Introducción al *machine learning*

¿Qué NO es el *machine learning*?

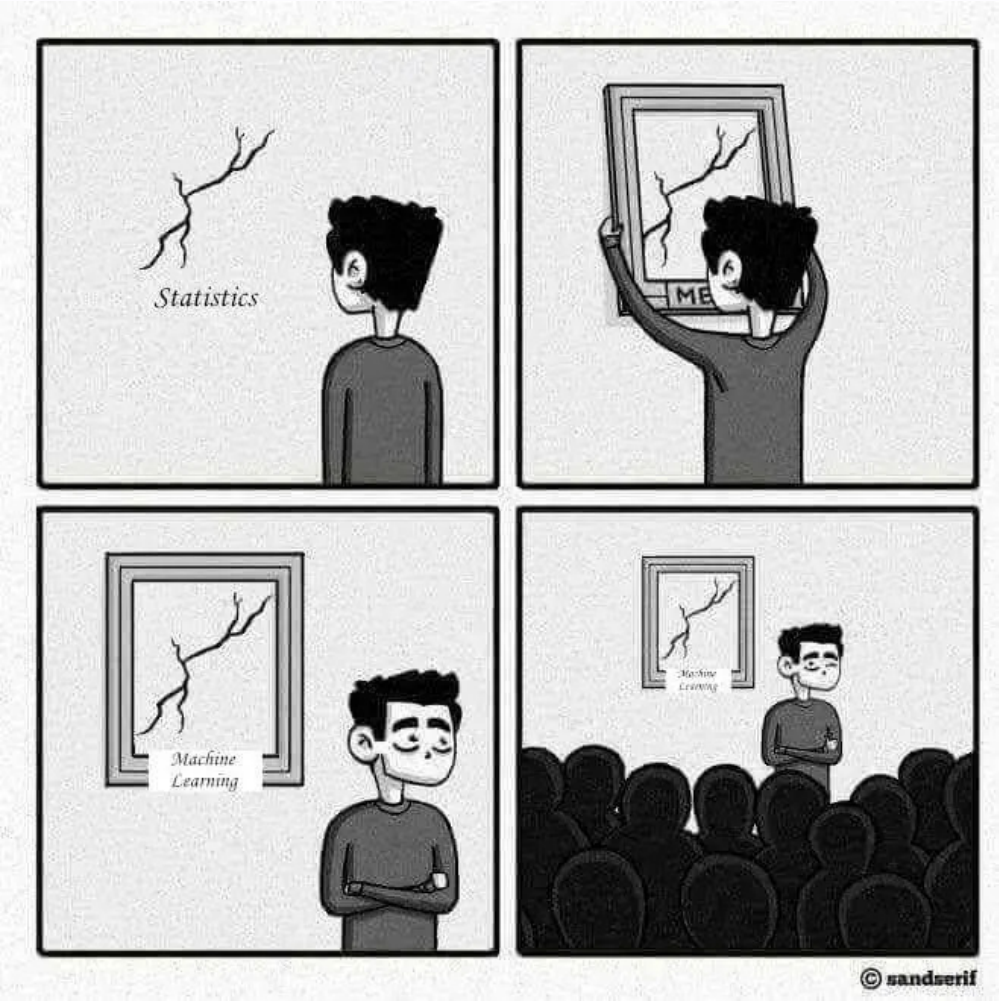
No es Hal-9000


No es T-800

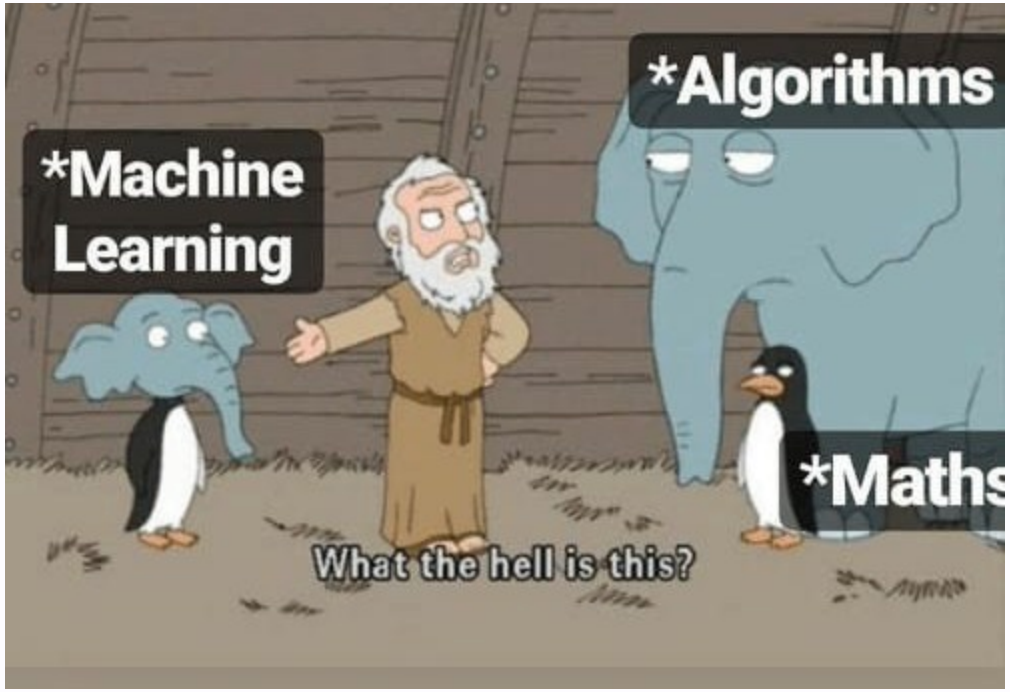
No es una alternativa a los seres humanos

¿Qué es el *machine learning*?

