DagsHub

Machine Learning in Production

What does ”Production” even mean?

Let’s talk about #Buzzwords

Hyperparameter Optimization

CI/CD/CT

Feature Stores

MLOps

AGI

Data-Centric AI

Active Learning

AutoML

Large Language Models

Meta Learning

AI

Metadata Store

Big Data

Let’s talk about #Buzzwords

Hyperparameter Optimization

AGI

CI/CD/CT

Feature Stores

MLOps

PRODUCTION

AutoML

Active Learning

AI

Meta Learning

Data-Centric AI

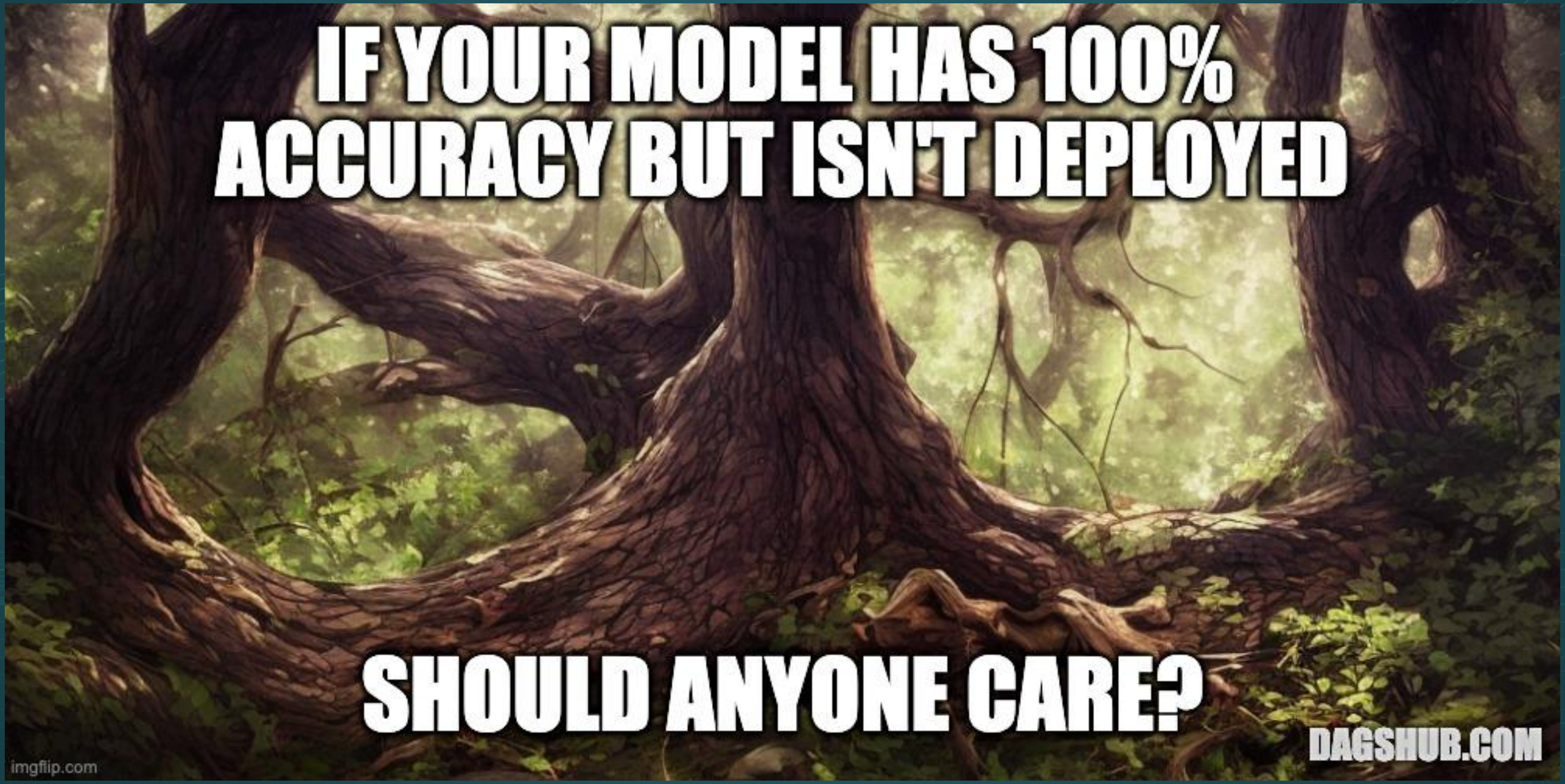
Big Data

Metadata Store

Let’s talk about #Buzzwords

PRODUCTION

Defining ML in production



About Me

Dean Pleban

Follow me:

@DeanPlbn

/DeanPleban

DagsHub

About Me

Dean Pleban

Building tools for ML teamwork

Follow me:

@DeanPlbn

/DeanPleban

DagsHub

About Me

Dean Pleban

Building tools for ML teamwork

Strongly believe in open source

Follow me:

@DeanPlbn

/DeanPleban

DagsHub

About Me

Dean Pleban

Building tools for ML teamwork

Strongly believe in open source

DagsHub

Follow me:

🐦 @DeanPlbn

in /DeanPleban

About Me

Dean Pleban

Building tools for ML teamwork

Strongly believe in open source

DagsHub

The MLOps Podcast

Follow me:

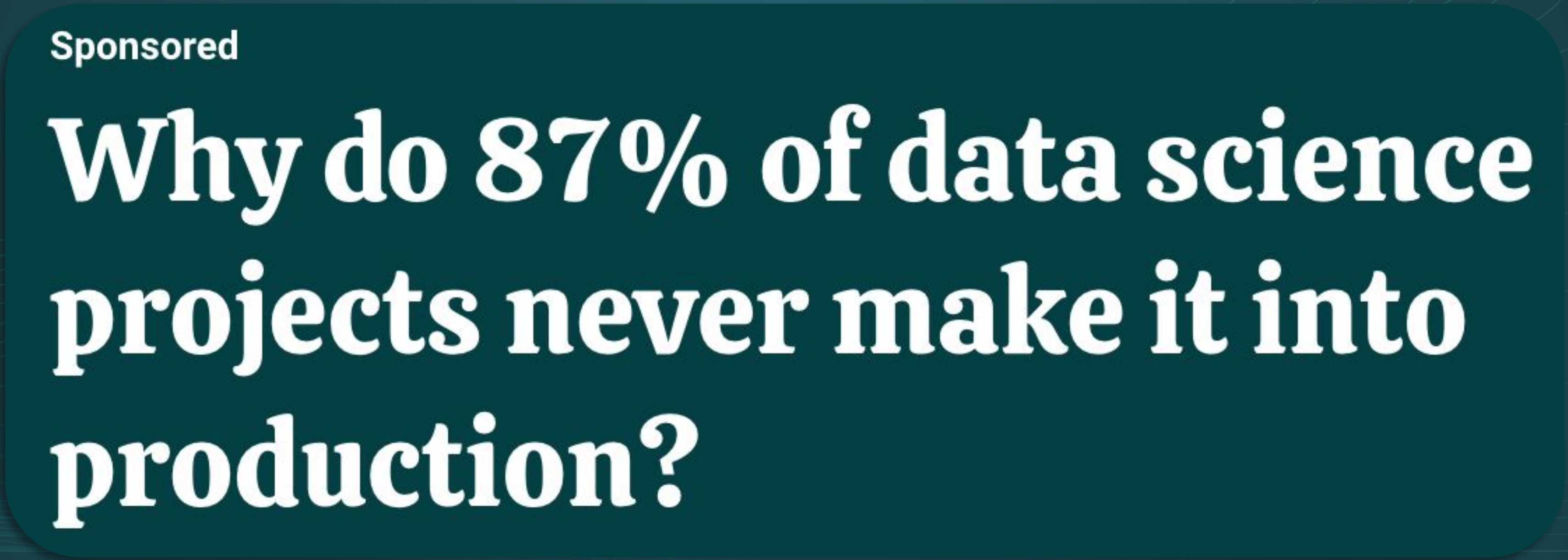
@DeanPlbn

/DeanPleban

Defining ML in production

DagsHub

Remember “that” statistic…



Sponsored

Why do 87% of data science projects never make it into production?

Why should you care?

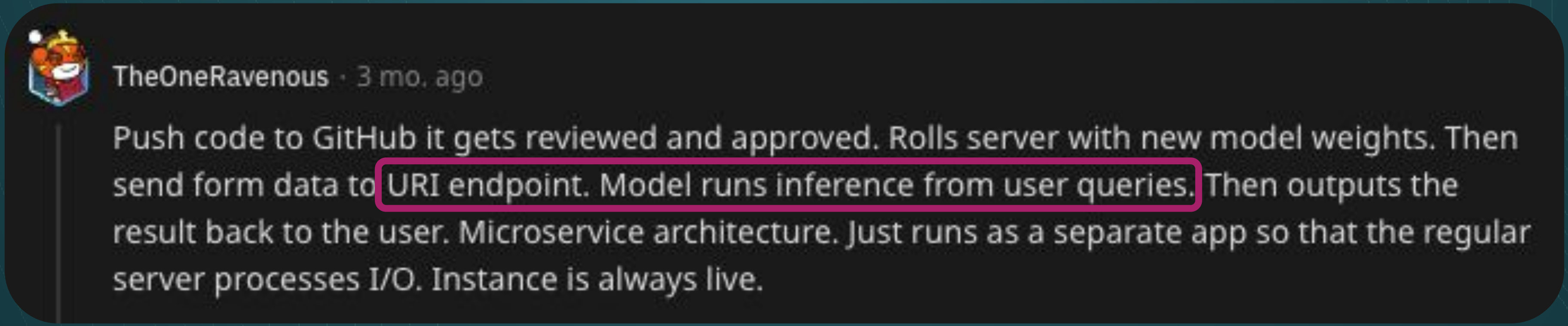
Understand what production means

Framework to think about buzzwords

First-principles guide to deployment

What types of production are there?

