

Let's go MLOps!



Be a MLOps Specialist!

C1_W1: Overview of the ML Lifecycle and Deployment

MLOps 이영전



본 PPT 슬라이드 제작에 사용된 PPT 자료들은 Coursera MLOps Specialist Course1강의를 참고했습니다.

- Introduction to Machine Learning in Production:
<https://www.coursera.org/learn/introduction-to-machine-learning-in-production>

Goal

- 01 Identity the Key Components of the ML Life Cycle
- 02 Define ‘concept drift’ as it relates to ML projects
- 03 Differentiate between shadow deployment, canary deployment and blue-green deployment
- 04 Compare and contrast the ML modeling iterative cycle with the cycle for deployment of ML products
- 05 List the typical metrics you might track to monitor concept drift



학습 목표

- 01 머신러닝 학습 주기(Life Cycle)의 핵심 요소에 대해 살펴봅니다
- 02 ML project와 관련된 'concept drift'를 정의합니다
- 03 다양한 배포 시나리오에 대해 알아봅니다
 - shadow deployment
 - Canary deployment
 - blue-green deployment
- 04 ML Modeling 반복 주기(iterative cycle)와 ML Product 배포 반복 주기를 비교·대조합니다
- 05 "concept drift"를 모니터링하기 위해 추적할 수 있는 일반적인 매트릭(metrics)에 대해 알아봅니다



Overview

01 About MLOps Specialization Course & Overview

A conversation with Andrew Ng, Robert Crowe and Laurence Moroney

02 The Machine Learning Project Lifecycle

03 Deployment



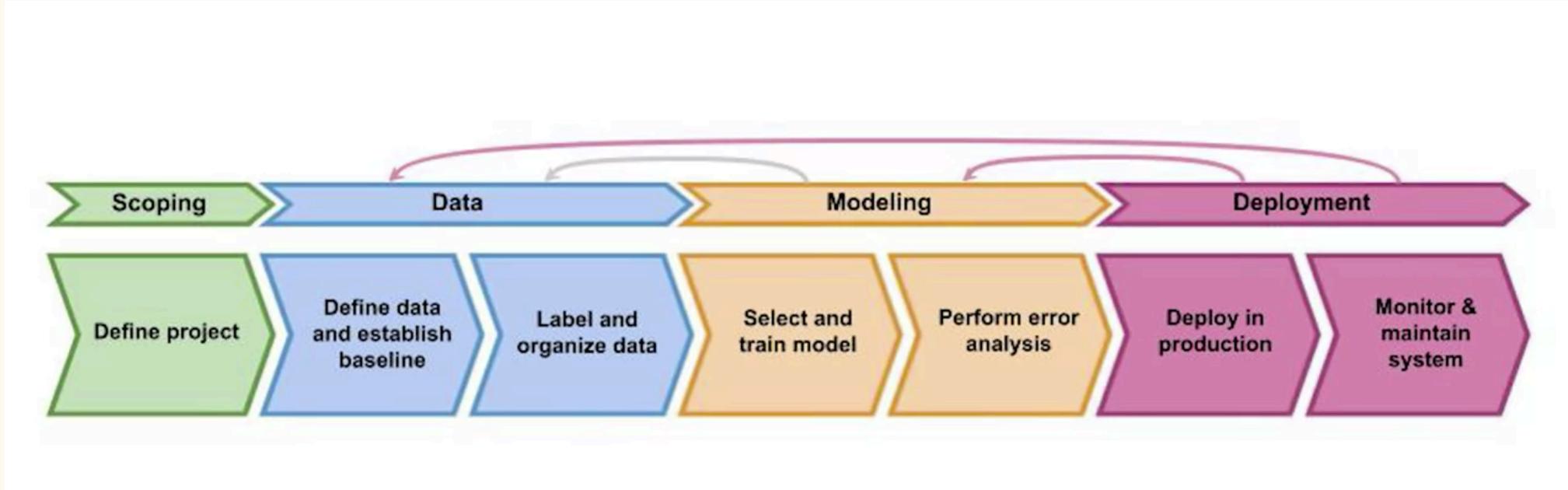
1



A conversation with Andrew Ng, Robert Crowe and Laurence Moroney

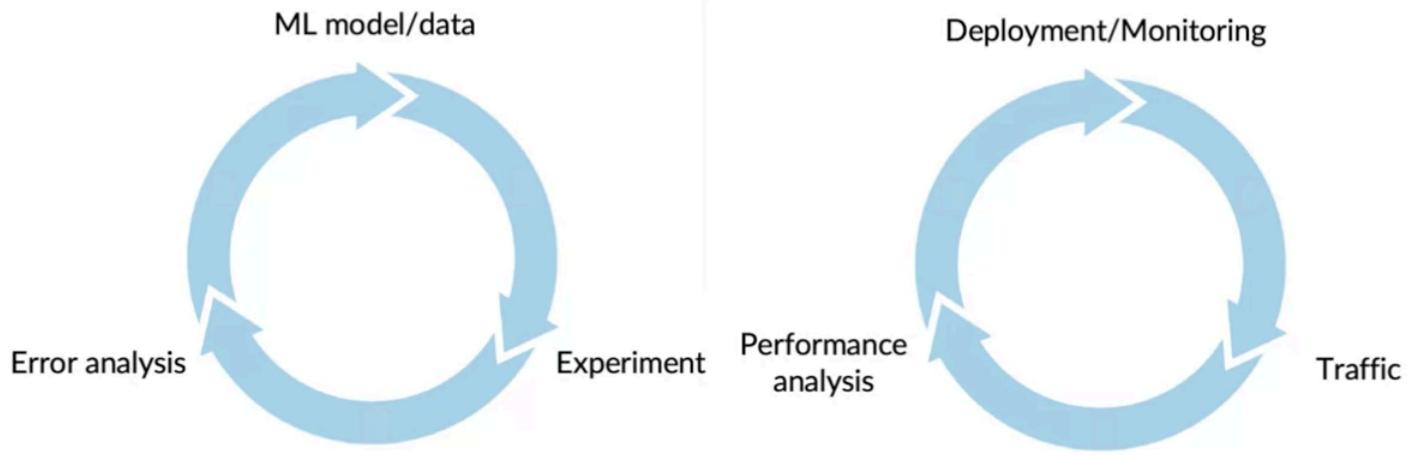
TL;DR

MLOps Specialization Course를 통해 머신러닝 모델 빌드, 배포, 관리로 이어지는 MLOps에 대한 실질적인 hands-on skill을 학습하실 수 있습니다.

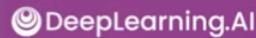


앞으로 자주 마주치게 될 다이어그램입니다! 익숙해지면 좋을 것 같아요😊

Just as ML modeling is iterative, so is deployment



Iterative process to choose the right set of metrics to monitor.