El *machine learning* es estadística disfrazada

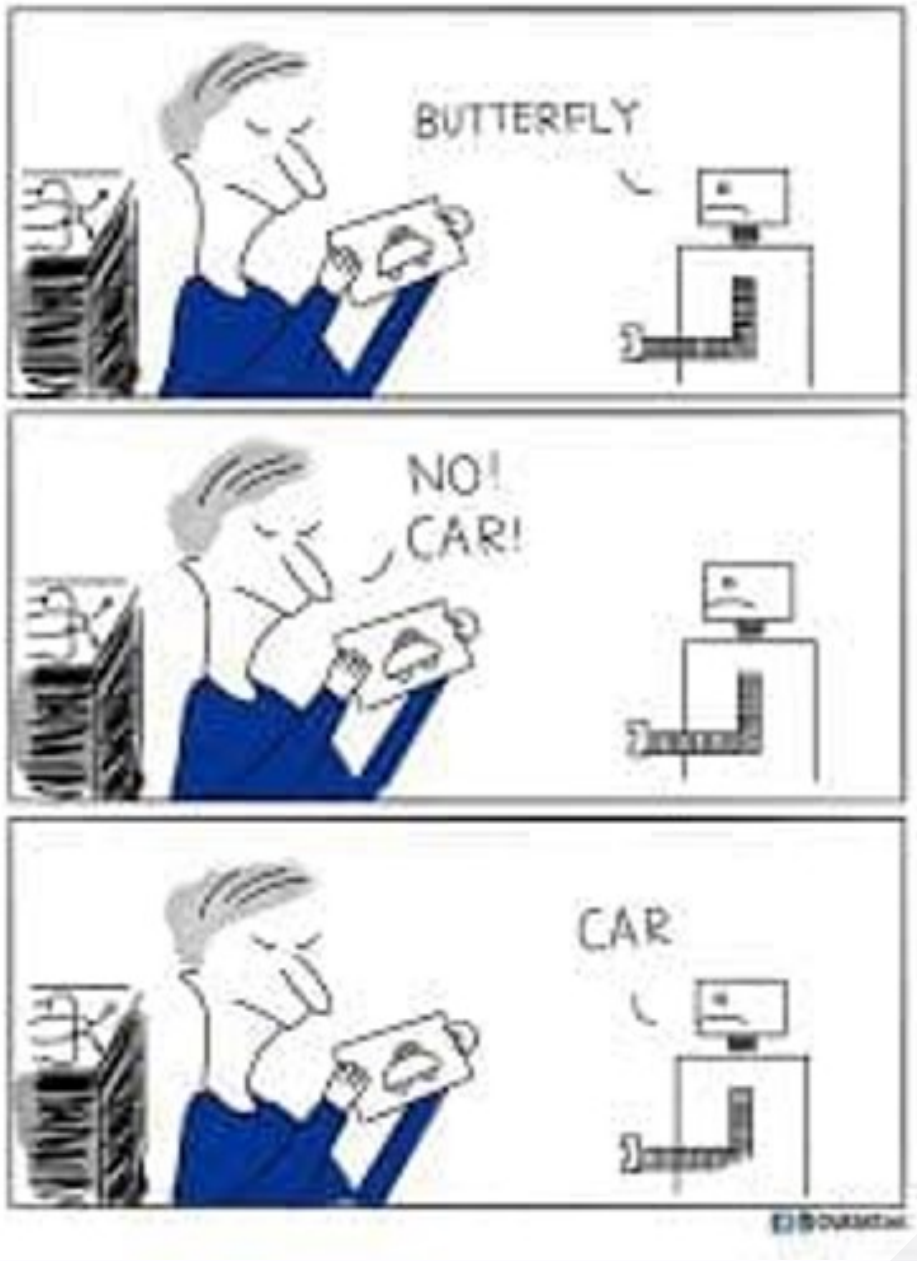


Machine learning = algorítmica + matemáticas



¿Machine learning?

Una mejor denominación para *machine learning* es *learning from data*.



1. Sin **datos** no hay **aprendizaje**
2. Sólo se **aprende** lo que está en los **datos**

¿Por qué ~~queremos~~ necesitamos el *machine learning*?


2019 This Is What Happens In An Internet Minute
facebook
1 Million
Logging In
18.1 Million
Texts Sent
YouTube
4.5 Million
Videos Viewed
390,030
Apps Downloaded
347,222
Scrolling Instagram
87,500
People Tweeting
1.4 Million
Swipes
tinder
188 Million
Emails Sent
1 Million
Views
twitch
41
Music
Streaming
Subscriptions
180
Smart Speakers
Shipped
amazon echo
4.8 Million
Gifs Served
GIPHY
41.6 Million
Messages
Sent
2.1 Million
Snaps
Created
$996,956
Spent Online
NETFLIX
694,444
Hours
Watched
3.8 Million
Search
Queries
Google
60
SECONDS
Created By:
@LoriLewis
@OfficiallyChadd

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463EB

of data will be created every day by 2025

IDC

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

95m

photos and videos are shared on Instagram

Instagram Business

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020

294bn

billion emails are sent

Radicati Group

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

3.9bn

people use emails

4TB

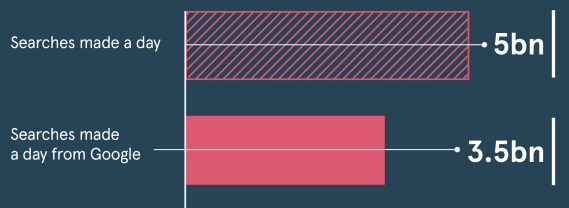
of data produced by a connected car

Intel

28PB

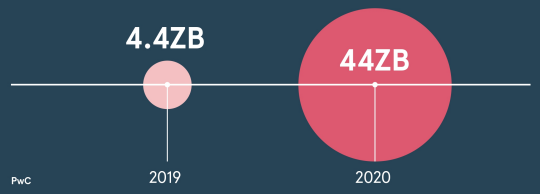
to be generated from wearable devices by 2020

