

Introducción a Machine Learning

- **¿Qué es Machine Learning?**
Definiciones: ML, DL, aprendizaje, atributos, observaciones, target o label..
- **El proceso de ML**
Feature engineering, modelos supervisados y no supervisados, train/test split, cross validation, evaluación de modelos
- **Casos de Aplicación**
Fraude, churn, upselling, cross-selling, segmentación, clustering geográfico.
- **Etapas de un proyecto de ML**
Framing del problema de negocio, integración de datos, modelado, puesta en producción y monitoreo.

Introducción a machine learning

¿Cómo podríamos hacer un programa para reconocer gatos en una foto?

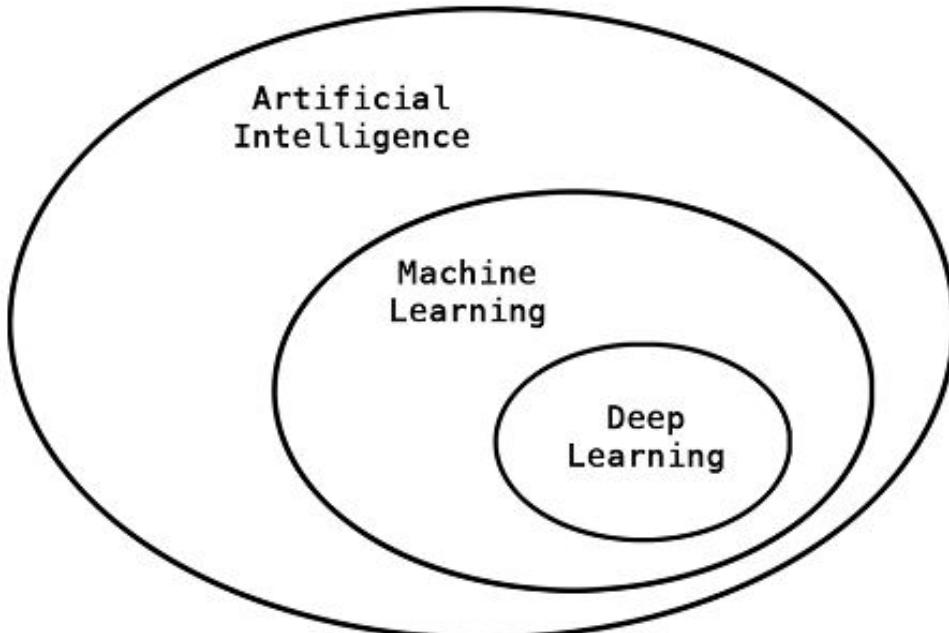


Decimos que un programa “aprende” si para cierta tarea, su performance mejora cuando crece la experiencia.

T. Mitchell

Introducción a Machine Learning

Machine Learning



- La “**Inteligencia Artificial Simbólica**” dominó el paradigma de IA desde 1950 hasta la parte de los ‘80. Su pico de popularidad: boom de los “**sistemas expertos**”.

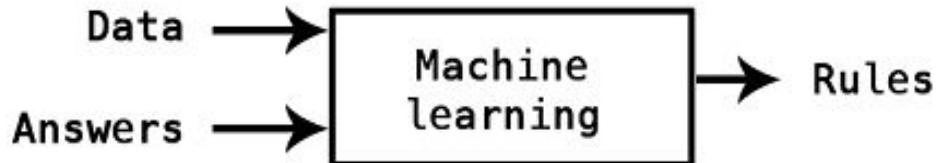
Introducción a Machine Learning

Machine Learning

Programación clásica (el paradigma de la IA simbólica): los humanos ingresan reglas (un programa), datos a ser procesados según esas reglas, y de este proceso resultan las respuestas esperadas.



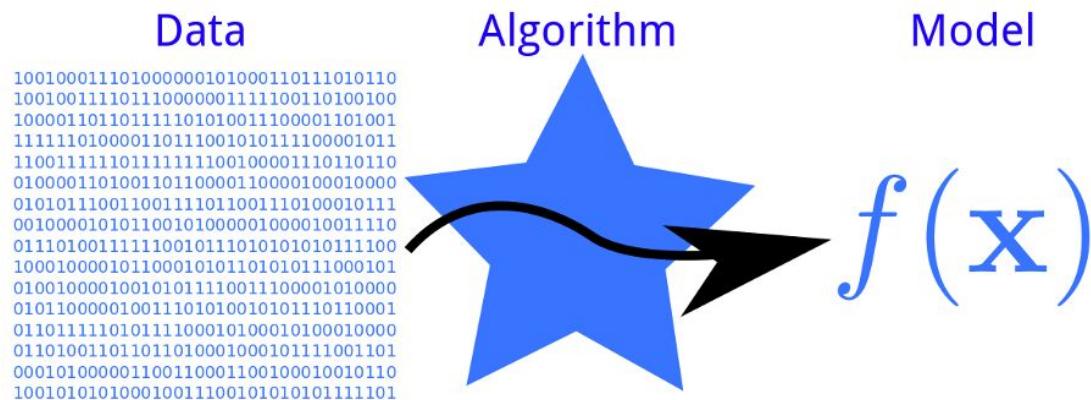
Machine Learning: los humanos ingresan datos como input además de las respuestas esperadas, y las reglas surgen como output. Estas reglas pueden luego ser aplicadas a nuevos datos para producir respuestas originales.



¿Qué es Machine Learning?

El aprendizaje entra en juego cuando le damos a estos modelos parámetros que pueden modificarse y que se adaptan a los datos observados de manera automática; de esta forma, puede considerarse que **el algoritmo aprende de los datos**.

Una vez que estos modelos han sido ajustados a datos observados previamente, pueden ser usados para **predecir y entender aspectos de datos nuevos**.



