Machine Learning Workflow

Michell S. Handaka

FOUNDER & CEO

michellsh

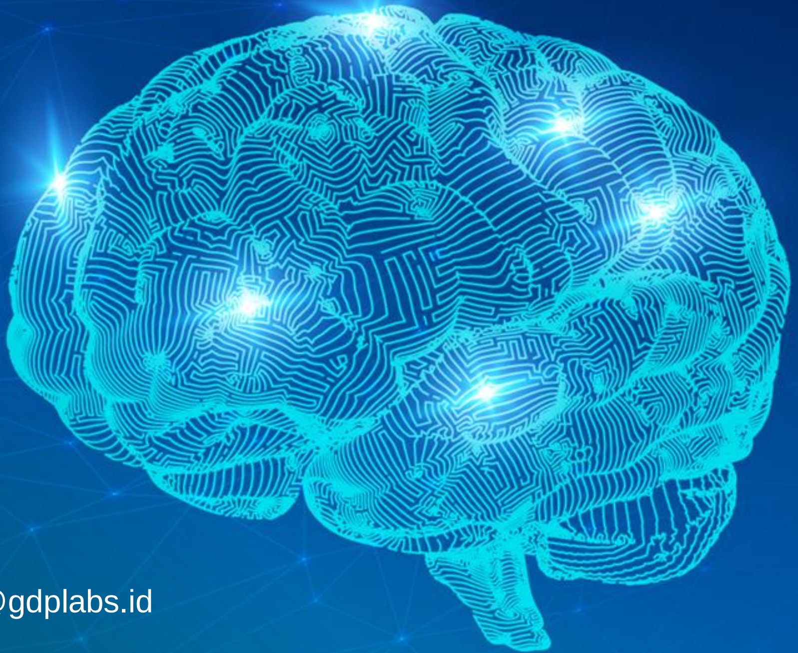
michell.s.handaka@glair.ai

Kevin Yauris

AI ENGINEER

kevinyauris

kevin.yauris@gdplabs.id



Contact Us glair | glair.ai | hi@glair.ai

OUTLINE

Introduction to GLAIR

01

glair Products and Services

glair Consulting Service

On demand talents capable of crafting customized solutions

glair Analytics Platform

An easy & intelligent way to turn your data into insights

glair Training Center

We are what we invest

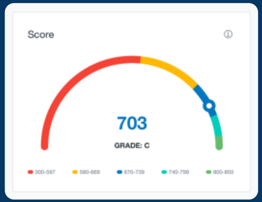
glair Consulting Service

Fraud Detection System

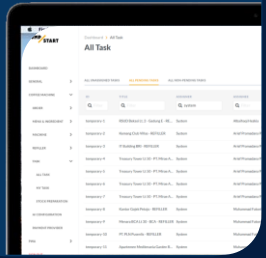
White Label Chat Platform

Chatbot

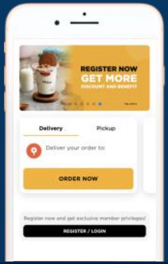
Credit Score



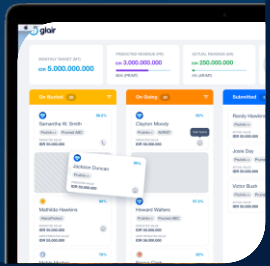
Refiller Assignment



PWA Ordering Online System

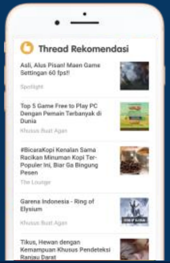


Propensity Modeling



glair Consulting Service

Recommendation System



OCR

Intelligent Extraction

Planogram Analytics

Sentiment Analysis & Topic Modeling

License Plate Recognition

Face Recognition

...

