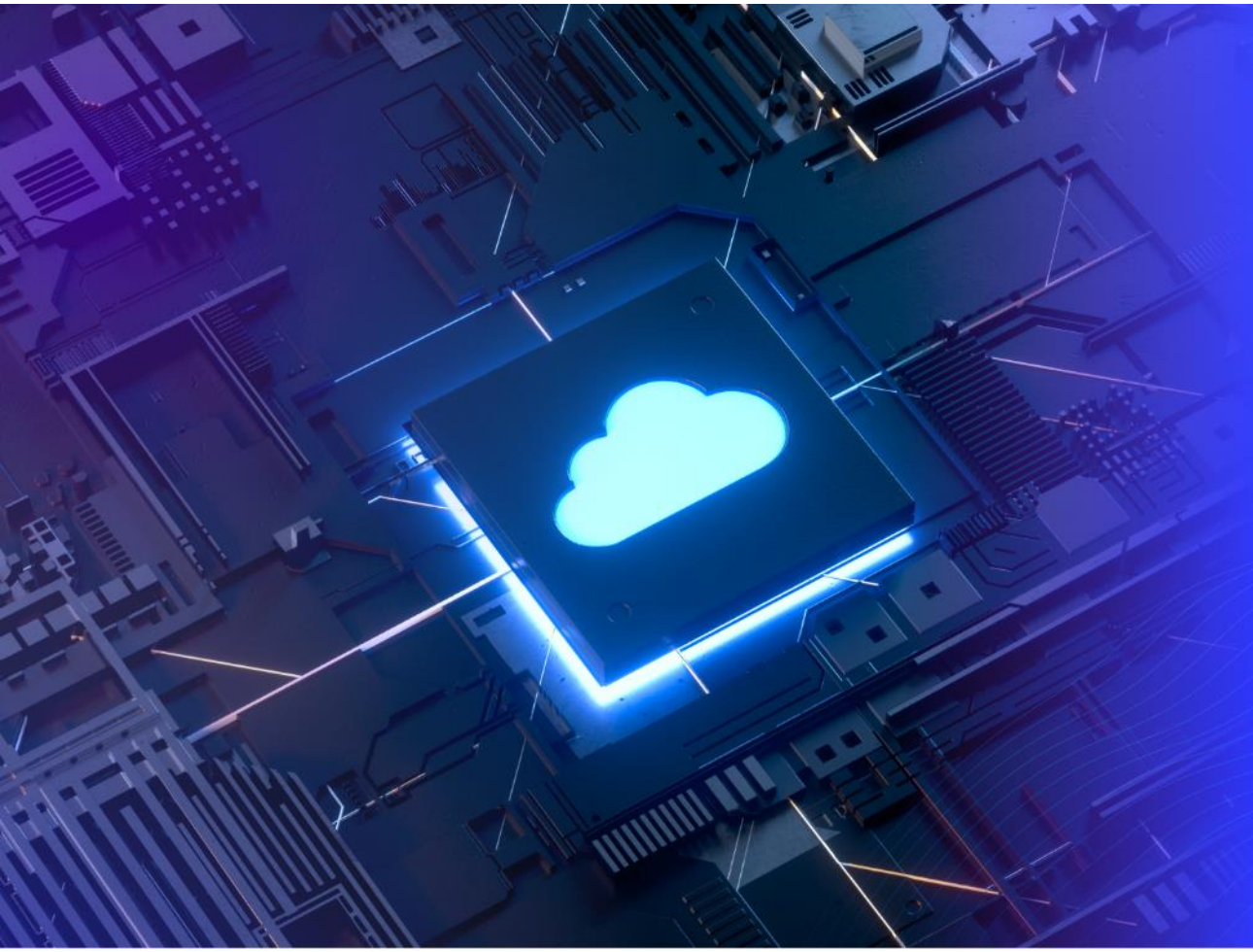
台灣積體電路製造股份有限公司
Taiwan Semiconductor Manufacturing Company, Ltd.



TSMC IT × NCTU CS 課號 5270

CLOUD NATIVE
Development Best Practice

INTRODUCING MLOPS

智慧平台暨品質部 | 蘇冠華 副理

May 25, 2022

Agenda

- What are DevOps and MLOps
- ML pipeline
- Challenges in production ML
- Importance of data
- Machine learning project: predicting Boston house prices
- (10 minute break)
- Deploying ML models in production
- Monitoring dashboard
- Model decay and detecting drift
- Summary

References

4 COURSE SPECIALIZATION

Machine Learning Engineering for Production (MLOps)

Offered by DeepLearning.AI

Enrolled | Go to Course | Save for Later

Sponsored by TSMC

About this Specialization

Understanding machine learning and deep learning concepts is essential, but if you're looking to build an effective AI career, you need production engineering capabilities as well.

Effectively deploying machine learning models requires competencies more commonly found in technical fields such as software engineering and DevOps. Machine learning engineering for production combines the foundational concepts of machine learning with the functional expertise of modern software development and engineering roles.

The Machine Learning Engineering for Production (MLOps) Specialization covers how to conceptualize, build, and maintain integrated systems that continuously operate in production. In striking contrast with standard machine learning modeling, production systems need to handle relentless evolving data. Moreover, the production system must run non-stop at the minimum cost while producing the maximum performance. In this Specialization, you will learn how to use well-established tools and methodologies for doing all of this effectively and efficiently.

In this Specialization, you will become familiar with the capabilities, challenges, and consequences of machine learning engineering in production. By the end, you will be ready to employ your new production-ready skills to participate in the development of leading-edge AI technology to solve real-world problems.

Shareable Certificate
Earn a Certificate upon completion

100% online courses
Start instantly and learn at your own schedule.

Flexible Schedule
Set and maintain flexible deadlines.

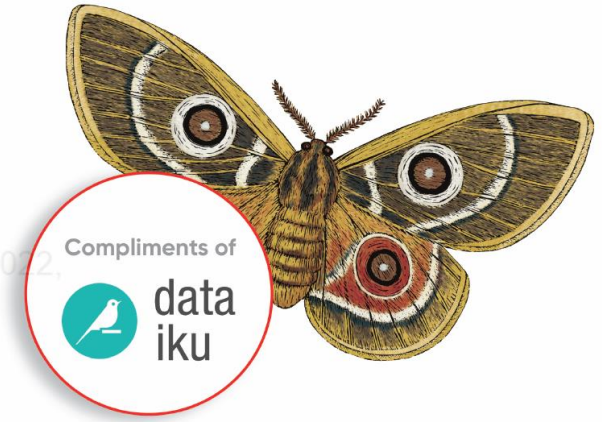
Advanced Level
Designed for those already in the industry.