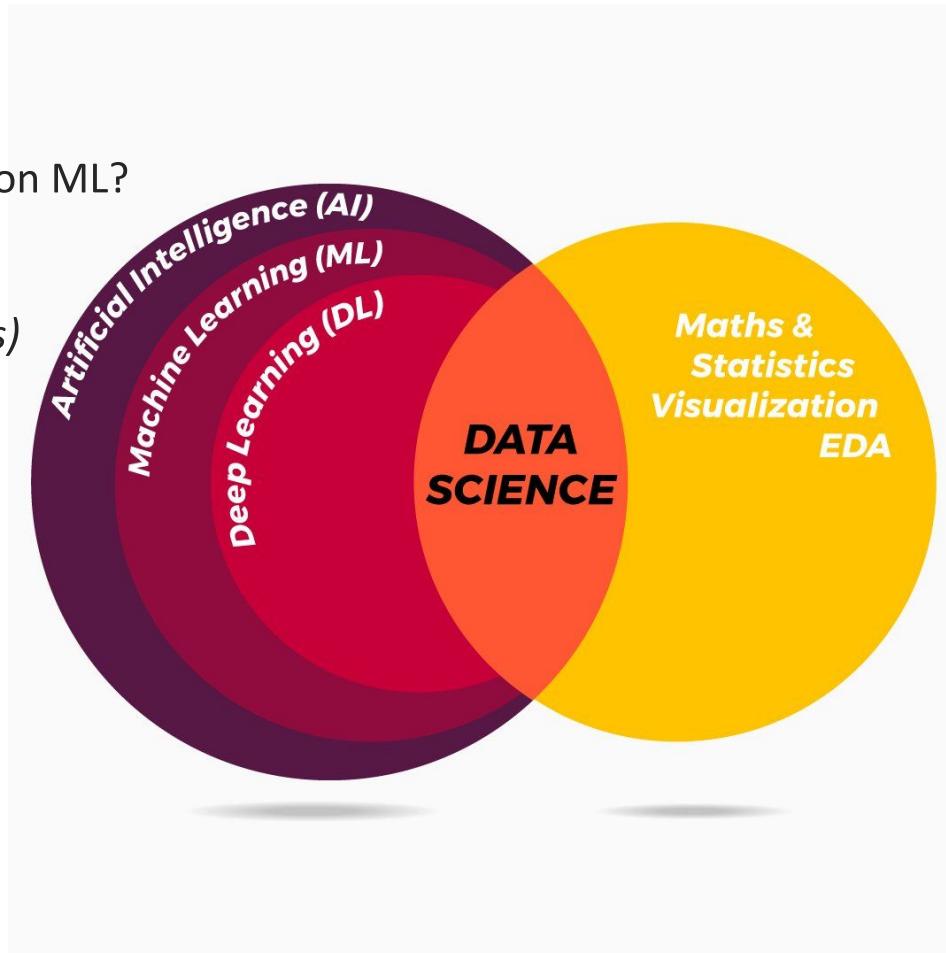
ML vs AI vs DS

¿Cuáles son algunos ejemplos de IA que no son ML?

- Sistemas basados en reglas
- Demostración automática (*proof solvers*)
- Optimización convexa o OR
- Visión por Computador

¿Y DS que no es ML?

Inferencia, Fourier, Métodos econométricos,
Biometría...



Introducción a Machine Learning

Sobre hombros de gigantes:

- Regla de Bayes (Bayes, 1763) de **probabilidad**
- Regresión por Mínimos Cuadrados (Gauss, 1795) de **astronomía**
- Máxima verosimilitud (Fisher, 1922) de **estadística**
- Redes neuronales artificiales (McCulloch/Pitts, 1943) de **neurociencias**
- Juegos Minimax (von Neumann, 1944) de **economía**
- Descenso de gradiente estocástico (Robbins/Monro, 1951) de **optimización**
- Búsqueda de costo uniforme (Dijkstra, 1956) de **algoritmos**
- Iteración de valores (Bellman, 1957) de **teoría de control**

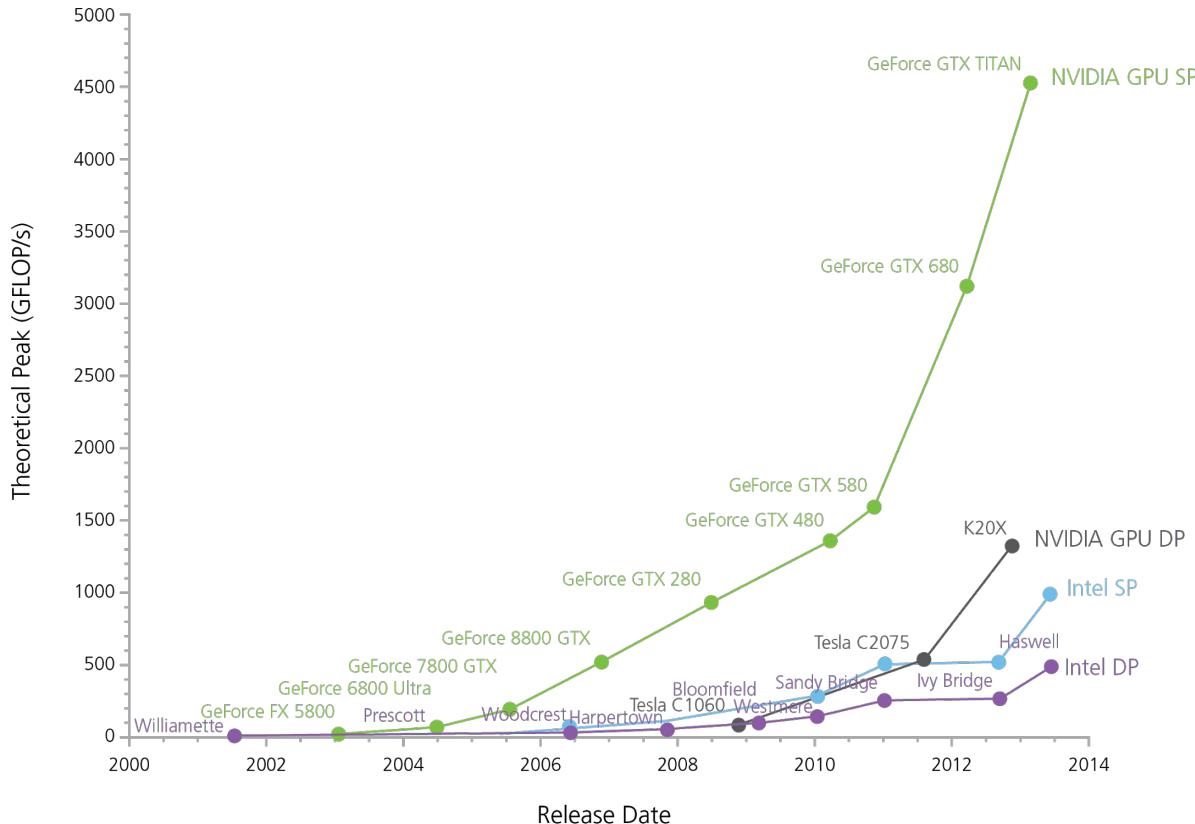
Disponibilidad de Datos

- Enorme cantidad de datos generados por minuto
- Posibilidad de trabajar sobre datos desestructurados
- Capacidad computacional

2019 *This Is What Happens In An Internet Minute*



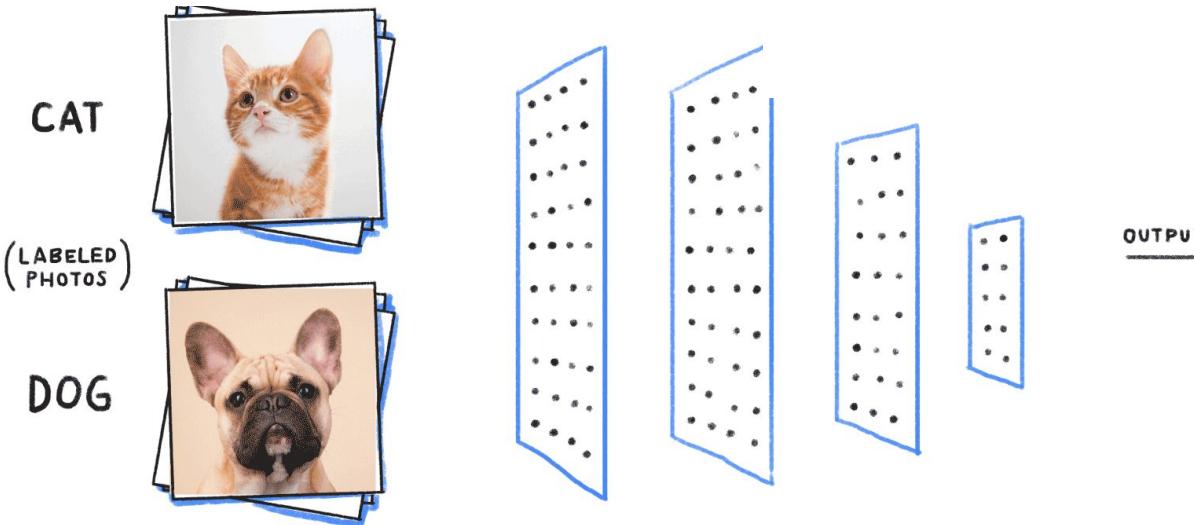
Cómputo



El proceso de aprendizaje

Machine Learning

Target



En clasificación, el proceso de ajuste consiste en hacer modificaciones a la función del modelo de forma tal de que, para cada input, el output se acerque al target correspondiente (resultado esperado).

