

Clase 1: Introducción

LABORATORIO DE PROGRAMACIÓN CIENTÍFICA PARA CIENCIA DE DATOS

MDS7202-1 – Otoño 2023

Objetivo de la clase

- Conocer al equipo docente
- Contexto
- ¿Para que es este curso?
- Metodología
- Información Administrativa
- Calendario

Equipo Docente



Pablo Badilla
Profesor

- Msc Ciencias de la Computación
- Data Scientist @BCI.
- Jefe de Desarrollo Dash-AI
- 3 iteraciones del curso.



Ignacio Meza
Profesor

- Ingeniero Civil Eléctrico
- Estudiante de Msc Ciencias de la Computación
- Data Scientist @BCI-MACH.
- 4 iteraciones del curso.



Sebastian Tinoco
Profesor Auxiliar

- Ingeniero Comercial
- Estudiante de Msc Ciencias de Datos

Hablemos de Vinos

Pregunta: ¿Cómo describirían un vino?



Pregunta: ¿Cómo describirían un vino?

Color

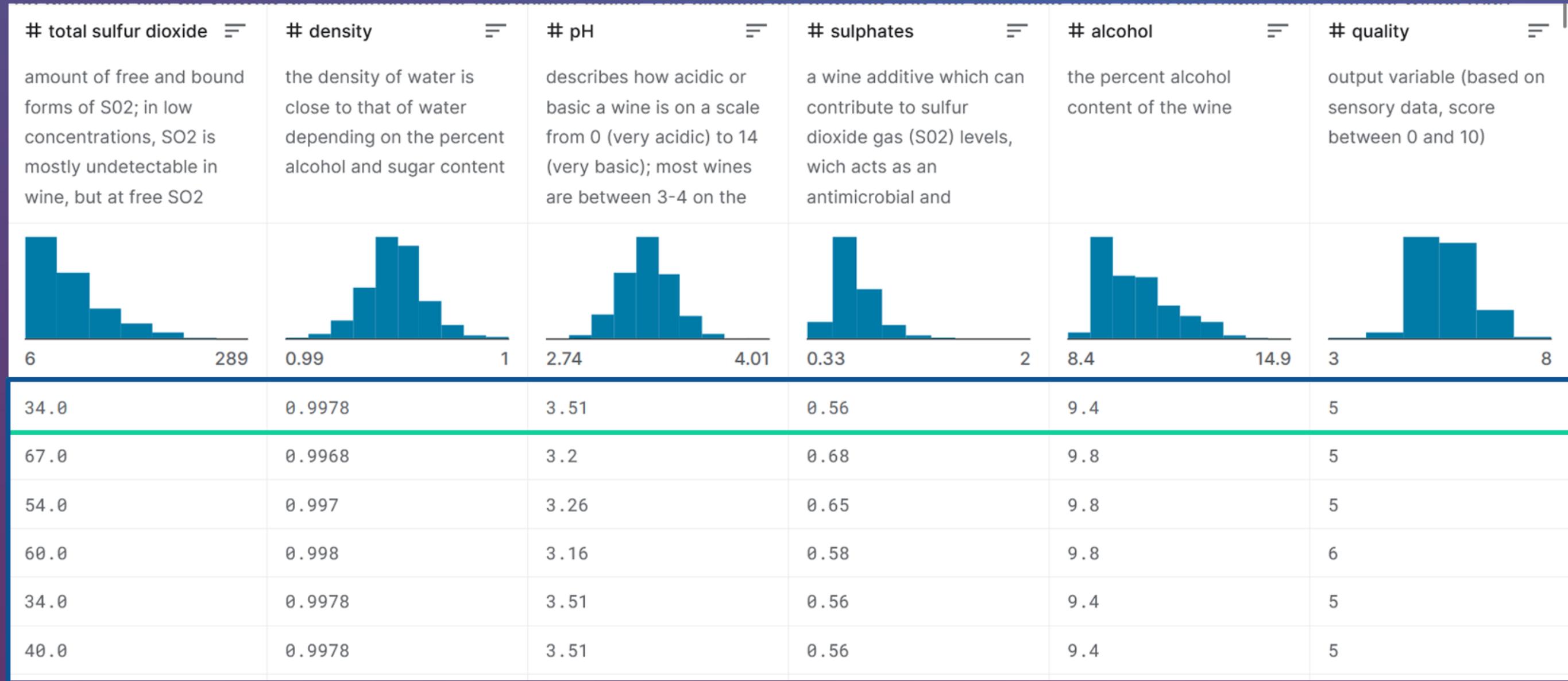
Sabor

Cepa

Año de origen ⏳
Grado alcohólico
Calificación del Vino



Pregunta: ¿Cómo describirían un vino?

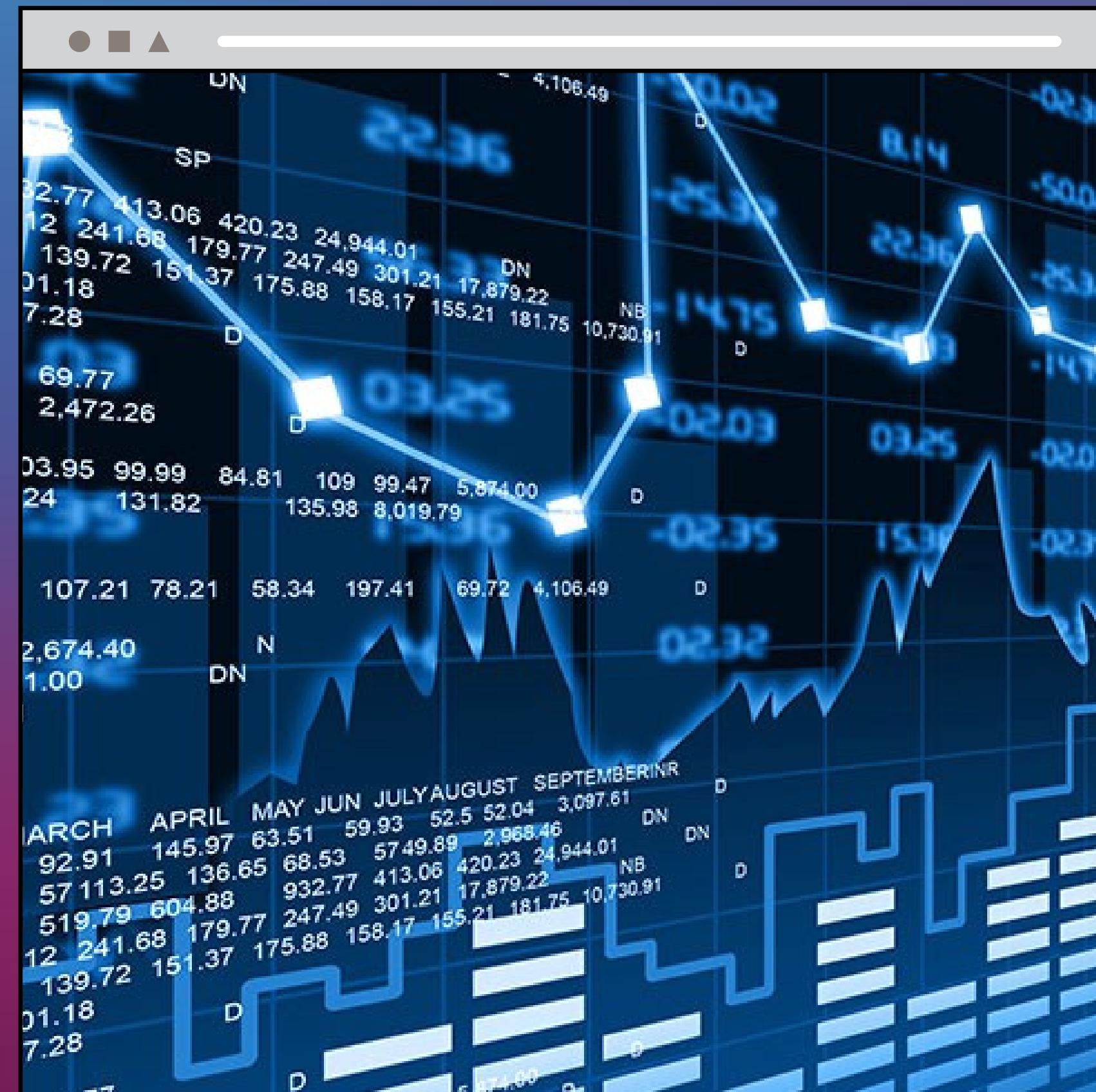


Datos

Dataset

Datos.....

- Colección de valores que describen o representan atributos o características de un objeto.



Fuentes de Datos

IOT

Teléfonos, smartwatches, domótica, electrodomésticos, drones, sensores industriales...



https://es.wikipedia.org/wiki/Internet_de_las_cosas

Aplicaciones, Servicios y Empresas

Almacenamiento y análisis de los patrones de los clientes.



https://elpais.com/tecnologia/2019/07/17/actualidad/1563358803_598879.html

Redes Sociales

En la práctica, todo lo que haces, buscas y ves desde que entras hasta que sales.

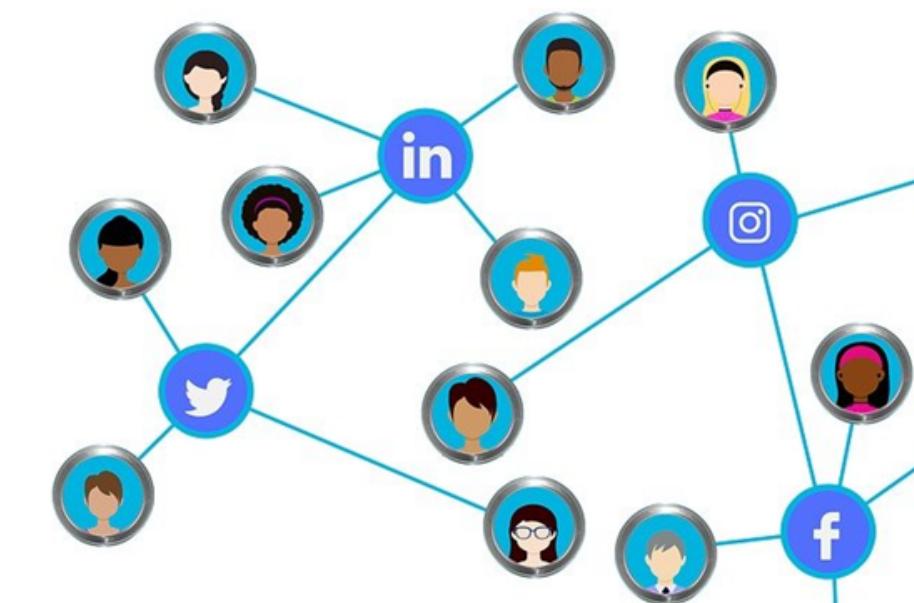


Image by [Pete Linforth](#) from [Pixabay](#)

Tipos de Datos...

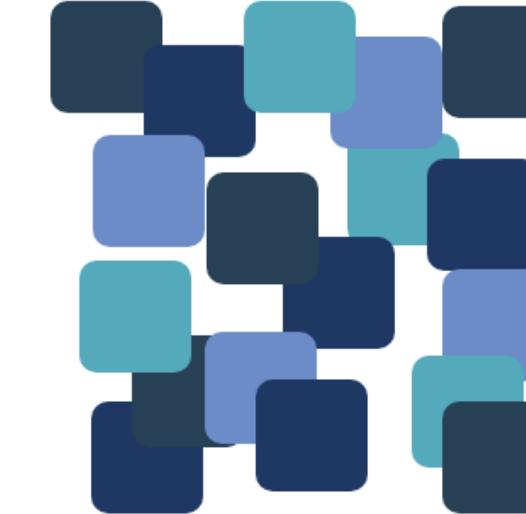


<https://en.wikipedia.org/wiki/Data>

STRUCTURED DATA



UNSTRUCTURED DATA



 Se ordena a través de filas, columnas y bases de datos relacionales.

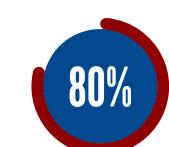
 Compuesta por números, fechas y strings.

 Alrededor del 20% de los datos que existen son estructurados.

 Requiere baja cantidad de almacenamiento (depende del caso)

 No puede ser ordenada a través de tablas, ya que sus datos no son estructurados.

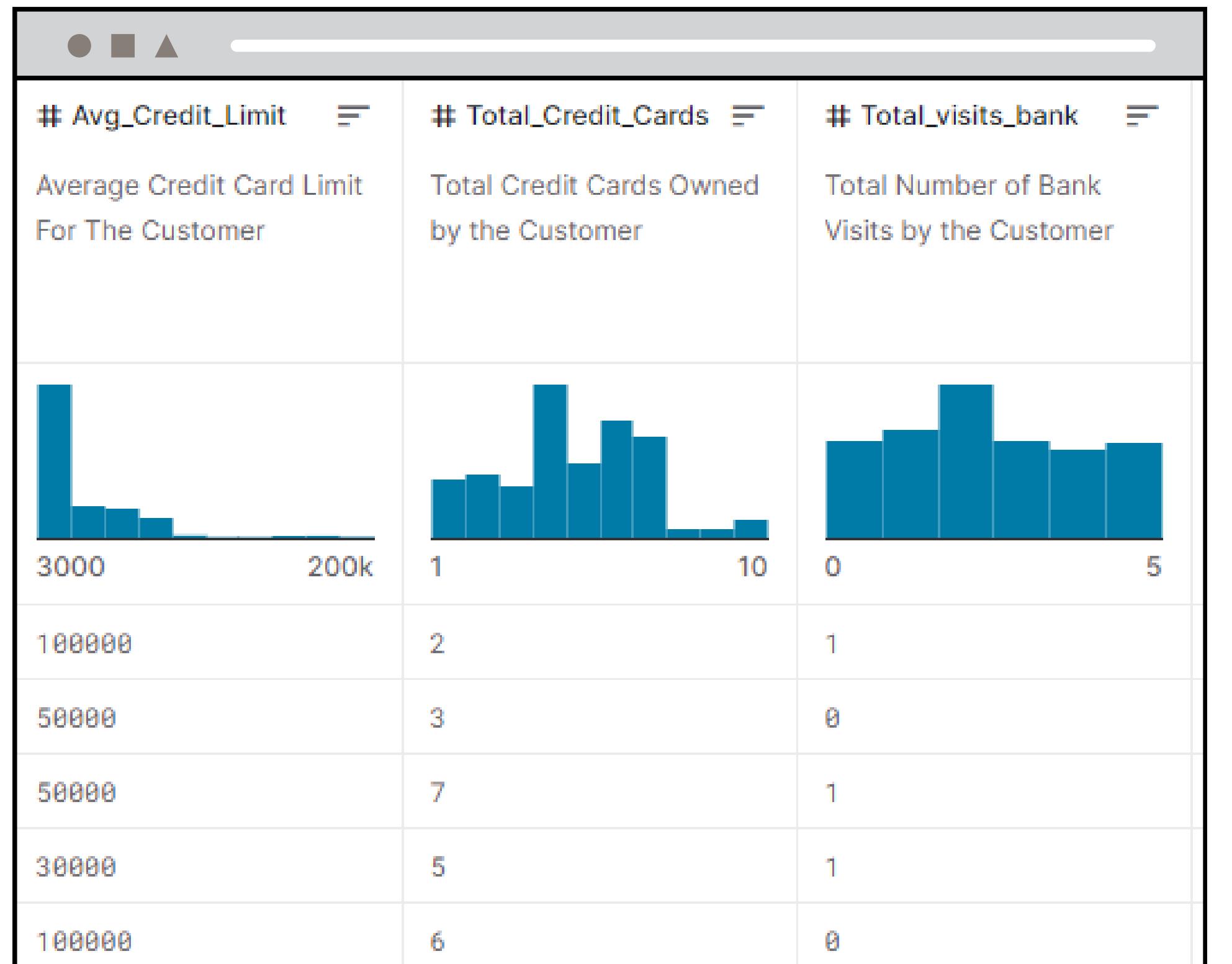
 Imágenes, videos, audio, etc.

