2022 SPE EUROPE ENERGY GEOHACKATHON

8. Advanced Python with ML

Kapetis, Dimos
Falcone, Federico

#DatafyingEnergy

5th October 2022

ABOUT US.

Dimos Kapetis

Bio

Data Scientist Senior Manager with 15 years of advanced analytics experience with focus on Machine Learning, Natural Language Processing, predictive modelling projects.

In his current role, he is responsible for developing and running of Artificial Intelligence solutions perform data analysis and evaluate the Machine Learning application. He is co-author of 50 scientific peer-reviewed scientific publications in data analysis.

Contacts

dimos.kapetis@accenture.com

[dimos-kapetis]()

Federico Falcone

Education

Data scientist manager in Accenture. He has a Master of Science degree in Computer Science and Engineering.

In his current role, he is involved as a lead for Artificial intelligence projects to design, develop and support clients with state-of-the-art solutions for their business. He has a strong knowledge of computer vision, natural language processing and advanced analytics machine learning topics.

Contacts

federico.falcone@accenture.com

[fedefalco92]()

[federicofalcone92]()

Accenture Applied Intelligence
Capabilities in Italy AI CODE

- Natural Language Processing
- Computer Vision
- TimeSeries
- ML Ops
- Agent-based Systems & Robotics
- Graph ML
- Predictive Customer Analytics
- Synthetic Data
- Simulation & Operations Research
- Speech & Audio Processing
- Reinforcement Learning
- Recommender Systems

Accenture Applied Intelligence has capabilities to **combine** state-of-the-art **Machine Learning models** with the expertise of the **delivery** E2E application, using **cloud agnostic** and **open-source technologies**. Additionally, we work on advisor activities and on awareness and culture of ML.

70+ — **# People**
Data Scientists, Data Engineers, Solution Architects, FE & BE Developers, PMs.

50+ — **# Different Clients**
Expertise in different domains and Industries: energy, utilities, telco, financial services, health, retail.

100+ — **# ML Delivered Projects**
Across multiple Machine Learning domains since 2016.

20+ — **# Technology Partners**
From cloud providers to software integration partners.

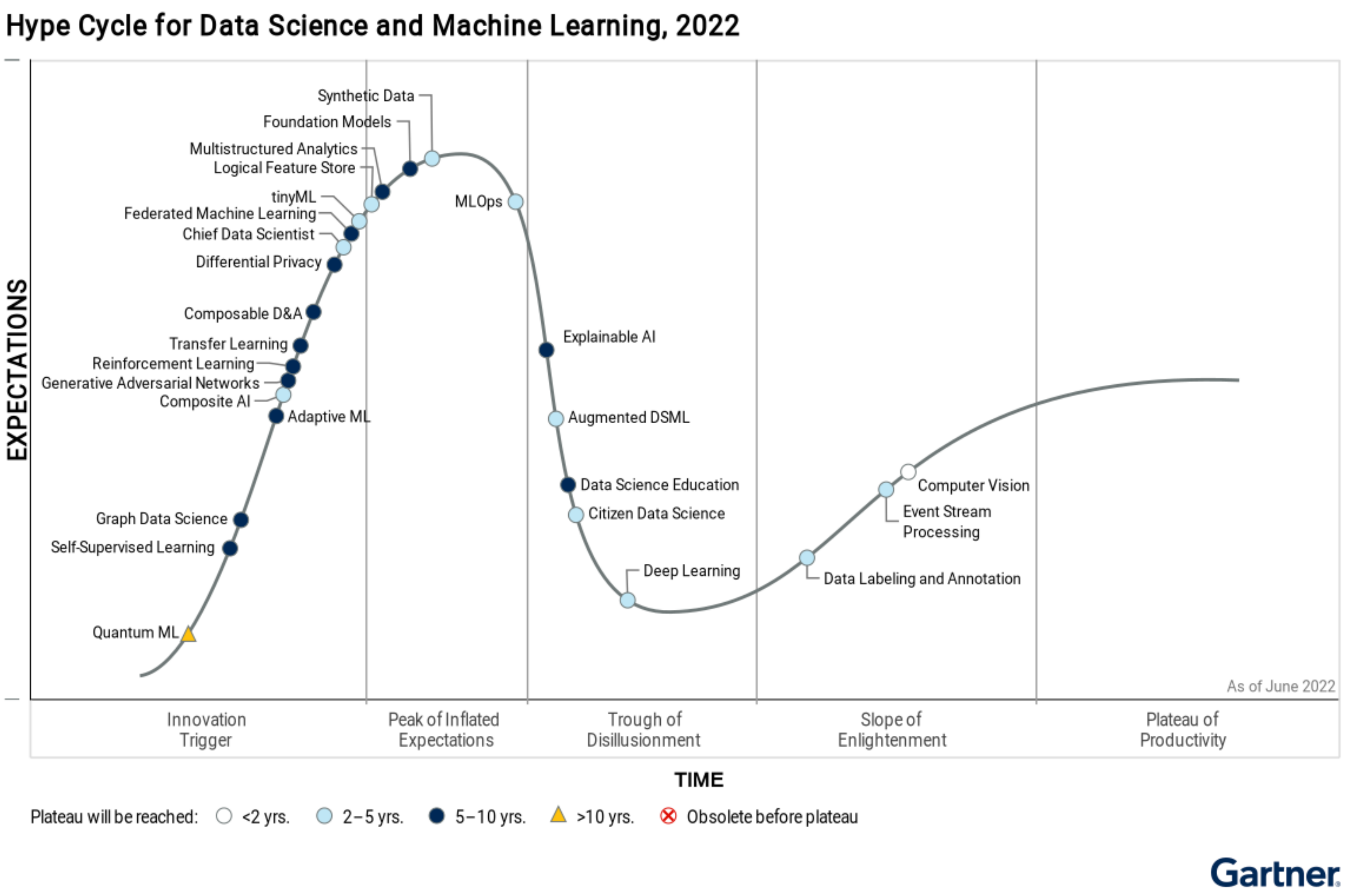
Agenda

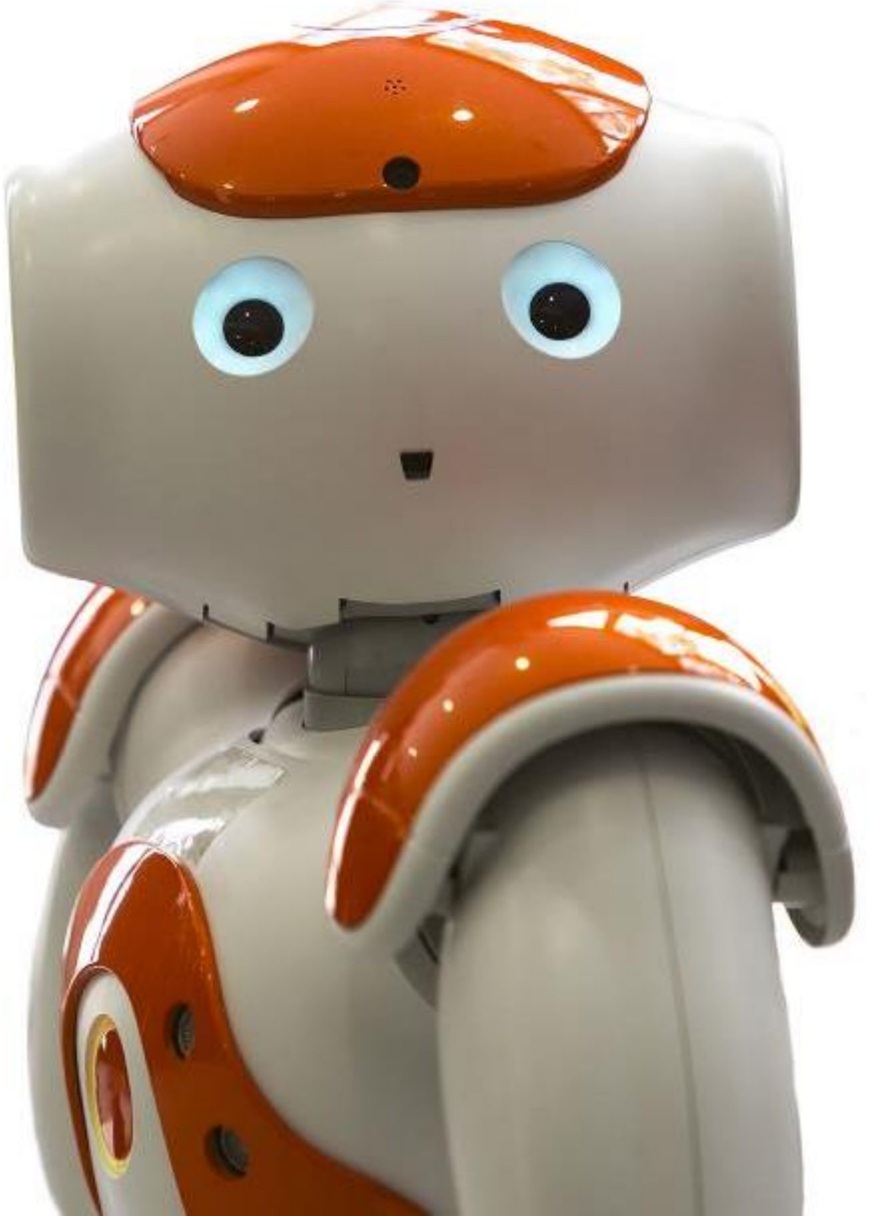
1. Machine Learning Introduction
2. Machine Learning Design Life Cycle
3. Prepare a ML Model with Python
4. Enterprise tips for python projects solution

