ACC-HPC June 2025

REGRESSION & CLASSIFICATION : Day 2

Machine Learning

Dr Kiran Waghmare
CDAC Mumbai

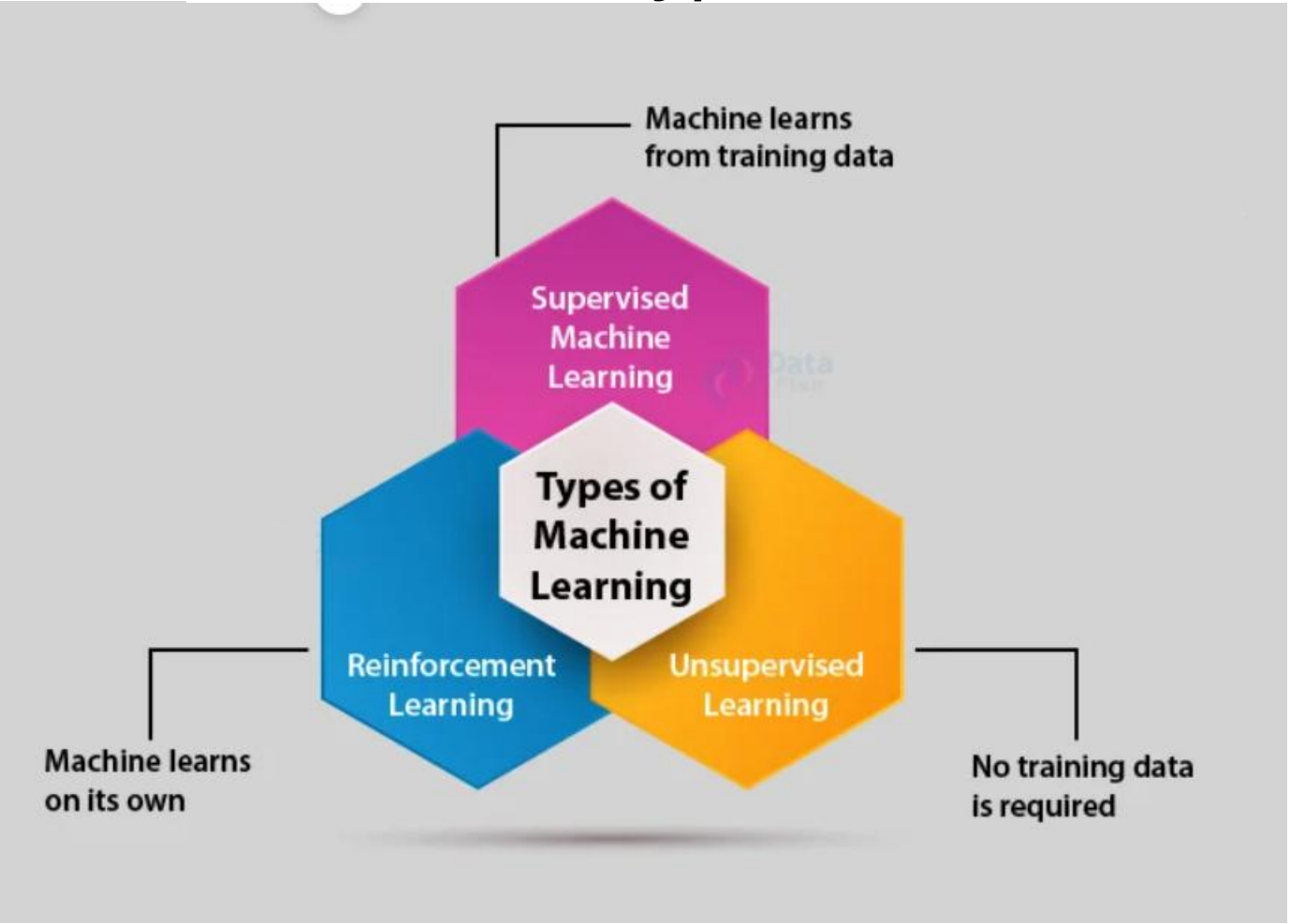
Comparison

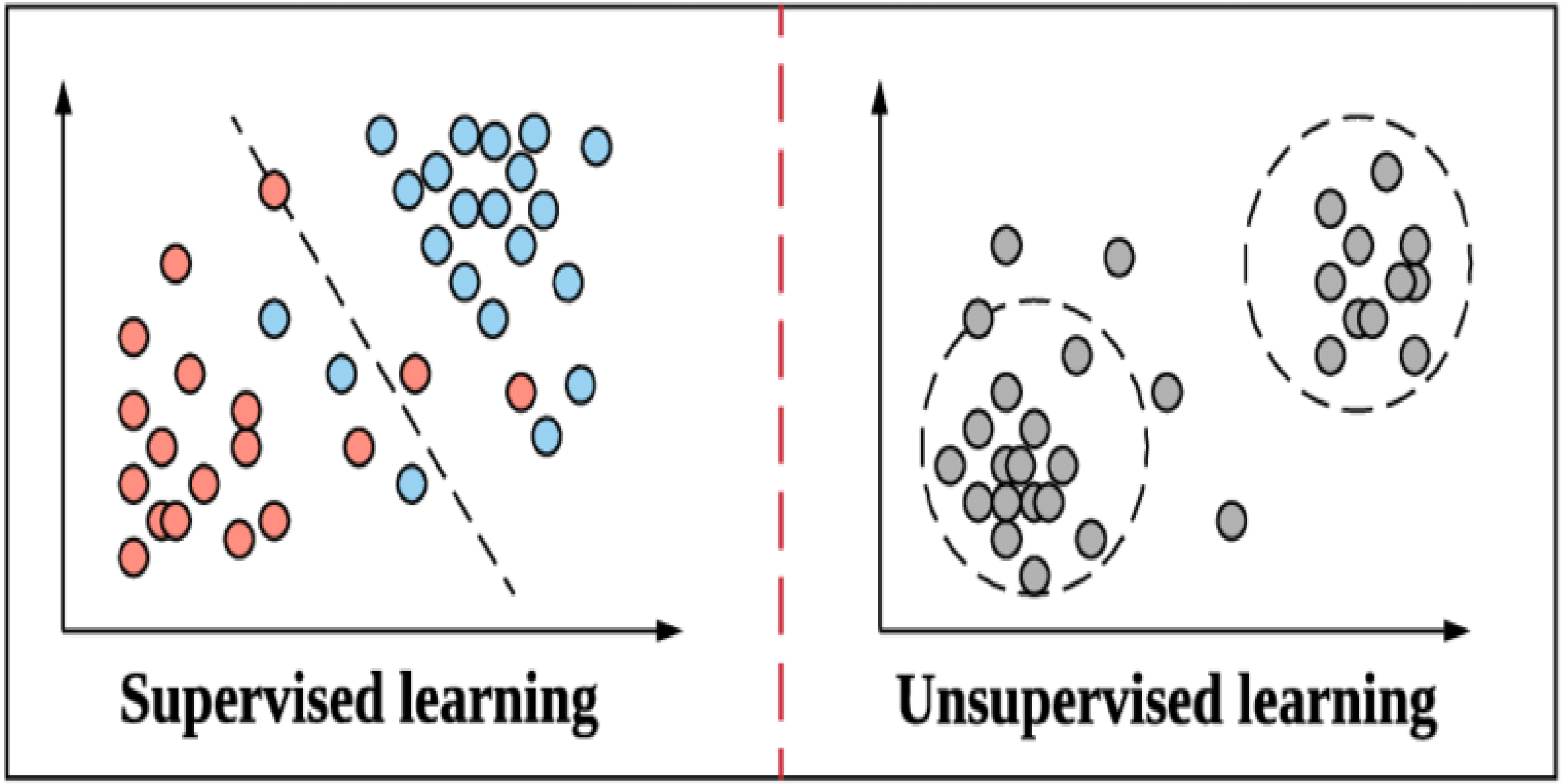
- **Traditional Programming**

- **Machine Learning**

Types of Learning algorithms

- Learning Algorithms can be classified into 3 types as follows –
  - **Supervised Learning**
  - **Unsupervised Learning**
  - **Reinforcement Learning**




Supervised learning
Unsupervised learning

Types of Machine Learning

MATER DYNAMICS

Supervised Learning

Train an algorith on a labeled data set to predict the crrect output value for unseen inputs.