English
Subtitles: English, French

O'REILLY®

Introducing MLOps

How to Scale Machine Learning in the Enterprise



Mark Treveil
& the Dataiku Team

TSMC IT × NCTU CS
2022 CLOUD NATIVE
Development Best Practice

Self Introduction

蘇冠華/ Allen Su
台大資工博士班(肄業)
8 years software engineer
4 years data analyst and machine learning engineer
4 years of experience in machine learning platform management
Certification: SCJP, SCWCD, PMP, Calibre DRC, Cadence Skill Programming
TSMC Patent : P20161397US00, P20191817US00
khsup@tsmc.com
kuanhua.su@gmail.com

陳沿任/ Jerry Chen
交大資工碩士
2 years AI & Platform engineer
chenyrc@tsmc.com

What are DevOps and MLOps?

- DevOps >> a set of practices >> shorten SDLC and provide CI, CD with high software quality.
- MLOps >> process of automating and production machine learning applications and workflows.
- So, MLOps = DevOps + Machine learning component.

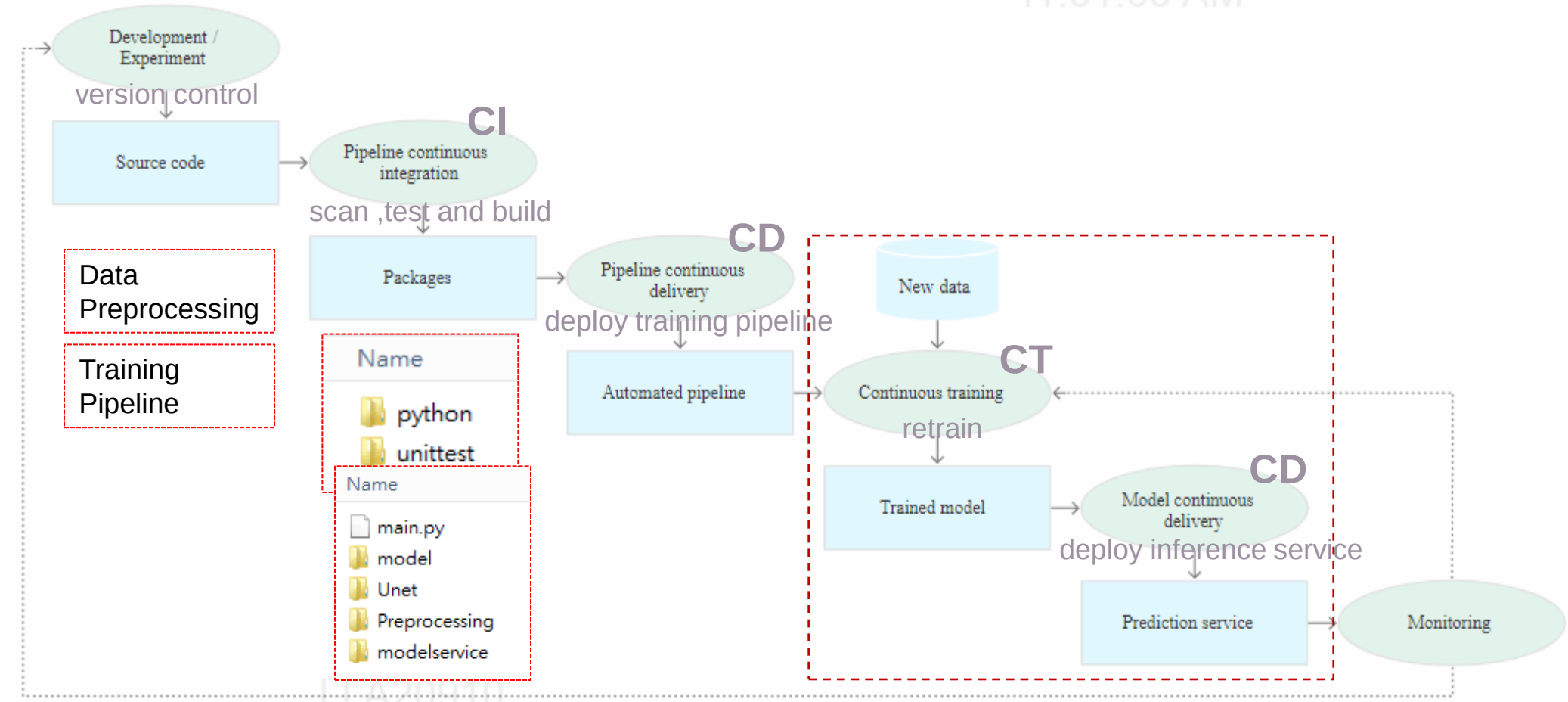
Recommendation Systems

- AIOps - AI for DevOps

  (software testing, failure forecasting, root cause analysis..)

Stages of the CI/CD automated ML pipeline.

