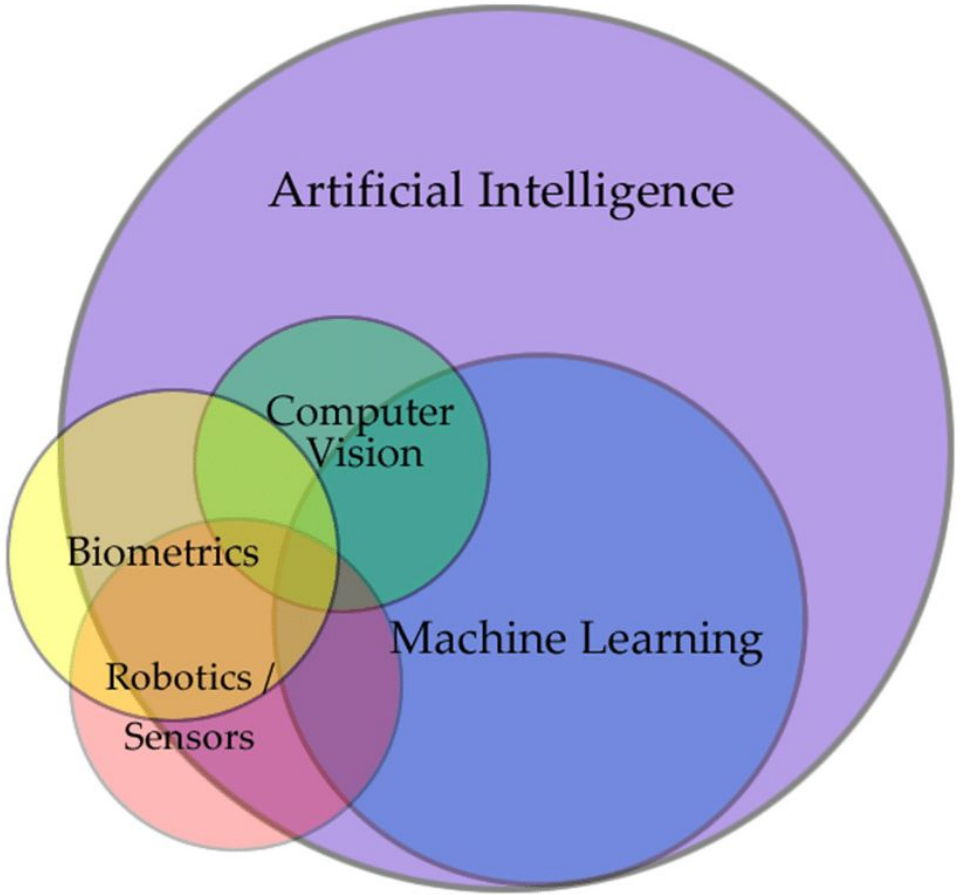
1. INTRODUCCIÓN A LA CIENCIA DE DATOS

Fecha:

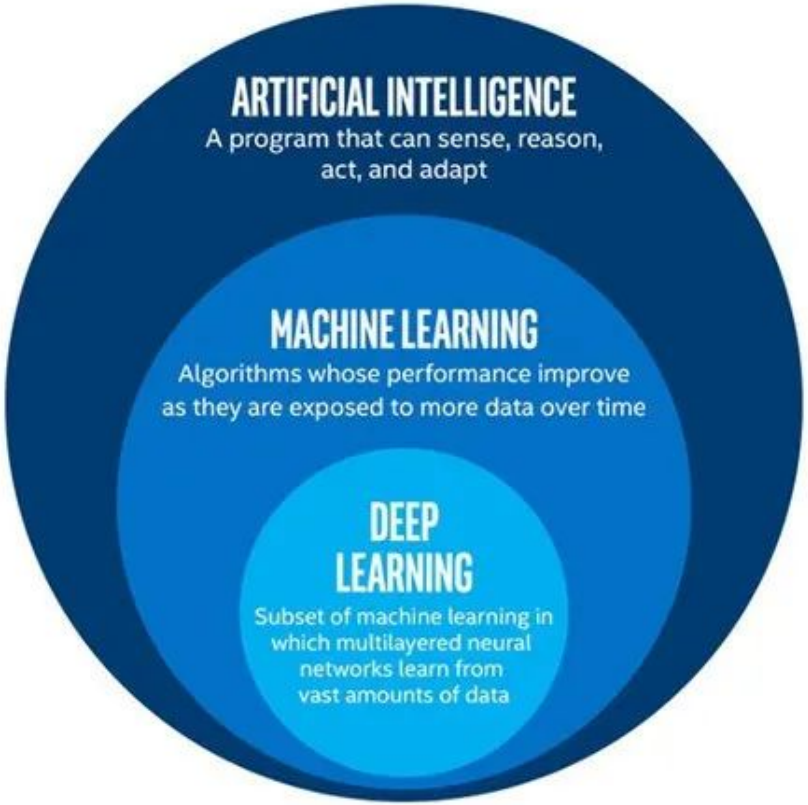
INTELIGENCIA ARTIFICIAL.

Disciplina que permite a las computadoras imitar a los humanos inteligencia como la toma de decisiones, el procesamiento de textos y percepción visual. Esta disciplina involucra muchas otras pero básicamente se desea que la máquina **sienta, razone, tome decisiones y que se adapte**.

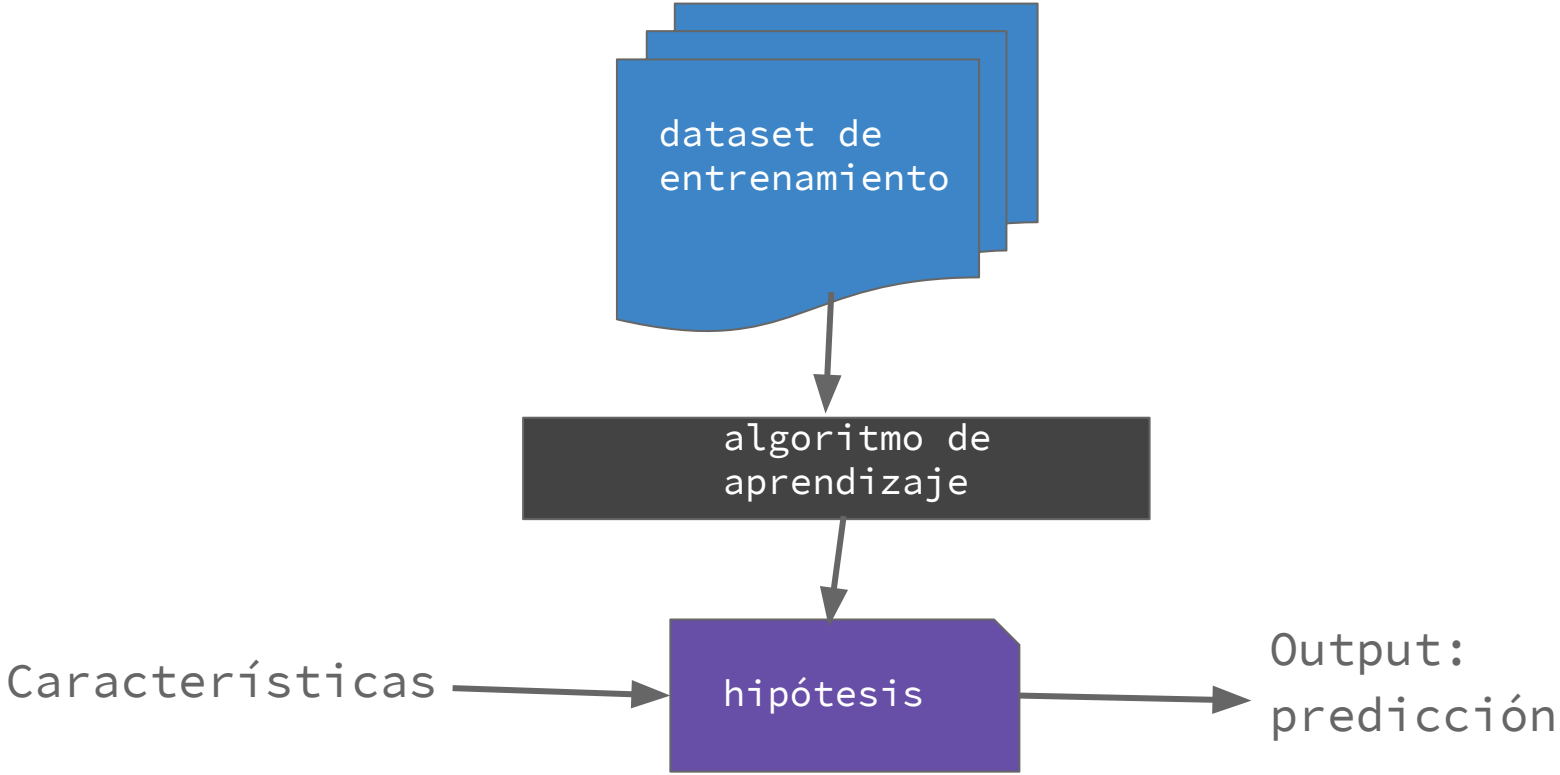


MACHINE LEARNING

El aprendizaje de máquina es un subcampo de la inteligencia artificial que permite a las máquinas tener la habilidad de aprender sin ser programadas explícitamente para ello. Se basan en la experiencia y con experiencia nos referimos a su exposición a más variedad de datos.

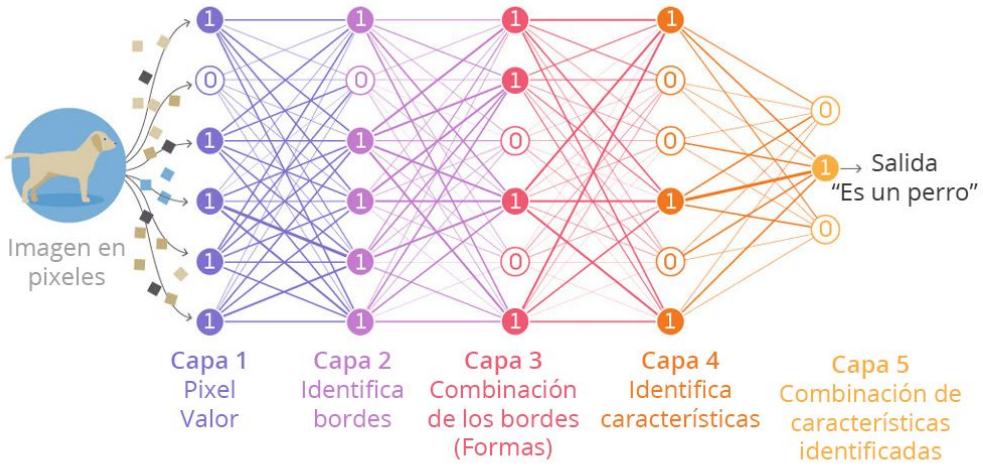


¿CÓMO FUNCIONAN LOS ALGORITMOS DE APRENDIZAJE?



DEEP LEARNING

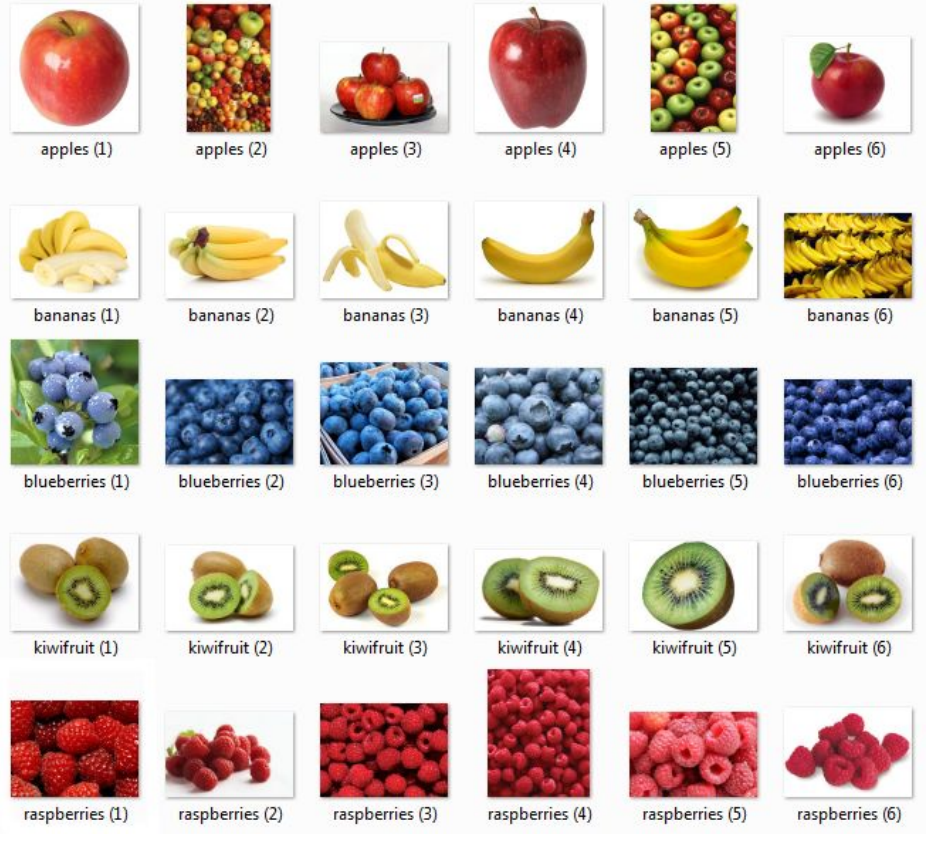
Es un campo especializado de ML que se basa en el entrenamiento de redes neuronales artificiales (ANNs) profundas utilizando un gran conjunto de datos, como imágenes o texto. Las ANNs son modelos de procesamiento de información inspirados en el cerebro humano.



