

BIG DATA E INTELIGENCIA ARTIFICIAL PARA LA CIENCIA SOCIAL

JORDI MORALES I GRAS

euskal
soziologia eta
zientzia
politikoaren
elkartea



asociación
vasca
de sociología
y ciencia política

CONTENIDOS DEL CURSO

SESIÓN 1

- El nuevo "datascape" de la Ciencia Social.
- Introducción a la Inteligencia Artificial, al Machine Learning:
 - Clarificación conceptual
 - Tipos de algoritmos para el Machine Learning
- Python y las Ciencias Sociales
- [Práctica con Orange Data Mining](#)

SESIÓN 2

- Técnicas de aprendizaje supervisado
 - Algoritmos para la clasificación o categorización
 - Algoritmos para la predicción de valores numéricos
- [Práctica con Orange Data Mining](#)

SESIÓN 3

- Técnicas de aprendizaje no supervisado y de Deep Learning: clusterización e incrustaciones
- [Práctica con Orange Data Mining](#)
- Justicia de datos

GUÍA DE PRÁCTICAS

Orange datamining

Funcionamiento de la herramienta

Limpieza y tratamiento de datos

Minando encuestas

Minando textos

Aprendizaje supervisado

Inferencia y predicción con datos de encuesta

Entrenamiento y predicción en textos

Análisis de sentimiento

Aprendizaje no supervisado

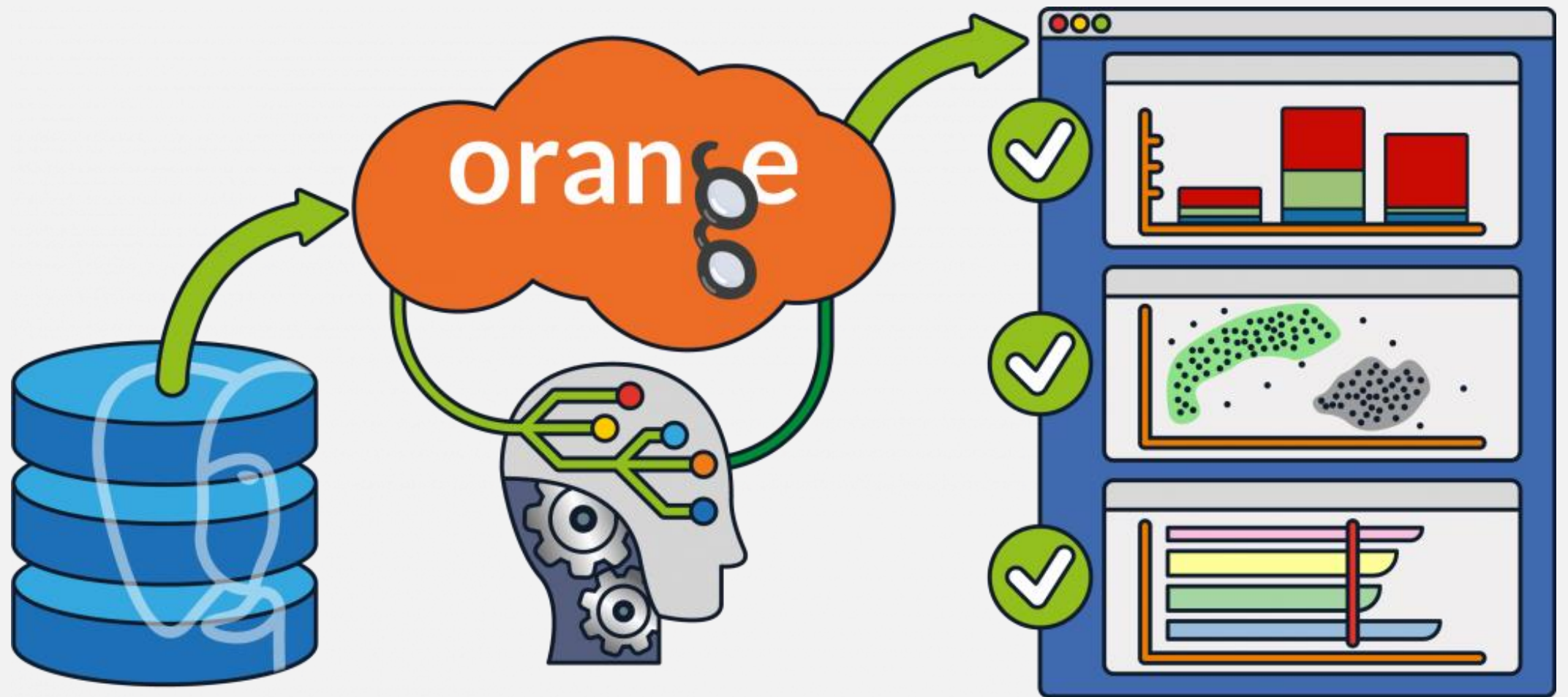
Clústering en textos (Topic Modelling)

Clústering para imágenes

HORARIOS DEL CURSO

- Inicio: 16:00
- Descanso (breve): 17:30 – 17:35
- Finalización: 18:50





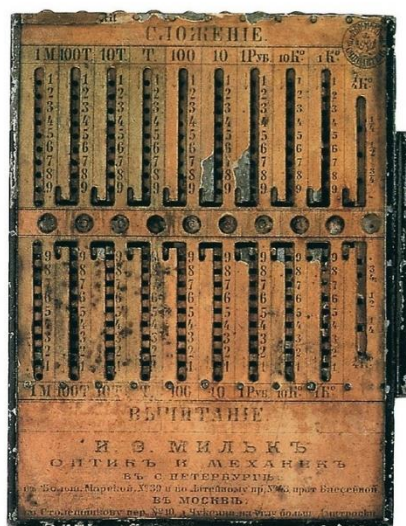
EL NUEVO DATASCAPE DE LA CIENCIA SOCIAL

Alain Desrosières
La politique
des grands nombres

Histoire de la raison statistique



La Découverte/Poche



IDEAS IN CONTEXT

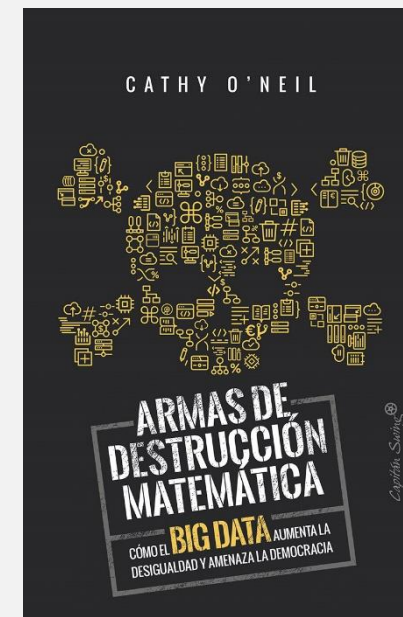
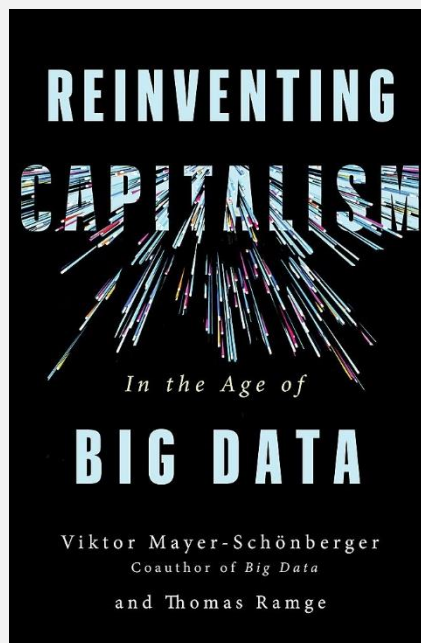
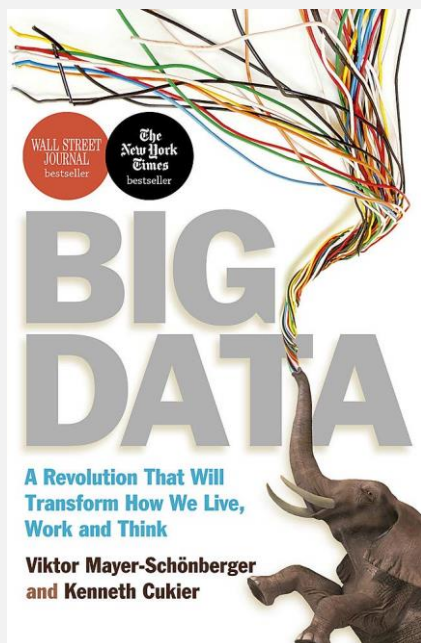
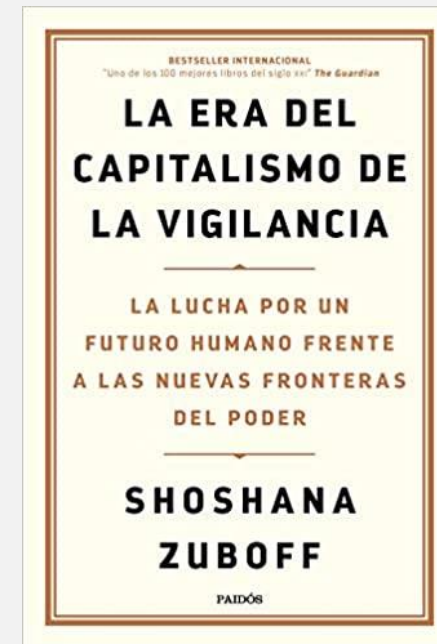
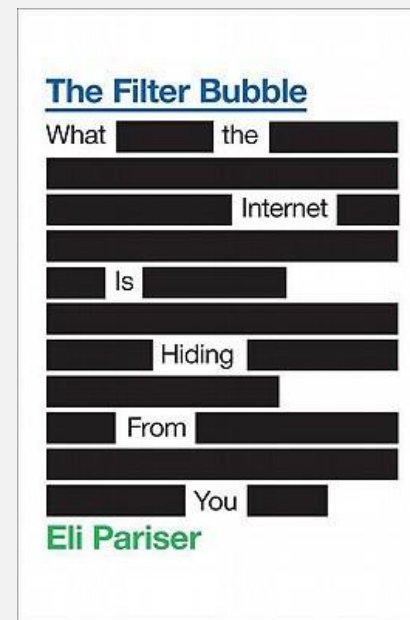
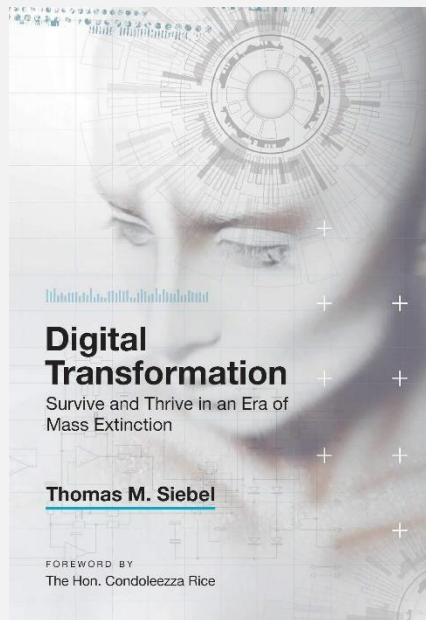
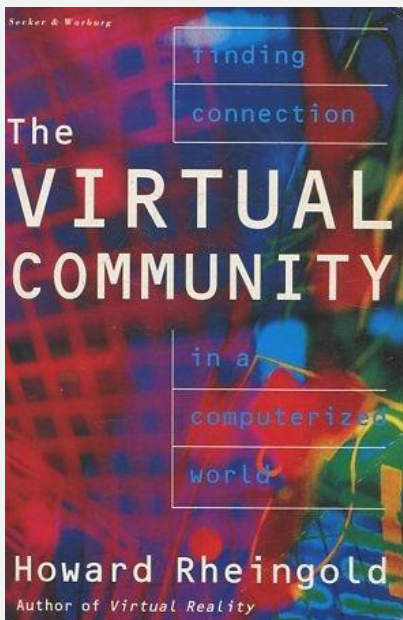
THE Taming
OF Chance



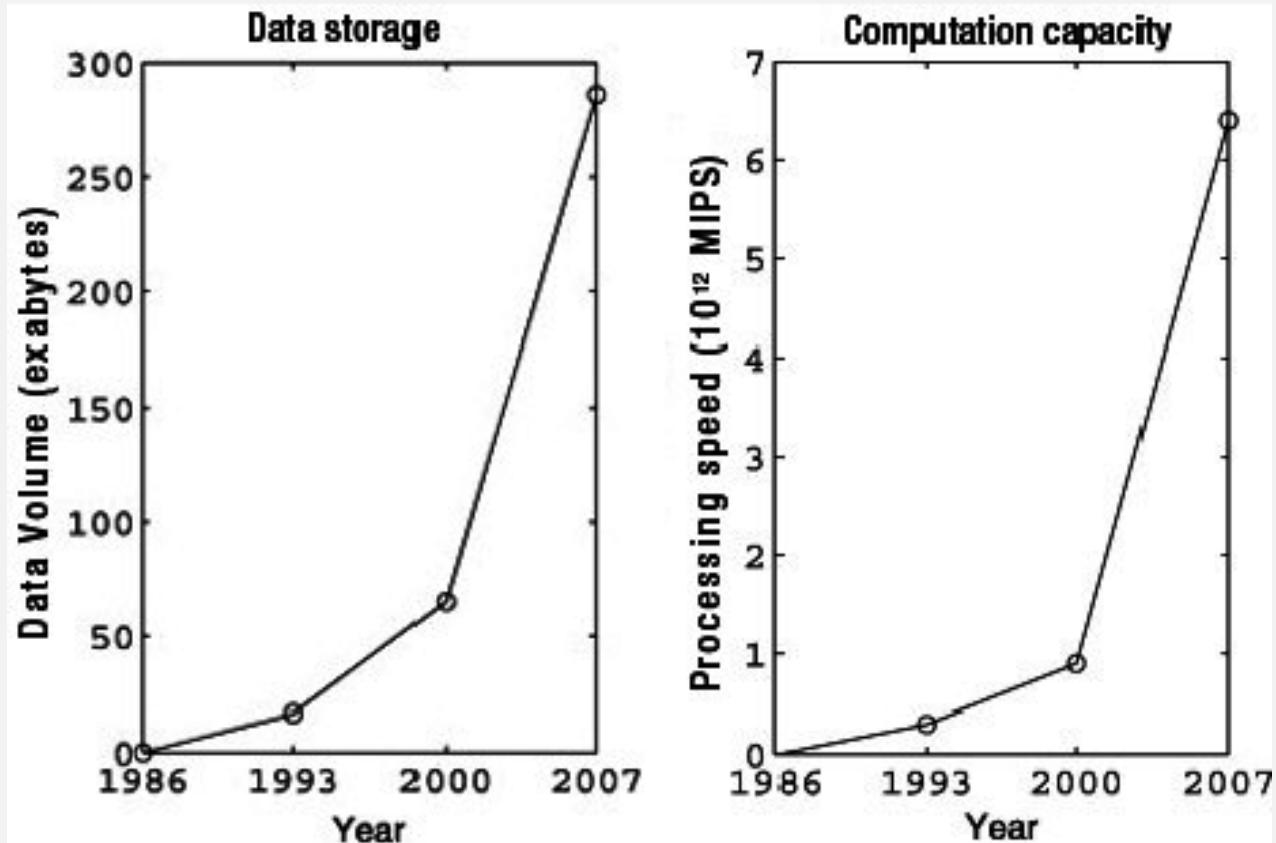
IAN HACKING



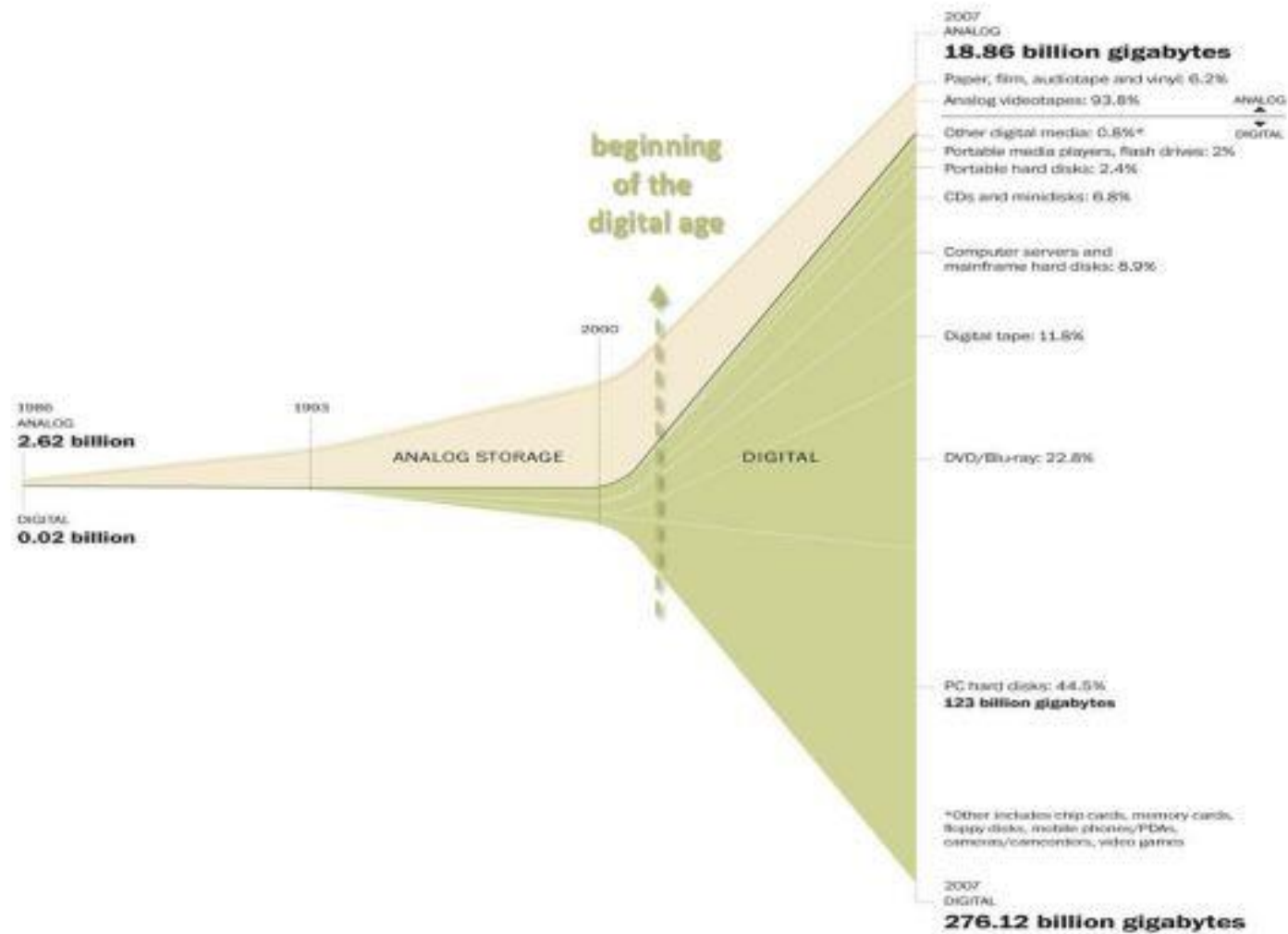
SAGE campus







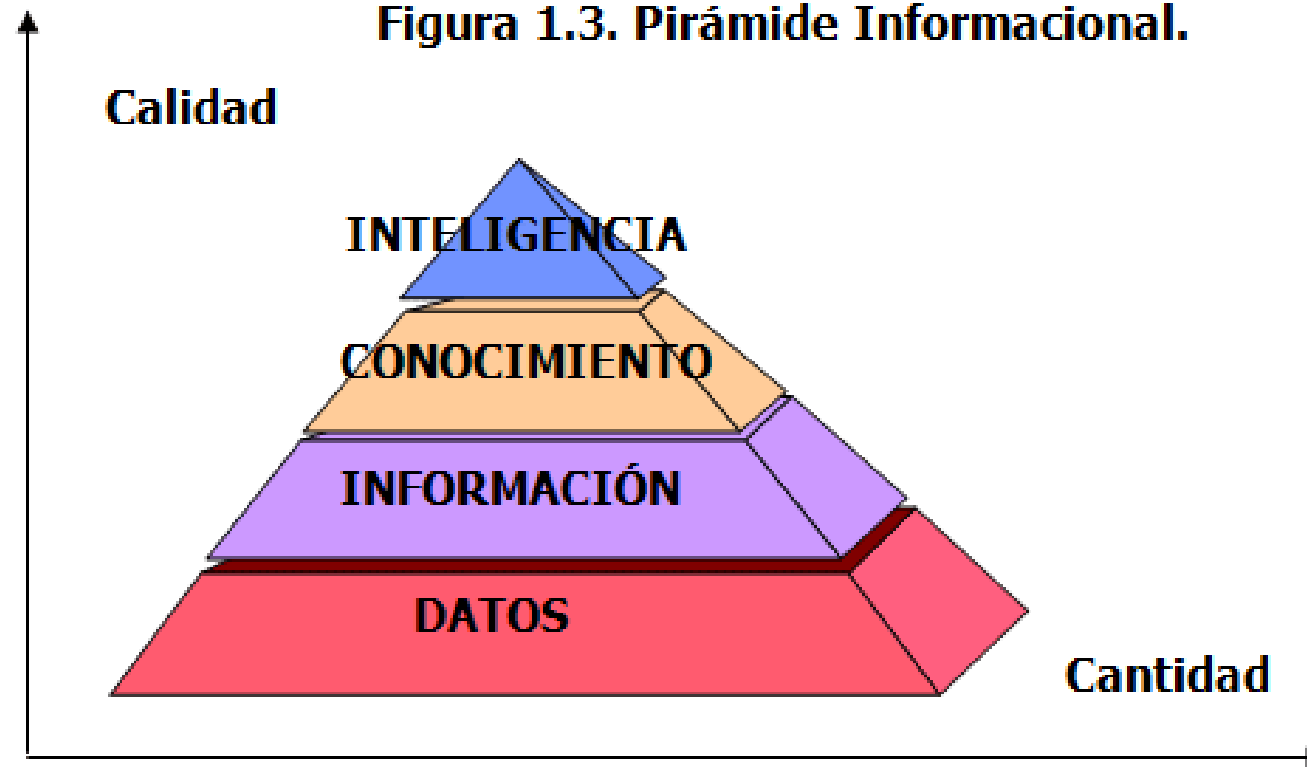
M. Hilbert and P. López, 2011.
The world's technological
capacity to store,
communicate, and compute
information.



