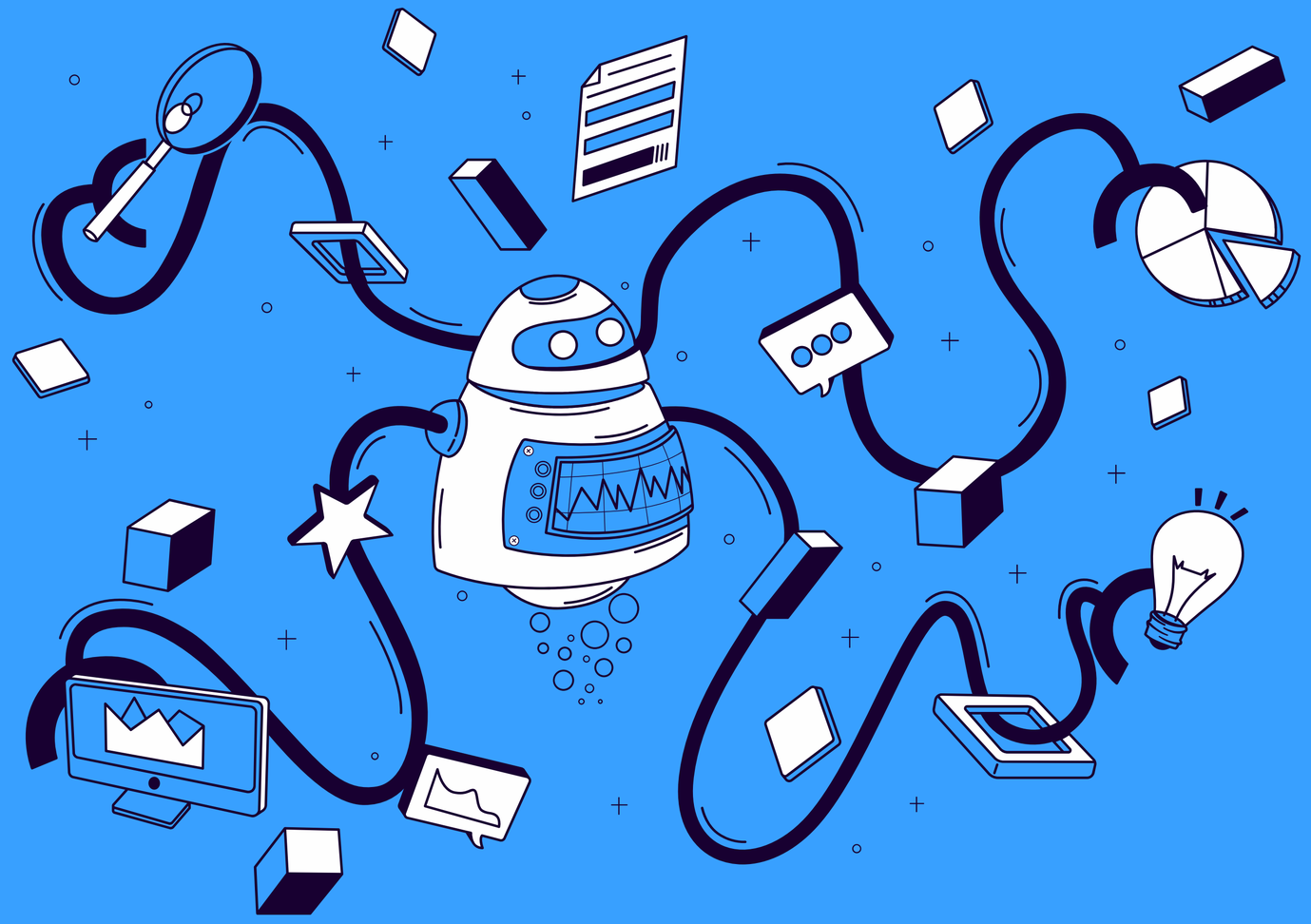
Introduction to Machine Learning

What is Machine Learning?



Computational Statistics

A method of analysis which utilizes statistical algorithms that iteratively learns from the data.

Statistical Learning

A method of statistical modelling that allows computers to learn from data and patterns.

Method of Analysis

The automation of building analytical models.

Introduction to Machine Learning

What is it used for?