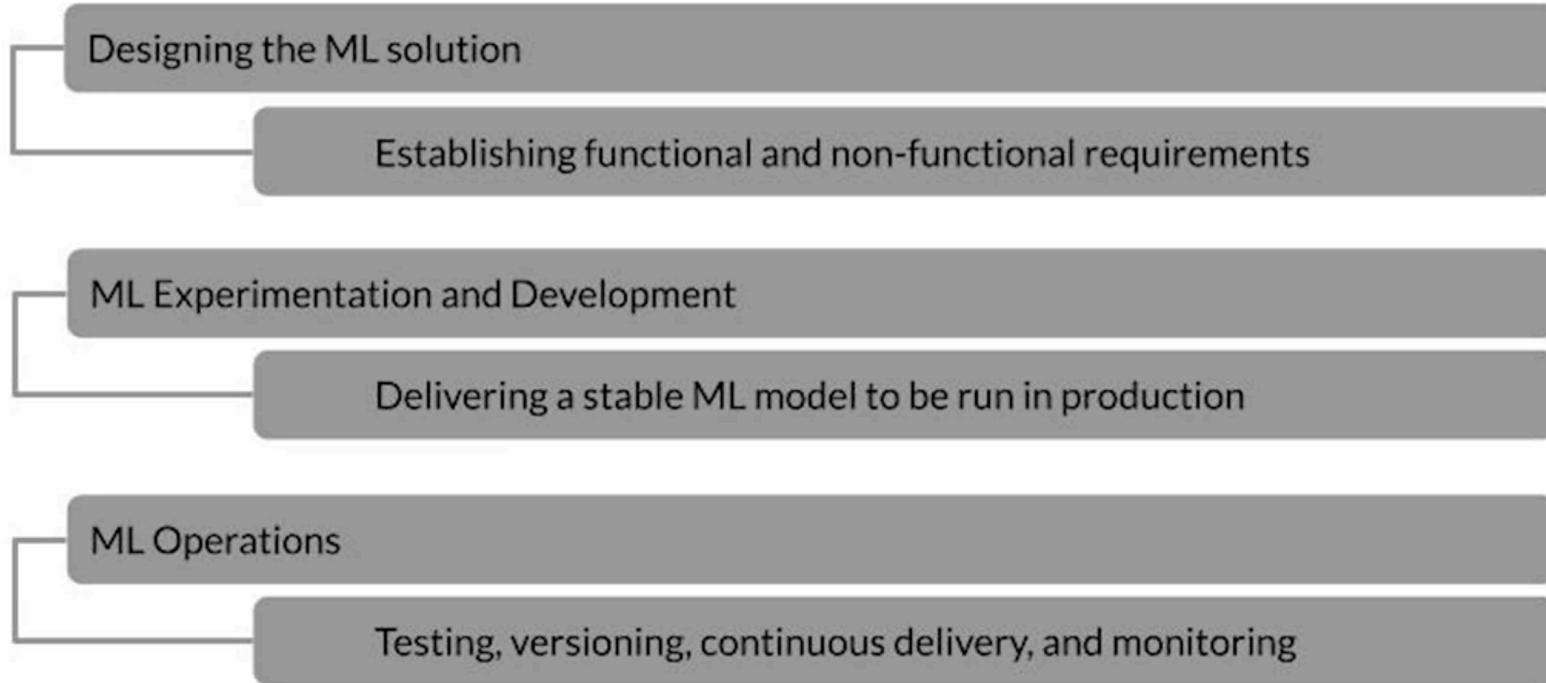
앞으로 자주 마주치게 될 다이어그램입니다! 익숙해지면 좋을 것 같아요😊

Q. 로컬 환경에서 머신러닝 모델을 구축하는 것과 머신러닝 배포는 같은게 아닌가요?
로컬 환경에서 구축한 모델에 배포(Deployment)만 추가한 게 아닌가요?

ML + Deployment = MLOps?

No!

The MLOps process



A. 로컬 환경에서 머신러닝 모델을 구축할 때와 같이 MLOps Product를 구축할 때도 머신러닝 challenge가 있다고 합니다

2

The Machine Learning Project Lifecycle



Case Study

특화과정에서 집중적으로 다룰 머신러닝 모델 배포(Deployment)에 대한 한 가지 사례를 살펴보겠습니다.

Deployment example

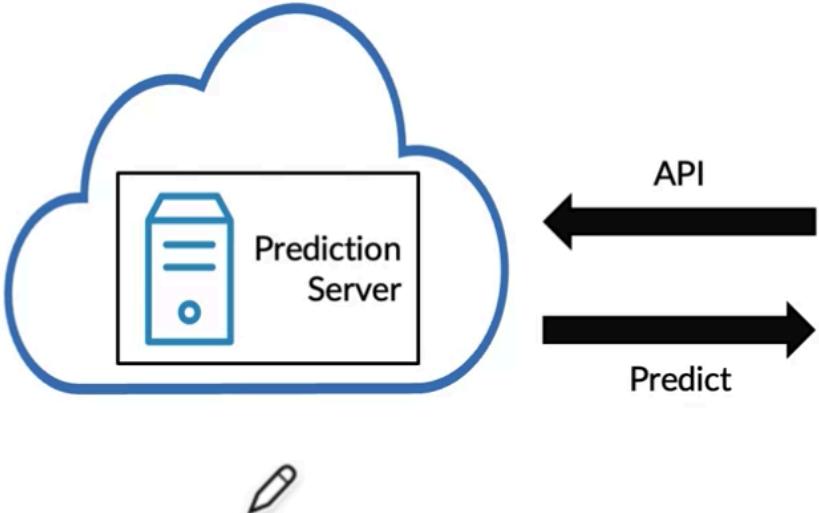
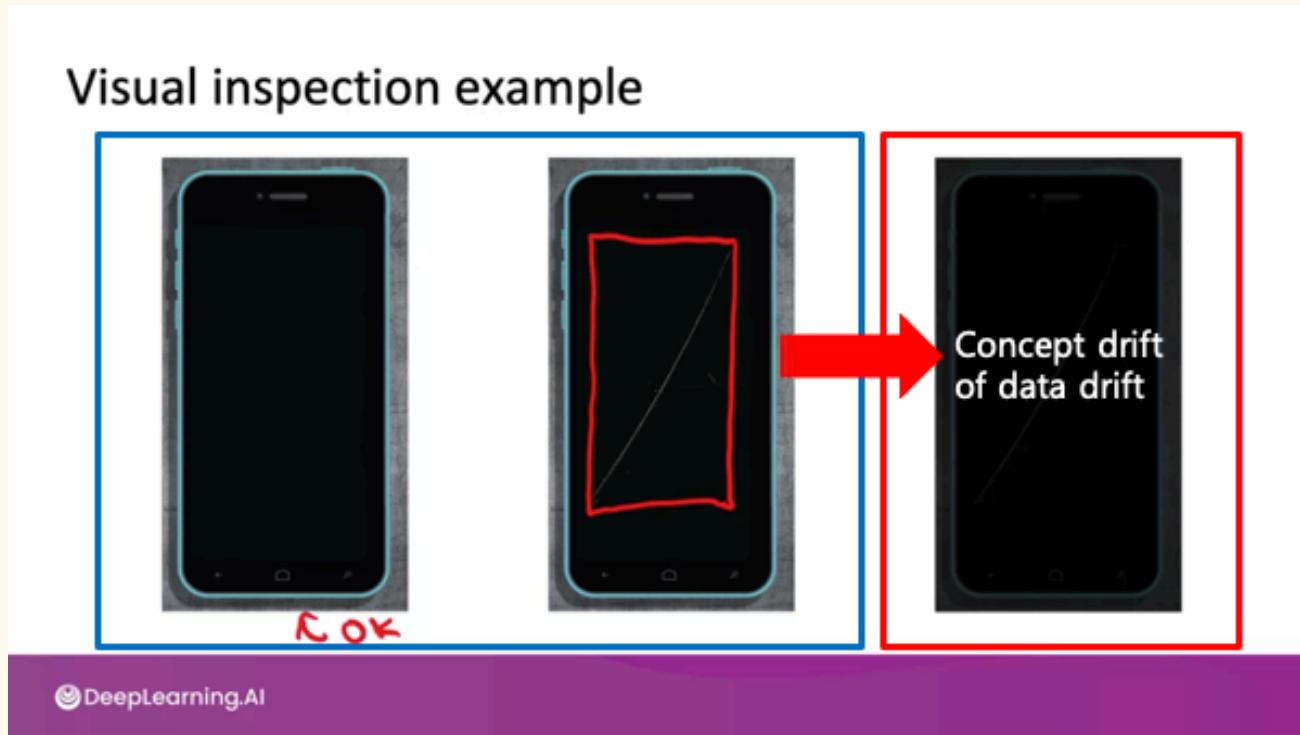


