Deploying Machine Learning Models to Flask

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Serializing trained model parameters to disk

Serializing scikit-learn models using JSON, pickle and joblib

Including pre-processing steps as a part of model serialization

Building a text pre-processing and classification pipeline in scikit-learn

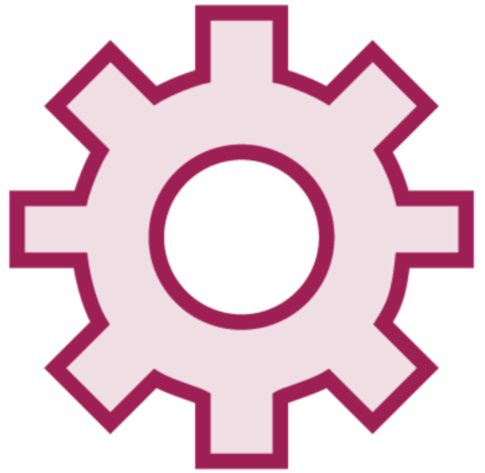
Deploying the classification model to a Flask web application

Model Serialization and Deployment

Serializing scikit-learn Models

JSON

Pickle



Joblib

JSON

Model parameters are accessible and intuitive

JSON

Fragile, requires knowledge of what parameters to serialize for each model

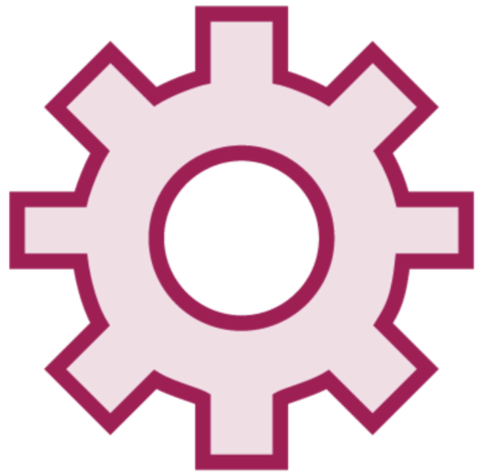
Pickle

Standard Python library for serializing and deserializing Python objects

Pickle

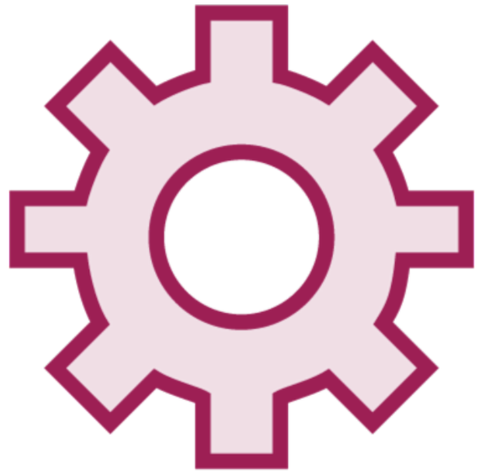
Can work with any Python object not just with model estimator objects

Joblib



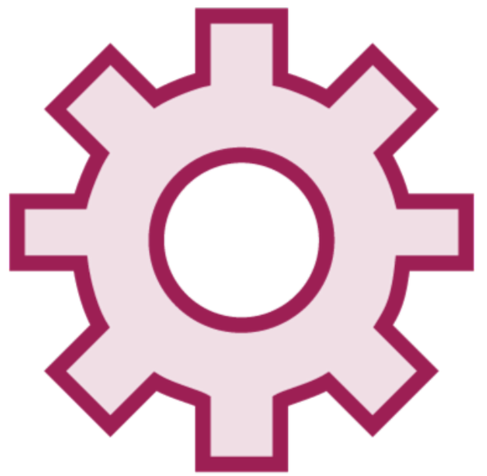
Set of tools to run Python functions as pipeline jobs to speed up long running applications

Joblib



Contains serialization and deserialization utilities for Python objects

Joblib



Is more efficient at storing large multi-dimensional NumPy arrays so works well for models

Mitigating Training-serving Skew

Training and prediction data should be processed using the same pipeline

Serialize data transformation operations along with the model parameters

Easy to do with scikit-learn using the Pipeline object

scikit-learn Pipeline

Estimator object that sequentially applies several transforms. Pipeline can be evaluated and tuned as a whole.

