Inteligencia artificial

MLOps, la operación de los procesos de Machine Learning

Cristina Jerez
UPCnet - Universitat Politècnica de Catalunya

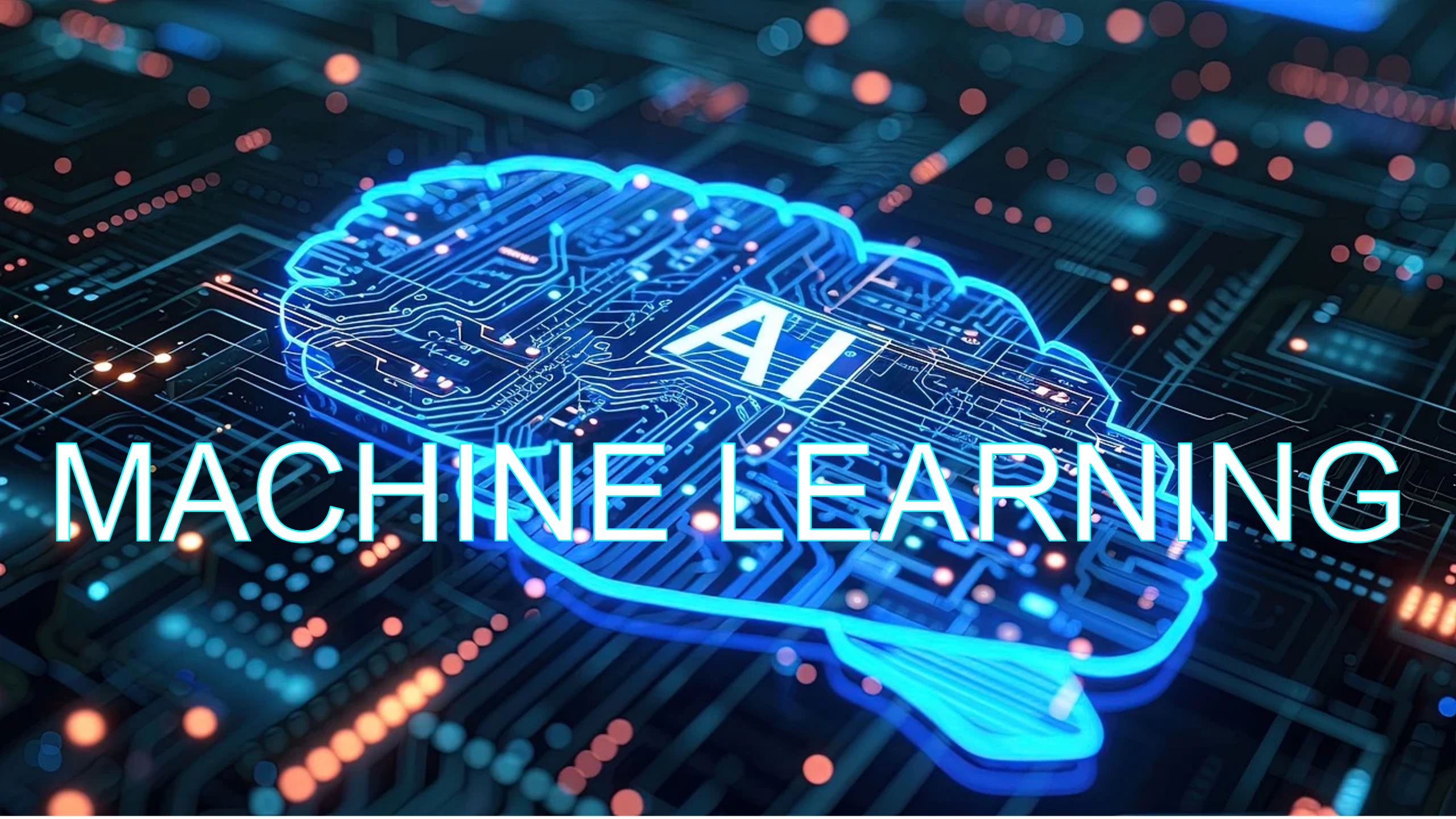
David Garcia
UPCnet - Universitat Politècnica de Catalunya

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

UPCnet

Mayo 2024


Transformers
Entrenamiento
Alucinación
Embedding
Inferencia
Fine-Tunning


AI
MACHINE LEARNING

Data Analysts

Los equipos de data analysts se encargan de recopilar, procesar, analizar e interpretar gran cantidad de datos para poder responder a preguntas o resolver problemas en diversos campos.