Statista



Smart insights

ACCUMULATED DIGITAL UNIVERSE OF DATA



PwC

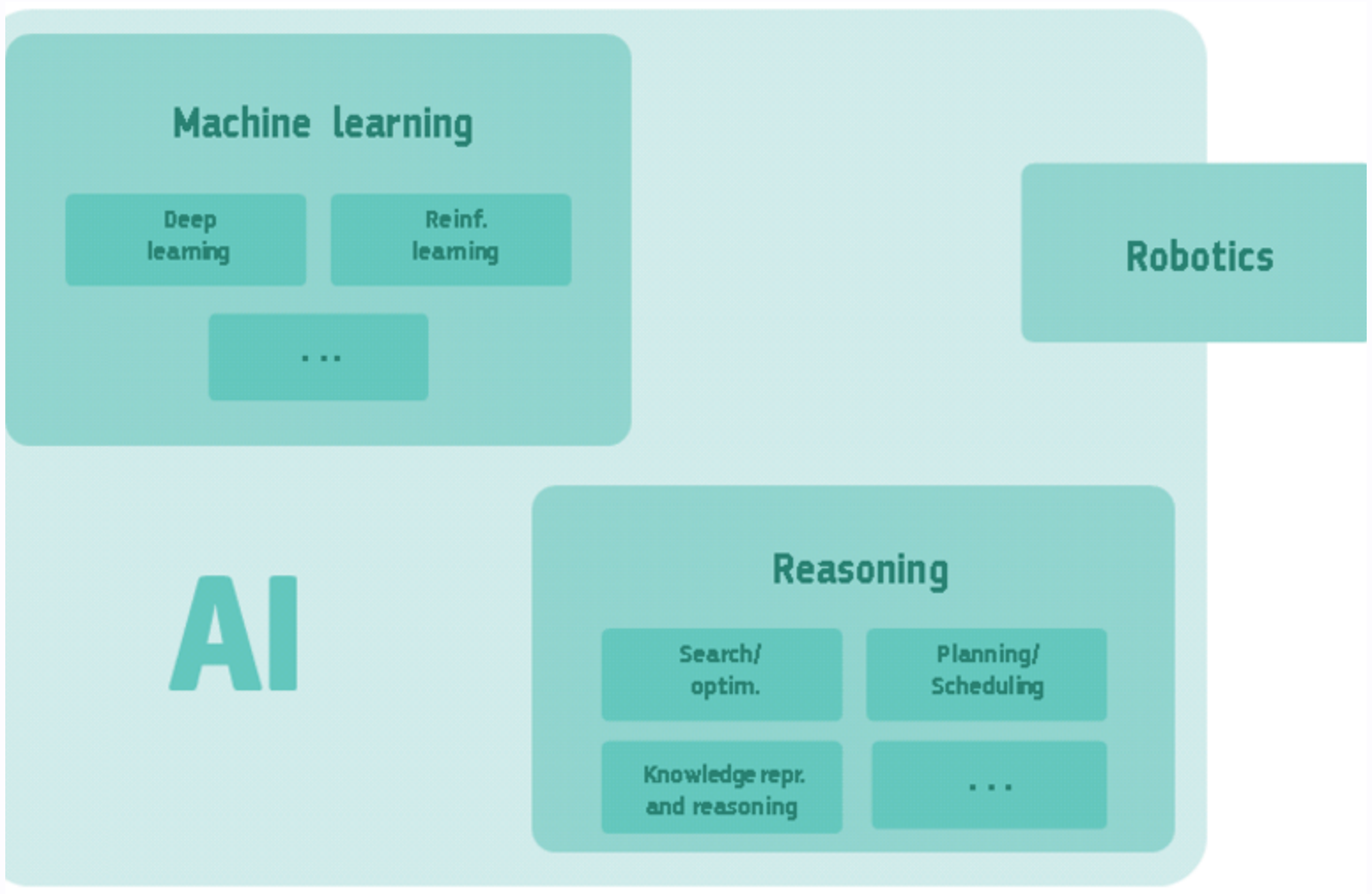
¿Por qué?

- Necesitamos obtener conocimiento de grandes volúmenes de datos
- Los ordenadores son muy buenos calculando:
  - Velocidad de ejecución
  - No incorporan sesgos
  - No se cansan, trabajando 24/7
  - Precisión en los cálculos

