



# Machine Learning

**Roberto Muñoz**

Doctor en Astrofísica

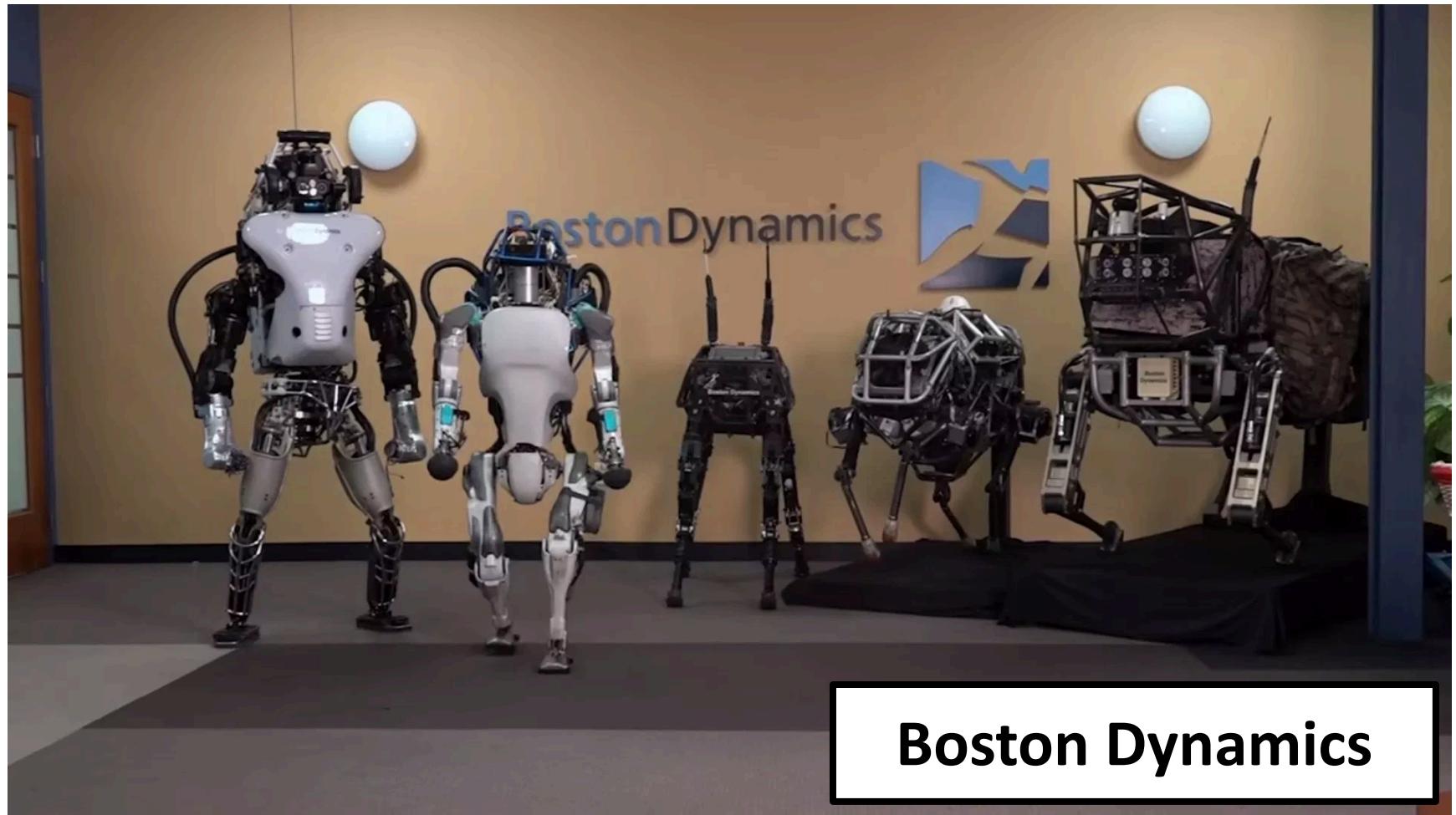
Director Centro I+D MetricArts



@RobertoKPax

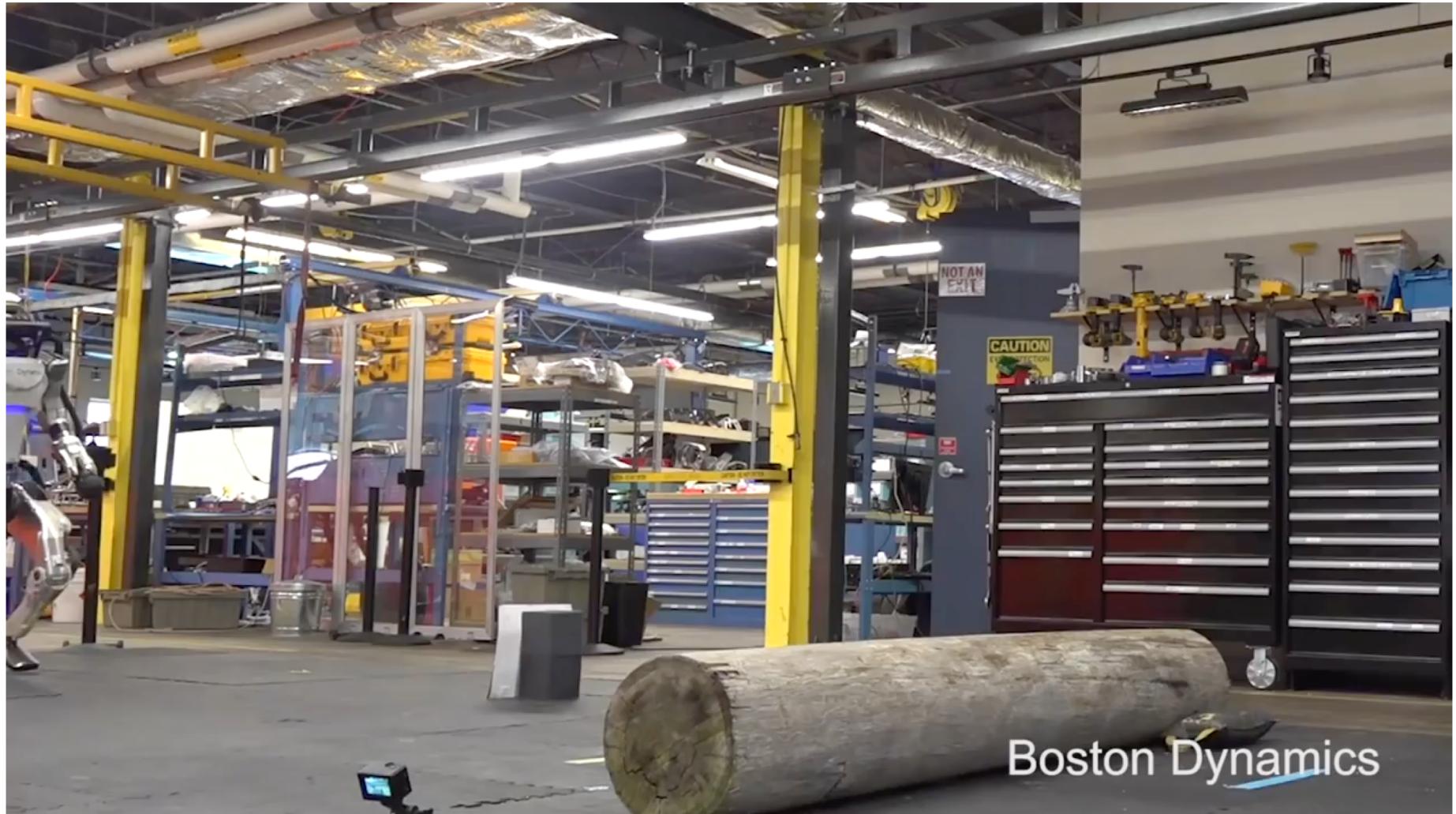
**METRICARTS**

# El ejército de Boston Dynamics



**Boston Dynamics**

# Atlas de Boston Dynamics



o amenza



The development of full artificial intelligence could spell the end of the human race.

— *Stephen Hawking* —

AZ QUOTES

El camino del Data Scientist

# **MACHINE LEARNING**

Survey

## ¿Qué es el Machine Learning?

Visitar siguiente link

<https://pollev.com/robertomunoz211>

# ¿Qué es el Machine Learning?

Es una disciplina que se ocupa de los sistemas necesarios para analizar grandes volúmenes de datos

Es una disciplina que se ocupa de los procesos necesarios para explorar y visualizar datos

Es un subcampo de las Ciencias de la Computación cuyo objetivo es desarrollar técnicas que permitan a los computadores aprender

Es un subcampo de la Inteligencia Artificial que comprende el uso de algoritmos que aprenden a partir de los datos y cuyo objetivo es hacer predicciones

