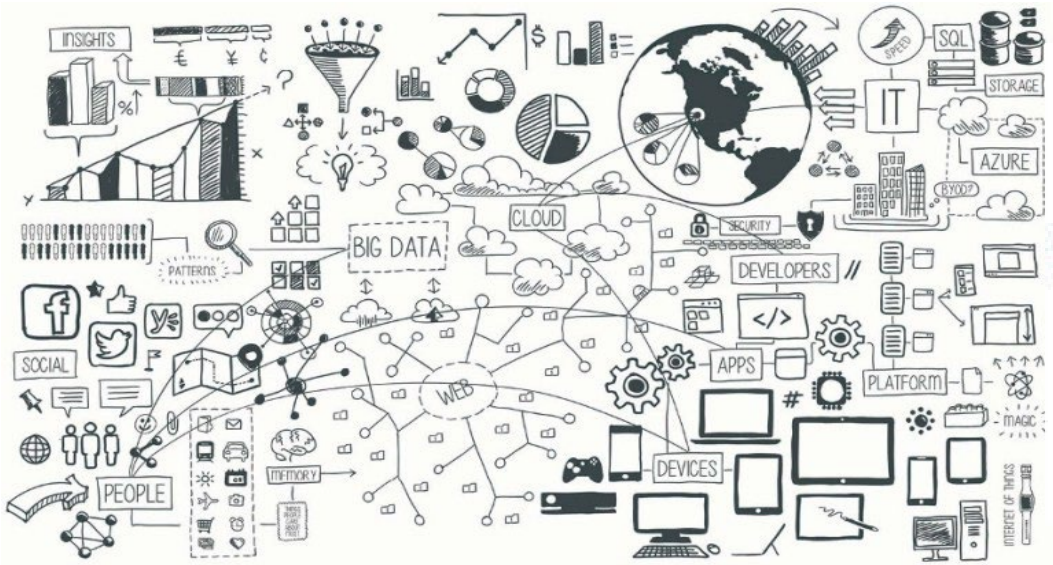
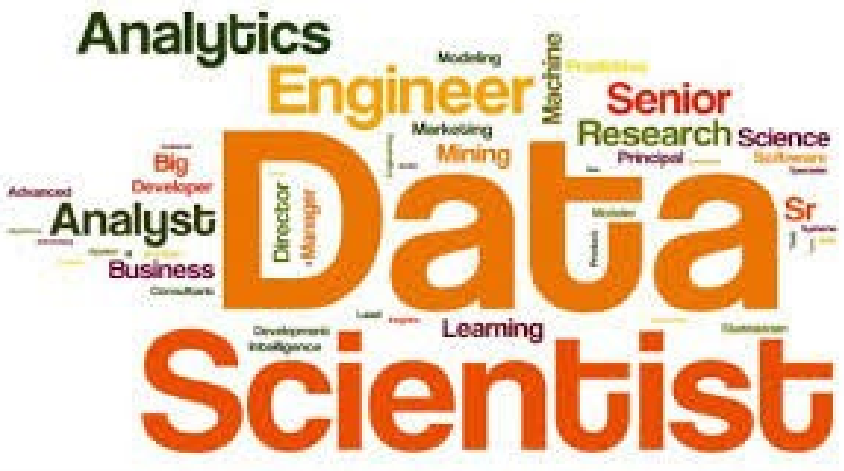
Data Mining (Minería de Datos)

INTRODUCTION AND HISTORICAL PERSPECTIVE





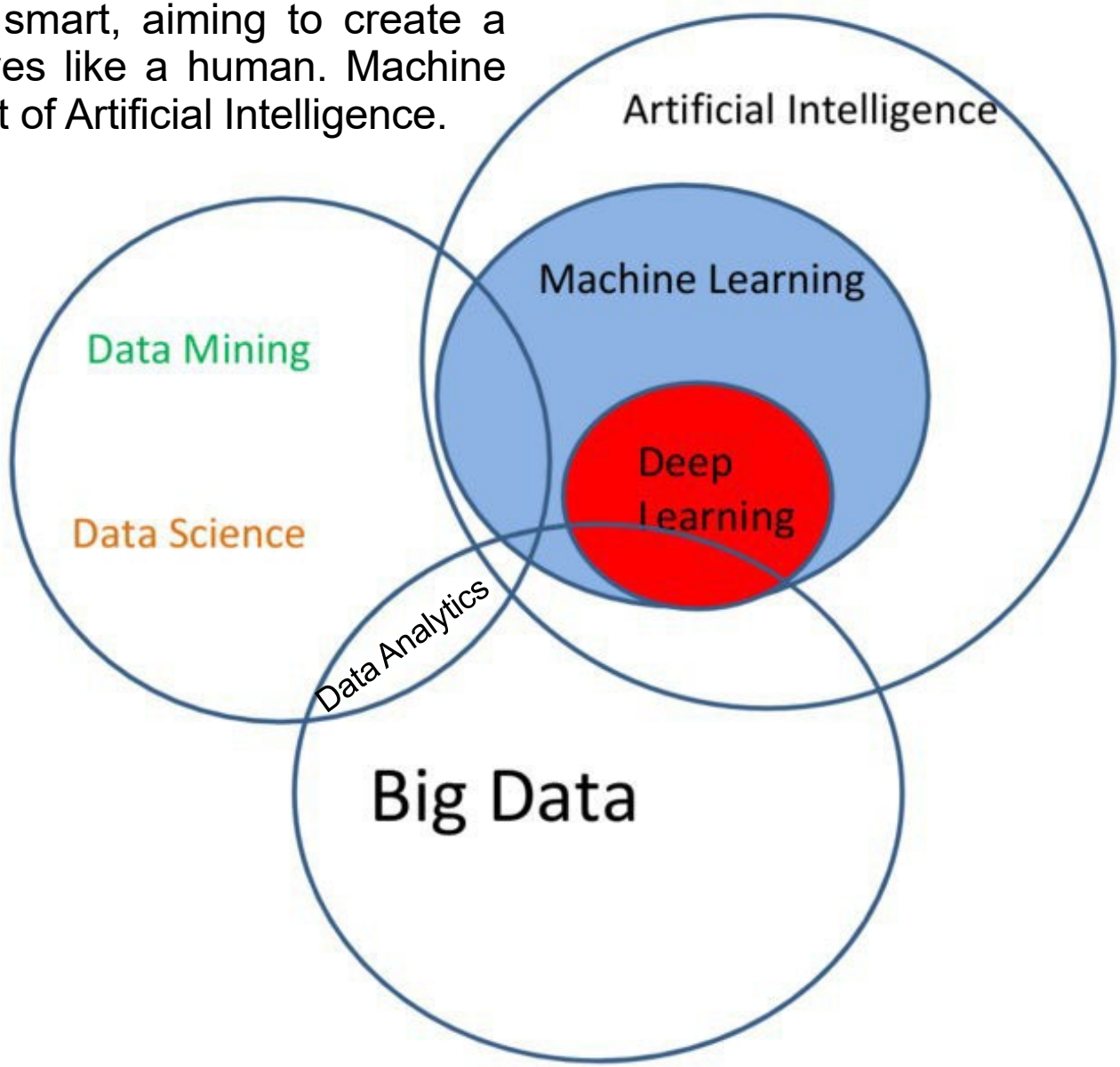
Sixto Herrera
(herreras@unican.es)

Matemática Aplicada y Ciencias de la Computación
Universidad de Cantabria

UNIVERSIDAD DE CANTABRIA

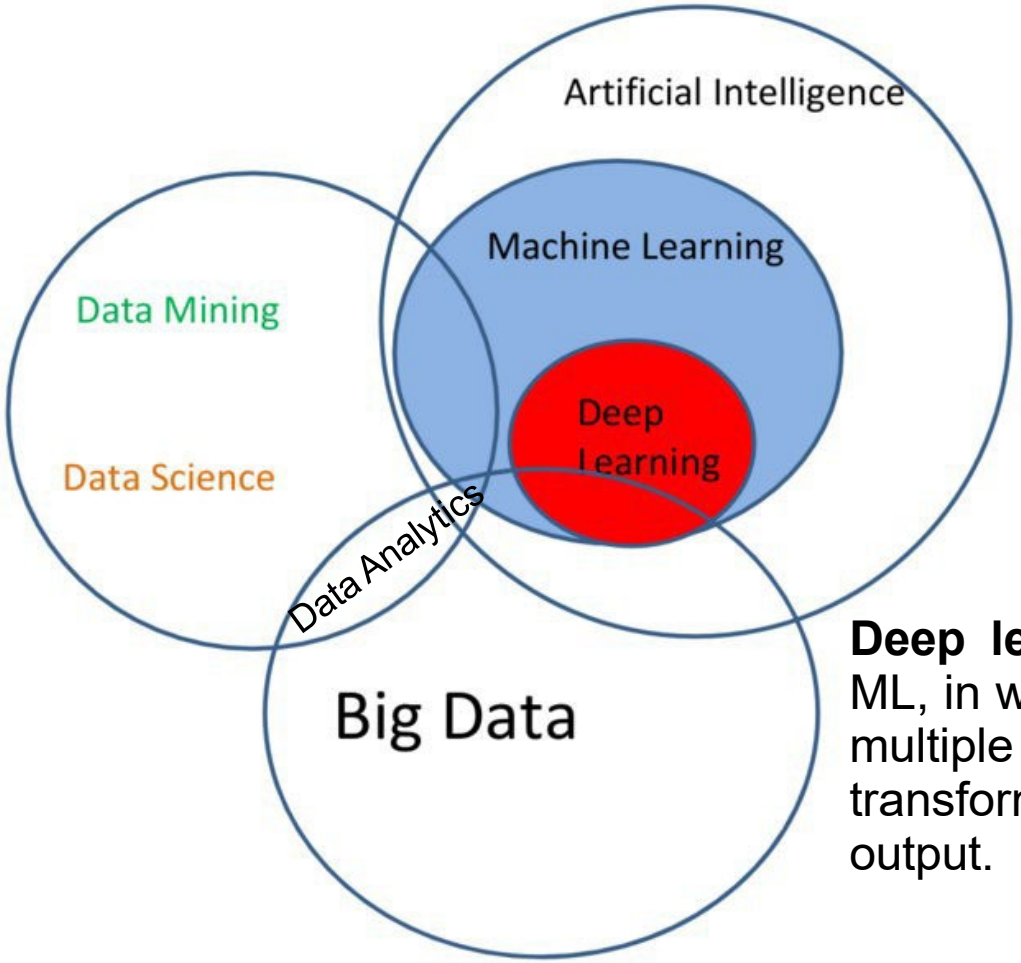
			M1966 – Data Mining
Oct	29	M	Presentación, Introducción y Perspectiva histórica (2h, T)
	30	X	Paradigmas de Aprendizaje, Problemas Canónicos y Datasets (2h, T-L)
	31	J	Reglas de Asociación (2h, T)
Nov	4	L	Reglas de Asociación (2h, L)
	6	X	Evaluación, Sobreajuste y Cross-Validation (2h, T)
	11	L	Cross-Validation (2h, L)
	13	X	Árboles de Clasficación y Decisión (2h, T)
	18	L	Árboles de Clasficación y Decisión (2h, L)
	20	X	Técnicas de Vecinos Cercanos, (k-NN) (2h, T)
	25	L	Técnicas de Vecinos Cercanos, (k-NN) (2h, L)
	27	X	Comparación de Técnicas de Clasificación (2h, L)
Dic	2	L	Árboles de Regresión (CART) (2h, T)
	4	X	Árboles de Regresión (CART) (2h, V, 17:30-19:30)
	9	L	Paquete CARET (2h, L, 17:30-19:30)
	11	X	Ensembles: Bagging and Boosting (2h, T)
	13	V	Random Forests (2h, L)
	16	L	Gradient Boosting (2h, T-L)
	18	X	Paquete CARET (2h, L, 17:30-19:30)
Ene	8	X	Reducción de la Dimensión (No lineal) (2h, T-L)
	13	L	Reducción de la Dimensión (No lineal) (2h, T-L)
	15	X	Técnicas de Agrupamiento (2h, T)
	20	L	Técnicas de Agrupamiento (2h, L)
	22	X	Predicción Condicionada (2h, L)
	24	V	Sesión de Repaso (2h, T-L)
	29	X	Examen//Cuestionario (2h, T-L)

Artificial intelligence (AI) is concerned with making machines smart, aiming to create a system that behaves like a human. Machine learning is a subset of Artificial Intelligence.



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).



Deep learning is a subset of ML, in which data is passed via multiple numbers of non-linear transformations to calculate an output.