Endpoint



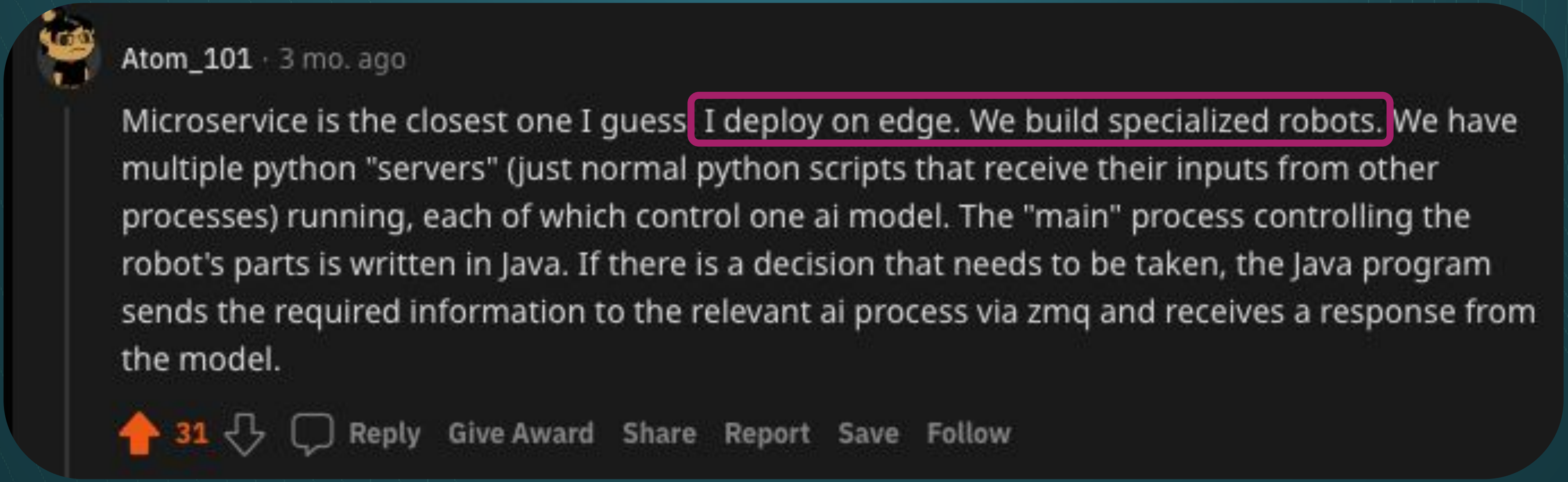
TheOneRavenous · 3 mo. ago

Push code to GitHub it gets reviewed and approved. Rolls server with new model weights. Then send form data to URI endpoint. Model runs inference from user queries. Then outputs the result back to the user. Microservice architecture. Just runs as a separate app so that the regular server processes I/O. Instance is always live.

What types of production are there?

Endpoint

Edge Device



Atom_101 · 3 mo. ago

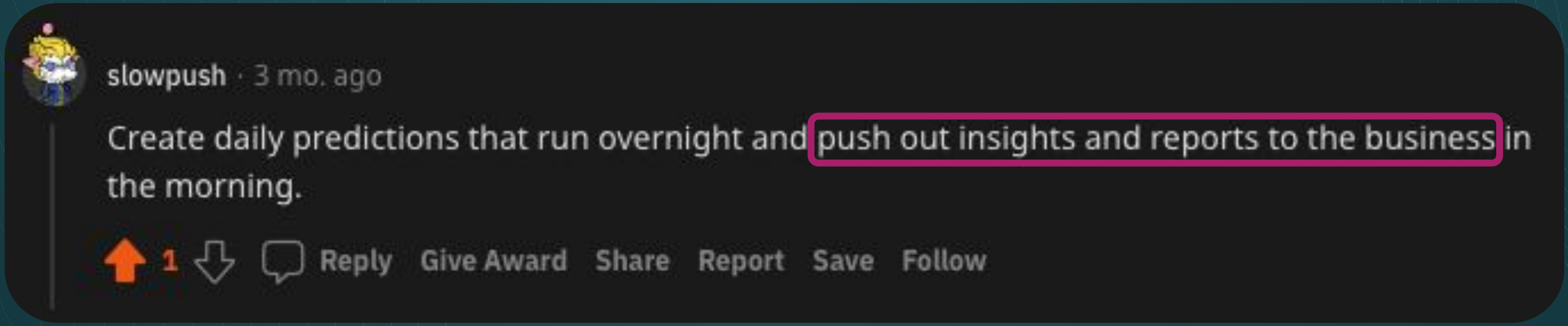
Microservice is the closest one I guess. I deploy on edge. We build specialized robots. We have multiple python "servers" (just normal python scripts that receive their inputs from other processes) running, each of which control one ai model. The "main" process controlling the robot's parts is written in Java. If there is a decision that needs to be taken, the Java program sends the required information to the relevant ai process via zmq and receives a response from the model.