Continuous integration (CI), continuous delivery (CD) and continuous training (CT) are at the core of Machine Learning Operation (MLOps)

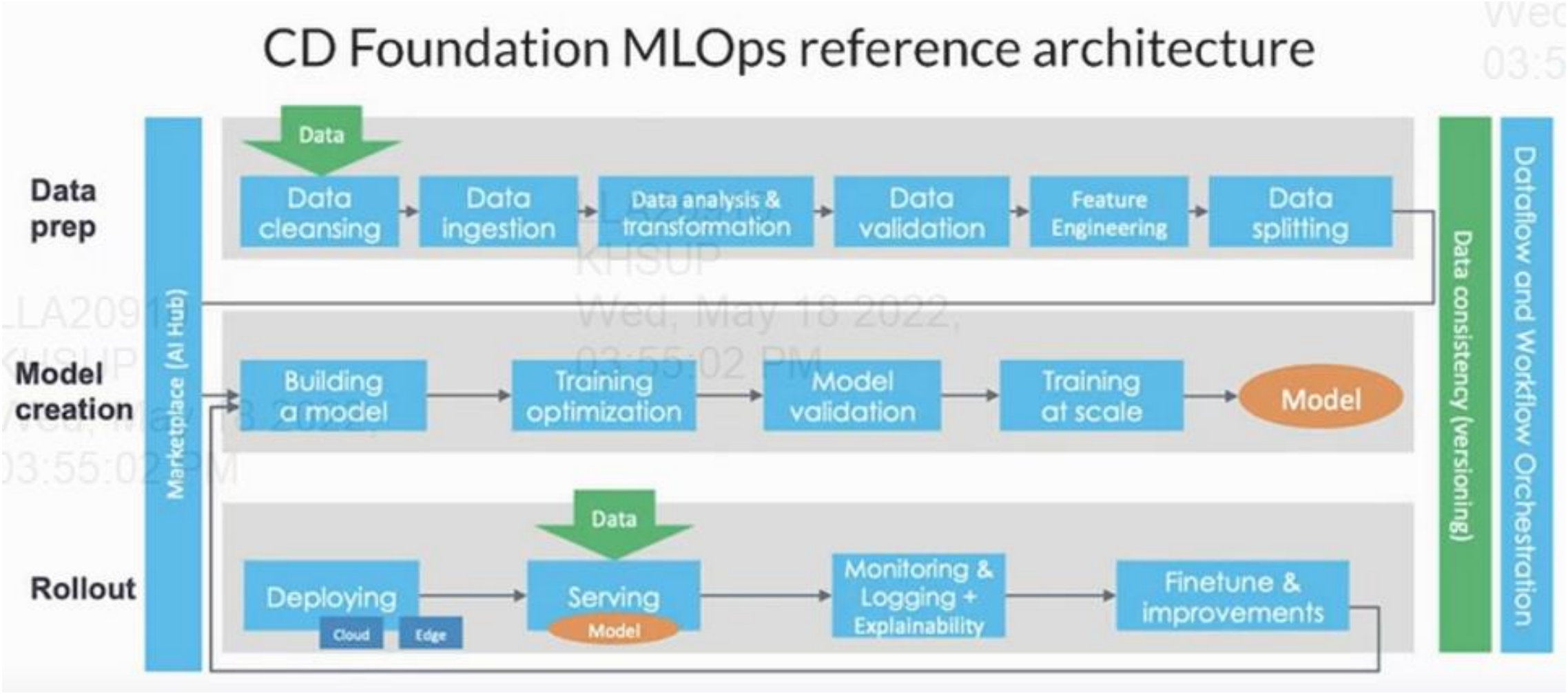


Ref: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

Challenges in Production ML

- **Integrated ML systems**
- **Continuously operate it in production**
- **Handle continuously changing data**
- **Optimize compute resource costs**

ML Pipeline

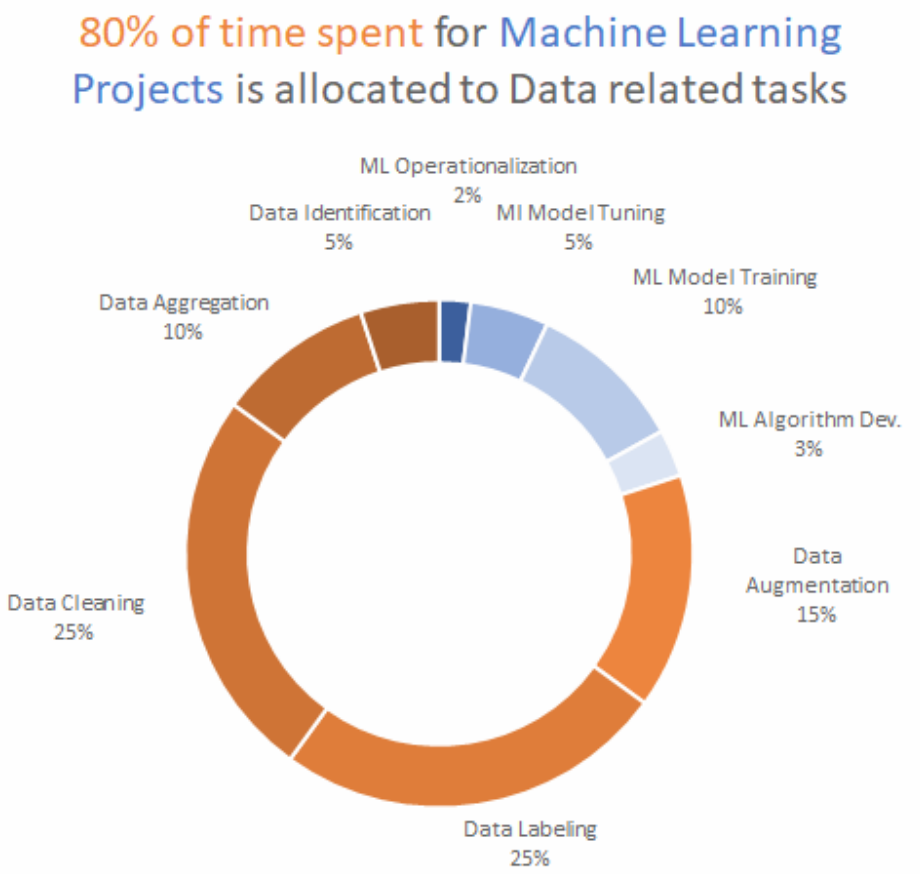


Ref: https://www.coursera.org/learn/introduction-to-machine-learning-in-production/

Importance of Data

- **Importance of data quality**
  - **Meaningful data:** maximize predictive content, remove non-informative data and check feature space coverage