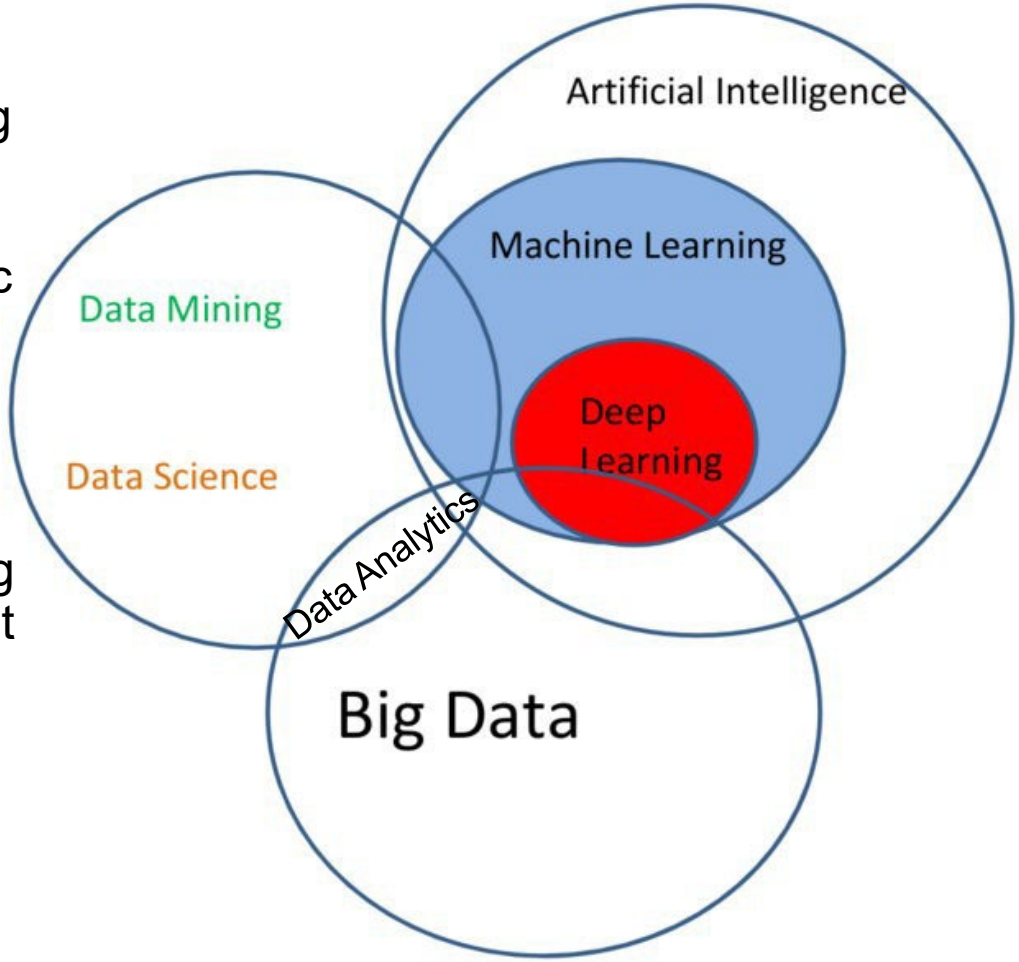
https://www.kdnuggets.com/2020/11/data-science-history-overview.html

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

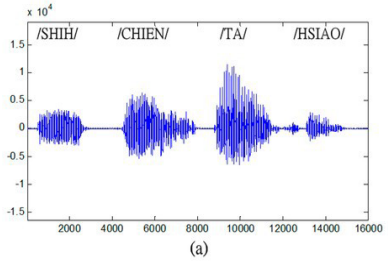
During this process machine learning algorithms are used (A. Flag).
ML Techniques → Generic
Data Mining → Understand some specific domain.

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

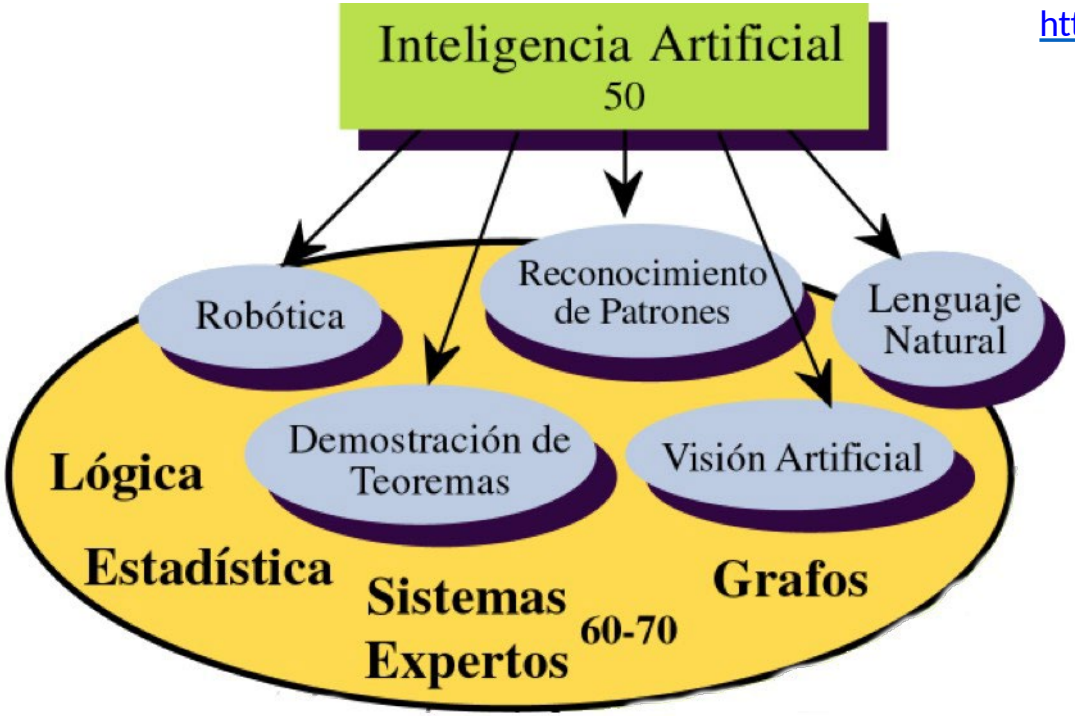
Machine Learning (ML) relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed (definition of A.Samuel).



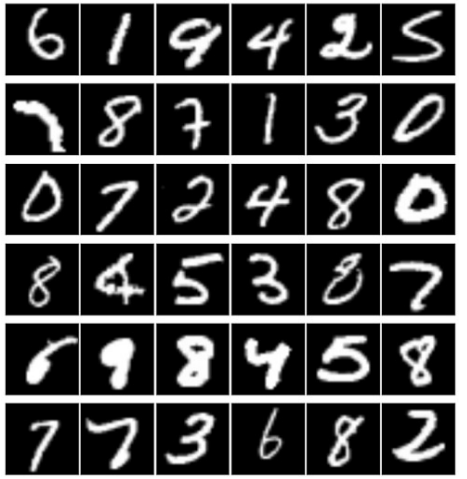
https://www.kdnuggets.com/2020/11/data-science-history-overview.html



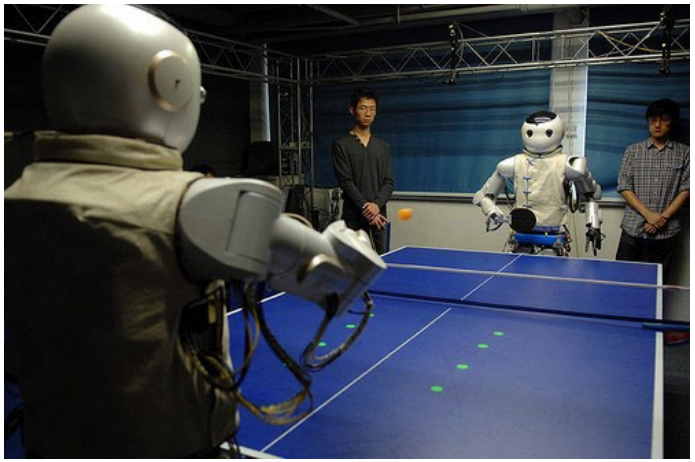
Overview of Natural Language Processing(NLP) with R and OpenNLP

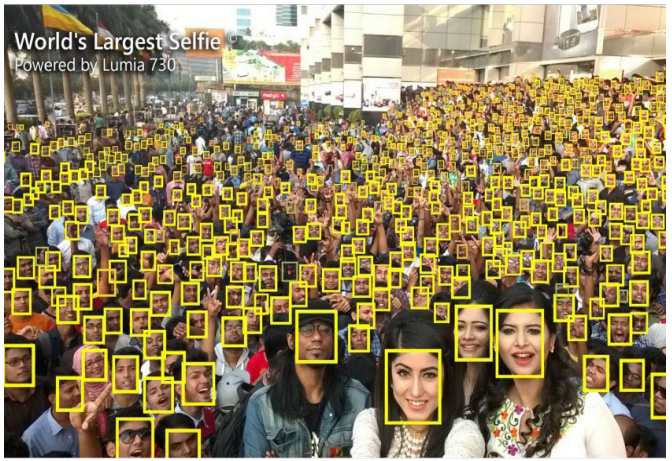


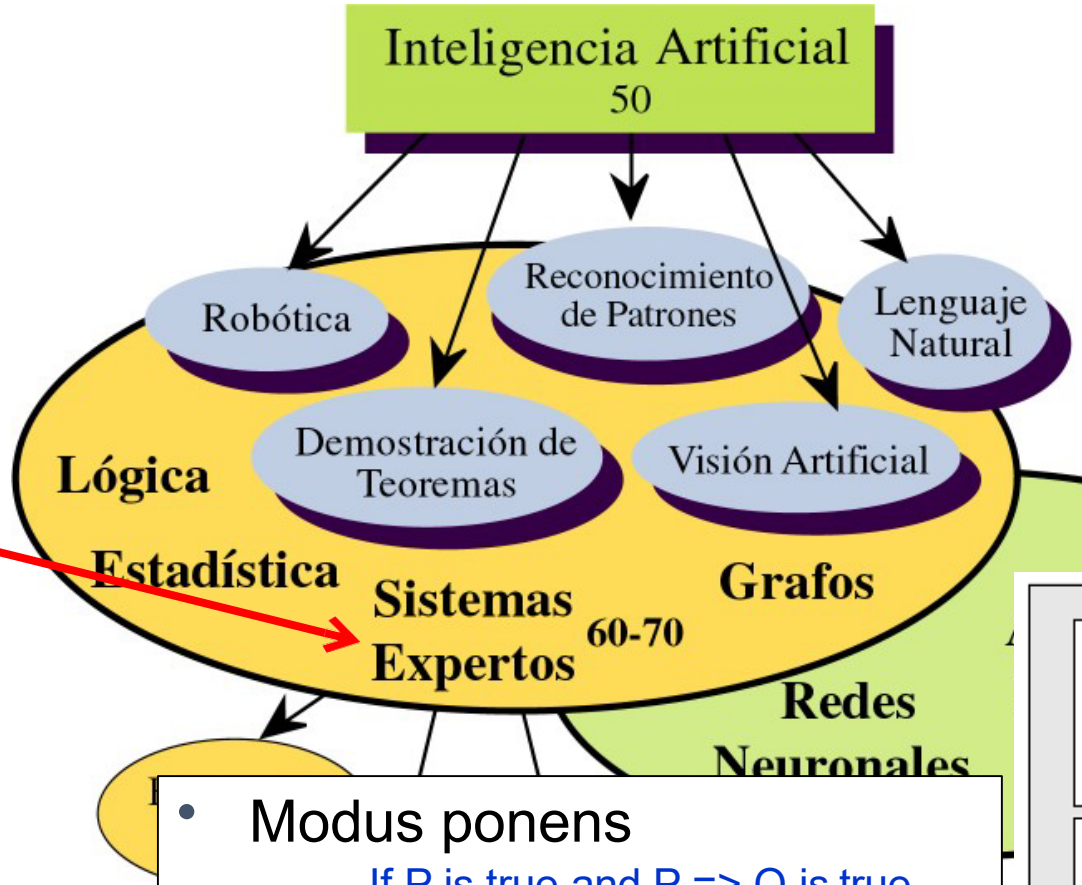
http://yann.lecun.com/exdb/mnist/
60000+10000 images 28x28
Labeled as {0,...,9}