What types of production are there?

Endpoint

Edge Device

Dashboard



slowpush · 3 mo. ago

Create daily predictions that run overnight and push out insights and reports to the business in the morning.

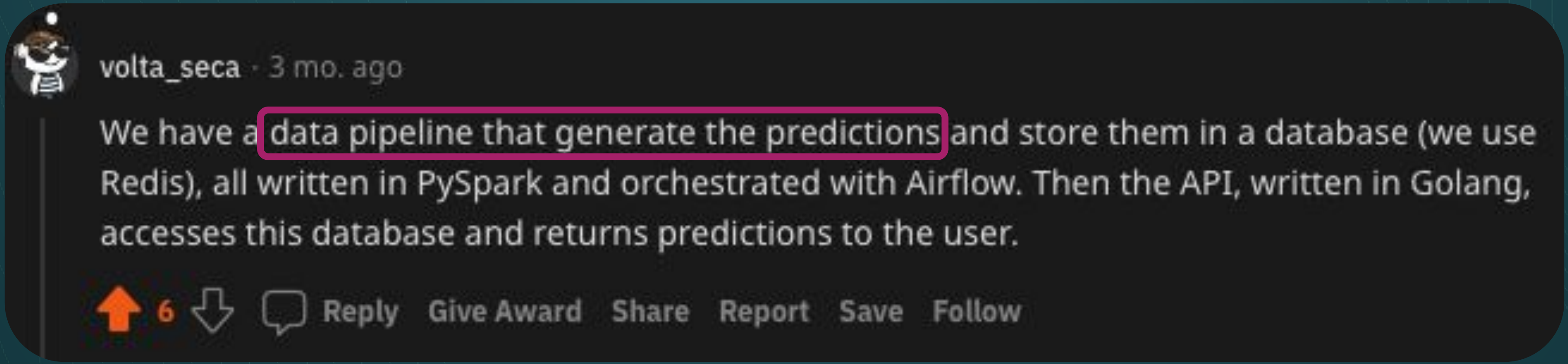
What types of production are there?

