

# FIȘA DISCIPLINEI

Denumire: **Mașini instruibile și recunoașterea formelor**

Număr de ore: Curs: 2, Laborator: 2

Număr de credite: 6

Condiționări la înscriere (discipline anterioare cerute):

- Inteligență Artificială

# OBIECTIVE

Cursul realizează o introducere în teoria și practica Mașinilor Instruibile și Recunoașterii formelor, ca parte a Inteligenței Artificiale.

Sunt tratate:

Învățarea conceptelor

Arbori de decizie

Rețele neuronale

Învățarea analitică

Învățarea bayesiană, etc.

Studentii sunt familiarizați și cu principalele abordări și realizări din domeniul recunoașterii formelor (abordarea statistică, abordarea neuronală, abordarea sintactică).

# Bibliografie

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer-Verlag, 2013. <http://www-bcf.usc.edu/~gareth/ISL/> (gratis online)
- Hal Daume III, [A Course in Machine Learning](#) (gratis online)
- Christopher Bishop. Pattern Recognition and Machine Learning, Springer, 2006.  
<http://research.microsoft.com/en-us/um/people/cmbishop/prml/>
- David Mackay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.  
<http://www.inference.phy.cam.ac.uk/itprnn/book.html>
- Tom Mitchell. Machine Learning. McGraw Hill, 1997.  
<http://www.cs.cmu.edu/~tom/mlbook.html>

# Bibliografie

- Dumitrescu, D., *Principiile Inteligentei artificiale*, Editura Albastră, Cluj, 2000.
- Ileană Ioan, Corina Rotar, Maria Muntean: *Inteligență Artificială*, Ed. Aeternitas, 2009.
- Ileană Ioan: *Rețele neuronale în tehnologie optoelectronică. Aplicații în recunoașterea formelor*, Ed. Aeternitas, Alba Iulia, 2002.
- Neagoe Victor, Stănășilă O.: *Recunoașterea formelor și rețele neuronale*, Ed. Matrix Rom, 1999.
- Nedevschi Sergiu: *Prelucrarea imaginii și recunoașterea formelor*, Editura Albastră, Cluj-Napoca, 1998.

# Introducere în domeniul mașinilor instruibile

# Introducere

- Ce sunt Mașinile Instruibile
- Istoric
- Aplicații / Exemple

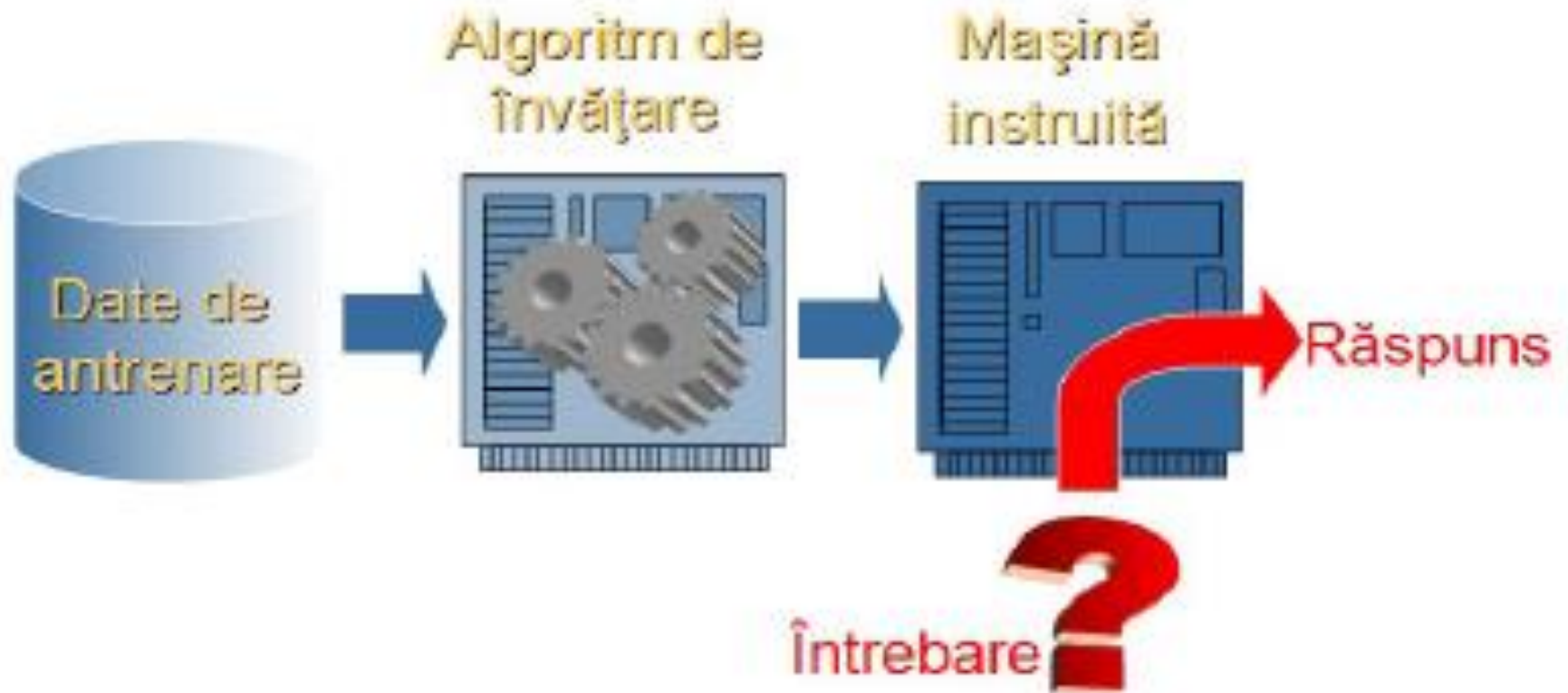
# Învățarea în sistemele artificiale inteligente

Inteligență  
Sisteme inteligente  
Învățarea automată

Definiție a procesului de învățare:

"Orice schimbare într-un sistem, care permite sistemului îmbunătățirea performanțelor în rezolvarea ulterioară a aceleiași probleme sau a unei probleme similare, reprezintă un proces de învățare".