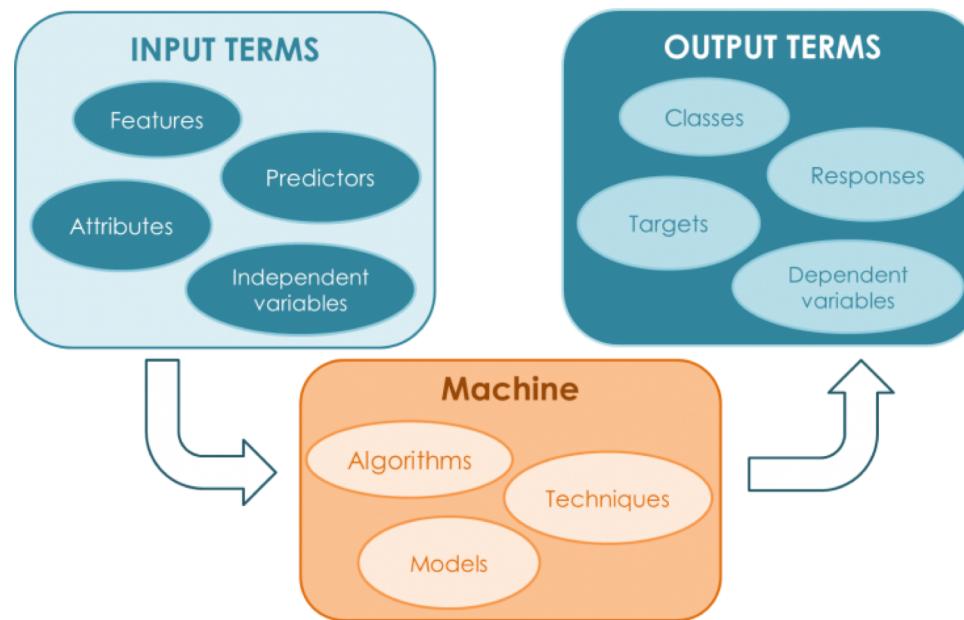
Es una disciplina que comprende el uso de redes neuronales profundas para construir modelos predictivos a partir del uso de datos

Es un subcampo que comprende el uso de métodos y técnicas computacionales/estadísticas que permiten generar modelos de manera automática mediante el uso de datos

Ninguna de las anteriores

# ¿Qué es el Machine learning?

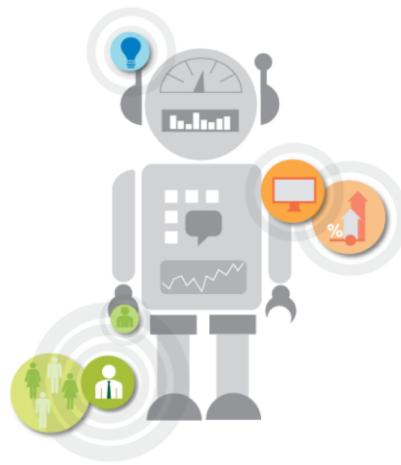
- El aprendizaje automático o *machine learning* es un subcampo de las Ciencias de la Computación cuyo objetivo es desarrollar técnicas que permitan a los computadores aprender.



# ¿Qué se necesita?

¿Qué se requiere para crear buenos sistemas de machine learning?

- Recursos de preparación de datos.
- Algoritmos – básicos y avanzados.
- Automatización y procesos iterativos.
- Escalabilidad.
- Modelado en conjunto.



¿Lo sabía?

- En el aprendizaje basado en máquina, un destino se conoce como etiqueta.
- En estadística, un destino se conoce como variable dependiente.
- Una variable en estadística se conoce como característica en el machine learning.
- Una transformación en estadística se conoce como creación de característica en el machine learning.