Necesidades

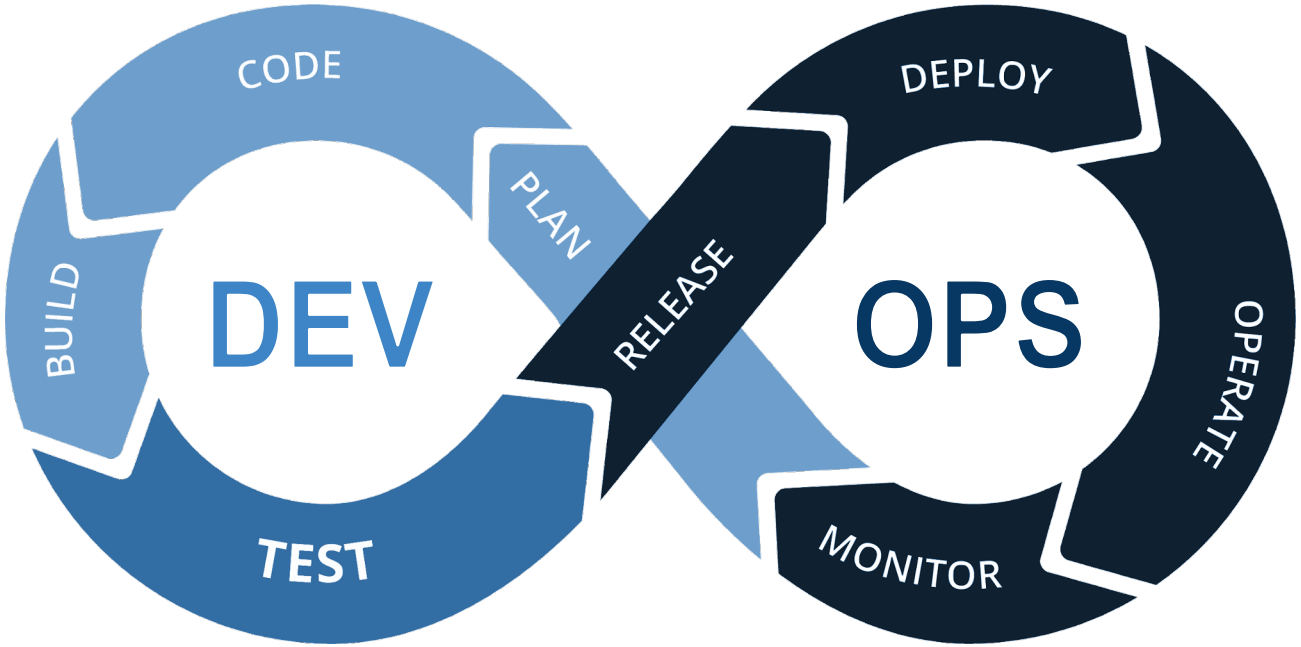
- **Recopilación y limpieza** de los datos con los que trabajar
- **Entornos de Programación** adaptados para trabajar con este ecosistema de datos
- **Facilitar el testeo del modelo** para la evaluación del correcto funcionamiento con i**nfraestructuras adaptables** a los requisitos de los modelos

DevOps…

Los equipos de Operaciones están ya en la rueda del funcionamiento de DevOps, trabajando con equipos de desarrollo generando un flujo de confianza

