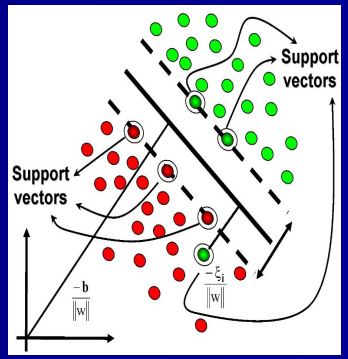
INTRO TO DATA MINING

-- AND RELATED CONCEPTS --

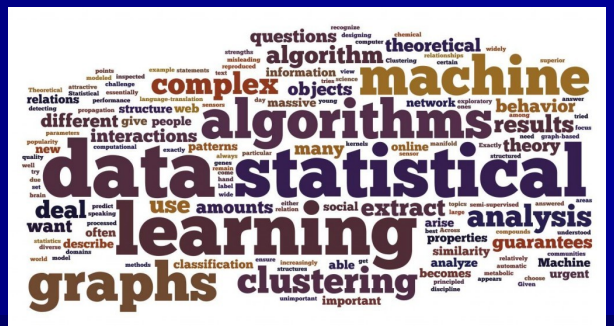
PRINCIPAL METHODS AND SUCCESSFUL APPLICATIONS

Iñaki Inza
Intelligent Systems Group, www.sc.ehu.es/isg
Computer Science Faculty
University of the Basque Country, Donostia - San Sebastian



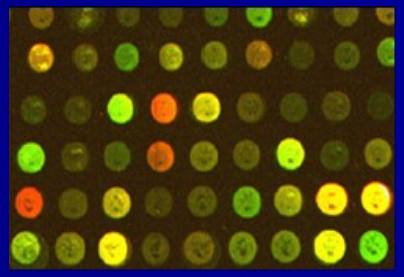
OUTLINE

- Intro to data science
- Main machine learning techniques
- Type of data matrix → Type of data analysis
- Data visualization
-
- Real-life applications for each type of analysis
- Other resources: business oportunities, big data, software tools



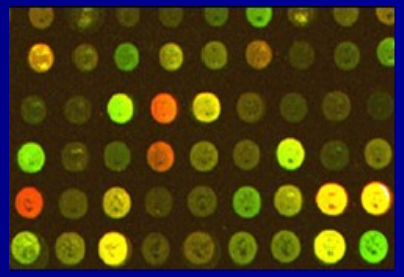
THE AGE OF DATA

- Data collected and stored at enormous speeds (GB/hour). **Data collections-floods** are being collected and warehoused:
  - social networks' activity
  - electronic purchases and transactions
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data (bioinformatics)
  - ...

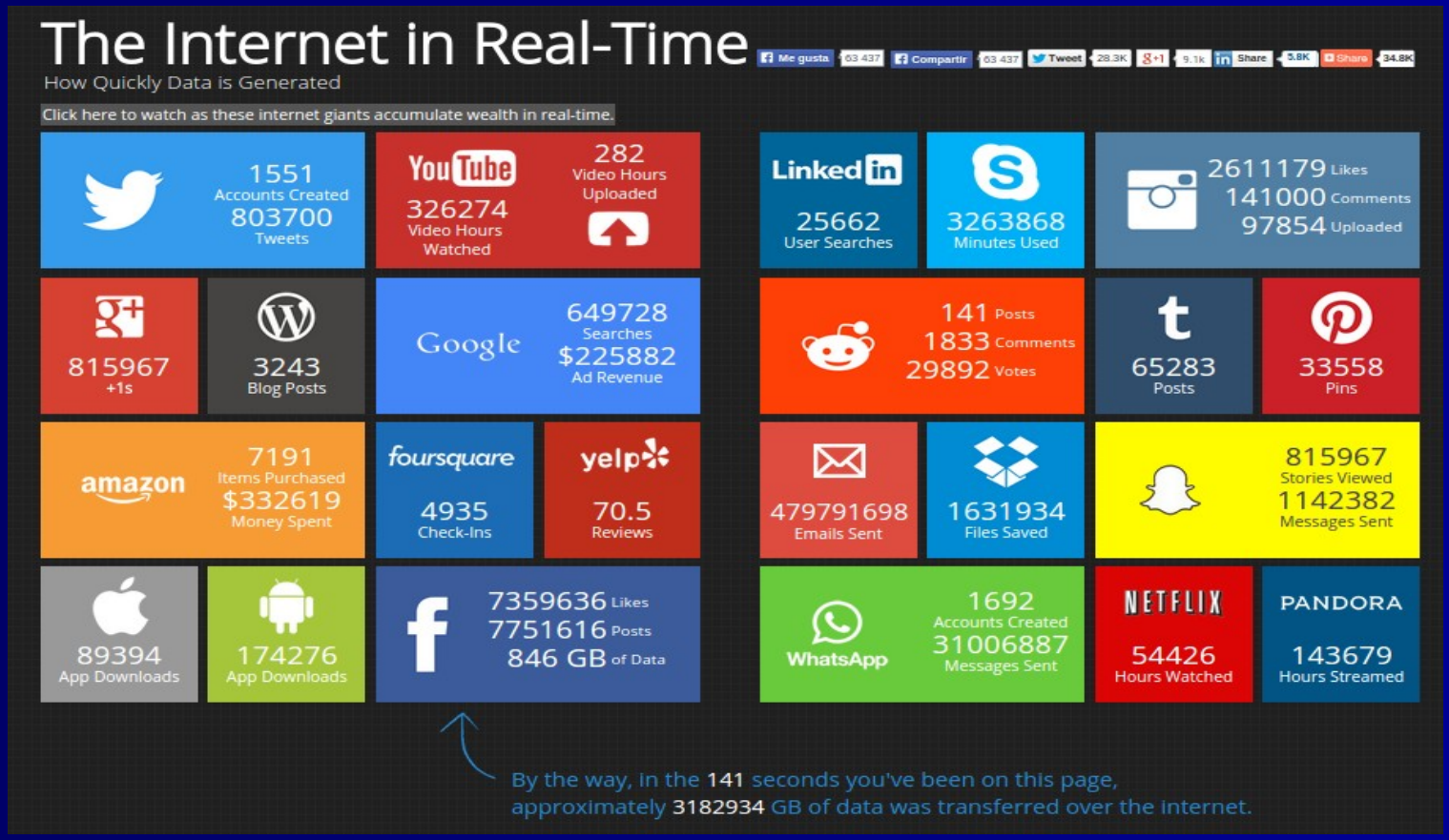


THE AGE OF DATA

- Computers and storage systems have become **cheaper** and more **powerful**
- Since 90's, much more data is being stored **than** analyzed (around 5-10%)
- "Data tsunami": in 2010 enterprises stored 7 exabytes = 7,000,000,000 GB
- **Traditiona**l data analysis techniques **unfeasible** for raw data



THE AGE OF DATA



THE AGE DATA SCIENCE

THE AGE DATA SCIENCE

Data Science is a team sport:
(the whole is greater than sum of its parts)

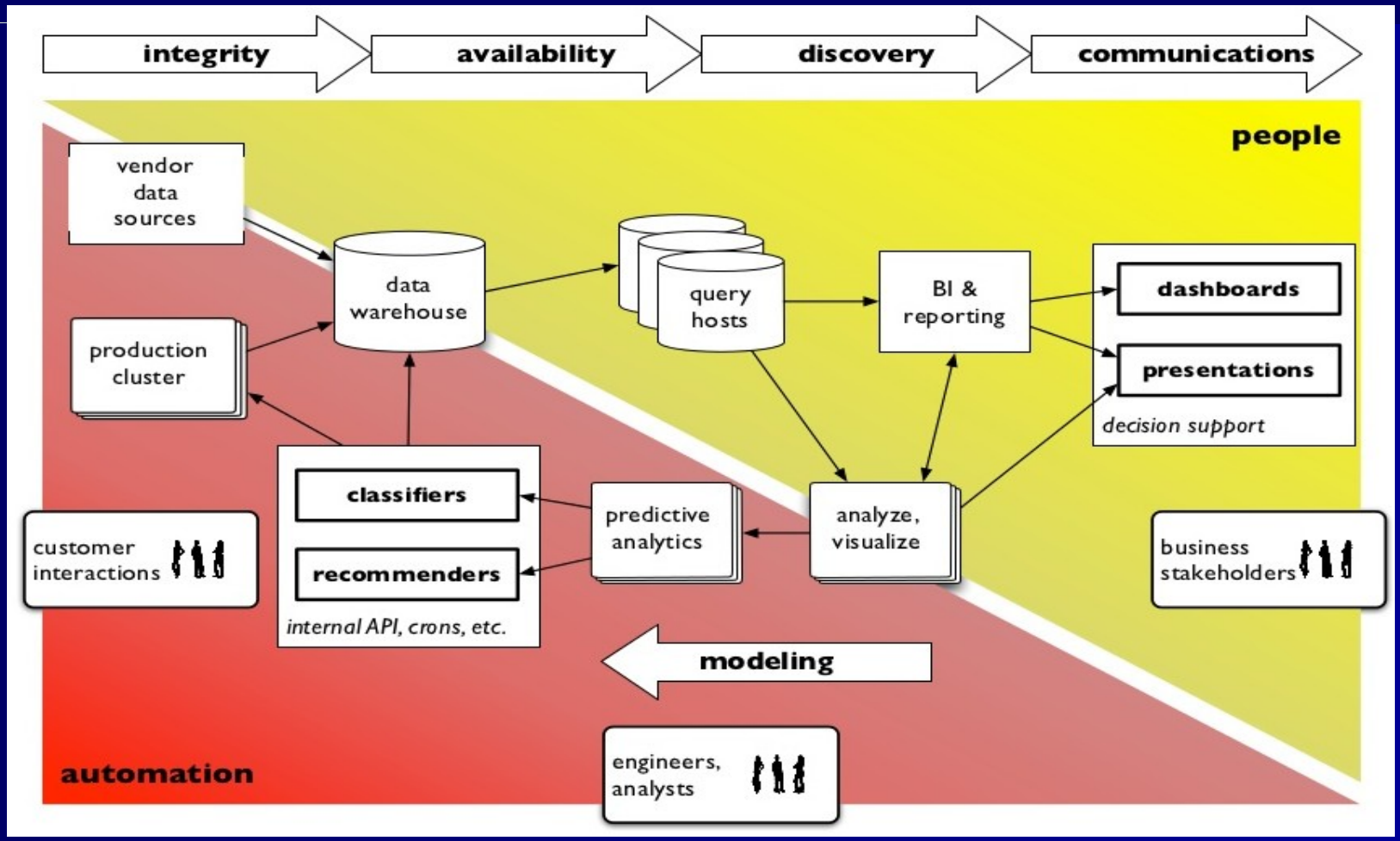
@KirkDBorne http://www.boozallen.com/datascience Booz Allen

Data Science is TRANSDISCIPLINARY Science!

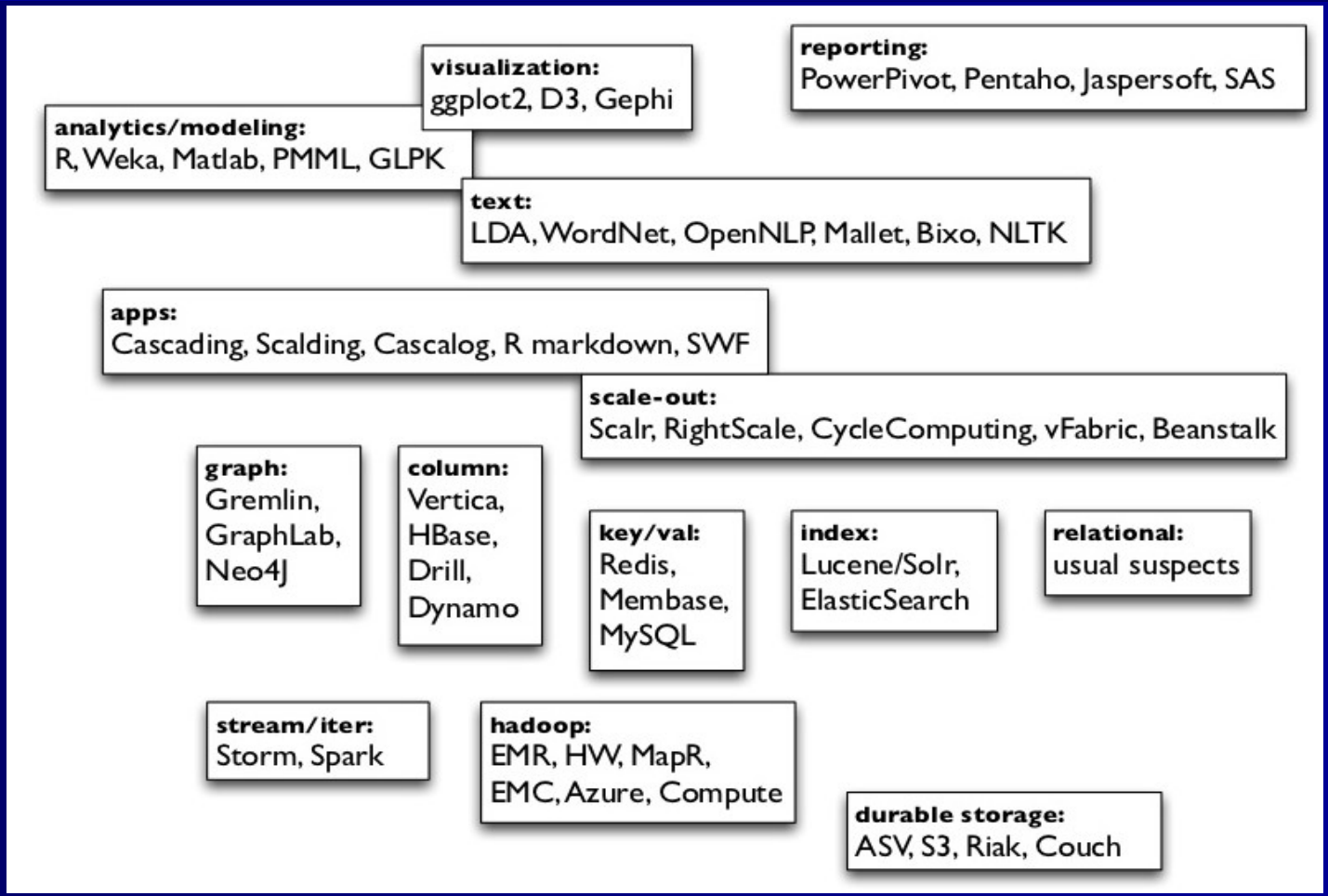
It is the collection of mathematical, computational, scientific, and domain-specific methods, tools, and algorithms that transcend discipline boundaries, applied to Big Data for discovery, decision support, innovation, and data-to-knowledge transformation:

- Advanced Database / Data Management & Data Structures
- Smart Metadata for Indexing, Search, & Retrieval
- Data Mining (Machine Learning) & Analytics (KDD)
- Statistics and Statistical Programming
- Data & Information Visualization
- Network Analysis and Graph Mining (Everything is a graph!)
- Semantics (Natural Language Processing, Ontologies)
- Data-intensive Computing (e.g., Hadoop, Cloud, …)
- Modeling & Simulation (computational data science)
- Domain-Specific Data Analysis Tools

