

Contenido



I. Ejemplo práctico de Aprendizaje de Máquina

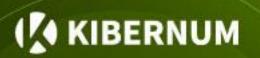
- Tipos de problema que se resuelven con aprendizaje de máquina
- Aplicación de un algoritmo de clasificación para resolver un problema

II. Fundamentos de la Ciencia de Datos y sus aplicaciones

- ¿Qué es la ciencia de datos?
- Rol y habilidades del científico de datos
- Problemas que resuelve la ciencia de datos



La Flor Iris











Iris



Iris es un género de plantas bulbosas de la familia Iridaceae con vistosas flores, cuyo nombre deriva del latín arco iris, refiriéndose a la extensa variedad de colores florales que poseen sus muchas especies y cultivares de jardín. Wikipedia

Nombre científico: Iris

Reino: Plantae

Orden: Asparagales

Clase: Liliopsida

Género: Iris; L., 1753 Categoría: Género

Clasificación inferior



Iris sibirica



Iris pseudaco...



Iris ensata



Iris pallida

Ver 35 más



lris versicolor

Variedades de Iris



Iris Setosa



Iris Versicolor



Iris Virginica



Formulación de la Pregunta

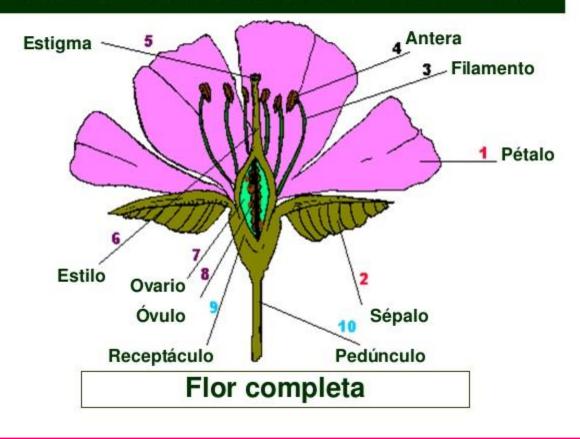
¿Cómo distinguir una especie de Iris?



En Ciencia de Datos, la **formulación de la pregunta** es la etapa más importante del proceso



MORFOLOGÍA DE LA FLOR



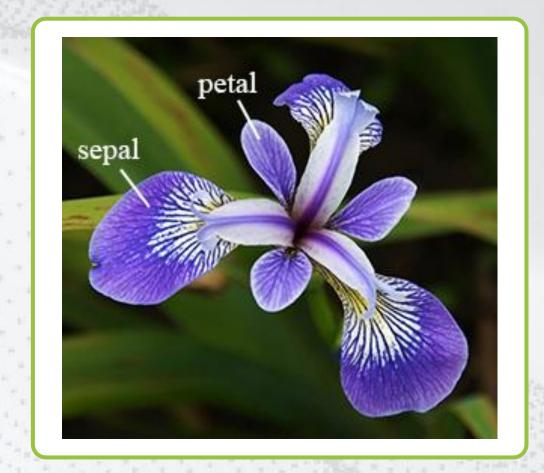
Conocimiento del Dominio del Problema



Identificar Características

Posibles características:

- Color del pétalo
- Color del sépalo
- Cantidad de colores
- Cantidad de pétalos
- Cantidad de sépalos
- Medidas del pétalo
- Medidas del Sépalo
- Largo del tallo
- Otras?

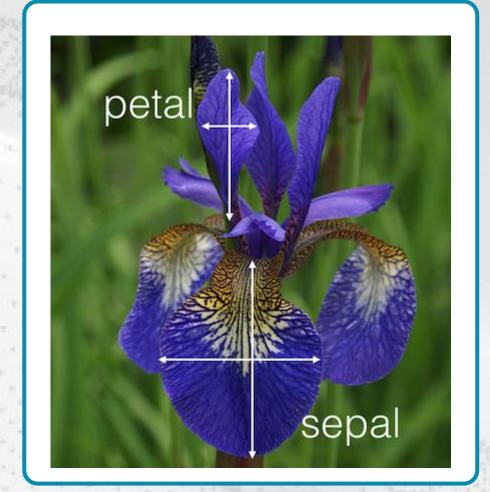




Nos interesa saber qué características permitirían distinguir una especie de otra

Hipótesis de Trabajo

- Vamos a postular como hipótesis que las siguientes características podrían permitirnos separar las distintas especies de Iris.
 - Largo sépalo
 - Ancho sépalo
 - Largo pétalo
 - Ancho pétalo





La **Definición de Características** es vital en el resultado de un modelo de aprendizaje de máquina.

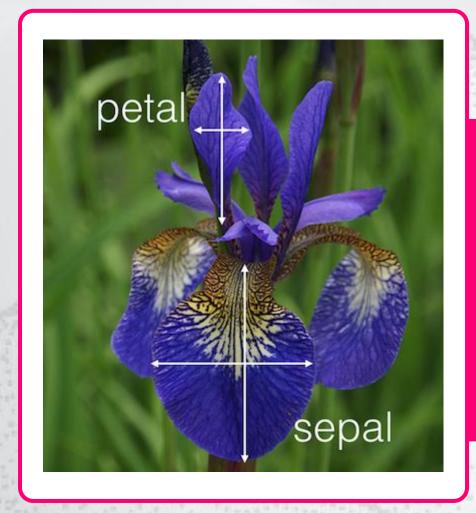


Recolección de Datos

En nuestro caso, recorreremos varios viveros de la ciudad, y haremos lo siguiente:

- Tomamos una especie de iris y medimos sus características.
- Tomamos muchas muestras de cada especie.





Recolección de Datos

Por ejemplo, tomaremos una instancia de flor en el vivero y registraremos lo siguiente:

Medición 1

• Largo pétalo: 1.4 cm

• Ancho pétalo: 0.2 cm

Largo sépalo: 7.0 cm

Ancho sépalo: 3.2 cm

• Especie: Versicolor



Preparar Set de Datos

 Repetimos las mediciones con muchas instancias de iris de distintas especies y tabulamos la información recolectada de esta forma:



La mayoría de las veces los datos vienen sucios y hay que lidiar con ellos (**Data Wrangling**) antes de utilizarlos.

Sepal length +	Sepal width +	Petal length +	Petal width \$	Species +
5.1	3.5	1.4	0.2	I. setosa
4.9	3.0	1.4	0.2	I. setosa
4.7	3.2	1.3	0.2	I. setosa
4.6	3.1	1.5	0.2	I. setosa
7.0	3.2	4.7	1.4	I. versicolor
6.4	3.2	4.5	1.5	I. versicolor
6.9	3.1	4.9	1.5	I. versicolor
5.5	2.3	4.0	1.3	I. versicolor
6.5	2.8	4.6	1.5	I. versicolor
7.2	3.0	5.8	1.6	I. virginica
7.4	2.8	6.1	1.9	I. virginica
7.9	3.8	6.4	2.0	I. virginica
6.4	2.8	5.6	2.2	I. virginica

(Valores en cms)

Preparar set de datos

Afortunadamente, este trabajo de recolección de información ya fue hecho y tenemos acceso a este set de datos desde la siguiente ubicación web.

https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data



About Citation Policy Donate a Data S

Repository

● We

View ALL

Iris Data Set

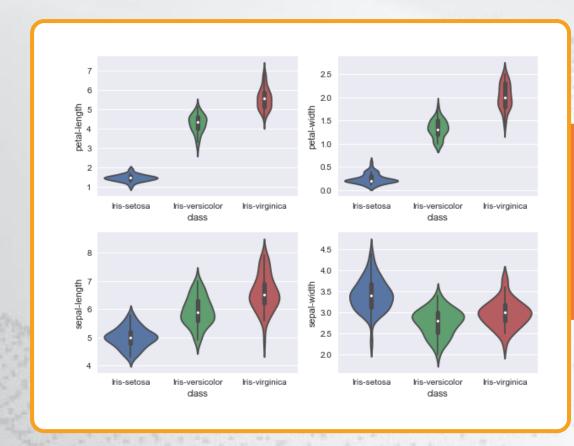
Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936



Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	1501408





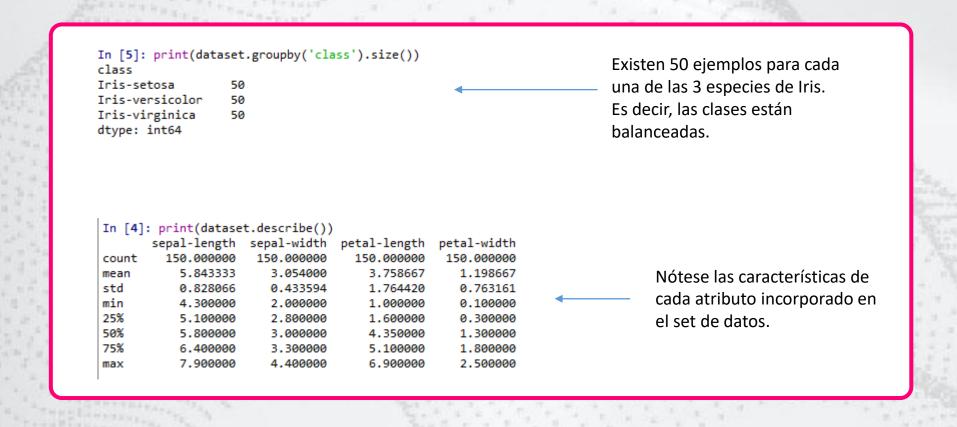
El EDA (Exploratory Data Analysis) está referido a los procesos críticos de realizar investigaciones sobre la data, ya sea para descubrir patrones, detectar anomalías, testear hipótesis y revisar supuestos, con la ayuda de la estadística descriptiva y el análisis visual.

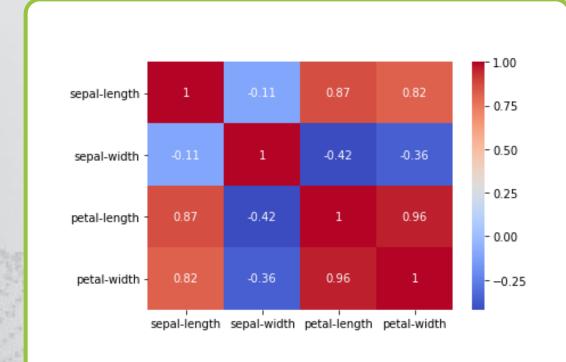


En la ciencia de datos, es fundamental el lograr un **entendimiento profundo** de los datos desde el inicio.

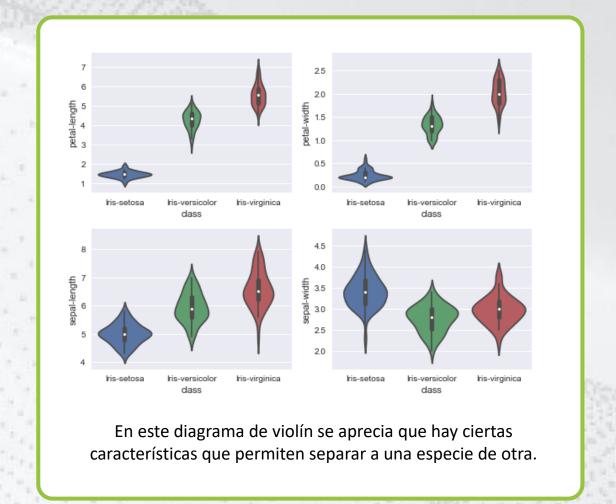


Acá se observa cómo la estadística descriptiva ayuda al entendimiento de los datos.

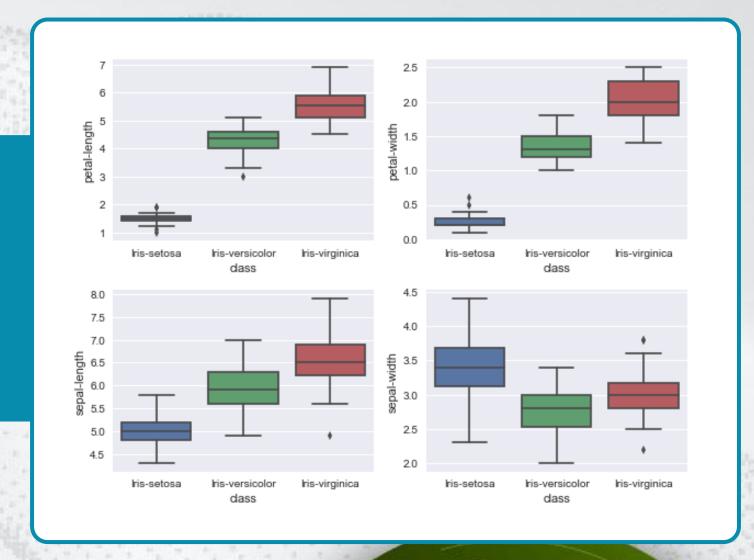




Esta es una matriz de correlación entre los distintos features (características) del set de datos, visualizado en un mapa de calor.



- Igual que en el caso anterior, pero visualizando un diagrama de caja y bigote.
- Note, por ejemplo, que la especie Setosa tiene un largo de pétalo más pequeño, mientras que la especia Virgínica tiene los mayores largos de tamaños de pétalo. Algo similar sucede con el ancho.



Mediciones

Etiquetas





Sepal length \$	Sepal width \$	Petal length +	Petal width \$	Species +
5.1	3.5	1.4	0.2	I. setosa
4.9	3.0	1.4	0.2	I. setosa
4.7	3.2	1.3	0.2	I. setosa
4.6	3.1	1.5	0.2	I. setosa



Nueva Data







7.2	3.0	5.8	1.6	
7.4	2.8	6.1	1.9	_
7.9	3.8	6.4	2.0	
6.4	2.8	5.6	2.2	

Formulación de un Modelo Predictivo

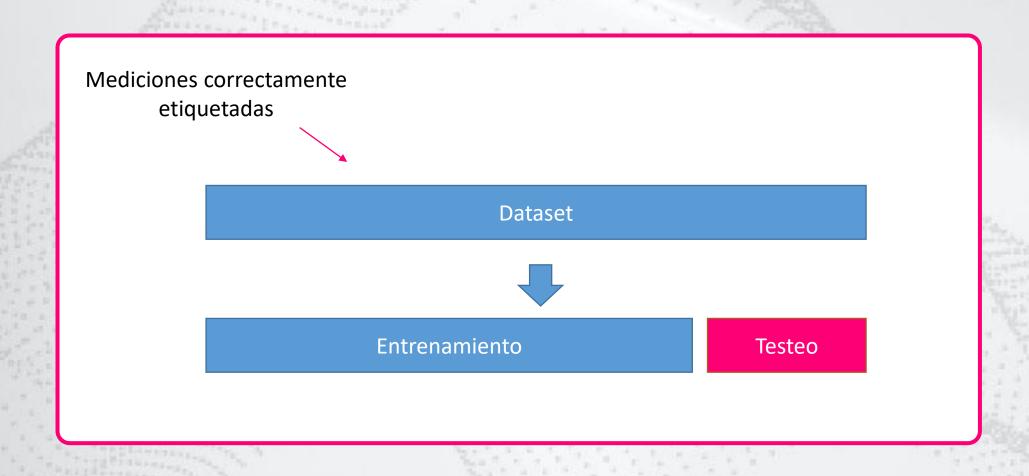
A partir de los datos de mediciones correctamente etiquetadas, entrenaremos un algoritmo que aprenda de los datos existentes con el objeto de etiquetar nuevos datos.

Es decir, el modelo entrenado nos permitirá realizar predicciones.



Validación Cruzada

Para poder medir el poder predictivo del modelo de aprendizaje, tomaremos 80% de la muestra (de forma aleatoria) para entrenar el algoritmo y el 20% para validarlo. A este proceso se le conoce como Cross Validation.



Entrenamiento

Existe una gran variedad de algoritmos de aprendizaje de máquina que nos pueden ayudar a resolver este problema. En este caso, debemos resolver una tarea de "Clasificación", puesto que necesitamos que asigne una "clase" o "categoría" a cada nueva medición.

