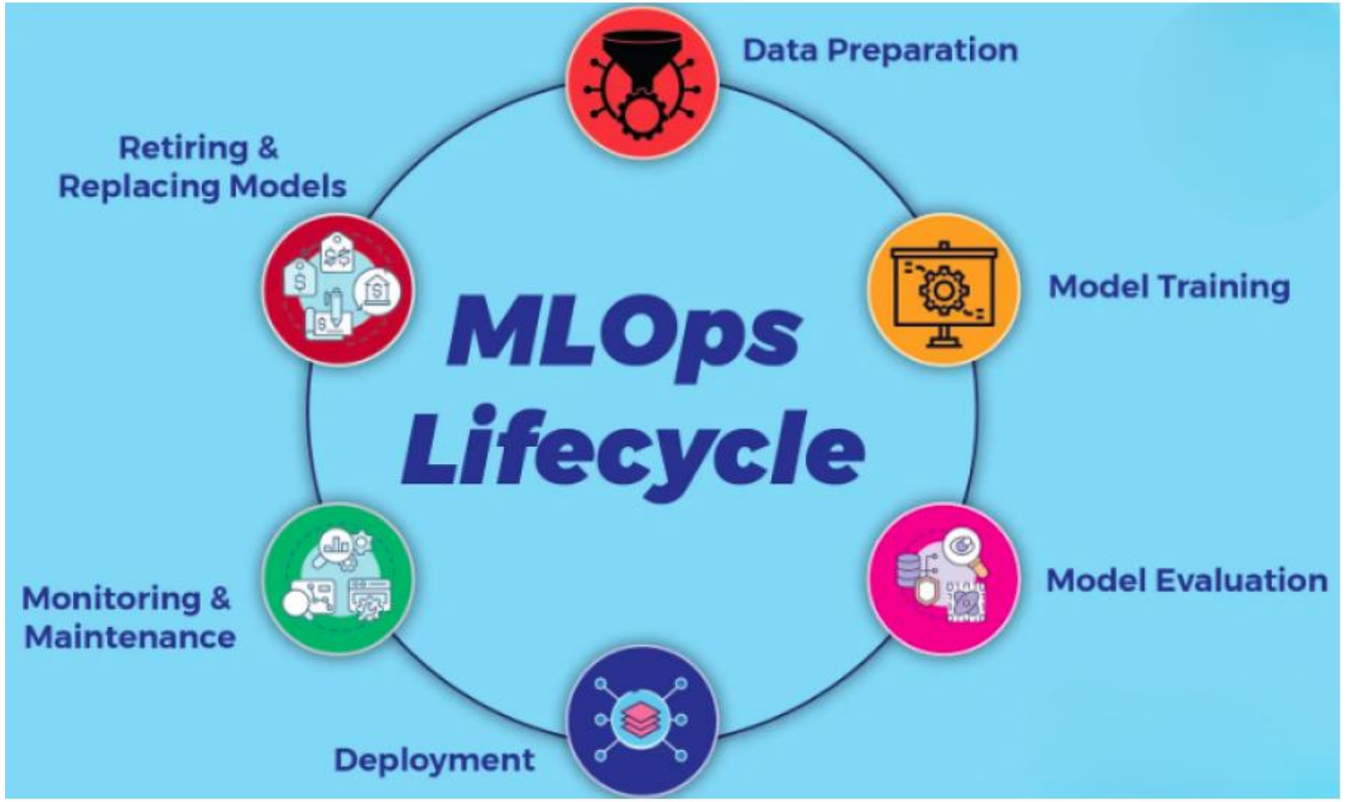
MACHINE LEARNING OPERATIONS

WEEK 12

UNLIMITED

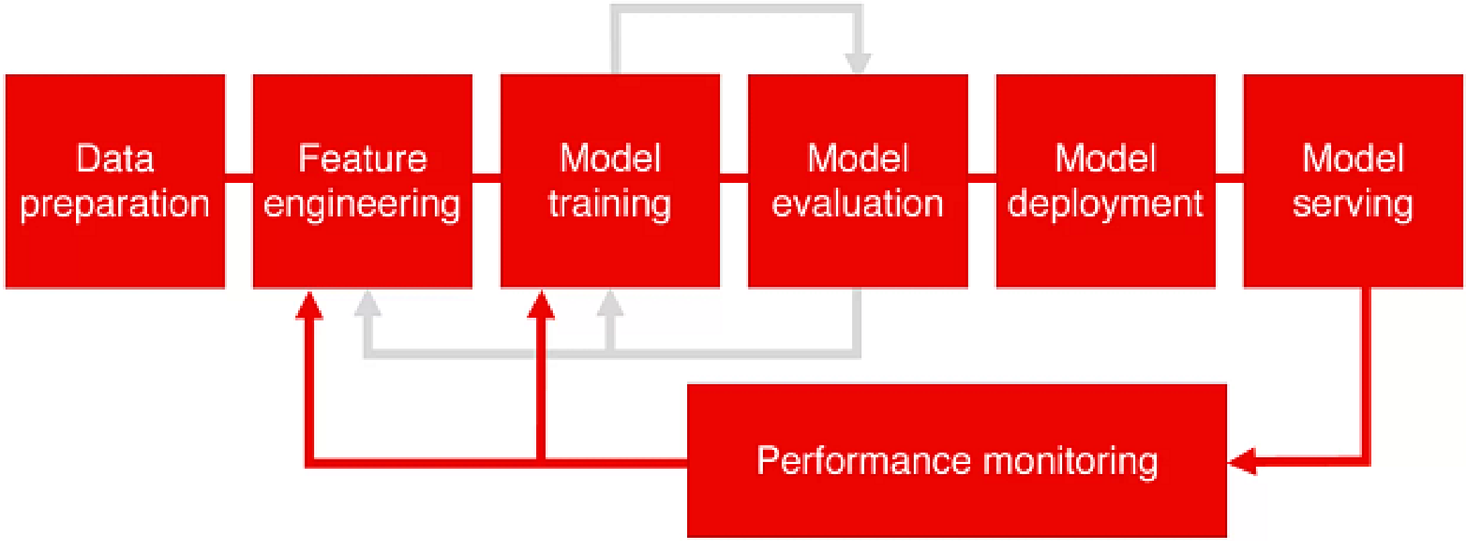
Presented by **Asst. Prof. Dr. Tuchsanai Ploysuwan**

Model monitoring with Evidently AI


MLOps
Lifecycle
Data Preparation
Model Training
Model Evaluation
Deployment
Monitoring &
