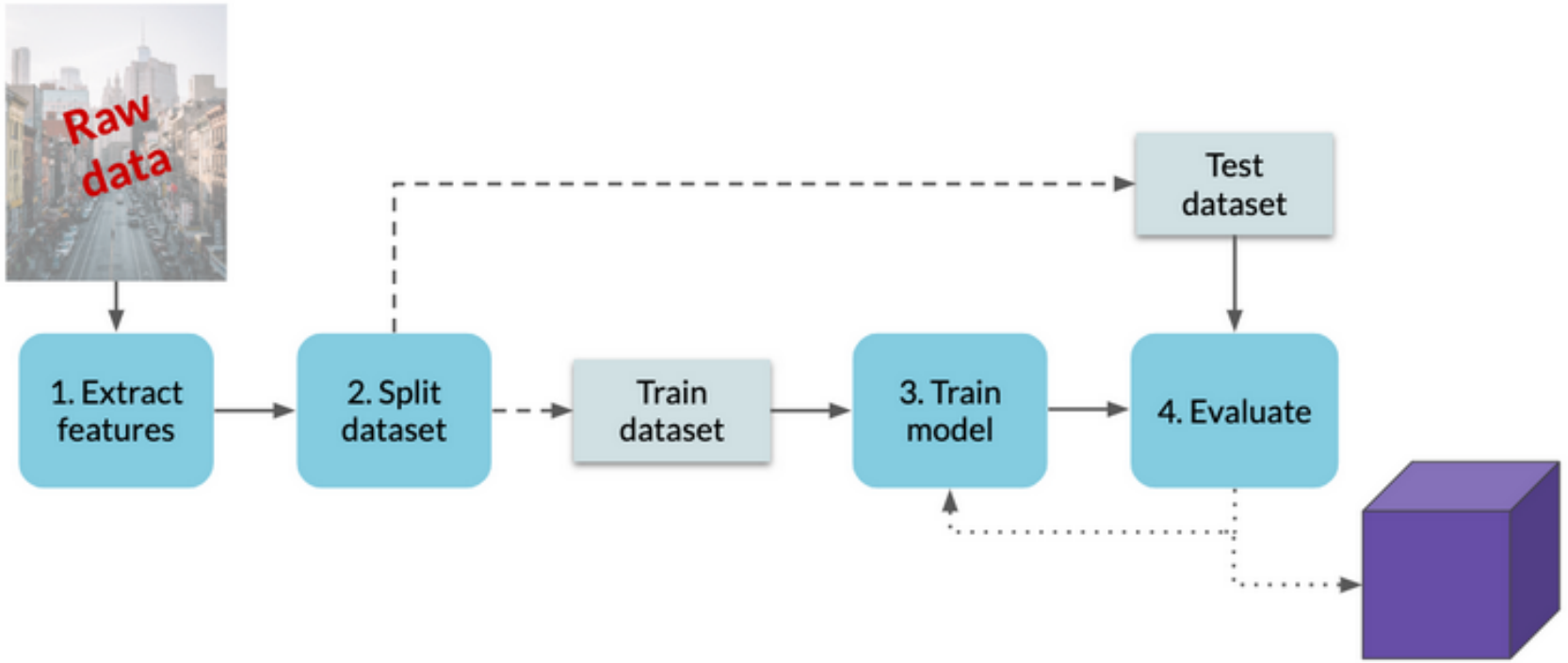
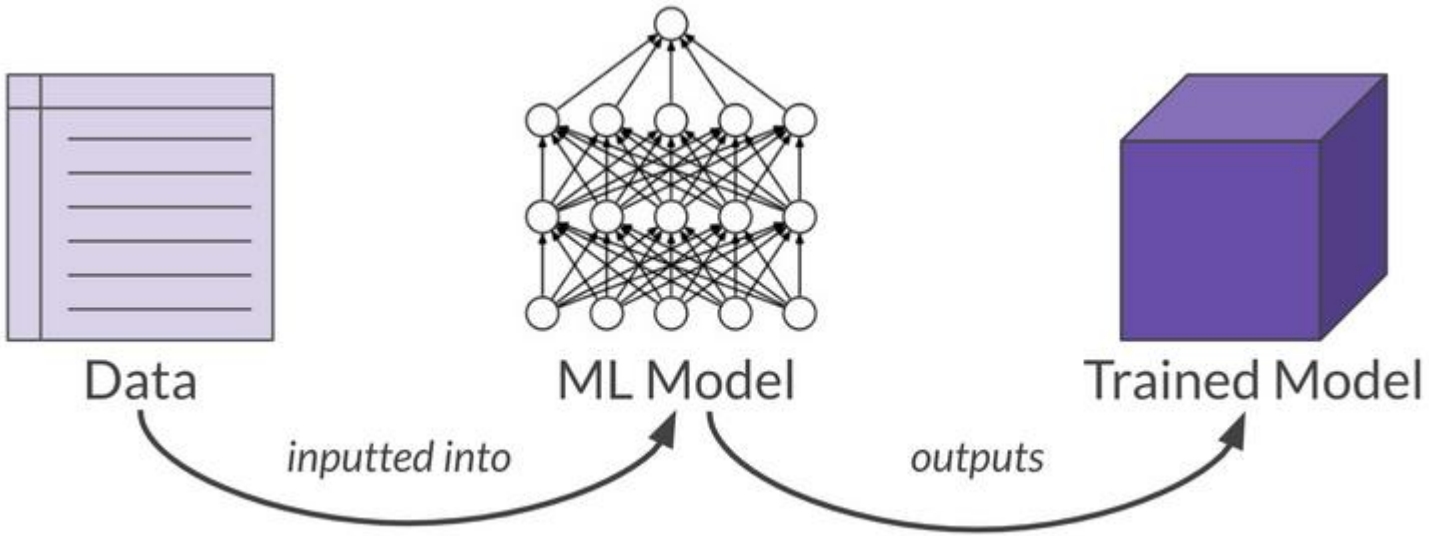
ML

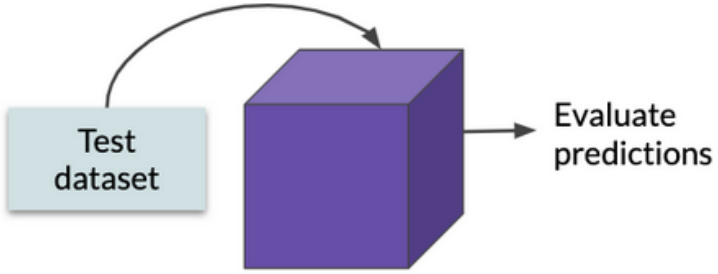
Fases de Machine Learning

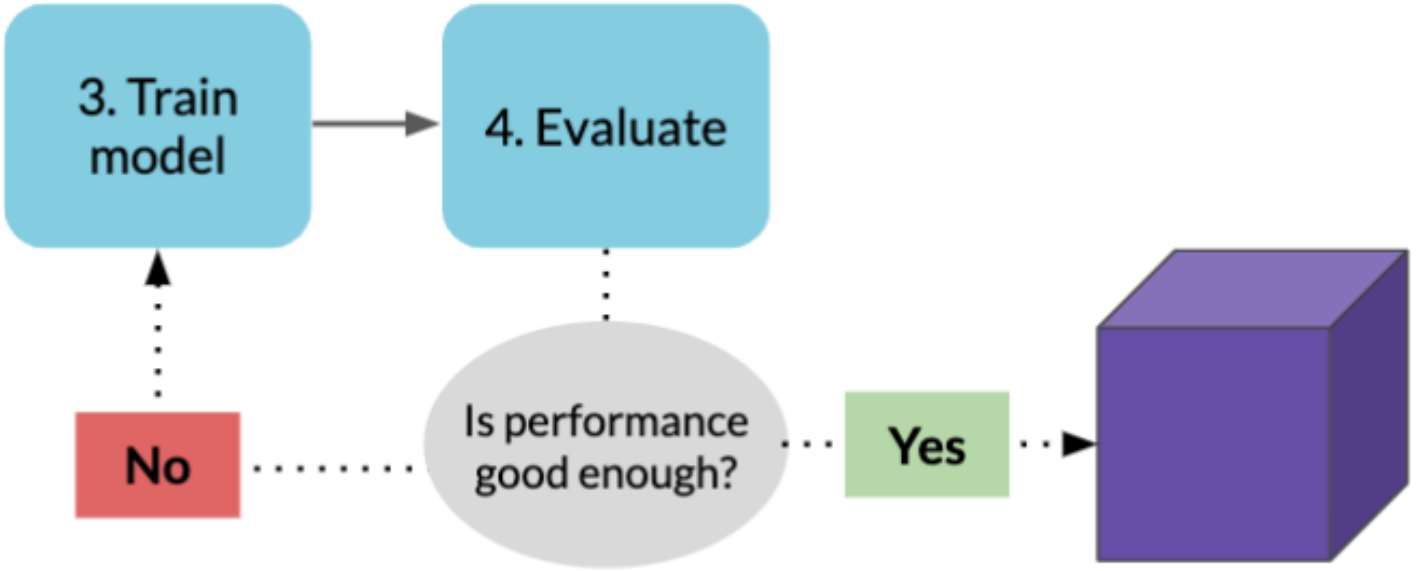


Train Model

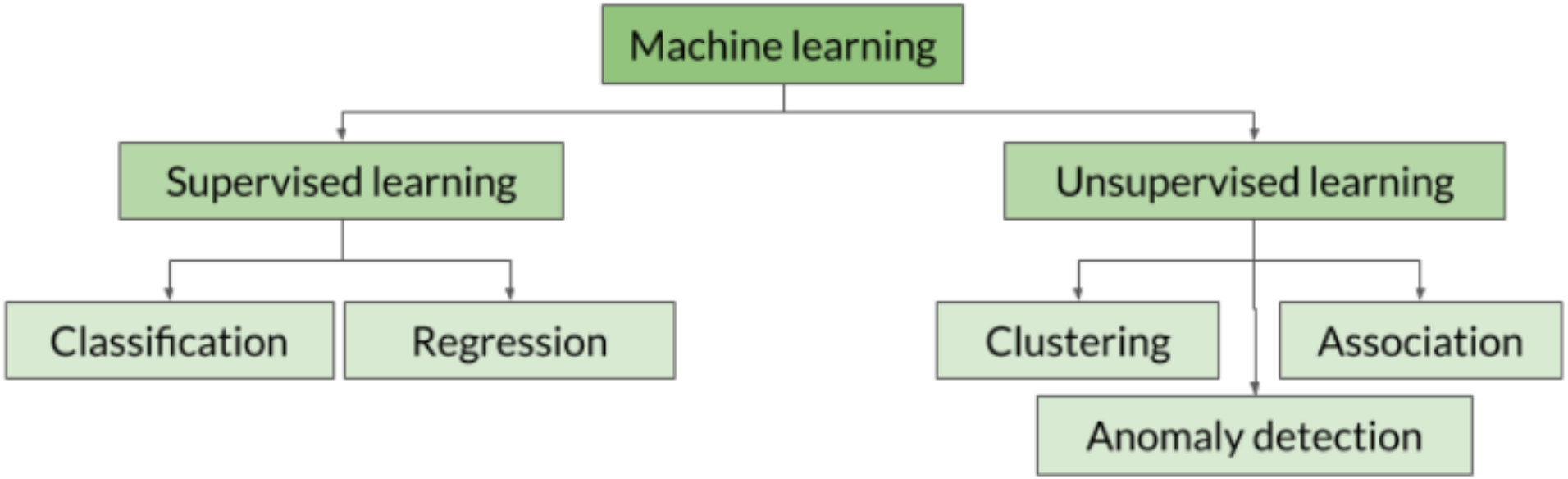


Evaluar la predicción





SUPERVISADO



SUPERVISADO

Supervisados

- Toma una observación y lo etiqueta
- Hay dos tipos de aprendizaje supervisado: clasificación y regresión.

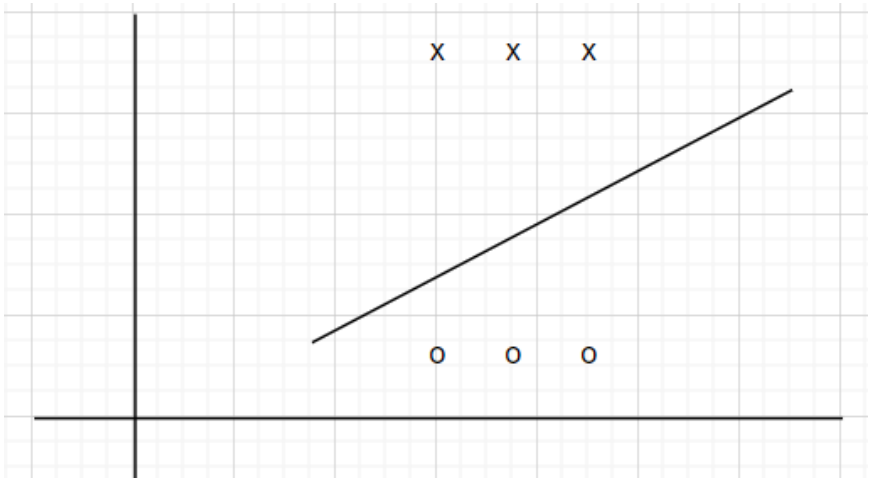
SUPERVISADO

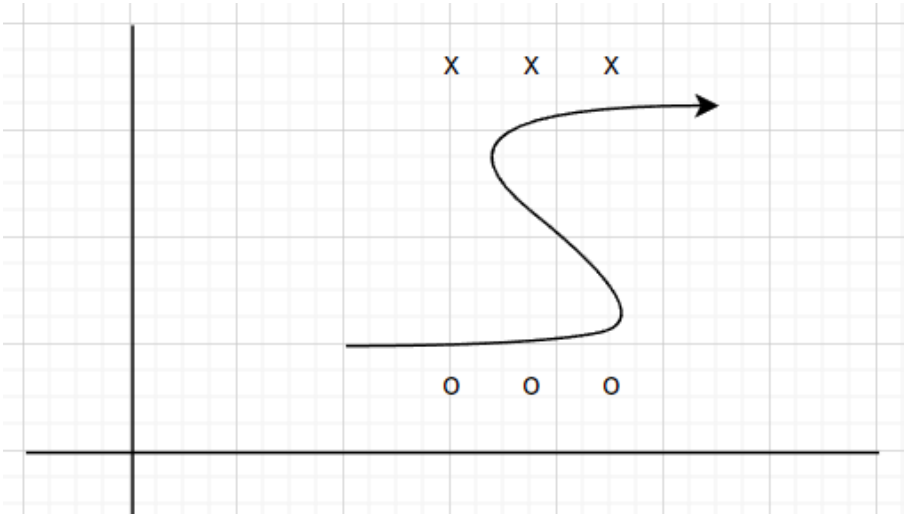
Clasificación

- La clasificación consiste en asignar una categoría a una observación. Es predecir una variable discreta, una variable que solo puede tomar unos pocos valores diferentes.
  - ¿Este cliente va a cancelar su suscripción o no?
  - ¿Este lunar es canceroso o no?
  - ¿Este vino es tinto, blanco o rosado?
  - ¿Es esta flor una rosa, un tulipán, un clavel, un lirio?