Para este ejemplo, utilizaremos los siguientes algoritmos:

- «Logistic Regression»
- «Random Forest»
- «Support Vector Machine»
- «Naive Bayes»

Posteriormente, evaluaremos cuál tuvo mejor performance en la resolución del problema.



Entrenamiento

Para el entrenamiento, podemos utilizar distintas herramientas y marcos de trabajo. A lo largo de este curso, utilizaremos dos alternativas:



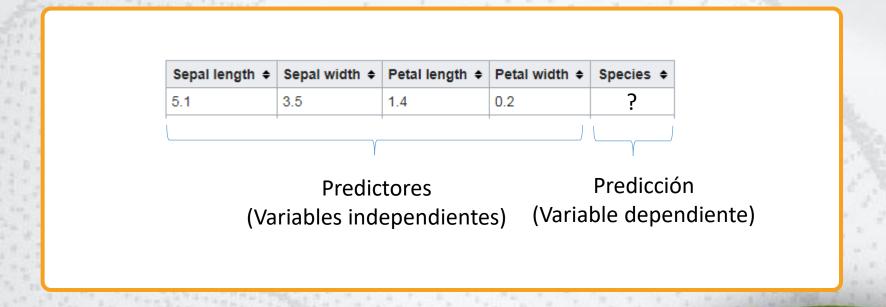
Directo al código, para quienes la programación y las librerías no es un problema.



Ambiente visual, permite comprender mejor los conceptos sin entramparse en la programación.

Hagamos Predicciones

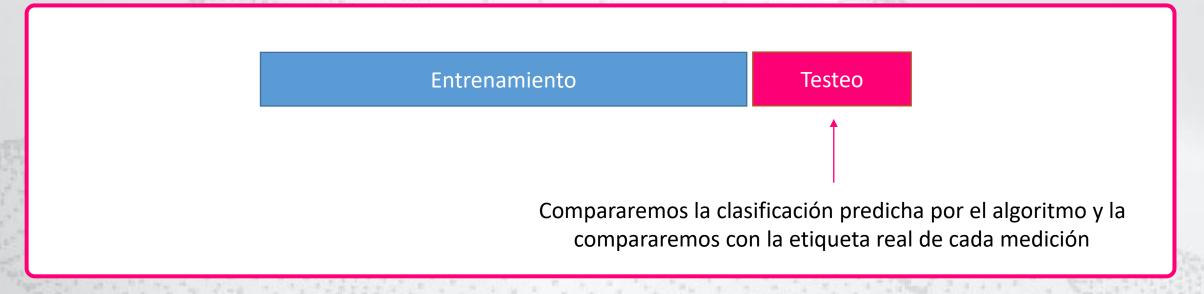
Una vez entrenado el algoritmo, tomaremos el «Set de Validación», y realizaremos algunas predicciones sobre la especie de iris.





Evaluemos el Algoritmo

También evaluaremos el desempeño del algoritmo en el set de Test, para así medir su poder predictivo.





Una métrica utilizada para evaluar el desempeño de un algoritmo es el **Accuracy**, que es la proporción entre aciertos y errores.

Medimos Accurracy Logistic Reg.

Comparamos en el «Set de Validación» los valores predichos con los valores verdaderos para tener una medida de exactitud.

Medimos Accurracy Random Forest

Ahora repetimos el proceso con el algoritmo Random Forest.

[[13 0 0] [0 15 1] [0 0 9]]

Accuracy: 0.973684210526

	precision	recall	f1-score	support
Iris-setosa	1.00	1.00	1.00	13
Iris-versicolor	1.00	0.94	0.97	16
Iris-virginica	0.90	1.00	0.95	9
avg / total	0.98	0.97	0.97	38

Medimos Accurracy SVM

Ahora repetimos el proceso con el algoritmo Support Vector Machine.

```
0 9]]
Accuracy: 0.973684210526
                             recall f1-score
                precision
                                                support
    Iris-setosa
                                         1.00
                                                     13
                     1.00
                               1.00
Iris-versicolor
                     1.00
                               0.94
                                         0.97
                                                     16
 Iris-virginica
                               1.00
                                         0.95
                     0.90
    avg / total
                     0.98
                               0.97
                                         0.97
                                                     38
```

Medimos Accurracy Bayes

Ahora repetimos el proceso con el algoritmo Naive Bayes.

```
[[13 0 0]
[ 0 16 0]
[ 0 0 9]]

Accuracy: 1.0

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 13
Iris-versicolor 1.00 1.00 1.00 16
```

1.00

1.00

1.00

1.00

1.00

1.00

38

Iris-virginica

avg / total





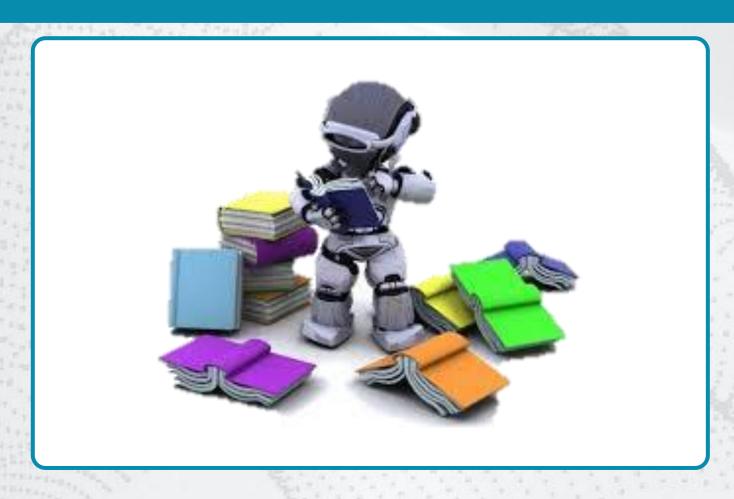






¿Qué hicimos?

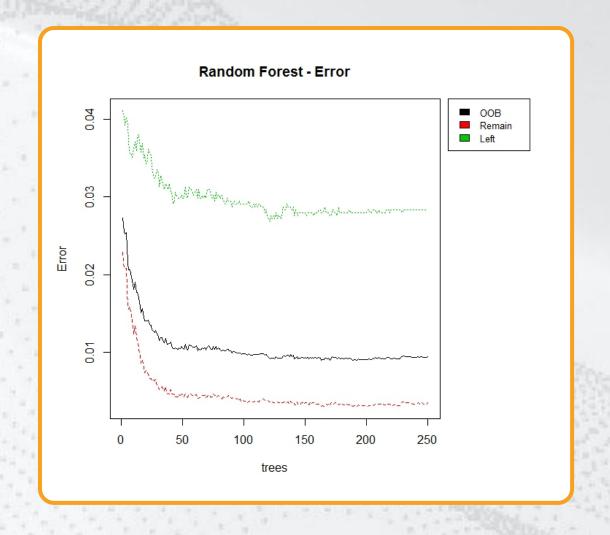
Hemos empleado algoritmos de **Machine Learning** para elaborar un modelo de clasificación que permita realizar predicciones.



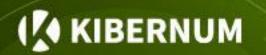
¿Cómo podríamos mejorar?

Los modelos de ML podrían mejorarse principalmente:

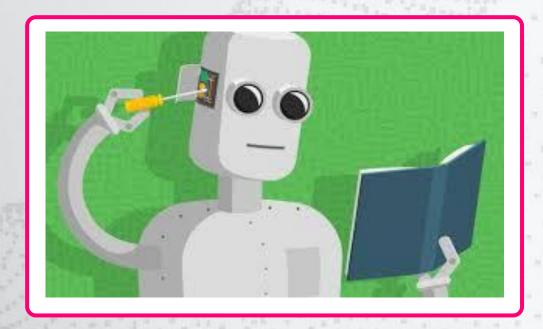
- Agregando más ejemplos al dataset
- Agregando nuevas características o predictores (o eliminando algunas)
- Optimizando el algoritmo (tunning)
- Utilizando otros algoritmos



Qué es Machine Learning



¿Qué es Machine Learning?



 Machine Learning es una rama de la inteligencia artificial en donde se le brinda a los computadores la habilidad de aprender sin ser explícitamente programados.