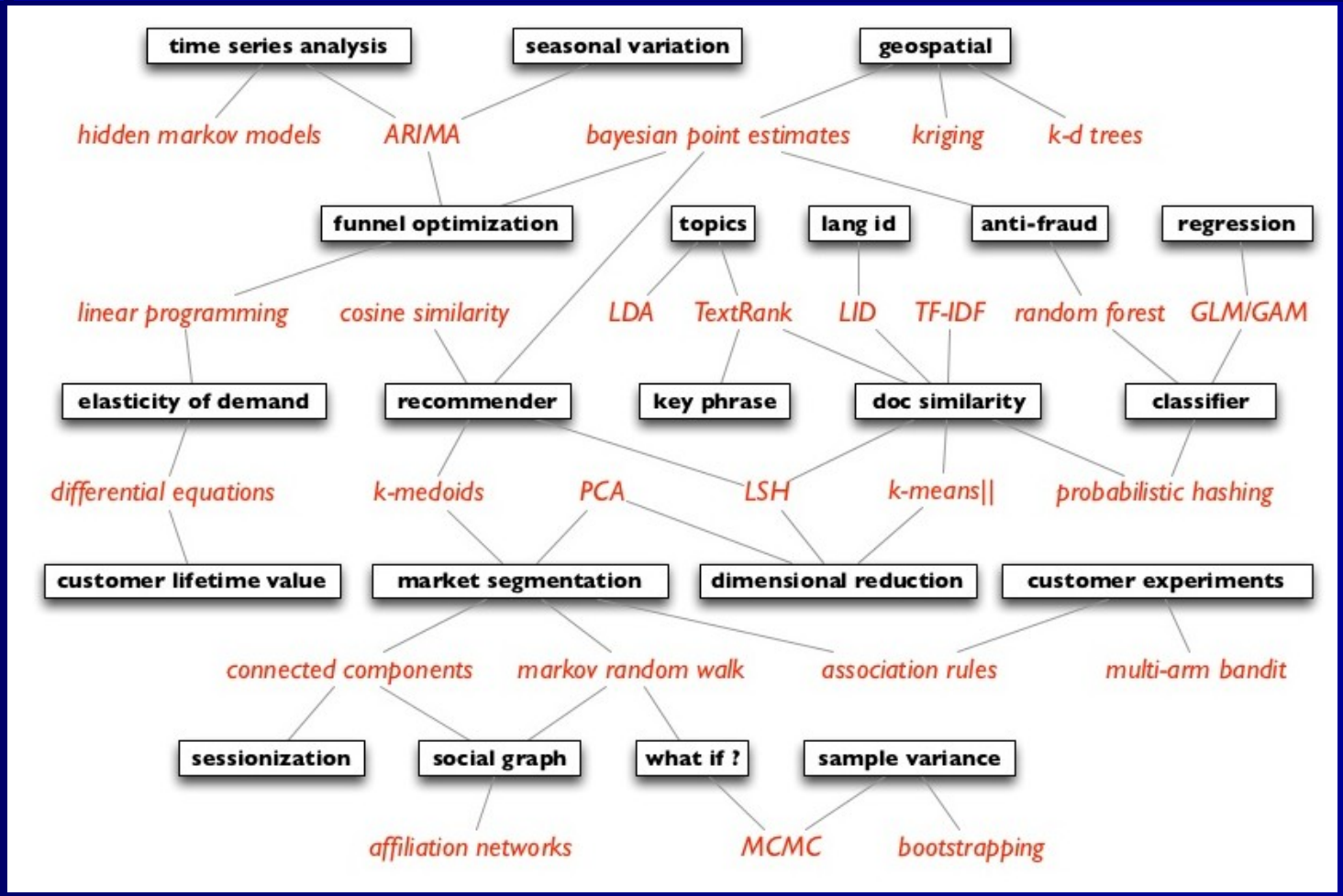
THE AGE DATA SCIENCE



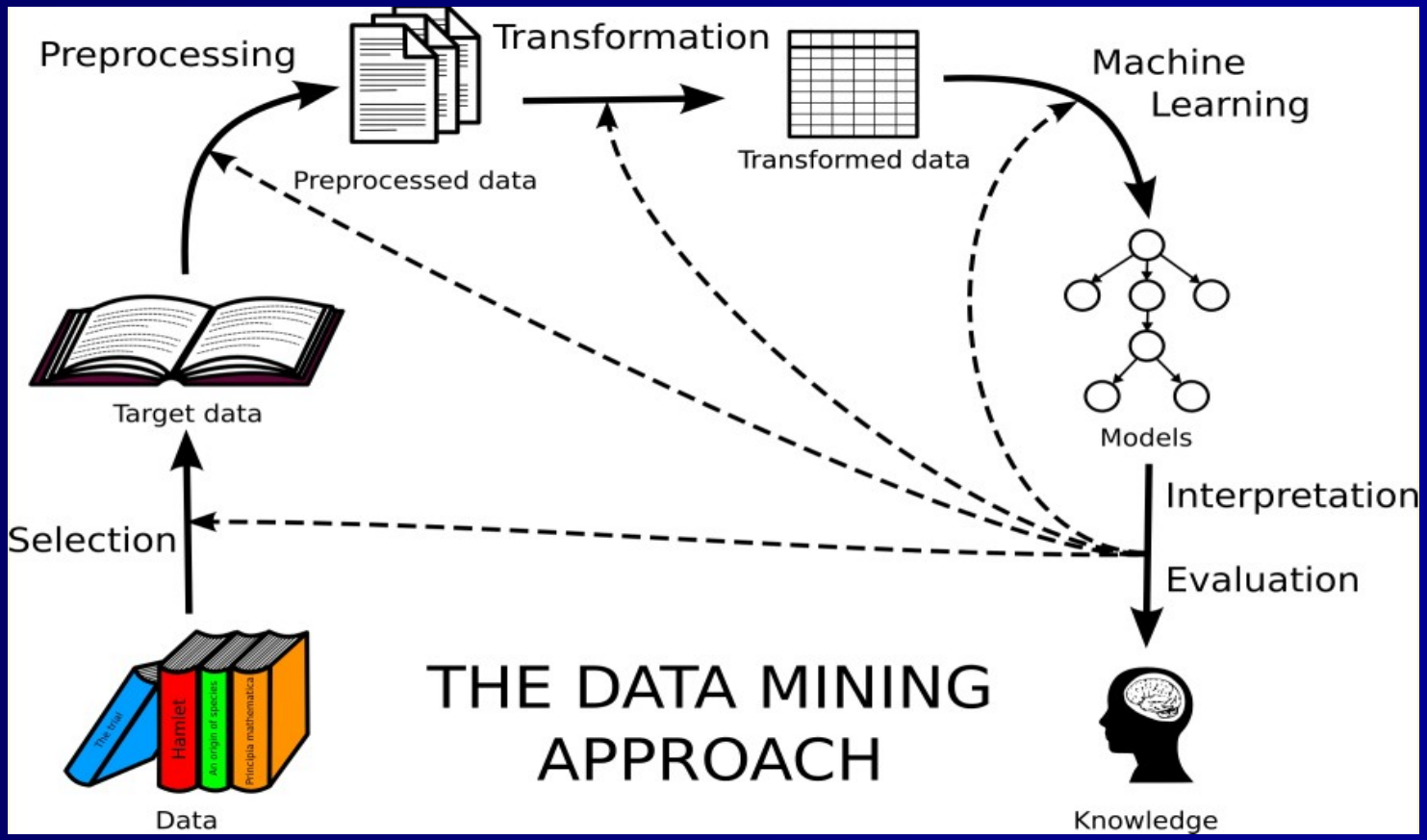
THE AGE DATA SCIENCE



THE AGE DATA SCIENCE



DATA MINING PROCESS: THE PIPELINE
KDD: KNOWLEDGE DISCOVERY IN DATABASES

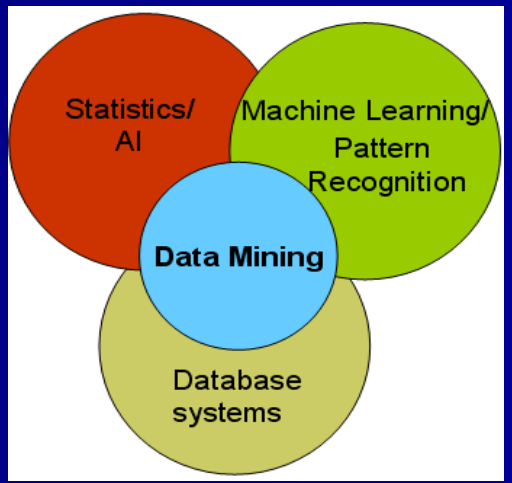


DATA MINING ROOTS

- **Traditional statistical techniques may be unsuitable to deal with:**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous**
  - **Data acquisition:**

    - Data Mining: may not be connected

    - Statistics: answers to specific questions



DEFINITION: DATA MINING

Definition (Fayyad et. al): The non-trivial discovery of *novel, valid , comprehensible* and potentially *useful* patterns from data.

What is a pattern? A relationship in the data. E.g.,

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data.

– *Data Mining* by Witten and Frank

- Technologies for analysis of data and discovery of (very) hidden patterns
- Uses a combination of statistics, probability analysis and database technologies
- Fairly young (<20 years old) but clever algorithms developed through database research

Data mining, also popularly referred to as *knowledge discovery in databases (KDD)*, is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, and other massive information repositories.

– *Data Mining: Concepts and Techniques* by Han and Kamber

- A one-page intro to Data Mining

BUSINESS INTELLIGENCE

- A new “culture” around the data management in the company
- Saving, analysis, sharing
- Objective: help the company in its decisions

BUSINESS OPORTUNITIES

Big Data, Big Impact:
New Possibilities for International Development

Executive Summary

A flood of data is created every day by the interactions of billions of people using computers, GPS devices, cell phones, and medical devices. Many of these interactions occur through the use of mobile devices being used by people in the developing world, people whose needs and habits have been poorly understood until now. Researchers and policymakers are beginning to realise the potential for channelling these torrents of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises for the benefit of low-income populations. Concerted action is needed by governments, development organisations, and companies to ensure that this data helps the individuals and communities who create it.

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

BUSINESS OPORTUNITIES

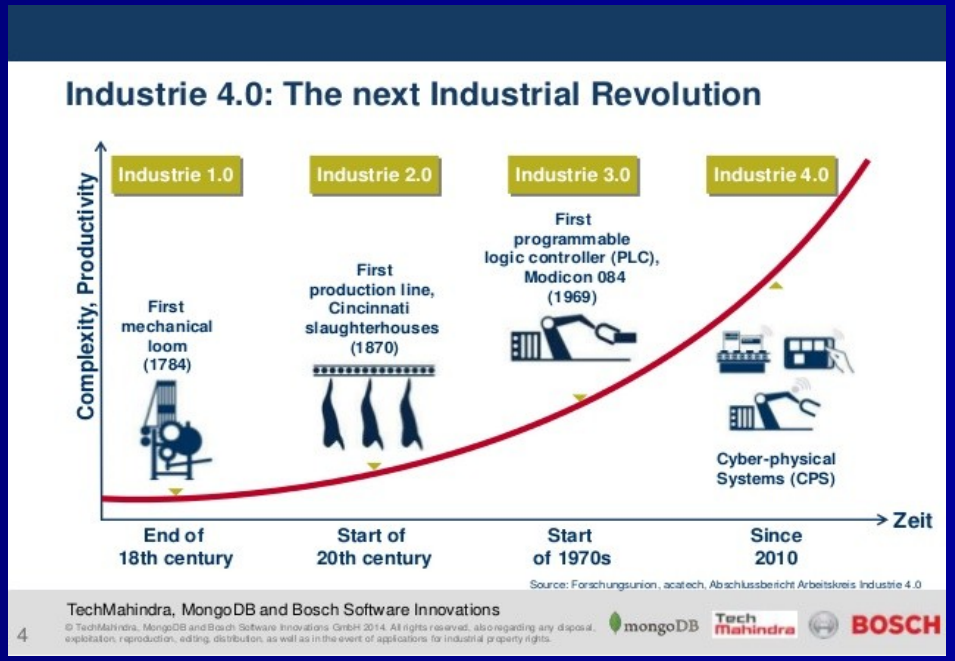
 **Online Product Sales** Predict the online sales of a consumer product based on a data set of product features.	$22,500	365	14 months ago
Boehringer Ingelheim **Predicting a Biological Response** Predict a biological response of molecules from their chemical properties	$20,000	703	14 months ago
Ford **Stay Alert! The Ford Challenge** Driving while not alert can be deadly. The objective is to design a classifier that will detect whether the driver is alert or not alert, employing data that are acquired while driving.	$950	176	2 years ago

World Economic Forum nominates kaggle.com as one of the 2014 economic pioneers [link]

kaggle.com

INDUSTRY 4.0

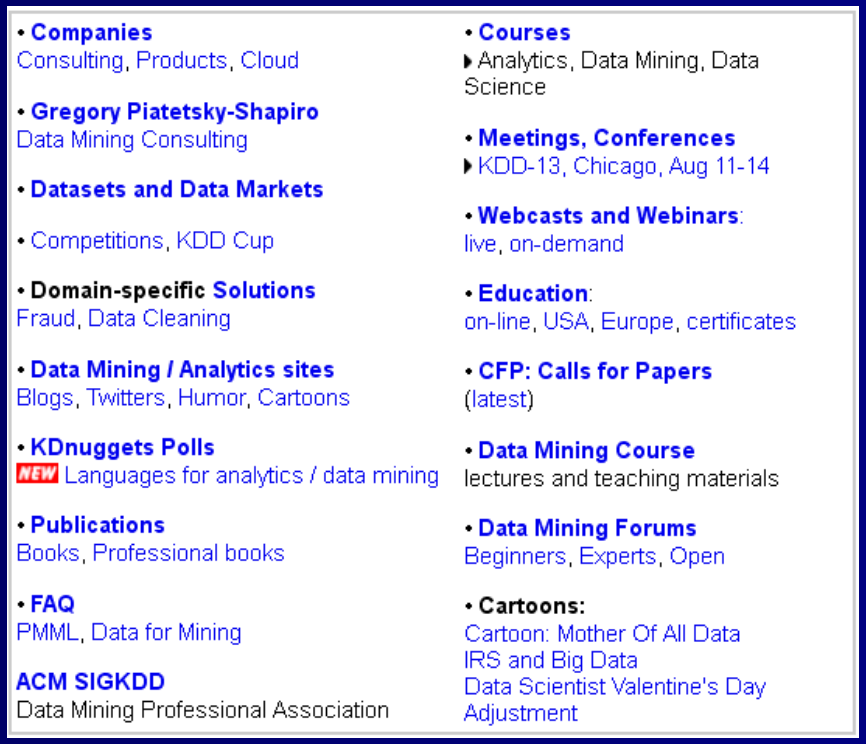
- 4th industrial revolution
- “Smart factories”
- Role of new technologies
- Role of big data and analytics
- “Basque Industry 4.0”

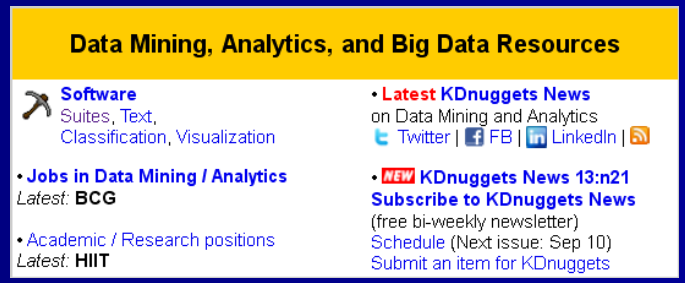


EMPLOYMENT OPORTUNITIES

- **Shortage by 2018 in the USA of 140-190,000 people with deep analytical skills**
- **1.5 M managers with the know-how to interpret the analysis of big data to make effective decisions (mckinsey.com)**
- **LinkedIn skills:**
  - **~ 75,000 users with Data Mining skill**
  - **~ 7,000 users with Predictive Modeling skill**

KEY RESOURCE
DATA MINING IN BUSINESS





Top 10 Data Analysis Tools for Business [link]

GLOBAL PULSE

- Iniciativa Naciones Unidas
- Datos: móviles, meteo, etc.

- Uso de este tipo de tecnologías:
  - emergencias humanitarias
  - situaciones de riesgo
  - pandemias, epidemias
  - desarrollo sostenible

- Respuesta inmediata
- Respuesta global

OPEN DATA GOVERNMENT

Data: Government, State, City, Local and Public

This is a directory of government, federal, state, city, local and other public datasets. See also Data APIs, Hubs, Marketplaces, Platforms, Portals, and Search Engines.

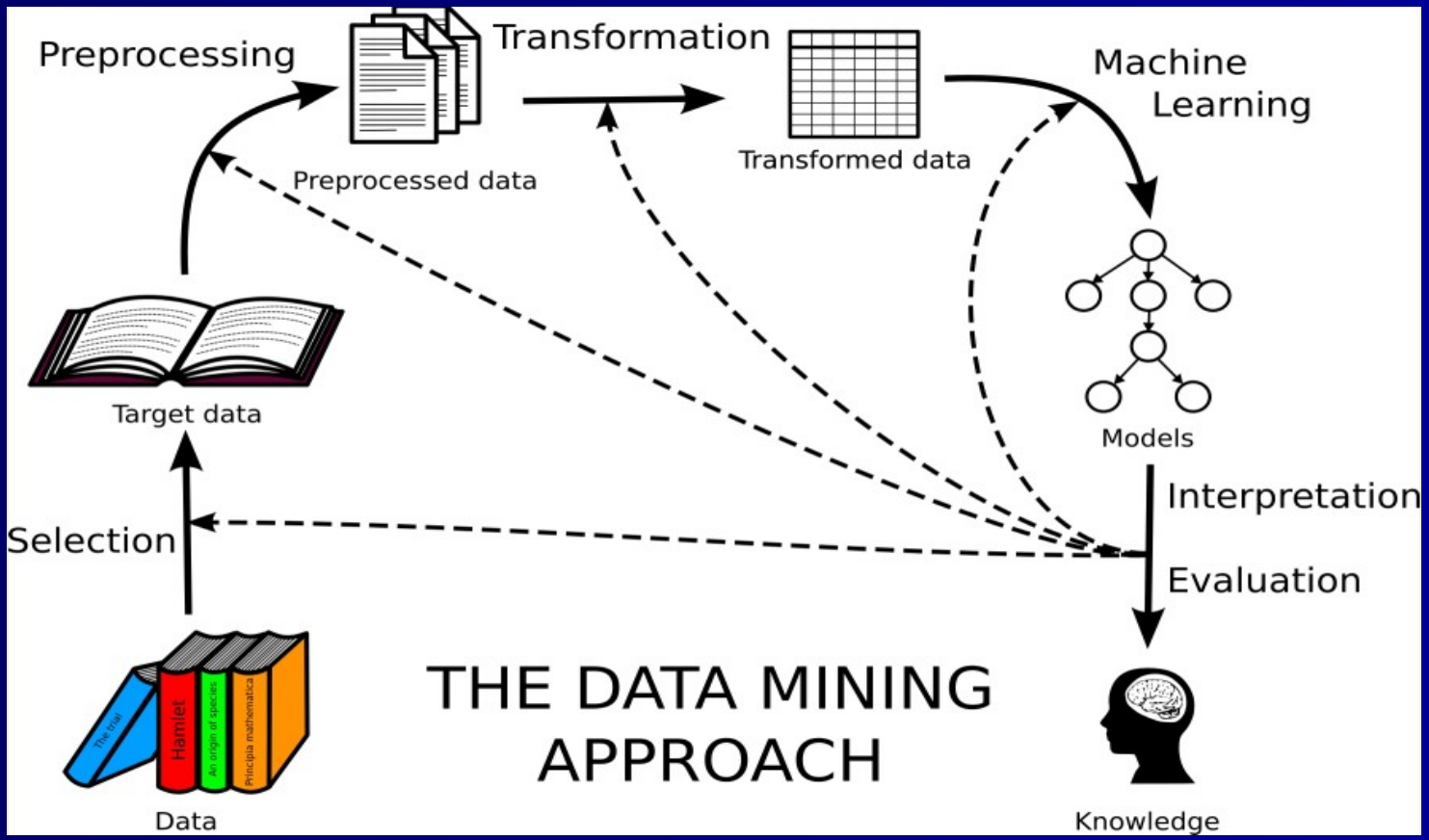
Portals | Global | USA | Canada | Europe | Asia | Australia, NZ and Pacific | Latin America | Africa | Middle East

Public data catalogs, portals, and services

- AWS (Amazon Web Services) Public Data Sets, provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications.
- Datacatalogs.org, open government data from US, EU, Canada, CKAN, and more.
- DataMarket, visualize the world's economy, societies, nature, and industries, with 100 million time series from UN, World Bank, Eurostat and other important data providers.
- datamob, Public data put to good use.
- Enigma, "Google for public data", provides easy access to government, NGO, and other public domain datasets.
- Freebase, a community-curated database of well-known people, places, and things.
- Google Public Data, with dynamic visualization and exploration tools.
- Knoema World Data Atlas, over 1000 indicators on all countries
- National Government Statistical Web Sites, data, reports, statistical yearbooks, press releases, and more from about 70 web sites, including

DATA MINING PROCESS: THE PIPELINE

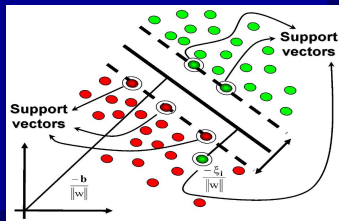
KDD PROCESS: KNOWLEDGE DISCOVERY IN DATABASES



DATA MINING: MAIN TASKS

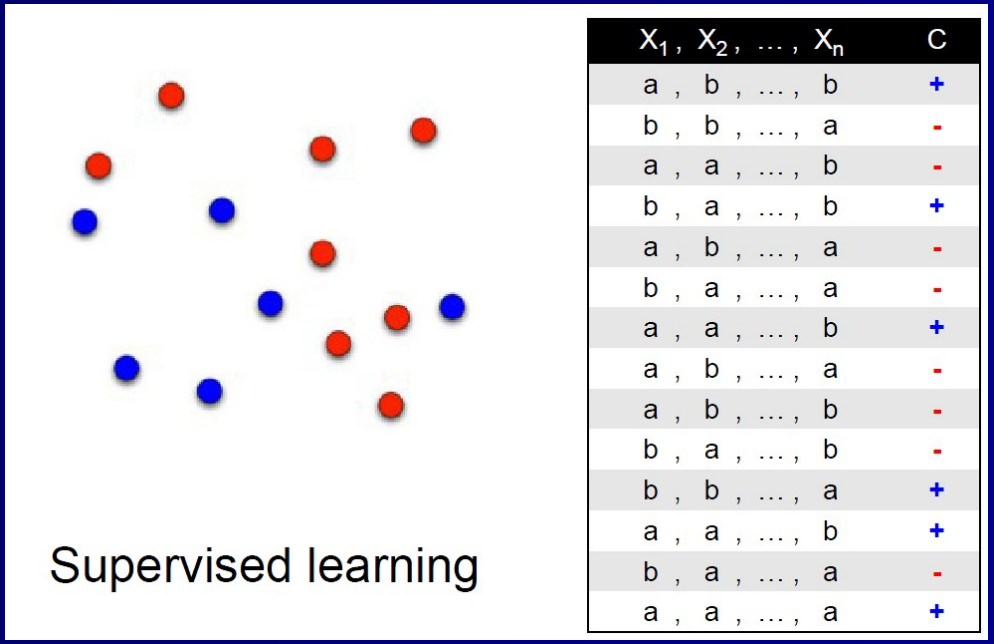
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables
    - *Supervised classification: nominal variable to be predicted*
    - *Regression: quantitative variable to be predicted*

