Despliegue de algoritmos

KEEPCODING
Tech School

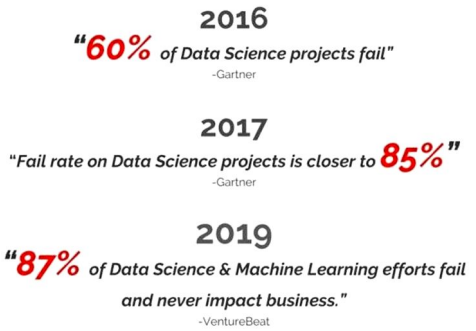
Acerca de mi

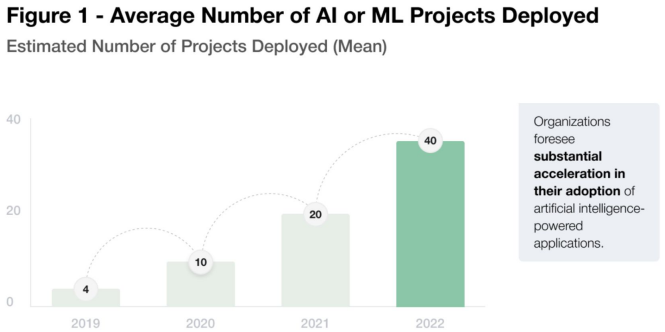
Grado en Ingeniería de materiales (UPM)
Master en Ingeniería Computacional y matemáticas (UOC)

Undergraduate Research Assistant - MIT
Data Scientist - NewCastle University
Sr Machine Learning Engineer - Tecnicas Reunidas
Sr Machine Learning Engineer - BASF

La importancia del despliegue de algoritmos

- Generar valor
- Predicciones de manera automática
- Usuarios sin conocimiento pueden usar los modelos de ML.





¿Qué problema vamos a aprender a solucionar?

Creating and deploying machine learning (ML) models supposedly takes too much time. Quantifying this problem is difficult, not least because there are so many job roles involved with a machine learning pipeline. With that caveat, let us introduce Algorithmia's "2020 State of Enterprise ML." Conducted in October 2019, 63% of the 745 respondents have already developed and deployed a machine learning model into production. On average, 40% of companies said it takes more than a month to deploy an ML model into production, 28% do so in eight to 30 days, while only 14% could do so in seven days or less.

Contenido

- MLOPs
- Tipos de inferencia
- Hardware para inferencia
- Monitorización con MLFlow
- Proveedores cloud
- Apache Beam
- Google cloud (GCP)
- FastAPI

Modelos y entrenamiento

Model Selection

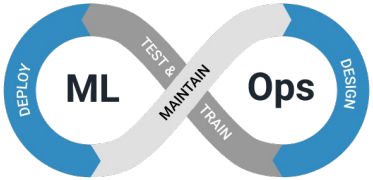
- Que no deberemos hacer: No seguir las buzzword.
- Qué debemos hacer: Escoger el modelo más simple, no él más llamativo.

Be solution-oriented, not technique-oriented

Problemas a la hora de analizar los modelos de ML

- Querer probar el potencial del DL sin mucha inversión.
- Difícil conseguir un buen rendimiento sin inversión en tiempo y dinero en el etiquetado de datos.