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







DATA MINING
1990

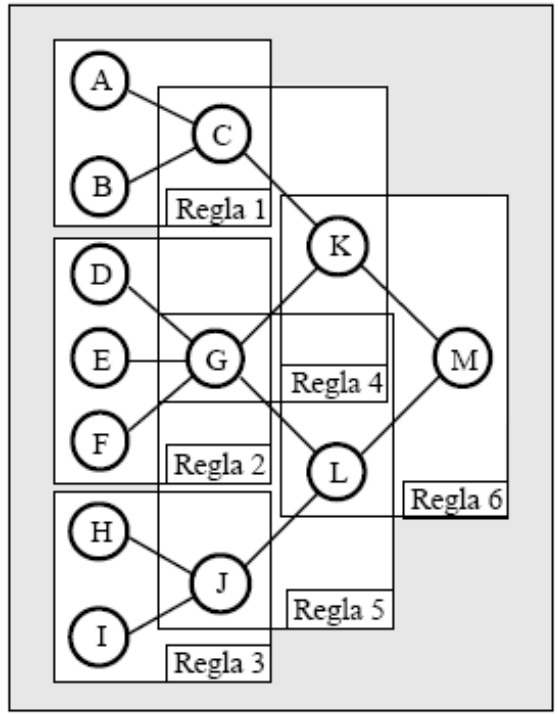
H, I cierto
L falso

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

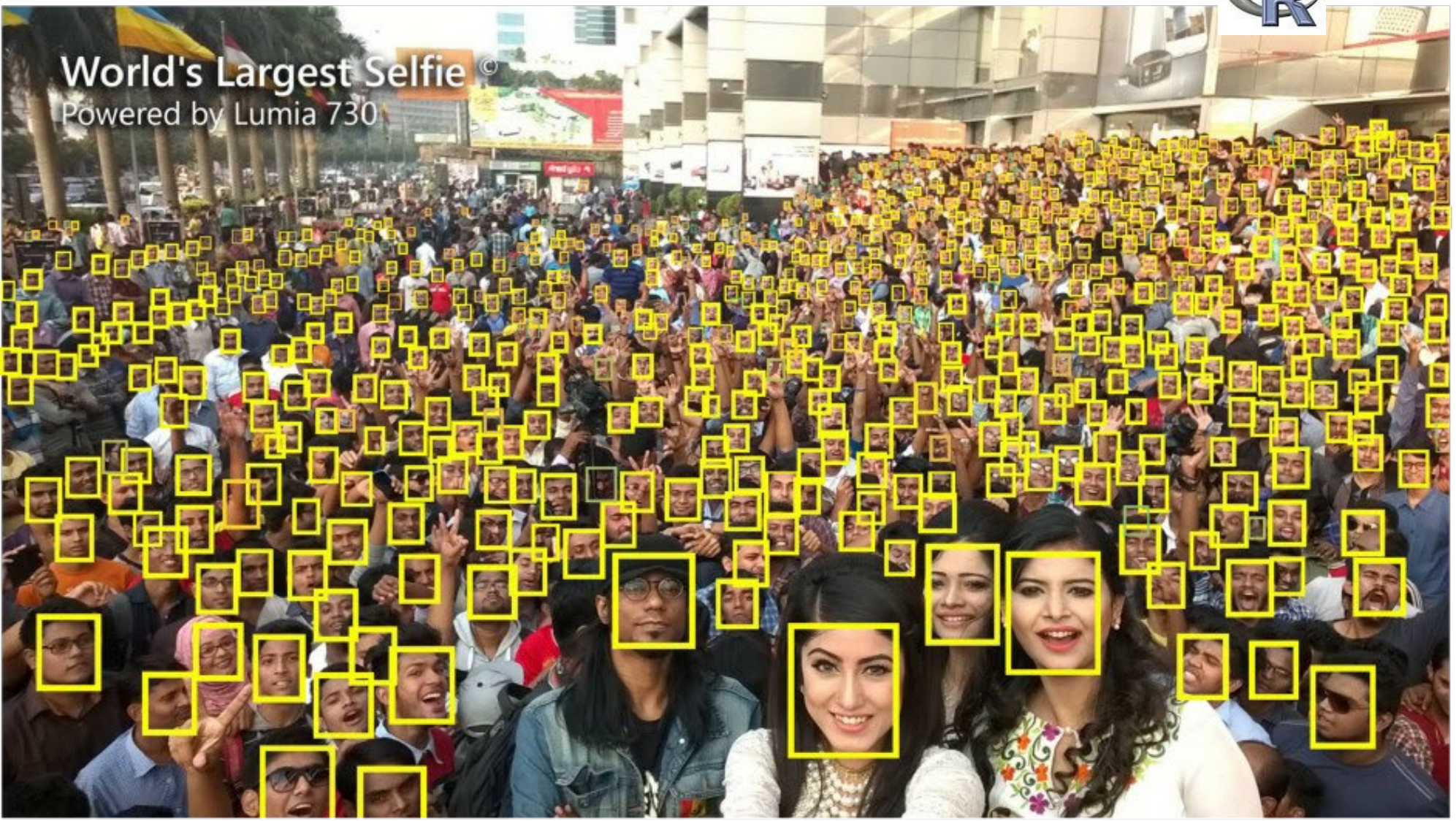


Human-like reasoning

Logical inference, look for relations on graphs, etc.

Serial processing
Limited and costly computing power

https://github.com/peiyunh/tiny

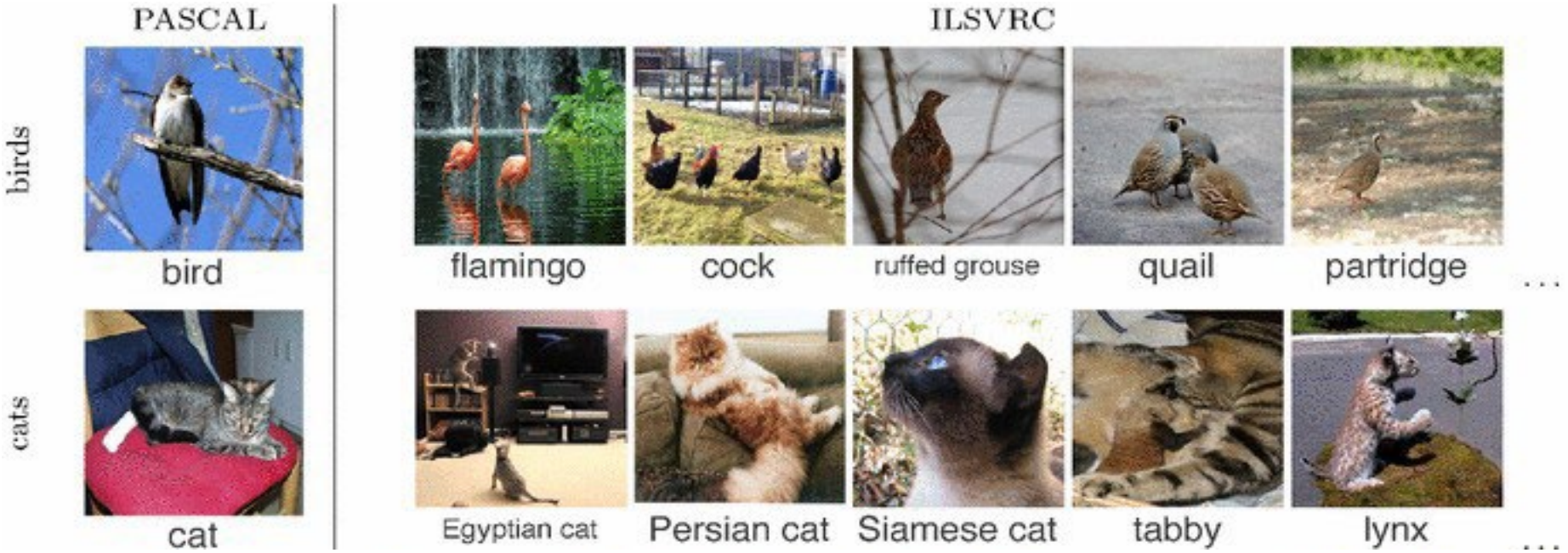


Develop a face detector (Tiny Face Detector) that can find ~800 faces out of ~1000 reportedly prese taking use of novel characterization of scale, resolution, and context to find small objects.

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

167,62 GB [kaggle]



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

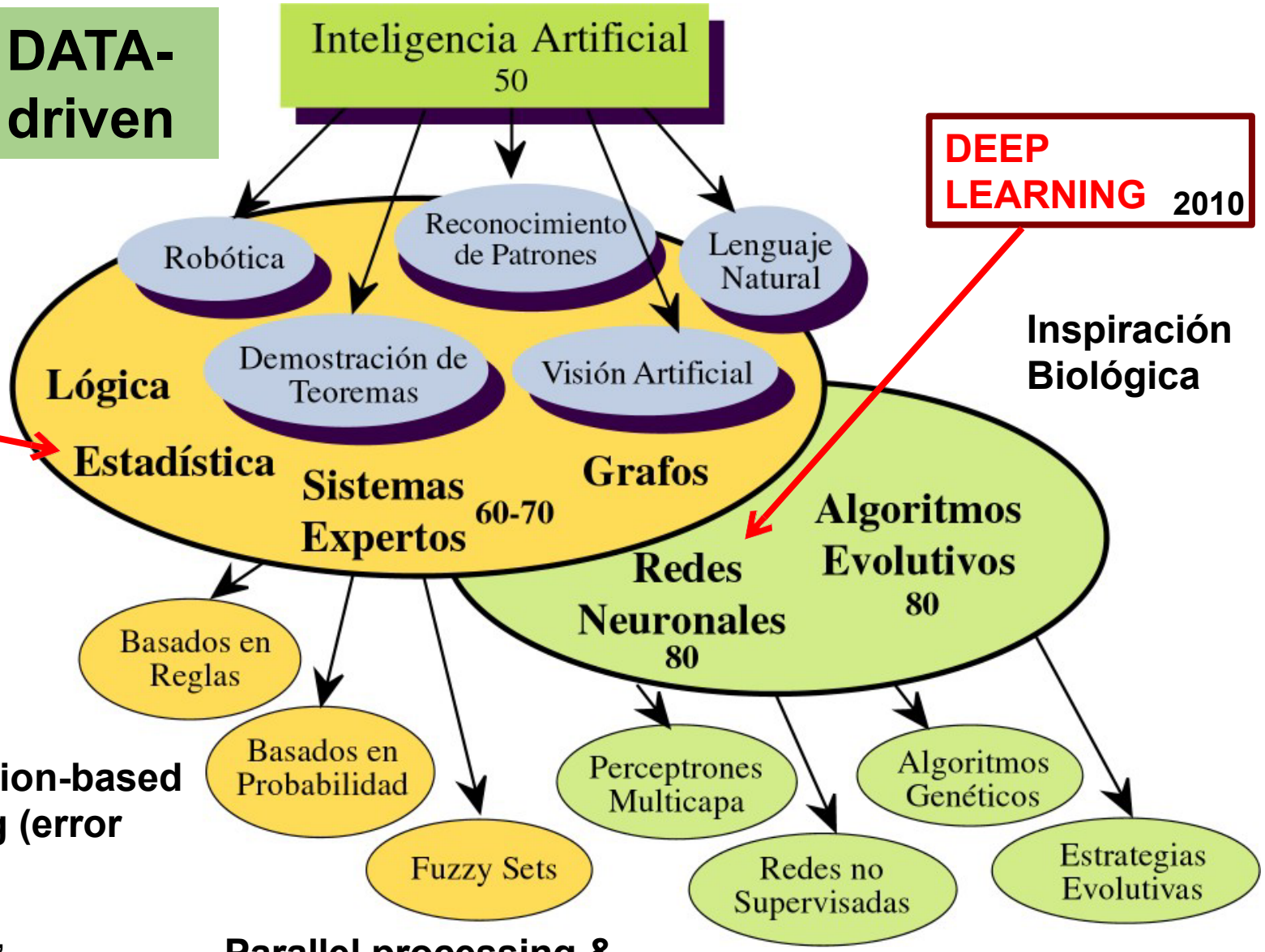
Nuevos Paradigmas DATA-driven

Statistical Inspiration

STATISTICAL LEARNING 2000

DEEP LEARNING 2010

Inspiración Biológica



Data driven using abstract representations

Kernels,
neural network, etc.

Optimization-based reasoning (error function).

Empirical risk,
gradient descend,
Backpropagation, etc.

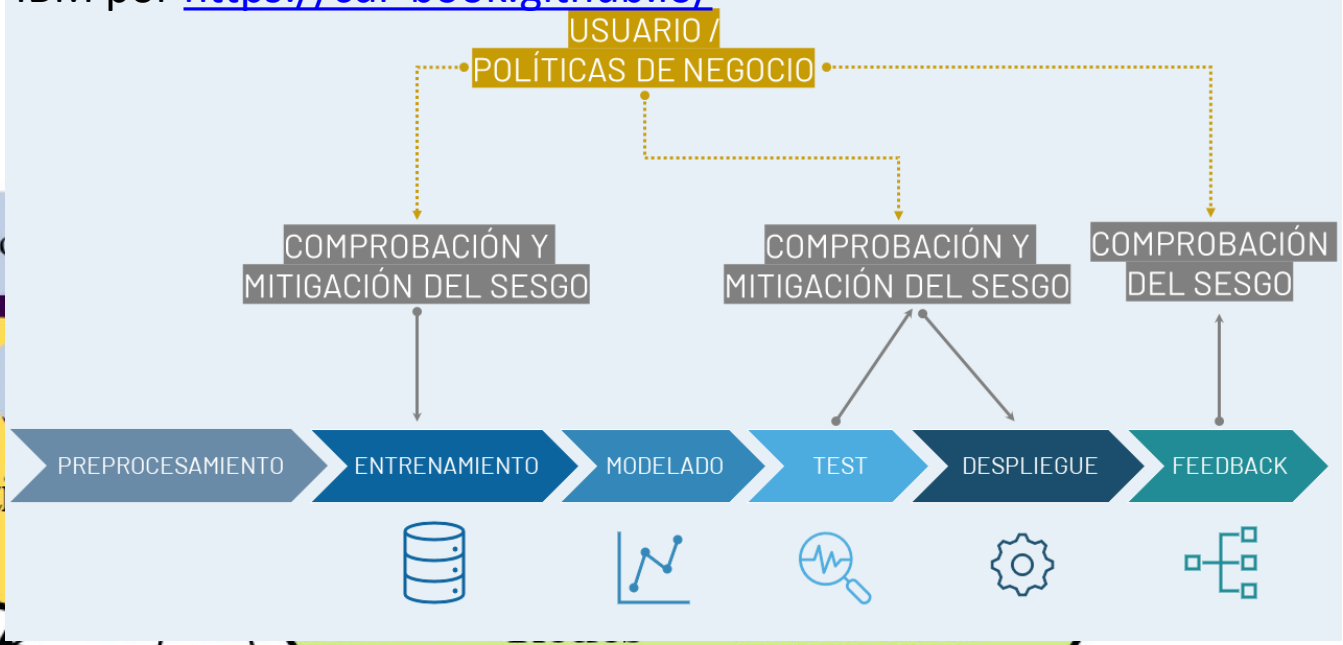
Parallel processing & distributed computing

HPC, GPUs, cloud, etc.

(a) how to parallelize computation?
(b) how to distribute data?
(c) how to handle failures?

Nuevos Problemas

Sesgos en el proceso de machine learning. Fuente: Adaptada de IBM por https://cdr-book.github.io/



Statistical Inspiration

STATISTICAL LEARNING 2000

Data driven using abstract representations

Kernels, neural network, etc.