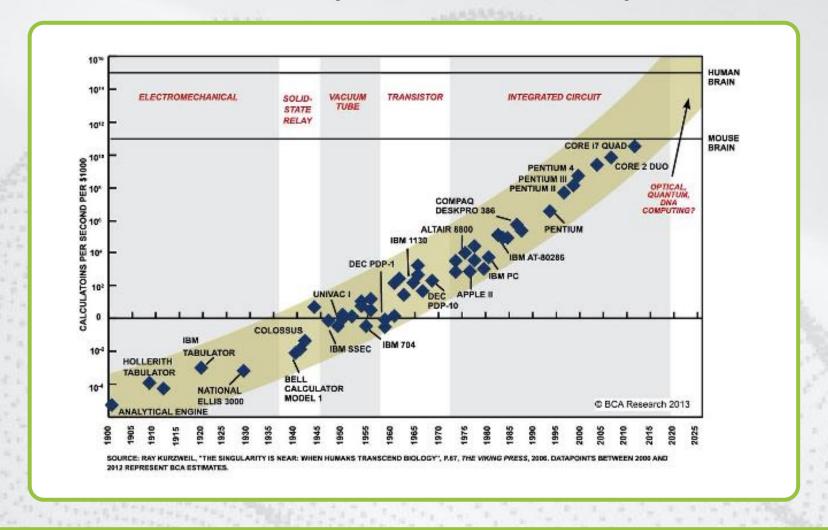
(Arthur Samuel, 1959)

• Se dice que un computador aprende de la experiencia E, con respecto a una tarea T y una medida de performance P, si su performance en T, medido por P, mejora con la experiencia E.

(Tom Mitchell, 1998)



Evolución Capacidad de Cómputo



El concepto de Machine Learning data de los años 50s, pero estos últimos 10 años ha cobrado mayor relevancia por el aumento de la capacidad de cómputo y de almacenamiento.

Análisis Exploratorio

- Se dice que un computador aprende de la experiencia E, con respecto a una tarea T y una medida de performance P, si su performance en T, medido por P, mejora con la experiencia E.
- Suponga que su programa de Email lo ve a usted marcando cuáles son los mails que corresponden a spam, y basado en esto, aprende a filtrar de mejor manera los correos. ¿Cuál es la tarea T en este caso?

- Clasificar los emails como spam o no spam.
- Observar las marcas de spam o no spam que aplicas sobre los correos.
- La cantidad o porcentaje de emails correctamente clasificados como spam/no spam.

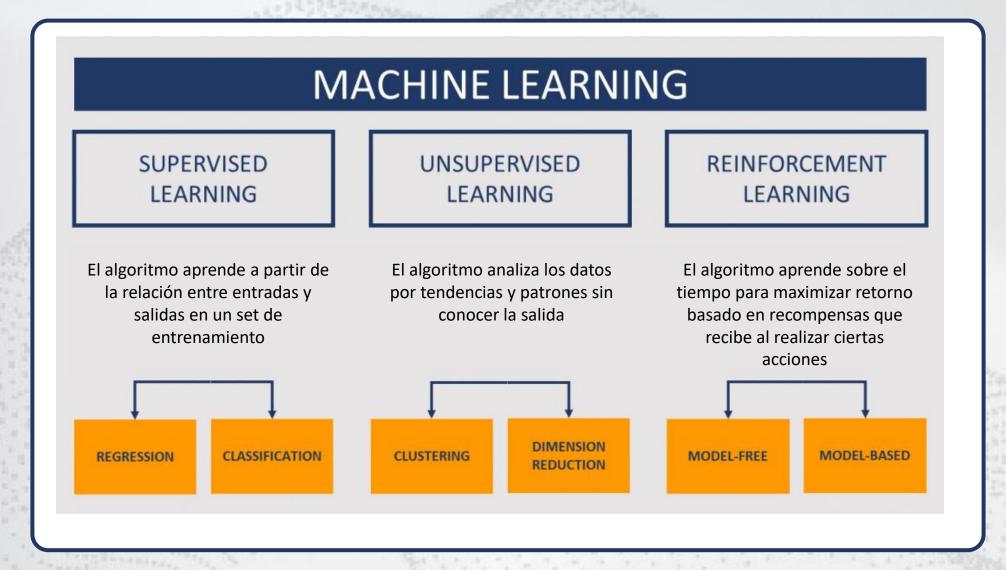


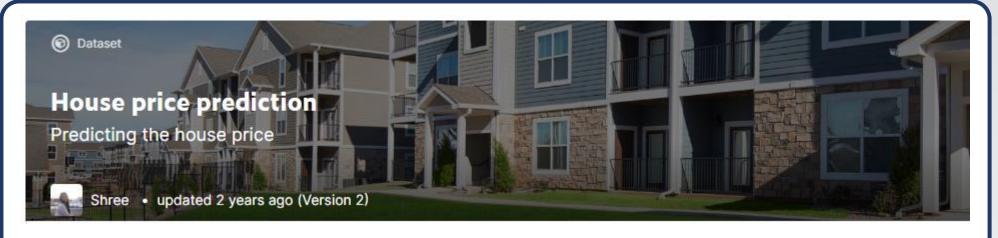
Aplicaciones de ML

- Detección de fraudes
- Detección de intrusos en la red
- Modelos de pricing
- Evaluación de riesgo en créditos
- Resultados de búsqueda
- Predicción de falla en equipos
- Publicidad en la red

- Sistemas de recomendación
- Segmentación de clientes
- Análisis de sentimiento
- Reconocimiento de imágenes
- Predecir la fuga de clientes
- Filtros anti spam
- Modelos financieros