Demo

Serializing model parameters to a JSON file

Demo

Serializing models to disk using pickle and joblib

Demo

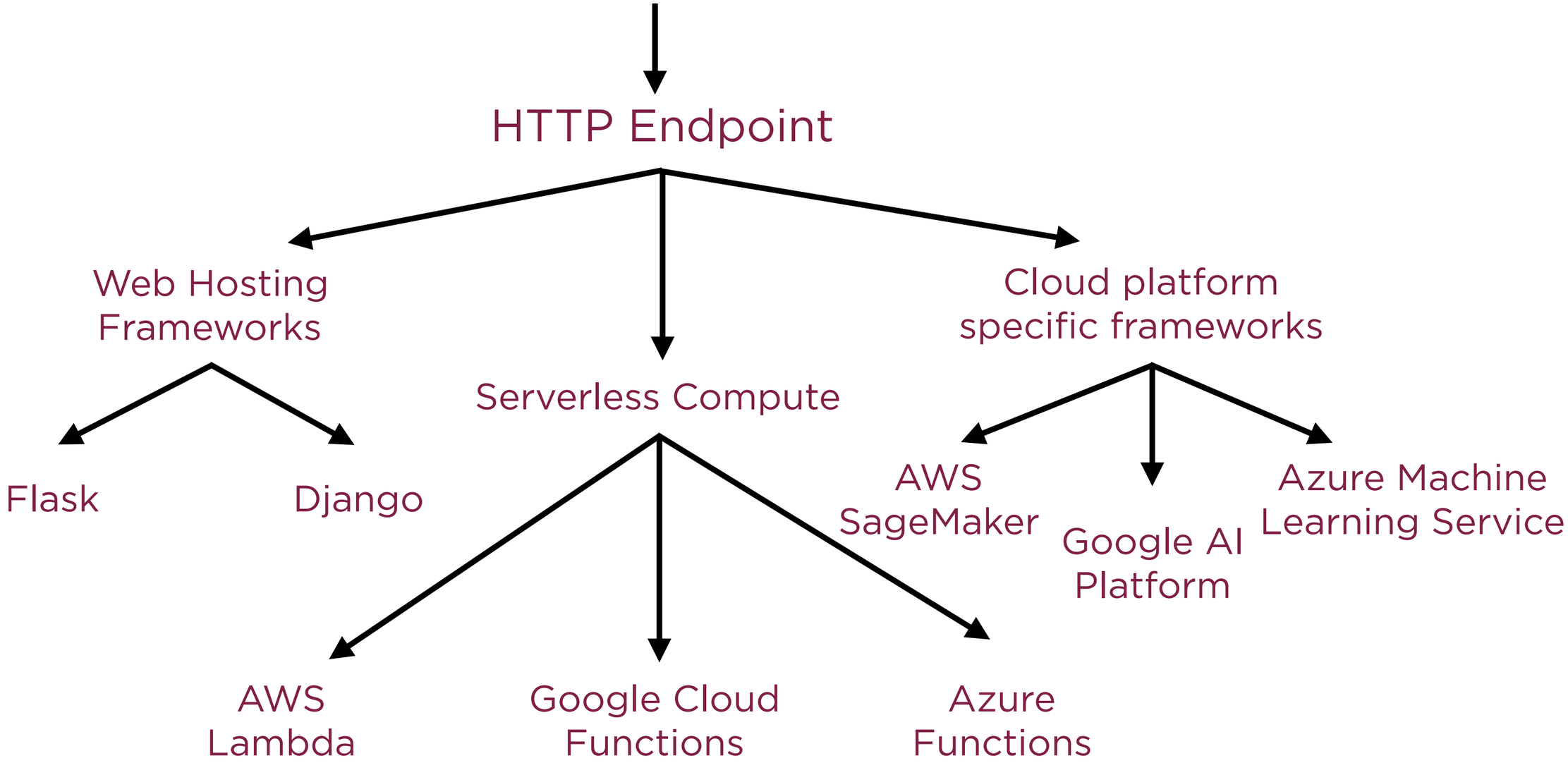
Saving and loading checkpoints for a model