Launching an AI initiative

02

6 Preliminary Steps Before Modeling

01
Define Objectives

02
Set Expectations

03
Understand the Data

04
Translate the Business Problem Into an AI Problem

05
Determine the Development and Deployment Type

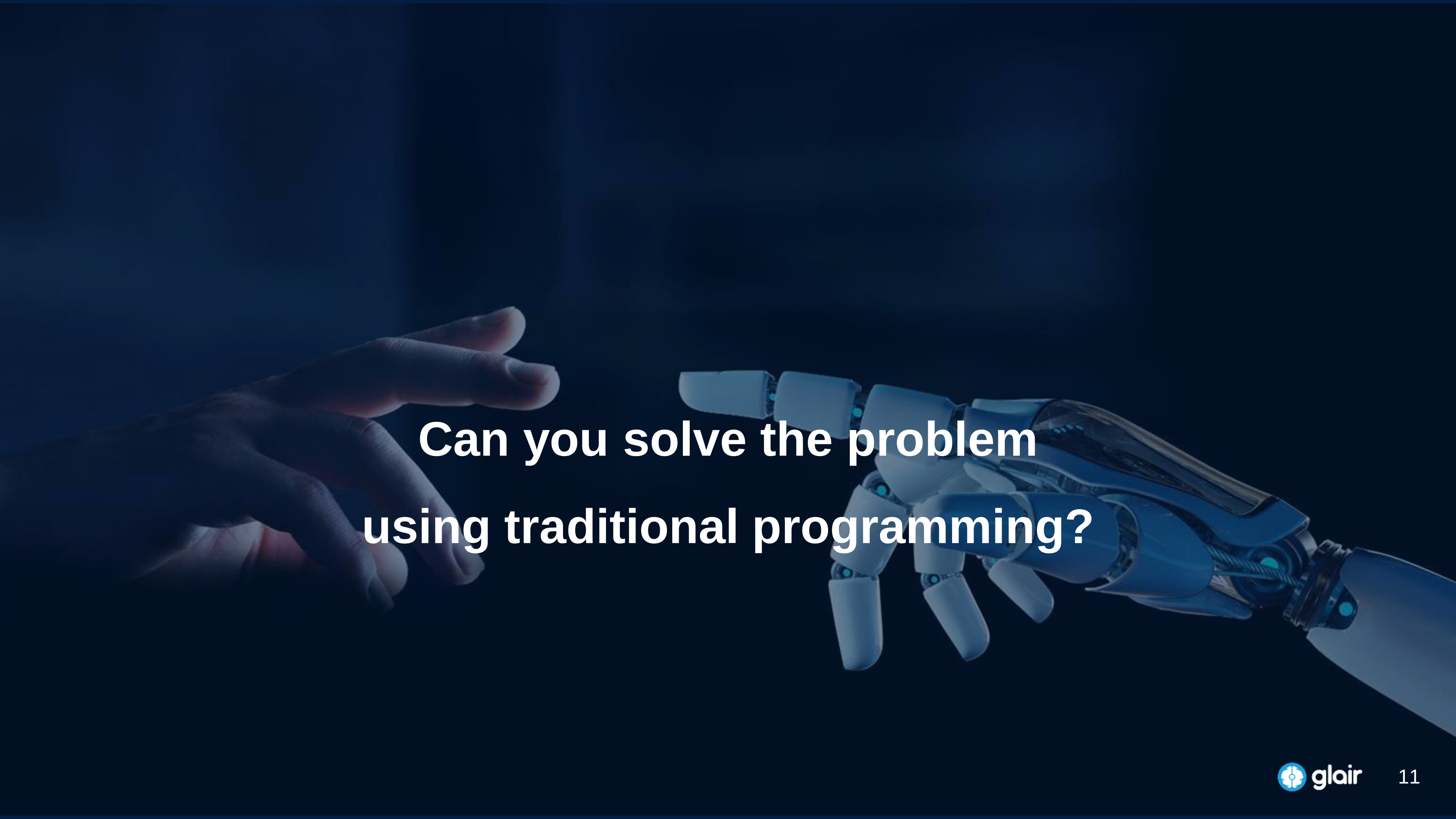
06
Set Success Criteria

01 Define Objectives

- Problems vs opportunities
- Increase revenue vs reduce cost
- Improvements vs innovations
- Cost and time vs differentiation

02 | Set Expectations

- Do you know the use cases in your industry?
- Have others approached and solved the problem using AI before?
  - What are the state of the art algorithms in the industry?
  - How about in academic setting?
  - What libraries are available?
  - What are the challenges others overcome?



Can you solve the problem using traditional programming?

Set Expectations

- Calculate your resources
  - Time
  - Engineers
  - Computing power

03 Understand the Data

- What data do you need and what data do you have?
- Do you have enough?
  - For image classification ~1000 images for each class
  - For structured data classification at least 100 rows, 10 rows for each class
- Where do you store the data and how do they flow?
- What is your data quality?

Data Size	Library
Big Data	Apache Spark, dmlc XGBoost
Small Data	pandas, scikit learn

03 | Understand the Data

What type of data do you have?