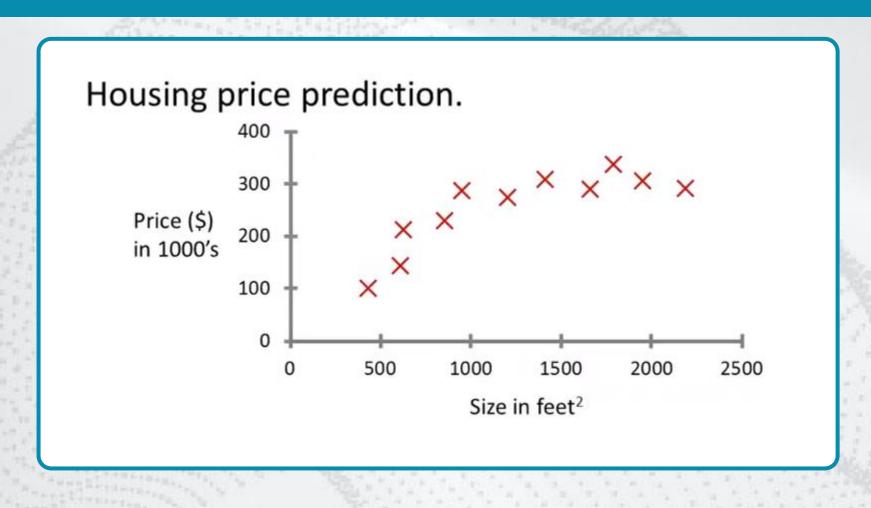
Tipos de ML



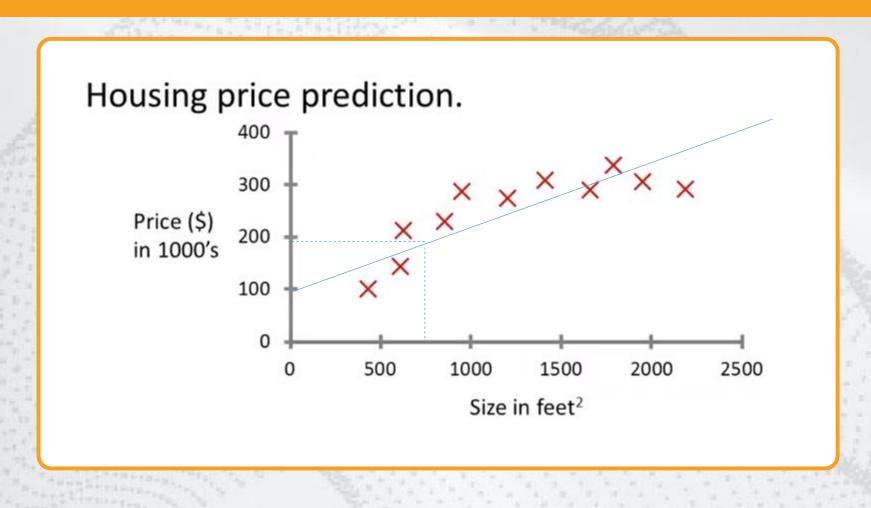


A	В	С	D	E	F	G	Н	1	J	K	L	M
date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basen	yr_built y
02-05-2014 0:00	313000.0	3.0	1.5	1340	7912	1.5	0	0	3	1340	0	1955
02-05-2014 0:00	2384000.0	5.0	2.5	3650	9050	2.0	0	4	5	3370	280	1921
02-05-2014 0:00	342000.0	3.0	2.0	1930	11947	1.0	0	0	4	1930	0	1966
02-05-2014 0:00	420000.0	3.0	2.25	2000	8030	1.0	0	0	4	1000	1000	1963
02-05-2014 0:00	550000.0	4.0	2.5	1940	10500	1.0	0	0	4	1140	800	1976
02-05-2014 0:00	490000.0	2.0	1.0	880	6380	1.0	0	0	3	880	0	1938
02-05-2014 0:00	335000.0	2.0	2.0	1350	2560	1.0	0	0	3	1350	0	1976
02-05-2014 0:00	482000.0	4.0	2.5	2710	35868	2.0	0	0	3	2710	0	1989
02-05-2014 0:00	452500.0	3.0	2.5	2430	88426	1.0	0	0	4	1570	860	1985
02-05-2014 0:00	640000.0	4.0	2.0	1520	6200	1.5	0	0	3	1520	0	1945
02-05-2014 0:00	463000.0	3.0	1.75	1710	7320	1.0	0	0	3	1710	0	1948
02-05-2014 0:00	1400000.0	4.0	2.5	2920	4000	1.5	0	0	5	1910	1010	1909
02-05-2014 0:00	588500.0	3.0	1.75	2330	14892	1.0	0	0	3	1970	360	1980
02-05-2014 0:00	365000.0	3.0	1.0	1090	6435	1.0	0	0	4	1090	0	1955
02-05-2014 0:00	1200000.0	5.0	2.75	2910	9480	1.5	0	0	3	2910	0	1939
02-05-2014 0:00	242500.0	3.0	1.5	1200	9720	1.0	0	0	4	1200	0	1965
02-05-2014 0:00	419000.0	3.0	1.5	1570	6700	1.0	0	0	4	1570	0	1956
02-05-2014 0:00	367500 0	4 0	3.0	3110	7231	2 0	0	0	3	3110	0	1997

¿Se podría predecir el precio de una propiedad de 750 ft2?



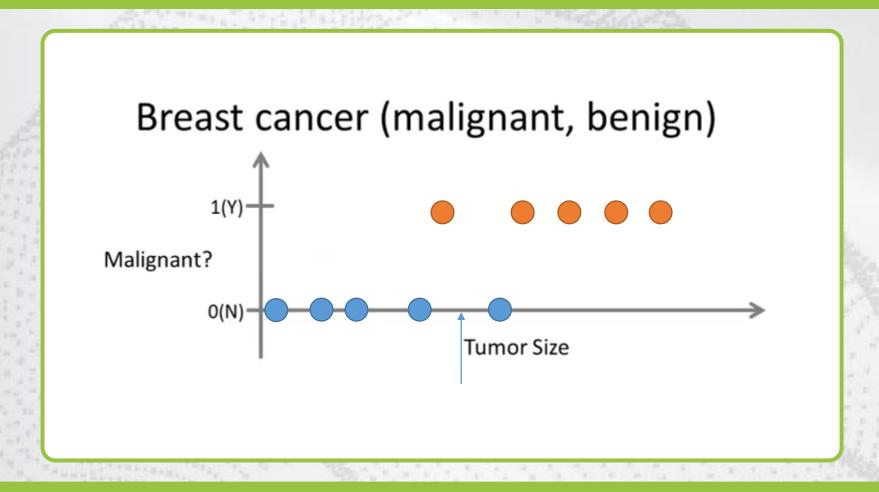
¿Se podría predecir el precio de una propiedad de 750 ft2?



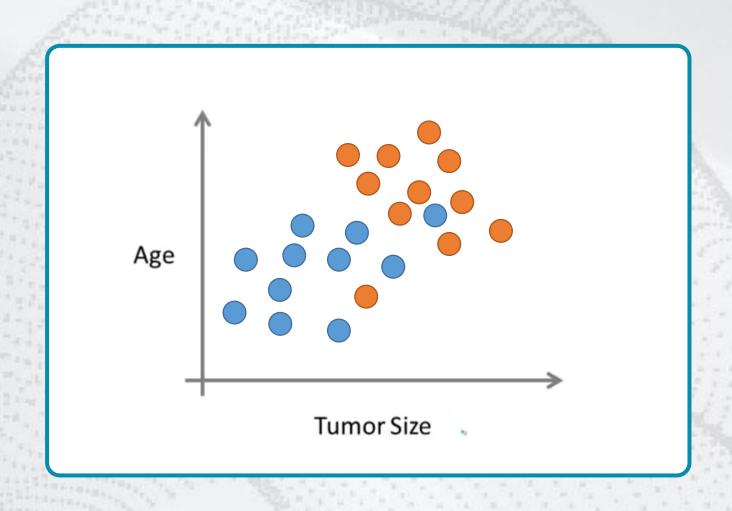
Es **supervisado**, porque ya se tienen "respuestas correctas".



Dado un tamaño de tumor, ¿puedo predecir si es cancerígeno?



Este es un problema de **clasificación**, porque se predicen valores discretos.



Pregunta

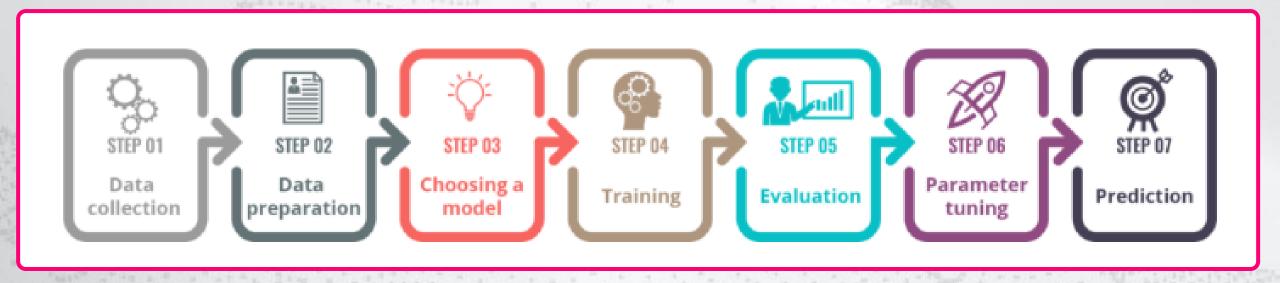
Eres científico de datos en una compañía y te piden que desarrolles un algoritmo para resolver los siguientes problemas:

- 1. Hay una gran cantidad de productos en stock, y se desea predecir cuántos ítems serán vendidos durante los siguientes tres meses.
- 2. Desarrollar un software que permita examinar las cuentas de los clientes para determinar si han sido hackeadas o no.

¿Clasificación o Regresión?



