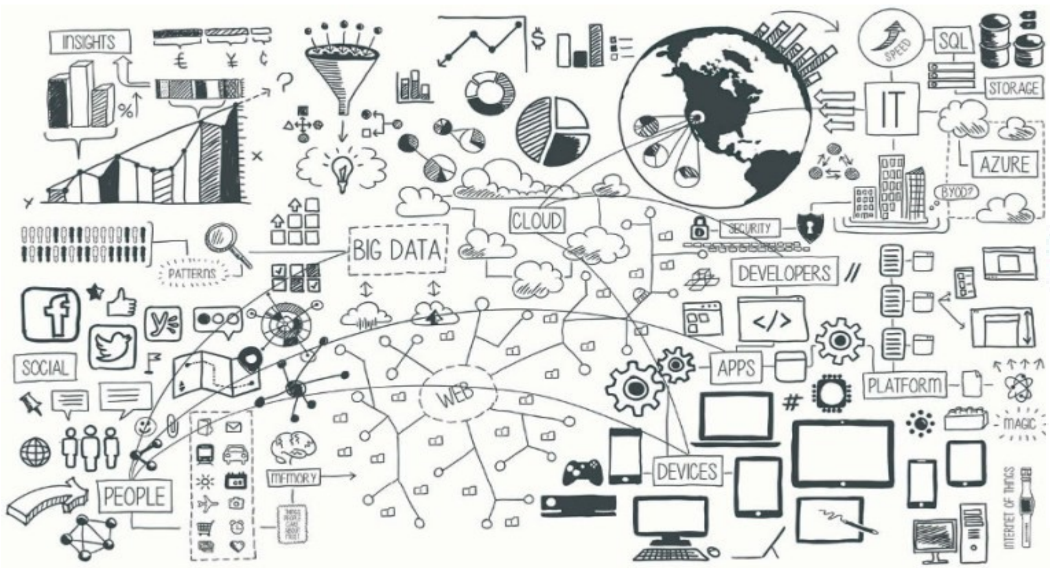
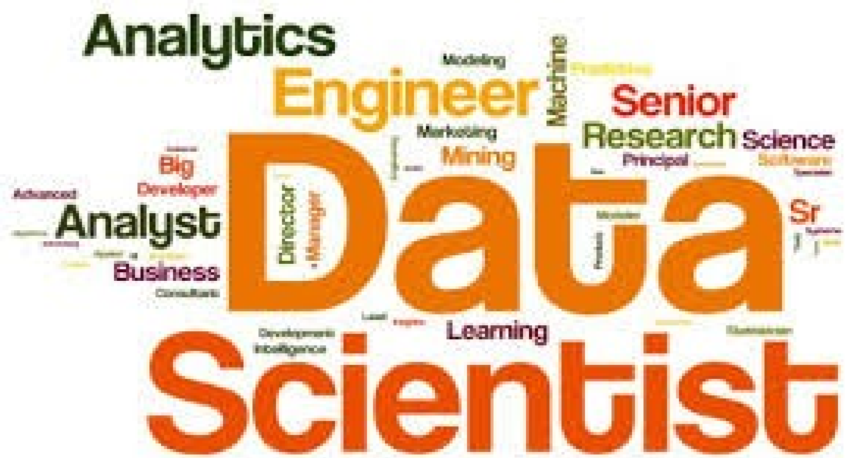
Data Mining (Minería de Datos)

INTRODUCTION AND HISTORICAL PERSPECTIVE

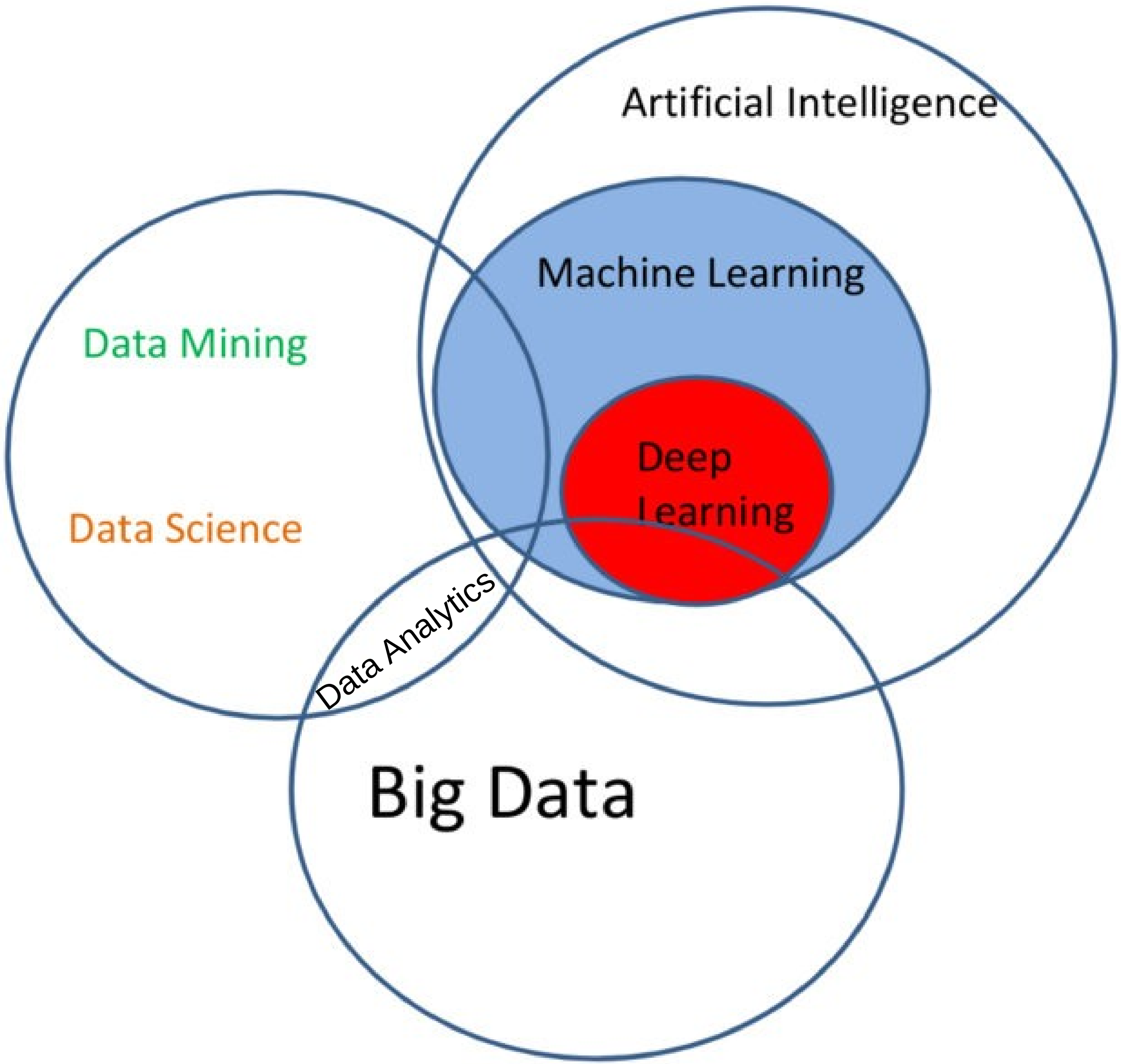




Sixto Herrera

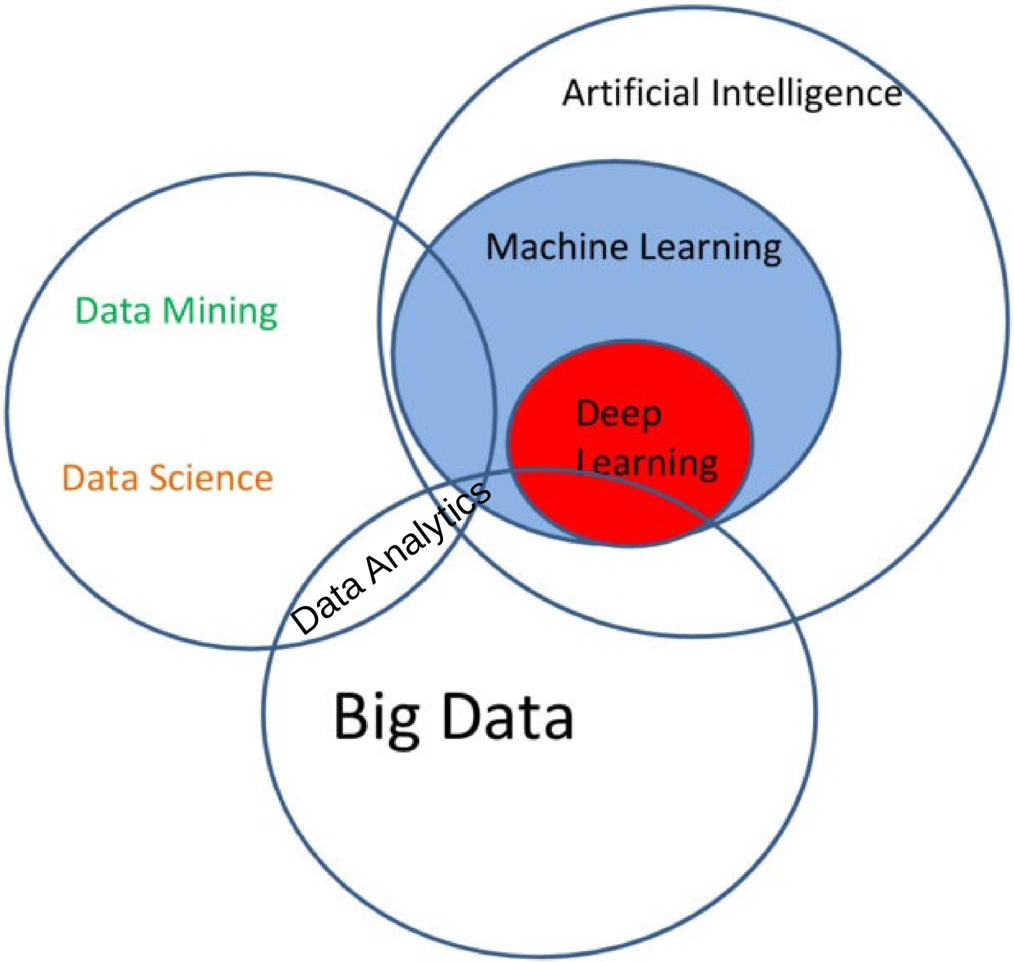
Grupo de Meteorología

Univ. de Cantabria – CSIC
MACC / IFCA


Artificial Intelligence
Machine Learning
Data Mining
Deep Learning
Data Science
Data Analytics
Big Data

Data Mining (DM) can be defined as the process that starting from apparently unstructured data tries to extract knowledge and/or unknown interesting patterns.

Machine Learning (ML) relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed (definition of A.Samuel).

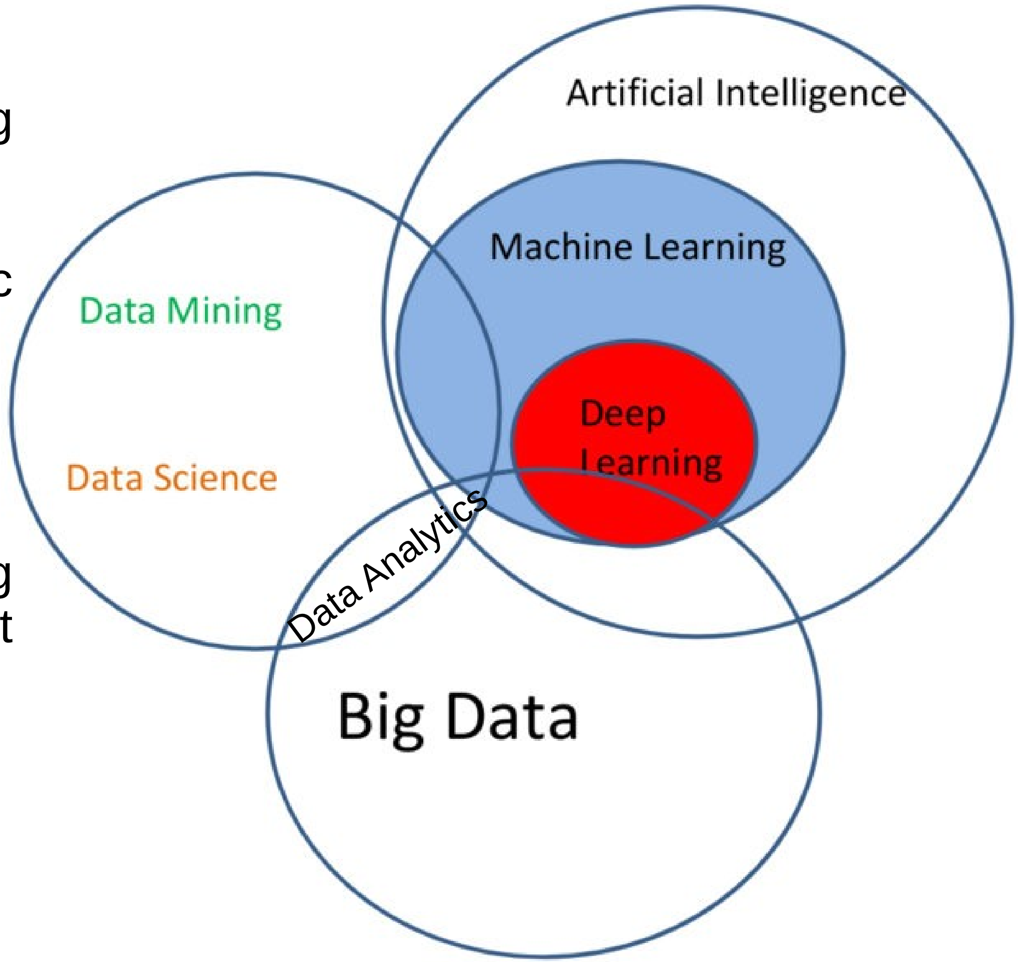


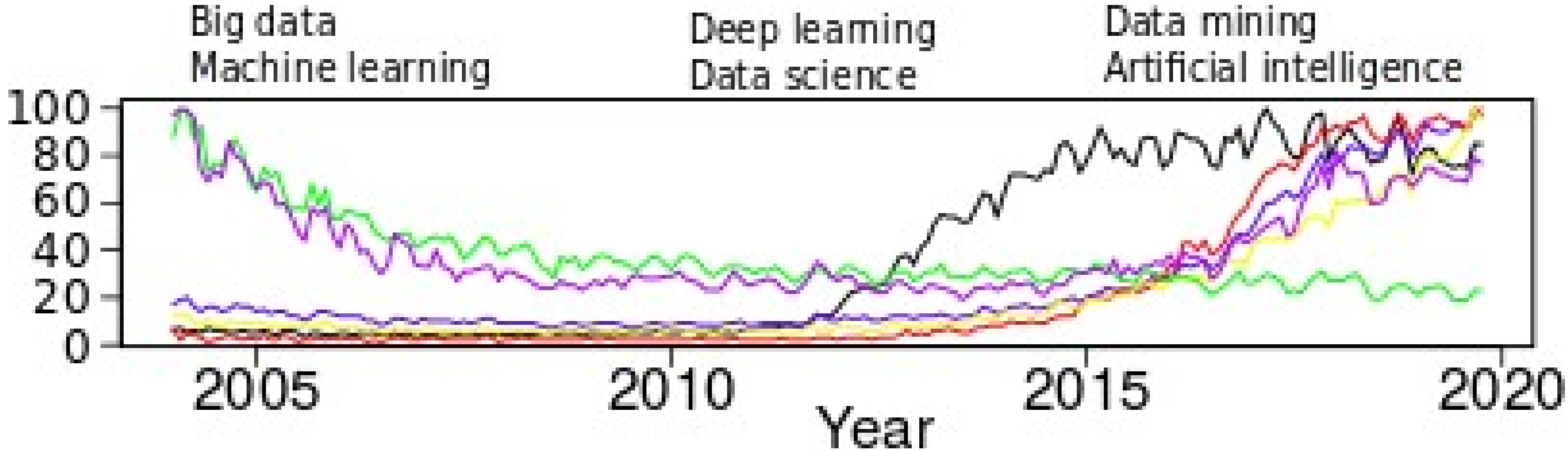
Data Mining (DM) can be defined as the process that starting from apparently unstructured data tries to extract knowledge and/or unknown interesting patterns.

During this process machine learning algorithms are used (A. Flag).
ML Techniques → Generic
Data Mining → Understand some specific domain.

While DM may utilize machine learning techniques, it may also drive the advancement of ML techniques/algorithms (P. Anantharam).