SUPERVISADO

Para la **clasificación RL** no funciona





y = ax + b

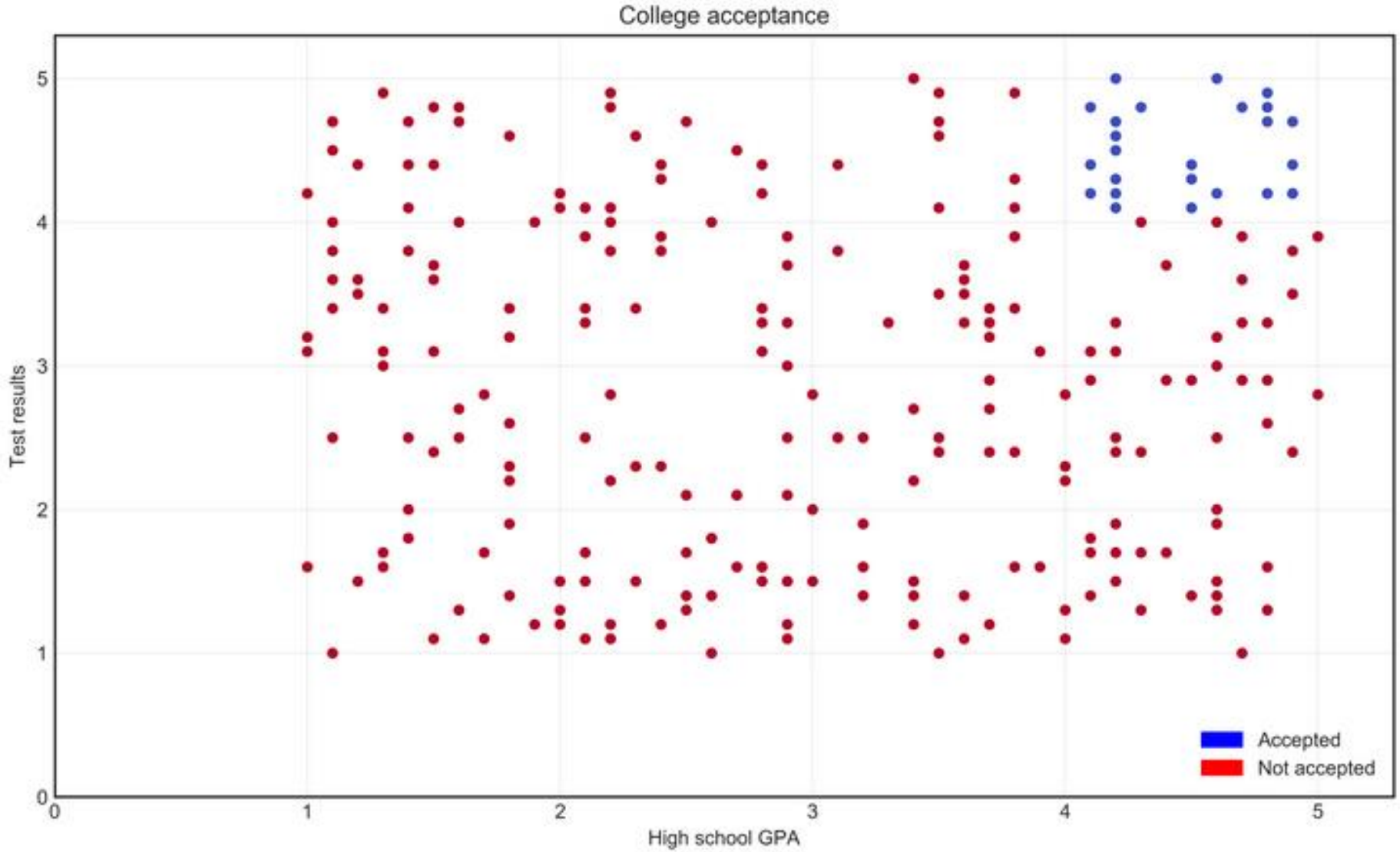
z = ax + by + c

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Función sigmoide o de Activación

Algoritmo de Regresión Logística

SUPERVISADO



SUPERVISADO

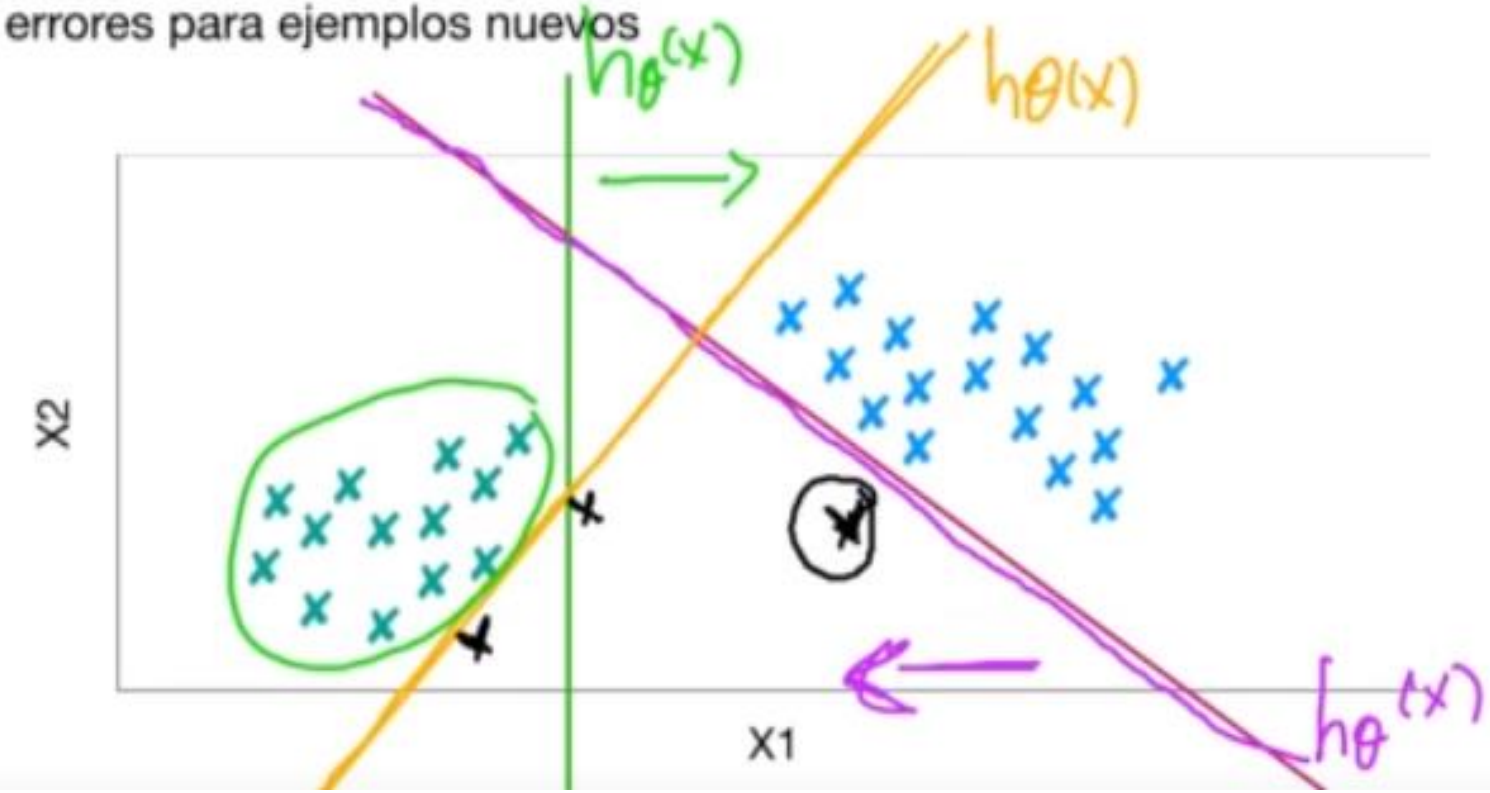
Support Vector Machine

Algoritmo de Clasificación

SUPERVISADO

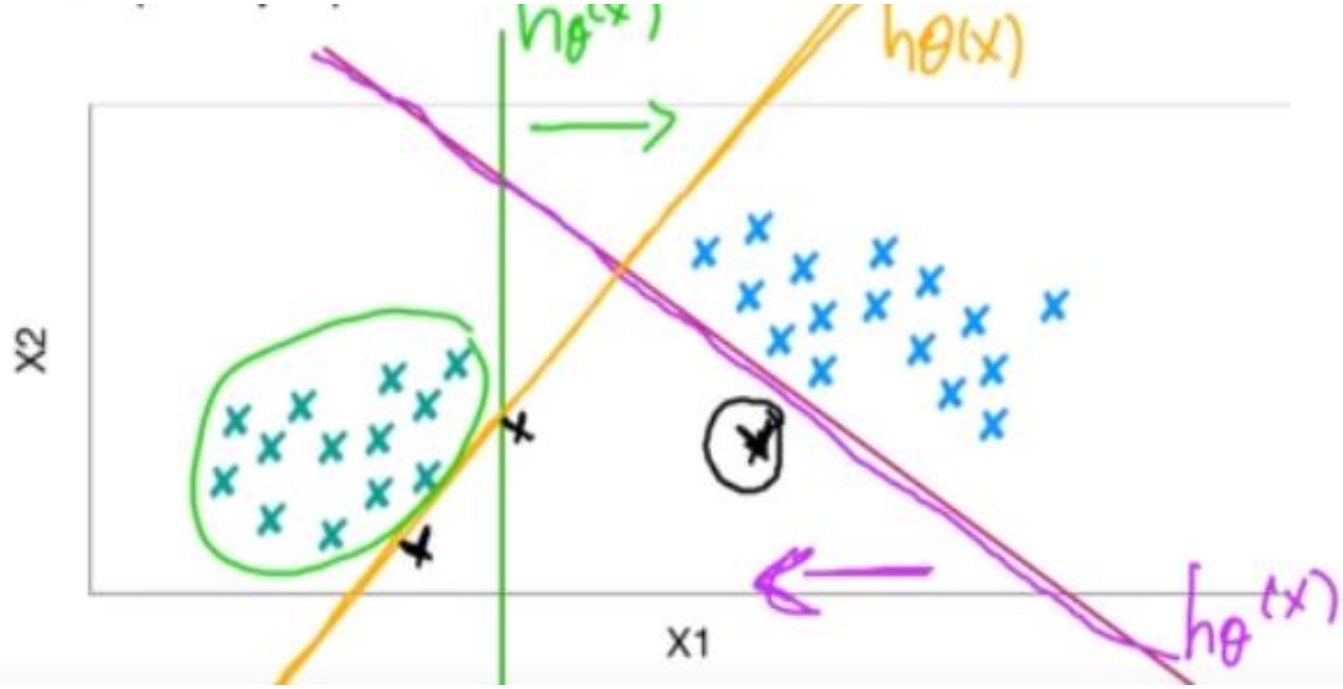
Funcionamiento del algoritmo

- Los modelos representados funcionan bien para el conjunto de datos de entrenamiento, pero el límite de decisión se encuentra tan cerca de algunos ejemplos que es probable que comentan errores para ejemplos nuevos



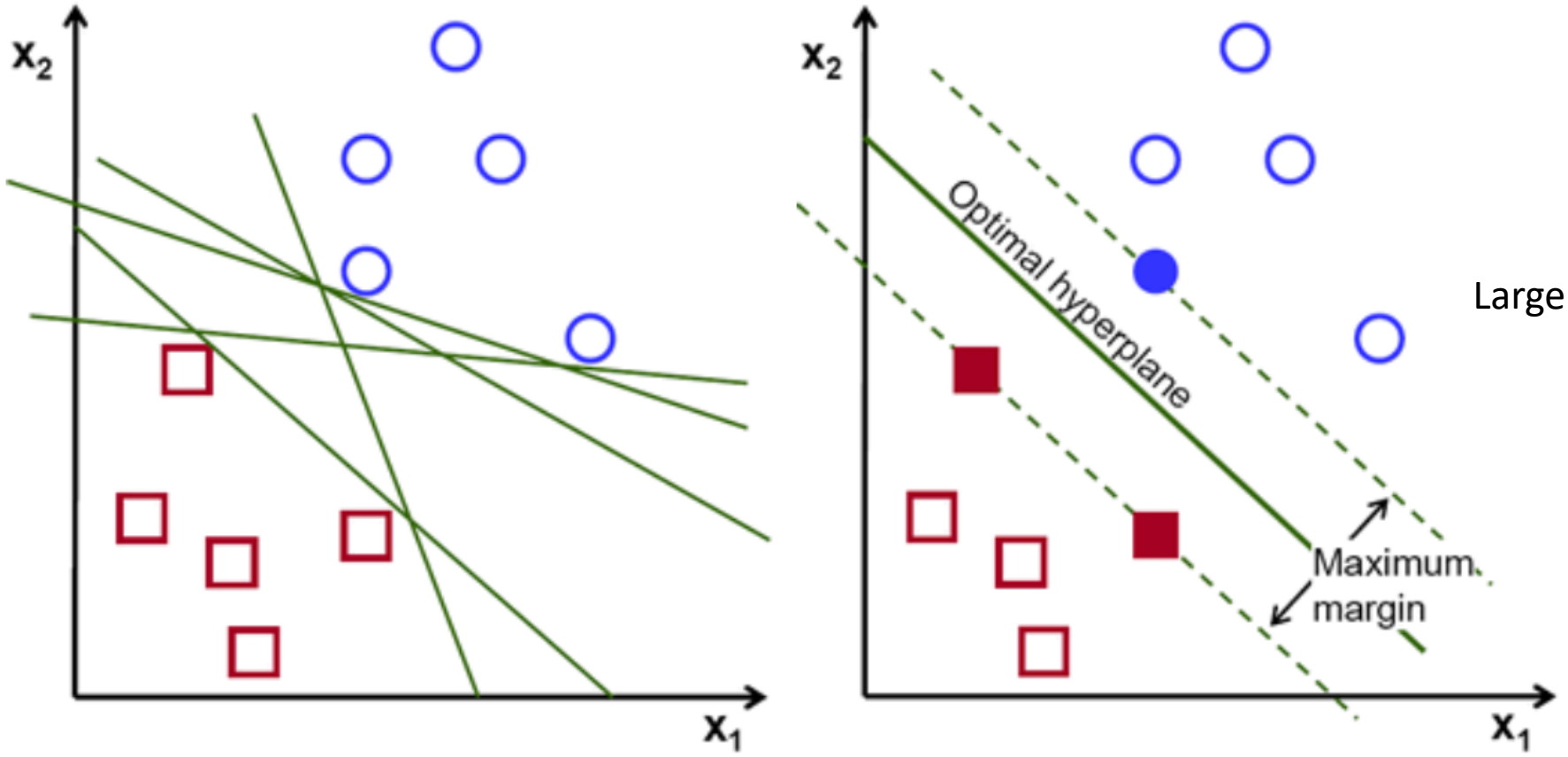
SUPERVISADO

Los modelos representados funcionan bien para el conjunto de datos de entrenamiento, pero el límite de decisión se encuentra tan cerca de algunos ejemplos que es probable que comentan errores para ejemplos nuevos.



SUPERVISADO

Nuestro objetivo es encontrar un plano que tenga el margen máximo, es decir, la distancia máxima entre puntos de datos de ambas clases.