KDnuggets™

How to Create Unbiased Machine Learning Models

https://www.kdnuggets.com/2021/07/create-unbiased-machine-learning-models.html

Código de Ética y Conducta Profesional de ACM:
https://www.acm.org/code-of-ethics/the-code-in-spanish
https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Nuevos Problemas

Alarmist messages:

Should be taken with care because:

The Mystery Behind Human Thought

Current Deep Learning Networks Are Not Robust

Machine Learning Systems Remain Inefficient

Deep Learning Networks are Hard to Improve

Short-Term Impact of AI Systems on Human Society

Ill-Adapted Infrastructure

Short-Term Future of Human Work

No Conceptual Breakthroughs

From ETL to ELT and (EL)T

https://www.kdnuggets.com/2020/12/future-etl-is-elt.html

https://www.kdnuggets.com/2018/02/current-hype-cycle-artificial-intelligence.html

KDnuggets™

How to Create Unbiased Machine Learning Models

https://www.kdnuggets.com/2021/07/create-unbiased-machine-learning-models.html

Necesidad de enormes cantidades de datos:

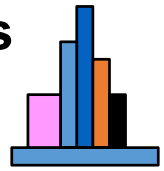
https://www.eldiario.es/tecnologia/curva-aprendizaje-inteligencia-artificial-riesgo-ansia-datos-infinita-no-obras-humanas_1_11271307.amp.html

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

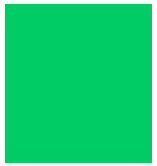
Knowledge Patterns



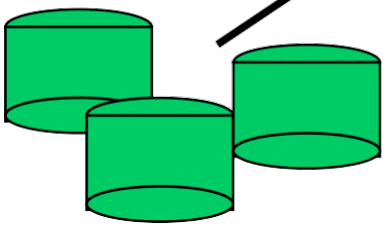
DATA MINING

Task-relevant Data

Data Warehouse



Data Cleaning



Databases

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

DATA ANALYTICS

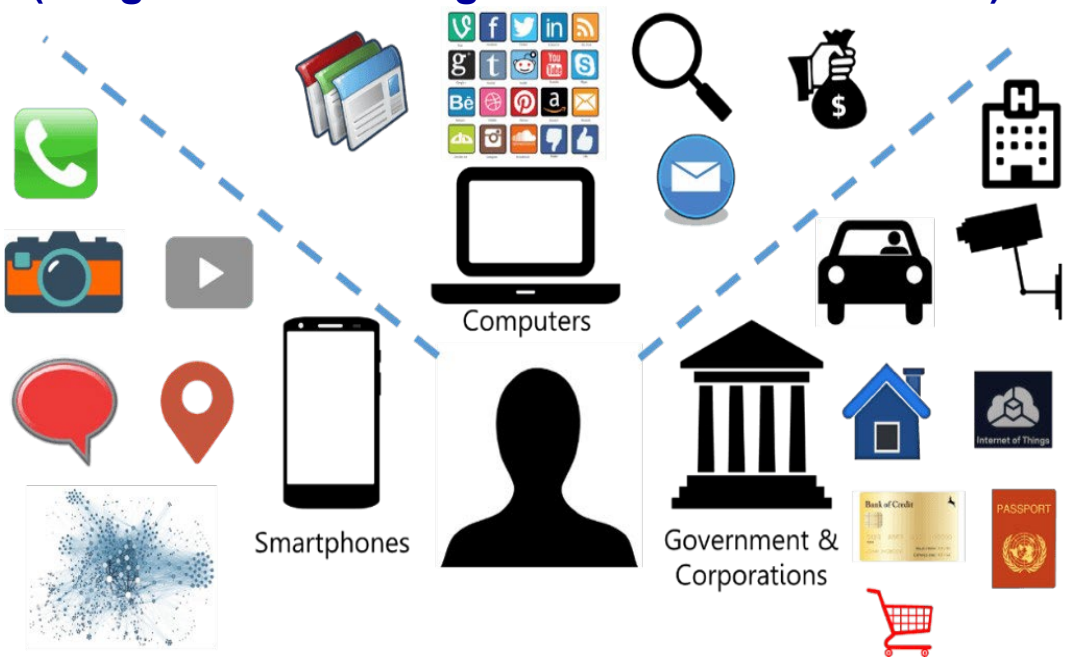
Data Discovery, Cleaning and Reduction

Modeling and Prediction

AI & Machine learning

Big data (large volume, variety and velocity at which data is being generated)

(integration of heterogeneous real-time sources)



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

2011

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

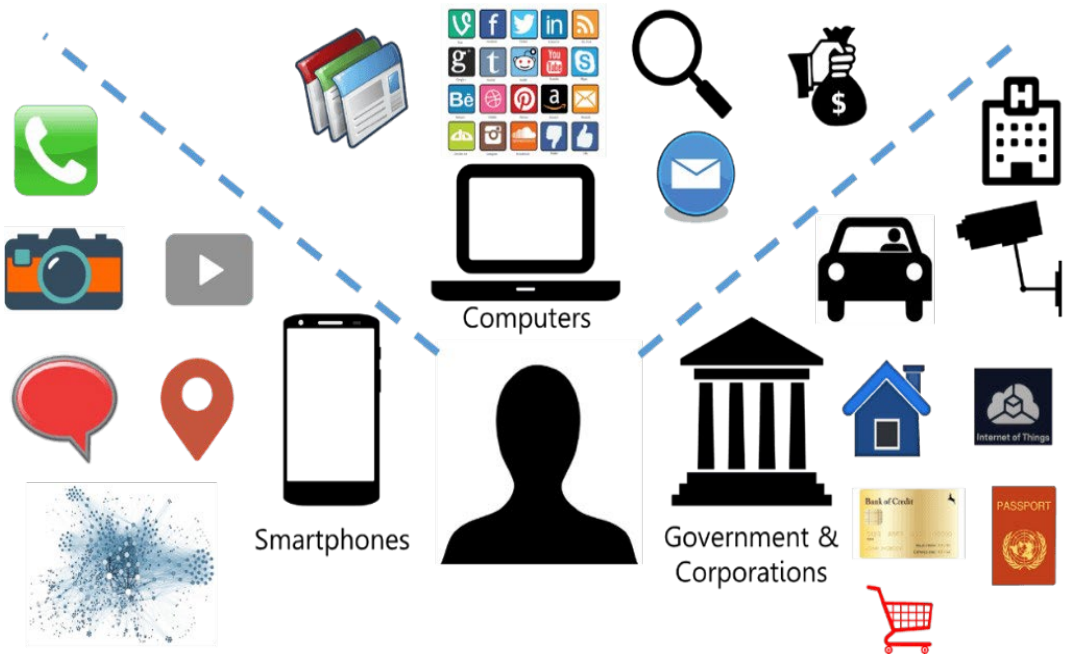
AI & Machine learning

DATA ANALYTICS

Data Discovery, Cleaning and Reduction

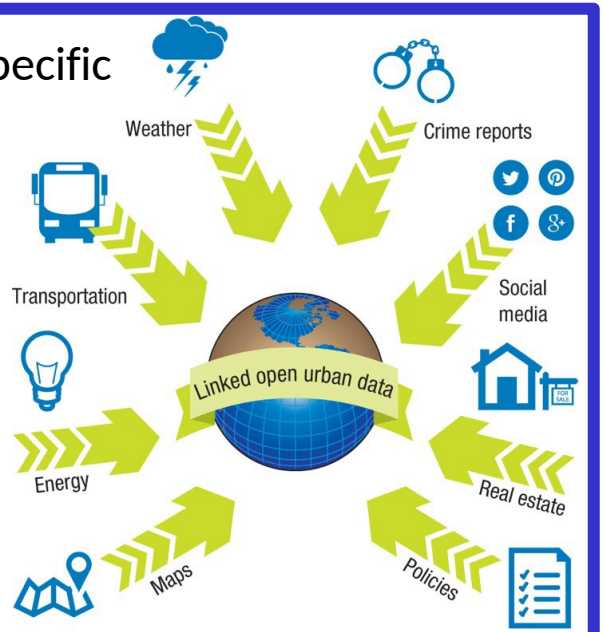
Modeling and Prediction

Big data
(integration of heterogeneous real-time sources)

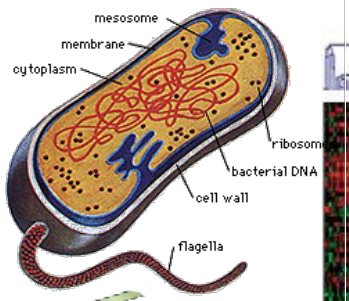


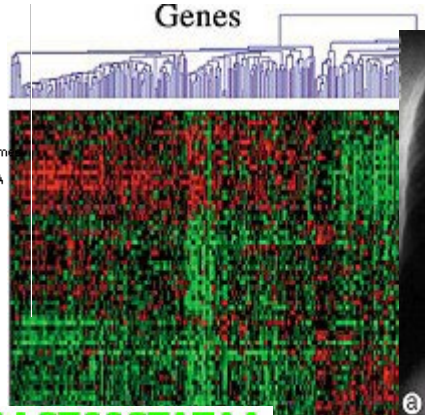
Domain-specific Problems

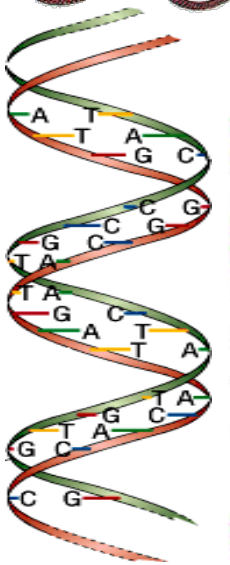
DataLabs



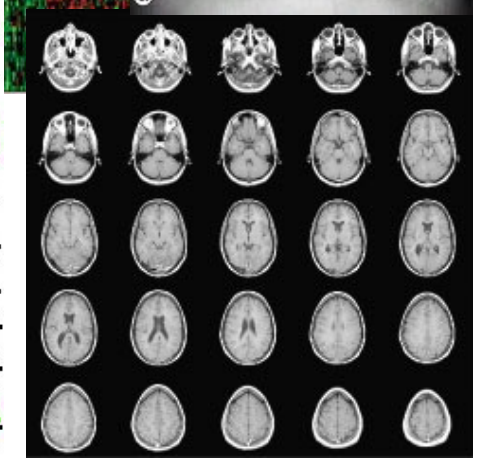
Financiero Seguros | Comercio y marketing | Industria y gestión empresarial | Tecnologías información y comunicaciones | **Sanitario y farmacéutico** | Meteorología, clima y medio ambiente

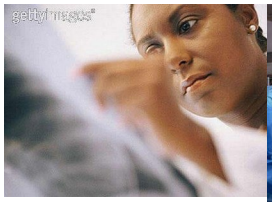


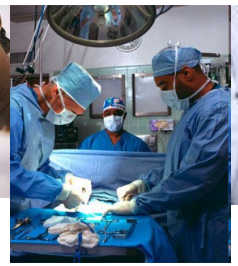




TCTAAGTCCGTATAA
AGATTCAGGCATATT
AGATTCAGGCATATT
TCTAAGTCCGTATAA
TCTAAGTCCGTATAA
AGATTCAGGCATATT
AGATTCAAGCATATT
AGATTCAAGCATATT
AGATTCAAGCATATT
AGATTCAGGCATATT
AGATTCAGGCATATT
TCTAAGTCCGTATAA
AGATTCAGGCATATT







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

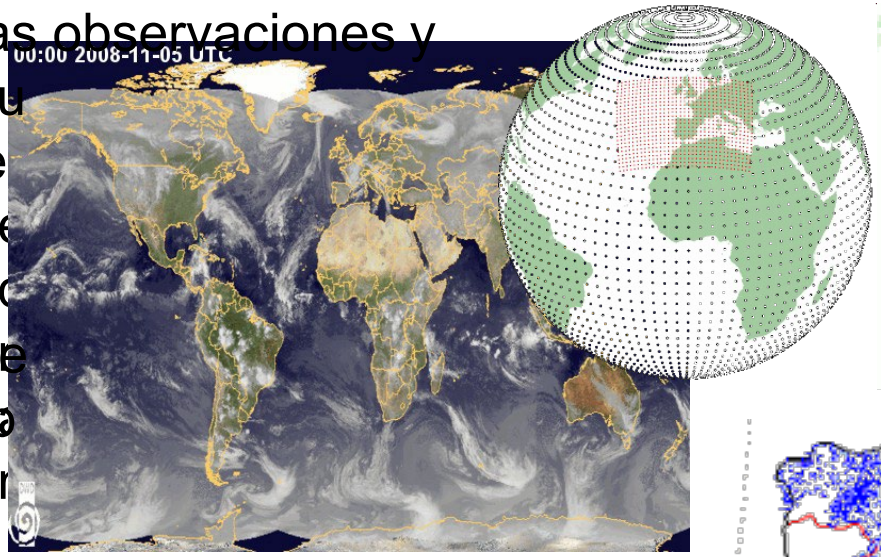
Industria y gestión empresarial

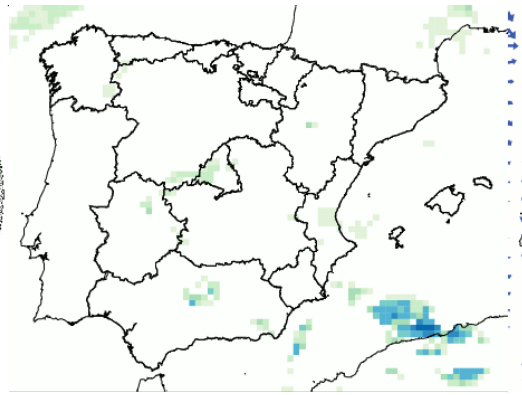
Tecnologías información y comunicaciones