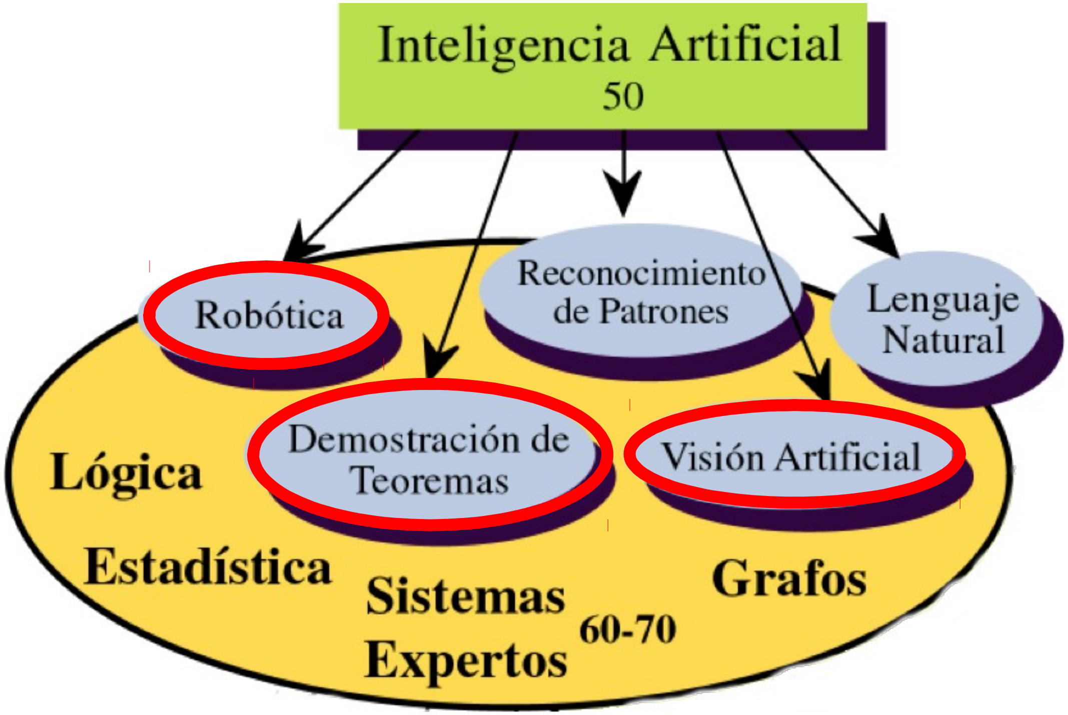
Machine Learning (ML) relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed (definition of A.Samuel).

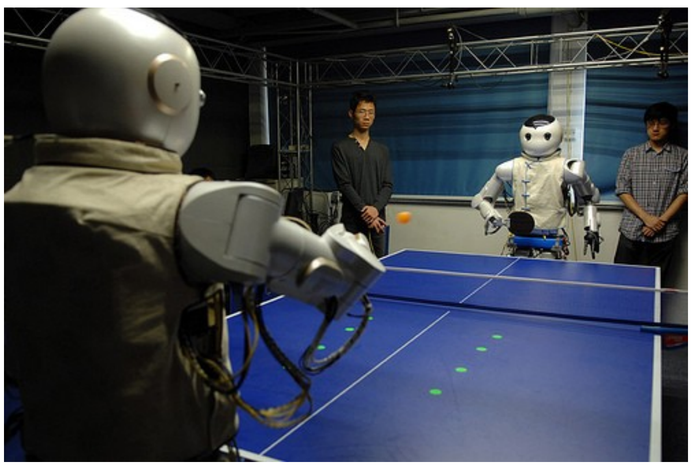


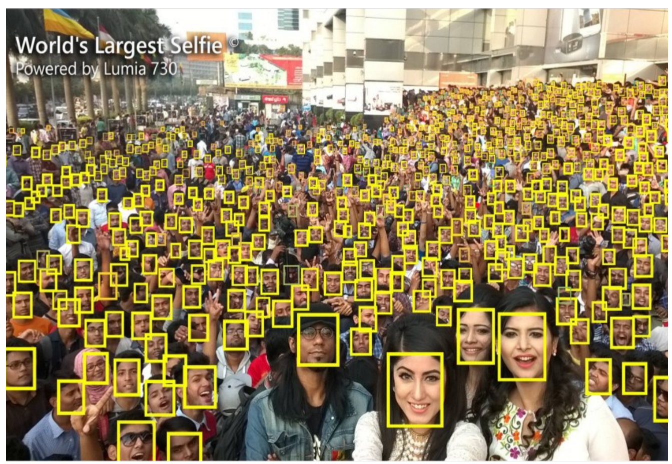



```
library(gtrendsR)
library(reshape2)
google.trends = gtrends(c("big data"), gprop = "web", time = "all")[[1]]
google.trends = dcast(google.trends, date ~ keyword + geo, value.var = "hits")
rownames(google.trends) = google.trends$date
plot(google.trends, type = "l")
google.trends = gtrends(c("machine learning"), gprop = "web", time = "all")[[1]]
google.trends = dcast(google.trends, date ~ keyword + geo, value.var = "hits")
rownames(google.trends) = google.trends$date
lines(google.trends, col = "blue")
                         ## Reproducir la figura anterior:
```

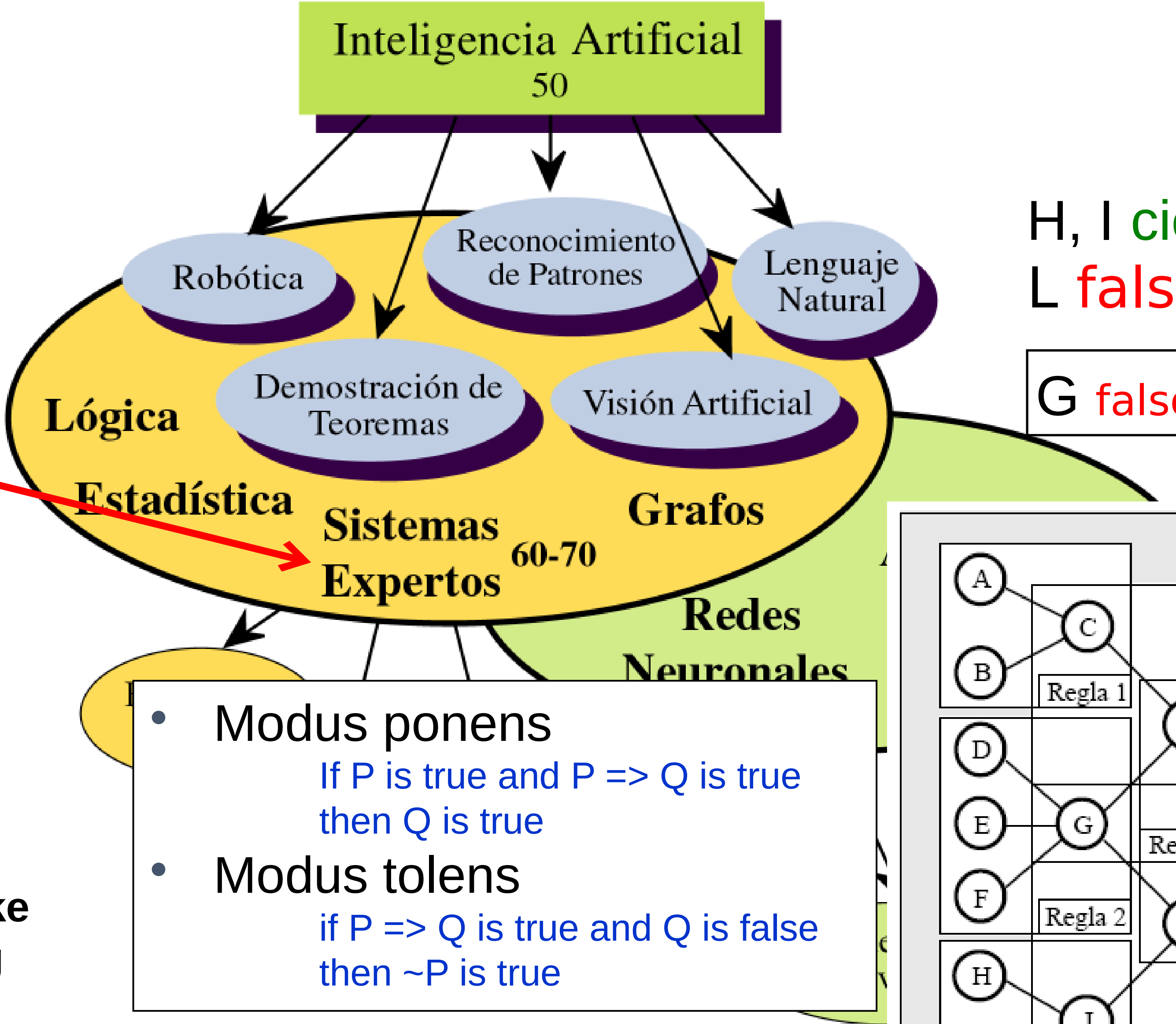
https://www.displayr.com/extracting-google-trends-data-in-r


Inteligencia Artificial
50
Robótica
Reconocimiento de Patrones
Lenguaje Natural
Demostración de Teoremas
Visión Artificial
Lógica
Estadística
Sistemas Expertos
60-70
Grafos




World's Largest Selfie
Powered by Lumia 730


1996
Games
Kasparov
Deep Blue
Subscribe
8
7
6
5
4
3
2
1
a b c d e f g h



DATA MINING 1990

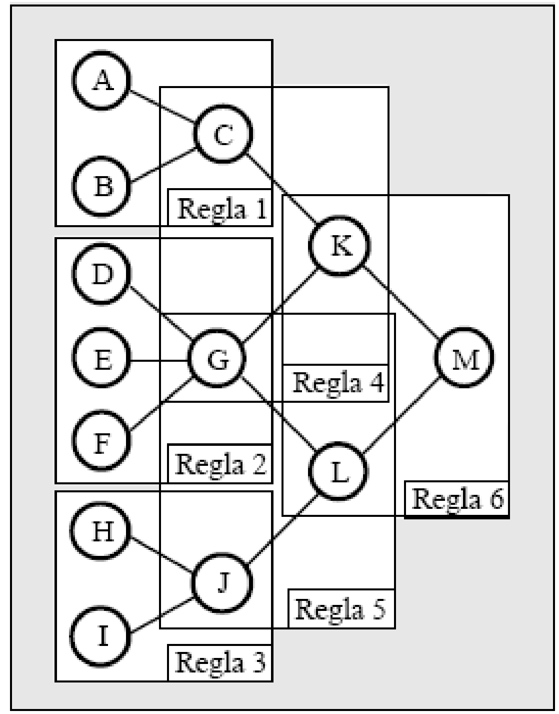
H, I cierto
L falso

G falso

Explicit representation of knowledge

Rules, graphs, etc.

- Modus ponens
  If P is true and P => Q is true then Q is true
- Modus tolens
  if P => Q is true and Q is false then ~P is true

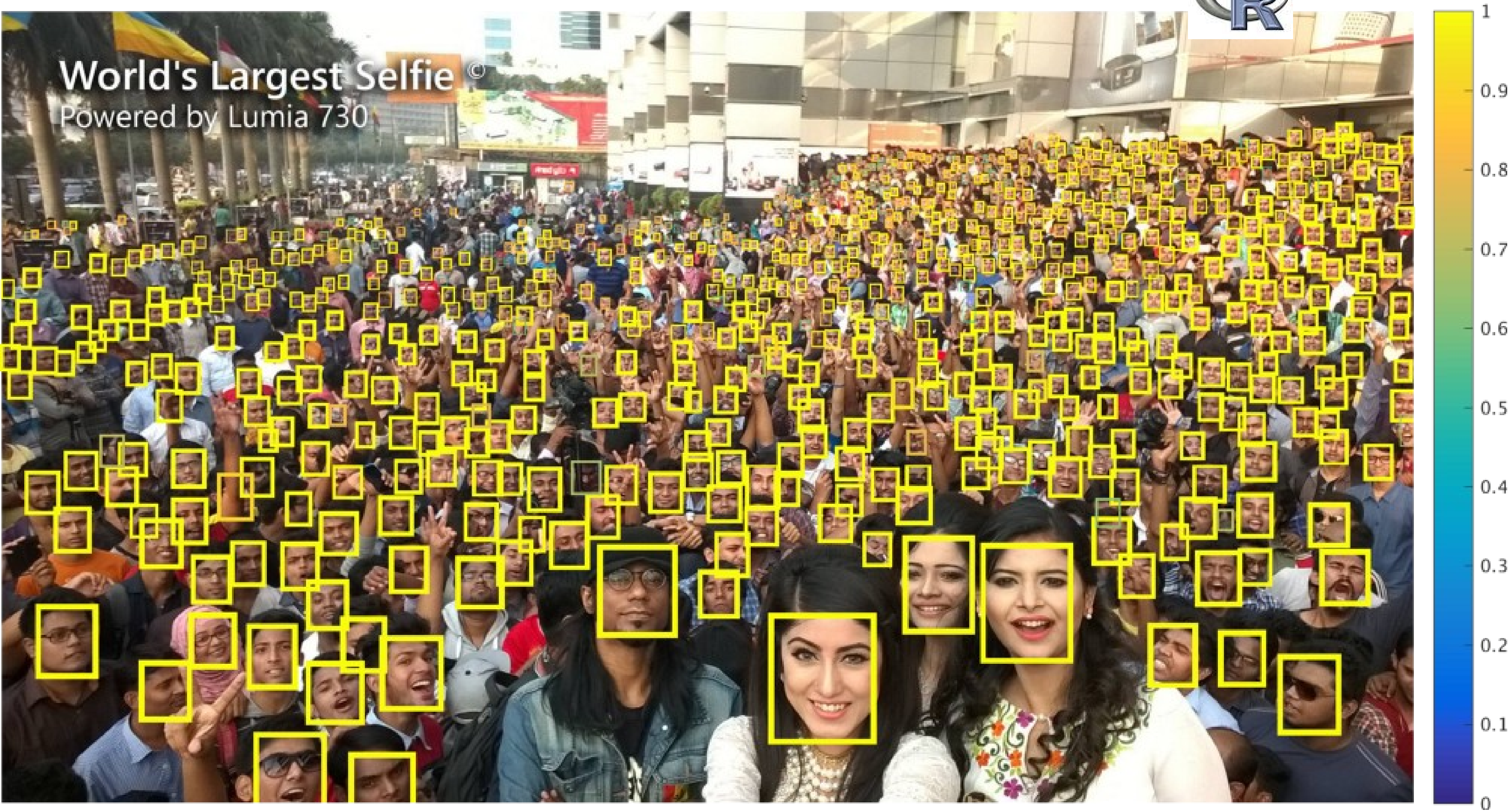


Human-like reasoning

Logical inference, look for relations on graphs, etc.

