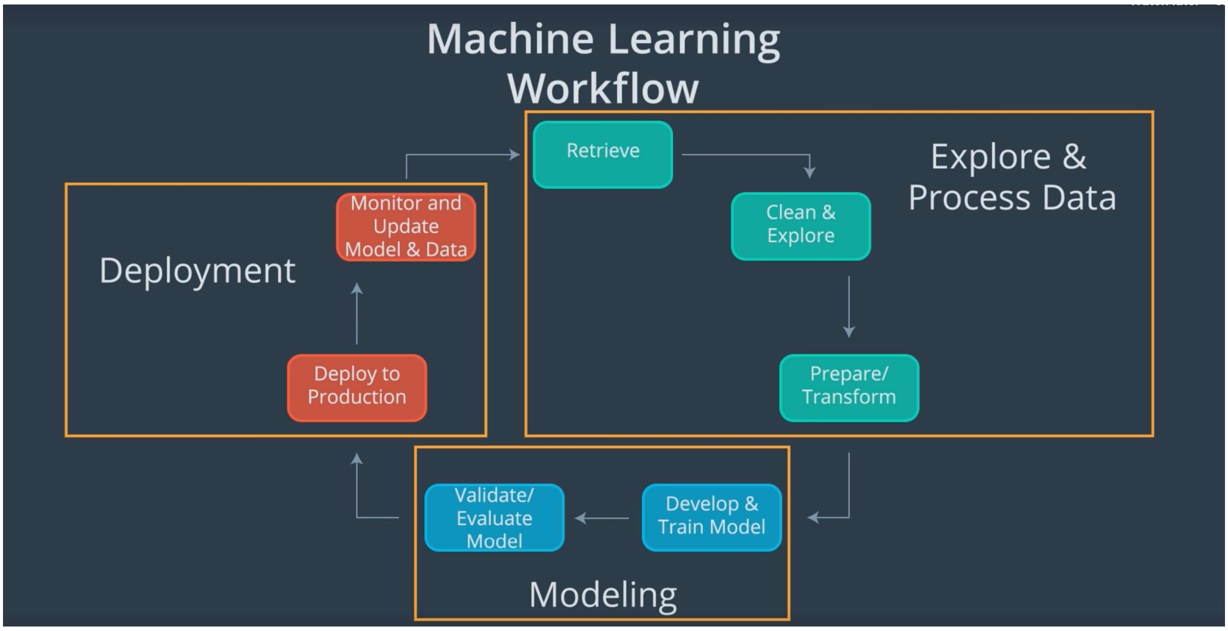
Deploymnet with Amazon Sagemaker

What's the machine learning workflow?



Source : Udacity

1. Retrieve : Download Data or gather
2. Clean and Explore (Outliers, Data that can be used)
3. Prepare and Transform (Data Structure, Normalizing, …) and split
4. Develop Architecture
5. Validate and Evaluate
6. Deploy to Production => App release
7. Monitor and Upload Model and Data => Retrain based on gathered Data (Tesla FSD)

Risks

1. (Potential) Increase in Security Vulnerabilities
2. Reduced Operational Governance Control (over cloud resources)

3. Limited Portability Between Cloud Providers
4. Multi-regional Compliance and Legal Issues

How does deployment fit into the machine learning workflow?

See diagram above.

What is cloud computing?

Cloud Information :

- Large Data Servers by Apple, Google, Amazon, Microsoft, ….

- Provides easier access to data, more resources, more flexibility
- Other Services in the cloud : Apps, Databases, VMs, other services like Sagemaker

Why would we use cloud computing for deploying machine learning models?

Benefits

1. Reduced Investments and Proportional Costs (providing cost reduction)
2. Increased Scalability (providing simplified capacity planning)
3. Increased Availability and Reliability (providing organizational agility)

Why isn't deployment a part of many machine learning curriculums?

Typically because the ML curriculums focus on Exploring and Processing Data as well as building the model. (Analyst approach) Usually the traditional SW developers focused on deployment, but today tools allow for analysts to deploy more easily. (for example no recoding to lower level code)

What does it mean for a model to be deployed?

Paths to Deployment:

1. Python model is recoded into the programming language of the production environment :
   Recode to Java or C++ for example, rarely used anymore
2. Model is coded in Predictive Model Markup Language (PMML) or Portable Format Analytics (PFA) :
   Code model in PMML or PFA standards for deploying to production environment, certain Software provides direct import of PMML, for example : IBM SPSS, Apache Spark, …

3. Python model is converted into a format that can be used in the production environment : -- -- Amazon Sagemaker way :

- Build Python model and use libraries and methods that convert into code used by production environment
- SW Frameworks like Pytorch, sci-kit learn have methods to convert Python to intermediate Standard Format like ONNX (Open Neural Network Exchange)
- ONNX can then be converted to a format suitable for the production environment
- Easiest way to move Python model from production to deployment
- Typically the way it is done nowadays
- Technologies like containers, endpoints and APIs also help to ease the work for deployment

What are the essential characteristics associated with the code of deployed models?

- Code is optimized to work on Production Environment.

What are different cloud computing platforms we might use to deploy our machine learning models?

AWS

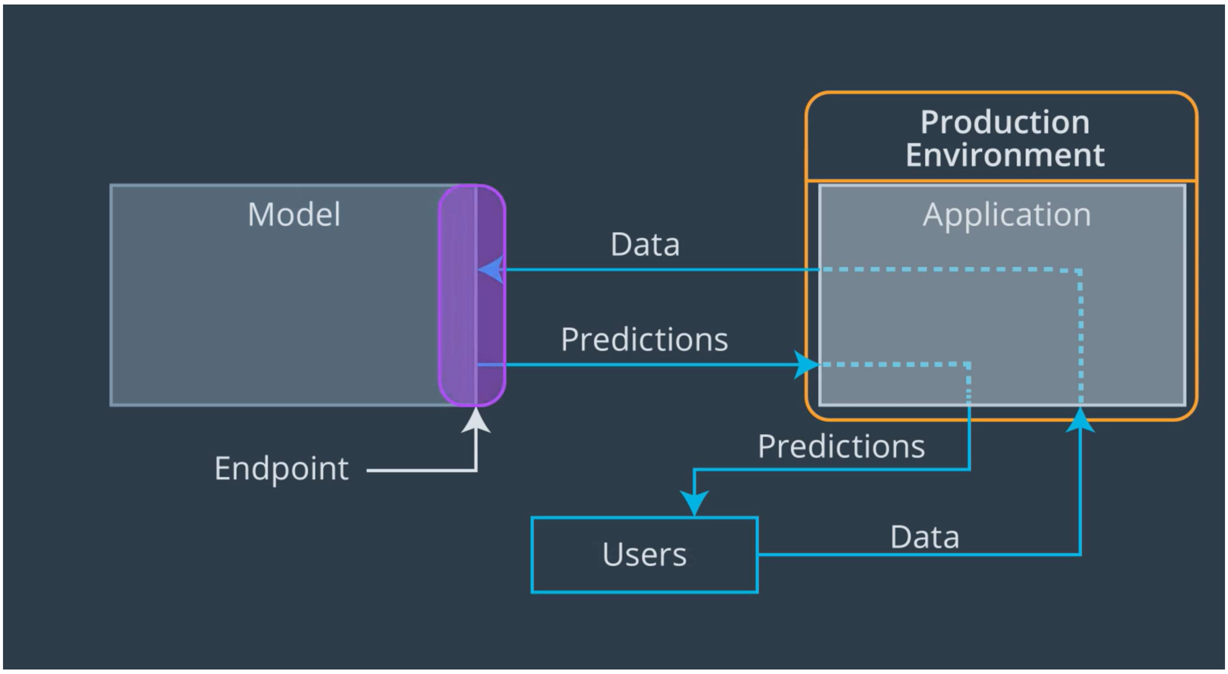
- Flexible in SW and Modeling as well as Deployment , built-in Algorithms, Custom Algorithms, Explore and Process Data

GCP

- Prediction costs, explore and Process data, ML SW, Flexibility in Modeling and Deployment

Others are Microsoft Azure, Paperspace (claims to offer more powerful and cheaper machines than the others), Cloud Foundry and IBM Watson

<u>Model Deployment Overview</u>



Source : Udacity

Endpoint :

- Allows App to send User Data to model
- Receives predictions back from the model and sends it to the App, based on User Data
- the endpoint itself is like a function call
- the function itself would be the model and
- the Python program is the application.

Endpoint and REST API

Communication between the application and the model is done through the endpoint (interface), where the endpoint is an Application Programming Interface (API).

- In this case, our API uses a REpresentational State Transfer, REST, architecture that provides a framework for the set of rules and constraints that must be adhered to for communication between programs.
- This REST API is one that uses HTTP requests and responses to enable communication between the application and the model through the endpoint (interface).
- Request is based on 4 parts : Endpoint, HTTP Method (POST method only in this case) , HTTP Headers (contains info about format, etc.) and the Message Data
- Respone is based on : HTTP Status Code (is reception successful?) , HTTP Headers and Message Data

As we learn more about RESTful API, realize that it's the application's responsibility:

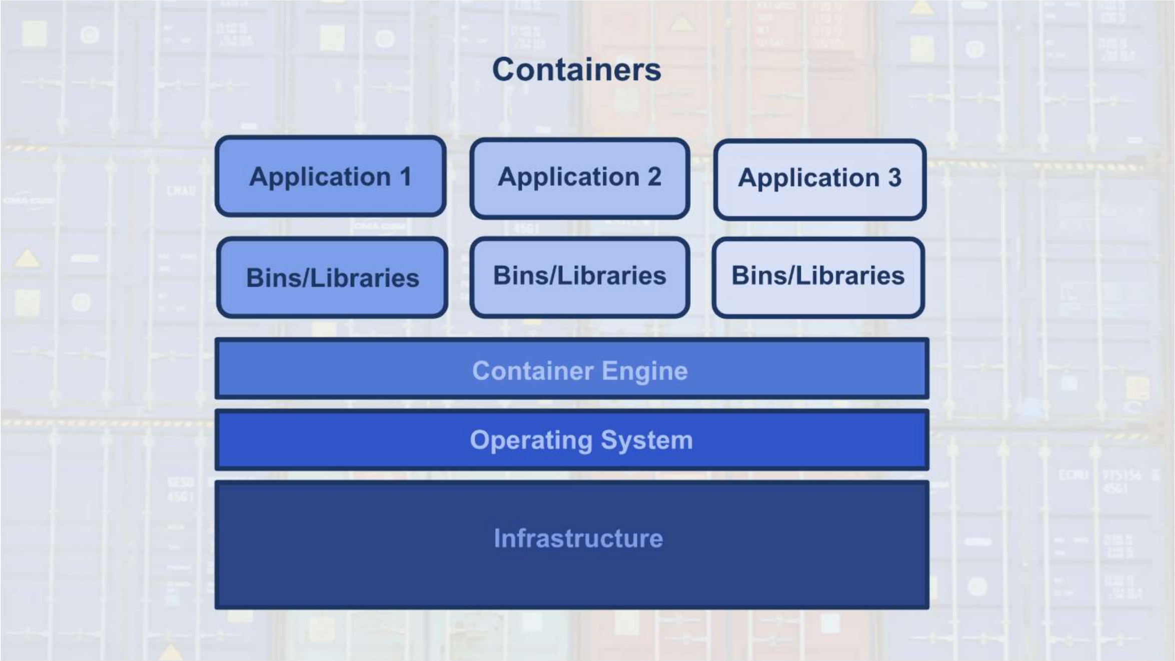
- To format the user's data in a way that can be easily put into the HTTP request message and used by the model
- To translate the predictions from the HTTP response message in a way that's easy for the application user's to understand
- Data and predictions will need to be in JSON or CSV format mostly

Containers

Both the model and the application require a computing environment so that they can be run and available for use. One way to create and maintain these computing environments is through the use of containers.

- Specifically, the model and the application can each be run in a container computing environment. The containers are created using a script that contains instructions on which software packages, libraries, and other computing attributes are needed in order to run a software application, in our case either the model or the application.
- Docker containers are similar to shipping containers :
  + Container can provide different types inside of it while making it easy to track, load and deploy as well as transport

  This architecture of containers provides the following advantages:

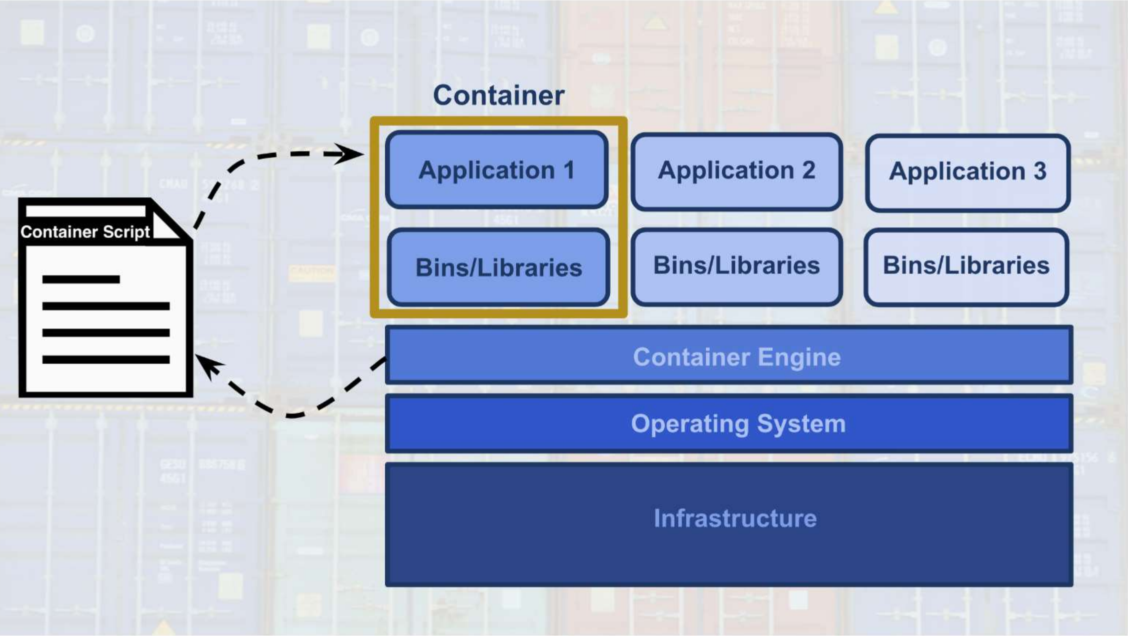


Source : Udacity

1. Isolates the application, which increases security.
2. Requires only software needed to run the application, which uses computational resources more efficiently and allows for faster application deployment.
3. Makes application creation, replication, deletion, and maintenance easier and the same across all applications that are deployed using containers.
4. Provides a more simple and secure way to replicate, save, and share containers.

Based on Container Script :

1. Can be shared for easy replication
2. Algo to create a container, for Docker : dockerfiles



Source : Udacity

Interview with Udacity expert :

- Mentions shipping container analogy
- Not a full VM, lighter than OS, low Overhead
- Udacity Workspaces environments are Containers
- Update build script, makes Container updating very easily and efficient

Characteristics of Deployment and Modeling

Hyperparameters :

- Not learned, must be set by developer
- Important part
- Automatic tuning provided by cloud services, sci-kit learn also offers this

Characteristics of Deployment

1. Model Versioning
2. Model Monitoring
3. Model Update and Routing to compare model versions against each other for Testing
4. Model predictions
5. On-Demand predictions
6. Batch predictions