Demo

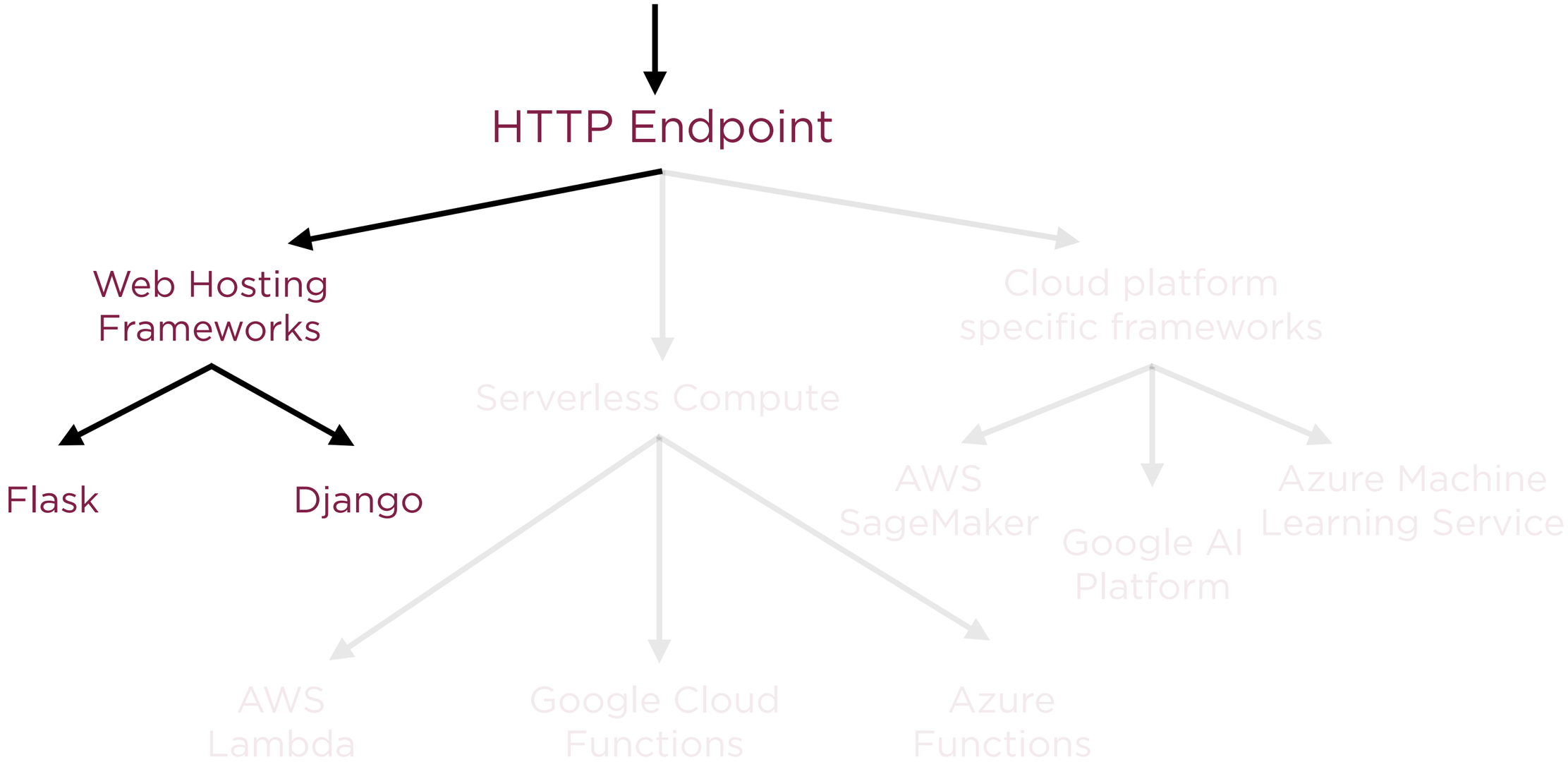
Serializing preprocessing objects and model parameters