Un modelo de ML es
una **función**
matemática

IA: hitos

Existen ya unos cuantos hitos acumulados de IA en los últimos años. En el programa Jeopardy! (IBM Watson, 2011), el juego Go (DeepMind, AlphaGo, 2016), el multijugador online masivo Dota 2 (OpenAI, 2019), Poker (CMU and Facebook, 2019).

También contamos con sistemas que rinden de manera excepcional en lectura y comprensión de lenguaje natural, reconocimiento del discurso, reconocimiento facial, o sistemas de imagen médicos.

Visión por Computadora

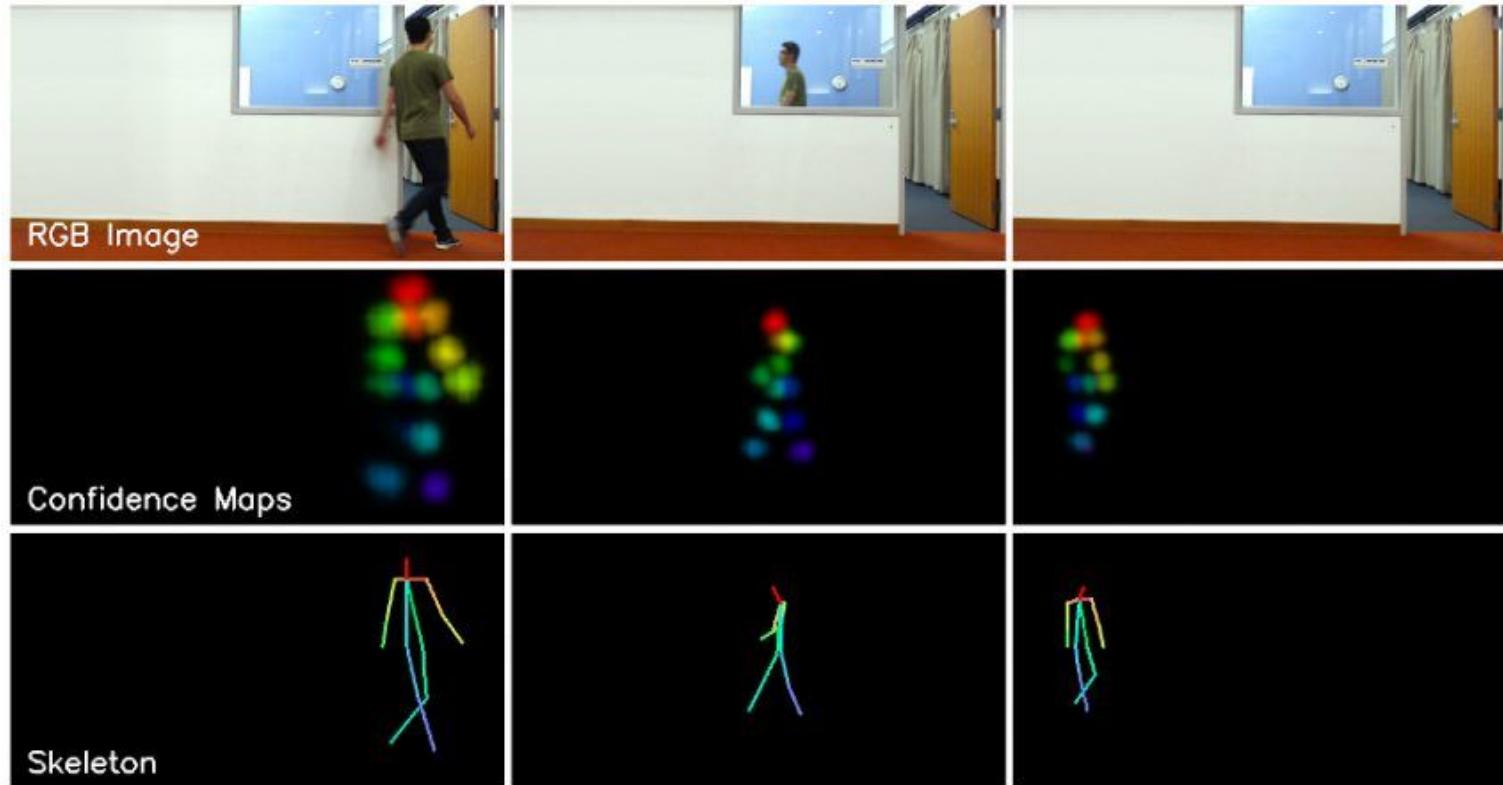
Ejemplos sci-fi

Robótica



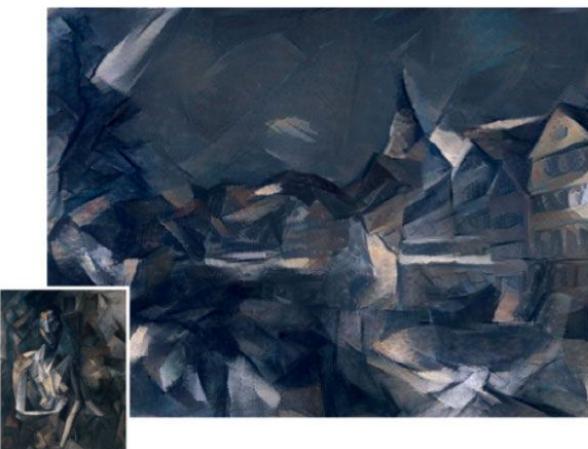
Ejemplos sci-fi

Ver a través de las paredes



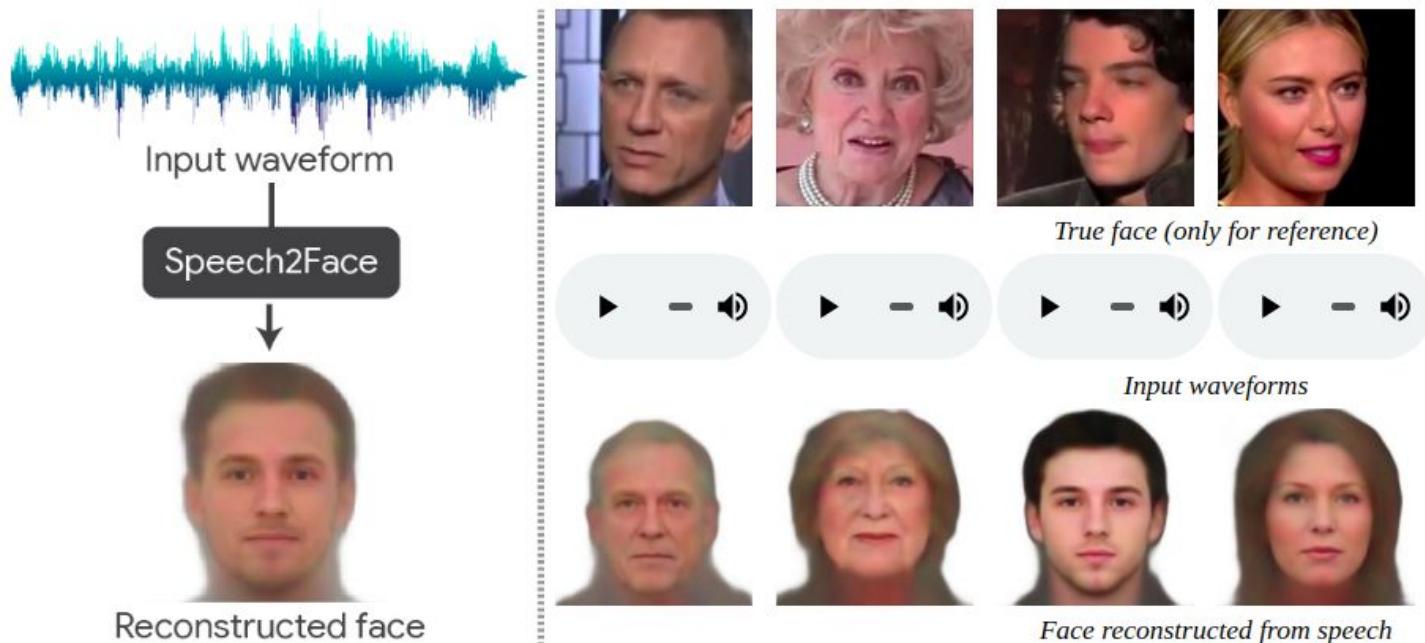
Ejemplos sci-fi

StyleTransfer



Ejemplos sci-fi

Speech2face



Ejemplos sci-fi

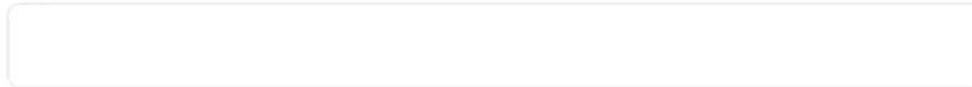
GPT-3

Describe a layout.

Just describe any layout you want, and it'll try to render below!

A div that contains 3 buttons each with a random color.

Generate



Ejemplos sci-fi

Neurociencias



Ejemplos sci-fi

Neurociencias

A T R C N S



Deep image reconstruction from human brain activity

Guohua Shen^{1,*}, Tomoyasu Horikawa^{1,*}, Kei Majima^{1,2,*}, and Yukiyasu Kamitani^{1,2}



