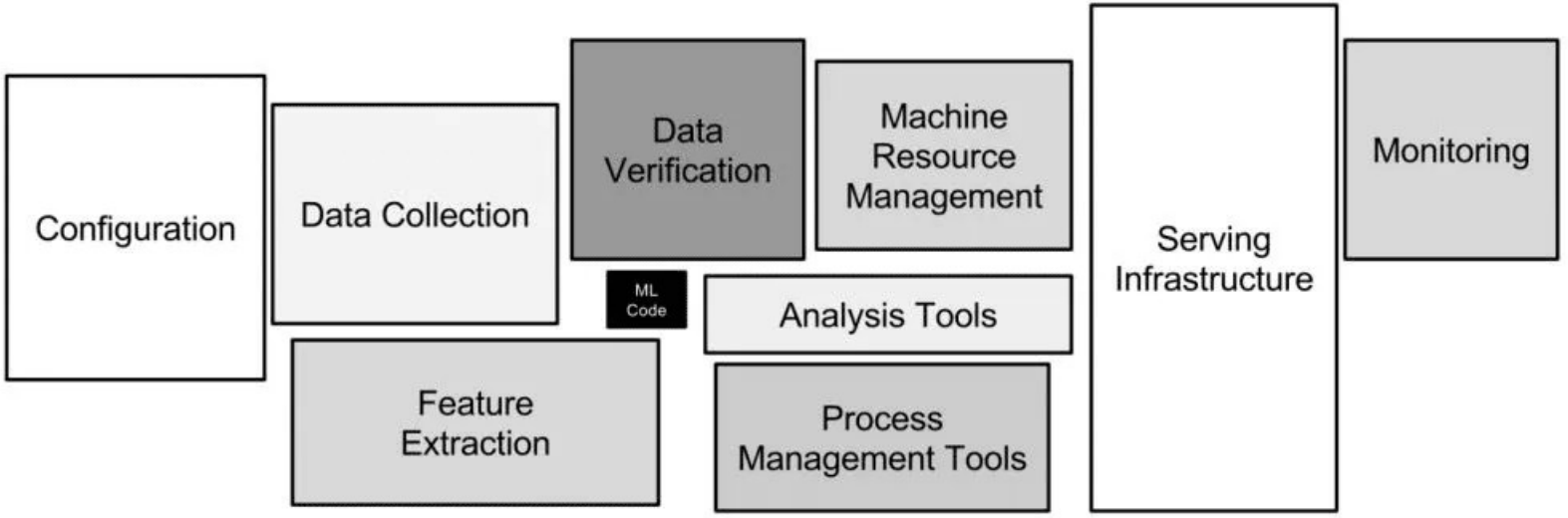
Intro to programming AI/ML Systems

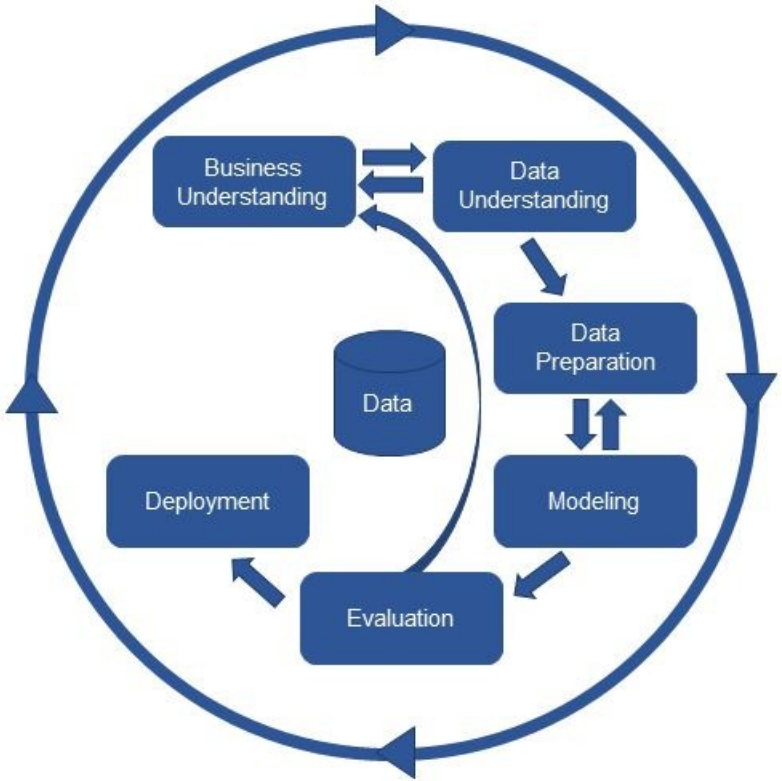
Typical AI ML Infrastructure

- The focus of this course is represented by the degree of “blackness” of each component

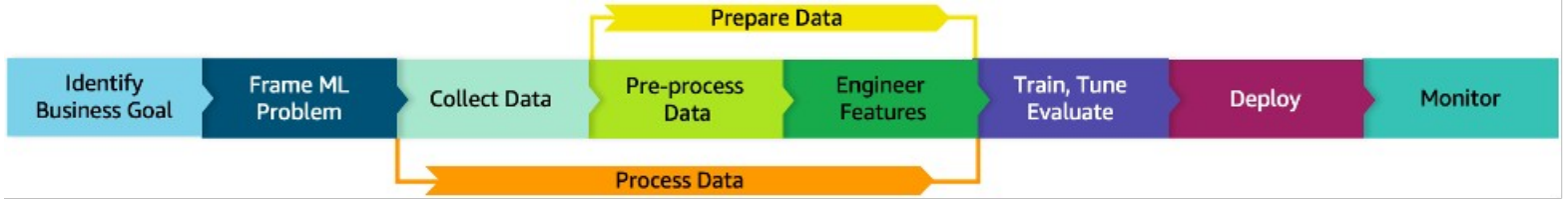


CRISP DM

- CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely adopted framework that provides a structured approach to data mining projects



ML life-cycle architecture diagram: AWS approach



- **Business goal -** An organization considering ML should have a clear idea of the problem, and the business value to be gained by solving that problem. You must be able to measure business value against specific business objectives and success criteria.
- **ML problem framing -** What is observed and what should be predicted (known as a label or target variable). Determining what to predict and how performance and error metrics must be optimized is a key step in this phase.
- **Data processing -** Training an accurate ML model requires data processing to convert data into a usable format. Data processing steps include collecting data, preparing data, and feature engineering that is the process of creating, transforming, extracting, and selecting variables from data.
- **Model development -** Model development consists of model building, training, tuning, and evaluation. Model building includes creating a CI/CD pipeline that automates the build, train and release to staging and production environments.
- **Deployment -** After a model is trained, tuned, evaluated and validated, you can deploy the model into production. You can then make predictions and inferences against the model.
- **Monitoring -** Model monitoring system ensures your model is maintaining a desired level of performance through early detection and mitigation

Business goal

Types of business goals

Financial
Specifies a financial target.

Growth
Aims to expand part of the business.

Employee
Helps with career advancement, life-work balance.