- ✓ Input / Output
- ✓ Labeled data
- ✓ "Replicate the right answers"
- ✓ Classification, prediction

Unsupervised Learning

Train an algorithm to find similari-ties on abnormalities in a data set.

- ✓ Input
- ✓ Unlabeled data
- ✓ "Find patterns in data"
- ✓ Clustering association
- ✓ Anomaly detection, custo-

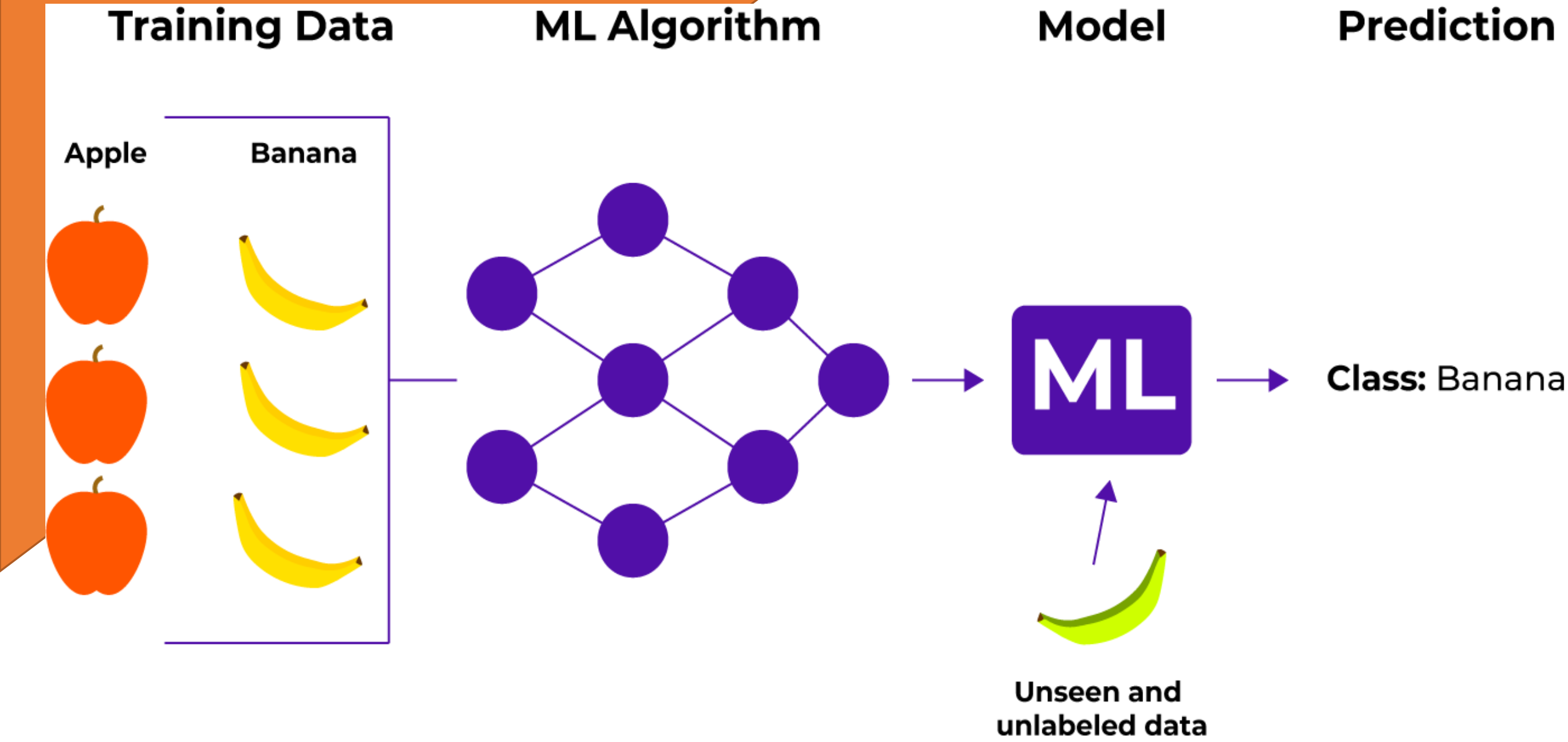
Reinforcement Learning

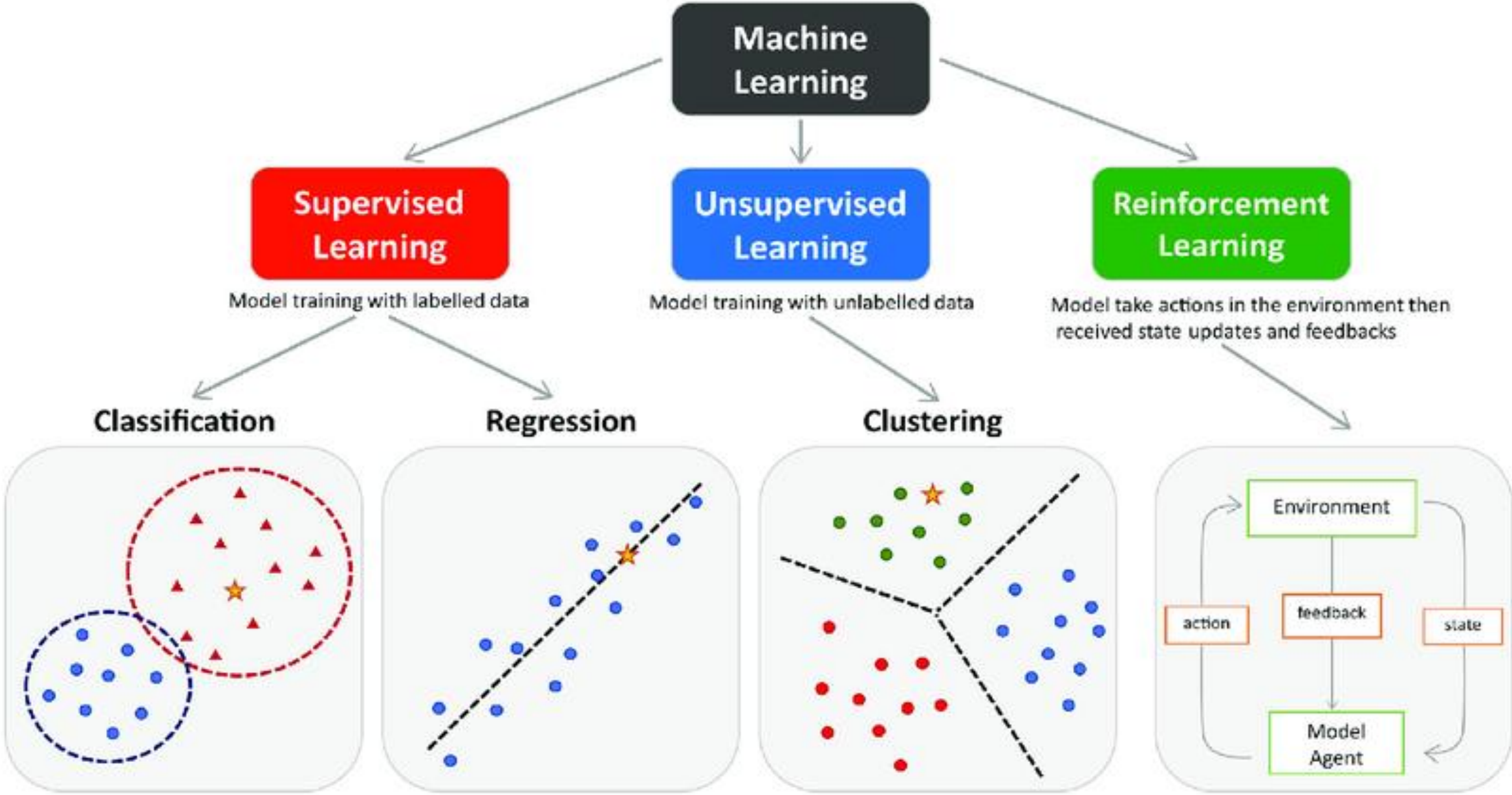
Learn through trial and error from interaction with an environment.