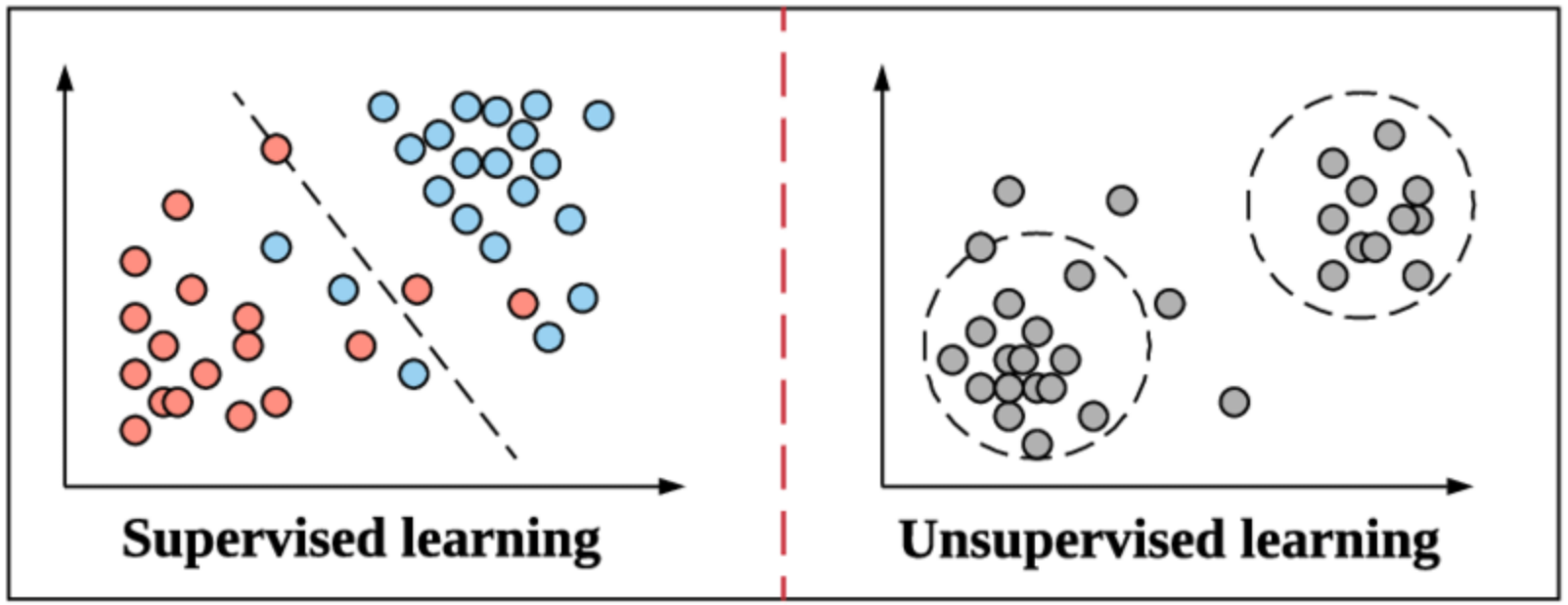
Algunas aplicaciones

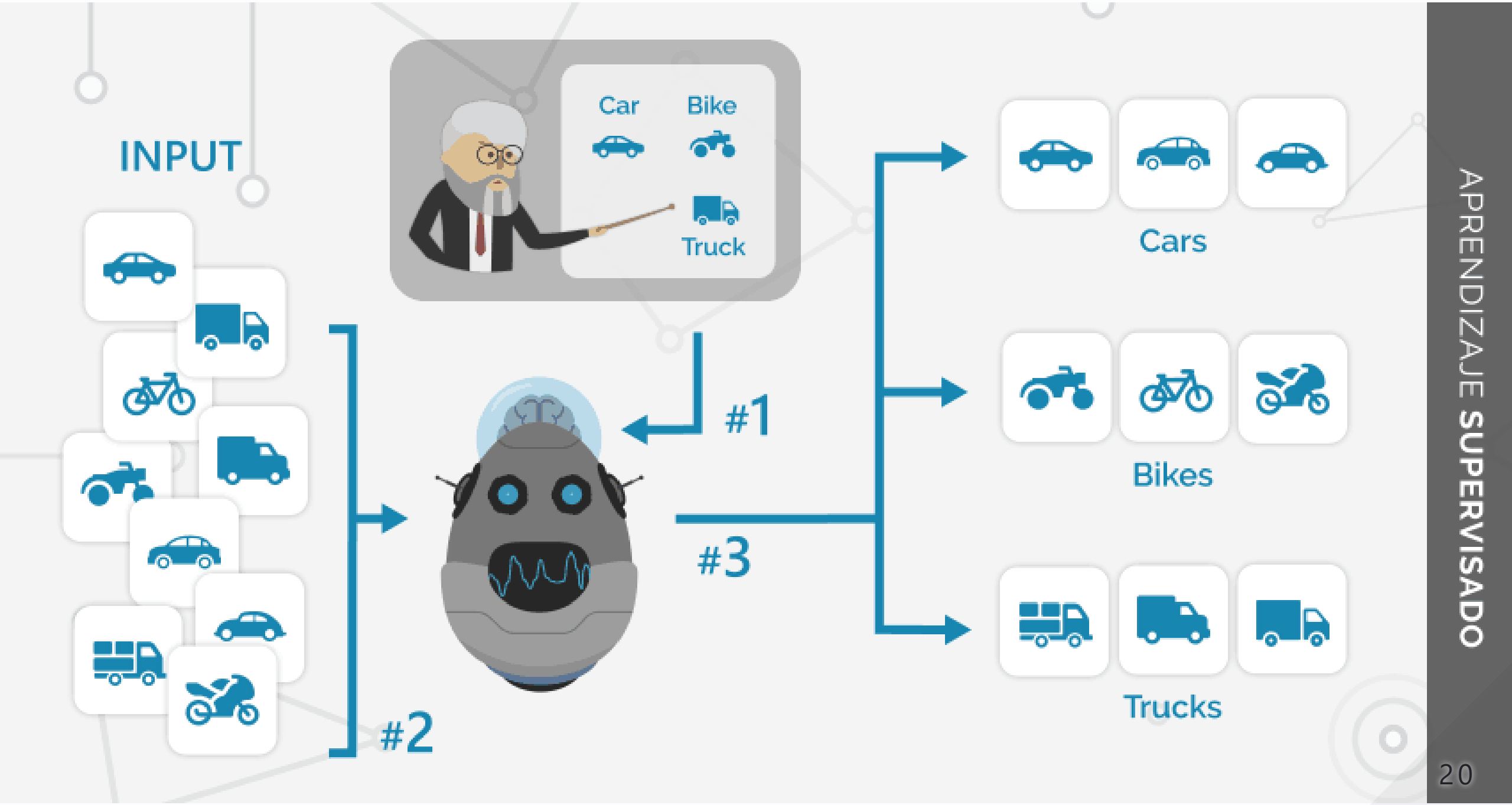
- Visión por computador
  - Conducción autónoma, detección de cáncer, ...
- Reconocimiento del habla
  - Siri, Alexa, OK Google, ...
- Procesamiento del lenguaje natural
  - Traducción automática, chat bots, ...
- Detección de patrones
  - Sistemas de recomendación, detección de fraude, análisis de comportamiento en webs, ...