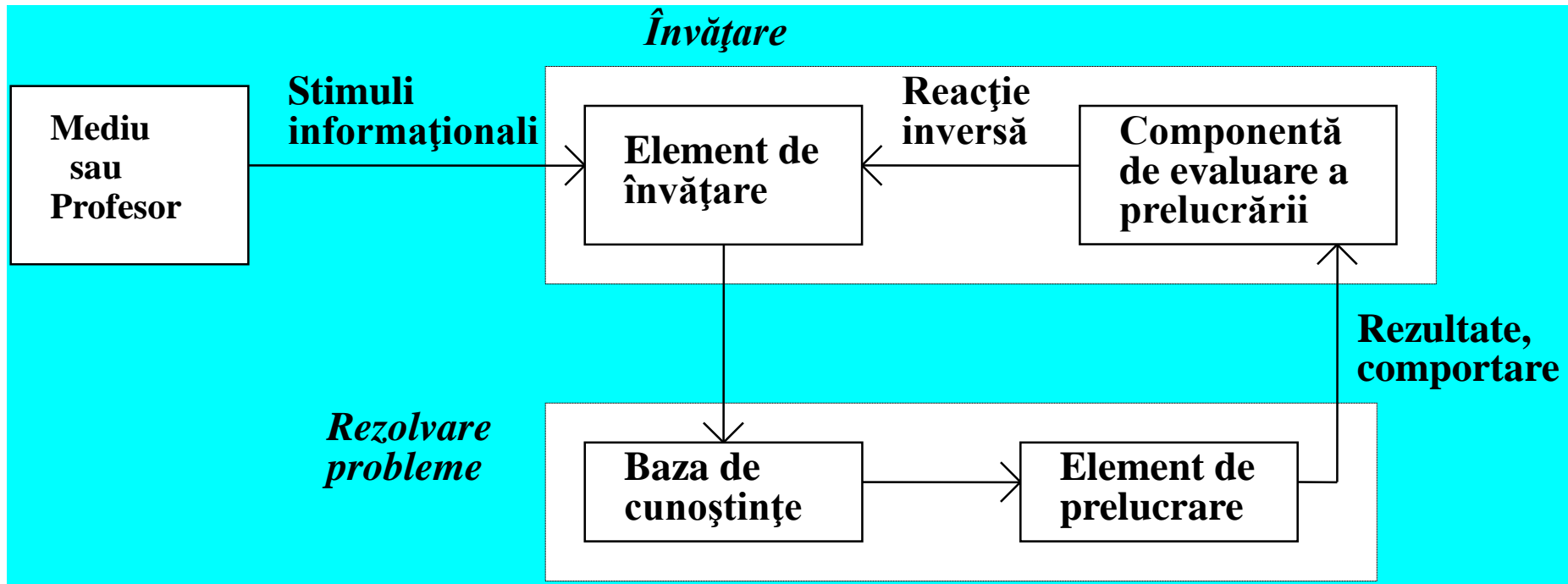
# Învățarea automată





# Clasificarea tipurilor de învățare

Un sistem de învățare automată poate fi văzut ca fiind format din cinci componente principale:



# Terminologie

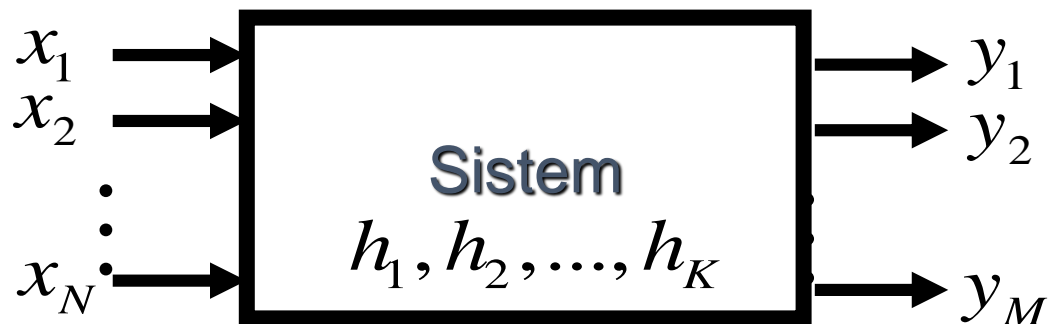
- Masini instruibile si recunoasterea formelor (MIRF)
- Învățare automată
- Instruire automată
- Machine learning

# Ce sunt Mașinile Instruibile?

- După dicționarul Webster, a învăța înseamnă a câștiga cunoștințe, înțelegere sau pricepere prin studiu, instruire sau experiență
- Tehnicile ML sunt algoritmi ce permit mașinilor să-și îmbunătățească performanțele în îndeplinirea unei anumite activități prin experiență
- Mașini instruibile / învățarea automată (**machine learning - ML**) = mașini care învață
- Învățare = Îmbunătățirea realizării unei acțiuni cu experiența
  - Îmbunătățirea acțiunii (task) T
  - În raport cu o măsură a performanței P
  - Bazată pe experiența E

# Ce sunt Mașinile Instruibile?

- Domeniul Mașinilor Instruibile se ocupă cu modul de construire a programelor de calculator care se autoperfecționează prin experiență
- ML învață structurile și regulile existente și descoperă noi înțelesuri și structuri
- Scopul învățării automate este de a construi sisteme de calcul care se pot adapta și pot învăța din experiența lor
- Algoritmii ML descoperă relațiile dintre variabilele unui sistem (intrări, ieșiri, variabile ascunse) direct din eșantioane ale sistemului.



# Când nu sunt necesari algoritmi ML?

- Când relațiile dintre toate variabilele sistemului (intrare, ieșire, ascunse) sunt complet cunoscute!
- Acesta NU e cazul pentru aproape nici un sistem real!

# Abordarea ML

- Învățarea prin exemple
- În loc să scriem un program pentru fiecare sarcină specifică, colectăm o mulțime de exemple care specifică o ieșire corectă pentru o anumită intrare.
- Un algoritm de învățare automată ia aceste exemple și produce un program care rezolvă problema.
- În cazul în care datele se schimbă, programul se poate schimba și el prin antrenarea pe noi date.
- Capacități masive de calcul sunt acum mai ieftine decât a plăti pe cineva să scrie un program de sarcini specifice.