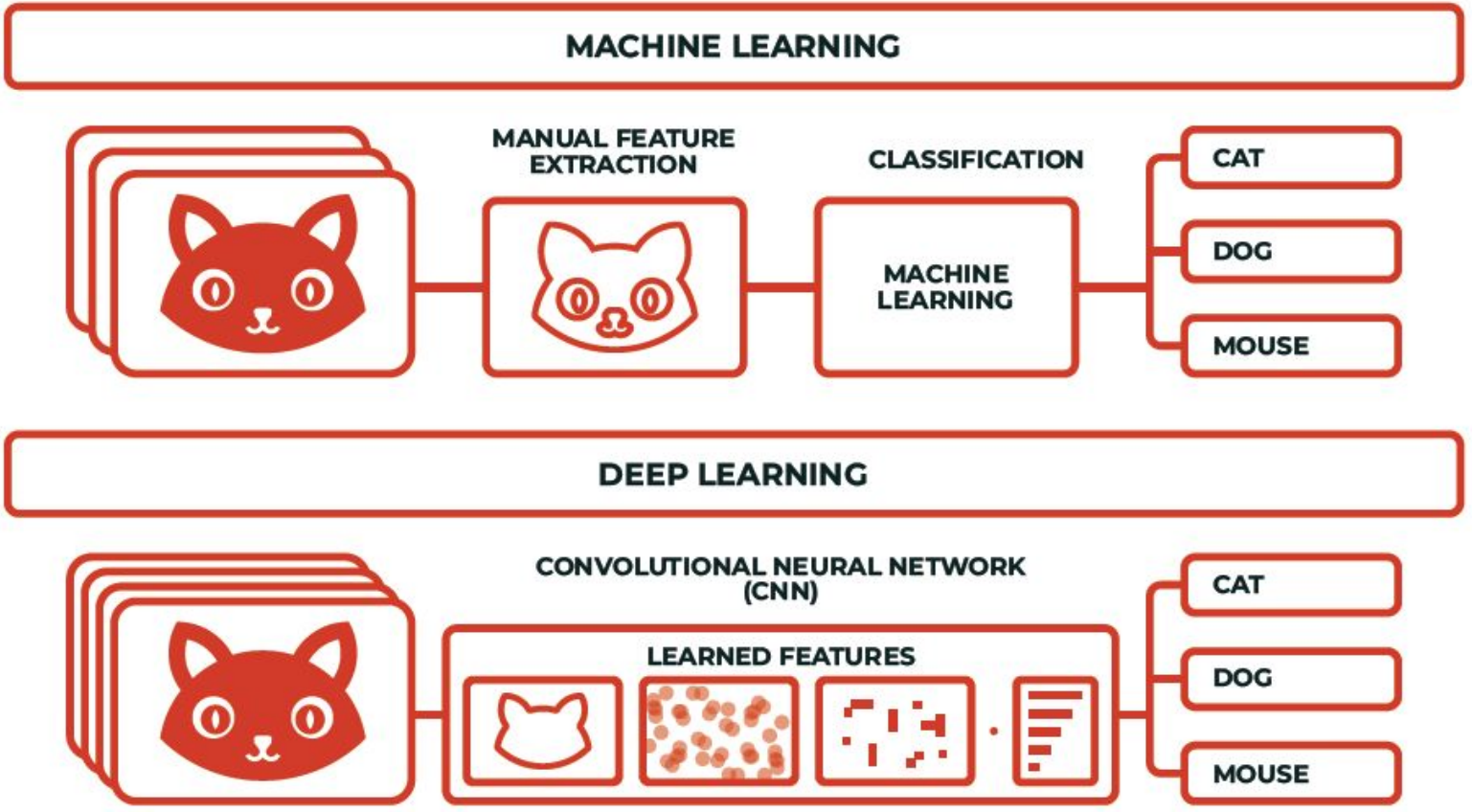
Fuente: https://www.quantamagazine.org/

¿CÓMO APRENDEMOS LOS HUMANOS?

All human learning is — observing something, identifying a pattern, building a theory (model) to explain this pattern and testing this theory to check if its fits in most or all observations.



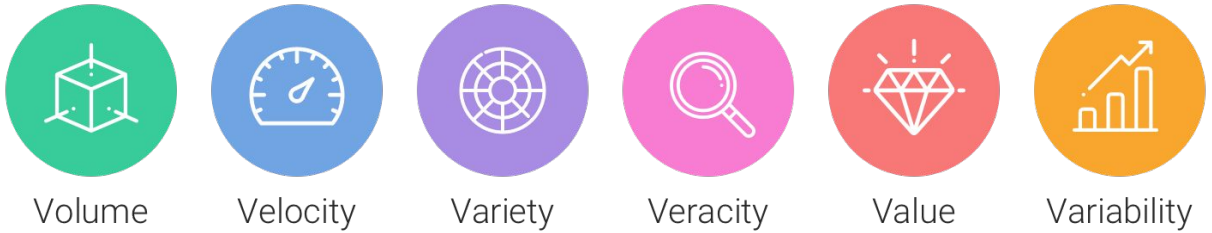
ML vs DL


MACHINE LEARNING
MANUAL FEATURE EXTRACTION
CLASSIFICATION
MACHINE LEARNING
CAT
DOG
MOUSE
DEEP LEARNING
CONVOLUTIONAL NEURAL NETWORK (CNN)
LEARNED FEATURES
CAT
DOG
MOUSE

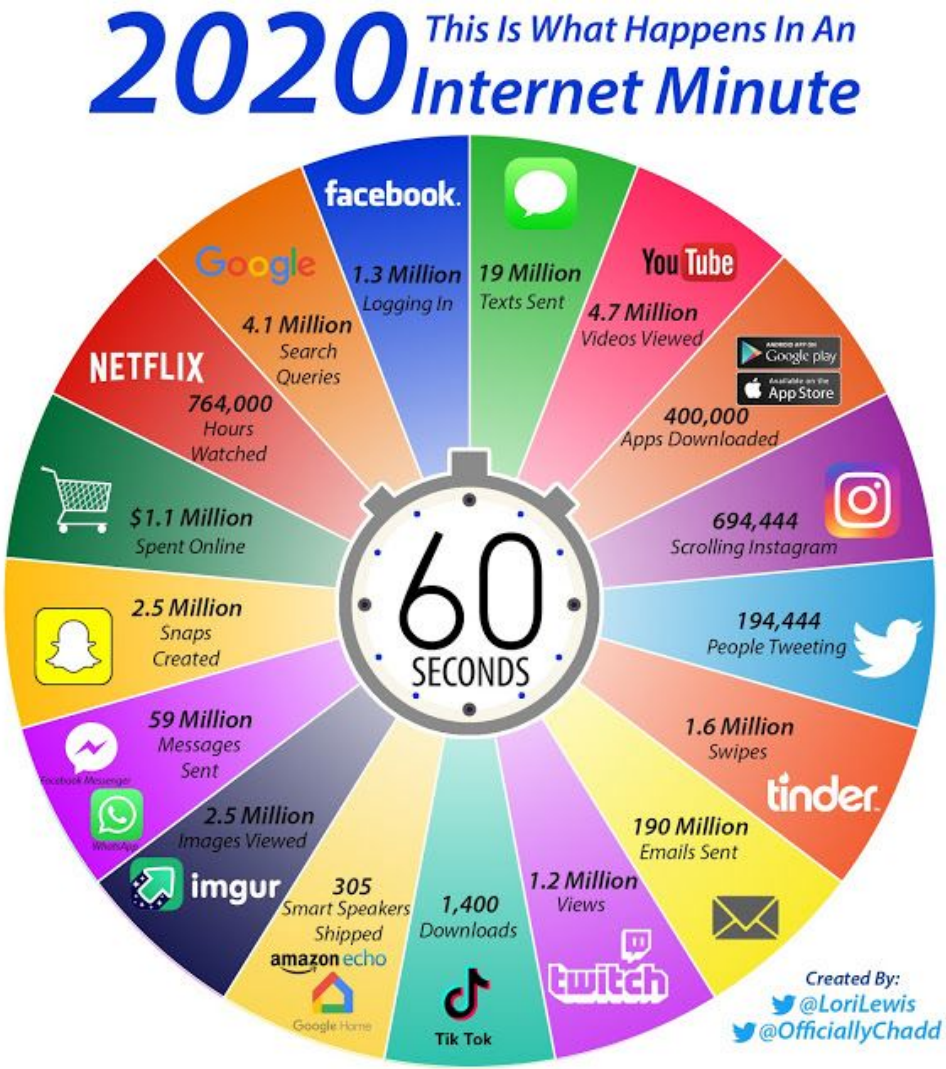
Big Data.

Conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño, complejidad (variabilidad) y velocidad de crecimiento dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales.

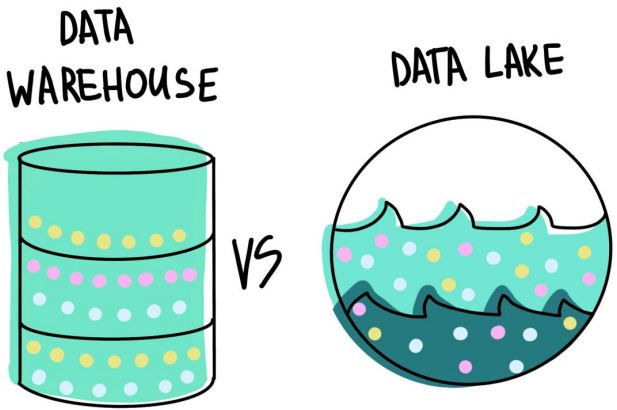
NO TODOS LOS PROYECTOS DE ML SON BIG DATA NI VISCEVERSA.

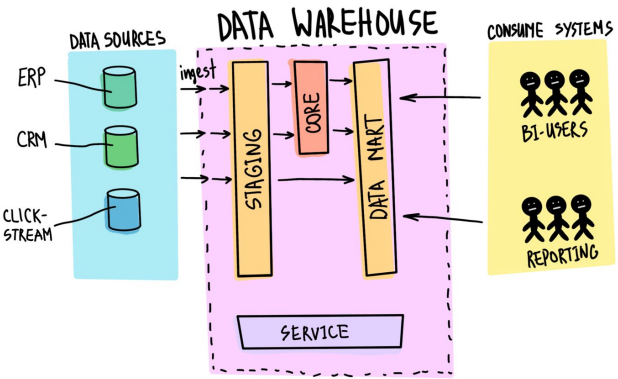


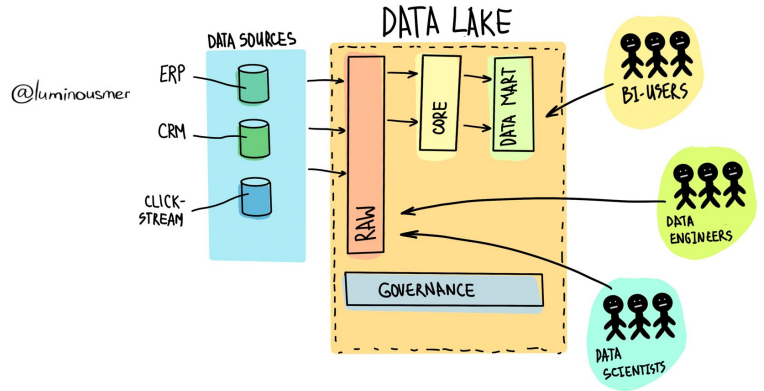
GENERAMOS MILLONES DE DATOS NUEVOS CADA MINUTO…



¿CÓMO LOS ALMACENAMOS?







Follow us!

Share it!

DATA LAKE VS DATA WAREHOUSE

DATA LAKE	DATA WAREHOUSE
• Data is kept in its raw form in Data Lake and here all the data are kept irrespective of the source of the data.	• Data Warehouse is composed of data that are extracted from transactional and other metrics systems.
• The main target for Data Lake is Data Scientists, Big Data Developers.	• The main target of Data Warehouse is the operational users.
• The main inputs to Data Lake are all kinds of data such as structured, semi-structured & unstructured data.	• The main inputs to Data warehouse are structured data that are coming from transactional systems.
• Comprises of raw data that may or might not be curated.	• It consists of curated data which is centralized and is ready to be used.

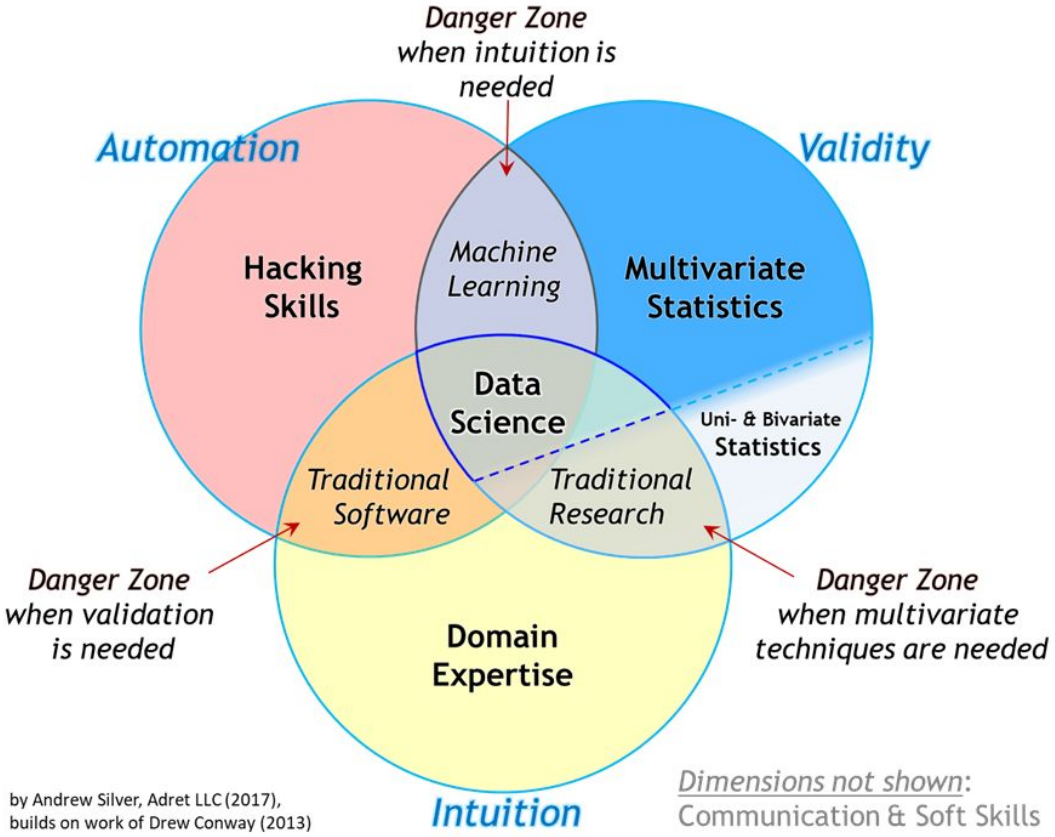
Like To Support

DataScienceLearn

Save or Regret!