# ¿Qué queremos que haga ML?

- Dada una imagen, predecir patrones complejos de nivel superior

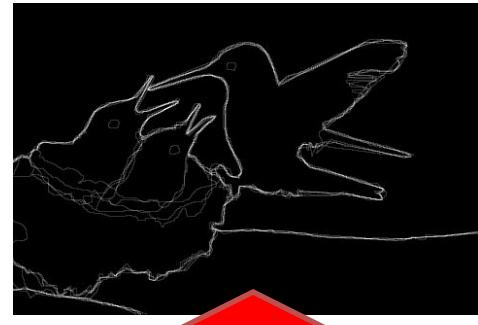
“Gato”



Reconocimiento



Detección



Segmentación  
[Martin et al., 2001]

# Tipos de Machine Learning

## Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

## Unsupervised Learning

- No labels/targets
- No feedback
- Find hidden structure in data

## Reinforcement Learning

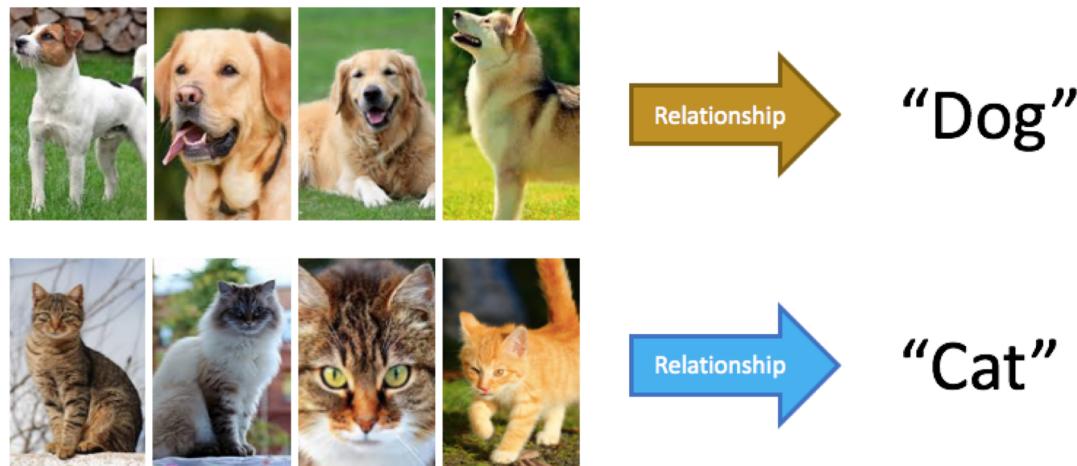
- Decision process
- Reward system
- Learn series of actions

# Tipos de Machine Learning

- **Aprendizaje supervisado**

El sistema aprende en base a datos estructurados o no estructurados. Clasificados previamente.

El algoritmo produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema.



## Métodos comunes

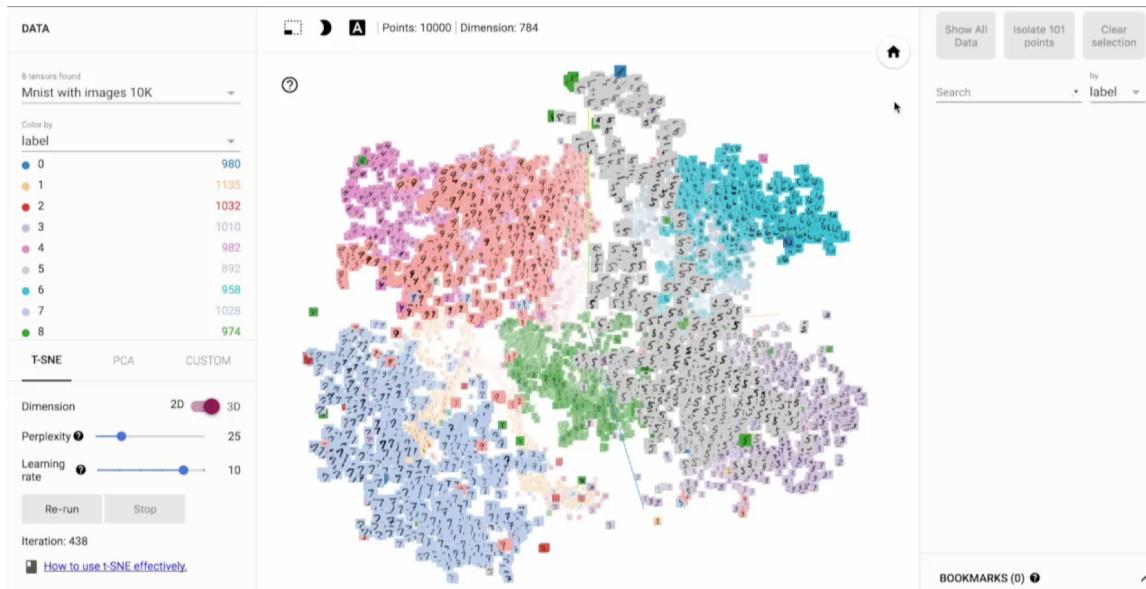
- Regresión Lineal
- Regresión Logística
- Support Vector Machine (SVM)
- Árboles de Decisión
- Random Forest

# Tipos de Machine Learning

- **Aprendizaje no supervisado**

Modelo se construye usando un conjunto de datos como entrada, los cuales no han sido clasificados previamente.