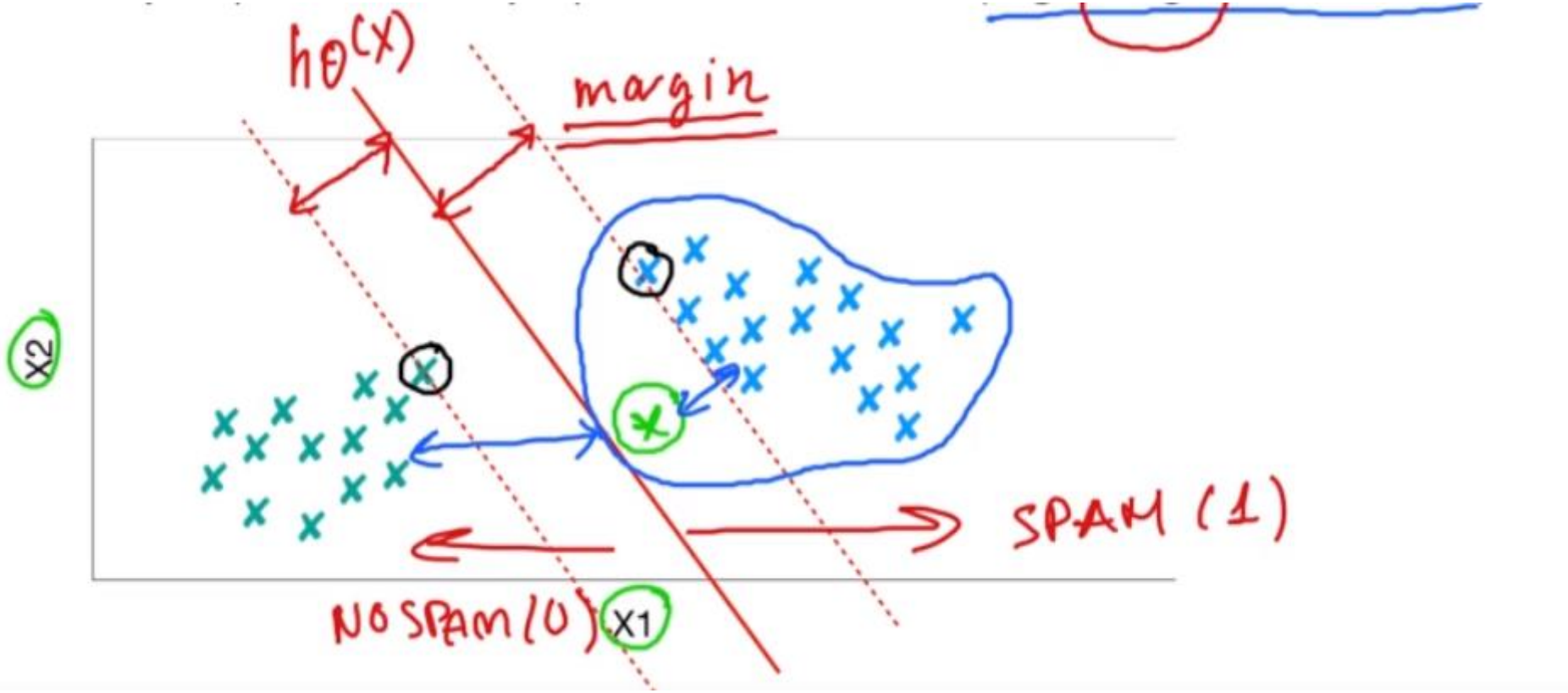


Large margin classification

Posibles hiperplanos

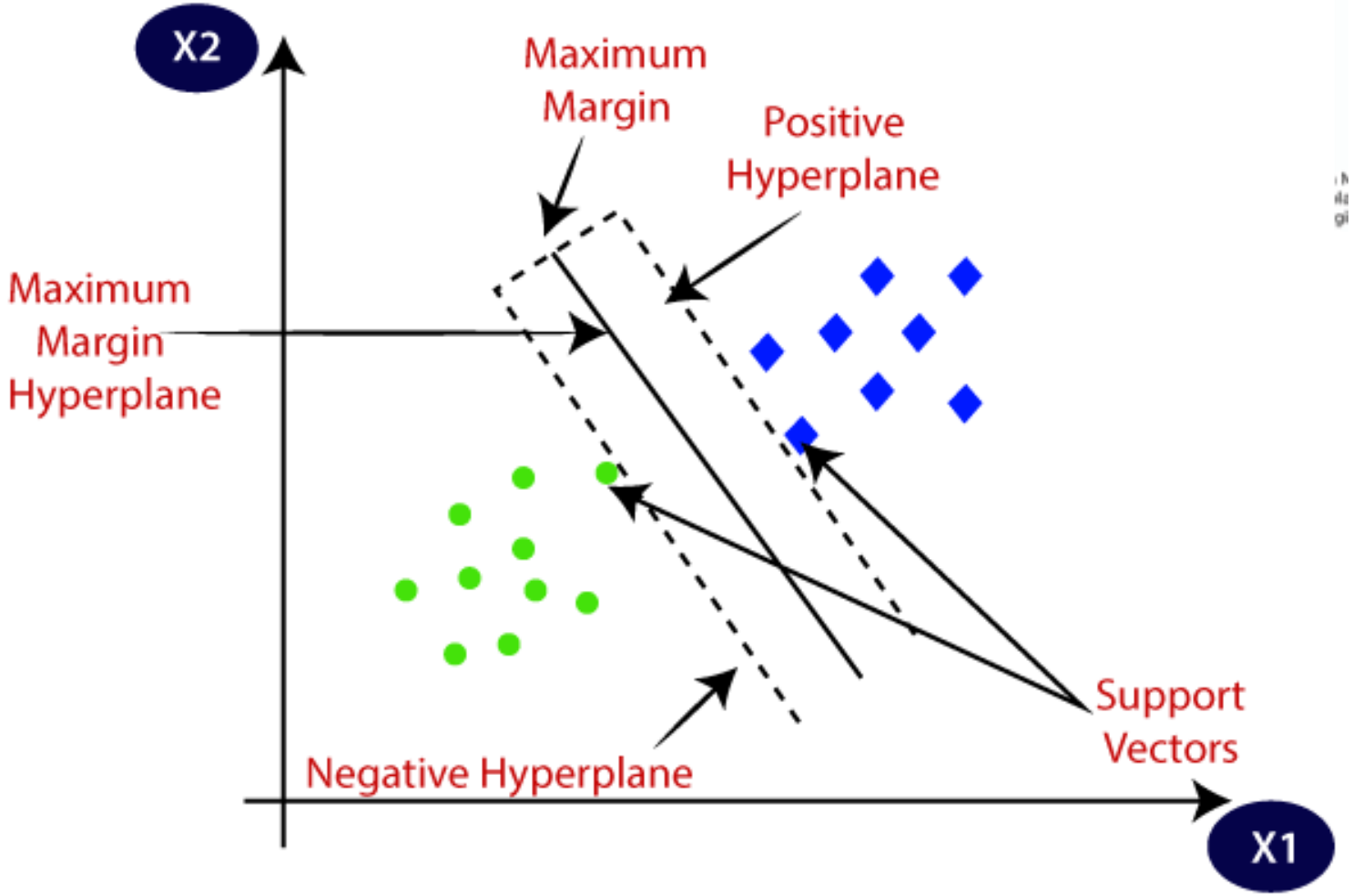
SUPERVISADO

SUPPORT VECTOR MACHINE



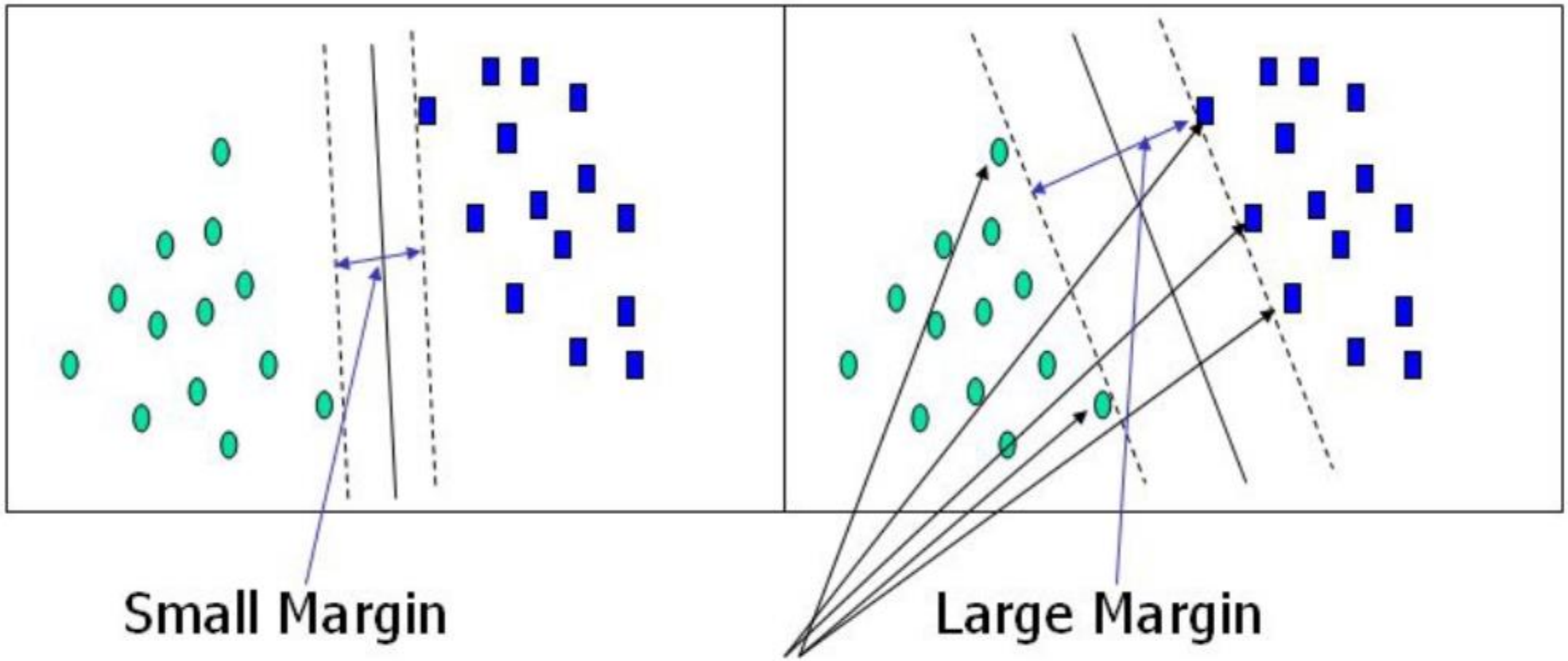
El objetivo del algoritmo es maximizar la distancia del límite de decisión desde los puntos de datos

SUPERVISADO



clasificación como de regresión

SUPERVISADO



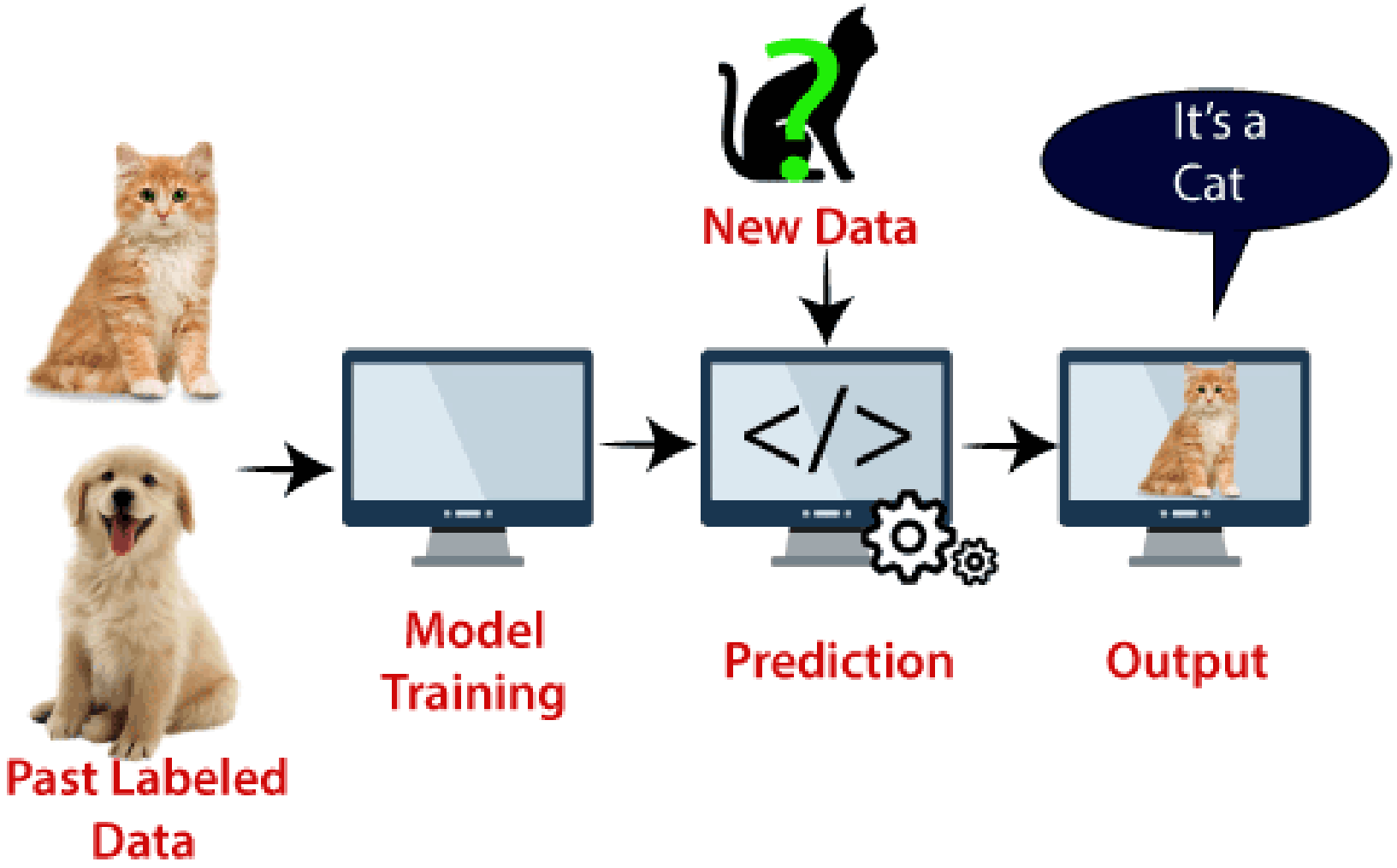
Vectores de apoyo

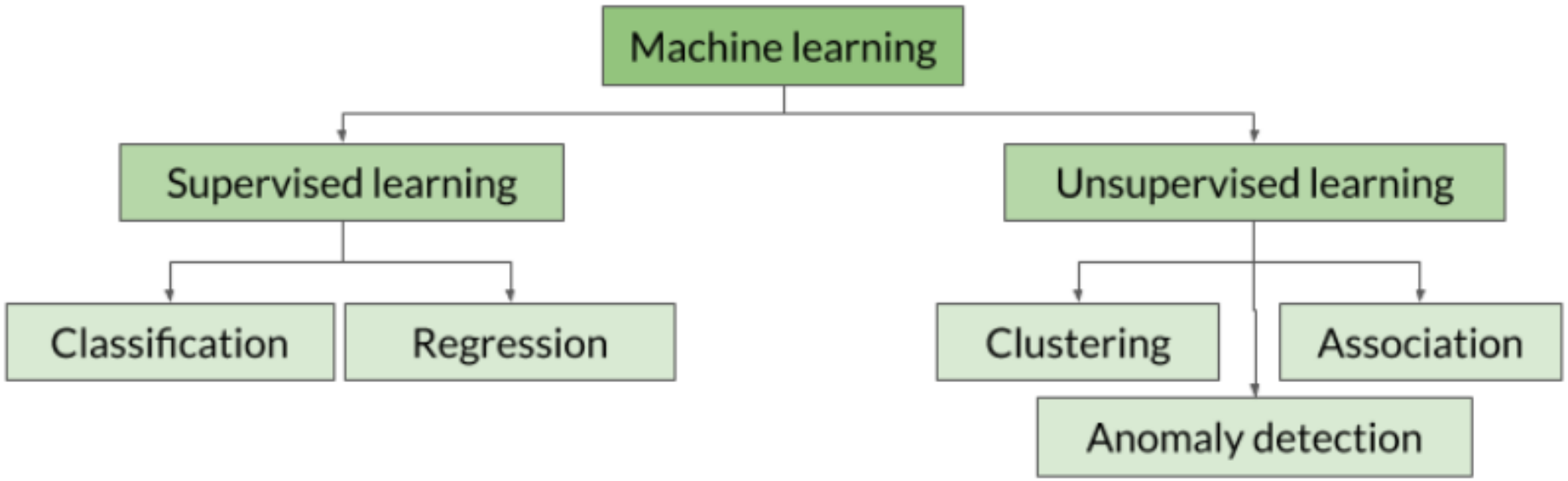
SUPERVISADO

SVM Características

- Algoritmo Supervisado con conjunto de datos etiquetados
- Producirá valores discretos y continuos
- Es un algoritmo capaz de realizar regresión y **clasificación** (lineal o no lineal)
- Funciona bien con datos complejos de tamaño pequeño o mediano
- Puede aplicarse de diferentes manera en función del conjunto de datos:
  - Linealmente separables
  - Linealmente no separables

SUPERVISADO




Machine learning
Supervised learning
Unsupervised learning
Classification
Regression
Clustering
Association
Anomaly detection

NO SUPERVISADO

No supervisados

- Idéntico al supervisado pero no tiene un objetivo específico, un destino.
- Aprende del conjunto de datos e intenta encontrar patrones
- Podemos encontrar información sin saber mucho de los datos
- Ejemplos: Detección de anomalías, asociaciones, tendencias, fidelizaciones.