Photo from camera

- 공장
 - 스마트폰 생산
 - 엣지 디바이스
 - 결함 탐지 소프트웨어
 - 서버
 - 행동(Action)

concept drift or data drift

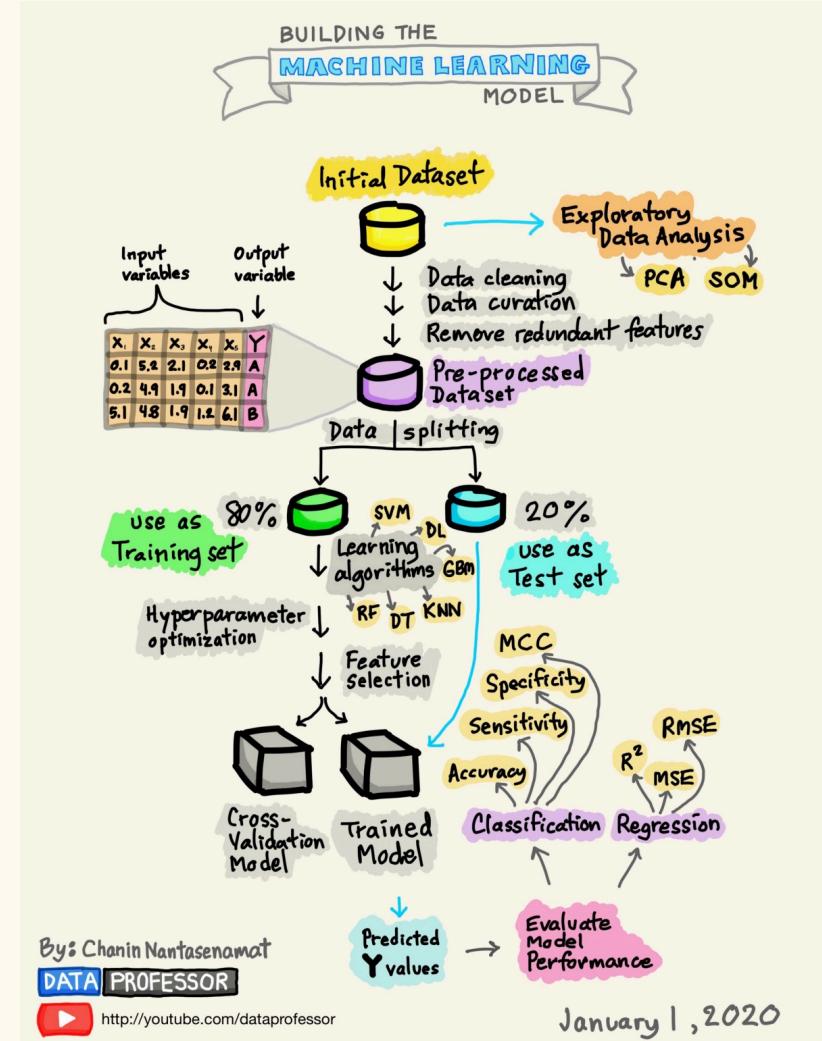
Q1. 학습때 보지 못했던 데이터를 마주치게 된다면 어떤 일이 일어날까요?



A2. 이러한 문제를 concept drift 또는 data drift라고 합니다

ML in production

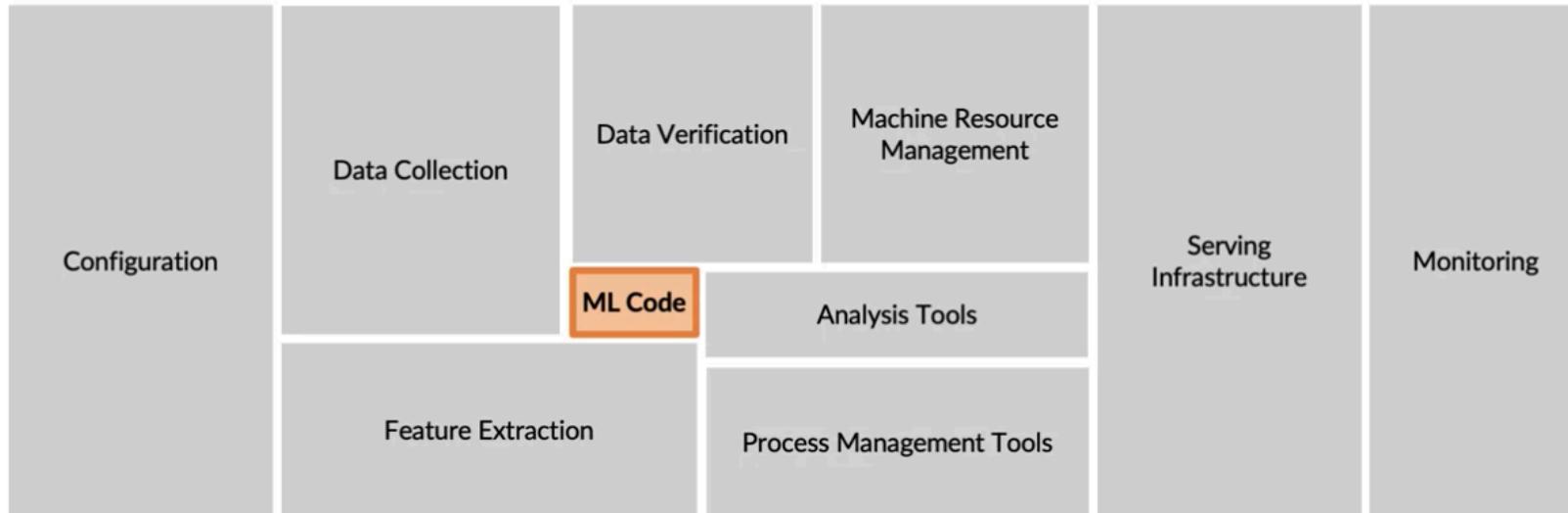
Q2. 다음 질문, Machine Learning Model 코드는 전체 프로세스에서 얼마만큼의 비중을 차지할까요?



ML in production

A2. 머신러닝 프로젝트 관점에서 바라봤을 때, 머신러닝 코드(Machine Learning Code)는 전체 프로젝트 코드의 일부분에 불과합니다

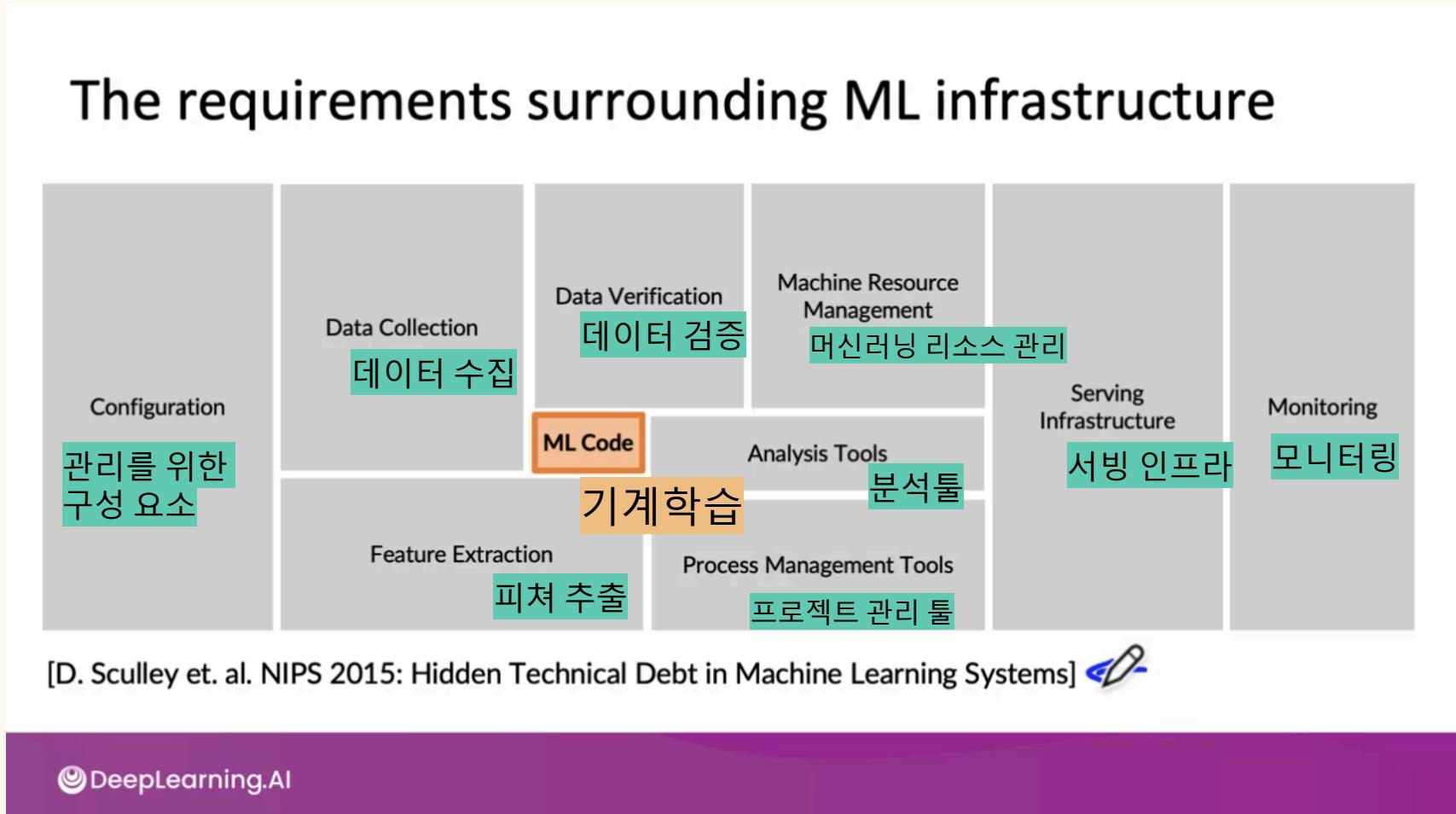
The requirements surrounding ML infrastructure