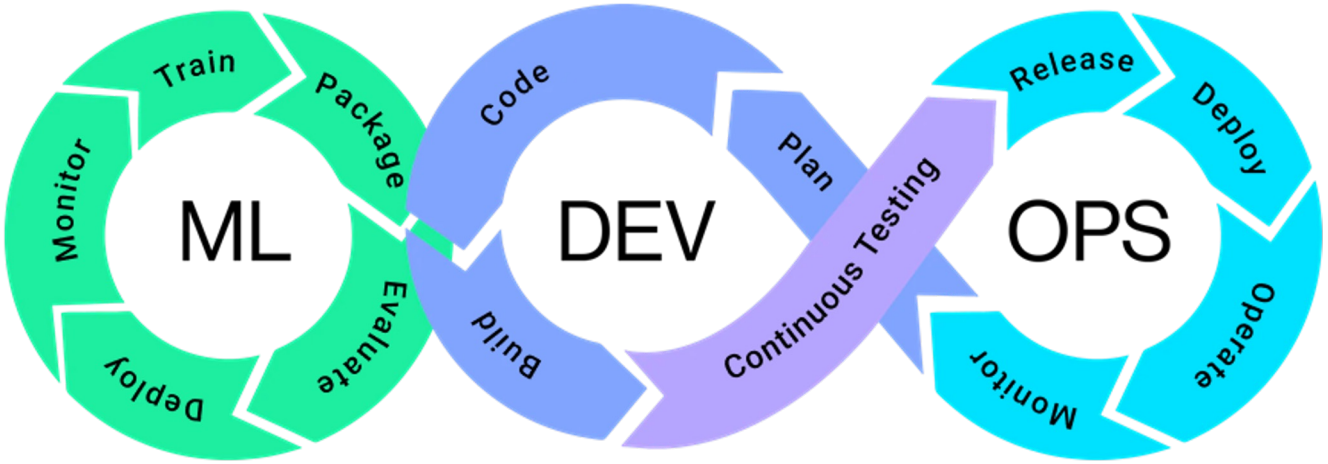
Nuevos retos

- **Data Analysts…** nuevas necesidades
- Entornos de **desarrollo** … con nuevas **tecnologías**
- Entornos **productivos** … con nuevas **capacidades**: GPU, accesibilidad…
- **Operación** de estos entornos … con nuevas preocupaciones vinculadas al **uso** y **seguridad**
- **Monitorización** de estos entornos … con **nuevos procesos**



MLOps

Debemos analizar los sistemas de AI desde una perspectiva de Operaciones de aprendizaje automático (MLOps) para aprender y definir un marco de común en el flujo de entrenamiento, desarrollo y operación para crear crear sistemas de aprendizaje automático robustos, escalables y seguros.



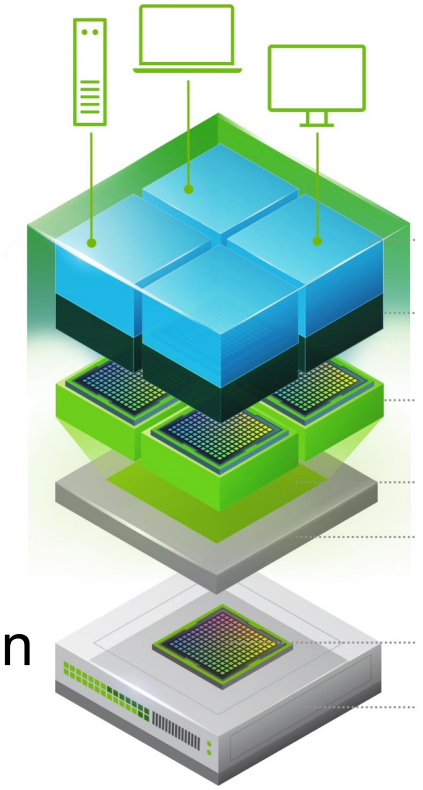
Entrenamiento → Empaquetado → Evaluación → Despliegue → Monitorización

Necesidades vs. Problema vs. Solucion

Entrenamiento	Empaquetado	Evaluación	Despliegue	Monitorización
GPU, necesidades hardware muy elevadas	CUDA, Frameworks de ML	Juego de pruebas representativo	Escalabilidad, seguridad, gestión del cambio…	¿Está respondiendo bien el modelo?
PCs sin recursos Procesos muy lentos	Imágenes contenedores muy pesadas y diferentes entre sí	Pocos datos o demasiado homogéneos	Los mismos que las aplicaciones	Calidad de los datos, drift de propósito, evaluar en base a las “predicciones”
Cloud Environment	Repositorio de imágenes	Recopilar más datos y mejorar hacer Fine-tunning	Incorporación en el flujo de DevOps	Servicios que controlen las alucinaciones y rendimiento