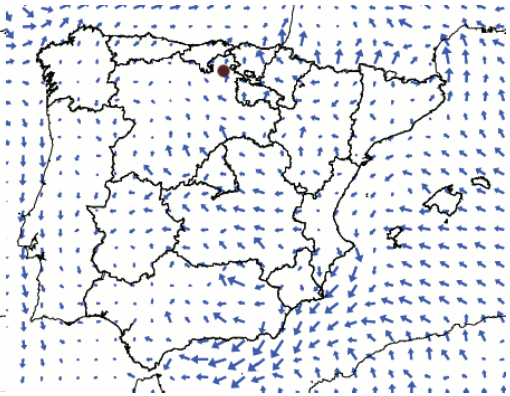
Sanitario y farmacéutico

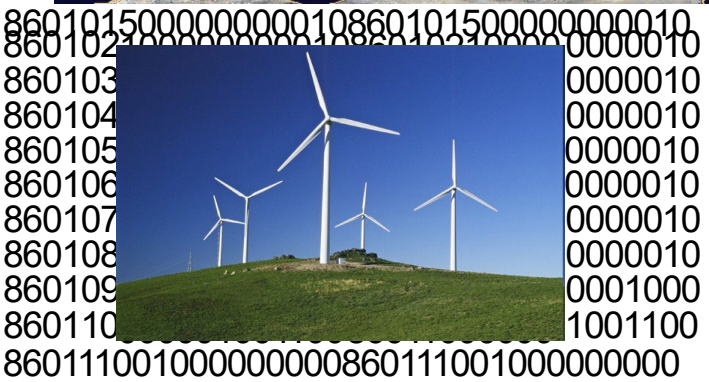
Meteorología, clima y medio ambiente

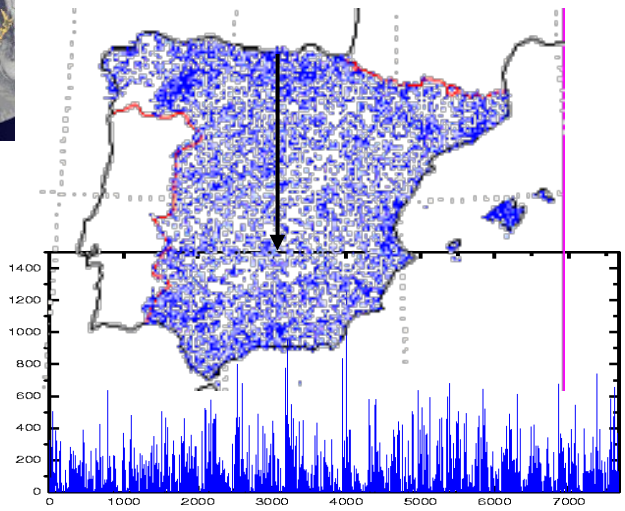
Las observaciones y simu
re
gene
c
hete
p
r

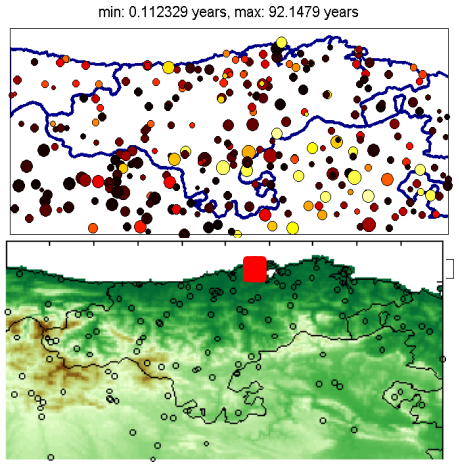




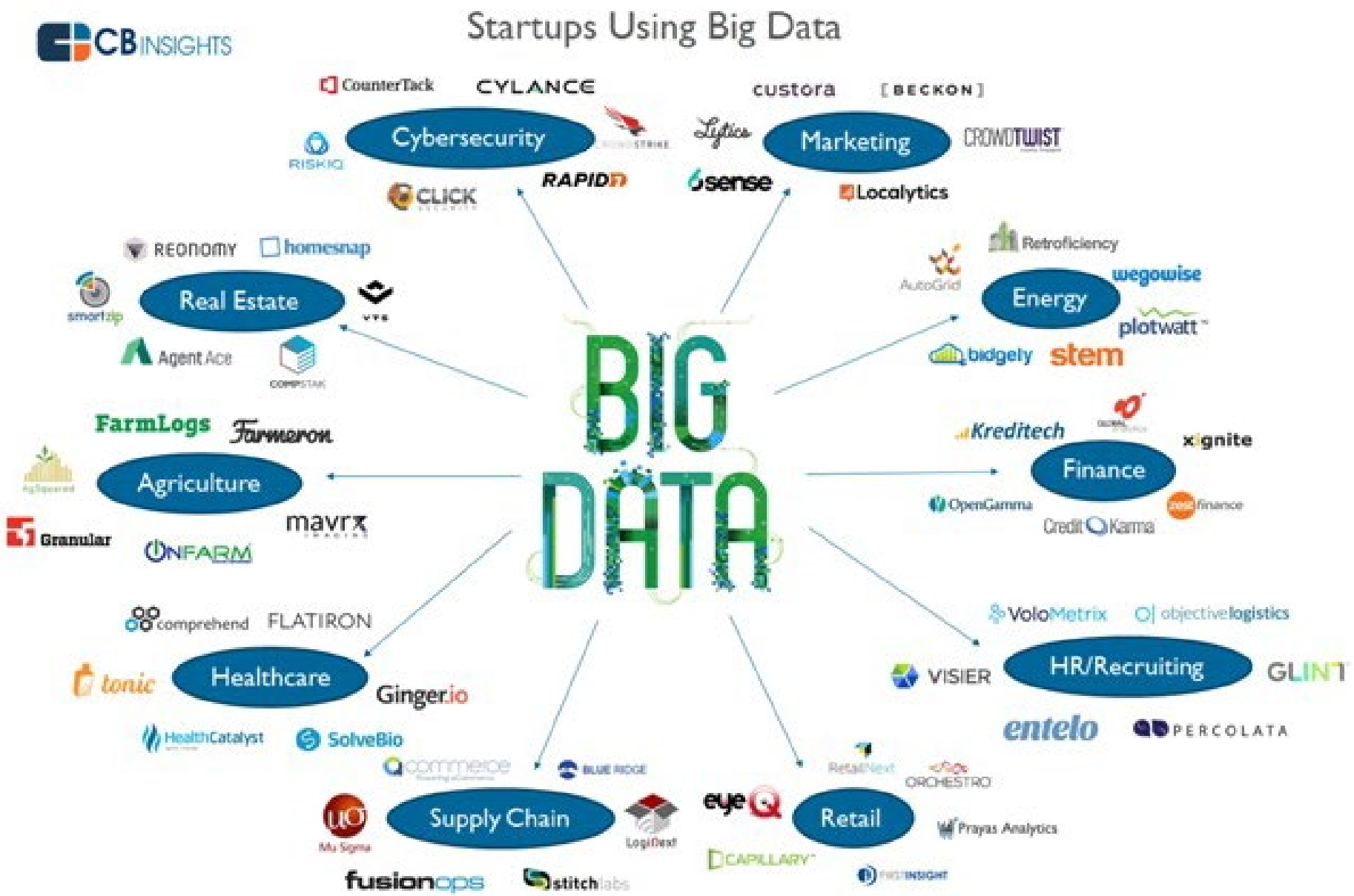








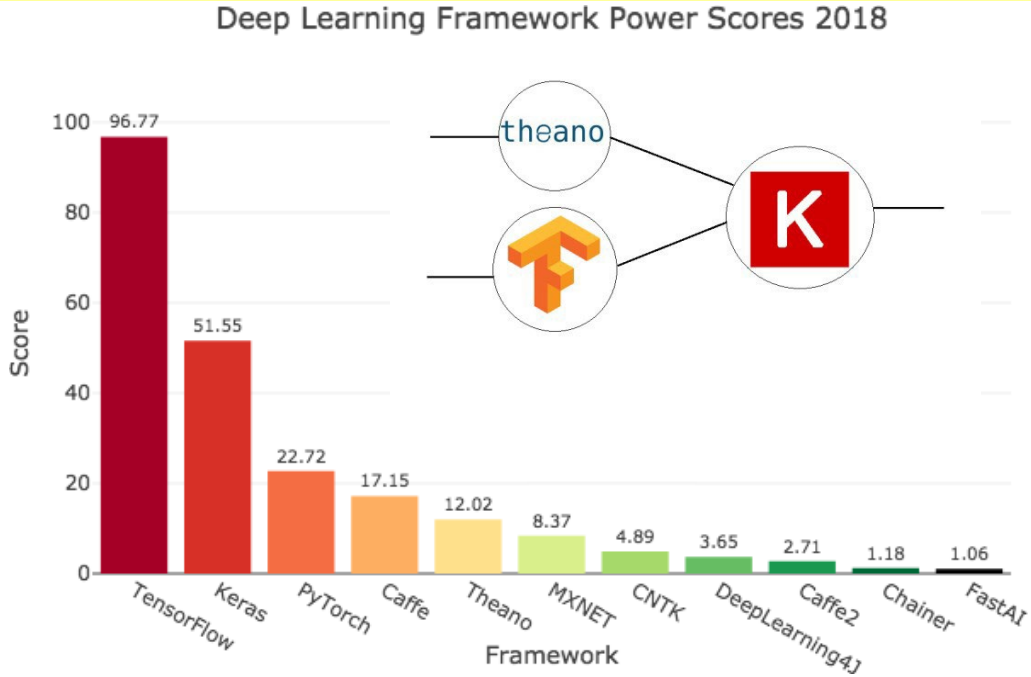
http://www.meteo.unican.es/downscaling

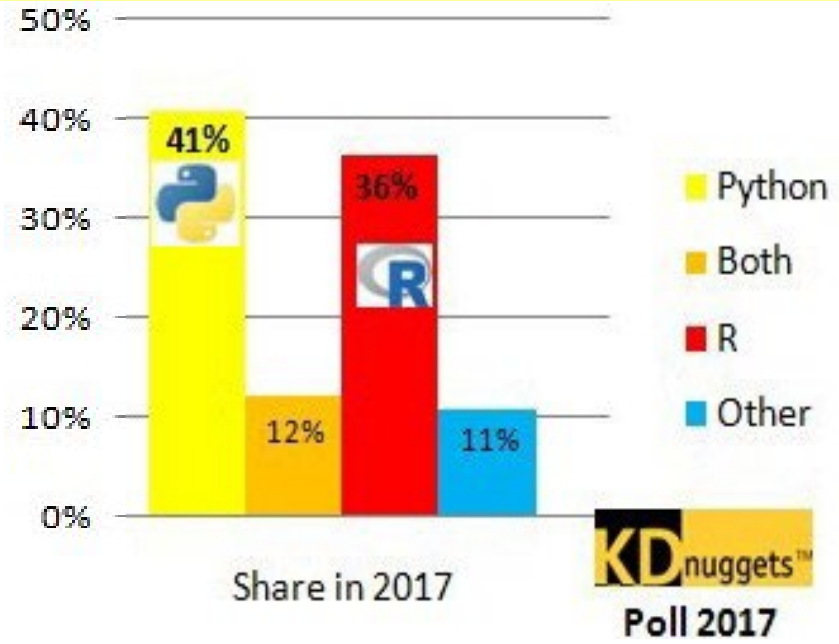

CB INSIGHTS
Startups Using Big Data
BIG DATA
Cybersecurity
CounterTack
CYLANCE
RISKIQ
CROWDSTRIKE
RAPID7
CLICK SECURITY
Marketing
custora
BECKON
Lytics
CROWDTWIST
6sense
Localytics
Real Estate
REONOMY
homesnap
smartzip
VTS
Agent Ace
COMPSTAK
Energy
Retroficiency
AutoGrid
wegowise
plotwatt
bidgely
stem
Agriculture
FarmLogs
Farmeron
AgSquared
Granular
mavrx
ONFARM
Finance
Kreditech
xignite
OpenGamma
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
INSIGHT

Key reasons for this hypergrowth including the effects of Moore's law, parallel and distributed computing, availability of Big Data (produce more than 8 quadrillion Gigabytes (i.e., 8 zetta bytes) of data by 2017), growing collaboration between academia and industry, and the amount of research that is being done in AI and its subfields.

https://www.kdnuggets.com/2018/02/domains-ai-rivaling-humans.html

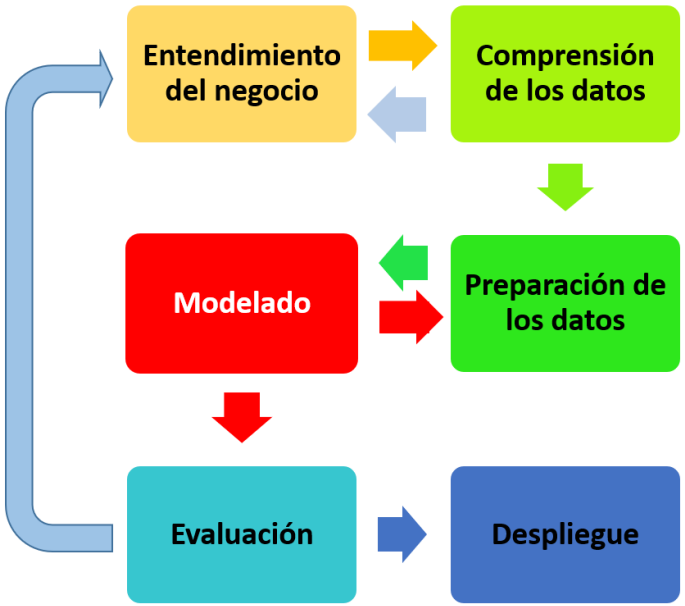
Key factors for the quick growth of data science are the open source software (In 2002, Torch was the first such ML software but others (e.g., Caffe, Theano, Keras, MXNet, DeepLearning4J, Tensorflow) have been introduced), and the efficient frameworks (and infrastructures) available:





https://www.analyticsinsight.net/how-python-and-r-dominate-the-data-science-landscape/
https://www.tiobe.com/tiobe-index/
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