1

Machine Learning Introduction

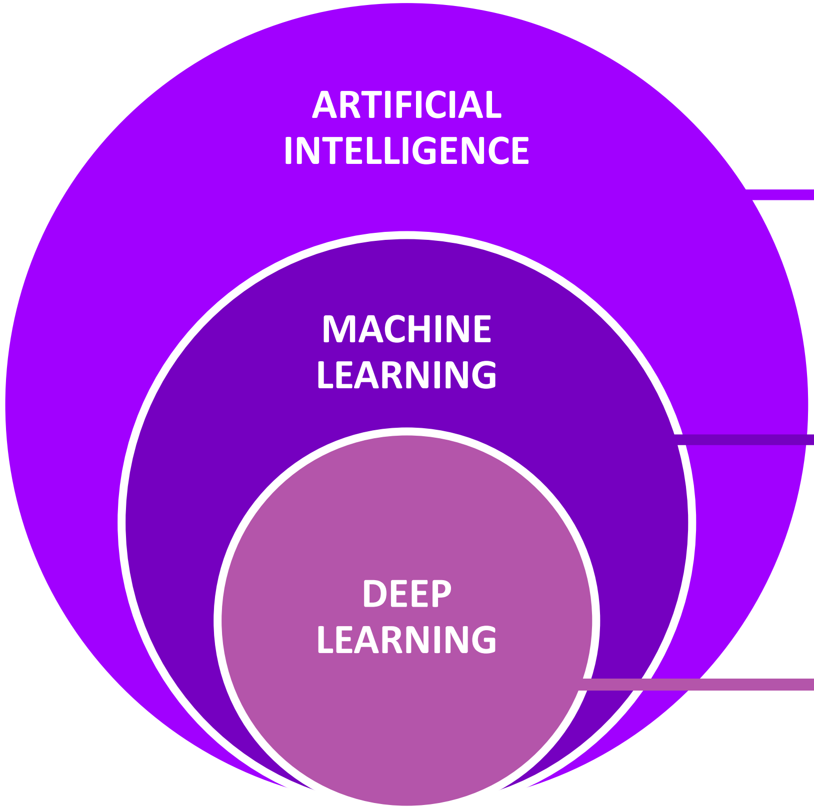
Hype Cycle for Data Science and Machine Learning





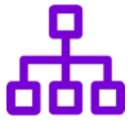
Artificial Intelligence Fields

From AI to Deep Learning



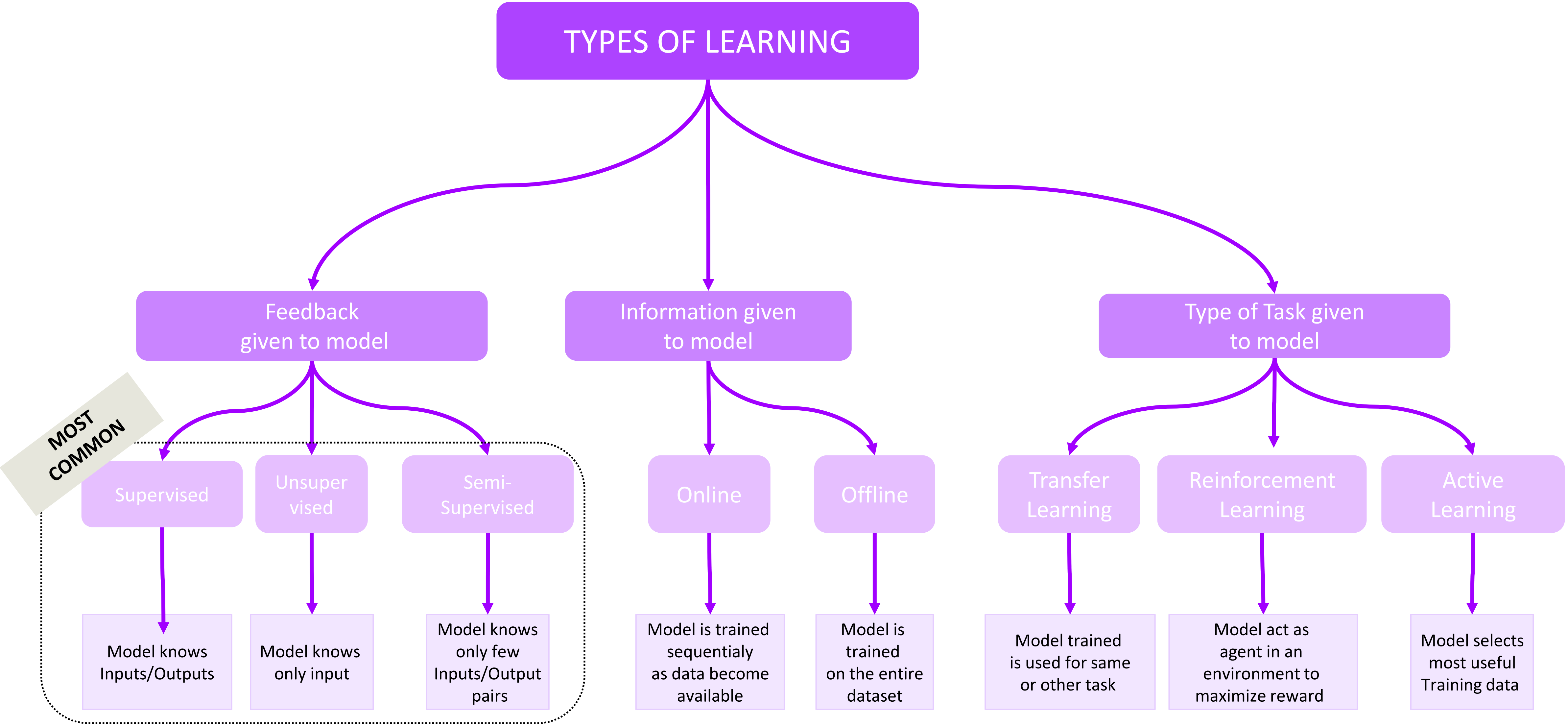
Artificial Intelligence (AI) is a collective term for various technologies that enables machines to imitate human behavior, e.g. by rules, decision trees or machine learning algorithms.

