Deploying Machine Learning Models to Google AI Platform

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Google AI Platform as re-branded umbrella service

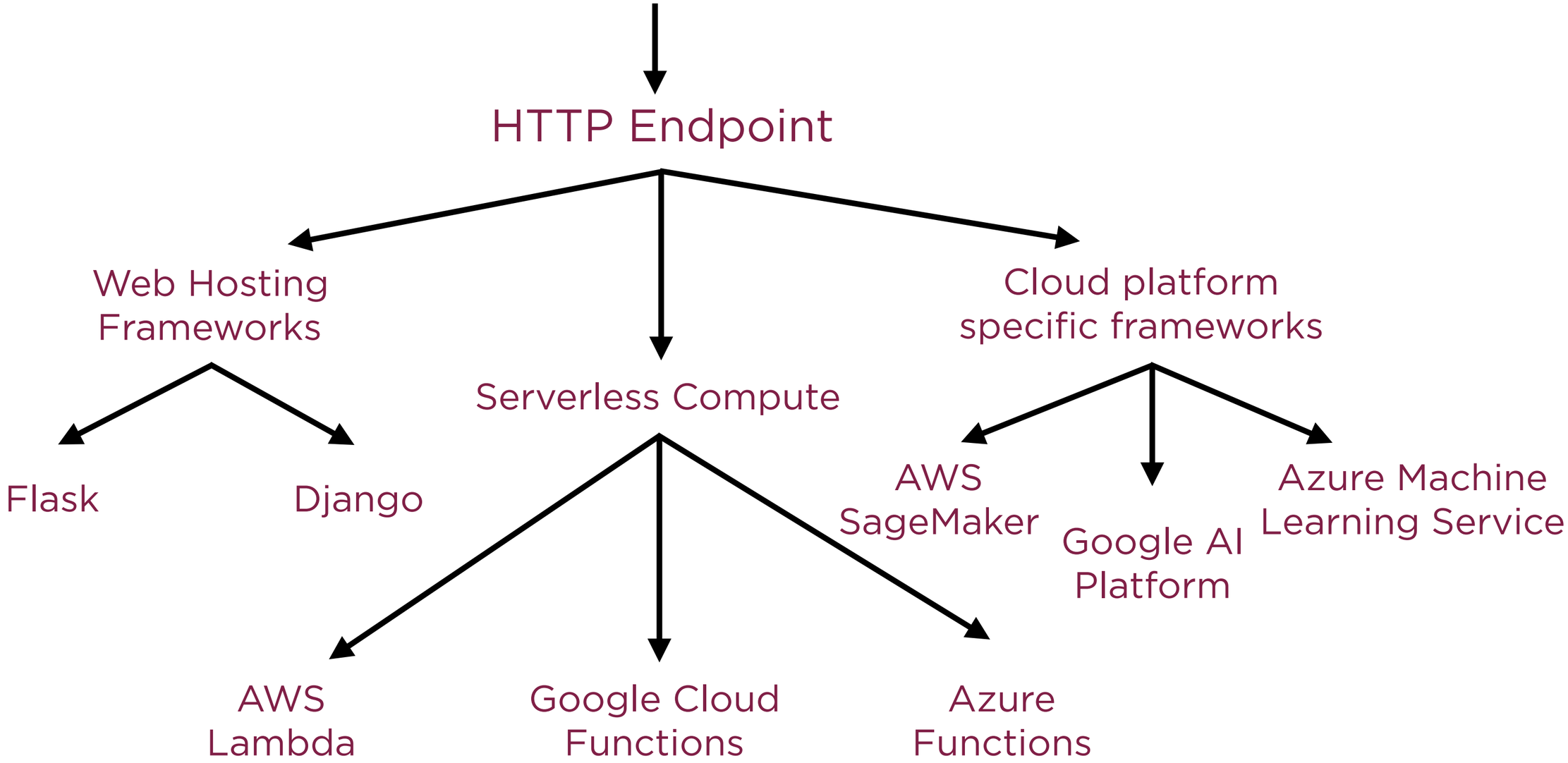
Training, prediction, data labeling service

Supported frameworks, models, model versions

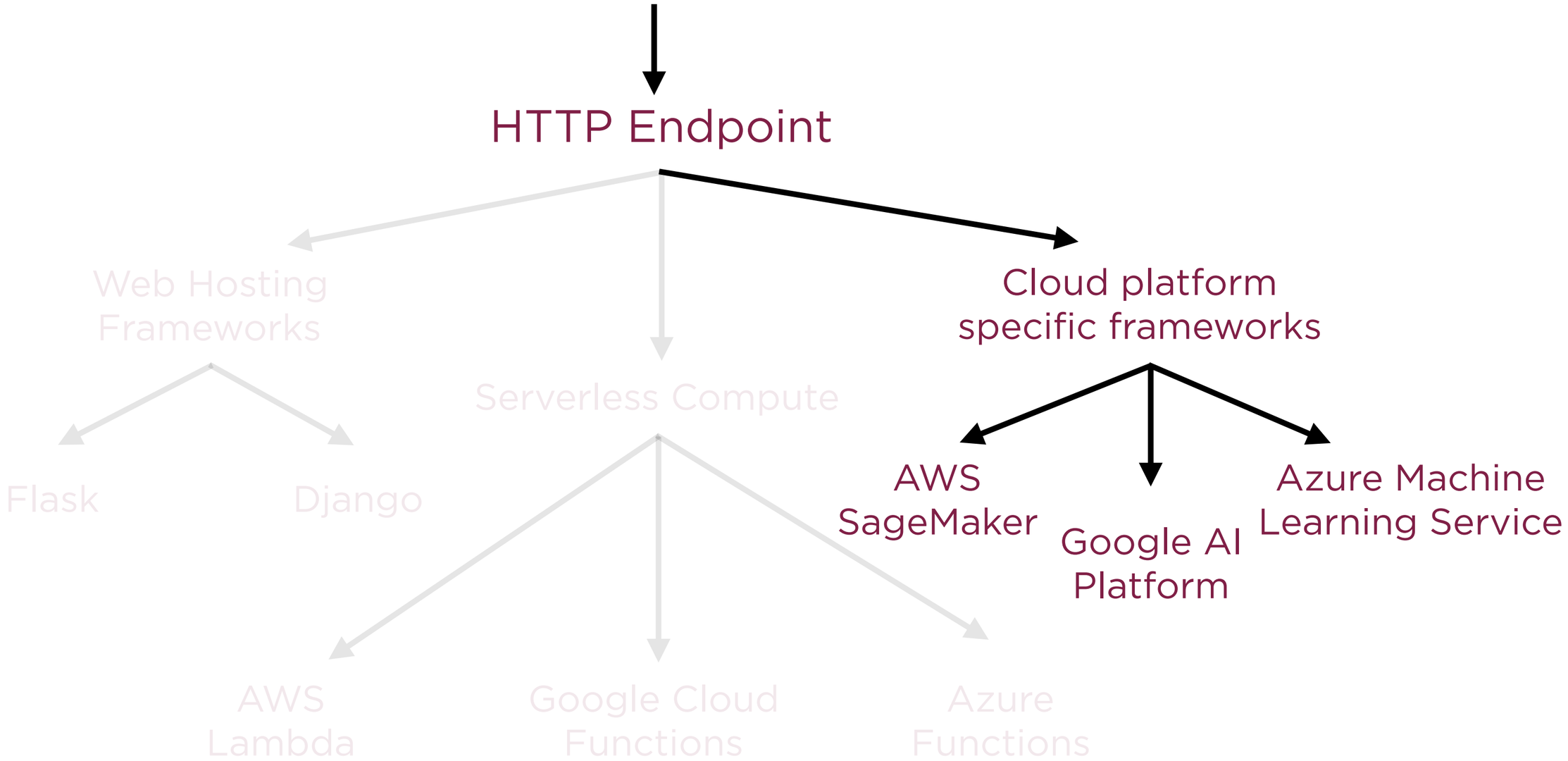
Deploying models using the web console and command line

Monitoring the model using Stackdriver

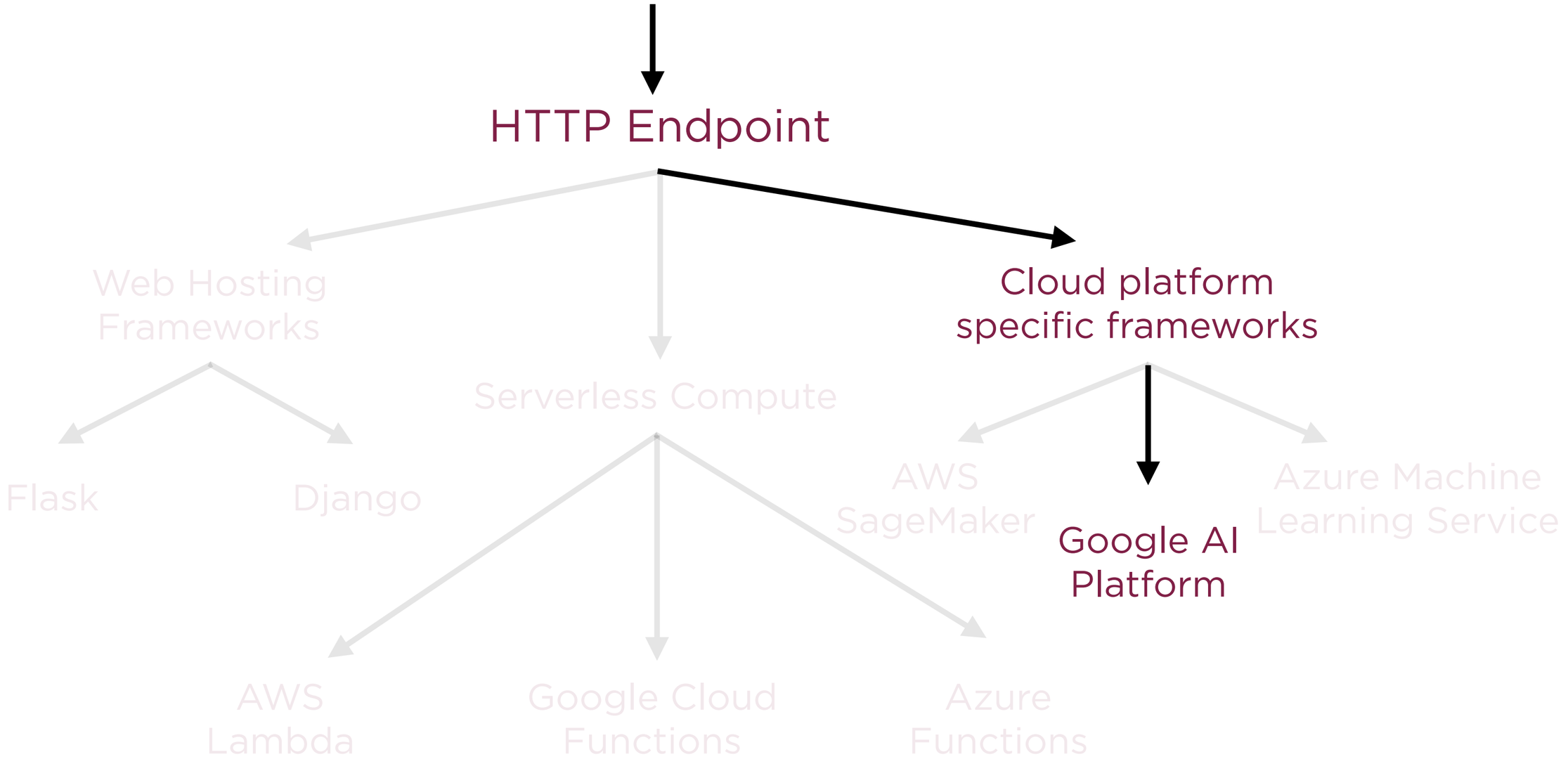
Deploying Models for Prediction



Deploying Models for Prediction



Deploying Models for Prediction



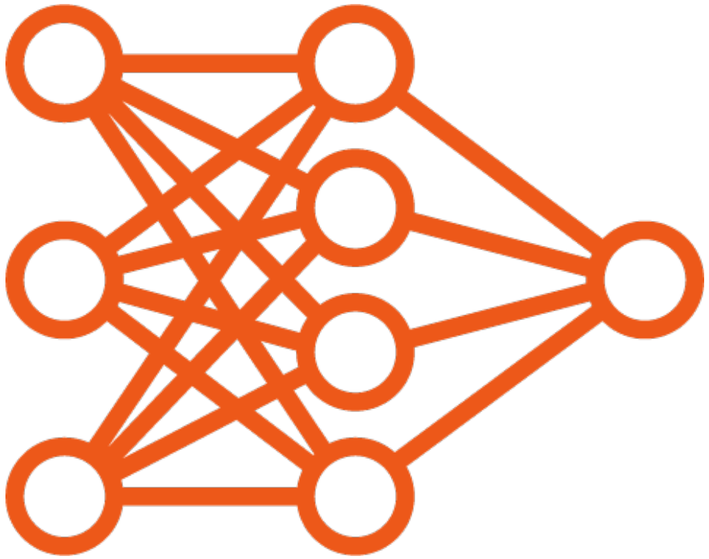
Cloud ML Engine

A managed service for building and deploying ML models on the GCP.

Google AI Platform

A new, re-branded umbrella service on the GCP that includes ML Engine as well as some other ancillary services.

Google AI Platform



ML Engine

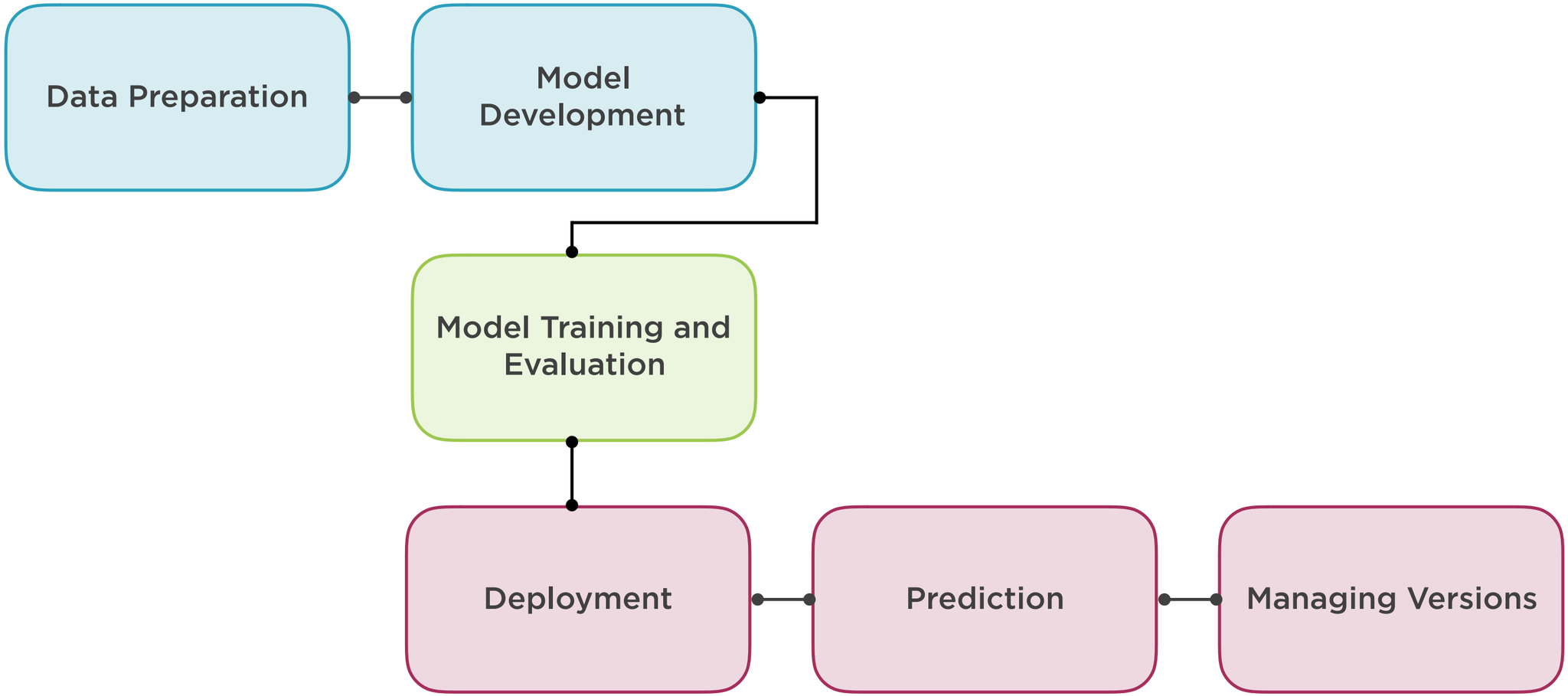
- Distributed training service
- Distributed prediction service

Deep learning VM images

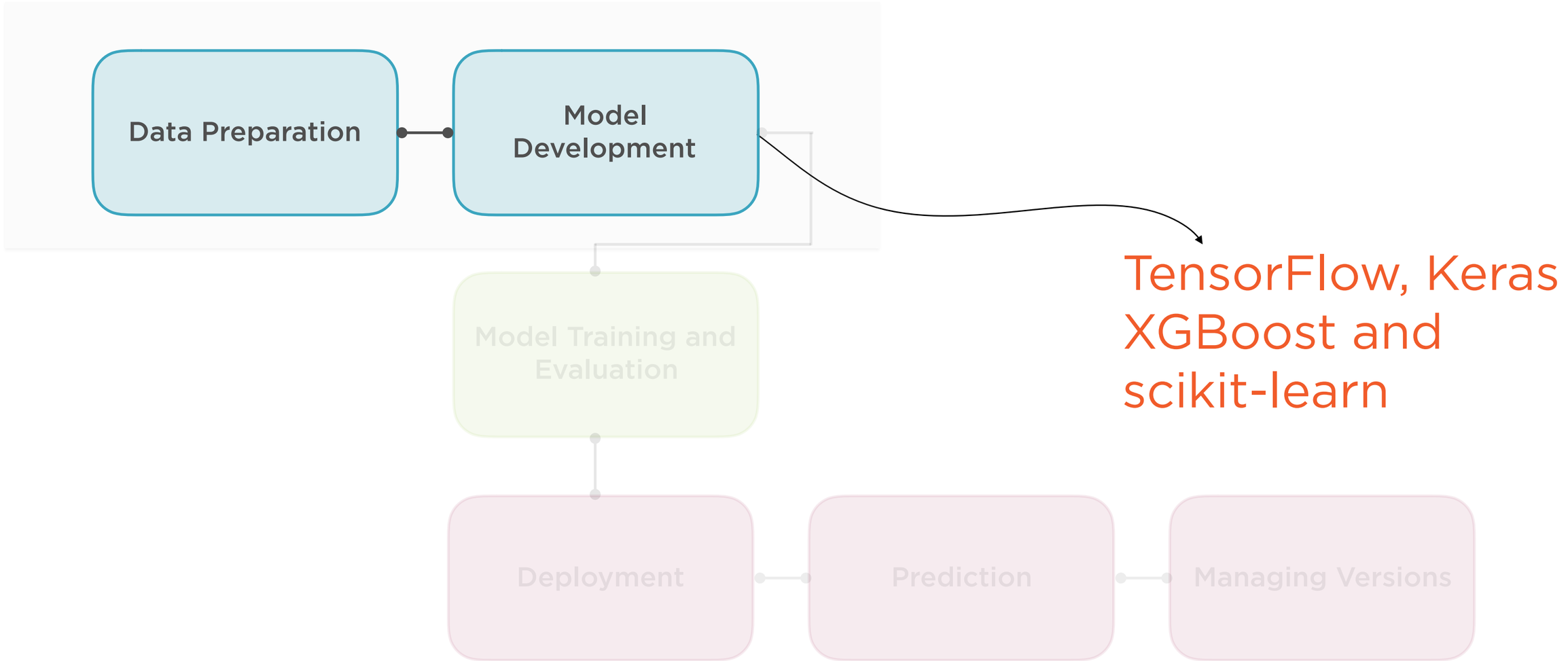
Data labeling service

JupyterLab notebooks

ML Engine in Production ML Workflow



ML Engine in Production ML Workflow



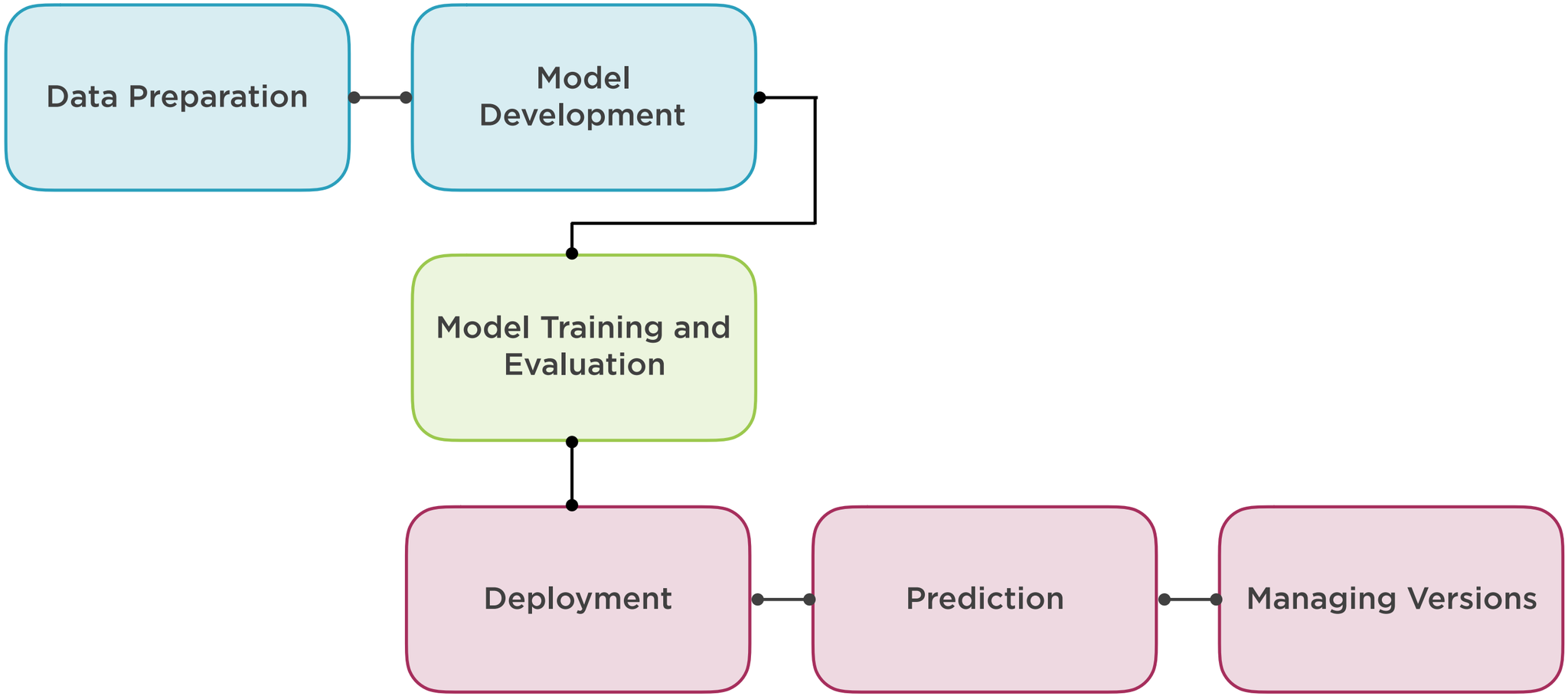
Model

Solution to a problem you're trying to solve. A recipe for predicting a value from data. Different ML techniques using different frameworks can be used to build a model.

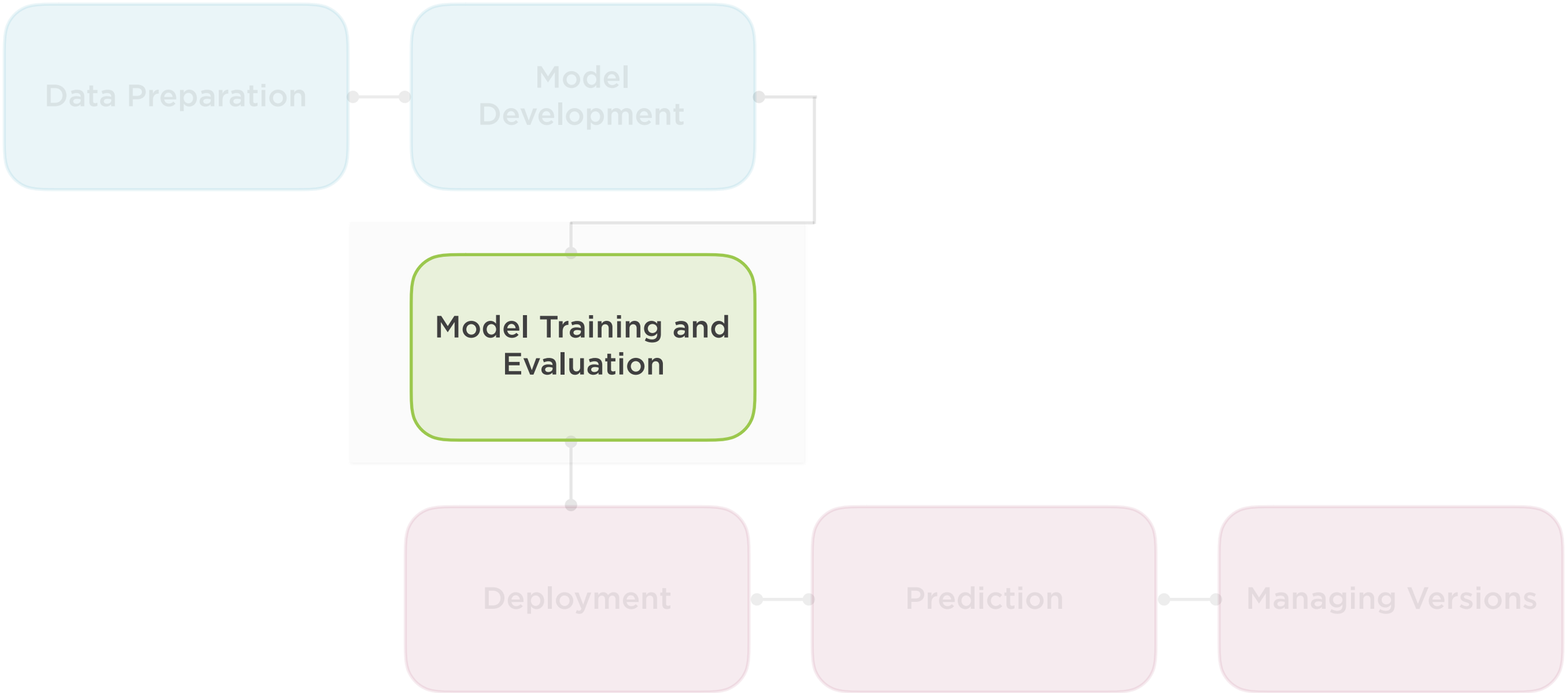
Model

Solution to a problem you're trying to solve. A recipe for predicting a value from data. Different ML techniques using different frameworks can be used to build a model.

ML Engine in Production ML Workflow



ML Engine in Production ML Workflow



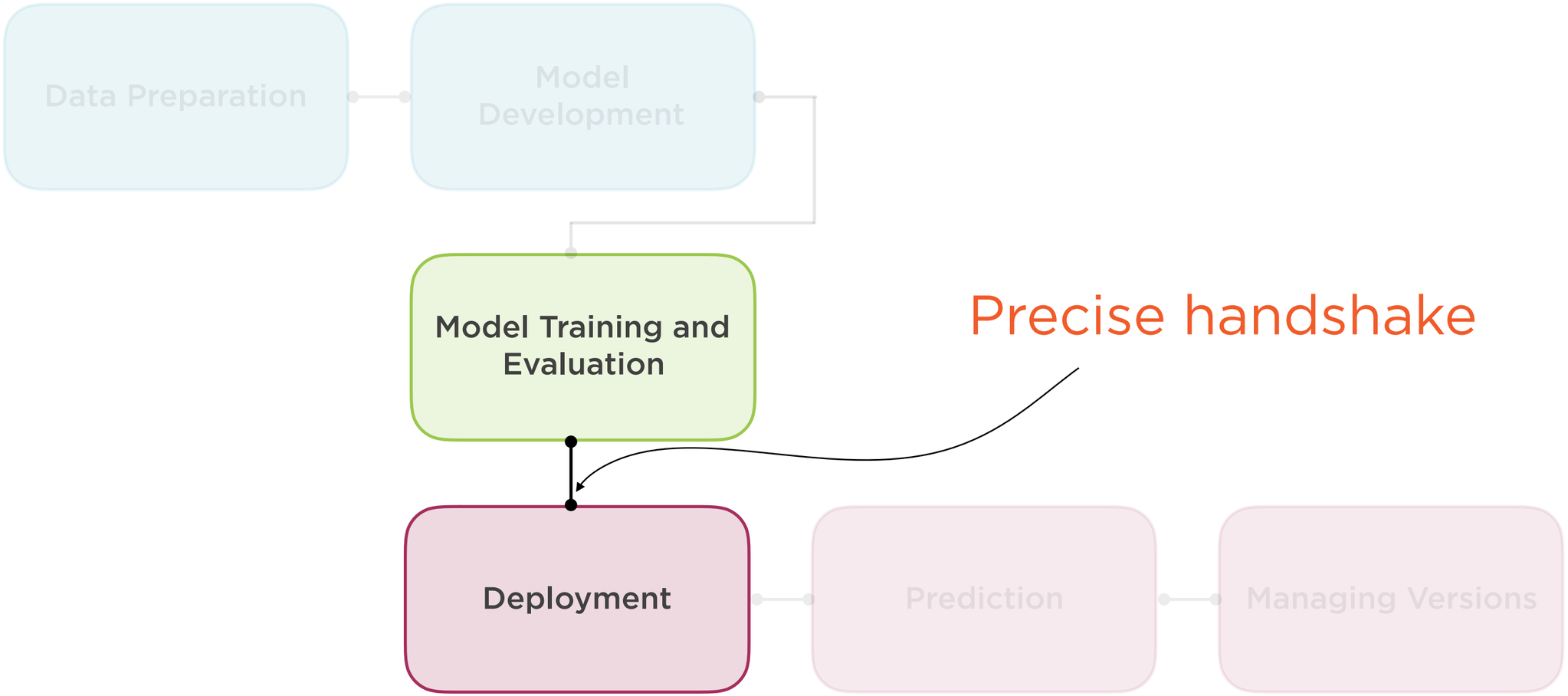
Trained Model

A model object after model parameters have been optimized to fit the training data.

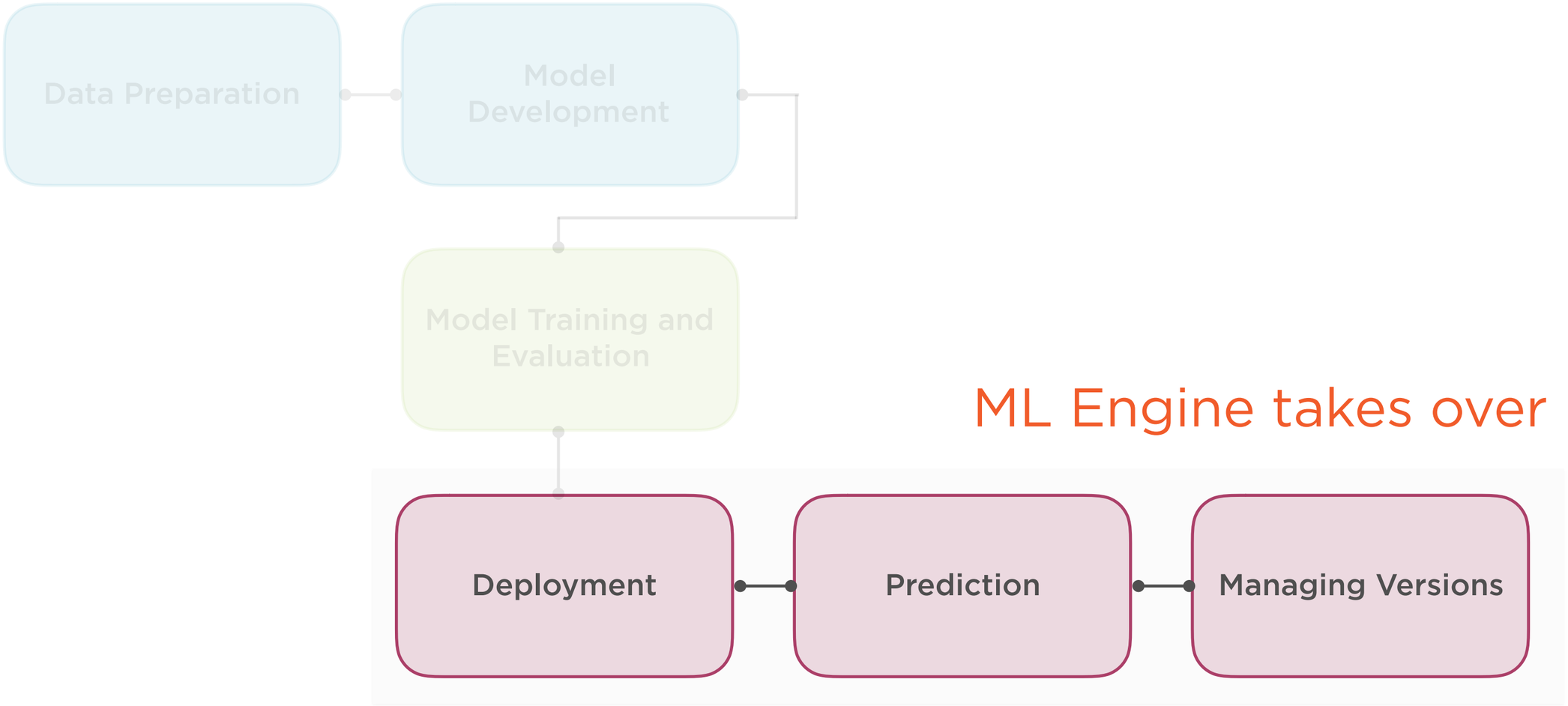
Model Evaluation

Process of tweaking model properties (hyperparameters) and selecting the best model after evaluating many candidates.