 Alrededor del 80% de los datos son de este tipo 

 Requieren mayor cantidad de almacenamiento.



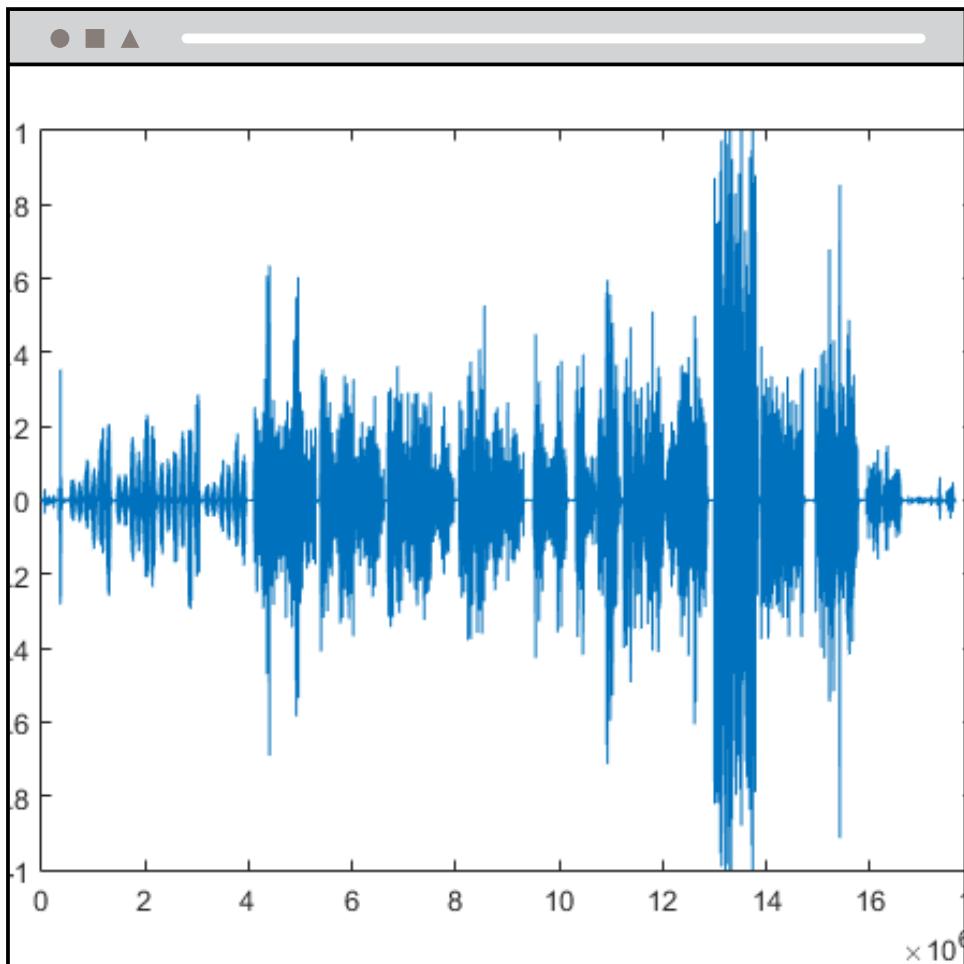
Tipos de Datos...



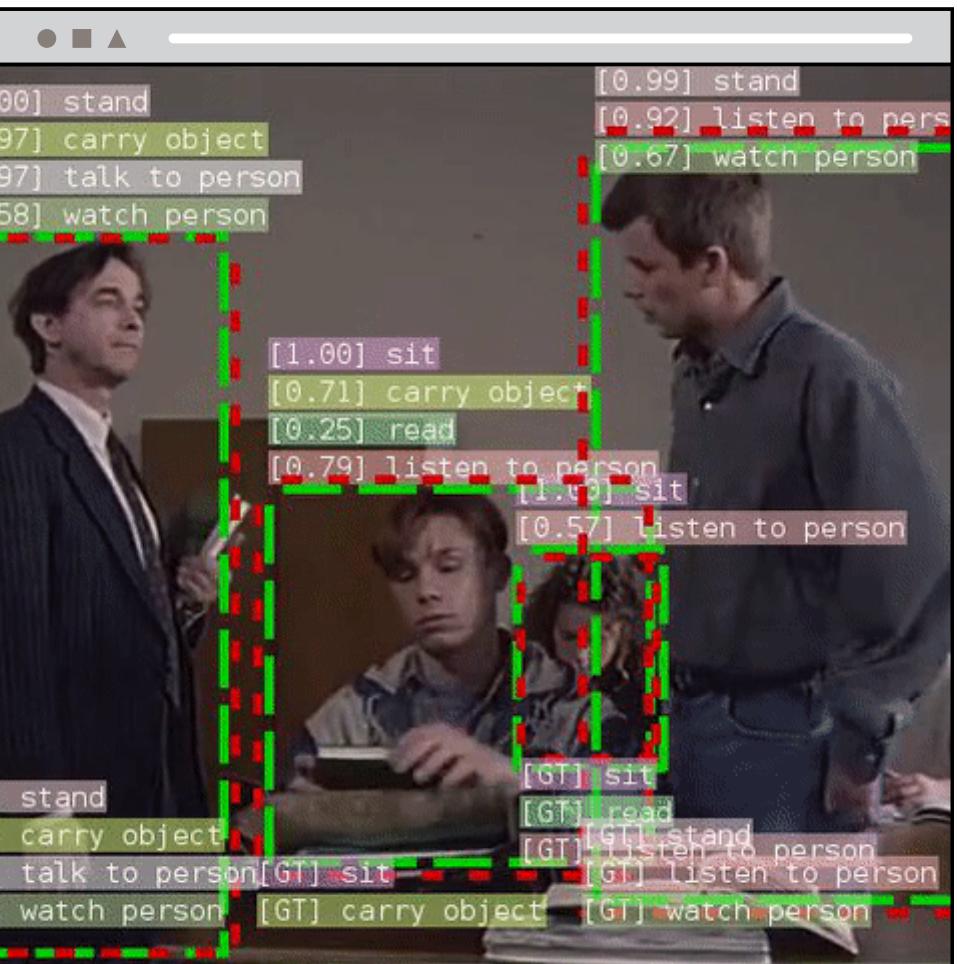


Tipos de Datos...

Audio



Videos

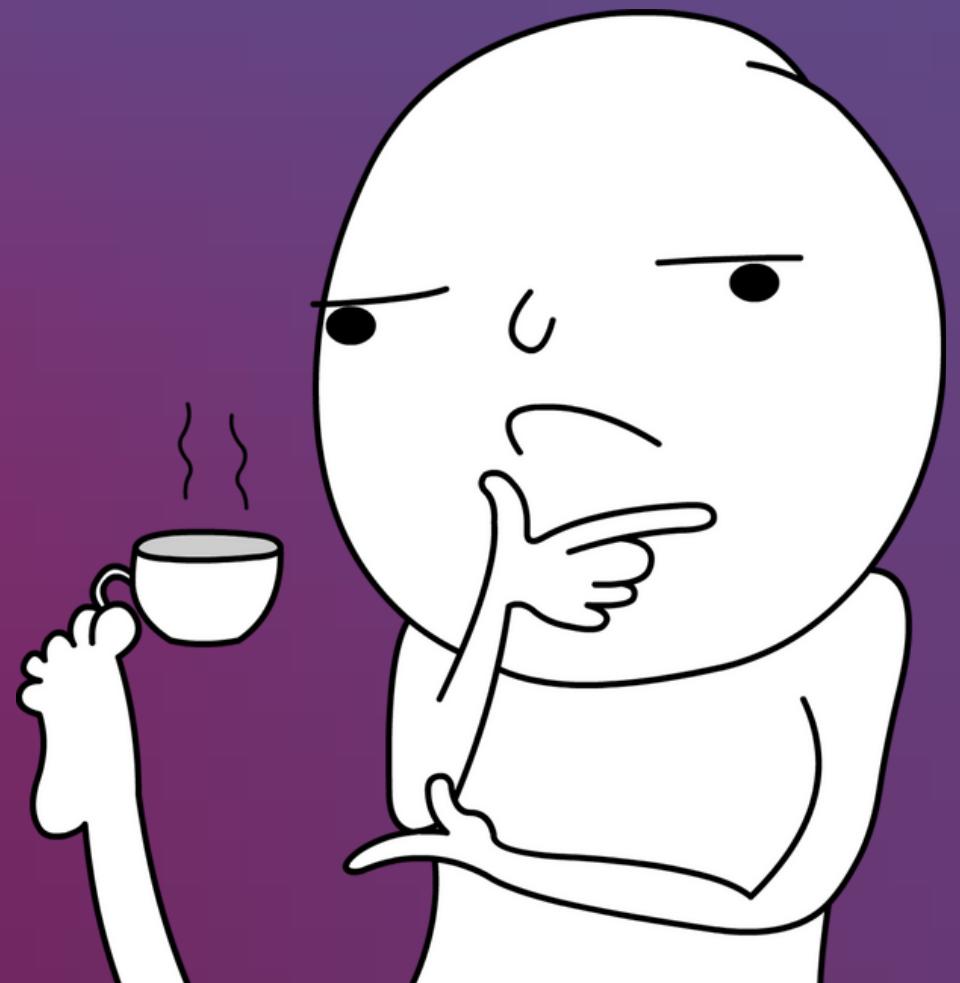


Texto

(string)	question (string)	answers (seq)
Giselle Carter...	"When did Beyonce start..."	{ "text": [late 1990s"
Giselle Carter...	"What areas did Beyonce compete..."	{ "text": ["singing and
Giselle Carter...	"When did Beyonce leave..."	{ "text": [], "answer_s
Giselle Carter...	"In what city and state did..."	{ "text": ["Houston, Te
Giselle Carter...	"In which decade did Beyonce..."	{ "text": [1990s"],...



Pero.... ¿Que hacemos con
los datos?....

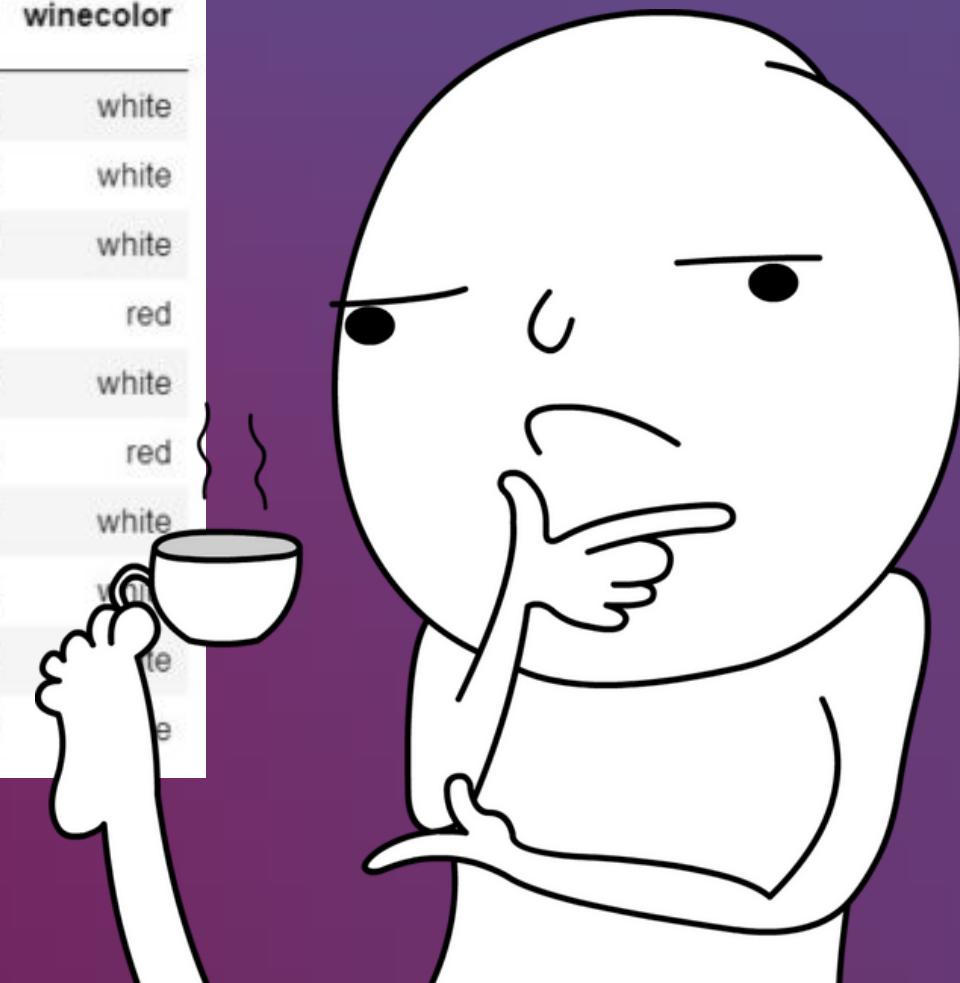


Volvamos a los datos de los vinos

¿Qué podríamos hacer con estos datos?

