
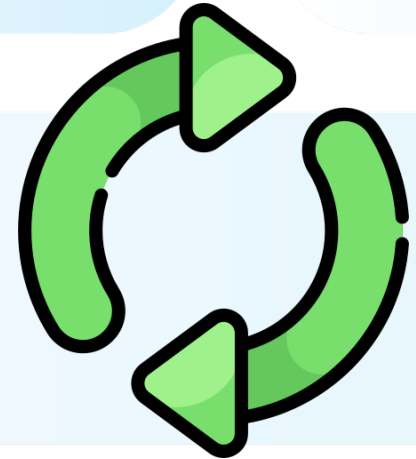


Iterate-on-data




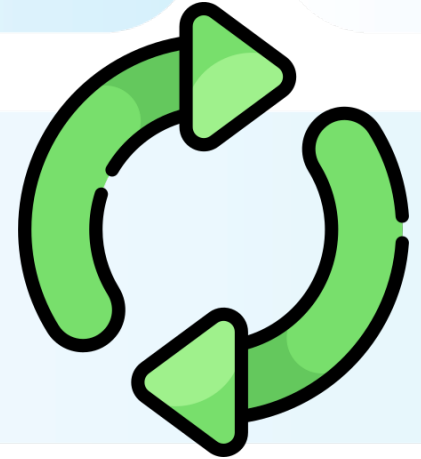
La creación de modelos es iterativa

- El machine learning es iterativo por naturaleza
- Realizas pequeños cambios, entrenas un modelo y analizas resultados
- Cambio en los hiperparámetros y la forma de procesar los datos:  **“Model-centric AI”**



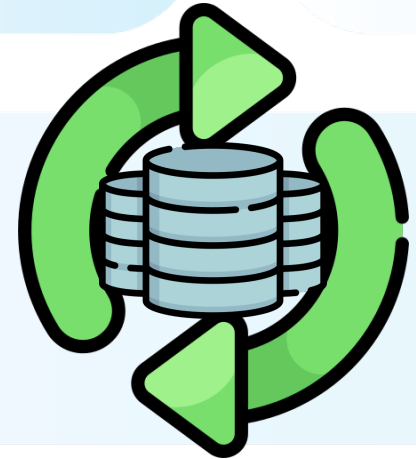
Iterate-on-data

- Mejorar los datos es otra forma de mejorar el modelo.
- “Data-centric AI” 
- Yo prefiero llamarle: itera sobre los datos, o **iterate-on-data**



Iterate-on-data

- Los datos son otro elemento que contribuye al modelo
- Los datos en el mundo real están en constante cambio
- Puedes comenzar con un dataset limitado e irlo mejorando con el tiempo



Iterate-on-data

- Más datos \neq mejores datos
- Mejores datos: más limpios, más relevantes o mejor pre-procesados
- Ya has lidiado con falta de datos o errores en tu entendimiento de los datos.
- Ya has hecho *iterate-on-data*



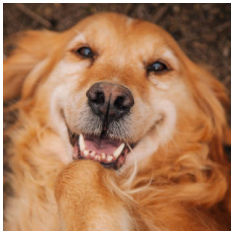
Mejores datos

- Más datos... pero de calidad
- Añadir datos solo por añadir puede aumentar el tiempo de entrenamiento sin incrementar el desempeño del modelo



Datos consistentes

- Consistencia en el dataset:



Perro



Animal

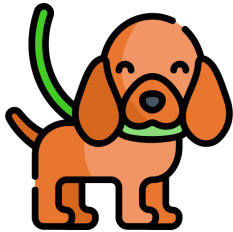


Canino

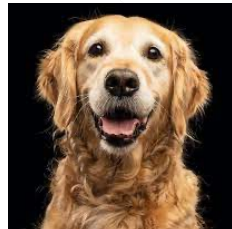
Etiquetado inconsistente



- Consistencia con el mundo real:



Entrenamiento



Mundo real



Datos de calidad

- Correctos
- Temporales
- Completos
- Granulares
- Confiables
- Disponibles
- Relevantes



Iterate-on-data

- Iterar sobre los datos es algo común
- Mi invitación es que lo hagas de forma consiente
- No debe reemplazar la iteración en el modelo