Clustering

- Document Grouping
- Customer Segmentation
- Patient Similarity

Classification

- Text Sentiment Analysis
- Medical Diagnosis
- Spam Filtering

Regression

- Pricing Models
- Prediction of Equipment Downtime
- Potency Predictiona

Introduction to Machine Learning

How does it differ from Neural Networks?

A

The Modelling

Neural Networks models the behavior of biological neurons.

B

Statistical Physics

Uses concepts in Statistical Learning and Physics to learn from patterns and solve tasks.

C

Data

Neural Networks tend to require large amount of datasets.

Introduction to Machine Learning

What are the Types of ML?

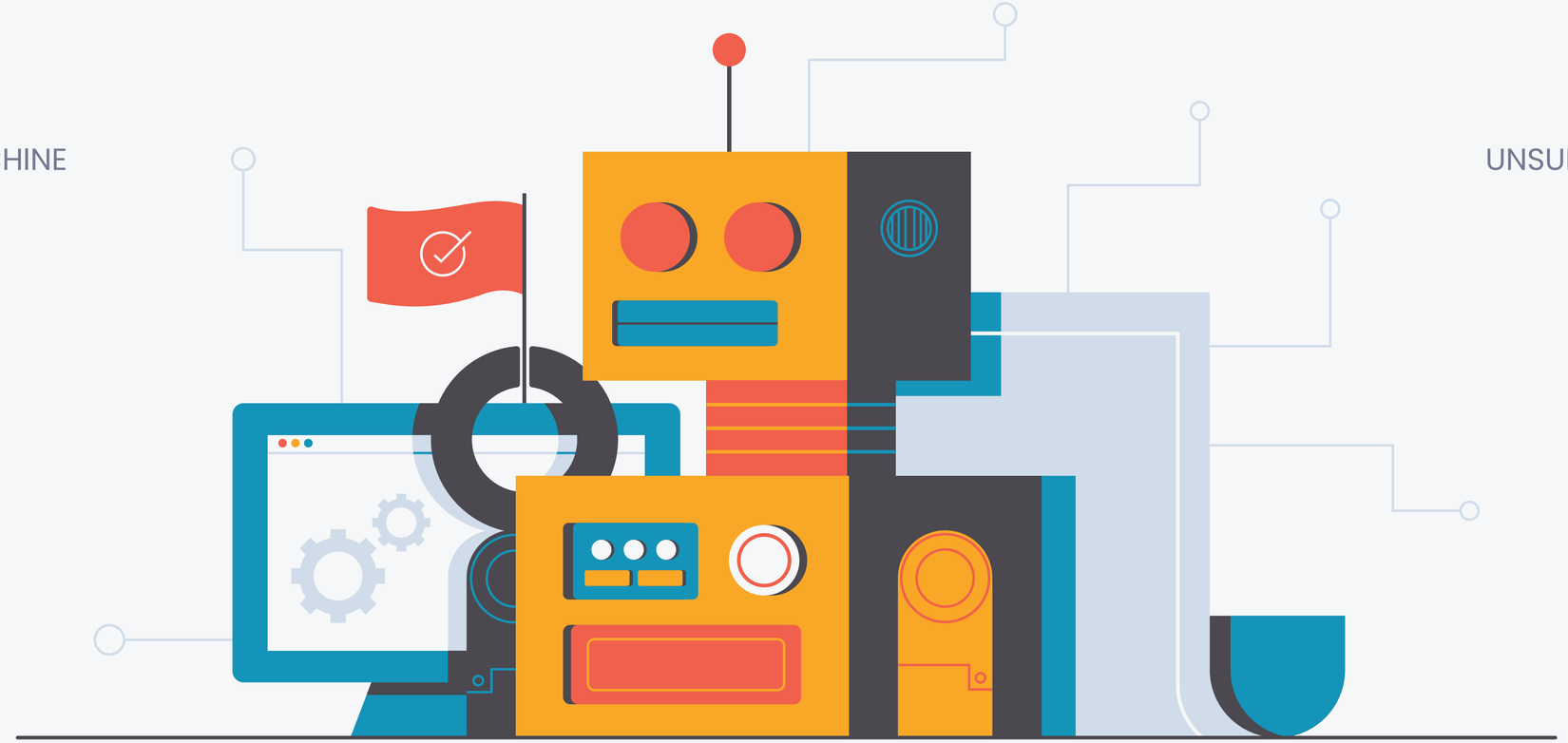
Training Data?