¿Qué área permite hacer cosas con esto?

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	winecolor
818	6.2	0.30	0.17	2.8	0.040	24.0	125.0	0.99390	3.01	0.46	9.0	5	white
3032	6.7	0.14	0.46	1.6	0.036	15.0	92.0	0.99264	3.37	0.49	10.9	5	white
2722	7.6	0.40	0.27	5.2	0.030	32.0	101.0	0.99172	3.22	0.62	12.3	7	white
5798	8.9	0.31	0.36	2.6	0.056	10.0	39.0	0.99562	3.40	0.69	11.8	5	red
757	6.8	0.22	0.37	15.2	0.051	68.0	178.0	0.99935	3.40	0.85	9.3	6	white
5662	9.1	0.68	0.11	2.8	0.093	11.0	44.0	0.99888	3.31	0.55	9.5	6	red
346	5.6	0.34	0.10	1.3	0.031	20.0	68.0	0.99060	3.36	0.51	11.2	7	white
3526	8.9	0.27	0.28	0.8	0.024	29.0	128.0	0.98984	3.01	0.35	12.4	6	white
4188	5.3	0.33	0.30	1.2	0.048	25.0	119.0	0.99045	3.32	0.62	11.3	6	white
2821	6.6	0.40	0.46	6.2	0.056	42.0	241.0	0.99680	3.50	0.60	9.9	5	white



Ciencia de Datos



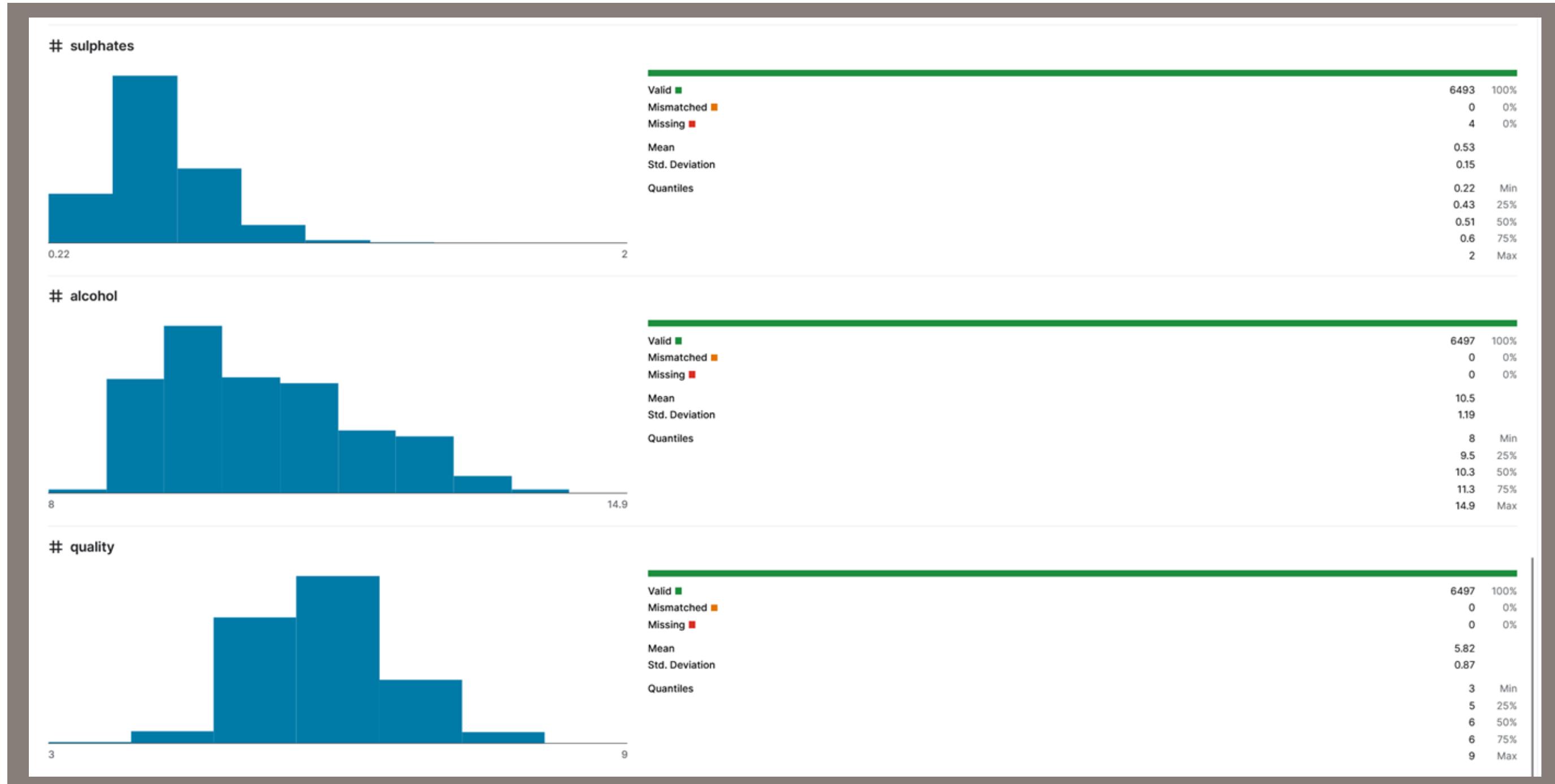
Ciencia de Datos

"Ciencia de datos es la combinación de matemáticas, estadísticas y aprendizaje automático (entre otros) que tienen por objetivo extraer conocimiento a partir de los datos."

Análisis Exploratorio de Datos

Análisis de un dataset a partir de el resumen y visualización de sus principales características.

Objetivo
Descubrir patrones y anomalías y probar hipótesis que tengamos sobre los datos.



Aprendizaje Automático

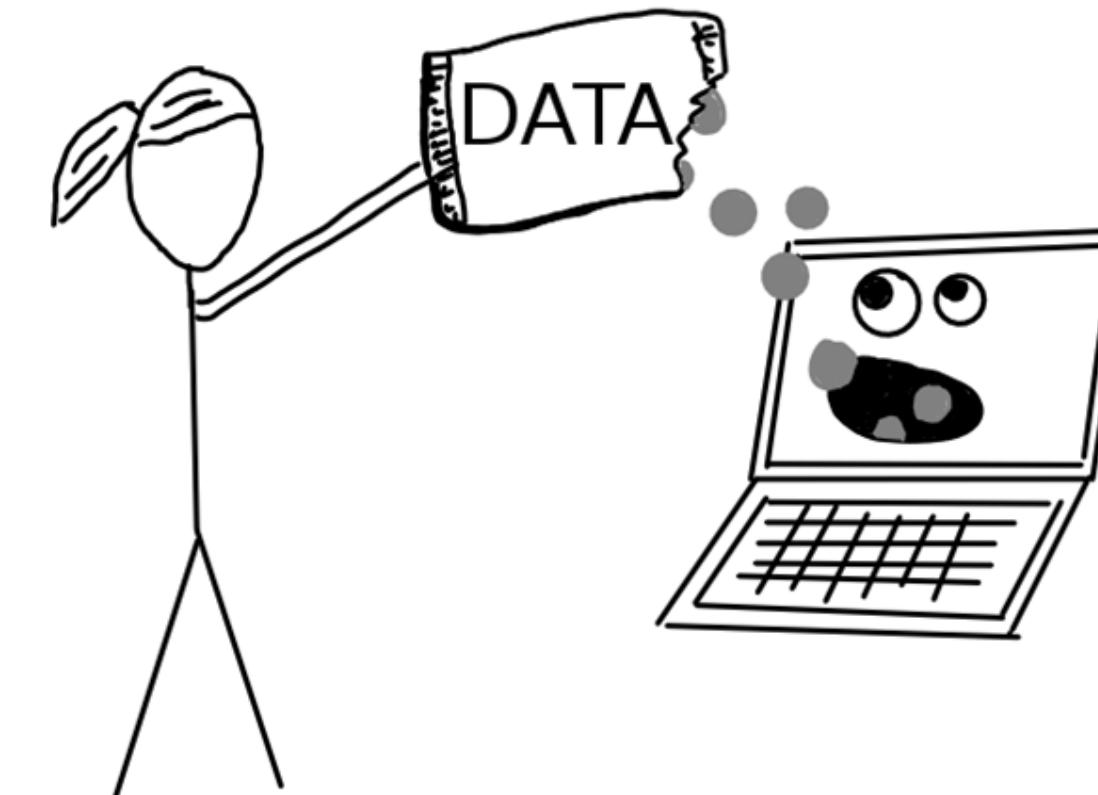
Estudio de algoritmos enfocados en generar modelos capaces de hacer predicciones / tomar decisiones sin que estos sean explícitamente programados para esto (“aprenden de los datos”).



Without Machine Learning



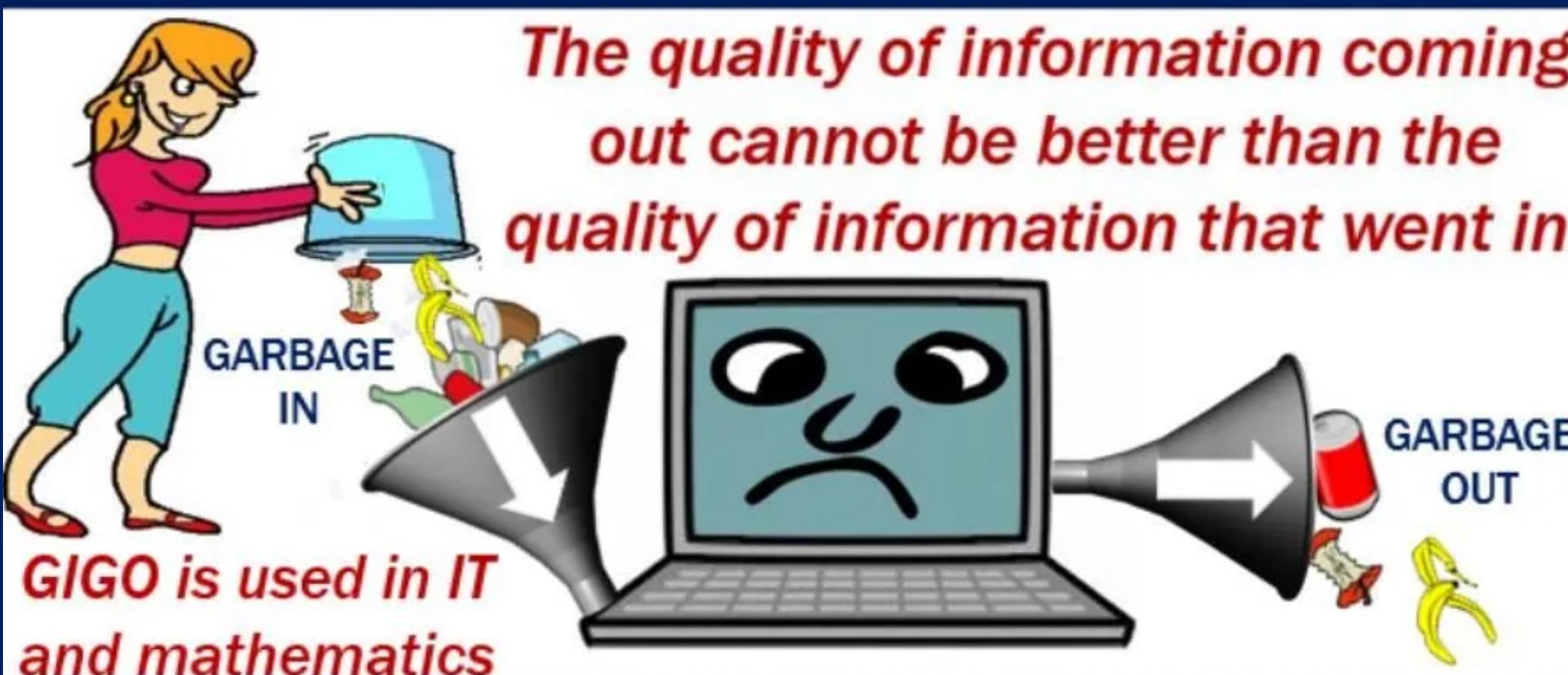
With Machine Learning



<https://christophm.github.io/interpretable-ml-book/terminology.html>

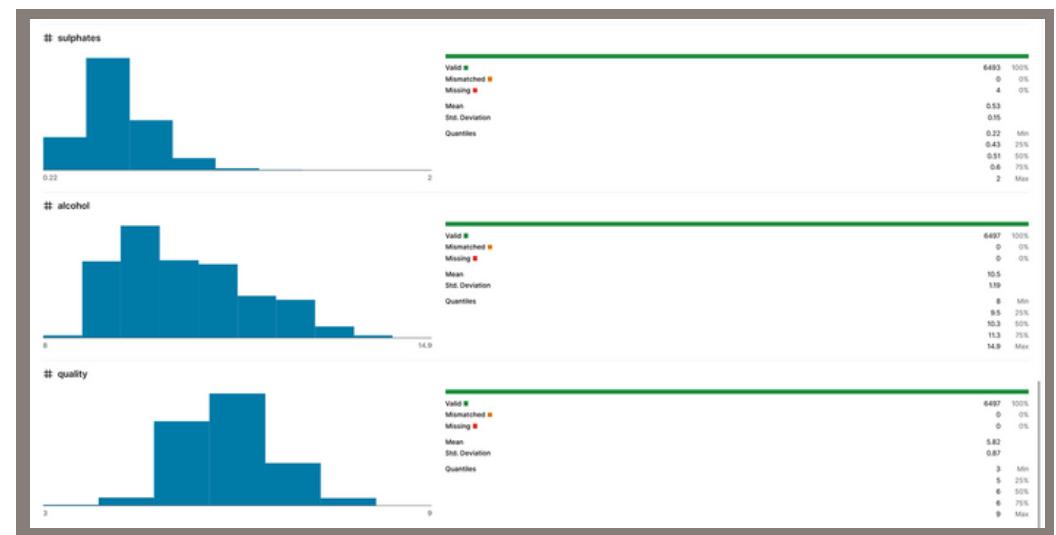
Pero ojo...