Machine Learning (ML) is a part of AI and includes a variety of statistical methods. Historical data are used to train machines to deal with input values.



Deep Learning (DL) is assigned to machine learning and uses, among other things, multilayer neural networks to solve complex problems with the help of large amounts of data.

MAIN TYPES OF LEARNING



Machine Learning

Supervised and unsupervised Learning

In **Supervised learning**, an AI system is presented with **data** which is **labeled**, which means that each **data tagged** with the **correct label**.

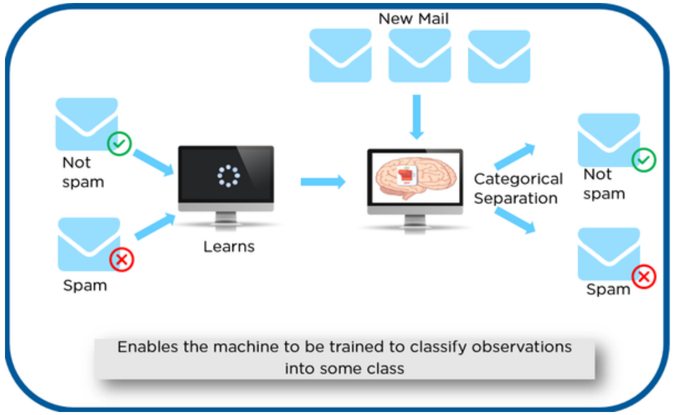
The goal is to **approximate** the **mapping function** so well that when you have **new input data** (x) that you can **predict the output variables** (Y) for that data.

CLASSIFICATION

A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.

REGRESSION

A regression problem is when the output variable is a real value, such as “dollars” or “weight”.



In **unsupervised learning**, an AI system is presented with **unlabeled**, **uncategorized data** and the system’s algorithms act on the data without prior training.

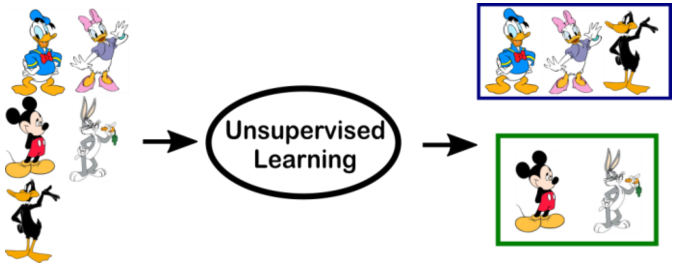
The **output** is dependent upon the **coded algorithms**. Subjecting a system to unsupervised learning is one way of testing AI.

CLUSTERING

A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

ASSOCIATION

An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.



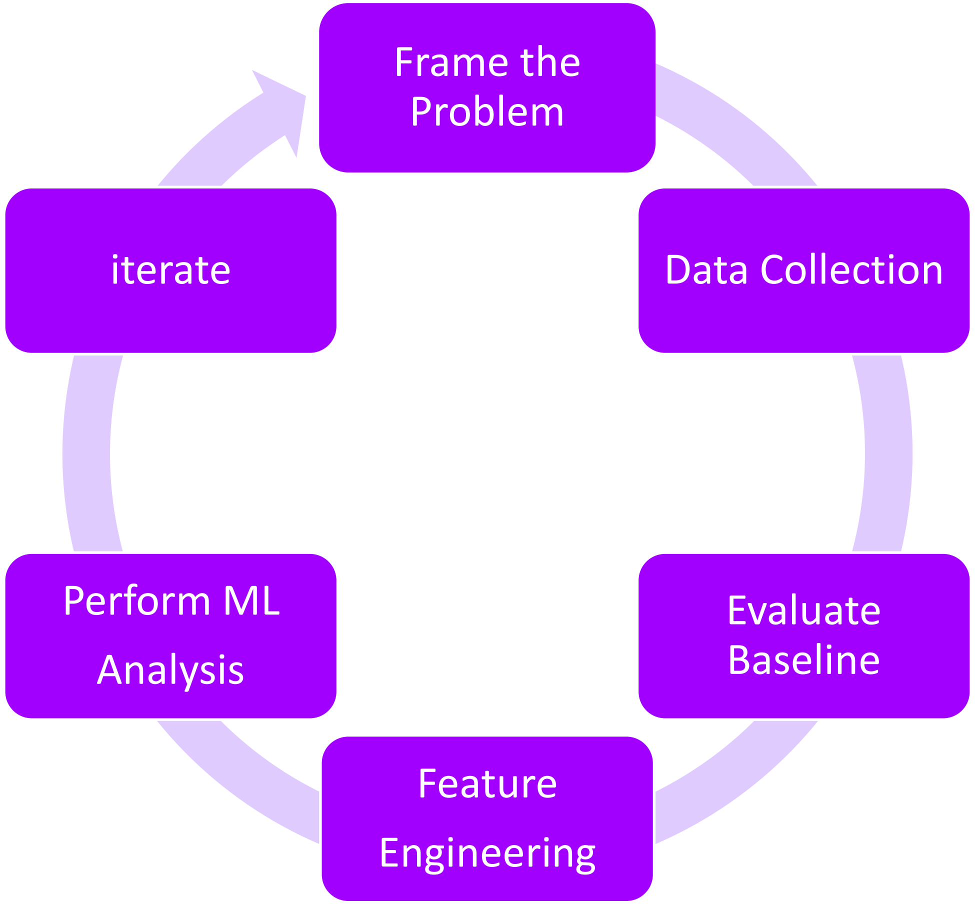
Machine Learning
Most common (supervised) algorithms

Model Type	Name	Description	Pros	Cons
Linear	Linear Regression	Find the "best fit" through all the data points. The forecast is composed of continuous numbers	Easy to understand: it is easy to see clearly which are the **main drivers** of the model	Model too simple to capture **complex relationships** between variables
Tree-based	Random Forest	Composed of a set of rules based on the characteristics of the data, it forms a tree that tries to combine all possible results to the prediction.	Provides a high-quality result. The model is fast to train but may not converge	The Model can be very large and difficult to interpret
	Gradient Boosting	Unlike Random Forest form set of rules in the form of a set of weak predictive models (error based)	Very high performance in computational terms	Difficult to interpet
Neural Networks	Neural Networks (Deep Learning)	"Interconnected neurons" that pass messages to each other at different levels of depth	Can handle extremely complex tasks - e.g.,image recognition	Very slow to train, because they often have a complex architecture. Model not easy to interpret

2

Machine Learning Design Life Cycle

AI DEVELOPMENT APPROACH
ML Development



If you goal is to use ML you need to define:

1. Happens during the ML Model
2. What happens before
3. Everything that happened next

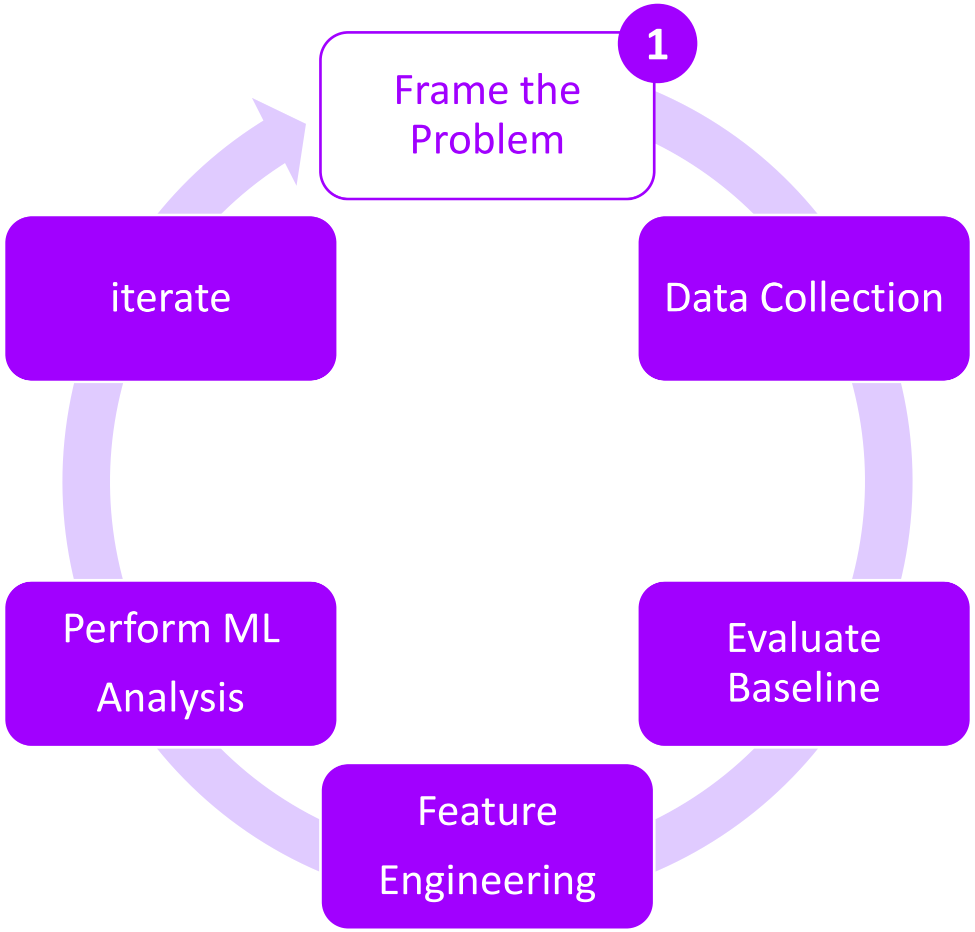
In Accenture we apply this methodology to achieve business and product goals, to help new Data Scientist to apply to real world problem.

Our team breaks down this into **6 steps**:

- Frame the problem
- Data Collection
- Evalute Baseline
- Features
- Perform ML Analysis
- Iterate

AI DEVELOPMENT APPROACH

Problem Definition – What problem are we trying solve?

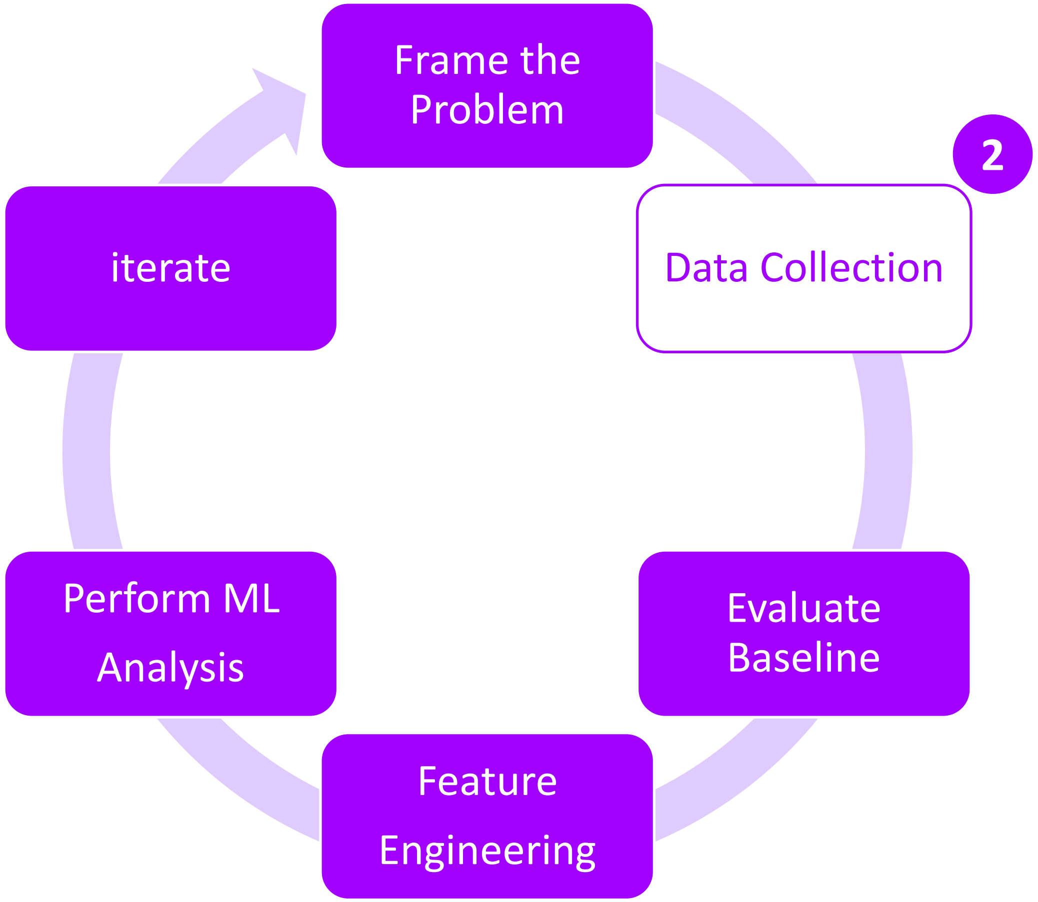


Business Problem

- Which **Data?** — Determine label and training
- Which **Features?** — Simple is better than complicated
- Which **Model?** — Determine the right task for your project (regression, classification, clustering, anomaly detection)
- What **Predictions** — How use and interpret model predictions
- **Improve business** — How this will impact business

AI DEVELOPMENT APPROACH

Data - What data do you have ?



1. Turn a person's requirements into data requirements

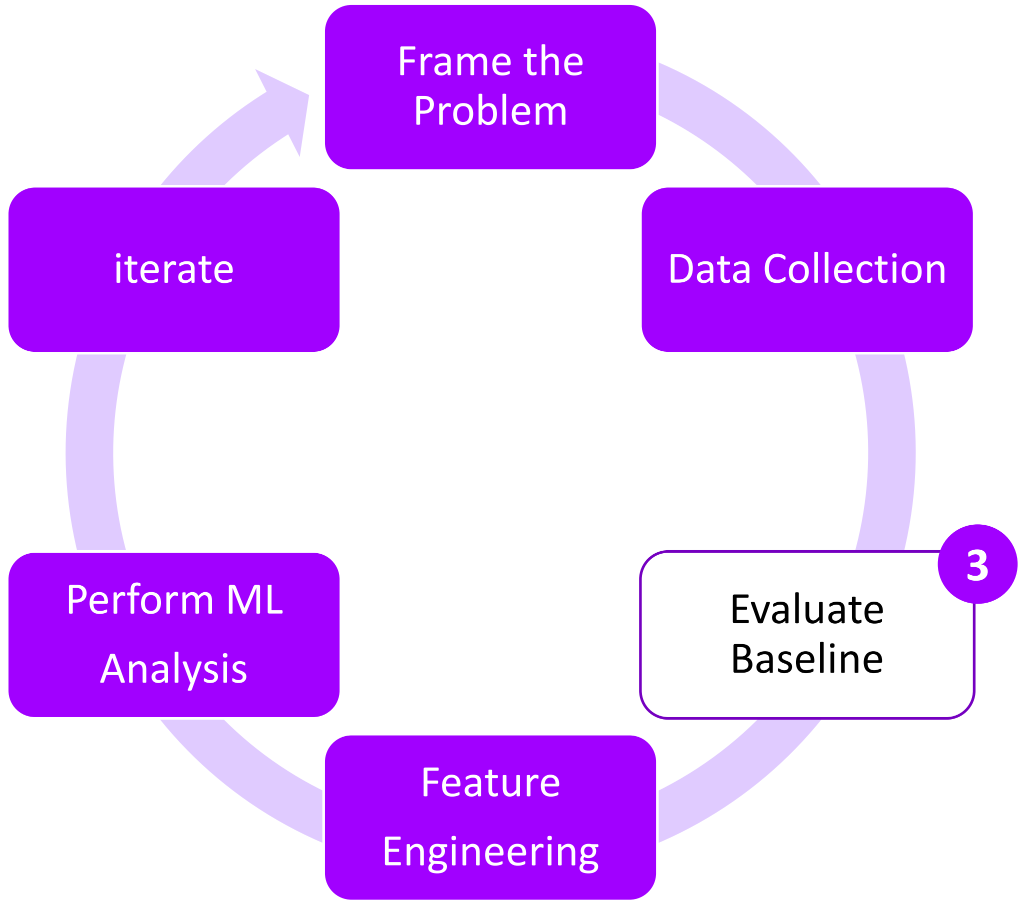
- Determine the type of data needed to train your model.
- You'll need to consider, predictive power, relevance, fairness, privacy, and security.