Washington Post
Hilbert and L

M. Hilbert and P. López, 2011.
The world's technological
capacity to store,
communicate, and compute
information.

Figura 1.3. Pirámide Informacional.



Fuente: Dante, Ponjuán, Gloria. Gestión de la información en las organizaciones. Principios, conceptos y aplicaciones. Santiago de Chile, 1998.



ESTRATEGIAS CLÁSICAS PARA LA OBTENCIÓN DE DATOS SOCIALES



FUENTES DE DATOS MASIVOS

MEDIOS SOCIALES

- Imagen, video, audio o texto de redes sociales virtuales

CLOUD

- Público, privado o corporativo

WEB

- Datos web, analytics

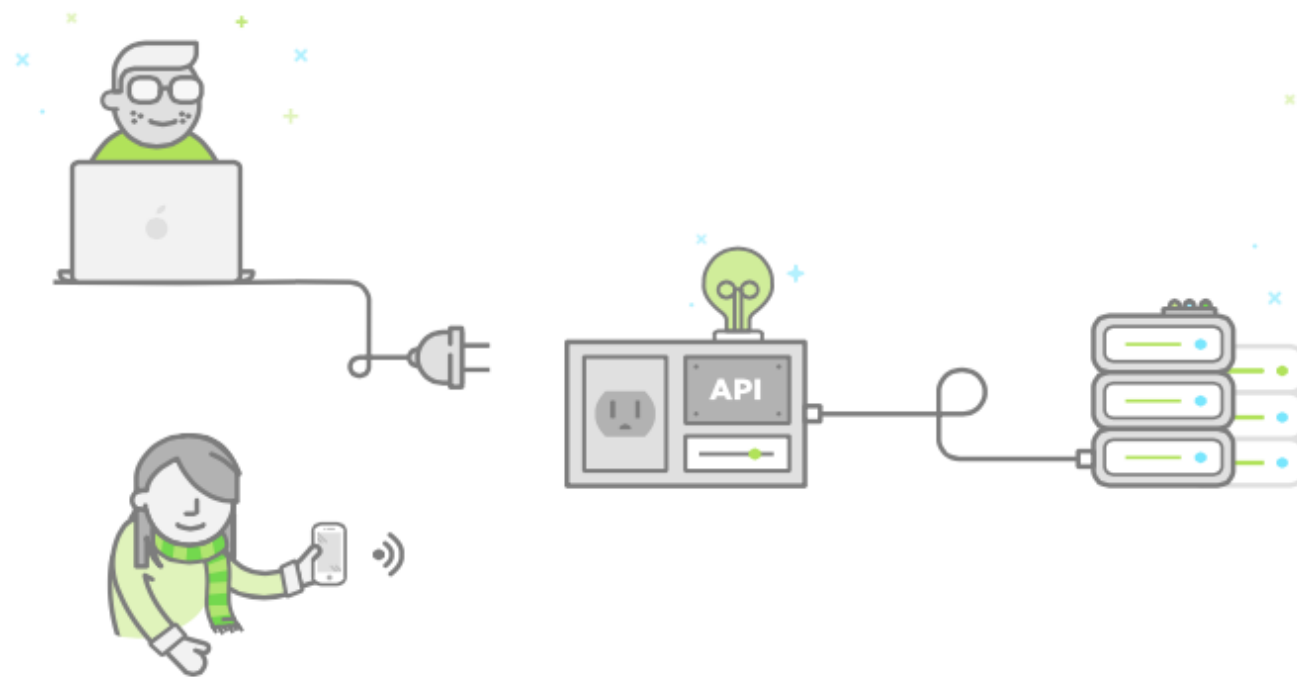
IoT

- Sensores y dispositivos conectados

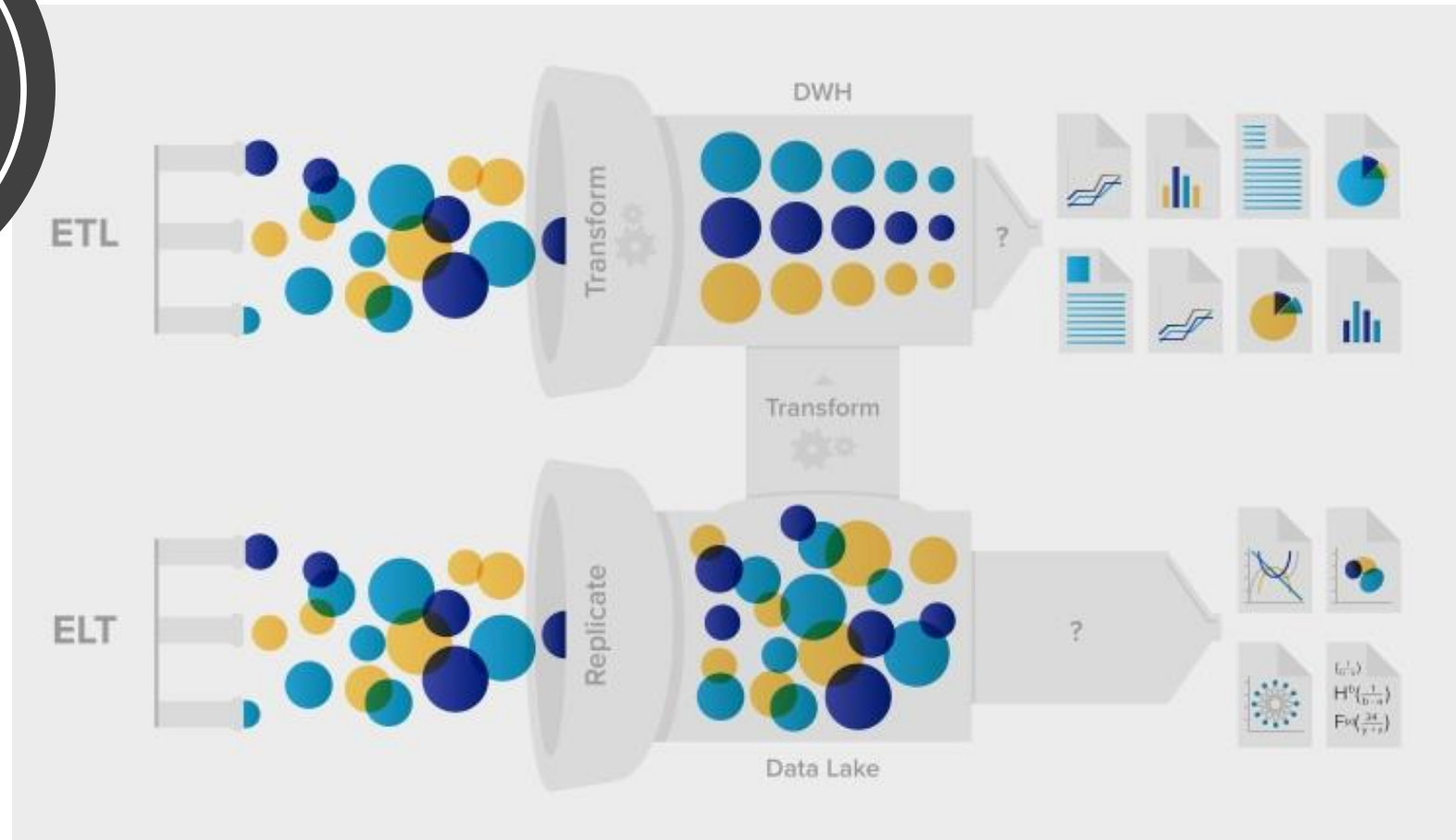
BASES DE DATOS

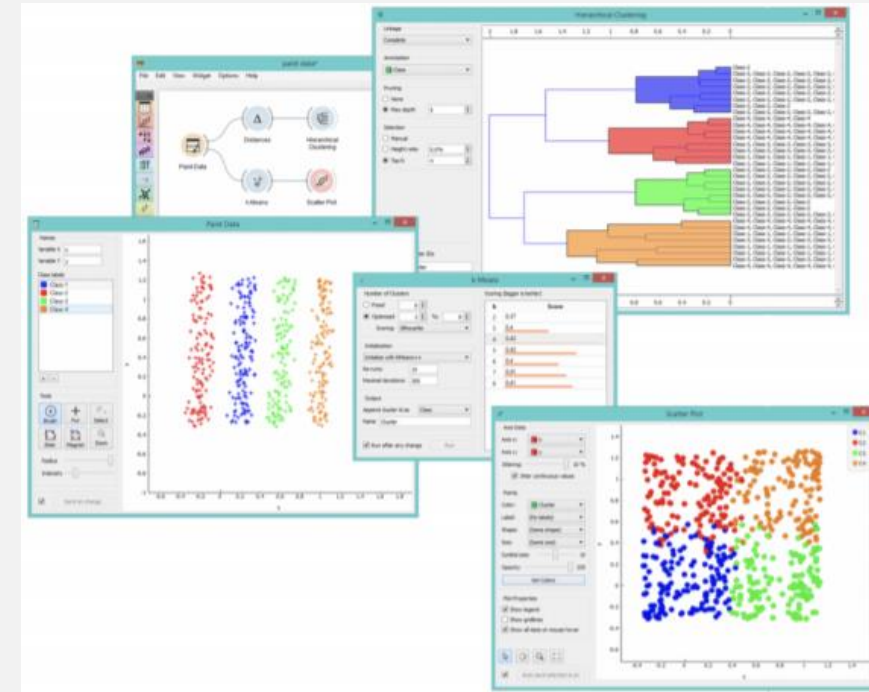
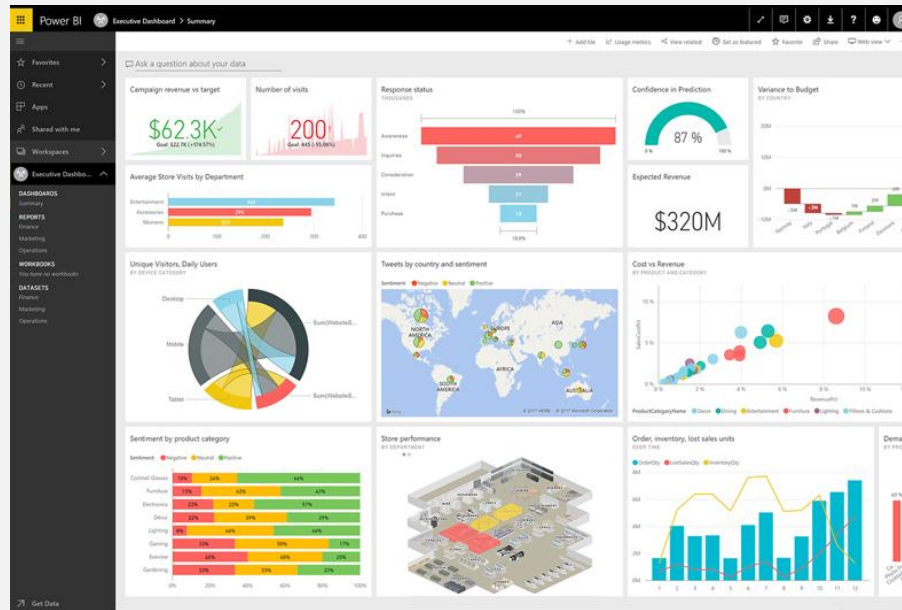
- Datos secundarios tradicionales

EL NUEVO ESCENARIO DE DATOS SOCIALES



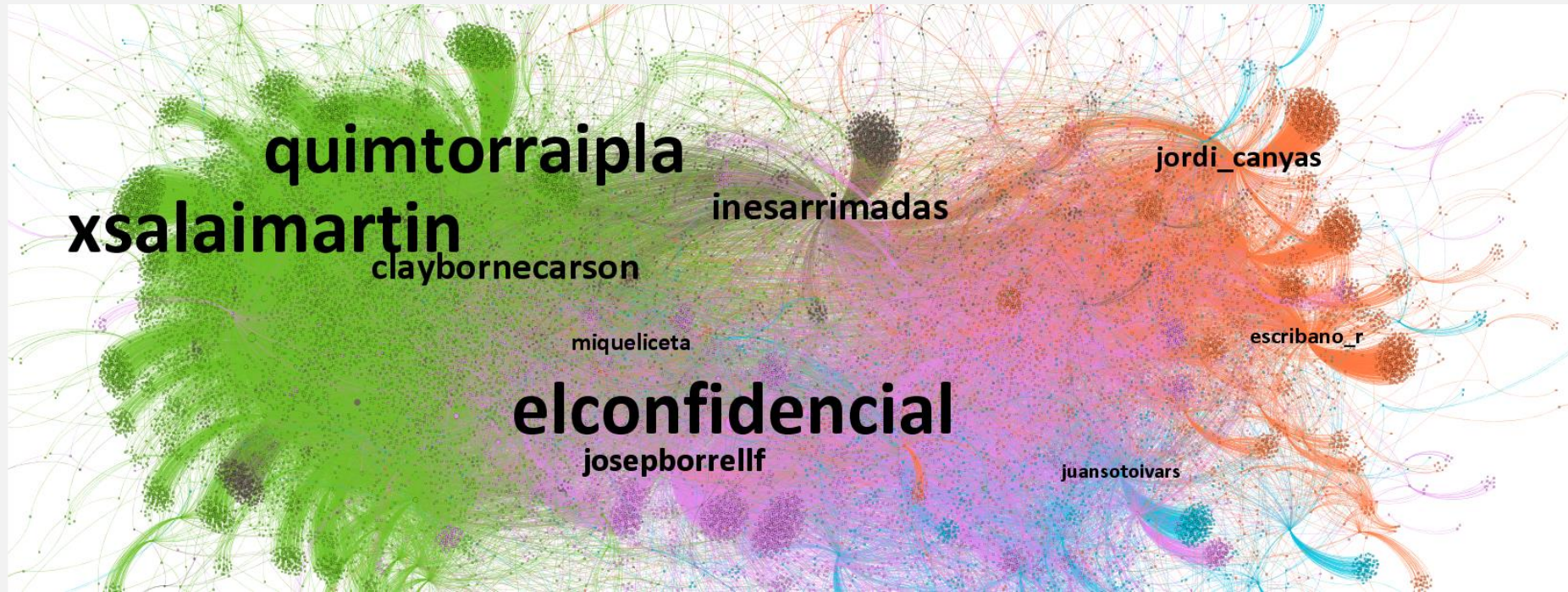
ETL VS ELT



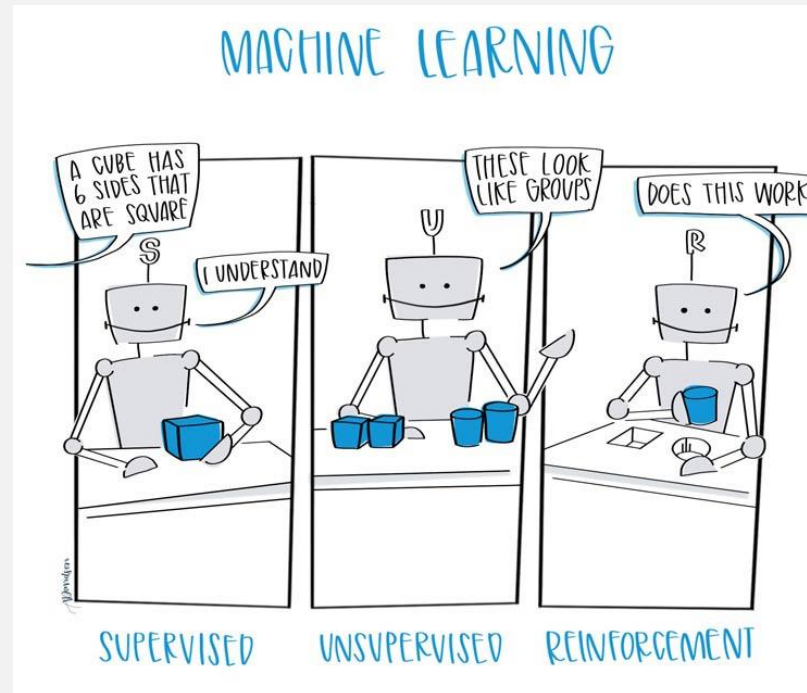


BI (ETL) VS. DATAMINING (ELT)

NUEVAS HERRAMIENTAS PARA LA CIENCIA SOCIAL



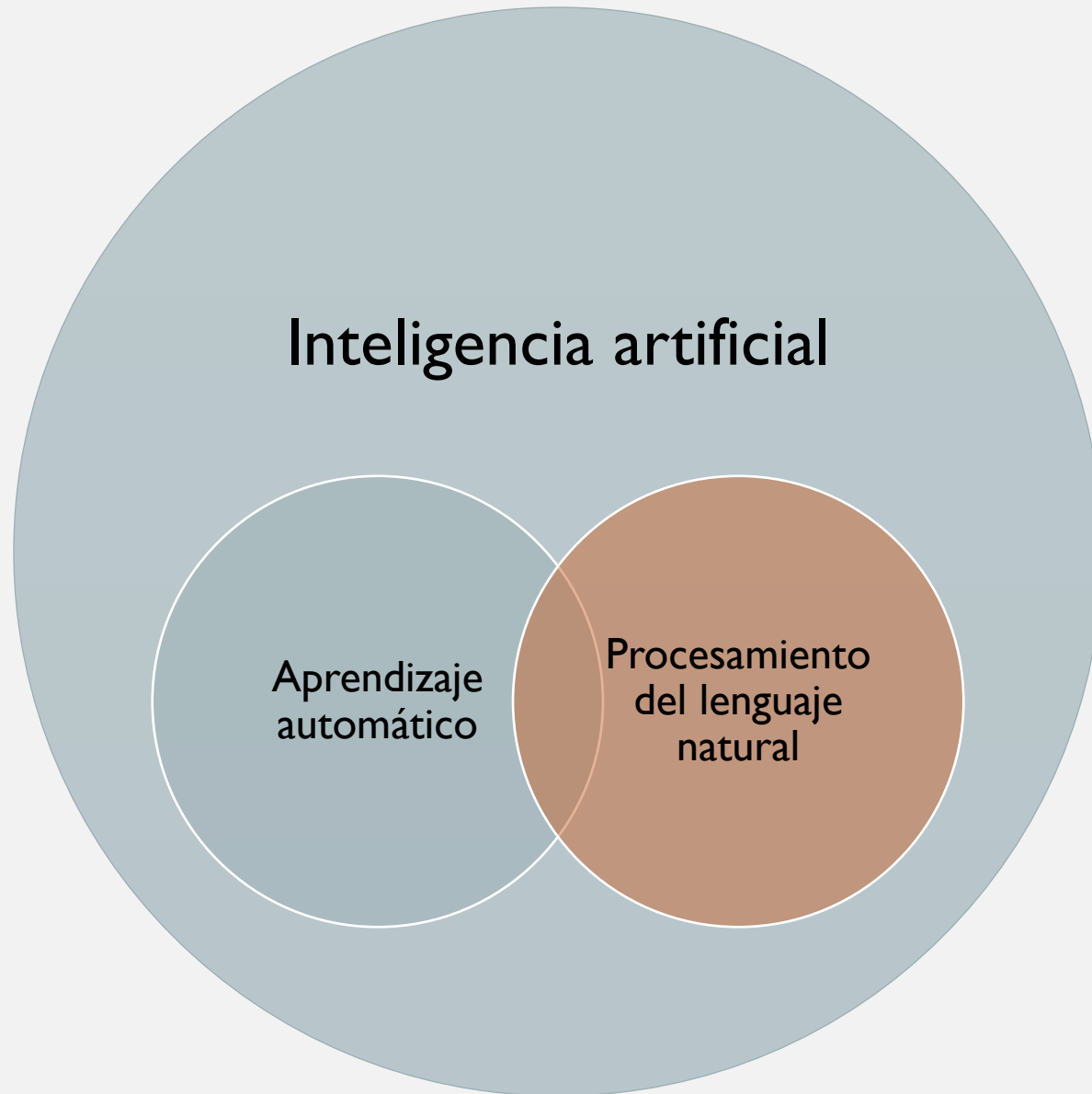
EL ANÁLISIS DE REDES SOCIALES



EL APRENDIZAJE AUTOMÁTICO O MACHINE LEARNING (ML)

CLARIFICACIÓN CONCEPTUAL

PROCESAMIENTO DEL LENGUAJE NATURAL



ALGORITMOS DE REGLAS HEURÍSTICAS

Tokenize on rules Let 's tokenize ! Is n't this easy ?

Tokenize on punctuation Let ' s tokenize ! Isn ' t this easy ?

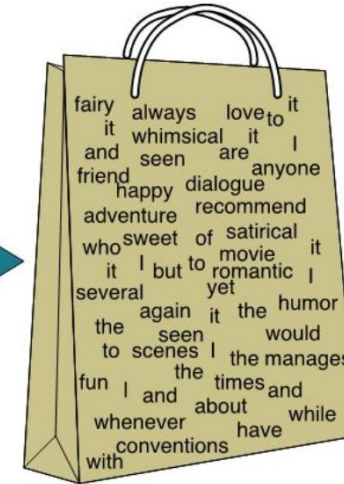
Tokenize on white spaces Let's tokenize! Isn't this easy?

Let's tokenize! Isn't this easy?