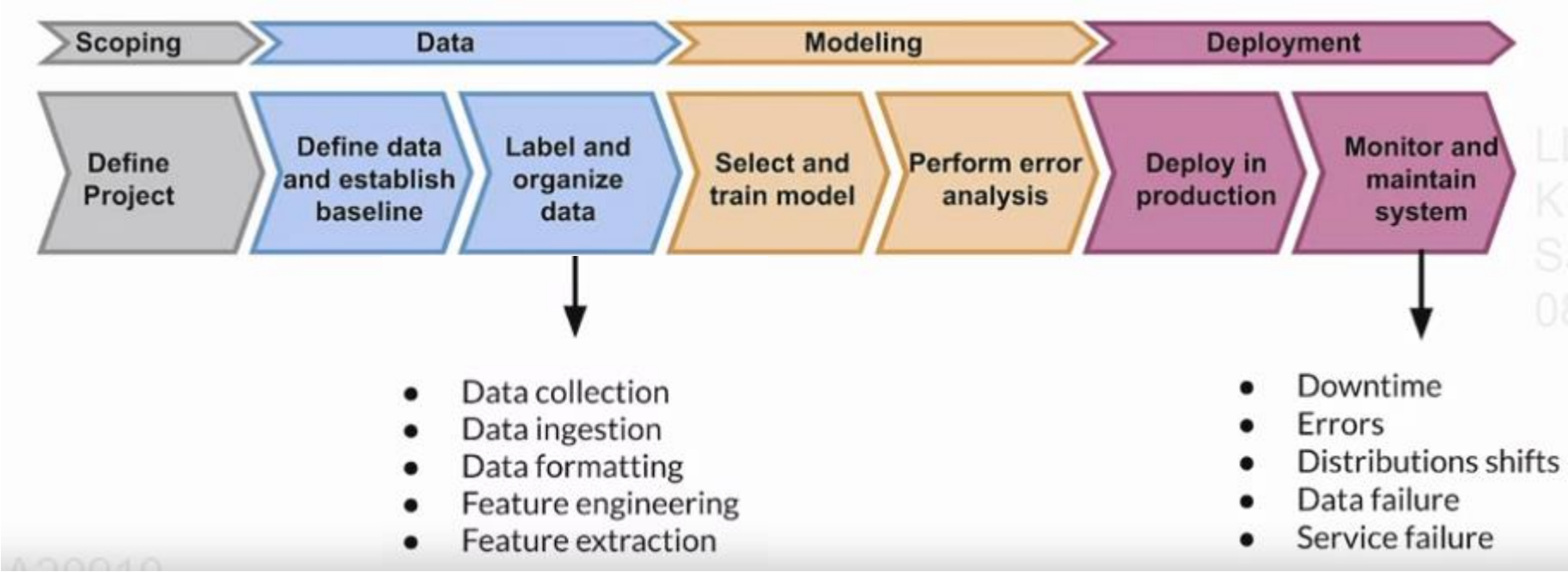
Models are not magic.
Garbage in, garbage out.



Ref: https://medium.com/whattolabel/data-labeling-ais-human-bottleneck-24bd10136e52

Importance of Data

- **Data pipeline: data collection, ingestion and preparation**
- **Data collection and monitoring**



data pipeline | modeling | deploy | monitoring | retrain

Machine Learning Project - Boston Housing

MEDV?：自住房的平均房價，以千美元計。

CRIM： 城鎮人均犯罪率。
ZN： 住宅用地超過 25000 sq.ft. 的比例。
INDUS： 城鎮非零售商用土地的比例。
CHAS： 查理斯河空變量（如果邊界是河流，則爲1；否則爲0）。
NOX： 一氧化氮濃度。
RM： 住宅平均房間數。
AGE： 1940 年之前建成的自用房屋比例。
DIS： 到波士頓五個中心區域的加權距離。
RAD： 輻射性公路的接近指數。
TAX： 每 10000 美元的全值財產稅率。
PTRATIO：城鎮師生比例。
B： 1000（Bk-0.63）^ 2，其中 Bk 指代城鎮中黑人的比例。
LSTAT： 人口中地位低下者的比例。

Quiz

- If we use this model to predict the current housing prices in Boston, will it be accurate? (feature space coverage)
- To predict the current housing prices in Hsinchu, will it be accurate?
- If we use the historical data of Hsinchu with the same features to build the model, will it be accurate?
- Assuming we find the corresponding features through data exploration and build a model, how do we ensure that it can be used over time? (data keeps changing over time)

ML Modeling vs Production ML

	ML Modeling	Production ML
Data	Static	Dynamic
Goal/KPI	Model accuracy	Stability of inference service
Model Training	Optimal tuning and training	Continuously assess and retrain
Challenge	Data pipeline & machine learning algorithm	All system

Deploying ML Models in Production

- **Model serving : batch inference and real-time inference**
- **Important metrics: latency, throughput and cost**

  (maximize throughput while minimizing latencies and cost)
- **There's a trade off between the model's predictive effectiveness and the speed of its prediction latency**

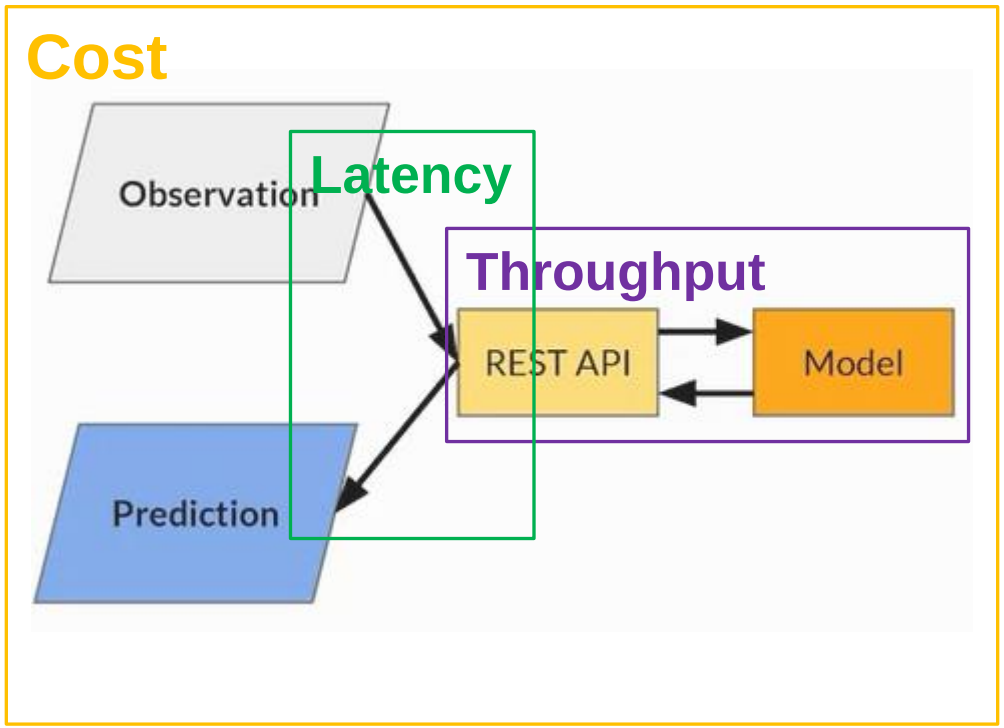
  (model complexity increases > cost increases)
- **To reduce the cost, we can have the following strategies**