- Description Methods
  - Find human-interpretable patterns that describe the data
    - *Clustering – unsupervised classification*
    - *Association rule discovery*
    - *Feature selection: discover the key predictors*
    - *Outlier detection*



SUPERVISED CLASSIFICATION

- Given a collection of records-samples (*training set*)
  - Each record contains a set of *attributes-features-predictors*
  - Each record belongs to a *class, our variable of interest (variable to be predicted)*

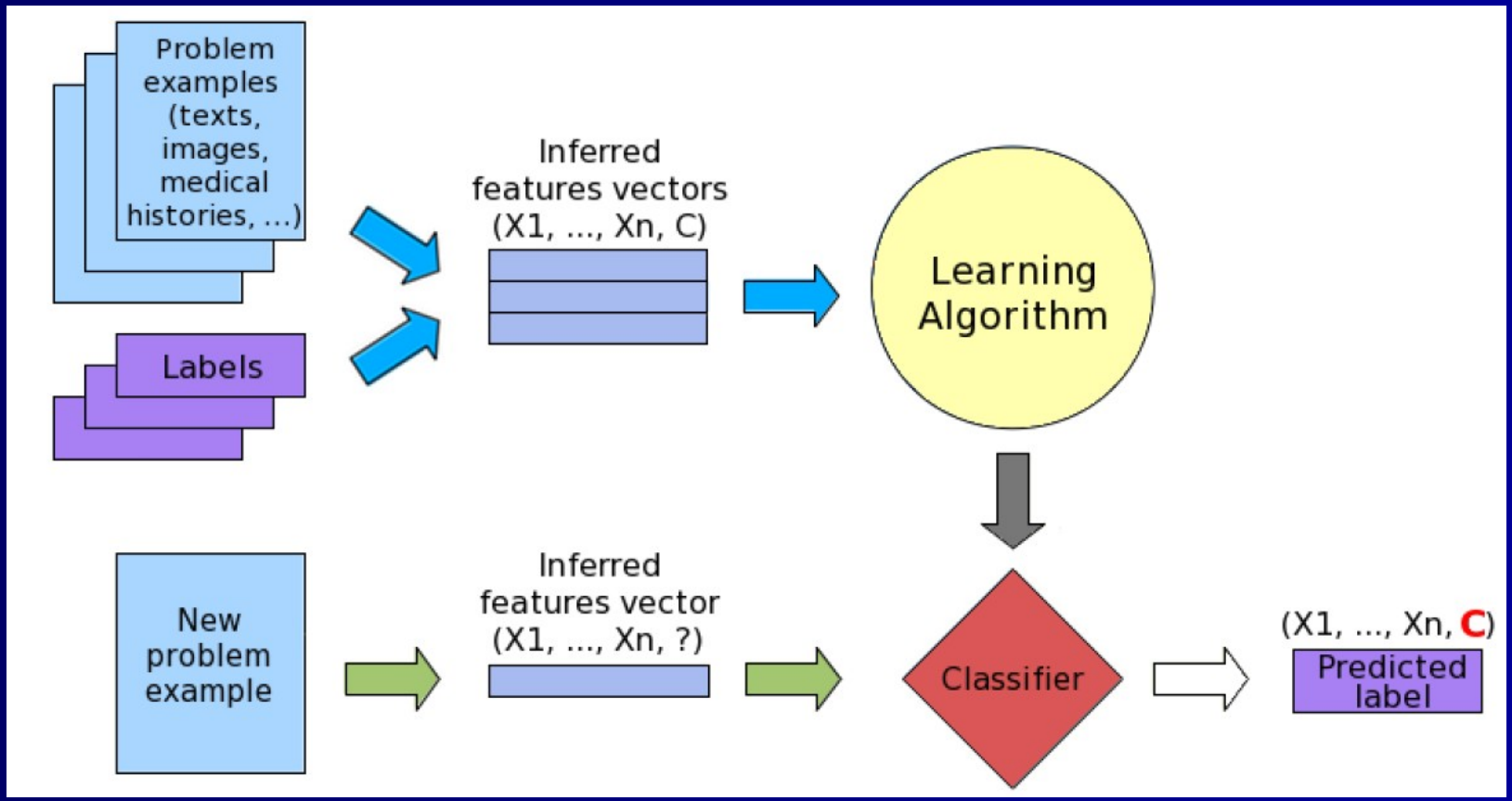


X_1 , X_2 , ... , X_n	C
a , b , ... , b	+
b , b , ... , a	-
a , a , ... , b	-
b , a , ... , b	+
a , b , ... , a	-
b , a , ... , a	-
a , a , ... , b	+
a , b , ... , a	-
a , b , ... , b	-
b , a , ... , b	-
b , b , ... , a	+
a , a , ... , b	+
b , a , ... , a	-
a , a , ... , a	+

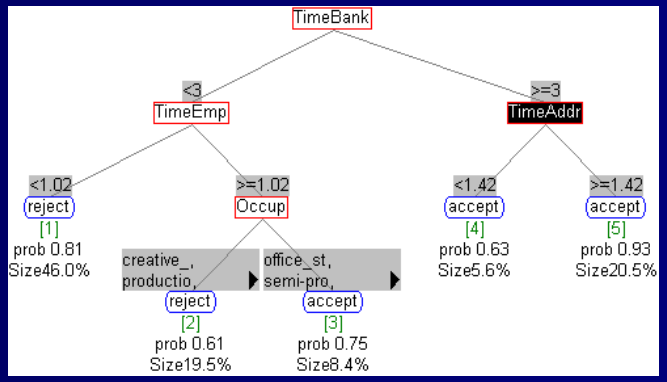
SUPERVISED CLASSIFICATION

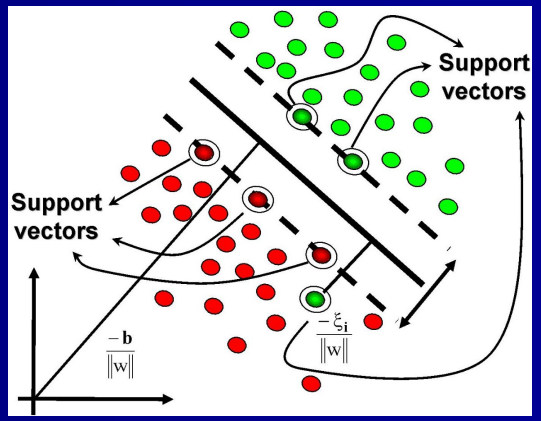
- Find a *model* for class attribute as a function of the values of other attributes. There is a broad range of model types:
  - Decision trees, Bayesian networks, neural networks…
- Goal: previously unseen records should be assigned a class as accurately as possible
  - A *test set* is used to *estimate the accuracy* of the model. There is a broad range of techniques for accuracy estimation: cross-validation, hold-out, bootstrap…

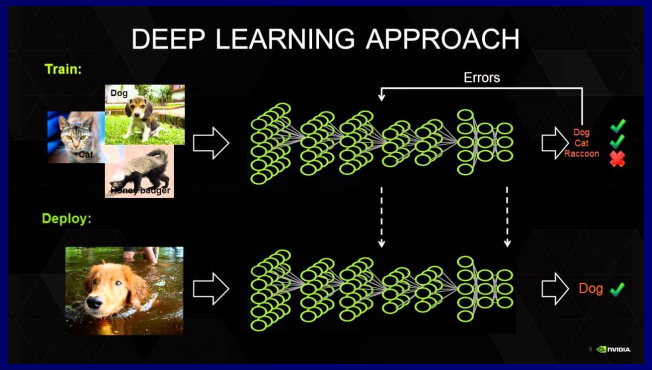
SUPERVISED CLASSIFICATION: the standard scenario

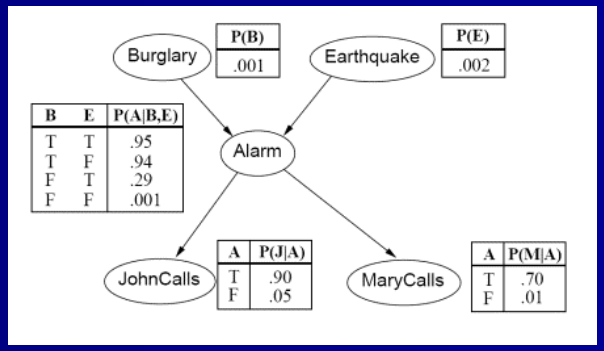


SUPERVISED CLASSIFICATION: models

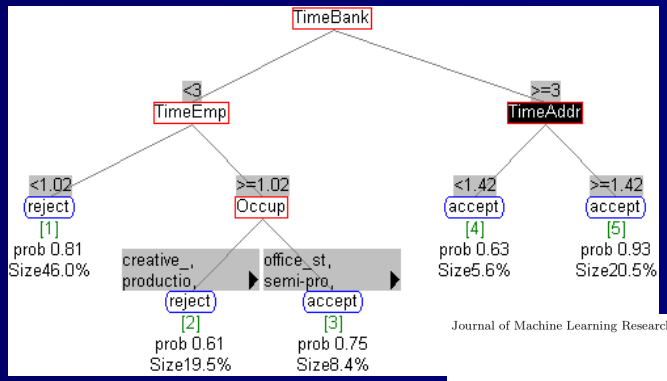


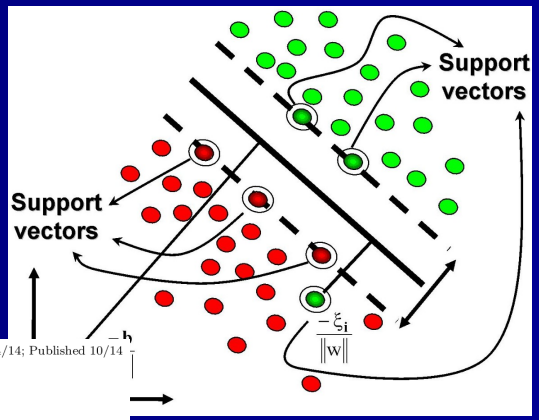






SUPERVISED CLASSIFICATION: models



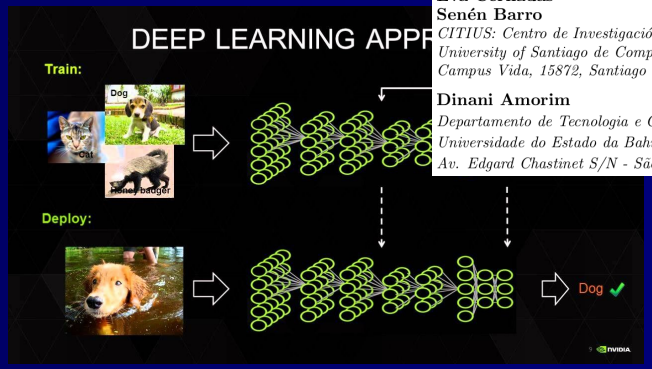


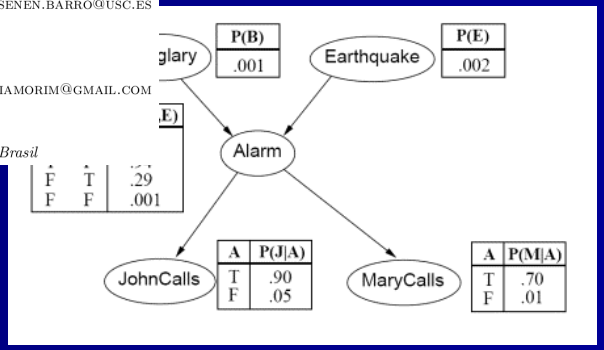
Journal of Machine Learning Research 15 (2014) 3133-3181 Submitted 11/13; Revised 4/14; Published 10/14

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

Manuel Fernández-Delgado MANUEL.FERNANDEZ.DELGADO@USC.ES
Eva Cernadas EVA.CERNADAS@USC.ES
Senén Barro SENEN.BARRO@USC.ES
CITIUS: Centro de Investigación en Tecnoloxías da Información da USC
University of Santiago de Compostela
Campus Vida, 15872, Santiago de Compostela, Spain

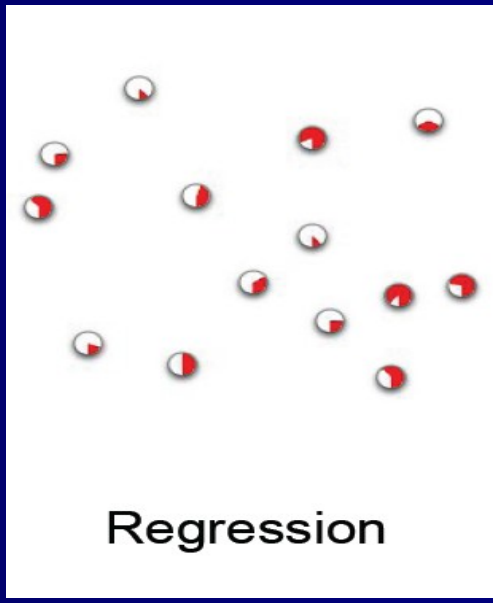
Dinani Amorim DINANIAMORIM@GMAIL.COM
Departamento de Tecnologia e Ciências Sociais- DTCS
Universidade do Estado da Bahia
Av. Edgard Chastinet S/N - São Geraldo - Juazeiro-BA, CEP: 48.305-680, Brasil





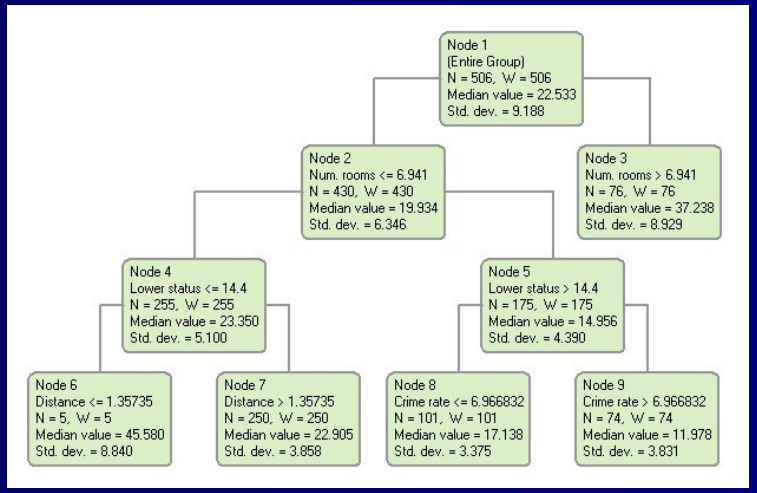
REGRESSION

- Given a collection of records-samples (*training set*)
- Each record is characterized by a set of *attributes-features-predictors*
- The *variable of interest* to be predicted is *quantitative*



X_1 , X_2 , … , X_n	C
a , b , … , b	1.1
b , b , … , a	0.1
a , a , … , b	6.3
b , a , … , b	2.0
a , b , … , a	3.3
b , a , … , a	2.1
a , a , … , b	5.5
a , b , … , a	4.0
a , b , … , b	0.5
b , a , … , b	5.6
b , b , … , a	4.3
a , a , … , b	2.2
b , a , … , a	1.2
a , a , … , a	3.1

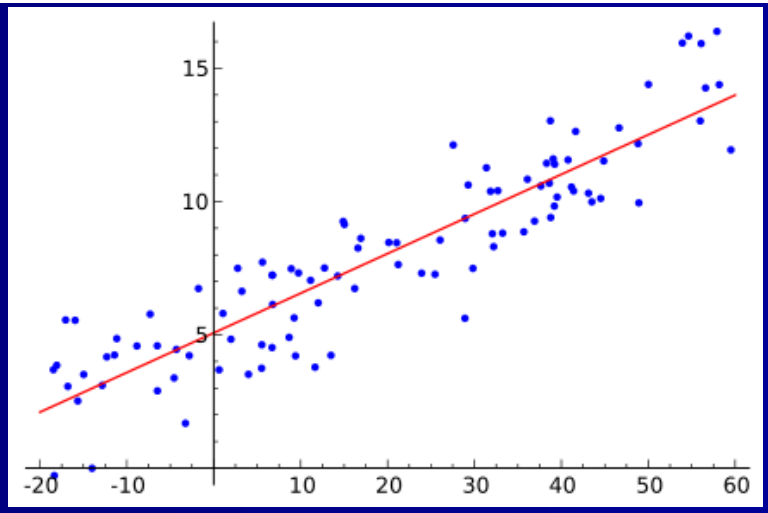
REGRESSION: models



Regression trees

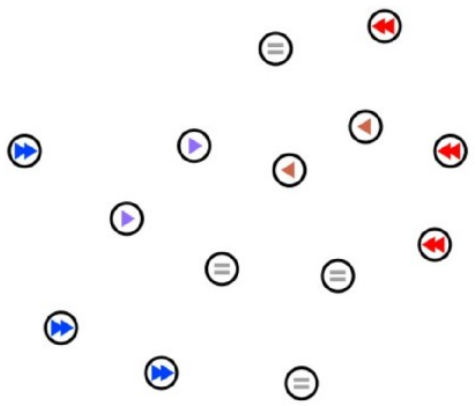
$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Linear regression



ORDINAL CLASSIFICATION

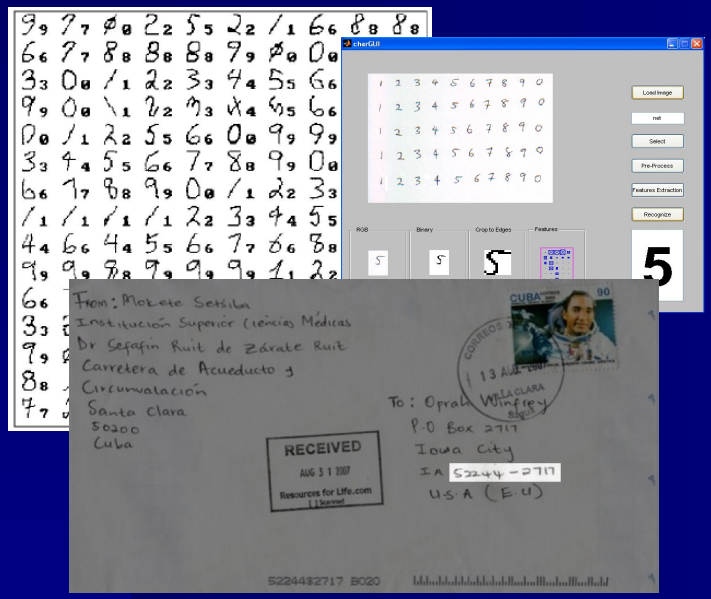
- The *variable of interest* to be predicted is *discrete, but ordered*