Serial processing

https://github.com/peiyunh/tiny



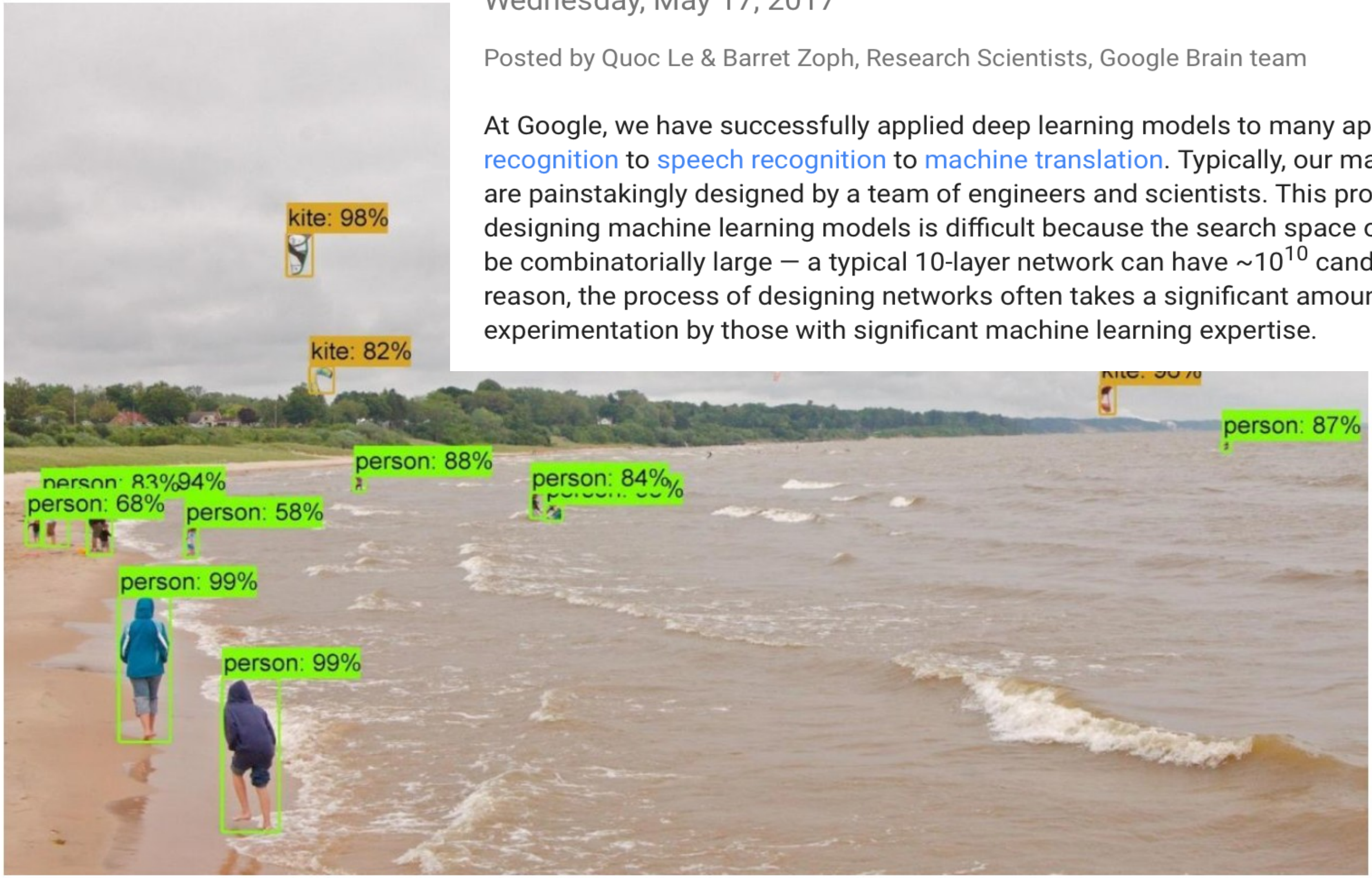
evelop a face detector (Tiny Face Detector) that can find ~800 faces out of ~1000 reportedly prese
aking use of novel characterization of scale, resolution, and context to find small objects.

Using Machine Learning to Explore Neural Network Architecture

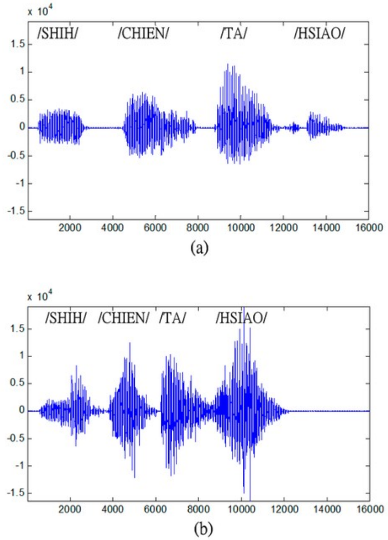
Wednesday, May 17, 2017

Posted by Quoc Le & Barret Zoph, Research Scientists, Google Brain team

At Google, we have successfully applied deep learning models to many applications, from image recognition to speech recognition to machine translation. Typically, our machine learning models are painstakingly designed by a team of engineers and scientists. This process of manually designing machine learning models is difficult because the search space of all possible models can be combinatorially large — a typical 10-layer network can have ~10^{10} candidate networks! For this reason, the process of designing networks often takes a significant amount of time and experimentation by those with significant machine learning expertise.

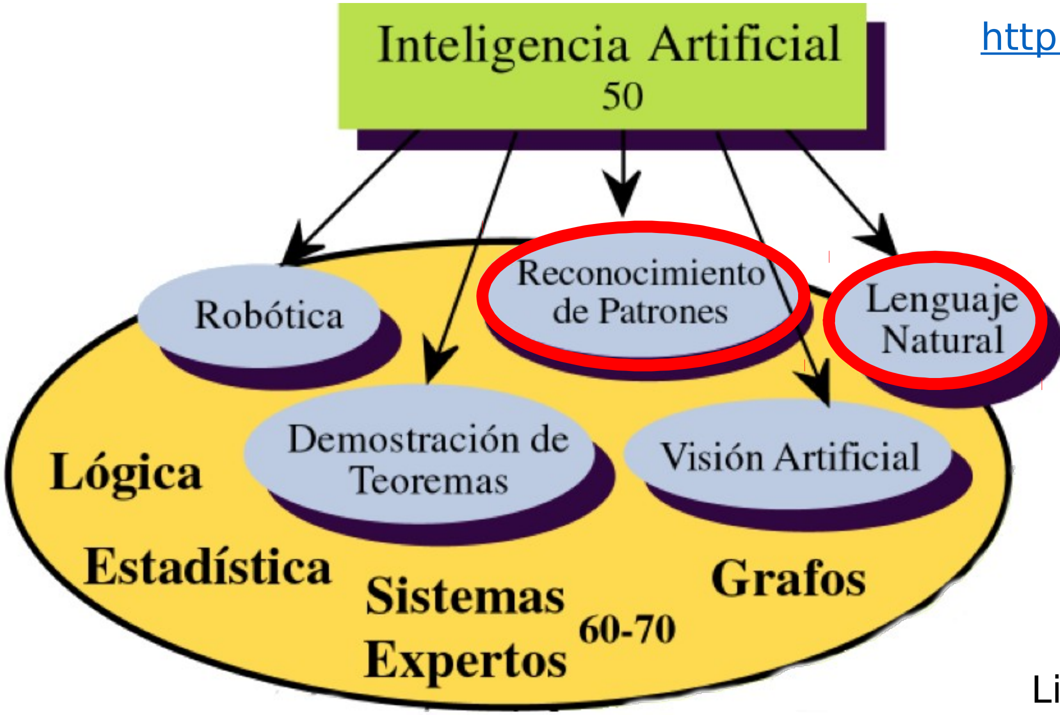


Un ejemplo de cómo identifica imágenes NASNet (Google Research)



R + openNLP

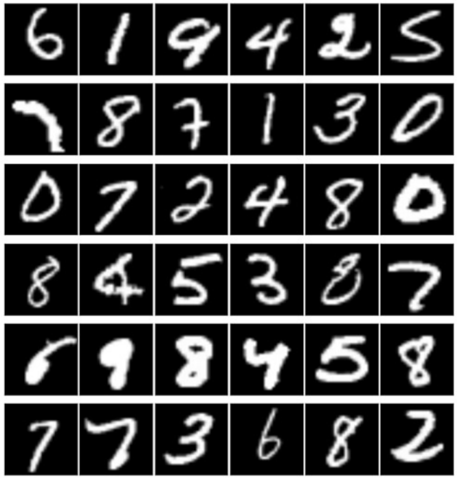
Overview of Natural Language Processing(NLP) with R and OpenNLP



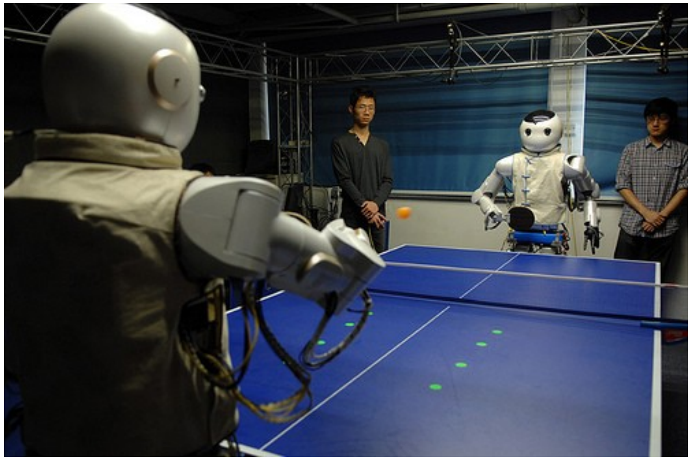
http://yann.lecun.com/exdb/mnist

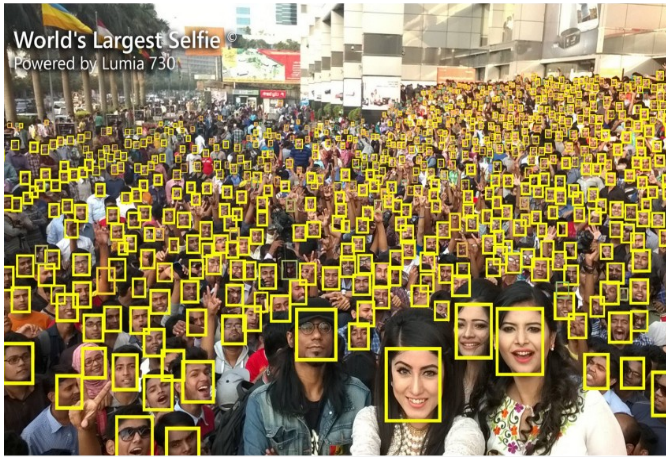
60000+10000 images 28x2

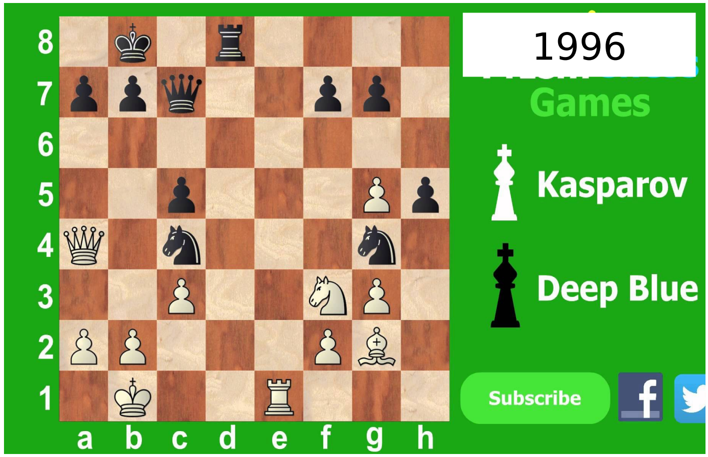
Labeled as {0,...,9



Lineal: 10%. k-NN: 3%. SVM: 1%. Deep: 0.3%



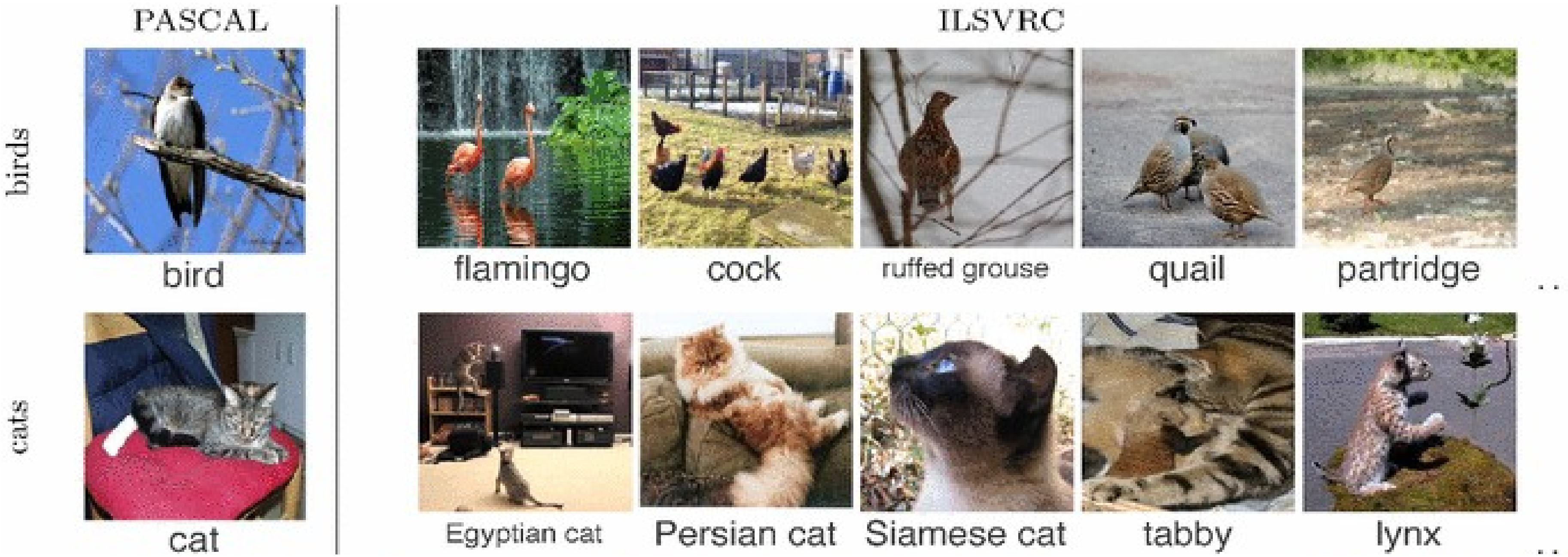




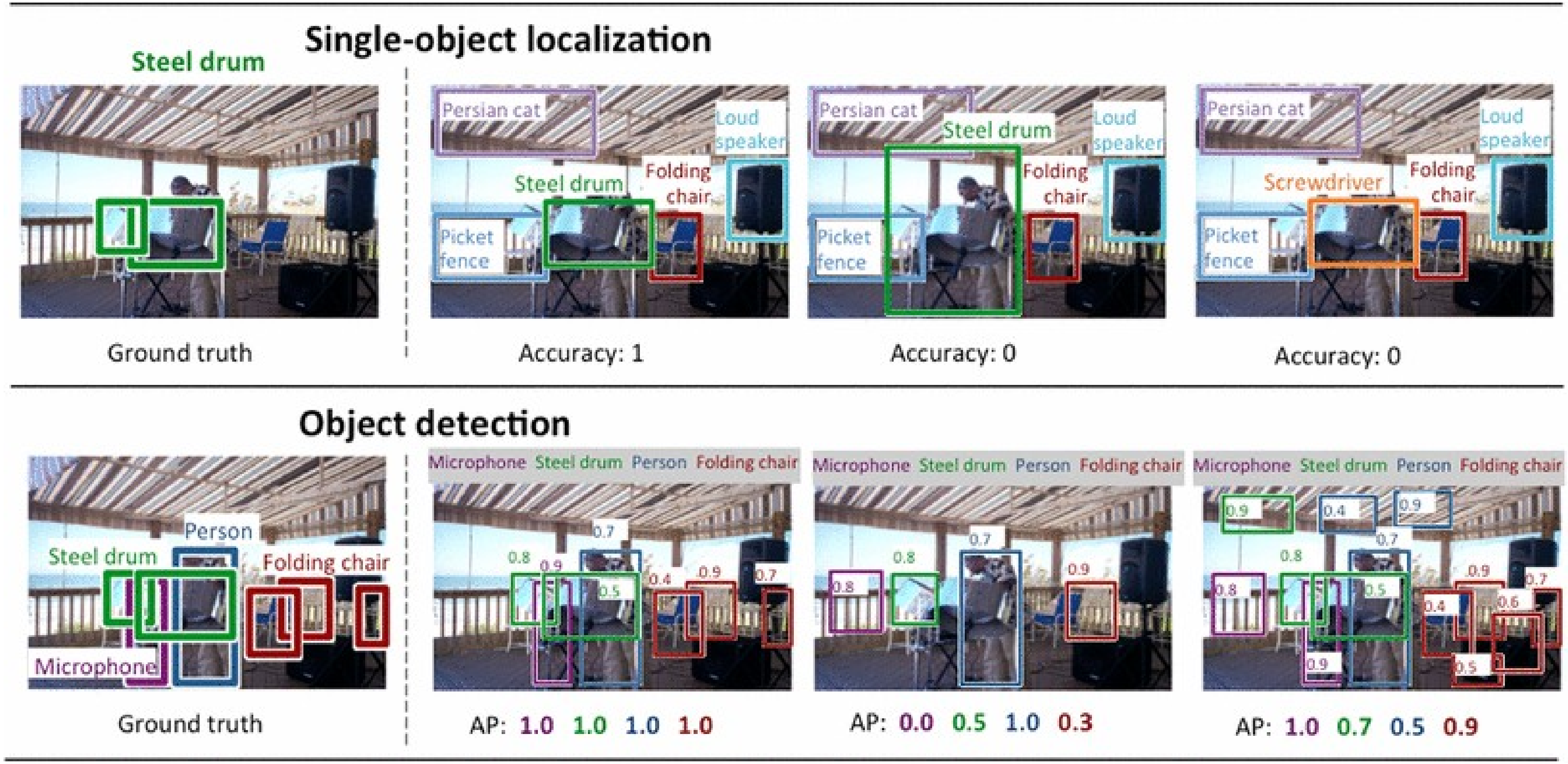
ImageNet is an image database organized according to the (nouns of the) [WordNet](#) hierarchy, in which each node of the hierarchy is depicted by an average of over five hundred images.

