile cumulate
2. Clasificarea traficului de frontier folosind caracteristici personale (date pasaport) si contextuale (geografie, tip transport, scopul vizitei, etc.)
3. Clasificare informatiilor de la agentii

Actiune:

1. Sporirea personalului de paza si protectie a ordinii publice la locul si timpul potrivit
2. Control selectiv amanuntit
3. Construirea unui caz pentru actiuni de contra-terorism

# Exemple de aplicatii [\[1\]](#), [\[2\]](#)

## DM in imagini

- Google cu image matching,
- recunoastere faciala, etc.
- Determinarea sentimentului pozitiv/negativ

## DM in text

Filtre SPAM – primul filtru folosea clasificatori Bayesieni

Retele Sociale, Targeted Marketing, etc.

# Administrative

## Etape Proiecte/Seturi de Date Disponibile

- Intelegerea problemei, evaluarea datelor brute si analiza initiala (partea descriptiva)
- Masarea datelor si aplicarea unei tehnici DM (parte predictiva)
- Prezentarea rezultatelor si a unor strategii de utilizare (partea prescriptiva)

## Echipe de lucru si roluri:

- “Analist” (1 loc)
- “Miner” (1loc)
- “Consultant” (1loc)

# Cum vom lucra?

## Python 3.\*

- numpy, pandas, scikit-learn, plotly
- dataframes, dictionaries, lists, json files
- Machine learning: clasificatori, regresori, validari incrucisate, linii de antrenare, validare, si cautare hiperparametrii

## Anaconda - Jupyter Notebooks

Laboratoarele se vor desfasura cu prezența fizică

Anaconda va fi instalat pe stațiile de lucru

Puteti salva in contul propriu temele de laborator

Puteti folosi laptopul vostru daca preferati

# Administrative

## Etapele examinarii

1. Saptamana 7 - prezentare analiza initiala, alegere model/tehnica, determinarea atributelor esentiale pentru modelare
2. Saptamana 14 - prezentarea finala a rezultatelor precum si a indicatiilor prescriptive