El *machine learning* es una disciplina de la **inteligencia artificial**



Dos enfoques:




INPUT
Car
Bike
Truck
#1
#2
#3
Cars
Bikes
Trucks

INPUT

"OK, let me try..."

#1

#2

Medium size, 4 wheels, glass...

Small size, 2 wheels...

Big size, 4+ wheels, glass...

Classical Machine Learning

Supervised Learning

(Pre Categorized Data)

Classification

(Divide the socks by Color)

Eg. Identity Fraud Detection

Regression

(Divide the Ties by Length)

Eg. Market Forecasting

Unsupervised Learning

(Unlabelled Data)

Clustering

(Divide by Similarity)

Eg. Targeted Marketing

Association

(Identify Sequences)

Eg. Customer Recommendation

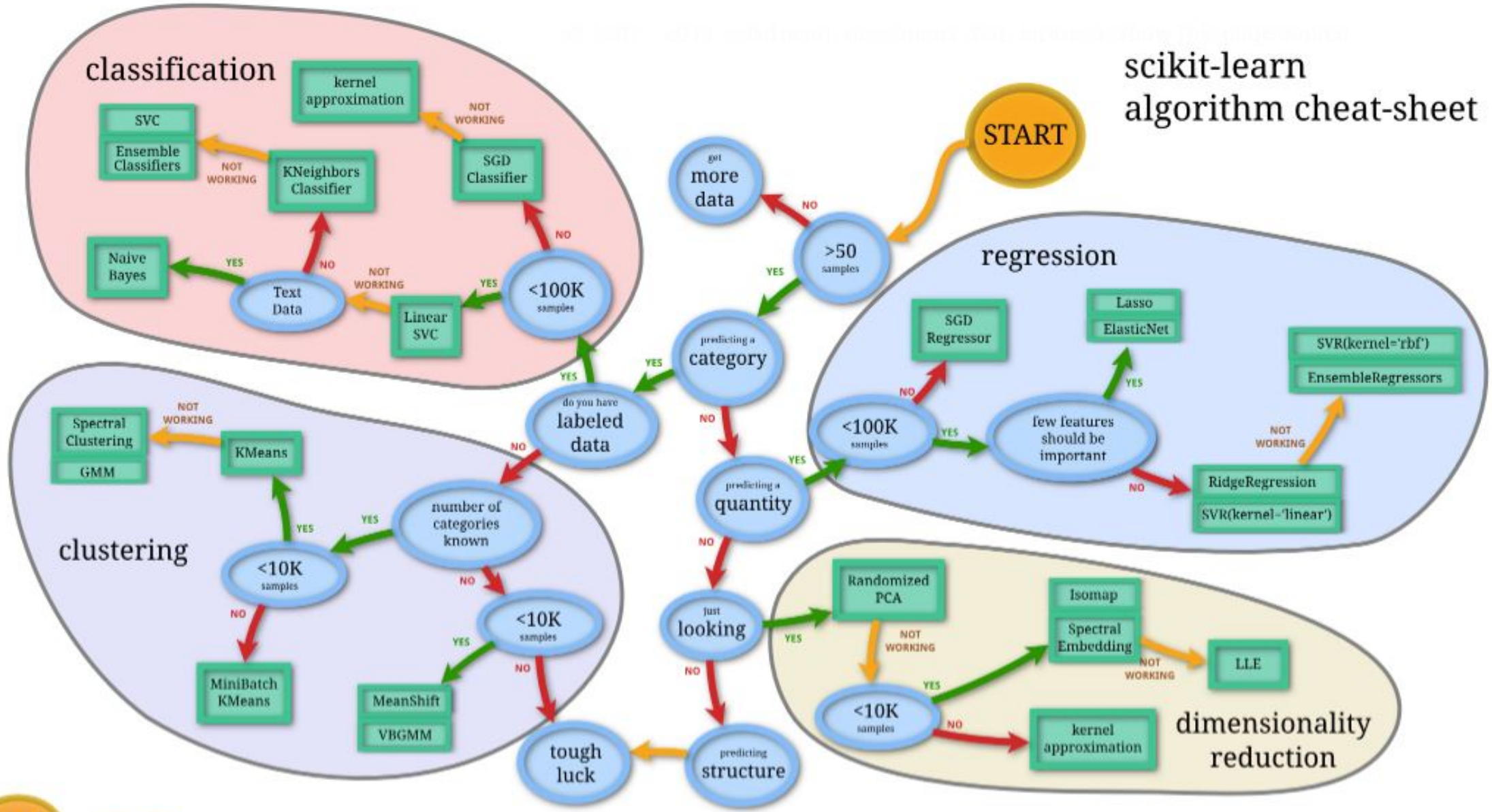
Dimensionality Reduction

(Wider Dependencies)