Volviendo a poner los pies en la tierra...

Conceptos Básicos

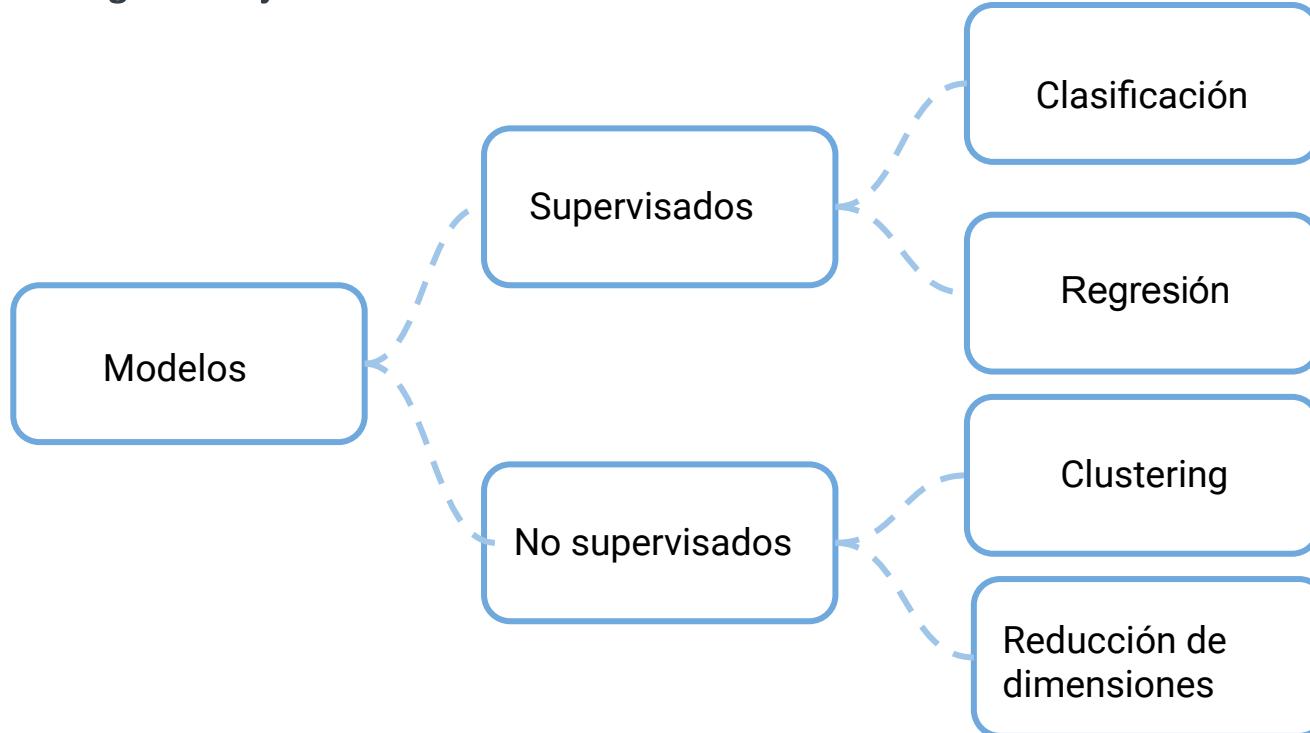
En nuestro caso la experiencia es un conjunto de **observaciones** con **atributos**. En algunos casos, estas observaciones tienen una variable **target** que queremos predecir. El target es optativo.

Para que los modelos de **Machine Learning** aprendan es necesario que la información esté organizada de manera matricial como a continuación:

upsell	user_id	mes	page_views	tiempo	compras previas
1	144332	5	23	15	3
0	634631	5	14	10	1
0	123126	5	10	8	0

Modelos supervisados y no supervisados

Clasificación según el objetivo



Feature Engineering



Los modelos de ML sólo son capaces de aprender de representaciones numéricas



¿Cómo podremos entonces trabajar con **texto, categorías o imágenes**?

Categorías: One Hot Encoding o variables dummies

Legible para un humano

Pet
Cat
Dog
Turtle
Fish
Cat

Legible para un algoritmo

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0



Texto: Bag of Word

Legible para un humano

Legible para un algoritmo

“Mary is hungry for apples.” →

MARY	IS	HUNGRY	HAPPY	FOR	APPLES	NOT	JOHN	HE
1	1	1	0	1	1	0	0	0
0	2	1	1	1	1	1	1	1

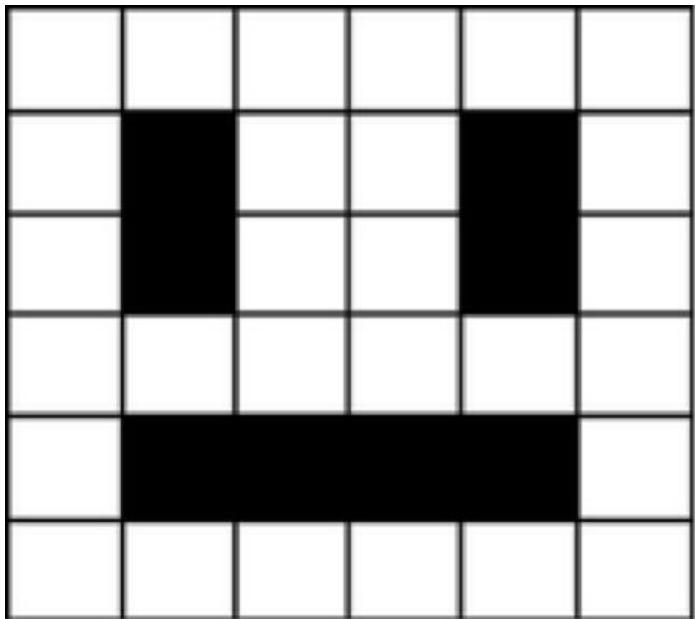
→ [1, 1, 1, 0, 1, 1, 0, 0, 0]