Etapas de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining).
Fuente: https://cdr-book.github.io/metodologia.html?q=crisp#met-crisp-dm



a) Entendimiento del negocio:
- a.1) Determinación de los objetivos de negocio
- a.2) Evaluación de la situación actual
- a.3) Determinación de los objetivos del proyecto
- a.4) Plan del proyecto

b) Comprensión de los datos:
- b.1) Recopilación
- b.2) Descripción
- b.3) Exploración (Análisis Exploratorio de Datos, AED)
- b.4) Verificación de la calidad

c) Preparación de los datos:
- c.1) Selección
- c.2) Limpieza
- c.3) Construcción
- c.4) Integración
- c.5) Formateo

d) Modelado:
- d.1) Selección de técnicas de modelado
- d.2) Generación de un diseño de evaluación
- d.3) Generación de modelos
- d.4) Validación del modelo

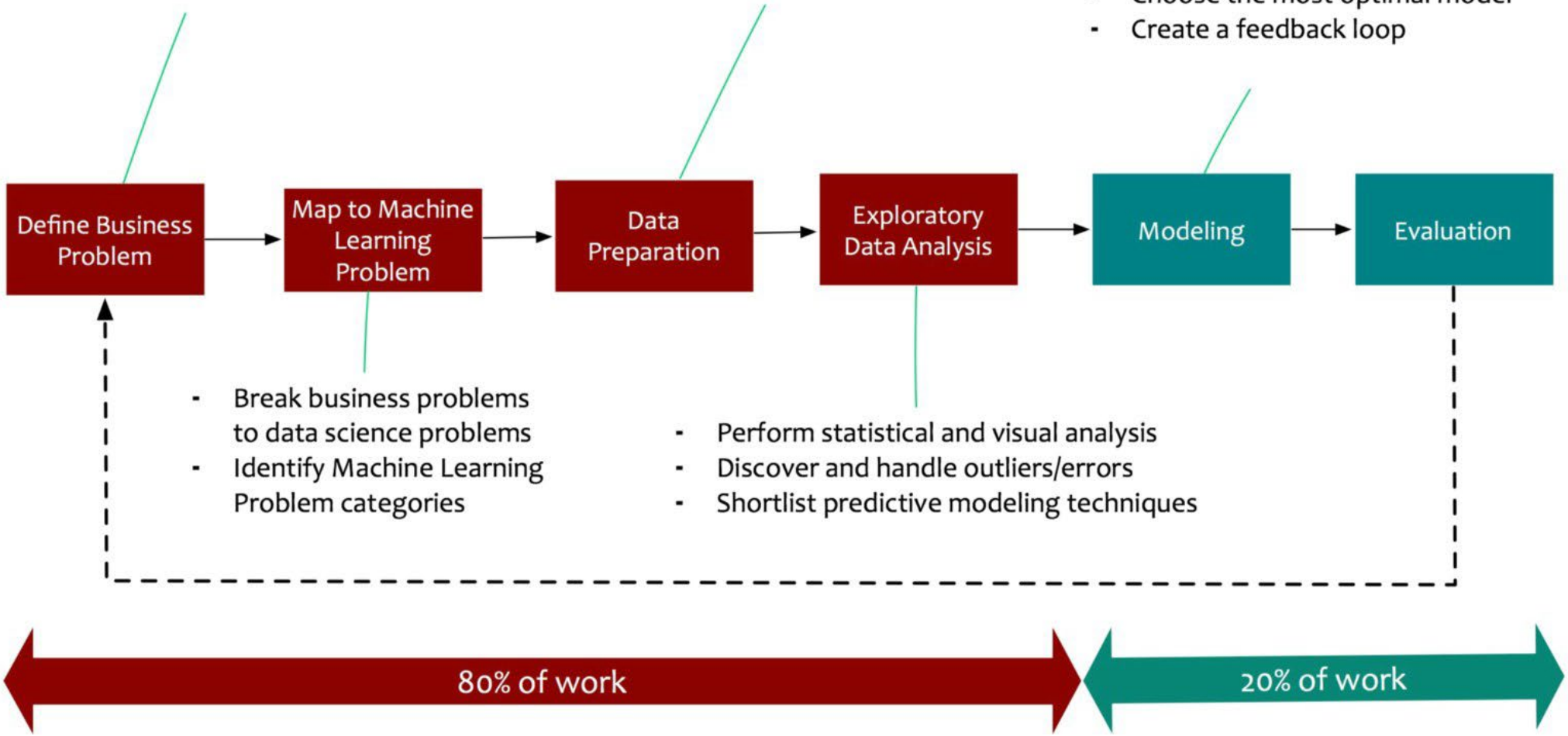
e) Evaluación. Se debe comprobar que el modelo final generado cumple las expectativas de negocio especificadas en la primera fase.

f) Despliegue:
- f.1) Planificación del despliegue
- f.2) Planificación del control y del mantenimiento.
- f.3) Informe final y Revisión final del proyecto

- 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

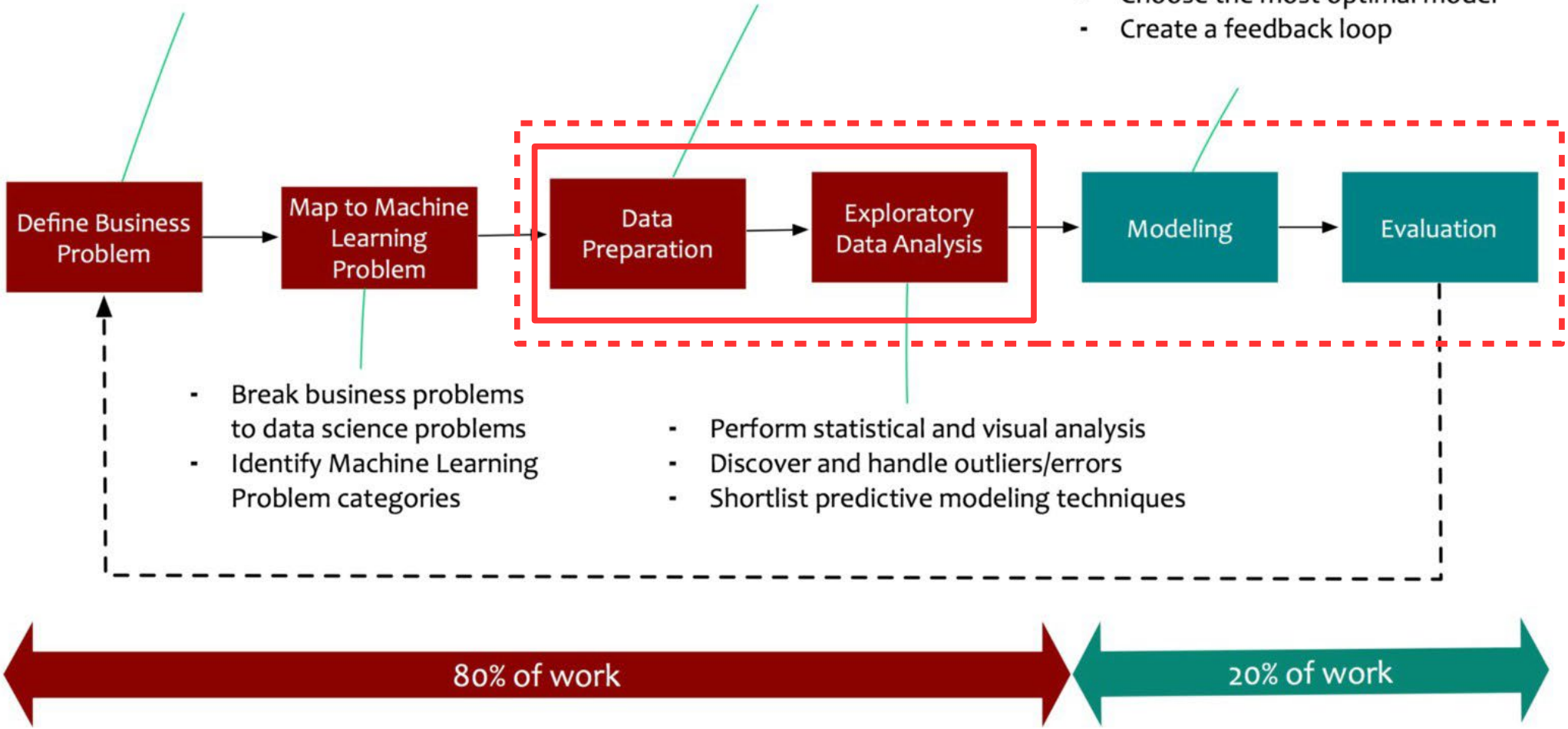


https://www.kdnuggets.com/2022/07/data-preparation-raw-data-machine-learning.html

- 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

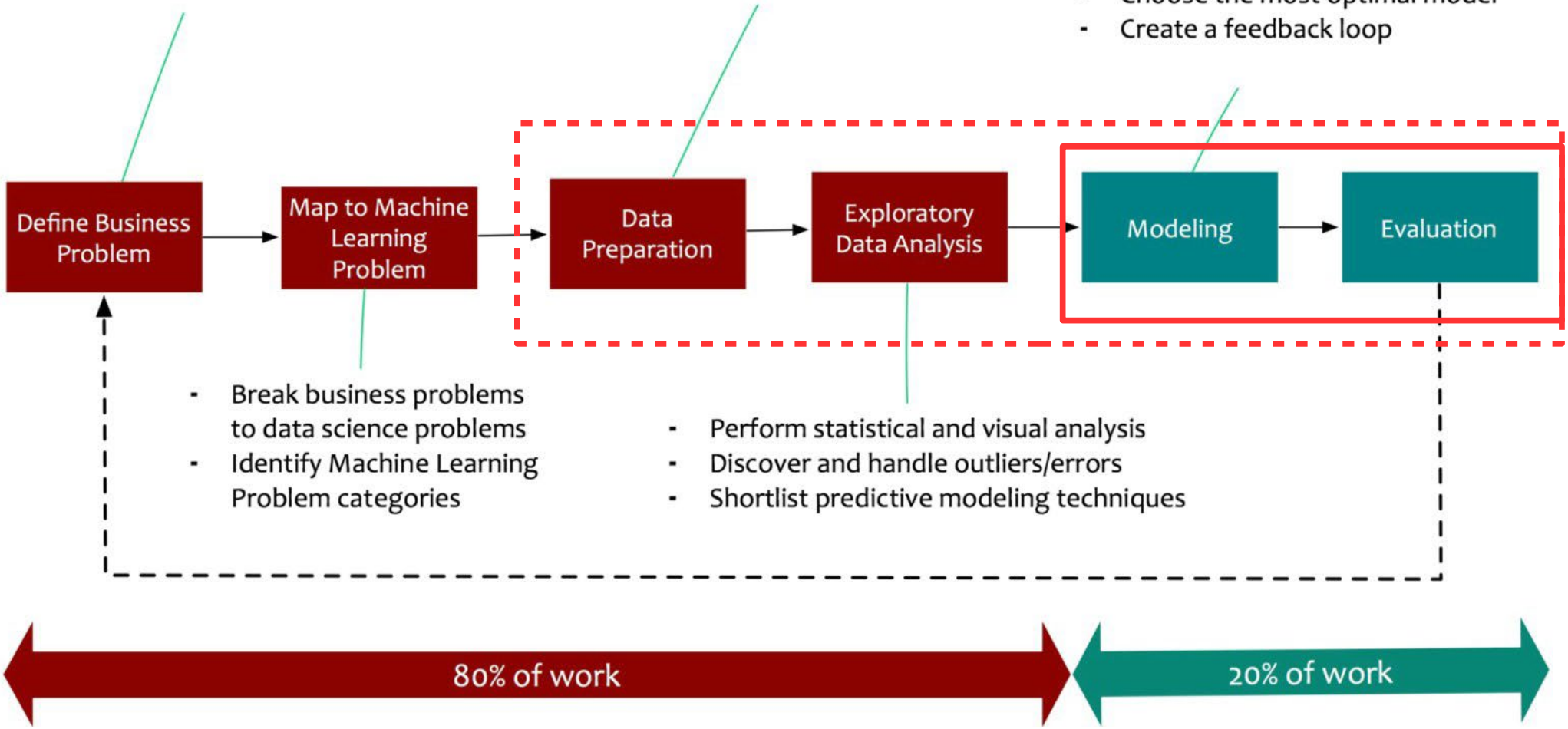


https://www.kdnuggets.com/2022/07/data-preparation-raw-data-machine-learning.html