What is GIGO?

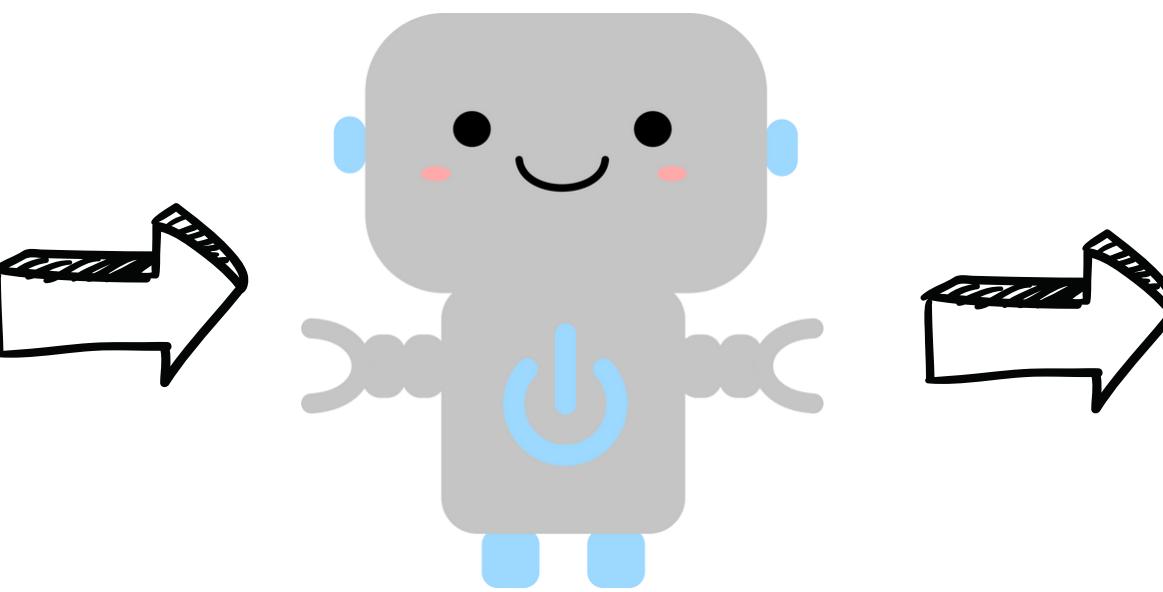


Garbage In, Garbage Out

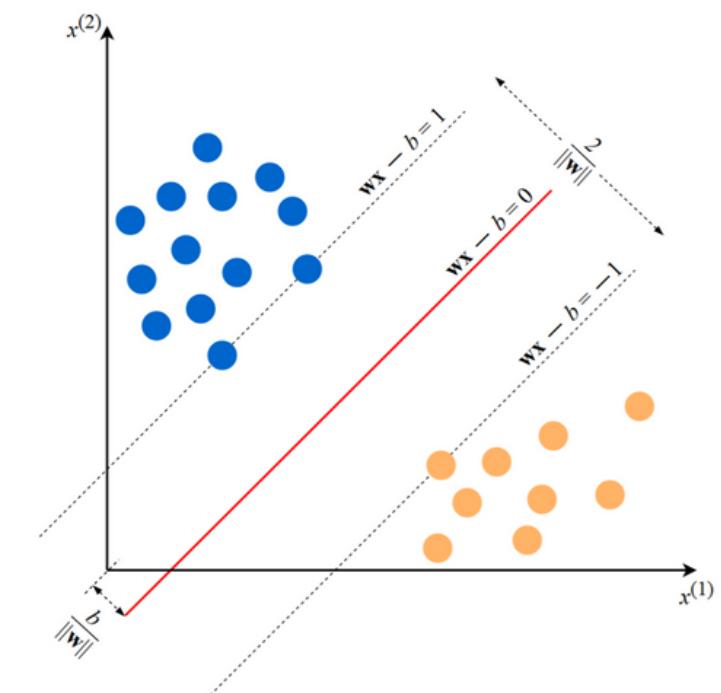
Aprendizaje Automático



Datos Crudos



Algún Modelo de Aprendizaje
Automático



Clasificación Automática

Aprendizaje Automático

Supervisado

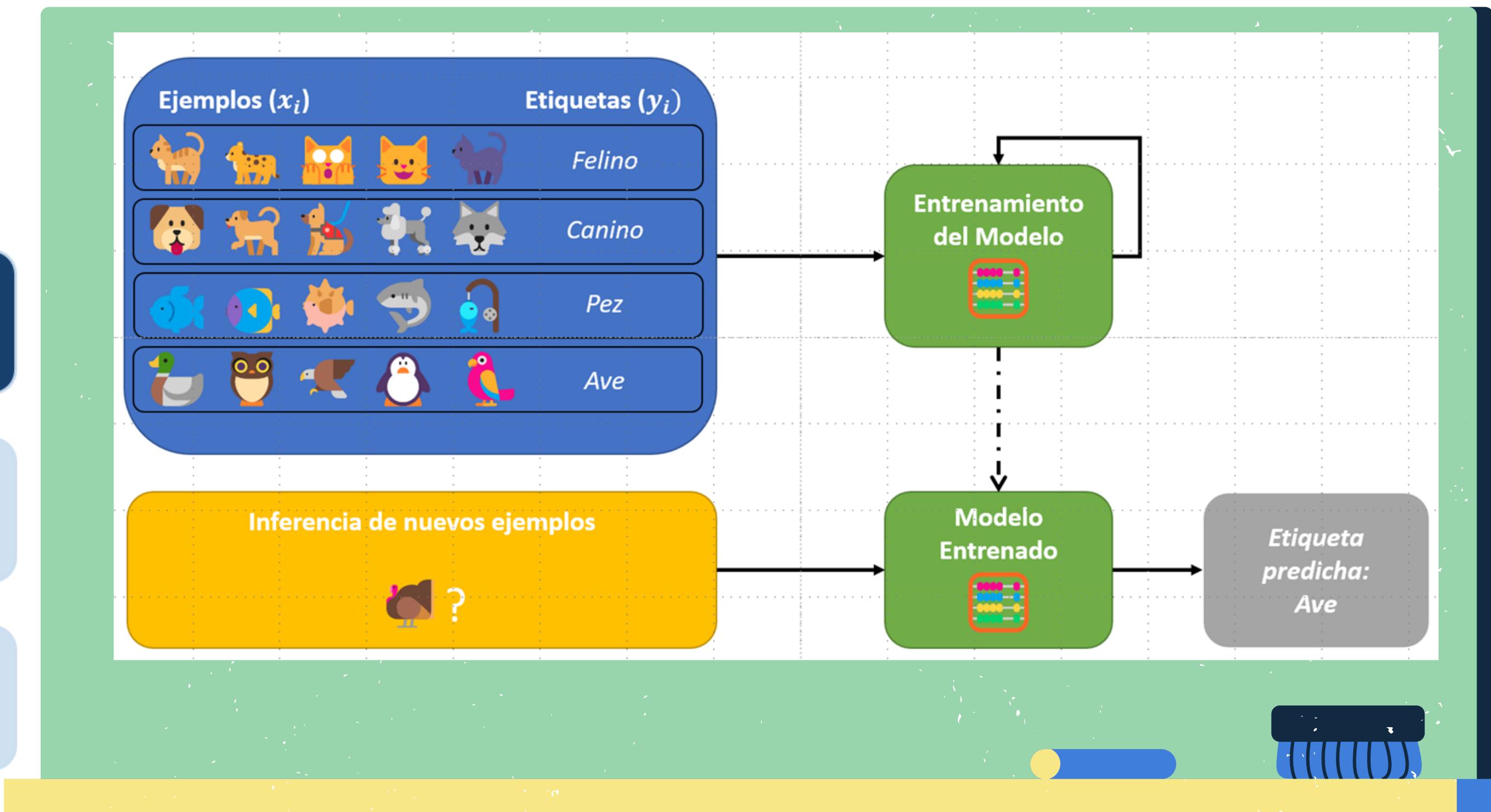
Aprende a través
de ejemplos
etiquetados.

**No
Supervisado**

Aprende patrones
a partir de datos no
etiquetados

Reforzado

Aprende a través de
experimentos en el
tiempo



Aprendizaje Automático

Supervisado

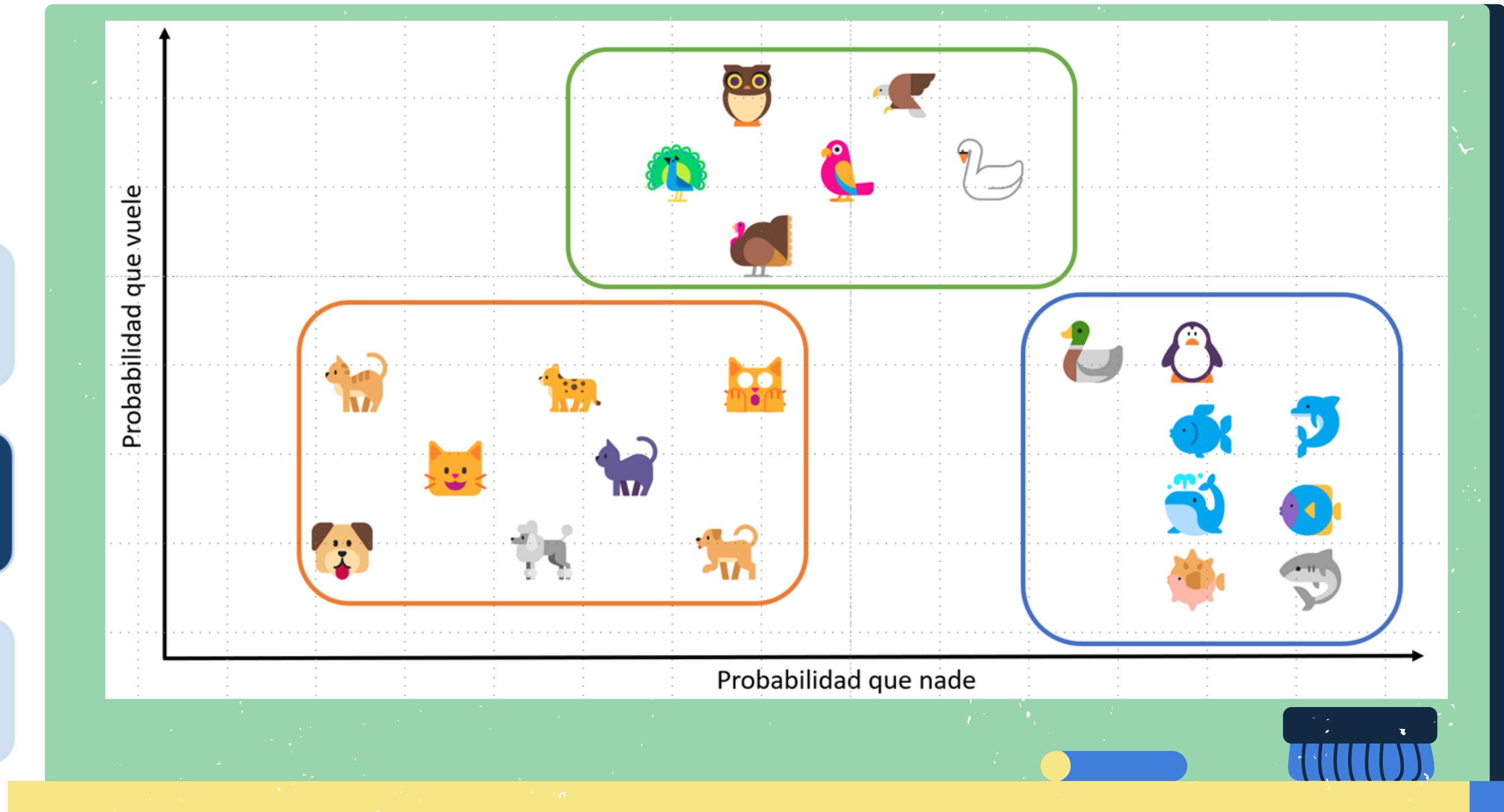
Aprende a través
de ejemplos
etiquetados.

**No
Supervisado**

Aprende patrones
a partir de datos no
etiquetados

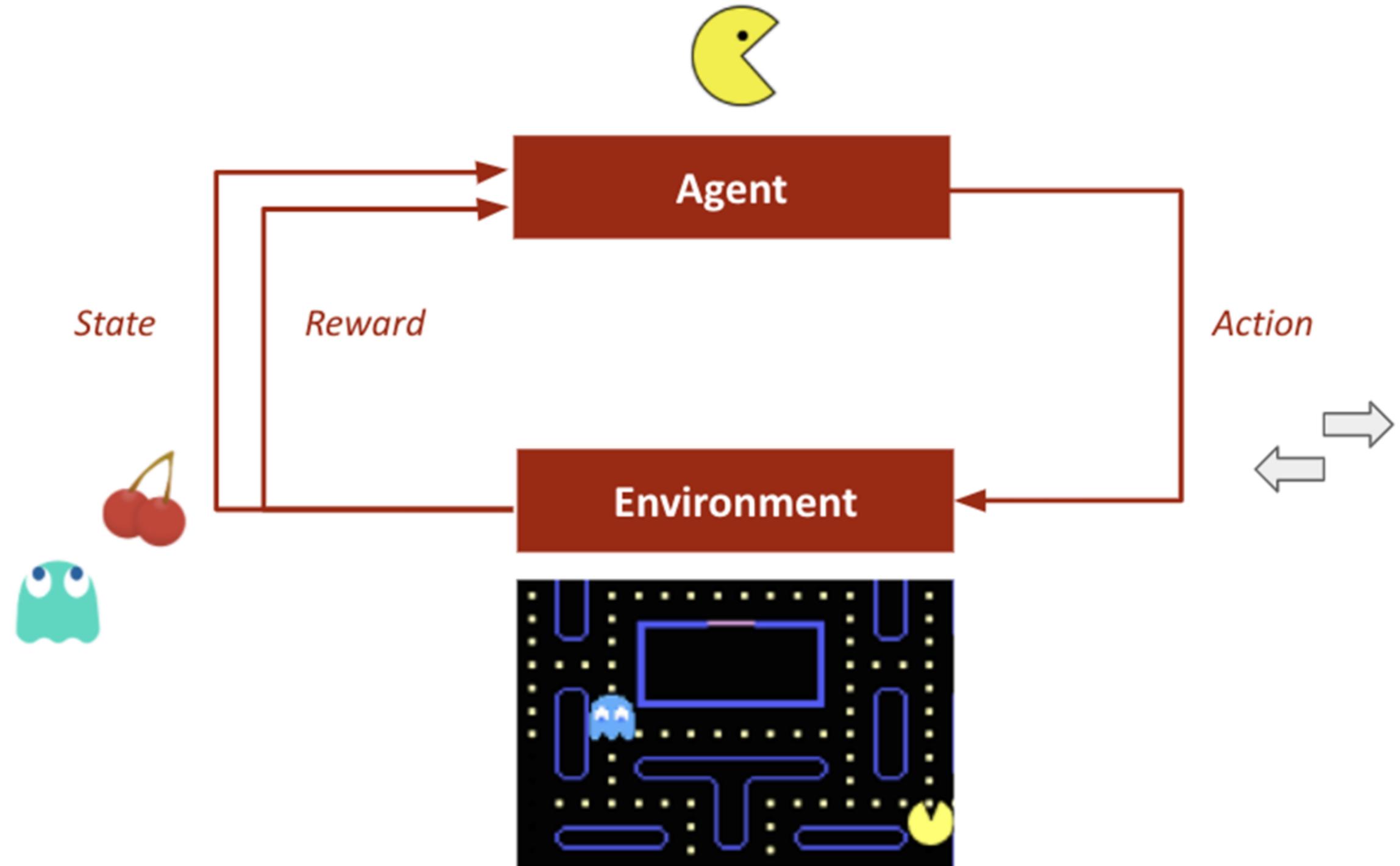
Reforzado

Aprende a través
de experimentos
en el tiempo



Aprendizaje Automático

Supervisado	Aprende a través de ejemplos etiquetados.
No Supervisado	Aprende patrones a partir de datos no etiquetados
Reforzado	Aprende a través de experimentos en el tiempo