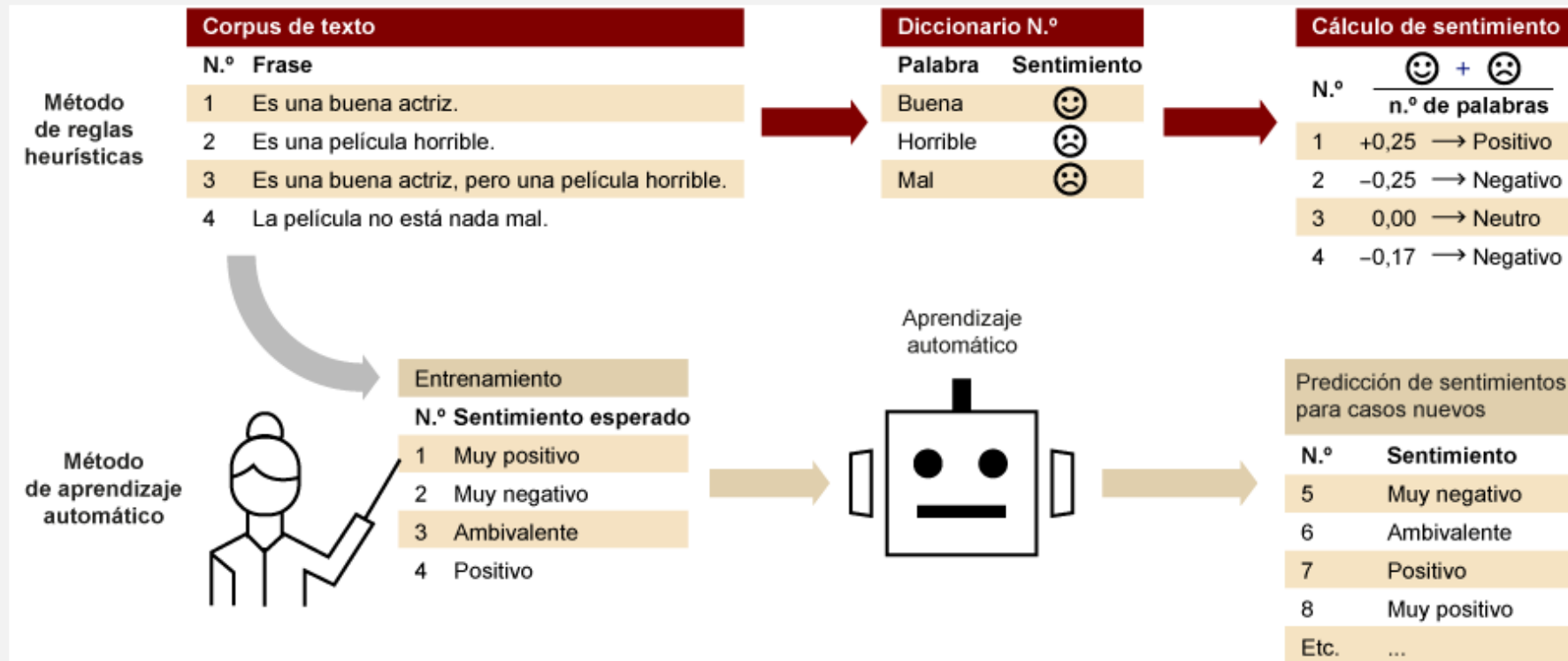
Word	Lemmatization	Stemming
was	be	wa
studies	study	studi
studying	study	study

The Bag of Words Representation

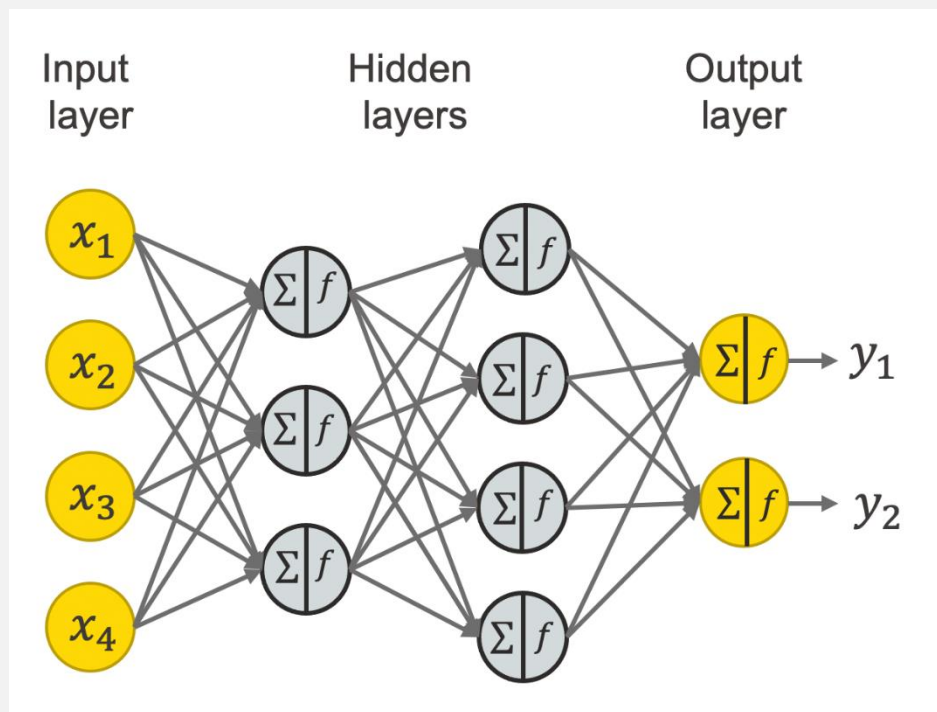
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...



DE LAS REGLAS HEURÍSTICAS AL
APRENDIZAJE AUTOMÁTICO



PLN + REDES NEURONALES

MACHINE LEARNING

STATISTICS



VS

**MACHINE
LEARNING**



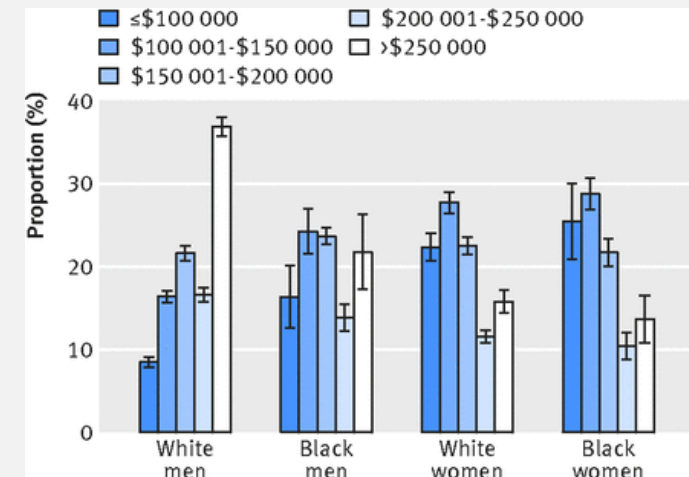


ESTADÍSTICA (EJEMPLO)

Sexo o fenotipo (raza) ¿Qué afecta más al sueldo?

Differences in incomes of physicians in the United States by race and sex: observational study (2016)

- 36.9% of **white male** physicians had adjusted annual income over \$250 000 compared with 21.8% of **black male** physicians ($P < 0.001$ for difference)
- 15.8% of **white female** physicians earned more than \$250 000 compared with 13.6% of **black female** physicians ($P = 0.09$ for difference)
- **White male** physicians earn substantially more than **black male** physicians, while **white and black female physicians** earn similar incomes to each other, but significantly less than their **male counterparts**



Modelo explicativo/inferencial

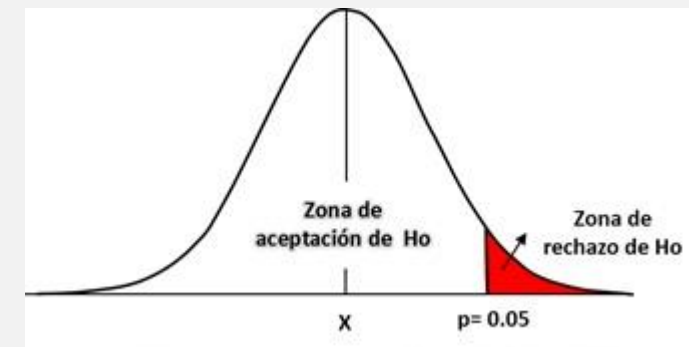
Sexo y Raza (Variables Independientes) → Sueldo (Variable dependiente)

Datos

2000-13 American Community Survey (61.327)

2000-08 Center for Studying Health System Change Survey (17.583)

Obtengo Respuestas



MACHINE LEARNING (EJEMPLO)

Quiero predecir el sueldo de un individuo (e.g. de un físico) en base a sus datos

- Sexo, Raza y Sueldo... Edad, Universidad(es) en las que ha estudiado, Becas disfrutadas, Años en activo, Sector de empleo, Ciudad de residencia, Profesión de los progenitores, Idiomas, Estado civil, N° de hijos...
- Amistades en Facebook y Twitter, Seguros contratados, Datos de consumo (tarjeta de crédito), Aficiones, Horas de navegación en Internet, Capital inmobiliario...
- Opción sexual, Creencias religiosas, Probabilidad de padecer cáncer...

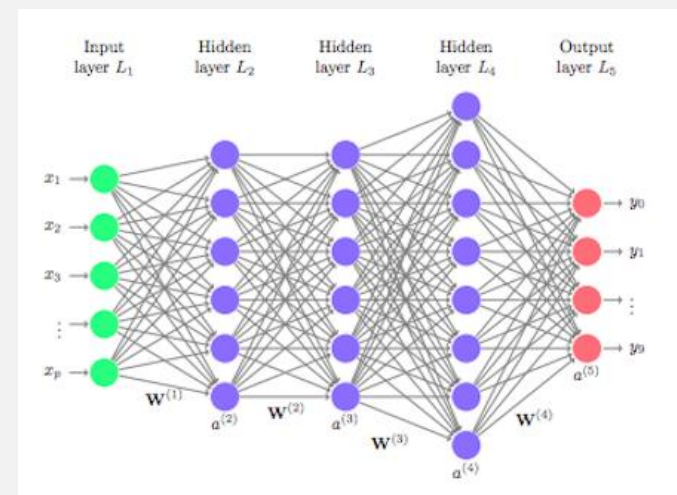
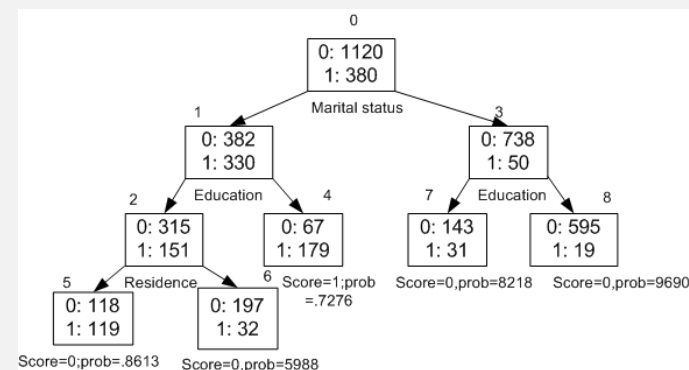
Datos

Introduzco tantas variables (*features*) como quiera (o pueda), de tantas fuentes de datos como sea capaz de reunir

Respuestas

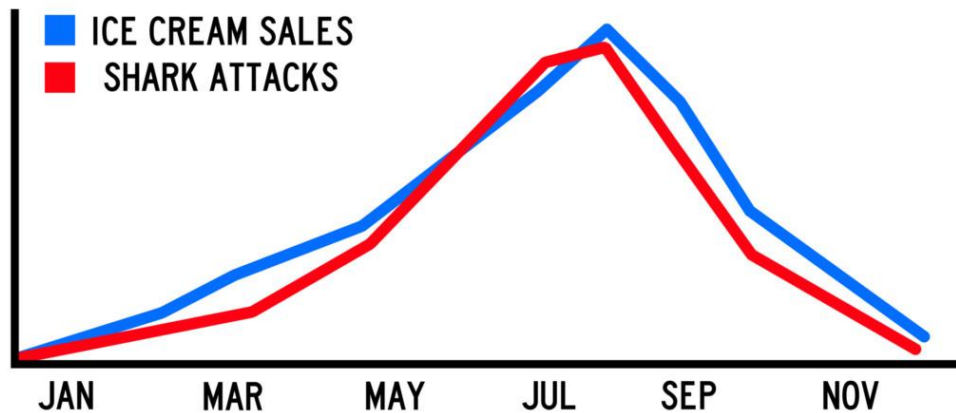
Entreno mi modelo para predecir la variable SUELDO (*target variable*) en base a casos conocidos y hago un test de validez

Obtengo un **modelo probabilístico-predictivo**

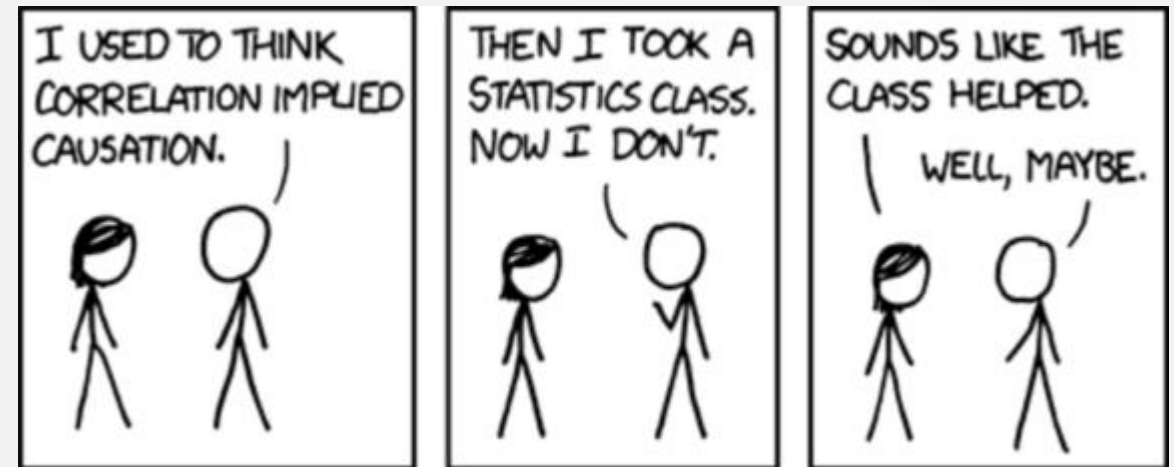


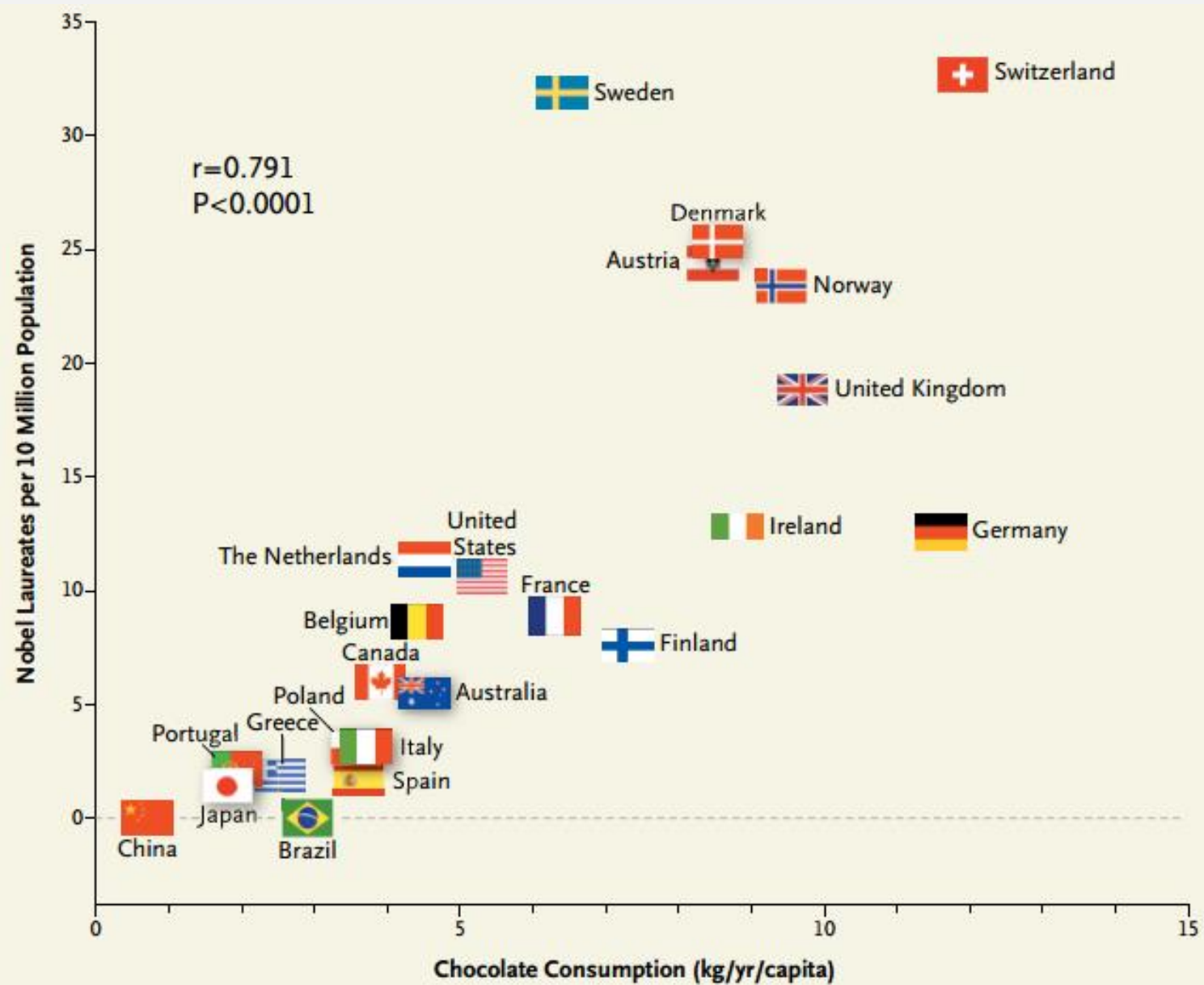
CORRELACIÓN Y CAUSALIDAD

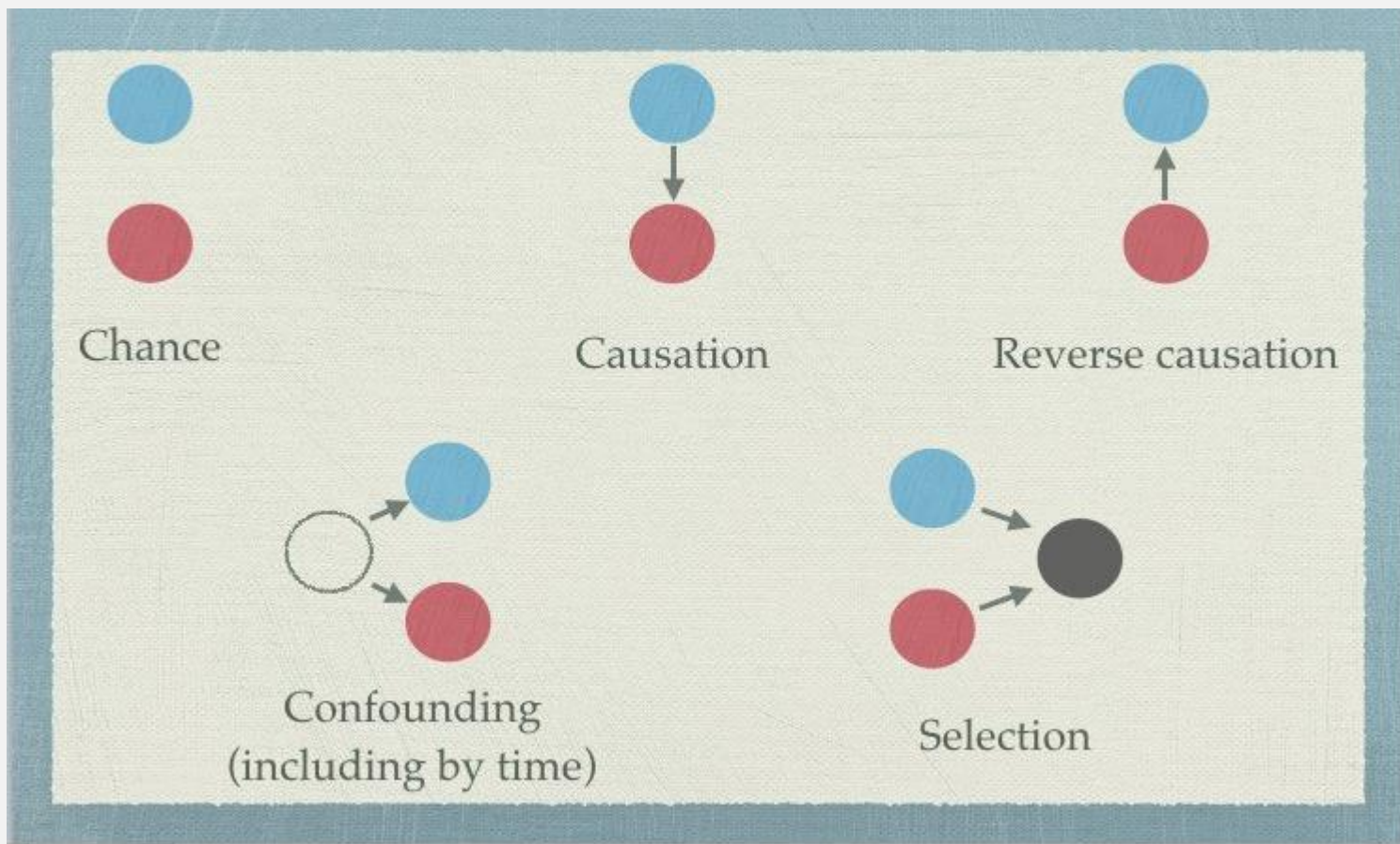
CORRELATION IS NOT CAUSATION!



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)







LA CAUSALIDAD
ES **UN TIPO DE**
CORRELACIÓN

ESTADÍSTICA VS. MACHINE LEARNING

ESTADÍSTICA

- + Explicación para datos disponibles (inferencia)
- + Orientada a la explicación de las relaciones entre variables (formalización, interpretación)
- + Principio de Parsimonia
- Despreocupación (o incredulidad) respecto a la capacidad predictiva de los modelos de datos

MACHINE LEARNING

- + Entrenamiento y testeo para datos futuros (predicción)
- + Orientado a la obtención de resultados replicables
- + Capacidad de computación
- Desinterés por la interpretación de los modelos de datos

MACHINE LEARNING, INTELIGENCIA ARTIFICIAL Y DEEP LEARNING

Artificial Intelligence



Any technique that enables computers to mimic human intelligence. It includes *machine learning*

Machine Learning



A subset of AI that includes techniques that enable machines to improve at tasks with experience. It includes *deep learning*

Deep Learning



A subset of machine learning based on neural networks that permit a machine to train itself to perform a task.

ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



MACHINE LEARNING

Machine learning begins to flourish.