# Motivația învățării automate

- Unele acțiuni (sarcini) pot fi specificate doar prin date sau exemple de antrenare
- Expertiza umană poate fi rară și/sau foarte scumpă
- Cantitatea de cunoștințe poate fi prea mare pentru o codificare explicită de către oameni
- Este necesară modelarea și estimarea unor parametri ascunși: Adesea sunt disponibile doar date din măsurători
- Puterea de calcul este în continuă creștere
- Cantitatea de date disponibile și capacitatea de stocare este tot mai mare

# Motivația învățării automate

- Progrese recente în algoritmi și teorie
- Un flux exponențial de date online
- Putere de calcul disponibilă

## Nișe pentru Mașini Instruibile

- Data Mining: se folosesc date istorice pentru îmbunătățirea deciziilor  
Înregistrări medicale → cunoștințe medicale
- Aplicații software pe care nu le putem programa direct:  
Conducerea autonomă a vehiculelor  
Recunoașterea vorbirii
- Programe care se auto - personalizează  
Cititoare de ziare care învață domeniile de interes ale utilizatorilor



# Introducere în domeniul mașinilor instruibile (învățării automate)

- Ce sunt Mașinile Instruibile
- **Istoric**
- Aplicații / Exemple

# Scurtă istorie a ML

- Anii 1950:
  - Programul lui Samuel de jucat dame
  - Perceptronul lui Rosenblatt
- Anii 1960:
  - Rețele neuronale
- Anii 1970:
  - Winston - ARCH
  - Meta-Dendral (Buchanan and Mitchell): reguli de prognoză bazate pe spectrometria de masă
  - ID3 (Quinlan): Reguli pentru încheierea jocului de șah
  - Michalski - AQ11: reguli de diagnostic (soia)
  - MACROPS: macro operatori în planificare în lumea blocurilor

# Scurtă istorie a ML

- Anii 1980: Progrese (câștiguri)
  - Teoria învățării
  - Algoritmi de învățare simbolică
  - Algoritmi de învățare conexionistă
  - Grupare (Clustering)
  - Învățare bazată pe explicații
  - Învățare inductivă ghidată de cunoștințe
  - Algoritmi genetici
- Anii 1990: Maturitate
  - Data mining
  - Învățare în grup (Ensemble learning: bagging, boosting) etc.
  - Metode kernel (Kernel methods)
  - Învățare cu întărire
  - Analize teoretice

- **Anii 2000**

- Mașini cu suport vectorial (SVM)
- Metode kernel
- Modele grafice
- Învățare statistică relațională
- Învățare prin transfer
- Etichetarea secvențelor
- Clasificare colectivă și ieșiri structurate
- Aplicații pentru sisteme de calcul
  - Compilatoare
  - Depanare
  - Grafică
  - Securitate (detectia intruziunilor, virușilor și viermilor)
- Managementul email-ului
- Asistenți personalizați care învață
- Învățare în robotică și vedere artificială

# Introducere în domeniul mașinilor instruibile (învățării automate)

- Ce sunt Mașinile Instruibile
- Istoric
- Aplicații / Exemple

# Discipline cu care ML interacționează

- Statistică
- Inteligența artificială
- Psihologie
- Medicină
- Teoria controlului
- Procesarea semnalelor și imaginilor
- .....

# Aplicații

- Nișe pentru Mașini Instruibile:

- Data Mining: se folosesc date istorice pentru îmbunătățirea deciziilor. Înregistrări medicale → cunoștințe medicale
- Aplicații software pe care nu le putem programa direct: conducerea autonomă a vehiculelor, recunoașterea vorbirii
- Programe care se auto - personalizează: cititoare de ziare care învață domeniile de interes ale utilizatorilor
- Medicină: analiza imaginilor, diagnostic și prognoză, descoperirea de medicamente
- Diferite aplicații există în domeniul securității: recunoașterea feței, verificarea semnăturii, a amprentelor, irisului, ADN-ului
- Recunoașterea modelelor/pattern-urilor: obiecte în scene reale, identități faciale sau expresii faciale, cuvinte vorbite
- Recunoașterea anomaliilor: secvențe neobișnuite de tranzacții pe cartile de credit, modele neobișnuite ale cititoarelor de senzori într-o centrală nucleară
- Predicție: prețurile acțiunilor viitoare sau ratele de schimb valutar, ce filme îi vor placea unei persoane?

# Exemple

- Clasificarea emailurilor: spam vs non-spam
- Predicții referitoare la riscul de a avea un atac de cord pe baza datelor demografice, clinice și dietă
- Identificarea cifrelor scrise de mână din codurile poștale scrise pe scrisori
- Clasificarea unei probe din țesut într-una dintre mai multe categorii de cancer pe baza expresiei genelor
- Legătura dintre venit unei persoane și alte caracteristici, printre care nivelul de educație, vârsta, etc
- Preprocesări ale unei imagini în vederea recunoașterii automate
- Netflix
- Competiții Kaggle ([www.kaggle.com](http://www.kaggle.com))