Arquitectura ML

Arquitecturas. Primera aproximación



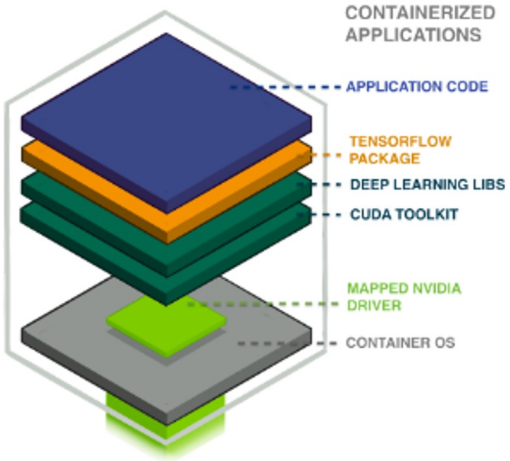
- ✗ Rigidez elección del hardware
- ✗ Gestión cambio poco práctica
- ✗ No escalable
- ✗ Difícil tener más de 1 modelo

CONS

PROS

- ✓ Implementación más sencilla
- ✓ Adquisición de conocimiento
- ✓ Facilidad para entrenar y hacer inferencia en un mismo entorno

Arquitecturas. Contenedores!



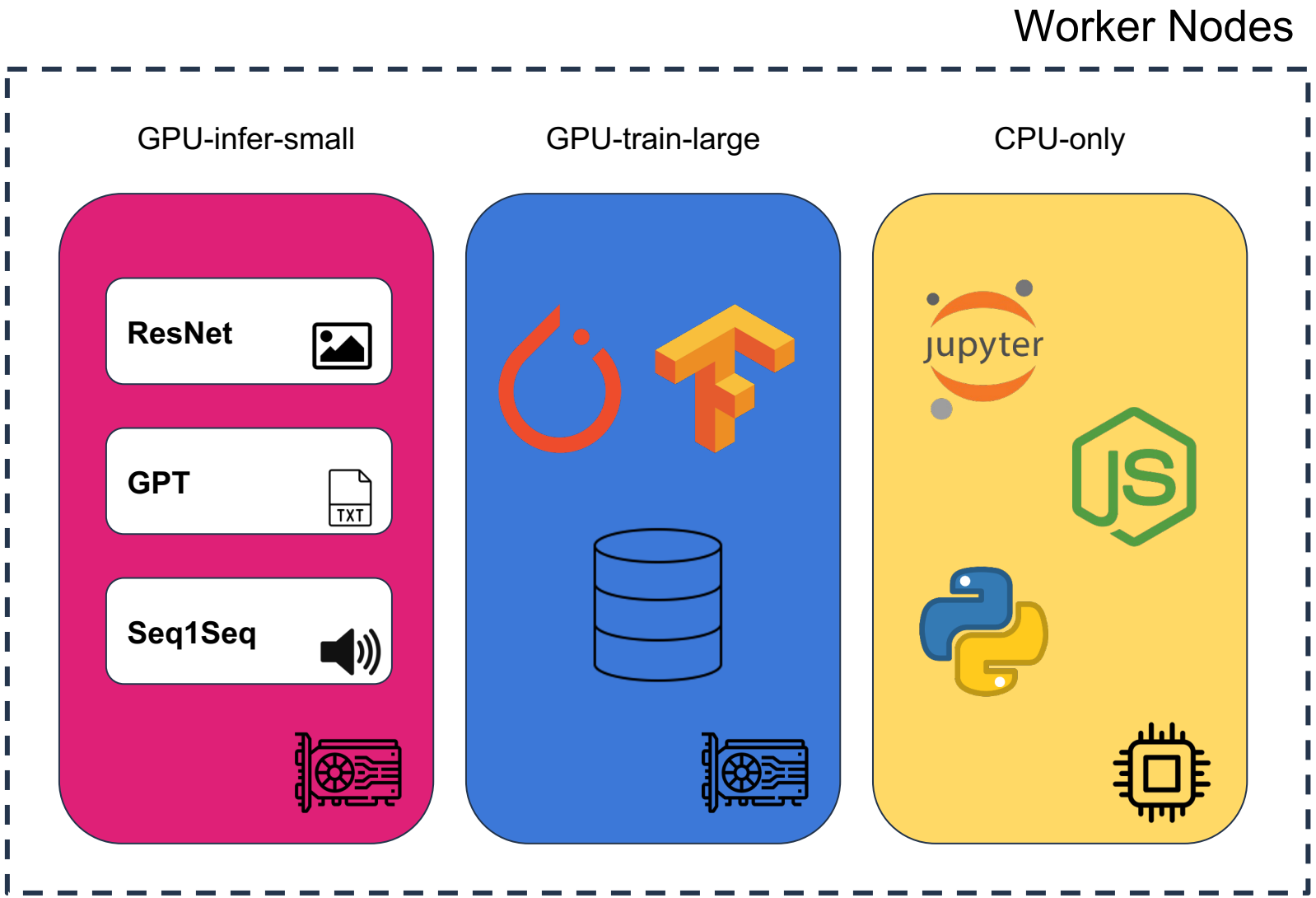
CONS

- ✗ Complejidad para encontrar documentación y requisitos software de los modelos
- ✗ Imagenes muy pesadas (~7-8Gb)
- ✗ Complejidad para hacer pruebas en local por requisitos hardware

PROS

- ✓ Escalabilidad
- ✓ Disponibilidad de imágenes de contenedores diferentes
- ✓ Flexibilidad de requisitos hardware y entornos de ejecución

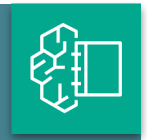
Arquitecturas. Entornos productivos



AWS IA Services

AWS dispone de diversos servicios que nos pueden dotar de capacidades de IA

SAGE MAKER STUDIO



Nos permite desarrollar modelos de forma ágil y sencilla.

Nos ofrece una suite de herramientas que permite ejecutar todos los pasos de creación y entrenamiento de un modelo.

SAGEMAKER

Nos ofrece una plataforma en la que ejecutar nuestros modelos para hacer predicciones.

Es la base para disponer de un entorno gestionado para operar modelos de inteligencia artificial.

JUMPSTART