Algunas tareas que se pueden resolver con ML

Supervisado	Clasificación	Asignar una etiqueta a un ejemplo.
	Regresión	Asignar un valor a un ejemplo.
	One Shot	Dado un solo ejemplo de una clase, reconocer si otro también pertenece a esta.
	Etiquetado de Secuencias	Asignar a una secuencia etiquetas.
No Supervisado	Clustering	Buscar grupos dentro de los ejemplos.
	Reducción de Dimensionalidad	Representar ejemplos usando menos características de estos.
	Detección de Outliers	Buscar datos fuera de los patrones comunes.
Otros Tipos	Optimización	Identificar los parámetros más óptimos de algún problema.
	Ranking	Dado un conjunto de ejemplos, ordenar estos según algún criterio.
	Recomendación	Dado un conjunto de ejemplos, sugerir uno de ellos.

Tareas específicas según área

Computer Vision

-  Image Classification
-  Image Segmentation
-  Zero-Shot Image Classification
-  Image-to-Image
-  Unconditional Image Generation
-  Object Detection
-  Video Classification

Natural Language Processing

-  Translation
-  Fill-Mask
-  Token Classification
-  Sentence Similarity
-  Question Answering
-  Summarization
-  Zero-Shot Classification
-  Text Classification
-  Text2Text Generation
-  Text Generation
-  Conversational
-  Table Question Answering

Audio

-  Automatic Speech Recognition
-  Audio Classification
-  Text-to-Speech
-  Audio-to-Audio
-  Voice Activity Detection

Multimodal

-  Feature Extraction
-  Text-to-Image
-  Visual Question Answering
-  Image-to-Text
-  Document Question Answering

Google/vit-base-patch16-224

like 91

Image Classification

PyTorch

TensorFlow

JAX

Transformers

Imagenet-1k

Imagenet-21k

arxiv:2010.11929

arxiv:2006.03677

vit

vision

AutoTrain Compatible

License: apache-2.0

Model card

Files and versions

Community 1

⋮

Train

Deploy

Use in Transformers

Vision Transformer (base-sized model)

Vision Transformer (ViT) model pre-trained on ImageNet-21k (14 million images, 21,843 classes) at resolution 224x224, and fine-tuned on ImageNet 2012 (1 million images, 1,000 classes) at resolution 224x224. It was introduced in the paper [An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](#) by Dosovitskiy et al. and first released in [this repository](#). However, the weights were converted from the [timm repository](#) by Ross Wightman, who already converted the weights from JAX to PyTorch. Credits go to him.

Disclaimer: The team releasing ViT did not write a model card for this model so this model card has been written by the Hugging Face team.

Model description

The Vision Transformer (ViT) is a transformer encoder model (BERT-like) pretrained on a

Downloads last month

185,459



Hosted inference API

Image Classification

Tiger



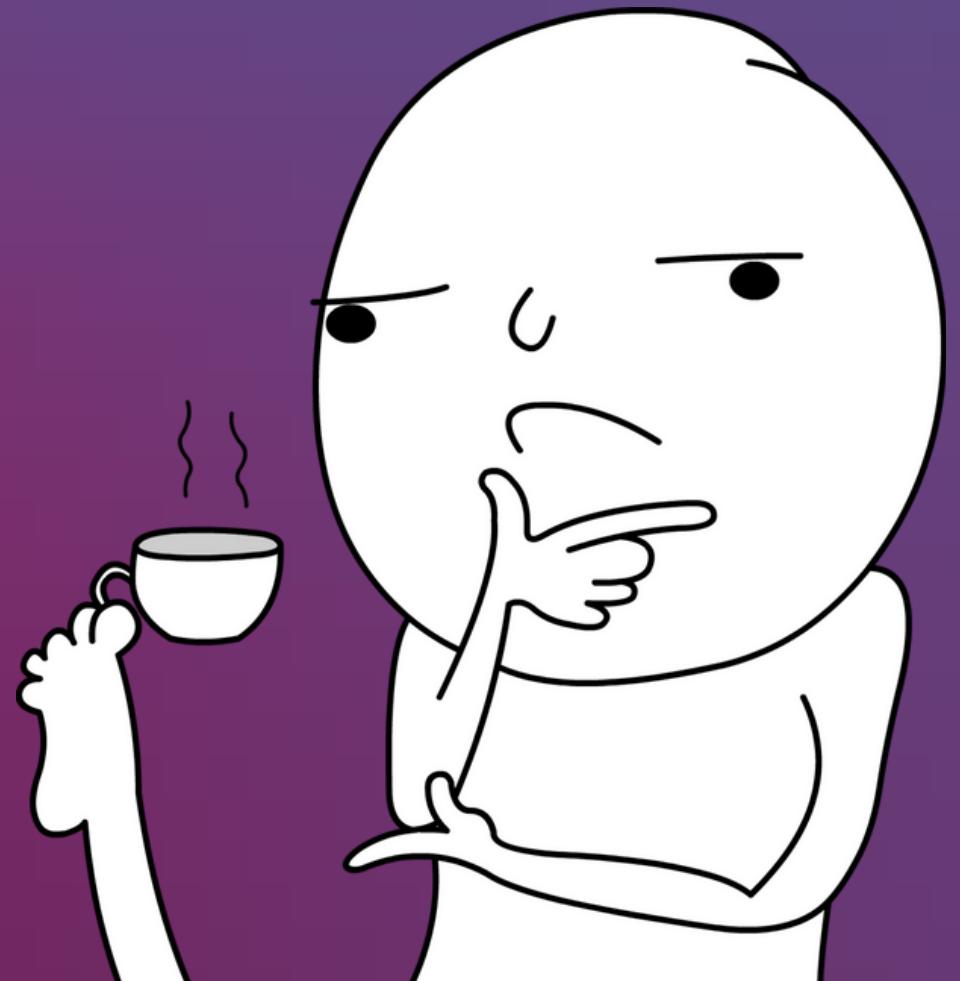
Computation time on cpu: cached

tiger, Panthera tigris

0.886

<https://huggingface.co/google/vit-base-patch16-224>

Por qué mi modelo tomo la decisión que tomó?

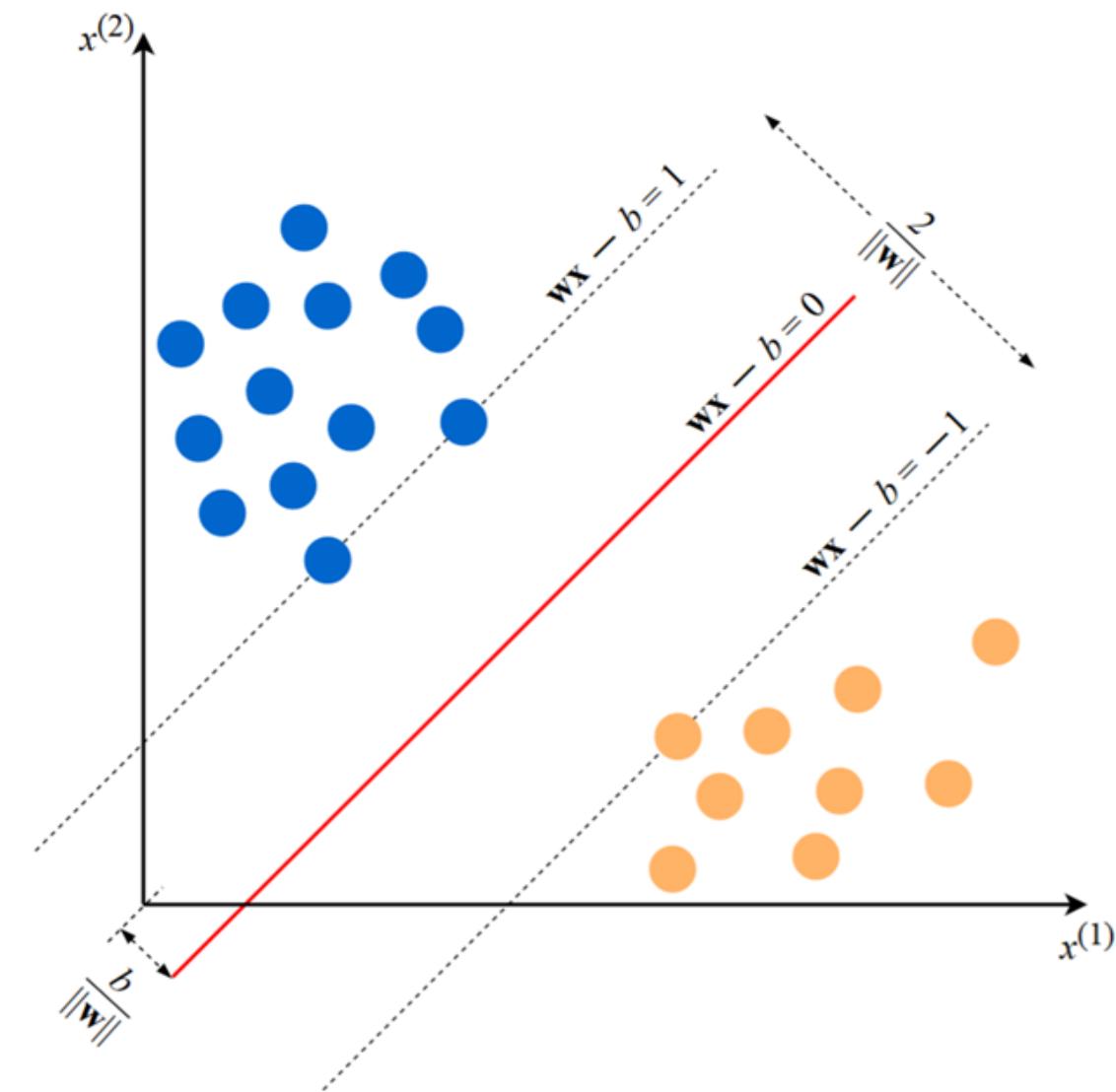


Interpretabilidad

Métodos que permiten entender el comportamiento de los sistemas de machine learning a los humanos.

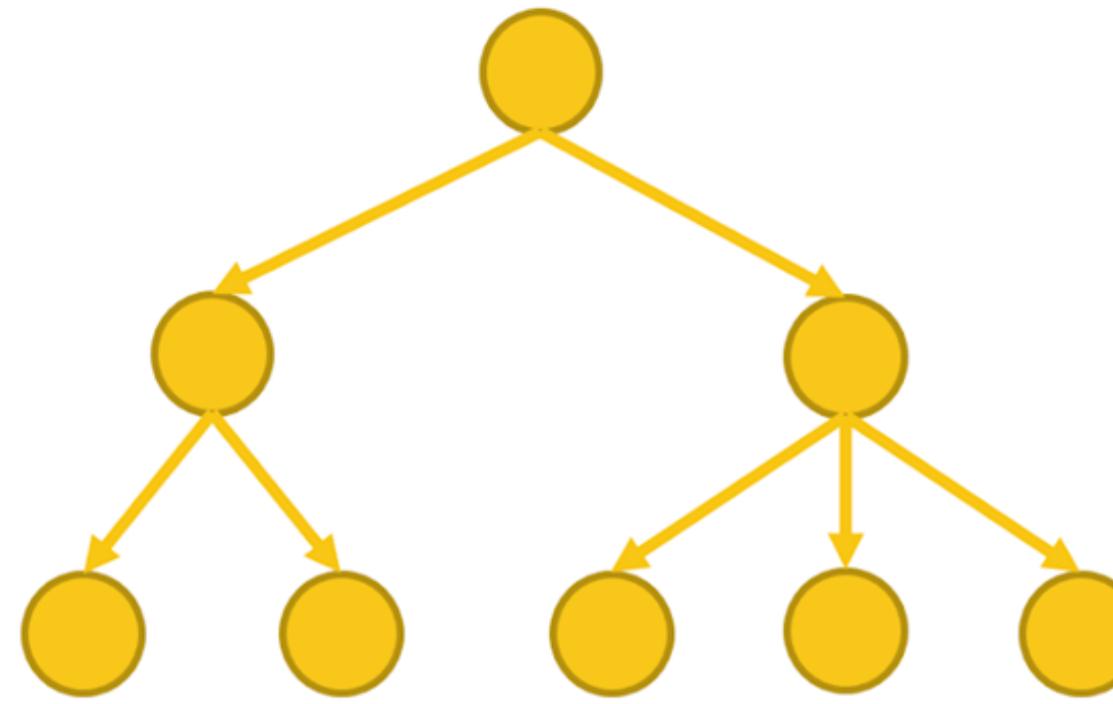
“El grado que un humano puede entender la causa de una decisión”