#synsets: 21841
#images: 14197122

150 GB [kaggle]



David G. Lowe, **Distinctive Image Features from Scale-Invariant Keypoints.** *International Journal of Computer Vision, 2004.*



Validation: top-5 error rate

2017 video included

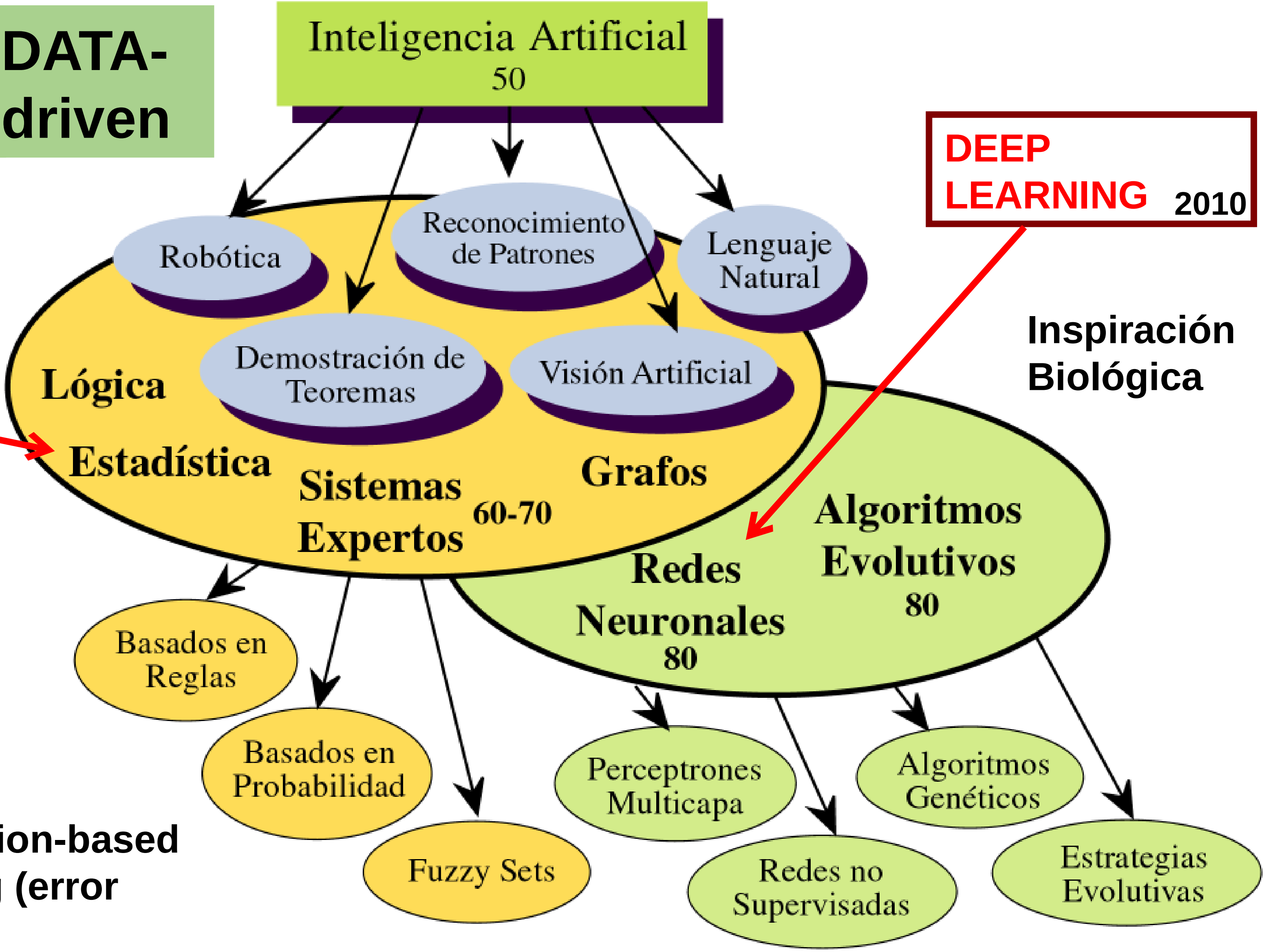
Inception-v3: 3.46% top-5 and 17.3% top-1 (25 million parameters). [Inception In kaggle]

O. Russakovsk (2015) ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision, 115, 211–252

Nuevos Paradigmas DATA-driven

Statistical Inspiration

STATISTICAL LEARNING 2000



Data driven using abstract representations

Kernels,
neural network, etc.

Optimization-based reasoning (error function).

Empirical risk,
gradient descend, etc.

Parallel processing

HPC, GPUs, cloud, etc.

the non trivial extraction of implicit, previously unknown, and potentially useful information from data

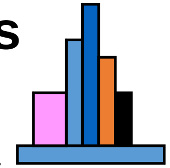
W. Frawley and G. ***Piatetsky-Shapiro*** *and C. Matheus, Knowledge Discovery in Databases: An Overview.*

AI Magazine, Fall ***1992****, 213-228.*

KDnuggets

AI & Machine learning

Knowledge Patterns

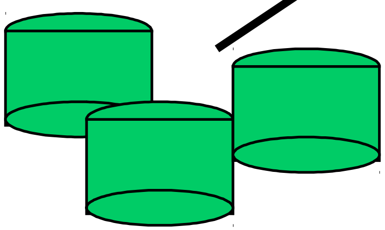


DATA MINING

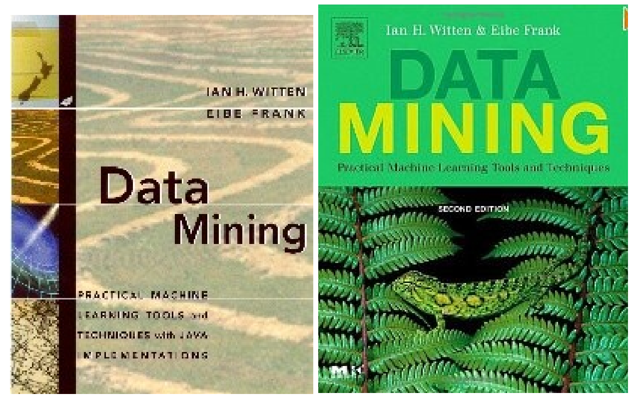
Task-relevant Data

Data Warehouse

Data Cleaning



Databases



Data Mining: Practical Machine Learning Tools and Techniques with Java Impleméntatenos

*Ian H. Witten, Eibe Frank (****1999****)*

WEKA The University of Waikato

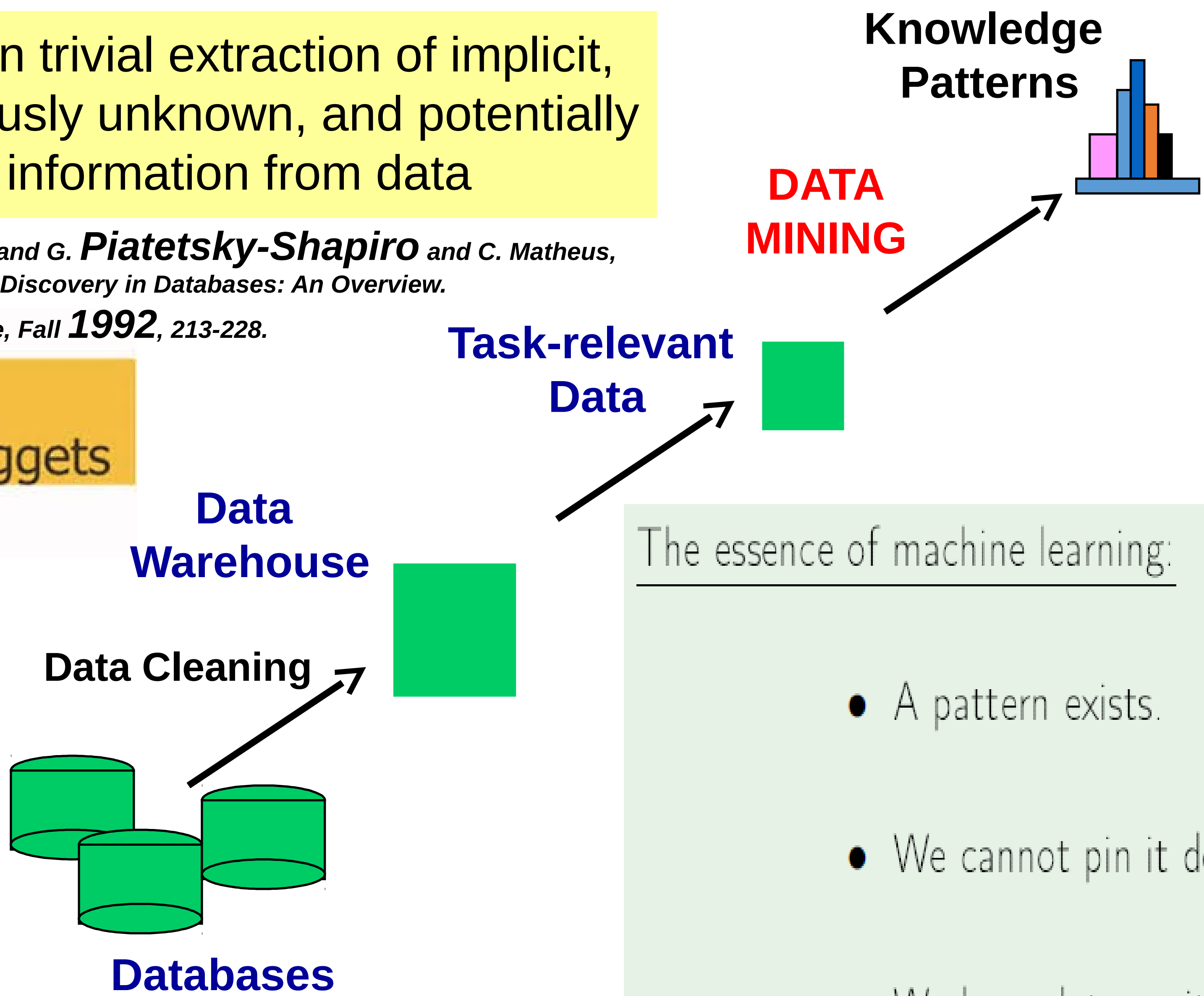
Machine Learning and Data Mining Open Soure Tools in Java

`http://www.cs.waikato.ac.nz/~ml/weka/`

the non trivial extraction of implicit, previously unknown, and potentially useful information from data

W. Frawley and G. ***Piatetsky-Shapiro*** *and C. Matheus, Knowledge Discovery in Databases: An Overview.*
AI Magazine, Fall ***1992****, 213-228.*

KDnuggets



AI & Machine learning

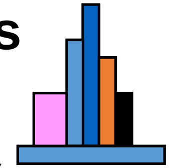
The essence of machine learning:

- A pattern exists.
- We cannot pin it down mathematically.
- We have data on it.

the non trivial extraction of implicit, previously unknown, and potentially useful information from data

S. Bryson et al., Visually exploring gigabyte data sets in real time.
Communications of the ACM, 42, 82-90,
Aug. ***1999***

Knowledge Patterns



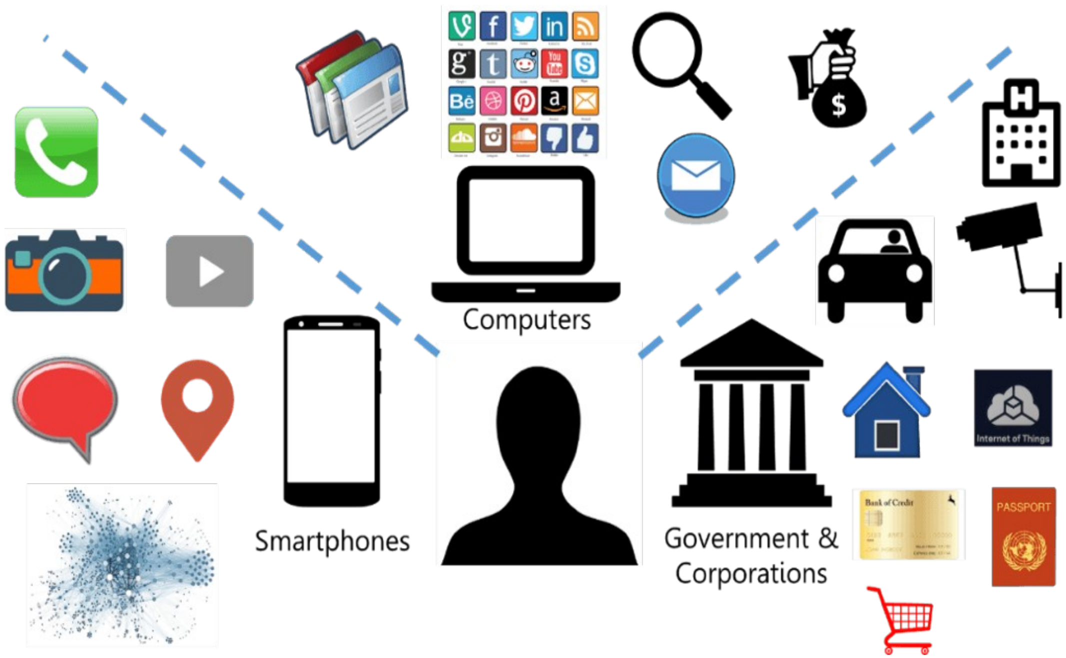
AI & Machine learning

DATA ANALYTICS

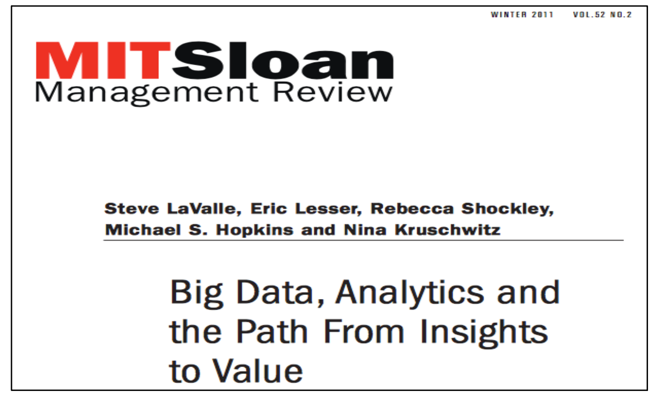
Data Discovery, Cleaning and Reduction

Modeling and Prediction

Big data
(integration of heterogeneous real-time sources)



2011


WINTER 2011 VOL.52 NO.2
MITSloan Management Review
Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S. Hopkins and Nina Kruschwitz
Big Data, Analytics and the Path From Insights to Value

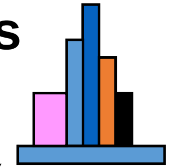
2014


Raghupathi and Raghupathi *Health Information Science and Systems* 2014, **2**:3
http://www.hissjournal.com/content/2/1/3
HEALTH INFORMATION SCIENCE AND SYSTEMS
REVIEW Open Access
Big data analytics in healthcare: promise and potential
Wullianallur Raghupathi[1*] and Viju Raghupathi[2]

the non trivial extraction of implicit, previously unknown, and potentially useful information from data