NO SUPERVISADO

Clusters, grupos

NO SUPERVISADO

Ejemplo Clustering

White Swiss Shepherd

Brown Japanese Bobtail

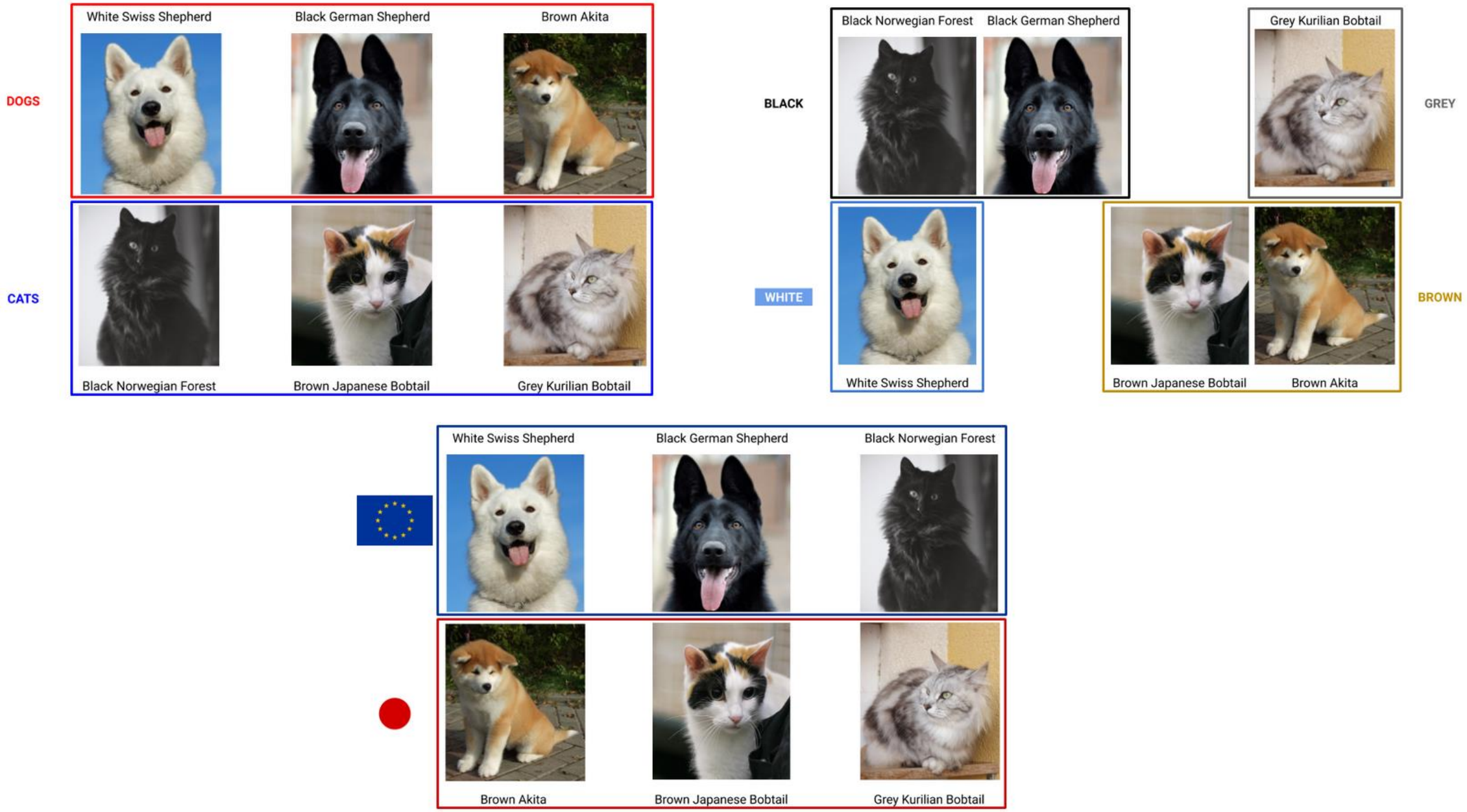
Brown Akita

Black Norwegian Forest

Black German Shepherd

Grey Kurilian Bobtail

NO SUPERVISADO

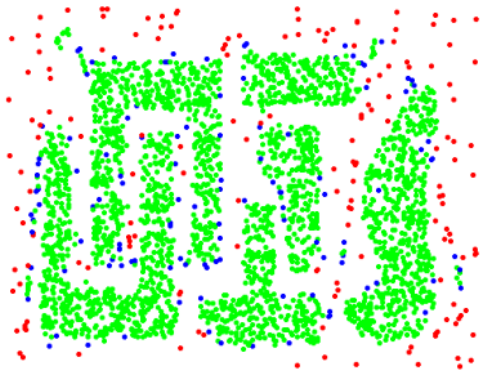


NO SUPERVISADO

Modelos de Cluster

- K-Means
  - Se puede especificar el número de clusters que le gustaría encontrar

- DBSCAN
  - Agrupación espacial basada en la densidad de aplicaciones con ruido
  - Hay que especificar que constituye el grupo, mínimo de obs del grupo

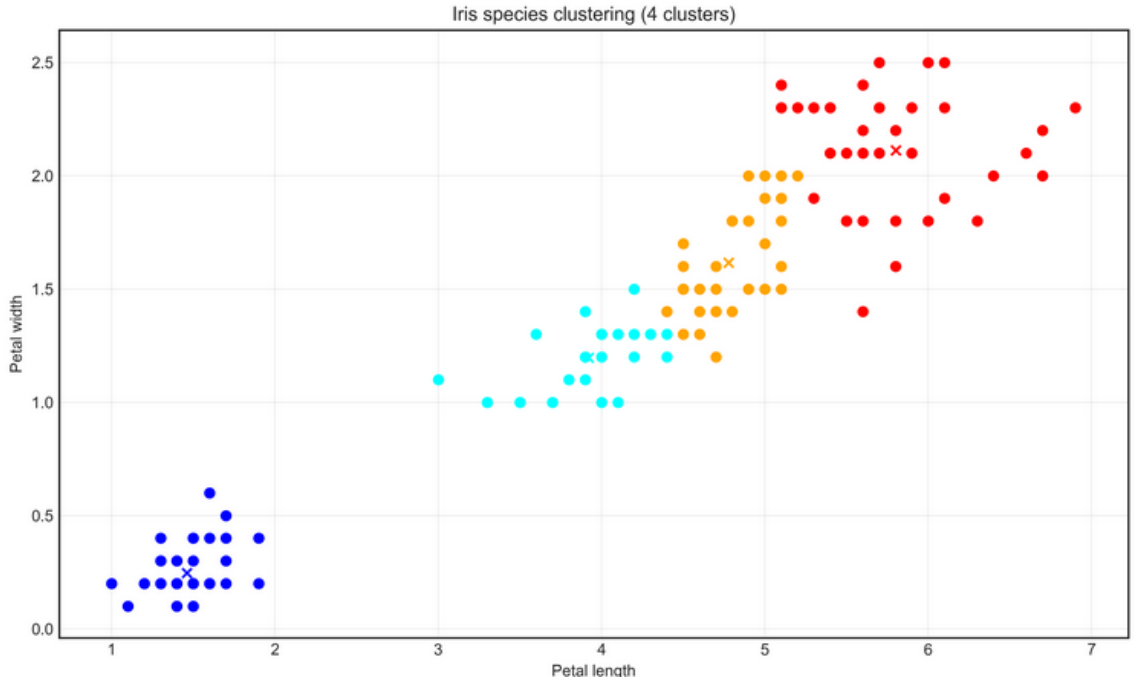


Datos de Flores de especies desconocidas, ancho y largo del pétalo

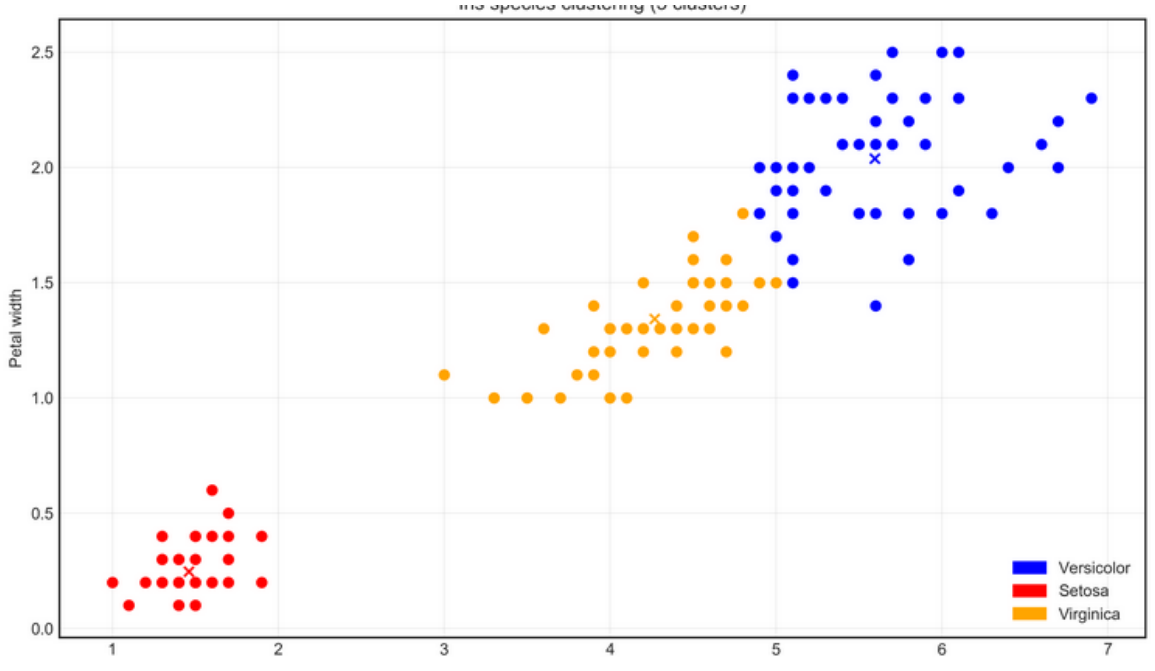
	Petal length	Petal width
0	1.4	0.2
1	1.4	0.2
2	1.3	0.2
3	5.1	1.9
...	...	...

*aún no tenemos etiqueta de la Especie, ni sabemos cuántas.

K- Means



K = 4

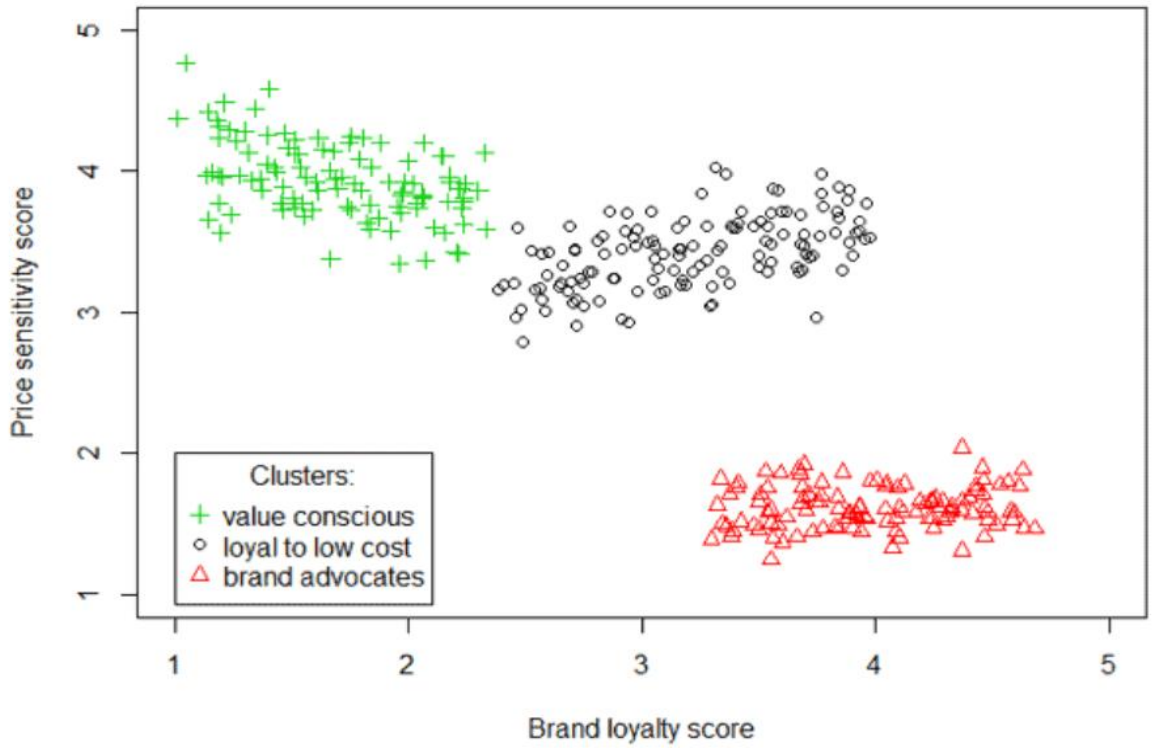


K = 3

Problemas de Agrupación

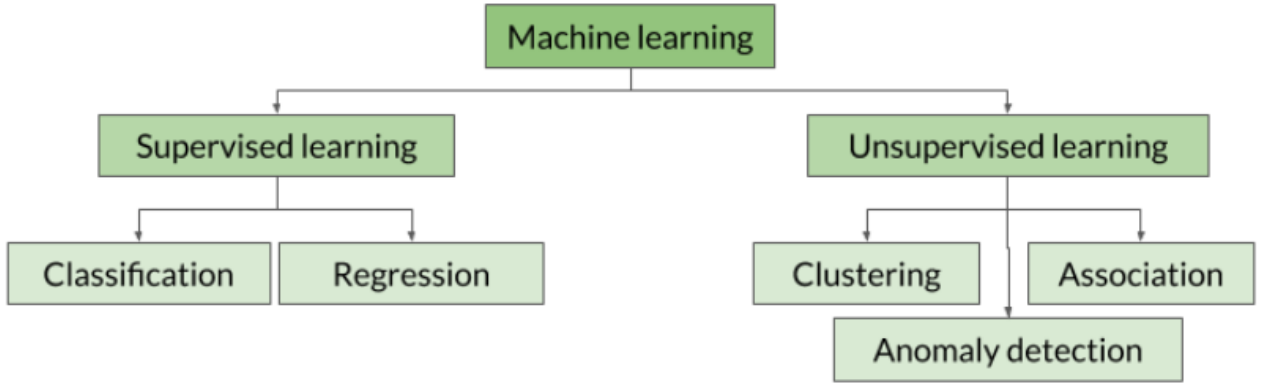
Los problemas de agrupación incluyen escenarios en los que las tendencias y las relaciones se descubren a partir de los datos. Se utilizan diferentes aspectos de los datos para agrupar ejemplos de diferentes maneras.