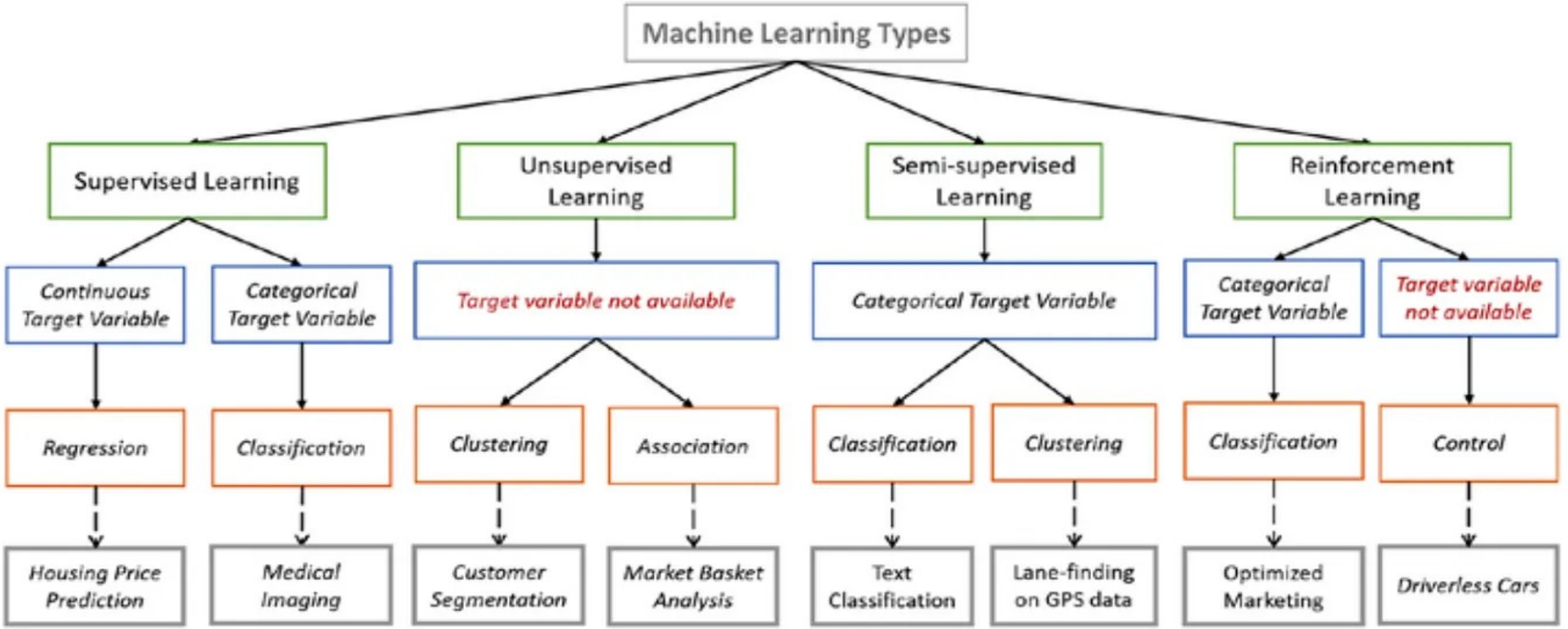
Process
Improves procedures that optimize the business.

Social
Promotes diversity or sets sustainability goals that improve the environment.

Time-based
Applies a timeline of days/weeks/months/years to accomplish the task.