“John is happy he is not
hungry for apples.” →

→ [0, 2, 1, 1, 1, 1, 1, 1, 1]

Imágenes: Brillo en los canales R,G y B

Legible para un humano



Legible para un algoritmo

0	0	0	0	0	0
0	1	0	0	1	0
0	1	0	0	1	0
0	0	0	0	0	0
0	1	1	1	1	0
0	0	0	0	0	0

Modelos de Machine learning



Ya tengo mis datos representados como vectores numéricos o “features” y etiquetas o “labels”

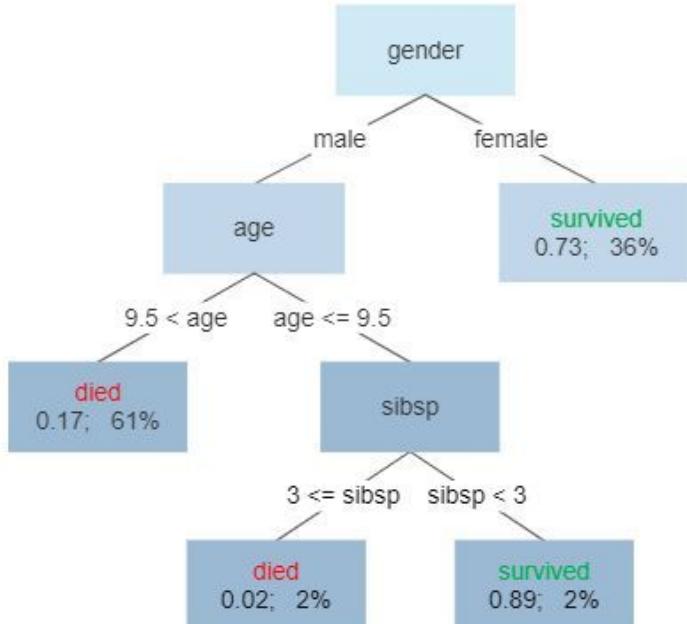


¿Cómo entreno un modelo que capture las relaciones entre los mismos?

Modelos supervisados

Modelos basados en árboles: en general los mejores predictores

Survival of passengers on the Titanic



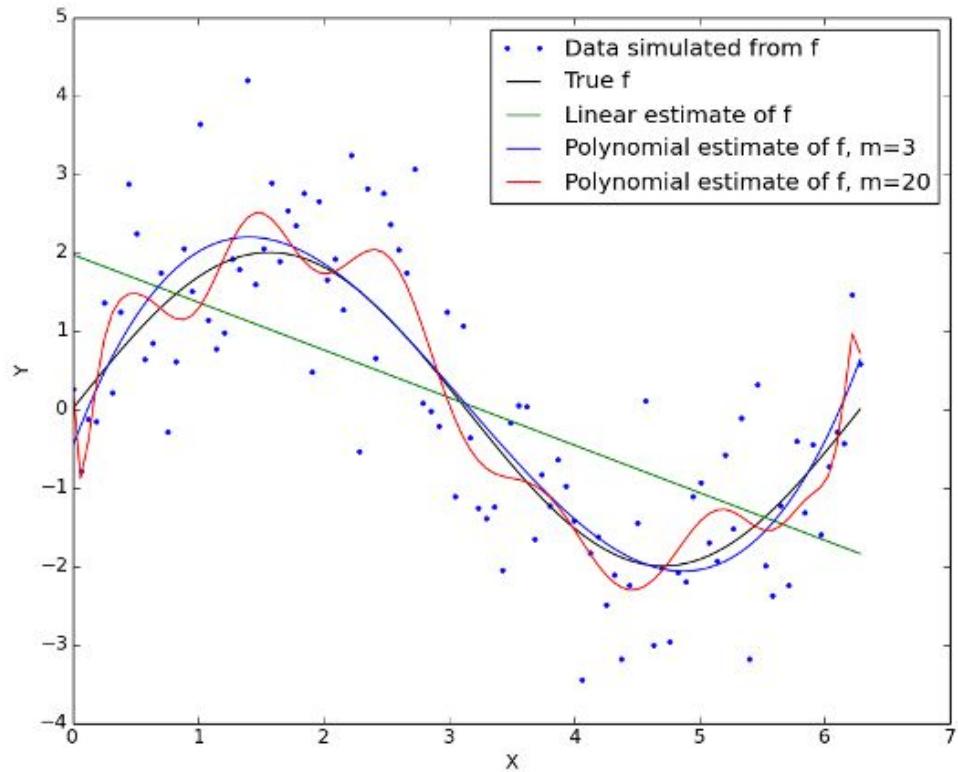
Los modelos de Machine Learning más performantes, se basan en “ensamblar” muchos árboles simples. Sirven tanto para **clasificación** como para **regresión**.

- **Random Forest**
- **XGBoost**
- **LightGBM**
- **Tree Gradient Boosting**

Modelos supervisados

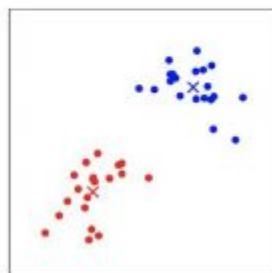
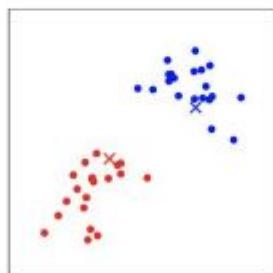
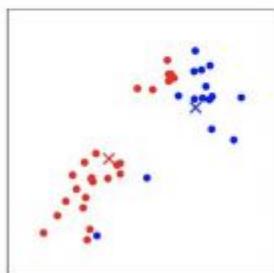
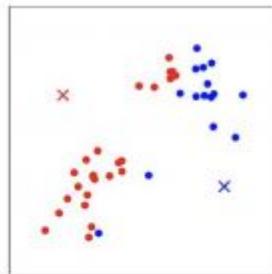
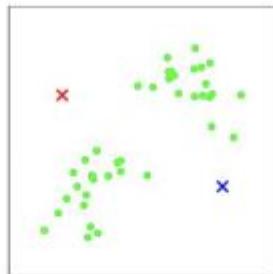
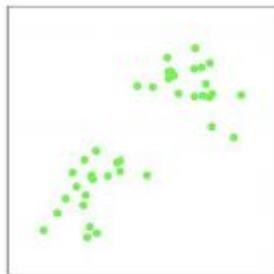
Otros modelos

- Regresión lineal
- Regresión logística
- KNN (K nearest neighbours)
- Y muchos más...