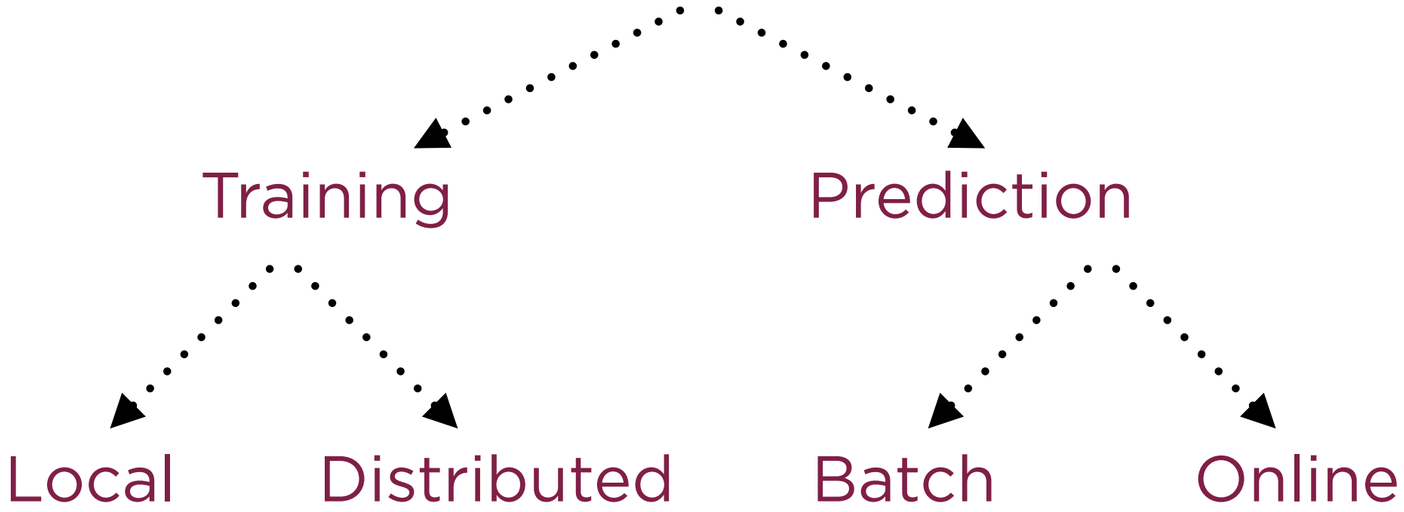
Framework and Engine



ML Engine in Production ML Workflow



Google Cloud ML Engine



Batch and Online Prediction

Online

Minimize latency

Batch

Maximize throughput

XGBoost and scikit-learn are only available in online mode

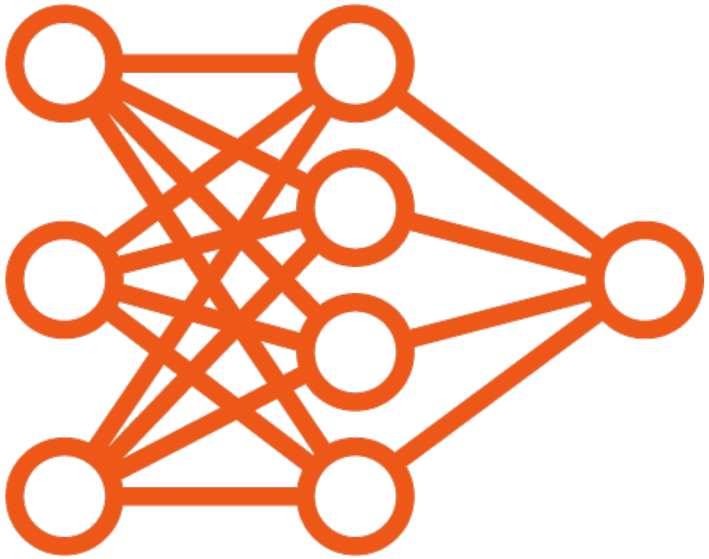
Model Version

When a saved model is handed to ML Engine for deployment, a version is used to uniquely identify that saved model. A single model may have several deployed versions.

Model Version

When a saved model is handed to ML Engine for deployment, a version is used to uniquely identify that saved model. A single model may have several deployed versions.

Google AI Platform



ML Engine Training Service is now called AI Platform Training Service

ML Engine Prediction Service is now called AI Platform Prediction Service

AI Platform Data Labeling Service

Human labelers will label a collection of data that you plan to use to train a custom machine learning model.

Data Labeling Service

Service in pre-release as of September 2019

Need to submit representative samples to human labeling team

Data Labeling Service

Dataset with representative samples

Annotation specification set of labels to be applied

Instructions about how to apply labels to dataset

Demo

Getting started with Google Cloud AI Platform

Demo

Creating and deploying a model and version on Cloud AI Platform

Demo

Creating an evaluation job to sample prediction instances

Demo

Testing the sentiment analysis model using GCP's web console

Demo

Using the gcloud command line utility to invoke the model for predictions

Demo

Deploying a new version of the model from the command line

Using curl to invoke model predictions

Demo

Monitoring a deployed model using Stackdriver

Summary

Google AI Platform as re-branded umbrella service

Training, prediction, data labeling service

Supported frameworks, models, model versions

Deploying models using the web console and command line

Monitoring the model using Stackdriver