Endpoint

Edge Device

Dashboard

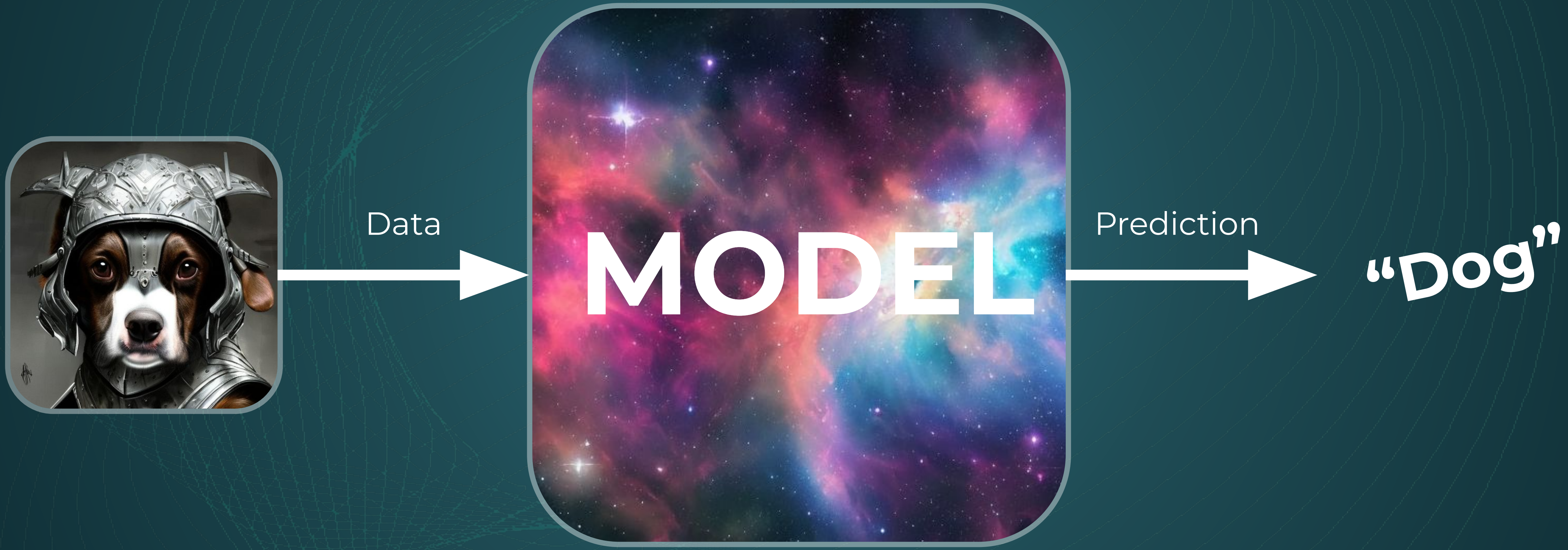
ETL / Pipeline



volta_seca · 3 mo. ago

We have a data pipeline that generate the predictions and store them in a database (we use Redis), all written in PySpark and orchestrated with Airflow. Then the API, written in Golang, accesses this database and returns predictions to the user.

Models vs. Pipelines



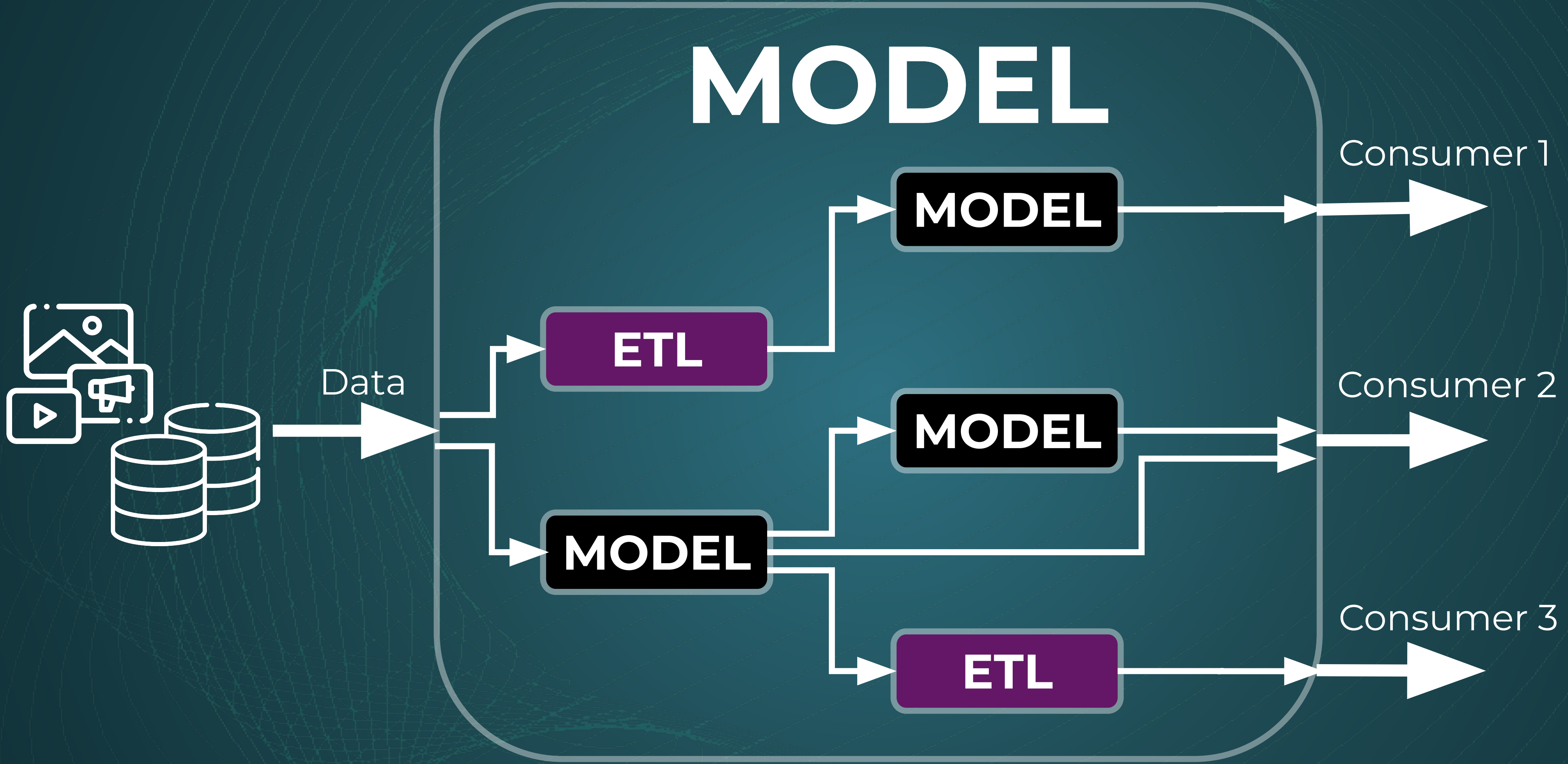
Data

MODEL

Prediction

"Dog"

Models vs. Pipelines



DagsHub

Models vs. Pipelines – Who cares?