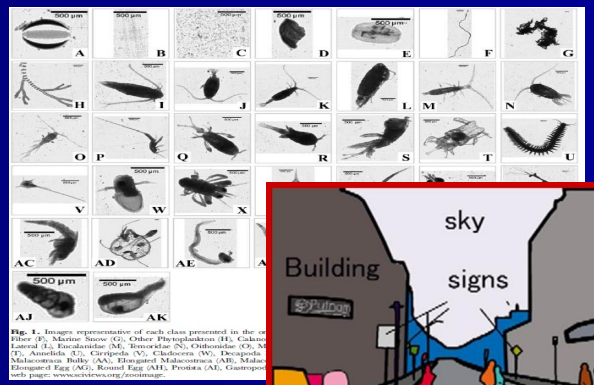


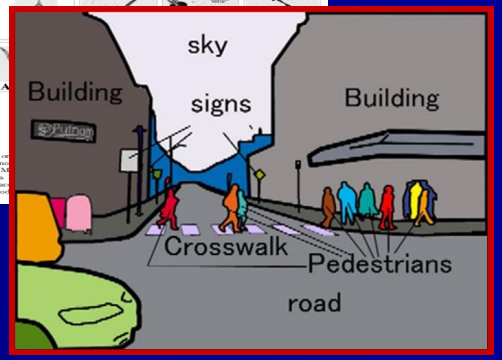
Ordinal learning

X_1 , X_2 , ... , X_n	C
a , b , ... , b	<<
b , b , ... , a	>
a , a , ... , b	=
b , a , ... , b	<
a , b , ... , a	>>
b , a , ... , a	>
a , a , ... , b	<
a , b , ... , a	=
a , b , ... , b	=
b , a , ... , b	<
b , b , ... , a	>>
a , a , ... , b	>
b , a , ... , a	<<
a , a , ... , a	=

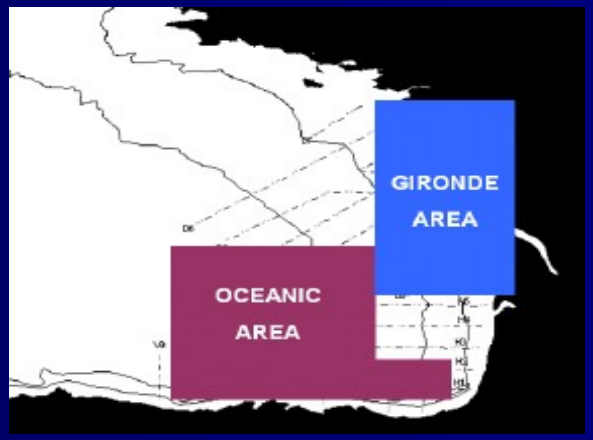
SUPERVISED CLASSIFICATION and REGRESSION: APPLICATIONS PATTERN RECOGNITION

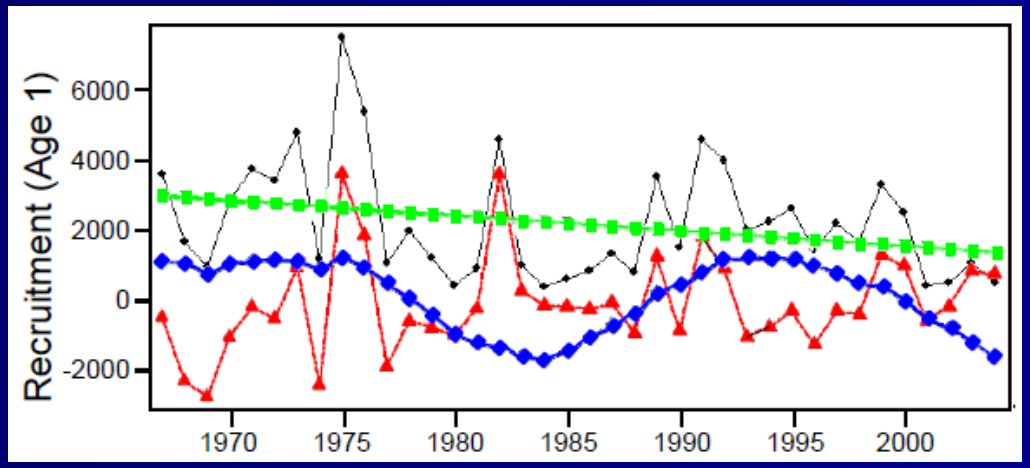






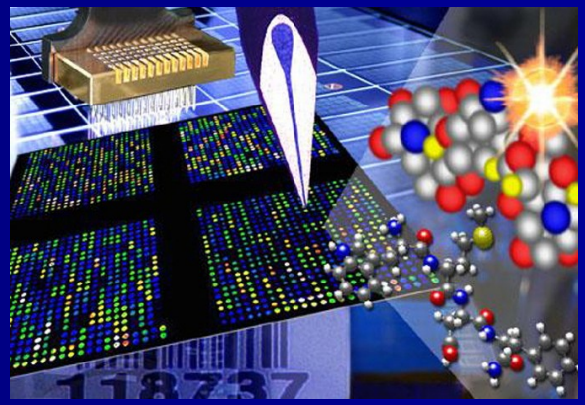
SEA SPECIES BIOMASS PREDICTION





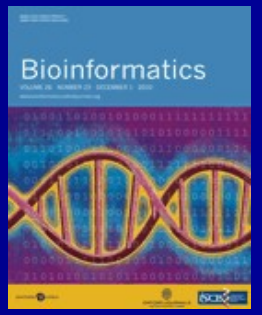
SPAM FILTERING
ANTIVIRUS DETECTION
MALWARE

BIOINFORMATICS
DIAGNOSIS AND PROGNOSIS OF DISEASES
BIOMARKER DISCOVERY



BIOINFORMATICS
DIAGNOSIS AND PROGNOSIS OF DISEASES
BIOMARKER DISCOVERY

Differential Micro RNA Expression in PBMC from Multiple Sclerosis Patients



Oncology
International Journal for Cancer Research and Treatment

Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues

Ensemble machine learning on gene expression data for cancer classification

nature

Genome Biology

Classification of Alzheimer's Disease and Parkinson's Disease by Using Machine Learning and Neural Network Methods

“FUGA” DE CLIENTES

- Telefonía
- Bancos

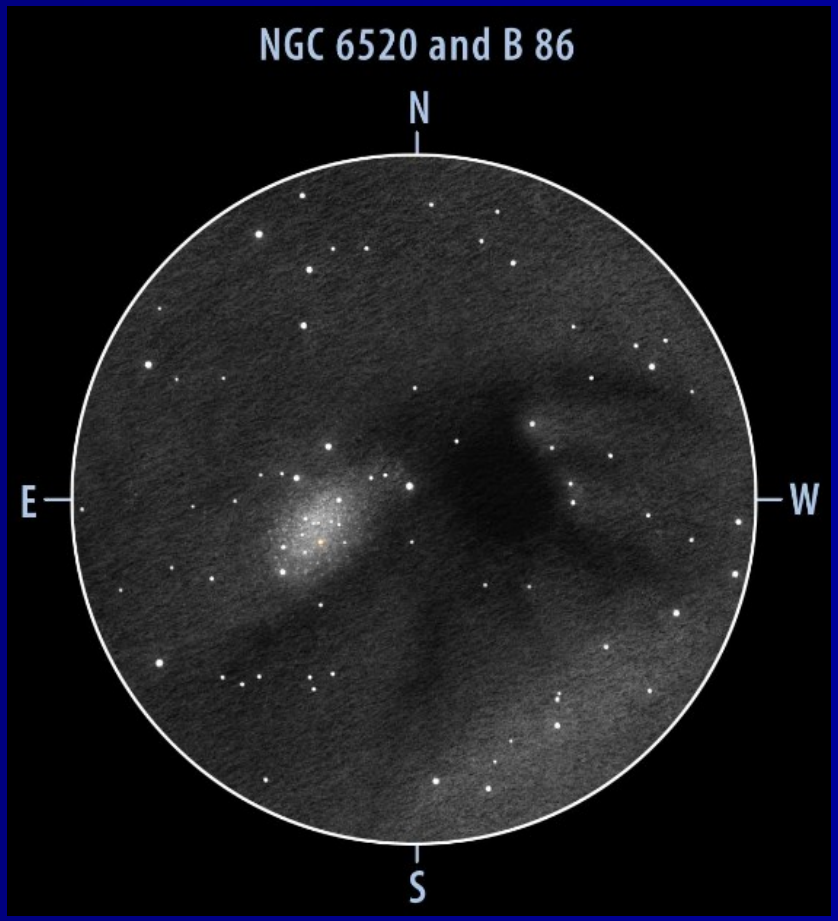
- Según mi “patrón” de llamadas, “quejas”, etc.

- ¿Riesgo de cambiar de operadora?

- ¿Ofertar?

CATALOGING SKY OBJECTS

NASA Machine Learning Group



PREDICCIÓN DE FALLOS EN MÁQUINAS

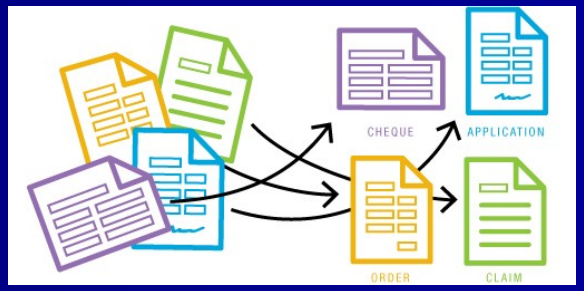
- Problema en Euskadi
- Fabricantes de máquina herramienta
- “Industry 4.0”

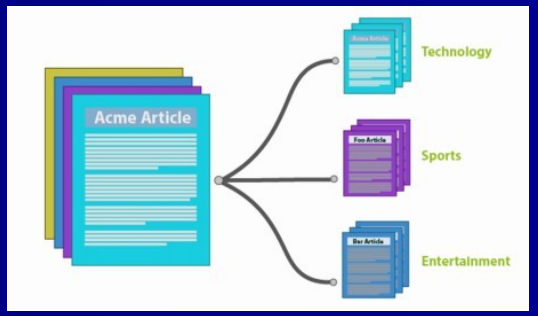
- Máquinas: sensórica

- Predicción: ¿cuándo puede fallar la máquina?
- “Systems fault diagnosis”
- “Mantenimiento proactivo”

CLASIFICACIÓN DE DOCUMENTOS

- “Natural Language Processing” (NLP)
- Tema - categoría
- Dificultad del texto
- Género autor

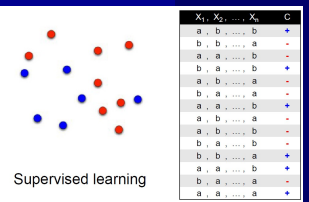


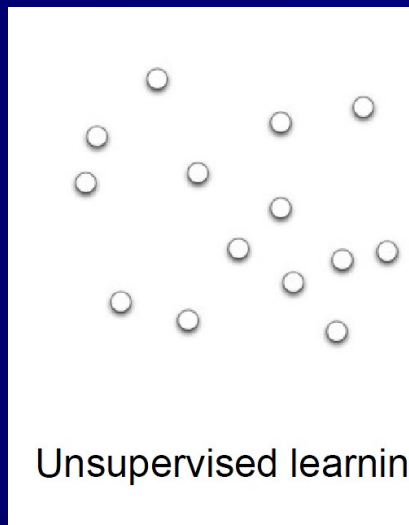


UNSUPERVISED CLASSIFICATION
CLUSTERING

- Given a collection of records-samples (*training set*)
  - Each record contains a set of *attributes-features-predictors*
  - *No "target feature" (class)* which supervises the learning process

- Find groups of cases with:
  - Large intra-group homogeneity: clustering similar samples
  - Large inter-groups heterogeneity

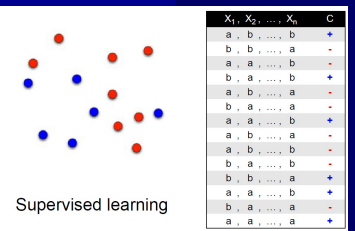




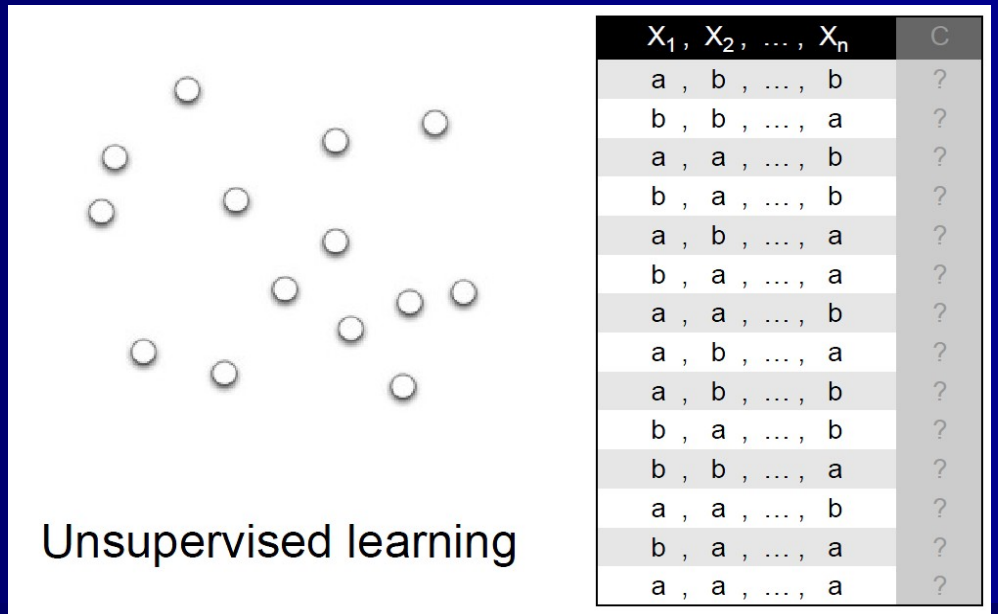
X_1 , X_2 , … , X_n	C
a , b , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , b	?
a , b , … , a	?
b , a , … , a	?
a , a , … , b	?
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?
a , a , … , a	?

UNSUPERVISED CLASSIFICATION CLUSTERING

- Difficult evaluation-measure of these properties → no recognition rate
- Number of groups… → deciding before-hand, difficult decision
- “Distance”-”similarity” function: e.g. Euclidean (ordinal), overlap (nominal: a=a→dist=0, a≠b→dist=1, a≠c→dist=1, b≠c→dist=1)



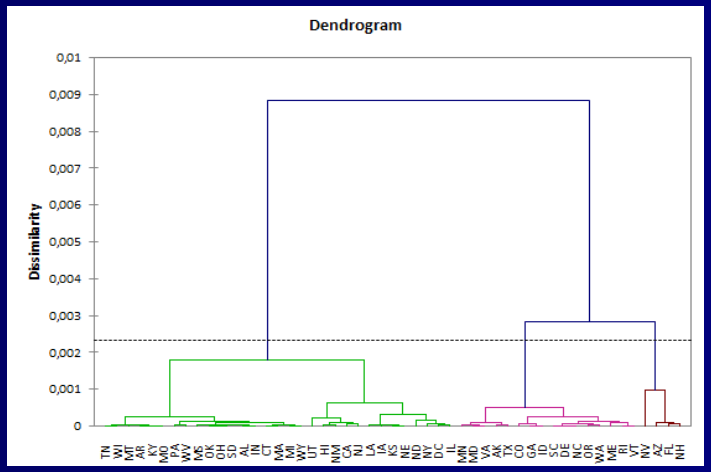
X_1 , X_2 , … , X_n	C
a , b , … , b	+
b , b , … , a	-
a , a , … , b	-
b , a , … , b	+
a , b , … , a	-
b , a , … , a	-
a , a , … , b	+
a , b , … , b	-
b , a , … , b	-
b , b , … , a	+
a , a , … , b	+
b , a , … , a	-
a , a , … , a	+



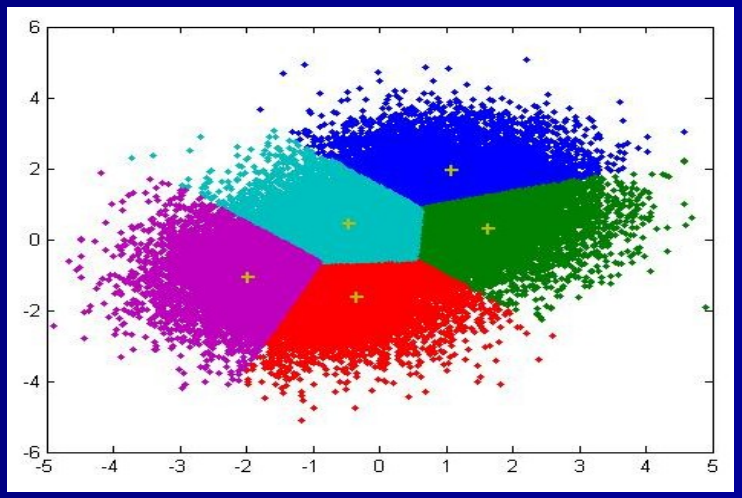
X_1 , X_2 , … , X_n	C
a , b , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , b	?
a , b , … , a	?
b , a , … , a	?
a , a , … , b	?
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?
a , a , … , a	?