SOURCE: SEAN MICHAEL KERNER; ICONS: PRESSUREUA/GETTY IMAGES

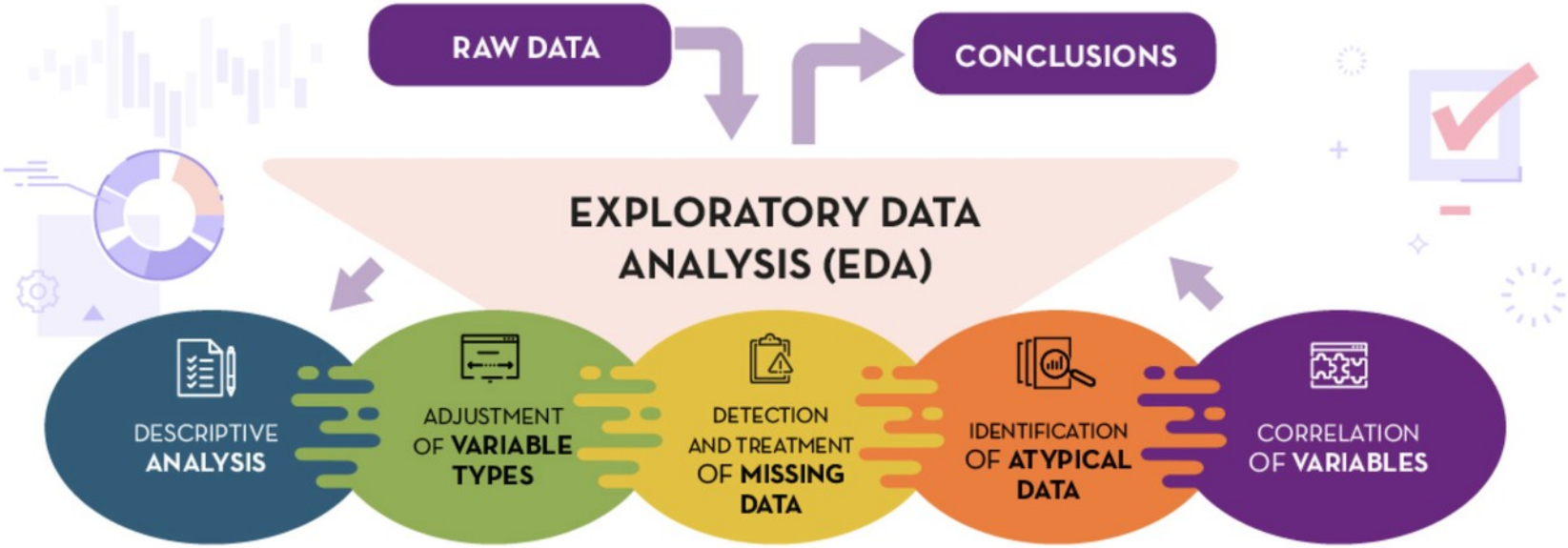
ML problem framing



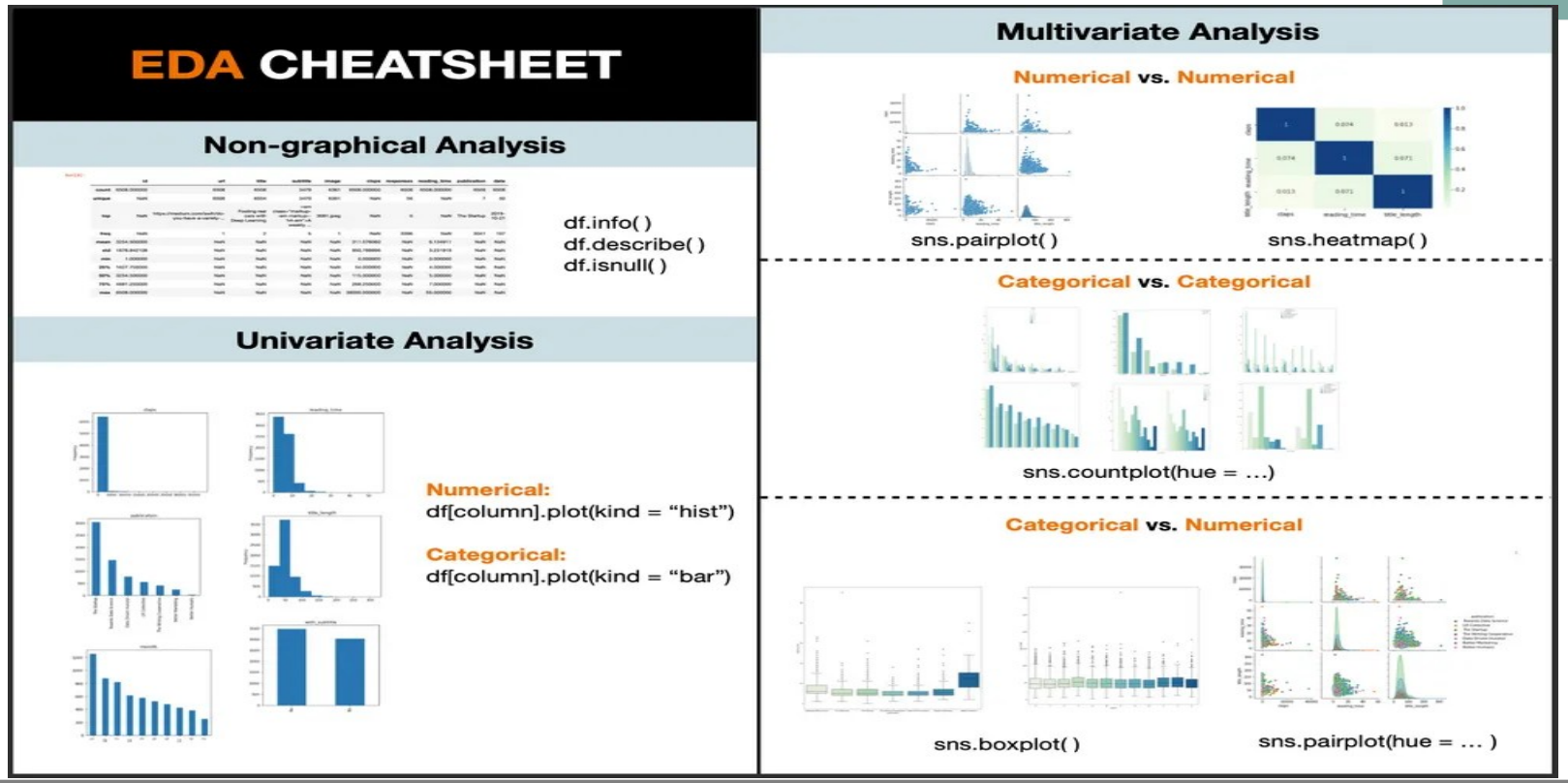
Typical Business Problem to ML Tasks framing