Modelos no supervisados

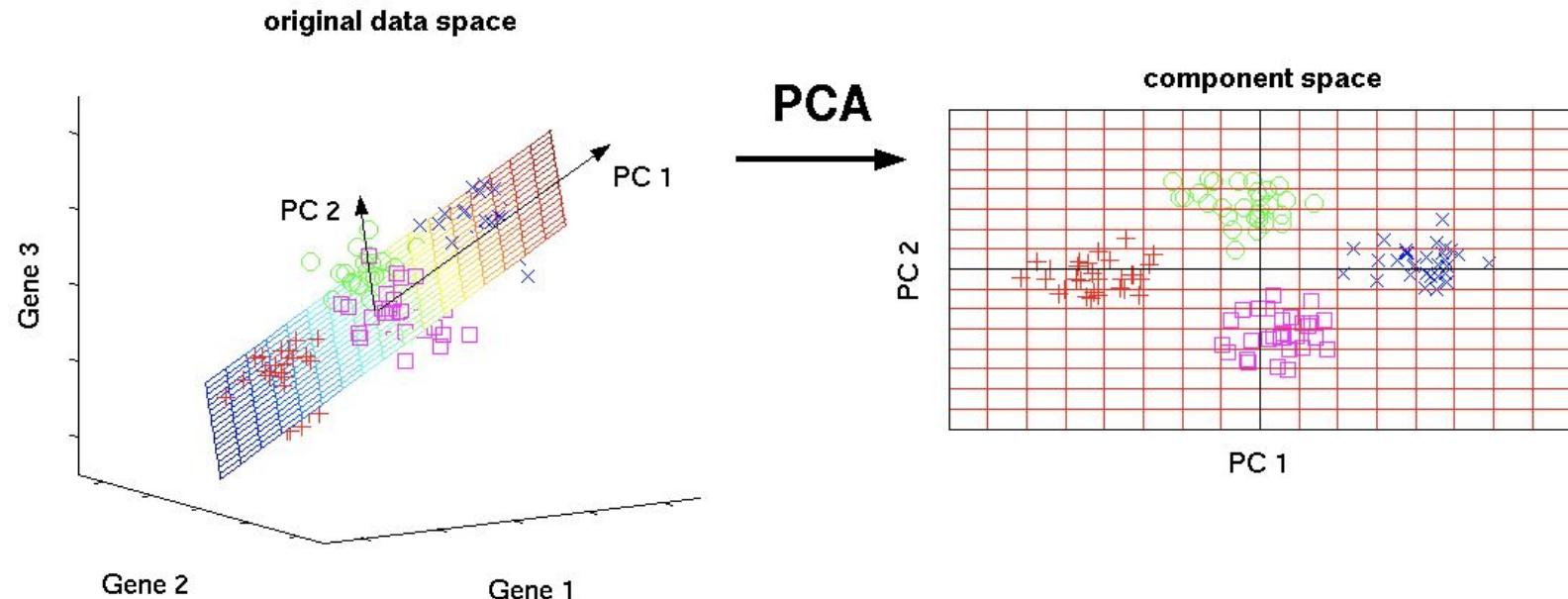
KMeans (clustering)



- Asignar centroides al azar
- Calcular qué punto pertenece a cada cluster
- “Promediar” todos los puntos en cada dimensión para recalcular el centroide
- Repetir hasta que los puntos dejen de cambiar de cluster.

Modelos no supervisados

PCA (Reducción de dimensiones)



Modelos de Machine learning



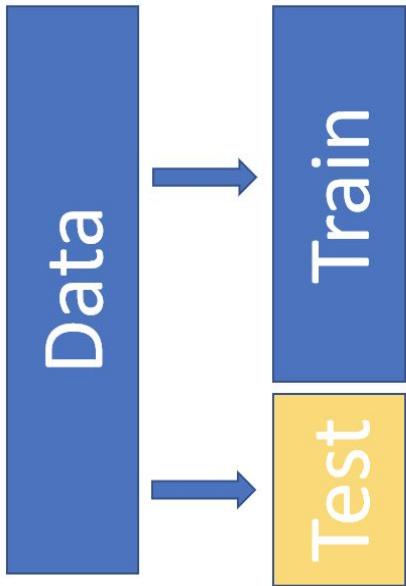
Ya tengo un modelo que me permite predecir, clusterizar o reducir dimensiones



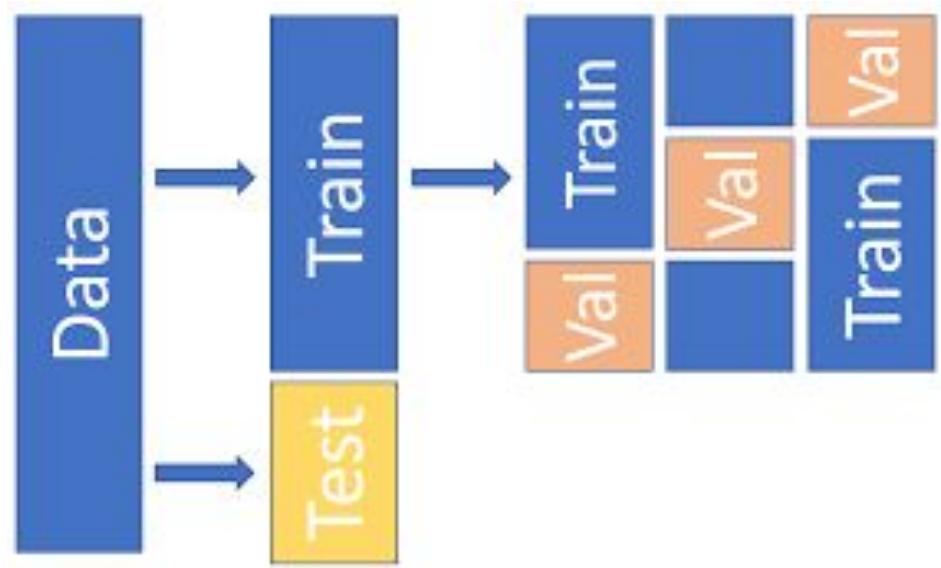
¿Cómo puedo saber cuán bien va a funcionar sobre datos nuevos?

Holdout sets para la evaluación

Train/Test Split

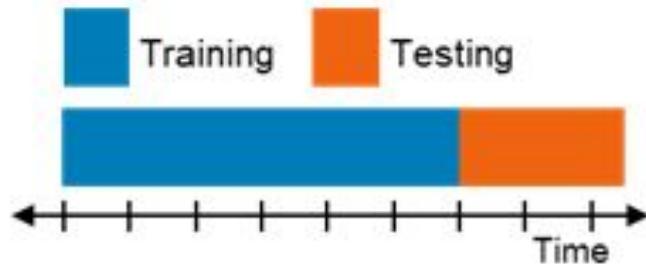


Cross Validation

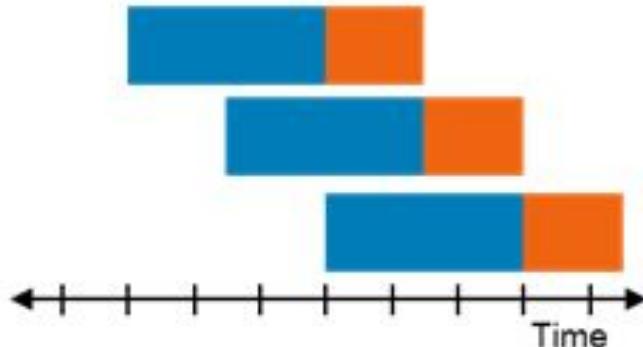


Holdout sets para la evaluación (series de tiempo y datos de panel)

Time-based Estimation



Time-based cross-validation



¡Cuidado! Es imposible usar el futuro para predecir el presente.

Cuando los modelos incorporan información que naturalmente no debería estar disponible este problema se llama "**information leak**".