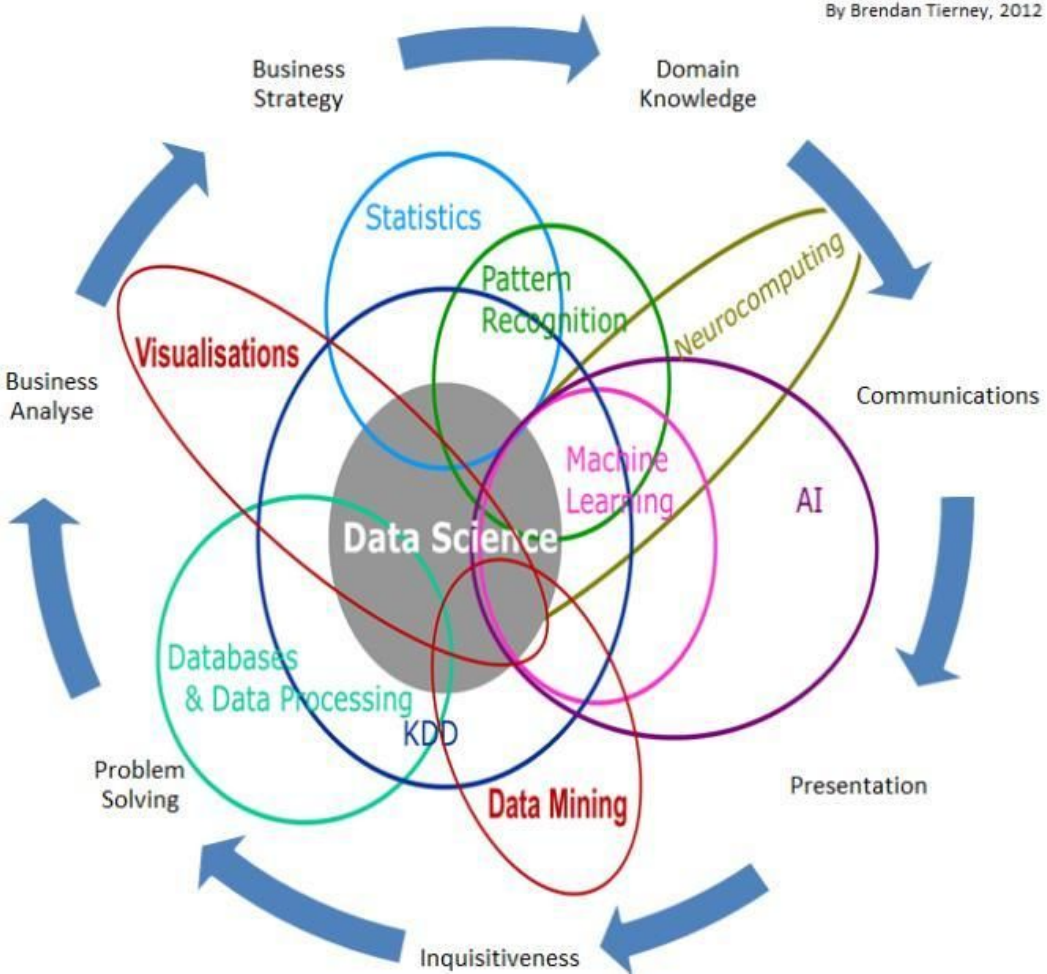
DATA SCIENCE

Es un campo interdisciplinario que se ocupa de los procesos y sistemas usados en la extracción de conocimiento a partir del análisis de datos. Es interdisciplinario porque requiere conocimientos de los campos de computación, matemáticas y del campo de estudio requerido.

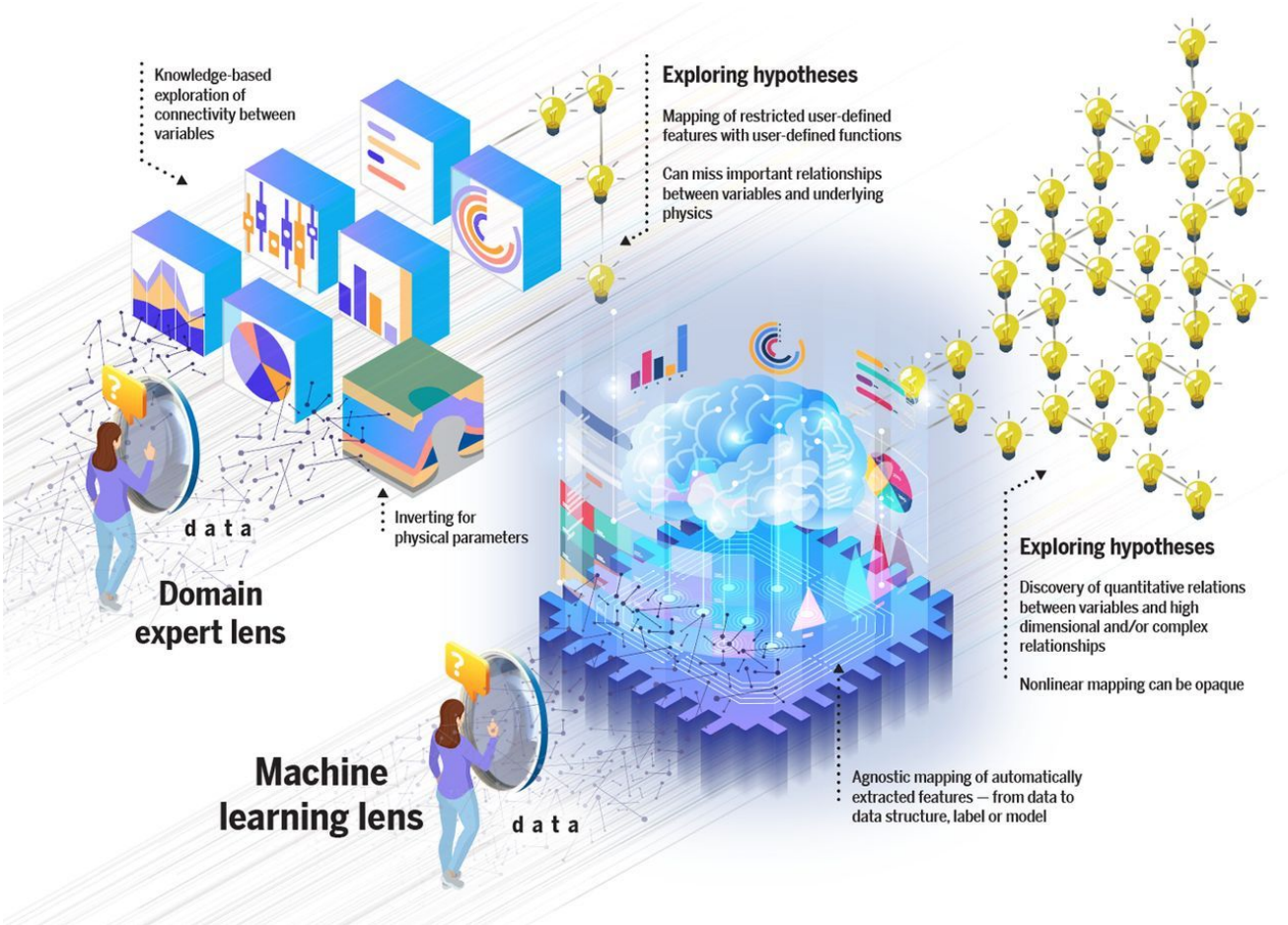


Data Science Is Multidisciplinary

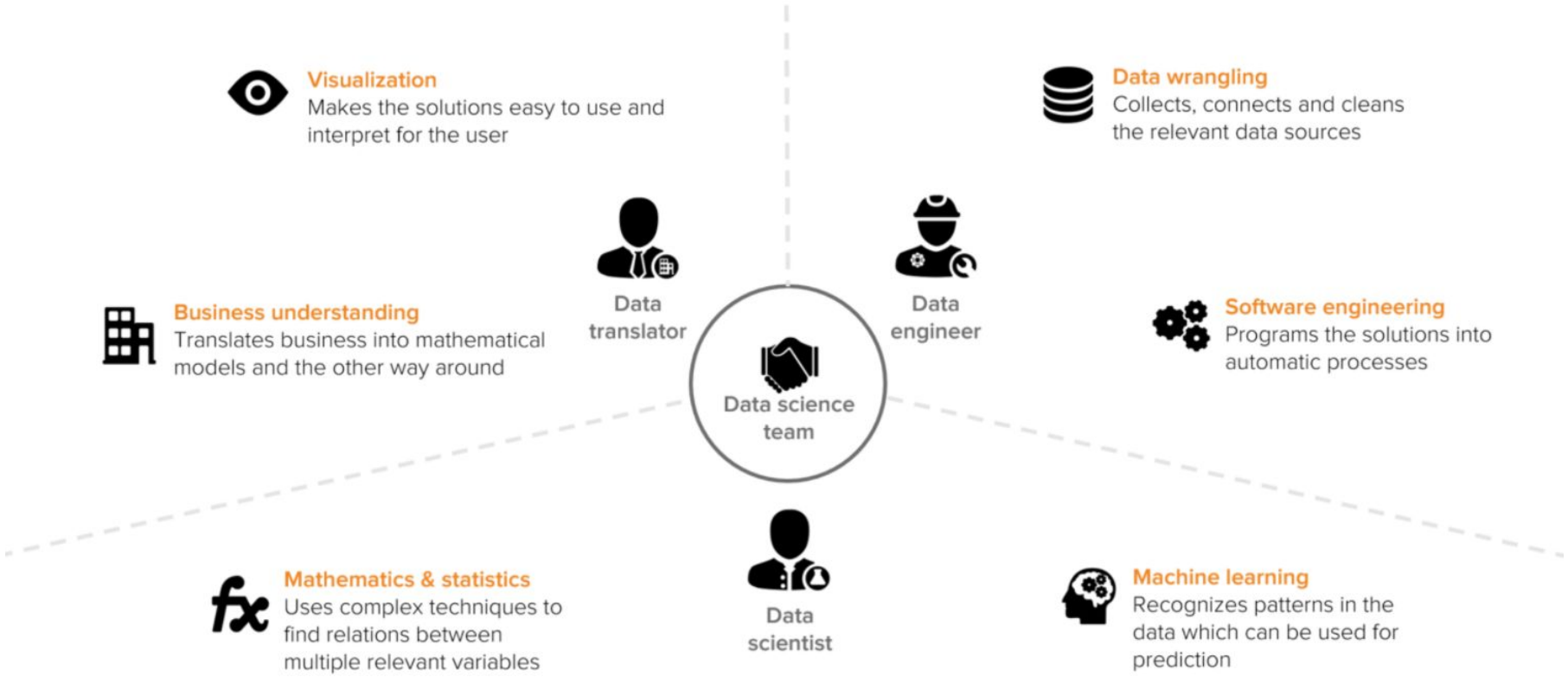
By Brendan Tierney, 2012



DOMINIO DE EXPERTO



ROLES EN CIENCIA DE DATOS



TIPO DE ANALÍTICA

DESCRIPTIVO

Ayuda a entender cómo van las cosas

- Analítica post hechos a través del análisis de datos históricos
- Mejora el entendimiento de la situación actual y la medición de resultados.

PREDICTIVO

Ayuda a predecir futuros rendimientos y resultados

- Supone el uso de la información para realizar predicciones, segmentaciones, optimizaciones y simulaciones
- Es necesario aplicar técnicas avanzadas de "data mining" y de algoritmos de predicción y forecasting.

PRESCRIPTIVO

Ayuda a sugerir un siguiente paso o acción

- Sintetiza las técnicas de Big Data, matemáticas y ciencias computacionales para aplicarles reglas de negocio y sugerir opciones de decisión.
- Pretende anticipar ventajas o prevenir riesgos y mostrar las implicaciones de cada decisión

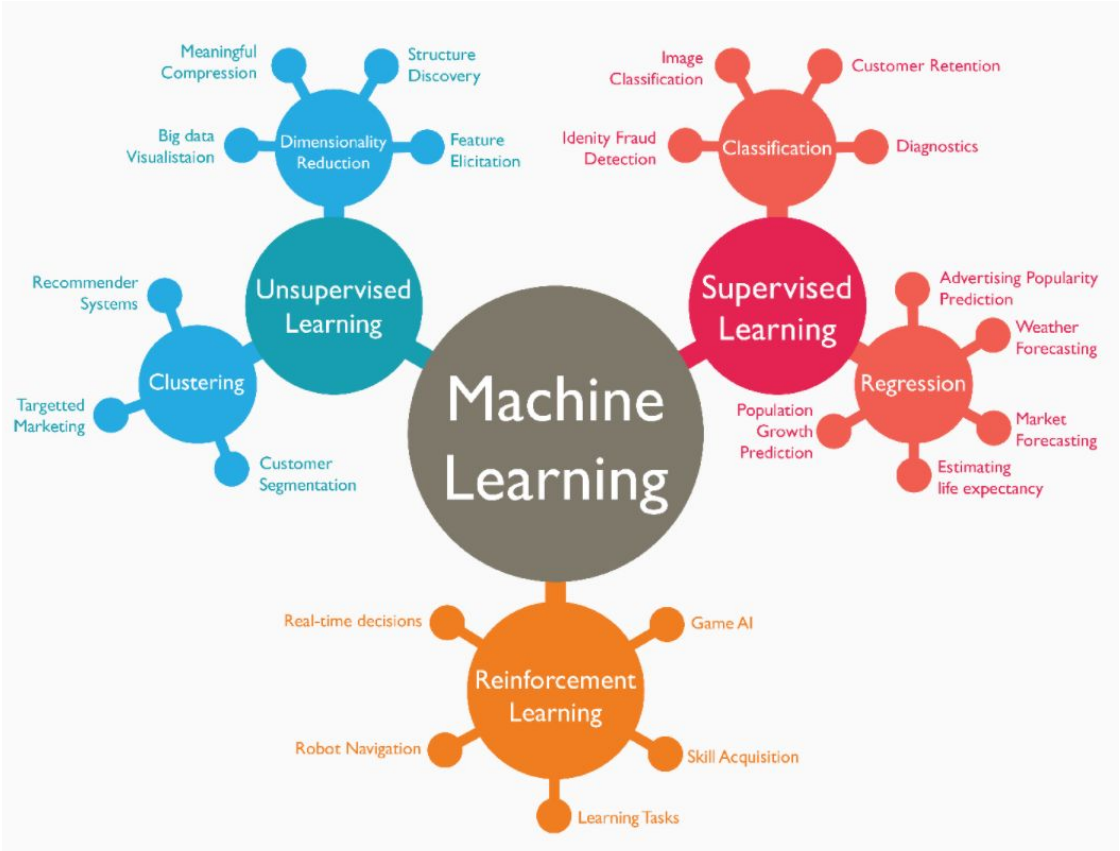
COGNITIVO

Ayuda a automatizar las tareas analíticas y las accesibiliza al público no experto