<https://christophm.github.io/interpretable-ml-book/interpretability.html>

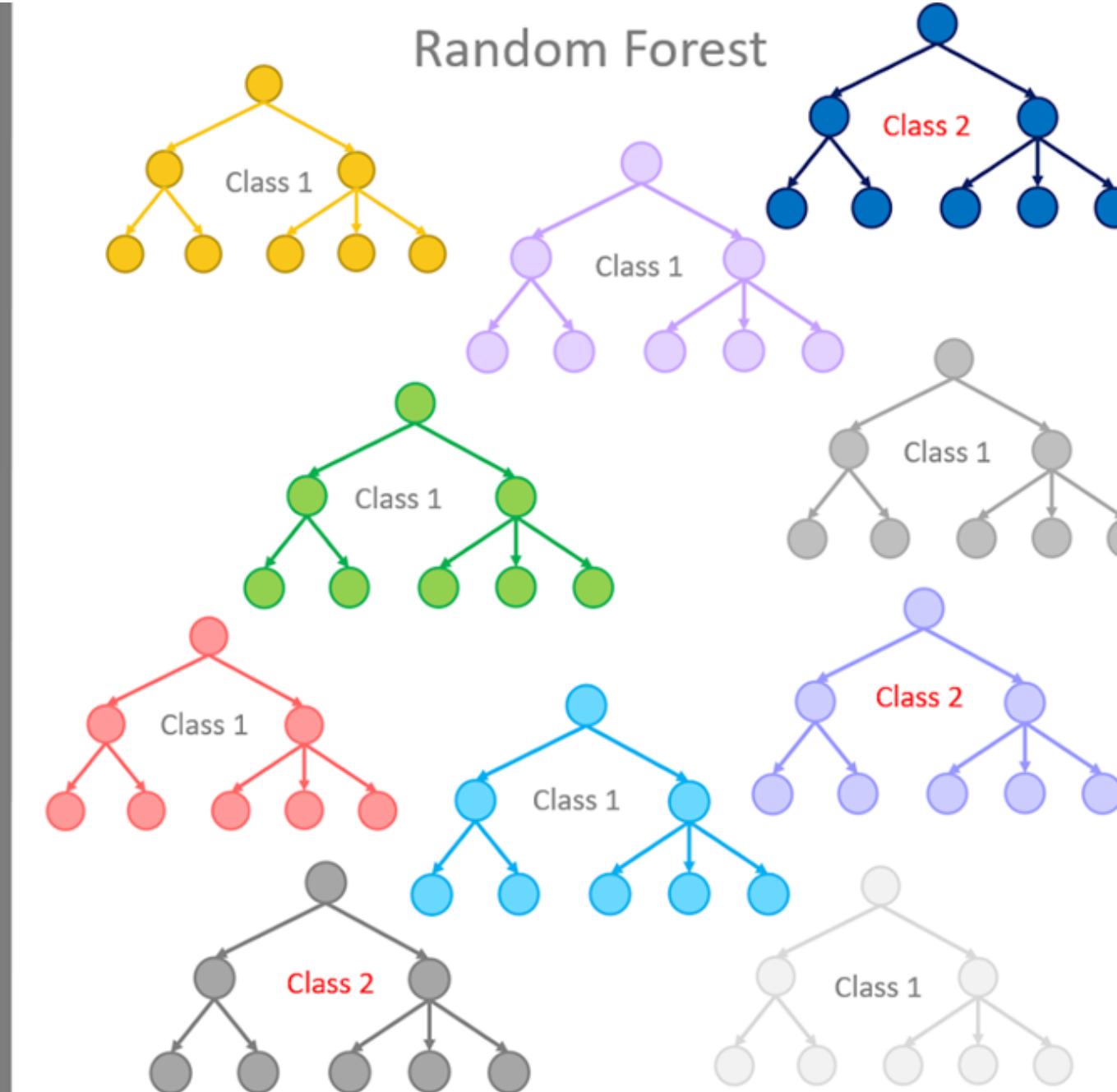


Interpretabilidad

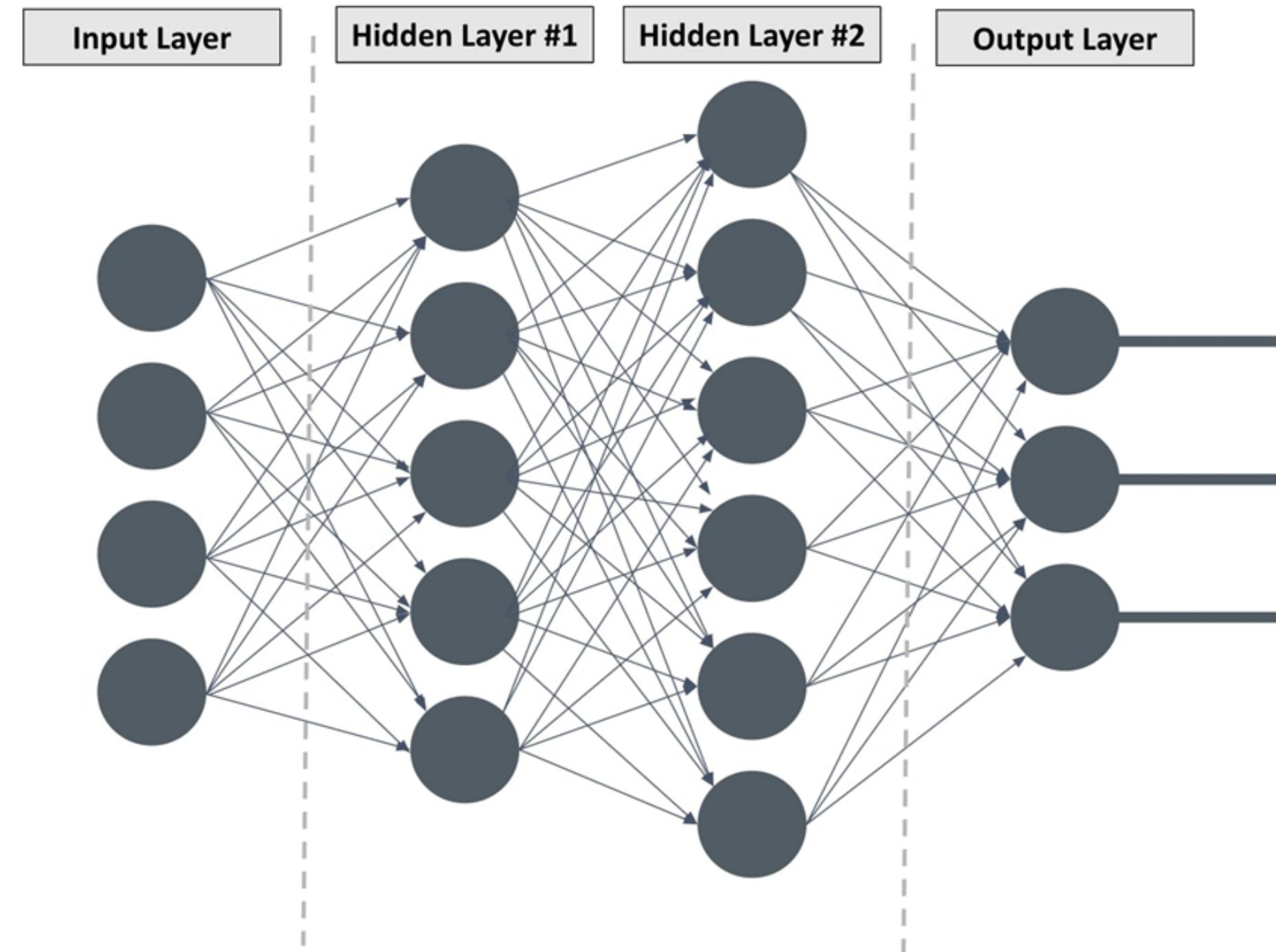
Single Decision Tree



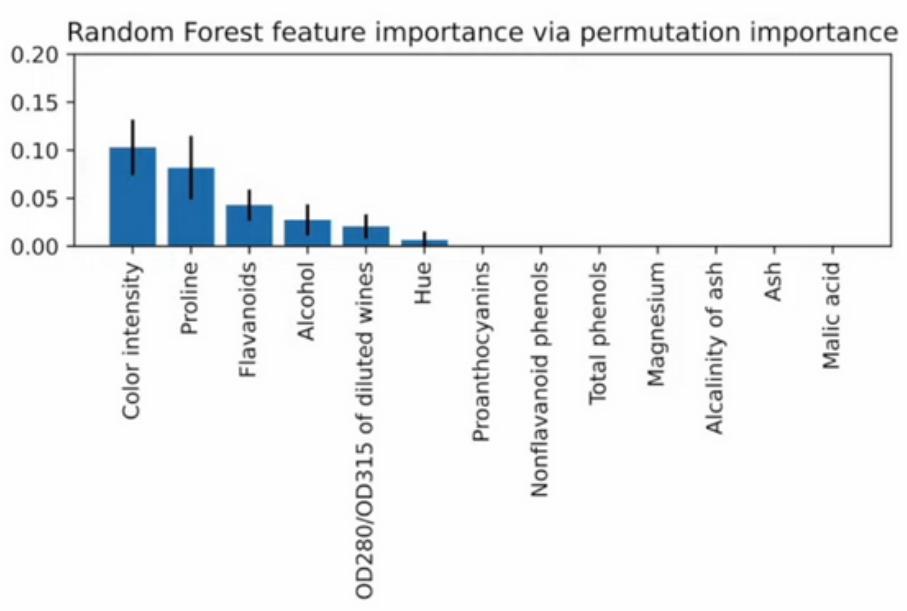
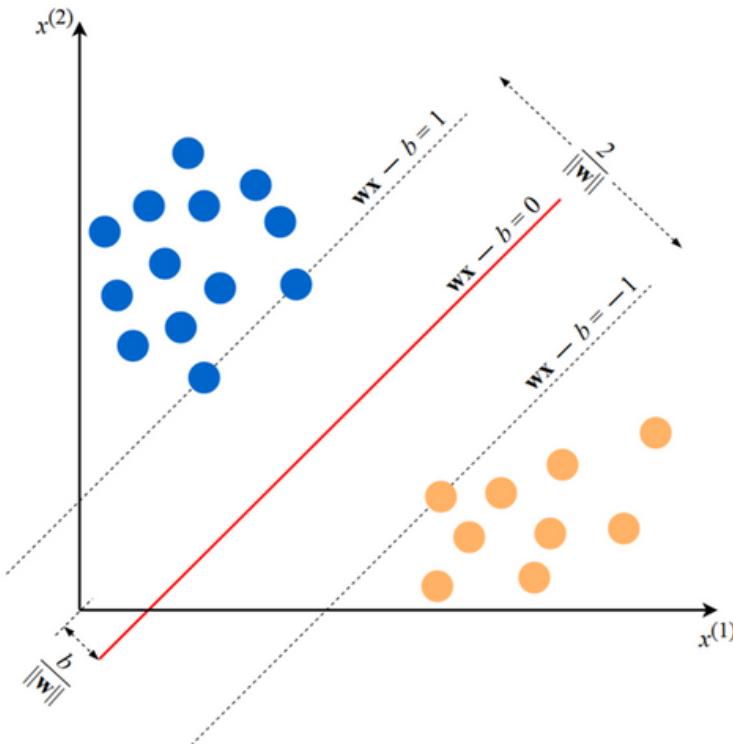
Random Forest



Intepretabilidad

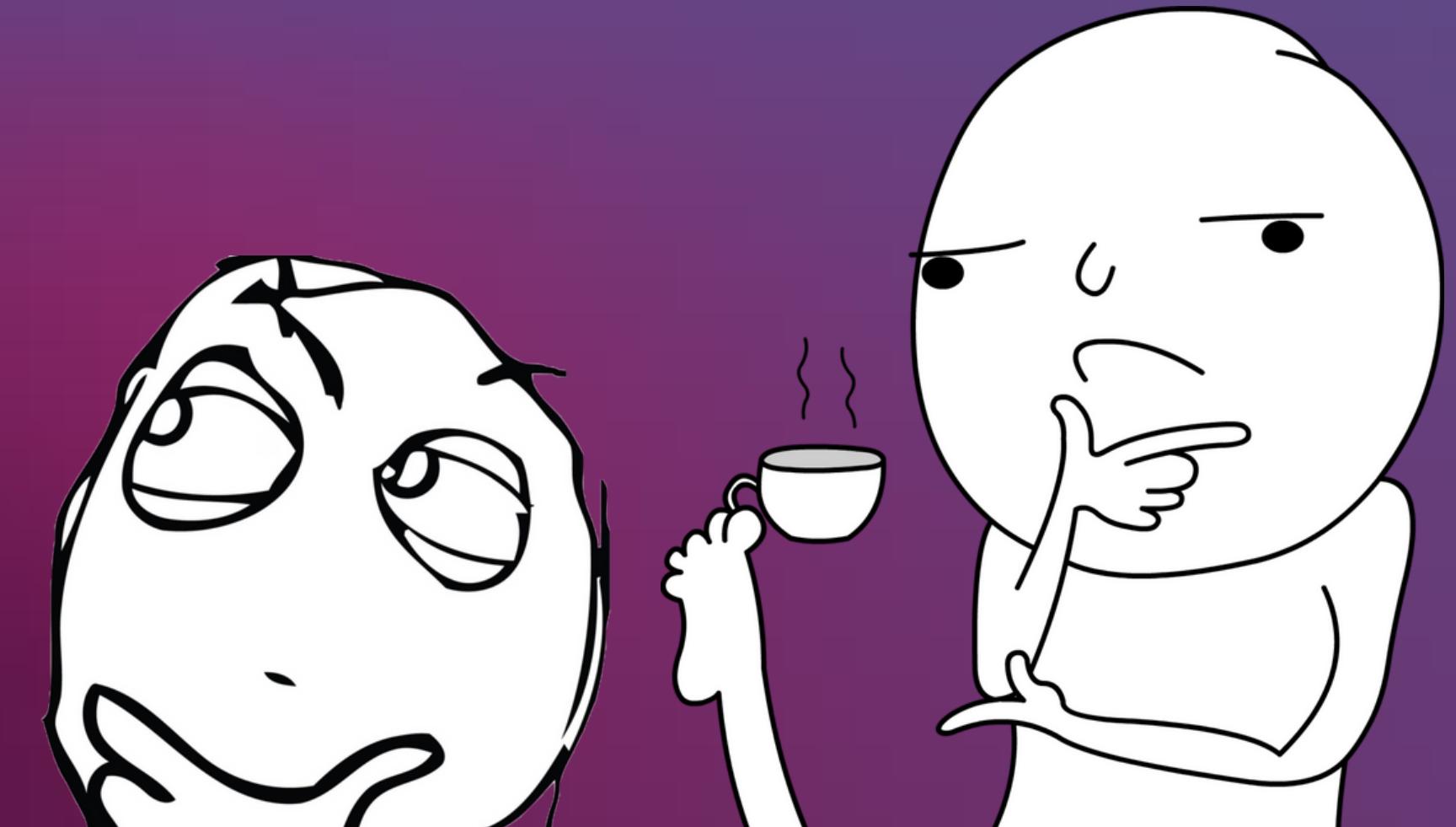


Ejercicio: Interpretar modelo de Vinos



Mientras mas explicación podremos darle a nuestros modelos, la producción del mismo será menos riesgosa y los pasos a producción significaran meno costosos.