DEEP LEARNING

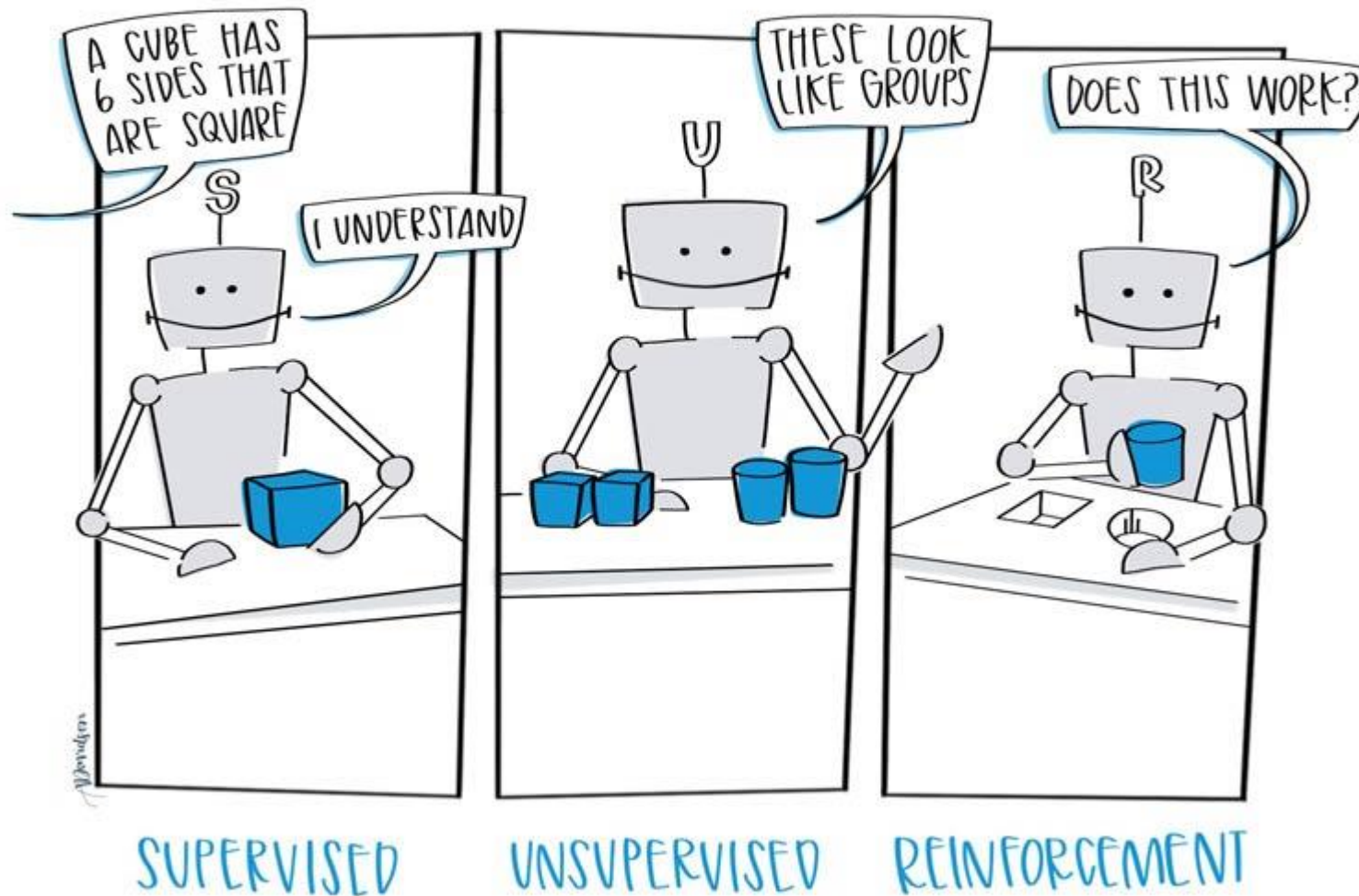
Deep learning breakthroughs drive AI boom.



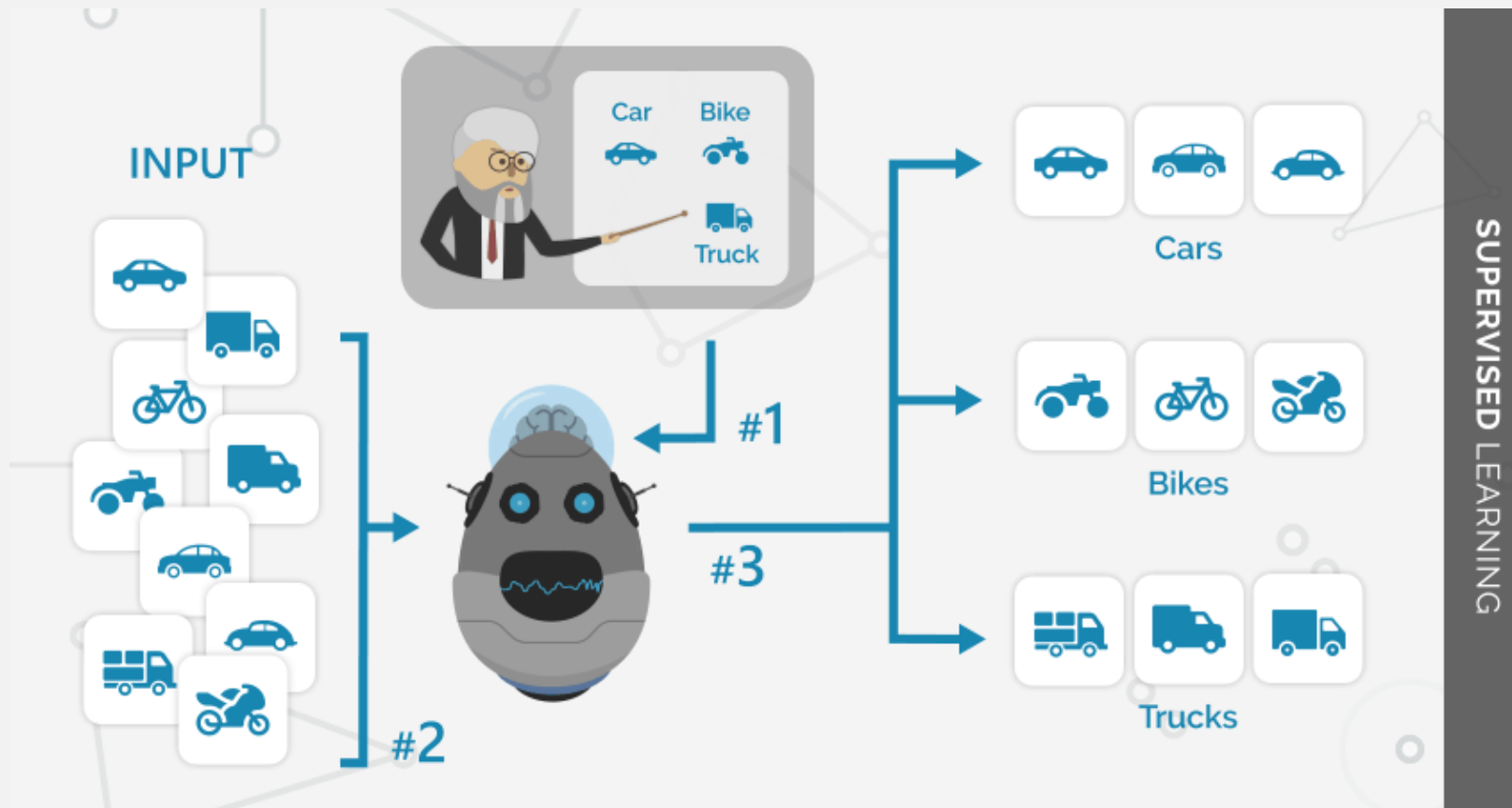
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

TIPOS DE ALGORITMOS PARA EL MACHINE LEARNING

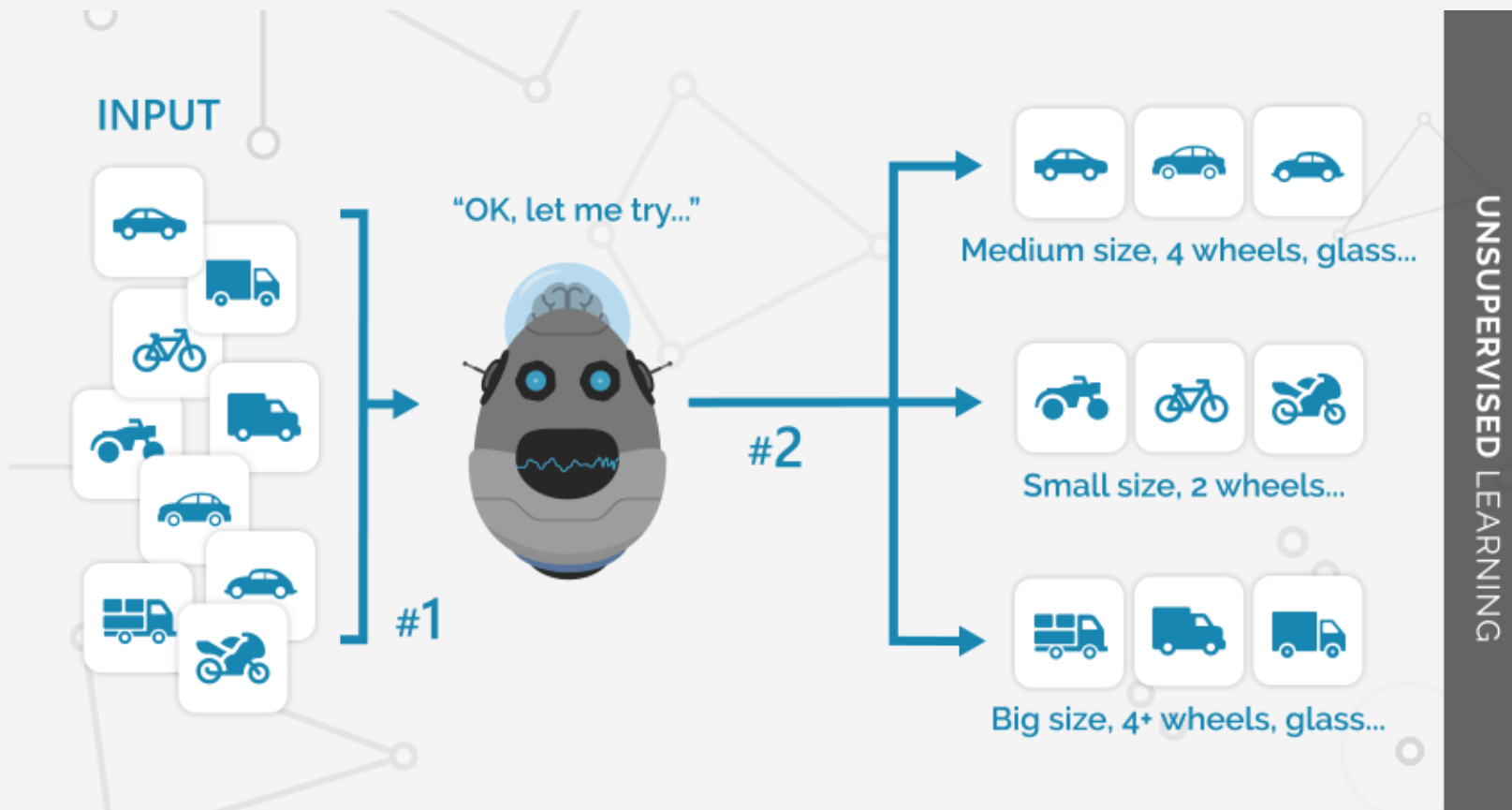
MACHINE LEARNING



APRENDIZAJE SUPERVISADO



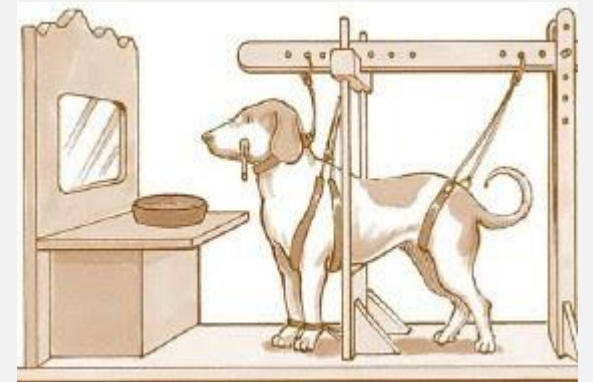
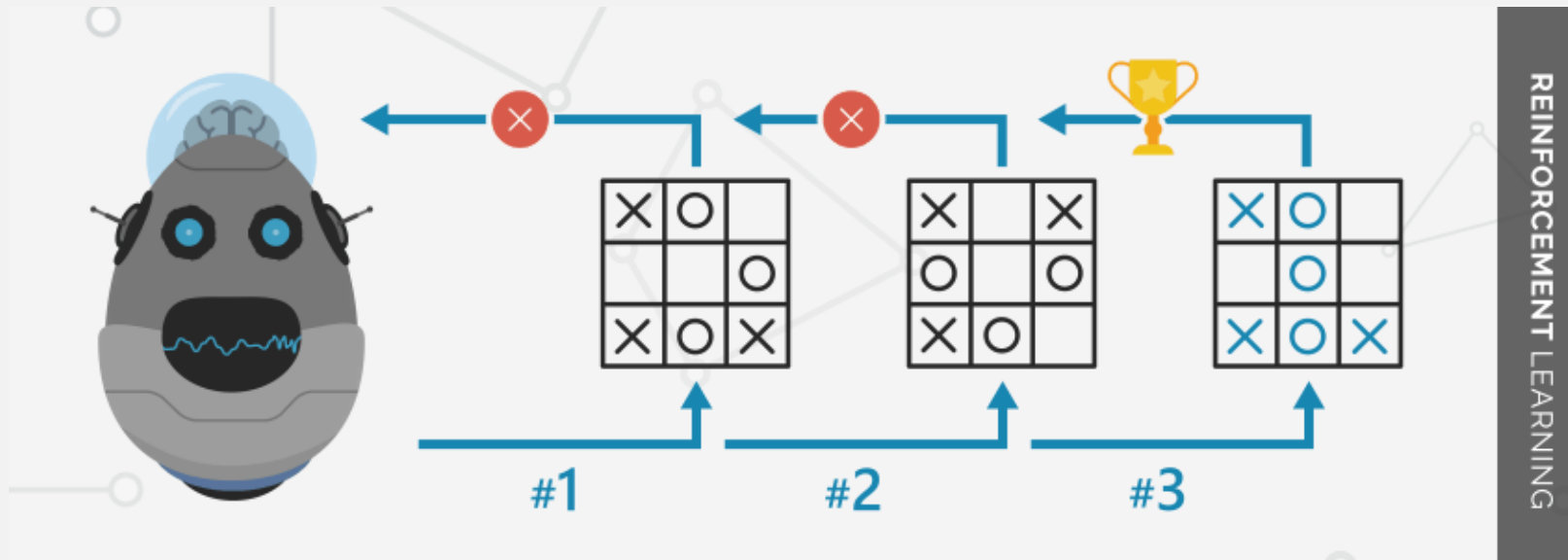
APRENDIZAJE NO SUPERVISADO



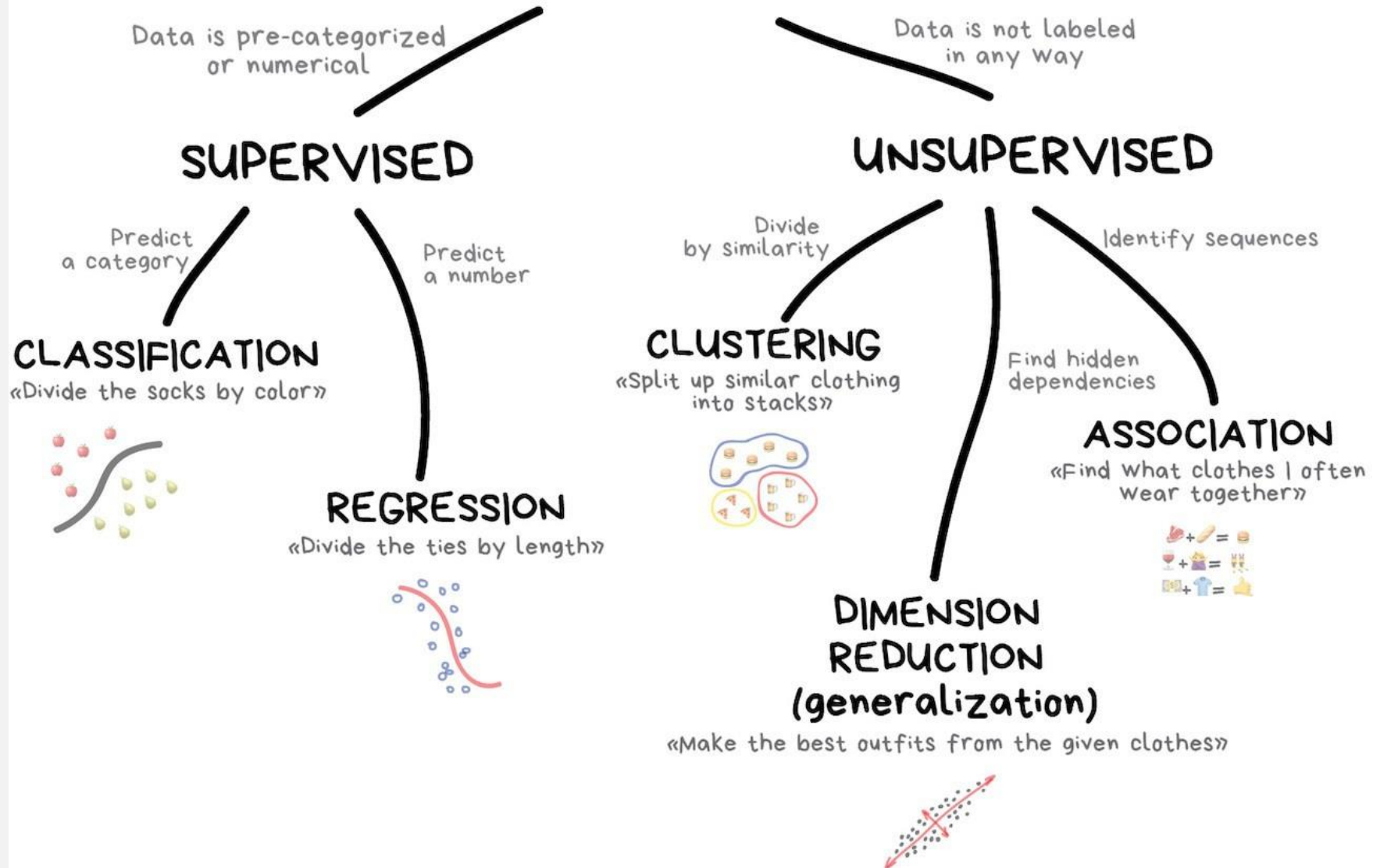
MÁQUINAS HECHAS A NUESTRA IMAGEN Y SEMEJANZA...



APRENDIZAJE POR REFUERZO



CLASSICAL MACHINE LEARNING



CLASSICAL MACHINE LEARNING

Data is pre-categorized
or numerical

SUPERVISED

Predict
a category

CLASSIFICATION

«Divide the socks by color»



Predict
a number

REGRESSION

«Divide the ties by length»



Data is not labeled
in any way

UNSUPERVISED

Divide
by similarity

CLUSTERING

«Split up similar clothing
into stacks»



Identify sequences

Find hidden
dependencies

ASSOCIATION

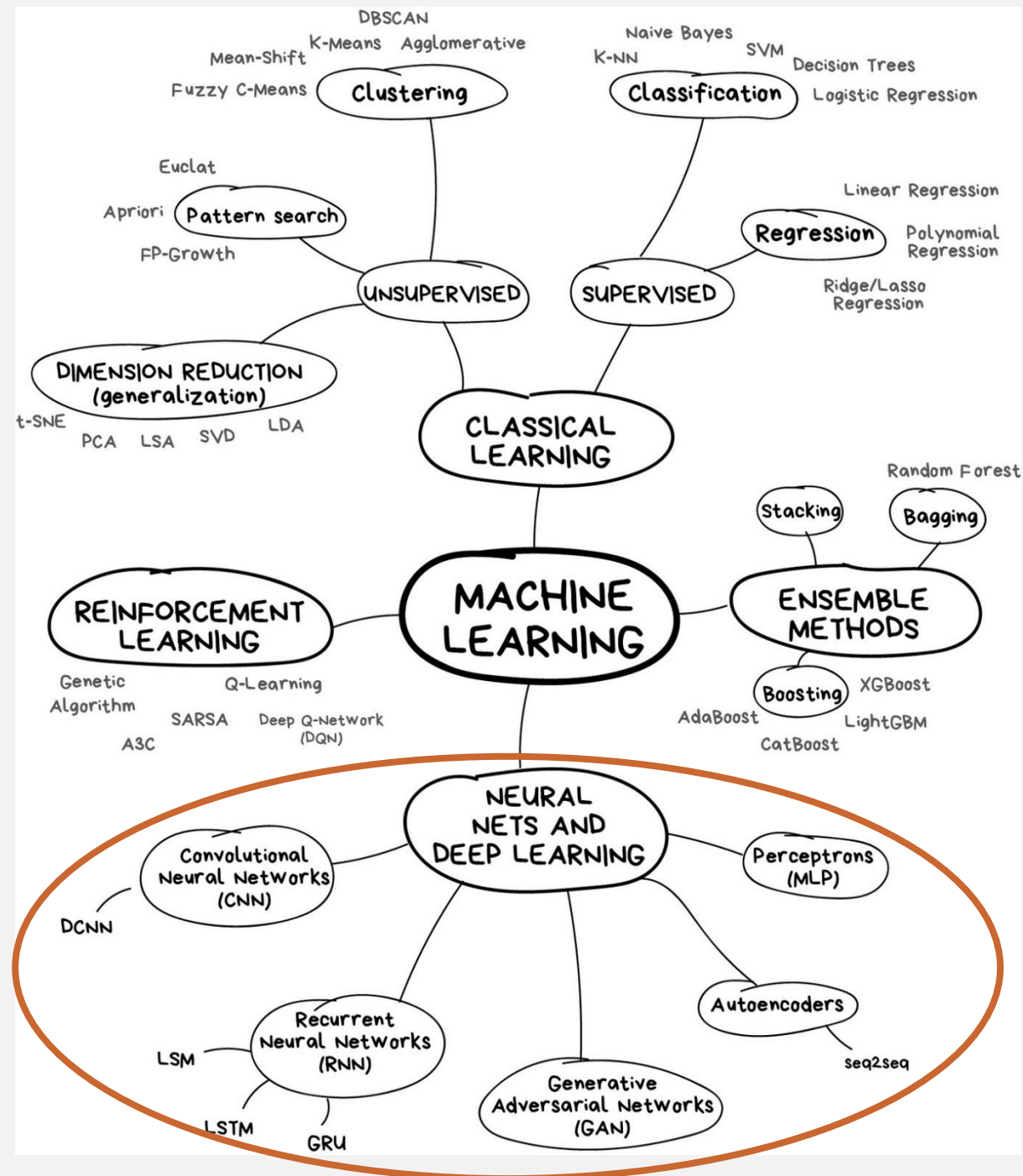
«Find what clothes I often
wear together»