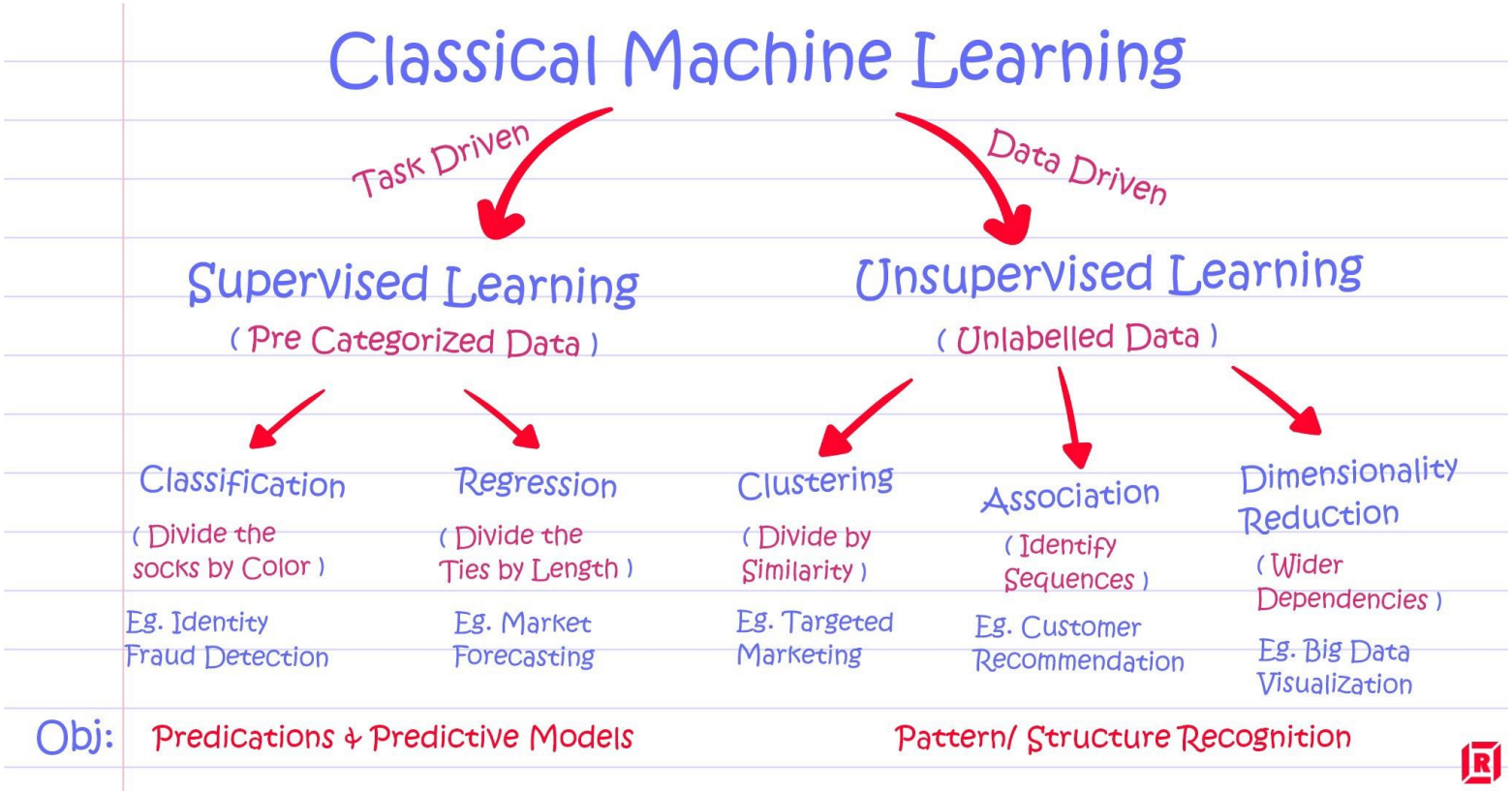
- Incorpora procesos de inteligencia de percepción, memoria, valoración, aprendizaje y razonamiento.
- Requiere el procesamiento de lenguaje natural

1.1 TIPOS DE APRENDIZAJE

Aprendizaje supervisado y no supervisado.

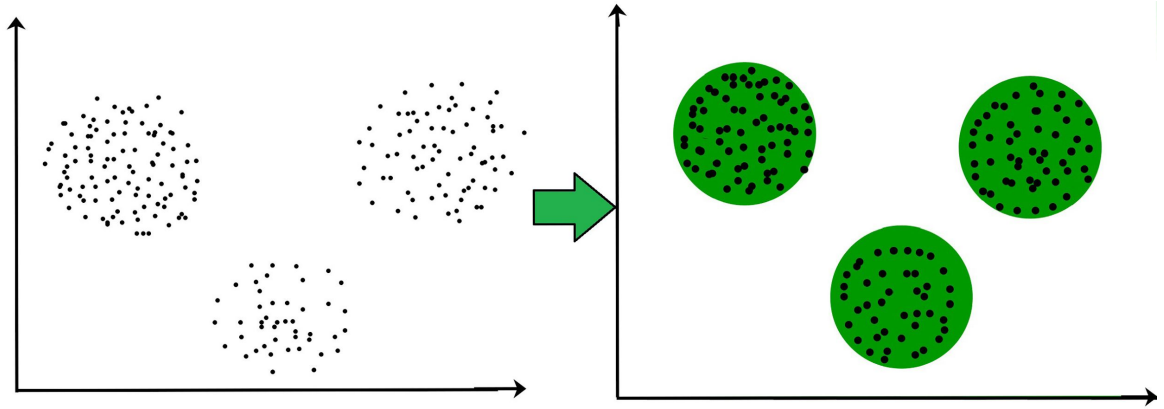


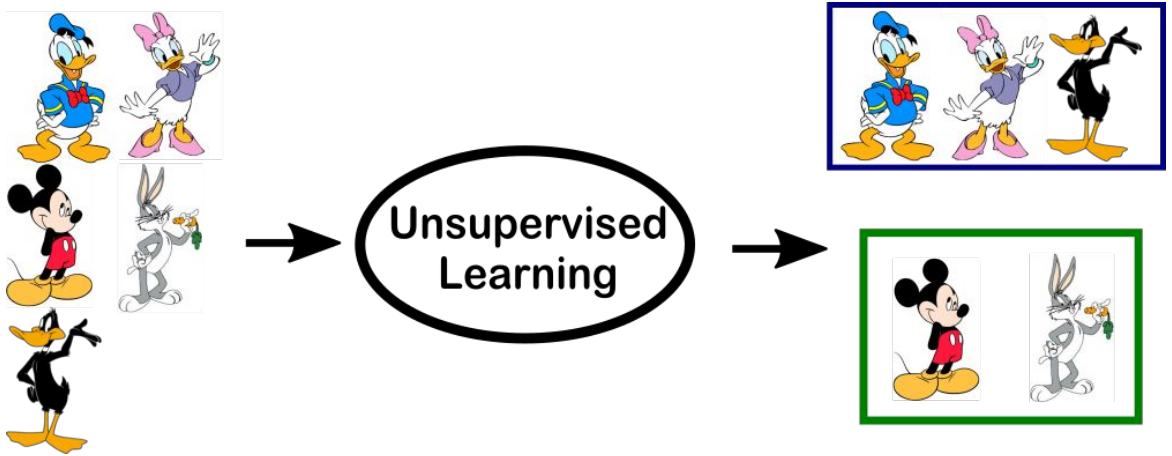
Aprendizaje supervisado y no supervisado.



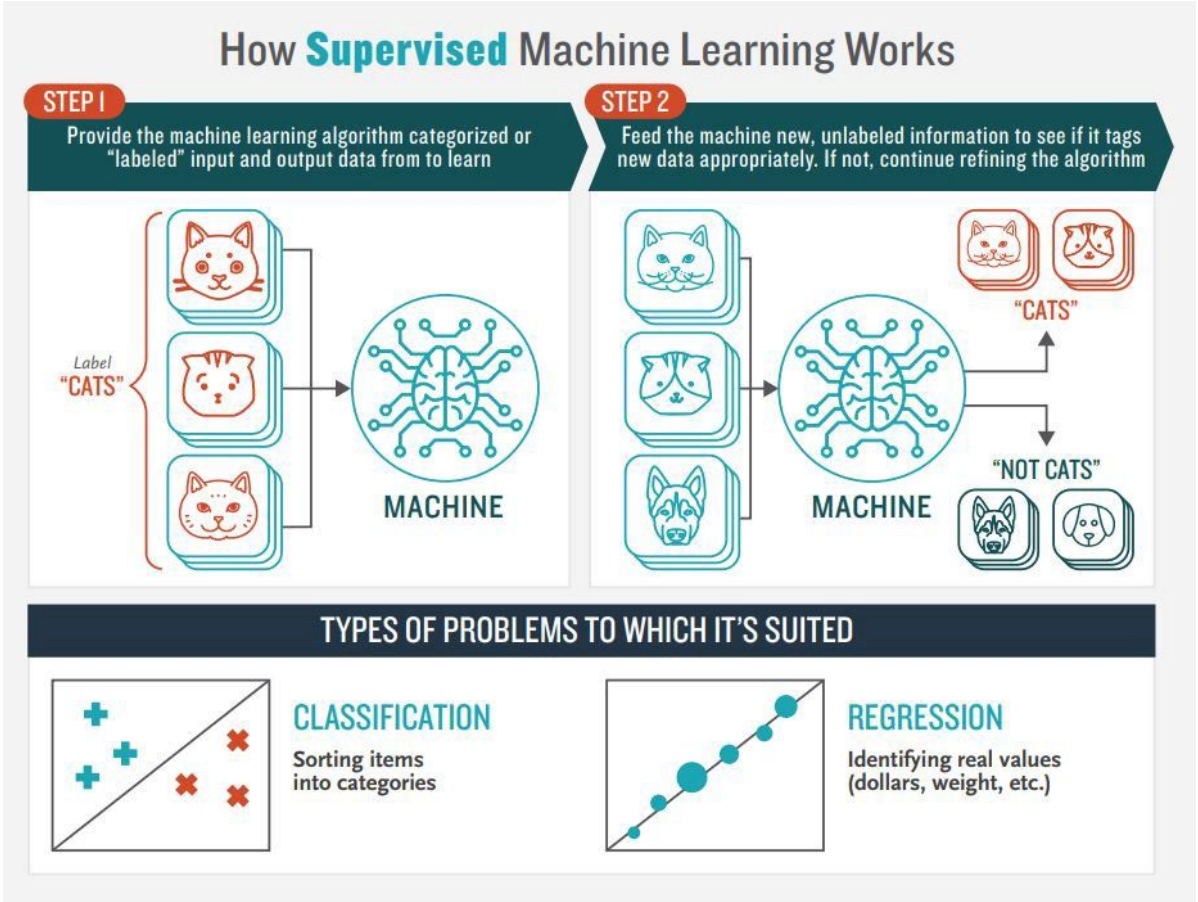
CLUSTERIZACIÓN

Consiste en alimentar a un algoritmo con datos NO etiquetados, el algoritmo debe ser capaz de **segmentar** estos datos en grupos que sean diferenciables entre sí.

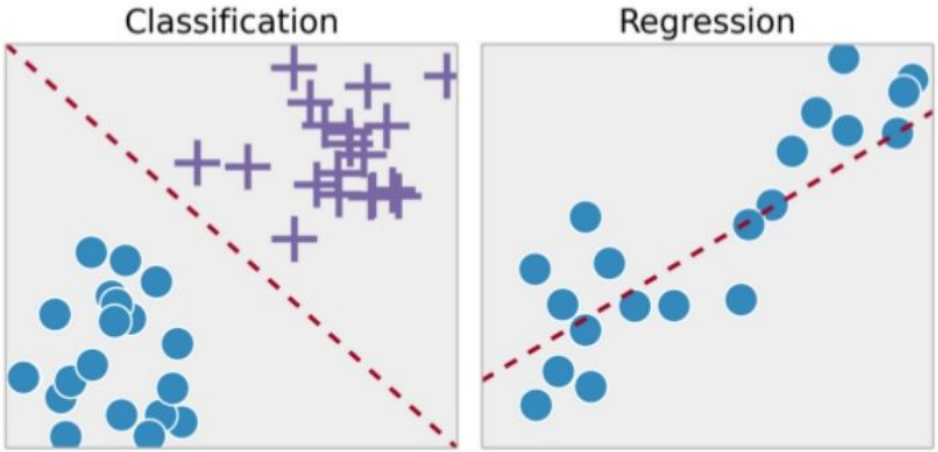




CLASIFICACIÓN Y REGRESIÓN.



CLASIFICACIÓN Y REGRESIÓN.



¿Hará frío o calor mañana?

Abril → Clasificador → Calor

Problema de clasificación

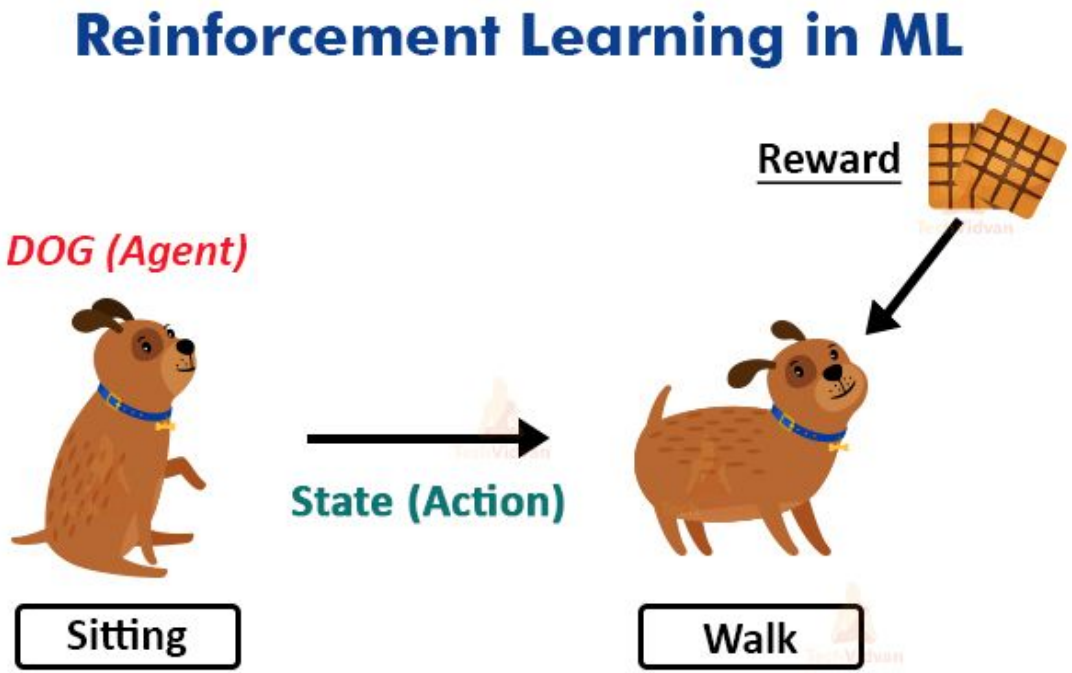
¿Qué temperatura hará mañana?

Abril → Regresor → 30º C

Problema de regresión

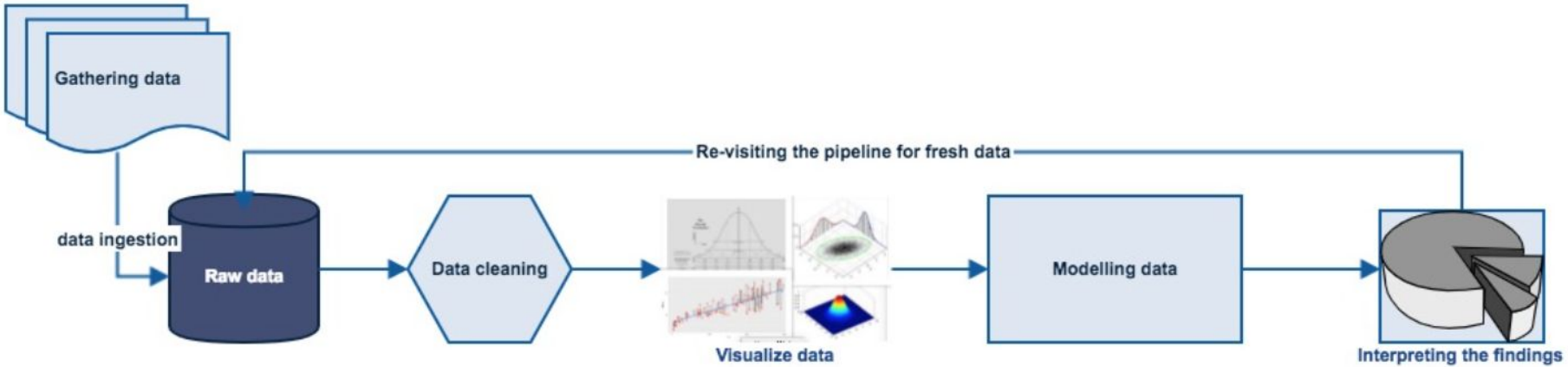
APRENDIZAJE POR REFUERZO

Se basa en la filosofía de dejar que el modelo aprenda esencialmente “haciendo lo que quiera”, “premiandolo” cuando se porte bien y “castigandolo” cuando se porte mal.



https://www.youtube.com/watch?v=fn3KWM1kuAw&ab_channel=BostonDynamics

Flujo de trabajo en proyectos de ciencia de datos.