  Use of serving framework for resource sharing, multi-model serving

  Use of accelerators in serving infrastructure

  Maintaining input feature lookup

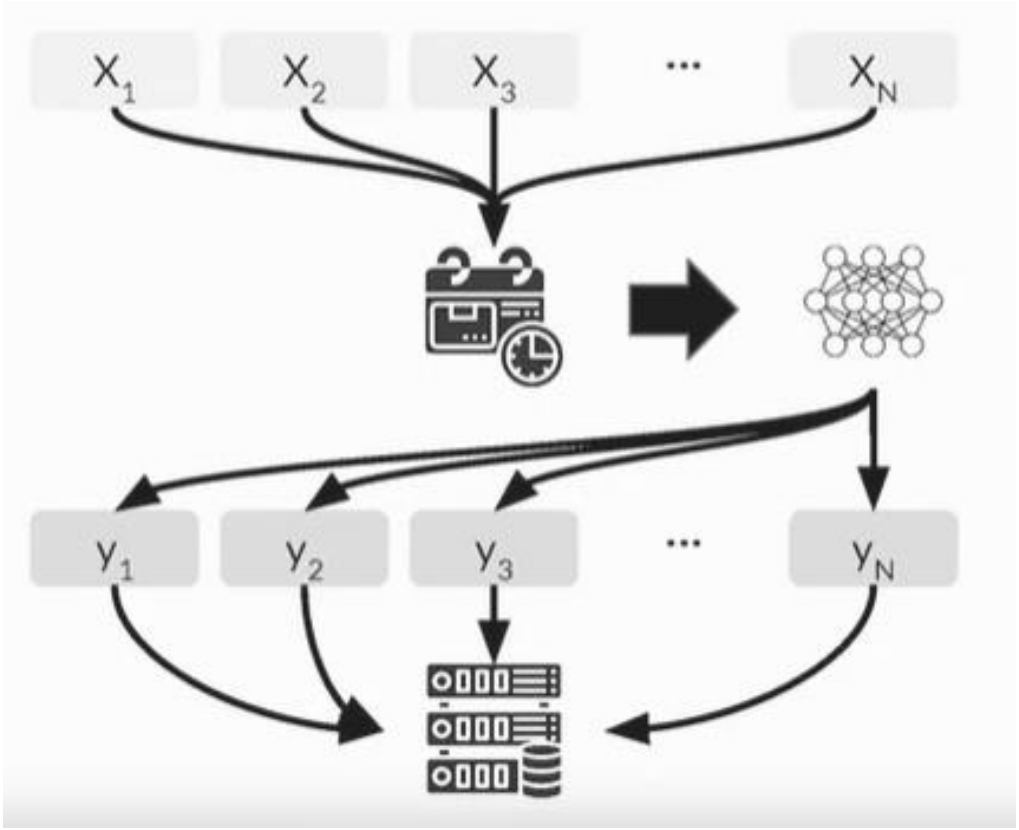
Real-time Inference



- Process of generating machine learning predictions in real time upon request.
- Predictions are generated on a single observation of data at runtime.
- Can be generated at any time of the day on demand.

data pipeline | modeling | deploy | monitoring | retrain

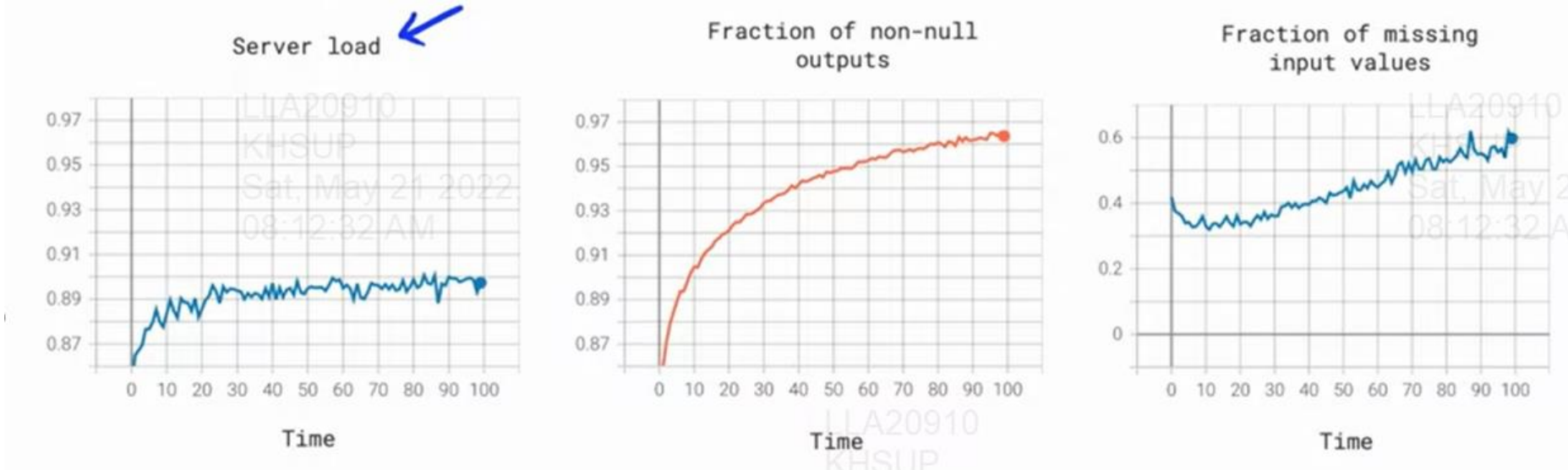
Batch Inference



- Generating predictions on batch of a observations.
- Batch jobs are often generated on some recurring schedule.
- Prediction are stored and made available to developers or end users.