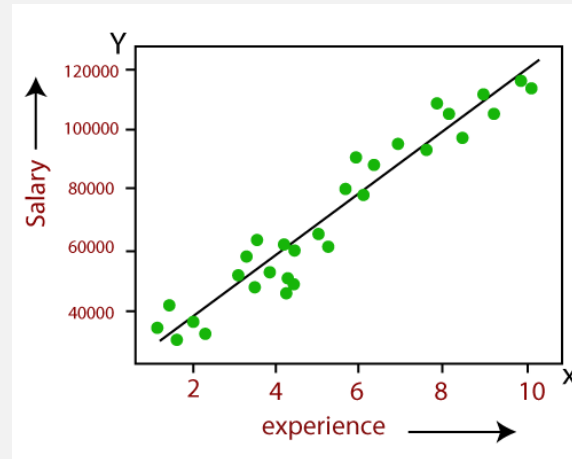
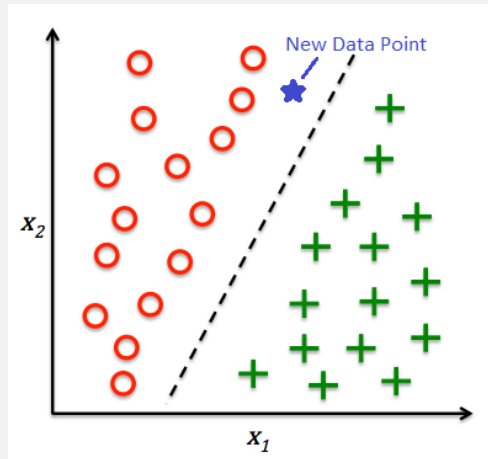


DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»



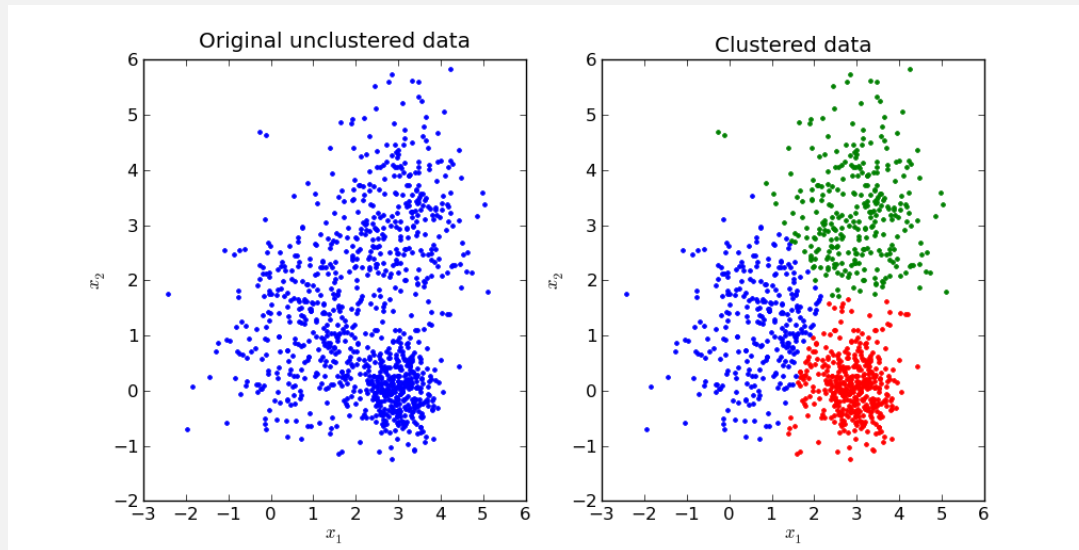




Aprendizaje supervisado

Objetivo principal: Categorizar nuevos datos en base a características conocidas (i.e. clasificar).

Algoritmos habituales: Regresión, Árbol de Decisión, Random Forest, SVM, AdaBoost, Redes Neuronales, Sentiment Analysis



Aprendizaje no supervisado

Objetivo principal: Categorizar datos en base a características endógenas subyacentes (i.e. clusterizar).

Algoritmos habituales: K-Means, Componentes Principales, Modularidad, Topic Modeling

PYTHON Y LAS CIENCIAS SOCIALES

COVID-19

Alarma en el Reino Unido por un fallo técnico en el recuento de contagios

- Unos 16.000 casos de coronavirus no fueron notificados a tiempo para rastrear los contactos

<https://www.lavanguardia.com/internacional/2020/1005/483861404090/alarma-reino-unido-recuento-fallo-tecnico-contagios.html>

SOFTWARE INADECUADO



MALOS DATOS



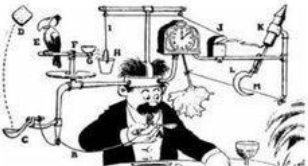
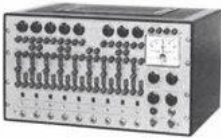
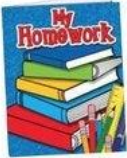
MALAS DECISIONES

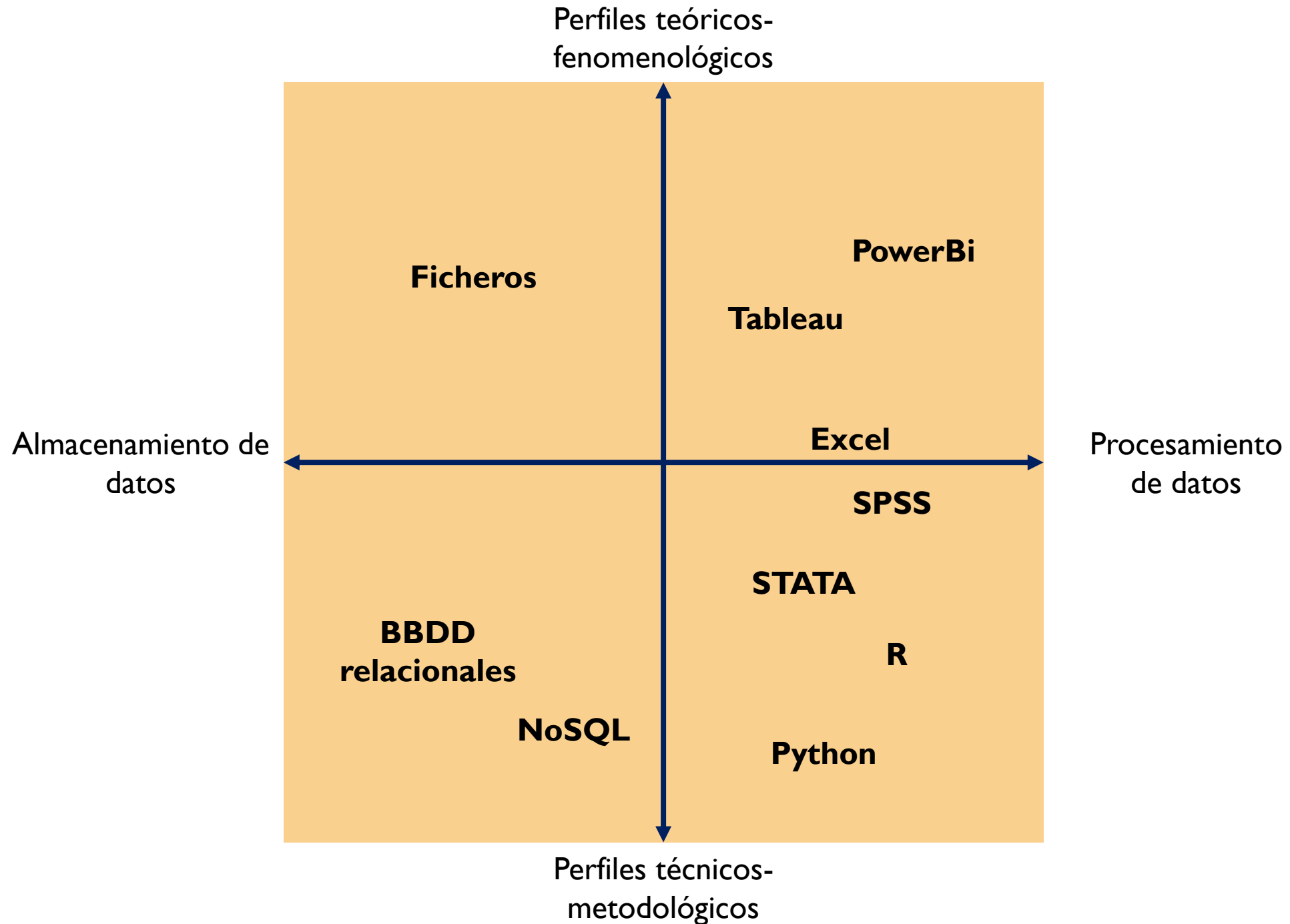
Un **fallo técnico** que implicó que unos **16.000 casos de coronavirus** en el **Reino Unido** no fueran notificados a tiempo ha retrasado los esfuerzos del Gobierno británico para rastrear los contactos de esas personas que dieron positivo, informan este lunes los medios nacionales.

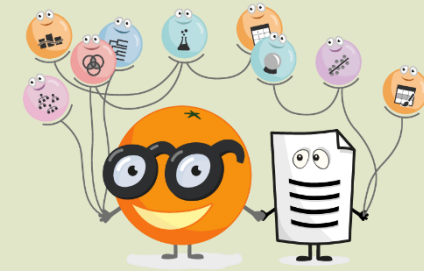
El error, según apuntan varios medios británicos, surgió desde un laboratorio que enviaba el reporte diario de los resultados de las PCR que realizaba en sus instalaciones en formato CSV (Coma Separated Values). Dicho formato es compatible con Excel, el programa que usa el PHE para indexar todos los casos. El archivo compartido con el departamento oficial incluía todo el histórico, no sólo los nuevos.

Día tras día, los responsables cargaban los nuevos datos al final del excel principal. Pero mientras que los archivos CSV pueden tener cualquier tamaño, los archivos de Microsoft Excel solo pueden tener una longitud de 1.048.576 filas. Cuando se abre un archivo CSV más largo que lo soportado en Excel, las filas inferiores se cortan y ya no se muestran, aunque el programa advierte al usuario que hay información que no cabe en el máximo establecido por Microsoft y, por ende, no será cargada en el documento.

AS SEEN BY USERS OF ...

	STATA	R	sas	python	SPSS
STATA					
R					
sas					
python					
SPSS					





LAS HERRAMIENTAS DEL ANALISTA DE DATOS

Data analyst jobs

3340 data analyst job ads were included in this analysis.

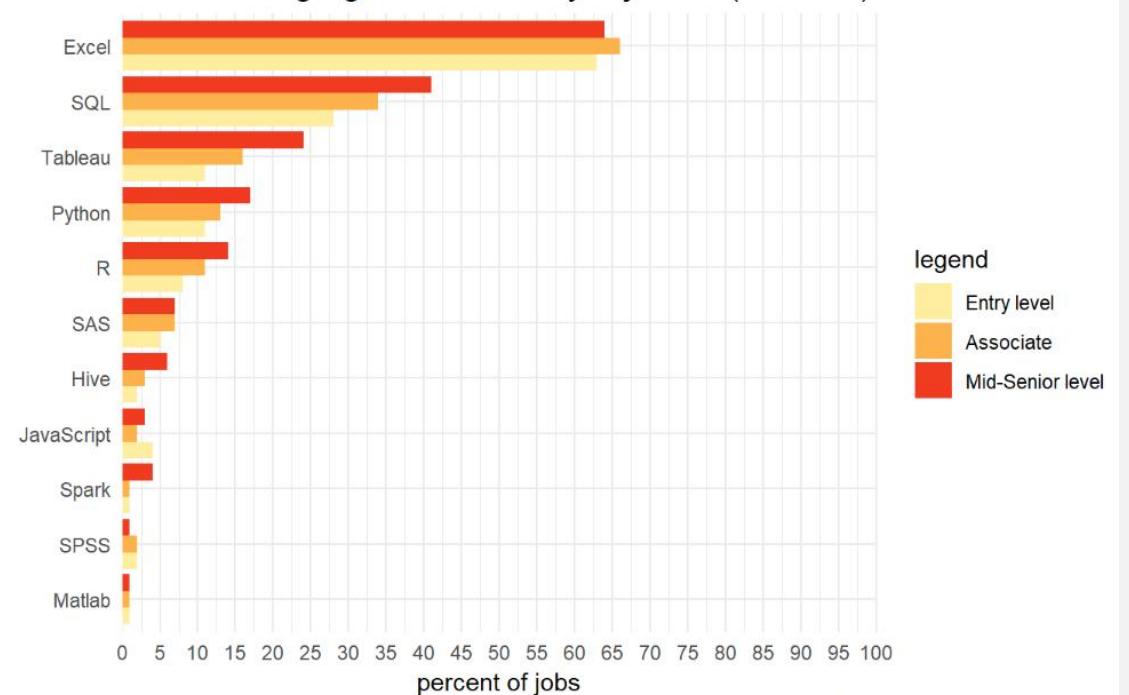
Tools / languages

Other skills

Degrees

Disciplines

Tools and languages in "Data Analyst" job ads (N = 3340)



Jobs were scraped from LinkedIn in Sept 2019
Locations included NYC, SF, Seattle, Boston, and Toronto
Positions were Entry, Associate, and Senior levels

FUNCIONAMIENTO DE LA HERRAMIENTA

LIMPIEZA Y TRATAMIENTO DE LOS DATOS



Data Table



Pivot Table



Concatenate



Merge Data



Select Rows



Select Columns



Preprocess



Feature Constructor



Purge Domain



Correlations



Feature Statistics

MINANDO DATOS DE ENCUESTA



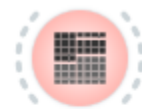
Feature Statistics



File



Scatter Plot



Sieve Diagram



Correlations



Box Plot



Mosaic Display

MINANDO TEXTOS



Import Documents



Twitter



Pubmed



Corpus



Preprocess Text



Bag of Words



Word Cloud



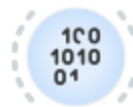
NY Times



The Guardian



Wikipedia



Similarity Hashing



Topic Modelling



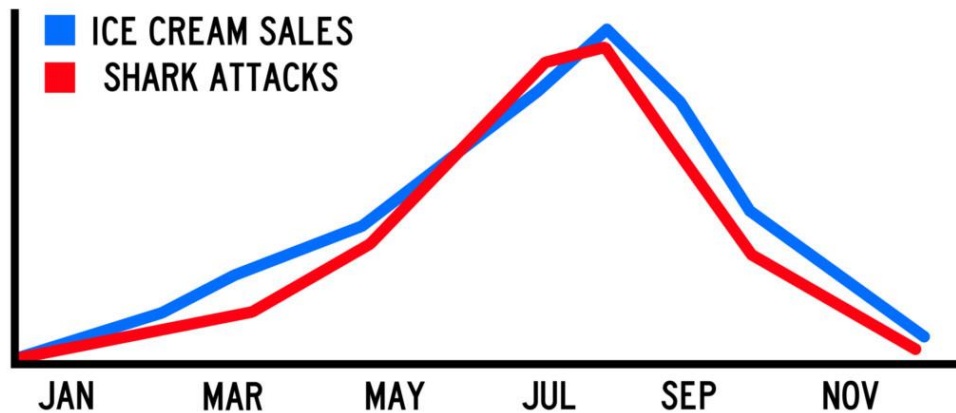
Sentiment Analysis

ALGORITMOS DE CLASIFICACIÓN Y DE PREDICCIÓN NUMÉRICA

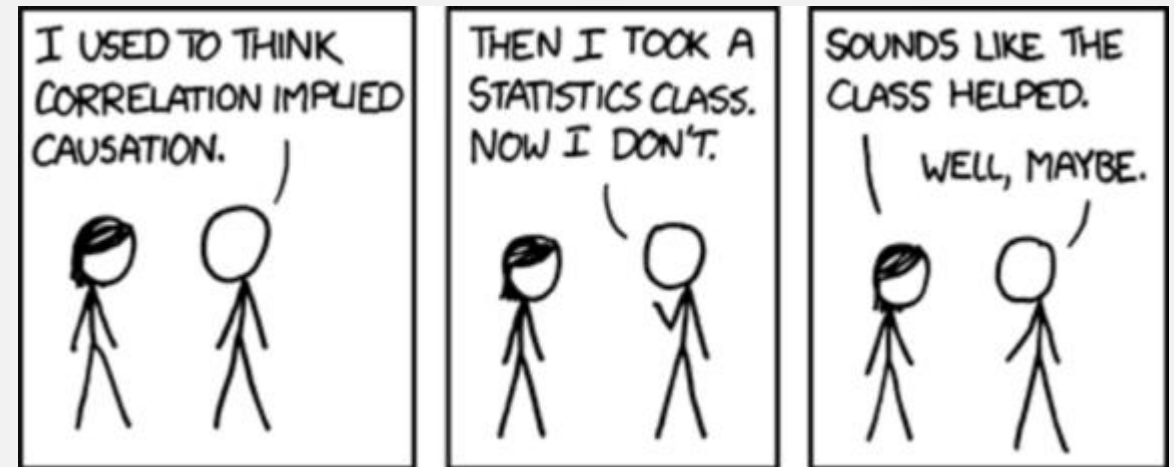
APRENDIZAJE SUPERVISADO

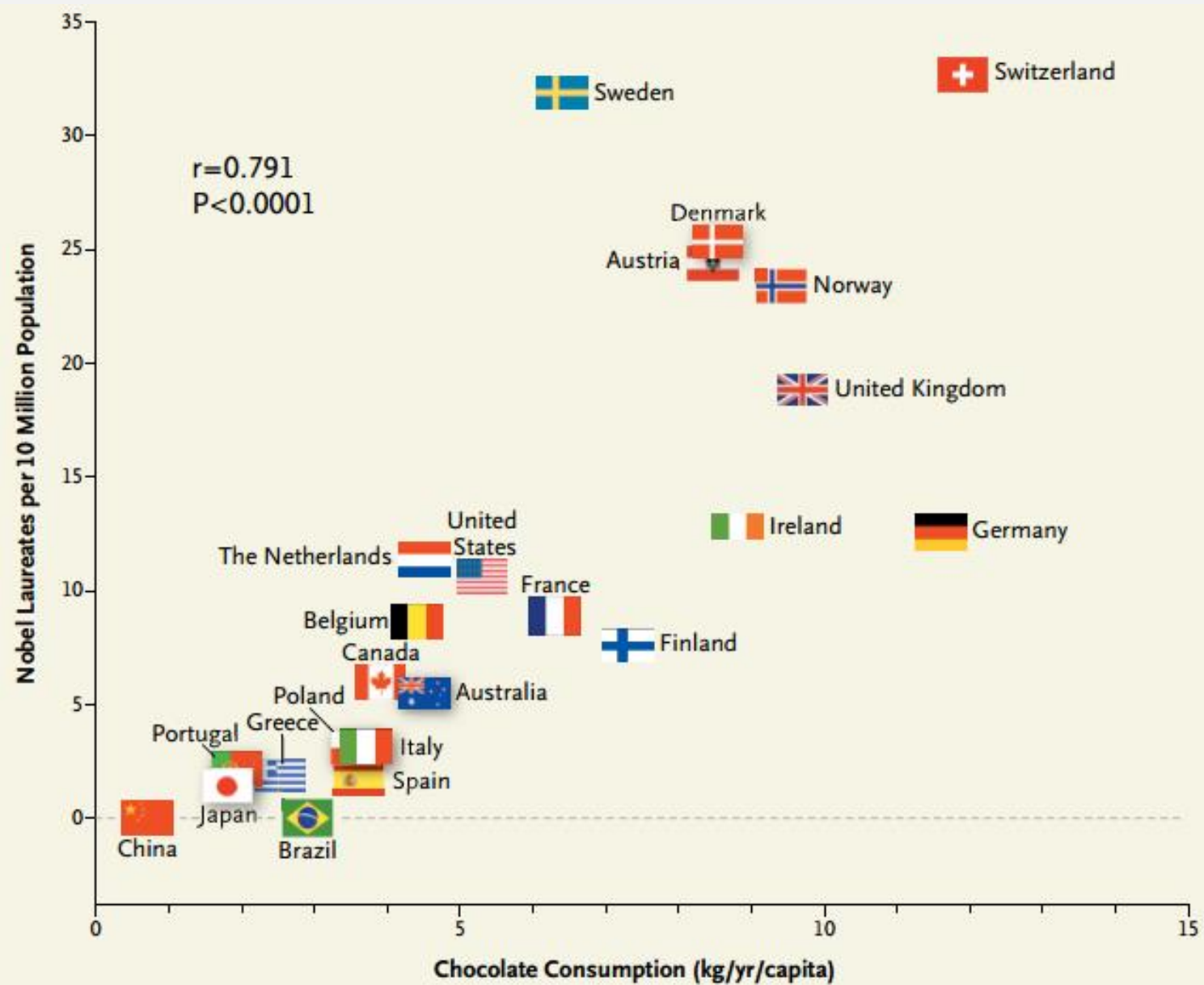
CORRELACIÓN Y CAUSALIDAD

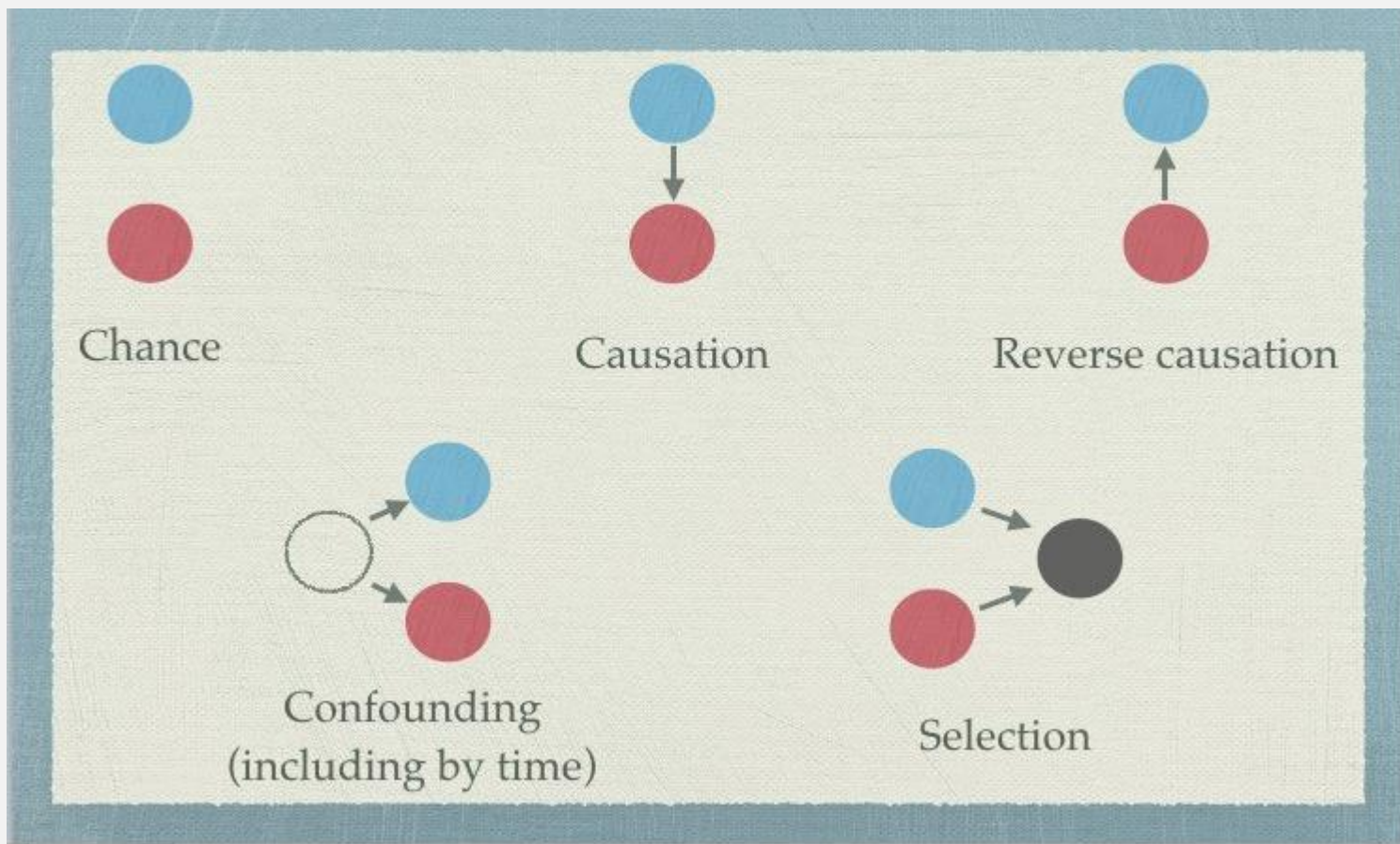
CORRELATION IS NOT CAUSATION!