Aquí un caso famoso de information leak que tomó conocimiento público:

<https://liaa.dc.uba.ar/es/sobre-la-prediccion-automatica-de-embarazos-adolescentes/>

Métricas de evaluación en problemas supervisados

Clasificación

Matriz de confusión

	Predicted: NO	Predicted: YES
Actual: NO	TN = ??	FP = ??
Actual: YES	FN = ??	TP = ??

$$Accuracy = \frac{\text{Casos Correctamente clasificados}}{\text{Casos Totales}}$$

$$Recall = TP / TP + FN$$

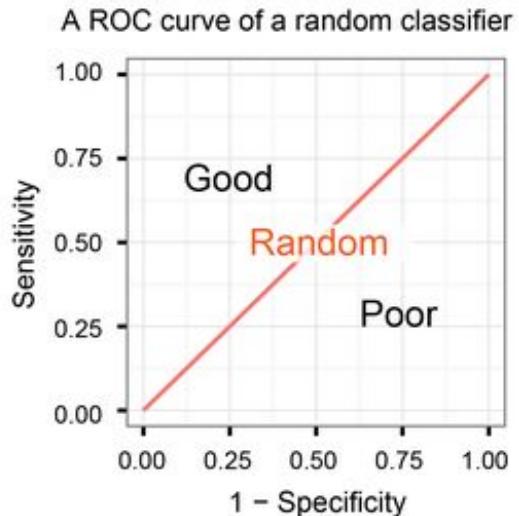
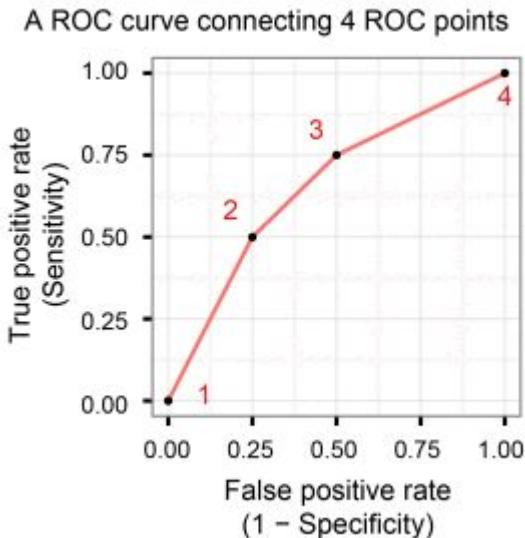
$$Precision = TP / TP + FP$$

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

$$F score = 2 \frac{precision * recall}{precision + recall}$$

Métricas de evaluación en problemas supervisados

Clasificación **binaria**: auc (área debajo de la curva)



Modelo perfecto:
auc=1

Ordenamiento de los casos al azar:
auc=0.5

Modelo que predice exactamente lo contrario a lo real:
auc=0

Métricas de evaluación en problemas supervisados

Regresión

$$MAE = \frac{1}{n} \sum |y - y_{predicha}|$$

$$MSE = \frac{1}{n} \sum (y - y_{predicha})^2$$

$$R^2 = \sum \frac{(y - y_{predicha})^2}{(y - y_{promedio})^2}$$

En resumen:

Para entrenar un modelo se necesita:

- Crear buenas características numéricas o “features” que representen las observaciones
- Conservar un conjunto de datos para evaluación
- Elegir un algoritmo que resuelva nuestro problema
- Evaluar los resultados
- Desplegar el modelo “en producción”

Casos de negocio: problemas supervisados

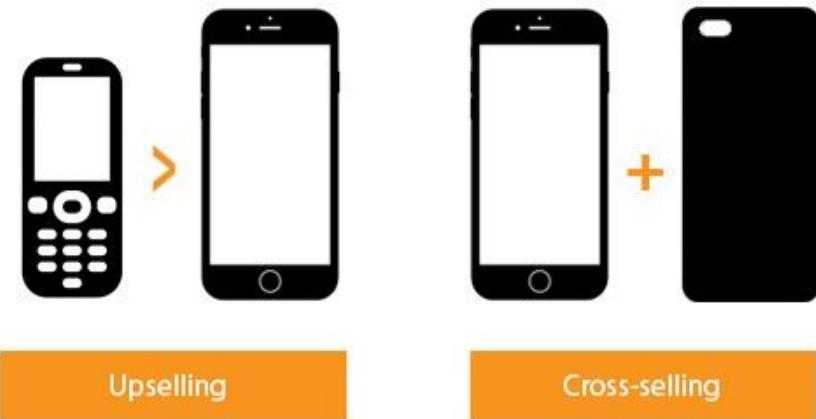
Predicción de propensión a conversión online



- **Objetivo:** predecir qué tan probable es que un visitante a una app realice una conversión. Ésta puede ser una compra online, un clic, completar un formulario, etc.
- Esto se usa para decidir cuánto pagar en una subasta, asignar un precio o producto dinámicamente, etc.

Casos de negocio: problemas supervisados

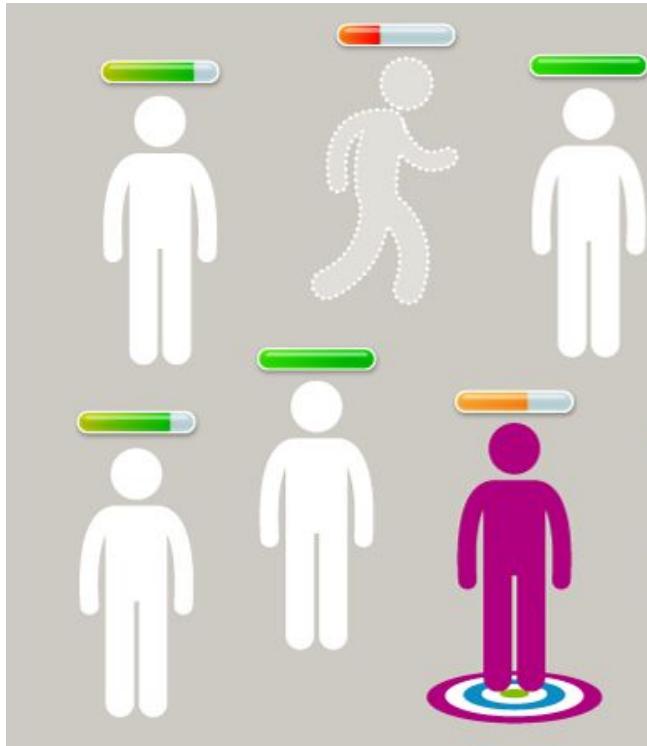
Modelos de upselling/cross-selling



- **Objetivo:** predecir la probabilidad de que un cliente compre un producto más caro (upselling) o complementario (cross-selling).
- Esto se puede integrar a un CRM o una herramienta de marketing para realizar campañas automáticamente.

Casos de negocio: problemas supervisados

Predicción de churn



- **Objetivo:** predecir la probabilidad de que un cliente se dé de baja en determinado período.
- Sabiendo quiénes son los más propensos podemos generar un incentivo para impedirlo.