data pipeline | modeling | deploy | monitoring | retrain

Monitoring Dashboard



- **Brainstorm the things that could go wrong. (HW, SW, ML)(input/output)**
- **Brainstorm a few statistics/metrics that will detect the problem.**
- **Using many metrics initially and continuous addition or removal.**

data pipeline | **modeling** | **deploy** | **monitoring** | **retrain**

Model Monitoring

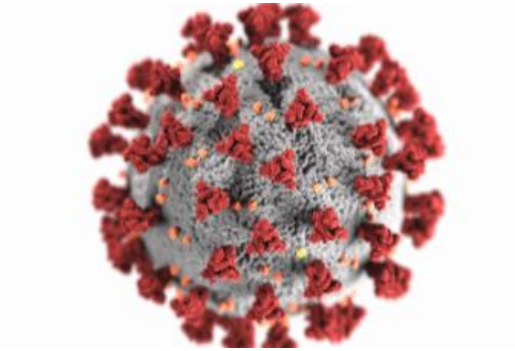
Basic: Input and output monitoring

- **Model input distribution(changes that may be associated with failures)**
- **Model prediction distribution (class imbalance or fairness issue)**
- **Model versions (how different versions perform)**
- **Input/prediction correlation(detect how changes in your inputs cause prediction failures)**

Logging for ML Monitoring

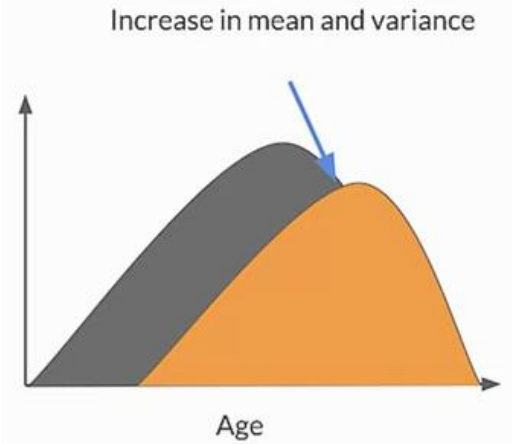
- **Log: is an immutable time stamped record of discrete events that happened over time. (including debugging and profiling messages)**
- **Logging advantages/disadvantages**
  - **(O) Easy to generate, providing valuable insight and focus on specific events (to help with root cause analysis)**
  - **(X) Impact system performance, aggregation can be expensive and setting up and maintaining this tooling carries with it a significant operational cost. (more cost)**

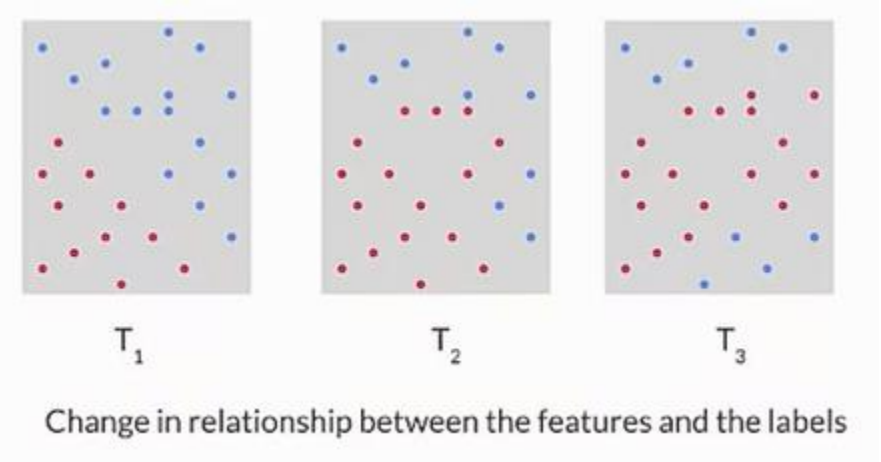
Model Decay



COVID-19 lockdowns in March 2020, which abruptly changed population behaviors all over the world.

- **Causes of model decay: data drift ,concept drift**
- **Data drift – statistical properties of input changes**
- **Concept drift – relationship between features and labels changes.**





data pipeline → modeling → deploy → monitoring → retrain

Detecting Drift

- **Detect data drift using unsupervised statistical methods**

  (compare your current data with your previous training data)
- **Using dashboards to monitor for trends and seasonality over time**
- **Continuous evaluation regularly sample's prediction input and output from deployed machine learning models**