- Structured (Tables)
- Unstructured (Image, Text, Speech)

The data type will affect the algorithm and library choices (including data processing)

Data Type	Algorithm	Library	SaaS
Structured	Logistic Regression, Decision Tree Family, GLM	Pandas, Apache Spark, Scikit-learn, XGBoost	Google AutoML Table, Microsoft Azure Machine Learning Studio
Image	CNN (Inception, ResNet, ...)	Tensorflow, Keras, OpenCV	Konvergen, Amazon Rekognition, Google Cloud Vision/AutoML, Microsoft Azure Computer Vision
Text	RNN, Transformer (BERT, GPT, ...)	Tensorflow, Keras, Spacy, Gensim	Prosa.ai, Amazon Comprehend, Google Cloud Natural Language/AutoML, Microsoft Text Analytics

04 Translate the Business Problem into an AI Problem

Business Problem	AI Problem
Credit Scoring	Binary Classification
Propensity Model	Binary Classification
Fraud Detection	Binary Classification
Demand Prediction	Regression
Sentiment Analysis	NLP: Text Classification
Face Recognition	CompVis: Image/Face Classification

04 Translate the Business Problem into an AI Problem

Business Problem	Business Metrics	ML Metrics
Credit Scoring	Revenue and NPL Rate	F1 Score
Propensity Model	Lift Metrics	PR-AUC
Fraud Detection	Precision and FP	PR-AUC
Demand Prediction	#Out of stock order	RMSE

05 Determine the Development and Deployment Type

- PoC vs production
- On-premise vs cloud
  - Affects library choices (e.g. cannot use cloud Saas on-premise!)

06 Set Success Metrics

- Useful for measuring progress
- Go for reasonable metrics (current state of the art)

Study Case: Propensity Modeling

for Bancassurance

03

Need to understand existing customers segmentation

Need to match right offers with right customers

INPUT

- Customer Demographic Data
- Product Specification Data
- Interaction Data: Transaction Data
- Contexts Data: Promo Events
- External Data: Mortality Table

OUTPUT

- Customer Segmentation
- Lead Generation (Classification)
- Premium Suggestion (Regression)

Business

Data Information

CUSTOMER & TRANSACTION DATA FOR THE LAST 18 MONTHS

- **Customer**
- **Interaction**
- **Transaction**
- **External**
- **Product**

Demographic data of existing bancassurance customers: Credit card, auto finance, mortgage, a “main” table that shows customers’ attributes

SAMPLE CASE

Current month : March 2021

Training data
September 2019 - February 2021

↓

Leads data
April 2021

For example:
If it is March 2021 and we have data for the last 18 months, then September 2019 - February 2021 will be used as our training data, and leads data will be obtained for April 2021.

Business Metrics

Lift Metric

Lift metric is a measure **comparing** the **relative performance** of a **propensity model** vs. a **random guess**

Lift Metric for Propensity

Sample data:

Rank	Actual	Approached?
1	**1**	v
2	0	v
3	**1**	v
4	**1**	v
5	0	v
6	0	x
7	1	x
8	0	x
9	0	x
10	0	x

Bucket size is number of customers that will be approached and determined according to the company's capabilities

The lift metric is defined as the ratio of model score to random guess score

1. **There are 10 customers**
2. **The model will rank these customers based on their probability to convert (higher rank indicates higher chance to convert)**
3. **Say we decide to approach 5 customers (the bucket size is 5)**
4. **After approaching the top 5 customers, only 3 of them converted**
   In this case, our model's score is: 3/5 = 0.6
5. **Assuming that we approach all 10 customers (random guess), we will have 4 converting customers. So the random guess score is: 4/10 = 0.4**
6. **The model's lift is then :** $\frac{0.6}{0.4} = 1.5$

