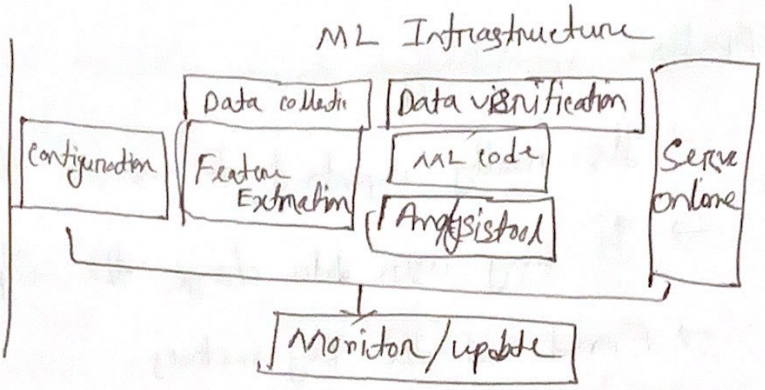
pipeLines → W1

→ Data drift problem



MLOps Lifecycle

Scoping → Data → Modelling ⟶ Deployment

Define project

Define Data

Label & organize

train mode

Error Analysis

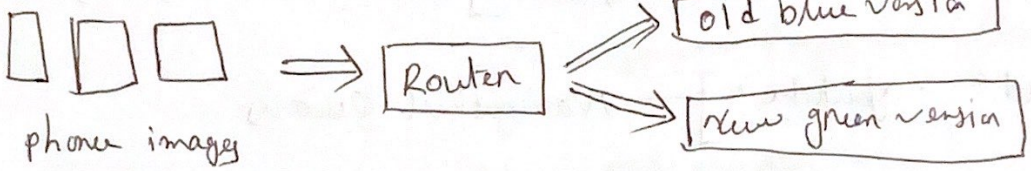
Deploy

monitor & maintainance.

Data Drift → your data change after your model deployed.

Canary development → deploy for a subset of user.

Blue green development



Monitor metrics | Moniton dashboard

→ Set thresholds for alarms

→ Adapt metrics and threshold over time.

~~Manual~~

Manual retraining

Automatic retraining

→ pipeline monitoring

→ Its really important to monitor the metrics.

→ If any user data change the output will change.

→ Find out the Key metrics.

W2

→ Model - centric AI development → NN models
→ Data - centric AI development → Dataset

⇒ AI System ⇒ Code/model + Data

→ web search example

"Apple pie Recipe" "Latest movie"] ⇒ Informational & Transactional Queries
"Wireless data plan" "Eid Festival"]

"Stanford" "Youtube"] — Navigational Queries

⇒ Sanity check for code & Algorithm
↳ try to use small amount of data at first

⇒ Error Analysis → → Add/improve data for specific categori
→ Skewed dataset (imbalance data)
→ Try to Brain storm what could go wrong?

Data Augmentation → create realistic example that human can check easily.

Experiment tracking →
- Algorithm / code
- Dataset used
- Hyperparameters
- Results

Tracking tools →
- Test file
- spread sheet
- Experiment tracking System
  - Tools
    - weights and biases
    - → Comet
    - → MLFlow
    - → Sage maker studio

F1 Score recommended for skewed dataset

W3 → Data

Data
- → Unstructured data
- → Structured data

⇒ Data and Label consistency

⇒ Human Level performance (HLP) → Estimate Bayes Error/irreducible error

↳ Beat the HLP is good for research purpose but for production its better to improve HLP so that model outperformance increase

#Data → Don't increase data by more than 10x at time

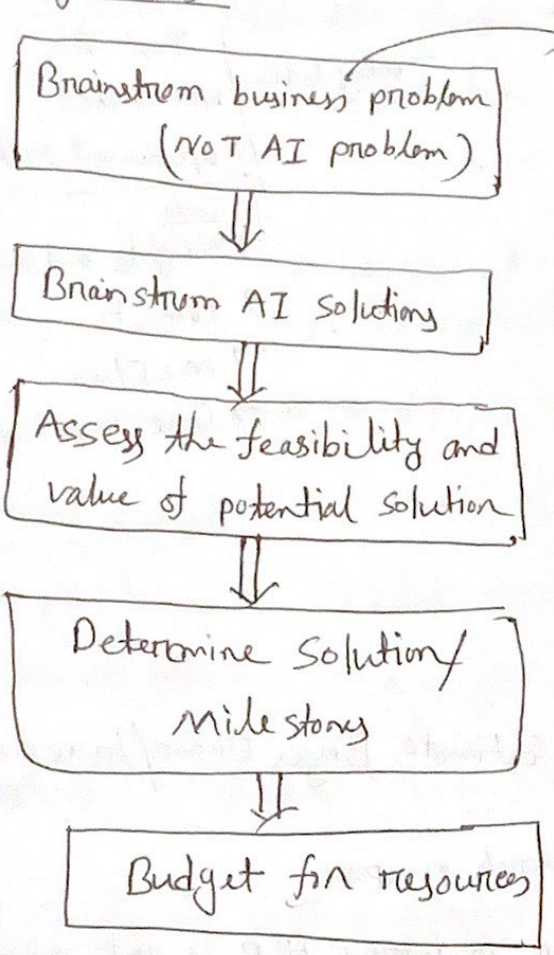
→ POC (proof of concept)
↳ Tensorflow Transform, Apache Beam, Airflow

meta-data, data provenance and lineage ⇒ Data pipeline

↳ File types, origin points — where data comes from — sequence of steps.

- A high HLP metric implies good label consistency.

Scoping process



→ Some question to ask
→ what thing you wish were working better
- Increase Conversion
- Reduce inventory
- Increase margin (profit)

→ Make sure your feature map to prediction

↳ → give past purchase, predict future purchase $x \to y$

X → Given DNA info, predict heart disease $(x \not\to y)$

→ Query-Level accuracy

Specialization #2

→ Machine learning Data Lifecycle

TF MetaData (w1)

↳ MLMD is a library to track of full lineage of entire ML like data ingestion, preprocessing, validation, training, evaluation, deployment and so on.

POC (proof of concept):

→ Goal is to decide if the application is workable and worth deploying

→ Focus on getting the prototype to work

→ pre-process data manually if needed. But take extensive notes/comments

Machine Learning pipeline (ML pipeline)

→ Pipeline Automation → Airflow, Argo, Celery, Luigi, KubeFlow

TensorFlow Extended (TFx)

E2E platform for ML pipeline

Data ingestion → Data validation → Feature Engineering → Train Model → validate Model → push if good → serve model.

* TensorFlow Data validation (TFDV) *

Chebyshev Distance

$$D_{chebyshev}(x, y) = \max_i (|x_i - y_i|)$$