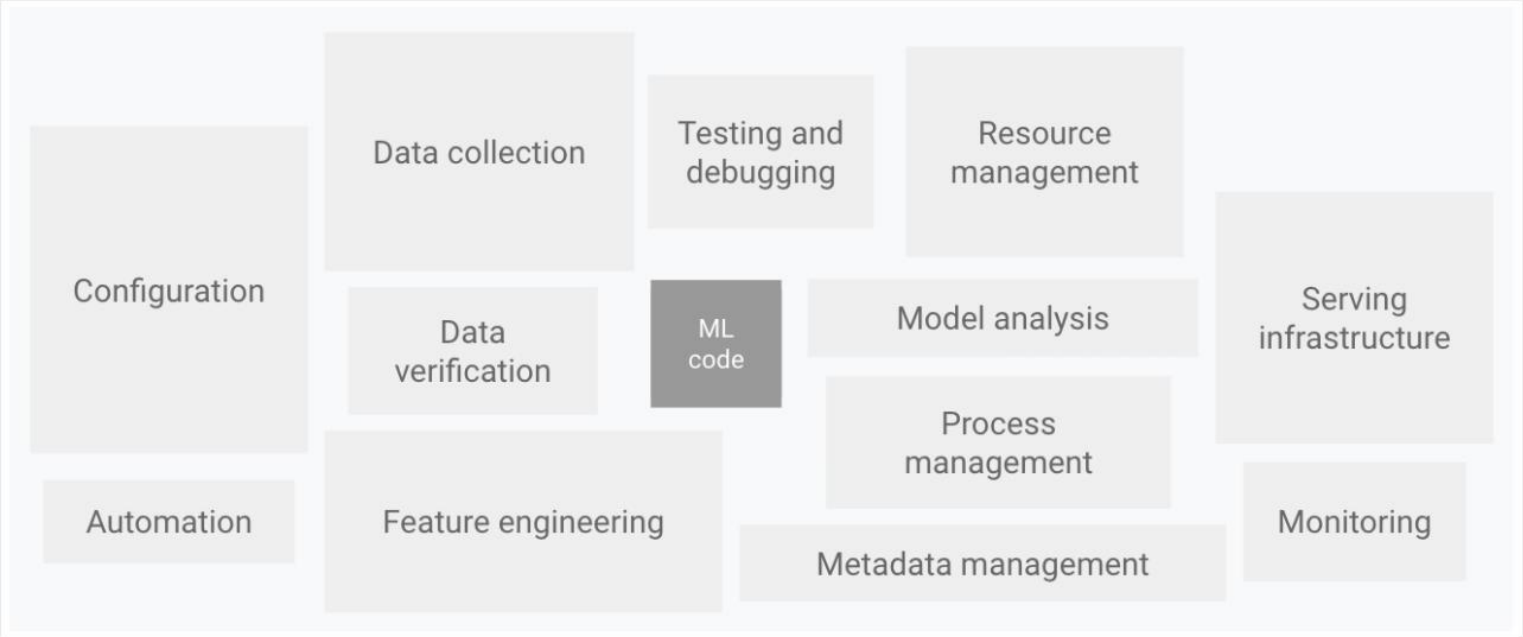
  (collect model's predictions with the ground truth to provide continual feedback on how well your model is performing over time)

Mitigate Model Decay

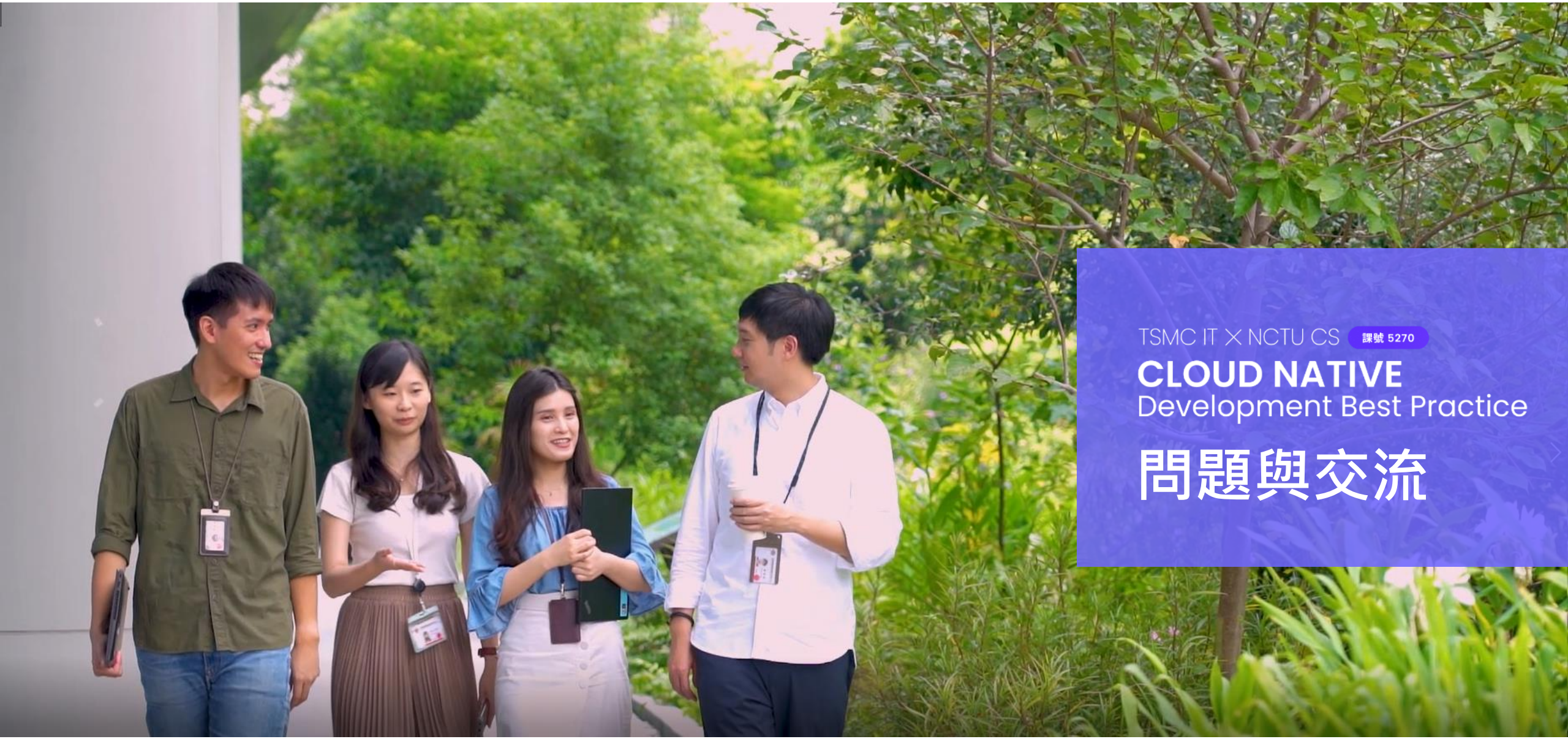
- You can either continue training your model using new data or reinitialize(re-build) your model
- Either approach is valid and depends on
  - How much new labelled data do you have?
  - How far has it drifted?
- Model retrain policy
  - On-demand (model decay): manually retrain your model
  - On a schedule :new labelled data is available at a daily, weekly..
  - Availability of new training data

Summary

- **In real-world, ML code is only a small fraction of ML system.**
- **The surrounding elements required are huge and complex.**
- **To develop and operate ML system, you must practice and build an MLOps environment to ensure the quality of you system through CI, CD, and CT.**