MLOPs



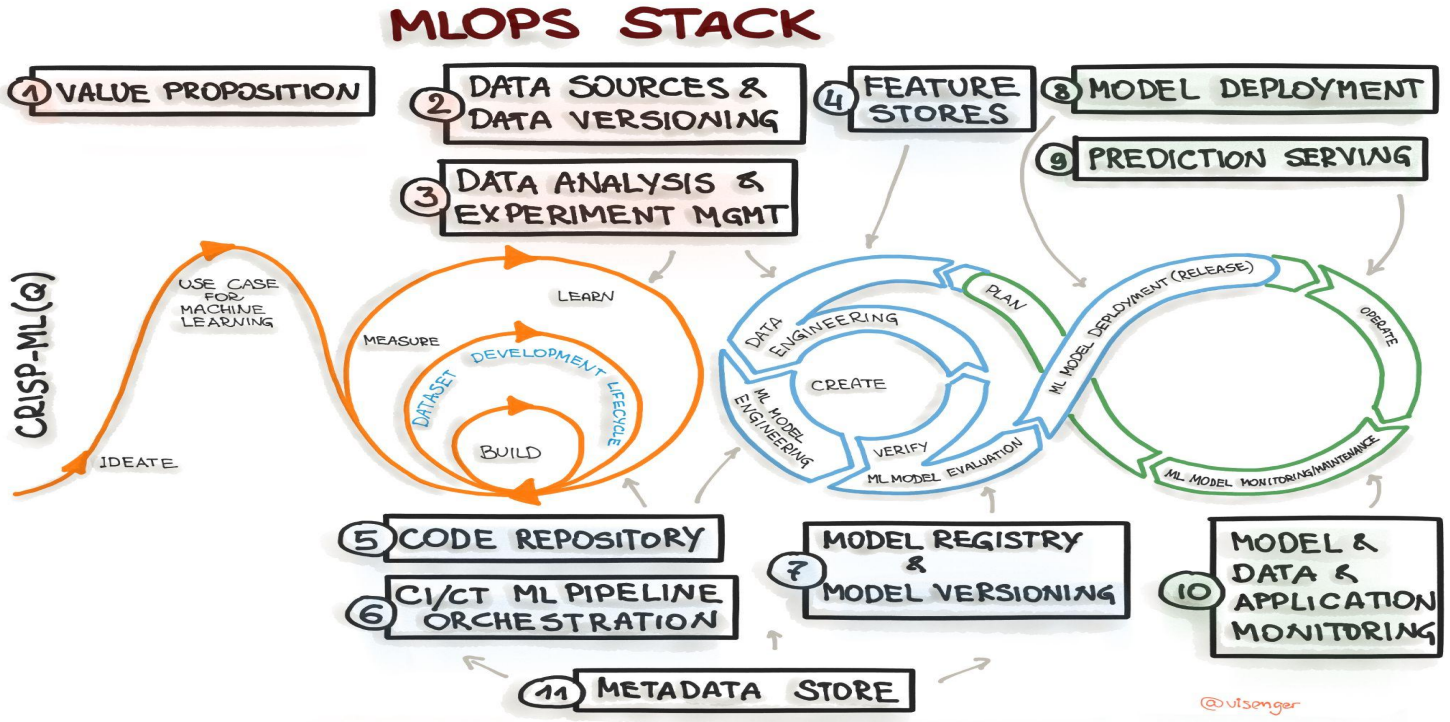
MLOPS

Concepto: MLOps es una práctica y cultura de la ingeniería de Aprendizaje Automático, cuyo fin es unificar e desarrollo (Dev) y las operaciones (Ops) del sistema de Machine Learning.

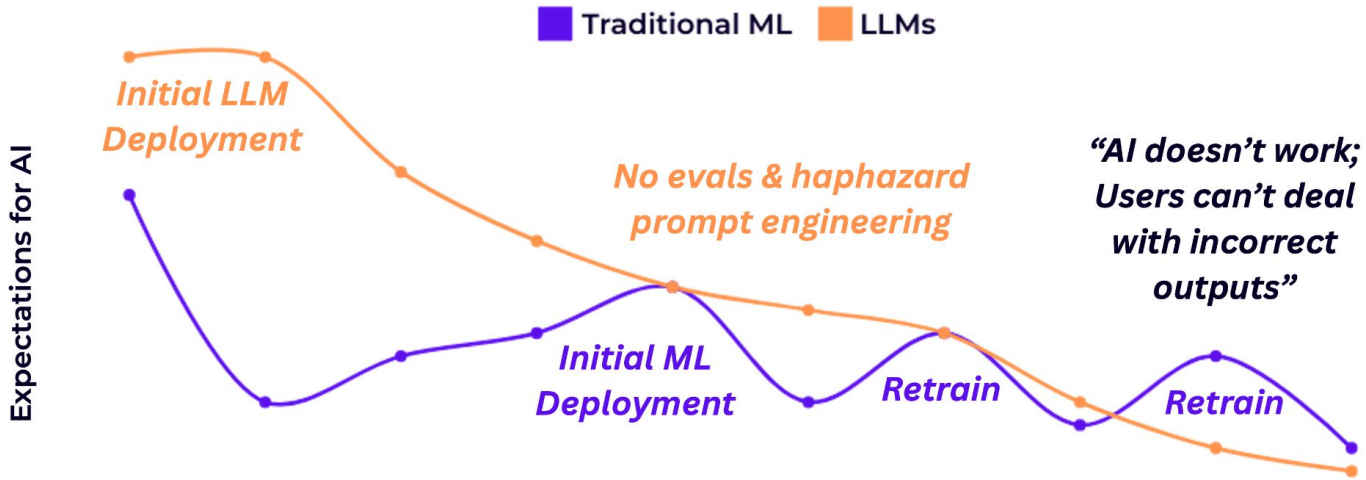
La práctica de MLOps implica abogar por la automatización y la supervisión en todos los pasos de la construcción del sistema de Aprendizaje Automático, incluida la integración, las pruebas, el lanzamiento, la implementación y la administración de la infraestructura.

