

Deployment



Introduction

1- Deploying machine learning models is crucial for putting them into real-world applications.

In this presentation, we'll explore the deployment process using three popular Python frameworks: FastAPI, Flask, and Django.



What is API?

- **Definition:** A set of rules and protocols facilitating communication between software applications.

- Purpose:

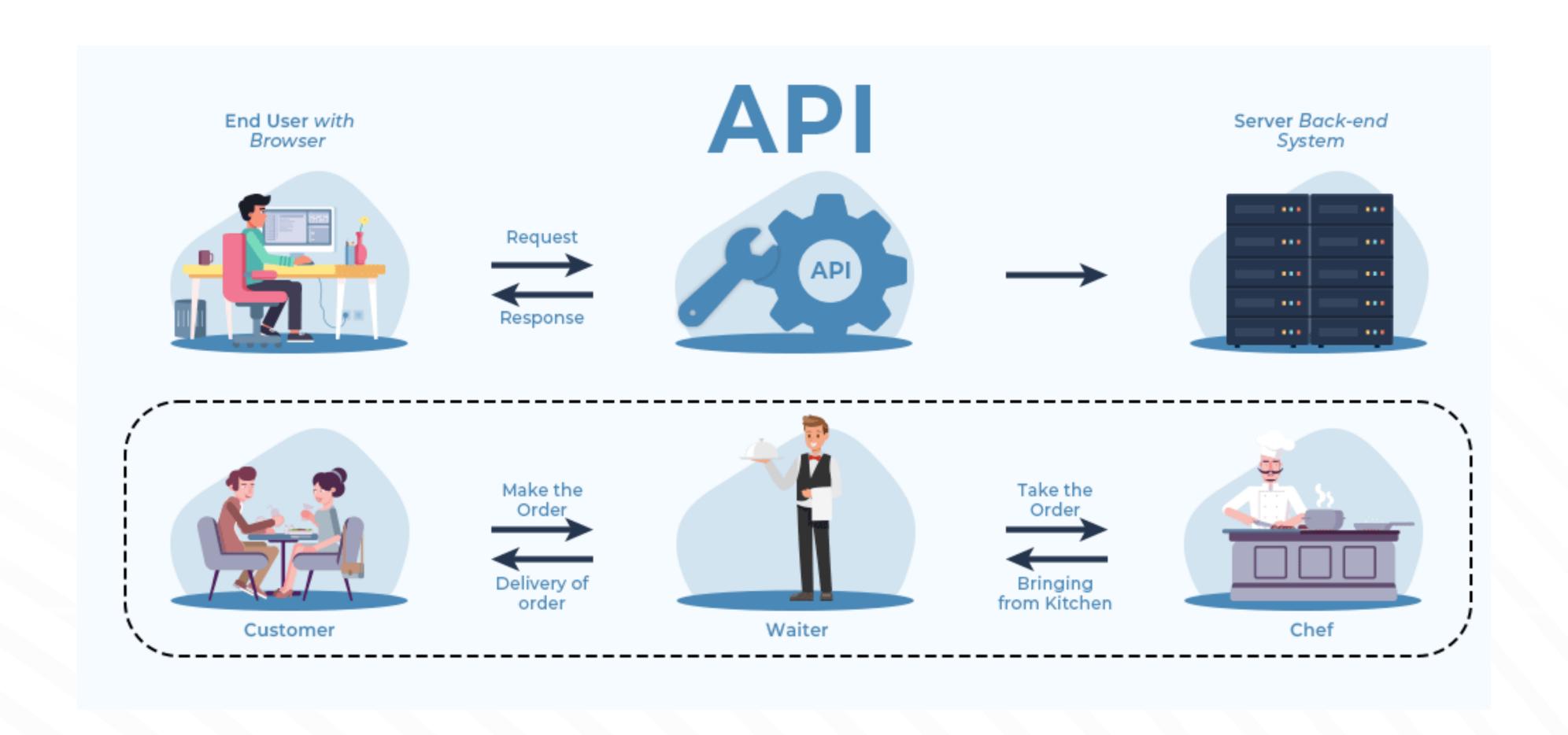
- Enables seamless interaction between different systems.
- Allows developers to access the functionality of a service without understanding its internal workings.

- Types:

- Web APIs: Use HTTP for communication (e.g., RESTful APIs).
- Library-based APIs: Provide functions for specific tasks.
- Operating System APIs: Enable interaction with the underlying OS.



What is API?