	Business Problem	ML Tasks		Business Problem	ML Tasks
1	Customer Churn Prediction	Classification: Predict churn vs. retention	9	Predictive Maintenance	Predictive Modeling: Forecast equipment failure
2	Fraud Detection	Anomaly Detection: Identify fraudulent transactions	10	Object Detection	Computer Vision: Detect objects in images or videos
3	Sentiment Analysis	Natural Language Processing: Analyze sentiment in text data	11	Personalized Medicine	Genetic Algorithms: Tailor treatments based on genetic data
4	Demand Forecasting	Time Series Forecasting: Predict future demand	12	Natural Language Translation	Sequence-to-Sequence Models: Translate text between languages
5	Credit Risk Assessment	Regression/Classification: Predict creditworthiness	13	Network Intrusion Detection	Classification: Detect and classify network intrusions
6	Image Classification	Computer Vision:Classify images	14	Time Series Anomaly Detection	Anomaly Detection: Identify abnormal patterns in time series
7	Recommendation Systems	Collaborative Filtering: Recommend items based on user behavior	15	Customer Segmentation	Clustering: Group customers based on similar characteristics
8	Speech Recognition	Deep Learning: Convert speech to text	16	Price Optimization	Reinforcement Learning: Optimize pricing strategies

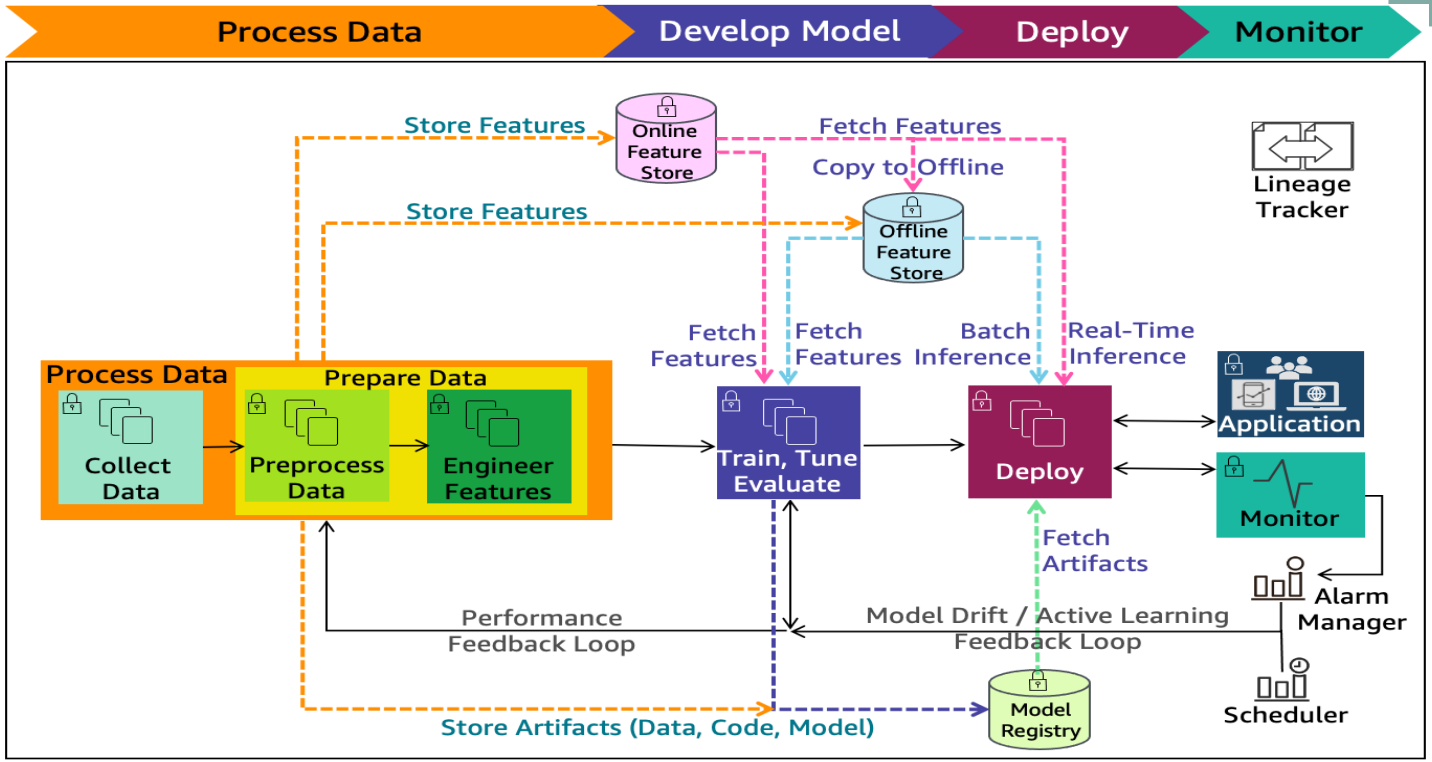
Data understanding: EDA tasks



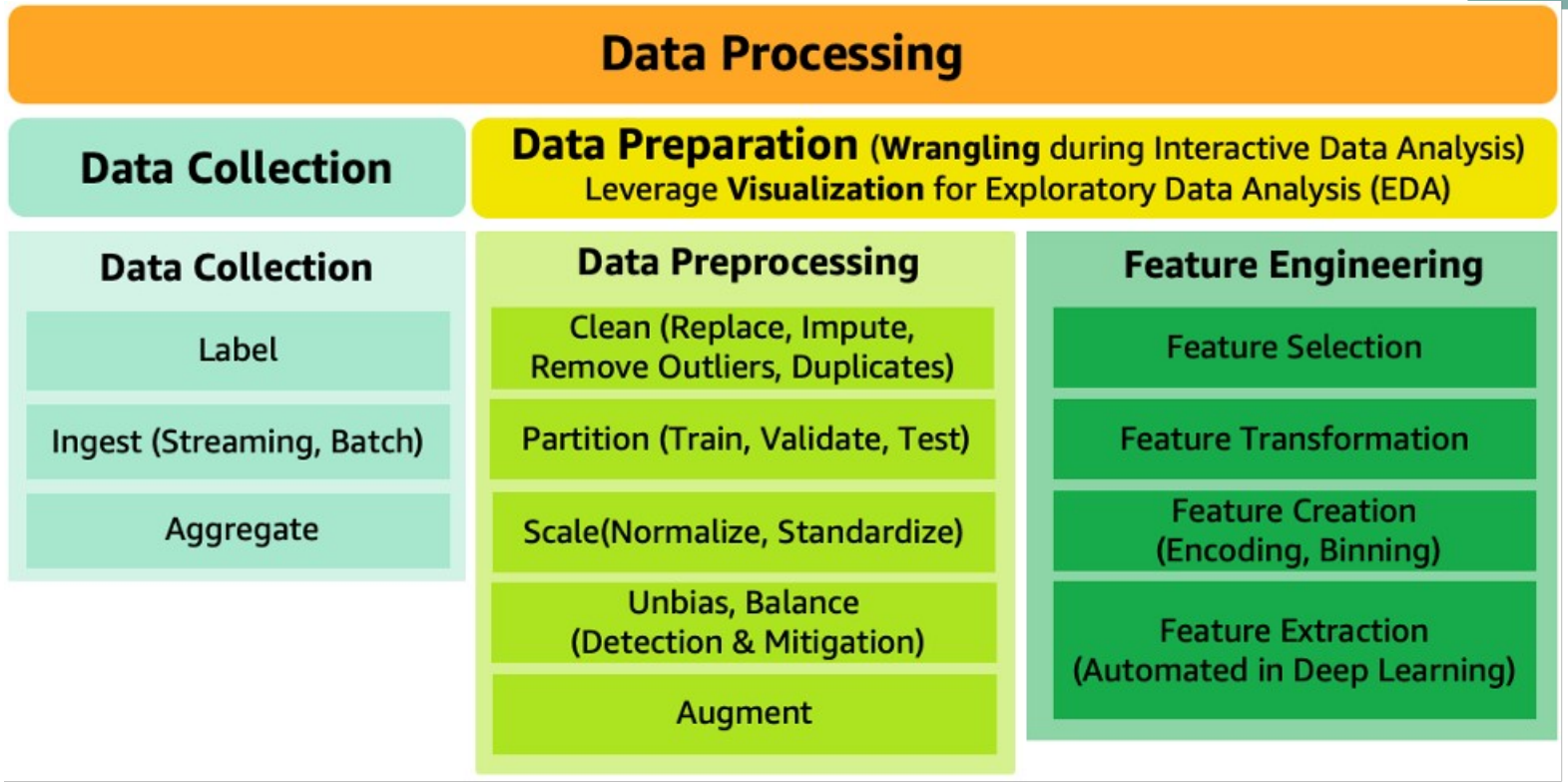
Data understanding: EDA tools



ML lifecycle phases



Data processing

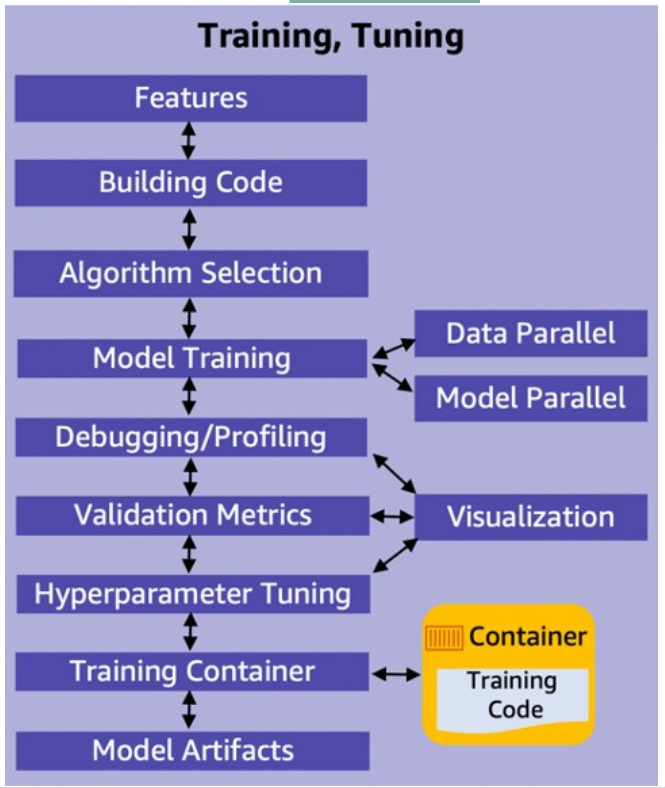


Model training and tuning

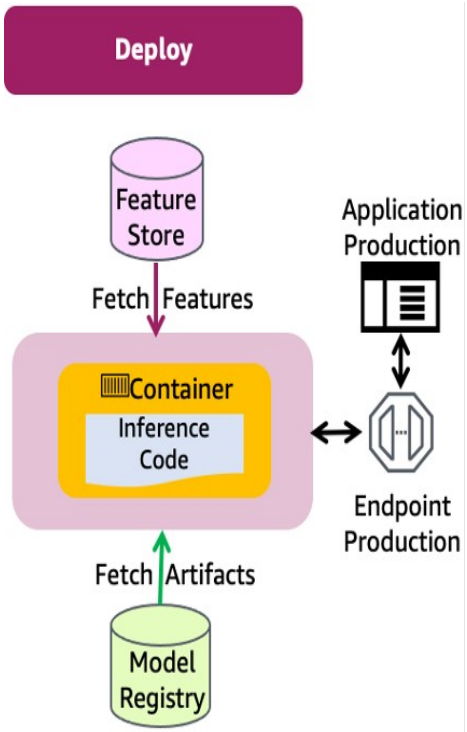
Features – Features are selected as part of the data processing after data quality is assured