Ser capaces de explotar nuestros modelos de manera automática adaptándonos a la necesidad de nuestro caso de uso

Que es MLOps?



Expectations While Building (Failed) AI Products



Cuales son los objetivos del MLOps

-**Automatización del despliegue**: Implementación automatizada y reproducible de los modelos en entornos de producción.

-Gestión del ciclo de vida: Gestionar eficientemente el ciclo de vida completo de los modelos.

-Colaboracion y comunicacion: Facilitar la colaboración efectiva entre equipos de desarrollo.

-Monitorización y mantenimiento continuo: Establecer sistemas robustos de monitorización para evaluar el rendimiento de los modelos en tiempo real.

Inferencia

Inferencia

-**Inferencia batch**: Realizan predicciones para un conjunto completo de datos de entrada de una sola vez. Es eficiente cuando se pueden procesar grandes cantidades de datos simultáneamente

- **Inferencia en tiempo real**: realizar predicciones a medida que llegan los nuevos datos, en lugar de esperar y procesarlos por lotes.

Inferencia en Batch

Ventajas: Eficiencia computacional, optimización de recursos y facilidad de implementación.

Desventajas: Tiempo de respuesta, actualizaciones asincrónicas y menos adecuado para trabajar con datos en tiempo real.

Inferencia en tiempo real

Ventajas: Respuestas inmediatas, adaptabilidad dinámica e interactividad

Desventajas: Requiere recursos inmediatos, menor eficiente para grandes volúmenes de datos, una planificación más compleja y un coste más elevado.

Elección del tipo de inferencia

- Eficiencia Computacional
- Tiempo de Respuesta
- Escalabilidad
- Costos Operativos

Hardware para inferencia

Unidades de procesamiento gráfico (GPUs): Las GPUs se utilizan mayoritariamente para ejecutar modelos de Deep Learning. Su uso tiene un coste muy elevado.

Unidades de procesamiento Central (CPUs): Mayor versatilidad que las GPUs pero menos eficiente a la hora de trabajar con modelos de Deep Learning. Pero cumple la mayoría de requisitos al trabajar con modelos kNN, SVM, XGBoost, etc.

Es realmente importante la inferencia?

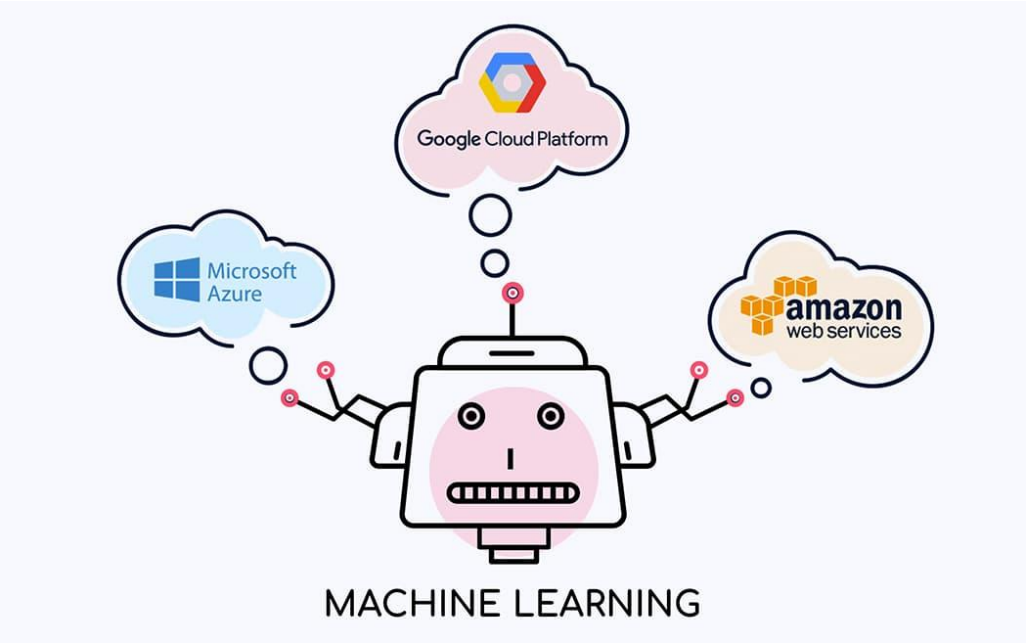
ChatGPT could cost over $700,000 per day to operate.