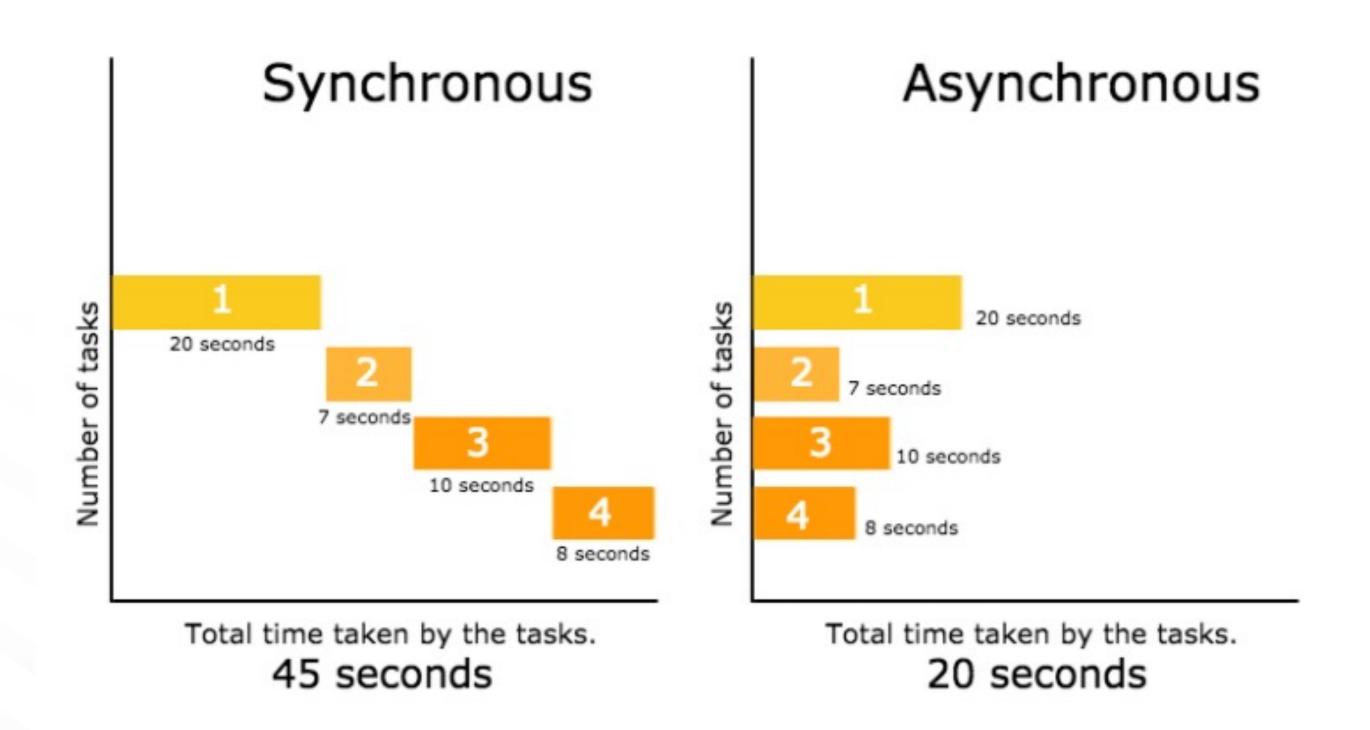


Which Better? FastAPI, Flask, Django

Aspect	FastAPI	Flask	Django
Framework	Asynchronous	Asynchronous	Asynchronous
Preformance	Fast	Lightweight	Moderate
Documentation	Automatic OpenAPI	Well-documented	Comprehensive
Flexibility	High	High	Moderate
Ease of Use	Easy	Easy	Moderate
Community	Growing	Large	Very Large



Synchronous Vs Asynchronous





What would we choose?

G FastAPI



Required Libraries

- FastAPI
- Uvicorn





Required Environment

- Model file (e.g., .h5 for Keras, .pkl for scikit-learn ..etc),
 the file contains the weights and parameters of your
 trained model.
- Preprocessing Script (If Applied)



Publishing on a Server

- Deploying your machine learning model onto a server is crucial for making it accessible over the internet.
- This section covers the process of publishing your model, enabling users or applications to interact with it remotely.



Hosting Services

- Alibaba cloud (SCCC)
- Amazon Web Services (AWS)
- Google Cloud
- Azure
- Heroku
- Render
- Digital Ocean



Resources

- How to Deploy FastAPI on AWS EC2: Quick and Easy Steps!
- Deploy Fast API to Azure App Service under 4min [Easy]



Thank You