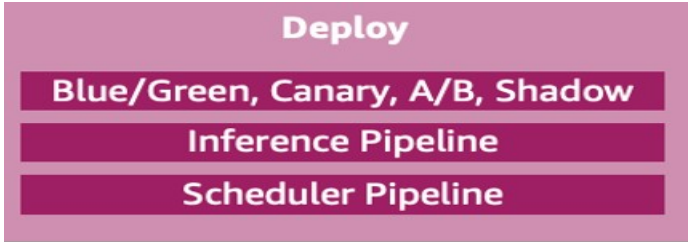
Building code – Model development includes building the algorithm and its supporting code.

- **Algorithm selection** – Run many experiments with parameter tuning across available options.
- **Model training** – The process of learning model parameters from training data.
- **Debugging/profiling** – A machine learning training job can have problems including: system bottlenecks, overfitting, saturated activation functions, and vanishing gradients. A debugger provides visibility into the ML training process through monitoring, recording, and analyzing data. It captures the state of a training job at periodic intervals.
- **Validation metrics –** Typically, a training algorithm computes several metrics such as loss and prediction accuracy. These metrics determine if the model is learning and generalizing well for making predictions on unseen data.
- **Hyperparameter tuning** – Settings that can be tuned to control the behavior of the ML algorithm are referred to as hyperparameters. The number and type of hyperparameters in ML algorithms are specific to each model.
- **Training code container** – Create container images with your training code and its entire dependency stack.
- **Model artifacts** – Model artifacts are the outputs that results from training a model. They typically consist of trained parameters, a model definition that describes how to compute inferences, and other metadata.



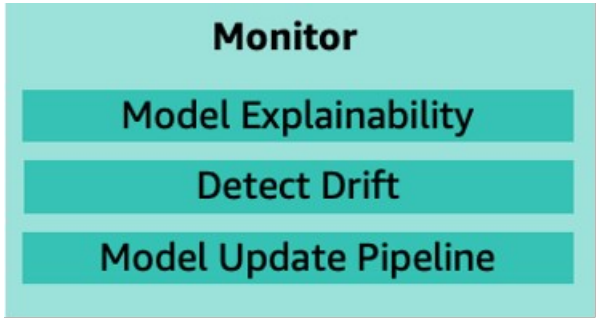
Model Deployment





- **Blue/green deployment -** provides two identical production environments (blue is the existing infrastructure and green is an identical infrastructure for testing). Once testing is done on the green environment, live application traffic is directed to it from the blue environment. Then the roles of the blue/green environments are switched.
- **Canary deployment -** a new model is deployed to a small group of users while other users continue to use the previous version. Once you're satisfied with the new results, you can gradually roll it out to all users.
- **A/B testing -** Direct a defined portion of traffic to the new model. Direct the remaining traffic to the old model. A/B testing is similar to canary testing, but has larger user groups and a longer time scale, typically days or even weeks.
- **Shadow deployment strategy -** the new version is available alongside the old version. The input data is run through both versions. The older version is used for servicing the production application and the new one is used for testing and analysis.
- **Inference pipeline** - pipeline that automates capturing of the prepared data, performing predictions and post-processing for real-time or batch inferences.
- **Scheduler pipeline** – periodical re-training minimizes the risk of data and concept drifts. A scheduler initiates a re-training at business defined intervals. Data preparation, feature pipelines will also be active during this process.

Monitoring



- **Model explainability** - evaluate the soundness of the model and if the predictions can be trusted.
- **Detect drift** - data and concept drifts, initiates an alert, and sends it to the alarm manager system. Data drift is significant changes to the data distribution compared to the data used for training. Concept drift is when the properties of the target variables change. Any kind of drift results in model performance degradation.
- **Model update pipeline** - model update pipeline for a re-train. The Data prepare and Feature pipelines will also be active during this process.