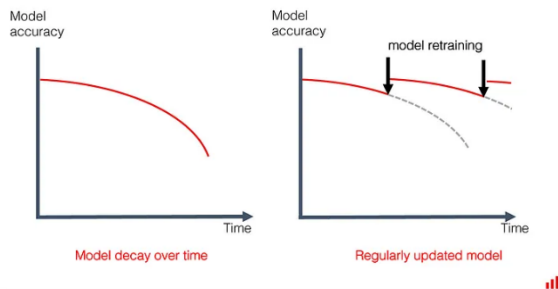


1주차

내가 생각하는 MLOps

: DevOps + ML → 모델을 이용한 서비스가 안정적인 성능으로 운영되도록 전반 환경 구축&관리



Common deployment cases

1. New product/capability
2. Automate/assist with manual task
3. Replace previous ML system

Key ideas:

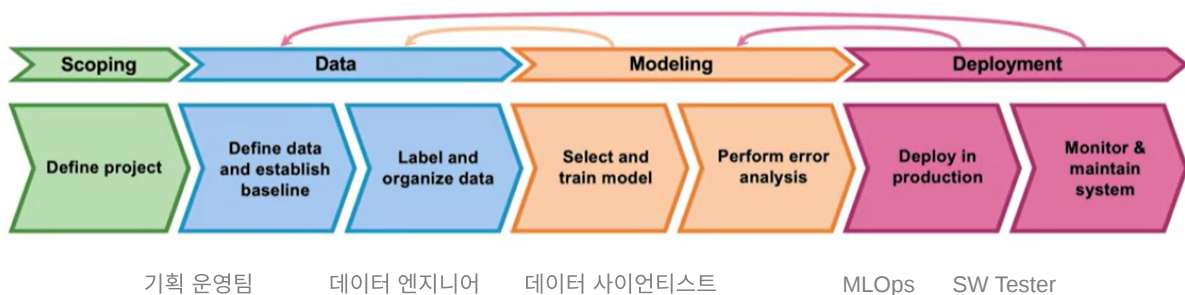
- Gradual ramp up with monitoring
- Rollback

그러기 위해 필요한 것

: 기획, 개발, 배포, 테스트, 운영, 모니터링, 유지보수

→ iterative process

The ML project lifecycle



Monitor

- SW Metrics: Memory, compute, latency, throughput, server load
- Input Metrics: Avg input len, Avg input vol, Num missing Values, Avg img brightness
- Output Metrics: time-null return, search, switches to typing

MLOps 업무 내용

- 데이터처리 파이프라인
- 자동화 CI/CD/CT 파이프라인 설계/개발/배포/운영
- ML 모델 추론/서빙 인프라 구축 및 운영
- 모델 실험, 관리, 모니터링 시스템 개발 및 운영
- 대규모 분산 학습/추론 프레임워크 설계 및 개발
- **Keywords** → 자동화, 비용 최적화


MLOps 기술 스택

- 클라우드: **AWS/GCP...**
- 컨테이너: **kubernetes**(with argoCD), **docker** ...
- 서버: kafka(데이터, 웹)/**Jenkins**(CI/CD)
- 공통: **python, git**
- 기타
 - kserve(모델 서비스를 위한 API 생성 툴)
 - bentoML(ML 배포 관리 도구)

ML serving using either kserve seldon or bentoml

What is better kserve, seldon core or bentoml ? and what are the advantages /disadvantages and feature of each one

Did a lot of research and can't find a clear answer

 <https://stackoverflow.com/questions/74232893/ml-serving-using-either-kserve-seldon-or-bentoml>



- Airflow(데이터 파이프라인 관리 및 오케스트레이션)
- gitlab(DevSecOps platform)
- shell script, YAML
- **Keywords**
 - → 클라우드 기반 대용량 실시간 서비스
 - → 분산 시스템
 - → 컨테이너(오케스트레이션)
 - → 새로운 프레임워크 적응에 능숙해야함

MLOps 프로젝트

- 운영 효율화 환경 구축
- 이상탐지 수행 파이프라인 구축
- 데이터~ 효율 파이프라인 구축

[https://fastcampus.co.kr/data_online_mlospj?
utm_source=google&utm_medium=cpc&utm_campaign=fassker^231016^233813&utm_content=mlops
강의&utm_term=&gad_source=1&gclid=Cj0KCQiAy9msBhD0ARIsANbk0A-
42e8kF2jgaxQUdv7xDvNGXMiFOGkz3tKo5_N3oiOZNcT-KhDTCxoaArLFEALw_wcB](https://fastcampus.co.kr/data_online_mlospj?utm_source=google&utm_medium=cpc&utm_campaign=fassker^231016^233813&utm_content=mlops강의&utm_term=&gad_source=1&gclid=Cj0KCQiAy9msBhD0ARIsANbk0A-42e8kF2jgaxQUdv7xDvNGXMiFOGkz3tKo5_N3oiOZNcT-KhDTCxoaArLFEALw_wcB)