- Lealtad a la marca (eje X)
- Sensibilidad al precio (eje Y)



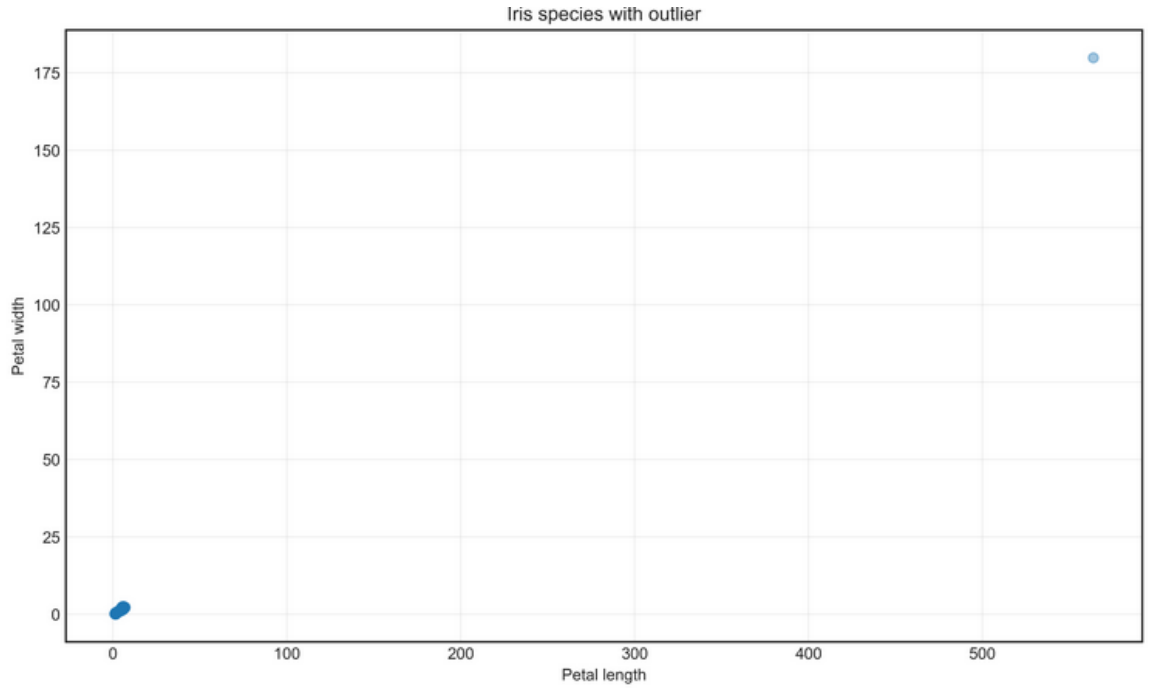
Fuente: https://select-statistics.co.uk/blog/customer-segmentation/

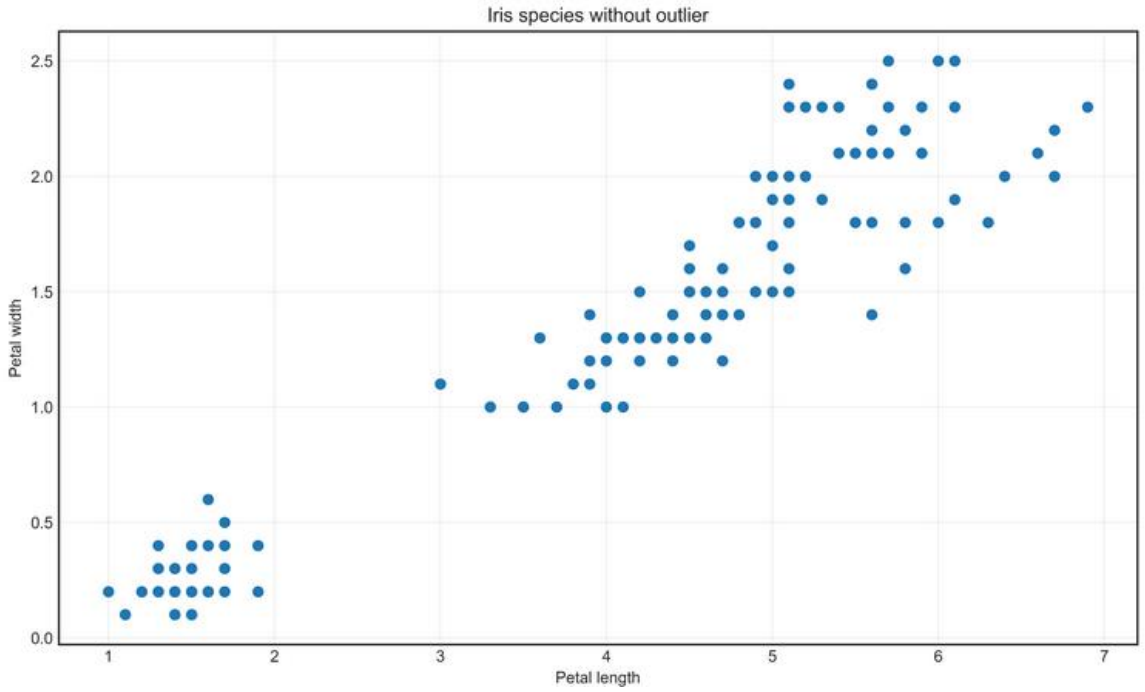
•(Verde) No son leales a la marca y son muy sensitivos al precio
•(Negro) Son leales a la marca pero sólo si es barato
•(Rojo): Son leales a la marca sin importar demasiado el precio



Detección Anormal

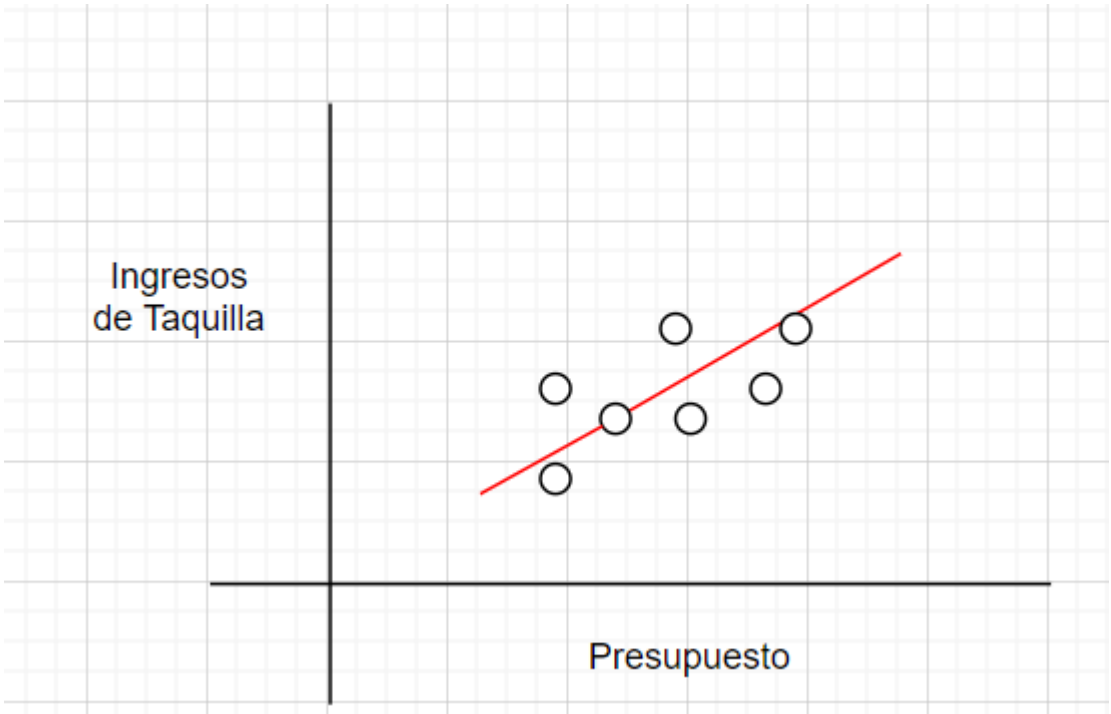
- Descubrir dispositivos que fallan más rápido o duran más
- Descubrir fraudes
- Descubrir pacientes que resisten sorprendentemente a una enfermedad mortal