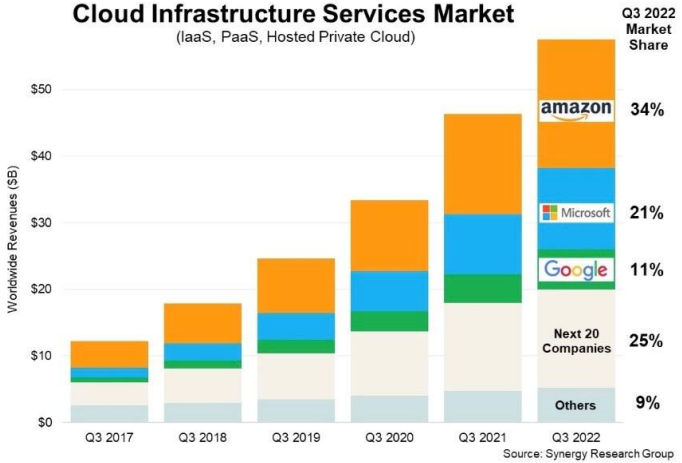
Microsoft is reportedly trying to make it cheaper.

https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4

Proveedores cloud

Proveedores cloud





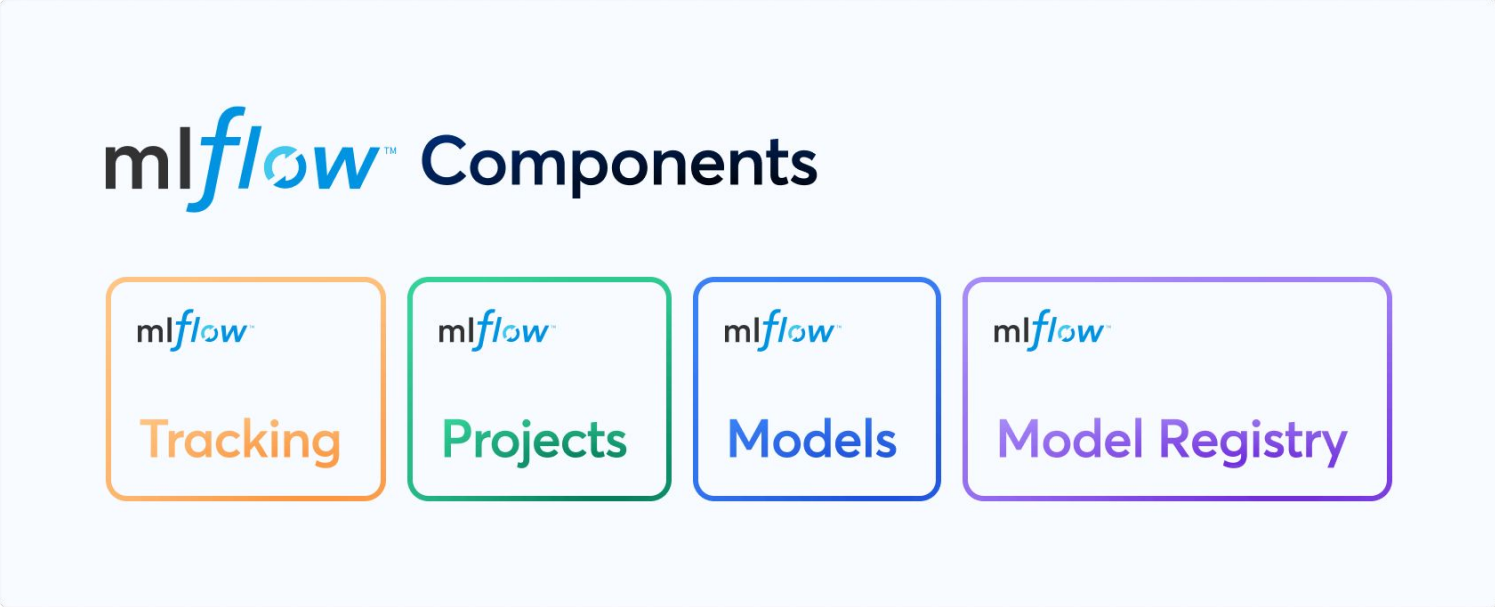
Ventajas y desventajas de usar servicios cloud

Ventajas: Escalabilidad, pago por uso, acceso global, variedad de servicios, mantenimientos y actualizaciones automatizados, facilidad de implementación y facilidad en la recuperación de desastres.