[D. Sculley et. al. NIPS 2015: Hidden Technical Debt in Machine Learning Systems]

ML in production

A2. 머신러닝 프로젝트 관점에서 바라봤을 때, 머신러닝 코드(Machine Learning Code)는 전체 프로젝트 코드의 일부분에 불과합니다



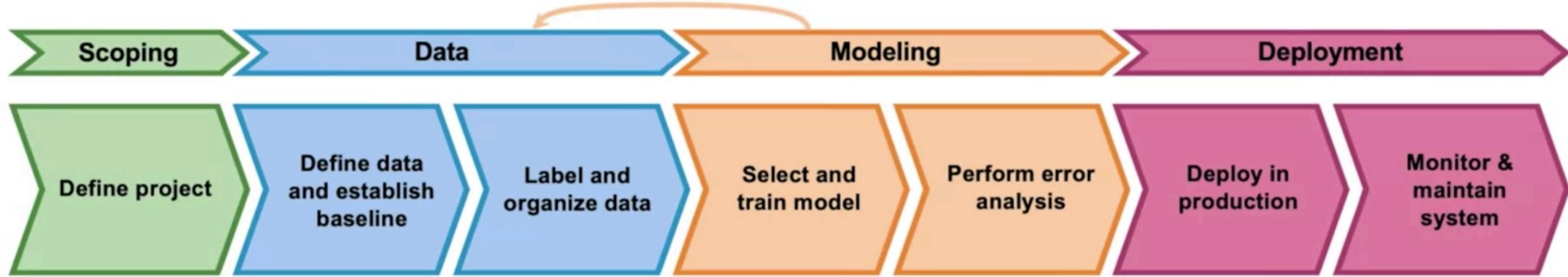
The ML project lifecycle

Q3. 그렇다면 머신러닝 프로젝트 *Life Cycle*은 어떻게 될까요?

The ML project lifecycle

Q3. 그렇다면 머신러닝 프로젝트 Life Cycle은 어떻게 될까요?

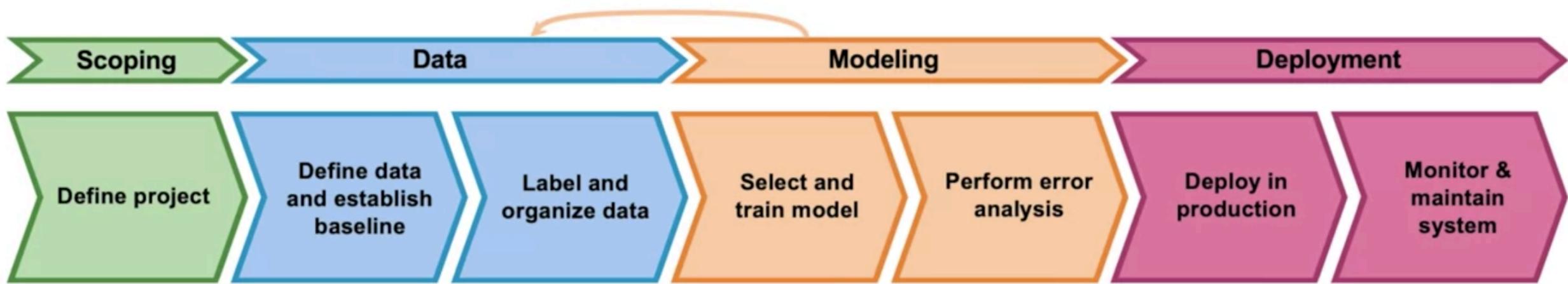
The ML project lifecycle



머신러닝 프로젝트 라이프 사이클은 위와 같습니다. 이후에도 몇 번이고 마주치게 될 것입니다!

The ML project lifecycle

The ML project lifecycle



Scoping:
문제를 정의하는 단계

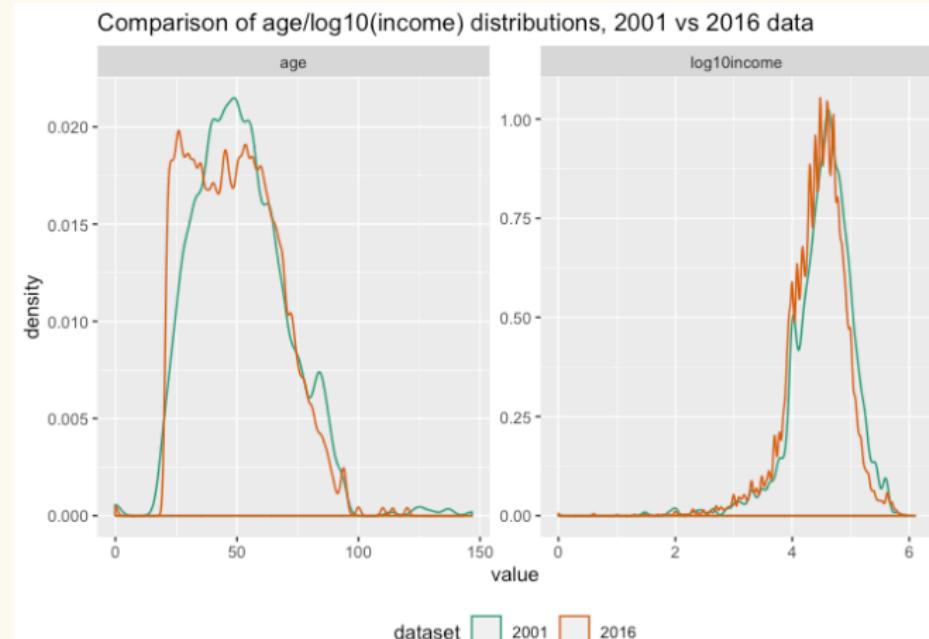
- Data:**
- 데이터에 대한 정의와 베이스 라인 설계
 - 데이터 레이블링 및 정리 (organize)

- Modeling**
- 모델 선택 및 학습
 - Error Analysis를 통한 모델 성능 검증

- Deployment**
- 배포 단계
 - 프로덕트 배포
 - 모니터링과 시스템 관리 (e.g. 데이터 분포가 바뀌었을 경우, 모델 업데이트)

The ML project lifecycle

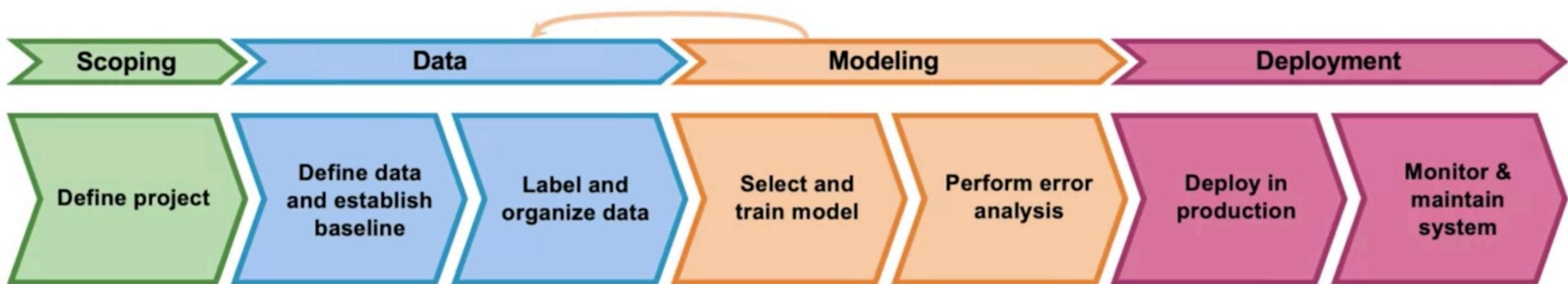
Q. 만약 배포후 데이터의 분포가 바뀌었다면 어떻게 해야 할까?