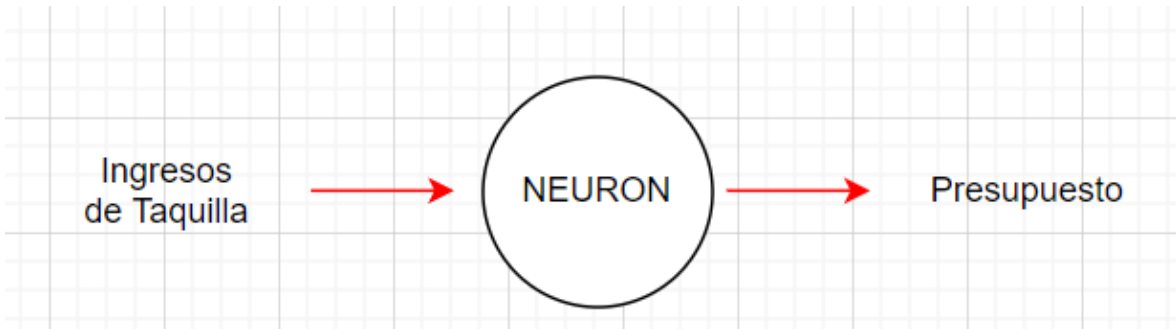


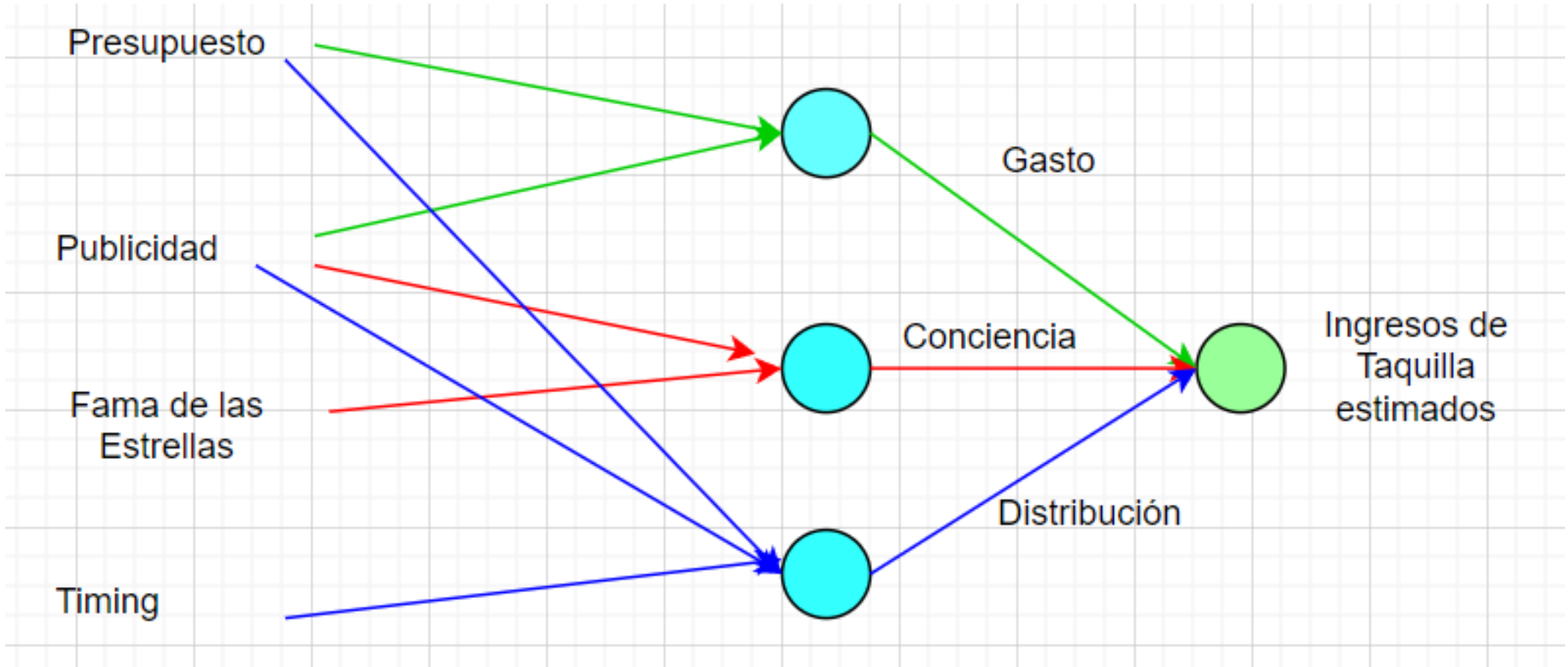


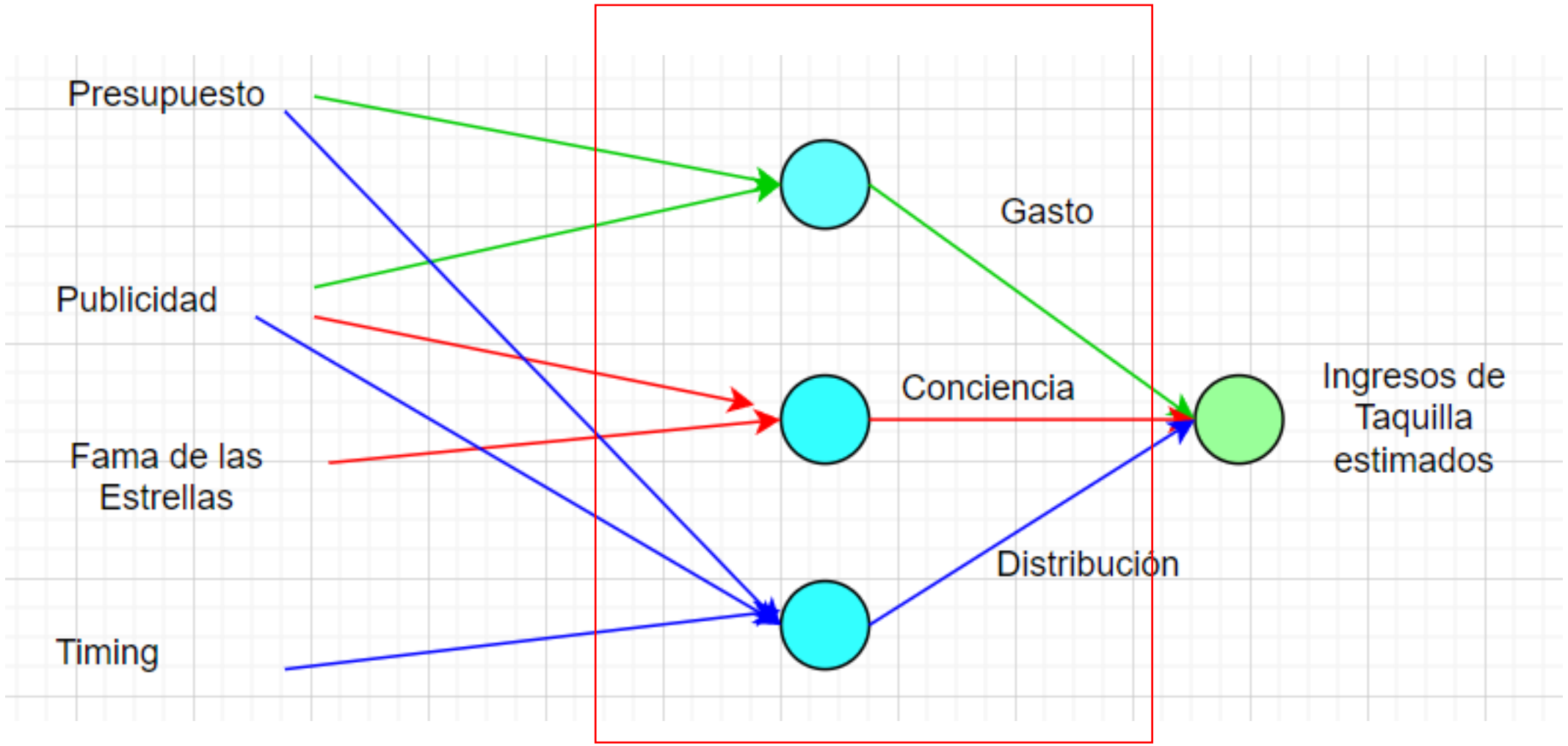
Deep Learning
Aprendizaje Profundo

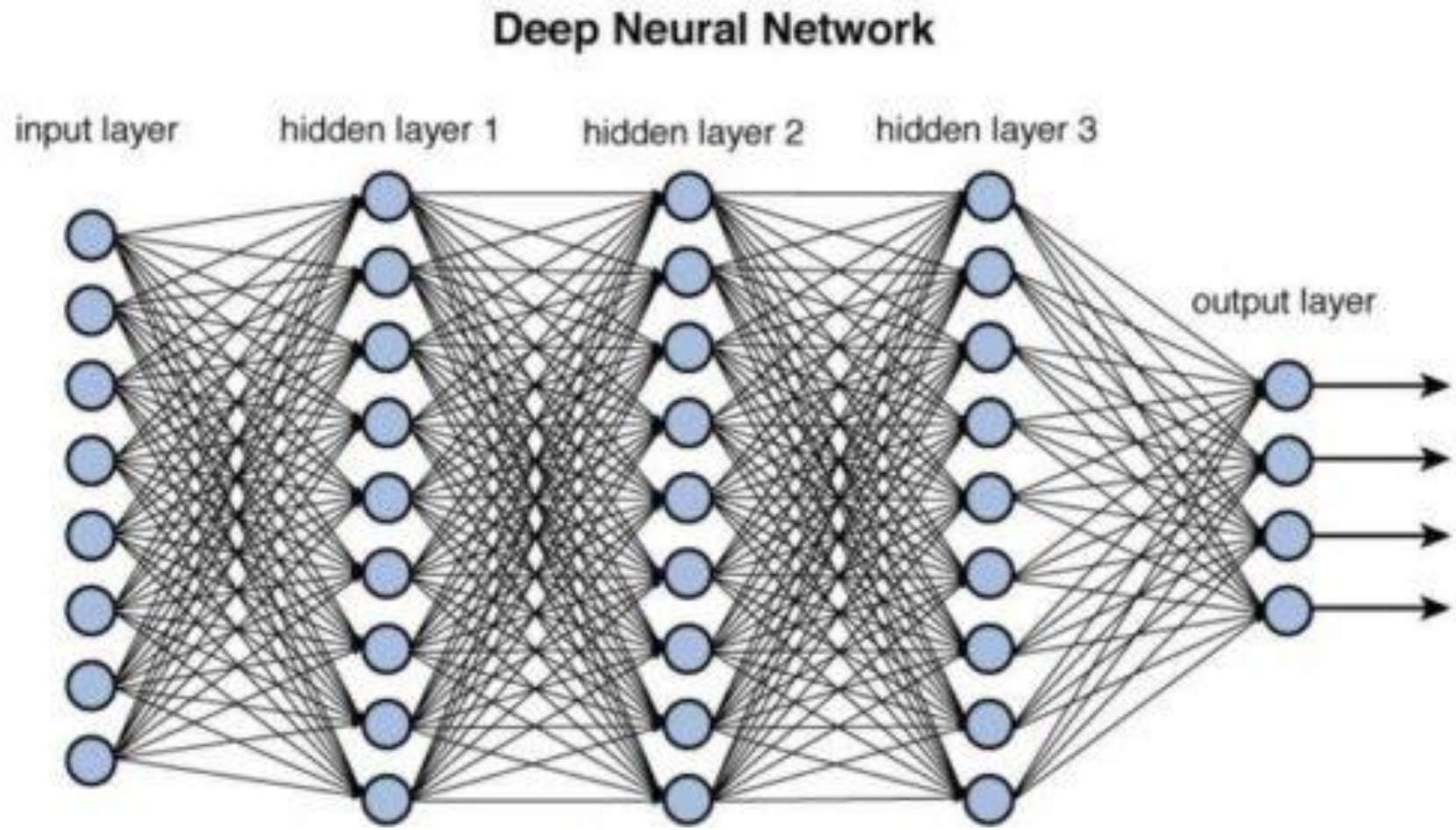
- Neural Networks
- Unidad básica: Neuronas (nodos)
- Requiere más datos
- Mejor usado cuando las entradas son textos o imágenes (menos estructurados)


Ingresos
de Taquilla
Presupuesto


Ingresos
de Taquilla
NEURON
Presupuesto


Presupuesto
Publicidad
Fama de las Estrellas
Timing
Gasto
Conciencia
Distribución
Ingresos de Taquilla estimados


Presupuesto
Publicidad
Fama de las Estrellas
Timing
Gasto
Conciencia
Distribución
Ingresos de Taquilla estimados


Deep Neural Network
input layer
hidden layer 1
hidden layer 2
hidden layer 3
output layer

Usos

- Google traduciendo grandes bloques de texto en cuestión de segundos.
- Galería de fotos de tu teléfono reconociendo caras automáticamente.

- Trabaja para un banco y desea ayudar a su empresa a otorgar préstamos más inteligentes. Le gustaría predecir la probabilidad de que un prestatario no cumpla con el pago de un préstamo en función de la información de su solicitud, como su historial de crédito, nivel de educación, ingresos y activos. Esta cantidad se puede utilizar para determinar la tasa de interés del préstamo. Tiene cientos de miles de ejemplos de préstamos de los últimos diez años.

  Por ahora, solo desea crear una prueba de concepto en su computadora portátil que demuestre que existe algún valor en la construcción de este modelo. El rendimiento no tiene que ser perfecto todavía.

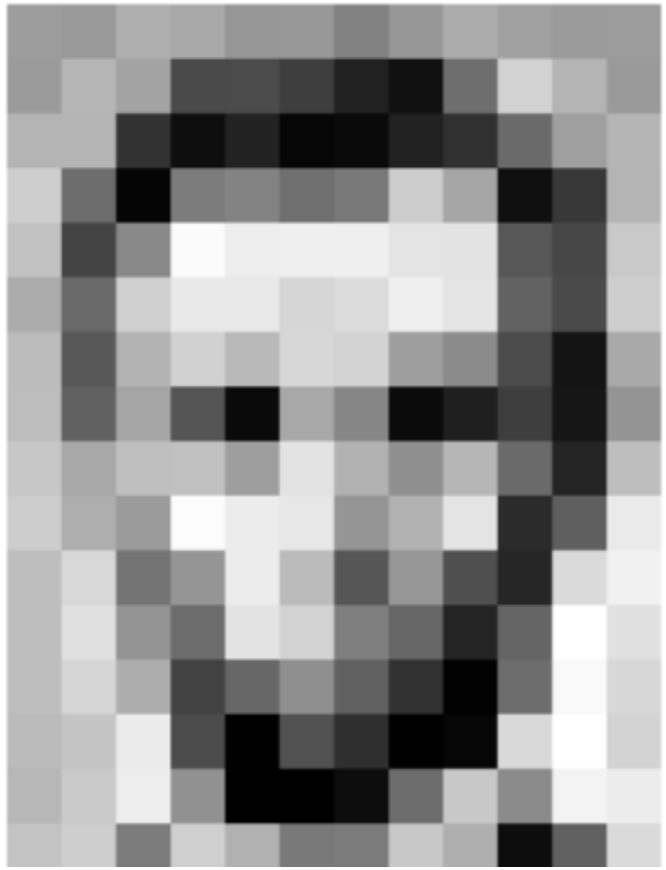
  Con base en esta información, ¿cómo debería resolver este problema?

a) Machine learning
b) Deep Learning

Redes neuronales procesando imágenes: Computer Vision

- El objetivo de la visión por computadora es ayudar a las computadoras a ver y comprender el contenido de las imágenes digitales.
- Es necesario para autos autónomos, para identificar imágenes externas al vehículo, señales de tránsito, cruceros, personas,..