S. Bryson et al., Visually exploring gigabyte data sets in real time. Communications of the ACM, 42, 82-90, Aug. 1999

Knowledge Patterns



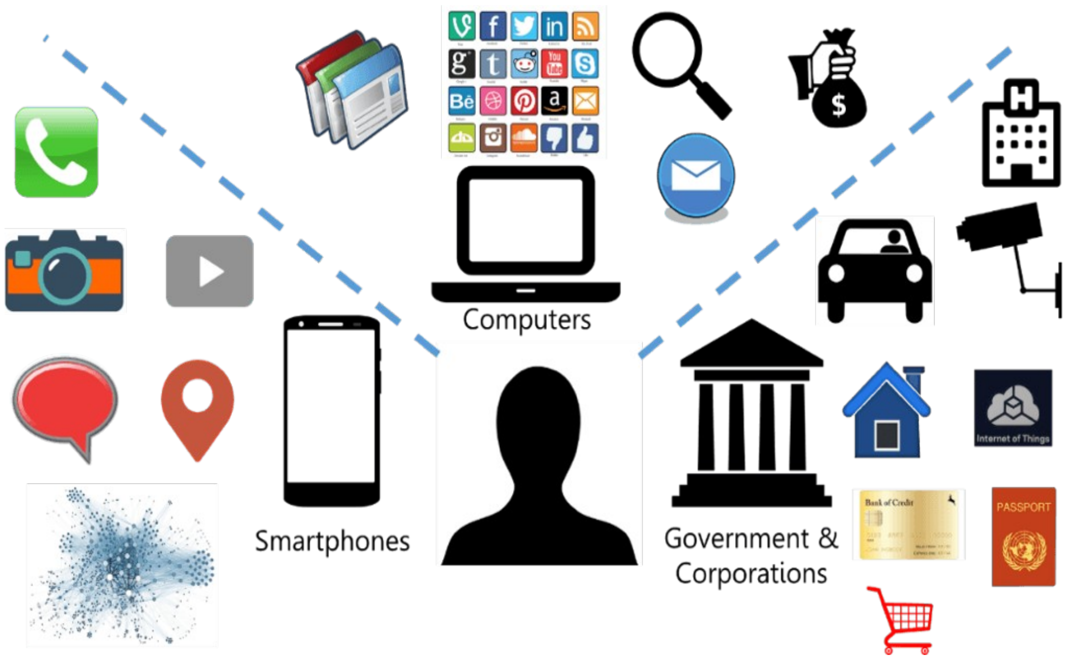
AI & Machine learning

DATA ANALYTICS

Data Discovery, Cleaning and Reduction

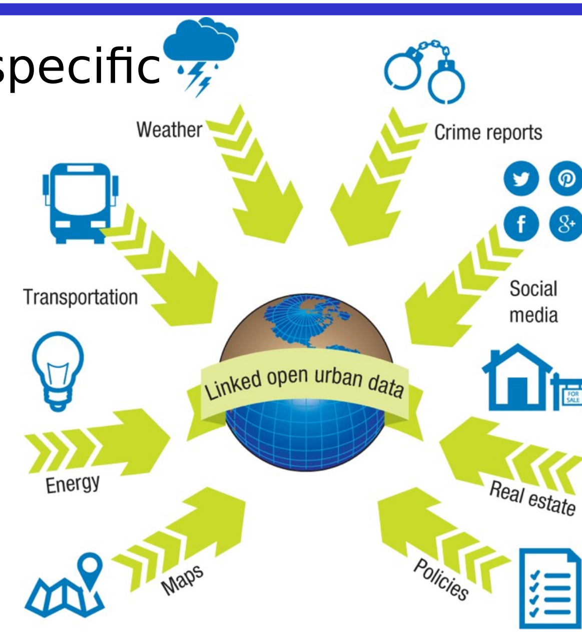
Modeling and Prediction

Big data
(integration of heterogeneous real-time sources)



Domain-specific Problems

DataLabs



Financiero Seguros

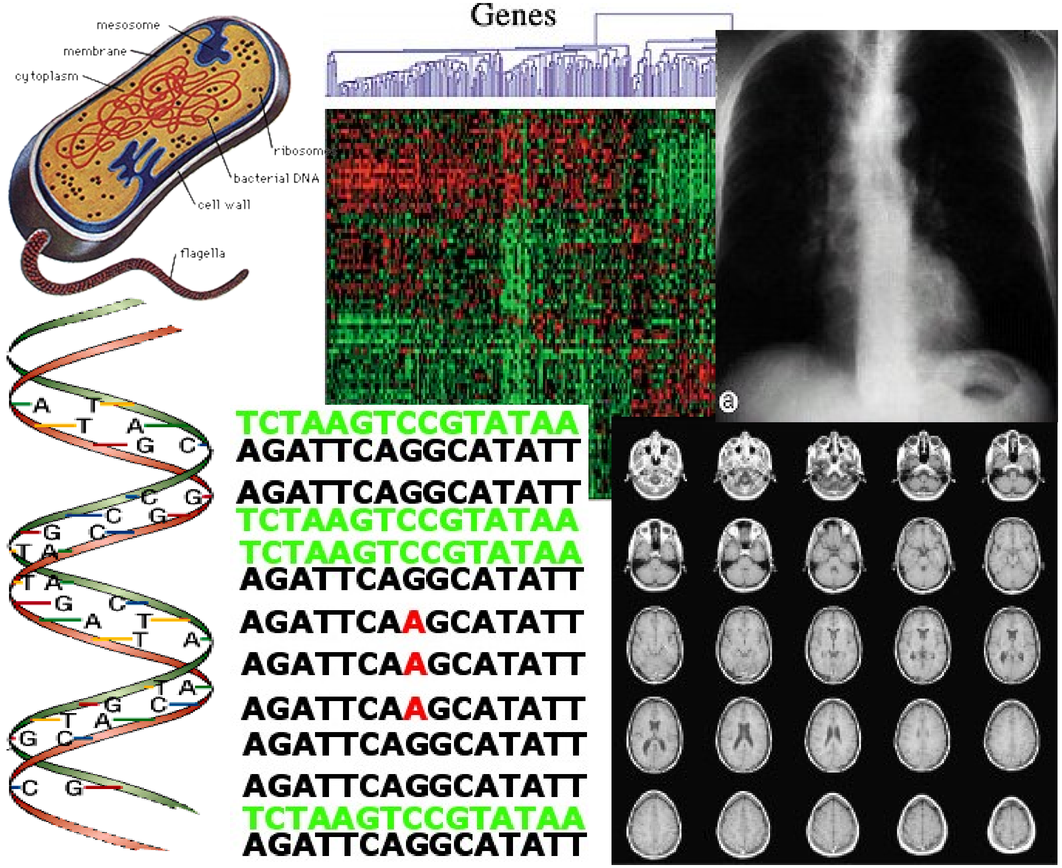
Comercio y marketing

Industria y gestión empresarial

Tecnologías información y comunicaciones

Sanitario y farmacéutico

Meteorología, clima y medio ambiente

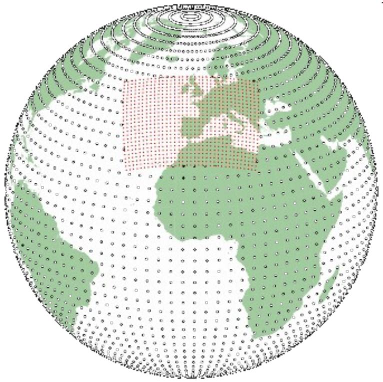


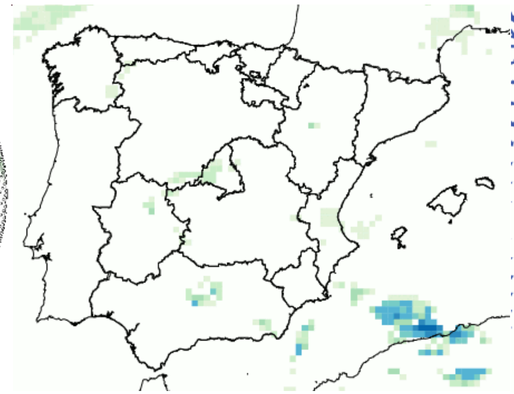
El SNS genera **5 millones** de altas hospitalarias al año almacenando datos sobre diagnósticos y procedimientos asociados a cada paciente que contienen información necesaria para la **gestión** del SNS.

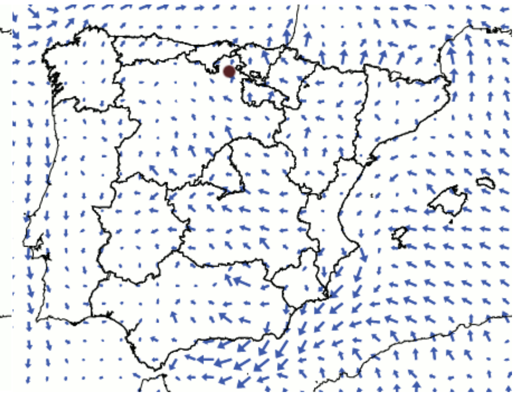
http://icmbd.es/

Financiero Seguros	Comercio y marketing	Industria y gestión empresarial	Tecnologías información y comunicaciones	Sanitario y farmacéutico	Meteorología, clima y medio ambiente

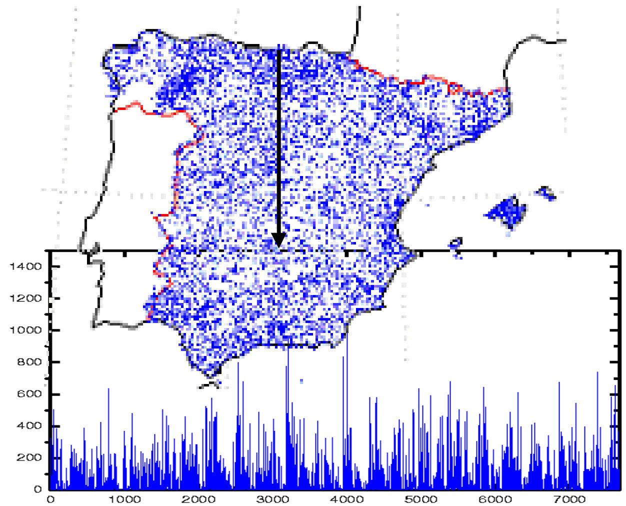
Las observaciones y simulaciones globales y regionales del clima generan **cientos de TB** de información heterogénea necesaria para la predicción meteorológica.

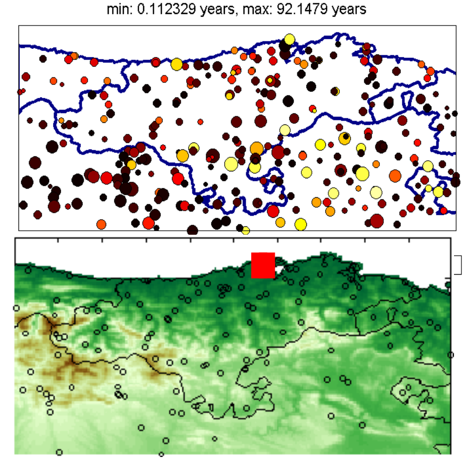




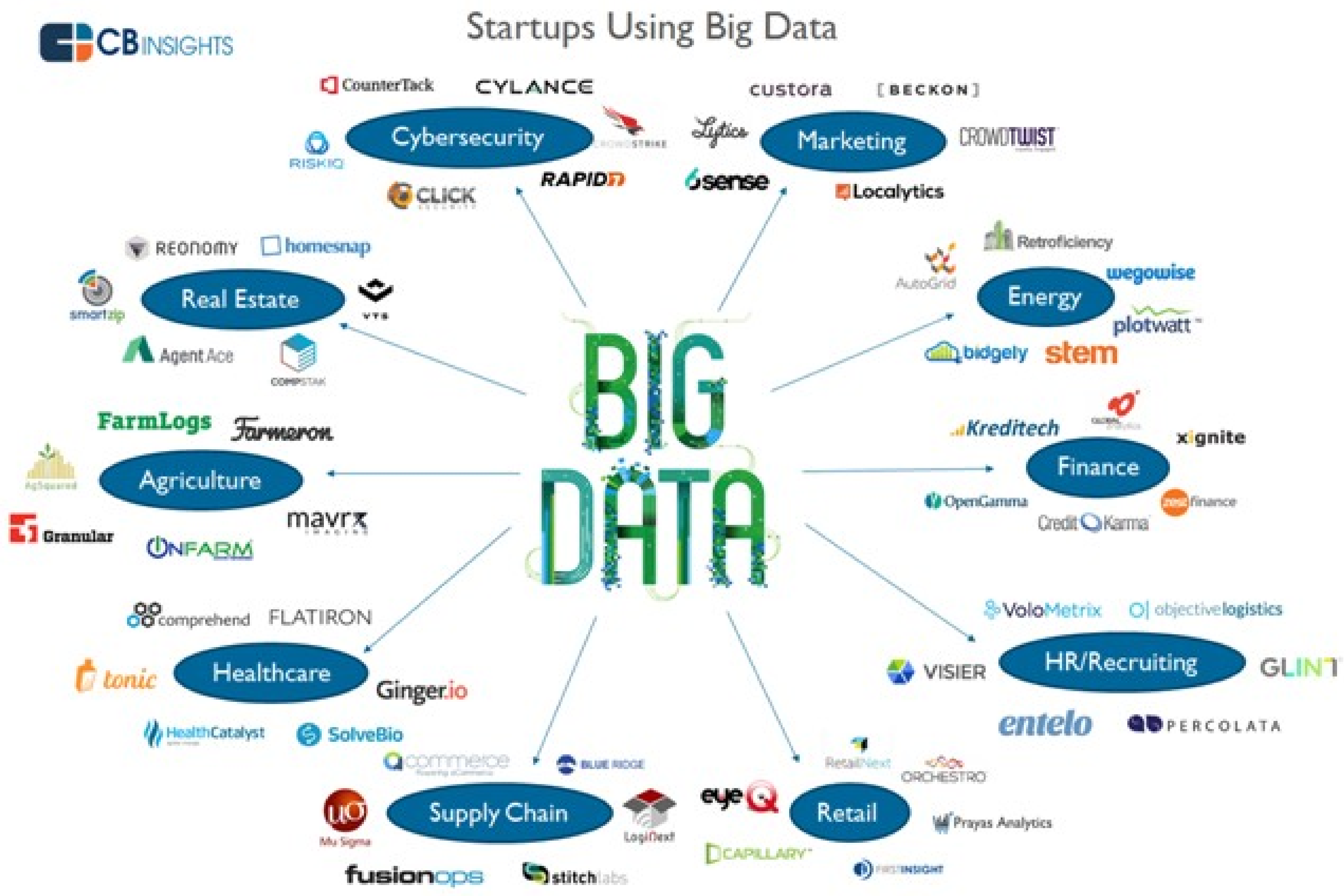


860101500000000010860101500000000010
860102 0000010
860103 0000010
860104 0000010
860105 0000010
860106 0000010
860107 0000010
860108 0000010
860109 0001000
860110 1001100
860111001000000000860111001000000000

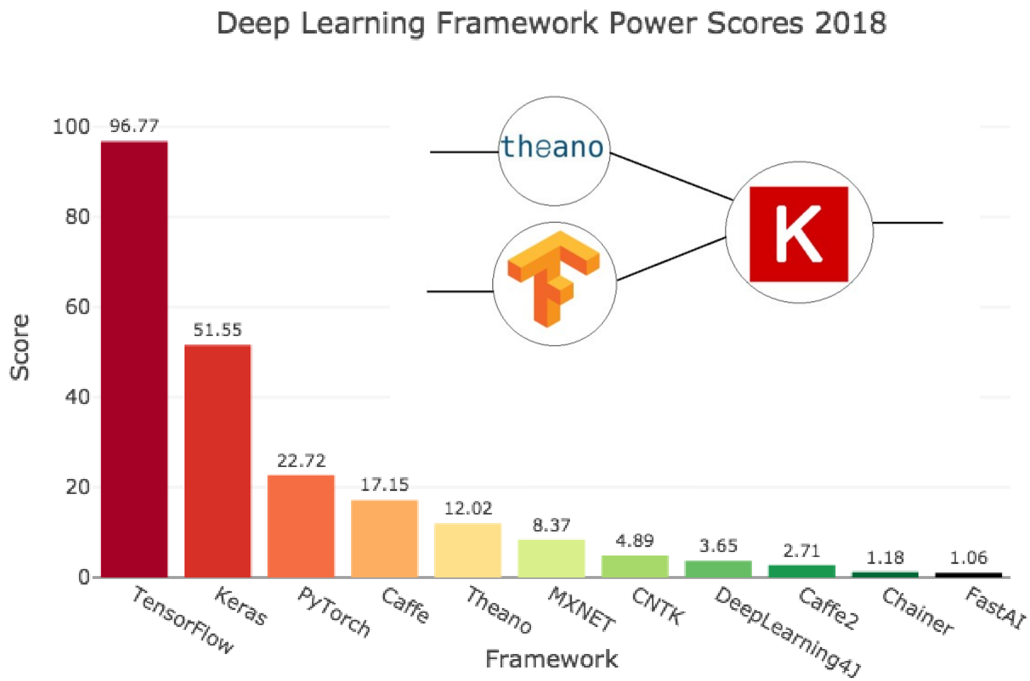


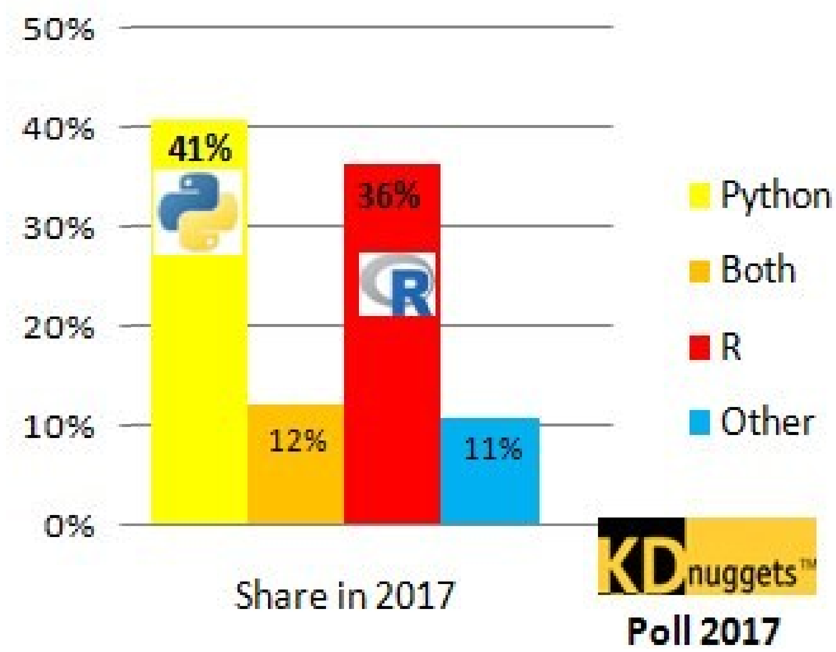


http://www.meteo.unican.es/downscaling


CB INSIGHTS
Startups Using Big Data
BIG DATA
Cybersecurity
CounterTack
CYLANCE
CROWDSTRIKE
RISKIQ
RAPID7
CLICK
Marketing
custora
[BECKON]
Lyftics
CROWDTWIST
6sense
Localytics
Real Estate
REONOMY
homesnap
smartzip
VTS
Agent Ace
COMPSTAK
Energy
Retroficiency
AutoGrid
wegowise
plotwatt
bidgely
stem
Agriculture
FarmLogs
Farmeron
AgSquared
mavrx
Granular
ONFARM
Finance
Kreditech
xignite
OpenGamma
zest finance
Credit Karma
Healthcare
comprehend
FLATIRON
tonic
Ginger.io
HealthCatalyst
SolveBio
HR/Recruiting
VoloMetrix
objectivelogistics
VISIER
GLINT
entelo
PERCOLATA
Supply Chain
commerce
BLUE RIDGE
Mu Sigma
LogiNext
fusionops
stitchlabs
Retail
RetailNext
ORCHESTRO
eyeQ
Prayas Analytics
CAPILLARY
FIRSTINSIGHT