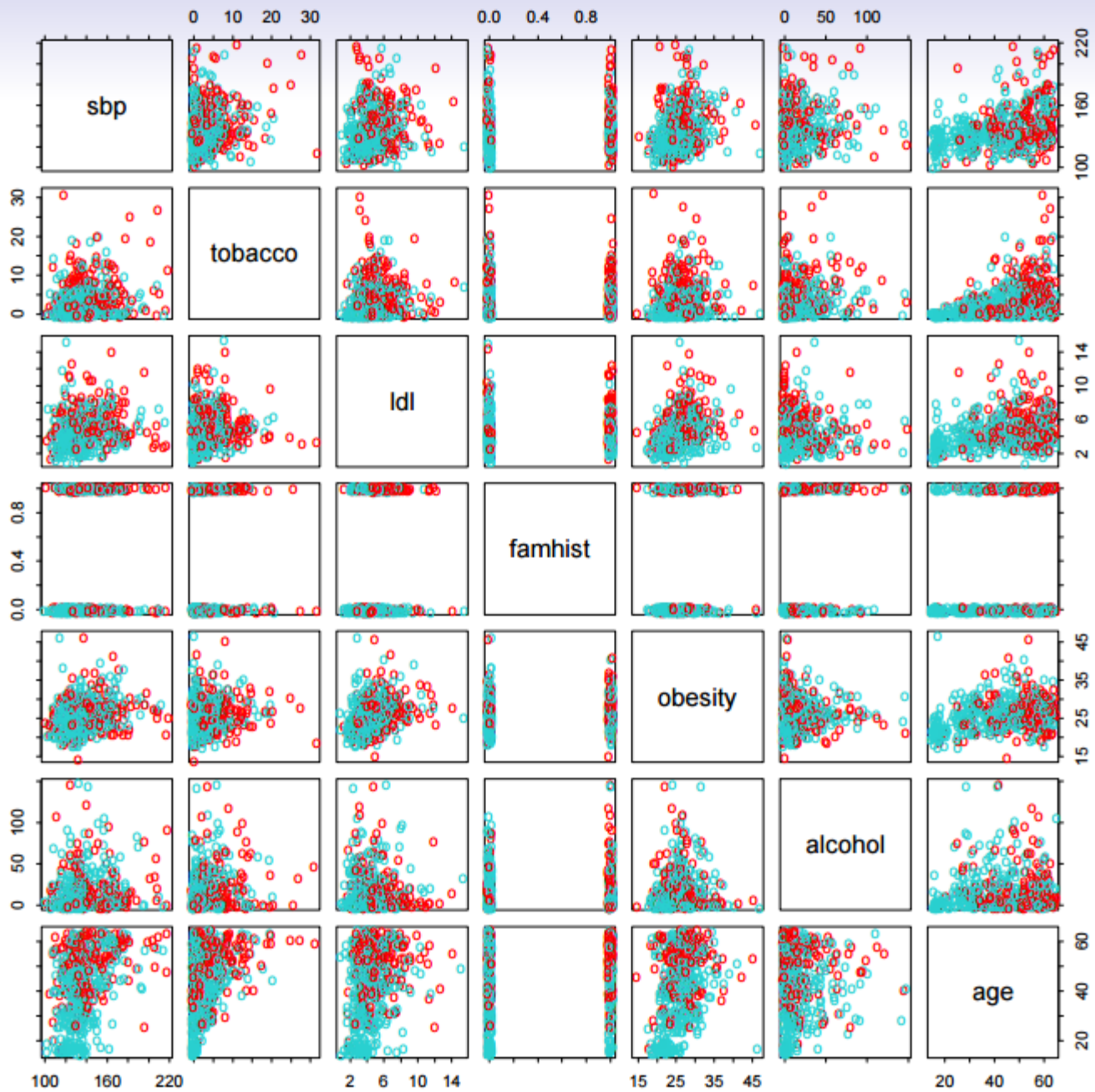
# Detectarea spam-urilor

- Datele conțin 4601 emailuri primite de către o persoană.
- Fiecare email e clasificat ca fiind spam sau nu.
- Scop: crearea unui filtru automat pentru detectarea spam-urilor
- Caracteristici de intrare: frecvențele relative ale celor mai frecvente 57 de cuvinte și semne de punctuație ce apar în aceste emailuri

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

# Exemple

- Clasificarea emailurilor: spam vs non-spam
- Predicții referitoare la riscul de a avea un atac de cord pe baza datelor demografice, clinice și dietă
- Identificarea cifrelor scrise de mână din codurile poștale scrise pe scrisori
- Clasificarea unei probe din țesut într-una dintre mai multe categorii de cancer pe baza expresiei genelor
- Legătura dintre venit unei persoane și alte caracteristici, printre care nivelul de educație, vârsta, etc
- Preprocesări ale unei imagini în vederea recunoașterii automate
- Netflix
- Competiții Kaggle ([www.kaggle.com](http://www.kaggle.com))