- ✓ States & actions
- ✓ No data set
- ✓ "Find actions that maximize reward"
- ✓ Decision making

Supervised learning

- machine learning task of learning a function that maps an **input** to an **output** supported example input-output pairs.
- In Supervised Learning, the dataset on which we **train our model is labeled**. There is a clear and **distinct mapping** of input and output. Based on the example inputs, the model is able to get **trained** in the **instances**.
- An example of supervised learning is **spam filtering**.
- Based on the **labeled data**, the model is able to determine if the data is **spam** or **ham.** This is an easier form of **training**.
- Spam filtering is an example of this type of **machine learning algorithm**.


Training Data
ML Algorithm
Model
Prediction
Apple
Banana
ML
Class: Banana
Unseen and
unlabeled data


Machine Learning
Supervised Learning
Model training with labelled data
Unsupervised Learning
Model training with unlabelled data
Reinforcement Learning
Model take actions in the environment then received state updates and feedbacks
Classification
Regression
Clustering
Environment
action
feedback
state
Model Agent

Reinforcement Learning

Yes

Reward +ve

(Strong Network)

No

Environment

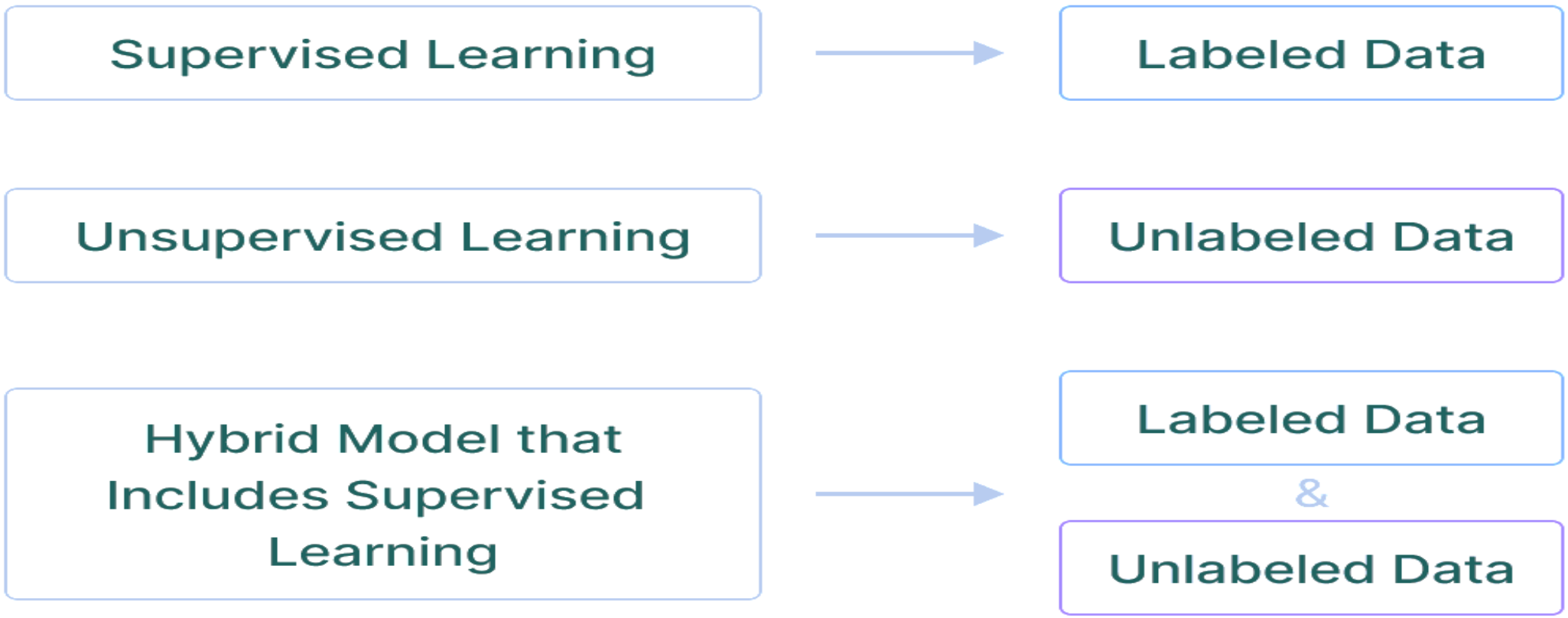
(Network Testing Zone)

Reward -ve

(Poor Network)

Keep searching until you get strong network

Data in Supervised vs. Unsupervised Learning



Machine learning models cheat sheet

Supervised learning

Data scientists provide input, output and feedback to build model (as the definition)

EXAMPLE ALGORITHMS:

Linear regressions
- sales forecasting
- risk assessment

Support vector machines
- image classification
- financial performance comparison

Decision tree
- predictive analytics
- pricing

Unsupervised learning

Use deep learning to arrive at conclusions and patterns through unlabeled training data.

EXAMPLE ALGORITHMS:

Apriori
- sales functions
- word associations
- searcher

K-means clustering
- performance monitoring
- searcher intent

Semi-supervised learning

Builds a model through a mix of labeled and unlabeled data, a set of categories, suggestions and exampled labels.

EXAMPLE ALGORITHMS:

Generative adversarial networks
- audio and video manipulation
- data creation

Self-trained Naïve Bayes classifier
- natural language processing

Reinforcement learning

Self-interpreting but based on a system of rewards and punishments learned through trial and error, seeking maximum reward.

EXAMPLE ALGORITHMS:

Q-learning
- policy creation
- consumption reduction

Model-based value estimation
- linear tasks
- estimating parameters

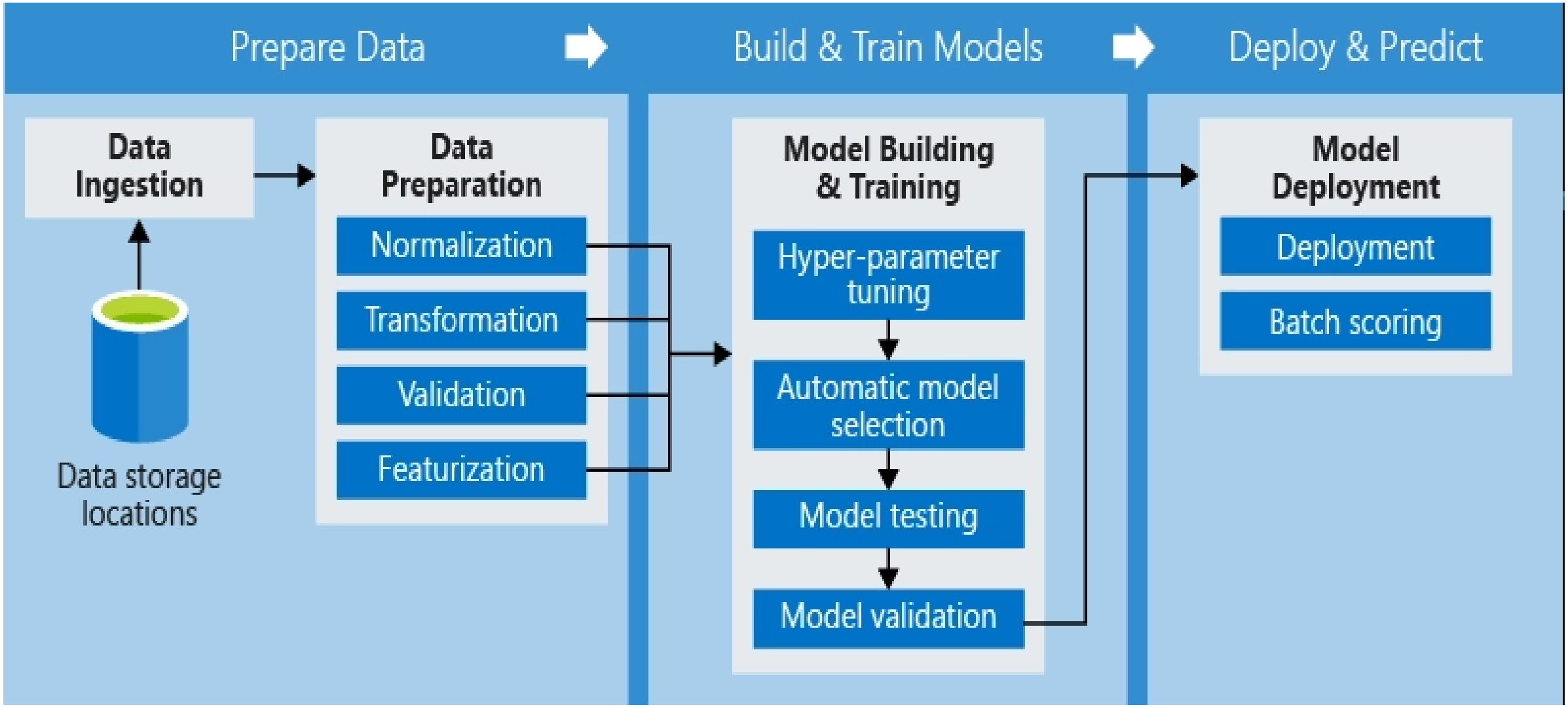
Fundamentals of Statistics

- Data science refers to dealing with data.
- Statistical analysis helps in enhancing predictability, pattern analysis, and concluding and interpreting the data.
- The two fundamental statistics concepts that play a key role in data science are descriptive and inferential statistics.