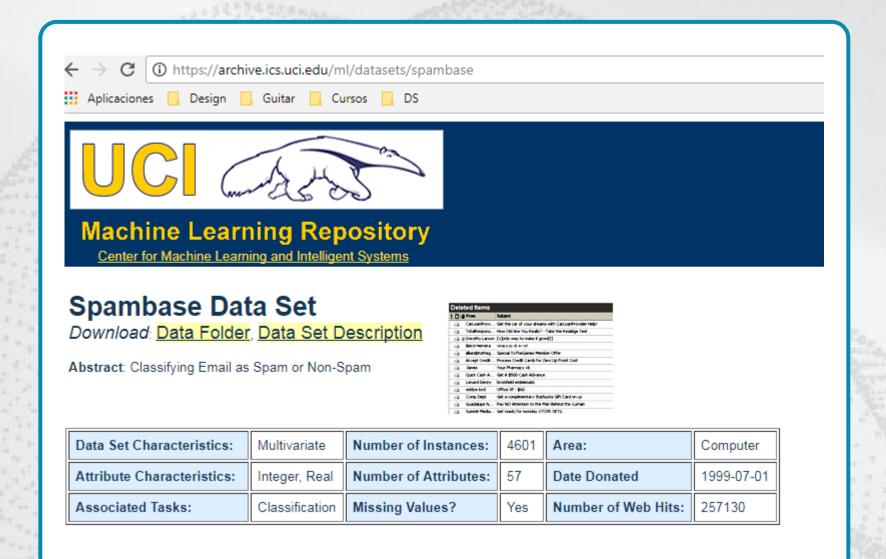
Pasos típicos de un problema ML

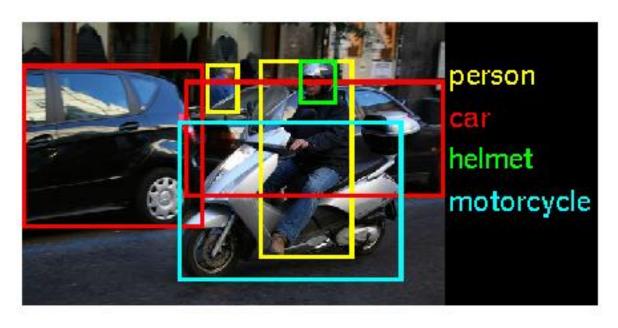


Volvamos en el tiempo a finales de la década de los 90s, en donde los mails SPAM son cada día más molestos.

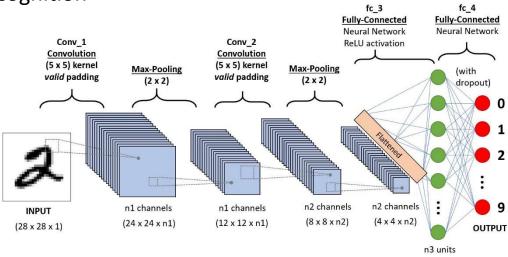
El Problema: Queremos identificar cuando un correo que llega a una cuenta corresponde a SPAM.



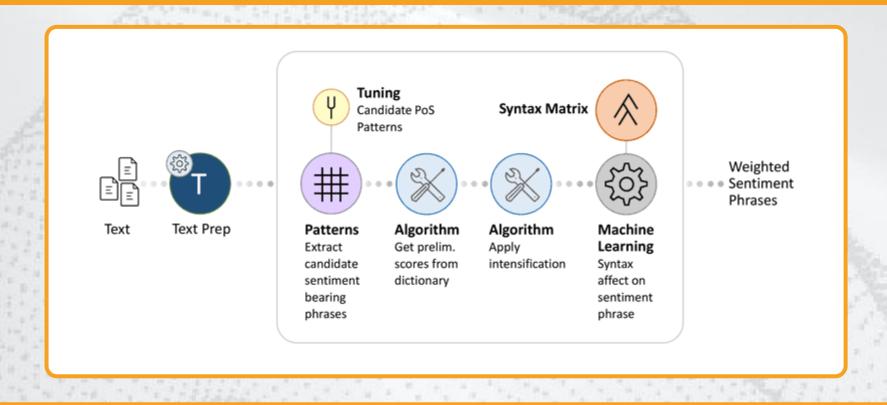




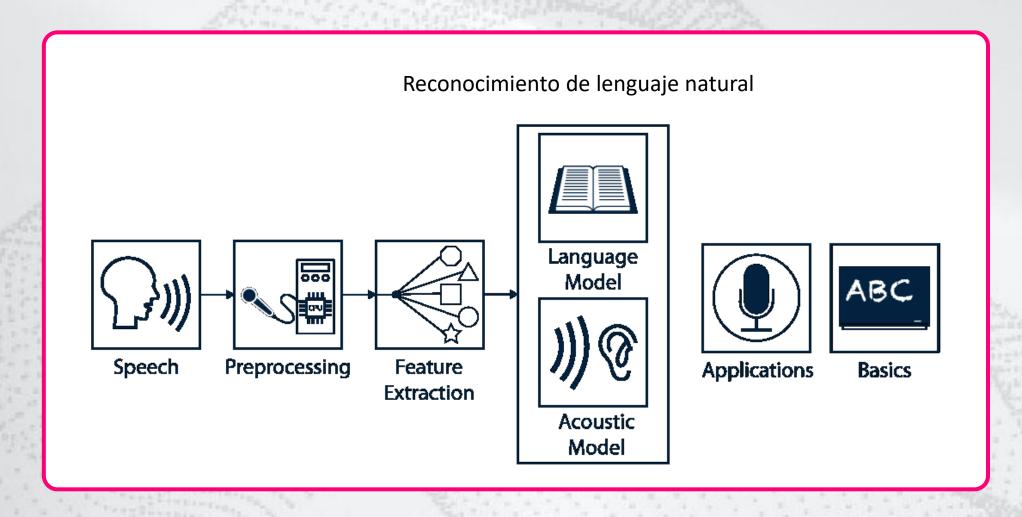
Visual Recognition



Análisis de sentimiento en las redes sociales.

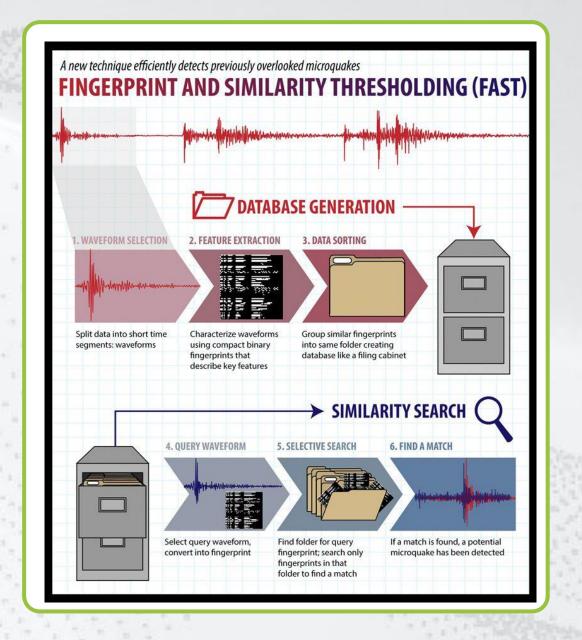


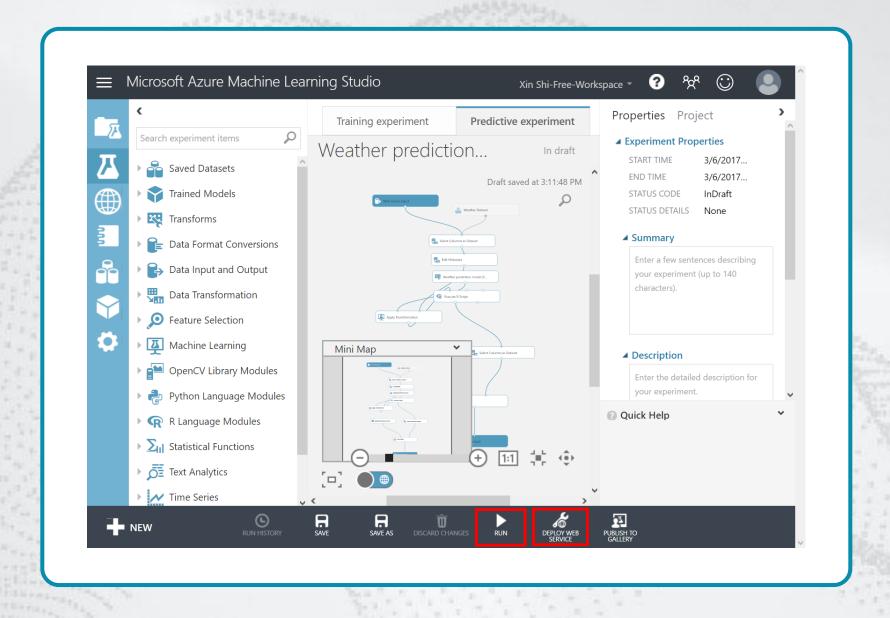
A partir de estos textos, el algoritmo es capaz de "aprender" a diferenciar el significado o polaridad de las opiniones y comentarios. En el caso del "**sentiment analysis**", los corpus de entrenamiento son previamente clasificados y ordenados entre positivos y negativos.