Ref: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning



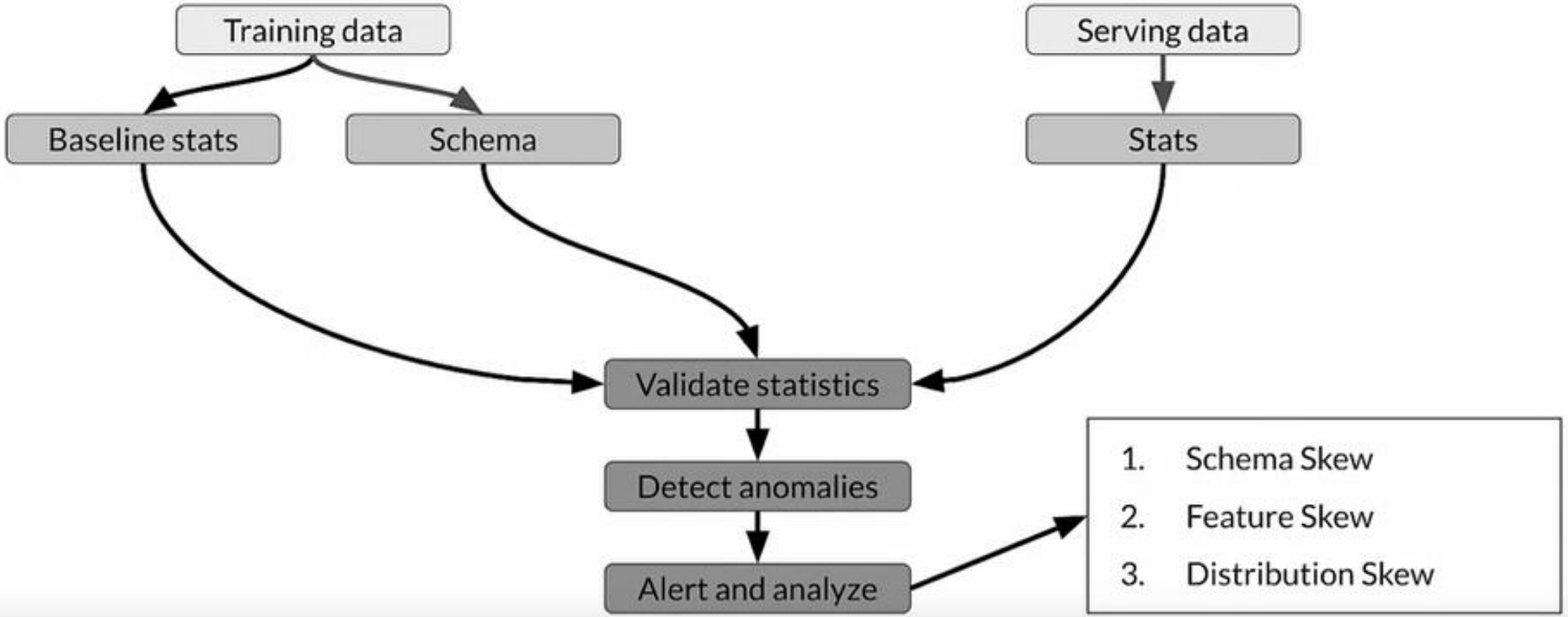
TSMC IT × NCTU CS 課號 5270

CLOUD NATIVE
Development Best Practice

問題與交流


TSMC IT × NCTU CS
課號 5270
CLOUD NATIVE
Development Best Practice
THANK YOU
FOR YOUR
ATTENTION

Skew detection - TFDV



Detecting Data Issues

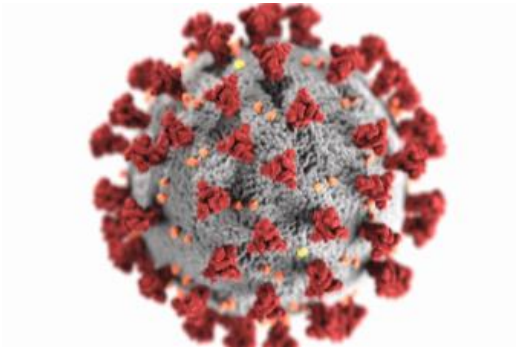
- **Two types of data issues: Drift(資料飄移) and Skew (資料偏斜)**
  - **Drift 資料隨著時間漸進或是突然的、規律的改變**
    - **Data drift : 輸入的數據 x 本身的改變.**

      → 原本 ML 系統在舊數據效果良好，伴隨著未知的數據出現，在未知數據區域表現效果差，ML 系統預測效果衰退。
    - **Concept drift : x 與 y 關係改變**

      → 當目標變量本身的統計屬性發生變化時，就會發生概念漂移。關係跟分布已經改變，模型需要更新。
  - **Skew**

    → 指兩個不同版本的資料，比較之下發生了 Schema 、特徵 x 及資料分布偏斜的差異。

Face ID to work with face masks



COVID-19 lockdowns in March 2020, which abruptly changed population behaviors all over the world.

Ref: https://ithelp.ithome.com.tw/m/articles/10261664

Concept Drift and Data Drift

Speech recognition example

Training set :

Purchased data, historical data with transcripts

Test set:

Data from a few months ago

How has the data changed?

Concept Drift: X→Y
Data Drift: X

TSMC IT & NCTU CS Meet up

- **Date/Time/Venue: 11/12(Fri) 14:30~15:30@F12 P1#924**
- **Attendees : TSMC IT x7, NCTU x7**

TSMC IT

- IT 林宏達 CIO
- BSID 林均彥 處長
- TSID 陳儒寬 副處長, 胡君怡 部經理
- AAID 沈文冰 副處長
- ICSD 陳守文/鄔國民 副處長
- 謝冬青 HR 經理

NCTU

- 曾煜棋院長 交大AI 學院/PAIR 中心主任
- 陳添福 交大資訊學院副院長
- 黃敬群 國際學位學程副主任
- 彭文志 交大資工系主任
- 黃俊龍 數據工程與科學研究所長
- 范倫達 資工系/電子資訊中心副主任
- 范瑀真 資工系 宣傳

TSMC IT × NCTU CS
2022 CLOUD NATIVE
Development Best Practice

Kubernetes on Personal Computer

Run local Kubernetes clusters using Docker Desktop on PC/Laptop.

TSMC IT will provide SOP for creation procedures.