Flujo de trabajo en proyectos de ciencia de datos.

1. **Definir un objetivo:** ¿Qué quiero hacer? ¿A dónde quiero llegar? ¿Qué quiero descubrir?

2. **Recolectar el conjunto de datos:** Buscar si existe información que me ayude a llevar a cabo mi objetivo.

3. **Explorar los datos:** Estadística descriptiva o Business Intelligence (BI).

4. **Limpieza de los datos:** Todo lo que involucre la transformación de la base de datos.

5. **Técnica de validación:** Dividir los datos en entrenamiento y prueba.

6. **Creación del modelo de inferencia.**

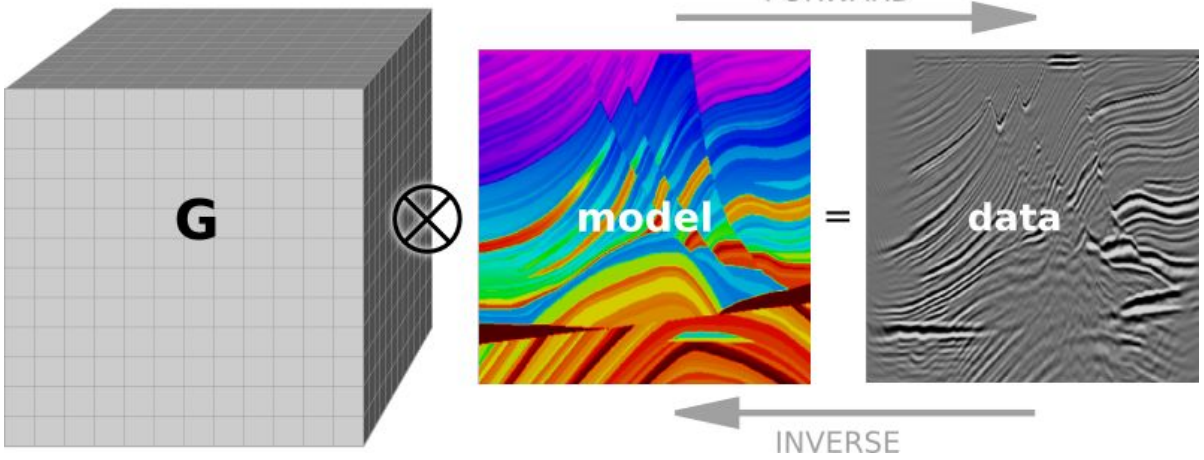
7. **Entrenamiento y prueba del modelo.**

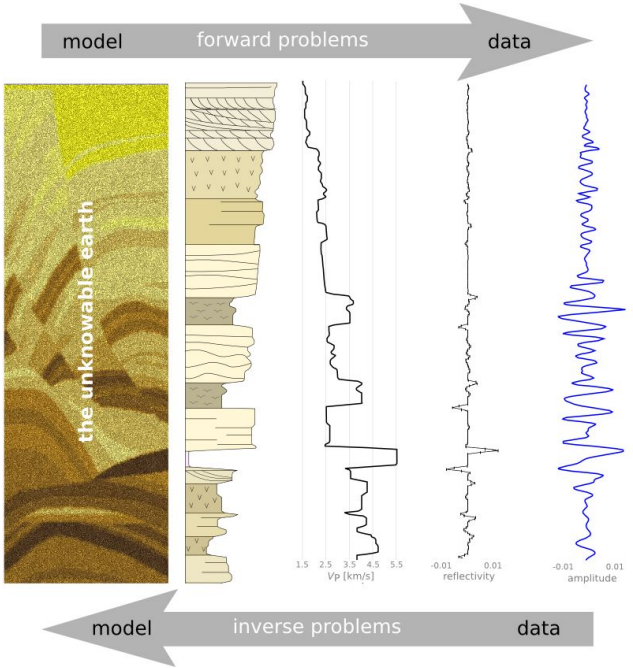
8. **Evaluación de su desempeño** con datos que no ha visto.

9. **Deployment del modelo:** Si todo lo anterior se cumple, es hora de usar este modelo para predecir datos nuevos!

1.2 EJEMPLOS

PROBLEMA INVERSO EN GEOFÍSICA



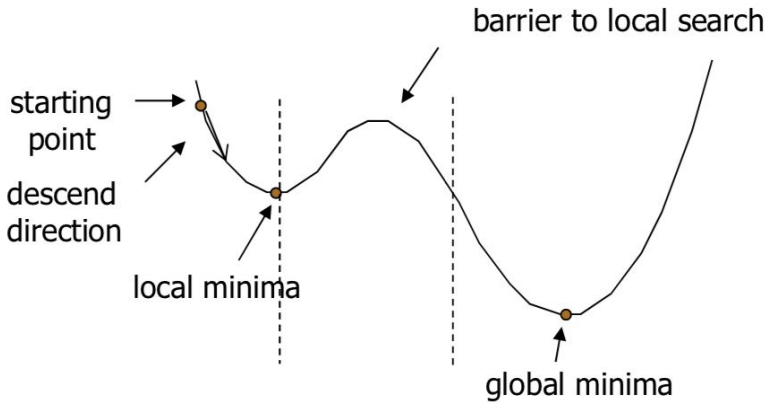


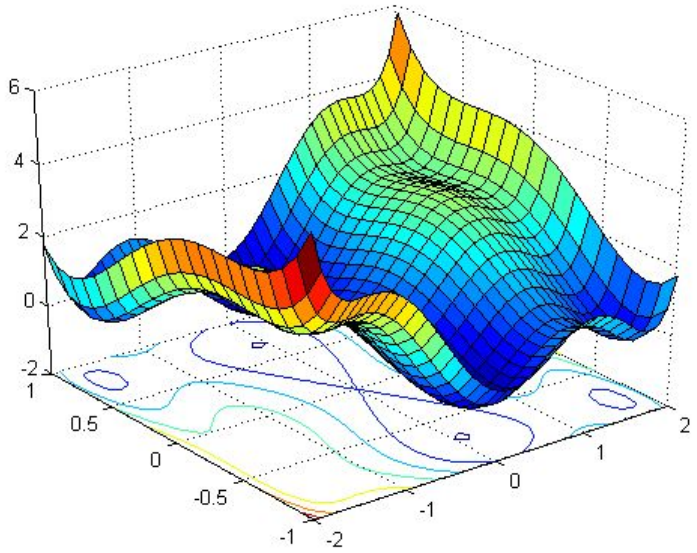
$$Gm = d$$

INTELIGENCIA ARTIFICIAL EN GEOFÍSICA

$$Gm = d$$

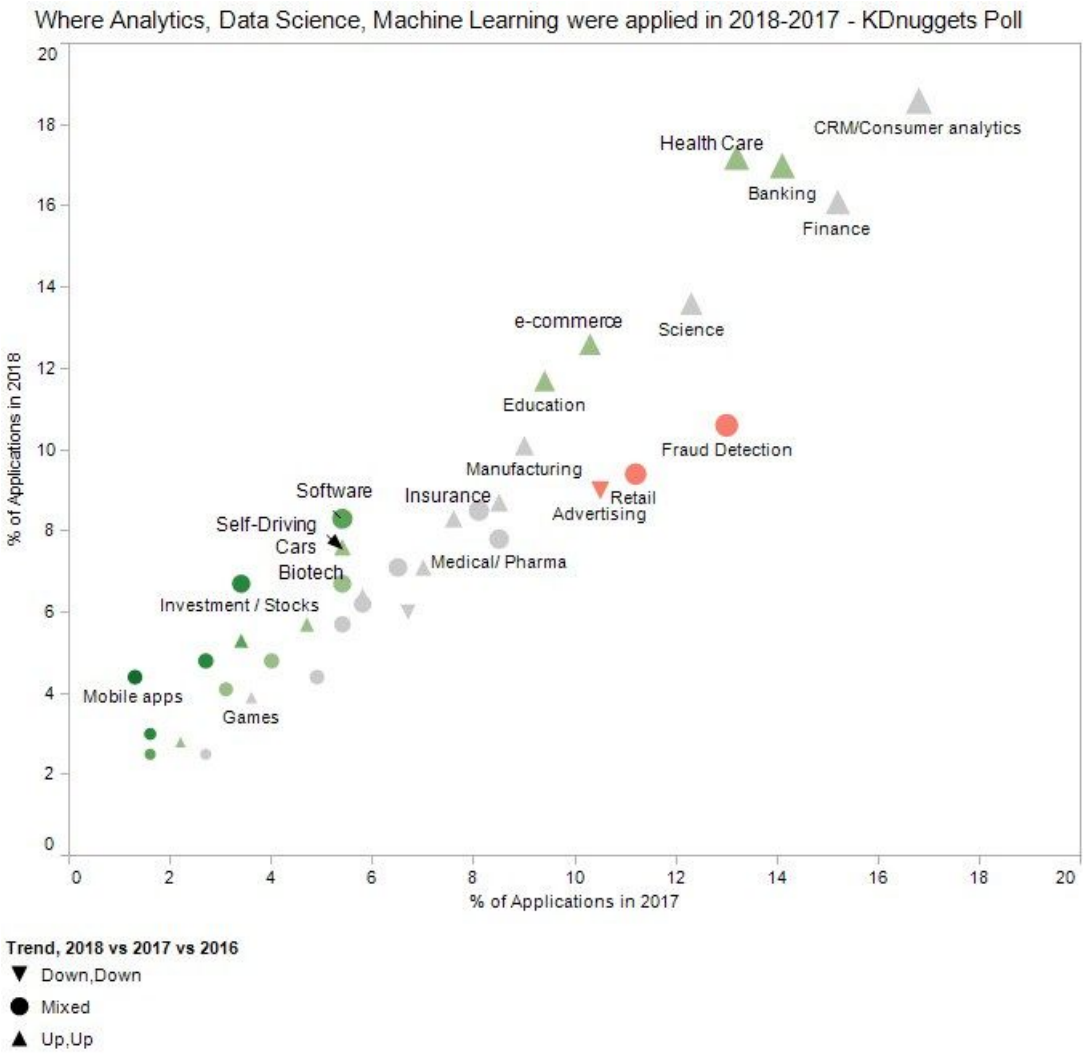
¿Simulated Annealing?
¿Algoritmos genéticos?





¿Estas estrategias son data driven?

¿DÓNDE SE APLICA LA CIENCIA DE DATOS?



STOCK PRICE PREDICTION

index	Date	Open	High	Low	Close	Volume
0	1/3/2012	325.25	332.83	324.97	663.59	7,380,500
1	1/4/2012	331.27	333.87	329.08	666.45	5,749,400
2	1/5/2012	329.83	330.75	326.89	657.21	6,590,300
3	1/6/2012	328.34	328.77	323.68	648.24	5,405,900
4	1/9/2012	322.04	322.29	309.46	620.76	11,688,800
5	1/10/2012	313.7	315.72	307.3	621.43	8,824,000
6	1/11/2012	310.59	313.52	309.4	624.25	4,817,800
7	1/12/2012	314.43	315.26	312.08	627.92	3,764,400
8	1/13/2012	311.96	312.3	309.37	623.28	4,631,800
9	1/17/2012	314.81	314.81	311.67	626.86	3,832,800
10	1/18/2012	312.14	315.82	309.9	631.18	5,544,000
11	1/19/2012	319.3	319.3	314.55	637.82	12,657,800
12	1/20/2012	294.16	294.4	289.76	584.39	21,231,800
13	1/23/2012	291.91	293.23	290.49	583.92	6,851,300
14	1/24/2012	292.07	292.74	287.92	579.34	6,134,400
15	1/25/2012	287.68	288.27			
16	1/26/2012	284.92	286.17			
17	1/27/2012	284.32	289.08			
18	1/30/2012	287.95	288.92			

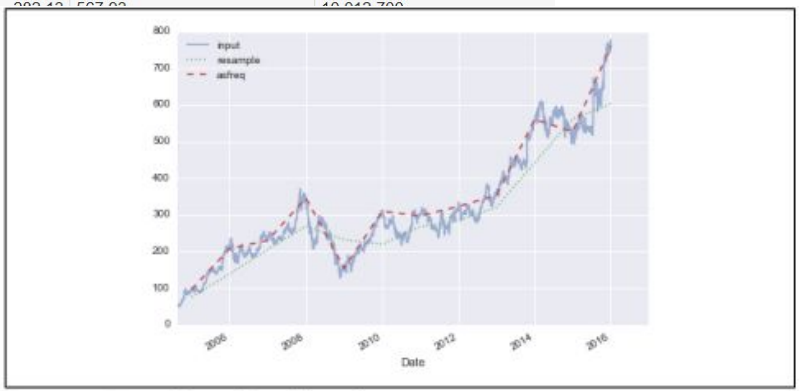
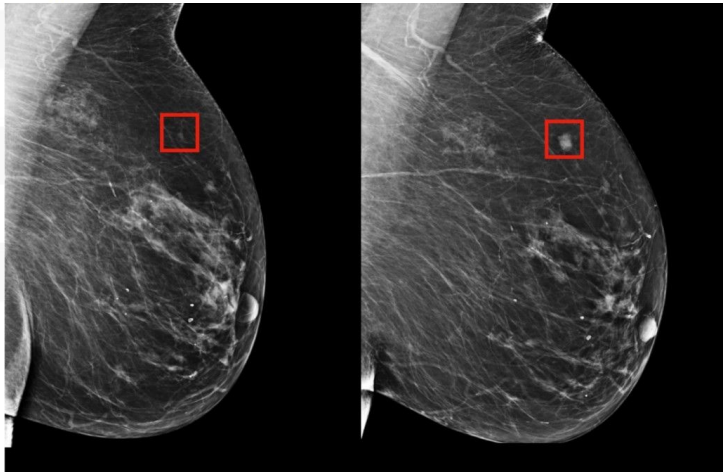