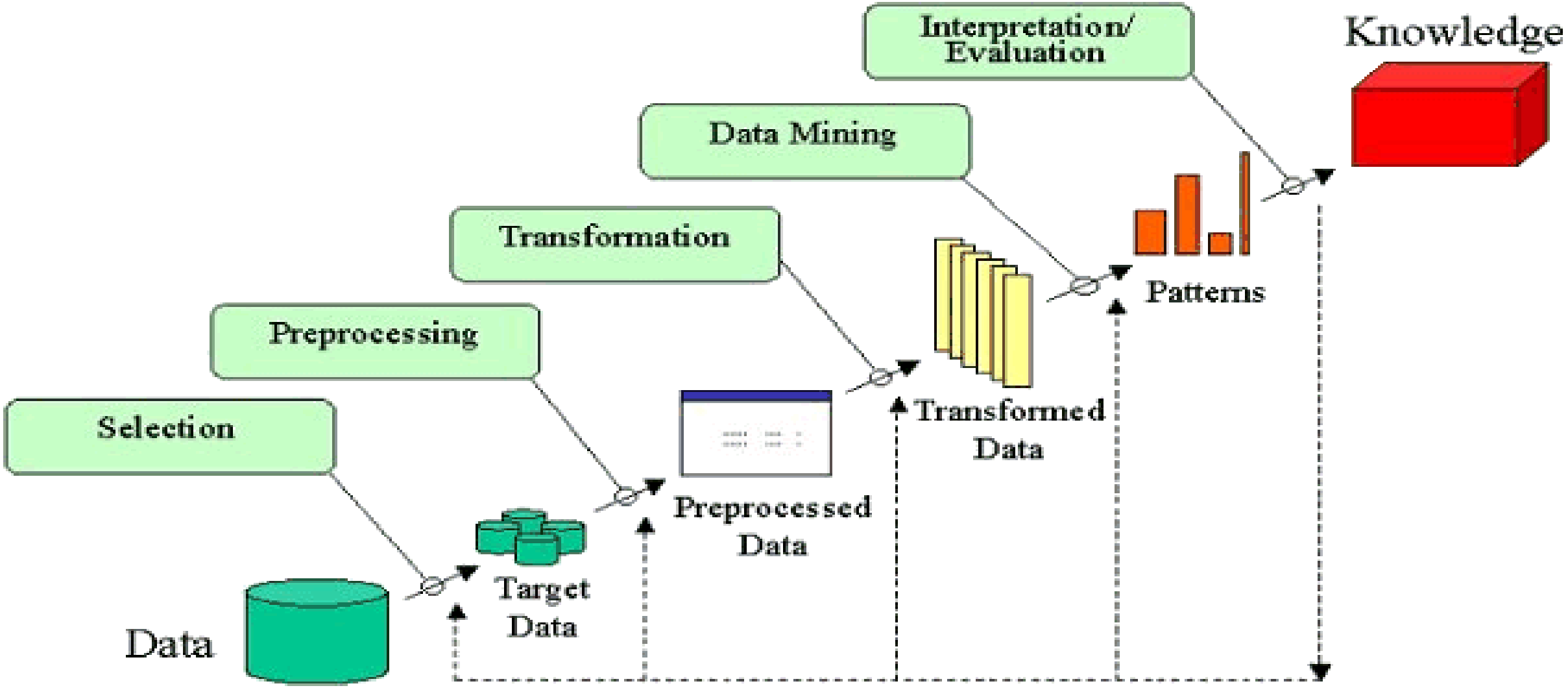
Steps in Python Machine Learning

- **We follow the following steps in Machine Learning Using Python-**
- **Collecting data.**
- **Filtering data.**
- **Analyzing data.**
- **Training algorithms.**
- **Testing algorithms.**
- **Using algorithms for future predictions.**