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)





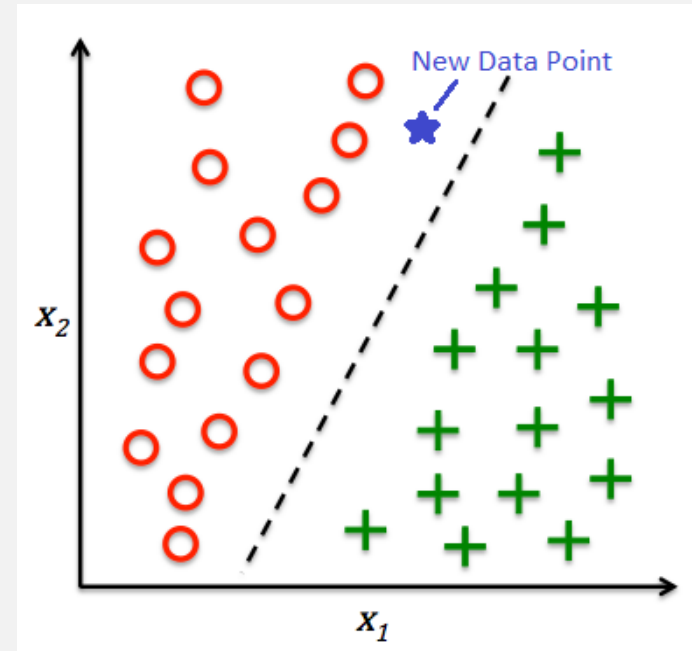


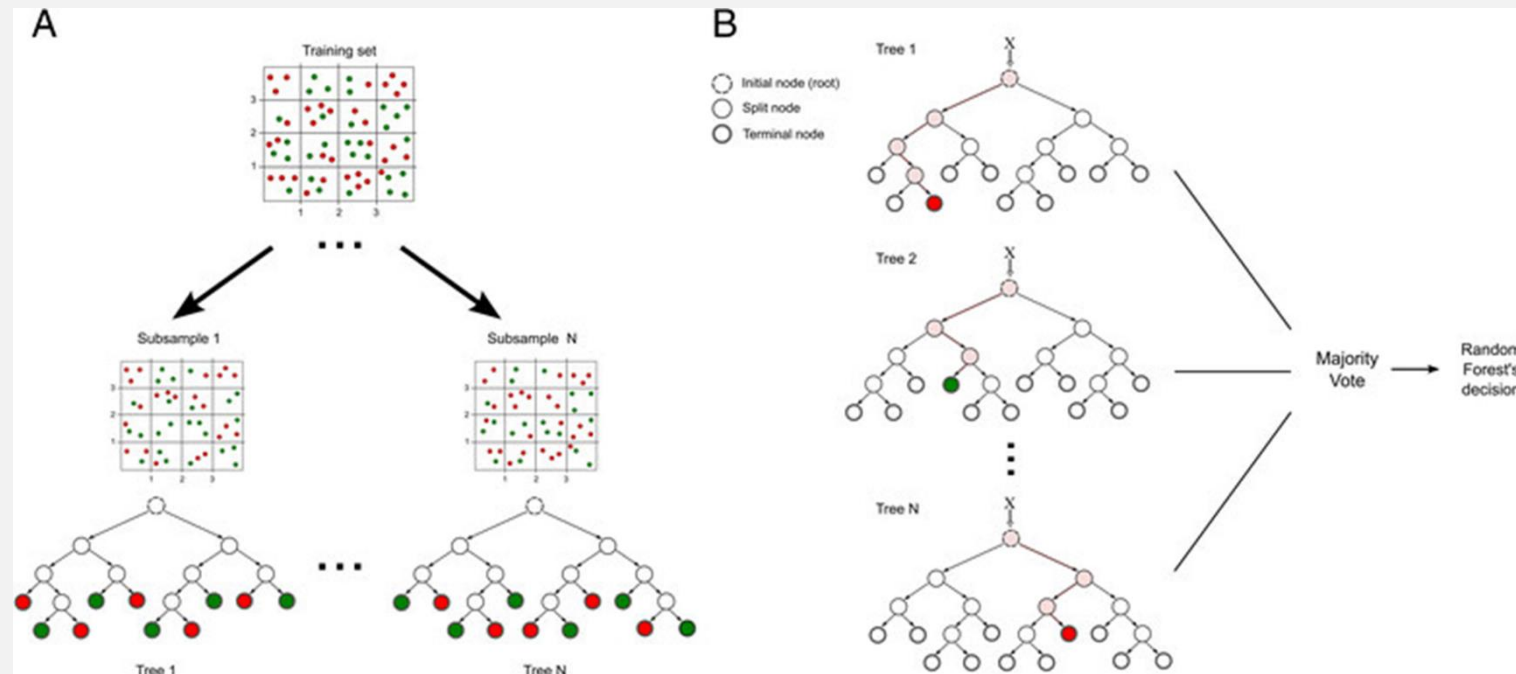
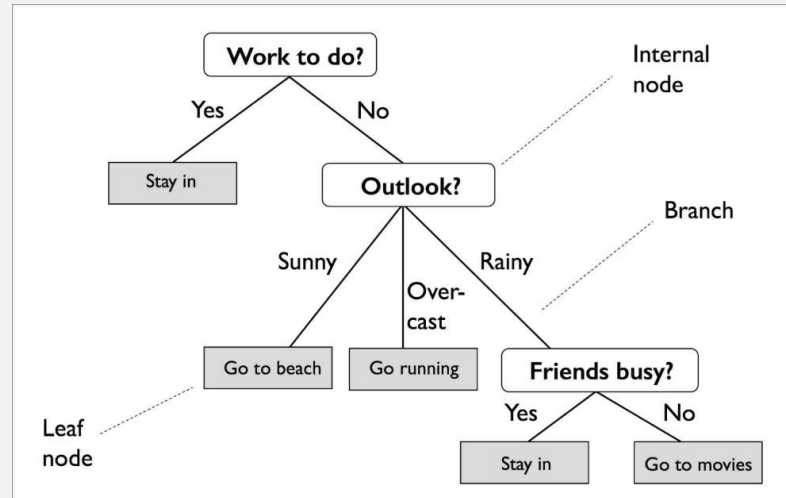
LA CAUSALIDAD
ES **UN TIPO DE**
CORRELACIÓN

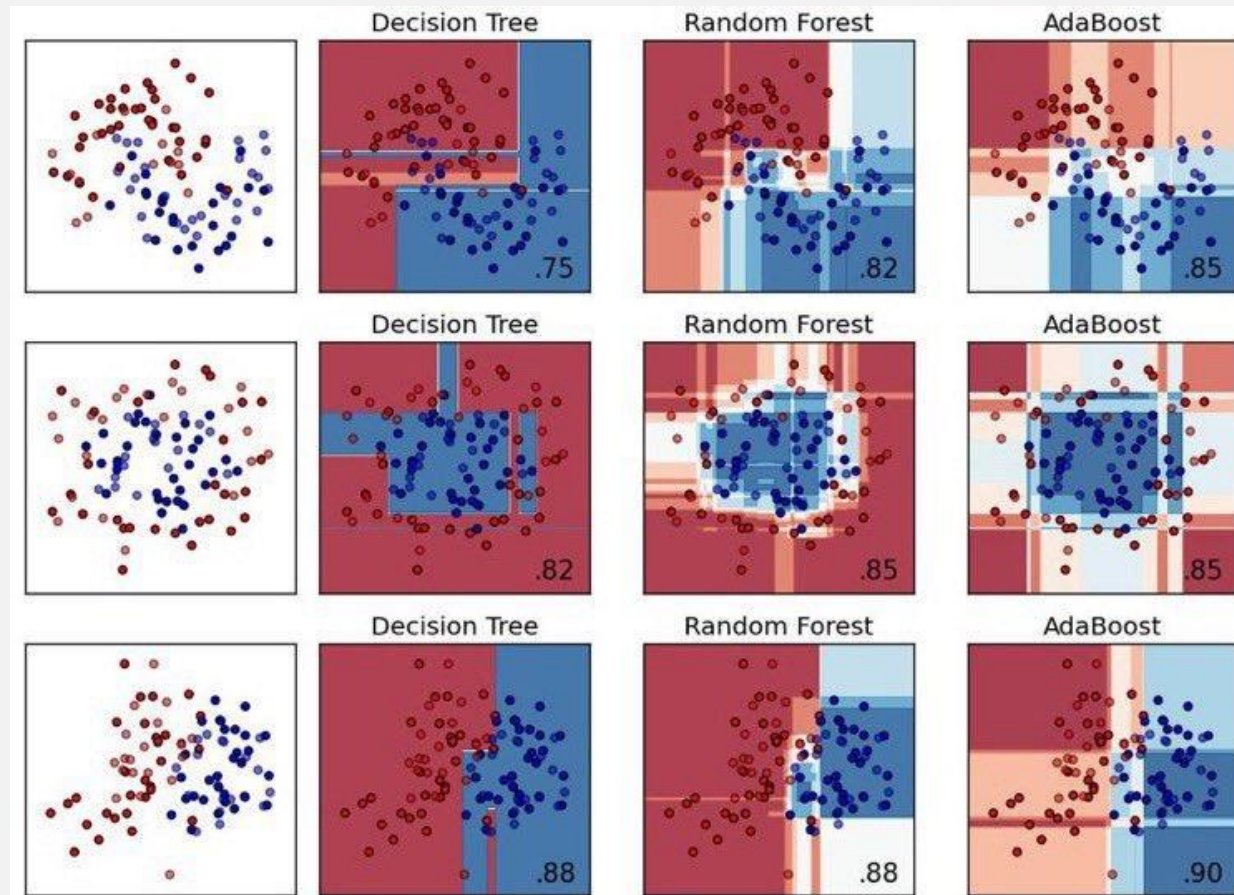
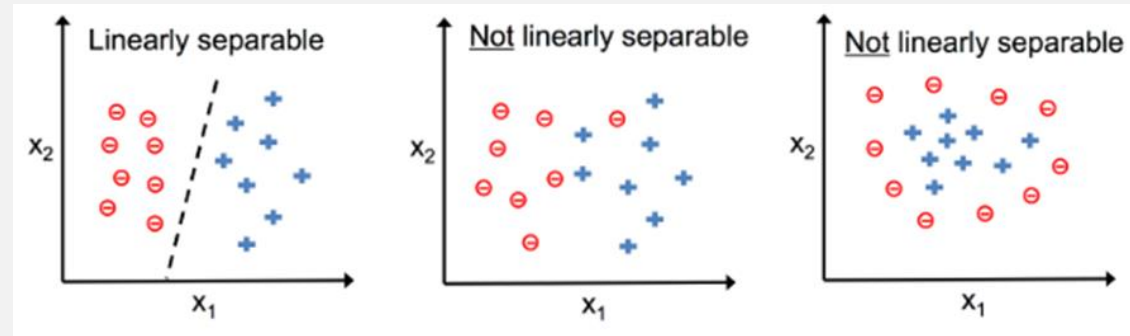
CLASIFICACIÓN (SUPERVISADA)

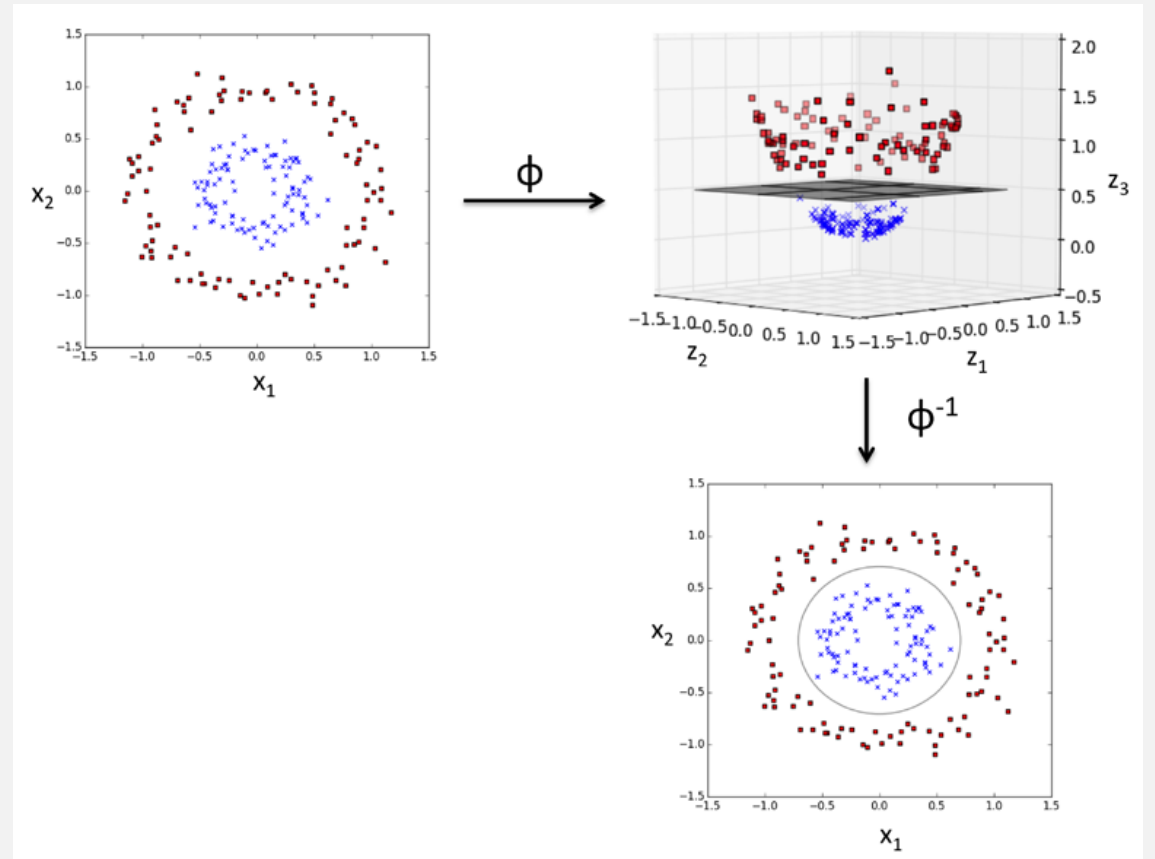
- Imputación de grupos en base a características conocidas o aprendidas
 - Filtros de SPAM en el correo
 - Concesión o no de una hipoteca
 - Detección de lenguas en textos
 - Análisis de sentimiento
 - Identificación de discursos de odio

El target de un algoritmo clasificador va a ser normalmente*** una variable discreta o categórica





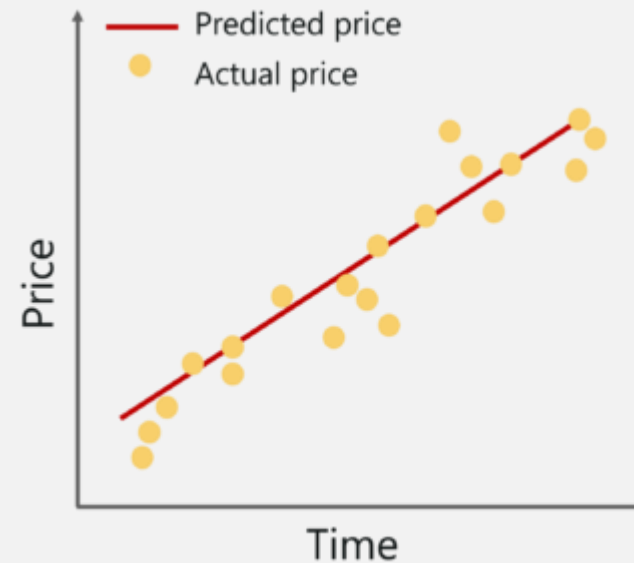




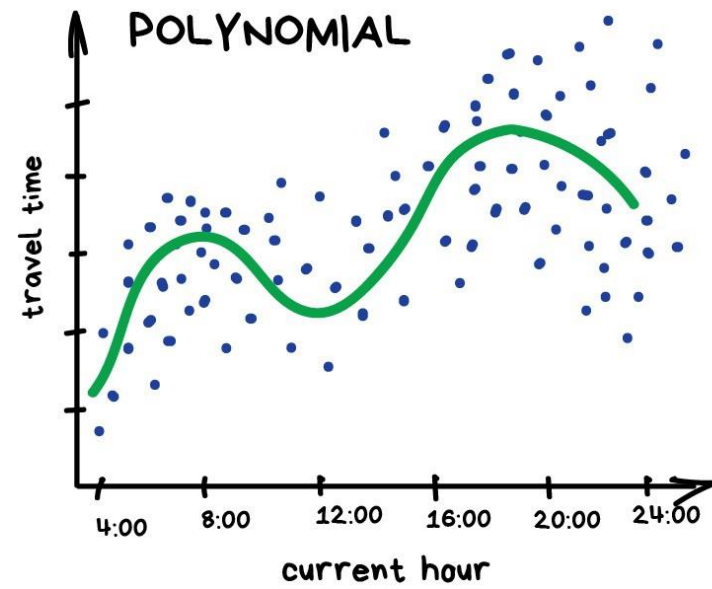
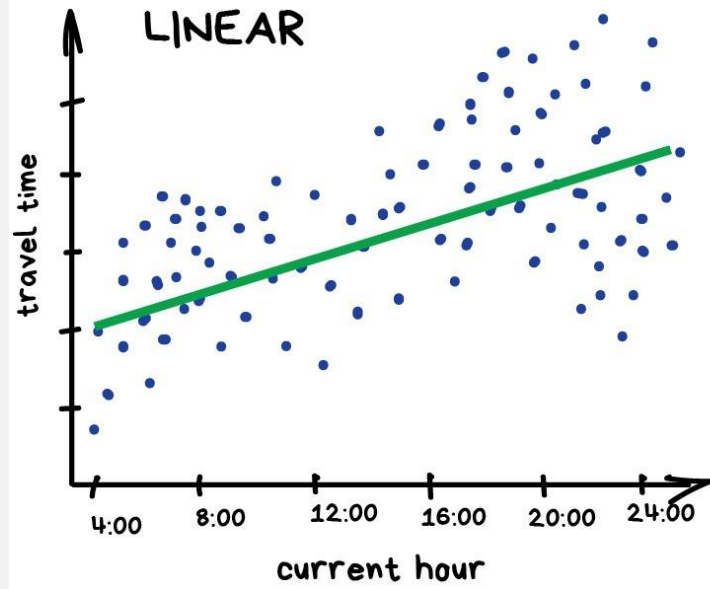
REGRESIÓN (SUPERVISADA)

- Predicción de valores numéricos en base a patrones conocidos
 - Predicción bursátil
 - Análisis de tendencias de consumo
 - Prospectiva comercial o financiera
 - Diagnóstico médico

El target de un algoritmo de regresión va a ser siempre una variable numérica y continua