Figure 3-6. Resamplings of Google's stock price

Diagnóstico médico

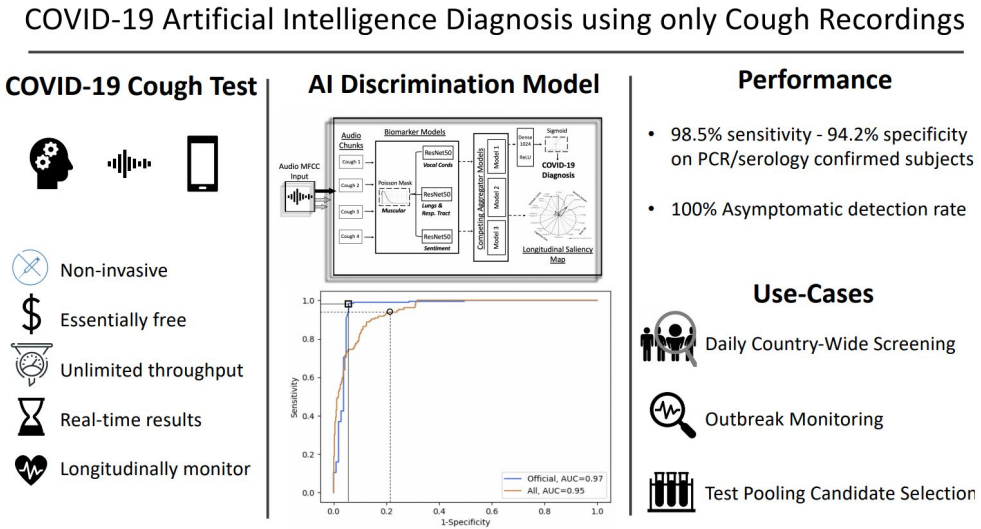
Cáncer de mama, Diabetes, autismo, etc.

mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
17.99	10.38	122.80	1001.0	0.11840	0
20.57	17.77	132.90	1326.0	0.08474	0
19.69	21.25	130.00	1203.0	0.10960	0
11.42	20.38	77.58	386.1	0.14250	0
20.29	14.34	135.10	1297.0	0.10030	
12.45	15.70	82.57	477.1	0.12780	
18.25	19.98	119.60	1040.0	0.09463	
13.71	20.83	90.20	577.9	0.11890	
13.00	21.82	87.50	519.8	0.12730	
12.46	24.04	83.97	475.9	0.11860	



DIAGNÓSTICO DE COVID-19

Detección de coronavirus a partir de tos.



RETAIL

Machine learning use cases in retail

Demand Prediction

Price Formation

Logistics

Merchandizing

Personalized Offers

Fraud Detection

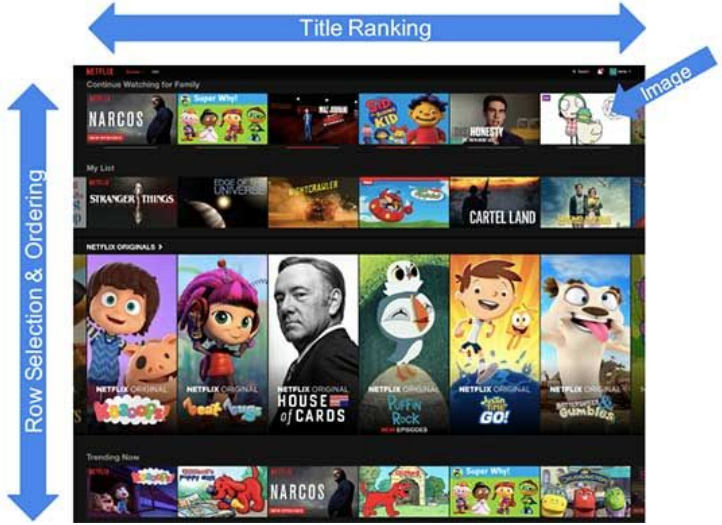
Churn Prediction

Location Optimization

Sentiment Analysis

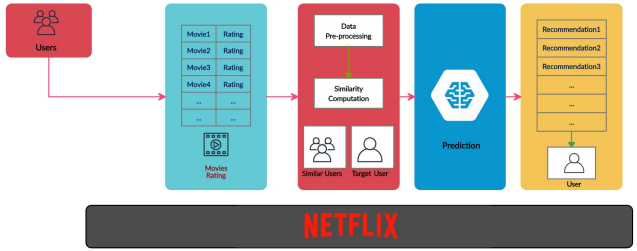
Document Work Automation

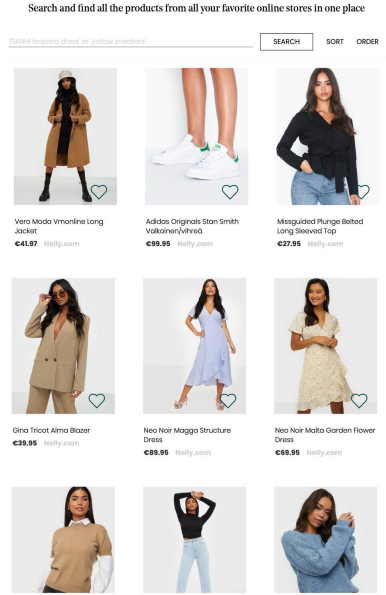
Everything is a Recommendation



Recommendations are driven by machine learning algorithms

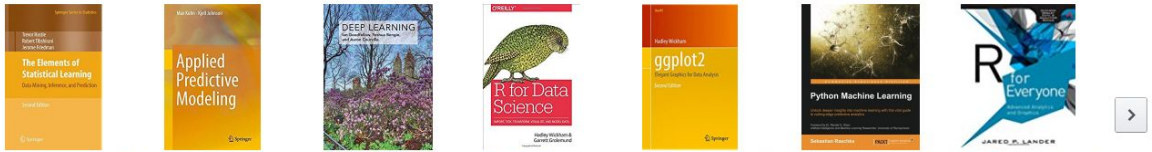
Over 80% of what members watch comes from our recommendations





Customers who bought this item also bought

Page 1 of 11



The Elements of Statistical Learning: Data Mining, Inference, and...
Trevor Hastie
17
Hardcover
CDN$ 49.90

Applied Predictive Modeling
Max Kuhn
9
Hardcover
CDN$ 85.09 prime

Deep Learning
Ian Goodfellow
8
Hardcover
CDN$ 92.40 prime

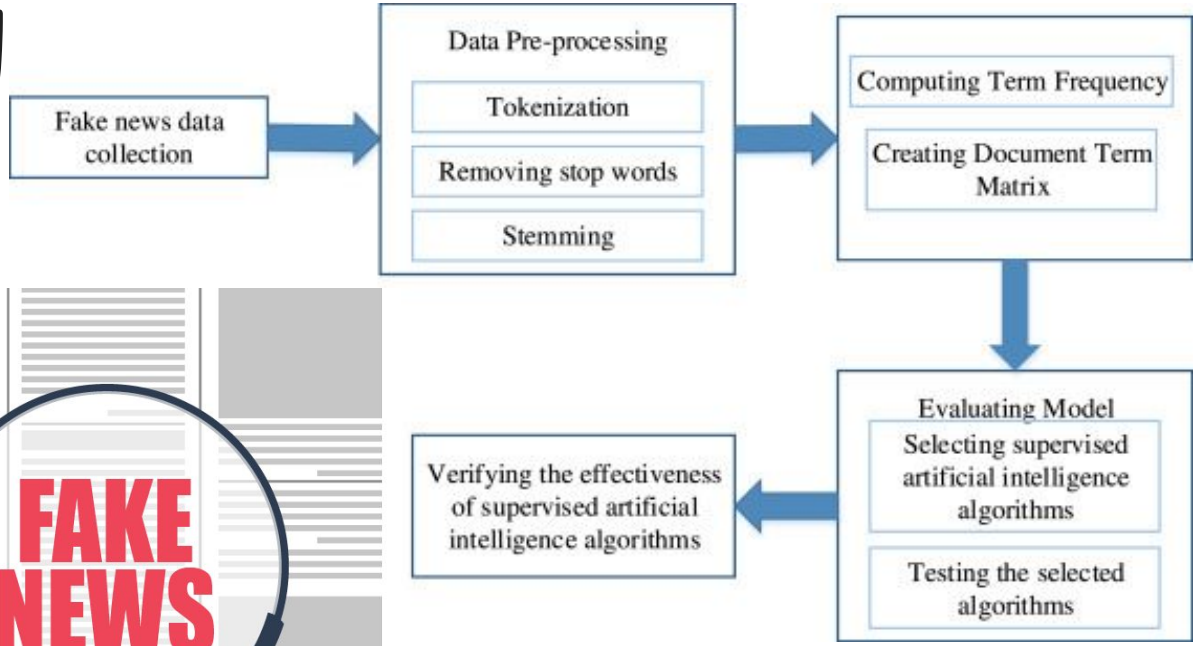
R for Data Science: Import, Tidy, Transform, Visualize, and Model Data
Hadley Wickham
7
Paperback
CDN$ 41.48 prime

ggplot2: Elegant Graphics for Data Analysis
Hadley Wickham
1
Paperback
CDN$ 55.65 prime

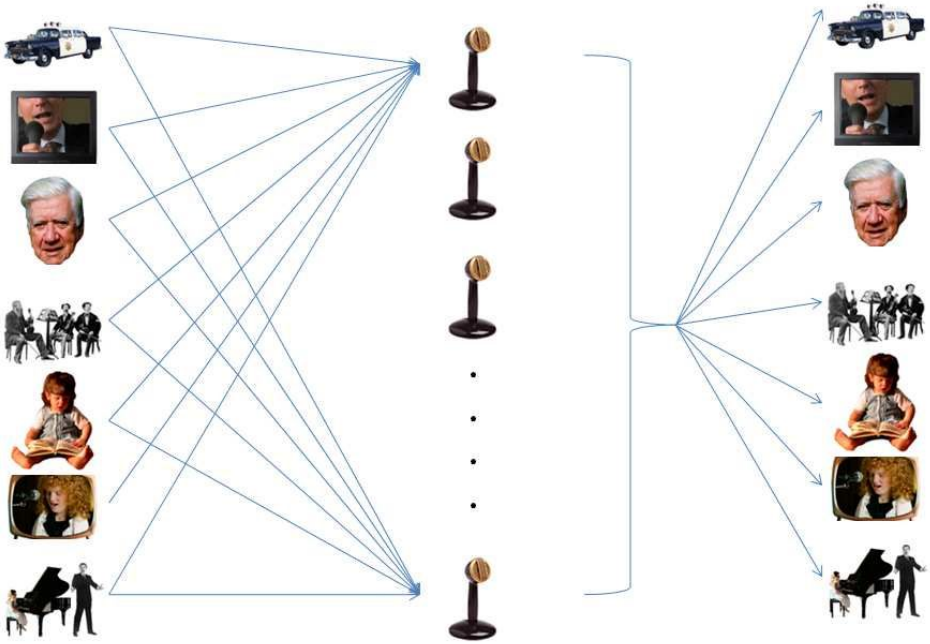
Python Machine Learning
Sebastian Raschka
7
Paperback
CDN$ 47.97 prime

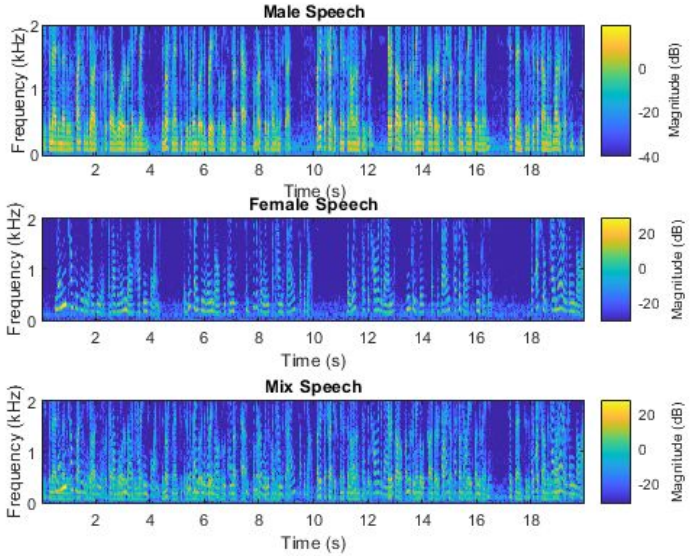
R for Everyone: Advanced Analytics and Graphics
Jared P. Lander
8
Paperback
CDN$ 38.43 prime