The ML project lifecycle

The ML project lifecycle

“Feel free to take a screen shot of this image and use it with your friends or by yourself to plan out your ML Project as well”



Q. 만약 배포후 데이터의 분포가 바뀌었다면 어떻게 해야할까?

A1. 모델을 업데이트한다 => more error analysis!

A2. Data 단으로 돌아간다 => update data -> model retrain -> deployment

3

Deployment



Deployment Overview

- Key challenges
- The Machine Learning Project Lifecycle
- Deployment patterns
- Monitoring
- Pipeline monitoring

Key challenges

머신러닝 모델 배포에 있어서 중요한 두 가지 문제점 들에 대해 살펴보겠습니다

1. machine learning or statistical issues
2. software engine issue

머신러닝 모델 배포에 있어서 어려운 점은(challenge) 크게 두 가지입니다.
첫째, 머신러닝 또는 통계적인 이슈(machine learning or statistical issues),
둘 째, 소프트웨어 엔진 이슈(software engine issue)입니다.

성공적인 배포를 위해 고려해야 할 두 가지 이슈들을 살펴보도록 하겠습니다

Key challenges

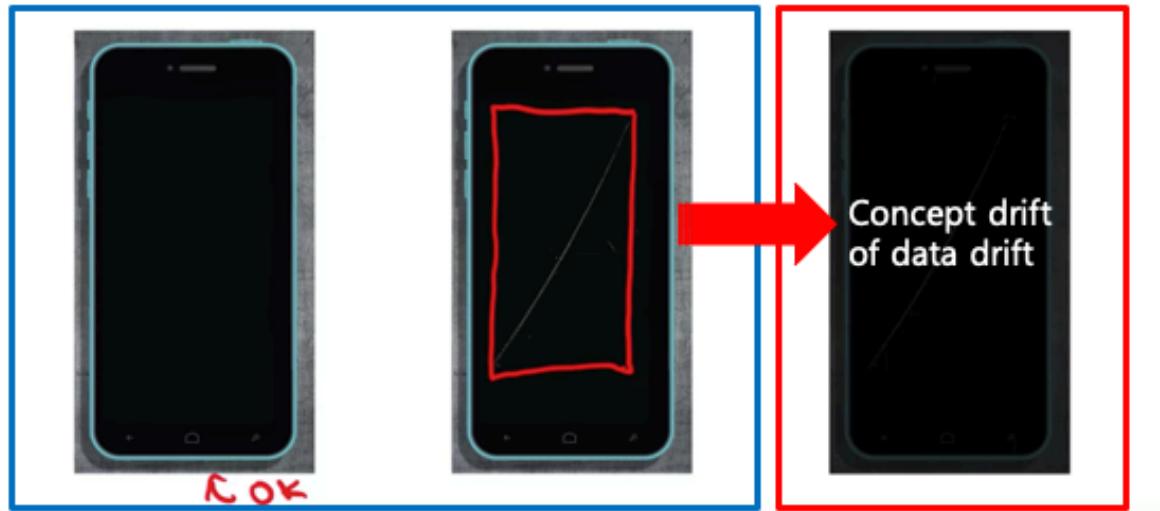
Concept drift and Data drift

Q. 머신러닝 시스템이 이미 배포된 후 데이터가 변경된다면 어떻게 해야 할까요?

Key challenges

Concept drift and Data drift

Visual inspection example



DeepLearning.AI

A. 앞서 살펴보았던 스마트폰 제조 공정에서 발생했던 데이터가 concept drift의 사례입니다.

결함 탐지 모델을 학습시켰을 때와 달리 공장에 이상이 생겨 조명이 어두워진다면 새로 마주하는 데이터는 기존 학습에 사용했던 데이터와 달라지게 됩니다.

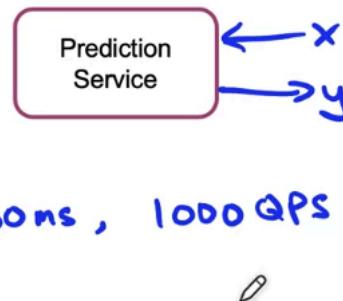
Key challenges

Software engineering issues

Software engineering issues

Checklist of questions

- Realtime or Batch
- Cloud vs. Edge/Browser
- Compute resources (CPU/GPU/memory)
- Latency, throughput (QPS)
- Logging
- Security and privacy



소프트웨어 엔지니어링과 관련된 이슈로는 아래와 같은 요소들이 있습니다.

- 머신러닝 모델을 배포시, Realtime(실시간)으로 배포할 것인지 또는 Batch(배치 단위)로 배포를 할 것인지에 대한 고려
- Cloud 환경에 배포를 할 것인지 Edge/Browser로 배포를 할 것인지에 대한 고려
- 컴퓨팅 리소스는 얼마나 사용할 것인지에 대한 고려
- 모델 응답 지연 시간은 얼마나 고려해야하는지, query 시간에 대한 throughput(처리량)에 대한 고려
- Logging: 최대한 많은 데이터를 고려하면 좋음
- 보안 및 privacy를 어떻게 고려해야 할지

위와 같이 다양한 이슈들을 고려해야 합니다

Deployment patterns

=> 지금부터는 다양한 배포 패턴에 대해 알아보겠습니다!

머신러닝 모델을 학습시킬 때, 배포를 하기 위한 최선의 방법은 모델을 작동시키고 모델이 최선을 다해 결과를 도출해주기를 기다리는 것이 아닙니다.

Deployment patterns

Common deployment cases

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

Key ideas:

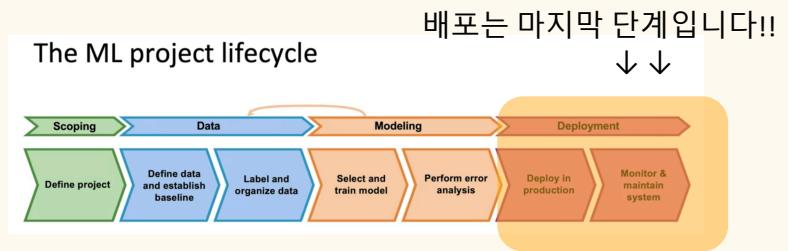
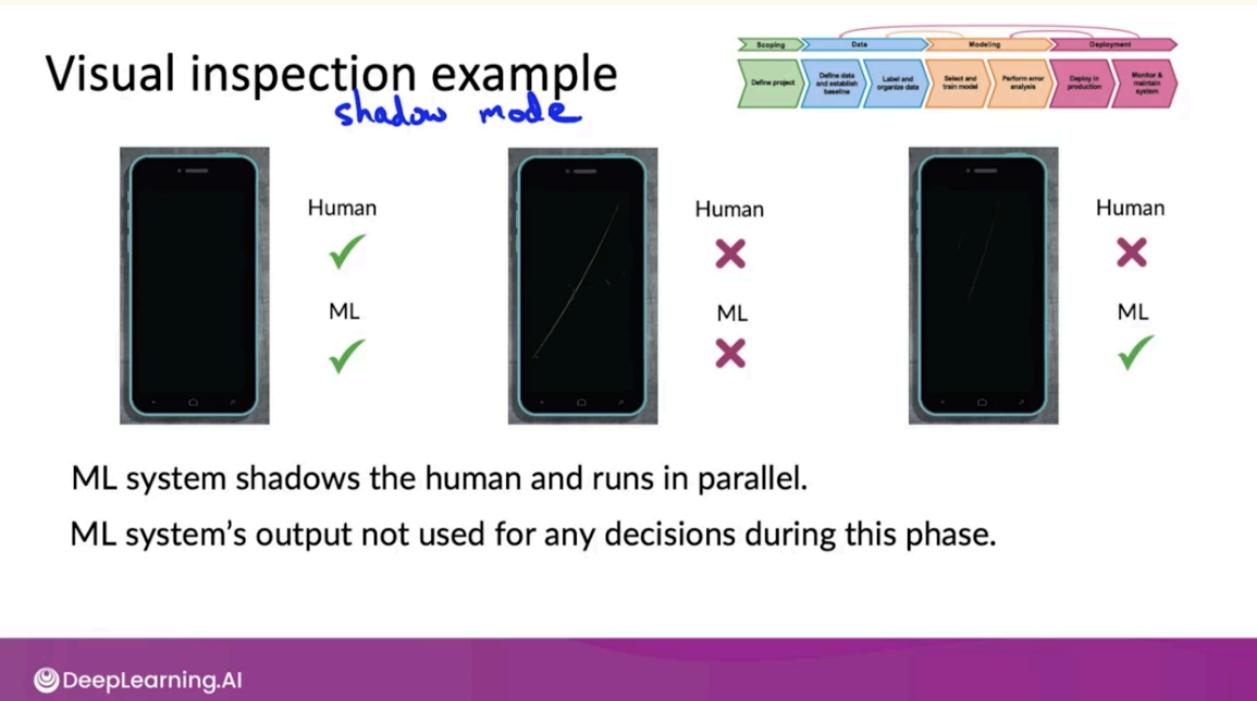
- Gradual ramp up with monitoring
- Rollback

