

ML en producción

**BENTOML**





## ¿Qué es BentoML?


Framework gratuito y de código abierto para servir y desplegar modelos de machine learning

Permite tomar un modelo entrenado y desplegarlo en infraestructura en la nube.





## ¿Por qué lo usaríamos?

- Es un “puente” entre el desarrollo del modelo y su puesta en producción.
  - Convierte un modelo de ML en una API en unas cuantas líneas de código Python\*.
  - Permite crear contenedores de Docker sin necesidad de tener mucho conocimiento.
  - Soporta todos los frameworks de machine learning populares.
- 



# Conceptos principales

## Modelo

Artefacto de machine learning para realizar inferencias

## Servicio

Permite montar un modelo como una API bajo una arquitectura basada en servicios (SOA)

## Runner

Permiten encapsular la ejecución del modelo de machine learning y permitir que el servicio lo ejecute si es necesario.

## Bento

Un paquete que contiene el servicio a desplegar, permite ser desplegado a la nube.



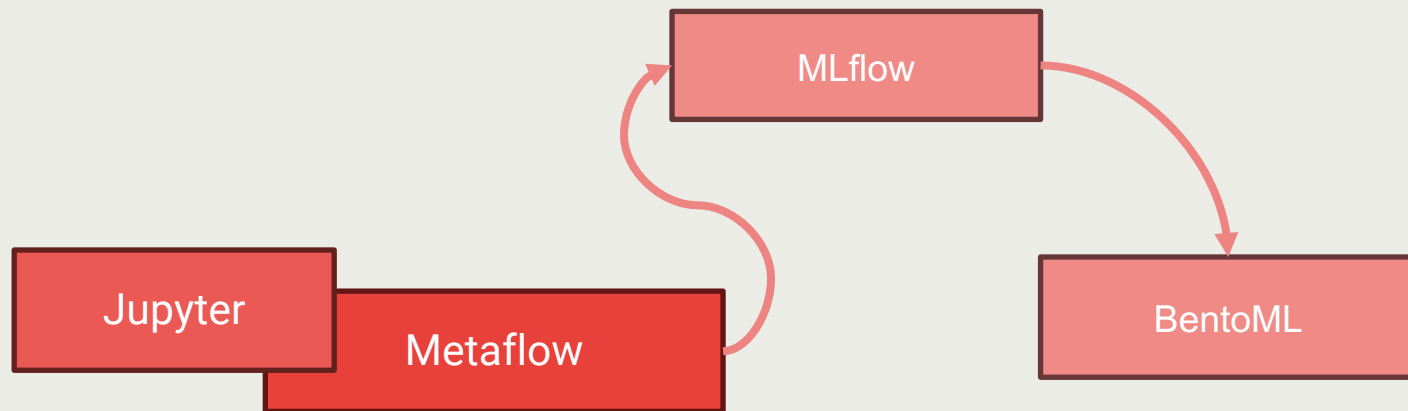
DEMO

[tcsq.dev/production-bento](https://tcsq.dev/production-bento)

Backup: [github.com/fferegrino/cf-mlops/tree/production-bento](https://github.com/fferegrino/cf-mlops/tree/production-bento)



# Nuestra “plataforma” de ML





# Preguntas y respuestas

Me puedes contactar después