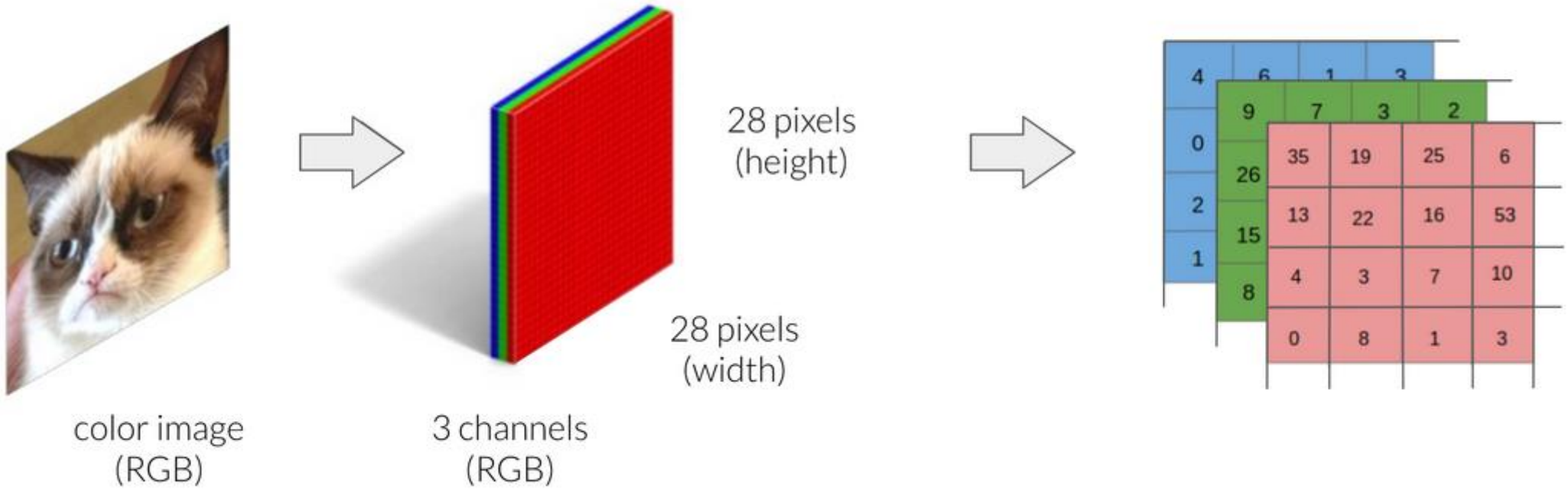
Image data

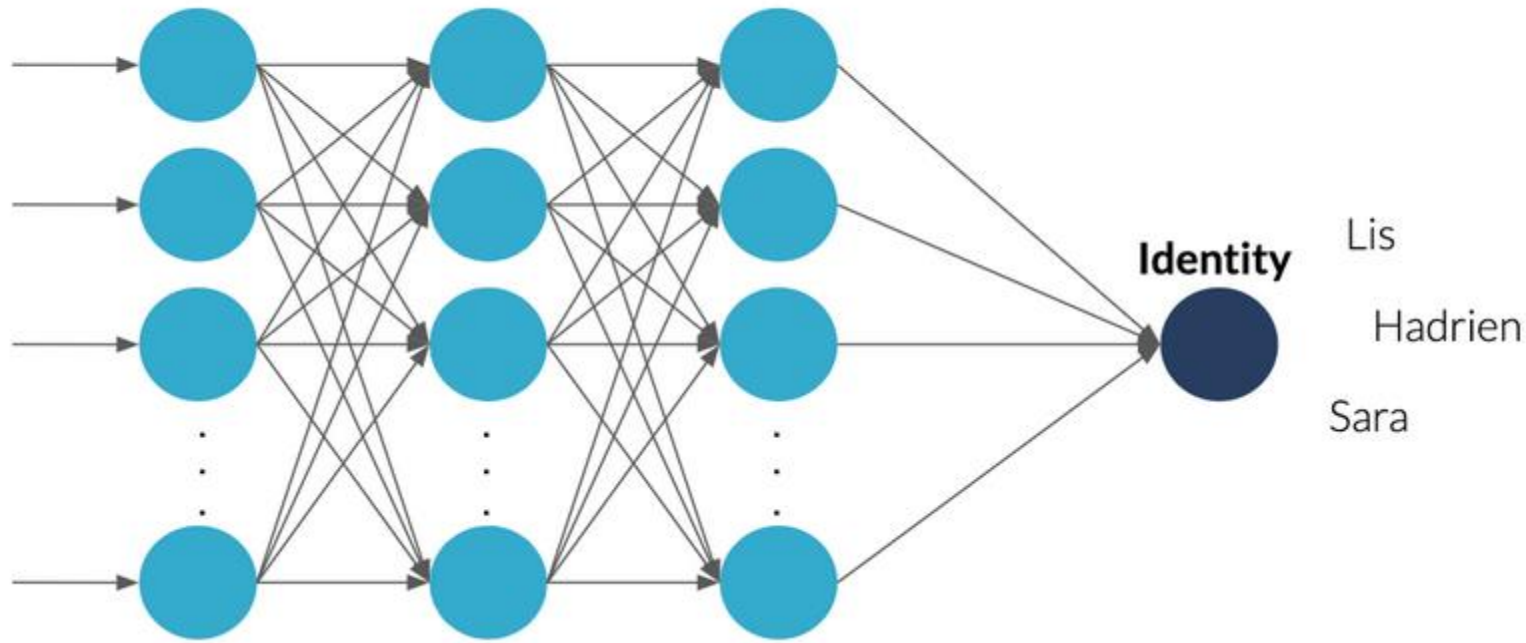


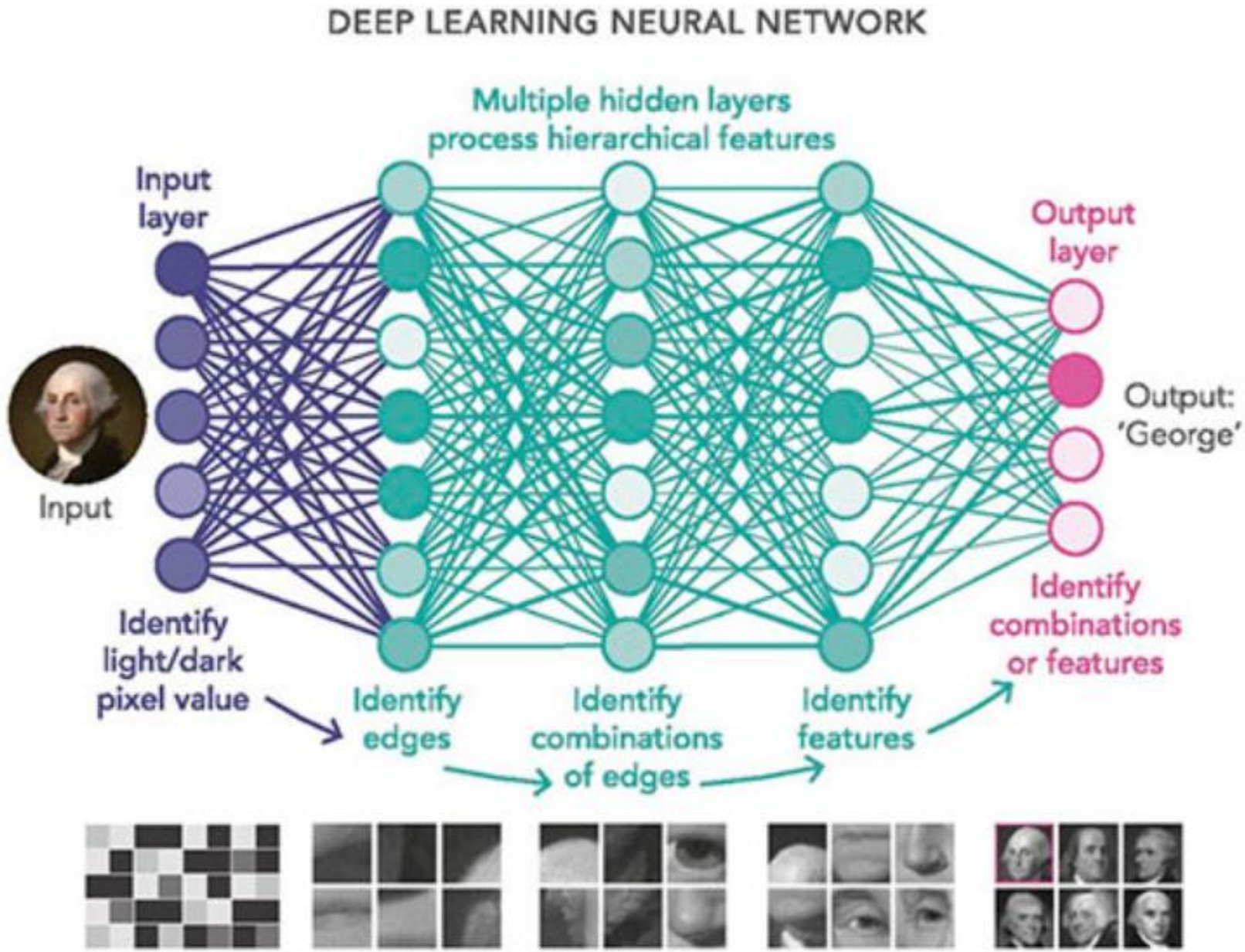
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

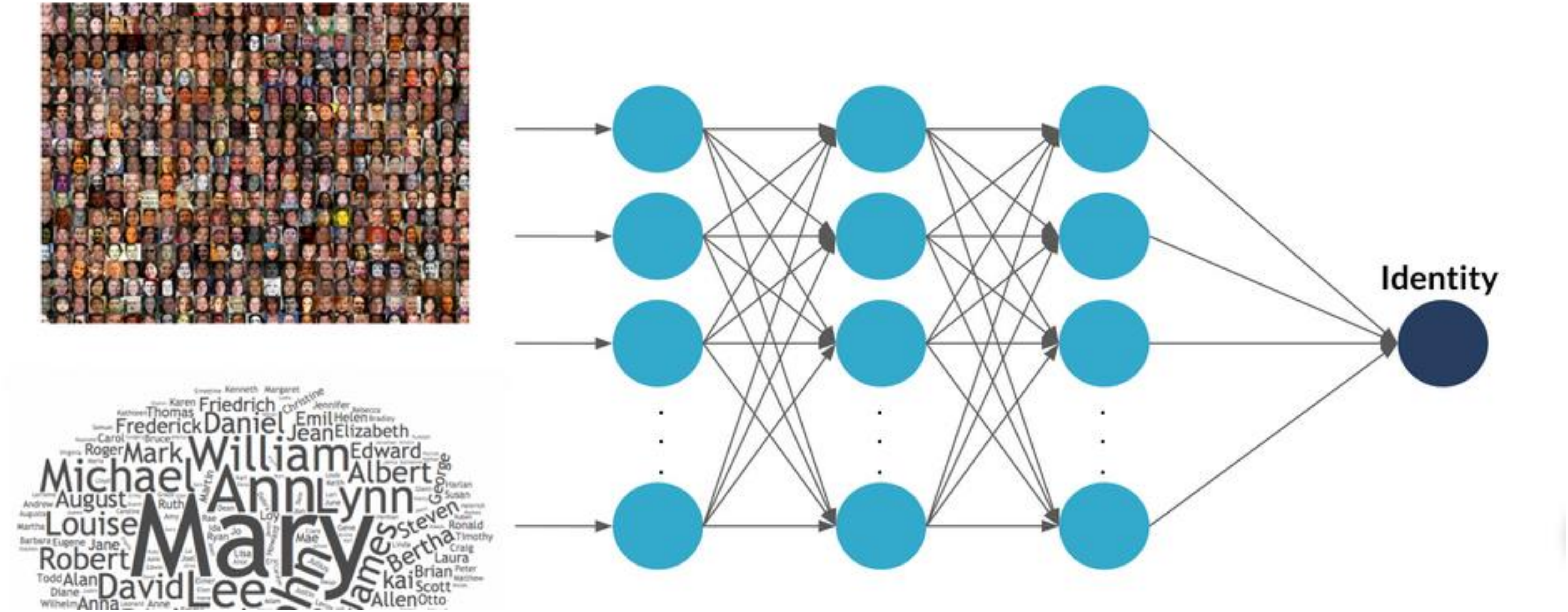
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Image data


28 pixels
(height)
28 pixels
(width)
color image
(RGB)
3 channels
(RGB)
4
6
1
3
9
7
3
2
0
26
35
19
25
6
2
15
13
22
16
53
1
8
4
3
7
10
0
8
1
3


Identity
Lis
Hadrien
Sara


DEEP LEARNING NEURAL NETWORK
Multiple hidden layers
process hierarchical features
Input
layer
Output
layer
Input
Output:
'George'
Identify
light/dark
pixel value
Identify
edges
Identify
combinations
of edges
Identify
features
Identify
combinations
or features


Identity
Kenneth
Margaret
Christine
Karen
Thomas
Friedrich
Jennifer
Rebecca
Frederick
Daniel
Emil
Helen
Jean
Elizabeth
Carol
Bruce
Roger
Mark
William
Edward
Michael
Albert
George
Ann
Lynn
Harlan
Susan
August
Ruth
Andrew
Louise
Steven
Ronald
Mary
Loy
Mae
Ryan
Jo
Lisa
Bertha
Timothy
Barbara
Eugene
Jane
James
Craig
Laura
Robert
Brian
Peter
Todd
Alan
David
Lee
kai
Scott
Diane
Anna
Anne
Allen
Otto

- Reconocimiento facial
- Vehículos automanejados
- Detección de tumores
- Detección de frutas con plagas
- Creación de imágenes, Deep fake, para poner personas en videos falsos donde la persona no está.

Cuando usar Deep Learning?

- Cuando se tiene gran cantidad de datos, superan a otras técnicas.
- Requieren computadoras con gran capacidad para su entrenamiento razonable.
- Cuando hay falta de conocimiento del dominio del problema para comprender sus características.
- Problemas complejos: visión artificial, procesamiento de lenguaje natural (NLP)

- Tu amigo acaba de enviarte una imagen por mensaje de texto. Sin embargo, cuando abre el mensaje, el contenido parece estar distorsionado. No se preocupe, tiene acceso a una red neuronal preentrenada que puede restaurar cualquier imagen.

  Antes de que pueda pasar la imagen a la red neuronal, debe convertirse en números para que tengamos características para ingresar. La resolución de la imagen es de 284×429 píxeles. Recuerde, cada píxel de color está representado por tres canales de color (rojo, verde y azul).

  ¿Cuántas características se pasarán al modelo para esta imagen en color?

a) 121836
b) 709
c) 365508

Las imágenes de los coches se transforman en números.

Las intensidades de los píxeles se introducen en una red neuronal.

Las neuronas aprenderán a detectar bordes

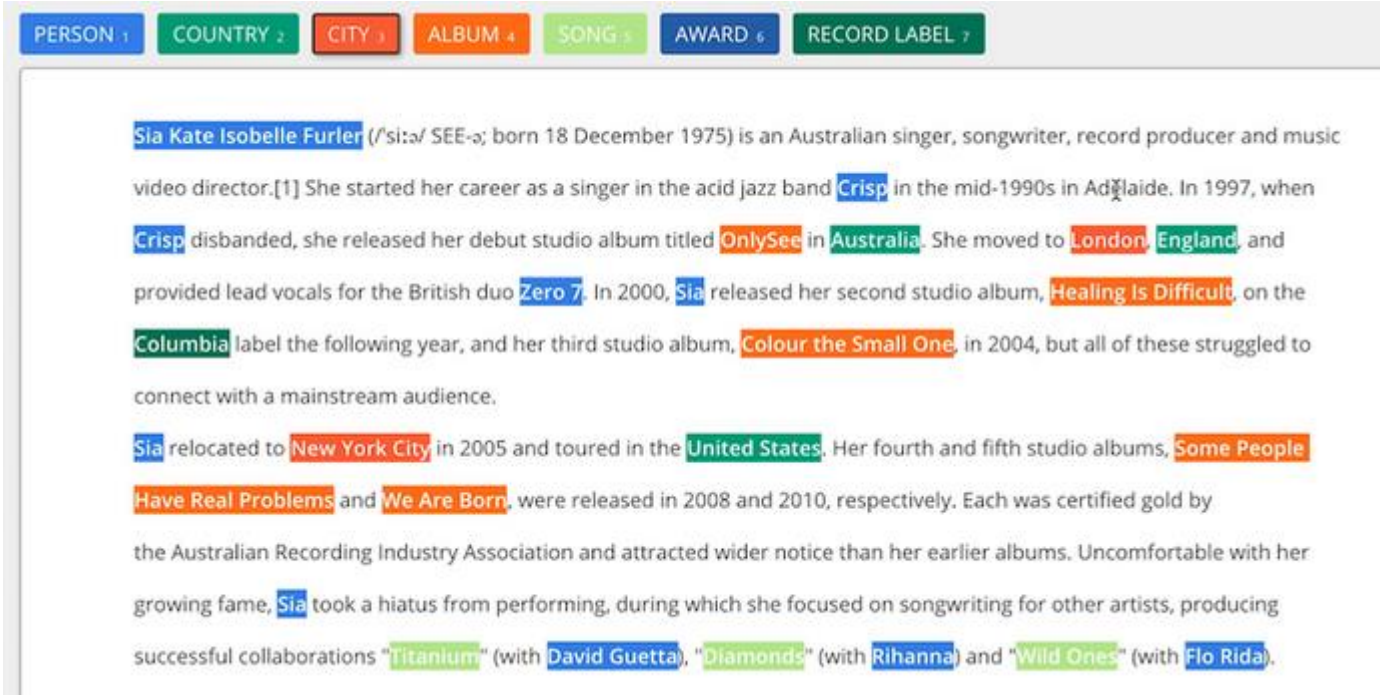
Las neuronas aprenderán a objetos más complejos como ruedas, puertas y ventanas.

Las neuronas aprenderán a detectar formas de vehículos

La imagen se clasifica como un coche o un camión.

NLP

- El procesamiento del lenguaje natural, o NLP, es la capacidad de las computadoras para comprender el significado del lenguaje humano.



Bolsa de Palabras

It is a period of civil war. Rebel spaceships, striking from a hidden base, have won their first victory against the evil Galactic Empire.

During the battle, Rebel spies managed to steal secret plans to the Empire's ultimate weapon, the DEATH STAR, an armored space station with enough power to destroy an entire planet.

Pursued by the Empire's sinister agents, Princess Leia races home aboard her starship, custodian of the stolen plans that can save her people and restore freedom to the galaxy....

the	7
to	4
rebel	2
plans	2
of	2
her	2
empire's	2
an	2
...	...

"U2 is a great band"

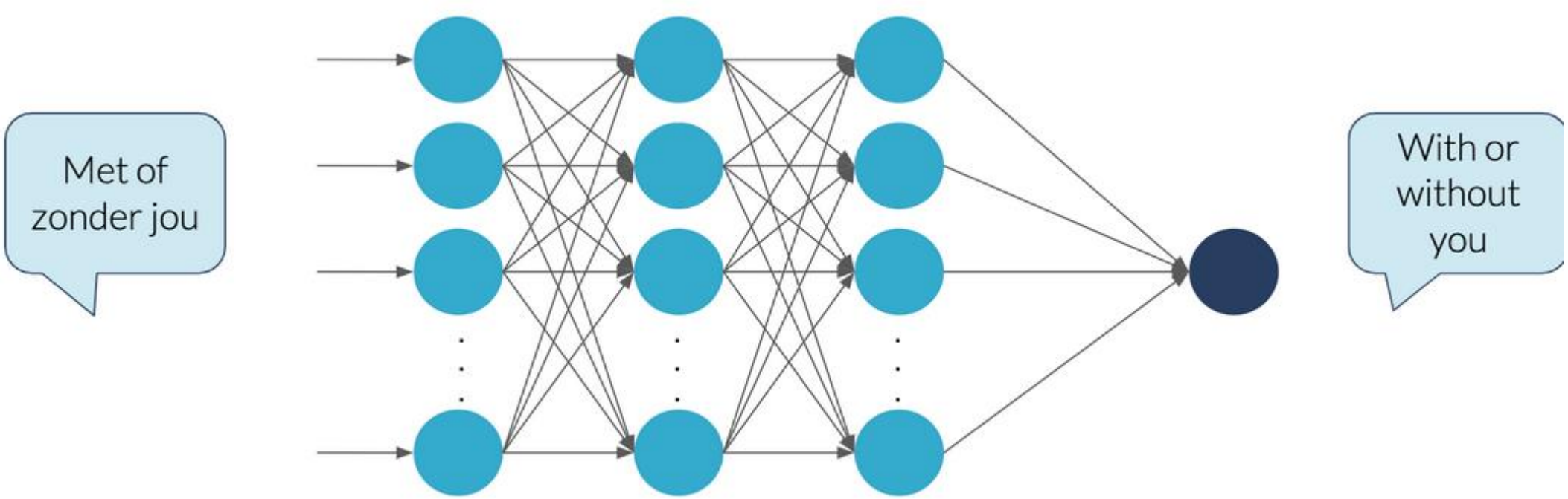
Word	Count
U2	1
Queen	0
is	1
a	1
great	1
band	1

"Queen is a great band"

Word	Count
U2	0
Queen	1
is	1
a	1
great	1
band	1

Ntwords[] = {1.,20,500, 30,..}
Indice: {0,1,2,34, ……}
Palabra: {“U2”, “Queen”, “Is”, “a”, “great”, “band”}

…
Palabra: {“U2 is a”, “Queen is a”,…..}



- Traductor
- Chatbot
- Asistente personal
- Análisis de sentimiento

Datos

- Tener datos de alta calidad
- Históricos con criterios en los hechos sobresalientes tener cuidado
- Fuentes confiables
- Relevancia de datos, revisiones de valores atípicos, excepciones, que destaque como sospechosa
- Dominio de expertos para analizar patrones de datos inesperados
- Documentación
- Explicabilidad: Los resultados del modelo hay que exponerlo y explicarlo en que consiste y verificar si se adhiere a las leyes del negocio y permitir detección de sesgos lo antes posible. (diabetes, letras a mano)