Models vs. Pipelines – Who cares?

End-to-end thinking

Models vs. Pipelines – Who cares?

End-to-end thinking

Better understand your requirements

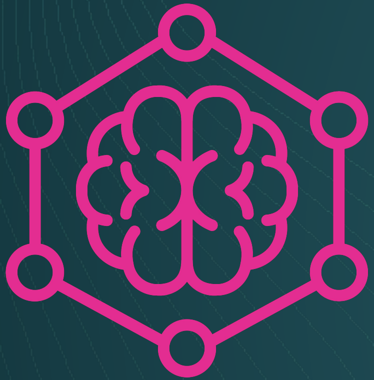
First-principles thinking

A woman thinking from **first-principles**, lost in thought, a painting by Van Gogh (By Stable Diffusion)

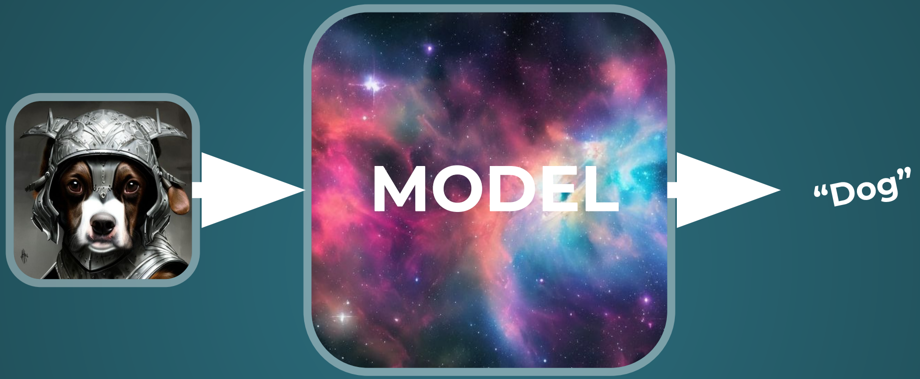
What we assume

Deploying a
single model

What we assume



Deploying a single model

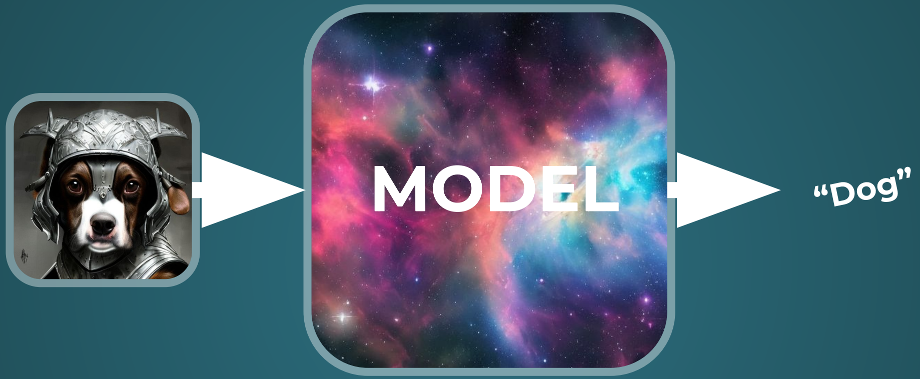


A simple flow (sort of)

DagsHub

What we assume

Deploying a single model



A simple flow (sort of)

The model is trained

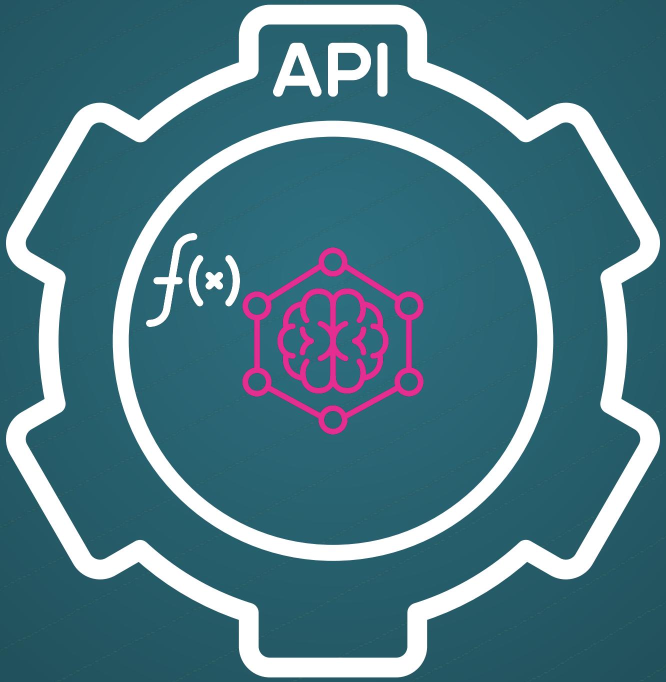
Breaking deployment down – **The hard part**