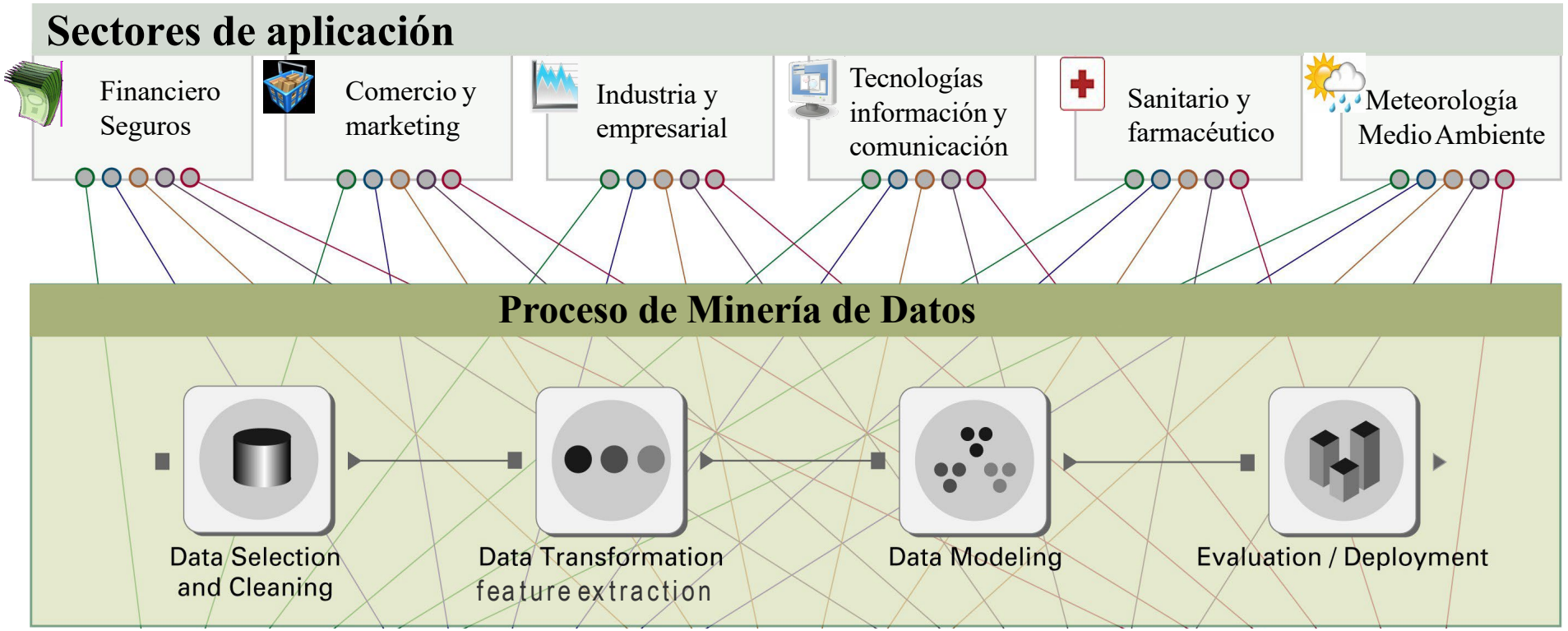
- Clearly defined business problem
- Set success criteria
- Define clear data science objectives

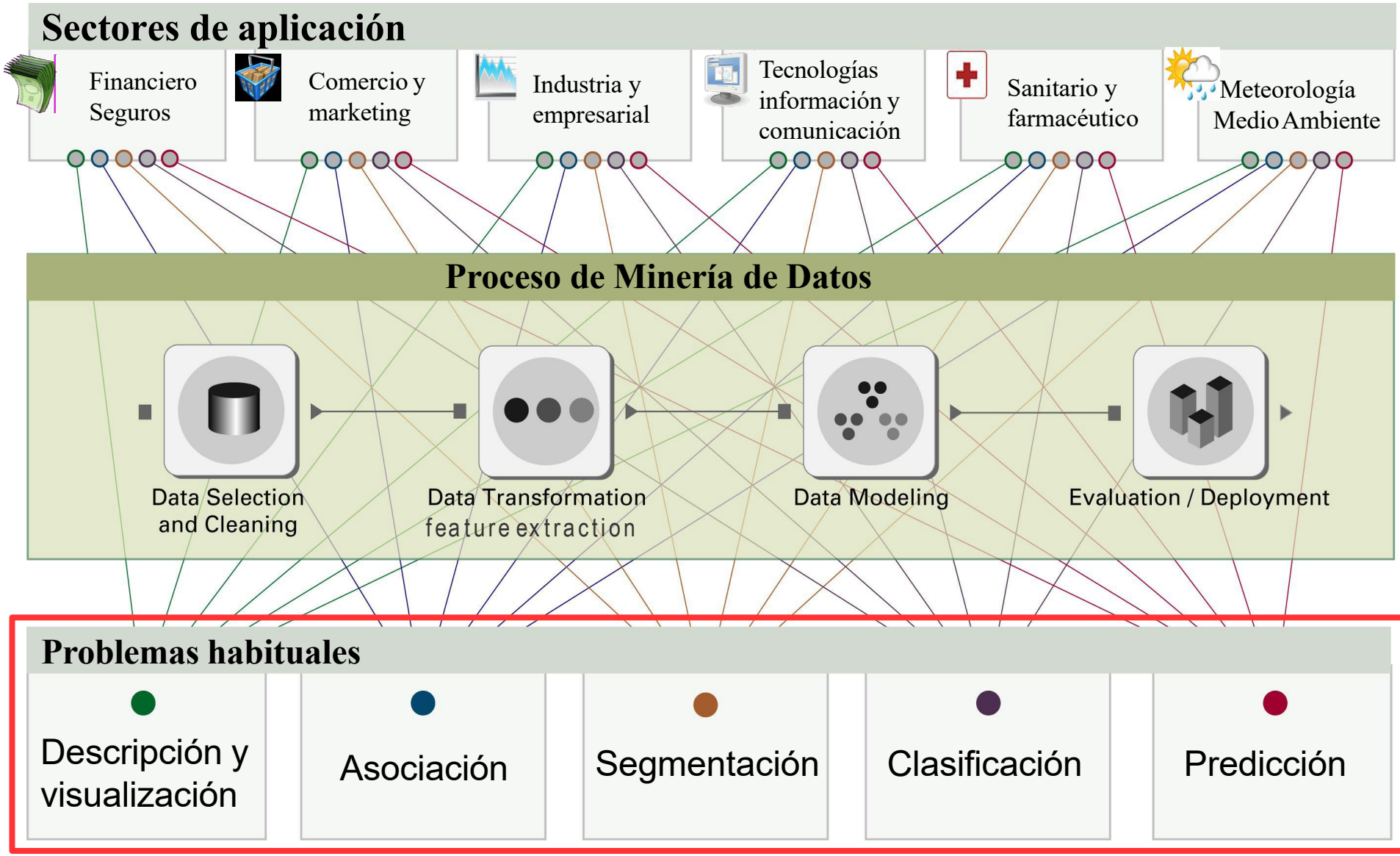
- Understand data points and constraints
- Formulate data analytics strategy
- Perform required transformation

- Experiment with multiple models
- Choose the most optimal model
- Create a feedback loop



https://www.kdnuggets.com/2021/05/essential-machine-learning-algorithms-beginners.html


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.

Sectores de aplicación

Financiero Seguros

Comercio y marketing

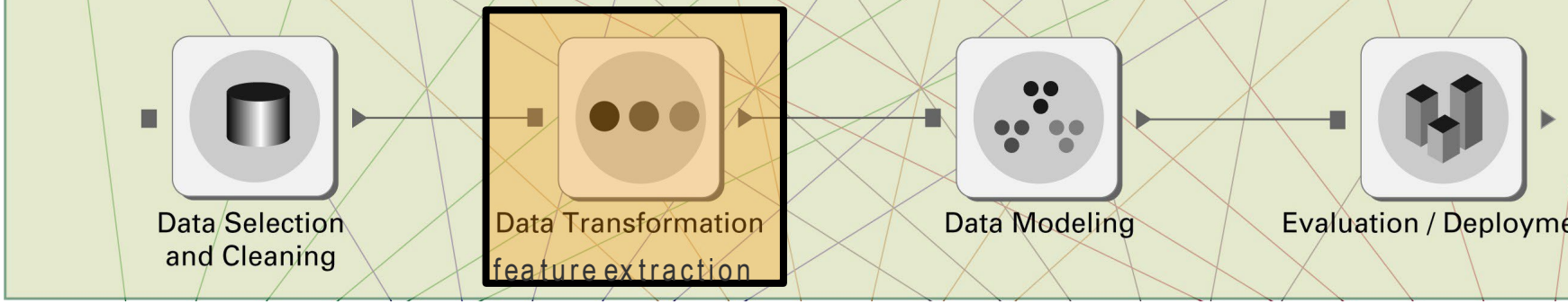
Industria y empresarial

Tecnologías información y comunicación

Sanitario y farmacéutico

Meteorología Medio Ambiente

Proceso de Minería de Datos



WIREs DATA MINING AND KNOWLEDGE DISCOVERY — WILEY

Explore this journal >

Advanced Review

Data discretization: taxonomy and big data challenge

ELSEVIER

Knowledge-Based Systems

Volume 86, September 2015, Pages 33-45

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

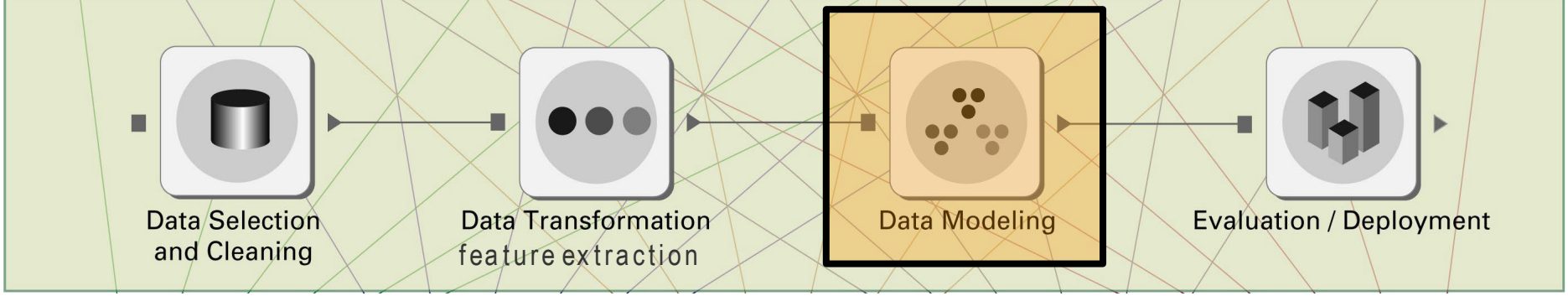
V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos

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

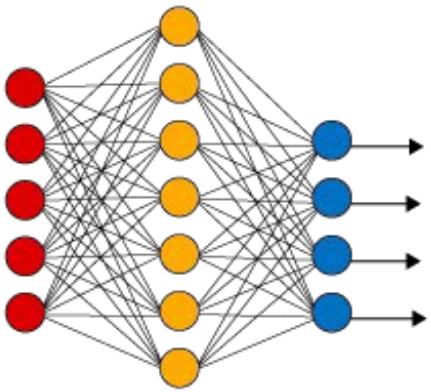
Sectores de aplicación

- Financiero Seguros
- Comercio y marketing
- Industria y empresarial
- Tecnologías información y comunicación
- Sanitario y farmacéutico
- Meteorología Medio Ambiente

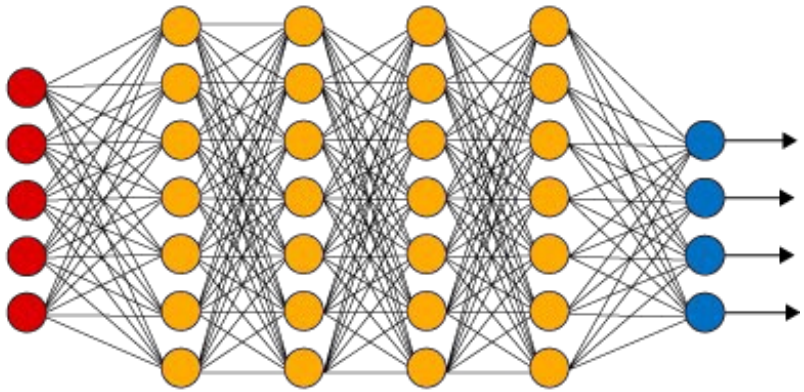
Proceso de Minería de Datos



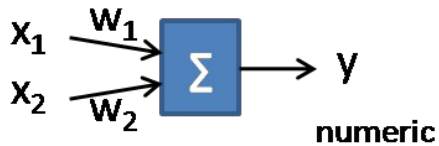
Simple Neural Network



Deep Learning Neural Network



Input Layer Hidden Layer Output Layer



$$y = w_0 + w_1x_1 + w_2x_2$$

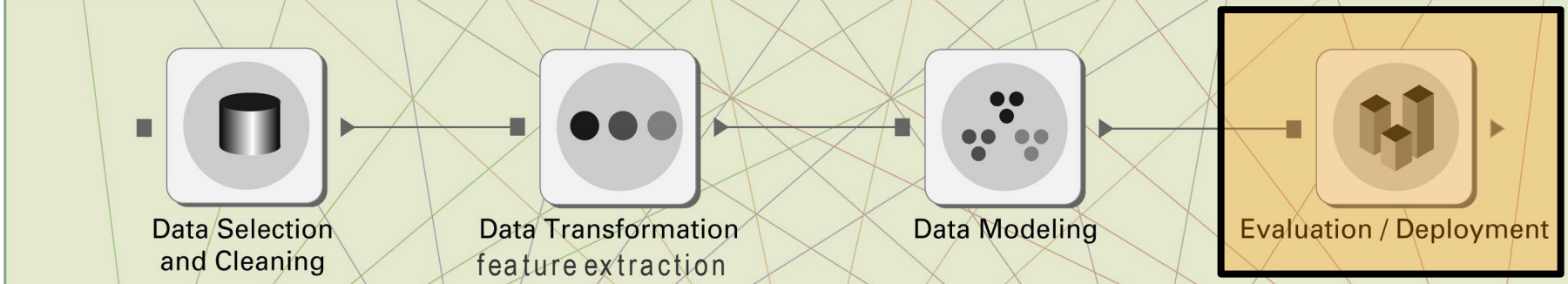
$$y = f(X,W) = X^T.W$$

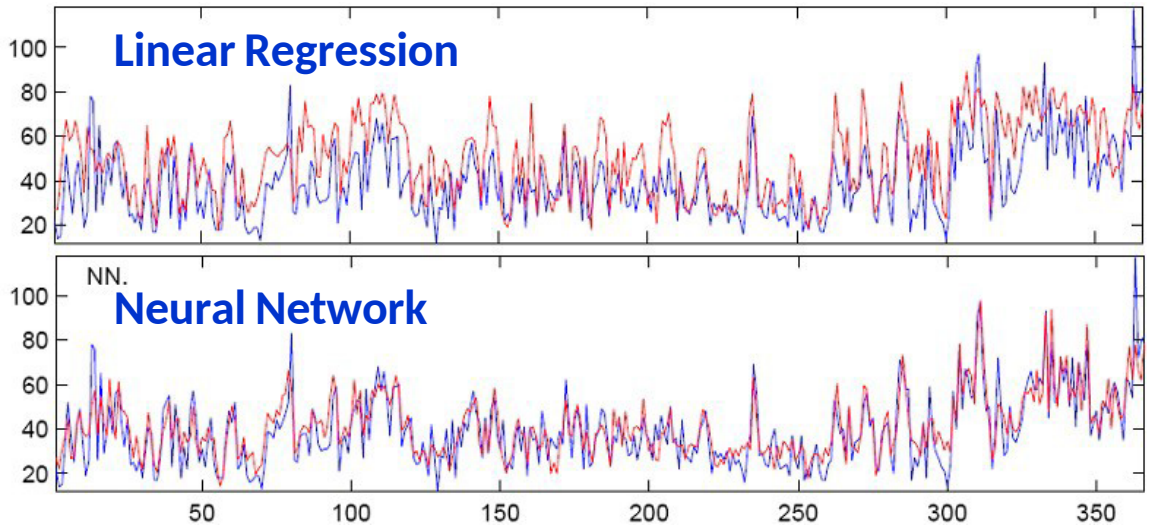
REGRESSION $W = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix}$ $X = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$