glair

Business

Benefits

Automated Pipeline

to create propensity models and make predictions anytime

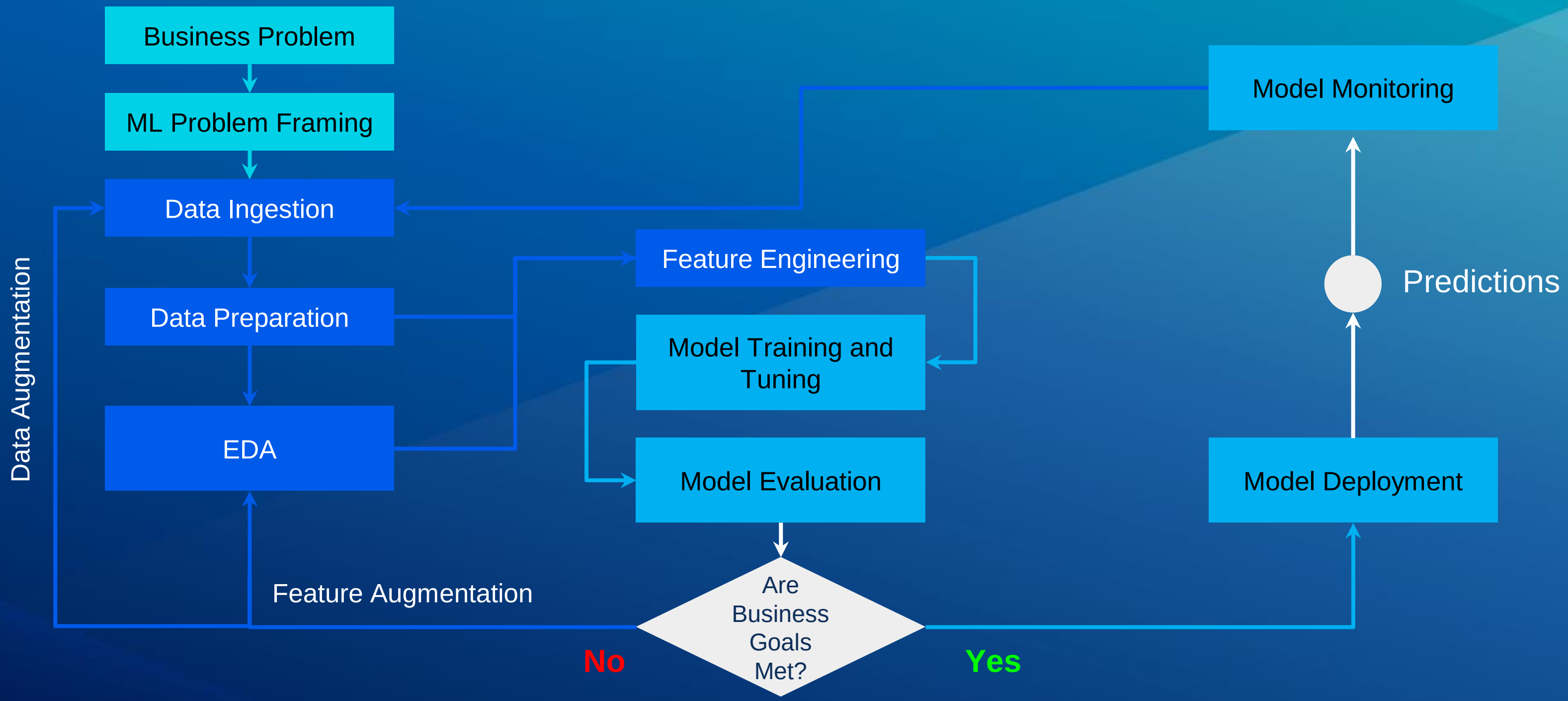
Robust

to changes in data distribution or behavior

Extensible

for new data or custom processes to be added easily

Machine Learning Workflow



Machine Learning Workflow

Business Problem

↓

ML Problem Framing

↓

Data Ingestion

01 Data Ingestion

- **Transportation of data from assorted sources** to a **storage medium** where it **can be accessed, used, and analyzed** by an organization.
- The destination is **typically a data warehouse, data mart, database, or a document store**.
- **Sources may be almost anything.**

01 | Data Ingestion Type

Batch processing

the ingestion layer p**eriodically collects and groups** source data
and sends it to the destination system.

- Groups may be processed based on any **logical ordering**,
  the activation of **certain conditions**,
  or a **simple schedule**.

Real-time processing

(stream processing or streaming) involves no grouping at all.

- Data is sourced, manipulated, and loaded **as soon as it's created or recognized by the data ingestion layer**.
- This kind of ingestion is **more expensive**, since it requires systems to **constantly monitor sources and accept new information.**

Machine Learning Workflow

Business Problem
↓
ML Problem Framing
↓
Data Ingestion
↓
Data Preparation

02 Data Preparation

- Step in which the **data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse** it.
- Purpose:

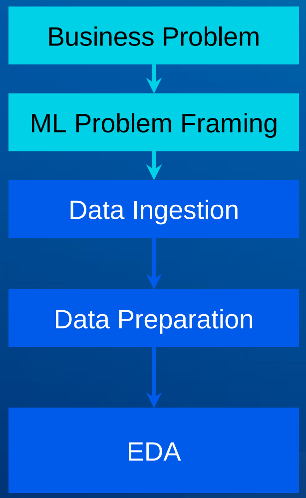
  The features of the data can now be easily interpreted by the algorithm.
  - Data cleansing or data cleaning:
    - **detecting and correcting** (or removing) corrupt or inaccurate records from a record set, table, or database
    - identifying **incomplete, incorrect, inaccurate or irrelevant parts** of the data and then replacing, modifying, or deleting the dirty data
  - **ETL (Extract, transform, load)**

02 Data Preparation - Extract, Transform, Load (ETL)

Extract, Transform, Load (ETL)

- Extract: step where **sensors wait for upstream data sources to land**. **Transport the data from their source locations to further transformations.**
- Transform: **apply business logic and perform actions such as filtering, grouping, and aggregation** to translate raw data into **analysis-ready** datasets.
- Load: **load the processed data** and **transport them to a final destination.**

Machine Learning Workflow



03 Exploratory Data Analysis (EDA)

- Exploratory Data Analysis

  Approach of analyzing data sets to **summarize their main characteristics**, often using **statistical graphics and other data visualization** methods.
- The objectives of EDA are to:
  - **Suggest hypotheses**
  - **Assess assumptions**
  - **Support the selection of appropriate** tools and techniques
  - **Provide a basis for further data collection**

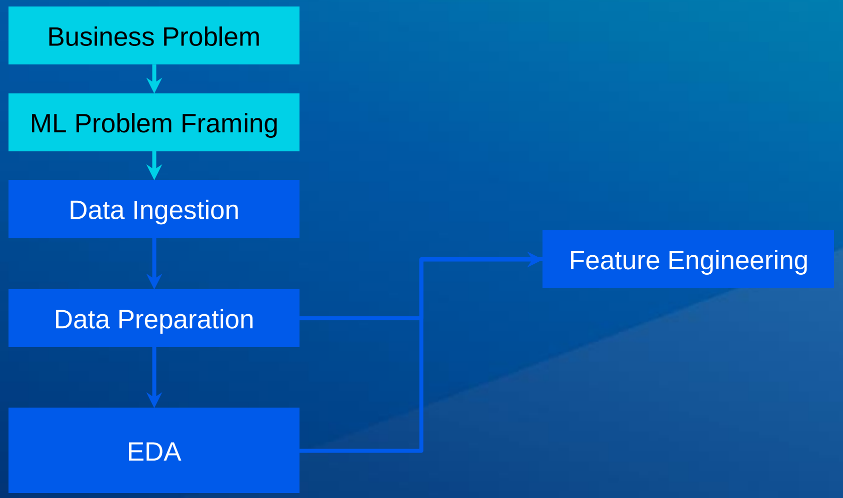
03 Exploratory Data Analysis (EDA) Technique and Tools

- Graphical / Visualization techniques
- Dimensionality reduction
- Numerical summaries
- Statistical testing

03 Exploratory Data Analysis (EDA) Visualization Techniques

Source: Developia.org

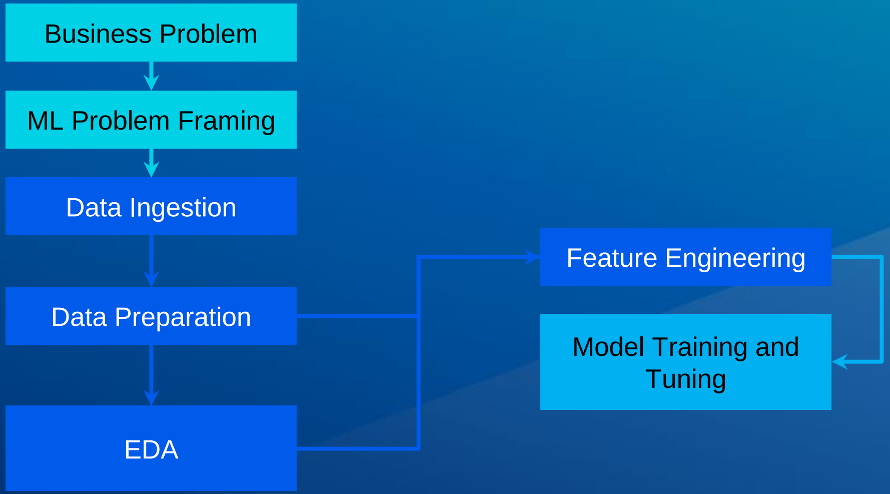
Machine Learning Workflow



04 Feature Engineering

- Feature engineering is the **process of using domain knowledge to extract features** from **raw data**
- Feature engineering efforts mainly have two goals:
  - **Preparing the proper input dataset**
  - **Improving the performance** of machine learning models.
- There is a lot the of feature engineering, some of them is:
  - Imputation
  - Handling outliers
  - Binning
  - Log transformation
  - Grouping operations

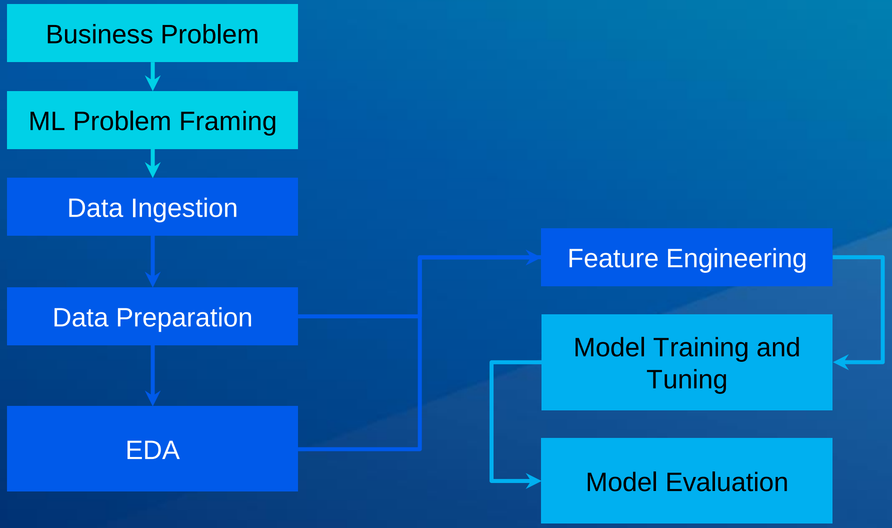
Machine Learning Workflow



05 | Model Training and Tuning

- Choose which model **to try**
- Better **try a simple model first with minimal feature engineering** and see how well it perform
- **Iteratively using a more complex model and features if needed**
- Having a **baseline model** to **make sure each complexity added into the model is worth to have**
- Consider **other constraints:**
  - **cost, explainability, and speed**
- **Maximize model performance** by doing **hyperparameter tuning**

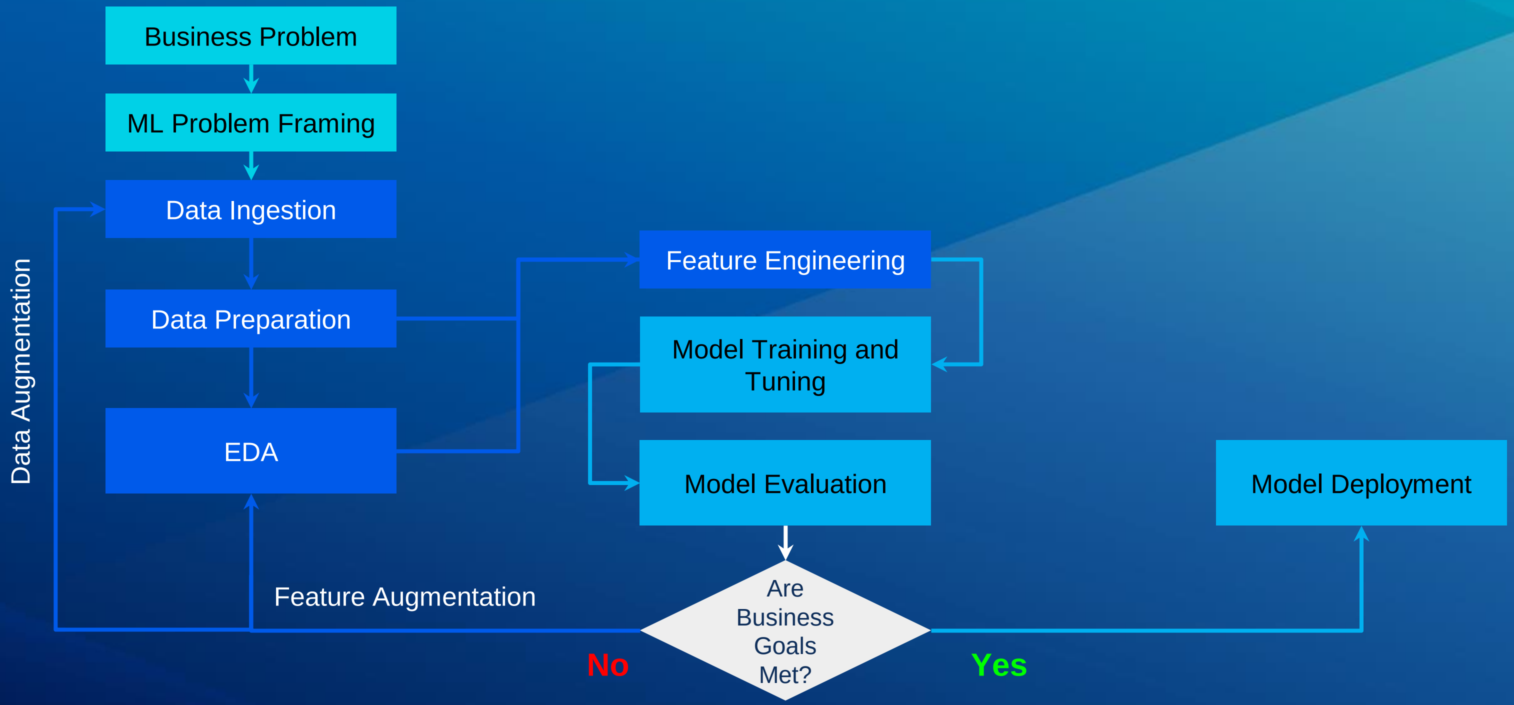
Machine Learning Workflow



06 Model Evaluation

- **Evaluate** the **model** with the **test data**
- Pick a **suitable metrics** for the problem.
  - There must be a **ML and business** metrics for the problem
- If the **data is unbalanced, precision or accuracy is not a suitable** metrics
- If the business metrics achieved then continue with model deployment, **if not reiterate** the model creation with **data augmentation** (adding more data) or **feature augmentation** (adding other features)

Machine Learning Workflow



07 Model Deployment

- **Enable the model to be used for inference**
- Considerations:
    - How to **wrap the prediction code as a production-ready service**?
    - **How to ship and load** the dumped model file?
    - Which **API / Protocol** to use?
    - **Scalability, Throughput, Latency**.
    - Deployments
        - How to **deploy new model versions**?
        - How to **rollback**?
        - Can we test it using **Canary Deployments** or **Shadow Deployments**?

ML Serving Frameworks

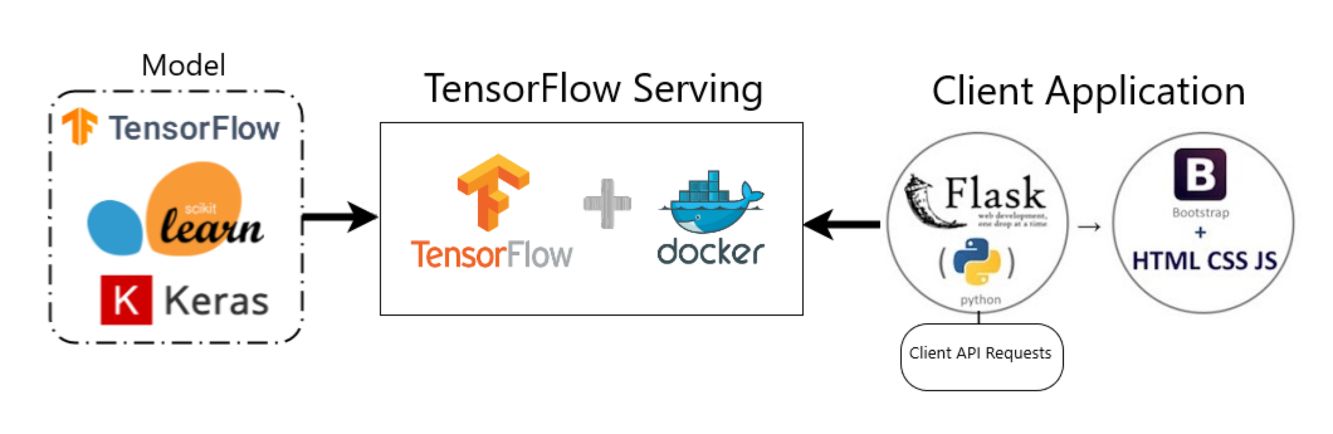
There are some several frameworks that provide solution

Just a REST API Wrapper “The K8 Model Serving Projects: KFServing and Seldon Core”	BentoML	Neptune	Tensorflow Serving
Cortex	Torchserve	KF Serving	

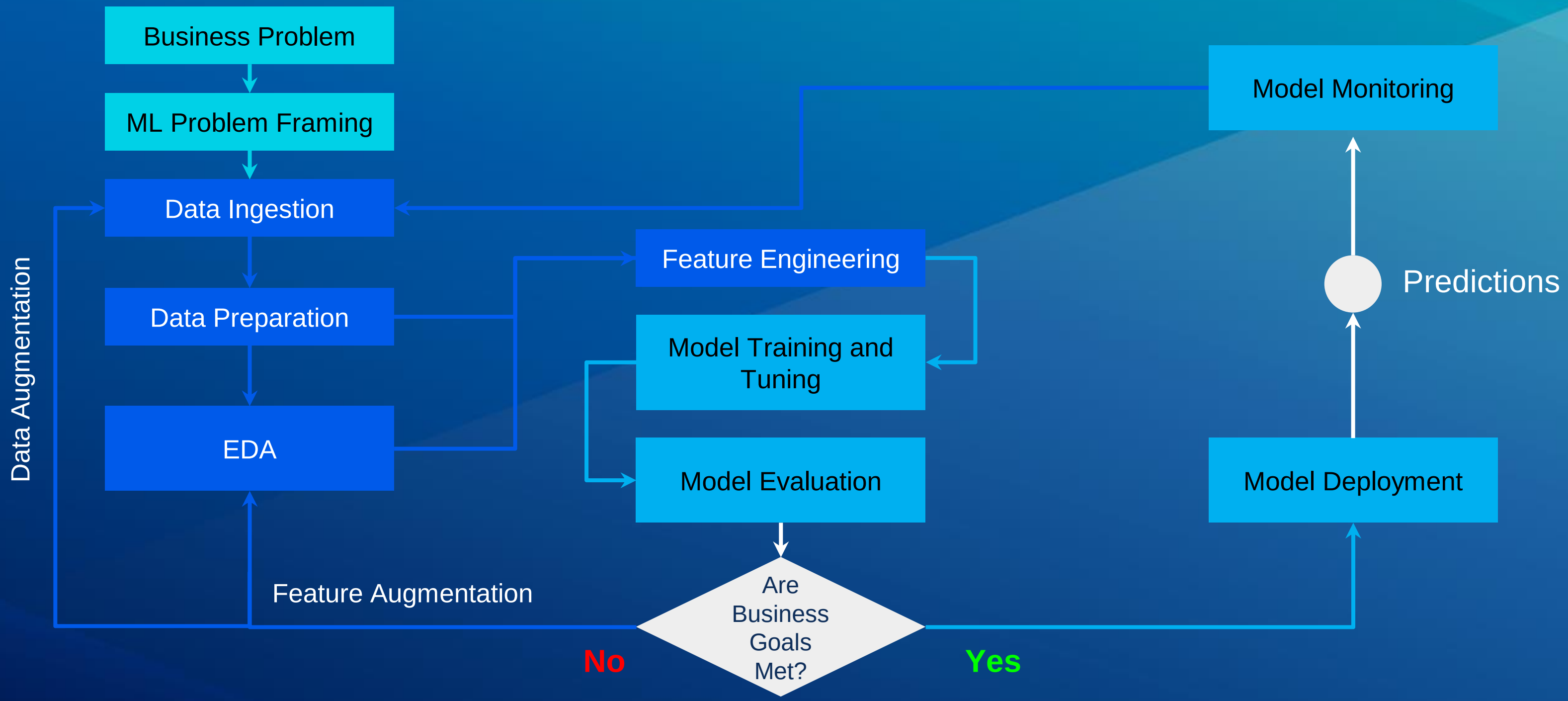
Source: Medium

TensorFlow Serving

- TensorFlow serving provides an easy integration for developers to incorporate AI in software systems and is already been used in productionizing a lot of google products.
- It can serve multiple models and multiple versions of the same model simultaneously



Machine Learning Workflow



08 Model Monitoring

- To make sure model still p**erform well when new data coming**.
- Things to be monitored:
  - **Service Health**
  - **Data Quality & Integrity**
  - **Data & Target Drift**
  - **Bias/fairness**

glair

Thank You!
Any Questions?

Michell S. Handaka

FOUNDER & CEO

michellsh

michell.s.handaka@glair.ai

Kevin Yauris

AI ENGINEER

kevinyauris

kevin.yauris@gdplabs.id

Contact Us glair | glair.ai | hi@glair.ai