Casos de negocio: problemas supervisados

Detección de fraude



- **Objetivo:** predecir la probabilidad de que una transacción sea fraudulenta

Casos de negocio: problemas supervisados

Sistemas de recomendación



- **Objetivo:** optimizar qué productos ofrecer en una plataforma de ventas online

amazon
NETFLIX

Casos de negocio: problemas no supervisados

Segmentación automática

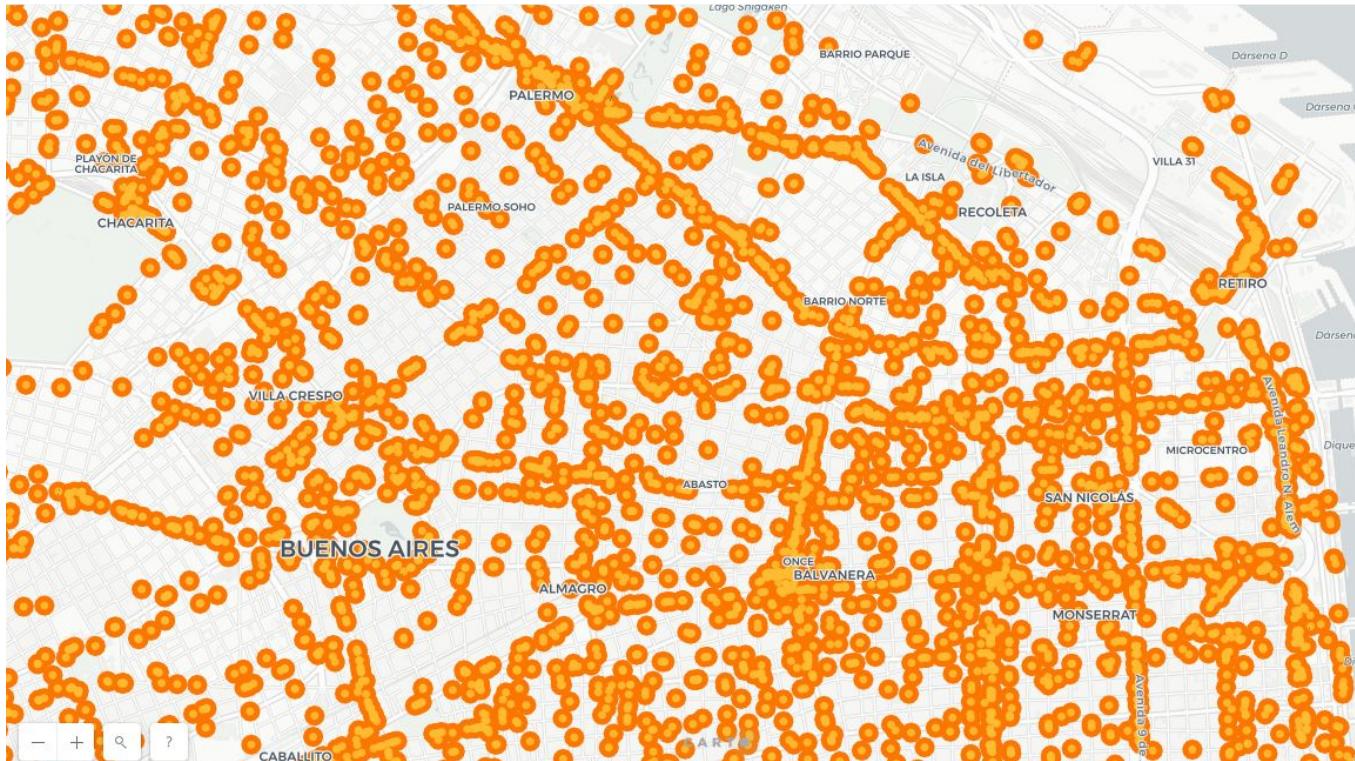
- Generación de segmentos a partir de información demográfica y de comportamiento.
- Con esta técnica se pueden generar K segmentos automáticamente, de modo de accionar de manera distinta sobre cada uno de ellos.



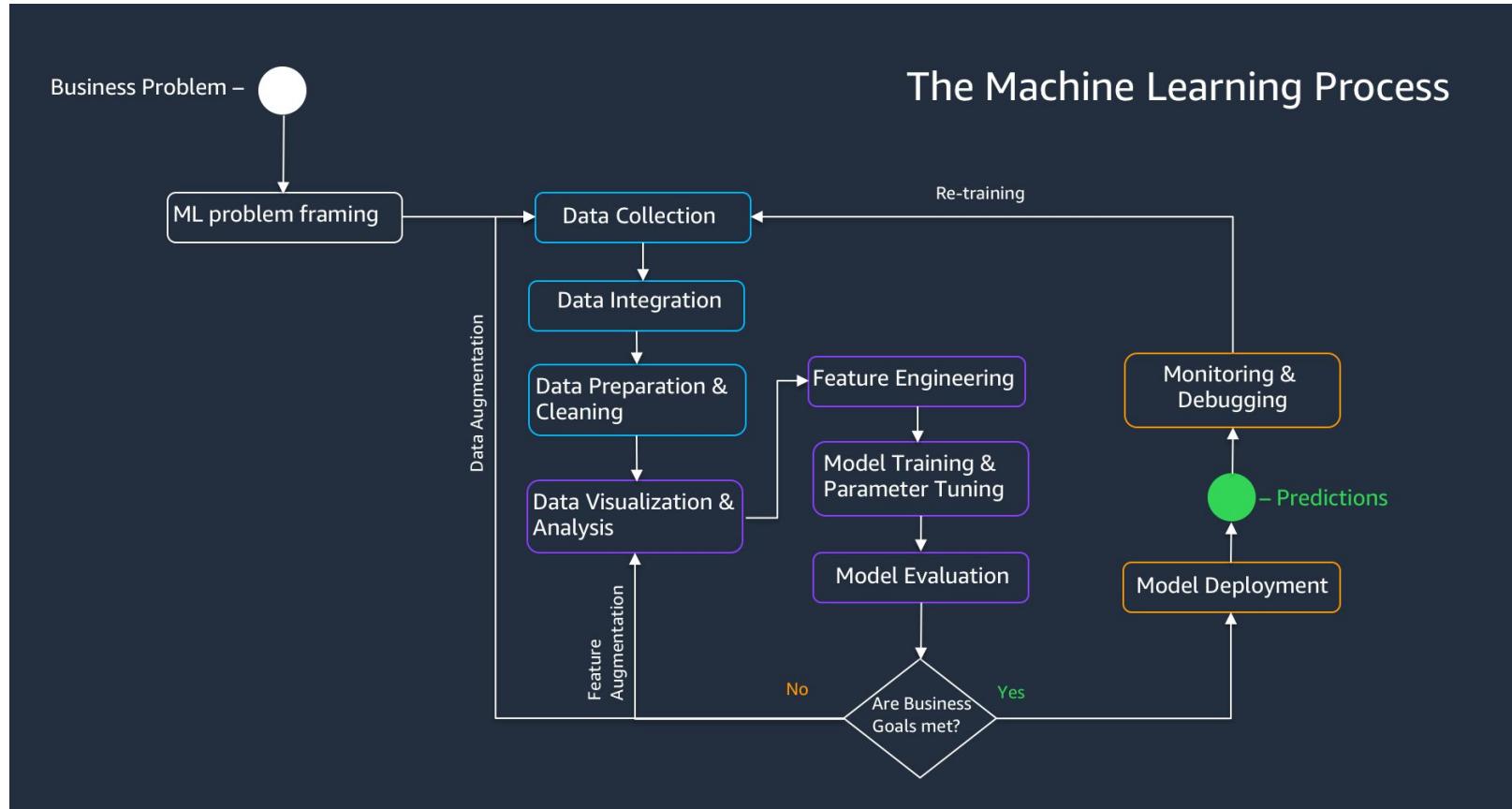
Casos de negocio: problemas no supervisados

Clustering geográfico I

- Encontrar patrones geográficos automáticamente.
- Por ejemplo: paradas de colectivos, transacciones de clientes,etc.



El proceso real de los productos basados en ML



Data Stack

Equipo de data



Ejemplo práctico

Sklearn