El sistema tiene que ser capaz de reconocer patrones para poder etiquetar las nuevas entradas.



## Métodos comunes

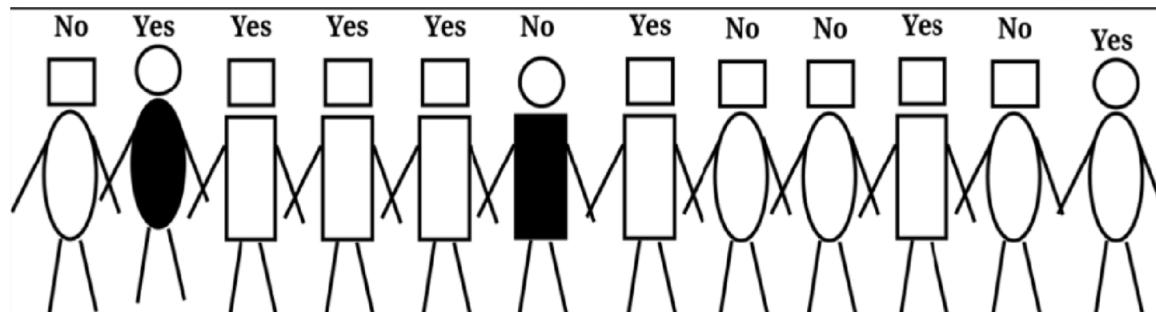
- K-means
- Mean Shift
- Expectation Maximization (EM)
- Autoencoders

Machine Learning

# **DEFINICIONES Y TERMINOLOGÍA**

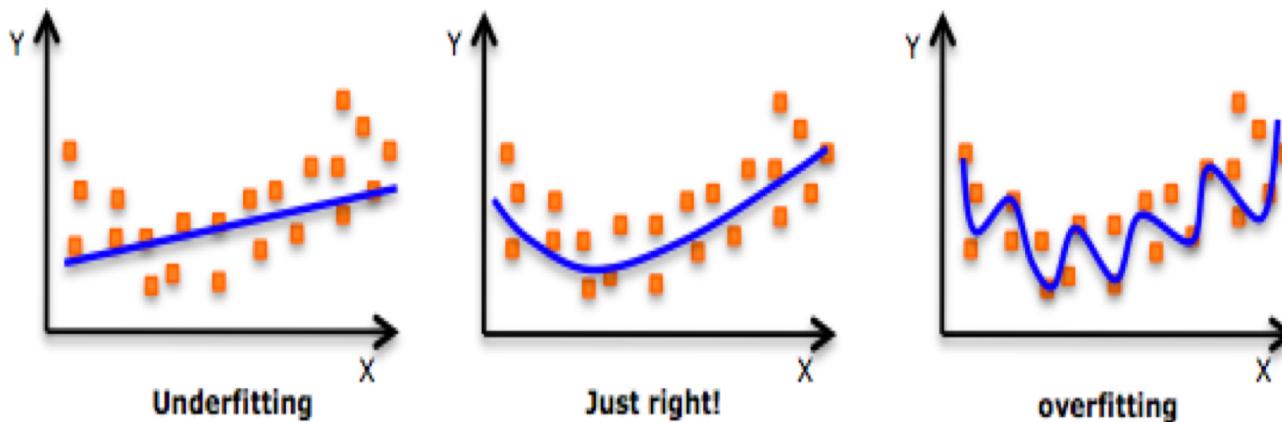
# Terminología

- Atributo (variable independiente):
  - Forma de la cabeza.
  - Forma del cuerpo.
  - Color del cuerpo.
- Objetivo o Etiqueta (variable dependiente): “Yes” o “No”
- Predicción: Son los valores que resultan de nuestro algoritmo (“Yes” o “No”).



# Over y Underfitting

- **Overfitting:** Hacerlo tan bien en un conjunto de datos que perdemos generalidad.
- **Underfitting:** Hacerlo mal en un conjunto de datos, perdiendo generalidad e incluso tendencias en los datos.