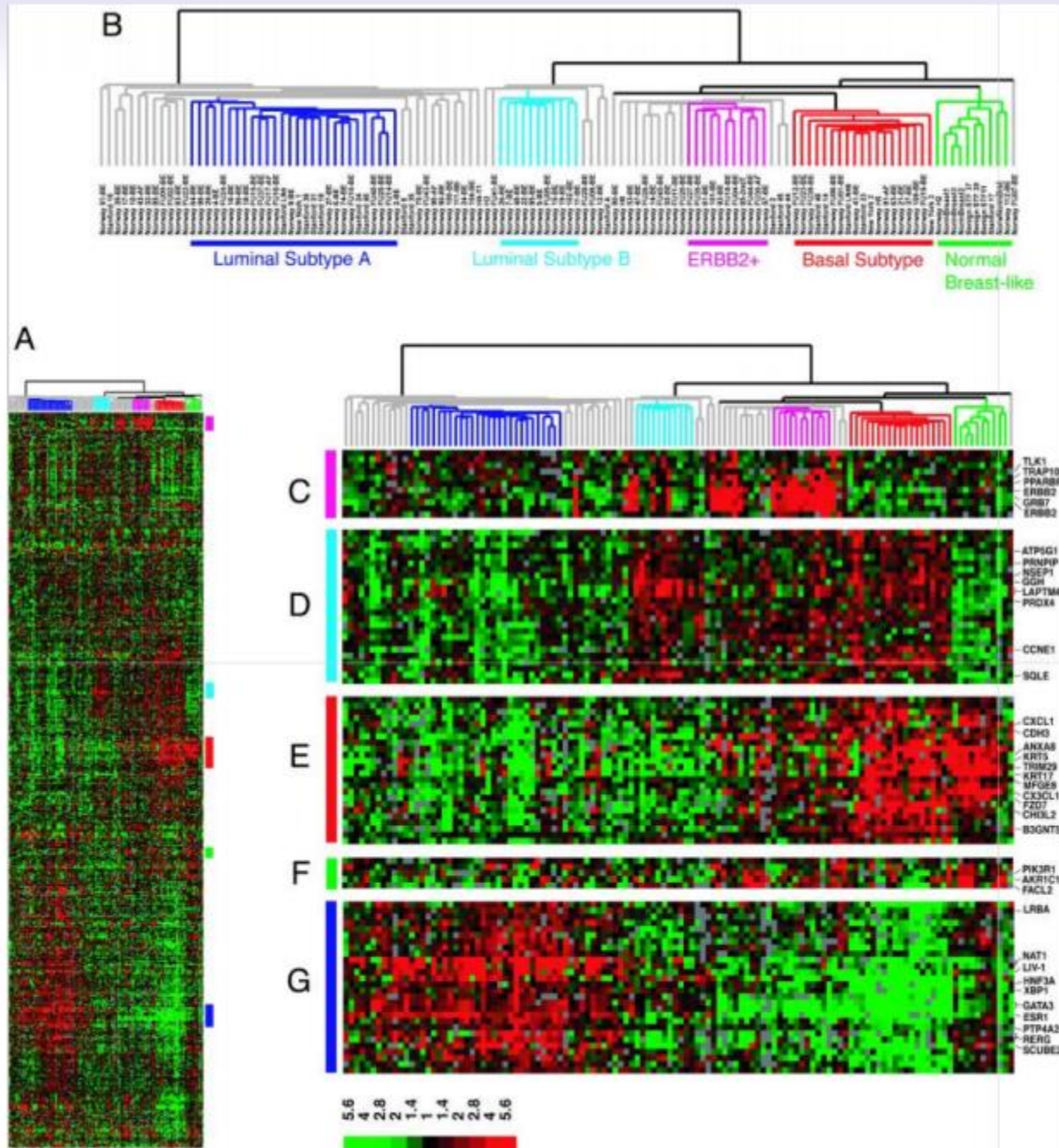
# Exemple

- Clasificarea emailurilor: spam vs non-spam
- Predicții referitoare la riscul de a avea un atac de cord pe baza datelor demografice, clinice și dietă
- **Identificarea cifrelor scrise de mână din codurile poștale scrise pe scrisori**
- Clasificarea unei probe din țesut într-una dintre mai multe categorii de cancer pe baza expresiei genelor
- Legătura dintre venit unei persoane și alte caracteristici, printre care nivelul de educație, vârsta, etc
- Preprocesări ale unei imagini în vederea recunoașterii automate
- Netflix
- Competiții Kaggle ([www.kaggle.com](http://www.kaggle.com))

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

# Exemple

- Clasificarea emailurilor: spam vs non-spam
- Predicții referitoare la riscul de a avea un atac de cord pe baza datelor demografice, clinice și dietă
- Identificarea cifrelor scrise de mână din codurile poștale scrise pe scrisori
- **Clasificarea unei probe din țesut într-una dintre mai multe categorii de cancer pe baza expresiei genelor**
- Legătura dintre venit unei persoane și alte caracteristici, printre care nivelul de educație, vârsta, etc
- Preprocesări ale unei imagini în vederea recunoașterii automate
- Netflix
- Competiții Kaggle ([www.kaggle.com](http://www.kaggle.com))

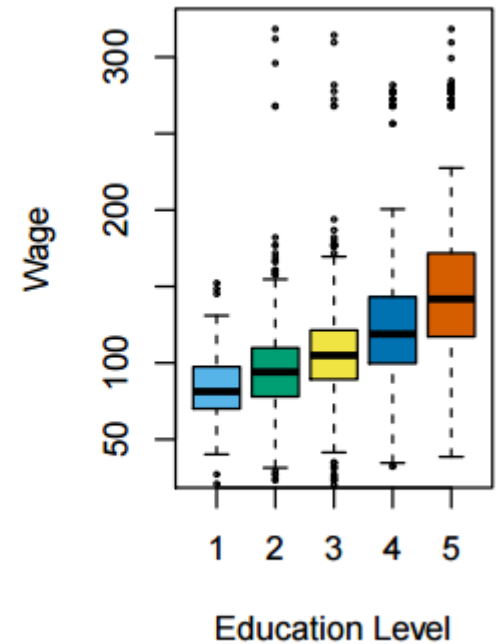
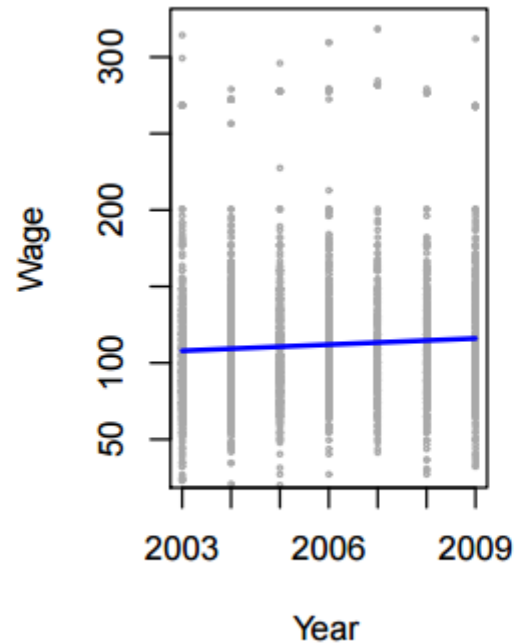
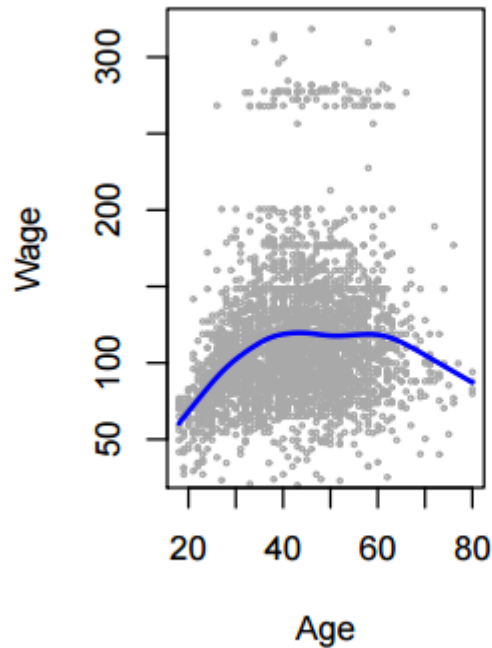


# Exemple

- Clasificarea emailurilor: spam vs non-spam
- Predicții referitoare la riscul de a avea un atac de cord pe baza datelor demografice, clinice și dietă
- Identificarea cifrelor scrise de mână din codurile poștale scrise pe scrisori
- Clasificarea unei probe din țesut într-una dintre mai multe categorii de cancer pe baza expresiei genelor
- Legătura dintre venitul unei persoane și alte caracteristici, printre care nivelul de educație, vârstă, etc
- Preprocesări ale unei imagini în vederea recunoașterii automate
- Netflix
- Competiții Kaggle ([www.kaggle.com](http://www.kaggle.com))