Ejemplo

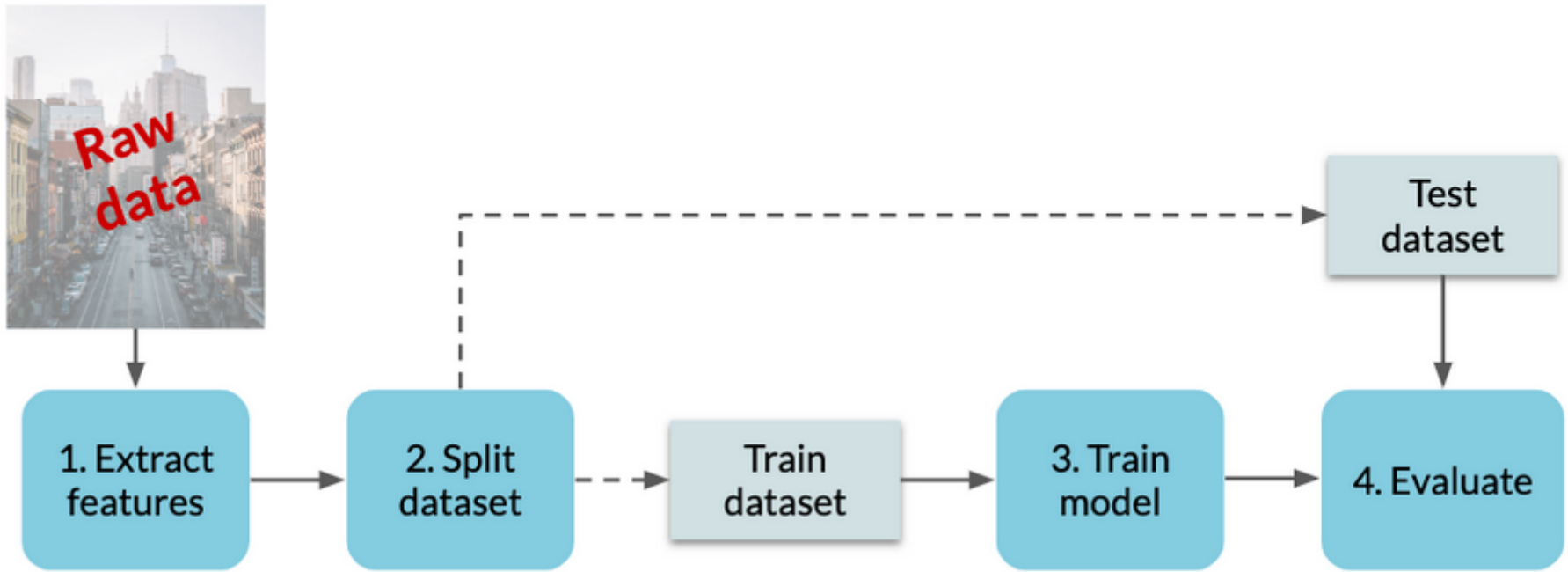
- Cuál necesita explicabilidad?
  a) Clasifica imágenes de mamografías con cáncer o sin cáncer
  b) Provee una lista de probables diagnósticos, dado un conjunto de factores del los pacientes.

***Los problemas en los que es útil saber por qué el algoritmo eligió una clasificación particular deben abordarse con IA explicable.**

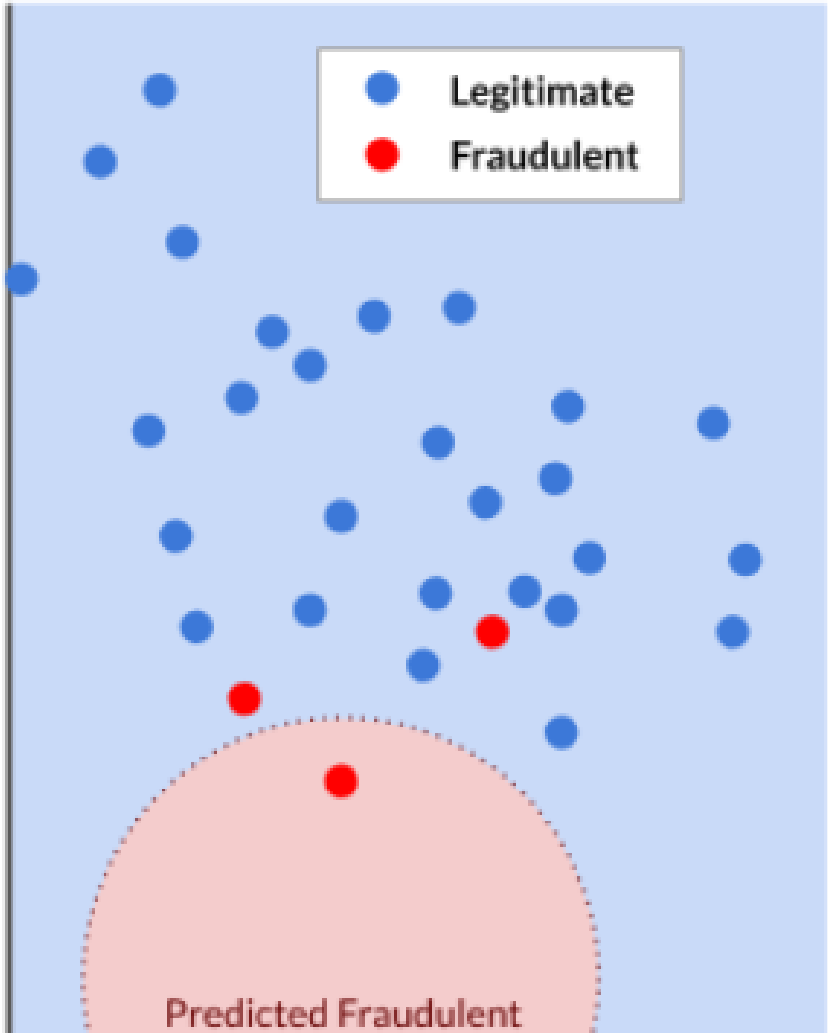
Evaluando la performance del Modelo

Evaluating performance

Evaluate step

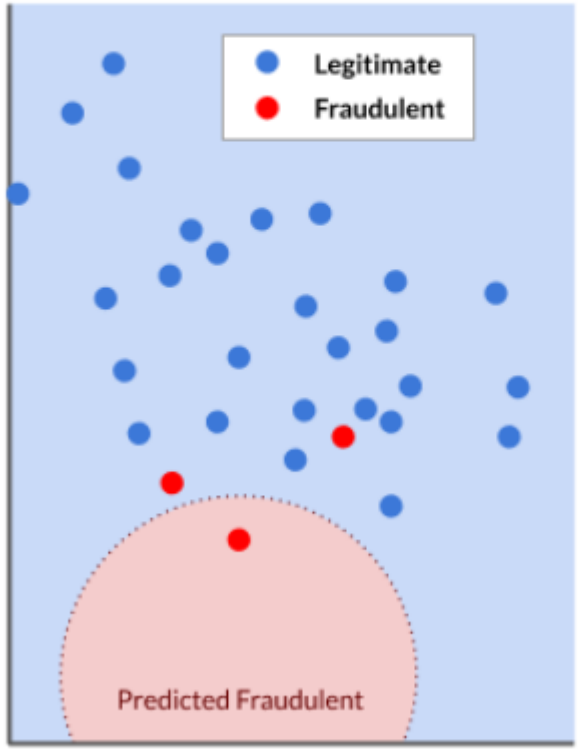


Accuracity (Precisión) = número correcto de ***clasificados*** / total de puntos
= 28 correctamente clasificados / 30 puntos totales = 93.33 %
*Esto omite la mayoría de Tx fraudulentas.



Matriz de Confusión

False positives, true negatives



		Actual values	
		Fraudulent	*Not Fraudulent*
Predicted	*Fraudulent*	1 true positives	0 false positives
	Not Fraudulent	2 false negatives	27 true negatives

Sum = 30 points

$$Sensitivity = \frac{true\ positives}{true\ positives + false\ negatives}$$

*Valora mejor la predicción

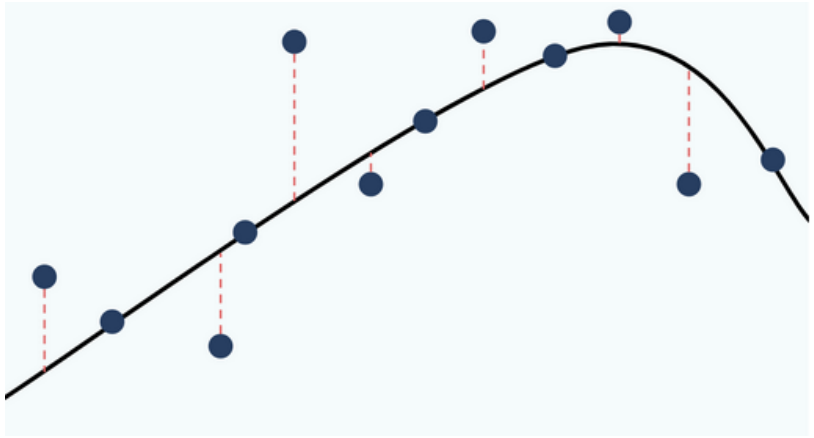
Sensitivity (Recall) = 1 / (1+2) = 33.34%

$$PuntajeF1 = \frac{2 * Precision * Sensibilidad}{Precision + Sensibilidad}$$

es de gran utilidad cuando la distribución de las clases es desigual, por ejemplo cuando el número de pacientes con una condición es del 15% y el otro es 85% , lo que en el campo de la salud es bastante común.

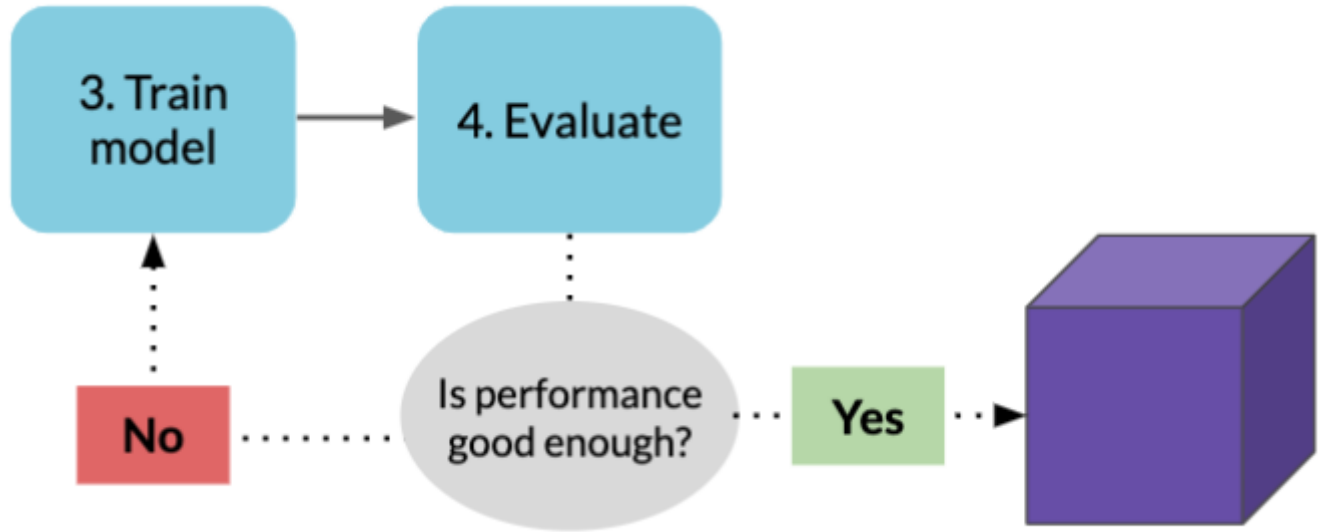
Evaluando la Regresión:

Esencialmente la diferencia entre el valor real y el valor predicho



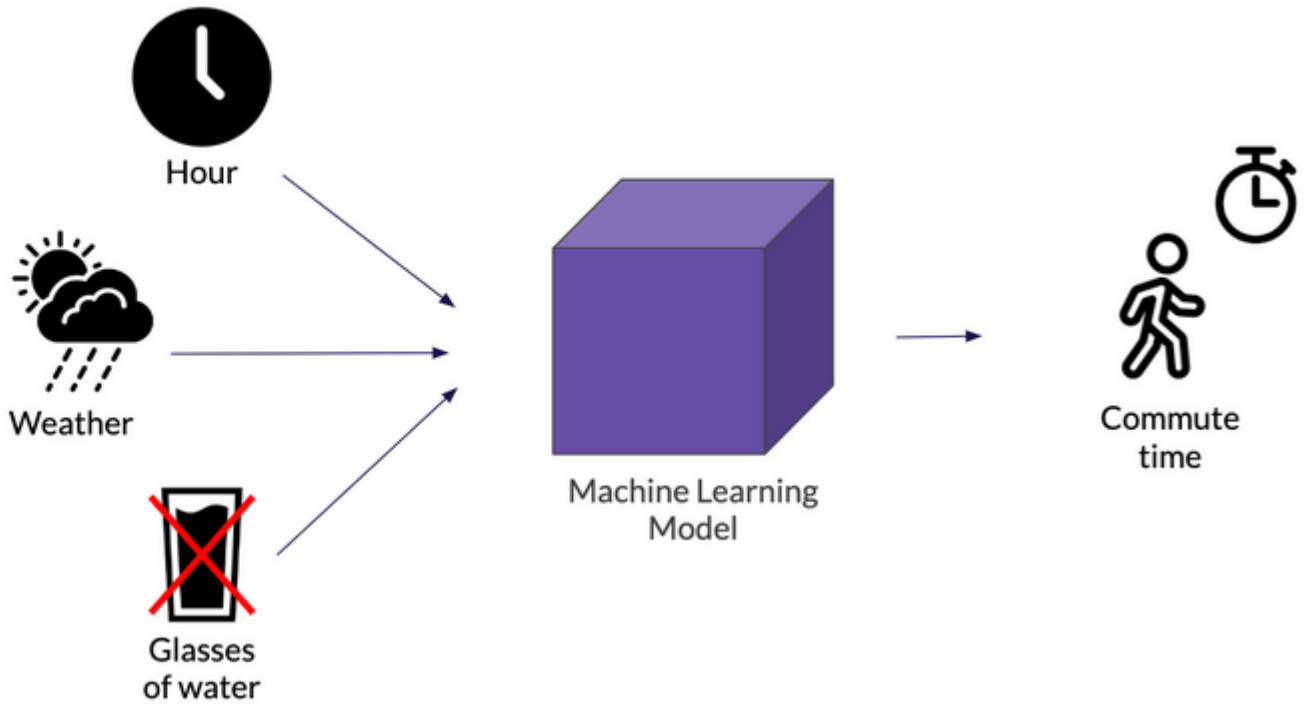
Error = distancia entre puntos del valor actual y el valor de la línea de predicción
Error cuadrático medio, error absoluto medio, entre otros.

Cómo mejorar el modelo?



Dimensionality reduction

Irrelevante: Algunas características no aportan información útil



Dimensionality Reduction

Correlación: Algunas características tienen similar información

Mantener una sola característica:
-Peso y Tamaño de zapatos

Colapsar múltiples características dentro de única **función** subyacente
-Si tenemos dos características altura y peso **->** Índice de Masa Corporal

Tunning de Hiperparámetros

Tamaño de la muestra, taza de aprendizaje, topología, tamaño de la red neuronal, todo lo que afecta al rendimiento, calidad y velocidad del modelo.

Métodos Ensamblados – Ensemble Methods

Combina varios modelos para producir uno óptimo.