Maintenance
Retiring &
Replacing Models

Maintain model quality

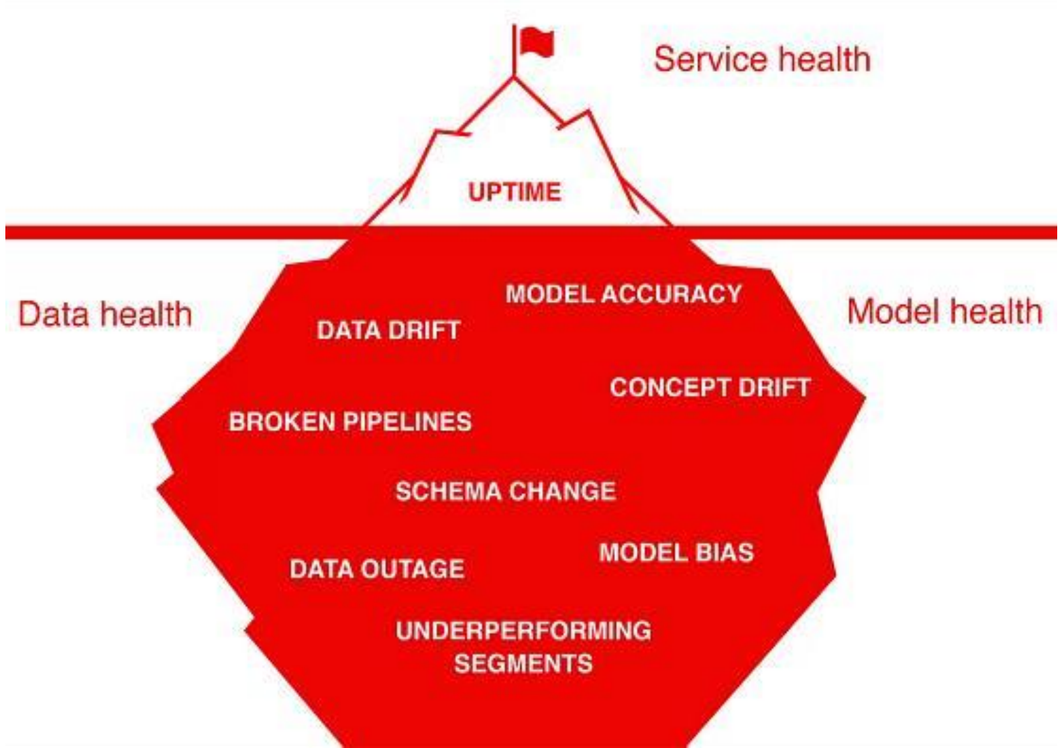
Model **performance**may **fluctuate**over time due to changes in the production dataset. Therefore, it is necessary to **monitor**the model and the service to ensure that it works as expected.