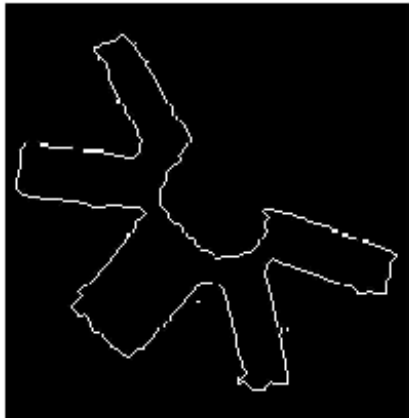
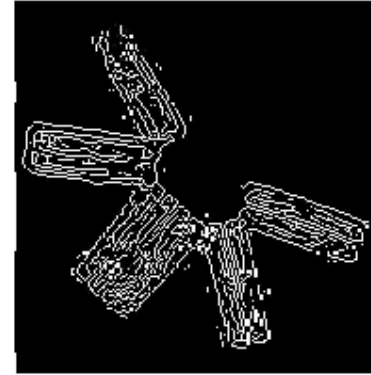
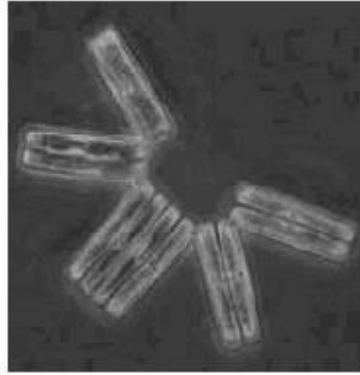
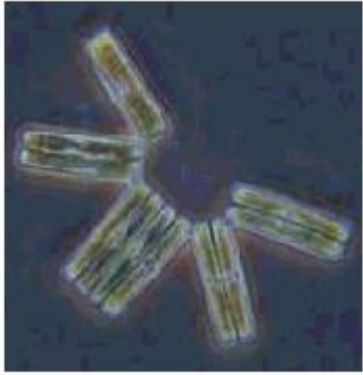
# Predicția venitului unei persoane pe baza unor caracteristici



# Exemple

- Clasificarea emailurilor: spam vs non-spam
- Predicții referitoare la riscul de a avea un atac de cord pe baza datelor demografice, clinice și dietă
- Identificarea cifrelor scrise de mână din codurile poștale scrise pe scrisori
- Clasificarea unei probe din țesut într-una dintre mai multe categorii de cancer pe baza expresiei genelor
- Legătura dintre venit unei persoane și alte caracteristici, printre care nivelul de educație, vârsta, etc
- Preprocesări ale unei imagini în vederea recunoașterii automate
- Netflix
- Competiții Kaggle ([www.kaggle.com](http://www.kaggle.com))

# Preprocesări ale unei imagini în vederea recunoașterii automate



Imagine color (fitoplancton)  
Imagine în tonuri de gri  
Detectia și liniarizarea conturilor  
Dilatarea conturilor  
Umplerea interiorului  
Eliminarea obiectelor ce depășesc marginea imaginii  
Segmentarea

# Exemple

- Clasificarea emailurilor: spam vs non-spam
- Predicții referitoare la riscul de a avea un atac de cord pe baza datelor demografice, clinice și dietă
- Identificarea cifrelor scrise de mână din codurile poștale scrise pe scrisori
- Clasificarea unei probe din țesut într-una dintre mai multe categorii de cancer pe baza expresiei genelor
- Legătura dintre venit unei persoane și alte caracteristici, printre care nivelul de educație, vârsta, etc
- **Competiția Netflix**
- **Competiții Kaggle ([www.kaggle.com](http://www.kaggle.com))**

# Netflix

- Competiție începută în oct. 2006. Datele de antrenare sunt formate din rating-uri a peste 18.000 de filme de la aprox. 400.000 de utilizatori
- Fiecare rating este un număr între 1 și 5
- Datele de antrenare sunt împrăștiate (aprox. 98% din rating-uri lipsesc)
- Obiectiv: predicții pe un set de date care nu apar în datele de antrenare (1 milion de perechi (utilizator, film))
- Premiu de 1 milion de dolari pentru o îmbunătățire cu 10% față de metoda curentă folosită de Netflix

# Netflix Prize

**COMPLETED**

[Home](#) [Rules](#) [Leaderboard](#) [Update](#)

## Leaderboard

Showing Test Score. [Click here to show quiz score](#)

Display top  leaders.

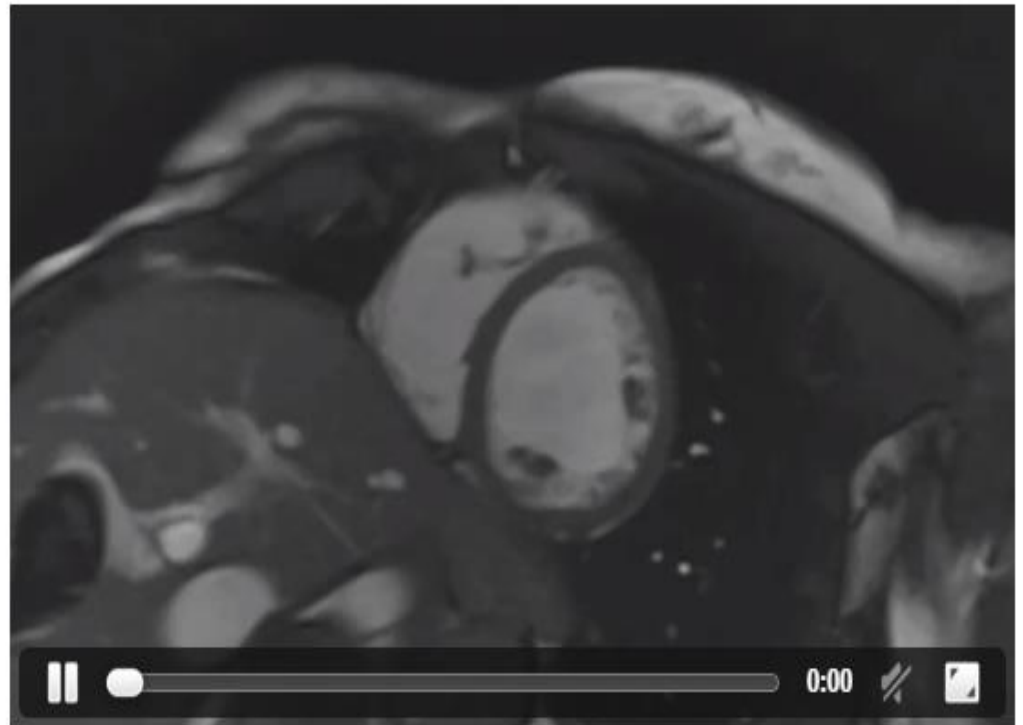
Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
<b>Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos</b>				
1	<a href="#">BellKor's Pragmatic Chaos</a>	0.8567	10.06	2009-07-26 18:18:28
2	<a href="#">The Ensemble</a>	0.8567	10.06	2009-07-26 18:38:22
3	<a href="#">Grand Prize Team</a>	0.8582	9.90	2009-07-10 21:24:40
4	<a href="#">Opera Solutions and Vandelay United</a>	0.8588	9.84	2009-07-10 01:12:31
5	<a href="#">Vandelay Industries !</a>	0.8591	9.81	2009-07-10 00:32:20
6	<a href="#">PragmaticTheory</a>	0.8594	9.77	2009-06-24 12:06:56
7	<a href="#">BellKor in BigChaos</a>	0.8601	9.70	2009-05-13 08:14:09
8	<a href="#">Dace</a>	0.8612	9.59	2009-07-24 17:18:43
9	<a href="#">Feeds2</a>	0.8622	9.48	2009-07-12 13:11:51
10	<a href="#">BigChaos</a>	0.8623	9.47	2009-04-07 12:33:59
11	<a href="#">Opera Solutions</a>	0.8623	9.47	2009-07-24 00:34:07
12	<a href="#">BellKor</a>	0.8624	9.46	2009-07-26 17:19:11