Breaking deployment down – **The hard part**

Breaking deployment down – **The hard part**

“A command line, someone is typing a command”

Breaking deployment down – **The hard part**



Breaking deployment down – **The hard part**

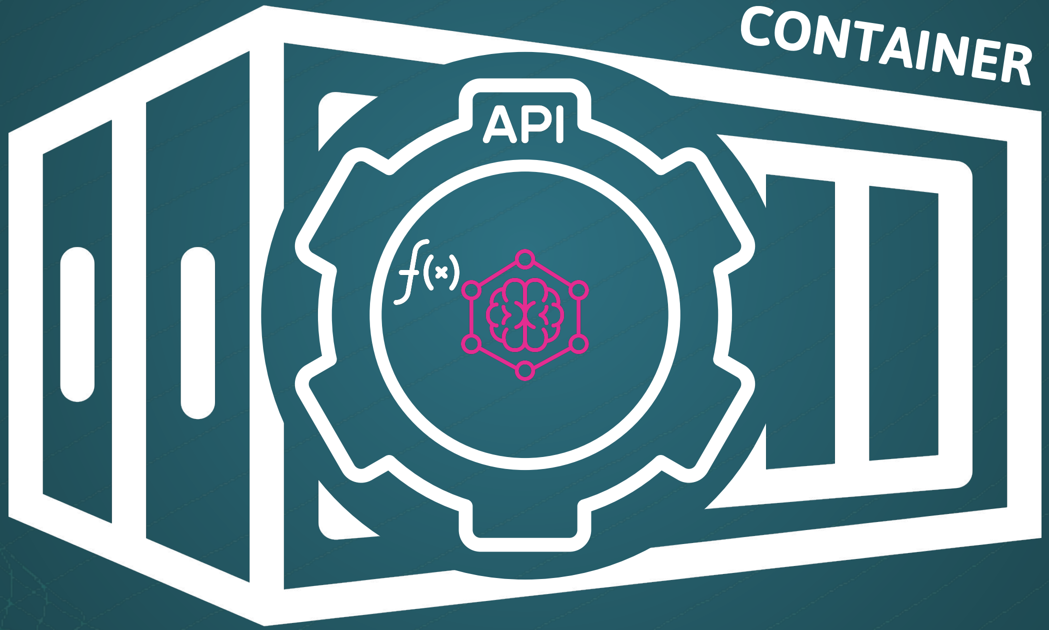
“A website UI on a computer”

Breaking deploym

DAGSHUB.COM
IT WORKS ON MY MACHINE
THEN WE'LL SHIP YOUR MACHINE
AND THAT IS HOW DOCKER WAS BORN

DagsHub

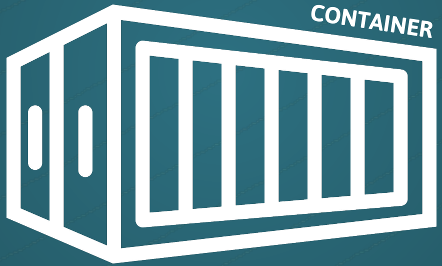
Breaking deployment down – **The hard part**