# Matriz de confusión

		Predicted class	
		$P$	$N$
$P$	$P$	True Positives (TP)	False Negatives (FN)
	$N$	False Positives (FP)	True Negatives (TN)

# Recall, Precision, Accuracy

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

- **Entrenamiento:** Ajustar los parámetros del algoritmo de forma tal de que se minimicen la cantidad de predicciones que no correspondan a la etiqueta original.
- **Recall:** Porcentaje de clasificados correctamente como positivos sobre todos los que realmente eran positivos.
- **Precision:** Porcentaje de clasificados correctamente como positivos sobre todos los clasificados como positivos.
- **Accuracy:** Porcentaje de clasificados correctamente.

## Entendimiento del problema

- Qué queremos predecir.
- Qué datos necesitamos.
- Qué datos podemos obtener.

## Preparación de los datos

- Tomar los datos previamente recopilados y darles un formato adecuado.
- Por ejemplo, la variable que describe del color de pelo de una persona puede tener varios valores (rubio, castaño, pelirrojo, negro, etc.)
- En algunos casos los algoritmos necesitan que se les entreguen los datos como números. Por ejemplo rubio=0, castaño=1, pelirrojo=2, negro=3.

## Conjunto de entrenamiento y evaluación

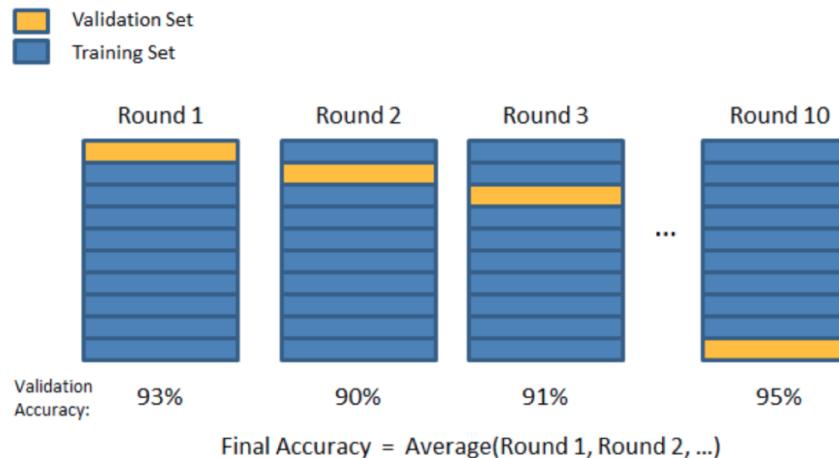
- Para poder realizar una buena comprobación del desempeño de nuestro entrenamiento es importante separar nuestros datos en los siguientes conjuntos:
  - Dataset de entrenamiento
  - Dataset de validación
  - Dataset de evaluación

## ¿Porqué usar estos datasets?

- Nuestro modelo será aplicado a datos que no han sido vistos previamente y nos gustaría tener una intuición de cuán bien trabaja el modelo sobre estos datos.
- Reservar una parte de los datos y no usarla en el entrenamiento. Nuestro algoritmo nunca los habrá visto anteriormente.
- Dado que sabemos el verdadero valor de la variable objetivo, tenemos la posibilidad de comparar lo que predecirá nuestro algoritmo y el verdadero valor.

# K-fold Cross Validation

- Los algoritmos tienen distintos parámetros que debemos determinar *a priori*. Por ejemplo, en un árbol de decisión sería la profundidad máxima del árbol
- Para evitar que el ajuste de esos parámetros nos lleve a un overfitting, podemos separar el conjunto de entrenamiento de la siguiente forma.



## Cómo seleccionar algoritmo de ML

- La selección del algoritmo de aprendizaje va a estar guiada por el tipo de etiqueta que se tenga (continua, discreta).
- La interpretación también es una herramienta que ayuda mucho, pues hay algoritmos que son muy fáciles de interpretar y usar (árbol de decisión).
- Una vez que elegimos nuestro algoritmo, entrenamos realizando el k-fold (siempre que sea posible) si su resultado es satisfactorio, entrenamos usando todo el training set como uno solo.
- Recomendación: Probar con distintos algoritmos y ver los resultados.

## Evaluación

- Hasta este punto nunca hemos tocado el test set, esto es sumamente importante pues queremos que en este paso el algoritmo nunca haya visto esos datos.
- En la validación obtenemos todos los indicadores (Accuracy, Precision, Recall) de los datos de Test. Esto nos permitirá saber cómo se comportará nuestro algoritmo en datos generales.