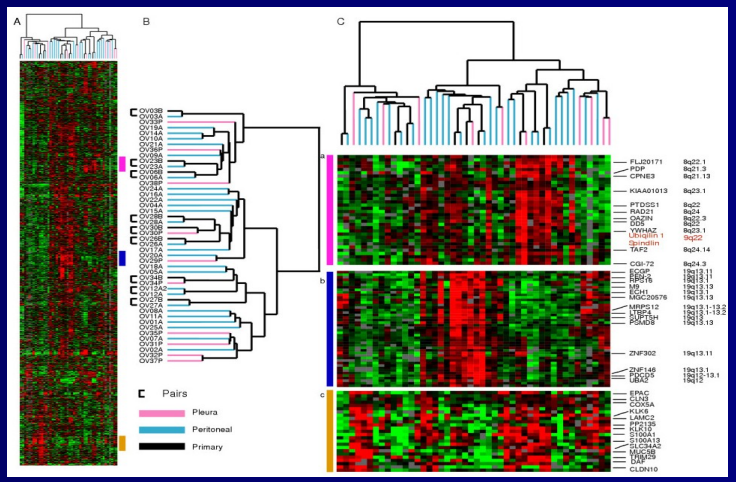
CLUSTERING: MODELS

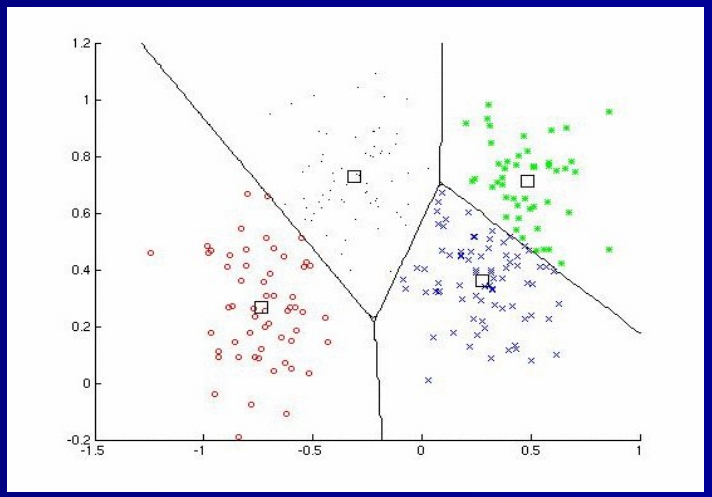
Hierarchical clustering



Partitional clustering (k-means)







CLUSTERING: APPLICATIONS
CUSTOMER SEGMENTATION

- Identify micro-markets and develop policies for each
- Targeted marketing
- Similar customers are grouped in the same cluster

COLLABORATIVE FILTERING RECOMMENDER SYSTEMS

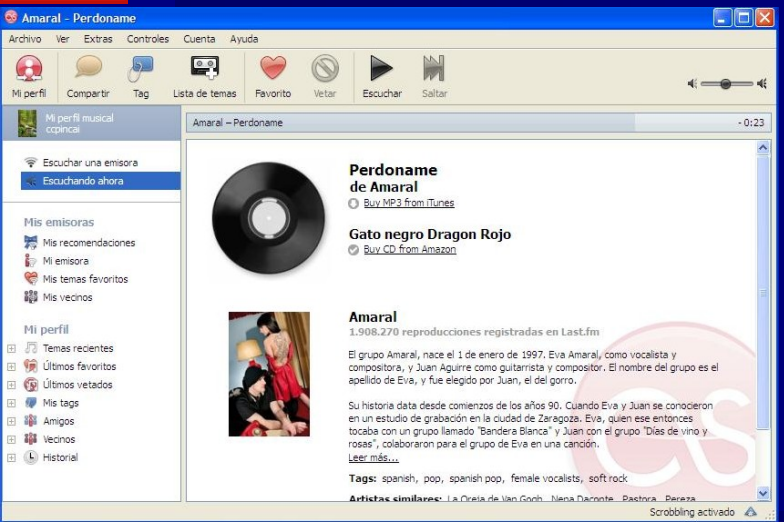
Group based on common items purchased, listened...

amazon

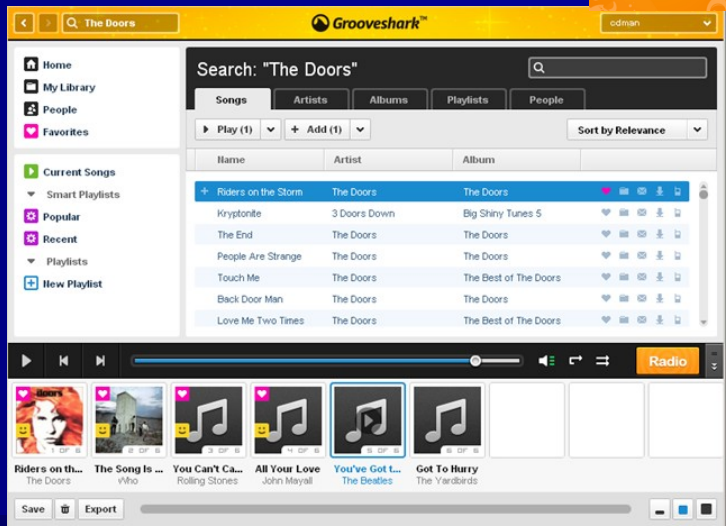
Customers Who Bought This Item Also Bought

	Angelica	Bill	Chan	Dan	Hailey	Jordyn	Sam	Veronica
Blues Traveler	3.5	2	5	3	-	-	5	3
Broken Bells	2	3.5	1	4	4	4.5	2	-
Deadmau5	-	4	1	4.5	1	4	-	-
Norah Jones	4.5	-	3	-	4	5	3	5
Phoenix	5	2	5	3	-	5	5	4
Slightly Stoopid	1.5	3.5	1	4.5	-	4.5	4	2.5
The Strokes	2.5	-	-	4	4	4	5	3
Vampire Weekend	2	3	-	2	1	4	-	-

last.fm



Grooveshark

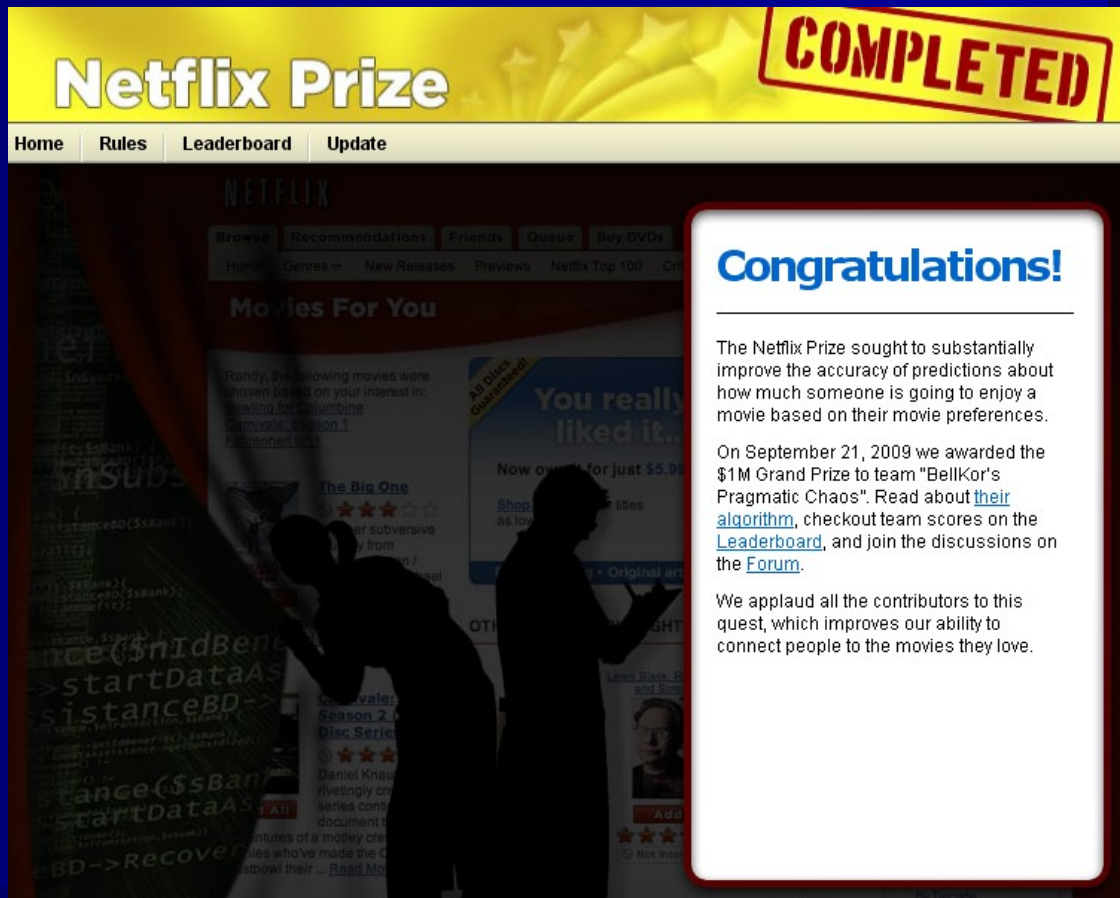


FILM RECOMMENDATION SYSTEMS

NETFLIX PRIZE

NETFLIX

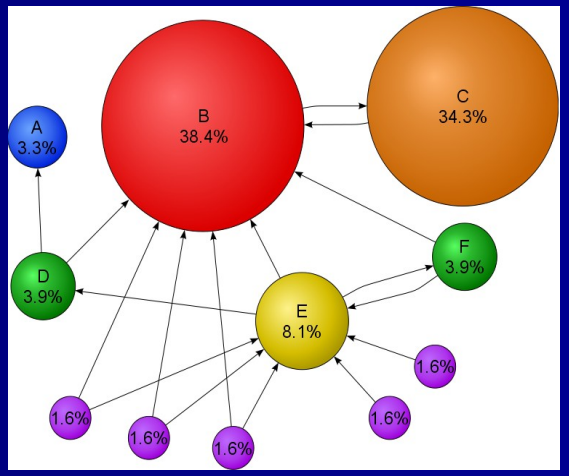
Watch TV shows & movies anytime, anywhere. For one low monthly price.



TEXT CLUSTERING

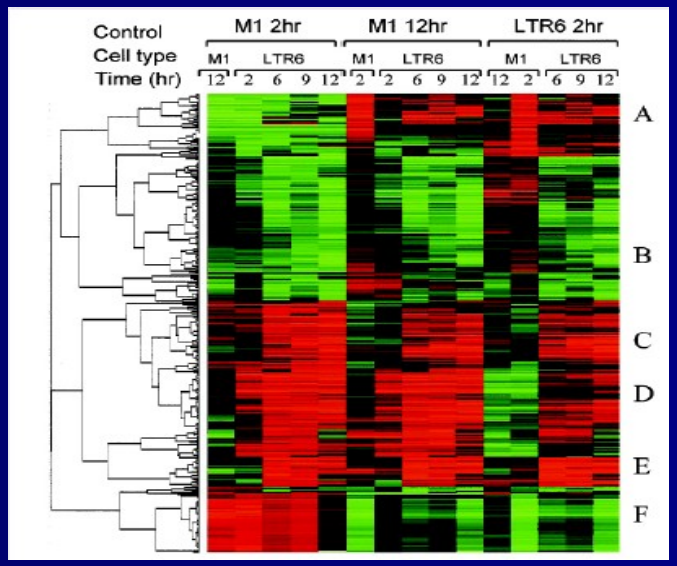
- Text clustering: documents that are similar → Google
  - PageRank
  - Diversity in search results

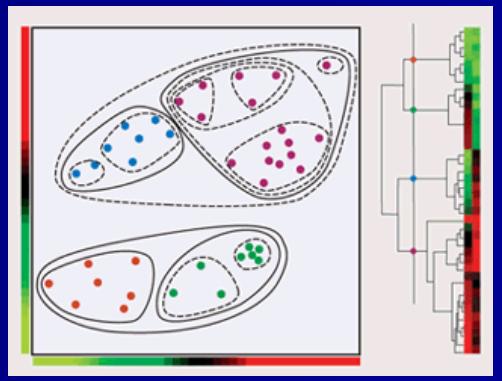
- Google, a key example of efficient data organization-”clustering”



DNA MICROARRAY CLUSTERING

- Find genes with similar expression profiles → a way to infer the function of genes whose function is unknown
- Biclustering… a classic concept in fashion again:
  - Hartigan JA (1972). "Direct clustering of a data matrix". *Journal of the American Statistical Association* **67** (337)
  - Finding a subgroup of samples with a similar pattern in a subgroup of variables (not in all the variables)

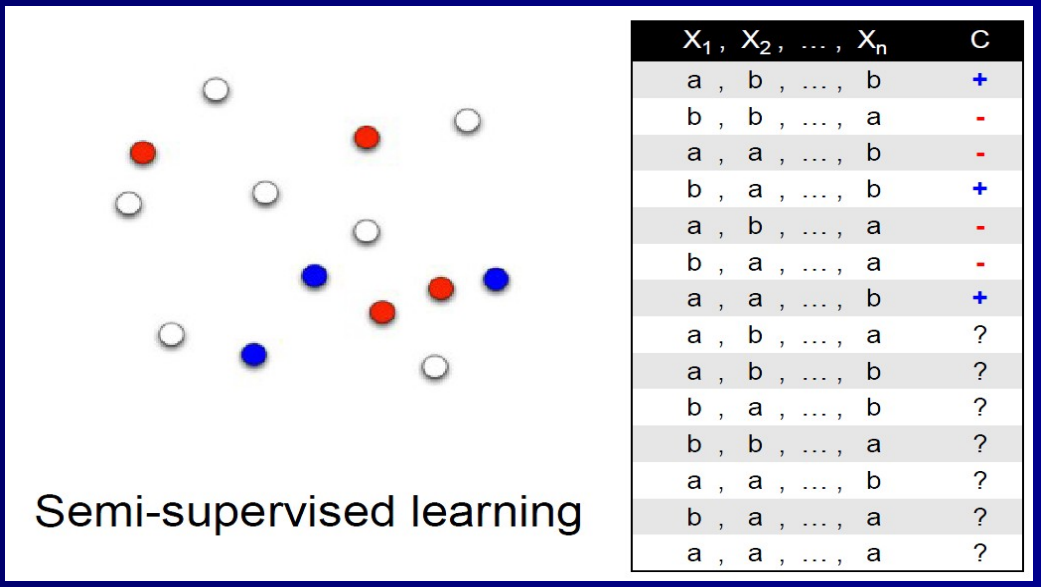




SEMI-SUPERVISED CLASSIFICATION

- Given a collection of records-samples (*training set*)
  - Each record contains a set of *attributes-features-predictors*
  - A small subset of the samples is categorized

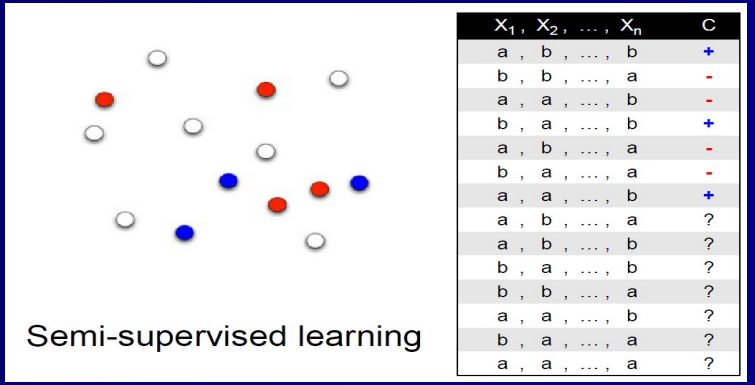
- "Weakly supervised classification"



X_1 , X_2 , … , X_n	C
a , b , … , b	+
b , b , … , a	-
a , a , … , b	-
b , a , … , b	+
a , b , … , a	-
b , a , … , a	-
a , a , … , b	+
a , b , … , a	?
a , b , … , b	?
b , a , … , b	?
b , b , … , a	?
a , a , … , b	?
b , a , … , a	?
a , a , … , a	?

SEMI SUPERVISED CLASSIFICATION

- Most of the samples do not show a class value. Why?
  - Categorization: human-time consuming task
  - No knowledge to categorize the samples
- Objective: learn a supervised model
- Can a learning process which takes advantage of unlabeled samples, construct a better supervised classification model?

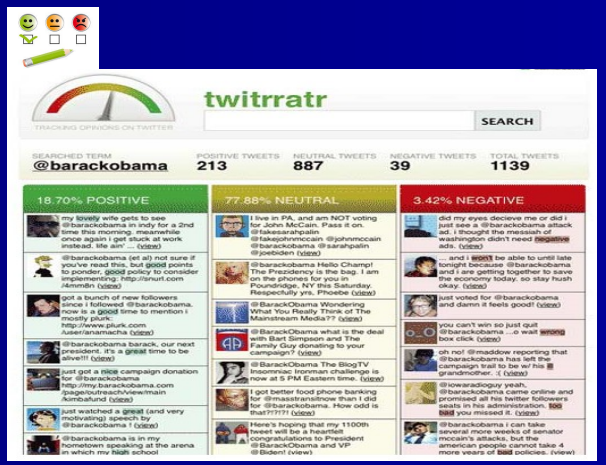


X_1 , X_2 , ... , X_n	C
a , b , ... , b	+
b , b , ... , a	-
a , a , ... , b	-
b , a , ... , b	+
a , b , ... , a	-
b , a , ... , a	-
a , a , ... , b	+
a , b , ... , a	?
a , b , ... , b	?
b , a , ... , b	?
b , b , ... , a	?
a , a , ... , b	?
b , a , ... , a	?
a , a , ... , a	?