A key factor for the quick growth of data science is the efficient frameworks (and infrastructures) available:



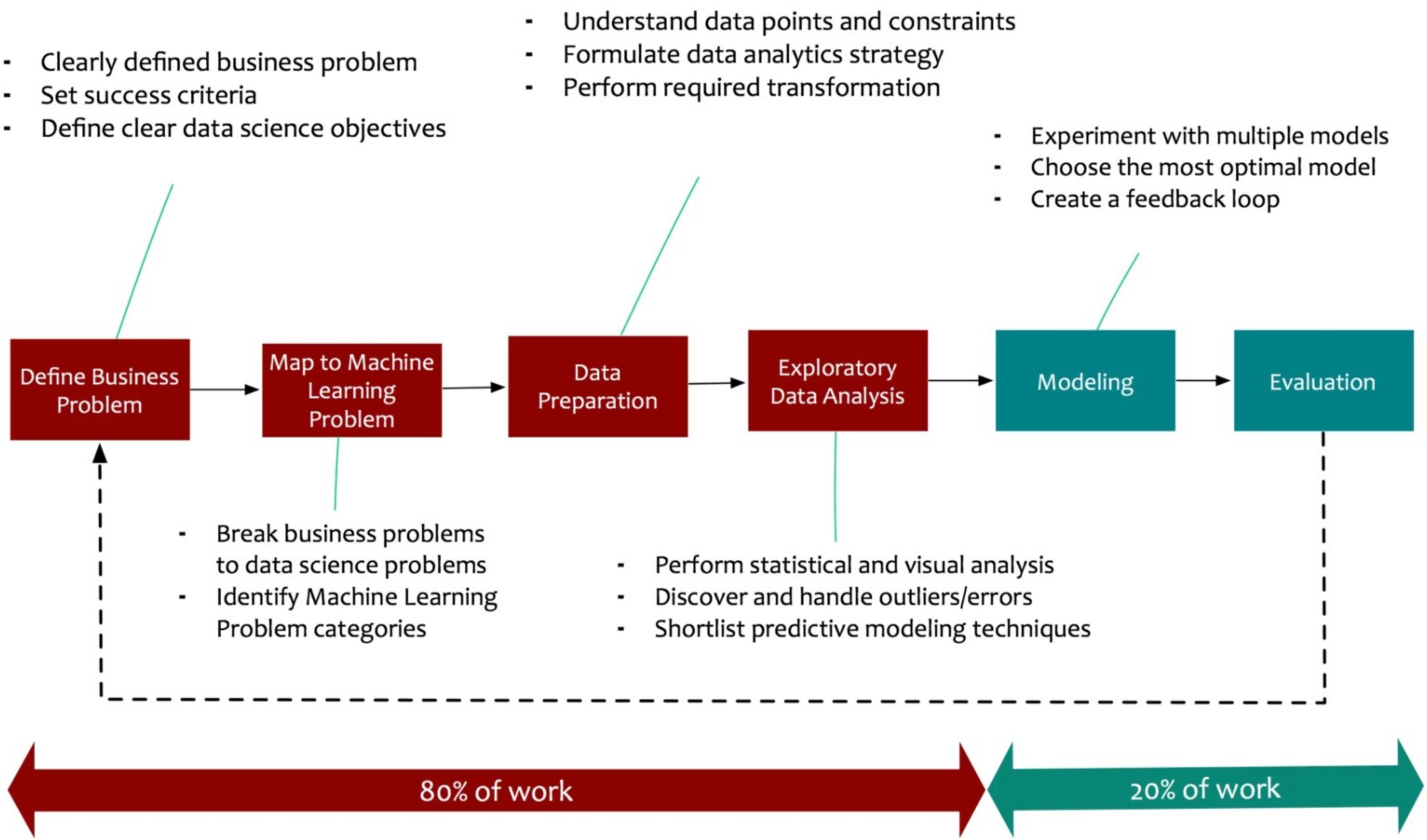


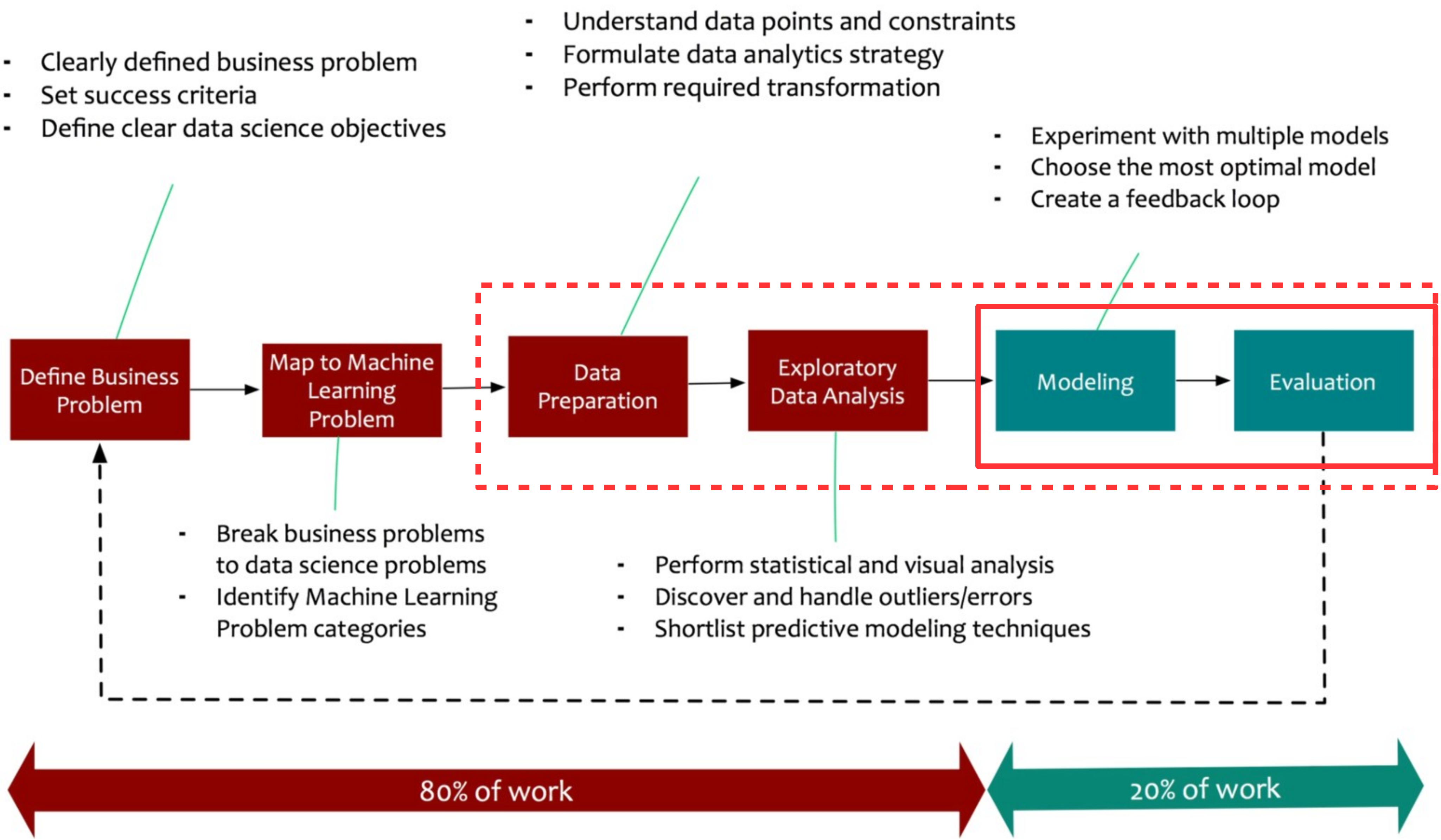
https://www.kdnuggets.com/2017/09/datacamp-keras-cheat-sheet-deep-learning-python.html

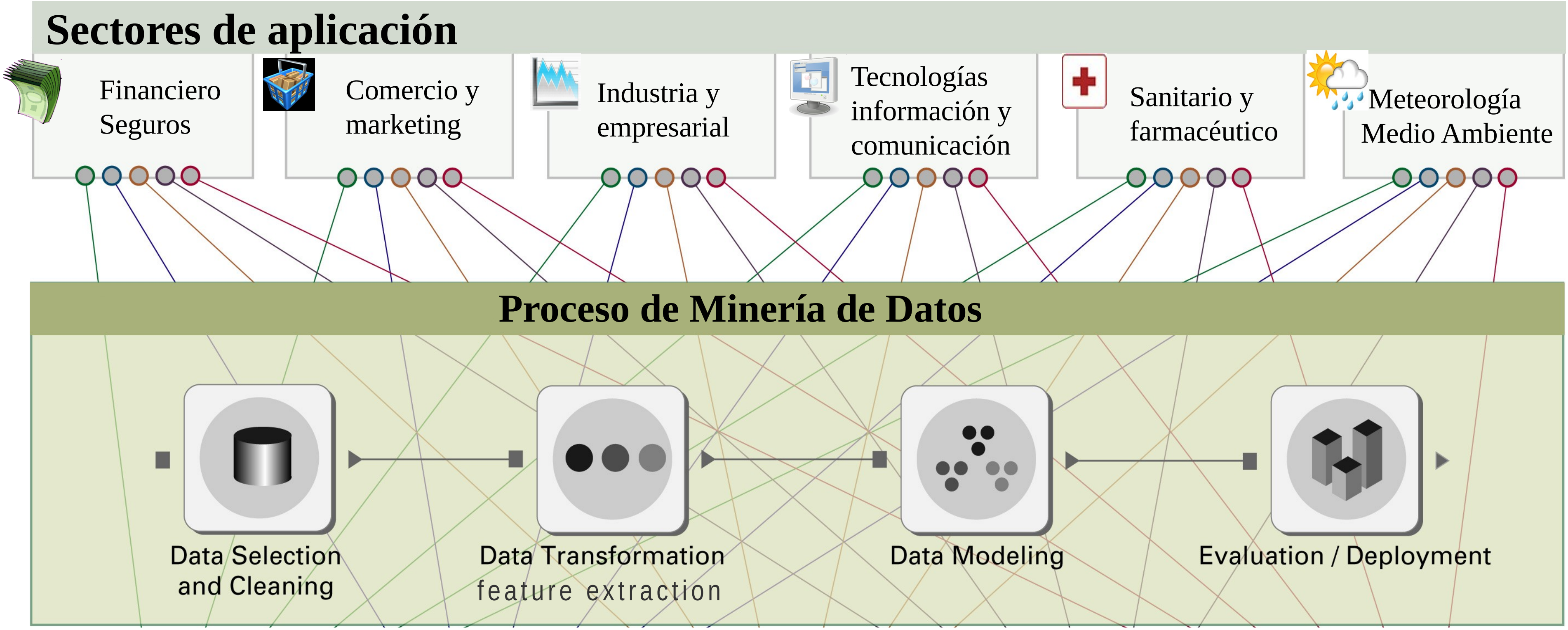
https://project.inria.fr/deeplearning/files/2016/05/DLFrameworks.pdf

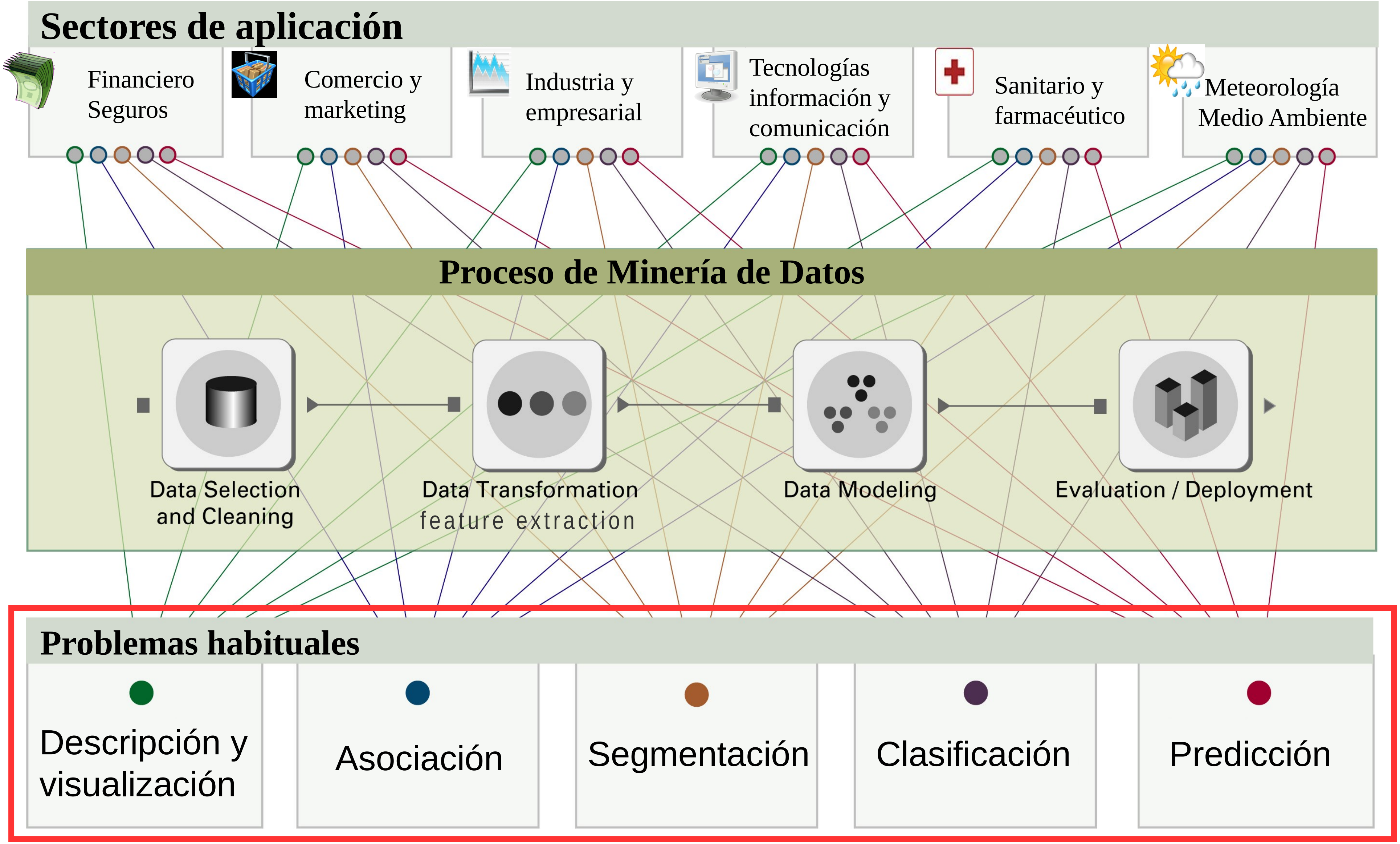
https://www.kdnuggets.com/2018/09/deep-learning-framework-power-scores-2018.html