Type of problems

There are different categories of problems that can occur to ML service:

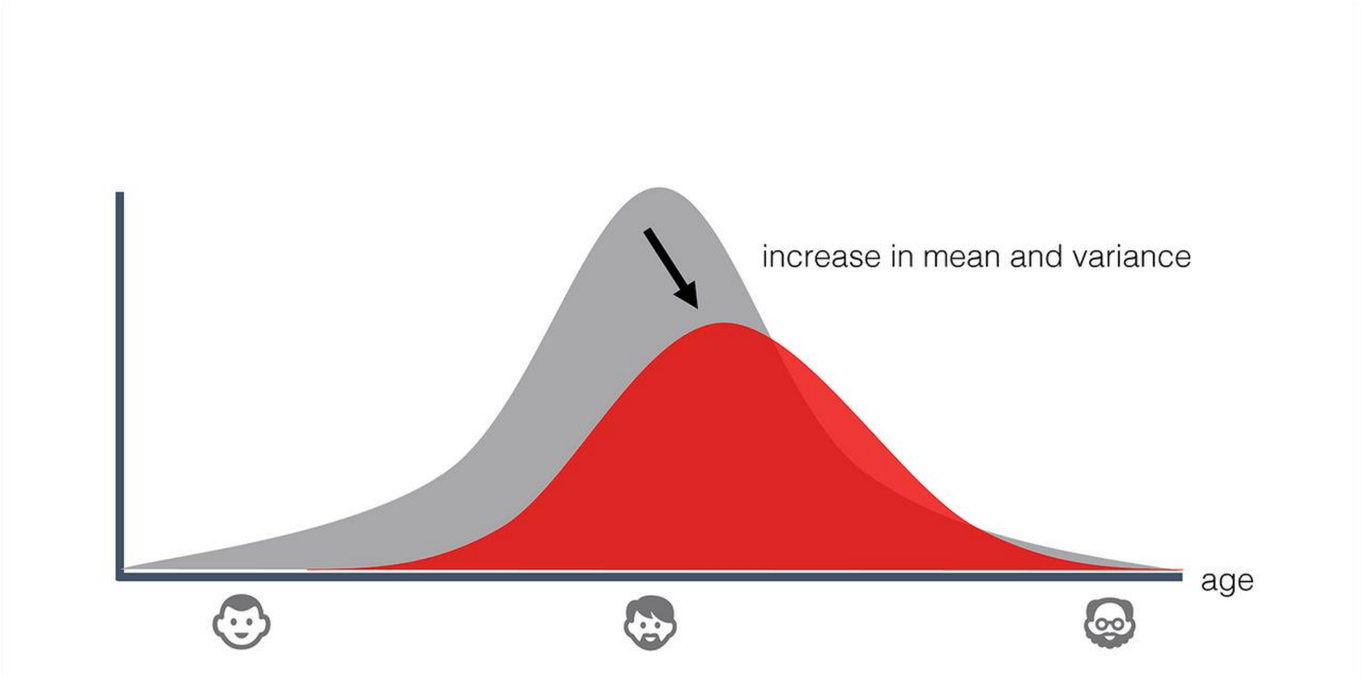
- **Poor data** quality, broken pipelines, or **technical**issues cause a drop in performance.
- **Data drift.** Itis the change in the distribution of data.
  Model works worse in unknown dataset regions.
- **Concept Drift**. The relationship between the target variable and input features changes.