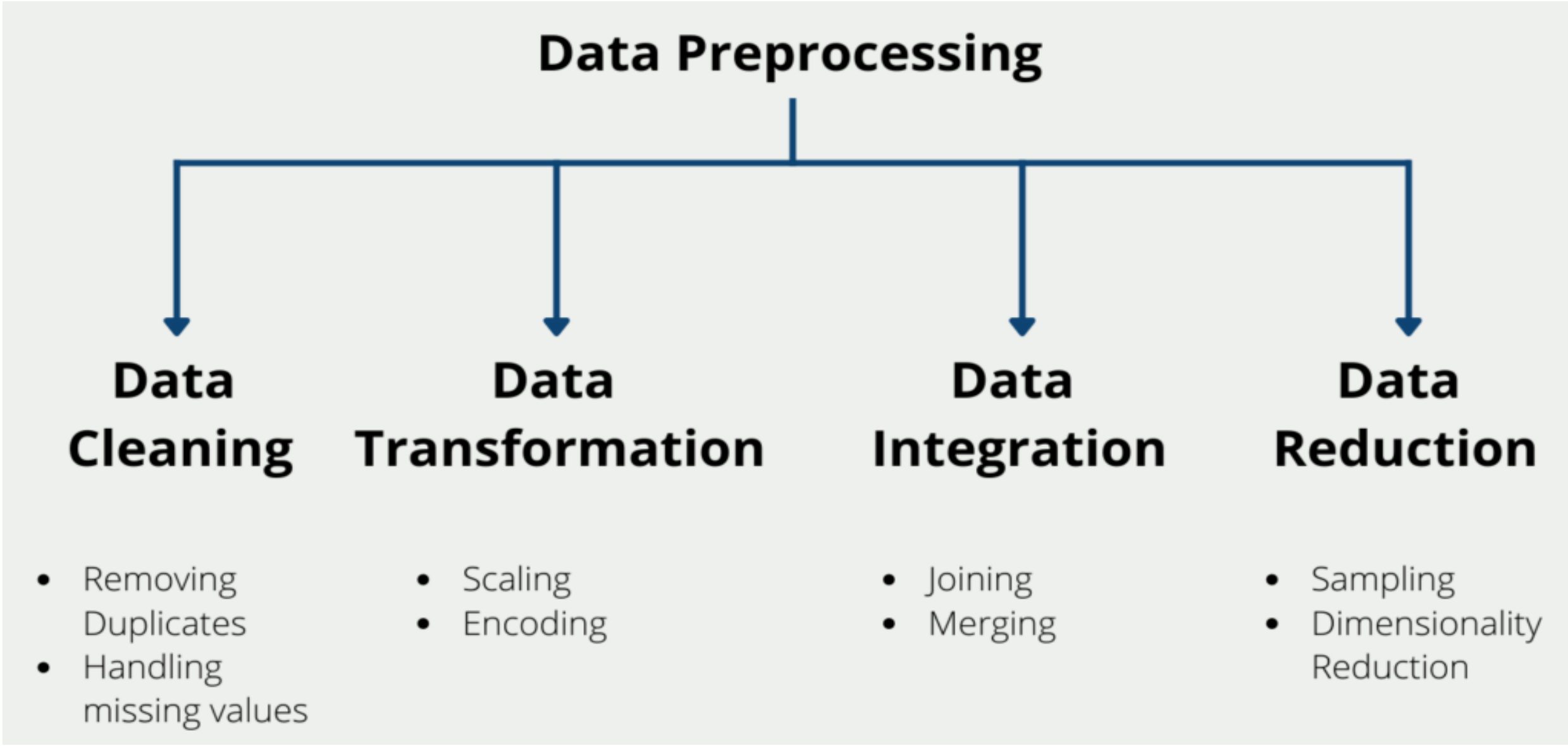
End-to-End Pipeline Diagram



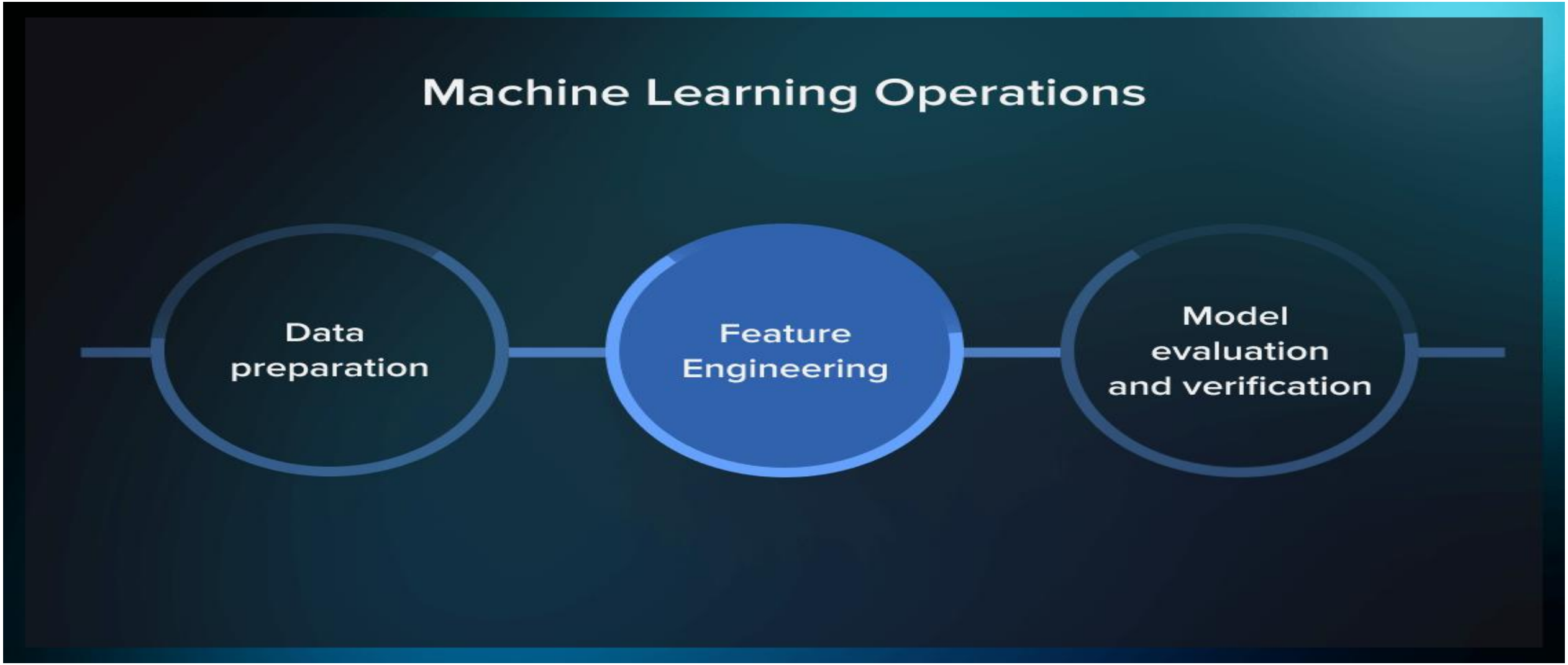
Data Pre-processing

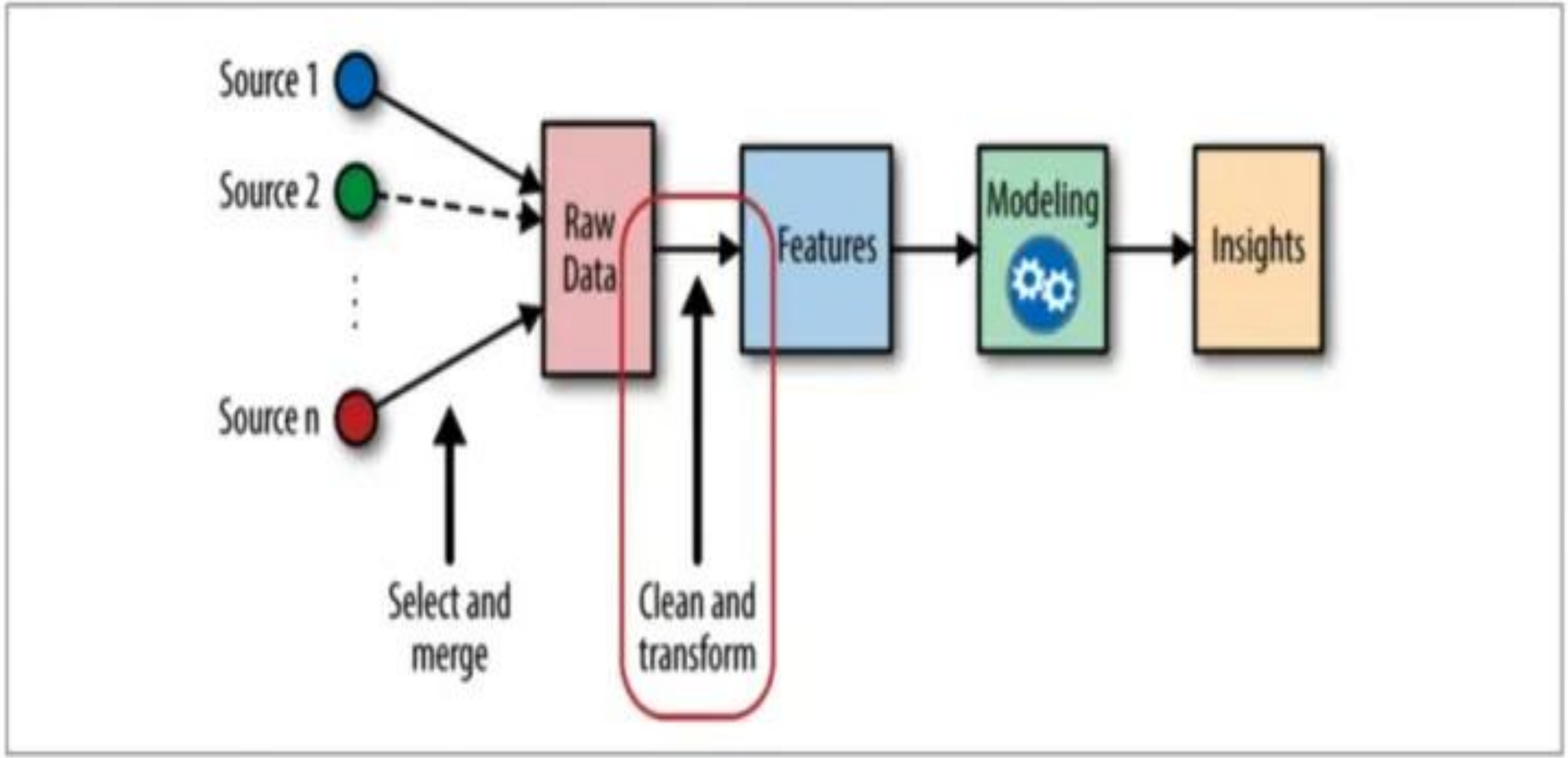