https://github.com/amueller/scipy_2015_sklearn_tutorial

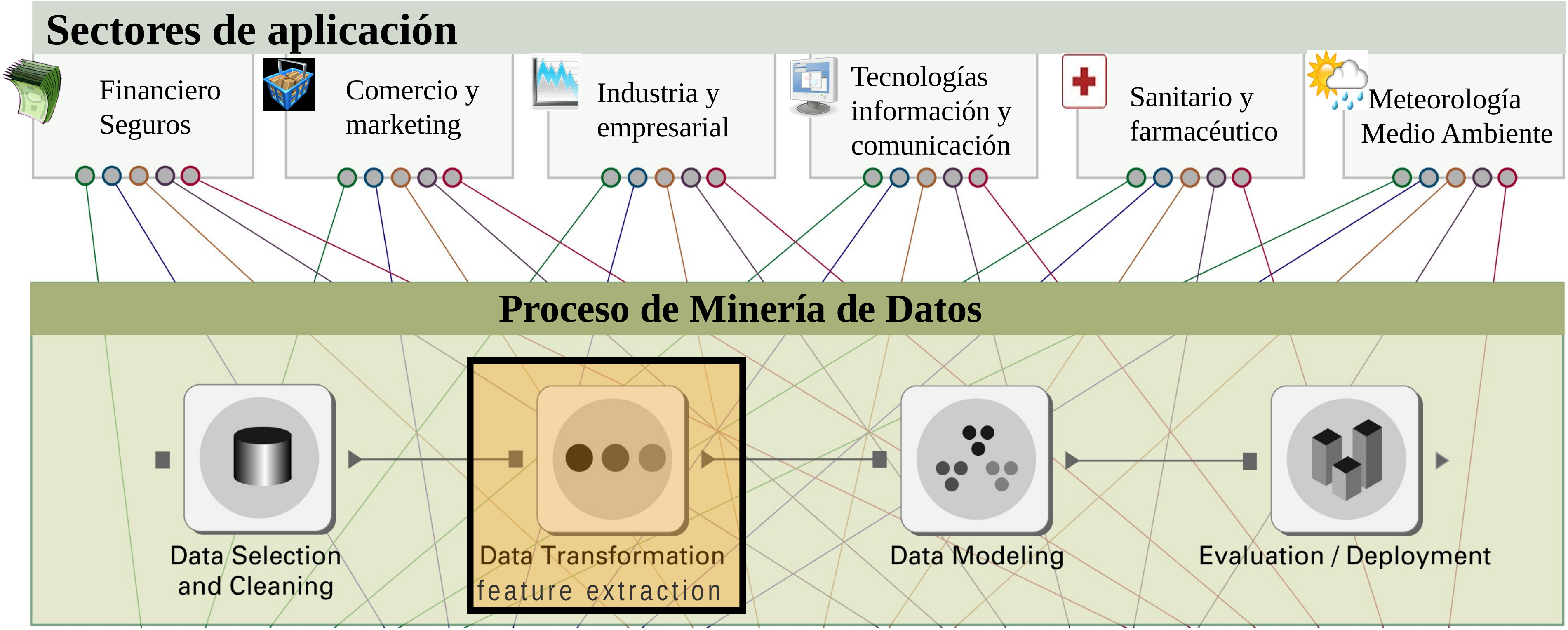

- Clearly defined business problem
- Set success criteria
- Define clear data science objectives
- Understand data points and constraints
- Formulate data analytics strategy
- Perform required transformation
- Experiment with multiple models
- Choose the most optimal model
- Create a feedback loop
Define Business Problem
Map to Machine Learning Problem
Data Preparation
Exploratory Data Analysis
Modeling
Evaluation
- Break business problems to data science problems
- Identify Machine Learning Problem categories
- Perform statistical and visual analysis
- Discover and handle outliers/errors
- Shortlist predictive modeling techniques
80% of work
20% of work


- Clearly defined business problem
- Set success criteria
- Define clear data science objectives
- Understand data points and constraints
- Formulate data analytics strategy
- Perform required transformation
- Experiment with multiple models
- Choose the most optimal model
- Create a feedback loop
Define Business Problem
Map to Machine Learning Problem
Data Preparation
Exploratory Data Analysis
Modeling
Evaluation
- Break business problems to data science problems
- Identify Machine Learning Problem categories
- Perform statistical and visual analysis
- Discover and handle outliers/errors
- Shortlist predictive modeling techniques
80% of work
20% of work


Sectores de aplicación
Financiero Seguros
Comercio y marketing
Industria y empresarial
Tecnologías información y comunicación
Sanitario y farmacéutico
Meteorología Medio Ambiente
Proceso de Minería de Datos
Data Selection and Cleaning
Data Transformation
feature extraction
Data Modeling
Evaluation / Deployment



Machine learning develop methods for data modelling and prognosis.



WIREs DATA MINING AND KNOWLEDGE DISCOVERY

Explore this journal >

Advanced Review

Data discretization: taxonomy and big data challenge

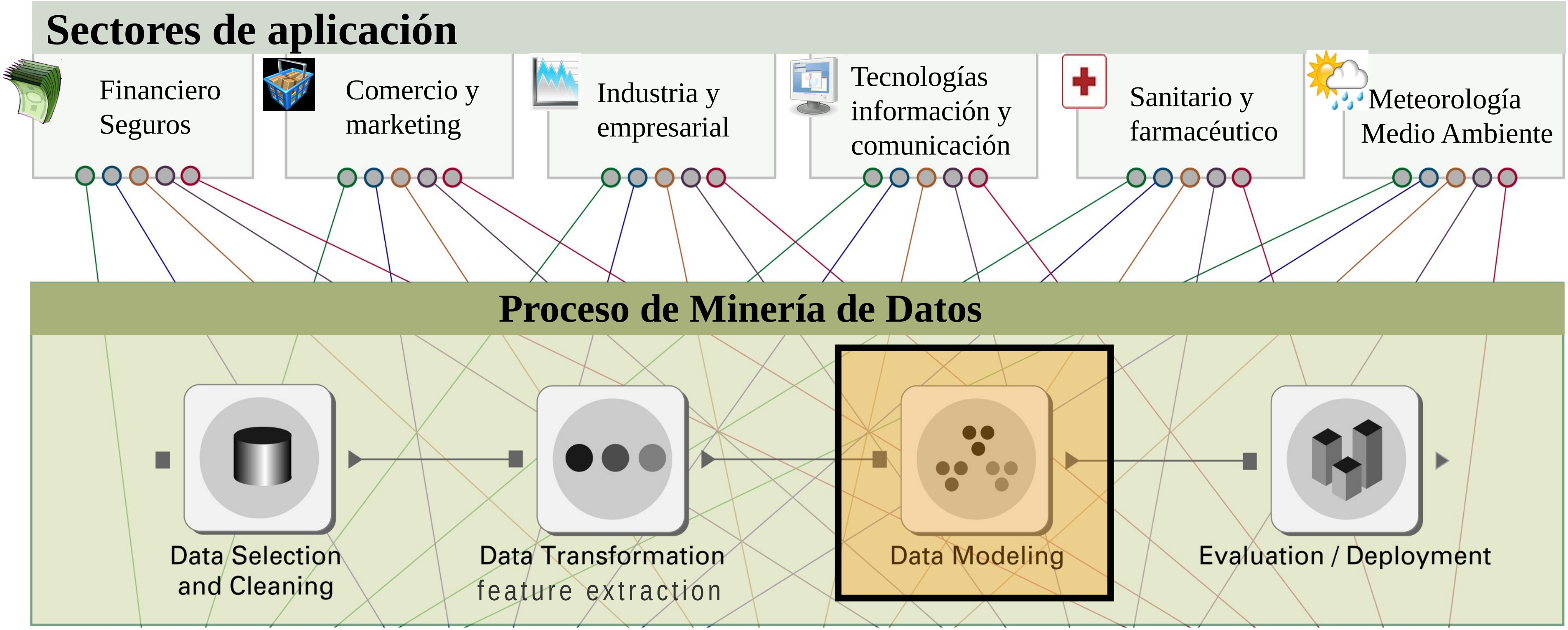
Knowledge-Based Systems

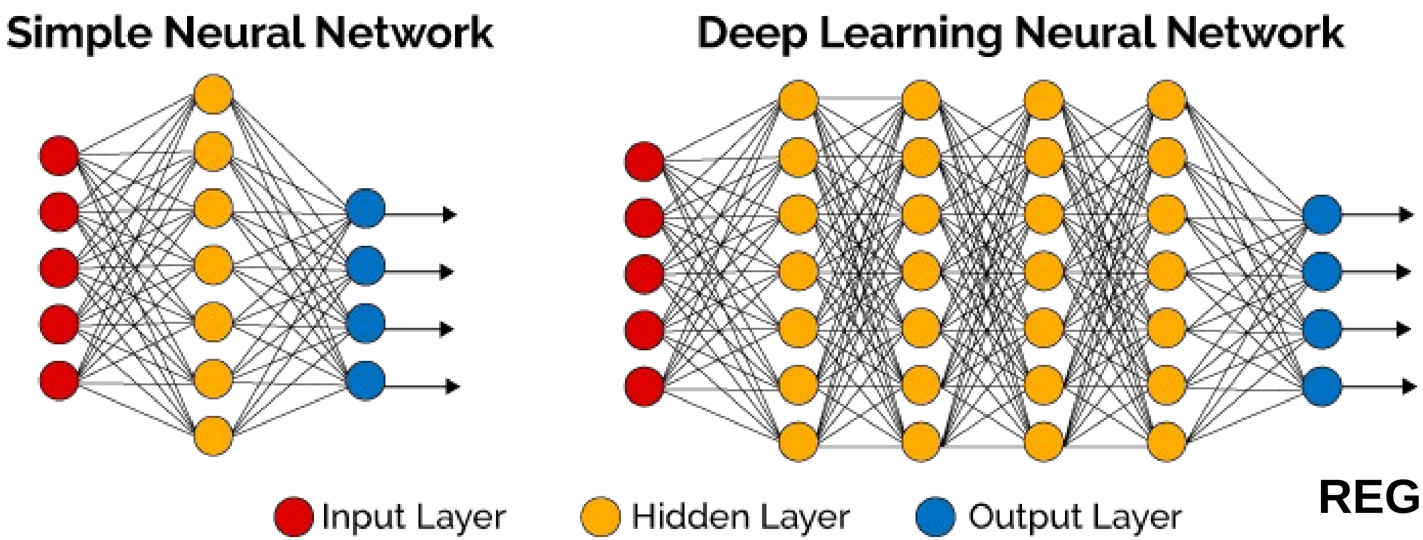
Volume 86, September 2015, Pages 33-45

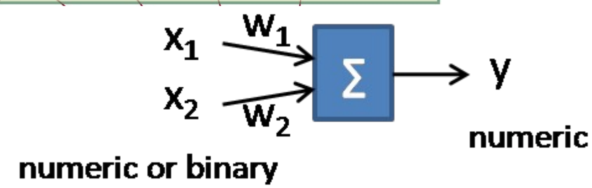
Recent advances and emerging challenges of feature selection in the context of big data

V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos

http://onlinelibrary.wiley.com/doi/10.1002/widm.1173/full



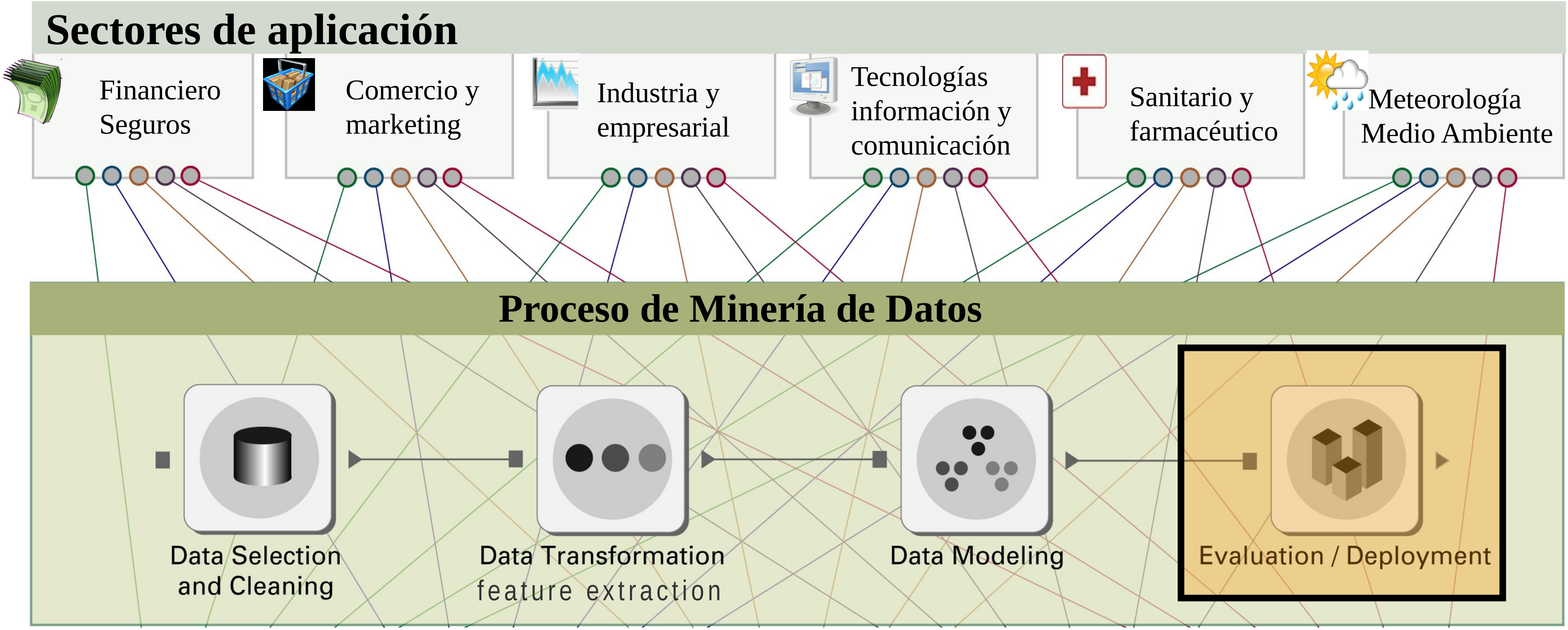


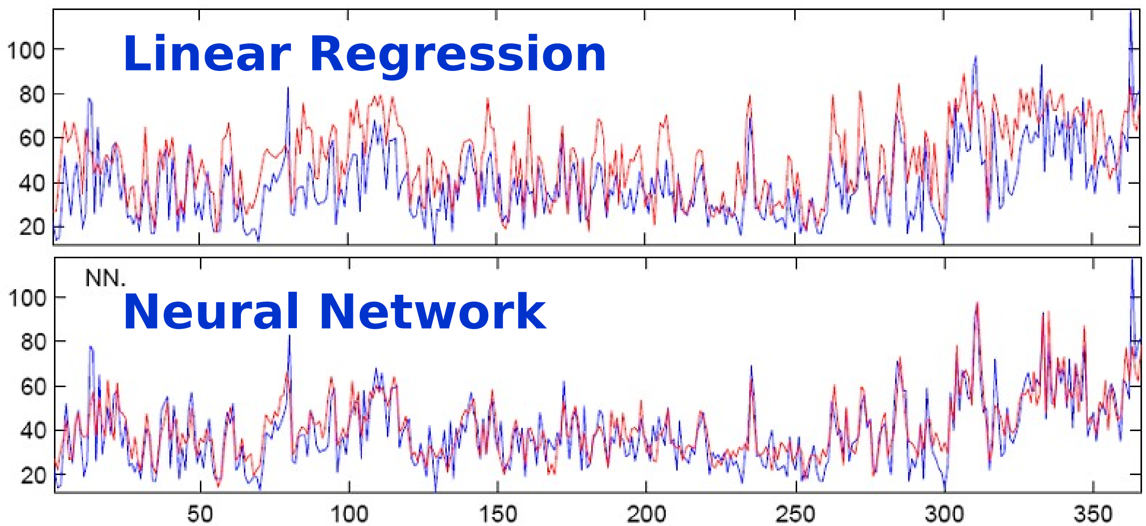


$$y = w_0 + w_1 x_1 + w_2 x_2$$

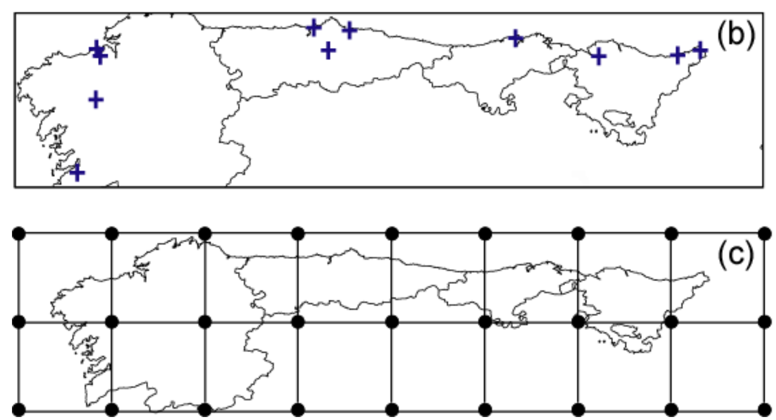
$$y = f(\mathbf{X},\mathbf{W}) = \mathbf{X}^T.\mathbf{W}$$