Data Drift

It occurs due to **changes**in the **input data**. To detect it you must observe the input data in production and compare it with the training data. Tests to detect changes in the distribution of the input data:

- Kolmogorov–Smirnov(**KS**) test
- Population Stability Index(**PSI**)
- **Z-score**



Methods for Detecting Data Drift

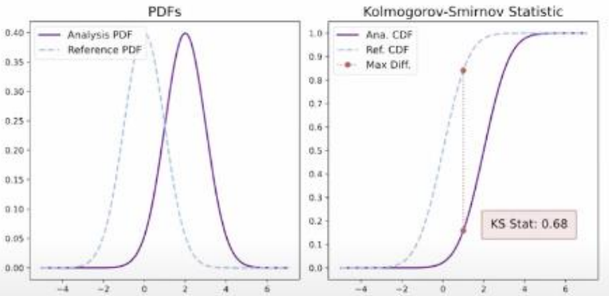
- All the methods for detecting data drift are lagging indicators of drift i.e. Only after they have processed enough data after any kind of drift that has occurred, that the actual drift is detected.

- **Kolmogorov-Smirnov (K-S) test**:
  a. The K-S test is a nonparametric test that compares the cumulative distributions of two data sets, the training data and the post-training data.
  b. The null hypothesis is that the two distributions are identical and the alternative is that they are not identical. If the null is rejected then we can conclude that there is a drift in the model.
  c. The test is valid for numerical columns.

- **Chi-squared test**:
  a. The chi-squared two-sample test is applied to the categorical features to identify data drift.

Kolmogorov-Smirnov Test

The Intuition



- Maximum distance of the cumulative distribution functions (CDFs)
- Prone to false positives, especially in bigger samples
- Outputs d-statistic and p-value

Population Stability Index

- **Population Stability Index**
  a. PSI was originally developed in the Banking and Finance industries for testing the changes in the distribution of a risk score over time.

  a. It's being used both for detecting numerical and categorical variables data-drifts. *How for categorical variables?* Dividing the data into buckets.

  a. KL(Kullback-Leibler) divergence is a good measure to find how much a observed distribution 'o' differs from it's ideal/reference 'r' version. $\sum_i r_i * \log\left(\frac{r_i}{o_i}\right)$

  a. KL divergence is not symmetrical though, i.e. if we permute 'o' and 'r', we won't necessarily find the same value.

Population Stability Index

- To address this issue, one can use a symmetrical version of the KL divergence, it's called the Jeffreys divergence, often known as the Population stability index (PSI). It's defined as the sum of KL divergence from 'o' to 'r' and the one from 'r' to 'o'.

$$\sum_i (r_i - o_i) * \log\left(\frac{r_i}{o_i}\right)$$

- When PSI<=0.1
  - This means there is no change or shift in the distributions of both datasets.

- When 0.1<PSI<0.2
  - This indicates a slight change or shift has occurred.

- When PSI>0.2
  - This indicates a large shift in the distribution has occurred between both datasets.

3. Z-Score Guidelines

The Z-score measures how many standard deviations an observation or statistic is from the mean. In the context of data differences, it can be used to standardize and compare individual data points or test statistics (e.g., from KS or PSI).

Purpose:

- Normalize differences to assess their significance relative to a distribution.
- Often used to compare a sample mean to a population mean or to evaluate outliers.

Formula:

$$Z = \frac{x - \mu}{\sigma}$$

Where (x) is the observed value, μ is the mean, and σ is the standard deviation.

Guidelines:

- **Thresholds:**
  - $|Z| < 1.96$: Within 95% confidence interval (not significant at α = 0.05).
  - (|Z| ≥ 1.96): Significant difference (assuming a two-tailed test at α = 0.05).
  - (|Z| ≥ 2.58): Highly significant (α = 0.01).

Practical Thresholds for Decision-Making:

Metric	No Action	Monitor	Take Action
KS (D)	D < 0.1, p ≥ 0.05	D = 0.1–0.2, p < 0.05	D > 0.2, p < 0.05
PSI	PSI < 0.1	PSI = 0.1–0.25	PSI ≥ 0.25
Z-Score		Z	< 1.96

Concept Drift

Concept Drift refers to the **change**in the relationships between input and output data in the underlying problem **over time**.You can detect it by looking at changes in the input prediction probabilities. Example: inflation in the prediction of house prices.

Prevent concept drift:

- Model **monitoring**
- Time based approach, **retraining**the model every X time
- **Continuous retraining**