Data Pre-processing Techniques



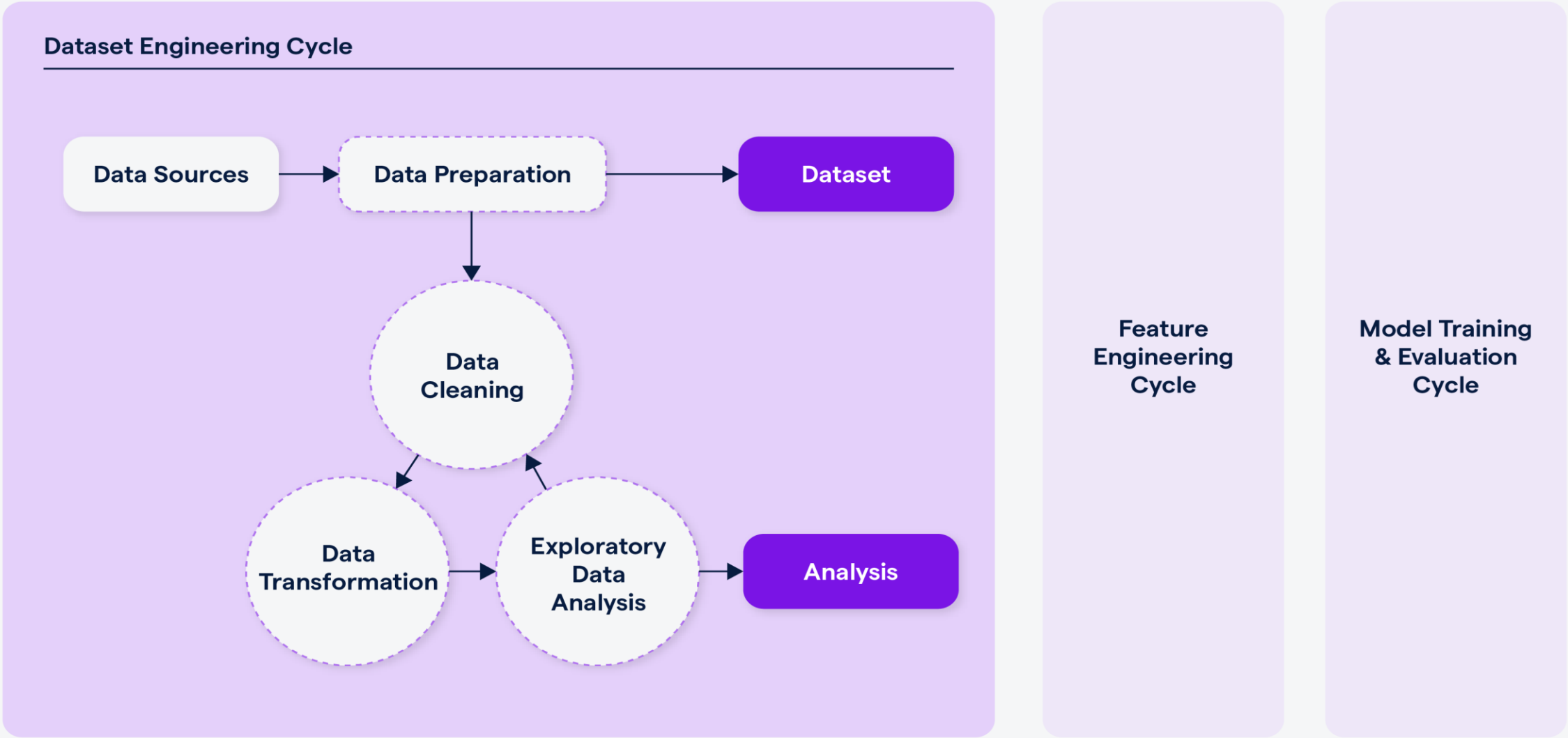
Feature Engineering



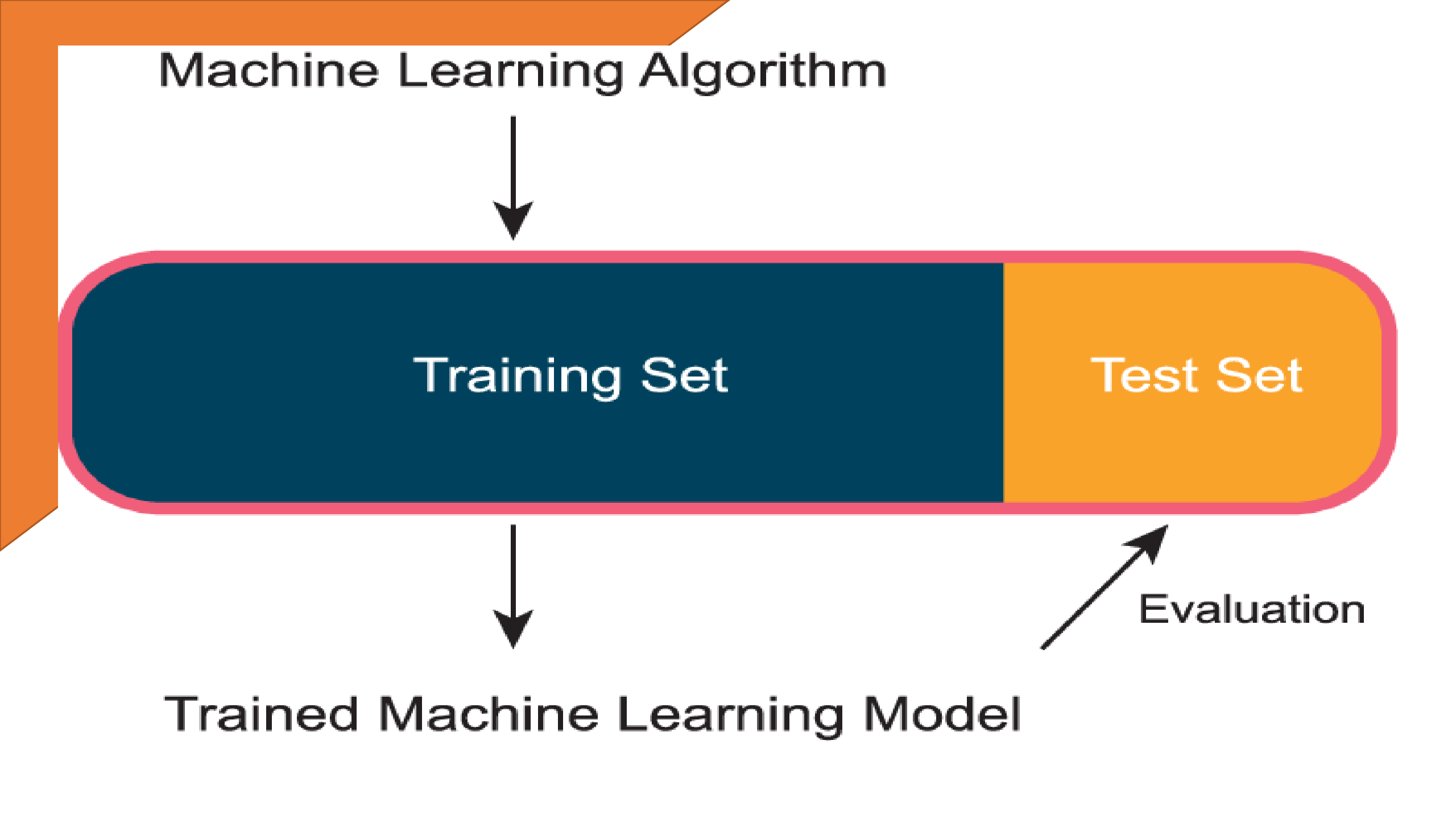


feature engineering goes here!