# Exemple

- Clasificarea emailurilor: spam vs non-spam
- Predicții referitoare la riscul de a avea un atac de cord pe baza datelor demografice, clinice și dietă
- Identificarea cifrelor scrise de mână din codurile poștale scrise pe scrisori
- Clasificarea unei probe din țesut într-una dintre mai multe categorii de cancer pe baza expresiei genelor
- Legătura dintre venit unei persoane și alte caracteristici, printre care nivelul de educație, vârsta, etc
- Netflix
- **Competiții Kaggle ([www.kaggle.com](http://www.kaggle.com))**

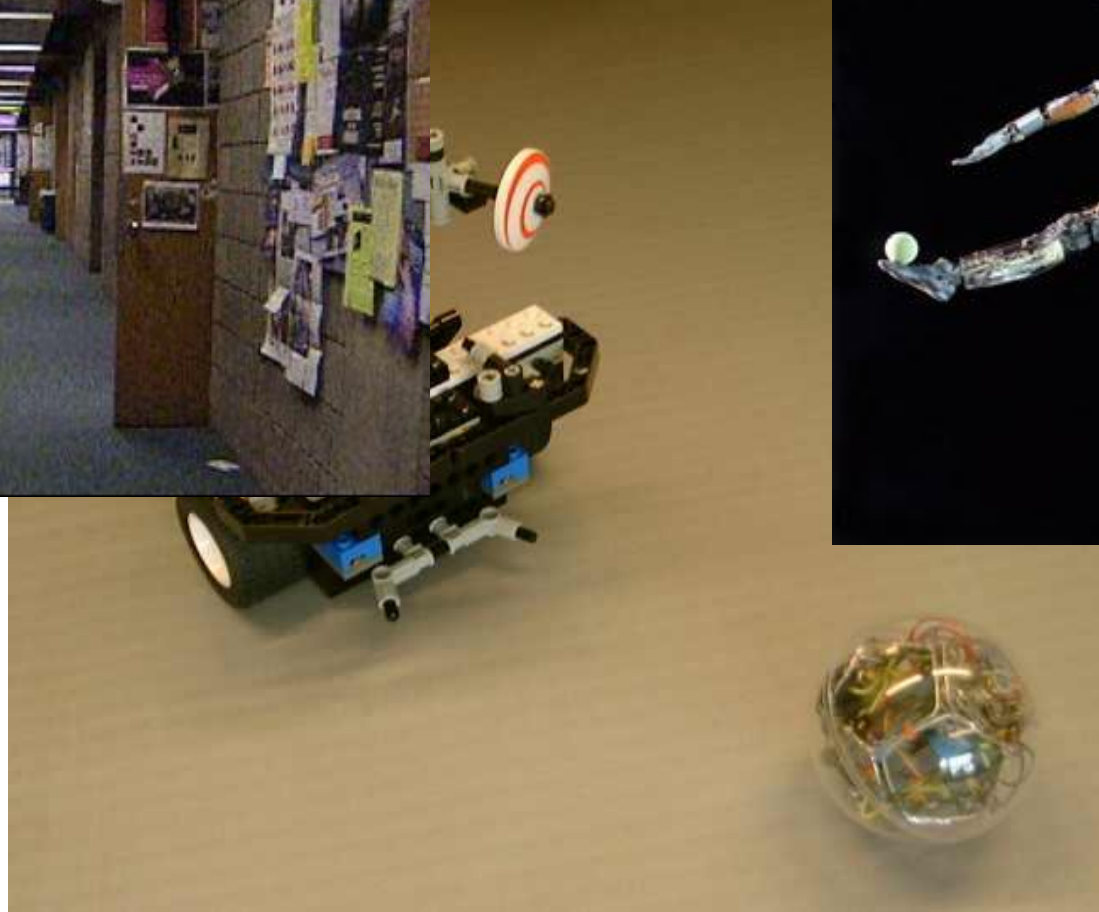
# Competiții Kaggle ([www.kaggle.com](http://www.kaggle.com))

- Diagnosticarea bolilor cardiace
- Datele de antrenare conțin mai multe imagini care reprezintă RMN-uri ale inimii
- Obiectiv: dezvoltarea unui algoritm care pe baza RMN-ului cardiac determină automat și cu precizie volumul de sânge dintr-un ciclu cardiac