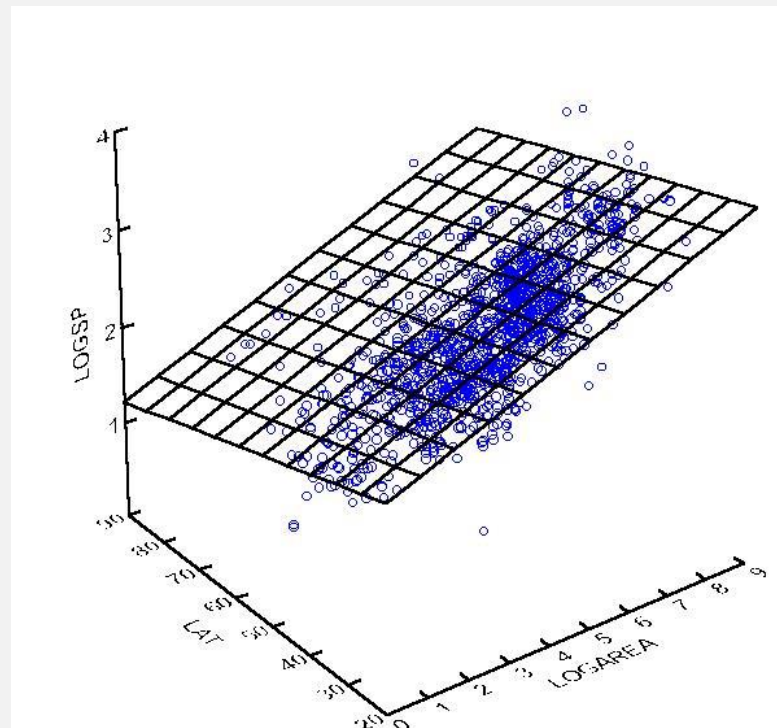


PREDICT TRAFFIC JAMS

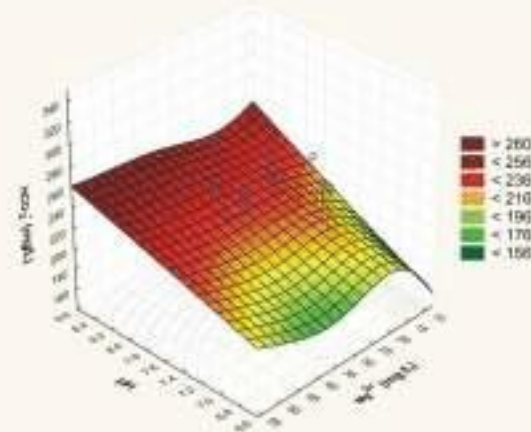


REGRESSION

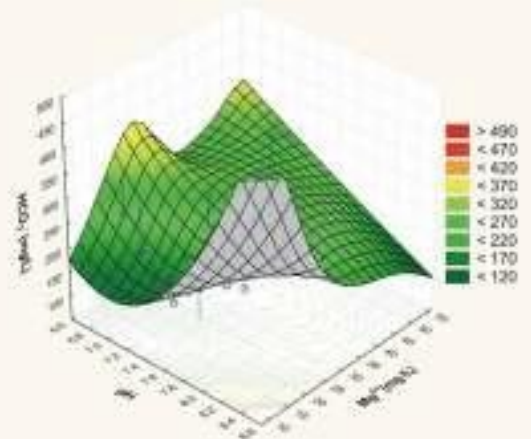




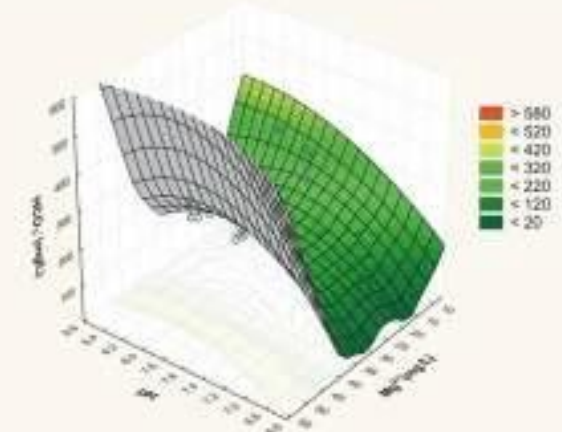
Surface plot - Distance Weighted Least Squares Method
 HCO_3^- concentration against pH and Mg^{2+} concentration - Winter



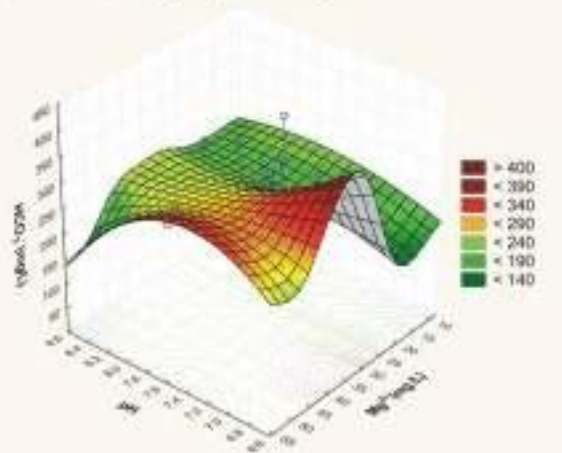
Surface plot - Distance Weighted Least Squares Method
 HCO_3^- concentration against Mg^{2+} concentration and pH - Summer



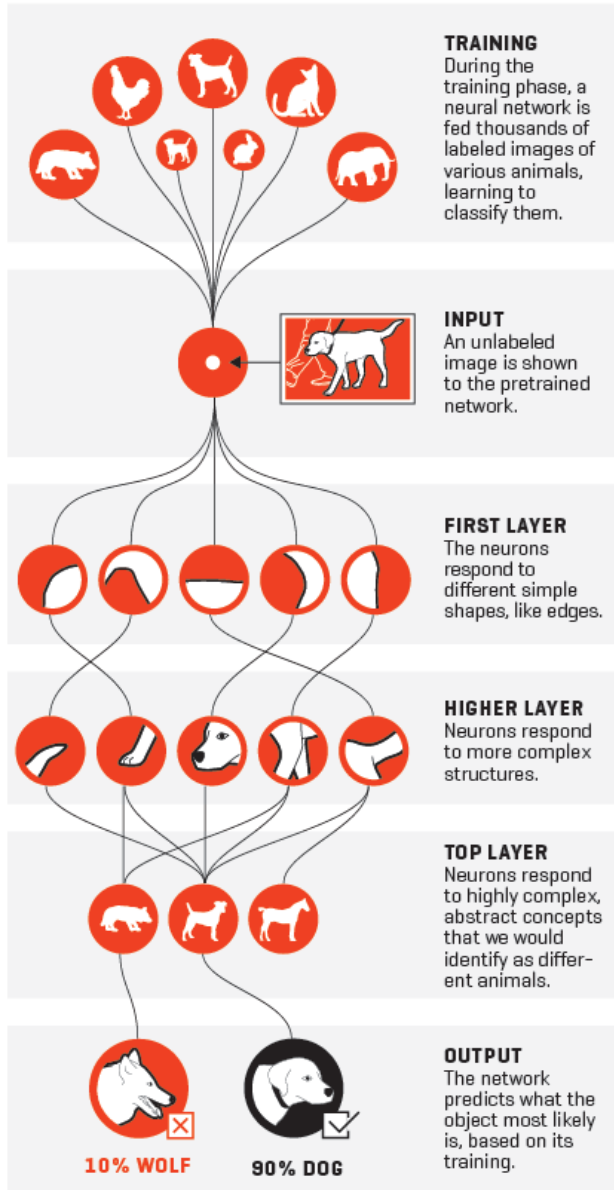
Surface plot - Distance Weighted Least Squares Method
 HCO_3^- concentration against pH and Mg^{2+} concentration - Spring



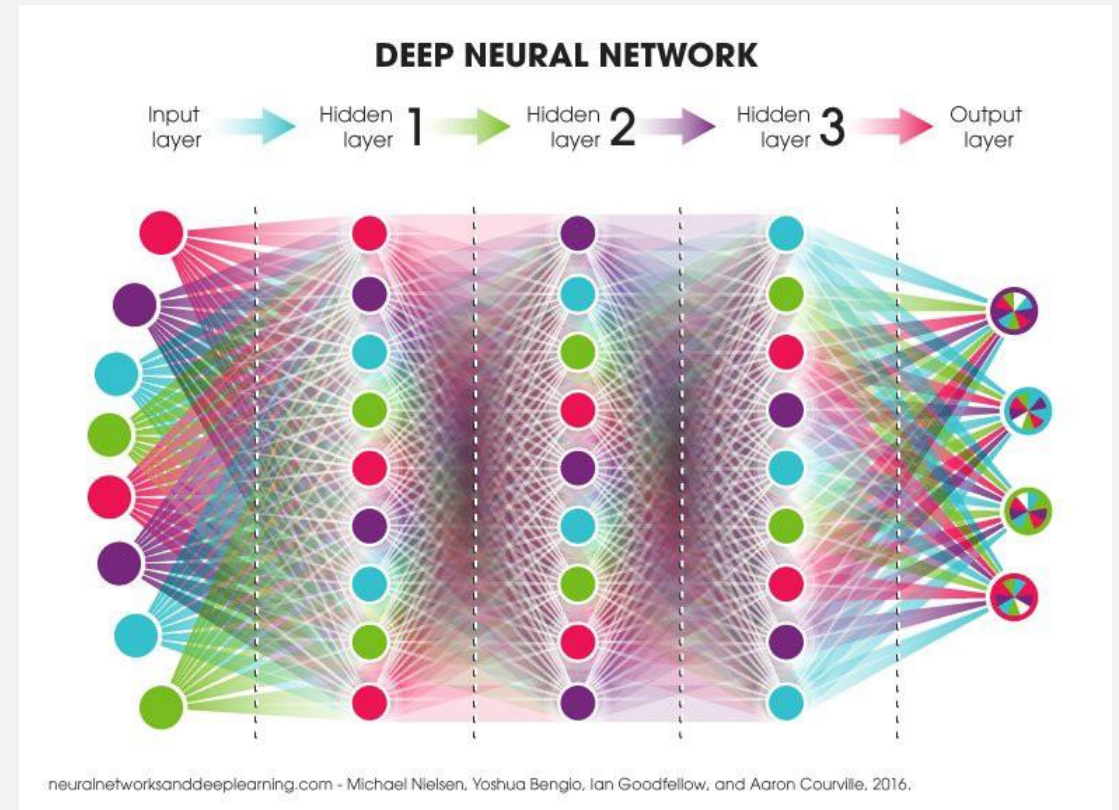
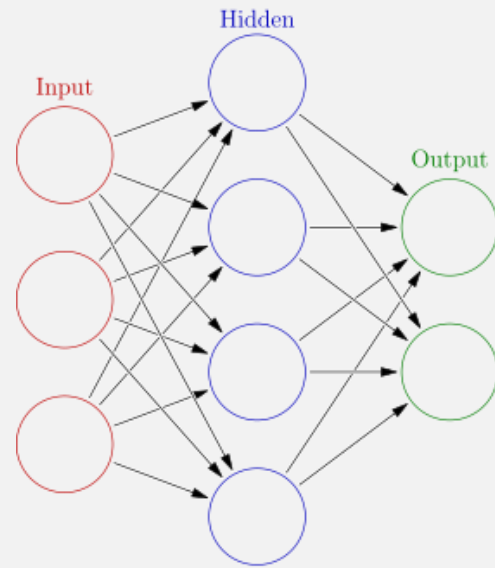
Surface plot - Distance Weighted Least Squares Method
 HCO_3^- concentration against pH and Mg^{2+} concentration - Autumn



HOW NEURAL NETWORKS RECOGNIZE A DOG IN A PHOTO



REDES NEURONALES Y DEEP LEARNING



INFERENCIA Y PREDICCIÓN CON DATOS DE ENCUESTA



Model	MSE	RMSE	MAE	R2
Neural Network	0.245	0.495	0.344	0.984
Random Forest	2.683	1.638	1.133	0.820
Tree	1.710	1.308	0.782	0.885
Linear Regression	10.683	3.268	2.368	0.284

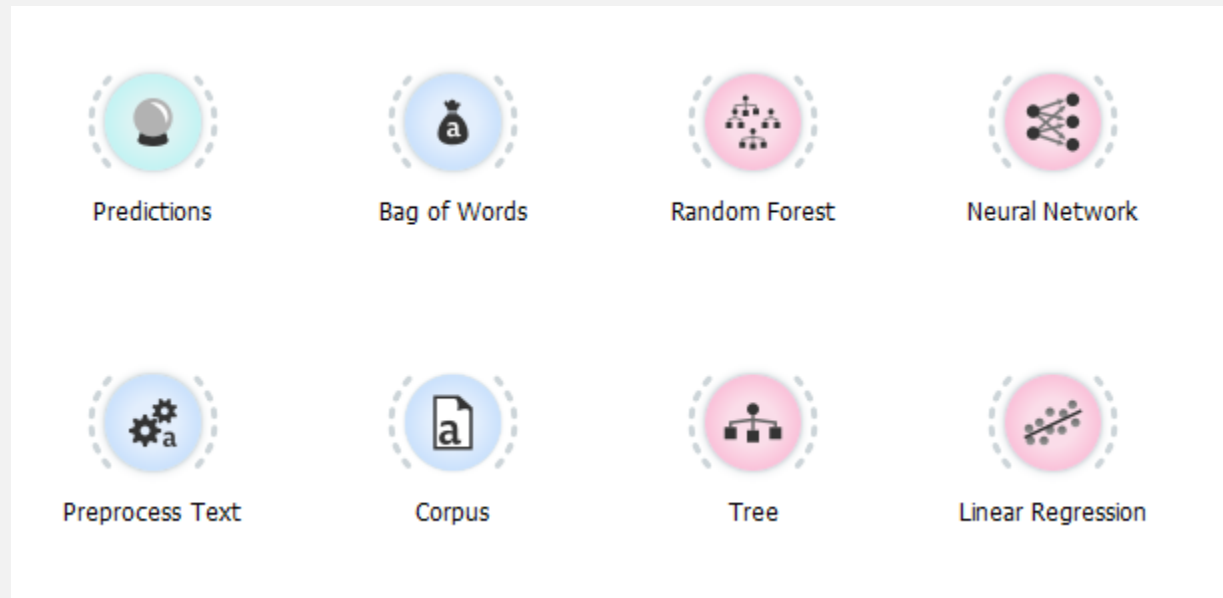
MSE: Error cuadrático medio
(cuanto menor, mejor)

RMSE: Raíz del error cuadrático
medio (cuanto menor, mejor)

MAE: Error absoluto medio
(cuanto menor, mejor)

R2: Coeficiente de determinación
(cuanto más cercano a 1, mejor)

ENTRENAMIENTO Y PREDICCIÓN DE TEXTOS



Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.944	0.937	0.918	0.932	0.937
Tree	0.783	0.944	0.931	0.941	0.944
Random Forest	0.998	0.962	0.956	0.962	0.962
Neural Network	1.000	0.999	0.999	0.999	0.999

AUC: Área bajo la curva (cuanto más cercano a 1, mejor)

CA: Exactitud (cuanto más cercano a 1, mejor)

F1: Valor F (cuanto más cercano a 1, mejor)

Precisión: Casos relevantes entre los recuperados (cuanto más cercano a 1, mejor)

Exhaustividad: Casos relevantes recuperados (cuanto más cercano a 1, mejor)

ANÁLISIS DE SENTIMIENTO



Corpus



Preprocess Text



Sentiment Analysis



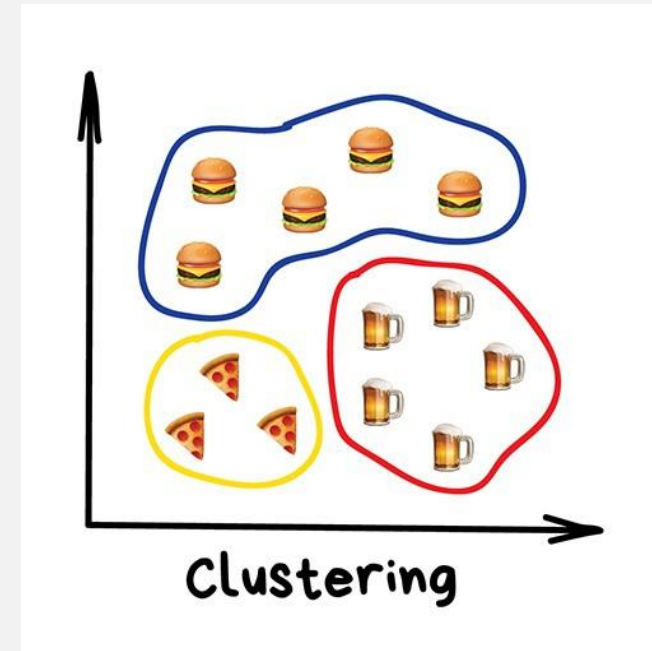
Tweet Profiler

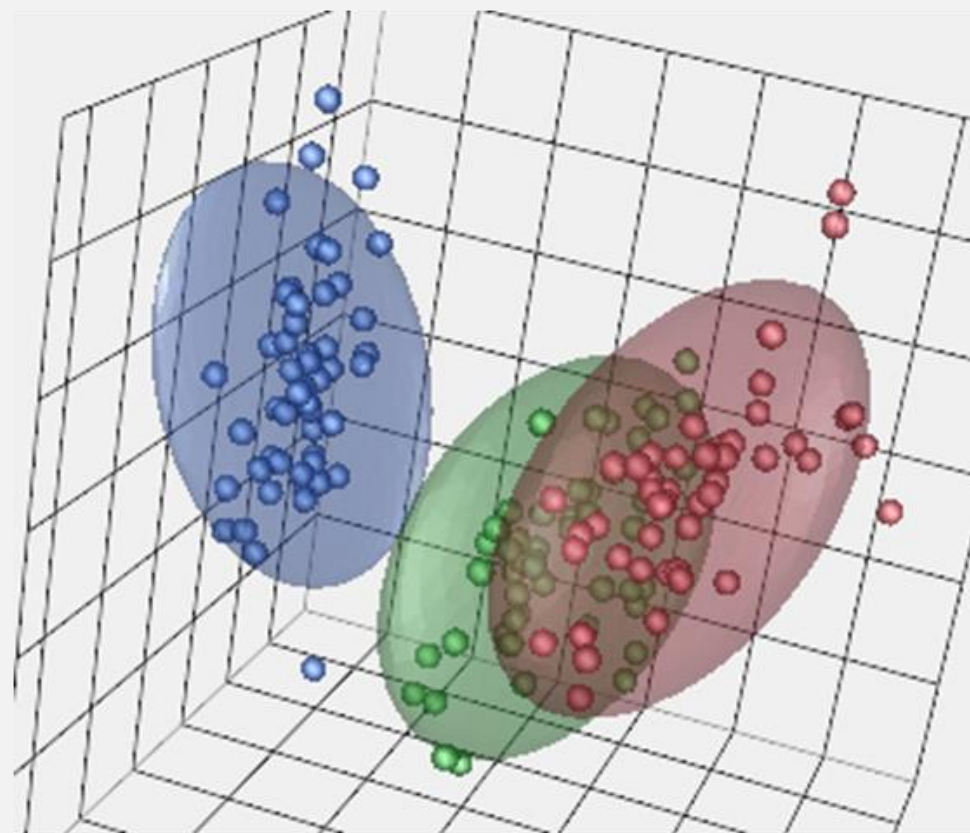
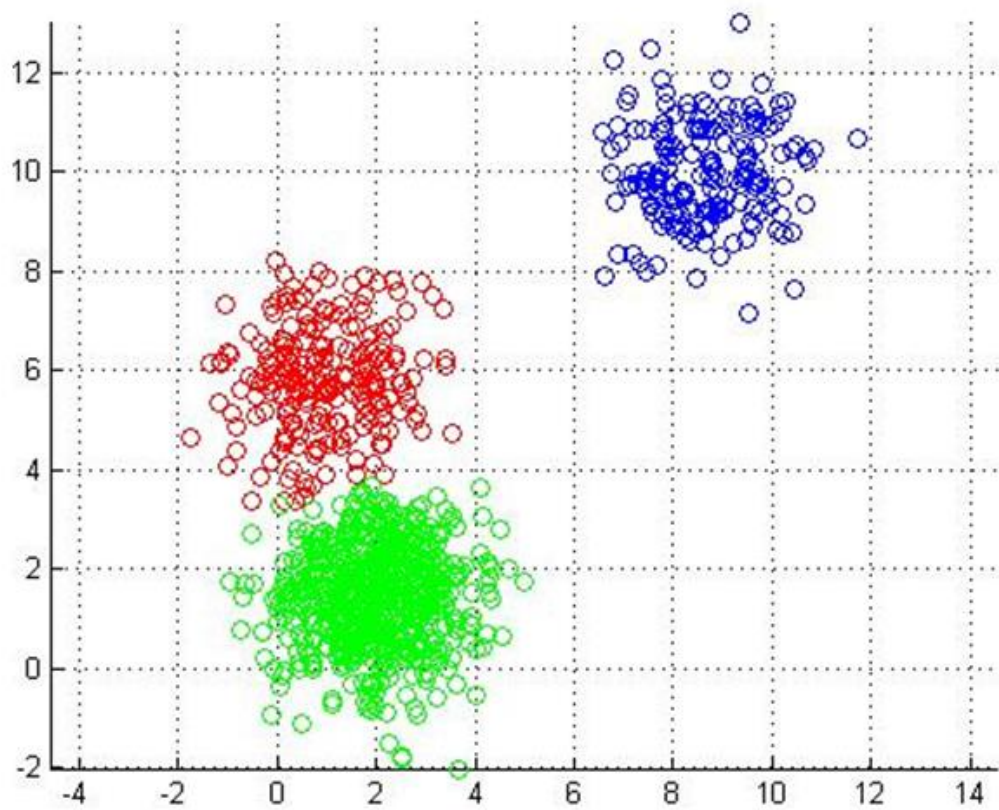
ALGORITMOS DE CLUSTERIZACIÓN

APRENDIZAJE NO SUPERVISADO

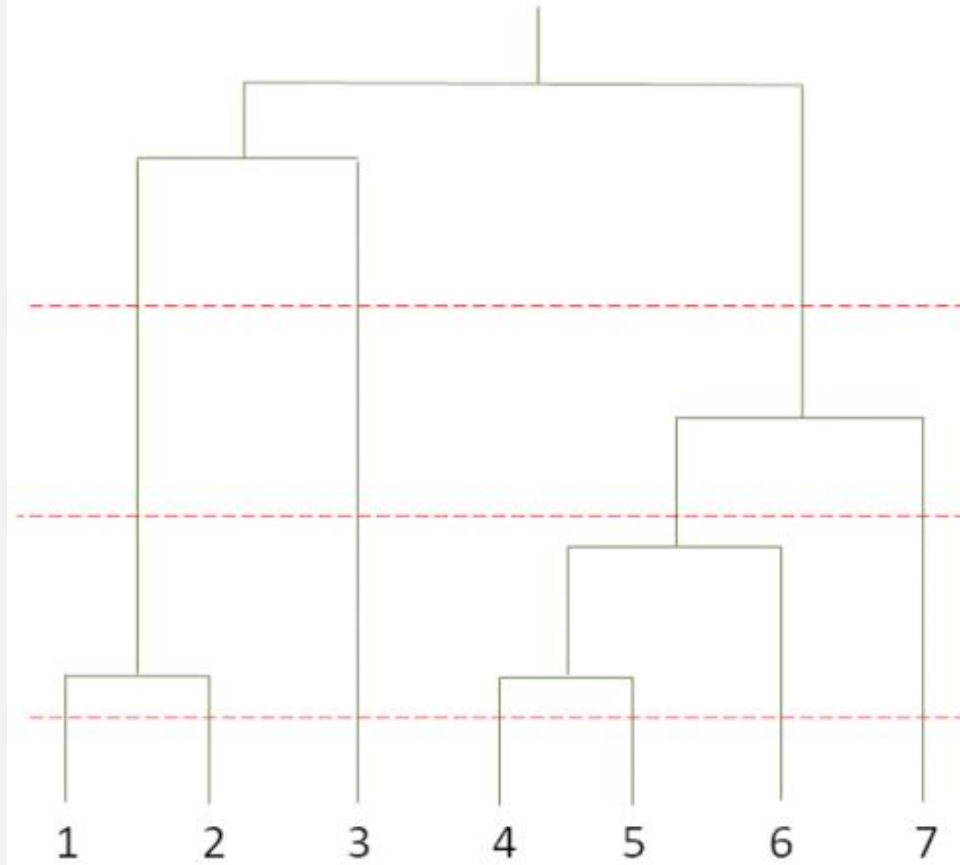
CLUSTERING (NO SUPERVISADO)