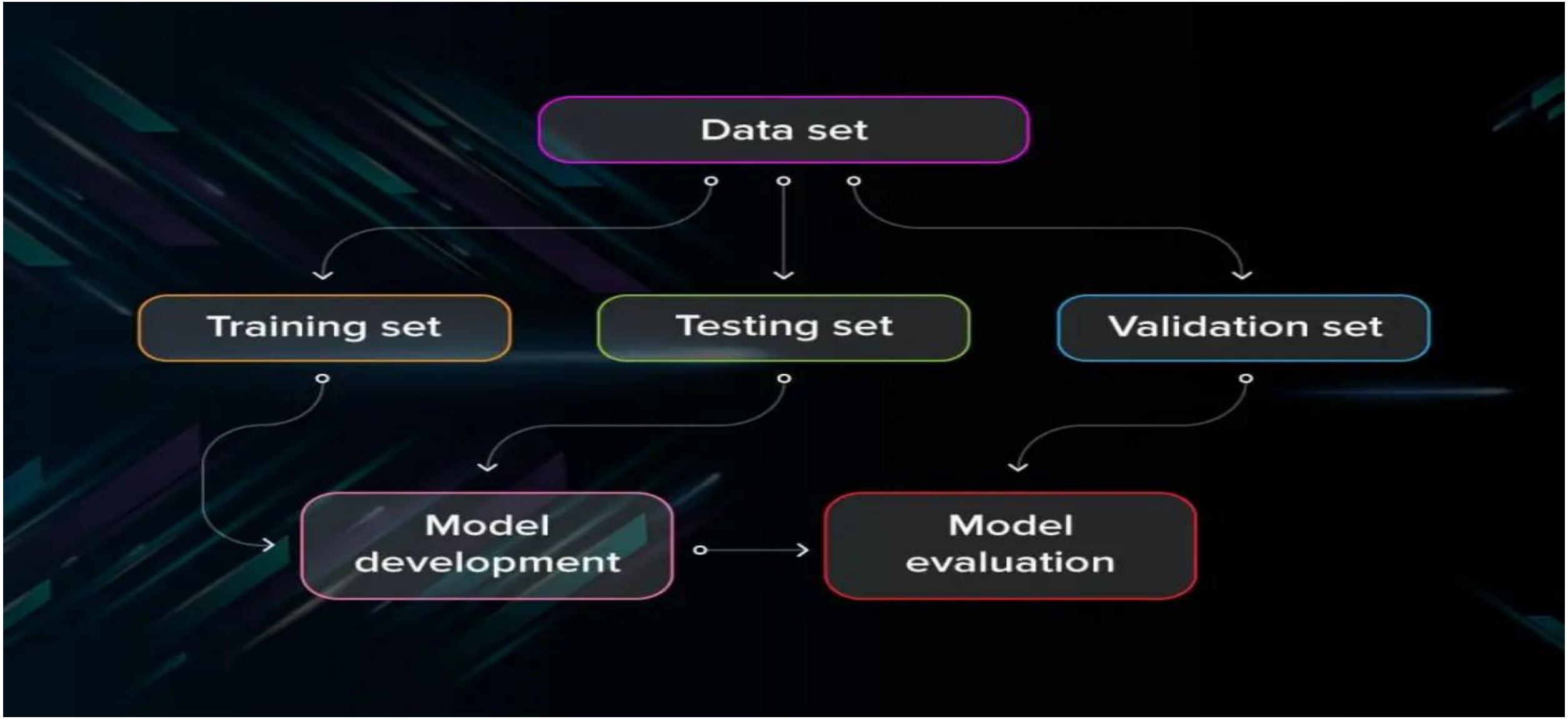
The Model Development Phase

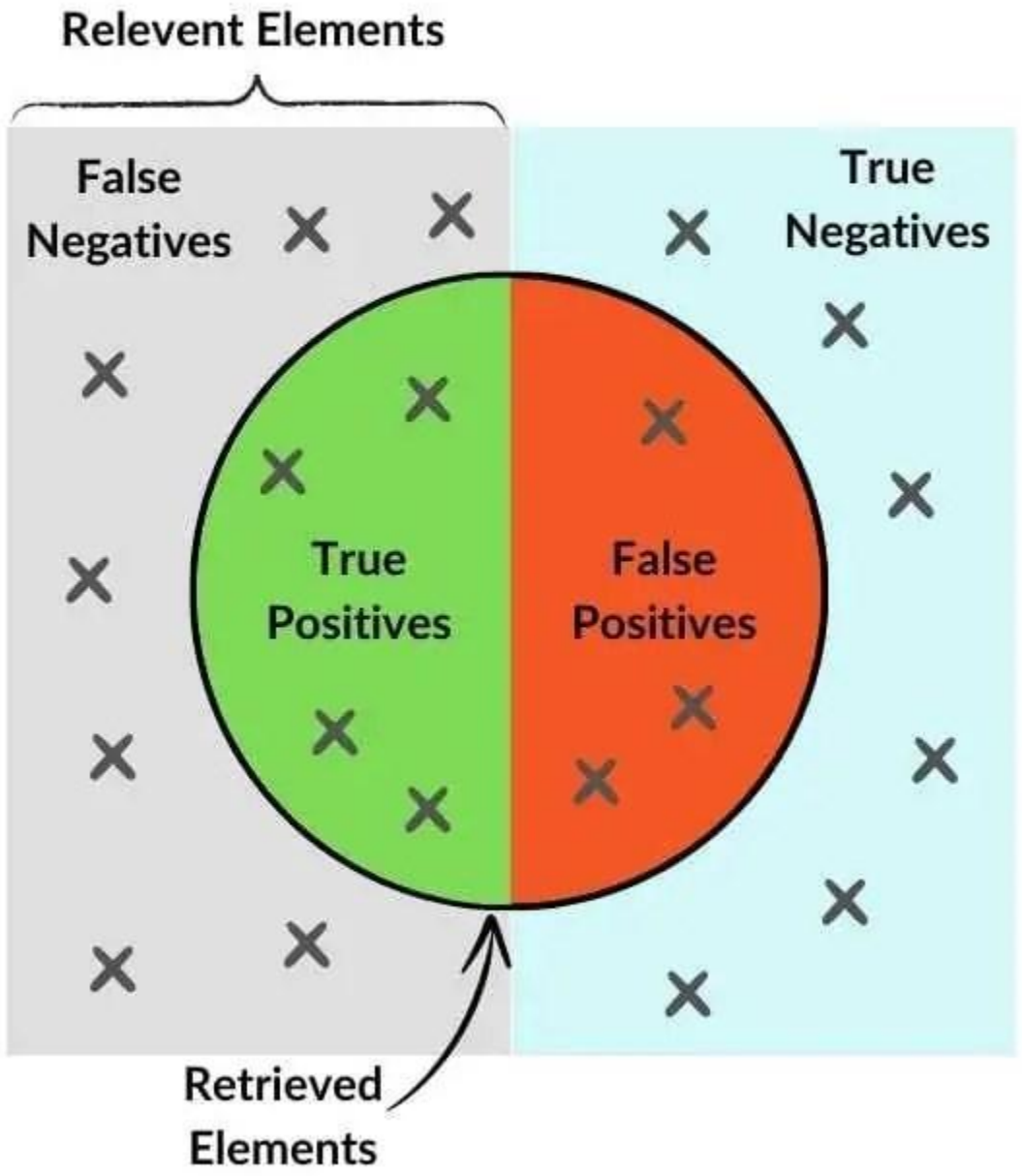


Feature engineering methods


Machine Learning Algorithm
Training Set
Test Set
Evaluation
Trained Machine Learning Model