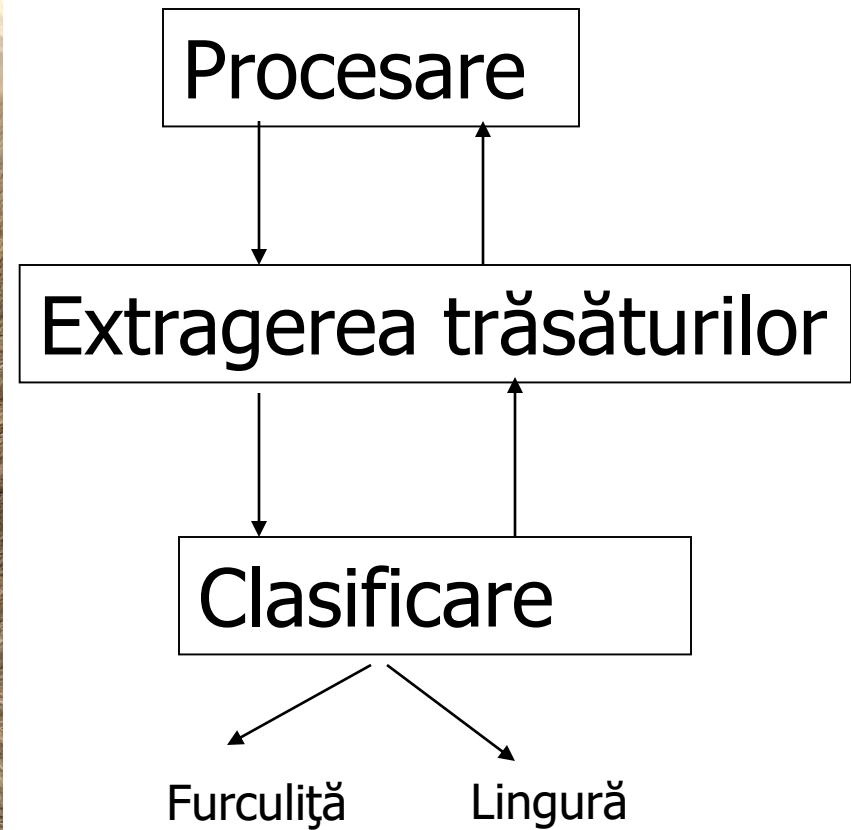


# Entități autonome

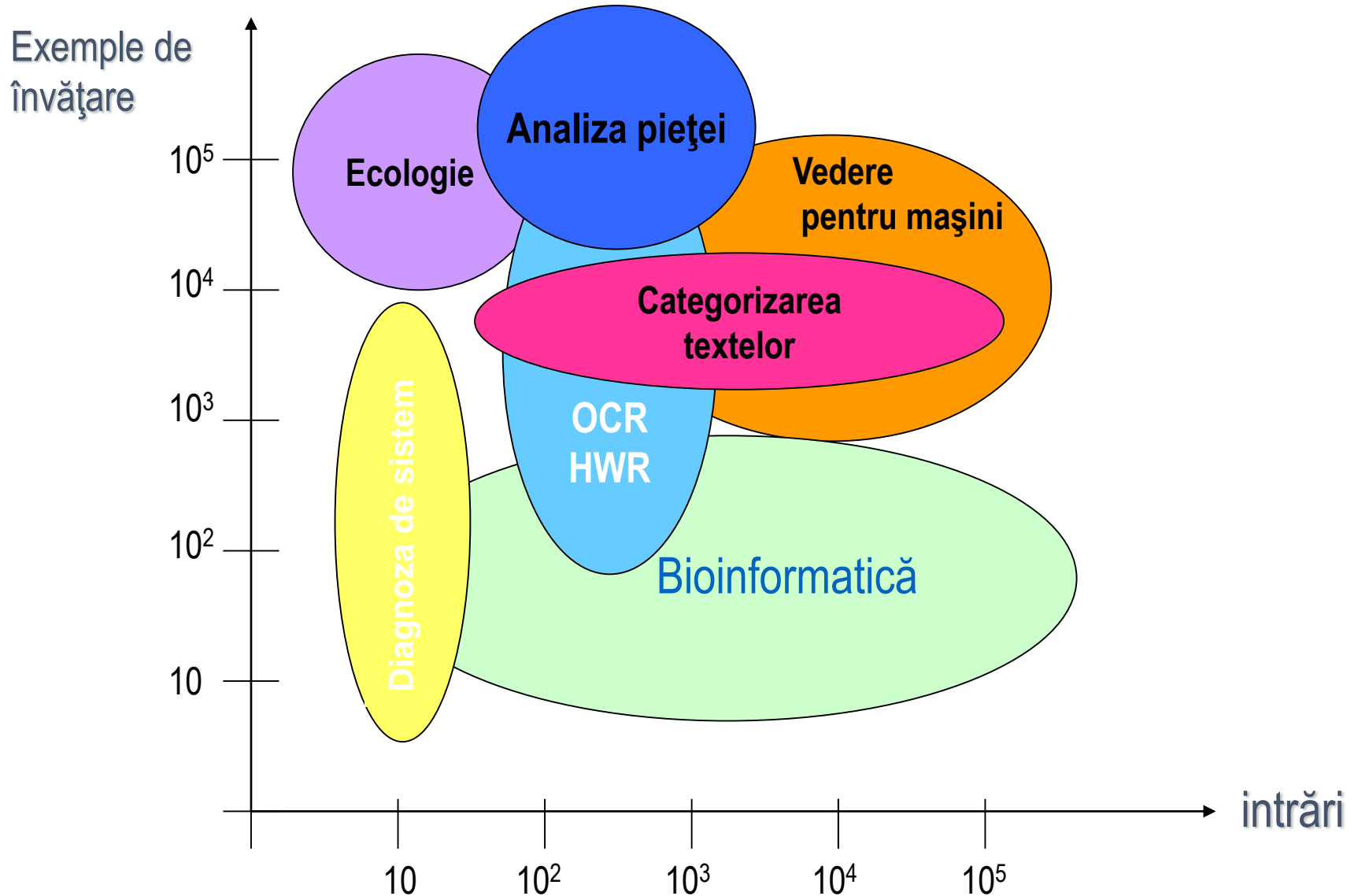


# Clasificarea formelor

## Dificultăți-exemple



# Aplicații



# Unde suntem?

## Astăzi:

- Algoritmi de primă generație: rețele neuronale, arbori de decizie, regresie....

## Mâine:

- Învățarea folosind date complexe (mixed-media) și baze de date multiple
- Învățare prin experimentare activă (active learning)
- Învățarea mai degrabă a deciziilor decât a prognozelor
- Învățare cumulativă, din mai multe surse (transfer, multi-task learning)