- Divide los casos sin basarse en características conocidas
 - Segmentación de mercados
 - Agrupación de textos (Topic Modeling)
 - Agrupación de imágenes (Image Embeddings)
 - Descubrimiento de patrones “ocultos” en los datos

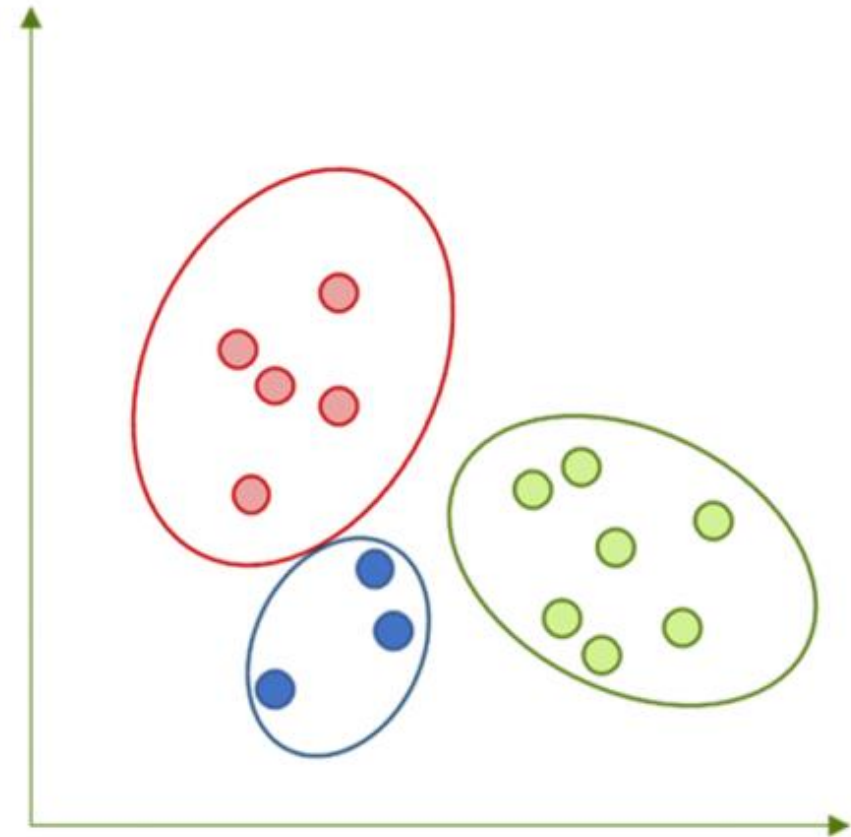




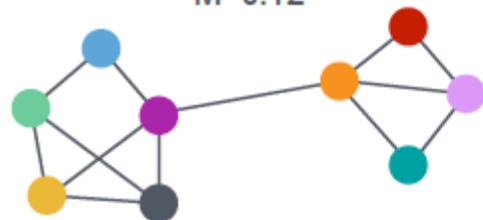
Hierarchical



Non-hierarchical



Negative Modularity
 $M=0.12$



Single Community
 $M=0$



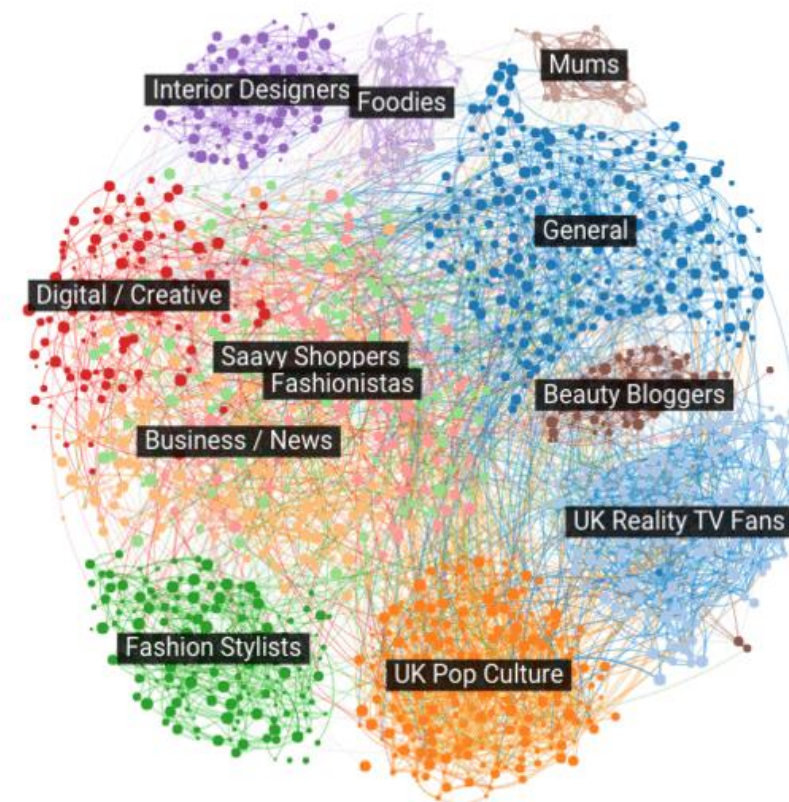
Suboptimal Partition
 $M=0.22$



Optimal Partition
 $M=0.41$



Modularity



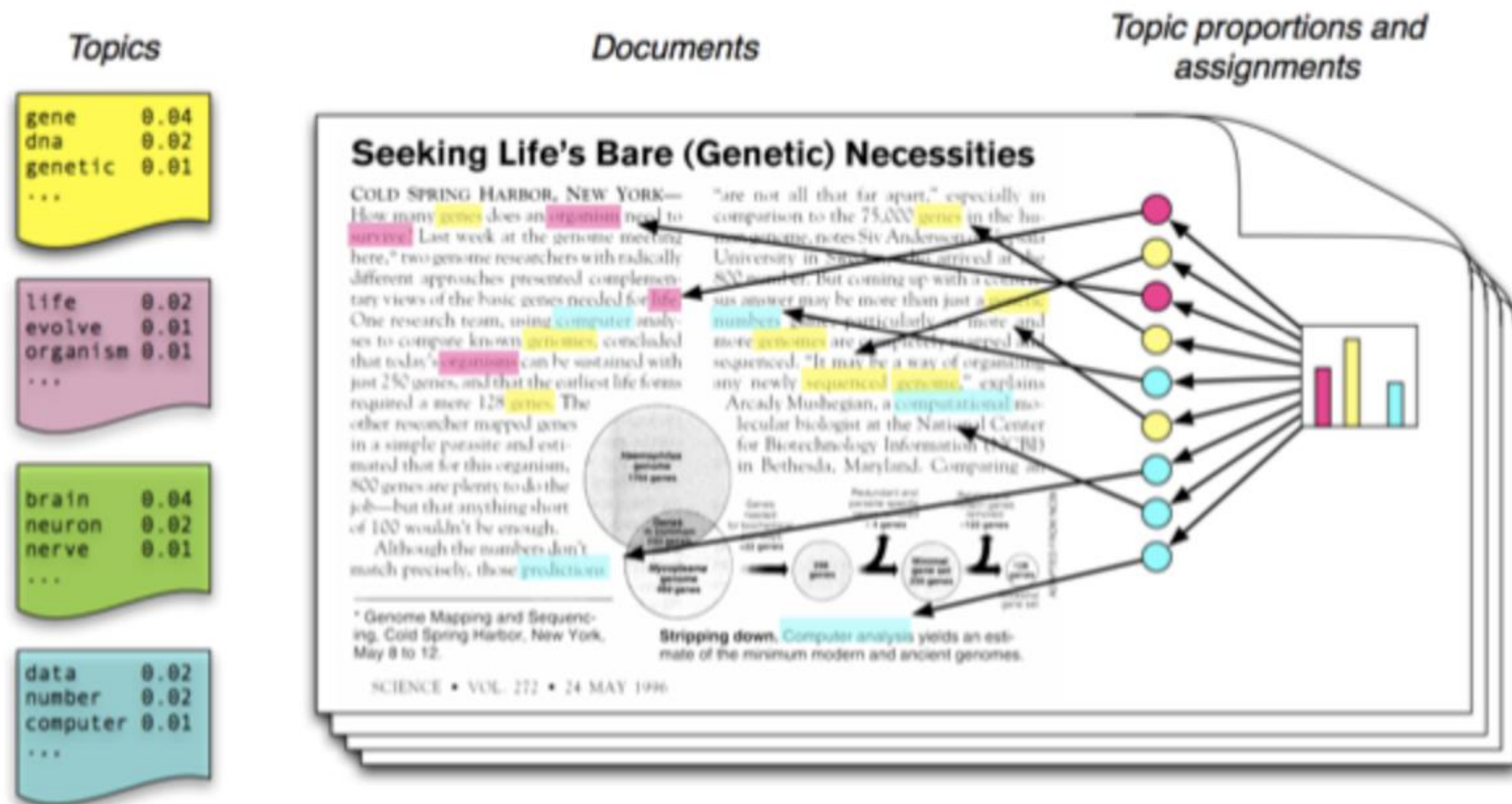
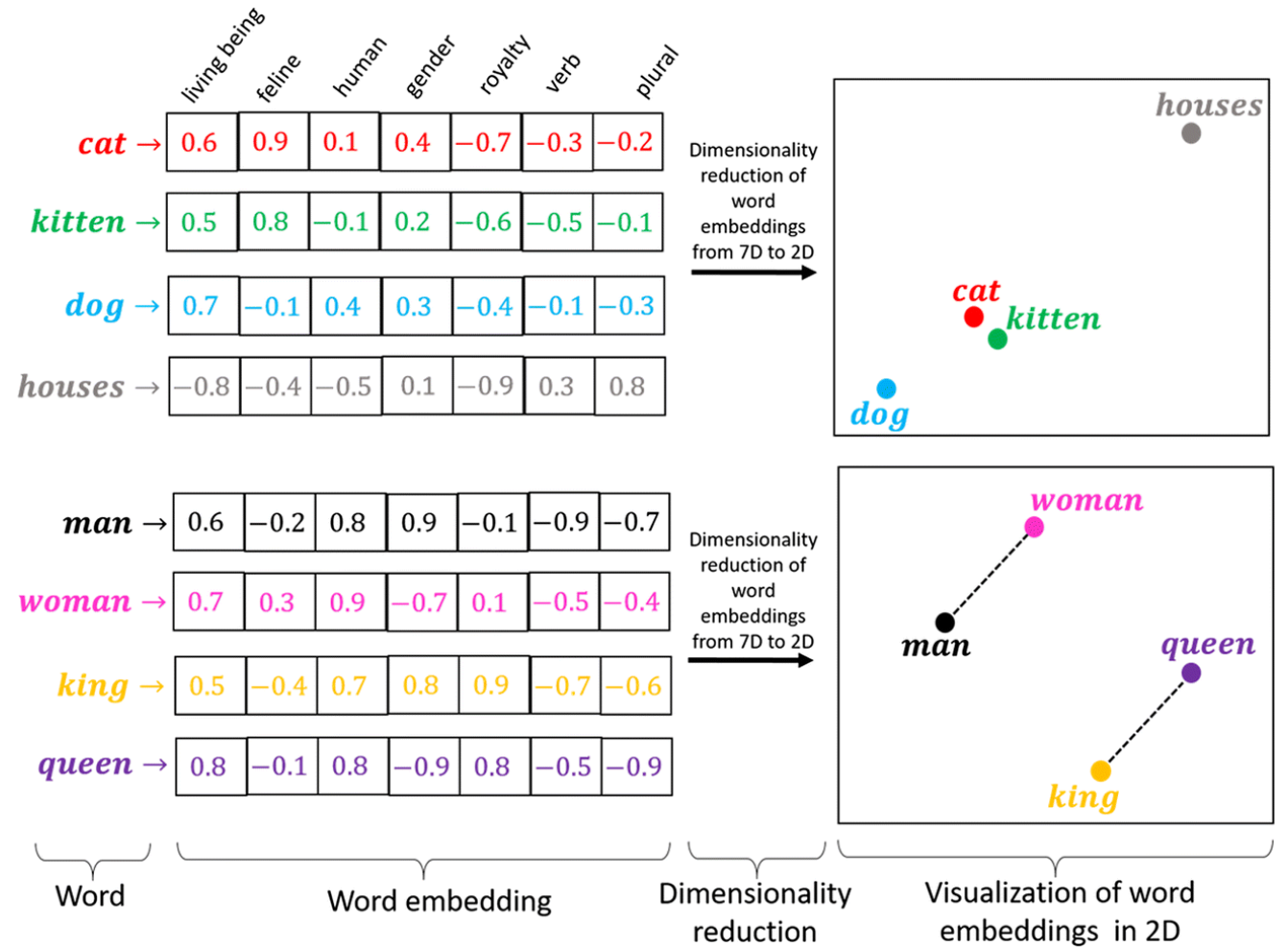
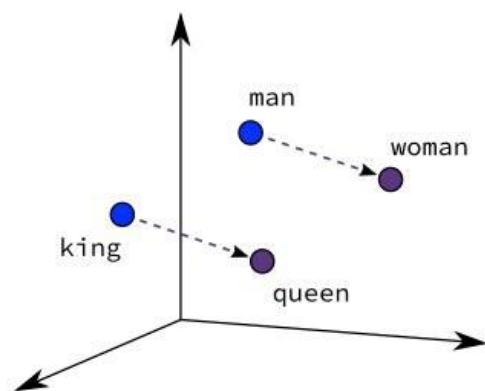


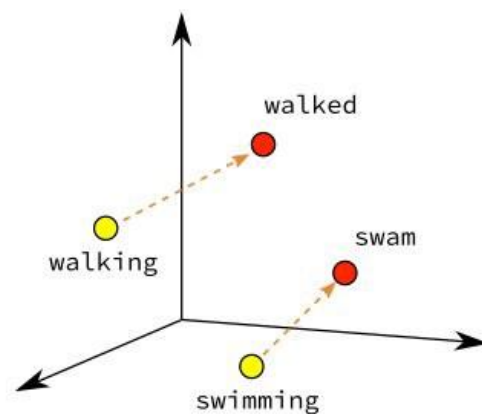
Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

INCRUSTACIONES (EMBEDDINGS)

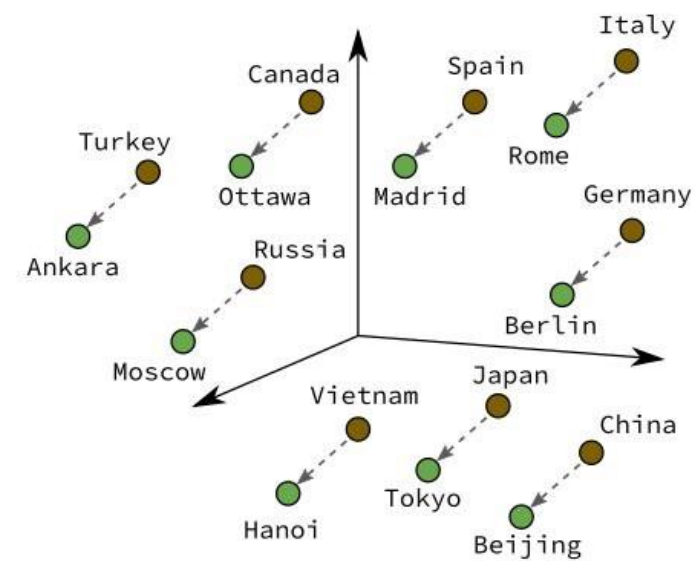




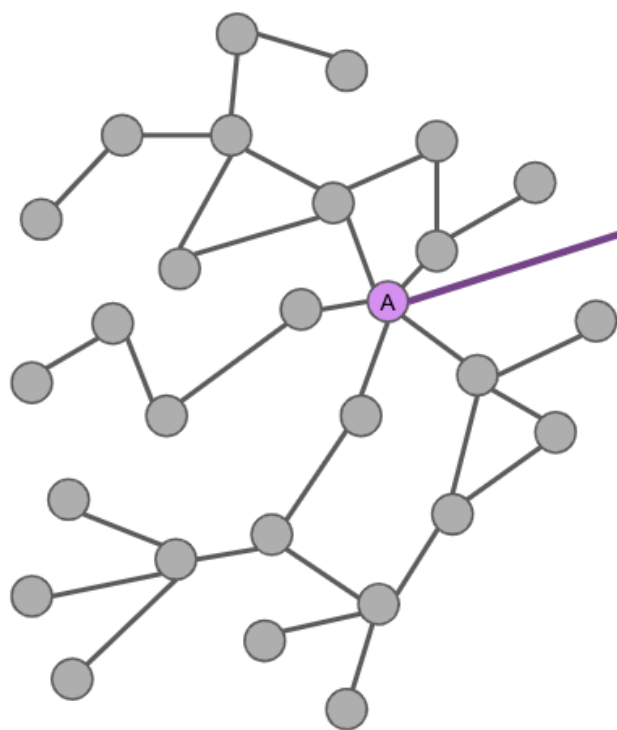
Male-Female



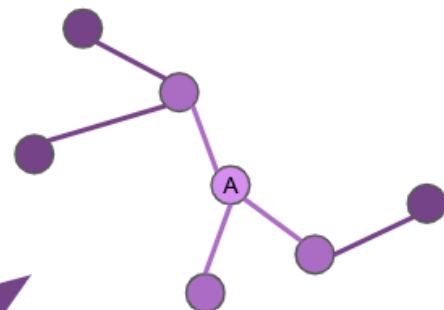
Verb Tense



Country-Capital



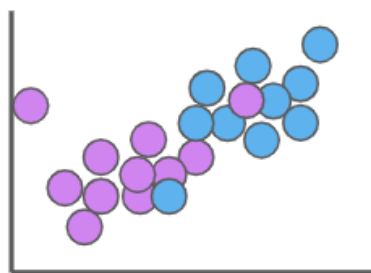
010...010
011..0011
..n
10011..00
1....010...
010011..0
011 ..n
10011..00
1....
010...010
011..0011
..n
10011..00
1....



Neighborhood
Prediction



Classification



Visualization

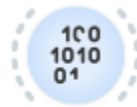
TOPIC MODELING



Corpus



Topic Modelling



Similarity Hashing

IMAGE EMBEDDINGS



Import Images



Image Embedding

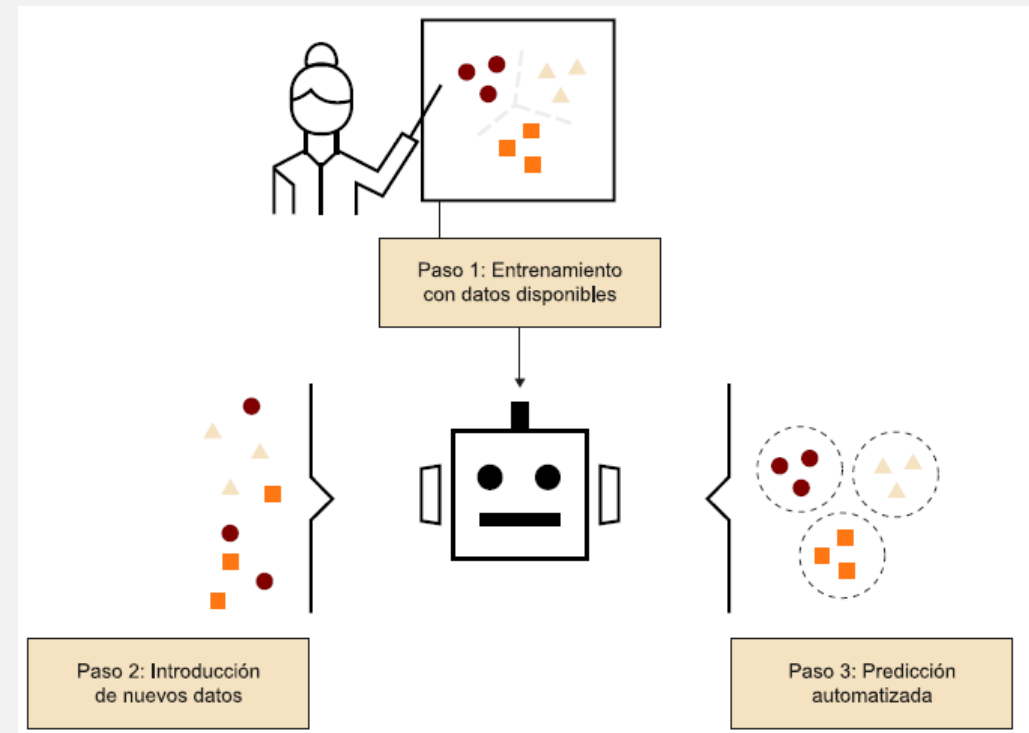
FAIR DATA TREATMENT / JUSTICIA DE DATOS

Los algoritmos no solamente pueden aprender de nuestros propios sesgos y de los sesgos contenidos en los datos, sino que son capaces de amplificarlos y magnificarlos.

Para corregirlos podemos implementar...

1. Intervenciones sobre los datos de entrenamiento.
2. Intervenciones sobre el algoritmo inductor.
3. Intervenciones sobre los resultados del algoritmo.

Edizel, Bora; Bonchi, Francesco; Hajian, Sara; Panisson, André; Tassa, Tamir (2020). «FaiRecSys: Mitigating algorithmic bias in recommender systems». *International Journal of Data Science and Analytics* (vol. 9, núm. 2, págs. 197-213). Nueva York: Springer.





ESKERRIK ASKO

morales.Jordi@gmail.com