Reconocimiento de temas musicales:

(Shazam, google, youtube)





Fundamentos de la Ciencia de Datos y sus Aplicaciones

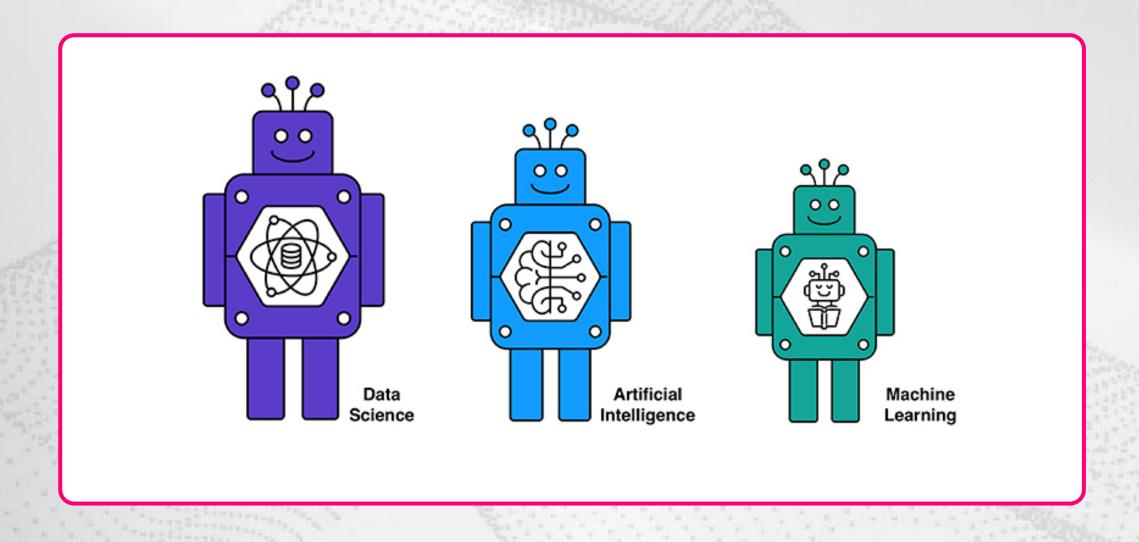


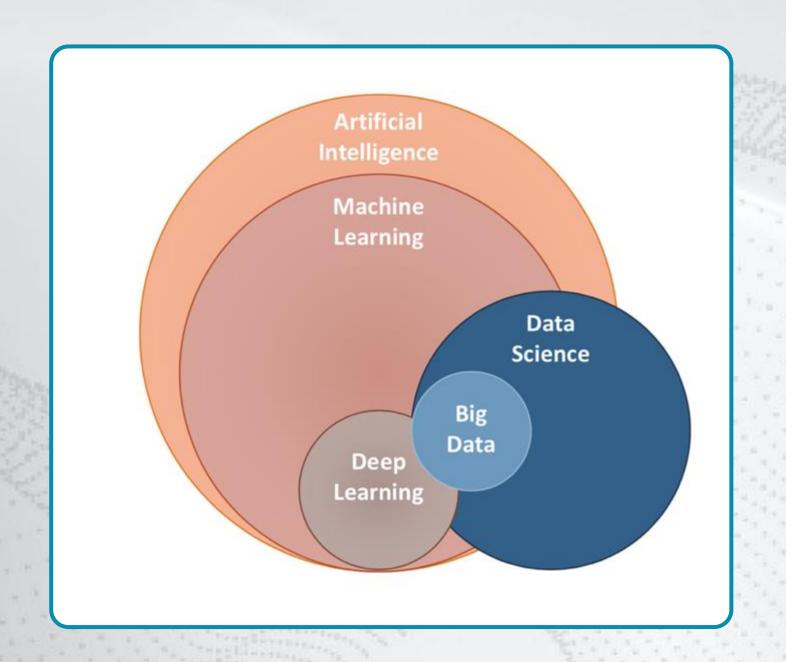
¿Qué es la Ciencia de Datos?

Una **ciencia** que permite **responder a las grandes preguntas** mediante el uso de técnicas computacionales, grandes volúmenes de datos, conocimiento estadísticas y conocimiento del negocio.