Este servicio nos permite ejecutar modelos fundacionales para realizar pruebas o hacer fine-tunning.

AWS ofrece un catálogo de modelos ya preparados para desplegarse de forma sencilla en SageMaker.

BEDROCK

El servicio serverless de SageMaker.

Nos permite disponer de varios modelos fundacionales siempre disponibles y sólo pagaremos por los tokens de entrada y salida.

ML Services

PROMPT MANAGEMENT

RESPONSIBLE AI

MONITORING & OBSERVABILITY

DATA LABELING

HUMAN IN THE LOOP

SECURITY

DATA DRIFT

INPUTS/OUTPUTS

IA MODEL

CI/CD

AUTOMATION & INTEGRATION FRAMEWORKS

DATA MANAGEMENT

MODEL EVALUATION

MODEL DEPLOYMENT

HALLUCINATIONS

MODEL SELECTION

COST

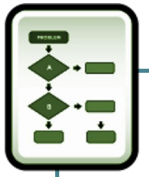
TRAINING INFRASTRUCTURE

MODEL MANAGEMENT

APPLICATION LAYER

API

Conclusiones y Recomendaciones

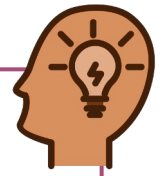


Incorpora los nuevos pasos a controlar dentro del flujo de trabajo

Creación del flujo **MLOPS**

Entender los **nuevos retos**

Comunicación con DataAnalysts para entender los nuevos retos

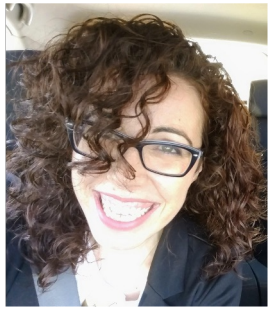


Identificar **servicios** según la **necesidad**

Conoce los diferentes servicios que pueden ayudar a necesidades concretas

Controlar el modelo desde el punto de vista de infraestructura y funcionamiento

Monitorización y **evolución**



Cristina Jerez Alvarez
cristina.jerez-alvarez@upcnet.es

David Garcia Domenech
david.garcia.domenech@upcnet.es