¿Es Machine Learning la solución a todos los problemas?



Cuando Usar ML

- **El problema es muy complejo para ser codeado.**
Ejemplo: Detector de SPAM
- **El problema es de tipo perceptivo (voz, texto, imágenes, ...).**
Imposibilidad de analizar de imágenes pixel a pixel. Mismo caso con voz o texto.
- **El problema nunca ha sido estudiado.**
Generar plan de medicinas para pacientes basados en genética y datos médicos disponibles de este.
- **El problema es muy simple.**
Problemas si/no.

Cuando No Usar ML

- **Todas las acciones deben ser explicables.**
Asignador automatizado de créditos.
- **El costo de un error es muy alto.**
Error al detectar alguna enfermedad mortal.
- **Obtener datos correctos es casi imposible.**
¿Puedo hacer un modelo para detectar fraudes en el SII si no tengo datos y definitivamente no me los van a pasar?
- **El problema puede ser resuelto haciendo software tradicional.**
- **El fenómeno tiene muchas posibles salidas.**

¿Las Máquinas Aprenden o piensan?



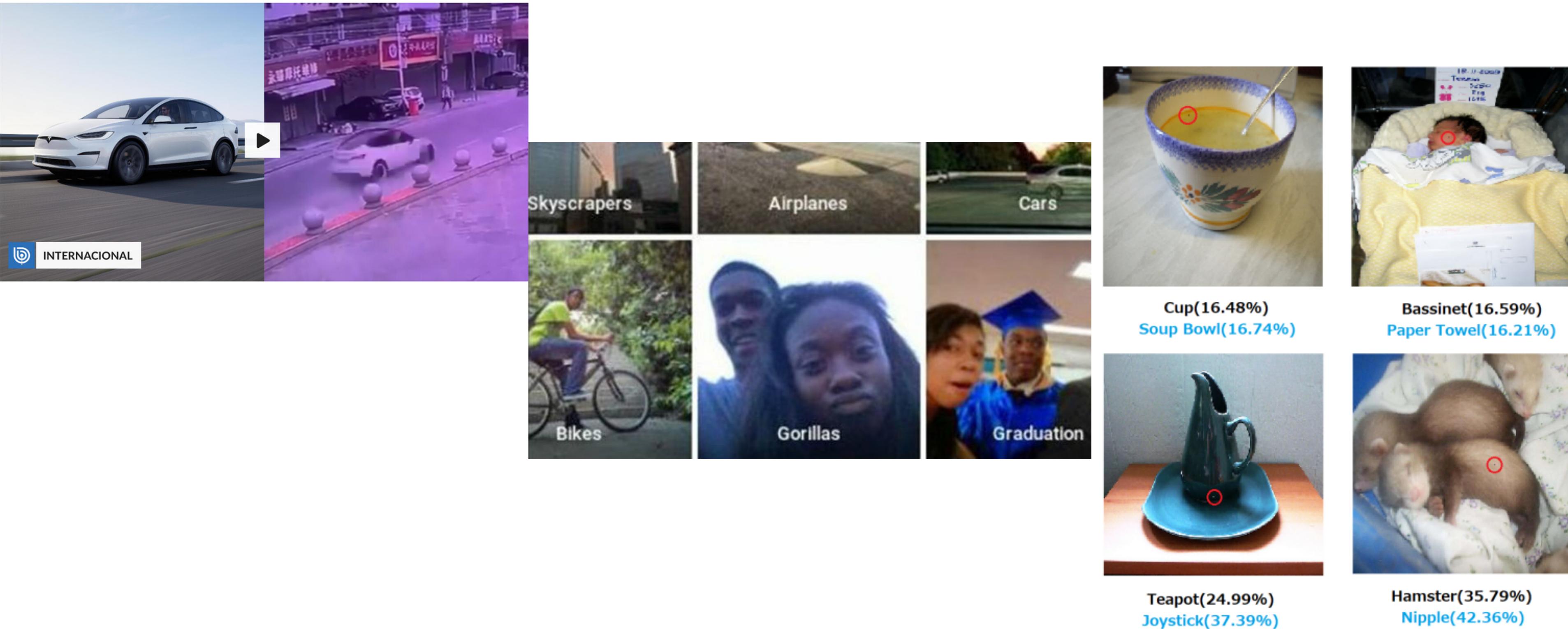
¿Las Máquinas Aprenden o piensan?

Probablemente no en el estado actual.

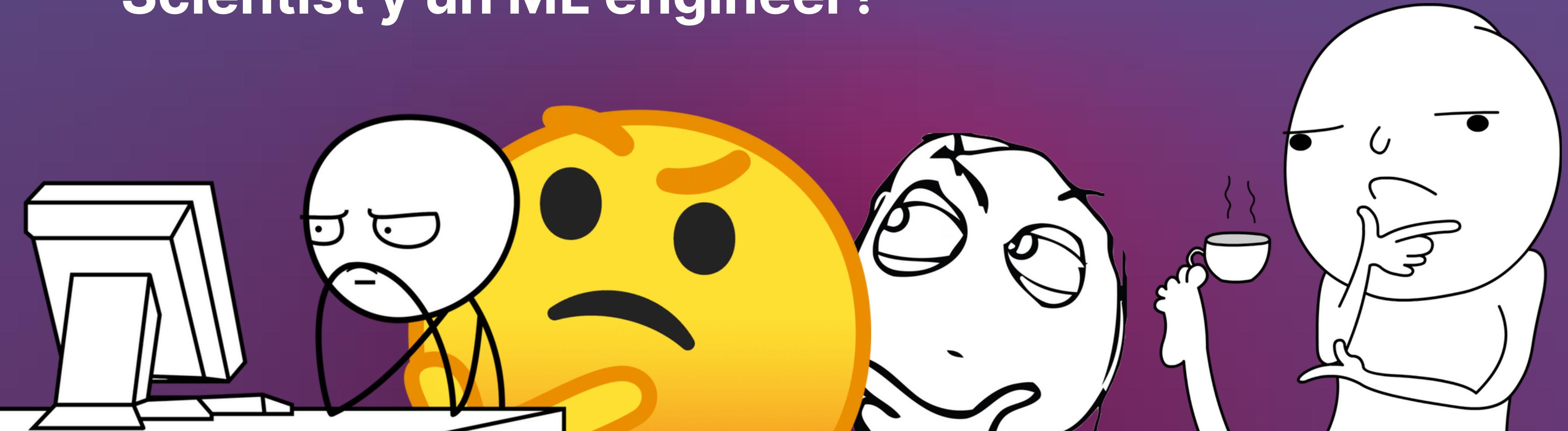
Machine Learning genera modelos matemáticos que infieren patrones a partir de los datos.

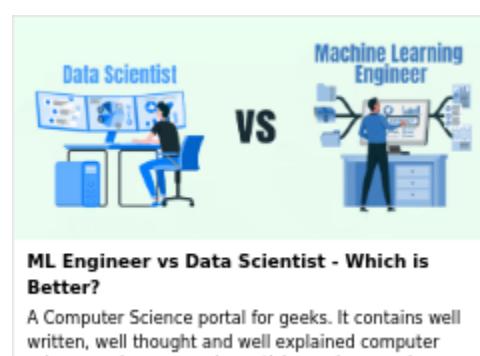
A diferencia de nuestra inteligencia, si la entrada cambia (aunque sea muy poco), las predicciones tenderán a equivocarse.

¿Las Máquinas Aprenden o piensan?



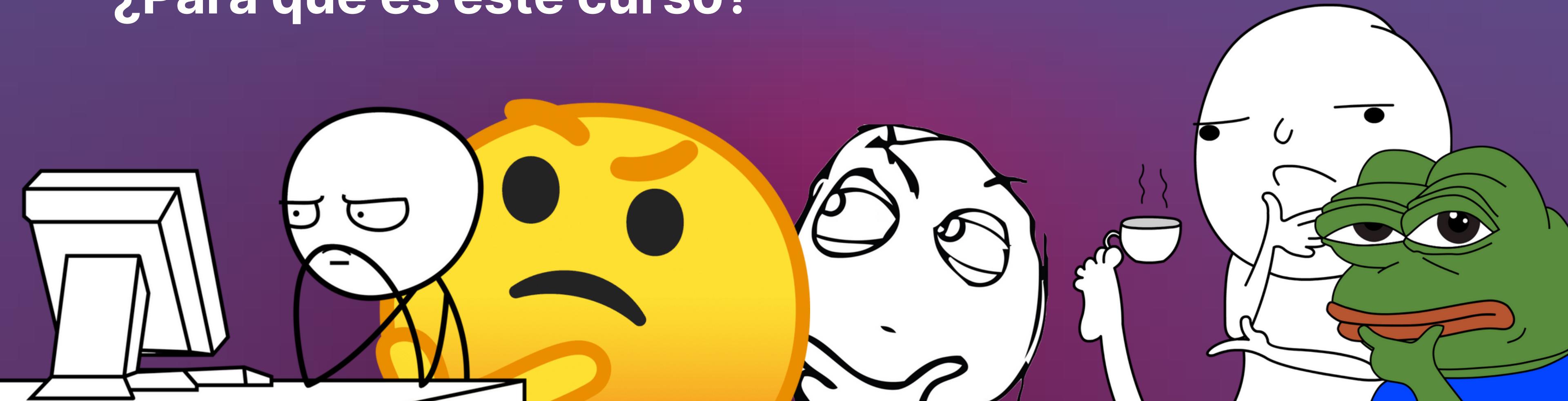
y.... ¿Cual es la diferencia entre un Data Scientist y un ML engineer?





Data Scientists	Machine Learning Engineers
Data Scientists are analytical experts who analyze and manage a large amount of data	Machine Learning Engineers are those who focus on researching, building, and designing self-reliant artificial intelligence (AI) systems to automate predictive models.
Skills needed to build a career as Data Scientist: Statistics, Data analysis, and visualization, Machine learning, Data Wrangling, SQL/NoSQL, Programming skills, Math, Probability, and Coding Skills (Python, R, etc.)	Skills needed to build a career as Machine Learning Engineer: Prototyping, Data Modeling, Programming skills (Python, SQL, Java, etc), statistics, Probability,
Helps in developing data annotation strategies	Helps in controlling the version of models, experiments, and metadata
Data Scientists help in the development of custom tools in order to optimize the complete modeling workflow	Machine Learning Engineer helps in the development of custom tools in order to optimize the complete deployment workflow
They work by visualizing and analyzing the data at various stages of Machine Learning lifecycle	They work by optimizing numerous models for memory, performance, throughput, and latency
Career Path: Data Engineer, Data Analyst, Data Scientist, Business Intelligence Analyst, Data Architect, etc	Career Path: Cloud Engineer, Machine Learning Engineer, AI Engineer, Human-centered AI systems designer, Computational linguist, etc
The average growth rate of Data Scientists: is 30.0% per year	The average growth rate of Machine Learning Engineer: 42.8% per year

Mucho contexto...
¿Para que es este curso?



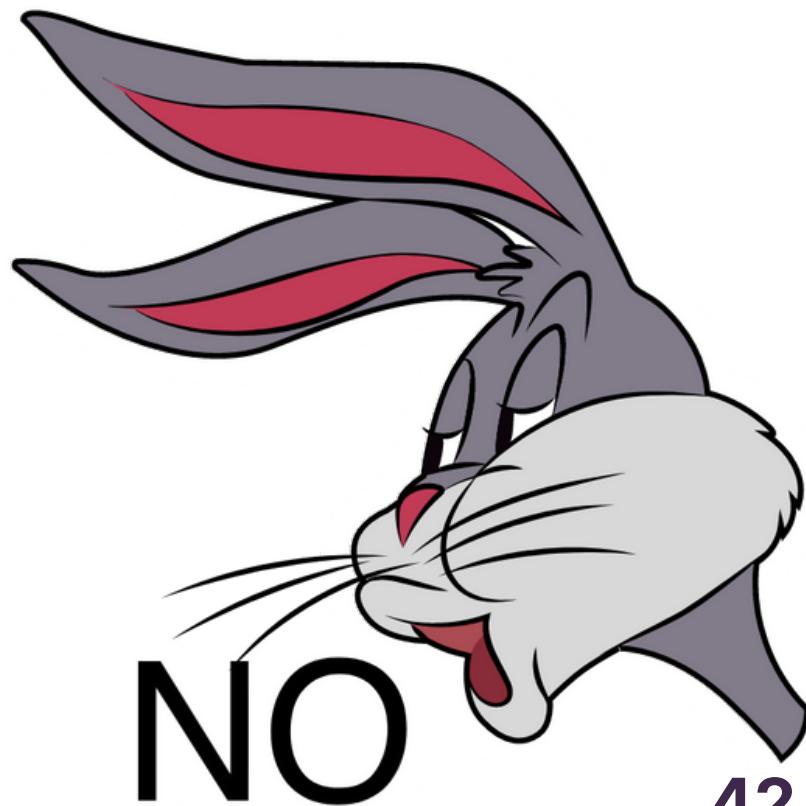
¿Qué no es este curso?