FAKE NEWS DETECTION

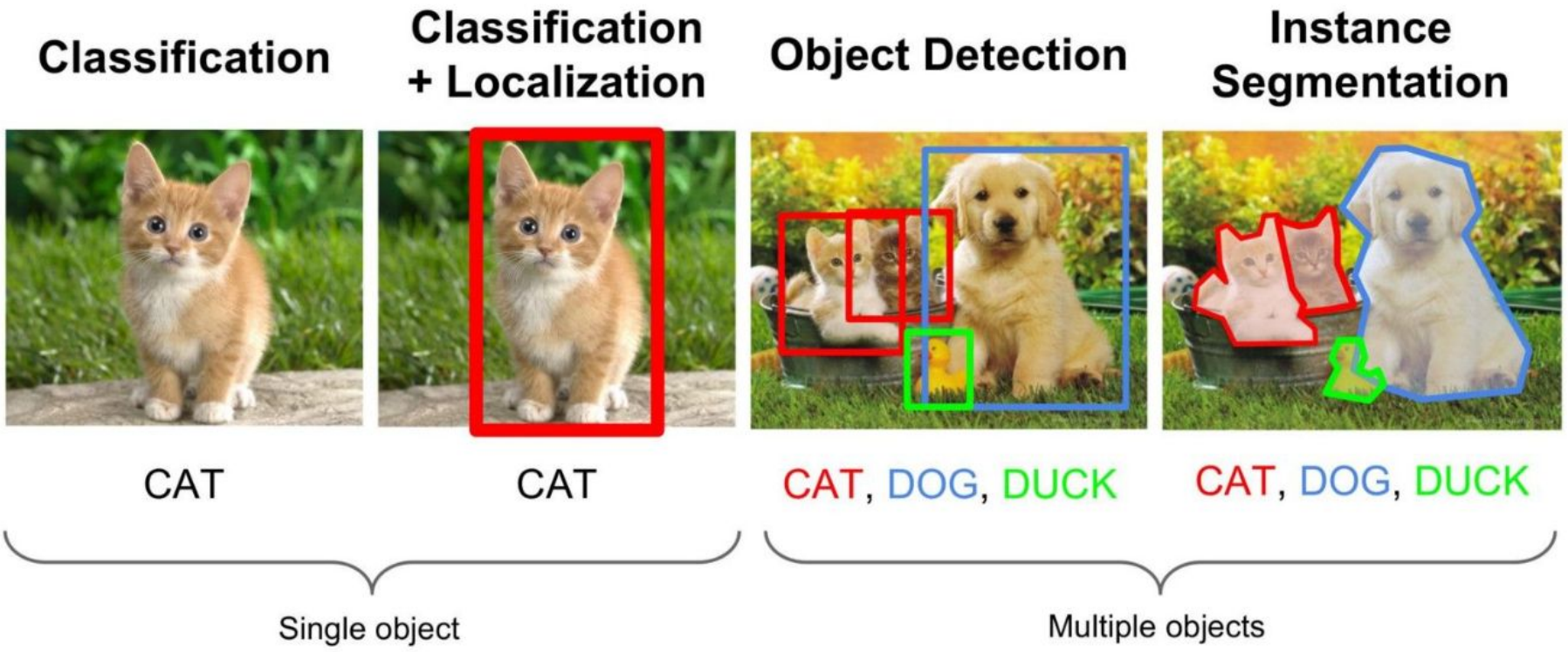


COCKTAIL PARTY PROBLEM



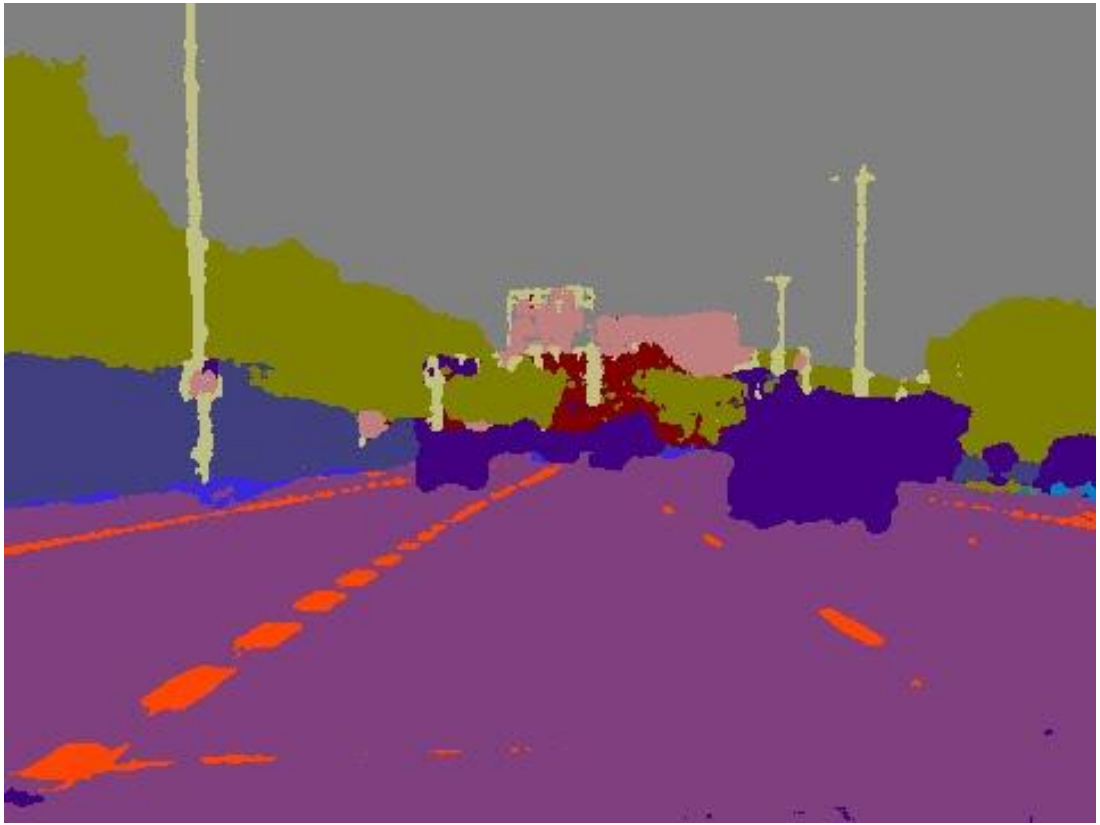


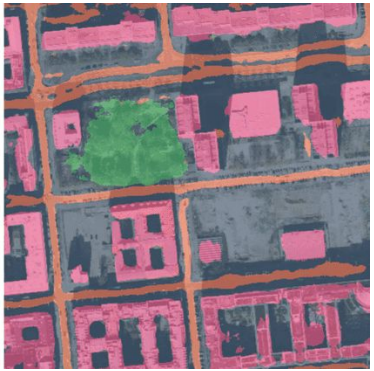
PROBLEMAS QUE SE RESUELVEN CON IMÁGENES...



DETECCIÓN DE OBJETOS

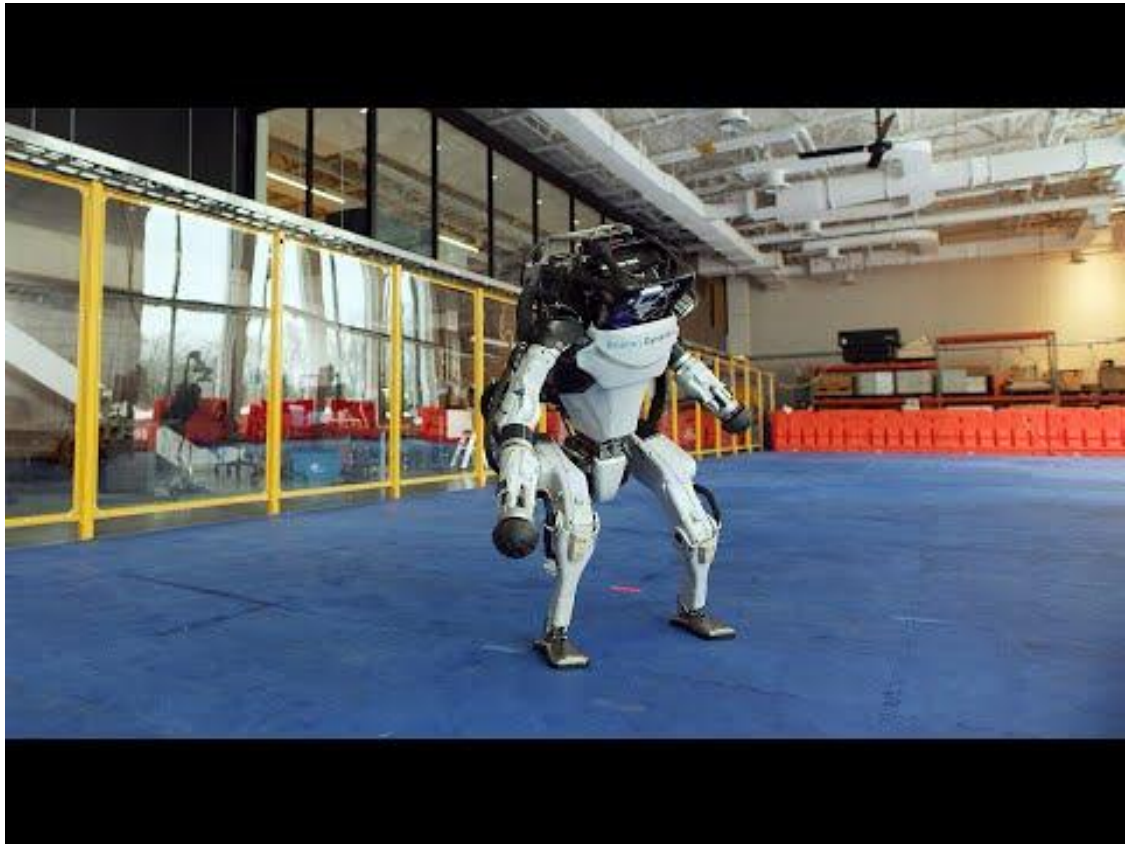
SEGMENTACIÓN



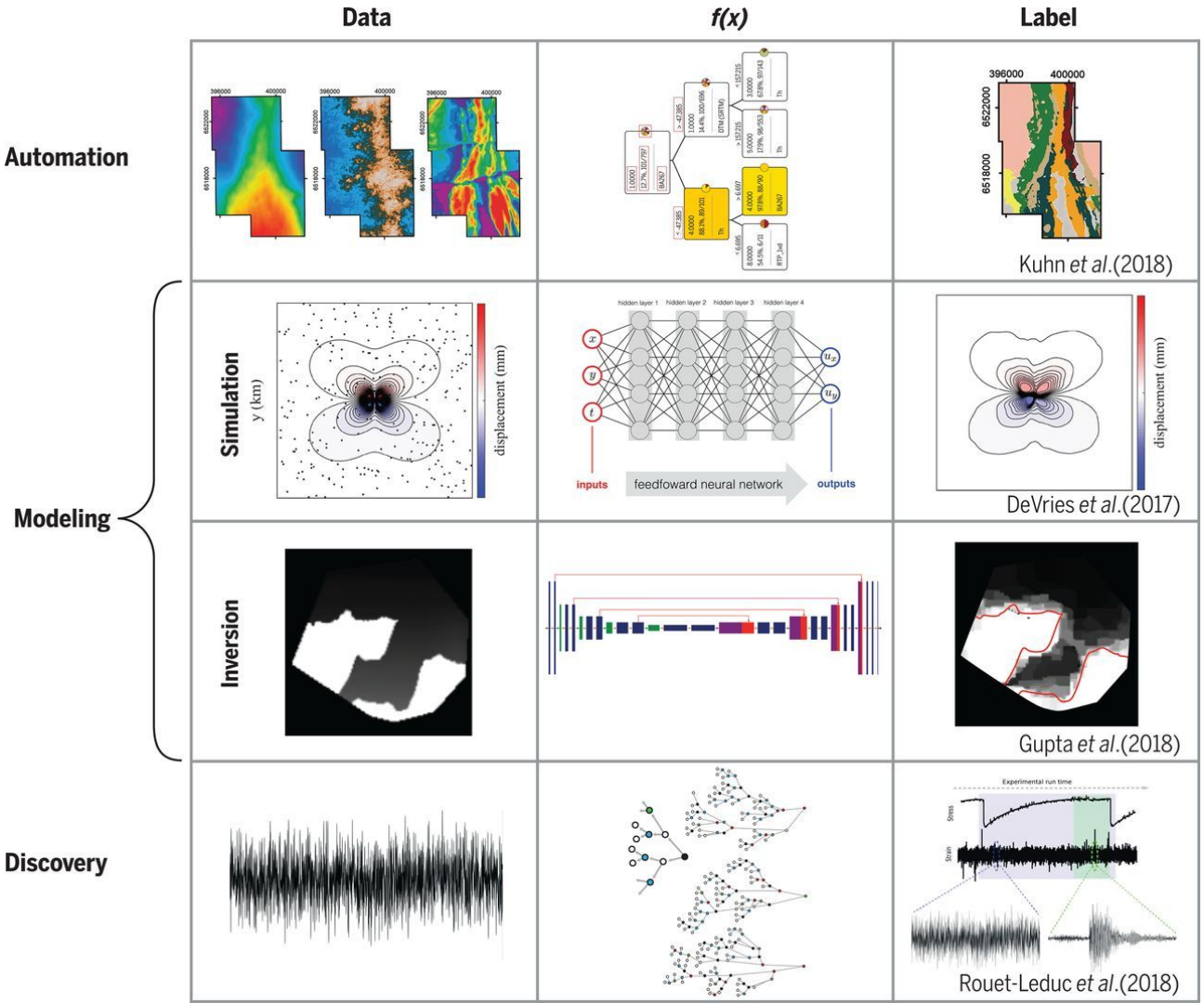


DEEP FAKES

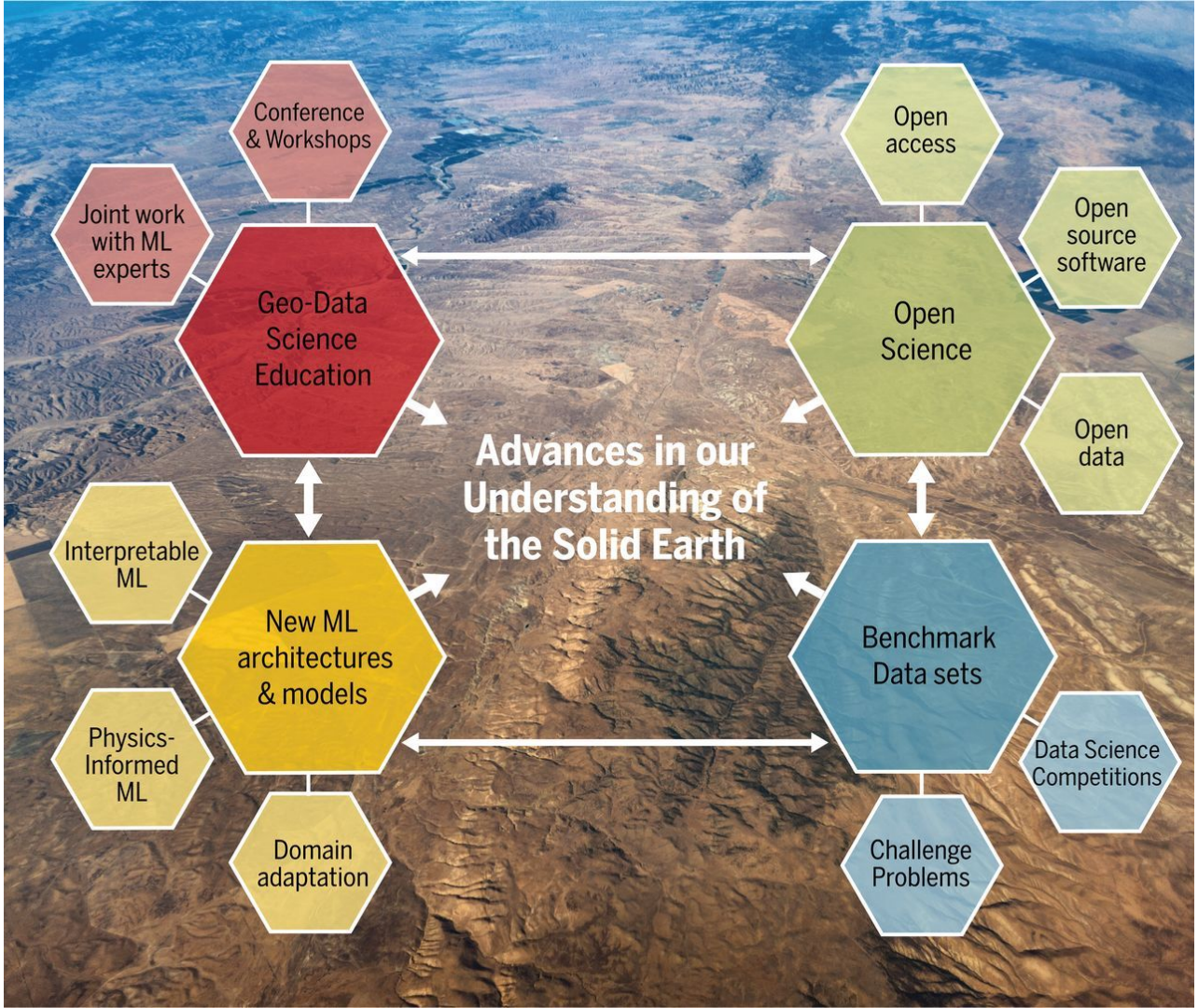
APRENDIZAJE POR REFUERZO



DATA SCIENCE EN GEOCIENCIAS

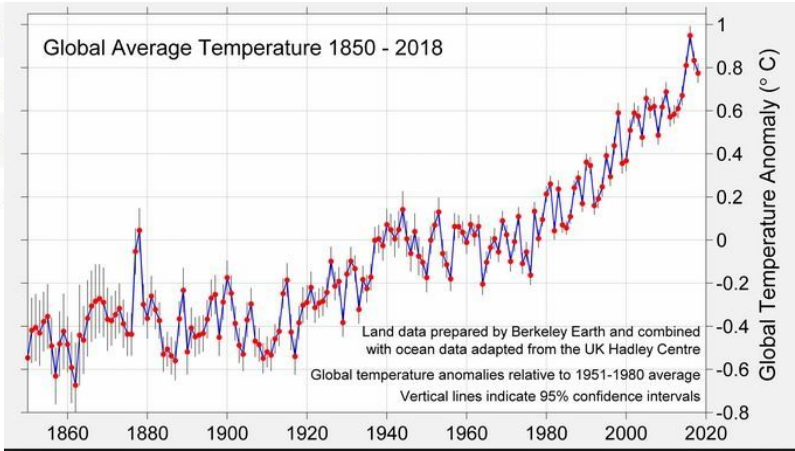


¿CÓMO SER DATA DRIVEN?