Eg. Big Data Visualization

Obj: Predications & Predictive Models

Pattern/ Structure Recognition


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
got more data
NO
>50 samples
YES
predicting a category
YES
NO
do you have labeled data
YES
NO
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
NO
RidgeRegression
SVR(kernel='linear')
predicting a quantity
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
YES
number of categories known
NO
<10K samples
NO
<10K samples
YES
NO
MiniBatch KMeans
MeanShift
VBGMM
just looking
YES
NO
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
predicting structure
tough luck
Back
scikit learn

Proyectos *machine learning*

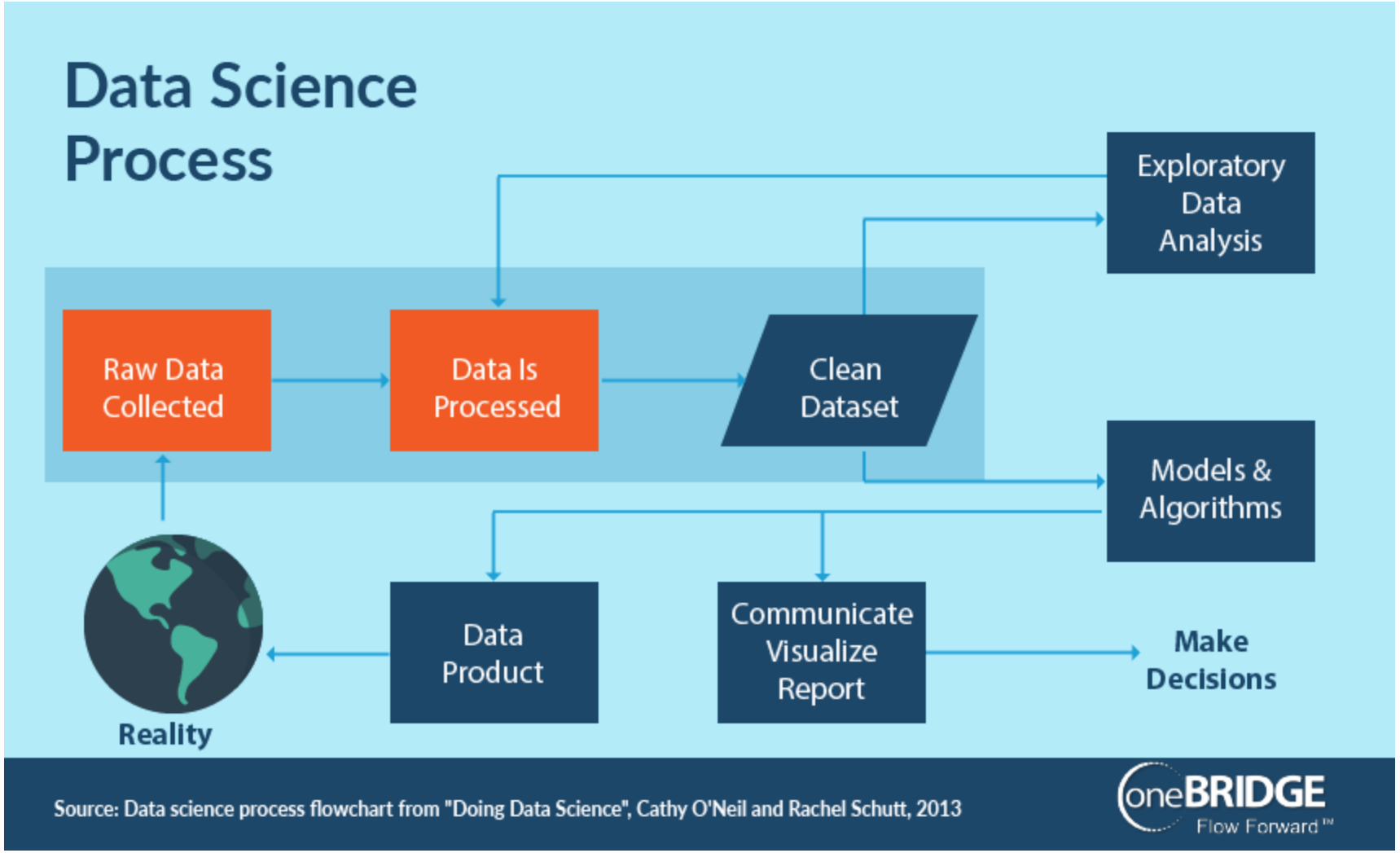
Nomenclatura de un *dataset*:

Feature 1	**Feature 2**	**...**	**Feature N**	**Label**
Observation 1,1	Observation 1,2	...	Observation 1,N	Label 1
Observation 2,1	Observation 2,2	...	Observation 2,N	Label 2
...	...	...	...	...
Observation M,1	Observation M,2	...	Observation M,N	Label M

Dataset de ejemplo:

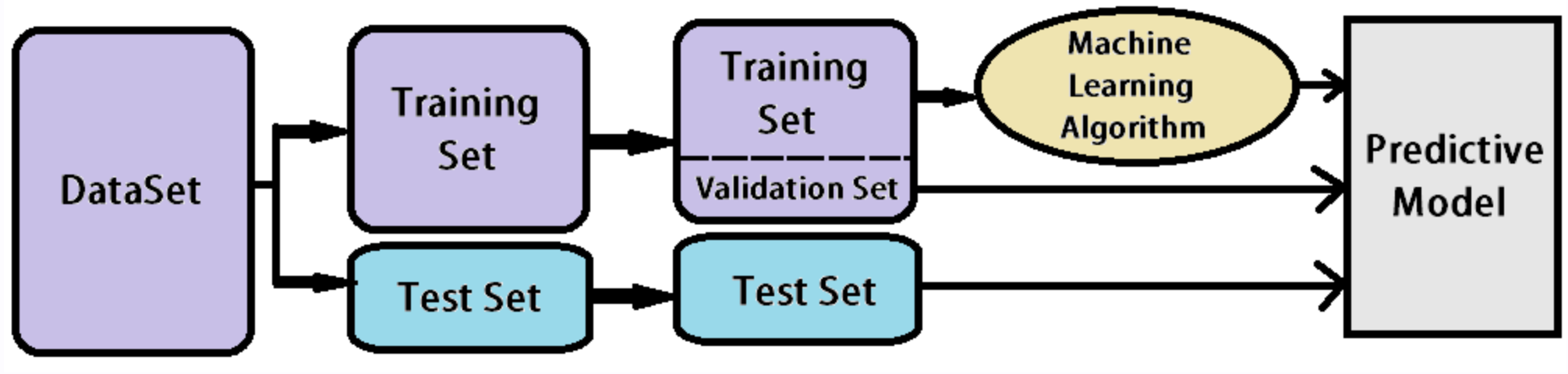
Latitude	**Longitud**	**Timestamp**	**Temperature**
40,3214873	-3,8123123	1587980647	18,3
18,9230112	15,2394502	1587984567	6,4
...	...	...	...
-7,2321231	10,7234433	1587968742	25,7

El proceso de la ciencia de datos



Machine learning paso a paso:

Proceso de evaluación:



Conseguir el modelo perfecto: