FourthBrain

MLE Program, Cohort 11 (MLE11)

Week 12: Introduction to MLOps

Last Week!

Concepts

- Encoder and Decoder Networks
- Bidirectional Encoder Representations from Transformers (BERT)
- General Pre-Trained Transformers (GPT-3)
- Fine-Tuning of Pre-Trained Transformers

🤖 This Week!

Concepts

- Introduction to MLOps
- MLOps Level 0: Manual
- Model Registries
- Model Servers
- Prediction Services

Hands on

- VS Code Onramp
- AWS Onramp
- Web App Health Check

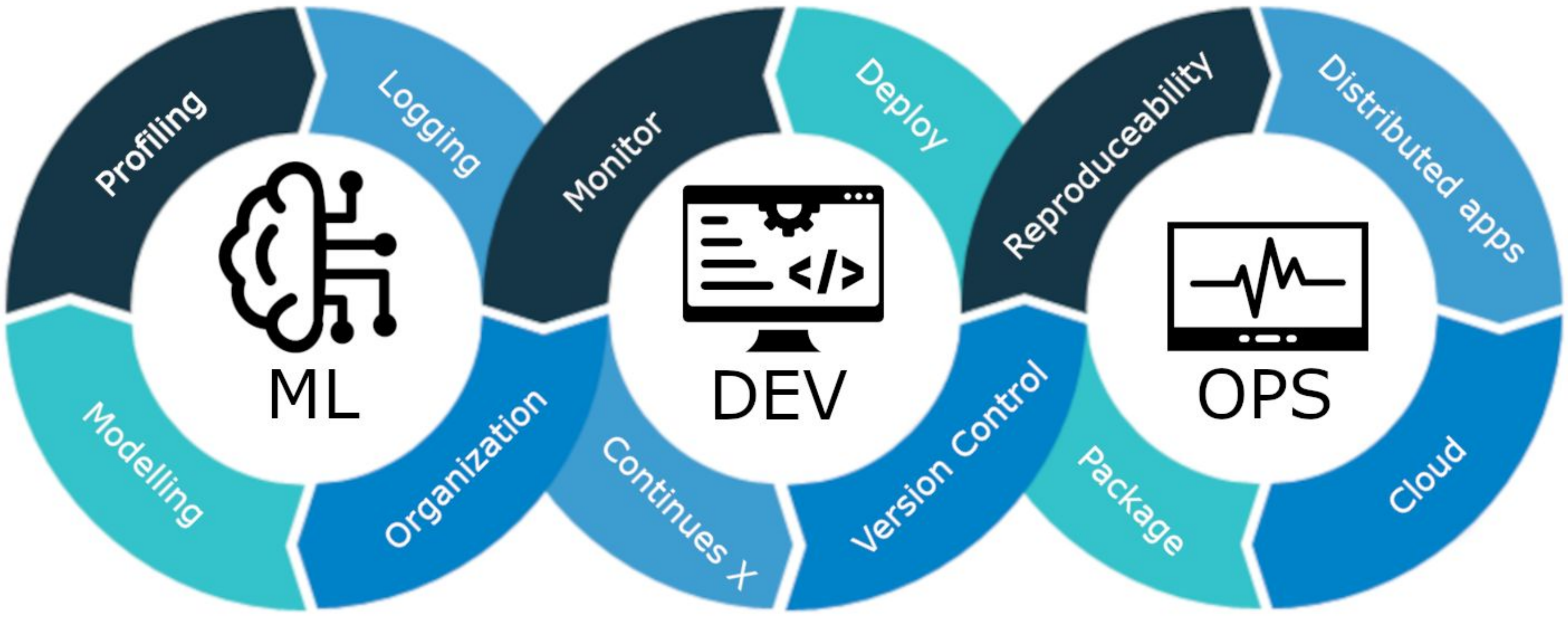
What questions do you have?

Note!

- Your Capstone code should be ready to be deployed as an app on the cloud
- Half of the class will be spent working on Capstones and deployment of your capstones

Introduction to MLOps

MLOps


Profiling
Logging
Monitor
Deploy
Reproduceability
Distributed apps
ML
DEV
OPS
Modelling
Organization
Continues X
Version Control
Package
Cloud

MLOps

- MLOps is a compound of machine learning and operations
- It is a practice for collaboration and communication between data scientists and operations professionals to help manage production ML (or deep learning) lifecycle
- MLOps empowers data scientists and app developers to help bring machine learning models to production

MLOps

- MLOps enables every asset in the ML lifecycle to be:
    - Tracked
    - Versioned
    - Audited
    - Certified
    - Re-used

- It provides orchestration services to streamline managing this lifecycle

MLOps | ModelOps | AIOps

- MLOps and ModelOps are largely being used interchangeably
- ModelOps could be more general than MLOps as it’s not only about machine learning models but any kind of models (i.e. rule-based models)
- AIOps could be more related to AI for DevOps (i.e. predictive maintenance for network failures)

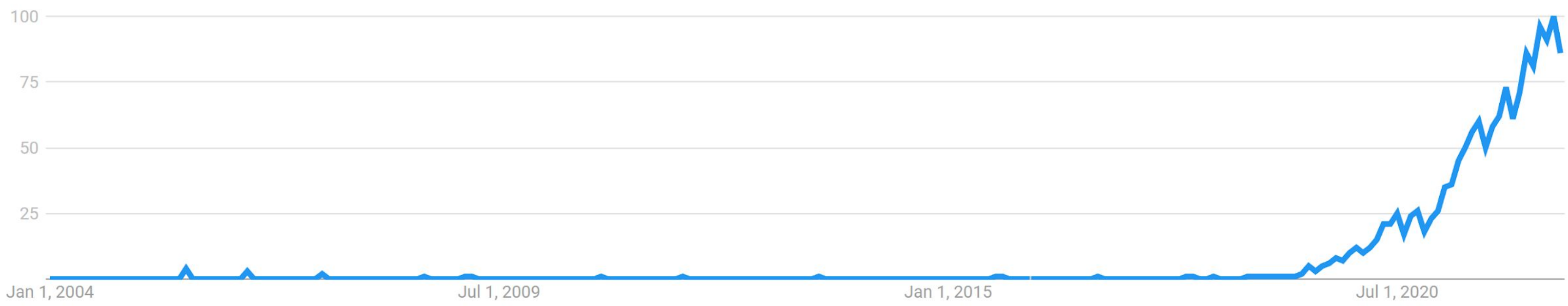
MLOps

- Machine learning Operations (MLOps) is quickly becoming a critical component of successful data science project deployment in the enterprise

- It is a process that helps organizations and business leaders generate long-term value and reduce the risk associated with data science, machine learning, and AI initiatives.

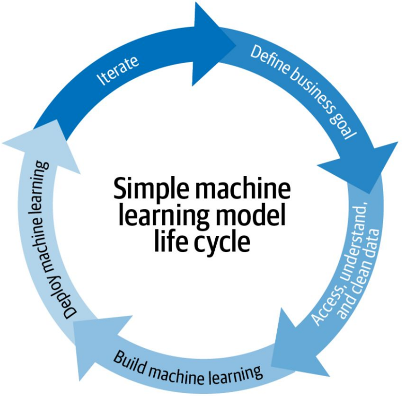
- It is a relatively new concept, but has been skyrocketing into the data science lexicon overnight?

MLOps | Interest over Time



MLOps | Definition

- At its core, MLOps is the standardization and streamlining of machine learning life cycle management
- For most traditional organizations, the development of multiple machine learning models and their deployment in a production environment is relatively new.



MLOps | Definition

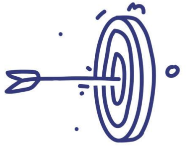
- Until recently, the number of models may have been manageable at a small scale, or there was simply less interest in understanding these models and their dependencies at a company-wide level.

- With decision automation models become more critical, and, in parallel, managing model risks becomes more important at the top level.

Group Discussion

What challenges associated with MLOps can you think of? How could they be resolved?

MLOps | Challenges



- There are many dependencies in MLOps
- Data is constantly changing and Business needs shift as well.
- Results need to be continually relayed back to the business to ensure that the reality of the model in production and on production data:
    - Aligns with expectations
    - Addresses the original problem or meets the original goal

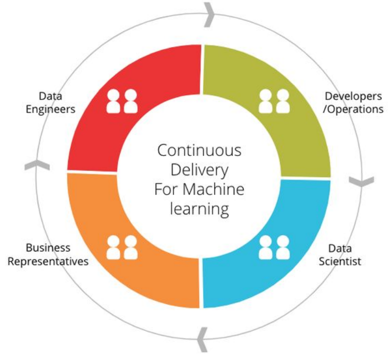
MLOps | Challenges

- Not everyone speaks the same language.

- Even though the machine learning life cycle involves people from the business, data science, and IT teams, none of these groups are using the same tools

- Each of the above-mentioned teams share different fundamental skills to serve as a baseline of communication

MLOps | Concepts



- Robust automation and trust between teams
- The idea of collaboration and increased communication between teams
- The end-to-end service life cycle (build, test, release)
- Prioritizing continuous delivery and high quality

MLOps | Mitigating Risk

- MLOps is important to any team that has even one model in production
- Depending on the model, continuous performance monitoring and adjusting are essential
- By allowing safe and reliable operations, MLOps is key in mitigating the risks induced by the use of ML models
- MLOps practices do come at a cost – a proper cost-benefit evaluation should be performed for each use case

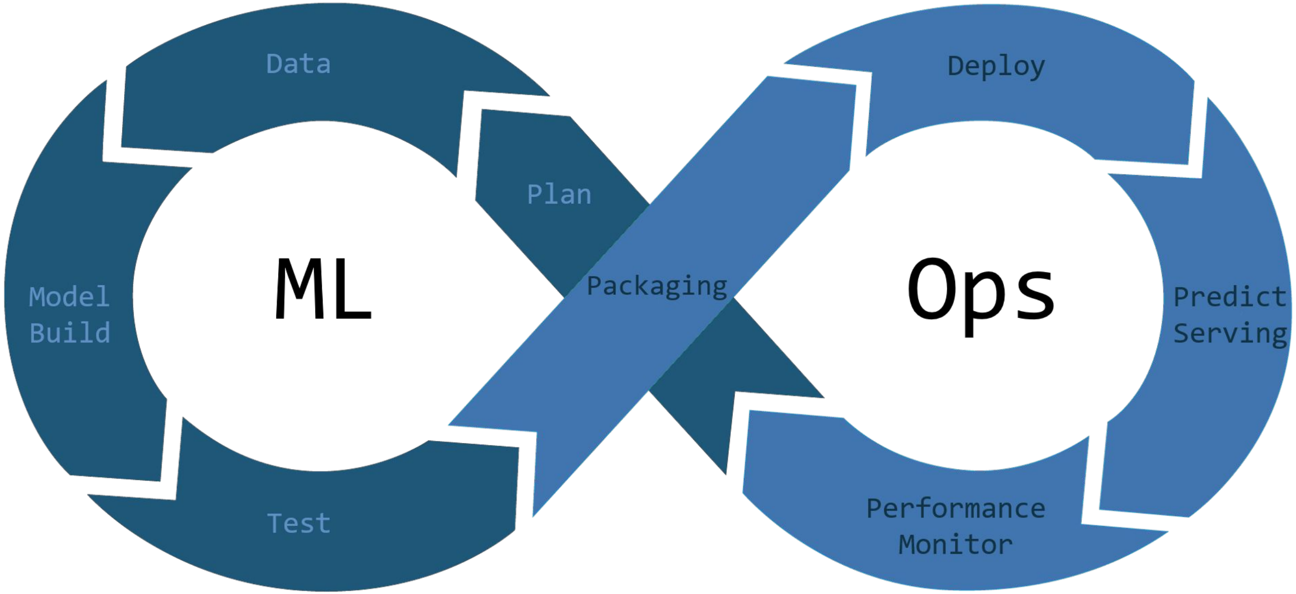
MLOps | Mitigating Risk

- When looking at MLOps as a way to mitigate the risks of a ML model, an analysis should cover:
    - The risk that the model is unavailable for a given period of time
    - The risk that the model returns a bad prediction for a given sample
    - The risk that the model accuracy or fairness decreases over time
    - The risk that the skills necessary to maintain the model (i.e., data science talent) are lost

MLOps | Key Benefits

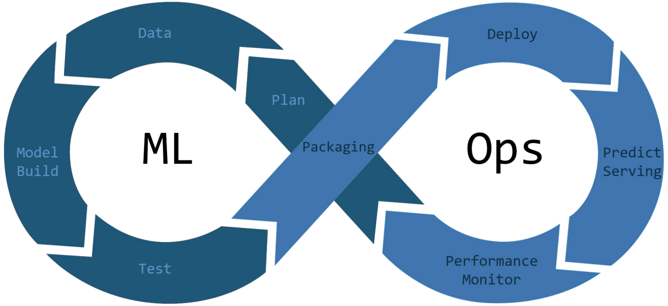
- Keep track of versioning, especially with experiments in the design phase
- Understand whether retrained models are better than the previous versions (and promote models that are performing better to production)
- Ensure (at defined periods—daily, monthly, etc.) that model performance is not degrading in production

MLOps | Cycle


Data
Plan
ML
Model
Build
Test
Packaging
Deploy
Ops
Predict
Serving
Performance
Monitor

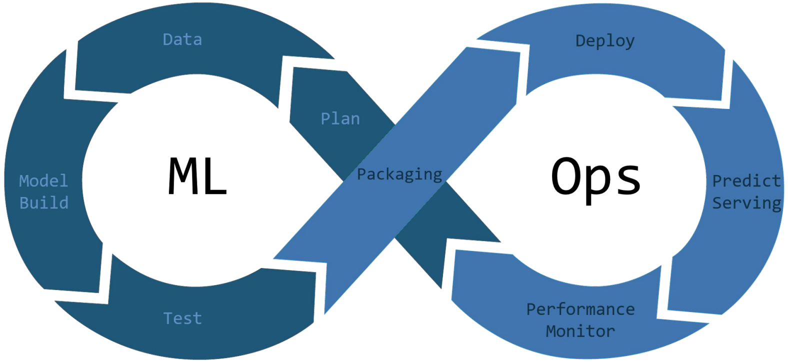
MLOps | Key Features

- Model Development
- Productionalization and Deployment
- Monitoring
- Iteration and Life Cycle
- Governance



MLOps | Model Development

- Establishing Business Objectives
- Data Sources and Exploratory Data Analysis
- Feature Engineering and Selection
- Training and Evaluation
- Reproducibility
- Responsible AI



MLOps | Establishing Business Objectives

- The process of developing a machine learning model typically starts with a business objective
- It can be as simple as reducing fraudulent transactions to $< 0.1\%$ or having to identify people’s faces on their social media photos
- Business objectives (that can be captured as KPIs):
    - Performance targets
    - Technical infrastructure requirements
    - Cost Constraints

MLOps | Data Sources and Exploratory Data Analysis

- Key questions for finding data to build ML models include:
  - What relevant datasets are available?
  - Is this data sufficiently accurate and reliable?
  - How can stakeholders get access to this data?
  - What data properties – features – can be made available by combining multiple sources of data
  - Will this data be available in real-time?
  - What platform should be used?
  - How will the data be updated once the model is deployed?

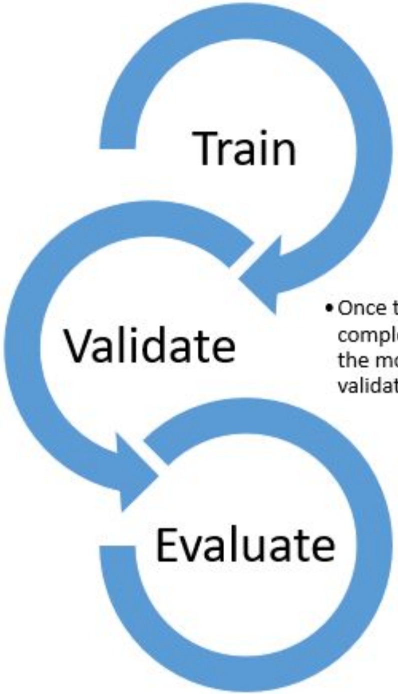
MLOps | Data Sources and Exploratory Data Analysis

- Key questions regarding data governance constraints
  - Can the selected datasets be used for this purpose?
  - What are the terms of use?
  - Is there personally identifiable information (PII) that must be redacted or anonymized?
  - Are there features, such as gender, that legally cannot be used in this business context?
  - Are minority populations sufficiently well represented that the model has equivalent performances on each group?

MLOps | Feature Engineering and Selection

- Exploratory Data Analysis leads naturally into feature engineering and feature selection
- Feature engineering is the process of taking raw data from the selected datasets and transforming it into “features” that better represent the underlying problem to be solved
- “Features” are arrays of numbers of fixed size, as it is the only object that ML algorithms understand
- Feature engineering includes data cleansing, which can represent the largest part of an ML project in terms of time spent

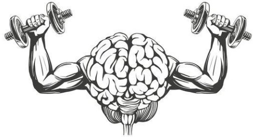
MLOps | Training and Evaluation



- Train different models with training dataset.
- Once training is completed, validate the models using validation dataset.
- At these phase, one must identify which model had a least error rate and high approximate rate.

MLOps | Training and Evaluation

- The process of training and optimizing a new ML model is iterative
- Several algorithms may be tested
- Features can be automatically generated
- Feature selections may be adapted
- Algorithm hyperparameters tuned
- Training is the most intensive step of the ML model life cycle when it comes to computing power



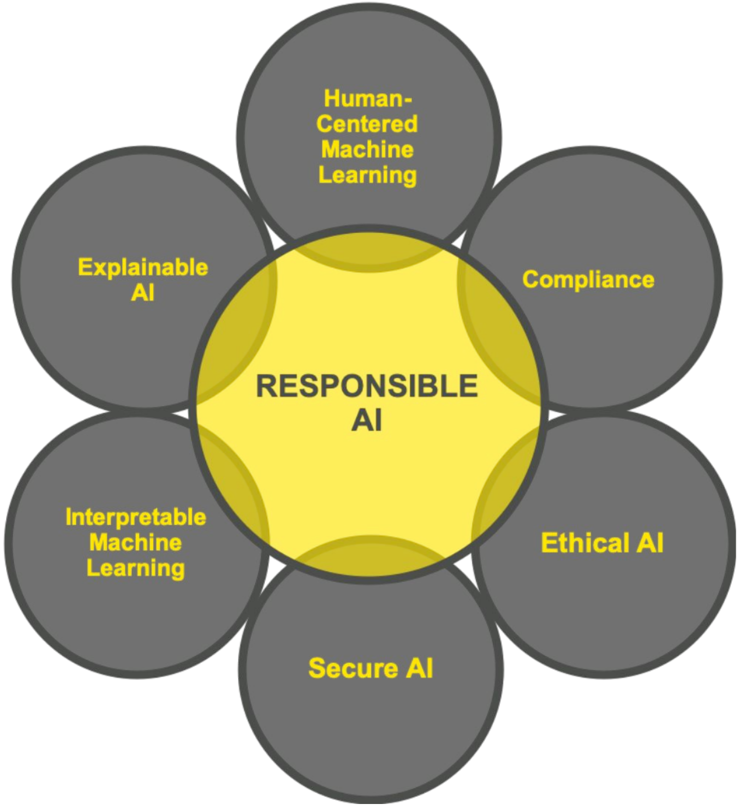
MLOps | Training and Evaluation

- Keeping track of the results of each experiment when iterating becomes complex quickly
- An experiment tracking tool can greatly simplify the process of remembering the data, the features selection process, and model parameters alongside the performance metrics
- These enable experiments to be compared side-by-side, highlighting the differences in performance

MLOps | Reproducibility

- While many experiments may be short-lived, significant versions of a model need to be saved for possible later use.
- The challenge is reproducibility, which is an important concept in experimental science in general.
- The aim of ML is to save enough information about the environment the model was developed in so that the model can be reproduced with the same results from scratch.

MLOps | Responsible AI



MLOps | Responsible AI

- Explainability techniques are becoming increasingly important as global concerns grow about the impact of unbridled AI

- The techniques most commonly used today include:
    - Partial dependence plots, which look at the marginal impact of features on the predicted outcome
    - Subpopulation analyses, which look at how the model treats specific subpopulations and that are the basis of many fairness analyses

MLOps | Responsible AI

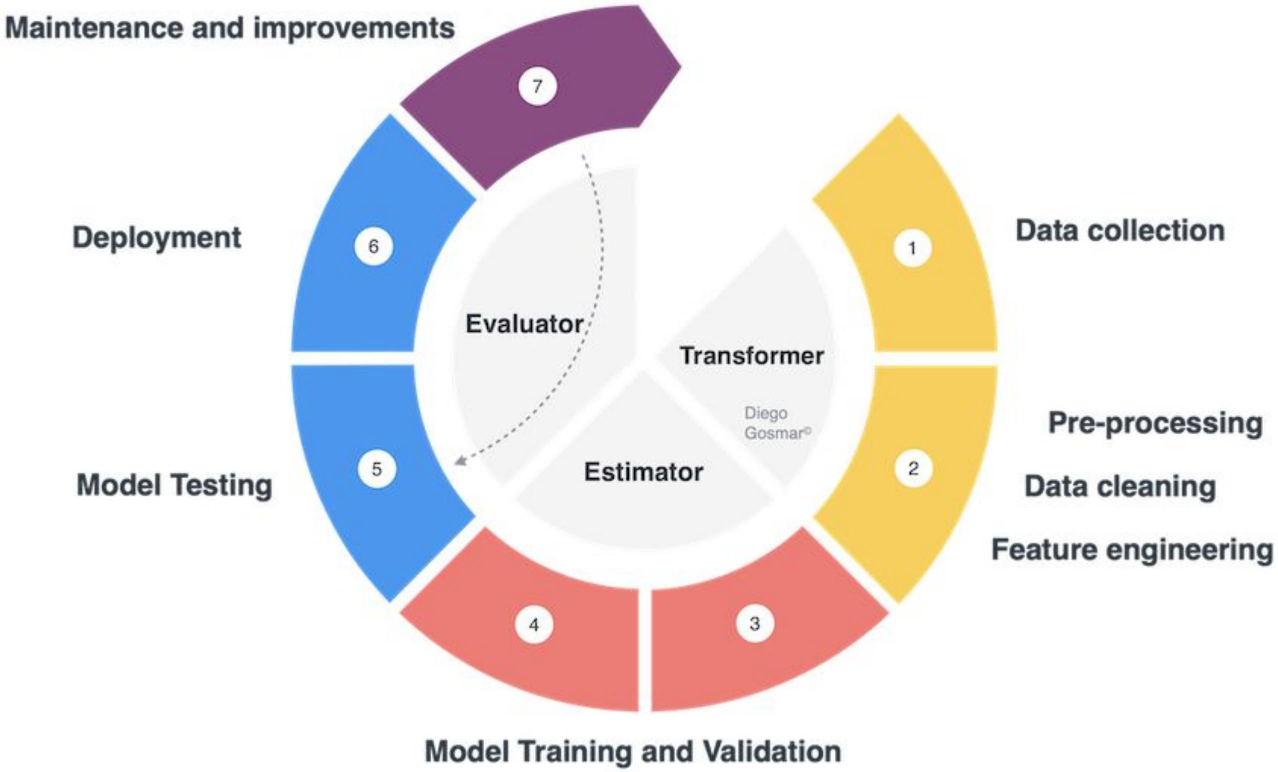
- Explainability techniques are becoming increasingly important as global concerns grow about the impact of unbridled AI

- The techniques most commonly used today include:

    - Individual model predictions, such as Shapley values, which explain how the value of each feature contributes to a specific prediction

    - What-if analysis, which helps the ML model user to understand the sensitivity of the prediction to its inputs

MLOps | Model Lifecycle



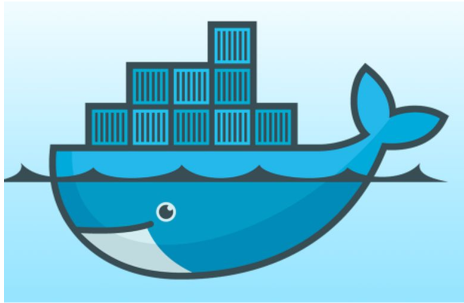
MLOps | Productionalization and Deployment

- Productionalizing and deploying models is a key component of MLOps that presents an entirely different set of technical challenges than developing the model.
- It is the domain of the software engineer and the DevOps team, and the organizational challenges in managing the information exchange between the data scientists and these teams must not be underestimated.
- Without effective collaboration between the teams, delays or failures to deploy are inevitable

MLOps | Model Deployment

- <u>**Containerization**</u> is an increasingly popular solution to the headaches of dependencies when deploying ML models.

- Container technologies such as Docker are lightweight alternatives to virtual machines, allowing applications to be deployed in independent, self-contained environments, matching the exact requirements of each model.



MLOps | Monitoring

- Once a model is deployed to production, it is crucial that it continue to perform well over time

- Good performance means different things to:
  - DevOps team
  - Data Scientists
  - Business

- Scalability of the compute resources can be an important consideration if you are retraining models in production

MLOps | Iteration and Life Cycle

- Developing and deploying improved versions of a model is an essential part of the MLOps life cycle

- There are various reasons to develop a new model version:
    - Model performance degradation due to model drift
    - Need to reflect refined business objectives and KPIs
    - Data scientists have come up with a better way to design the model

MLOps | Iteration and Life Cycle

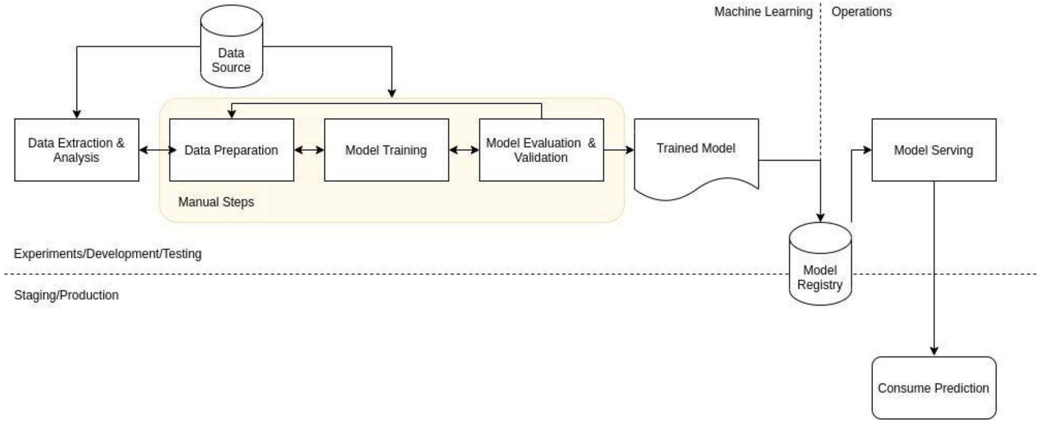
- In some fast-moving business environments, new training data becomes available every day.
- Daily retraining and redeployment of the model are often automated to ensure that the model reflects recent experience as closely as possible.

MLOps Level 0: Manual

MLOps – Level 0

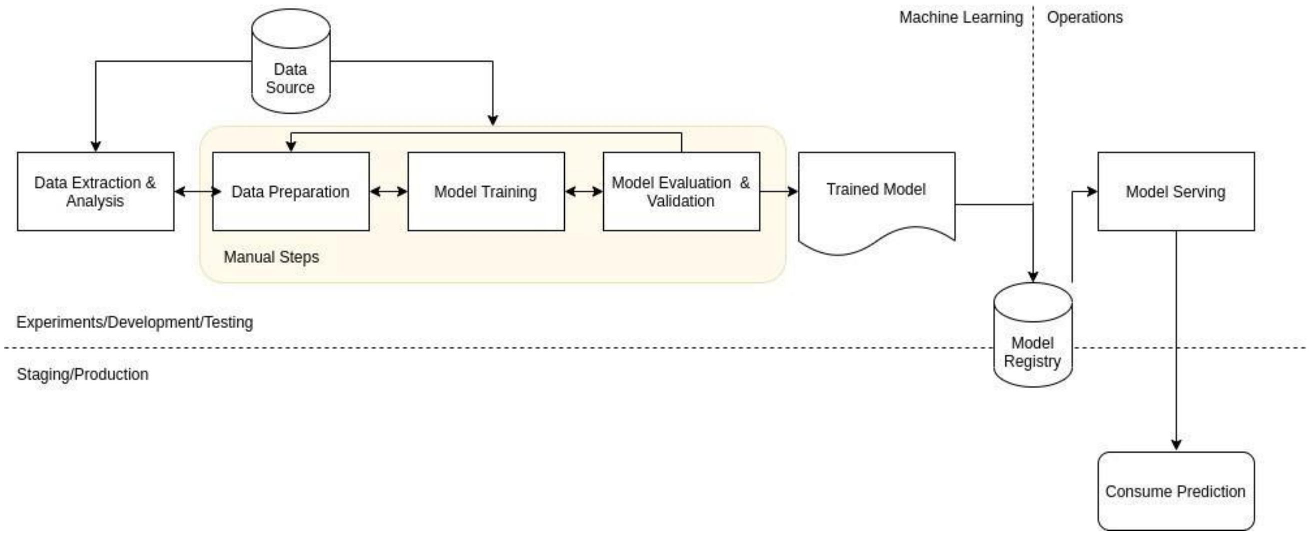
- The level of automation of the MLOps steps determines how mature the Machine Learning process is

- The first level (level 0) of MLOps is the basic level of maturity.



MLOps – Level 0 | Characteristics

- Every step in the workflow is manual.
- Typically work is done in notebooks such as Jupyter / JupyterLab / Zeppelin, and the code is still considered experimental.



MLOps – Level 0 | Characteristics

- The machine learning and operations component of the machine learning system is disconnected
- Data scientists will typically do all the work:
    - Data sourcing
    - Data extractions
    - Data Analysis
    - Data preparation
    - Model Training
    - Model Evaluation
    - Model Validation
    - Model Registry
    - Deploying the model with low latency serving

MLOps – Level 0 | Characteristics

- Model Releases are infrequent
- Continuous Integration of code is non-existent
- Testing is done inside notebooks or during the execution of scripts
- The code for training and visualization will typically be source controlled
- Continuous deployment is non-existent
- Deployment of this workflow is all about getting the model into a prediction service (i.e. REST API)
- No performance monitoring/tracking (which leads to difficulties to determine if a model has degraded and a re-training process must be done)

MLOps – Level 0 | Challenges Solutions

- Monitor the quality of the model in production
  - Detect model performance degradation and model staleness
  - Determine when the re-training process is needed
- Frequent re-training of the models
  - Data changes over time (with high velocity)
  - The model in production needs to be trained with the most recent data seen in production
- Continuous experimentation
  - Try different feature engineering variables, model architectures and hyperparameters

AWS and preparing for the assignment

Group Discussion

5 min

How does API work? Provide some examples.

What is FastAPI?

App Structure

- How should we structure the code?
  - main.py, model.py and any other supplementary directories
- Capstone code structure
  - one repository for all of the code
  - directory structure:
    - data, images, source code, dvc, model monitoring...
  - multiple repositories for the code
    - each part comes with its own directory

Cloud and Deployment

- Running app locally during the development of the app
- FourthBrain mainly uses AWS for our assignments
- AWS EC2:
  - Elastic Compute Cloud
  - Various virtual environments names “instances”
  - Maximum customization capabilities - from security, memory, storage, CPU power

VSCode as our main IDE going forward!

Demo

AWS login

EC2 setup

VSCode setup

VSCode to AWS

Hello World Fast API deployment

Prediction Services

Prediction Services

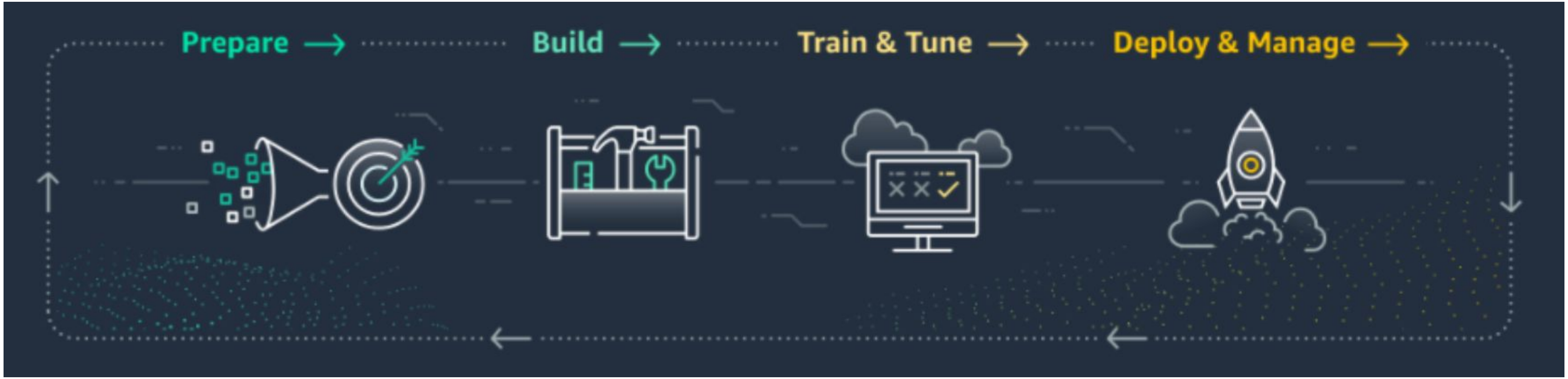
Prediction Services

- Amazon SageMaker could be a good example for prediction Services

- It is a ML service enabling data scientists, data engineers, MLOps engineers, and business analysts to build, train, and deploy ML models for any use case,

- It does not require heavy ML expertise

Prediction Services | AWS SageMaker



Prediction Services | AWS

- **Amazon Rekognition** – Computer Vision
    - Analyze Images and Videos
    - Catalog assets
    - Automate workflows
    - Extract meaning from media and applications

- **Amazon Lookout for Vision** – Detect defects and automate inspection
    - Identify missing product components
    - Identify vehicle and structure damage
    - Identify irregularities for comprehensive quality control

Prediction Services | AWS

- **AWS Panorama** – Utilize computer vision at the edge
  - Improve operations with automates monitoring
  - Find bottlenecks and assess manufacturing quality and safety
- **Amazon Textract** – Extract Text and Data
  - Pull valuable information from millions of documents at speed
- **Amazon Comprehend** – Acquire Insights
  - Maximize the value of unstructured text with NLP

Prediction Services | **AWS**

- **Amazon A2I** – Control Quality
  - Add humans to the review process to ensure accuracy and compliance of sensitive data

- **Amazon Lex** – Build chatbots and Virtual agents
  - Create automated conversation channels to improve customer service

- **Amazon Transcribe** – Automate speech recognition
  - Enhance applications and workflows with automatic speech recognition

Prediction Services | AWS

- **Amazon Polly** – Give your apps a voice
  - Convert text into life-like speech
  - Improve user experience and accessibility
- **Amazon Kendra** – Find accurate information Faster
  - Enhance websites and applications with Natural Language speech
  - Help users quickly search for what they need

Prediction Services | AWS

- **Amazon Personalize** – Personalize online experiences
  - Use ML to customize applications and websites to each individual user
- **Amazon Translate** – Engage audiences in every language
  - Expand your reach and accessibility with
    - Fast translation
    - Accurate translation
    - Customizable translation

Prediction Services | AWS

- **Amazon Forecast** – Forecast business metrics
  - Harness unique data types and time series data to create accurate end-to-end prediction models
- **Amazon Fraud Detector** – Detect online fraud
  - Stop adversaries and identify potential attacks with technology honed through years of use on amazon.com

Prediction Services | AWS

- **Amazon Lookout for Metrics** – Identify data anomalies
  - Detect and identify root causes of unexpected changes in metrics such as revenue and retention
- **Amazon DevOps Guru** – Improve application availability
  - Simplify operational performance measurement and reduce application downtime

Prediction Services | AWS

- **Amazon CodeGuru Reviewer** – Automated code reviews
    - Detect bugs and assess critical issues and vulnerabilities fast for higher quality code
- **Amazon CodeGuru Profiler** – Eliminate costly inefficient code
    - Use runtime behavior analysis to improve application performance and decrease compute costs

Reminder

- Code Freeze:
  - March 26th
  - after that all of your code should be done, and only minor deployment and documentation should be worked on
  - **If this is not possible - you MUST let us know**
  - **Cramming will not be possible.**
  - April 4th

What questions do you have?

Feedback on Lecture and Concepts?

See you next week!