No es un curso teórico:

- Veremos aspectos teóricos solo en cuanto se necesite. Pero se pondrá énfasis en la práctica.
- Múltiples opciones para profundizar (<http://mds.uchile.cl/programa/>):

Electivos (catálogo 2020)

- EL7006 Redes Neuronales y Teoría de Información para el Aprendizaje
- EL7007 Introducción al Procesamiento Digital de Imágenes
- EL7024 Teoría de Información: Fundamentos y Aplicaciones
- EL7037 Computación Evolutiva
- EL7014 Diagnóstico y Pronóstico de Fallas
- EL7031 Robotics, Sensing and Autonomous Systems
- EL7021 Seminario de Robótica y Sistemas Autónomos
- CC6204 Deep Learning
- CC5212 Procesamiento Masivo de Datos
- CC5213 Recuperación de Información Multimedia
- CC7220 La Web de Datos
- CC5208 Visualización de Información
- CC5113 Aprendizaje Automático Bayesiano
- CC5615 Business Analytics
- CC5509 Reconocimiento de Patrones



¿Qué no es este curso?

Proyecto de Ciencia de Datos (aka, el otro curso core del magister):

- No tendremos un único proyecto grande con un cliente real.
- No tendrán la experiencia de trabajar directamente en la industria.

Sin embargo, veremos en detalle paso a paso de la metodología y las tecnologías que se ocupan en estos para resolver proyectos de ciencia de datos.

¿Qué no es este curso?

Lugar para usar software cerrado:

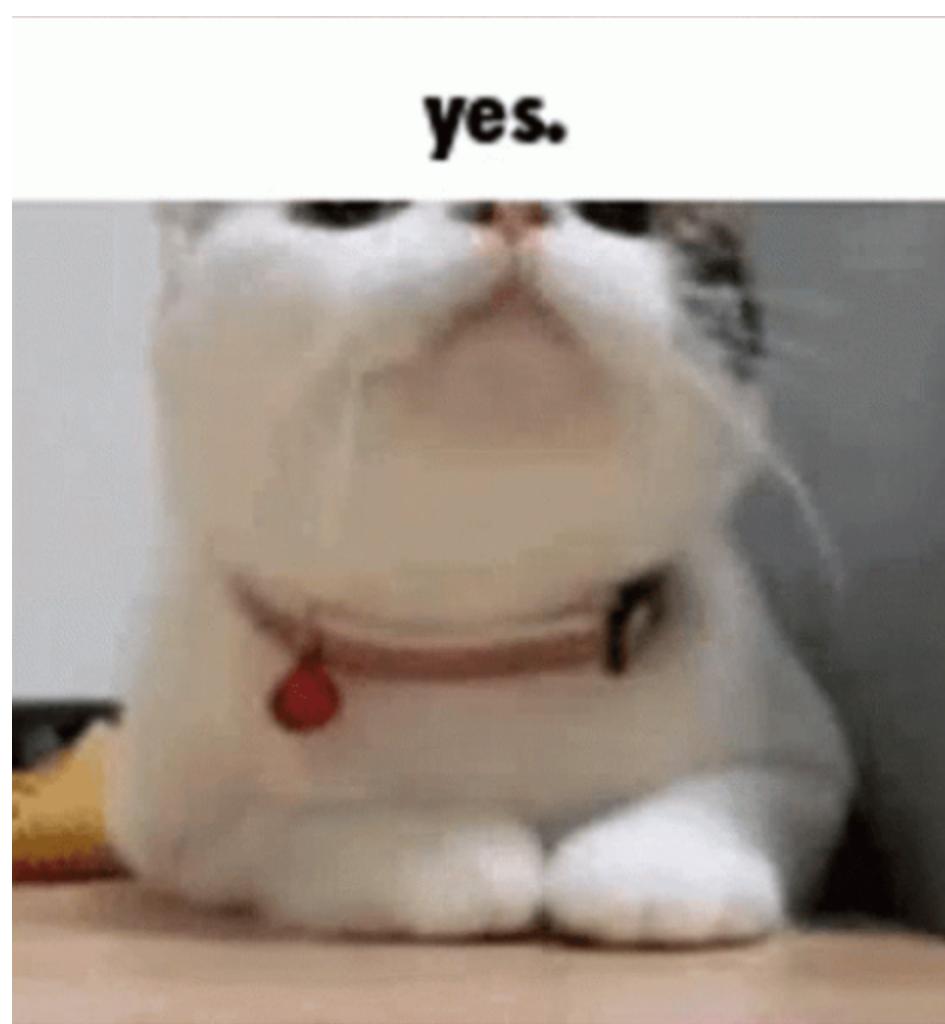
- Trabajaremos con un conjunto de software open-source definido
- No usaremos ninguna solución cerrada pero distribuida gratuitamente (freeware) ni que se venda (software privativo).

Tampoco es un muestrario de open-source software:

- Usaremos las librerías con un fin.
- Las librerías probablemente sean reemplazables por alguna alternativa. Pensar siempre que podrían quedar desactualizadas a futuro.

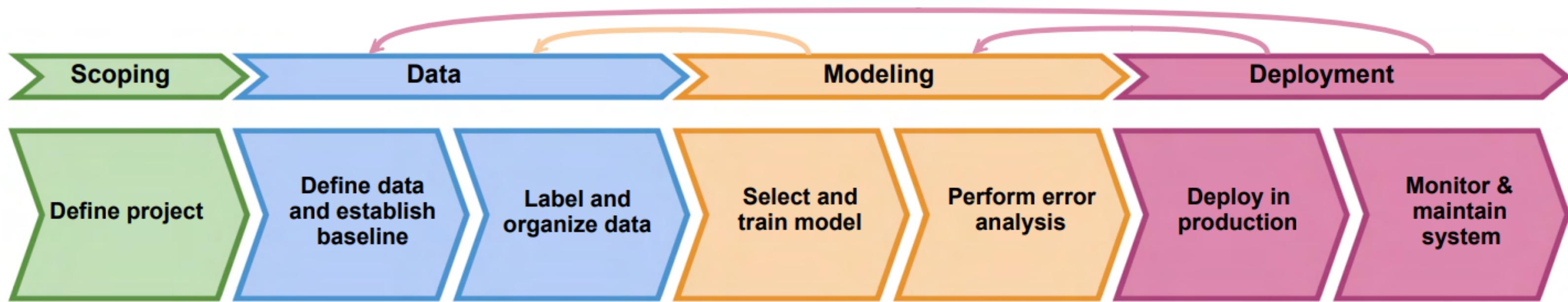
¿Qué **si** es este curso?

El curso estará enfocado en entregar las herramientas necesarias, tanto teóricas como prácticas, para el análisis, modelamiento, resolución y puesta en marcha de proyectos en ciencia de datos.



Metodología del Curso

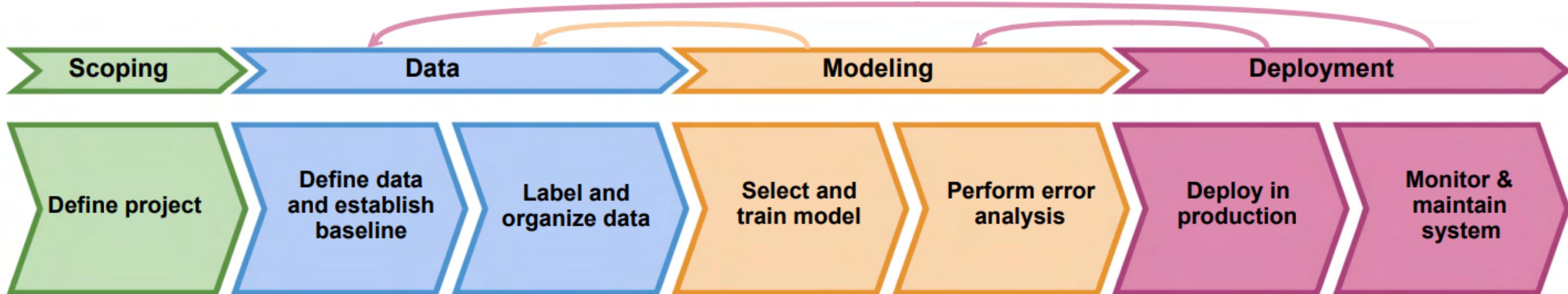
Metodología del Curso



Metodología del Curso

- **Unidad 1:** Introducción a las Herramientas para Ciencia de los Datos (3 sem).
- **Unidad 2:** Manejo de Arreglos Multidimensionales y computación Científica (1 sem).

- **Unidad 4:** Modelos de Aprendizaje Automático (5 sem).
- **Unidad 5:** Optimización de Código y Despliegue (3 sem).



- **Unidad 3:** Manejo de Datos Tabulares y con Pandas (2 sem).
- **Unidad 4:** Visualizaciones y Análisis Exploratorio de Datos (1 sem).

Herramientas por Revisar



Información Administrativa

Reglas del curso ✓



Reglas del curso ✓

Curso Presencial

Todas las cátedras serán tutoriales en donde resolveremos problemas desafiantes usando distintas tecnologías aplicadas a la ciencia de los datos.

¡La idea es que participen!

- Cátedras de 1.30 hora.
- Pausas de 5 minutos a los 40 minutos de clases.
- No hay asistencia obligatoria a las cátedras.
- La colaboración es fundamental para resolver problemas complejos. Este será un concepto fundamental en el curso. Todas las evaluaciones serán grupales.

Evaluaciones

Las evaluaciones este semestre serán:

12 Laboratorios

1 Proyecto que será dividido en 2.

Repositorio

Labs

Evaluación corta que evalúa las herramientas vistas en la semana.

- Serán por ZOOM  en horario de auxiliar. Cada equipo tendrá su propia sala y canal para trabajar.
- El Auxiliar estará presente para contestar dudas.
- Estarán diseñados para que se puedan contestar en el horario del auxiliar más un pequeño tiempo de trabajo en casa.

Labs

Reglas:

- **Grupos de 2** (puede ser individual, pero no se recomienda).
- **No se aceptaran laboratorios** de alumnos que no hallan asistido a la sesión.
- **6 días de plazo + 3 días de atraso** con descuento de **1 punto por día**. Entregas el Jueves a las 23.59.
- Días bonus: **5 días para atrasarse sin penalización** y distribuibles como ustedes quieran.

marzo de 2023							▲	▼
lu	ma	mi	ju	vi	sá	do		
27	28	1	2	3	4	5		
6	7	8	9	10	11	12		
13	14	15	16	17	18	19		
20	21	22	23	24	25	26		
27	28	29	30	31	1	2		

Labs

Reglas:

- Se pide **respeto por los horarios hábiles**. No se responderá el sábado/domingo en la noche.
- **Consultas por correo**.
- Consultas del lab anterior permitidas en el lab de la semana actual, pero evitarlas.

Incentivamos conversar con sus compañeros acerca de la soluciones de los labs, pero:

- **Está prohibida la copia** (obviamente).
- También que solo uno del equipo trabaje. La idea es que ambos resuelvan en conjunto el lab!
- **Si existen problemas con sus compañeros, comentarlo al equipo docente.**

Proyecto

Evalúa el contenido general del curso.

- Constan de 2 entregas (notebooks con informe + código).

Reglas:

- Grupos de 2.
- Plazo y problema por determinar. Sin atrasos.
- Para la segunda entrega habrá una competencia en donde se medirán con sus compañeros. Los mejores modelos serán recompenzados.

Repositorio

Contiene todas las evaluaciones del curso.

La idea es que integren esta herramienta de trabajo dentro de sus metodologías usuales.

Se evaluará:

- Rama master o main en donde estarán todas las evaluaciones ya terminadas. Las entregas serán revisadas por este medio.
- Ramas lab{n} en donde deberán tener la el desarrollo del lab y proyecto actuales. La idea es que a medida que trabajen vayan guardando su lab/proyecto no finalizado en esta rama.

Se irá registrando en cada evaluación el correcto uso de git.

Evaluaciones

La nota final se calcula como:

$$\mathbf{NF = 70\% \, labs + 25\% \, proyecto + 5\% \, Repositorio}$$

Observación: Se debe aprobar labs y proyecto por separado para pasar de curso.

¡Recuerden que el objetivo es aprender!
Por ende, está totalmente prohibida la copia

Canales de Comunicación

Foro de U-cursos:

- Cualquier duda de los contenidos del curso y administrativas.
- Todos los mails con consultas al equipo docente serán redirigidos al foro.
(Nuevamente, colaboración...)

Github: Repositorio Oficial: Todo el material del curso se encontrará ahí, incluido el calendario.



Calendario

💡 Podrá estar **sujeto a cambios**, tanto por la pandemia como por cualquier otra eventualidad. La versión actual la podrán encontrar en el repositorio del curso.

Semana	Módulo	Fecha	Temas	Detalle	Evaluación
1	Introducción a las Herramientas para Ciencia de los Datos	3/13/2023	Introducción al curso. IDEs y Jupyter. Markdown. Ambientes de Ejecución	Primera parte: Explicación del objetivo del curso, los contenidos, los métodos de evaluación y las reglas. Segunda parte: Trabajo con Jupyter Notebook/Lab y Markdown como entorno de trabajo para Data Science. Tercera Parte: Creación y administración de ambientes de ejecución de Python.	
		3/15/2023	Control de versiones con Git	Fundamentos de Git, repositorios, ciclos de vida de los archivos en un repositorio, Commits, Branches y colaboración y repositorios remotos.	
		3/17/2023			
2		3/20/2023	Python 1: Introducción a Python	Introducción a Python: Sintaxis, variables, tipos de datos básicos, expresiones y operaciones, control de flujo, colecciones (listas, tuplas, conjuntos, diccionarios), iteraciones.	
		3/22/2023	Python 2: Programación Funcional y Modular.	Funciones, scope, unit testing. Decoradores. librerías built-in y programación modular.	
		3/24/2023			Lab 1: Git
3	Manejo de Arreglos Multidimensionales y computación Científica	3/27/2023	Python 3: Programación Orientada a Objetos y Excepciones	Clases, objetos, constructores, abstracción y encapsulación, herencia, polimorfismo. Excepciones y manejo de estas. Quizas generadores.	
		3/29/2023	Arreglos Multidimensionales con Numpy	Creación de arreglos multidimensionales, vectorización de operaciones, atributos de los arreglos, indexado, operaciones básicas, documentación, funciones universales,	
		3/31/2023			Lab 2: Python

Clase 1: Introducción

LABORATORIO DE PROGRAMACIÓN CIENTÍFICA PARA CIENCIA DE DATOS

MDS7202-1 – Otoño 2023