Flask for Model Deployment

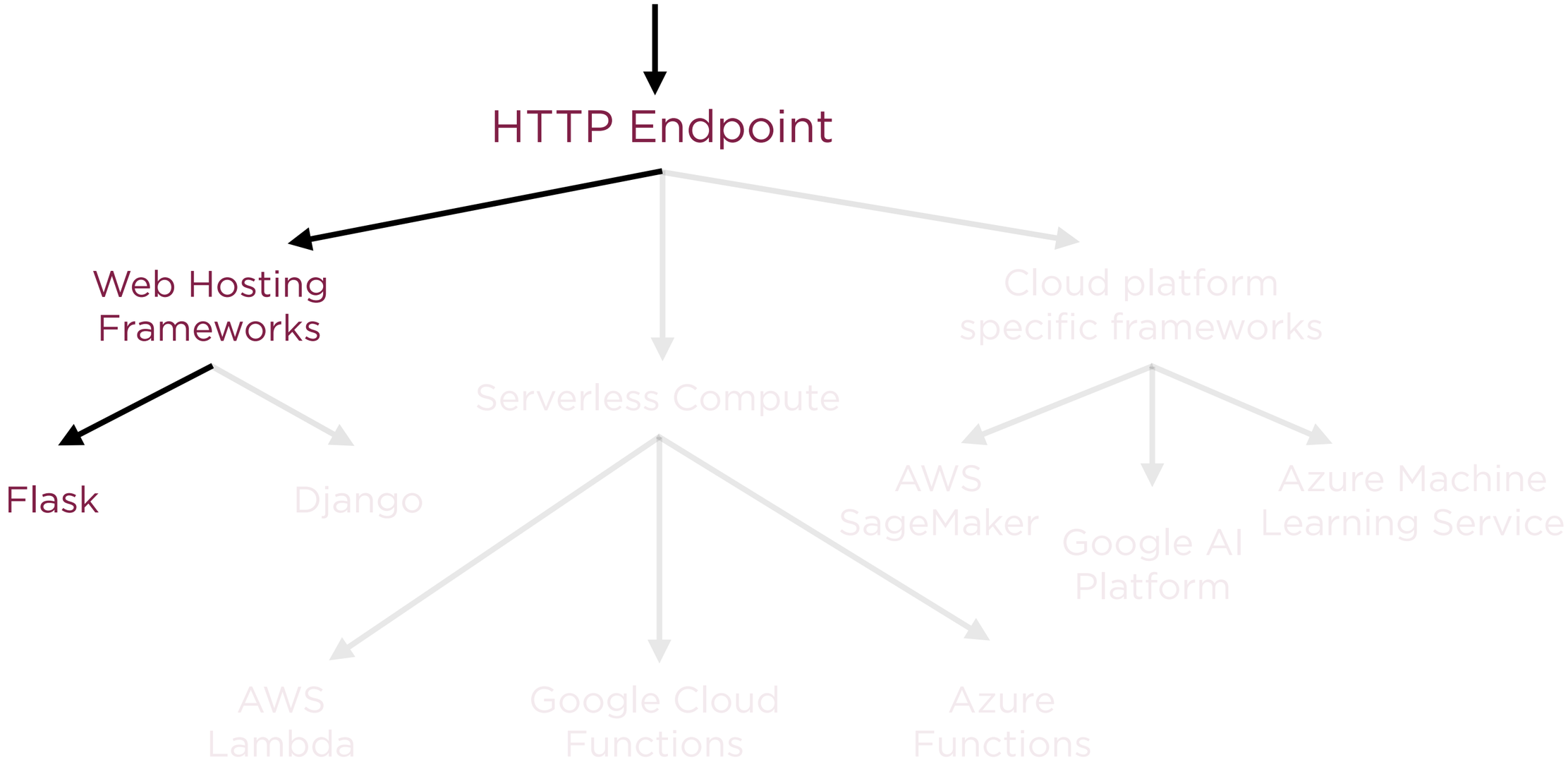
Deploying Models for Prediction



Deploying Models for Prediction

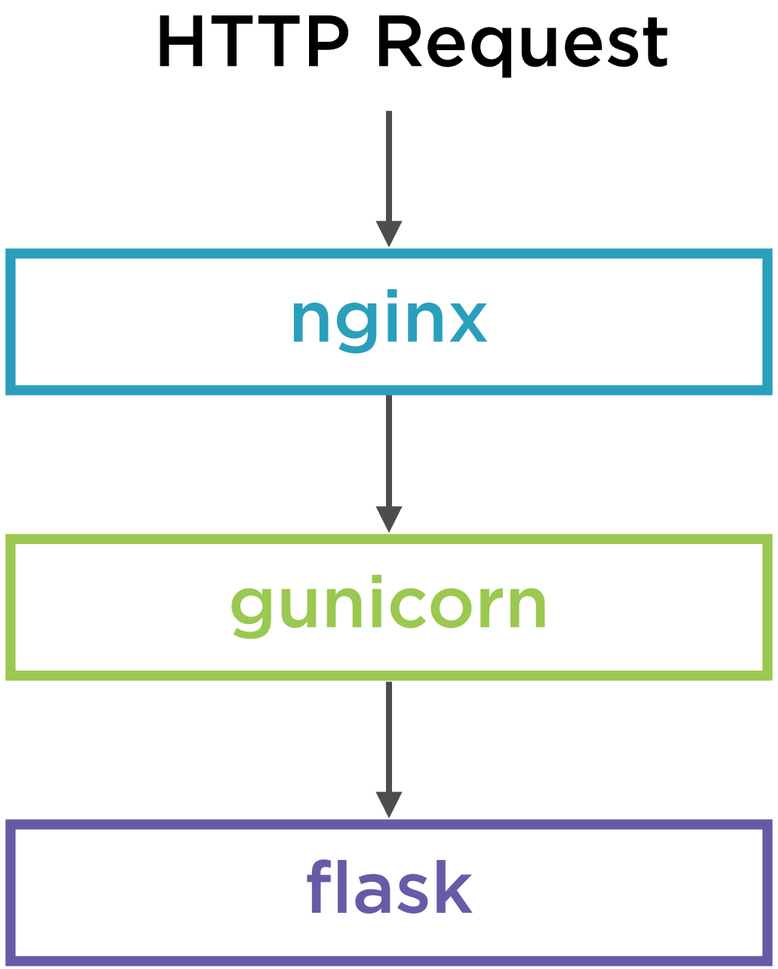


Deploying Models for Prediction



Flask: Lightweight web framework for making models available as HTTP endpoints

Hosting



nginx

Open source software for web serving, reverse proxying, caching, load balancing

Reverse proxy:

- Sits behind a firewall and directs requests to the appropriate backend
- Additional level of abstraction between client and server

gunicorn

Web server for Unix

WSGI HTTP Server:

- WSGI (Web Server Gateway Interface) is a Python standard which determines how a web server communicates with applications