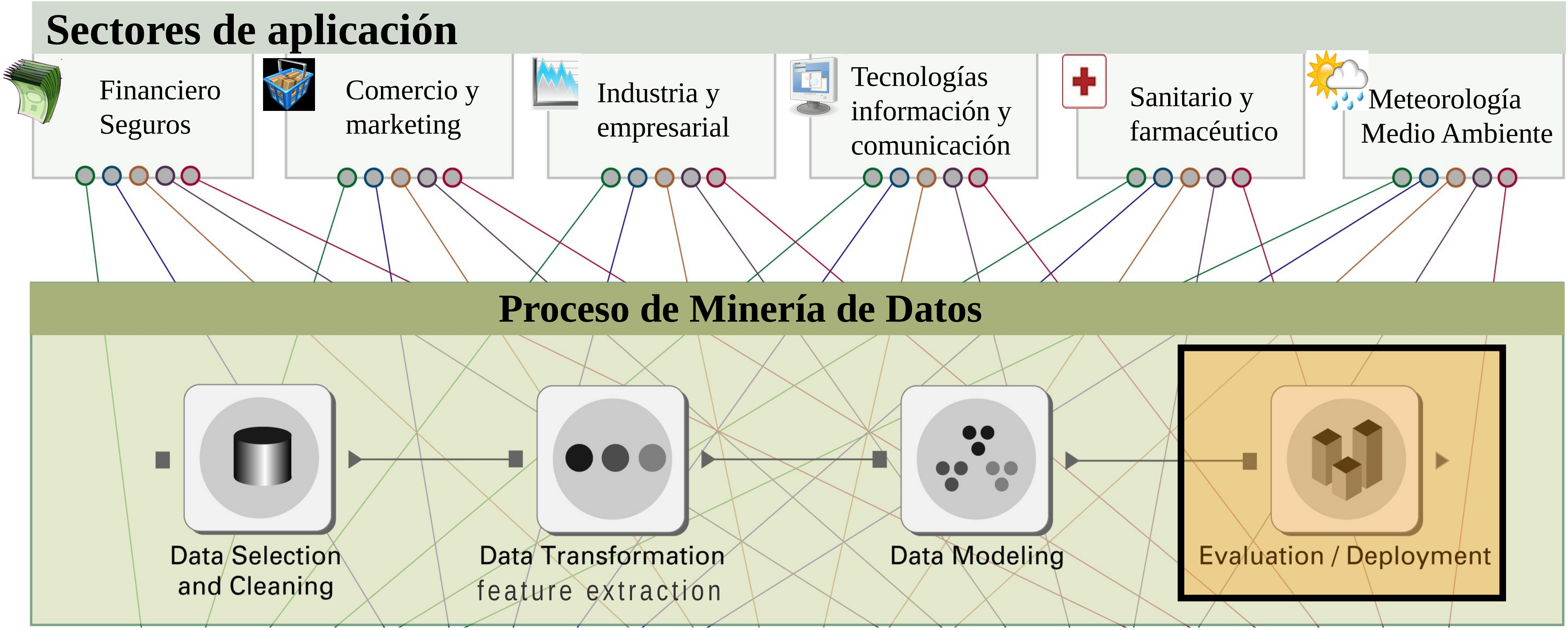
REGRESSION $\mathbf{W} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix}$ $\mathbf{X} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$

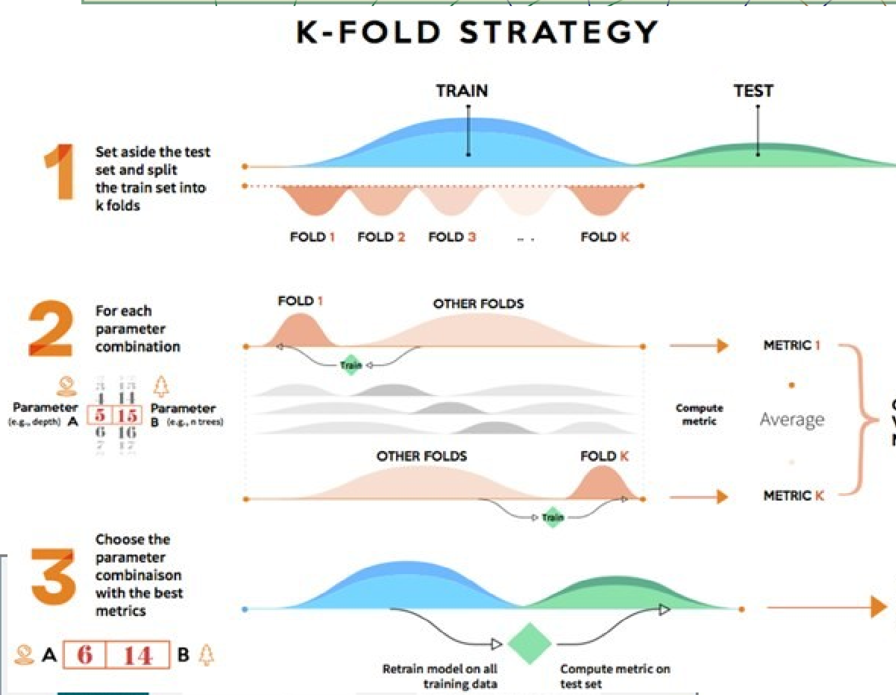

Sectores de aplicación
Financiero Seguros
Comercio y marketing
Industria y empresarial
Tecnologías información y comunicación
Sanitario y farmacéutico
Meteorología Medio Ambiente
Proceso de Minería de Datos
Data Selection and Cleaning
Data Transformation feature extraction
Data Modeling
Evaluation / Deployment

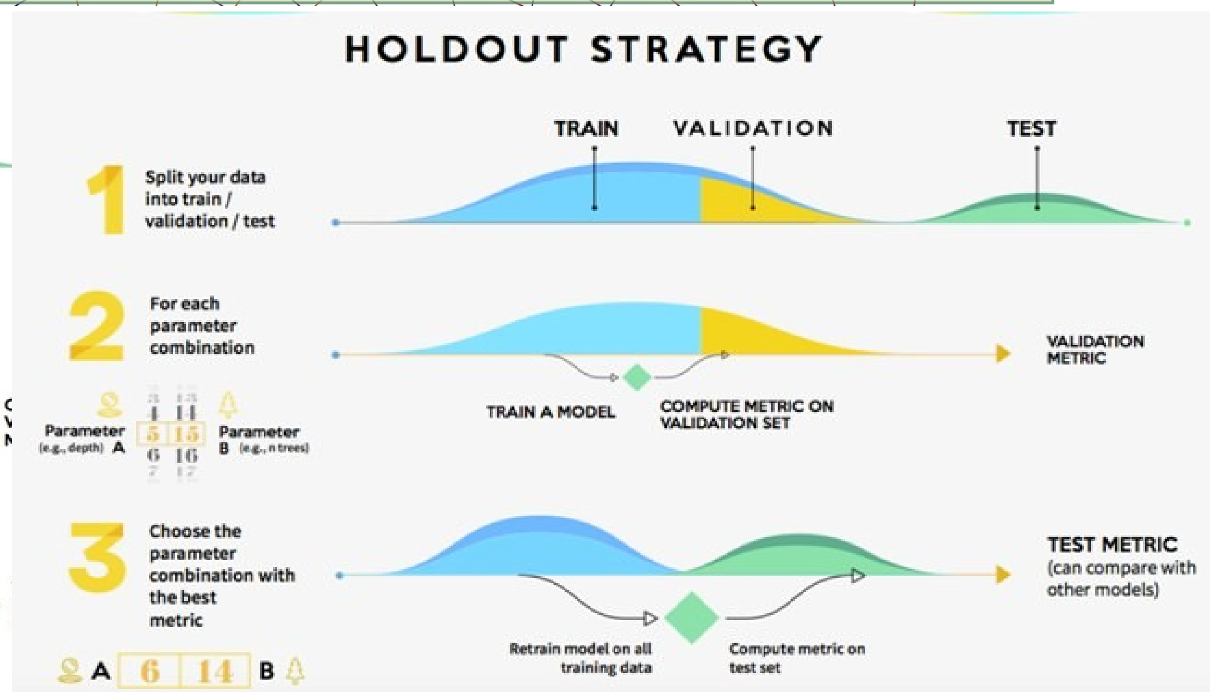

Linear Regression
NN.
Neural Network
100
80
60
40
20
50
100
150
200
250
300
350

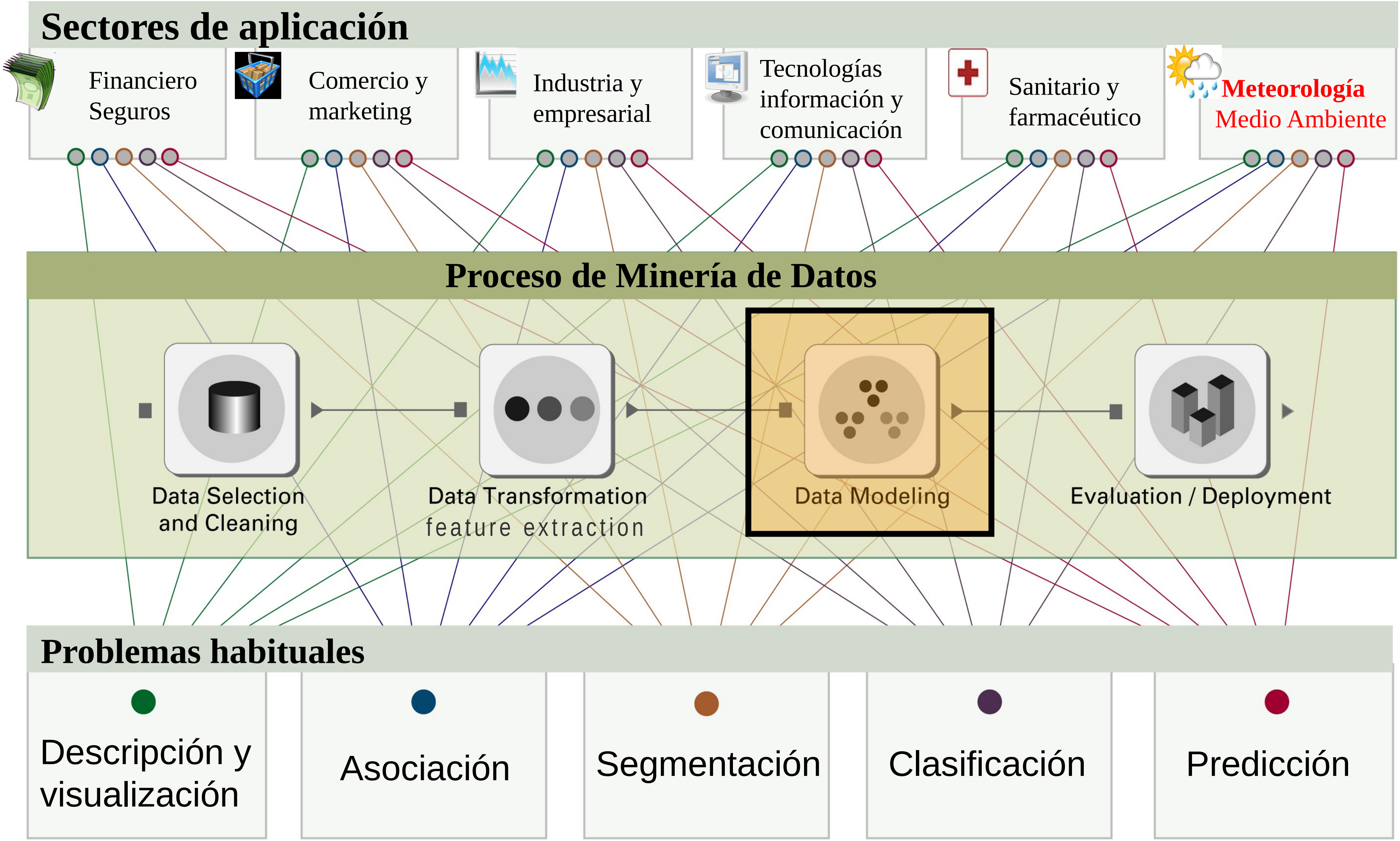
Wind Speed


(b)
(c)

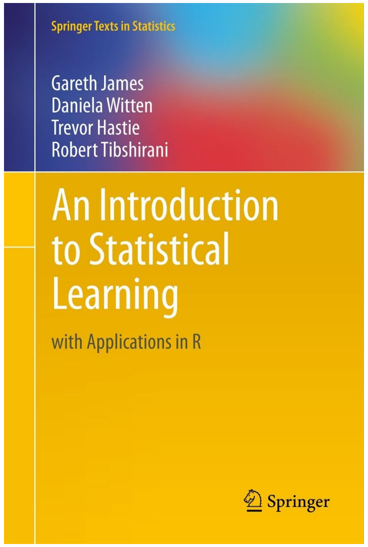

Sectores de aplicación
Financiero Seguros
Comercio y marketing
Industria y empresarial
Tecnologías información y comunicación
Sanitario y farmacéutico
Meteorología Medio Ambiente
Proceso de Minería de Datos
Data Selection and Cleaning
Data Transformation feature extraction
Data Modeling
Evaluation / Deployment


K-FOLD STRATEGY
TRAIN
TEST
1
Set aside the test set and split the train set into k folds
FOLD 1
FOLD 2
FOLD 3
FOLD K
2
For each parameter combination
FOLD 1
OTHER FOLDS
Train
METRIC 1
Parameter (e.g., depth) A
5 15
6 16
Parameter B (e.g., n trees)
Compute metric
Average
OTHER FOLDS
FOLD K
METRIC K
3
Choose the parameter combinaison with the best metrics
A 6 14 B
Retrain model on all training data
Compute metric on test set


HOLDOUT STRATEGY
TRAIN
VALIDATION
TEST
1
Split your data into train / validation / test
2
For each parameter combination
VALIDATION METRIC
TRAIN A MODEL
COMPUTE METRIC ON VALIDATION SET
Parameter (e.g., depth) A
5 15
6 16
Parameter B (e.g., n trees)
3
Choose the parameter combination with the best metric
TEST METRIC (can compare with other models)
A 6 14 B
Retrain model on all training data
Compute metric on test set



Machine learning develop methods for data modelling and prognosis.



An Introduction to Statistical Learning: With Applications in R

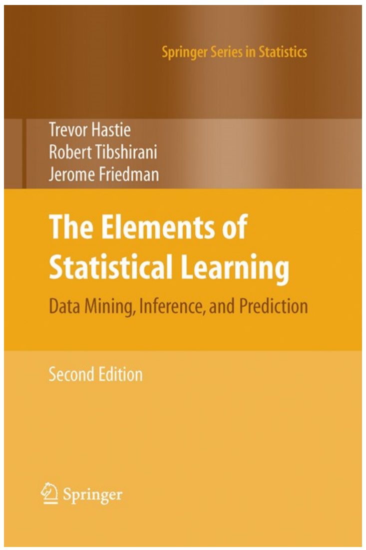
James, G., Witten, D., Hastie, T., Tibshirani, R.

Springer (2013)

```
require(ISLR)
```

http://www-bcf.usc.edu/~gareth/ISL

[PDF]



The Elements of Statistical Learning

Trevor Hastie, Robert Tibshirani, Jerome Friedman

Springer (2nd ed. 2009, Corr. 9th printing 2017)

https://web.stanford.edu/~hastie/ElemStatLearn/

[PDF]

KDnuggets

Gregory Piatetsky-Shapiro

https://www.kdnuggets.com