일반적인 배포 패턴은 다음과 같습니다



Deployment patterns

Pattern1. Shadow mode Deployment



Shadow mode deployment

- 머신러닝 학습 알고리즘이 수행되는 방식과 인간의 판단이 비교되는 방식에 대한 데이터 수집을 위한
- 즉, 학습 알고리즘의 이전 구현(older implementation of a learning algorithm)과 현재 개발한 머신러닝 알고리즘의 결과를 비교하는 것입니다

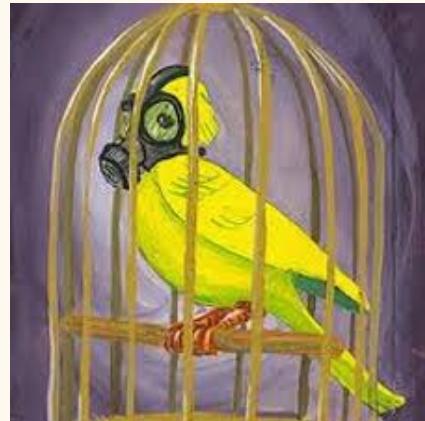
ML system's output not used
for any decisions during this phase!

Deployment patterns

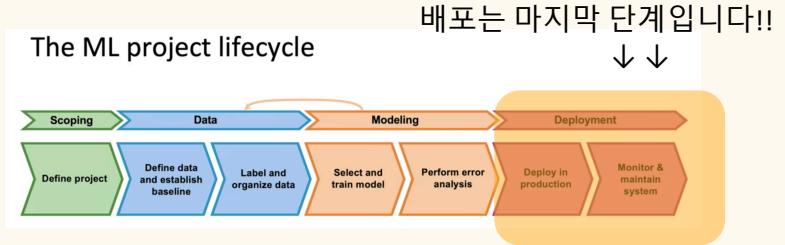
Pattern2. Canary Deployment



↑ He is coal miner



↑ This is canary



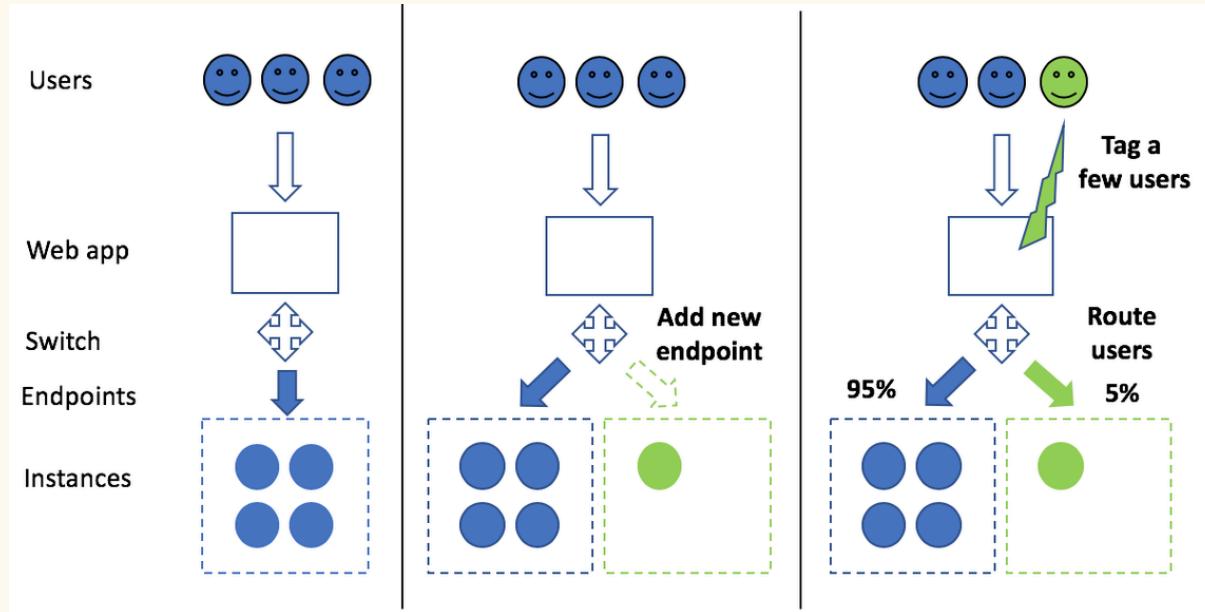
Canary deployment

- Canaries were **iconically used in coal mines to detect the presence of carbon monoxide.**

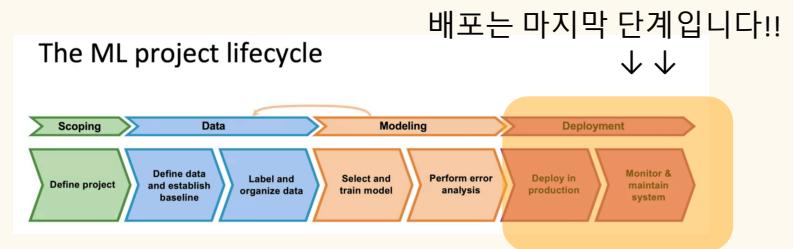
‘탄광의 카나리아는 다가온 위험을 먼저 알려주는 대상을 가리키는 말입니다. 탄광에서 나오는 유독 가스에 죽거나 다치는 일을 피하고자 광부들이 유독 가스에 민감한 카나리아를 데리고 강도로 내려간 일에서 유래했습니다.’

Deployment patterns

Pattern2. Canary Deployment



- Roll out to small fraction (say 5%) of traffic initially
- Monitor system and ramp up traffic gradually



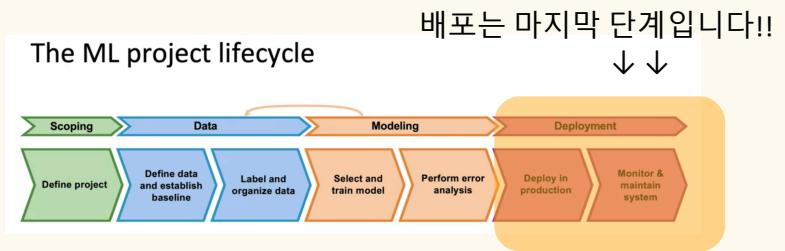
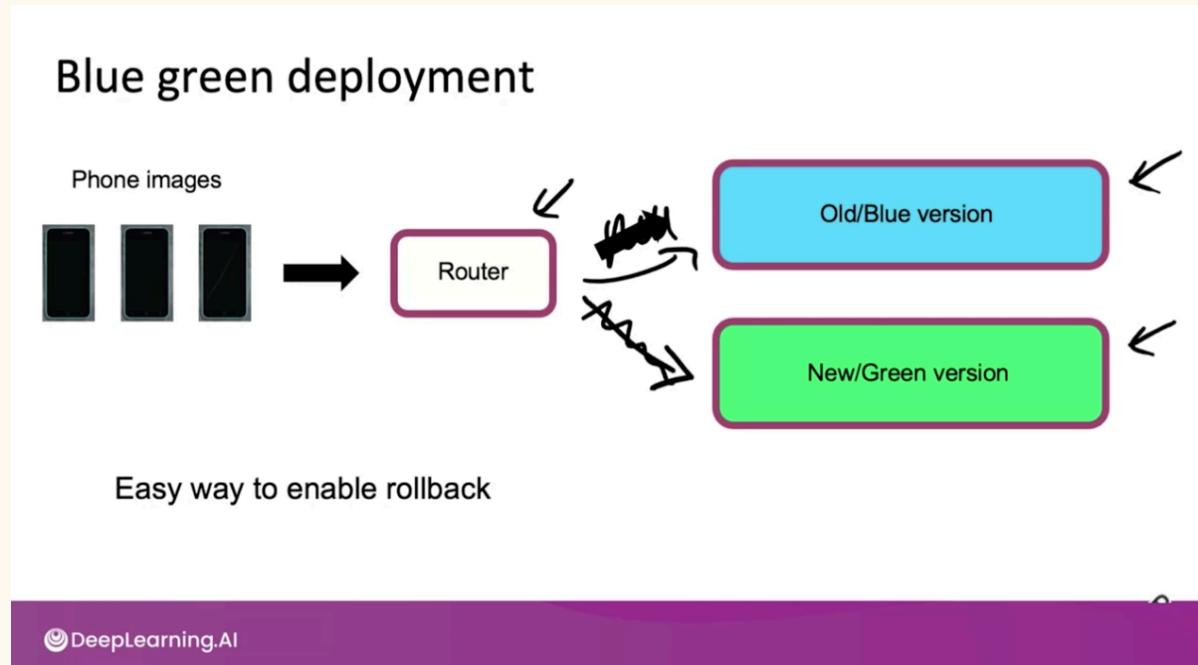
Canary deployment