Desventajas: Costos variables, dependencia del proveedor, latencia, limitaciones de personalización y conectividad a internet.

ML Flow

Componentes principales de ML FLow



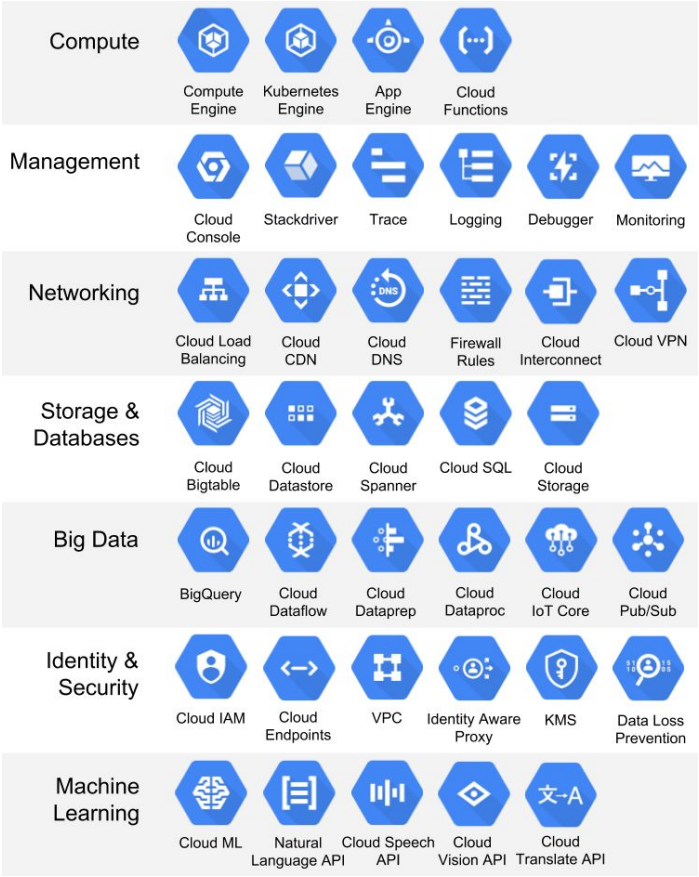
GCP

Google Cloud

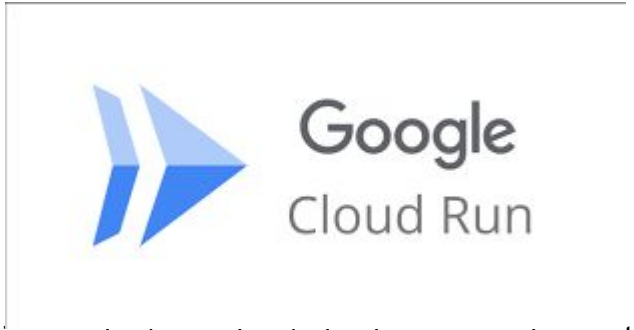
AI Platform

- AI Platform te permite entrenar modelos de aprendizaje automático a gran escala, alojar tu modelo entrenado en la nube y hacer predicciones sobre nuevos datos.
- Construye modelos con datos de cualquier tamaño usando infraestructura de entrenamiento distribuido.
- Integración con Cloud Dataflow, Cloud Storage y Cloud Datalab.

Entorno GCP



GCP Cloud Run



Entorno sin servidor: Permite ejecutar aplicaciones en un entorno sin servidor (serverless), donde no necesitas gestionar la infraestructura.

Escalado automático: Escala automáticamente tus aplicaciones según la demanda, adaptándose al tráfico que reciban.

Despliegue sencillo: Puedes desplegar tus aplicaciones fácilmente a partir del código fuente o de imágenes de contenedor.

Ideal para microservicios y APIs: Es ideal para implementar microservicios, APIs, sitios web y más, proporcionando una forma eficiente y flexible de gestionar aplicaciones.

FastAPI

-FastAPI es un framework de Python para la creación rápida de API.
-Permite integrar nuestros modelos de ML/DL en aplicaciones.

Desplegar nuestros modelos de manera sencilla, rápida y escalable

Llamadas a API. GET, PUT, POST, DELETE.

GET: Recuperamos datos.
POST: Envia datos para crear un nuevo recurso.
PUT: Actualiza los datos existentes en él servidos.
DELETE: Elimina datos del servidor.

KEEPCODING

Tech School

Madrid | Barcelona | Bogotá