Model Evaluation



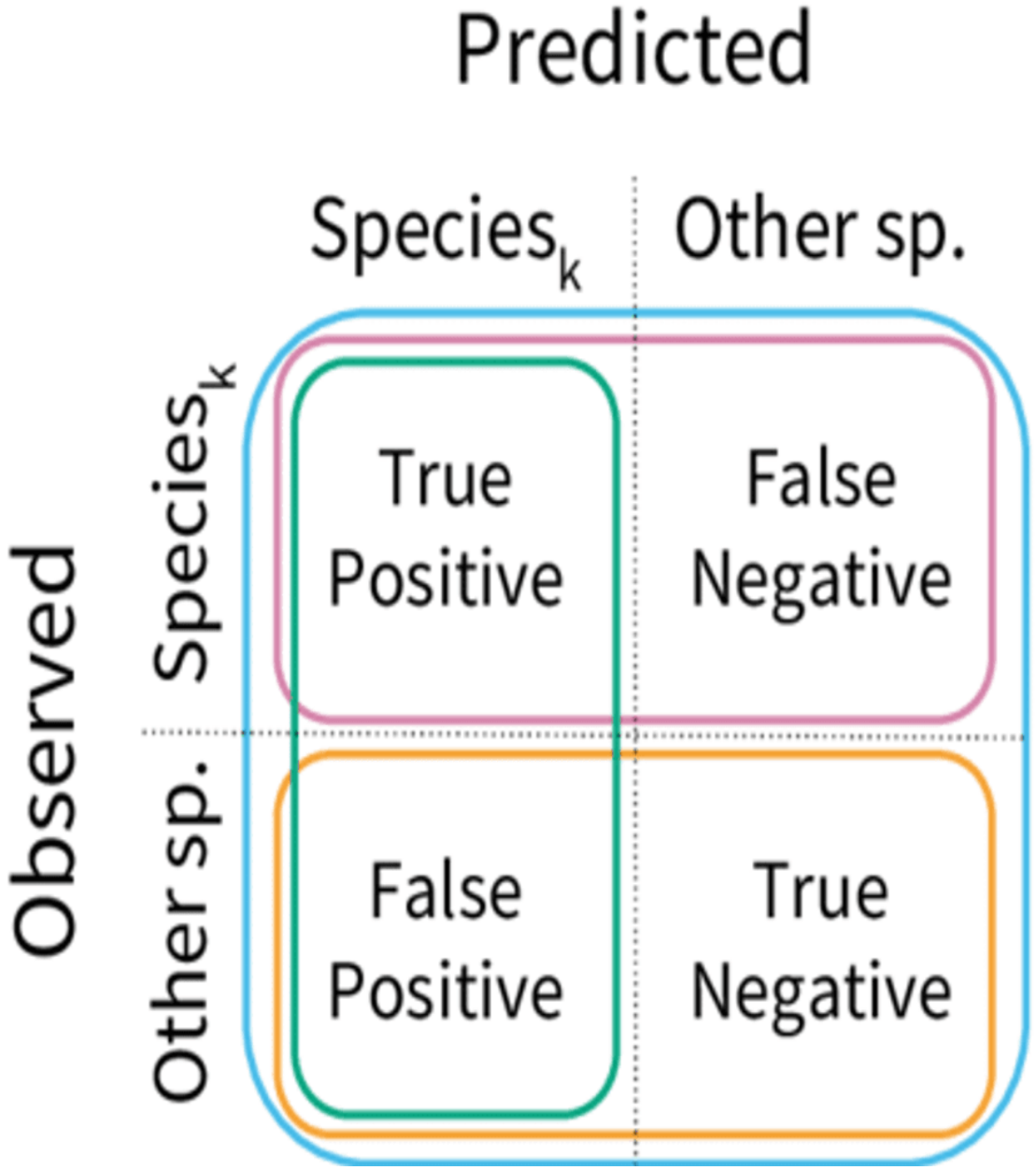


How many retrieved elements are relevant?

Precision = $\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$

How many relevant elements are retrieved?

Recall = $\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$



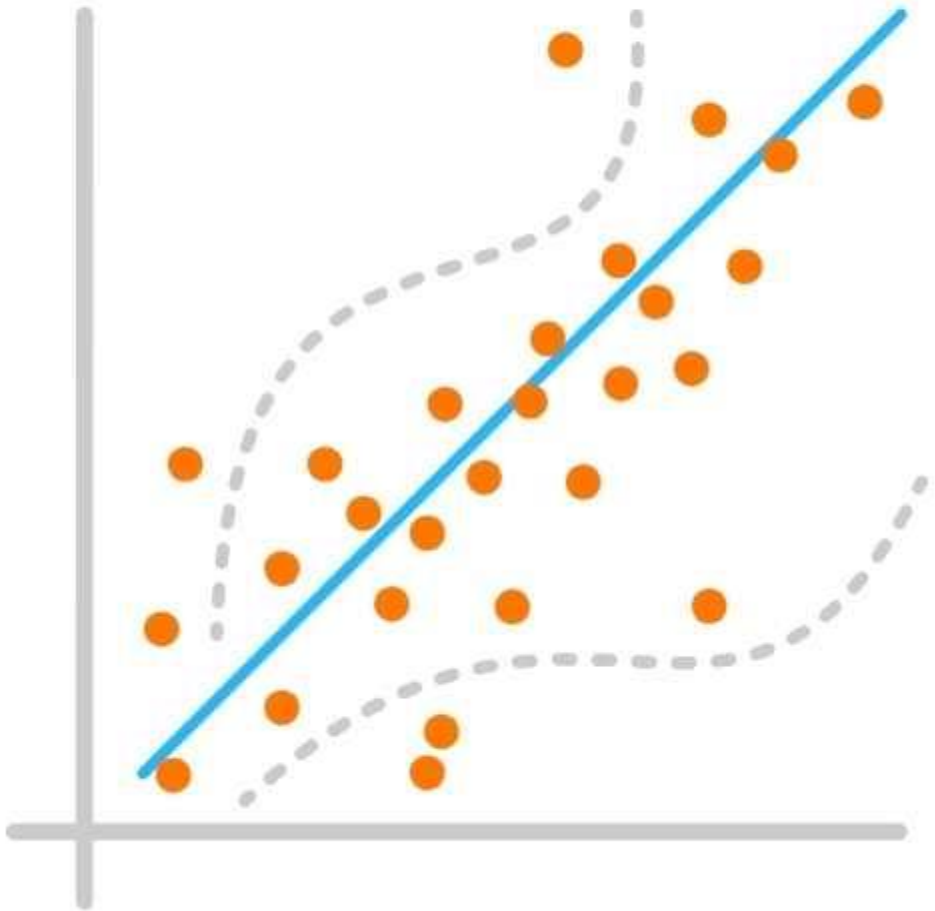
Accuracy $= \frac{TP + TN}{TP + TN + FP + FN}$

Specificity $= \frac{TN}{TN + FP}$

Precision $= \frac{TP}{TP + FP}$

Recall $= \frac{TP}{TP + FN}$

4 Common Regression Metrics



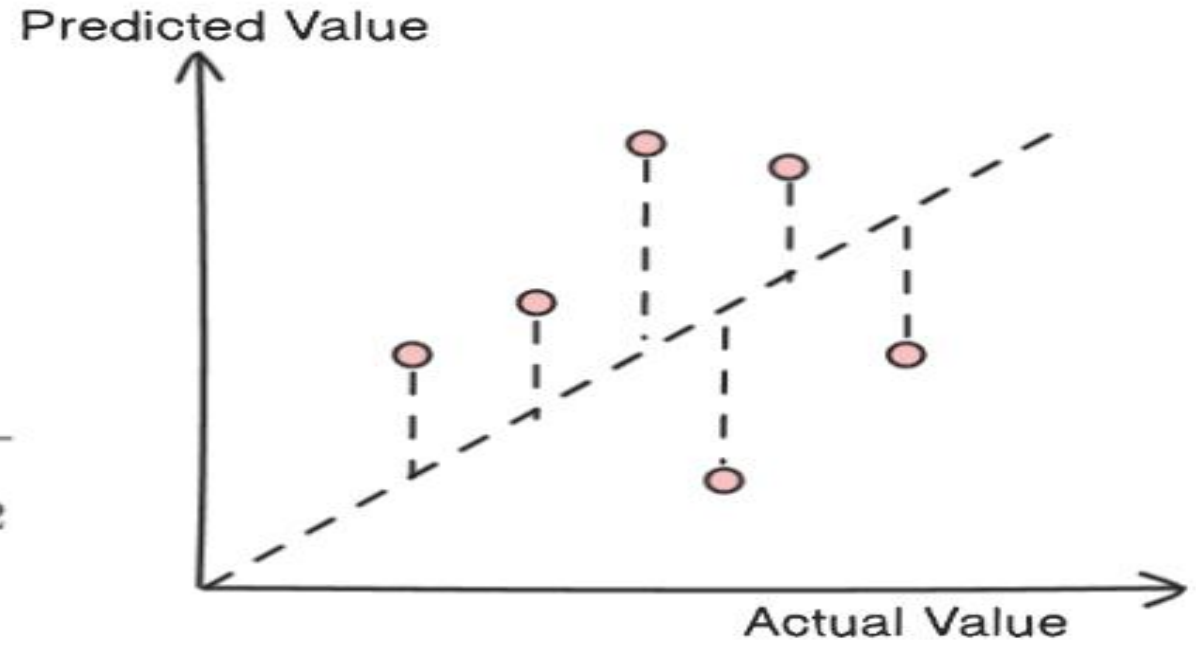
1. Mean Squared Error (MSE)
2. Root Mean Squared Error (RMSE)
3. Mean Absolute Error (MAE)
4. R-squared (R^2)

Regression Metrics

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$



$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

CROSS VALIDATION, EXPLAINED