- Simple, lightweight, fast and works with many web frameworks

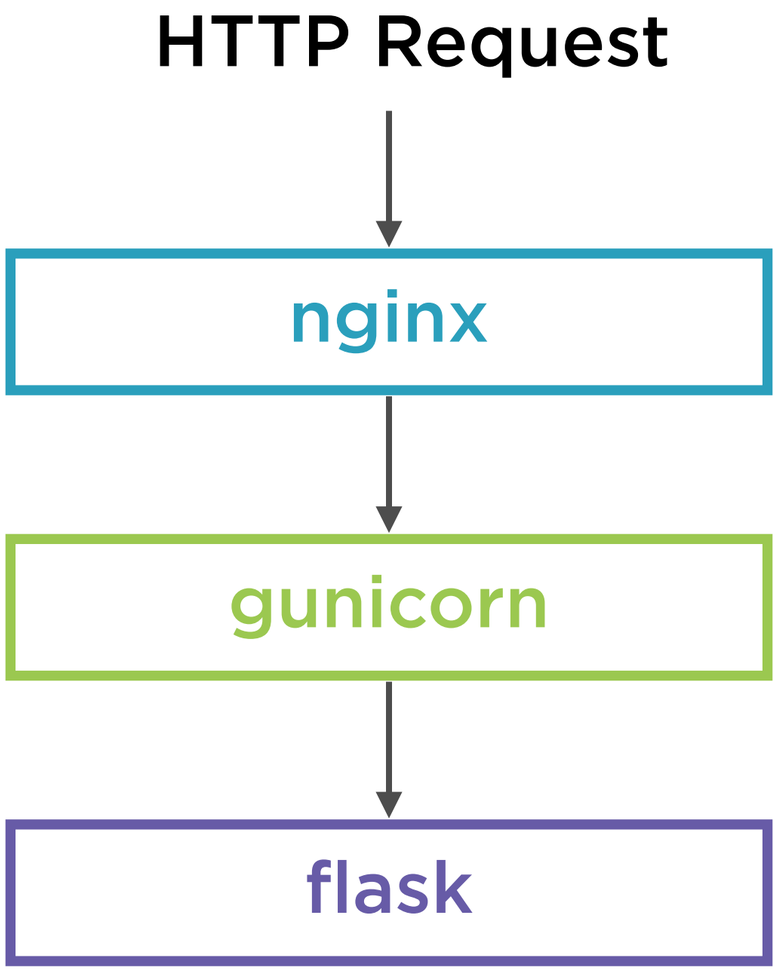
flask

Microframework for Python web app development

Worker:

- The actual instance of the application which hosts the inference code
- Loads the trained model and returns prediction results

Hosting



Demo

Deploying a scikit-learn pipeline to Flask

Summary

- Serializing trained model parameters to disk
- Serializing scikit-learn models using JSON, pickle and joblib
- Including pre-processing steps as a part of model serialization
- Building a text pre-processing and classification pipeline in scikit-learn
- Deploying the classification model to a Flask web application