Sectores de aplicación

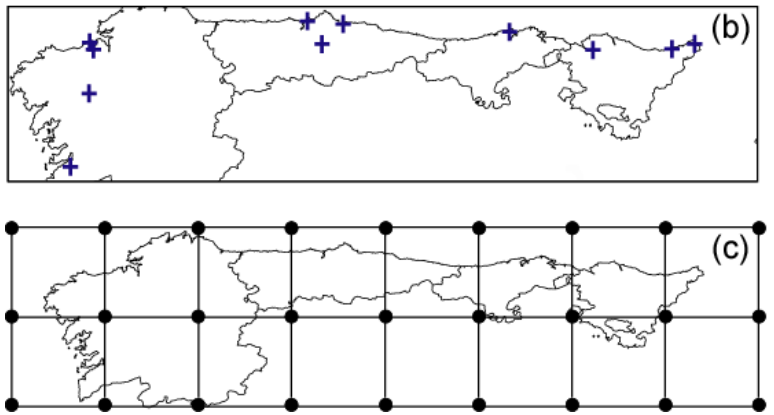
- Financiero Seguros
- Comercio y marketing
- Industria y empresarial
- Tecnologías información y comunicación
- Sanitario y farmacéutico
- Meteorología Medio Ambiente

Proceso de Minería de Datos





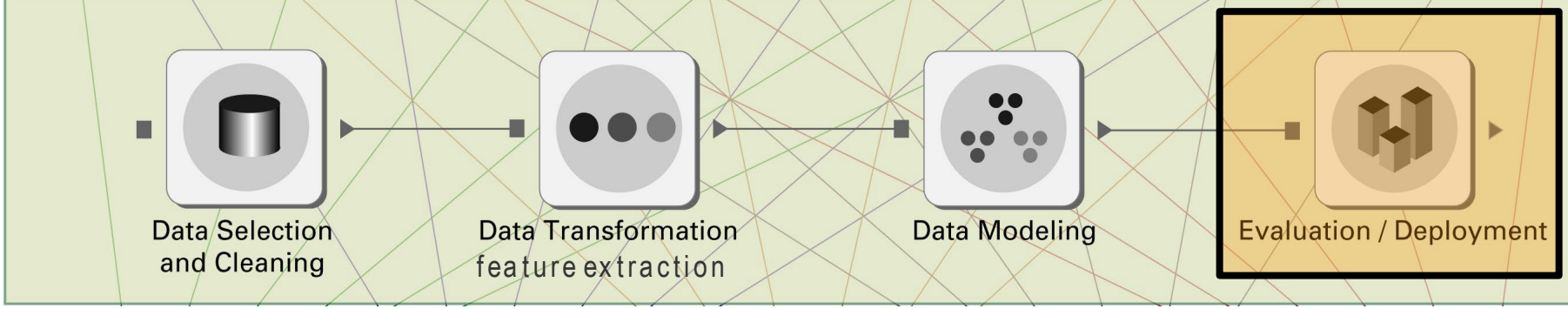
Wind Speed

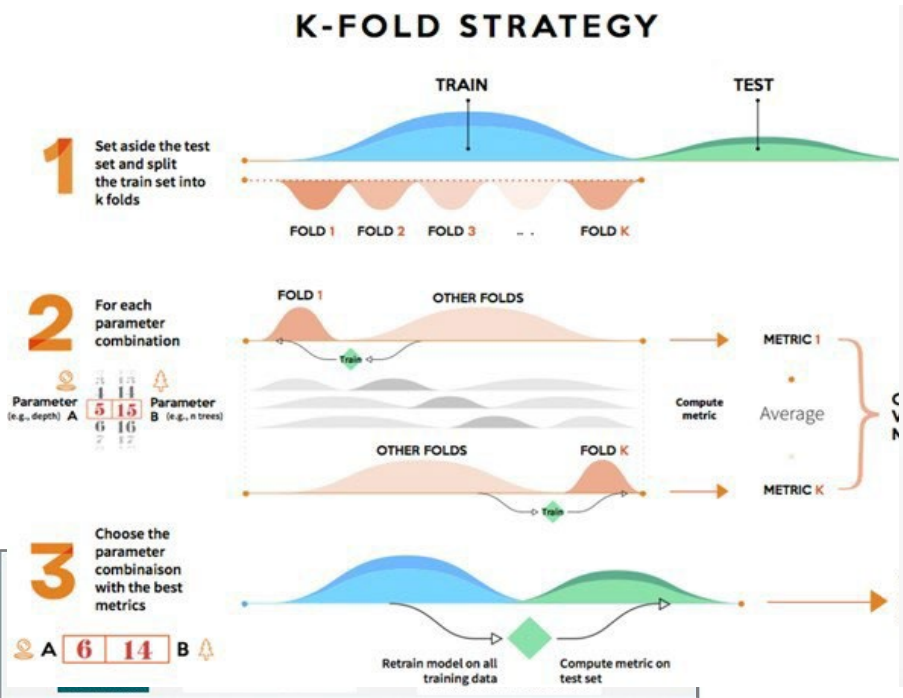


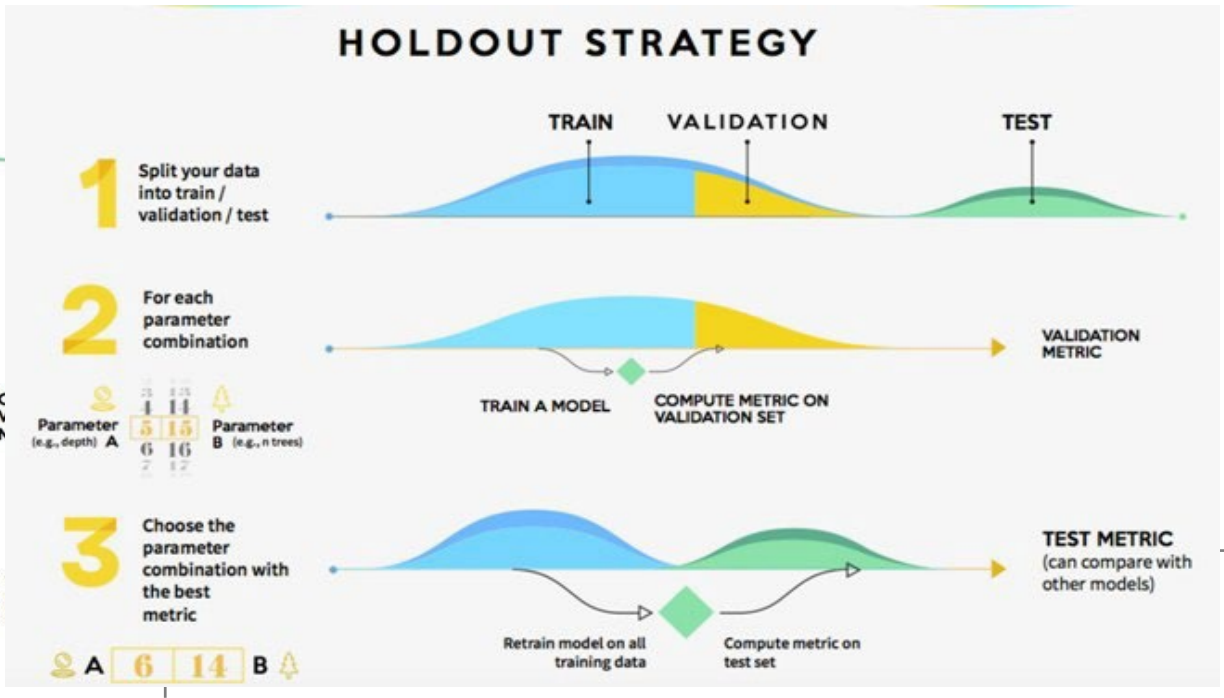
Sectores de aplicación

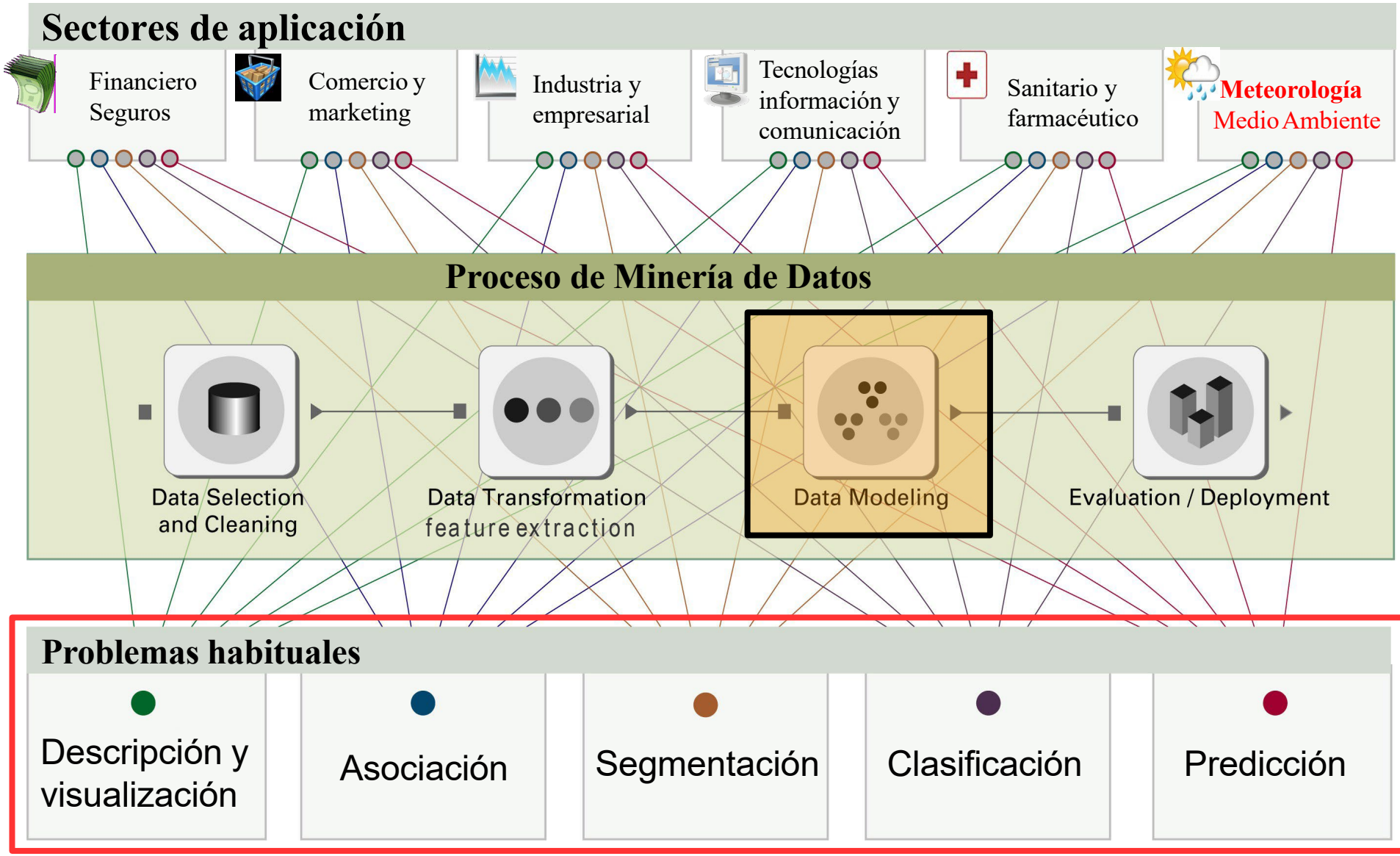
- Financiero Seguros
- Comercio y marketing
- Industria y empresarial
- Tecnologías información y comunicación
- Sanitario y farmacéutico
- Meteorología Medio Ambiente

Proceso de Minería de Datos









Machine learning develop methods for data modelling and prognosis.