2. Identify all available datasets

- Evaluate your data and their collection method to ensure they're appropriate for your project.
- External (open data) or Internal data that determine that formulate you problem

AI DEVELOPMENT APPROACH

Evaluate – How could work?



Evaluate Baseline:
Build a first possible model to act as a baseline for feature model and future work (e.g. use simple regression model to predict the future values, or energy consumption)

Dataset Partition (Train/test/validation)

The golden standard is to split the data into 3 sets:

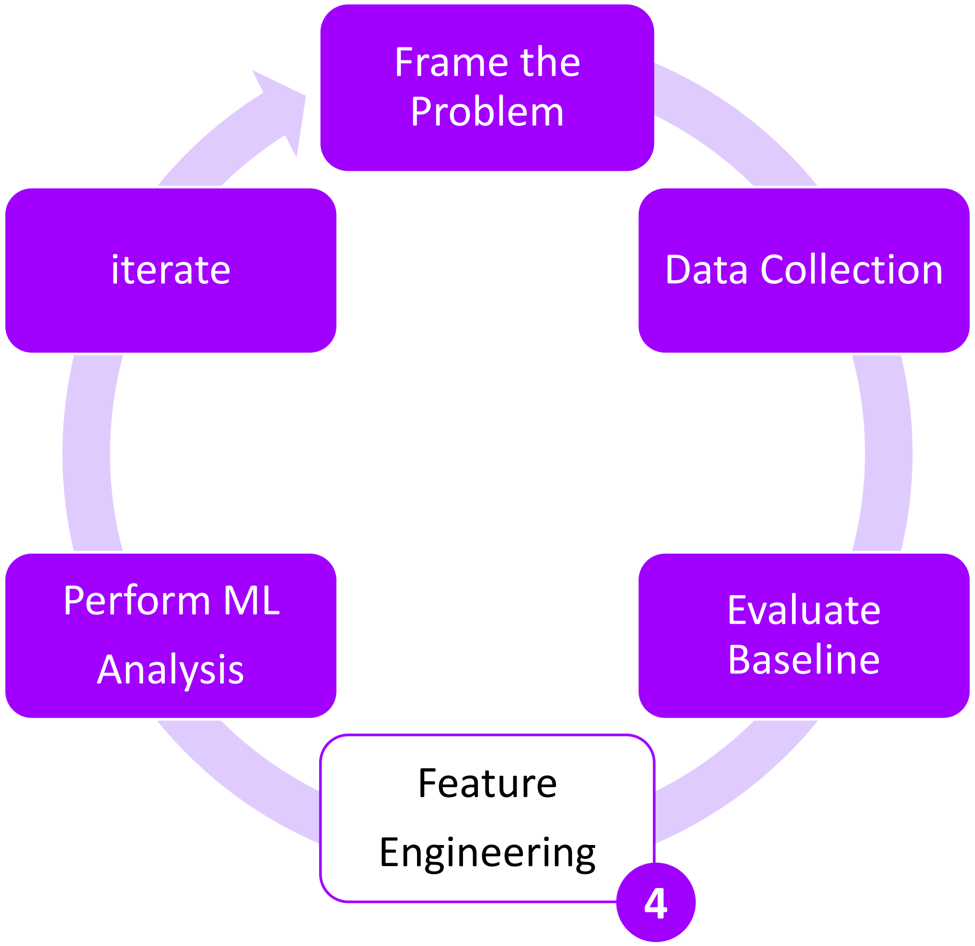
a. Training (model building)
b. Evaluation (hyper parameter of the model)
c. Test set (not used for training or tuning)

Data Partition Type

a. Random
b. Sequential (used for time series)
c. Stratified

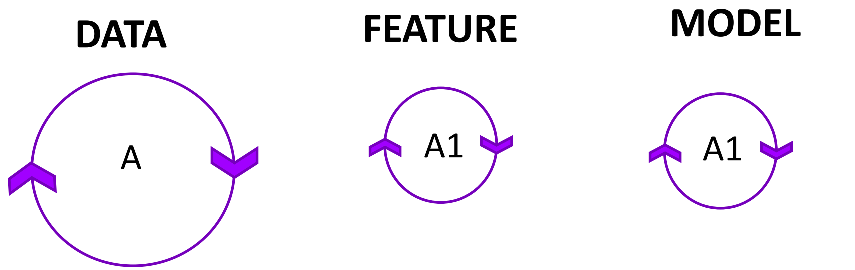
AI DEVELOPMENT APPROACH

Features – What features of data best align with model metrics?



Feature engineering is how we take domain specific knowledge, and encode it into a format that our model can leverage effectively. Considered as one most **important phase** of AI model development

LIFE CYCLE ITERATION



Feature and **Model** has small iteration Life cycle vs **Data**

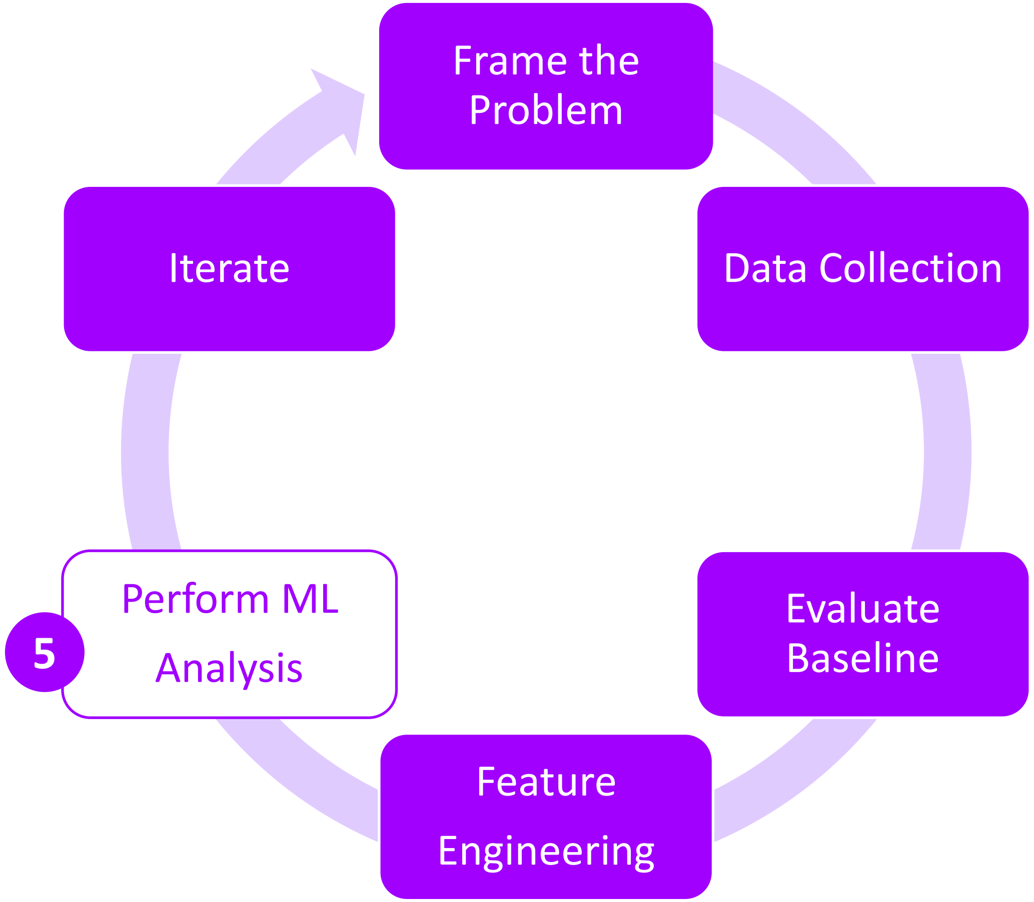
FEATURE ENGINEERING

- Transform features (e.g. one-hot encoding)
- Scale features (e.g. min-max scaling, Z-score)
- Build new (may impact semantic meaning)

AI DEVELOPMENT APPROACH

Model - what model best suits the problem and data?

5

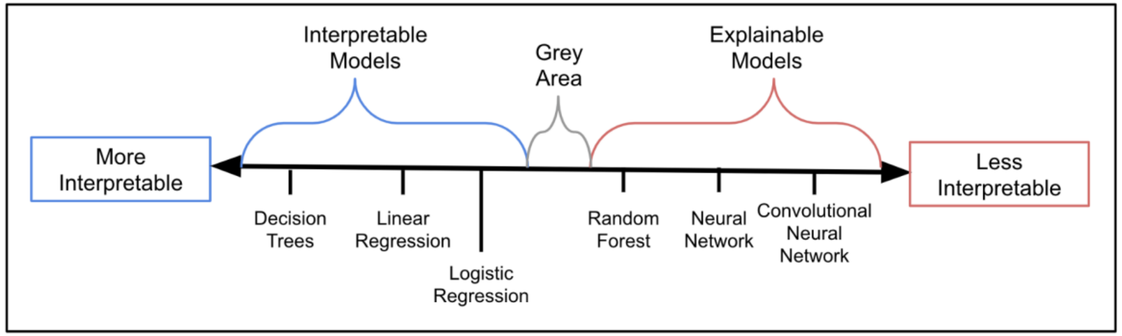


Create a Predictive Model

- Use feature from previous step

Factors that govern choice of model:

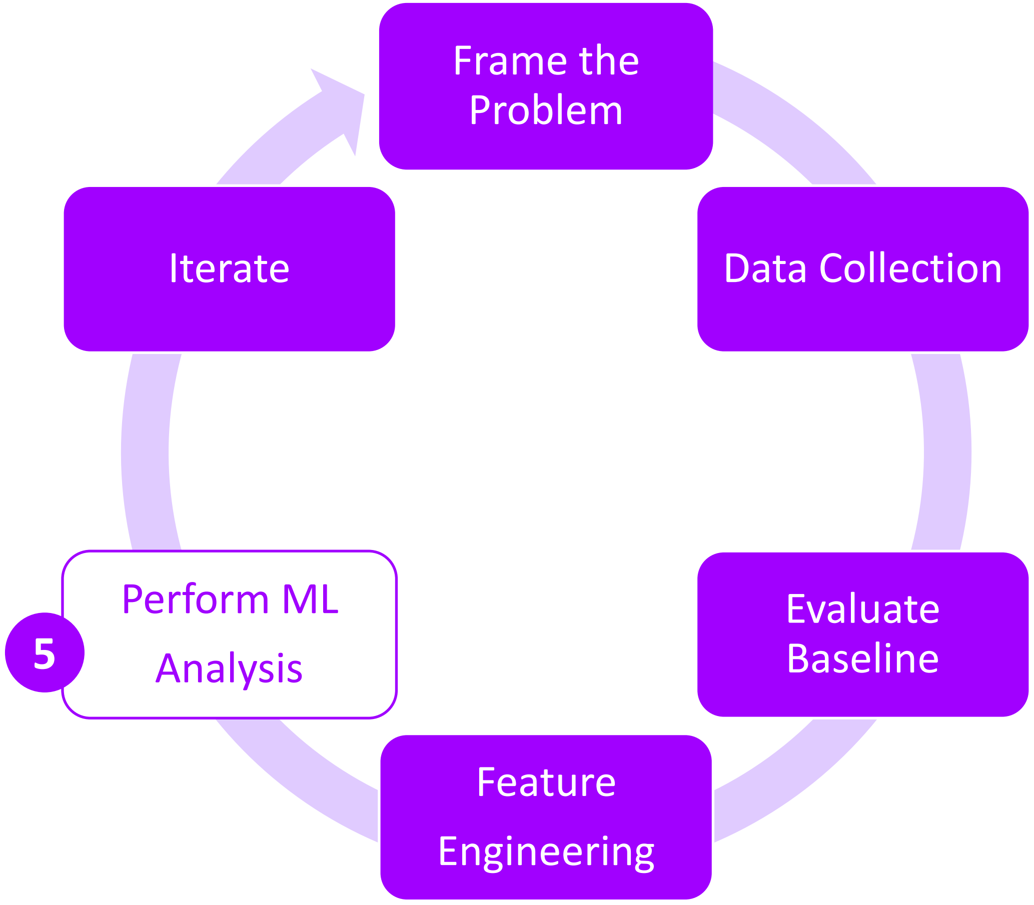
- Features
- Data Volume
- Interpretability



AI DEVELOPMENT APPROACH

5

Model - what model best suits the problem and data?



Create a Predictive Model
- Use feature from previous step

Factors that govern choice of model:
- Features
- Data Volume
- Interpretability

Model Tuning Types

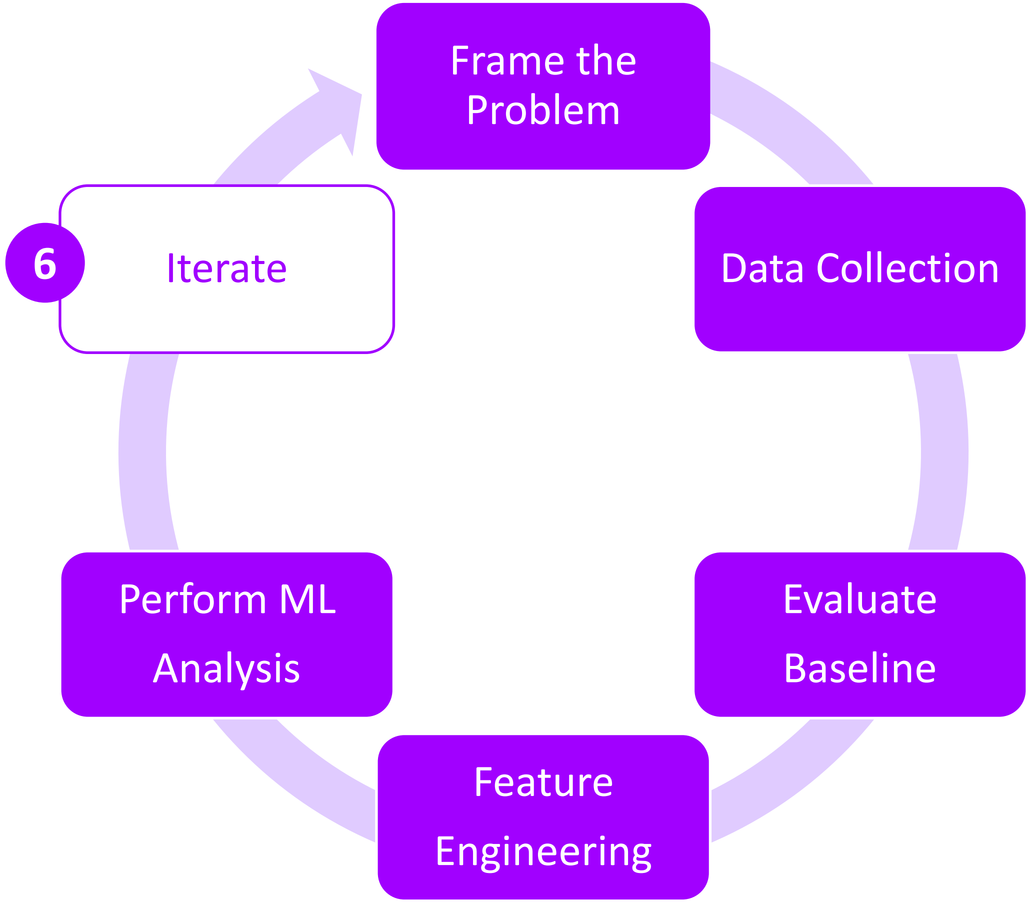
A- Hyperparameter tuning like learning rate and regularization parameter

B- Model architecture settings
- Feature interactions for linear model
- Number of leave/trees for tree based
- Number, type and width of layers for neural network

AI DEVELOPMENT APPROACH

6

Iterate- how can you iterate and improve upon the previous steps



- This step involves all the previous steps.
- Your **goal** should be minimising the time between offline experiments (experiment phase) and online experiments (production)

Poor performance on training data means:
-Model hasn't **learned properly**. Try a different model, improve the existing one, collect more data, collect better data.

-Model doesn't **generalise well**. Your model may be overfitting the training data. Use a simpler model or collect more data.

3

Prepare a Machine Learning Model with Python

Demo on Energy Efficiency on buildings

Demo hands-on
Energy efficiency on buildings

CONTEXT

Energy efficiency on buildings is a trend topic that people pay more attention in the last year, due to energy crisis.

The possibility to **predict** the **heating** and **cooling load** by the usage of **building information** could open many **scenarios of application**: for instance, cost-optimization problems of building renovation could consider which buildings get more advantages in term of energy saving.

DATASET: Energy efficiency

- Original Dataset: https://archive.ics.uci.edu/ml/datasets/Energy+efficiency
- Kaggle Dataset: https://www.kaggle.com/datasets/winternguyen/energy-efficiency-on-buildings

TOOLS

- Google Colab (suggested): it does not require any additional installation (https://colab.research.google.com)
- Jupyter Lab: it requires installation

COMPOSITION

Input Variables

- Relative Compactness
- Surface Area - m^2
- Wall Area - m^2
- Roof Area - m^2
- Overall Height – m
- Orientation - 2:North, 3:East, 4:South, 5:West
- Glazing Area - 0%, 10%, 25%, 40% (of floor area)
- Glazing Area Distribution (Variance) - 1:Uniform, 2:North, 3:East, 4:South, 5:West

Target Variables

- Heating Load - kWh/m^2
- Cooling Load - kWh/m^2

TASK

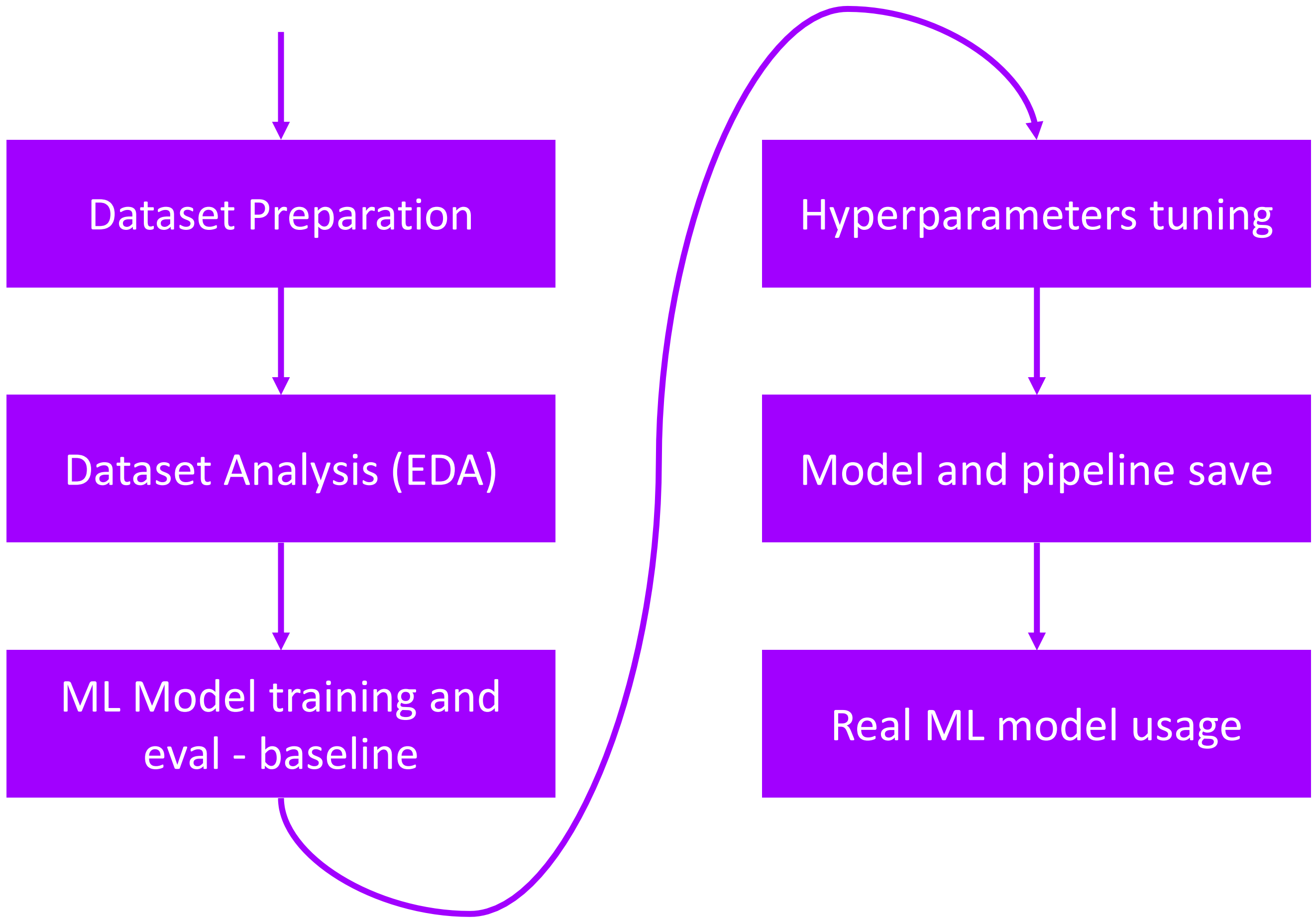
Regression: predict the heating and cooling load with a ML model, giving the input and target variables.

Demo hands-on
Steps and objective

OBJECTIVE

During the lab session, we will focus on the classical **steps** to **build** and **train** ML model, with a focus on **EDA** (**Exploratory Data Analysis**). Then, we demonstrate how to **use** the **model** in **real** context.

We also include and discuss about some useful **libraries** that can help and accelerate your development work.

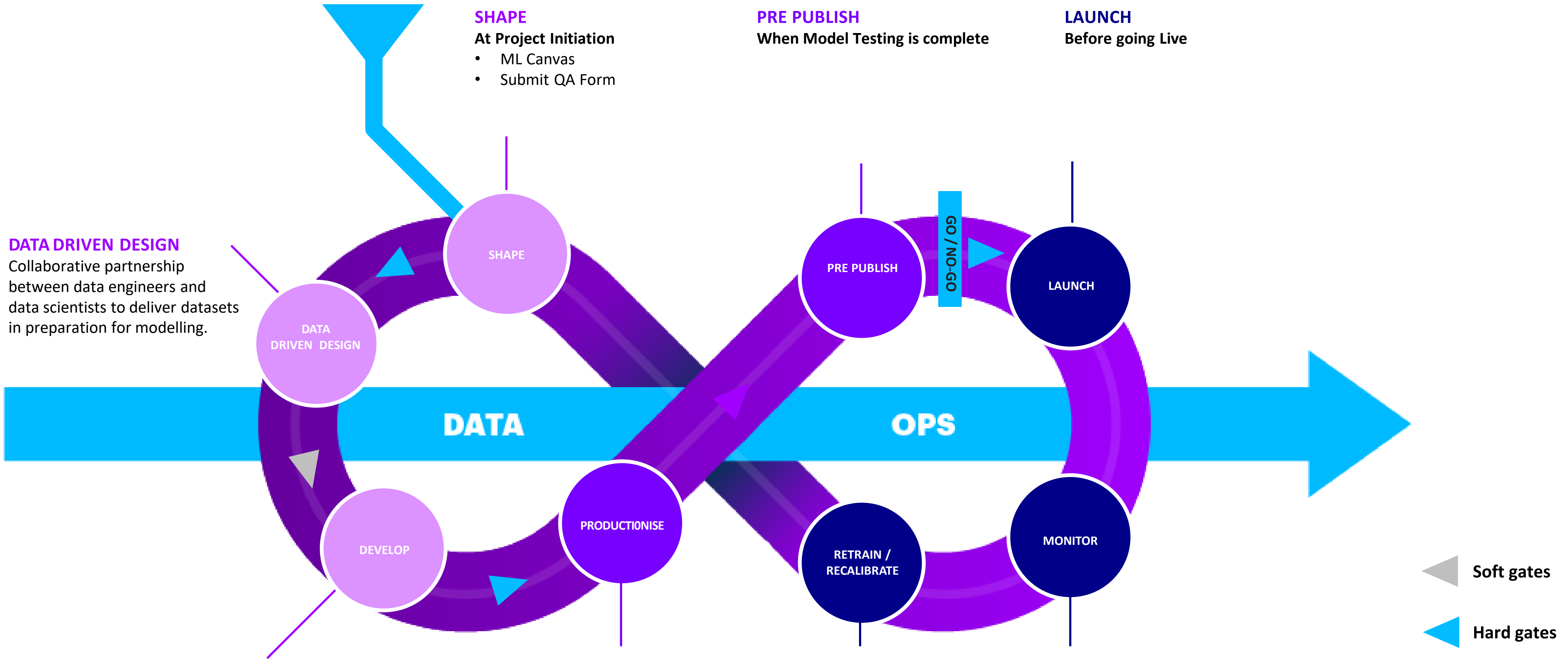


4

Enterprise tips for Python solutions

From dev to production

MACHINE LEARNING
LIFECYCLE



SHAPE
At Project Initiation
- ML Canvas
- Submit QA Form

PRE PUBLISH
When Model Testing is complete

LAUNCH
Before going Live

DATA DRIVEN DESIGN
Collaborative partnership between data engineers and data scientists to deliver datasets in preparation for modelling.

DEVELOP
Iterative model and feature development in agile sprints, cycling between model training, scoring, testing, and re-training.

PRODUCTIONISE
Hardening and testing of data pipelines and models for deployment into live environments.

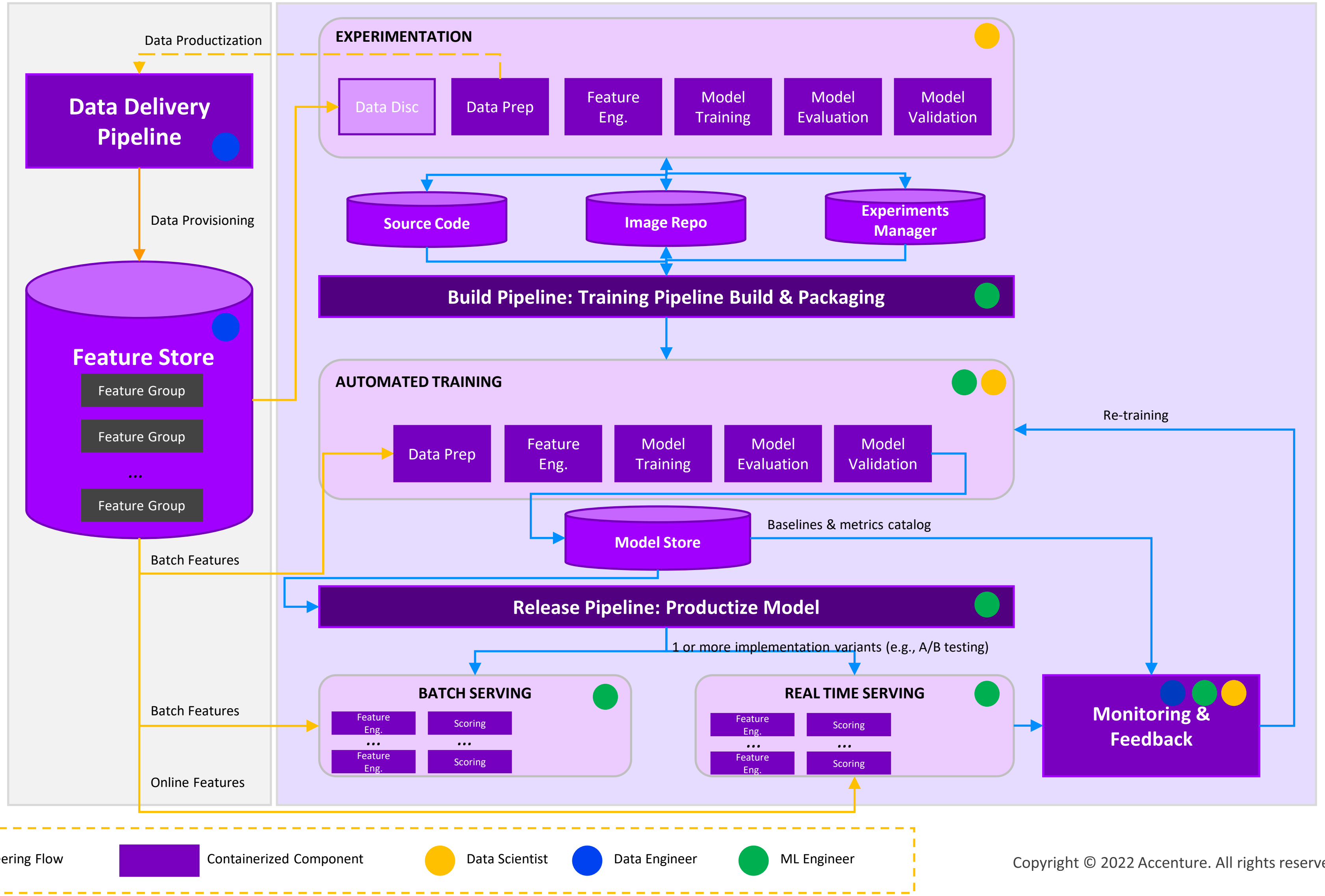
MONITOR
Monitor and manage the performance of production models over time, identifying any that need to be re-trained or re-calibrated.

MACHINE LEARNING
CONCEPTUAL ML ARCHITECTURE

A data scientist's model without any integration in **production context** is not useful.

A **ML model** should be considered as a **software** and all steps of a **software engineering development** should be applied: therefore, a **ML application** is based on the conjunction of **code**, **data** and **models**.

The operations behind ML applications are defined as **MLOps**.



ENERGY GeO Hackathon
Society of Petroleum Engineers

SPE International Italian Section
SPE International London Section
SPE International Netherlands Section
SPE International Romanian Section
SPE International Copenhagen Section
SPE International Geothermal Technical Section

Q&A

Geological Survey of Denmark and Greenland GEUS
accenture
CGG
MathWorks
Microsoft
NVIDIA
Oilfield International
SAMSUNG
TNO innovation for life
TRE ALTAMIRA

ENERGY GeO Hackathon

Society of Petroleum Engineers

www.spehackathon-eu.com

#DatafyingEnergy