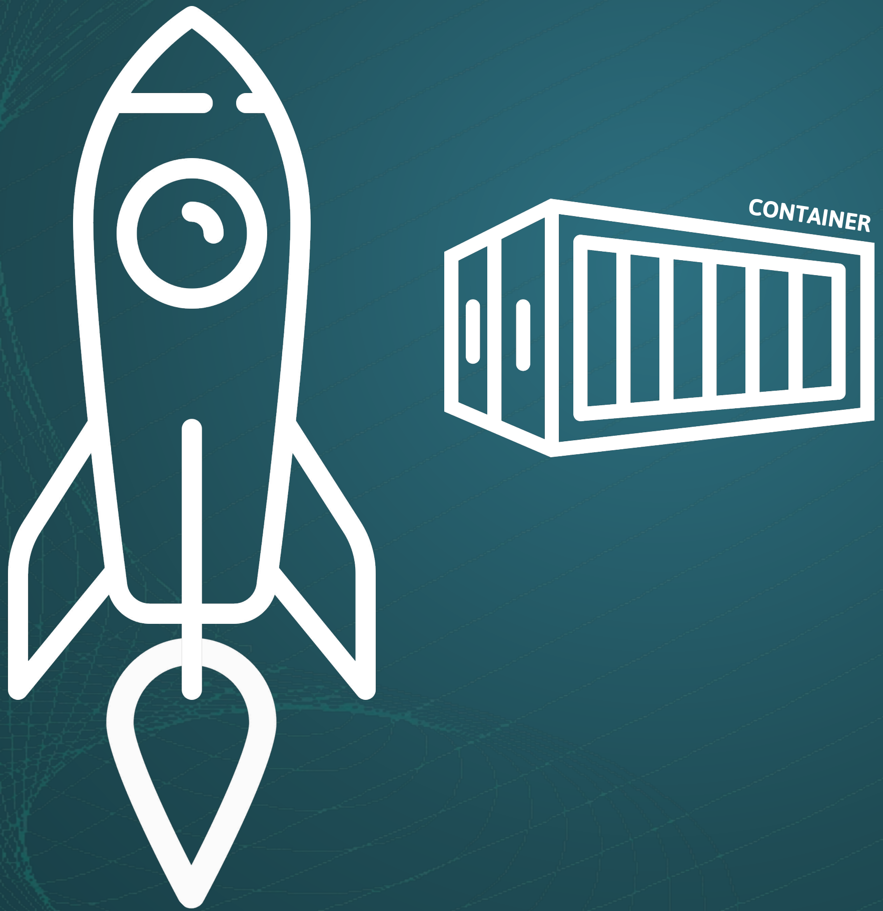
DagsHub

Breaking deployment down – **The hard part**

“Two computers running the same program”



Breaking deployment down – **The hard part**



Breaking deployment down – **The hard part**

Breaking deployment down – **The hard part**

"Many computers connected to the cloud"

Breaking deployment down

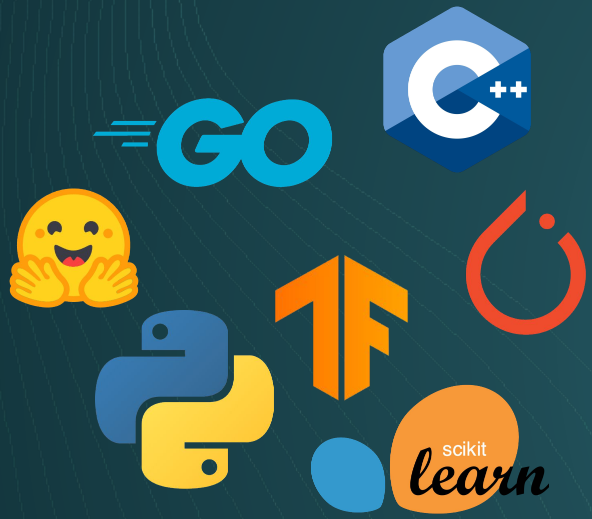
1. Wrap the model in a prediction function

2. Wrap the function in an API

3. Put everything in a suitable environment

4. Provision infrastructure to host the environment

1. The prediction function



Recommended
Tools

1. The prediction function

Recommended
Tools

ONNX

Model
formats

1. The prediction function

GO
C++
scikit
learn

Recommended Tools

ONNX

Model formats

Define a class or interface

2. The API wrapper

FastAPI

Flask
web development,
one drop at a time

Requests
http for humans

Recommended
Tools

2. The API wrapper

FastAPI

Flask
web development,
one drop at a time

Requests
http for humans

Recommended Tools

Define the right endpoints

2. The API wrapper

FastAPI

Flask
web development,
one drop at a time

Requests
http for humans

Recommended Tools

Define the right endpoints

Authentication

3. The environment container

Recommended
Tools

3. The environment container

Recommended Tools

Steps 2+3 in one

4. The infrastructure

Recommended
Tools

4. The infrastructure

Necessary

Recommended

Tools

4. The infrastructure

Necessary

Recommended

Tools

GPUs

Further Reading

1. Building an API for ML models: https://towardsdatascience.com/step-by-step-approach-to-build-your-machine-learning-api-using-fast-api-21bd32f2bbdb

2. Authentication with FastAPI: https://fastapi.tiangolo.com/tutorial/security/

3. Docker for data science: https://dagshub.com/blog/setting-up-data-science-workspace-with-docker/

4. Deploy GPU Accelerated Applications with ECS and Docker: https://www.docker.com/blog/deploy-gpu-accelerated-applications-on-amazon-ecs-with-docker-compose/

Thank You!

Follow me:

🐦 @DeanPlbn

in /DeanPleban

Follow DagsHub:

🐦 @TheRealDAGsHub

in /company/dagshub/