Deploying Deep Learning Models to AWS Sagemaker

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

The machine learning workflow with Amazon SageMaker

Develop training script for distributed training

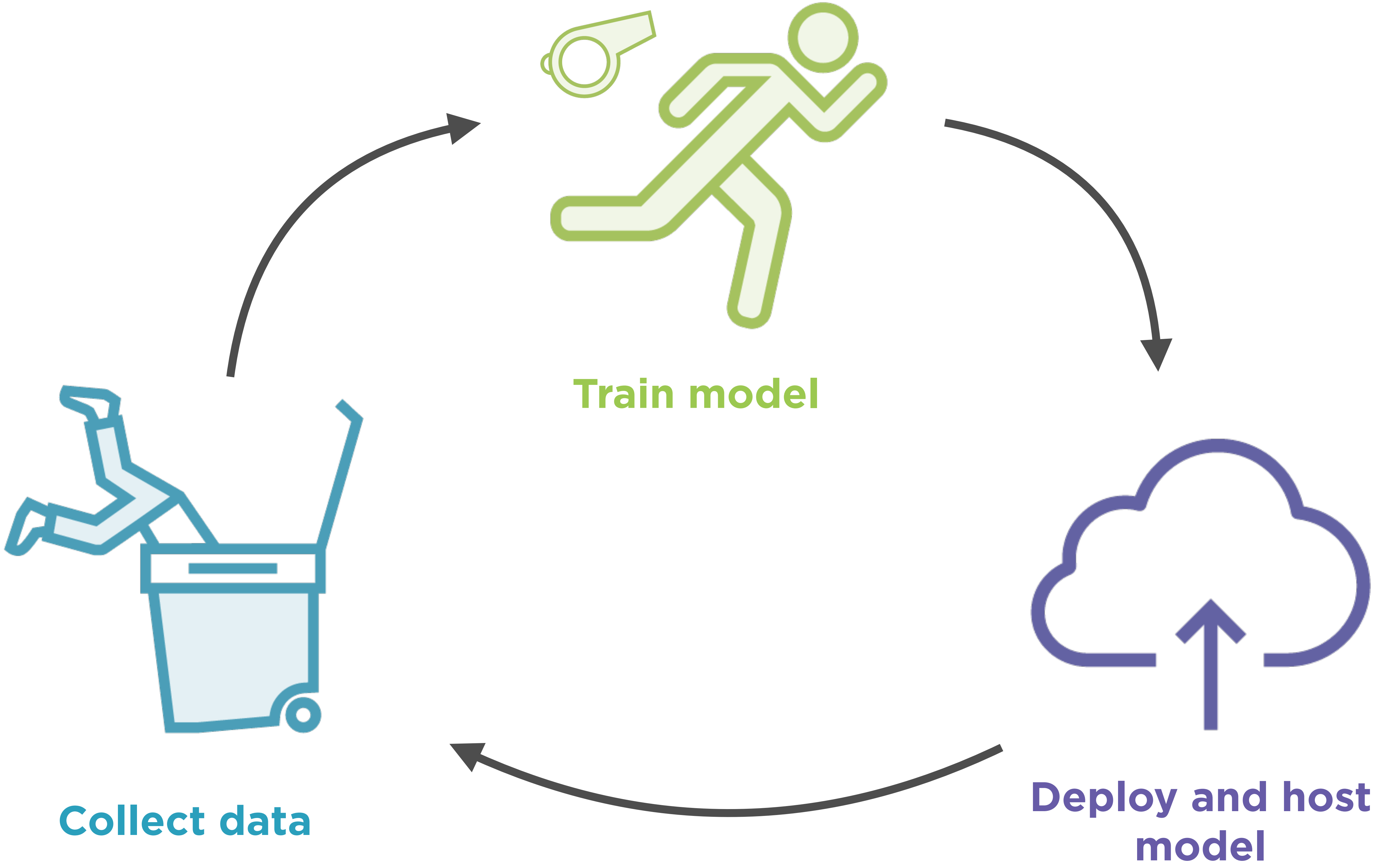
Run distributed training for models using high-level estimators

Deploy and host models on Amazon Hosting Services for online prediction

Enable CloudTrail to track events for auditing and compliance

Amazon SageMaker for Deep Learning

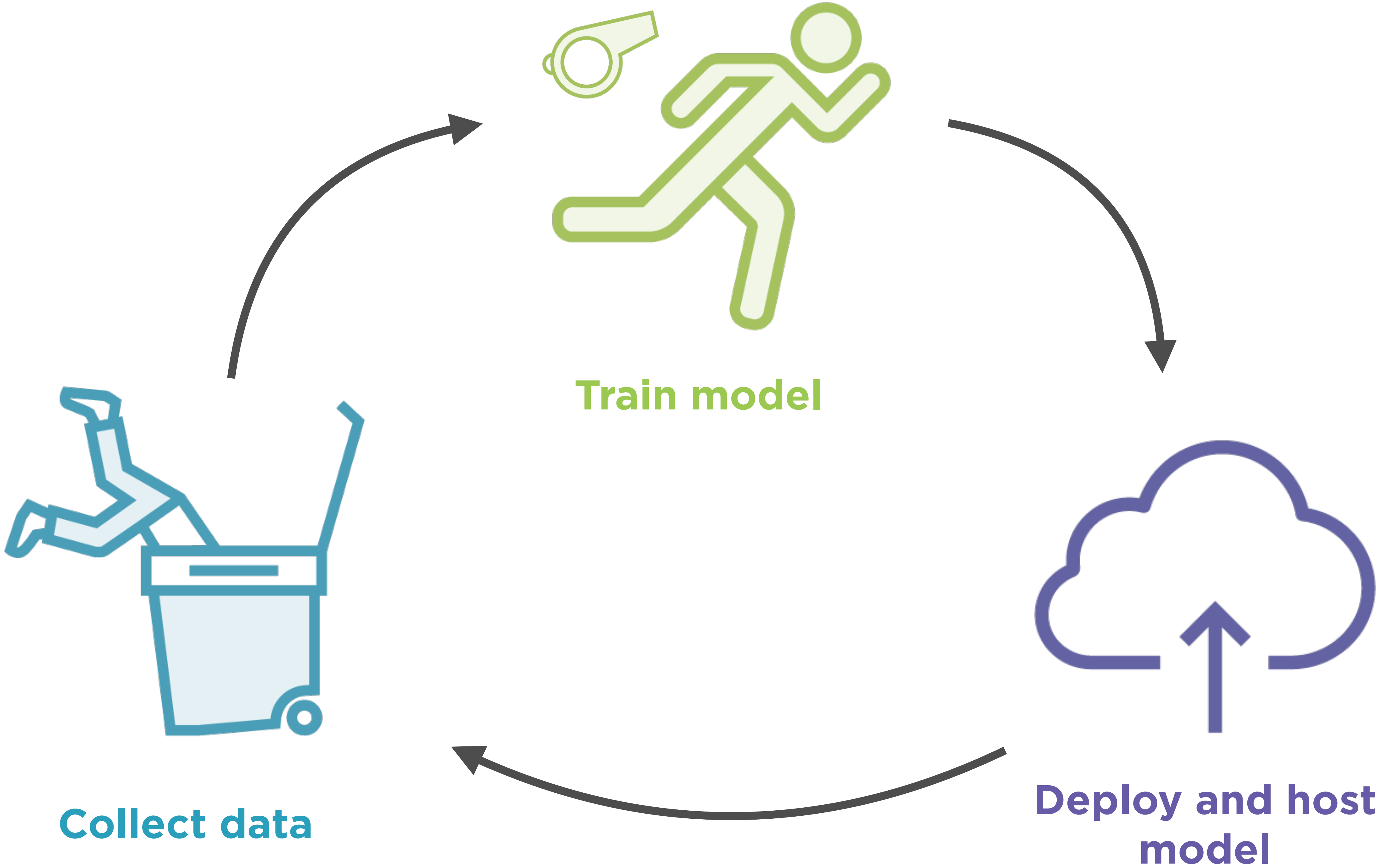
Machine Learning Workflow



Data Preparation

SageMaker runs Jupyter notebooks on instances in the cloud to explore and prepare data

Machine Learning Workflow



Model Training

Machine learning algorithms

- Traditional models, neural networks

Allocate compute resources

- VMs, memory, scaling parameters, GPUs/ CPUs

Evaluate the model

- AWS SDK for Python, Jupyter notebooks

ML Algorithms on SageMaker

Built-in algorithms

Out-of-the-box models hosted on containers on the AWS cloud

Bring your algorithm

Develop your own code in TensorFlow, Apache MXNet etc.

ML Algorithms on SageMaker

Built-in algorithms

Out-of-the-box models hosted on containers on the AWS cloud

Bring your algorithm

Develop your own code in TensorFlow, Apache MXNet etc.

ML Algorithms on SageMaker

Bring your algorithm

Develop your own code in TensorFlow, Apache MXNet etc.

Bring your own code

Bring your own model

Bring your own container

ML Algorithms on SageMaker

Bring your algorithm

Develop your own code in TensorFlow, Apache MXNet etc.

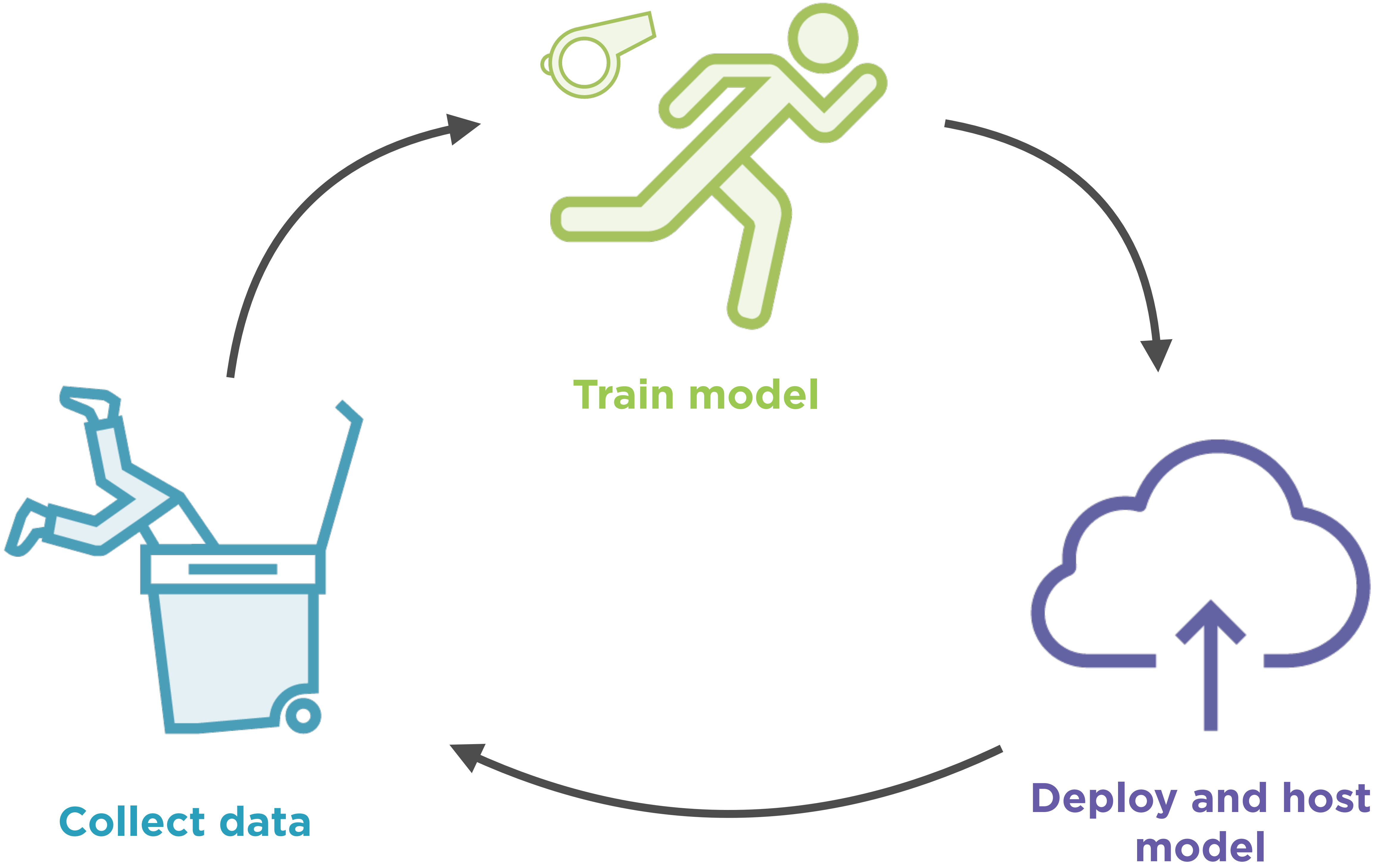
Bring your own code

Bring your own model

Bring your own container

Training a Model on SageMaker

Machine Learning Workflow

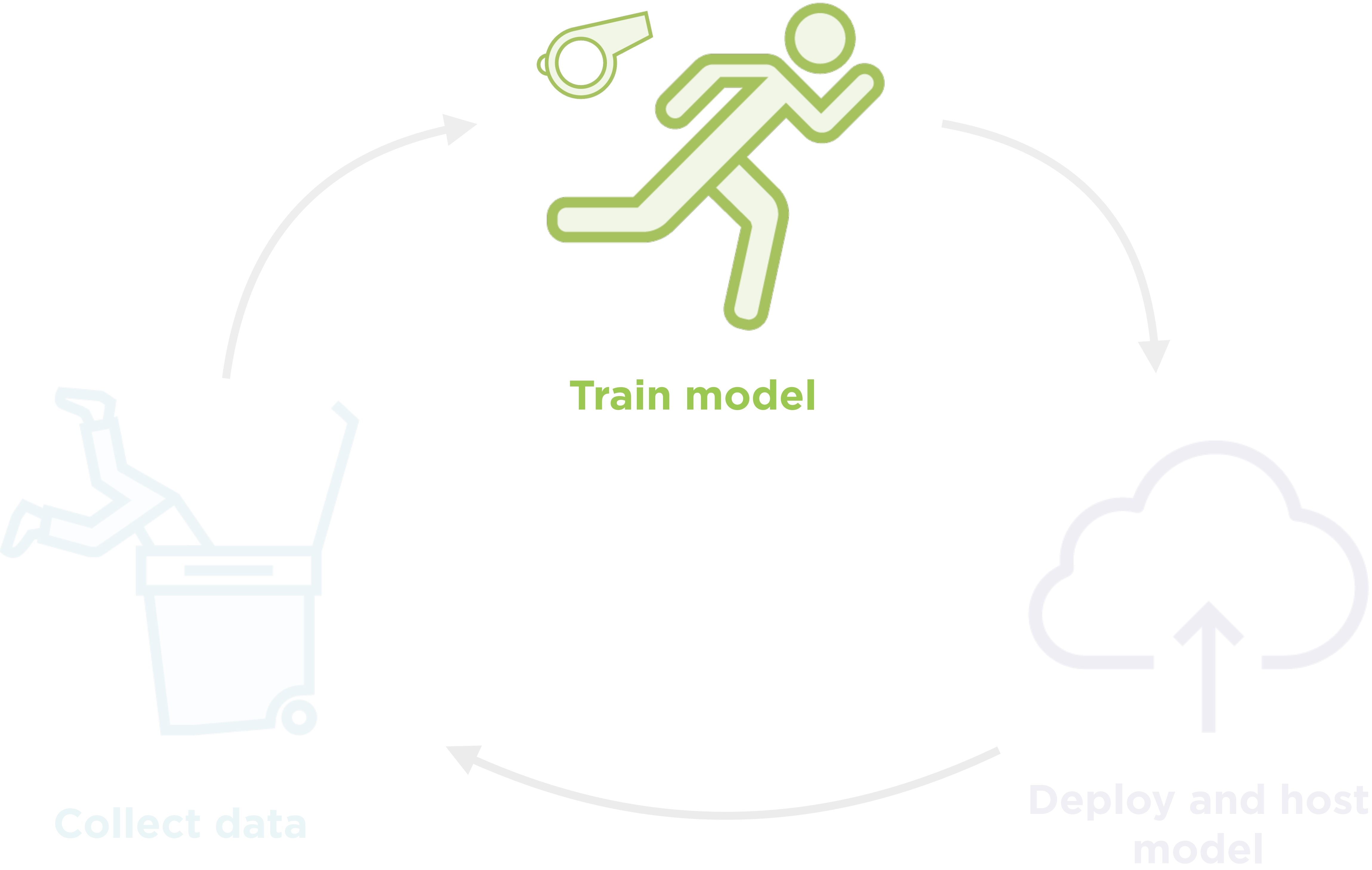


Train model

Deploy and host model

Collect data

Machine Learning Workflow

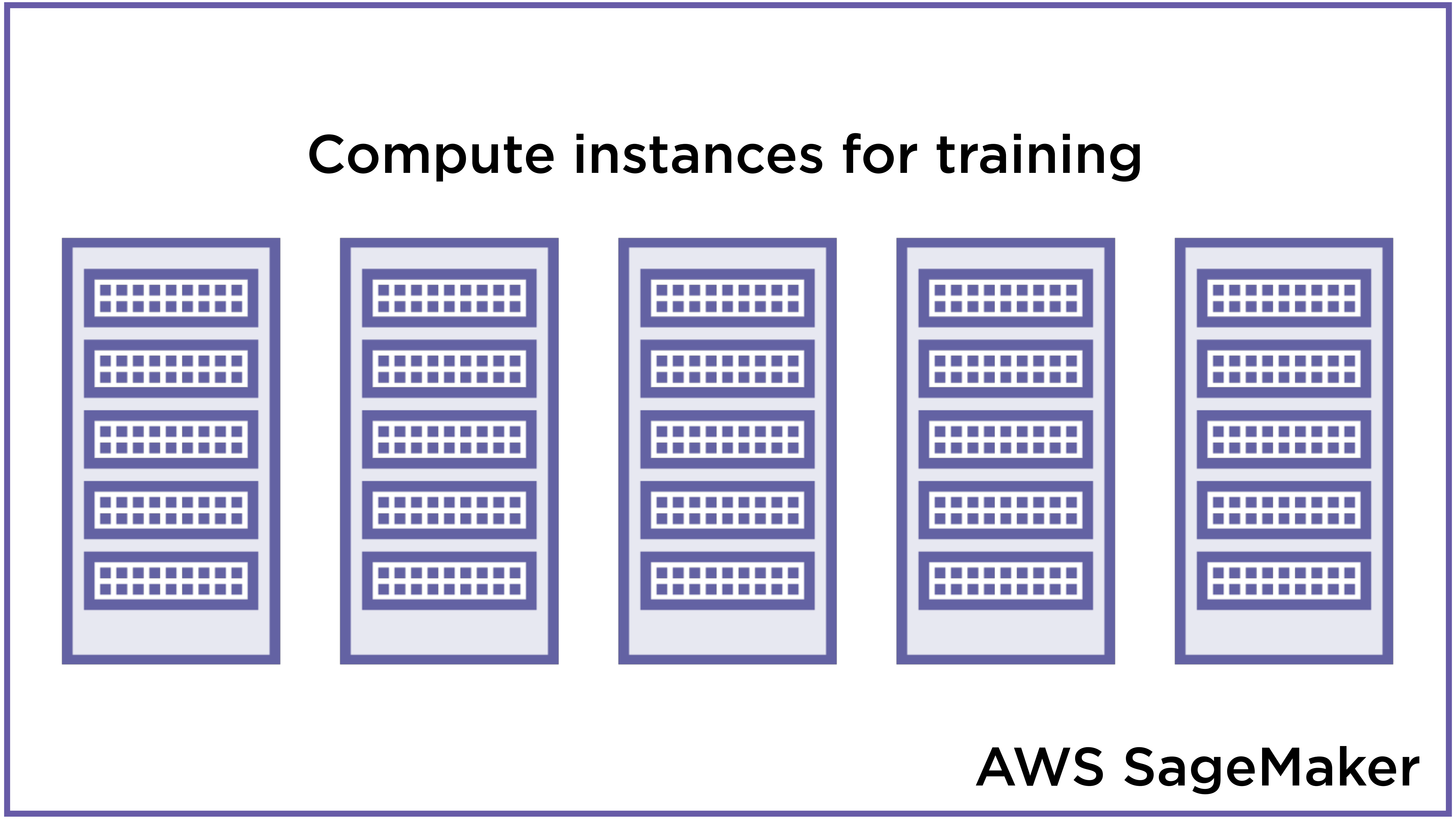


Train model

Deploy and host model

Collect data

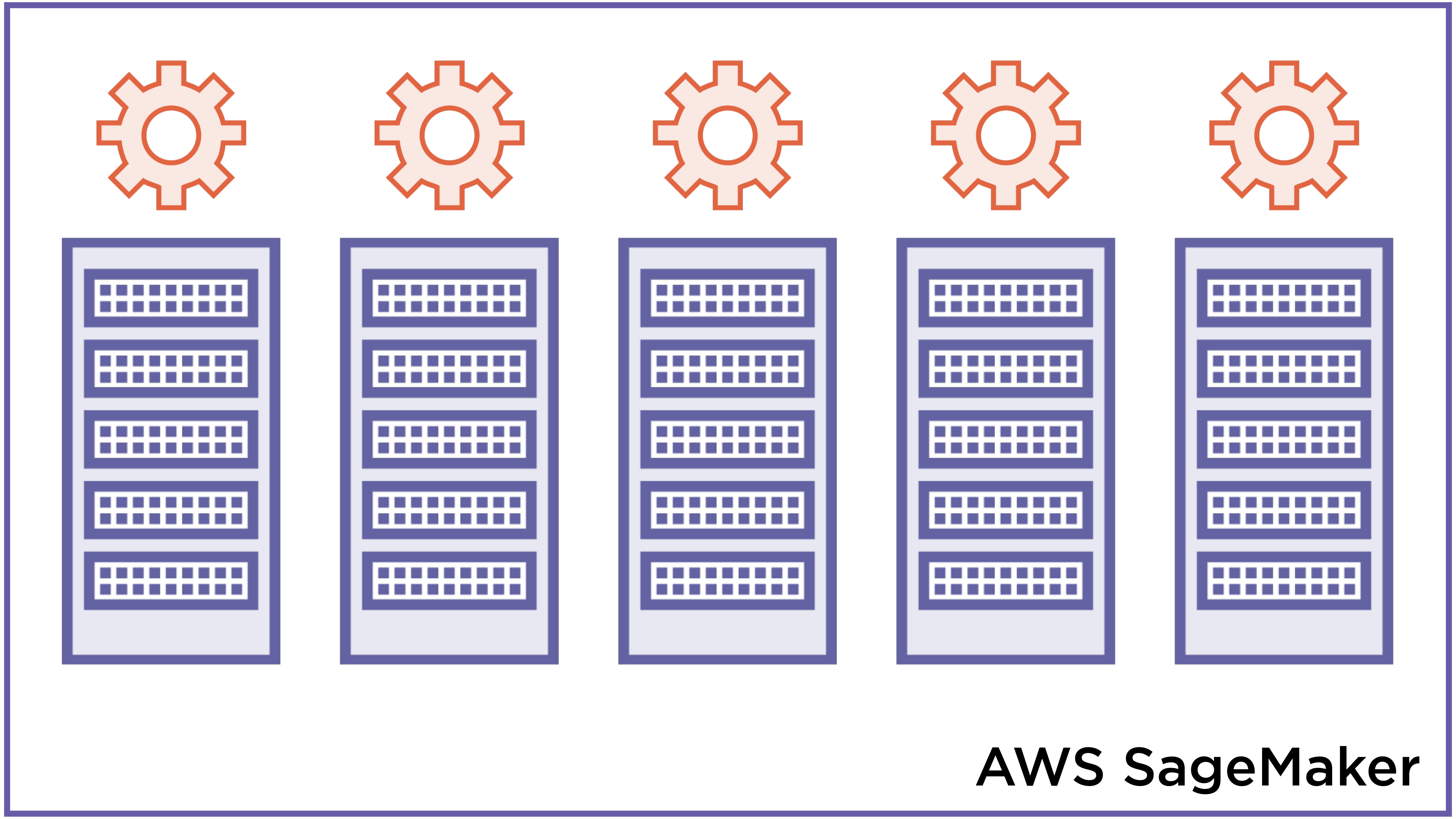
Training a Model



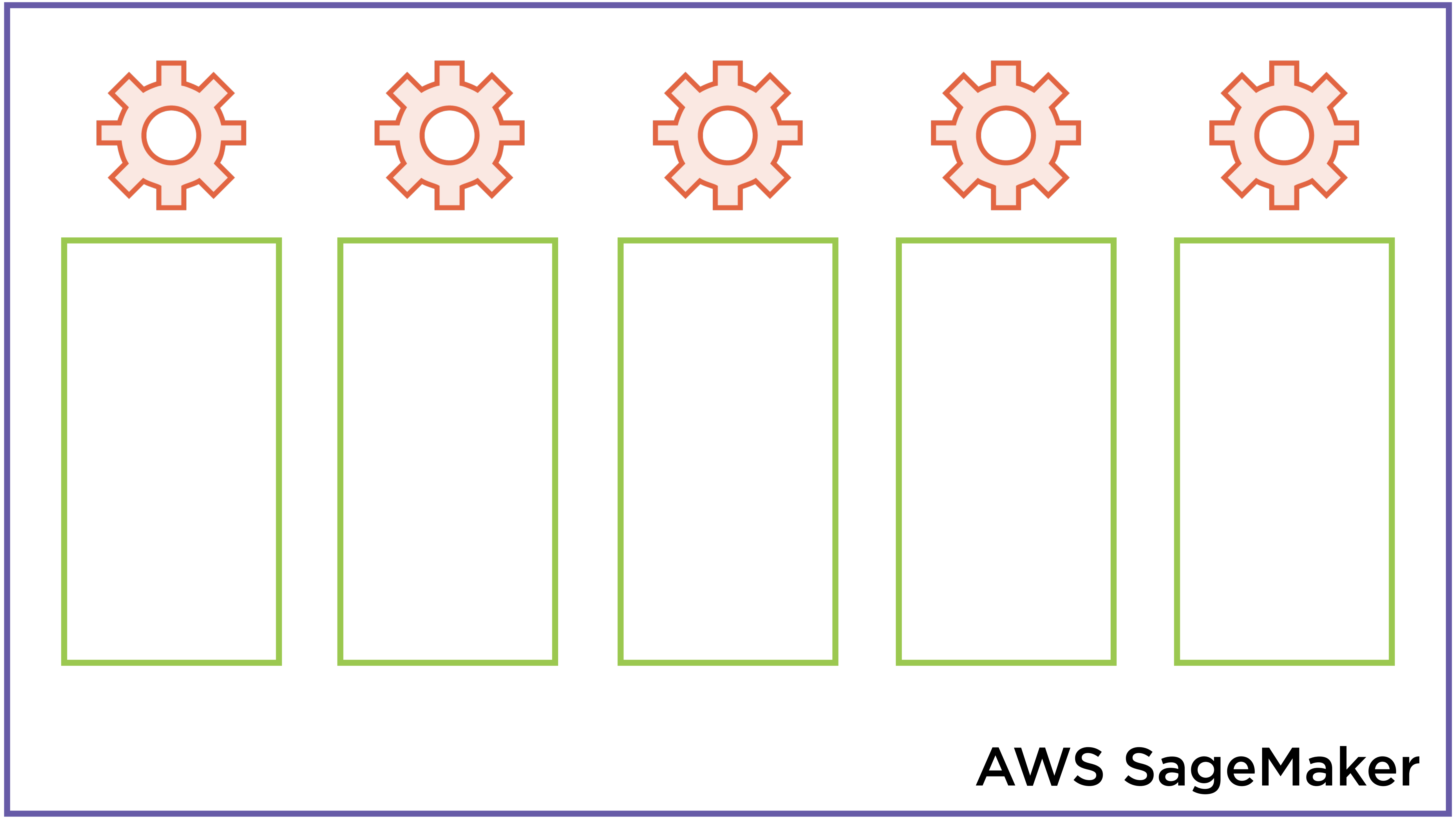
Compute instances for training

AWS SageMaker

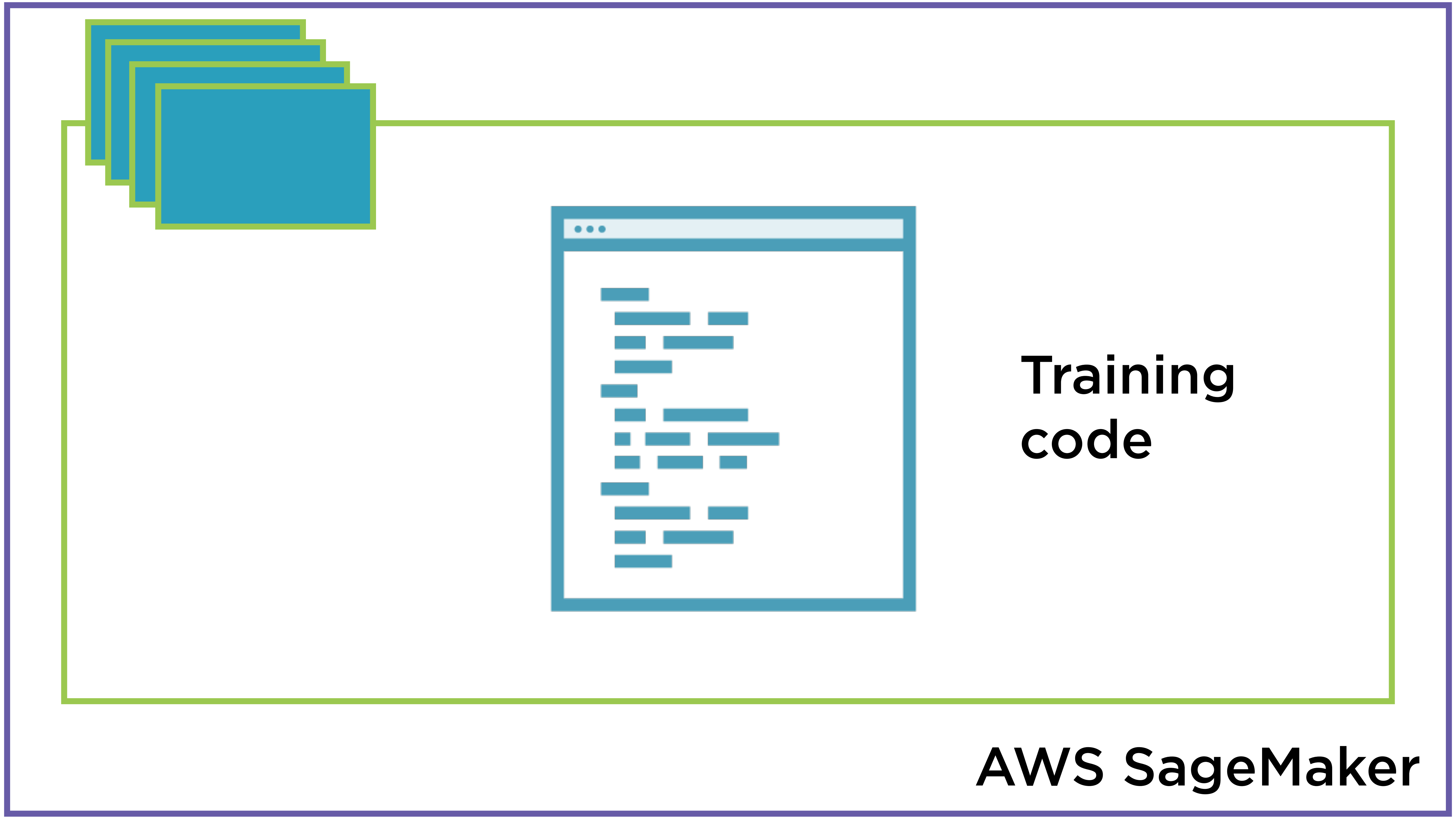
Training a Model



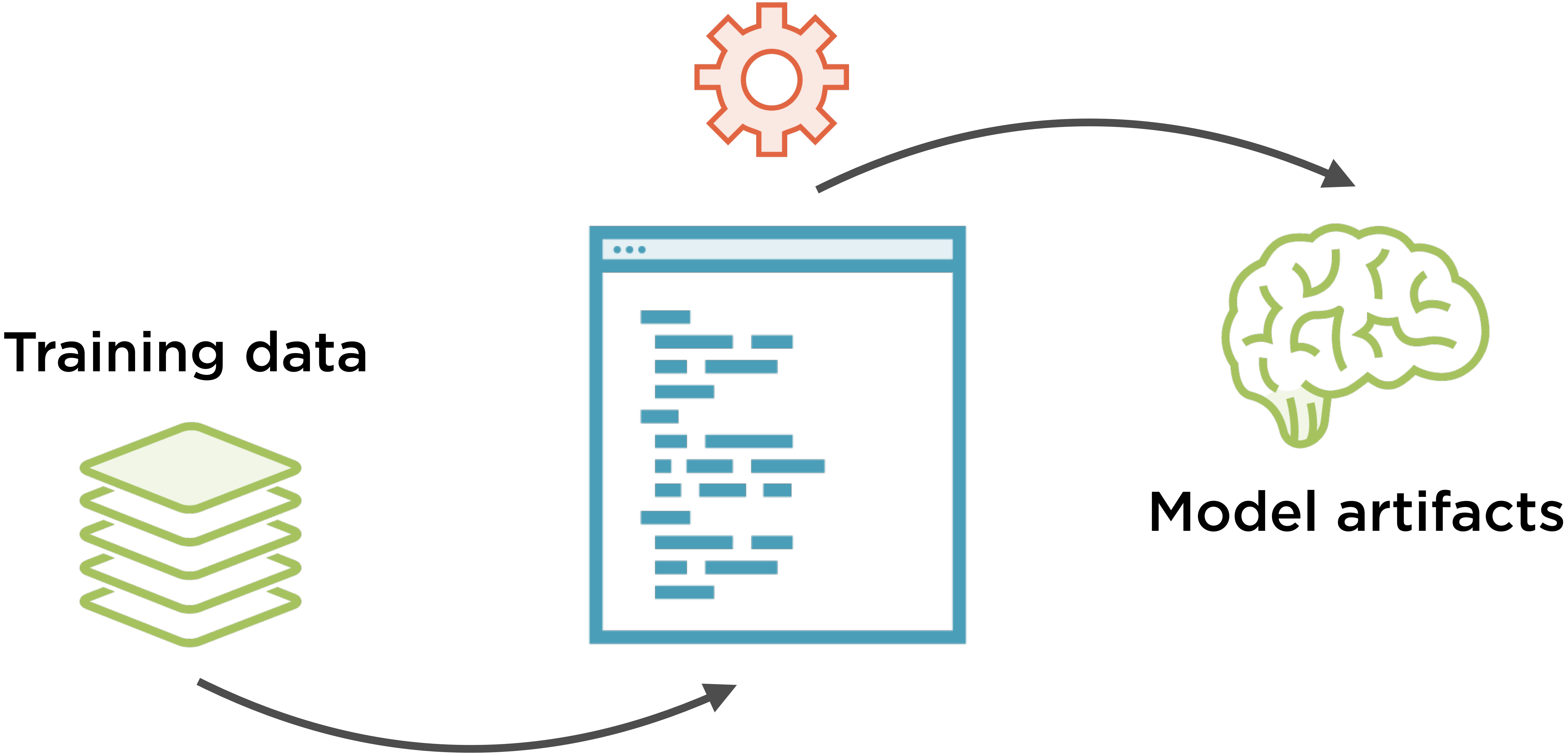
Training a Model



Training a Model



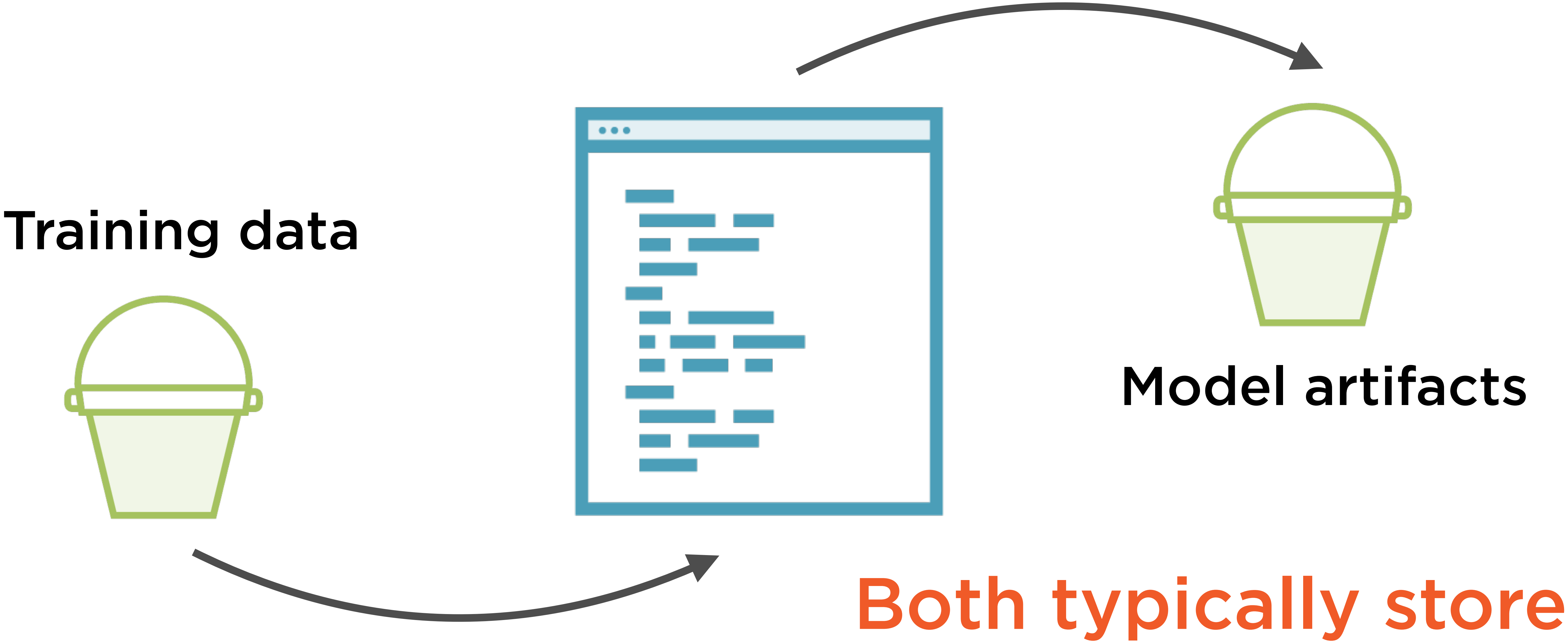
Training a Model



Training data

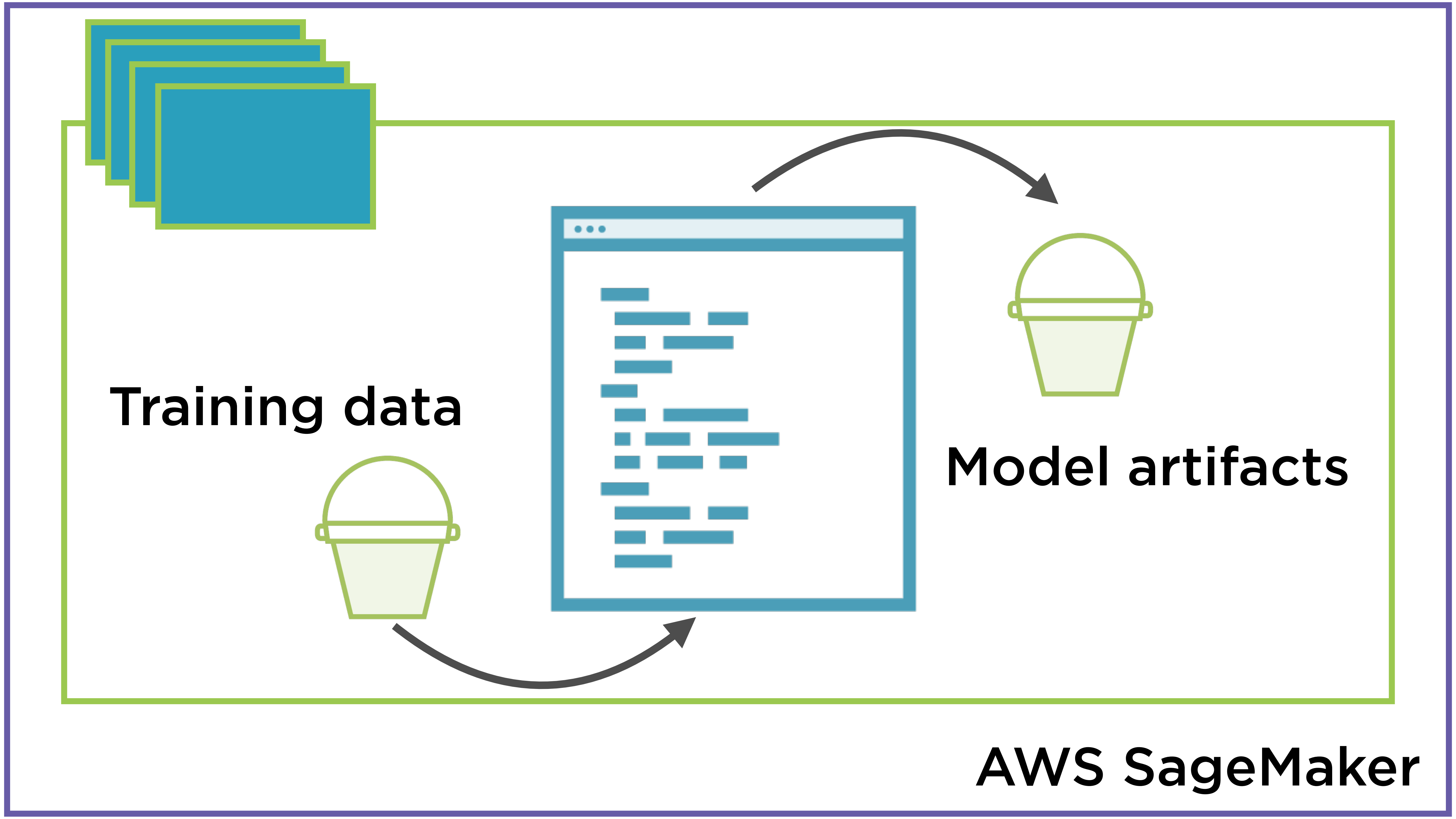
Model artifacts

Training a Model



Both typically stored in S3 buckets

Training a Model

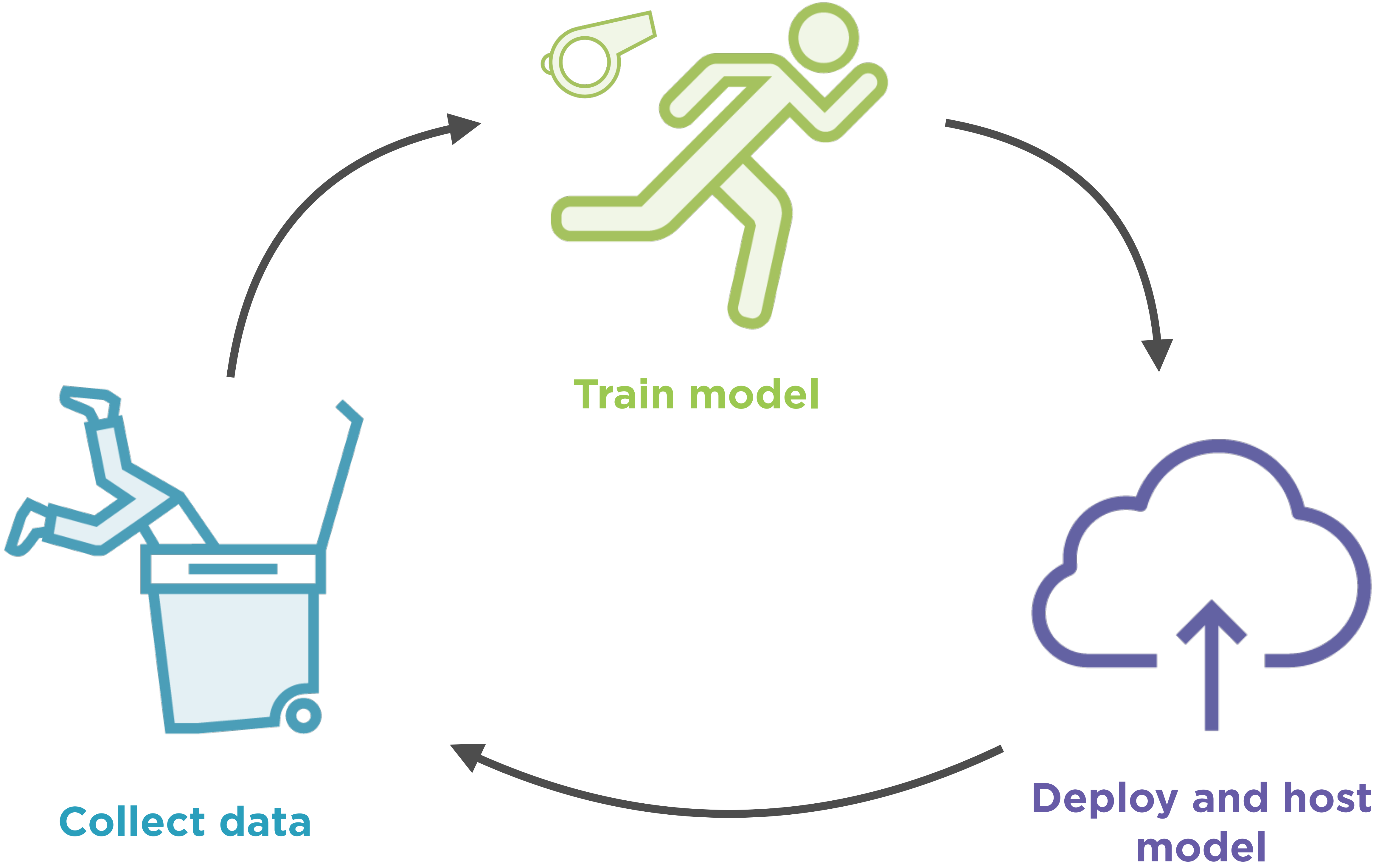


Estimator

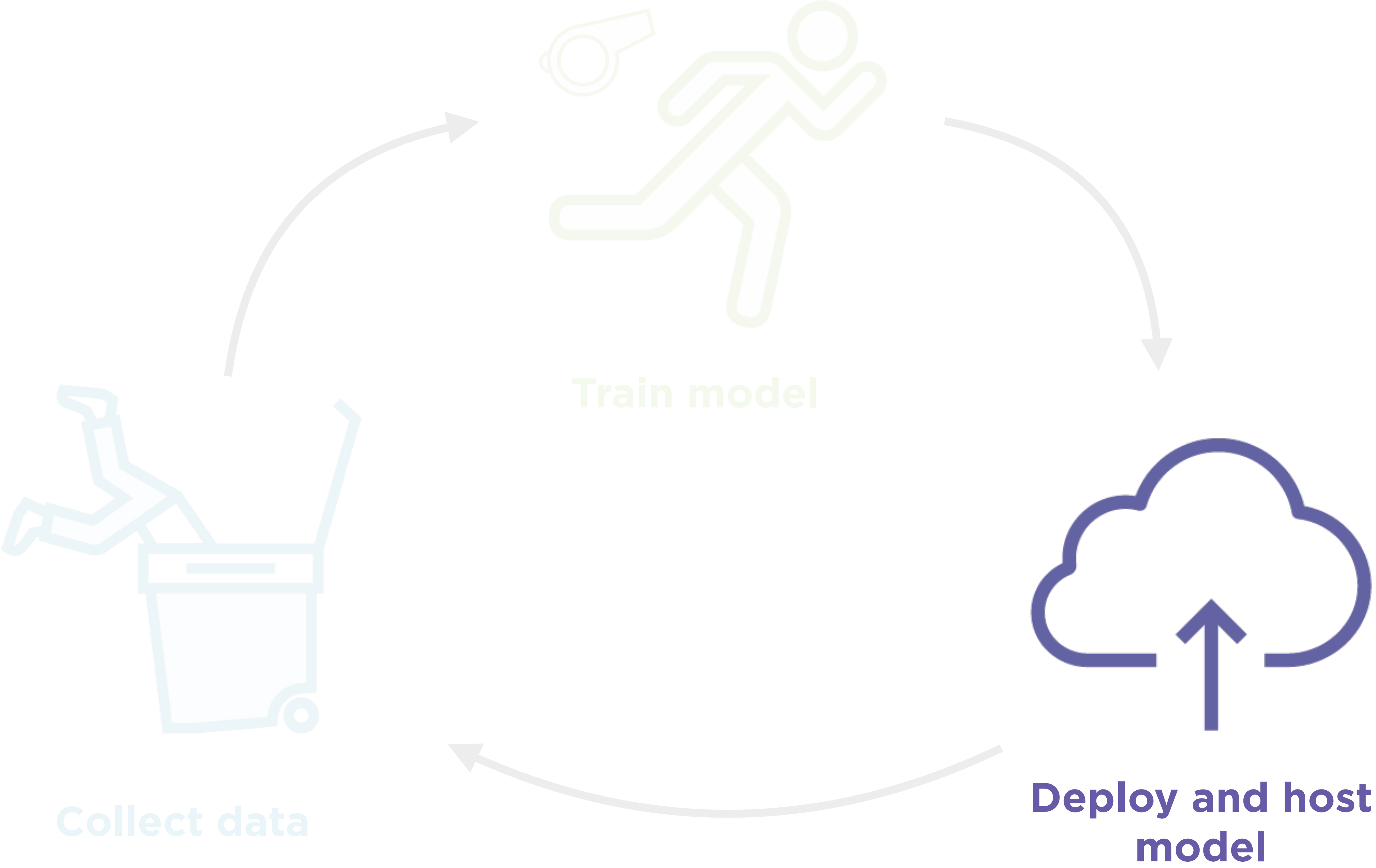
High-level API, specific to a cloud platform, that helps build, train, and deploy models

Deploying a Model on SageMaker

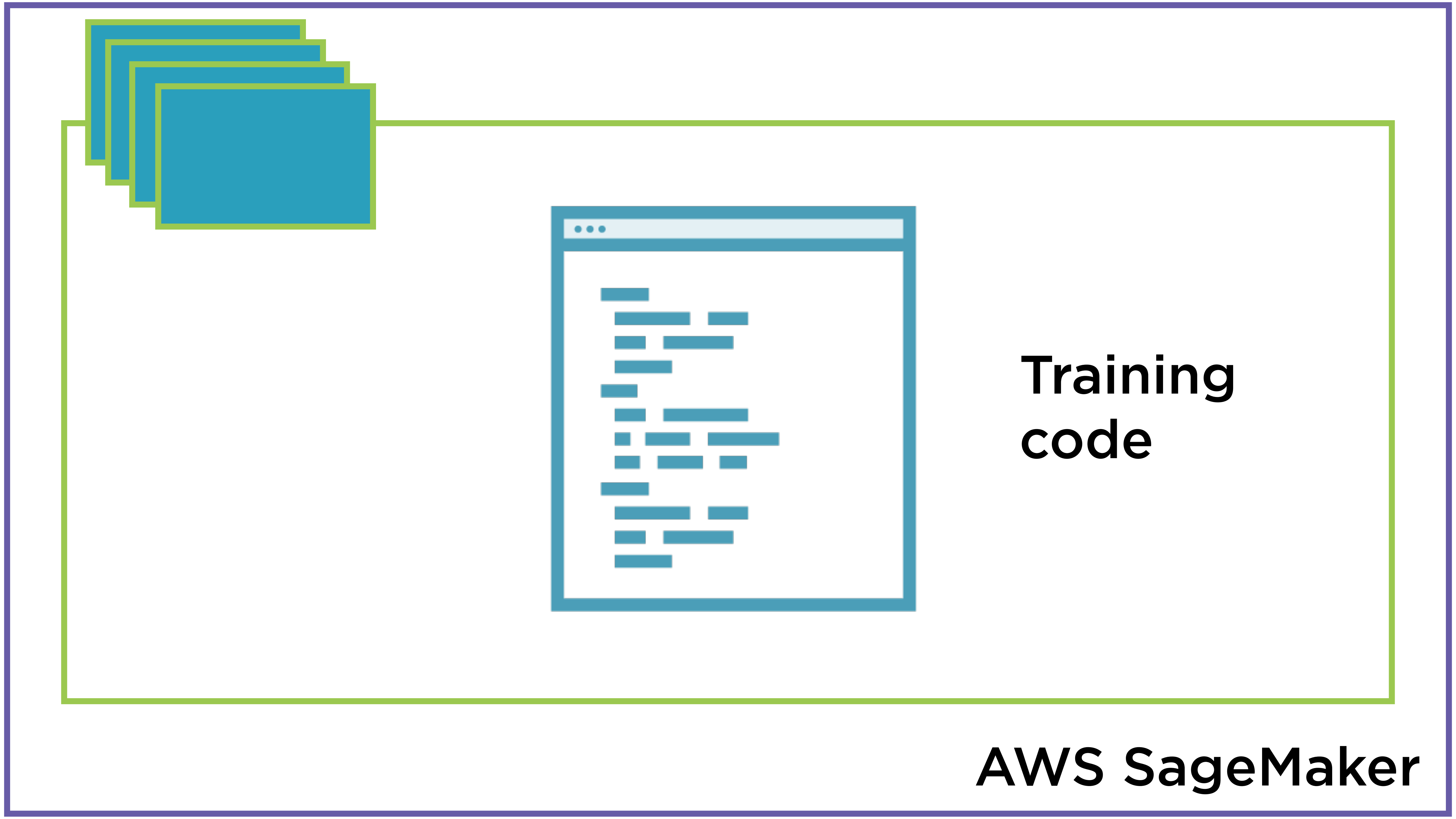
Machine Learning Workflow



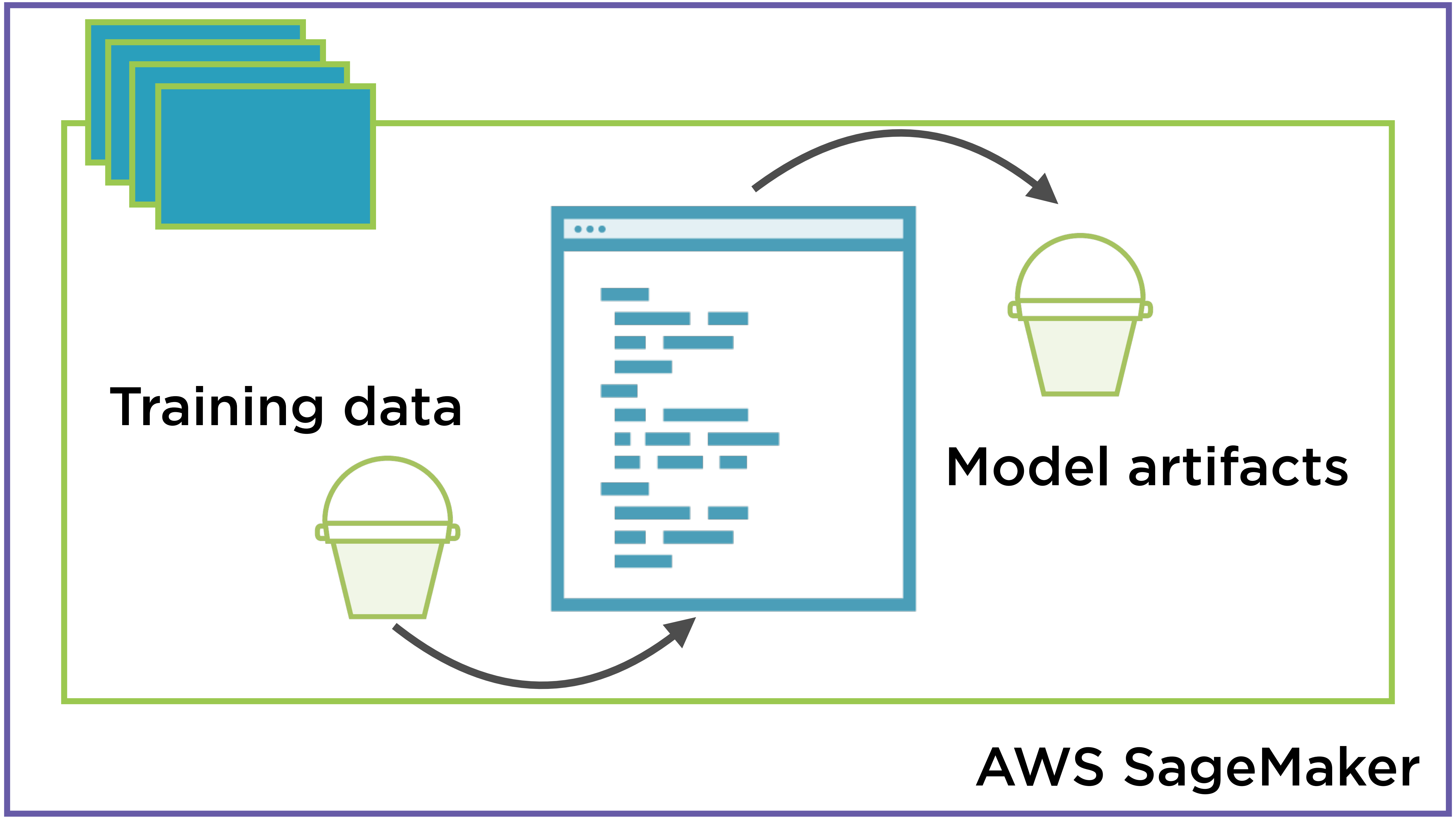
Machine Learning Workflow



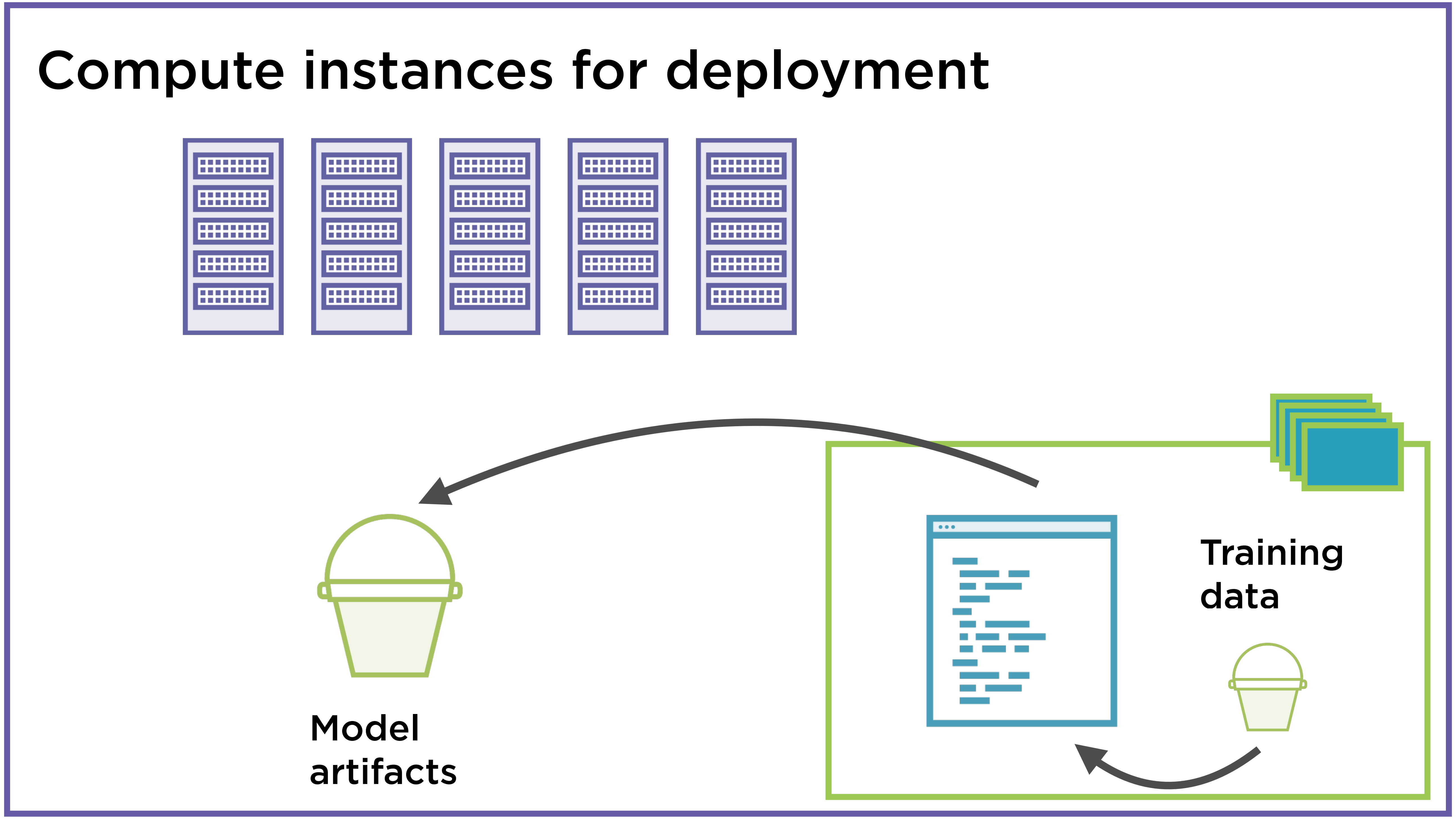
Training a Model



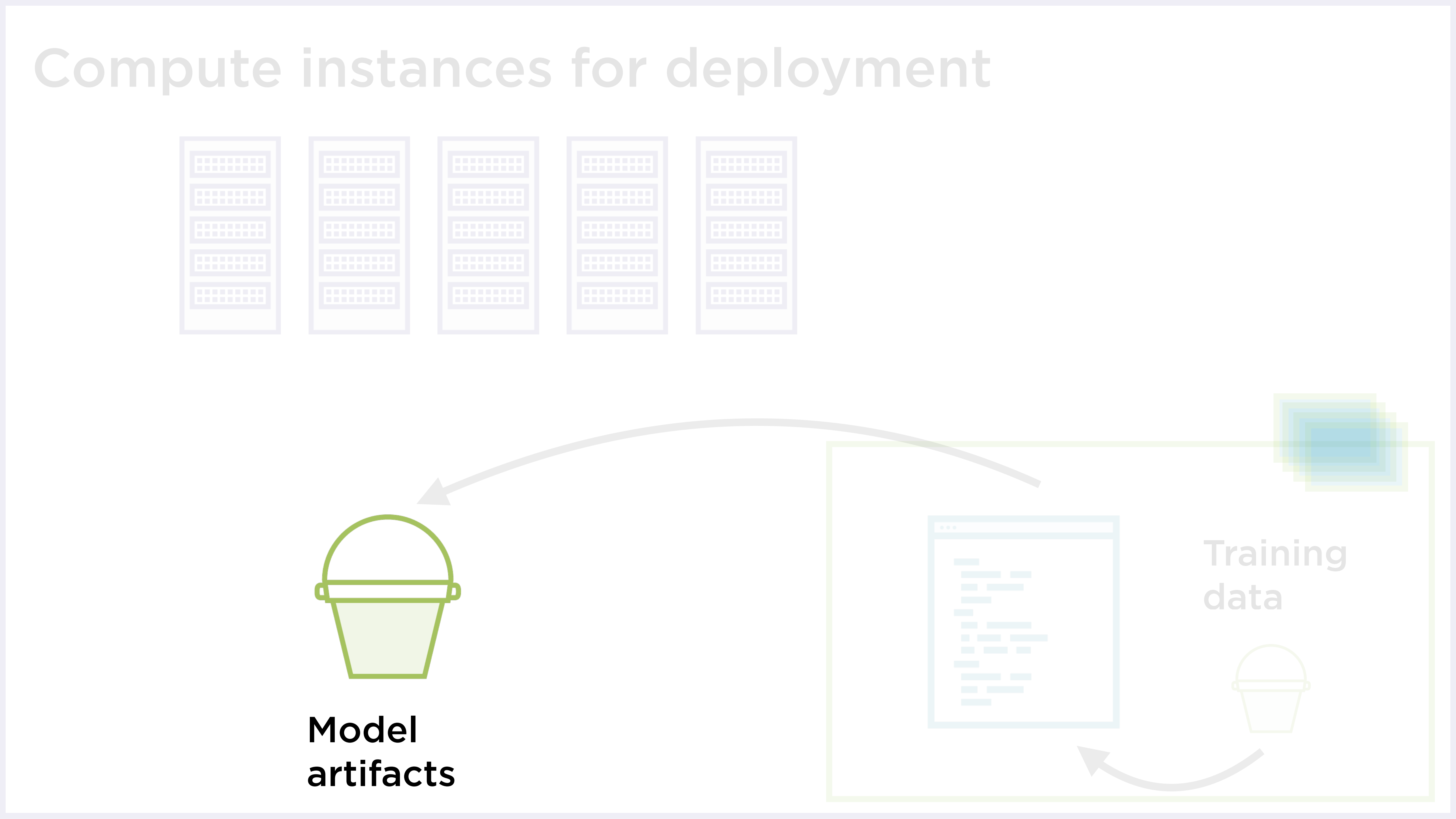
Training a Model



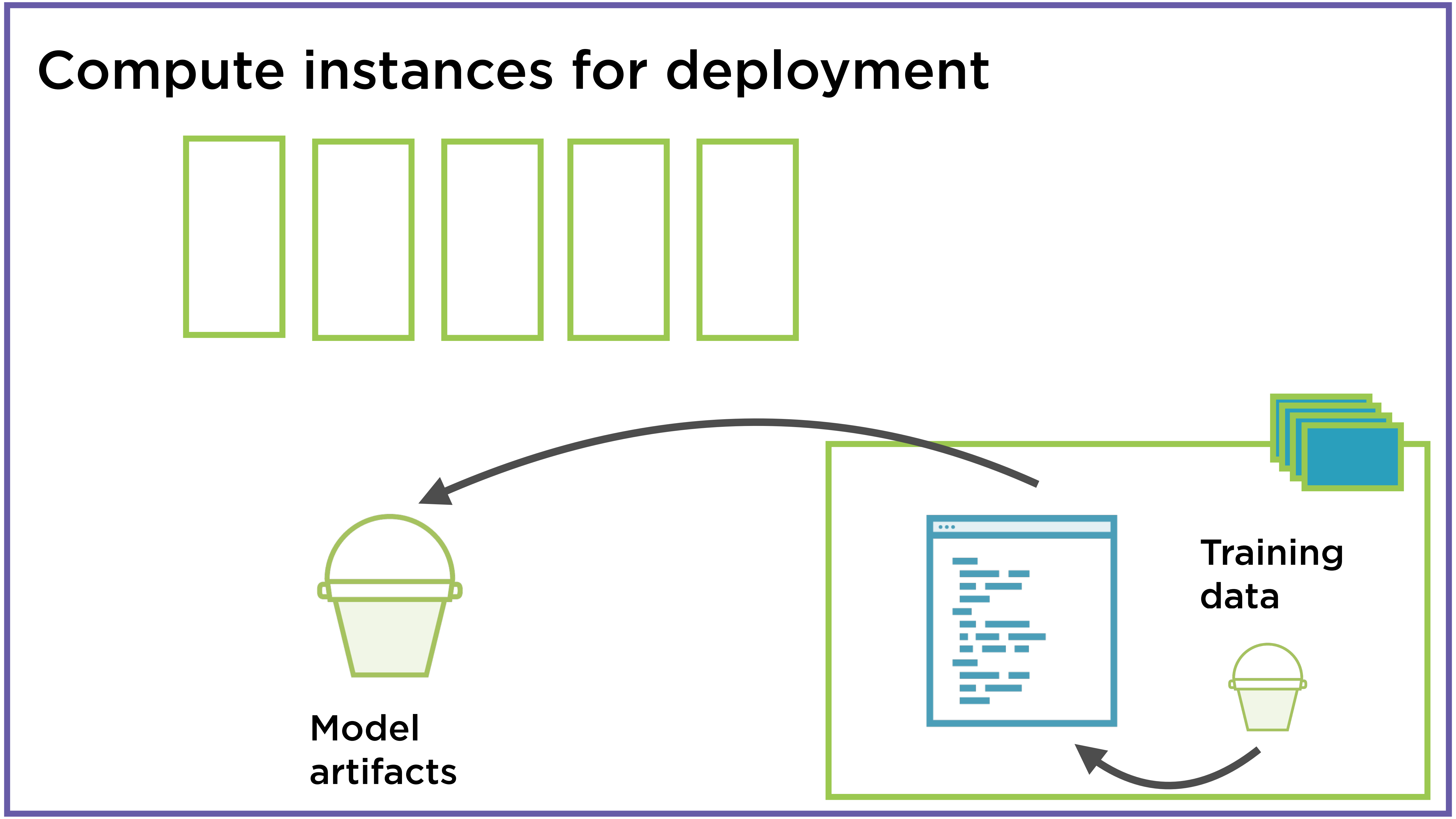
Deploying a Model



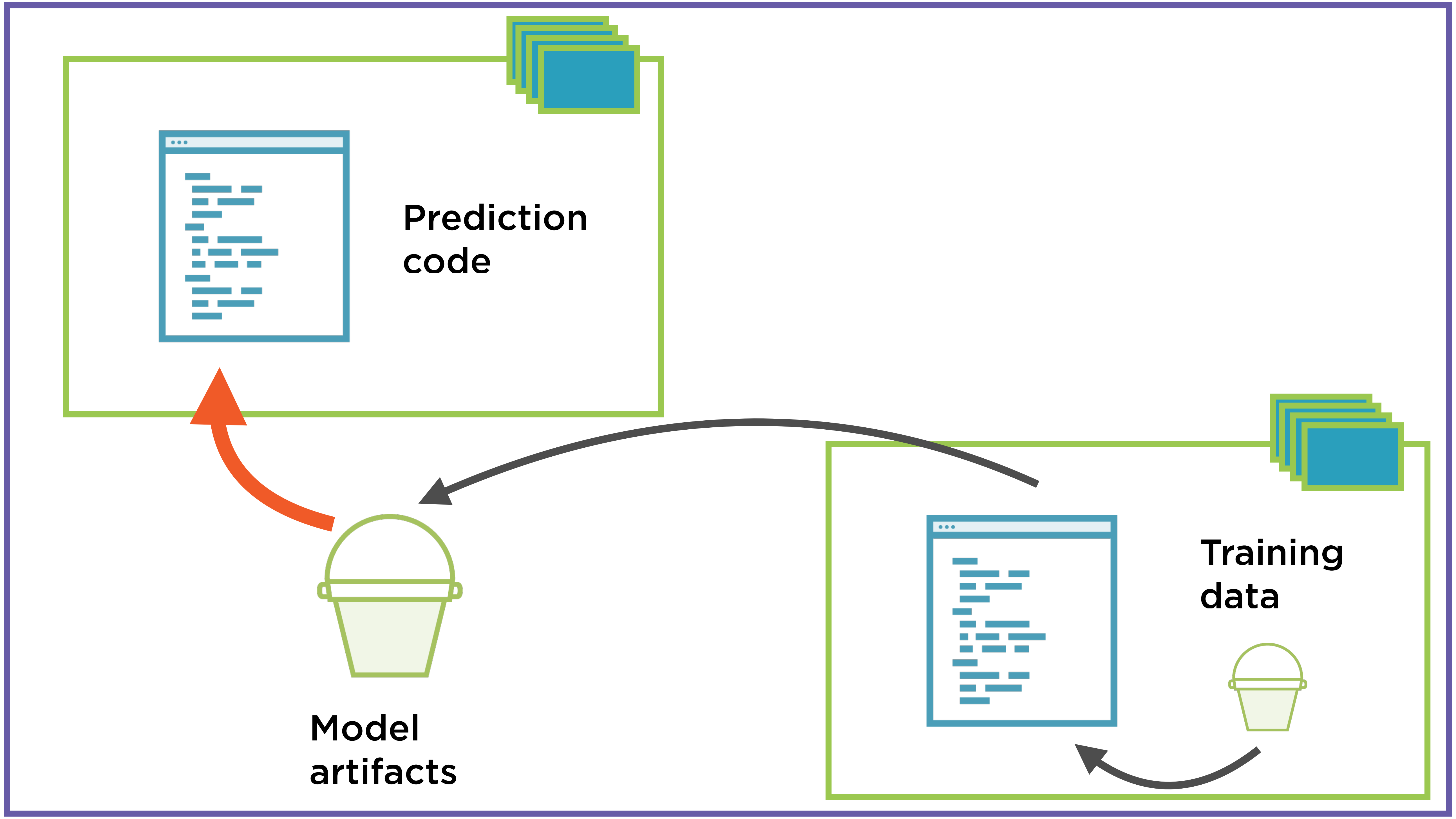
Deploying a Model



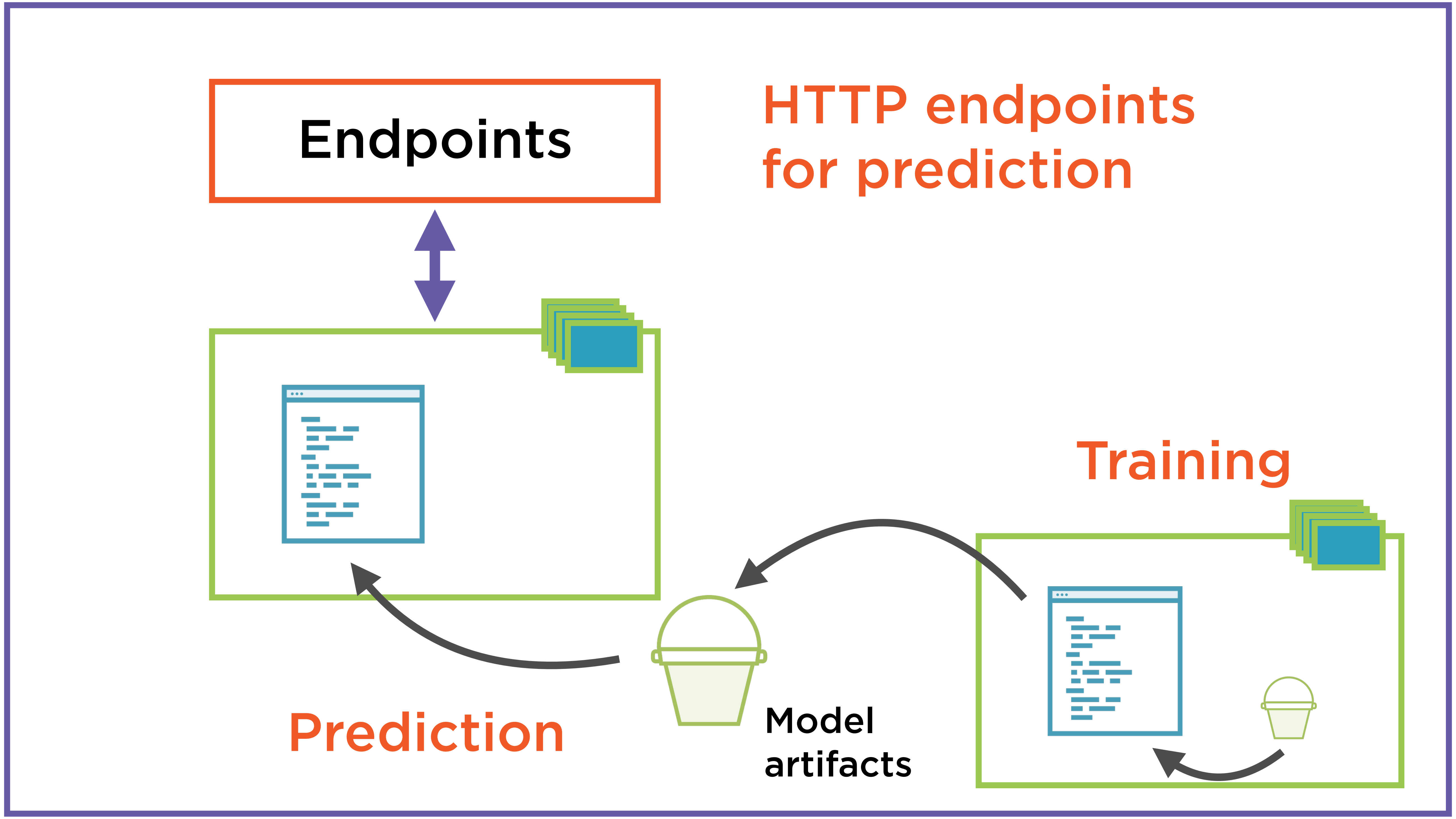
Deploying a Model



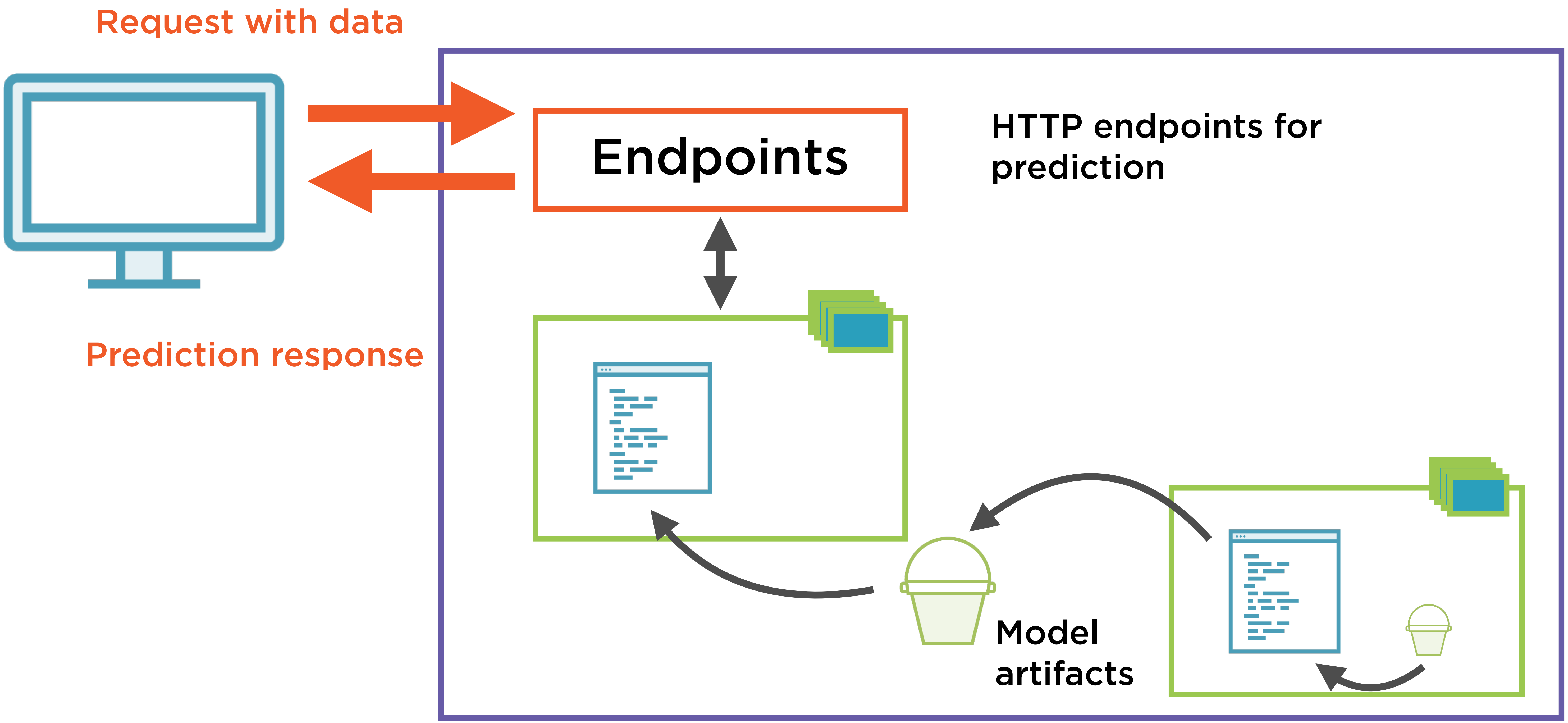
Deploying a Model



Deploying a Model



Deploying a Model



Steps in Deploying a Model

Create a model

Specify model artifacts, give model name

Create an endpoint configuration

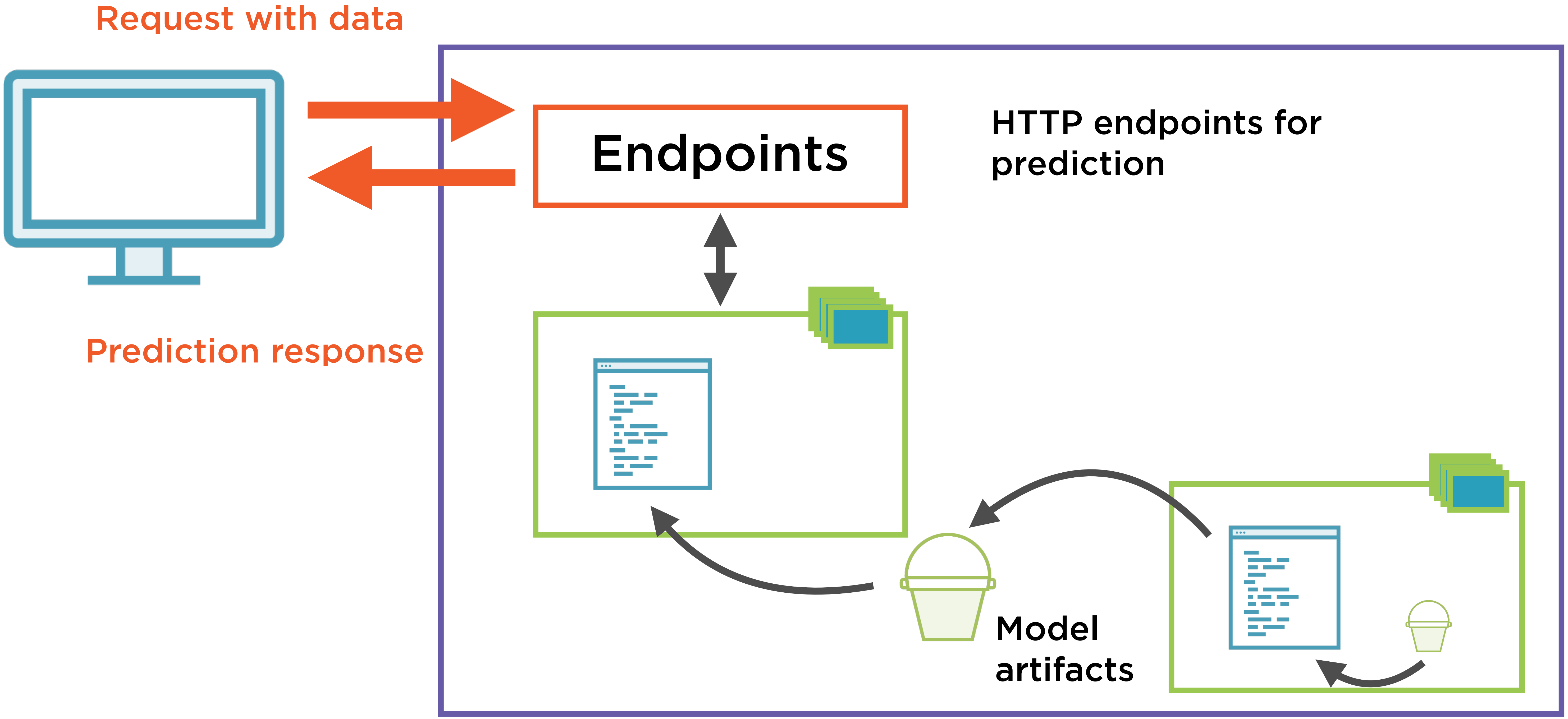
Specify model name and compute instances

Create an HTTPs endpoint

Provide endpoint config to SageMaker

These steps are wrapped into higher level abstractions when using the estimator API

Deploying a Model



Request with data

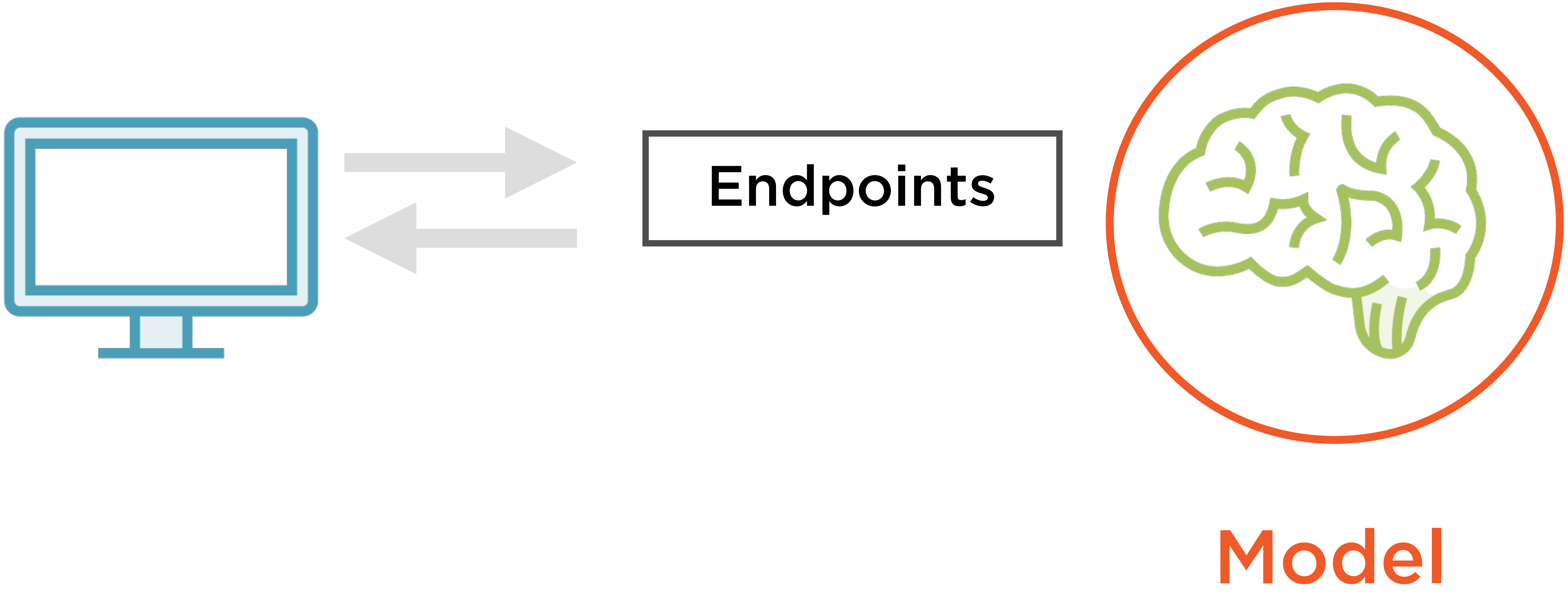
Endpoints

HTTP endpoints for prediction

Prediction response

Model artifacts

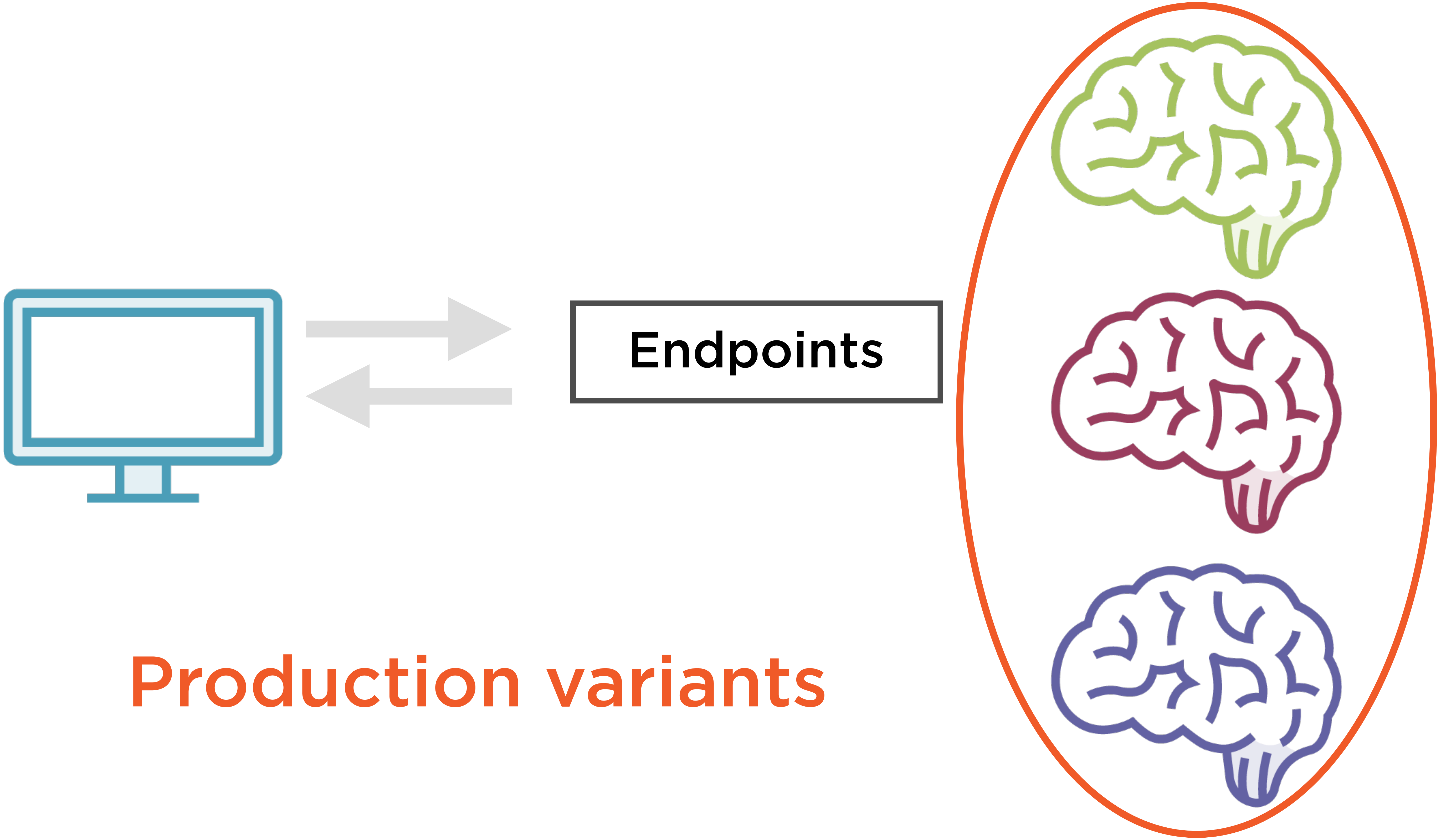
Deploying a Model



Endpoints

Model

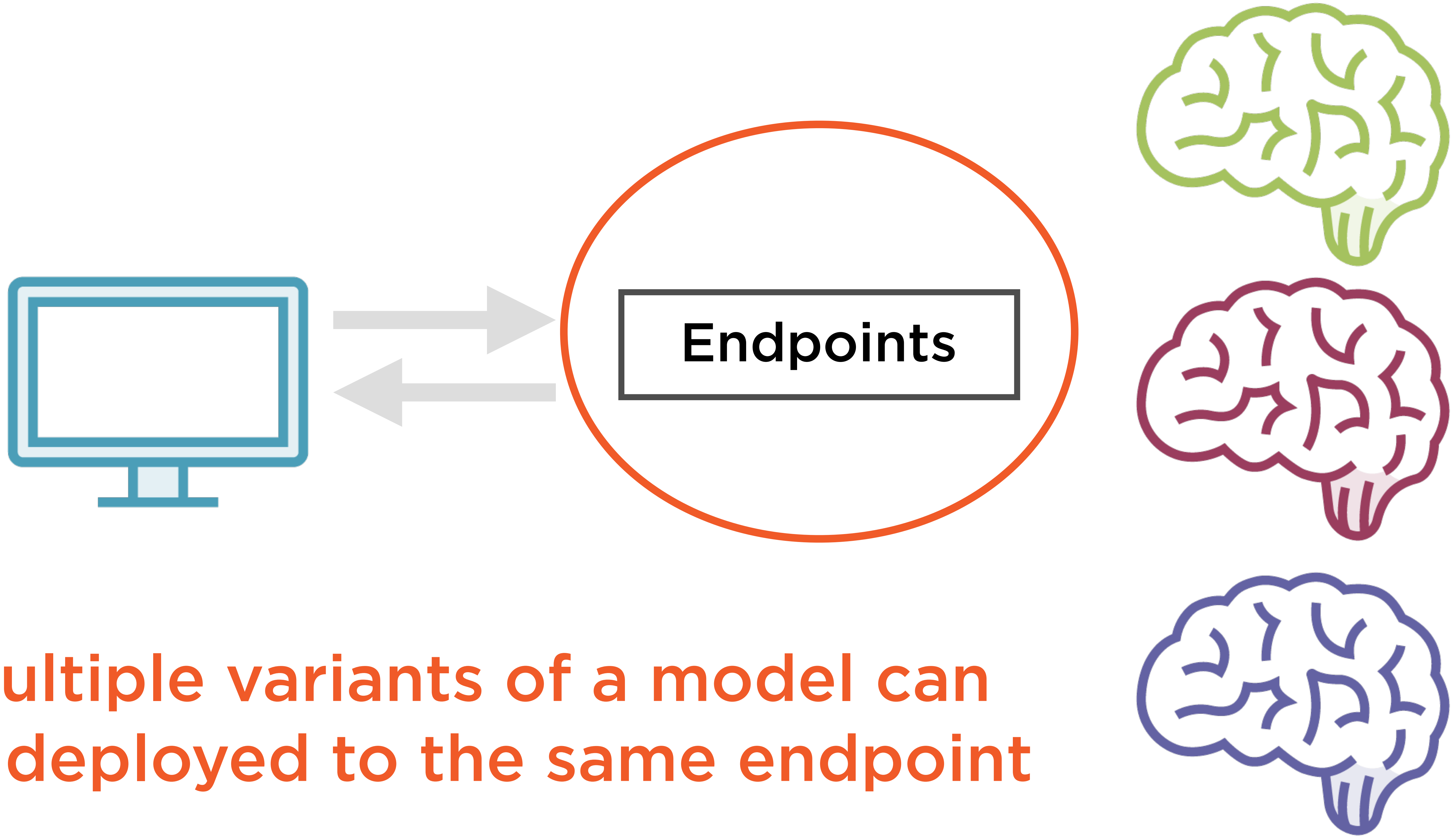
Multiple Variants of a Model



Endpoints

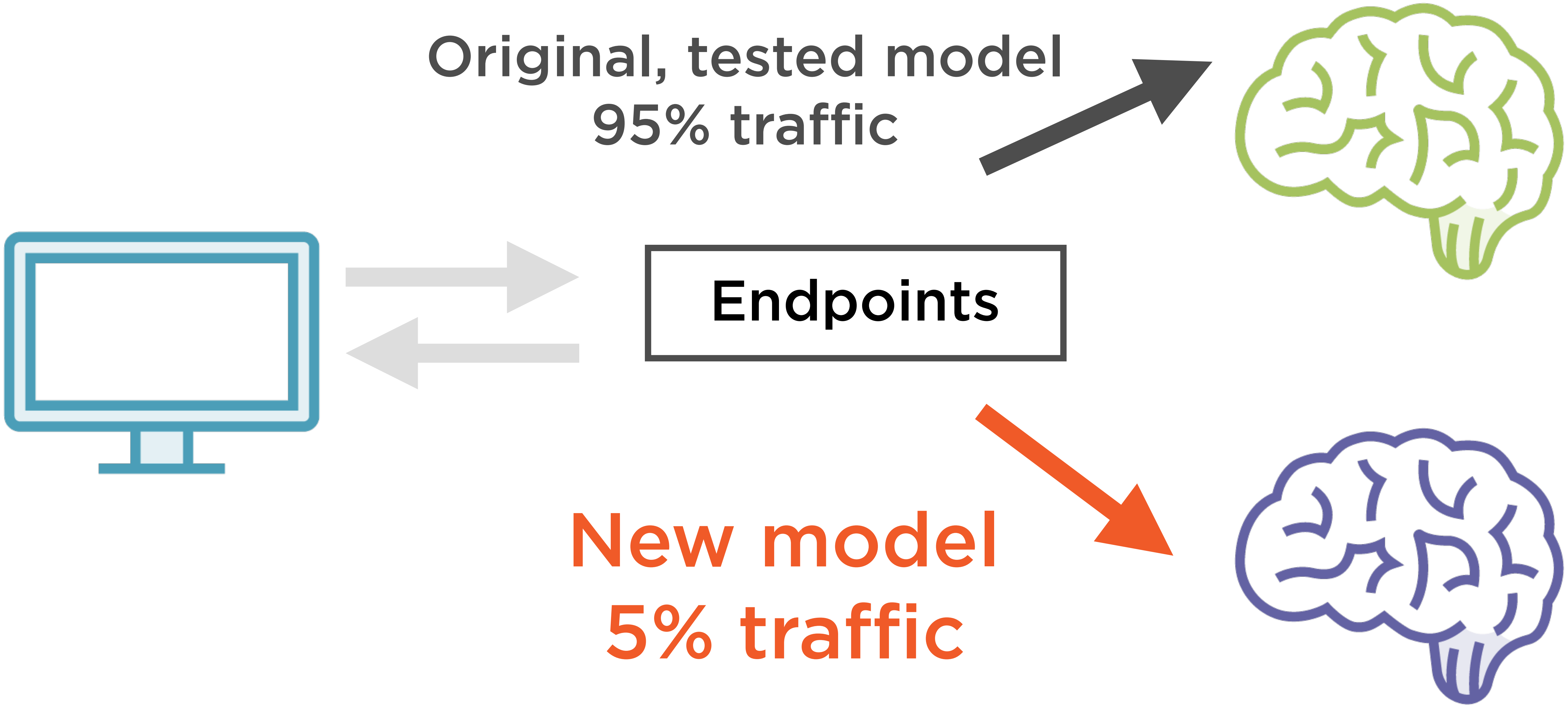
Production variants

Multiple Variants of a Model



Multiple variants of a model can be deployed to the same endpoint

Test Model Variants in Production

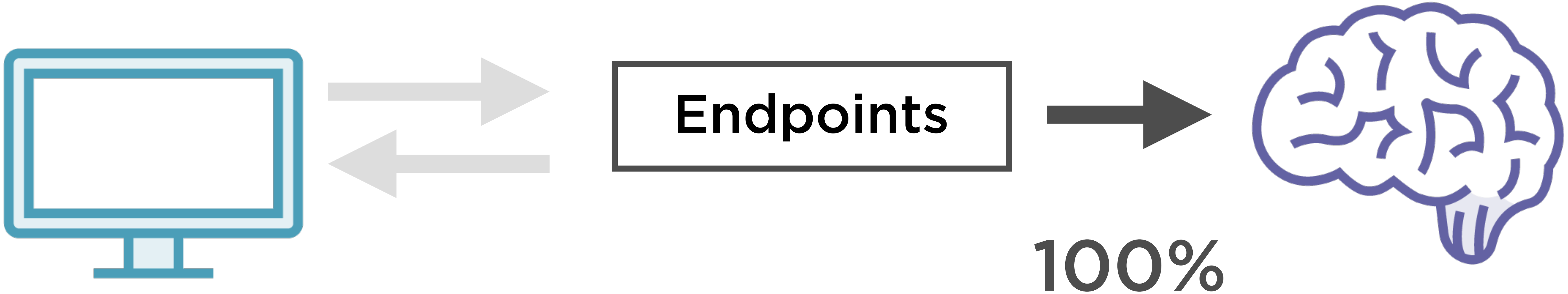


Original, tested model
95% traffic

Endpoints

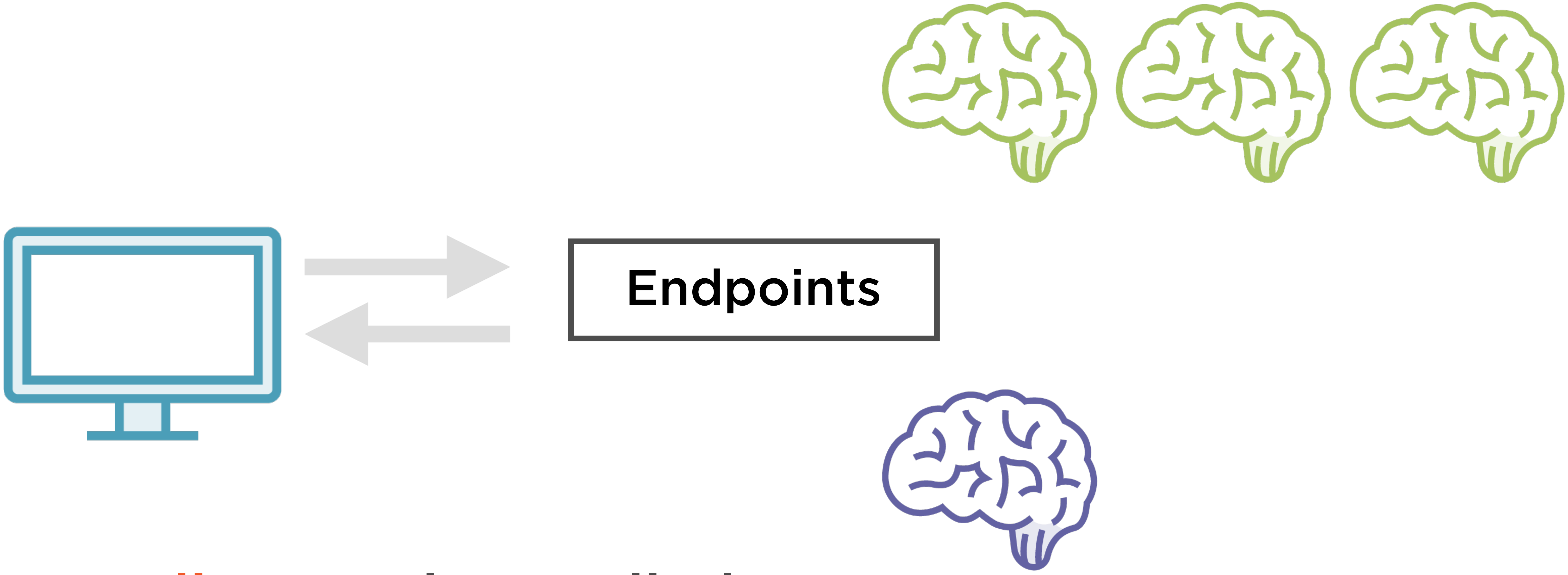
New model
5% traffic

Deploy New Models Without Downtime



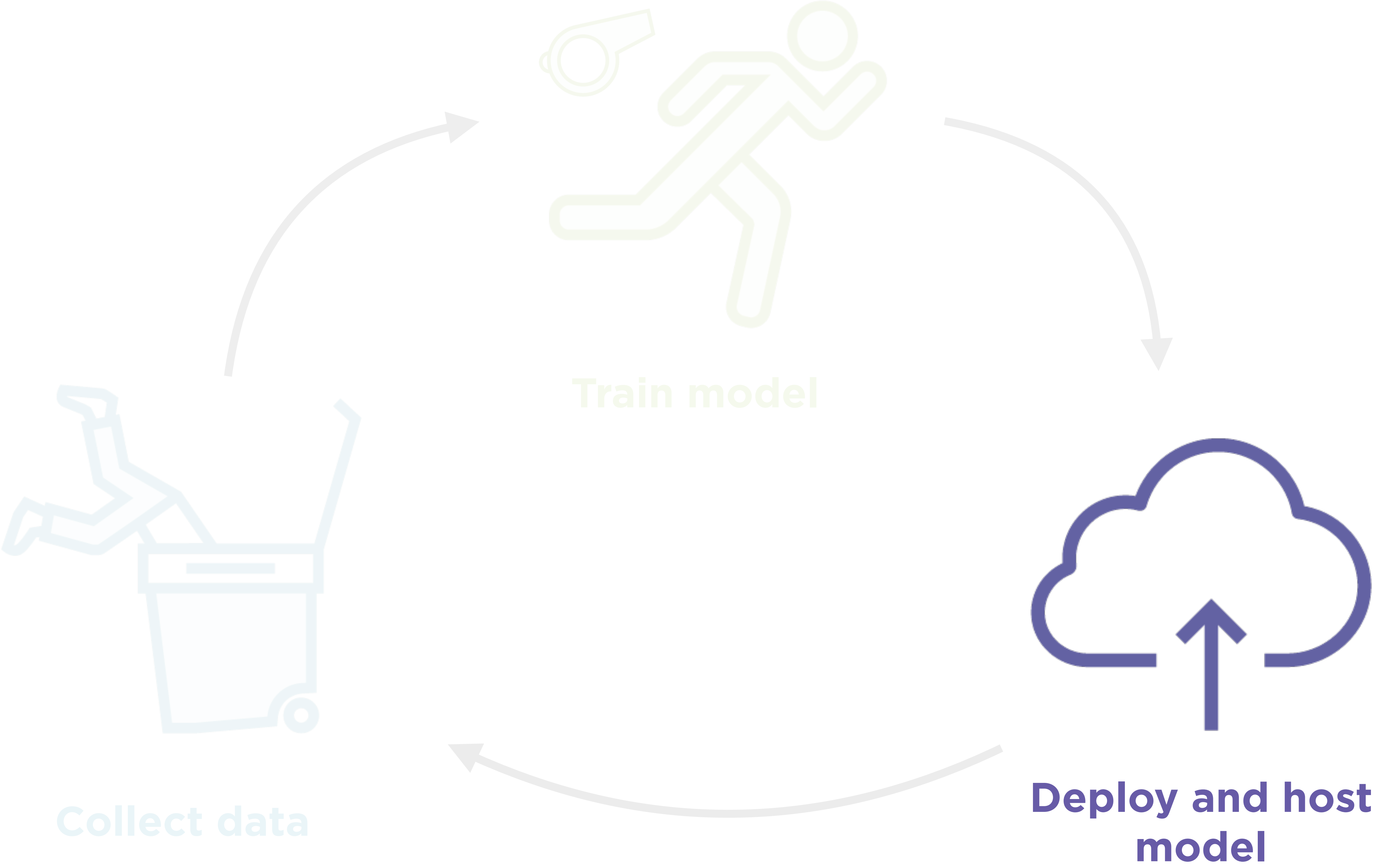
Slowly move 100% of the traffic to the new model

Autoscale Model Variants



Autoscaling can be applied to any production variant

Machine Learning Workflow



Demo

Creating an Amazon SageMaker notebook instance

Demo

Setting up a TensorFlow training script for distributed training

Demo

Performing distributed training using the SageMaker TensorFlow estimator

Demo

Deploying the model for prediction to AWS Hosting Services

Demo

Enabling CloudTrail to track events for auditing and compliance

Summary

The machine learning workflow with Amazon SageMaker

Develop training script for distributed training

Run distributed training for models using high-level estimators

Deploy and host models on Amazon Hosting Services for online prediction

Enable CloudTrail to track events for auditing and compliance

Related Courses

Deploying PyTorch Models in Production: PyTorch Playbook

Scaling scikit-learn Solutions