Predicción de clima

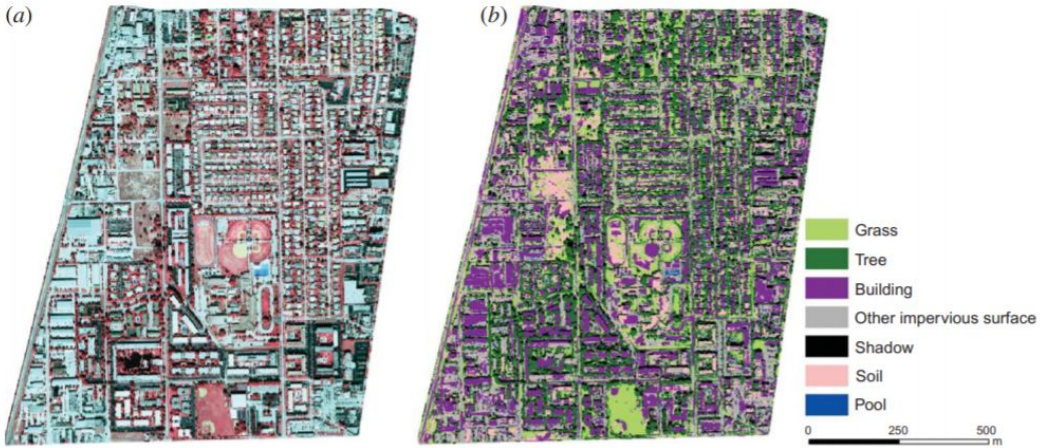
dt	AverageTemperature	AverageTemperatureUncertainty	City	Country	Latitude	Longitude
2013-05-01	14.139	0.196	Berlin	Germany	52.24N	13.14E
2013-06-01	17.473	0.236	Berlin	Germany	52.24N	13.14E
2013-07-01	20.901	0.161	Berlin	Germany	52.24N	13.14E
2013-08-01	19.335	0.265	Berlin	Germany	52.24N	13.14E
2013-09-01	NaN	NaN	Berlin	Germany	52.24N	13.14E
1824-01-01	20.116	1.370	Bogotá	Colombia	4.02N	74.73W
1824-02-01	19.797	2.109	Bogotá	Colombia		
1824-03-01	20.044	1.548	Bogotá	Colombia		
1824-04-01	19.766	1.786	Bogotá	Colombia		
1824-05-01	19.555	1.371	Bogotá	Colombia		



Identificación tipo de suelo urbano

	label	BrdIndx	Area	Round	Bright	Compact	ShpIndx	Mean_G	Mean_R	Mean_NIR	SD_G	SD_R	SD_NIR
0	car	1.27	91	0.97	231.38	1.39	1.47	207.92	241.74	244.48	21.41	20.40	18.69
1	concrete	2.36	241	1.56	216.15	2.46	2.51	187.85	229.39	231.20	6.57	6.97	7.02
2	concrete	2.12	266	1.47	232.18	2.07	2.21	206.54	244.22	245.79	6.16	4.93	5.53
3	concrete	2.42	399	1.28	230.40	2.49	2.73	204.60	243.27	243.32	5.76	5.56	5.46
4	concrete	2.15	944	1.73	193.18	2.28	4.10	165.98	205.55	208.00	11.46	8.90	9.77
5	tree	3.11	169	1.47	172.22	2.49	3						
6	car	1.20	44	0.79	208.80	1.14	1						
7	car	1.00	88	0.22	234.51	1.11	1						
8	building	1.59	1737	0.67	219.61	1.30	1						
9	tree	2.37	153	1.30	120.24	2.85	2						

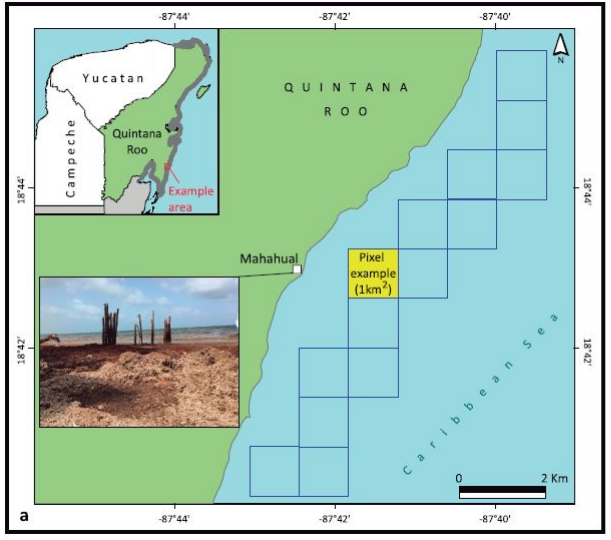
10 rows × 148 columns

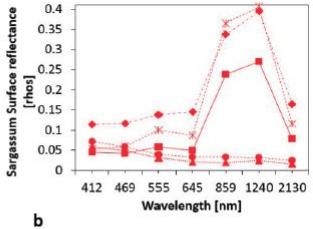


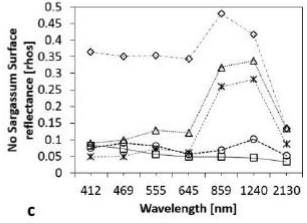
IDENTIFICACIÓN DE SARGAZO

Table 2 Mean of the rhos band. Average values of surface reflectance (rhos), at the different wavelengths (λ), used to this study. Units of the wavelengths are nanometers (nm).

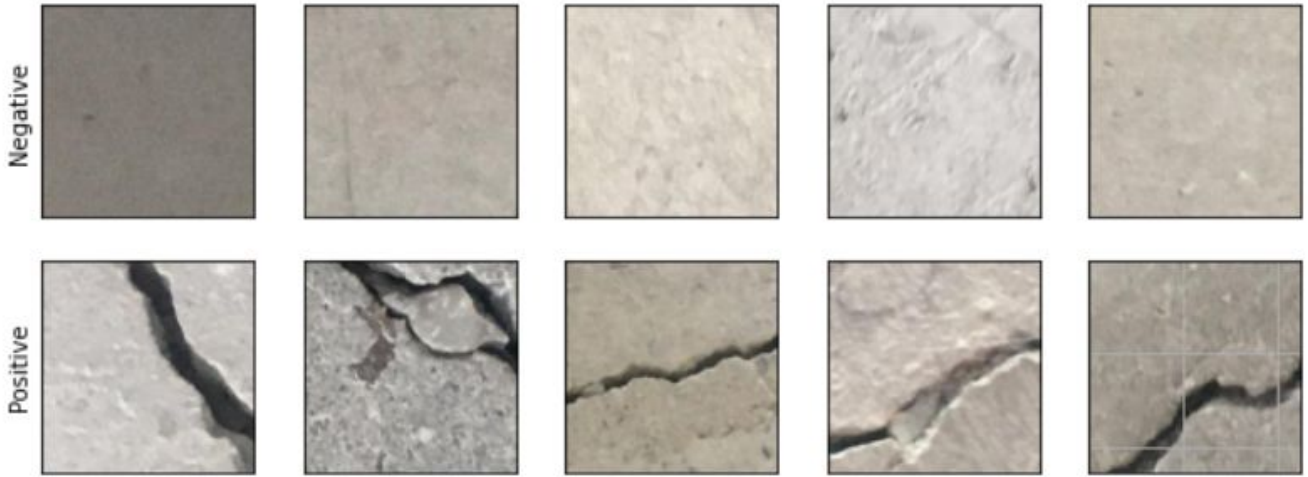
	rhos_412	rhos_469	rhos_555	rhos_645	rhos_859	rhos_1240	rhos_2130
Without *Sargassum*	0.131517	0.13489	0.141123	0.124477	0.227052	0.207291	0.085164
With *Sargassum*	0.114489	0.12090	0.133097	0.116607	**0.247237**	**0.233480**	0.084166







Detección de fracturas



Bases de datos

Open datasets:

- Kaggle https://www.kaggle.com/
- UCI Repository https://archive.ics.uci.edu/ml/index.php

Public Government & Finance, Economics Datasets for Machine Learning:

- School System Finances:

  https://catalog.data.gov/dataset/annual-survey-of-school-system-finances
- World Bank Open Data: https://data.worldbank.org/

Image Datasets for Computer Vision

- ImageNet http://www.image-net.org/
- Google’s Open Images: https://ai.googleblog.com/2016/09/introducing-open-images-dataset.html

1.3 Frameworks

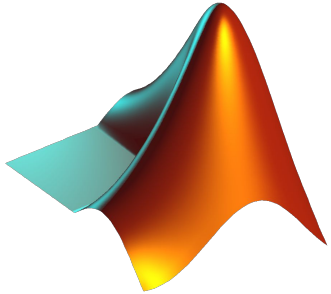
Frameworks para ML y DL.

DataRobot

scikit learn

TensorFlow + Keras

PyTorch



orange

sas

alteryx

TIBCO®

Microsoft Azure

Caffe2

IBM

H_2O.ai

Open for Innovation KNIME

rapidminer

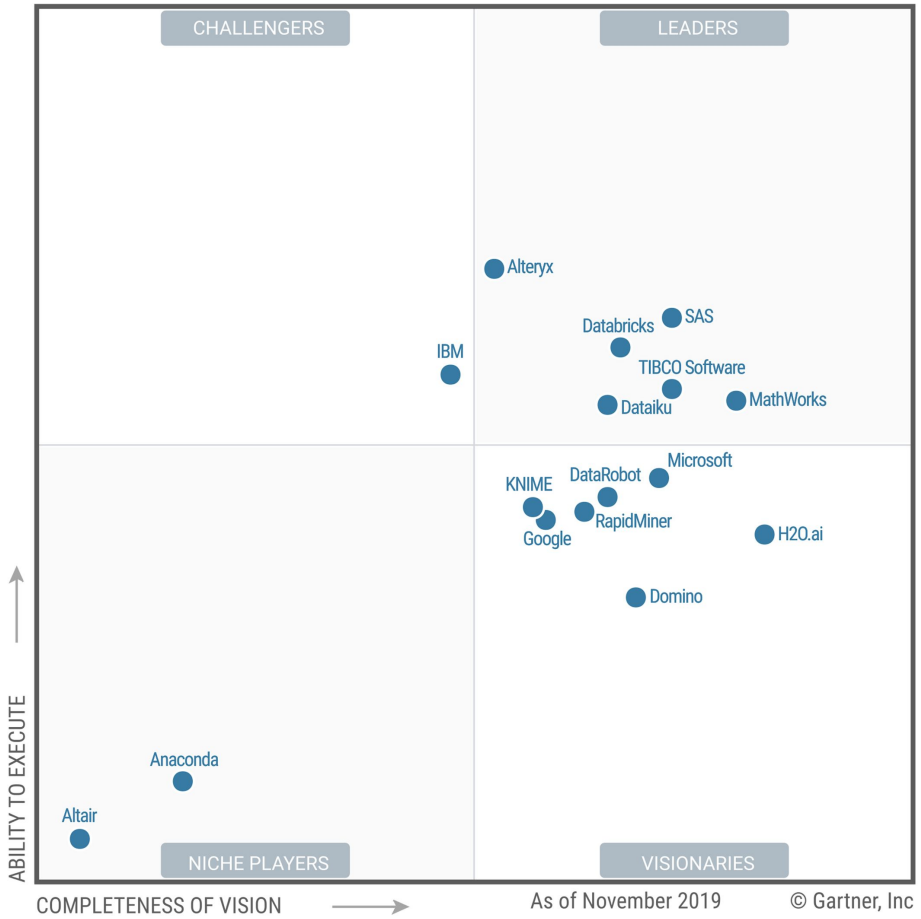
R

… más

¿CUÁLES SON LOS FRAMEWORKS MÁS USADOS?

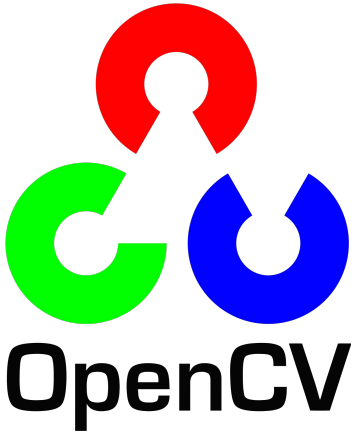
Gartner es una compañía dedicada a dar calificación a las herramientas de tecnologías de la información.

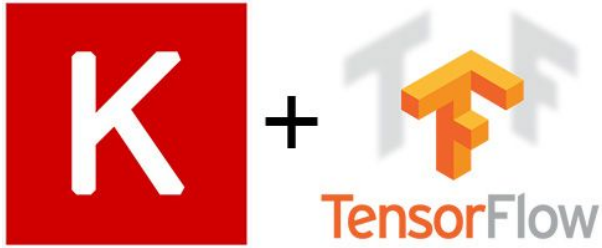
Figure 1. Magic Quadrant for Data Science and Machine Learning Platforms



Source: Gartner (February 2020)

LO QUE USAREMOS...





ACTIVIDADES DURANTE EL CURSO

- Clasificación de cultivos con imágenes satelitales.
- Predicción de temperatura y a partir de registros de calidad de aire.
- Predicción de esfuerzo en concreto.
- Clasificación de cobertura de suelos mediante imágenes satelitales.
- Clustering de sismos históricos.
- Detección de fracturas en imágenes de concreto.
- Detección de sargazo.
- Clasificador de tipos de rocas a partir de imágenes.

Propuestas de proyecto

Como se observó en los ejemplos, hay múltiples aplicaciones de problemas de regresión y clasificación, así como miles de datasets disponibles.

➔ Se deberá proponer un tema de interés personal sobre geociencias para desarrollarlo a lo largo del curso.