SENTIMENT ANALYSIS

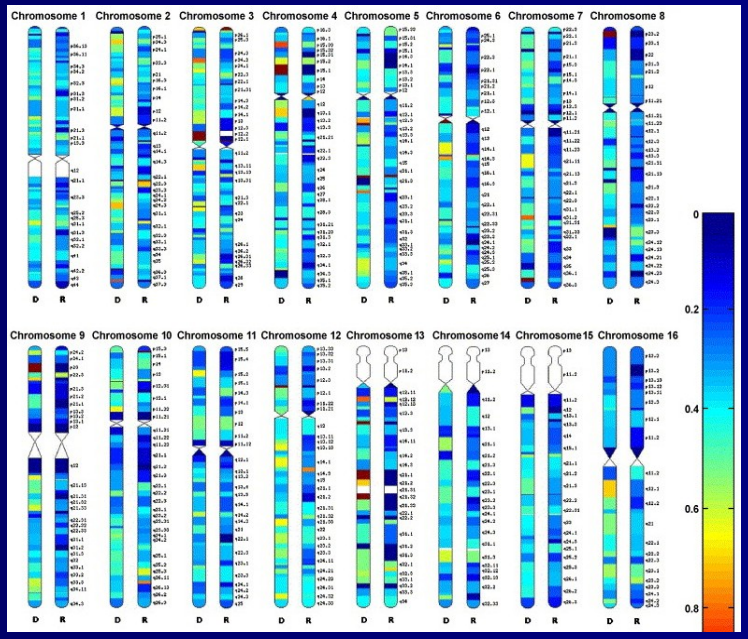
- Companies: reputation
- Opinions about its products:
  - social networks
  - blogs
  - forums...
- Automatically classify the written opinion: {+, -, neutral}
- NLP: “Natural Language Processing”

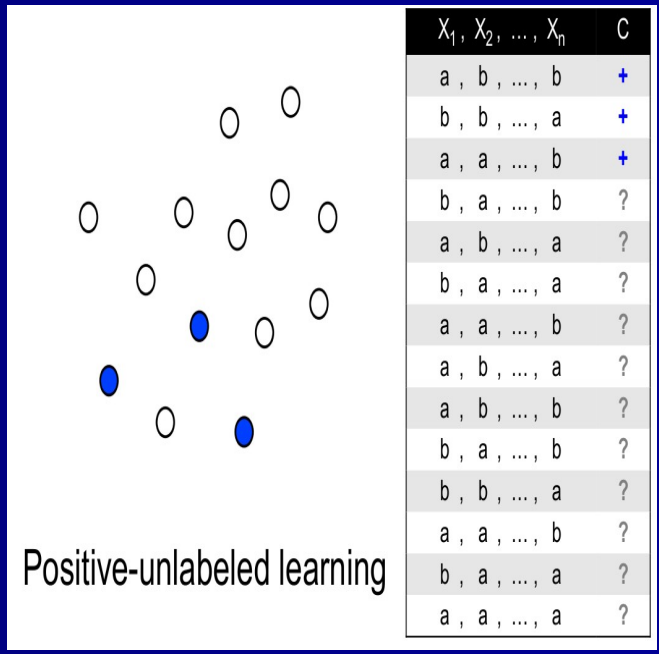
```
1728 </DOCUMENTO>
1729 <DOCUMENTO ID_REF="291896" ID="275">
1730  debe estar bien pero... 116 + las llamadas a moviles...
     es mucha pasta
1731
1732 aunque ono no es mas barato con sus 24mb.
1733 </DOCUMENTO>
1734 <DOCUMENTO ID_REF="291898" ID="276">
1735  Pues yo estuve el sábado por la tarde en la Tienda
     Telefónica de Gran Vía (Madrid) y la verdad que no sé
     qué bitrate tendrá el Imagenio ese de las narices, pero
     se veía como el culo, como un DivX cutrillo.
1736 </DOCUMENTO>
```



PREDICTION OF GENES RELATED TO CANCER

- It is already known that certain genes are related to cancer
- Rest of the genes: it can not be stated that they are not related to cancer
- Helpful to prioritize, for oncogenic experts, the study of specific genes
- More difficult than semi-supervised classification: positive-unlabaled



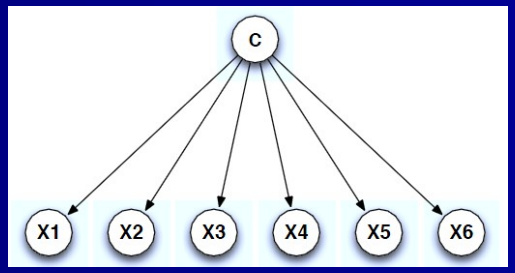


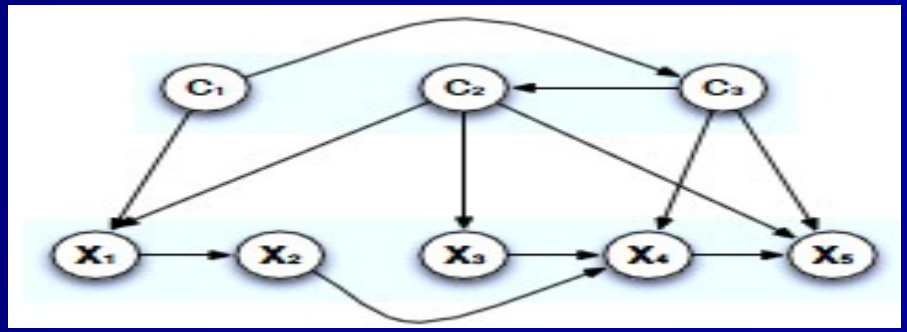
$X_1, X_2, \ldots, X_n$	C
a , b , ... , b	+
b , b , ... , a	+
a , a , ... , b	+
b , a , ... , b	?
a , b , ... , a	?
b , a , ... , a	?
a , a , ... , b	?
a , b , ... , a	?
a , b , ... , b	?
b , a , ... , b	?
b , b , ... , a	?
a , a , ... , b	?
b , a , ... , a	?
a , a , ... , a	?

OTHER TYPES OF CLASSIFICATION PROBLEMS
MULTIDIMENSIONAL CLASSIFICATION

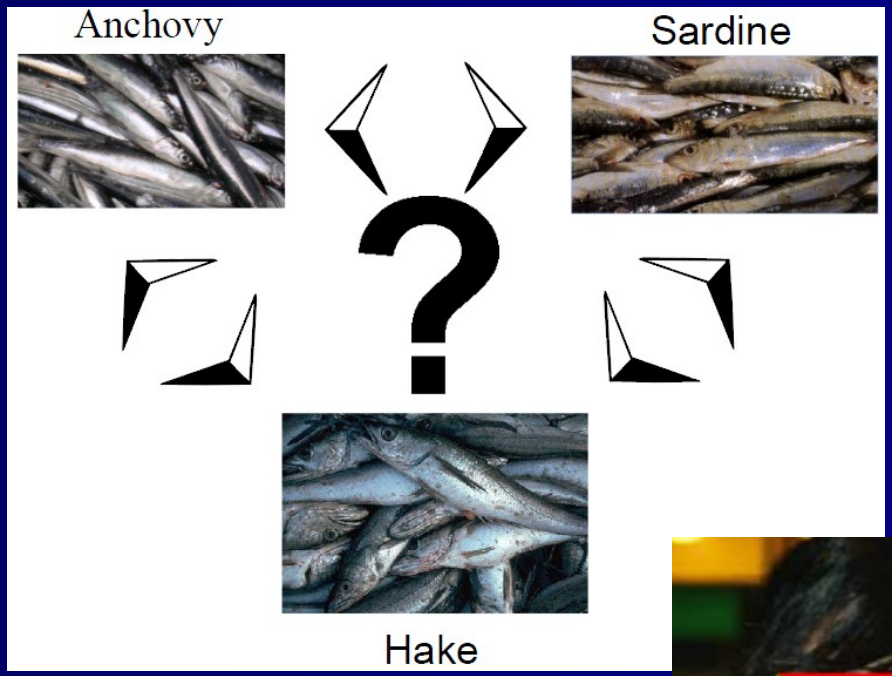
- Several class variables to be jointly predicted
- Learn relationships between class variables
- New term: Joint accuracy

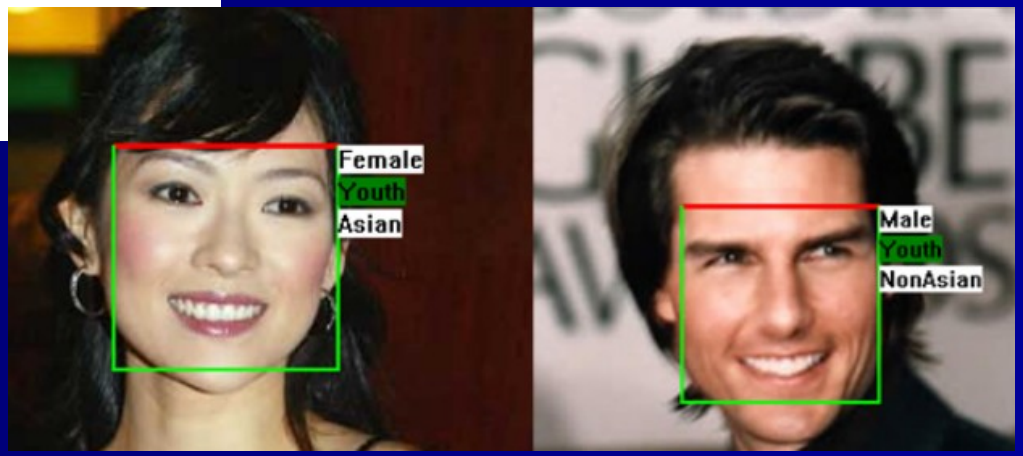
X_1	X_2	...	X_n	C_1	C_2	...	C_m
$x_1^{(1)}$	$x_2^{(1)}$	...	$x_n^{(1)}$	$c_1^{(1)}$	$c_2^{(1)}$	...	$c_m^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	...	$x_n^{(2)}$	$c_1^{(2)}$	$c_2^{(2)}$	...	$c_m^{(2)}$
...	...	...	...	...	...	...	...
$x_1^{(N)}$	$x_2^{(N)}$	...	$x_n^{(N)}$	$c_1^{(N)}$	$c_2^{(N)}$	...	$c_m^{(N)}$



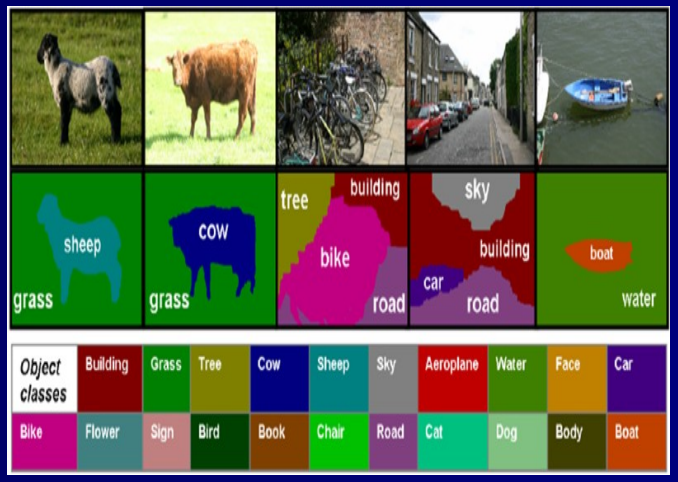


MULTIDIMENSIONAL CLASSIFICATION APPLICATIONS





MULTILABEL CLASSIFICATION



N.	Film	Year	Genre
1	Cadena perpetua	1994	Crime, Drama
2	El padrino	1972	Crime, Drama
3	El padrino. Parte II	1974	Crime, Drama
4	El bueno, el feo y el malo	1966	Adventure, Western
5	Pulp Fiction	1994	Crime, Thriller
6	12 hombres sin piedad	1957	Drama
7	La lista de Schindler	1993	Biography, Drama, History, War
8	El caballero oscuro	2008	Action, Crime, Drama, Thriller
9	El señor de los anillos: El ret...	2003	Action, Adventure, Drama, Fantasy
10	El club de la lucha	1999	Drama

X1	X2	...	Xn	C
0	1	...	0	a,c
1	0	...	0	b
1	0	...	1	b,c
0	0	...	1	a,b
1	1	...	0	a,b,c
0	1	...	1	a,b
0	0	...	0	b,c

X1	X2	...	Xn	C
1	1	...	1	?

MULTIPLE INSTANCE LEARNING

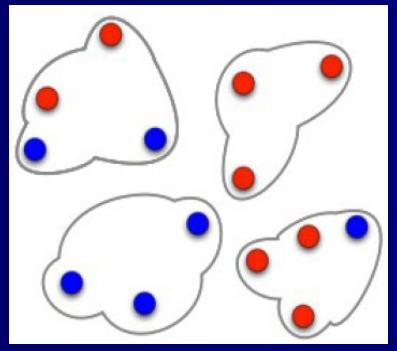
X_1 , X_2 , ... , X_n	C
a , b , ... , b	+
b , b , ... , a	-
a , a , ... , b	-
b , a , ... , b	+
a , b , ... , a	-
b , a , ... , a	-
a , b , ... , a	-
a , a , ... , b	+
a , b , ... , b	-
b , a , ... , b	-
b , a , ... , a	-
a , a , ... , b	+
b , b , ... , a	+
a , a , ... , a	+

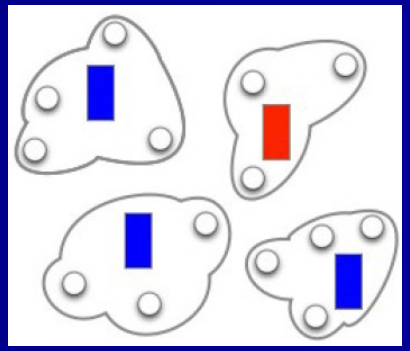
Bag label:

+ At least one instance in the bag is positive.

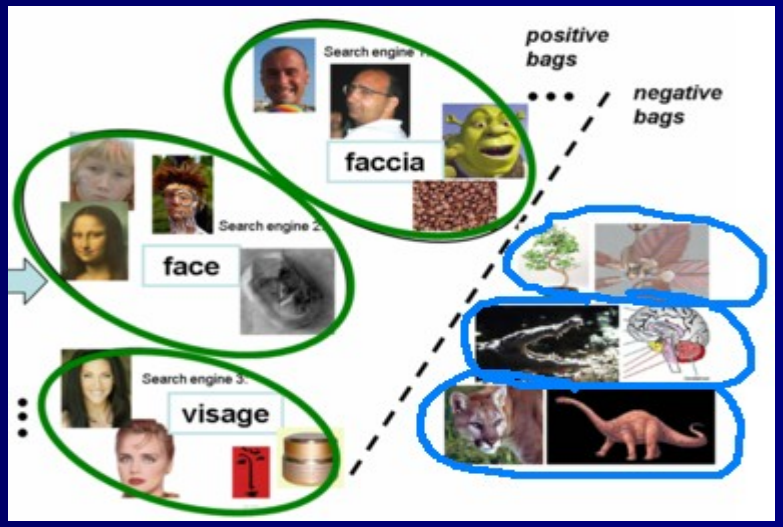
- Otherwise

X_1 , X_2 , ... , X_n	C
a , b , ... , b	
b , b , ... , a	+
a , a , ... , b	
b , a , ... , b	
a , b , ... , a	
b , a , ... , a	-
a , b , ... , a	
a , a , ... , b	
a , b , ... , b	
b , a , ... , b	+
b , a , ... , a	
a , a , ... , b	
b , b , ... , a	+
a , a , ... , a	

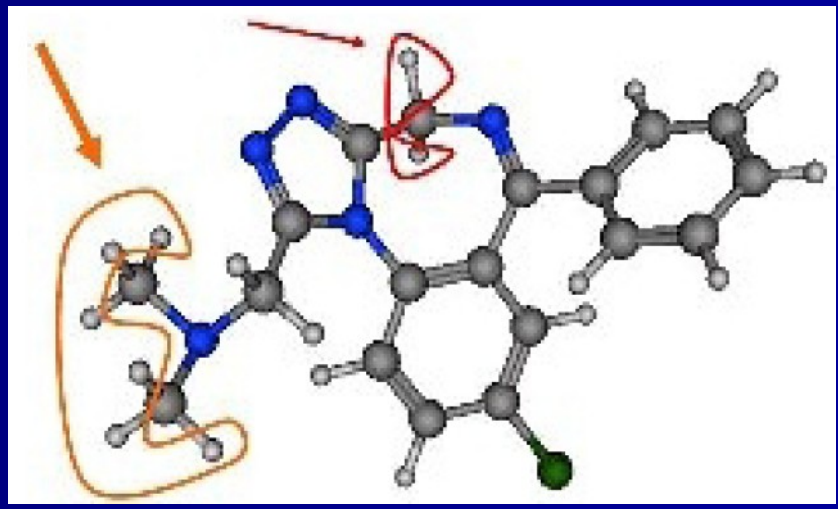




MULTIPLE INSTANCE LEARNING

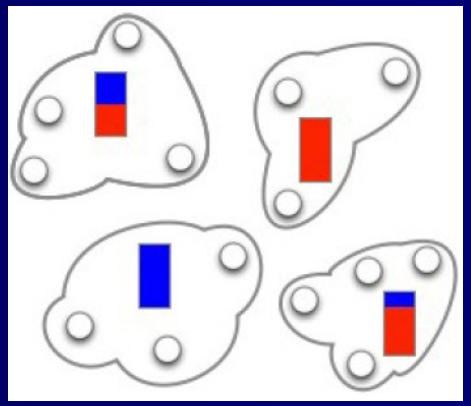


Are all the images “faces”?



Are all foldings of the protein of the same type?

LEARNING with LABEL PROPORTIONS



Bag label proportions:

p_+ Proportion of positive (+) instances in the bag

p_- Proportion of negatives (-)

$X_1, X_2, \ldots, X_n$	C
a , b , … , b	
b , b , … , a	0.5
a , a , … , b	0.5
b , a , … , b	
a , b , … , a	
b , a , … , a	0.0
a , b , … , a	1.0
a , a , … , b	
a , b , … , b	0.25
b , a , … , b	0.75
b , a , … , a	
a , a , … , b	
b , b , … , a	1.0
a , a , … , a	0.0

LABEL PROPORTIONS APPLICATIONS

Embryo selection in Assisted Reproductive Technologies (ART)

Two steps:

- **Transfer**: step in which one or several embryos are placed into the uterus of the patient.
- **Implantation**: step in which pregnancy is established (by one or several embryos).

Application	MILIp problem
Transfered embryos	Dataset
Implanted or not	Class labels
ART process	Bag
Number of children	Label proportions

Possible voters based on previous election results

- It involves any situation related with an election that can be organised as follows:

Application	MILIp problem
Census	Dataset
Candidates	Class labels
Polling station	Bag
Election results	Label proportions

CLASSIFICATION WITH PARTIAL LABELS

Each instance comes annotated with several class labels but only one of them is valid.

$X_1, X_2, \ldots, X_n$	C
a , b , … , b	a,b,c
b , b , … , a	a,c
a , a , … , b	d
b , a , … , b	b,c
a , b , … , a	a,d
b , a , … , a	a,b,d
a , a , … , b	b,c,d
a , b , … , a	c
a , a , … , b	b,c
b , a , … , a	b
a , a , … , a	a,b

FULL-SET CLASSIFICATION

X_1 , X_2 , ... , X_n	C
a , b , ... , b	b
b , b , ... , a	c
a , a , ... , b	c
b , a , ... , b	b
a , b , ... , a	a
b , a , ... , a	a
a , a , ... , b	c
a , b , ... , a	c
a , a , ... , b	b
b , a , ... , a	b
a , a , ... , a	a

Training

X_1 , X_2 , ... , X_n	C
b , b , ... , a	Y_1
a , a , ... , b	
b , a , ... , b	
a , b , ... , a	Y_2
b , a , ... , a	
a , b , ... , a	
a , a , ... , b	Y_3
a , b , ... , b	
b , a , ... , b	
a , a , ... , b	Y_4
b , b , ... , a	
a , a , ... , a	

Test

$Y_i = (y_{i1}, y_{i2}, y_{i3}) =$ PermutationOf{a,b,c}

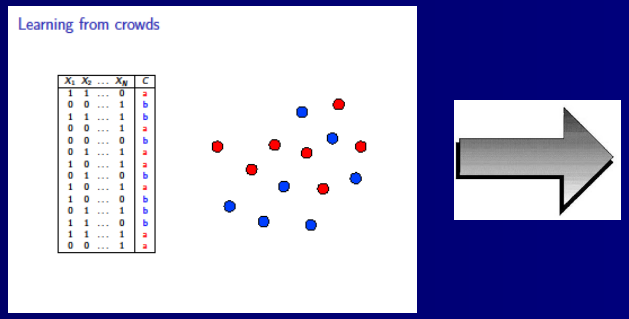
Permutations						
y_{i1}	a	a	b	b	c	c
y_{i2}	b	c	a	c	a	b
y_{i3}	c	b	c	a	b	a

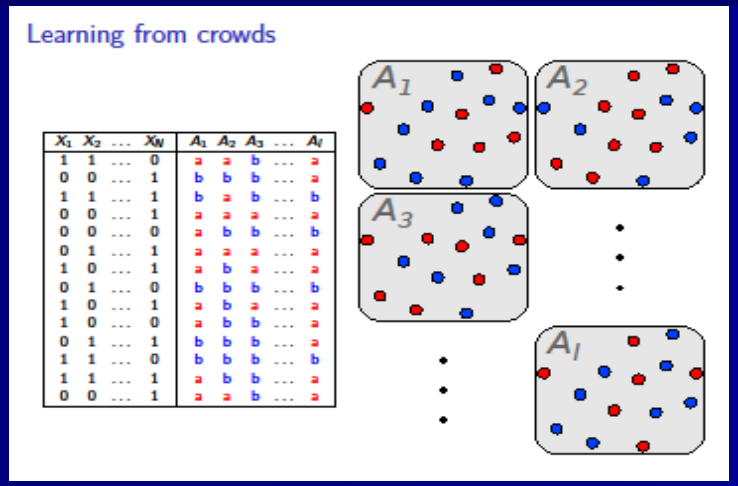
- identify different known objects in a group
- identify the location of the students in the classroom