- 전체 트래픽이 100이라고 했을 때, 초기 트래픽의 5% 정도만 roll out하고 알고리즘이 실제 결정을 내리도록 함
- 즉, 모델이 트래픽의 작은 비율에서만 실행될 수 있기에 모델의 실수가 트래픽의 작은 부분에만 영향을 미치게 됨
- 따라서 모델의 문제를 조기에 발견할 수 있음

다시 말해 Canary Deployment를 사용하면 학습 알고리즘을 배포하는 과정에서 지나치게 큰 결과가 발생하기 전에 문제를 조기에 발견 할 수 있습니다.

Deployment patterns

Pattern3. Blue green deployment



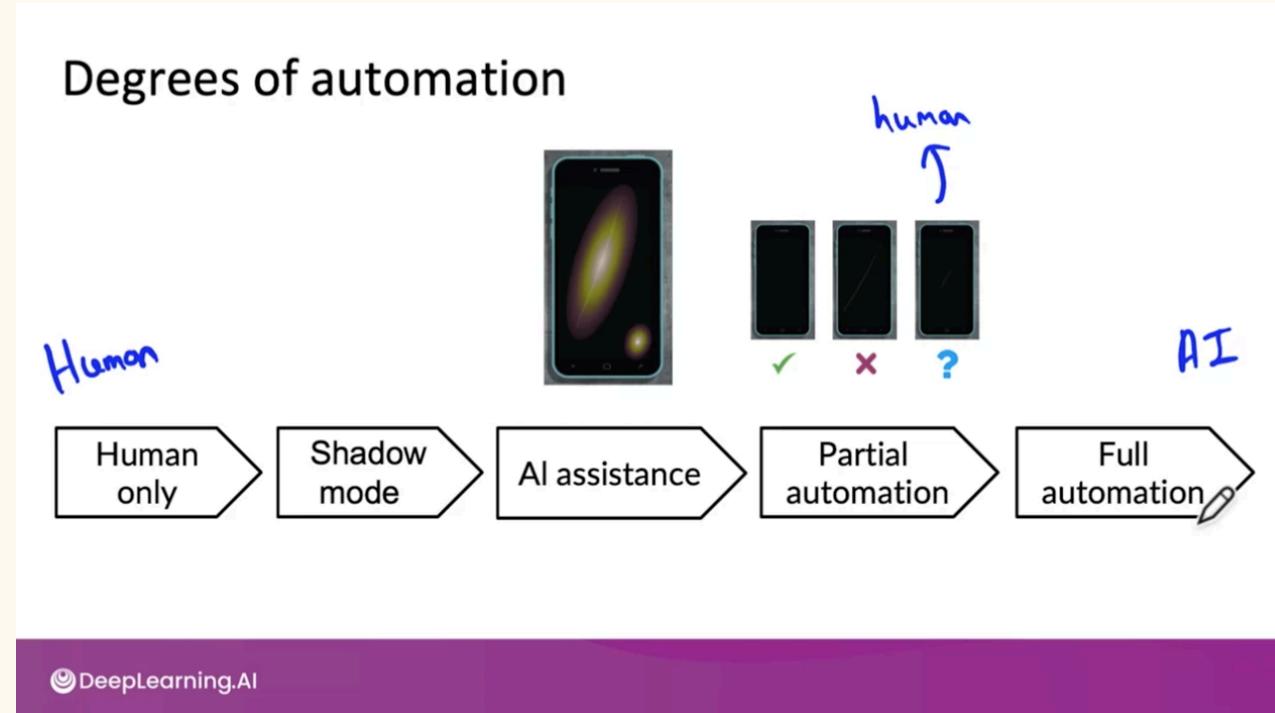
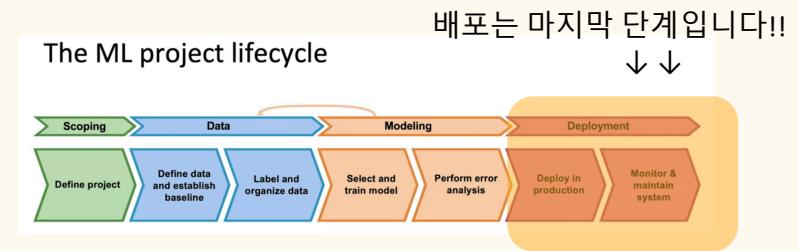
Blue green deployment

- 용어에서와 같이 Old Version software는 Blue version, New version software는 Green version으로 이해하시면 됩니다
- 학습자가 원하는 방식으로 라우터를 연결하여 입력 값에 대한 결정을 내릴 수 있습니다
- 오래된 모델부터 최신 모델에 이르기까지 쉽게 롤백이 가능한 장점이 있습니다

롤백이란 이전 버전 혹은 원하는 버전으로 소프트웨어를 새로 설치하는 것입니다.

Deployment patterns

Degress of automation



가장 왼쪽인 Human only(인간에 완전 의존적인) 방법부터 Full automation(완전 자동화?)에 이르기까지 다양한 정도가 있는 것을 확인하실 수 있습니다

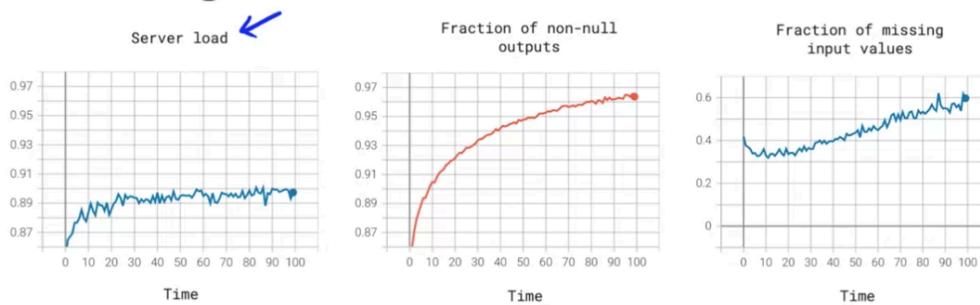
Monitoring

Q. 머신러닝 시스템이 우리가 설정한 기대치를 만족시키는지 어떻게 확인할 수 있을까?

Monitoring

A. Andrew Ng 교수님께서는 세 가지 방법을 추천해주셨습니다

Monitoring dashboard



- Brainstorm the things that could go wrong.
- Brainstorm a few statistics/metrics that will detect the problem.
- It is ok to use many metrics initially and gradually remove the ones you find not useful.

Monitoring dashboard

1. Brain Storming: 팀원들과 모여 앉아 가능한 많은 모든 요소 고려하기
2. Few statistics/metrics 살펴보기
3. 모니터링 초반에는 다양한 지표 살펴보기. 이후 지표를 줄여나가기

Monitoring

Q. 머신러닝 시스템이 우리가 설정한 기대치를 만족시키는지 어떻게 확인할 수 있을까?

Examples of metrics to track

Software metrics:

Memory, compute, latency, throughput, server load

Input metrics:

Avg input length
Avg input volume
Num missing values
Avg image brightness

Output metrics:

times return " " (null)
times user redoes search
times user switches to typing
CTR

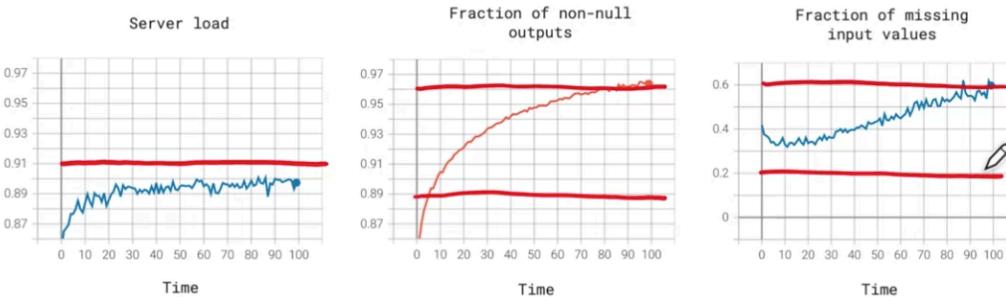
Examples of metrics to track

왼쪽 그림에서와 같이 metrics, Input metrics, output metrics를 종류별로 살펴보시면 도움이 된다고 합니다!