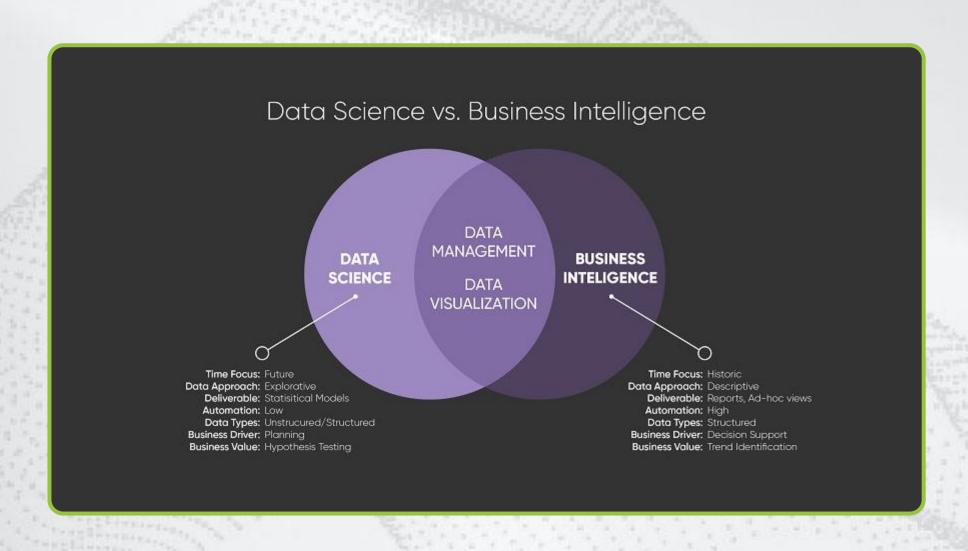
ML, DS, AI



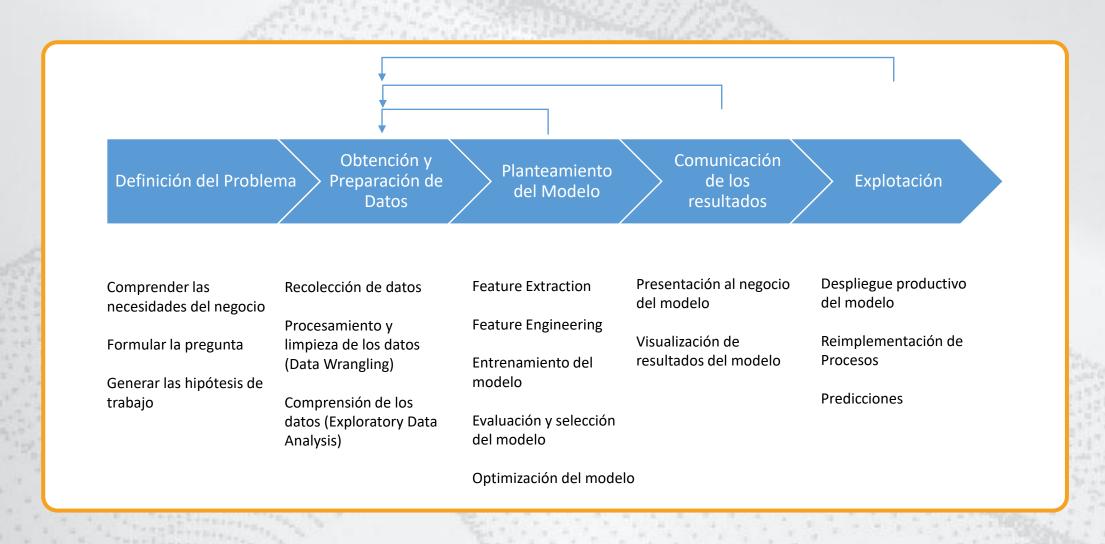


ML, DS, Al

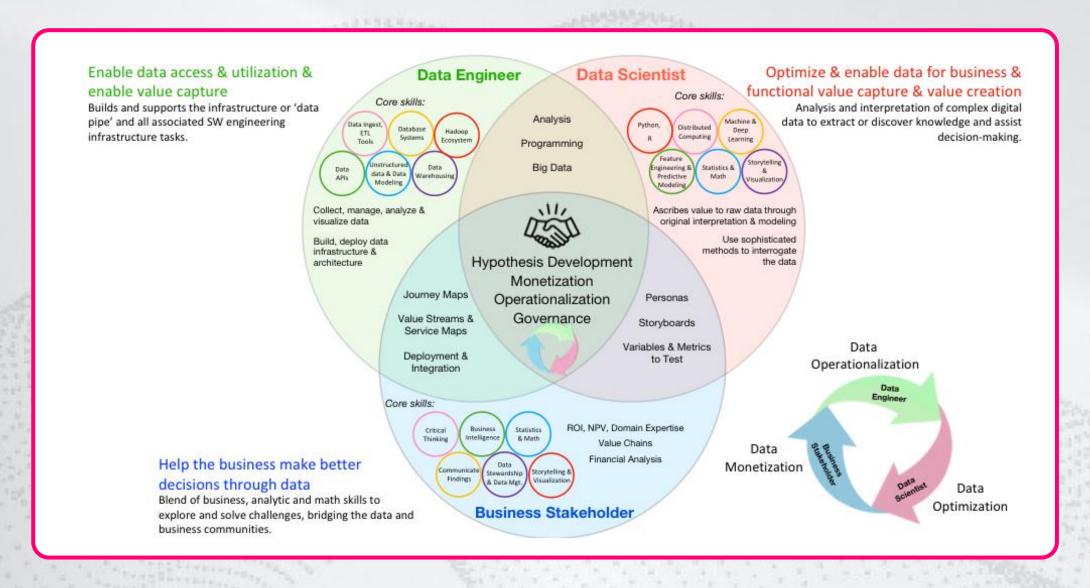
Data Science v/s Business Intelligence



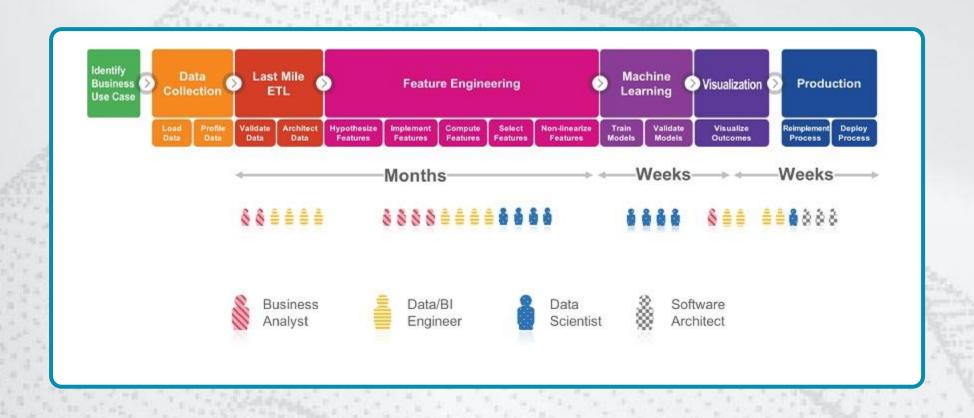
Ciclo de Vida de un problema de Ciencia de Datos



Roles en un Proyecto de DS



Tiempo y Esfuerzo de un Proyecto de DataScience



Skills Críticos para un DS



Skills Críticos para un DS





¿Podemos predecir, a partir de la información de las fichas médicas de la clínica, quién se va a atender el próximo mes?



Everything is a Recommendation



Over 80% of what people watch comes from our recommendations

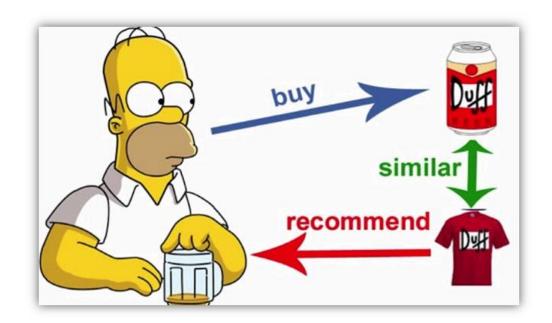
Recommendations are driven by Machine Learning ¿Y qué preguntas?

NETFLIX

¿Podemos predecir qué película de seguro te gustará?



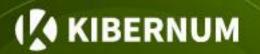


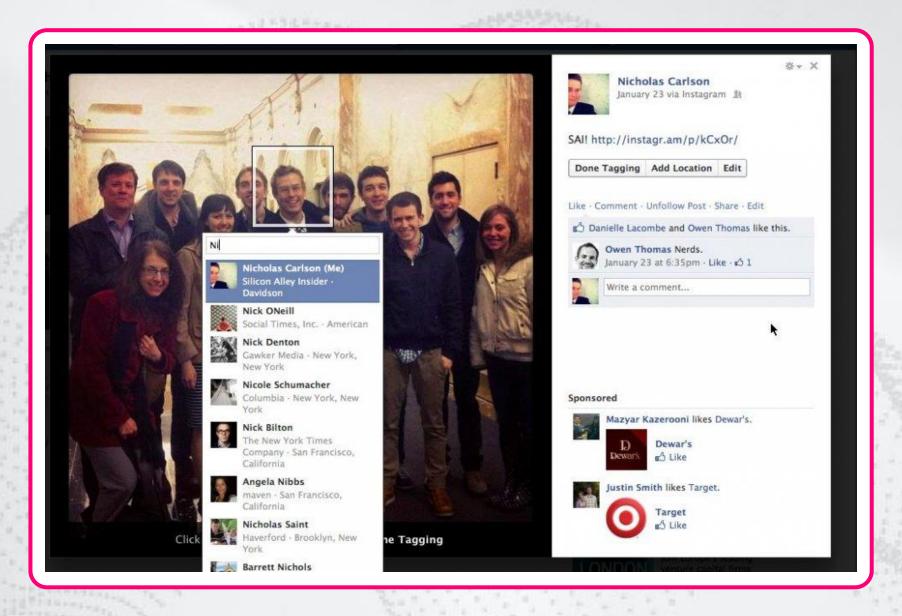


¿Podemos predecir qué artículo de seguro te interesará?

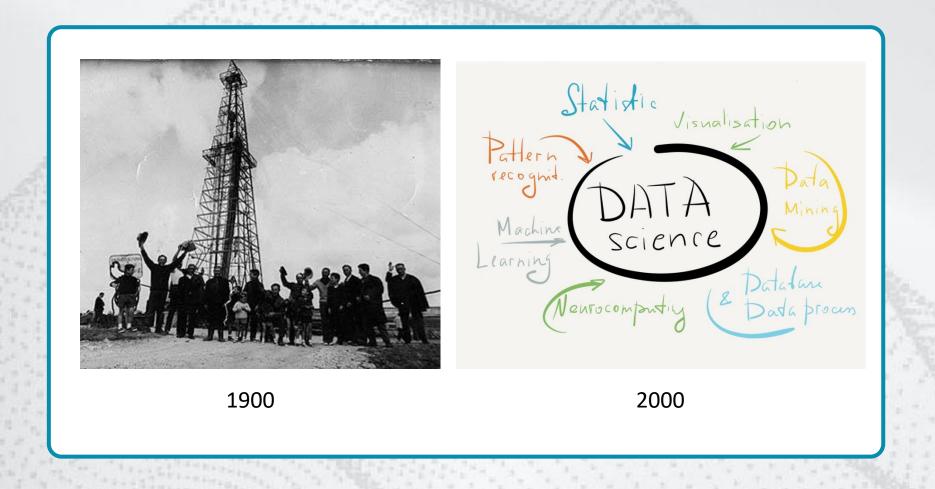


- ¿Podemos predecir quiénes son tus amigos?
- ¿Podemos predecir qué publicidad o contenido estarías interesado?





Por último





Gracias