LEARNING FROM CROWDS

- real class for each object is not known: no “golden truth"

- domain experts (A_j) annotate their opinion about the label of each object





Learning from crowds

Motivation

- Expensive/difficult expert labeling
- Recent availability of cheap (non-expert) labeling sources

Data collection

- Social networks, games, etc.
- Specific platforms (e.g. Amazon Mechanical Turk)

ASSOCIATION RULES

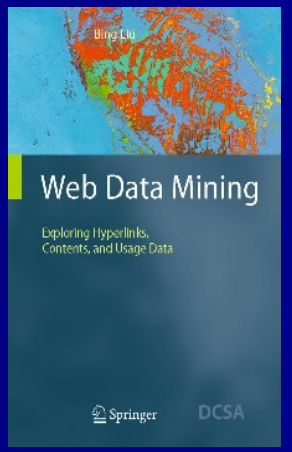
- Given a set of records each of which contain some number of items from a given collection;
  - Depedendency rules which will predict occurrence of an item based on occurrences of other items.
  - Rules are composed of “antecedent” and “consequence” parts: IF-THEN form
  - No “class” concept: any item can be in the “antecedent” or “consequence” part
  - **“Support**” and **“Confidence”** concepts

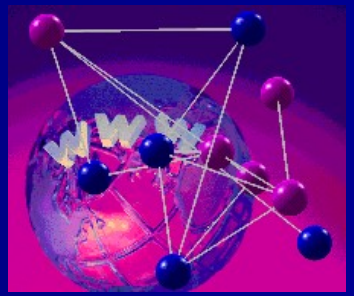
t1: {bread, cheese, fluidmilk}
t2: {apple, eggs, salt, yogurt}
t3: {bananas, eggs, saladvegetable}
.....
tn: {biscuit, eggs, fluidmilk}

1	ANTECEDENT	==>>	CONSEQUENCE	Support (%)	Confidence (%)
2	Pizza & Tomato	==>>	Grated cheese	5%	82%
3	Pizza & "Man"	==>>	Beer	3%	75%
4	SaladVegetable & Meat	==>>	Wine	10%	68%
5	Milk & Bread	==>>	Jam	18%	61%
6	Diaper & "Man"	==>>	Beer	4%	44%
7	Coke & Nachos	==>>	Paper serviette	2%	40%

MARKET BASKET ANALYSIS

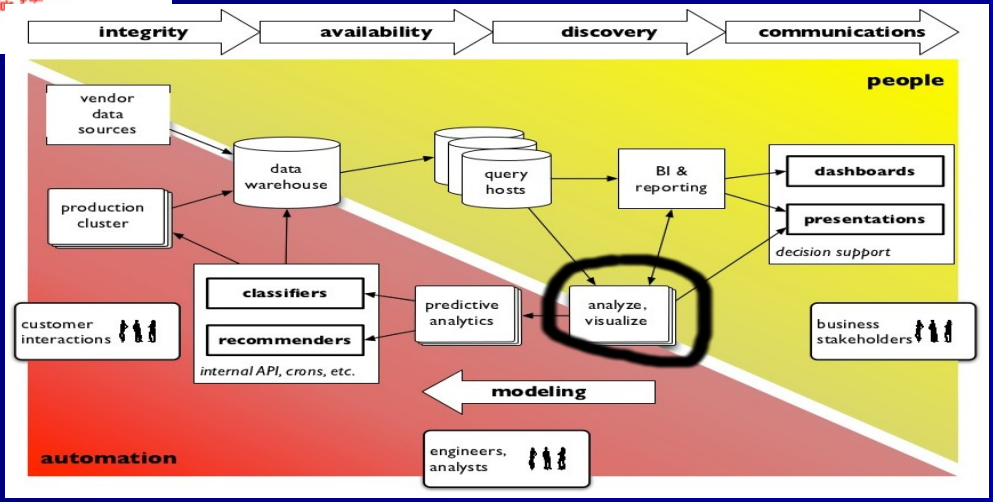
WEB USAGE MINING - CLICKSTREAM ANALYSIS



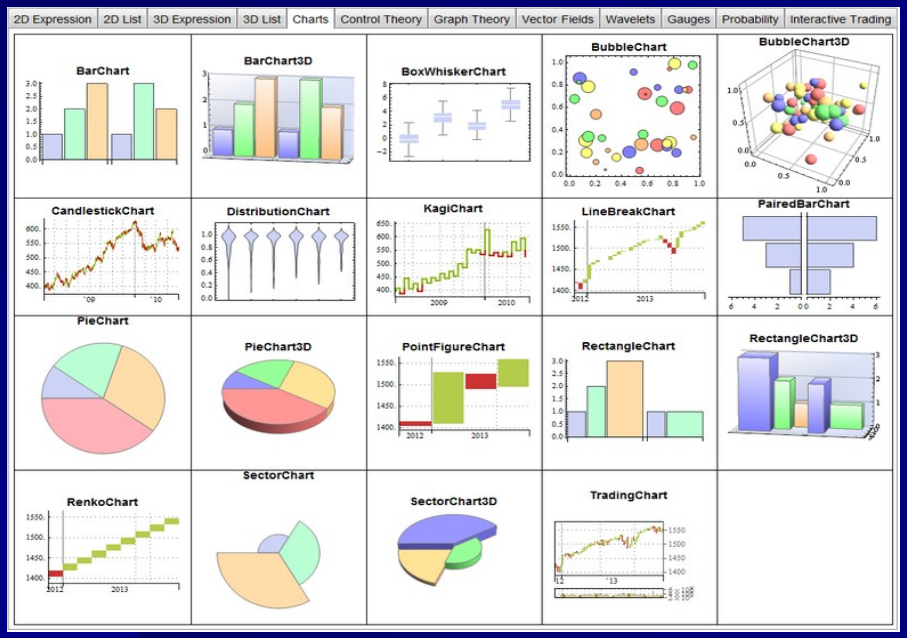


- Association rules and the Apriori algorithm: A first tutorial

DATA VISUALIZATION
DATA EXPLORATION



DATA VISUALIZATION
DATA EXPLORATION



- List of commercial and free visualization tools
- 11 steps for data exploration in R (with codes)

TOP-10 DATA MINING ALGORITHMS

Contents [hide]

1. C4.5
2. k-means
3. Support vector machines
4. Apriori
5. EM
6. PageRank
7. AdaBoost
8. kNN
9. Naive Bayes
10. CART
Interesting Resources
Now it's your turn...

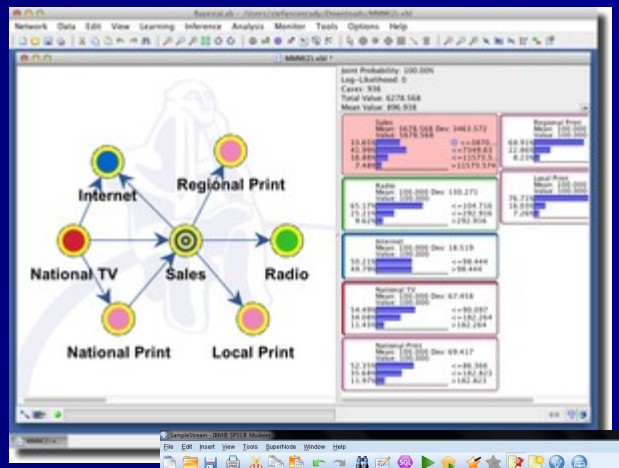
http://rayli.net/blog/data/top-10-data-mining-algorithms-in-plain-english/

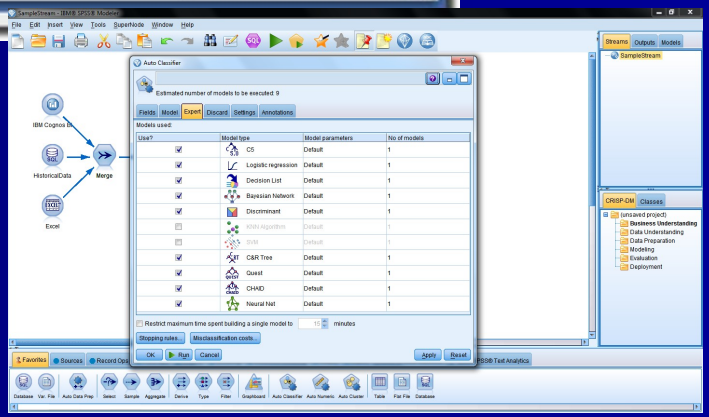
DATA REPOSITORIES

- Public datasets on Github
- UCI machine learning repo sitory
- kaggle.com competitions

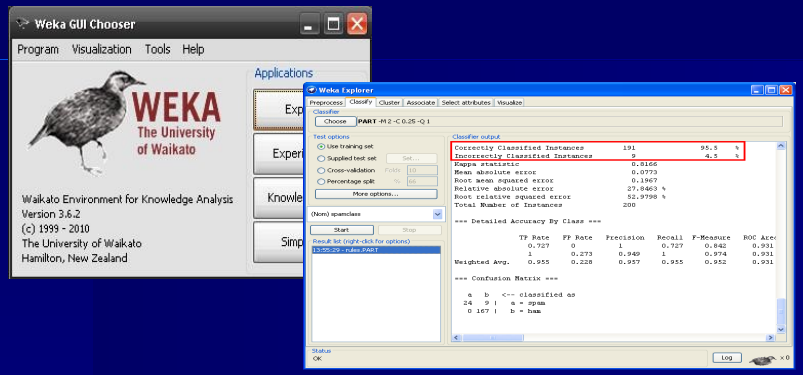
kaggle

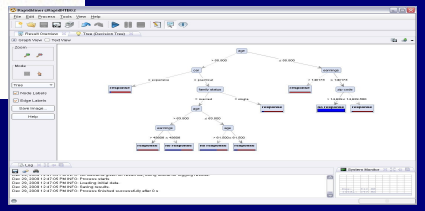
COMERCIAL SOFTWARE FOR DATA MINING





FREE SOFTWARE FOR DATA MINING



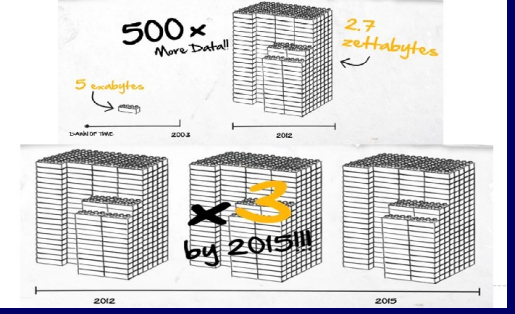


- Software suites for data mining, analytics and knowledge discovery
- 11 open source tools to make the most of machine learning
- Top 10 machine learning projects in GitHub
- 50 useful machine learning & prediction APIs
- Classification software: a list
- Top 15 frameworks for machine learning experts
- Bayesian networks and Bayesian classifier software

KDnuggets

POLLS' RESULTS
kdnuggets.com

- Application field
- Analyzed data types / sources
- Primary programming language for data mining
- Used software tools
- Complete list of polls

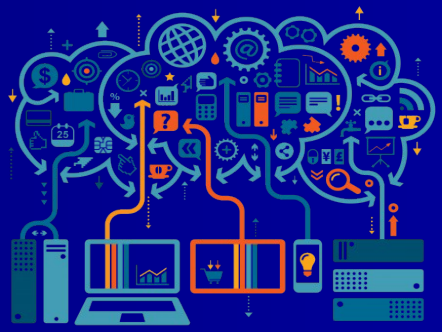


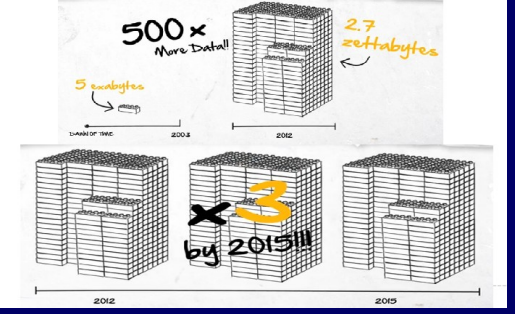
BIG DATA

- New technological concept, new generation of technologies
- In part, “inheritor” of data mining

- Related to the challenges exposed to manipulate massive datasets (petabytes, exabytes):
  - Capture and storage
  - Processing and computing
  - Analysis and mining

- Social networks, Electronical purchases, financial companies, GPS systems, sensors, images
- …





BIG DATA

- **Demands the development of new architecture platforms (hardware and database) and updated analysis techniques because of data size, diversity and complexity**

- **“4V” definition: volume, variety, velocity, value**
- **Unestructured data, diverse sources**

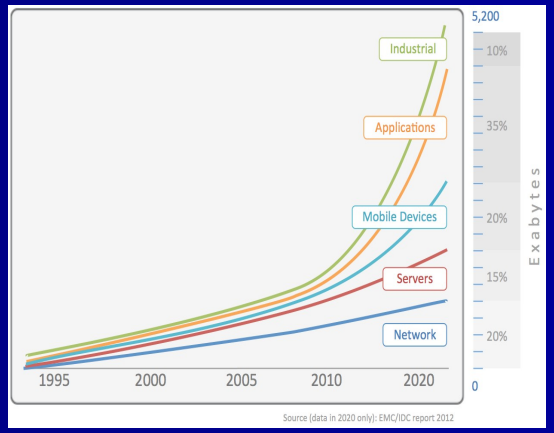
- **Business oportunity**

Big Data, Big Impact:
New Possibilities for International Development

WORLD ECONOMIC FORUM
COMMITTED TO IMPROVING THE STATE OF THE WORLD

WHY “BIG”?

- Megabytes, Gigabytes… → Terabytes, Exabytes
- Classic numerical matrices … → images, text, links, localizations
- PC's … → more advanced computing platforms (e.g. cloud)
- Excel sheets, databases … → more advanced hosting database systems (e.g. MongoDB)



BIG DATA

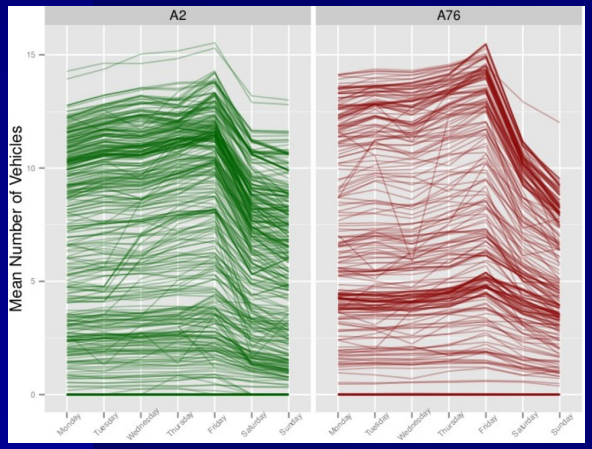
mongoDB

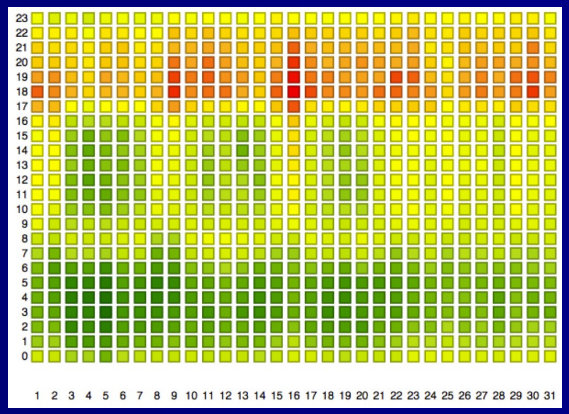
- **Application fields: credit card transactions, telecommunication companies, web processing and traffic, social networks…**
- **Parallel computation**
- **Massive database storage**
- **Grid and cloud computing, GPU…**
- **NoSQL, MongoDB non-relational…**
- **Big data “chain” for data: generation, acquisition, storage, analysis**

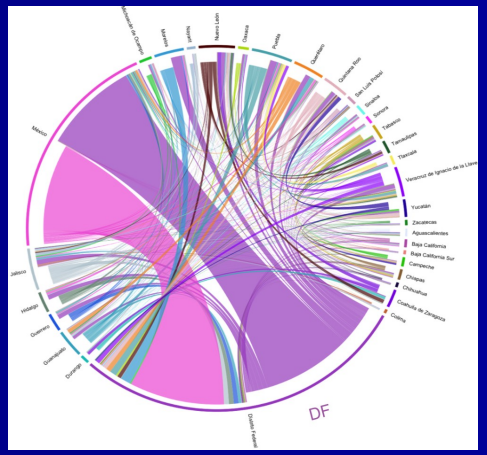
APPLICATIONS - EuroStat

- **EuroStat – UNECE: SandBox benchmark platform for big data [link]**

- **Traffic loops: hourly traffic intensity**
- **Electricity consumption**
- **Geolocalized tweets: people movement**

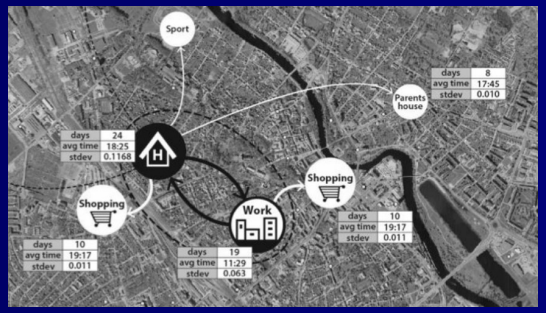


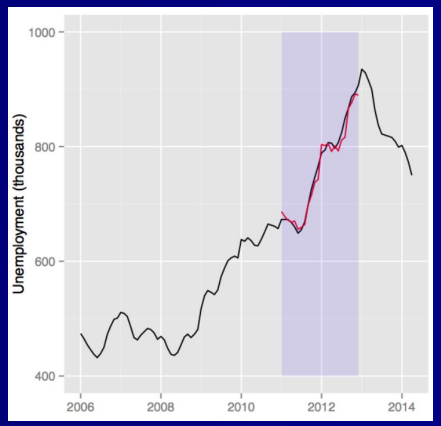




APPLICATIONS - EuroStat

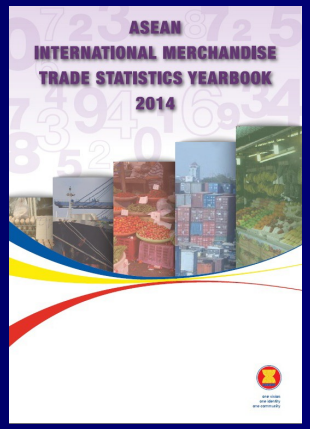
- Phones' usage: people movement and tourism indicators
- GoogleTrends topics: relation with unemployment
- Wikipedia pages' views: relation with tourism