Monitoring

Q. 머신러닝 시스템이 우리가 설정한 기대치를 만족시키는지 어떻게 확인할 수 있을까?

Monitoring dashboard



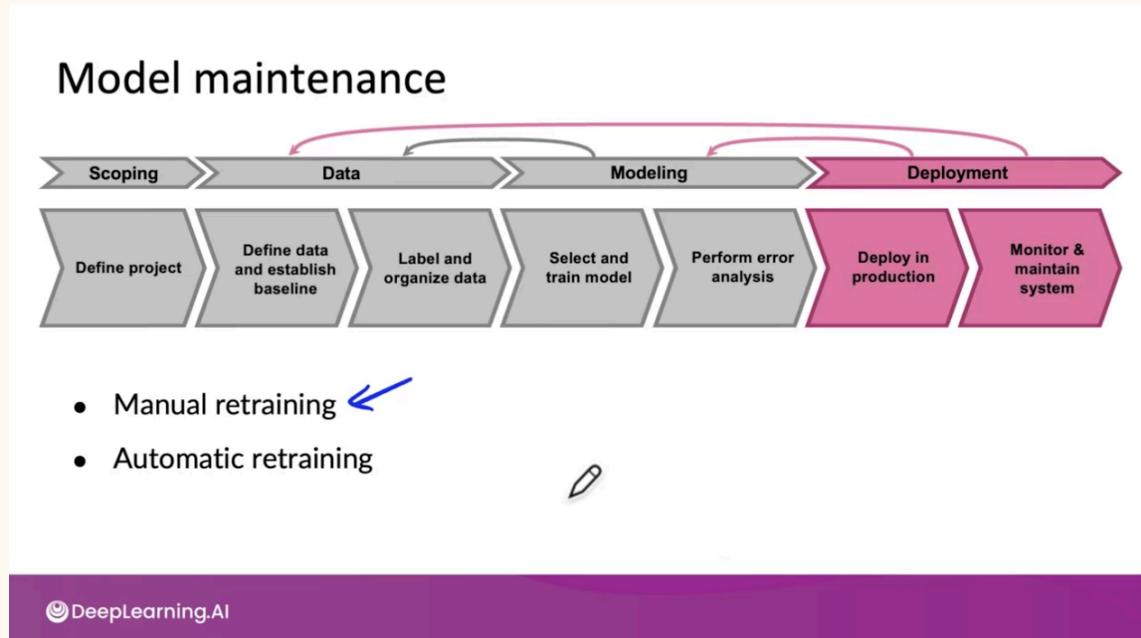
- Set thresholds for alarms
- Adapt metrics and thresholds over time

Monitoring dashboard

메트릭을 선택한 이후에는 **alarm**을 설정하고 모니터링을 해주시면 됩니다

Monitoring

Q. 머신러닝 시스템이 우리가 설정한 기대치를 만족시키는지 어떻게 확인할 수 있을까?



Model maintenance

“모니터링의 핵심은 시스템 모니터링을 통해 심층적인 오류 분석을 수행할 수 있다는 점입니다”

모니터링을 통해 시스템 성능을 유지하거나 개선하기 위해 모델을 업데이트 할 수 있습니다

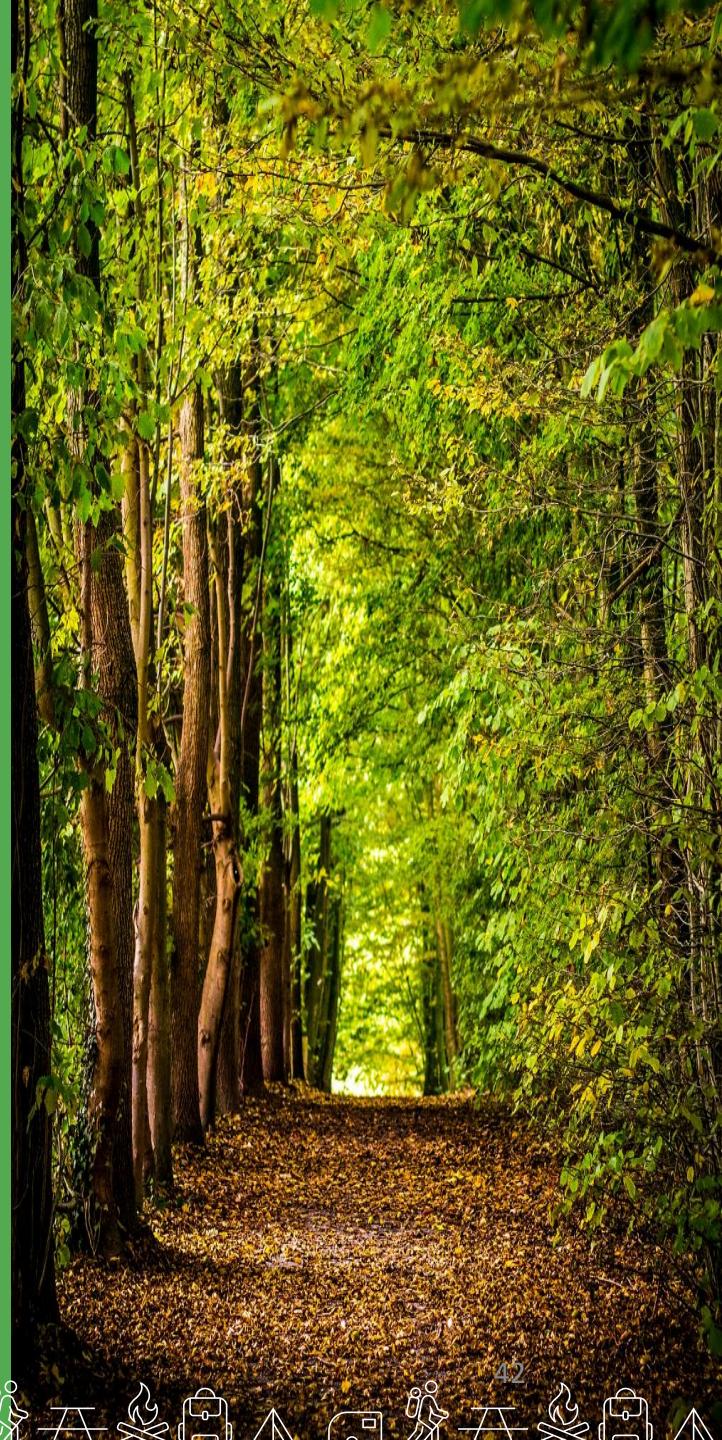
Pipeline Monitoring

Q. 그렇다면 파이프라인 모니터링을 위해서는 어떤 지표를 살펴봐야 할까요?

다양한 지표들이 있겠지만 Software metrics, Input metrics, Output metrics 등을 모니터링을 하면 좋다고 합니다

Review

- 01 머신러닝 학습 주기(Life Cycle)의 핵심 요소에 대해 살펴보았습니다
 - Scoping, Data, Model, Deployment
- 02 ML project와 관련된 'concept drift'를 정의할 수 있게 되었습니다
- 03 다양한 배포 시나리오에 대해 알아보았습니다
 - shadow deployment
 - Canary deployment
 - blue-green deployment
- 04 ML Modeling 반복 주기(iterative cycle)와 ML Product 배포 반복 주기를 비교·대조할 수 있게 되었습니다
- 05 "concept drift"를 모니터링하기 위해 추적할 수 있는 일반적인 매트릭(metrics)에 대해 알아보았습니다.





끝



Q.

질문 있으신가요?



감사합니다!