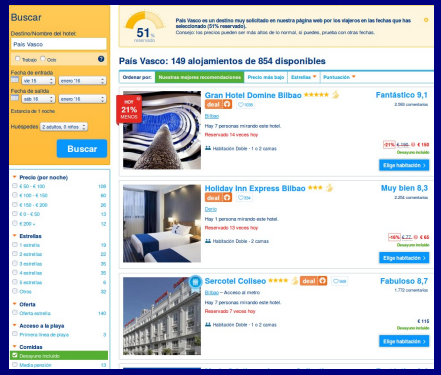


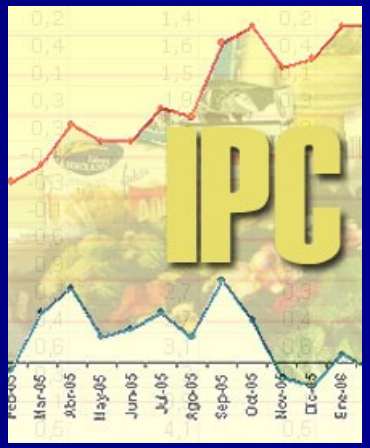


APPLICATIONS - EuroStat

- International Merchandise Trade Statistics
- Daily hotel prices in websites
- Consumer price index calculation from Internet



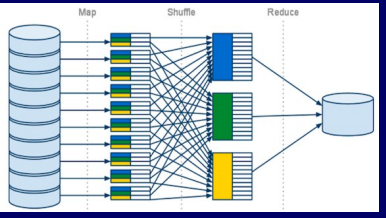




APPLICATIONS

- Where big data is?
  - - Sensors
  - - Private companies
  - - Internet: web scraping
  -
- European “open government data portal” [link]
- EuroStat: types of big data [link]

- Analysis:
  - - Visualization: maps, evolution, time series...
  - - When a classification problem exists: machine learning
  - - Alternative calculation of indexes
  - - “React to events”

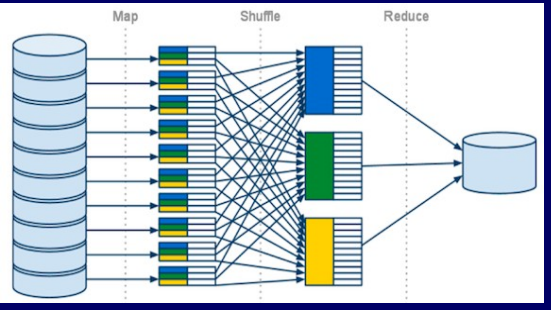


BIG DATA TECHNOLOGY

Principal big data programming framework: "Map and Reduce"

- **Designed by Google and offered to the community in 2004**
- **Extended as free software and known as "Hadoop"**

- **Used by Yahoo!, Facebook, Amazon...**
- **Hadoop big data platforms: Amazon Web Services, Google Cloud Platform, Microsoft ML Azure**
- **Hadoop as a service: 18 cloud options**
- **Hadoop: top 6 questions answered**



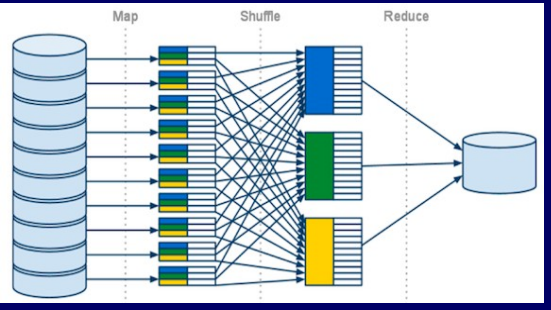
BIG DATA

- Parallel data programming
- Hadoop Distributed File System (HDFS)

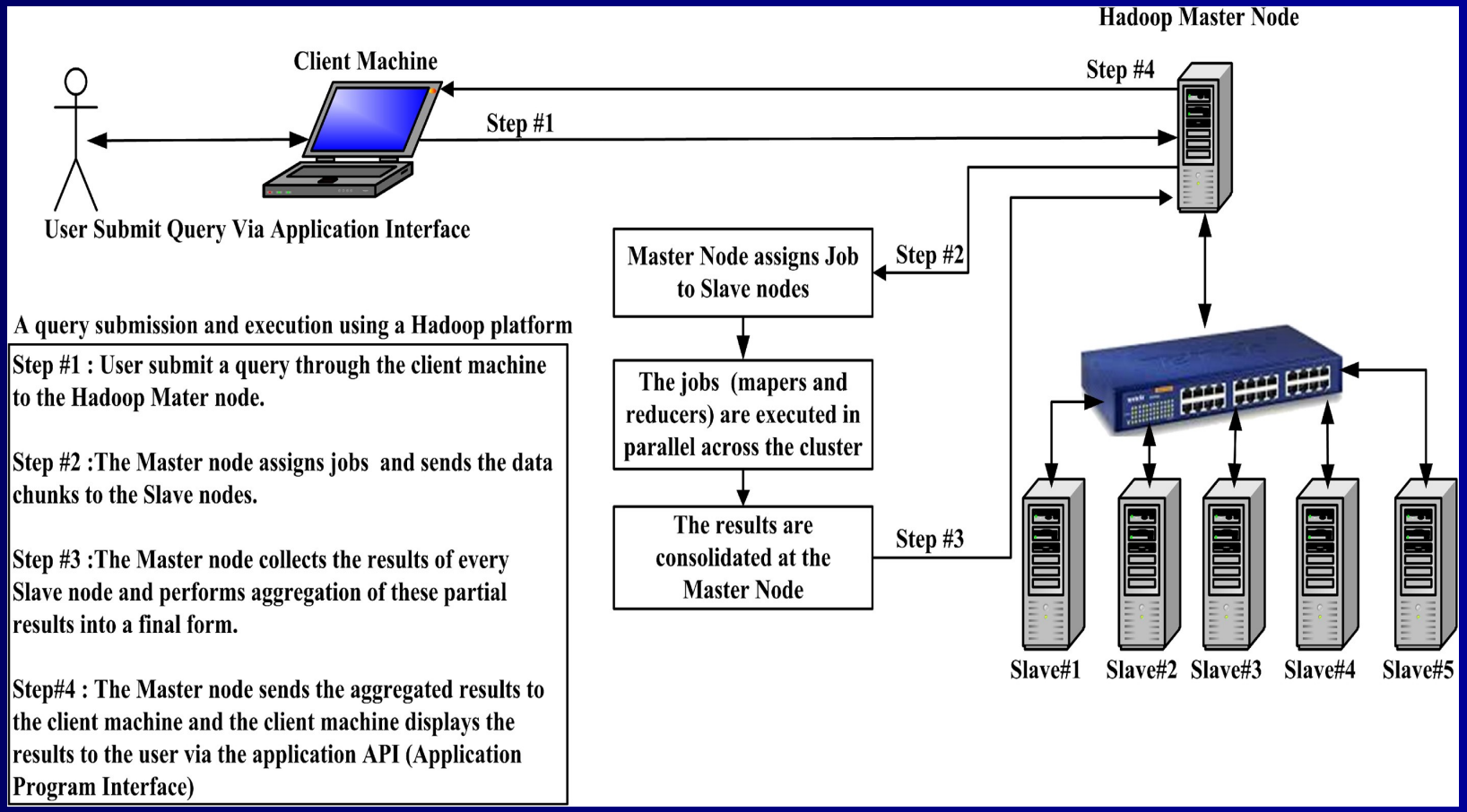
- Programming simplicity
- Transparent to the user

- Other similar and popular programming framework:

Apache Spark, Apache Flink...



BIG DATA



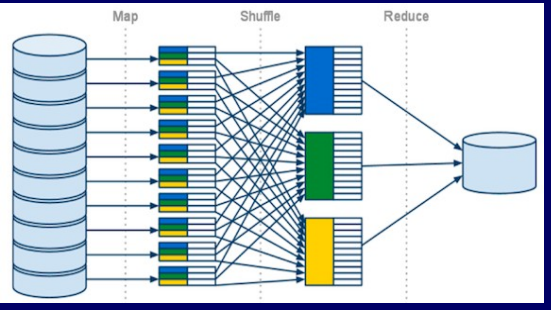
A query submission and execution using a Hadoop platform

Step #1 : User submit a query through the client machine to the Hadoop Mater node.

Step #2 :The Master node assigns jobs and sends the data chunks to the Slave nodes.

Step #3 :The Master node collects the results of every Slave node and performs aggregation of these partial results into a final form.

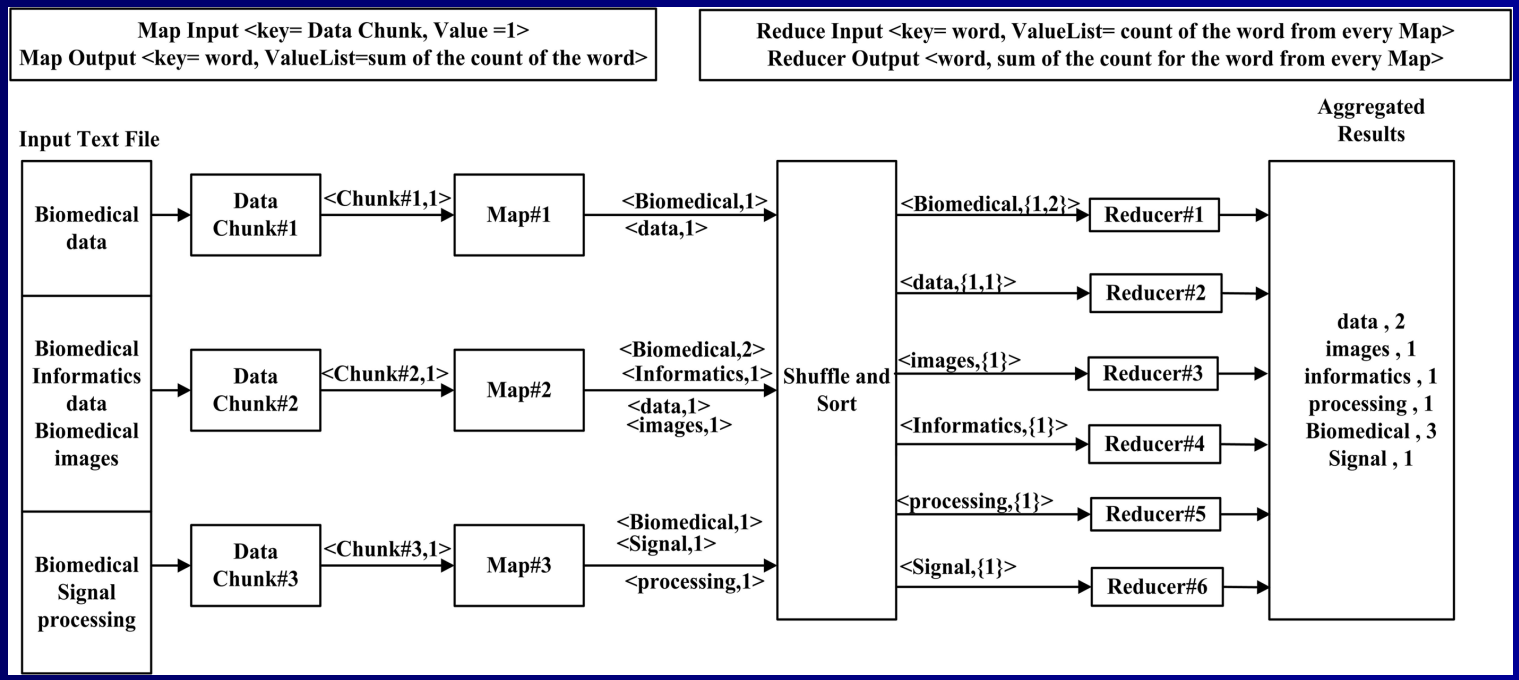
Step#4 : The Master node sends the aggregated results to the client machine and the client machine displays the results to the user via the application API (Application Program Interface)



BIG DATA

- “Map and Reduce” example: WordCount in a text

- Phases: “#Data chunk”, “#Map”, “#Shuffle and Sort”, “#Reduce”, “#Aggregation”



BIG DATA

Hadoop related projects – Hadoop “ecosystem”

Hadoop related project and technology	Description
Avro	• Avro is a framework for performing remote procedure calls and data serialization.
Flume	• Flume is a tool for harvesting, aggregating and moving large amounts of log data in and out of Hadoop.
HBase	• Based on Google’s Bigtable, HBase is an open-source, distributed, versioned, column-oriented store that sits on top of HDFS. HBase is column-based rather than row-based, which enables high-speed execution of operations performed over similar values across massive datasets.
HCatalog	• An incubator-level project at Apache, HCatalog is a metadata and table storage management service for HDFS.
Hive	• Hive provides a warehouse structure and SQL-like access for data in HDFS and other Hadoop input sources
Mahout	• Mahout is a scalable machine-learning and data mining library.
Oozie	• Oozie is a job coordinator and workflow manager for jobs executed in Hadoop, which can include non-MapReduce jobs.
Pig	• Pig is a framework consisting of a high-level scripting language (Pig Latin) and a run-time environment that allows users to execute MapReduce on a Hadoop cluster.
Sqoop	• Sqoop (SQL-to-Hadoop) is a tool which transfers data in both directions between relational systems and HDFS or other Hadoop data stores, e.g. Hive or HBase.
ZooKeeper	• ZooKeeper is a service for maintaining configuration information, naming, providing distributed synchronization and providing group services.
YARN	• YARN is a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users’ applications.
Cascading	• Cascading is an alternative API to Hadoop MapReduce. Cascading now has support for reading and writing data to and from a HBase cluster.
Twitter Storm	• Twitter Storm is a free and open source distributed real time computation system.
High performance computing cluster (HPCC)	• HPCC is an open source, data-intensive computing system platform developed by LexisNexis Risk Solutions
Dremel	• Dremel is a scalable, interactive ad-hoc query system for analysis of read-only nested data

Hadoop key terms: explained [link]

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014 97

Data Mining with Big Data

Xindong Wu, *Fellow, IEEE*, Xingquan Zhu, *Senior Member, IEEE*,
Gong-Qing Wu, and Wei Ding, *Senior Member, IEEE*

BIG DATA MINING

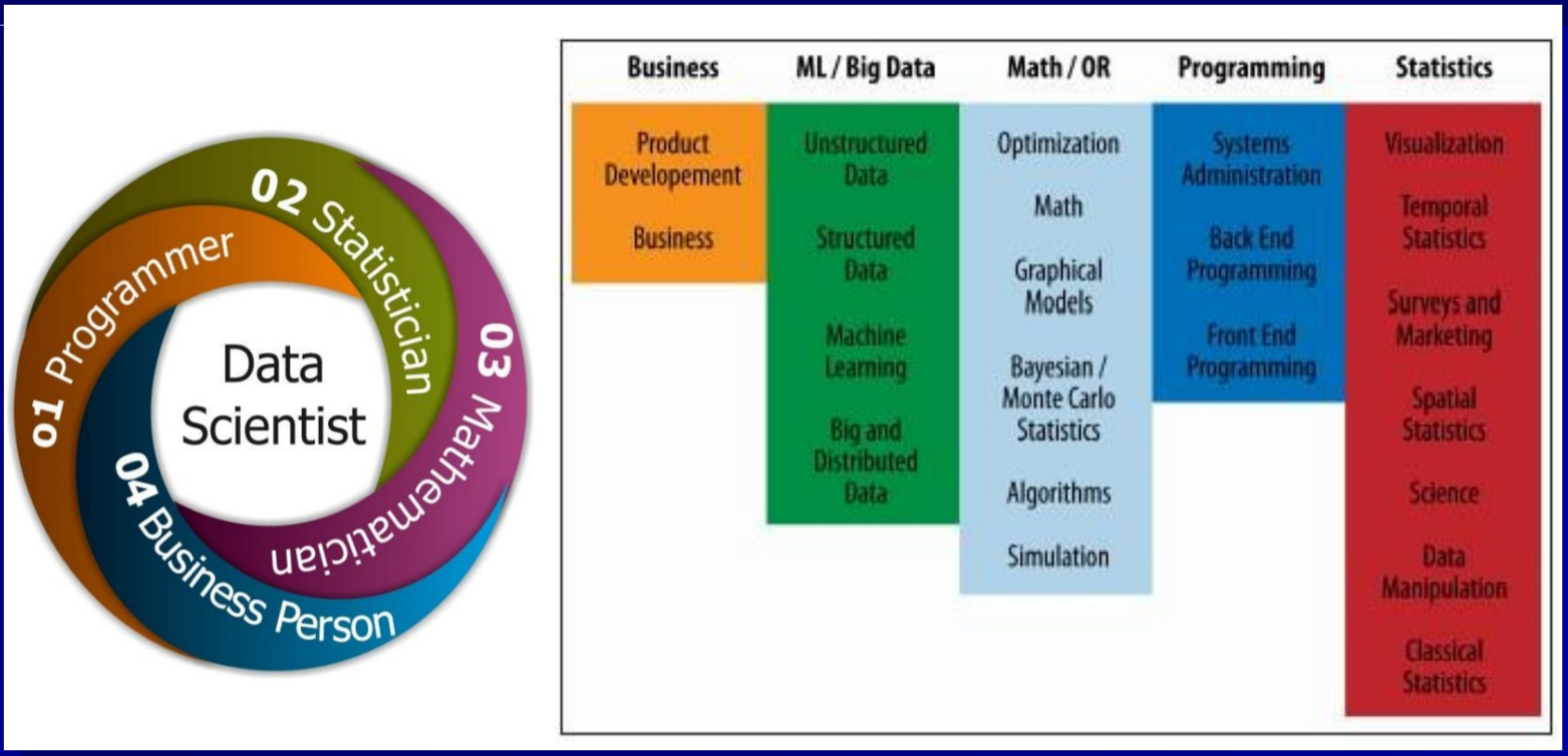
- **Analyzing data taking advantage of the *"Map and Reduce"* schema**
  - **Learning algorithms executed in parallel: data is commonly in different sources**
  - **Merged-collected in the "Reduce" step: global model is constructed by combining local submodels**

- **Different initiatives:**
  - **Machine learning with R and H2O [link]**
  - **Apache Mahout: scalable machine learning [link]**
  - **Microsoft ML Azure [link]**
  - **TensorFlow: Google's latest machine learning system [link1] [link2]**
  - **"The cloud machine learning work" [link]: Amazon, IBM Watson, Microsoft Azure**
  - **6 cloud based machine learning services [link]**

USEFUL LINKS FOR BIG DATA

- IEEE Access'2014: “Towards scalable systems for big data analytics: a technology tutorial” [paper]
- Presentations on BigData: top on SlideShare [link]
- The big data question: Hadoop or Spark?
- IBM tutorial: Hadoop, open source big data for the impatient
- Introduction to big data with Apache Spark
- Apache Spark, the hot new trend in big data
- OnePageR: Data Science with R, dealing with big data [link]
- 100 open source Big Data architecture papers for data professionals

WHO IS A DATA SCIENCIST



- 7 steps for learning data mining and data science
- Most viewed YouTube videos on data mining
- Tour of real-world machine learning problems

KDD PROCESS
KNOWLEDGE DISCOVERY IN DATABASES