YES

SUPERVISED MACHINE LEARNING

NO

UNSUPERVISED MACHINE LEARNING



Introduction to Machine Learning

Supervised Machine Learning

Training Data

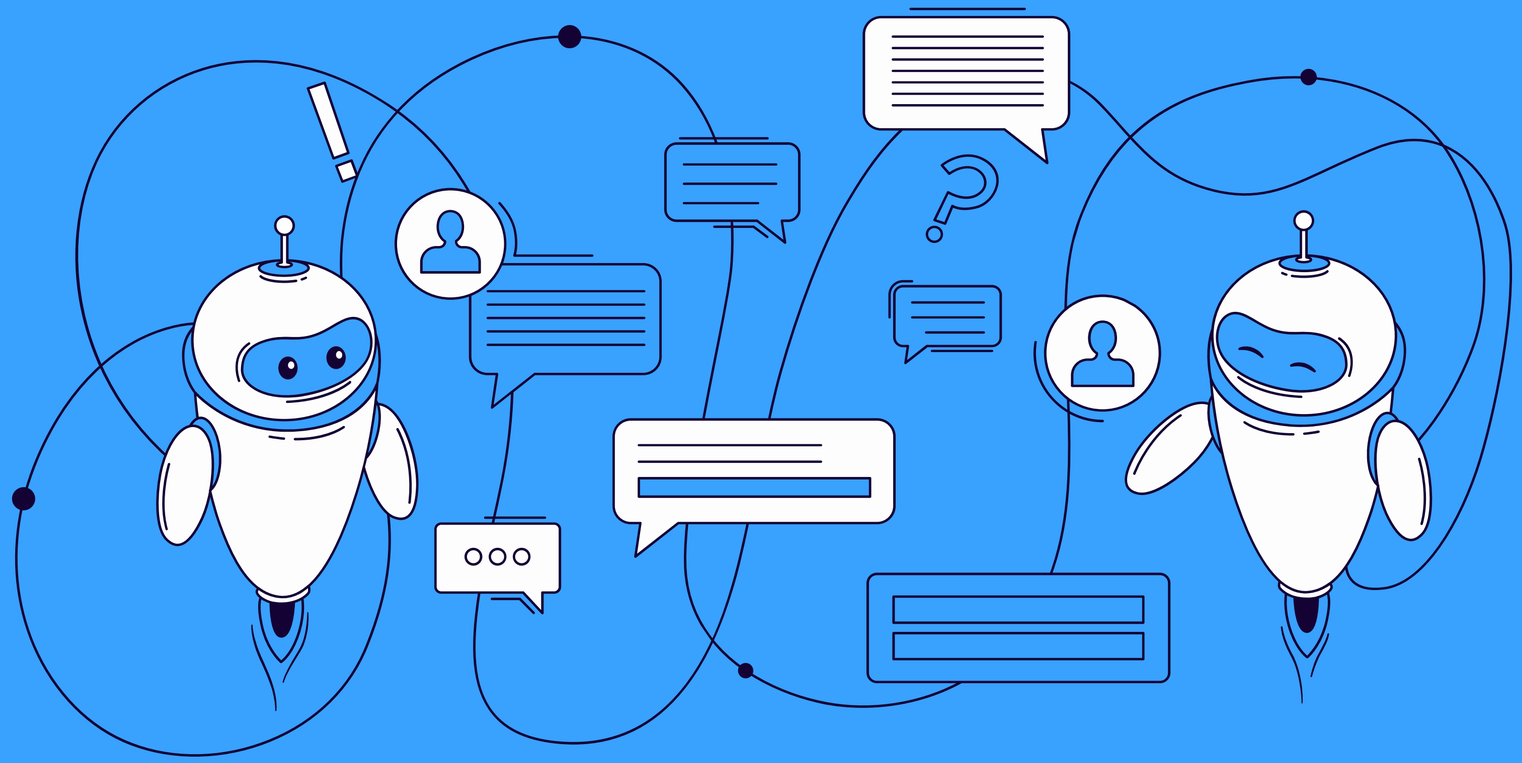
- Labelled examples
- X are the set of features.
- Y is the set of outputs/labels

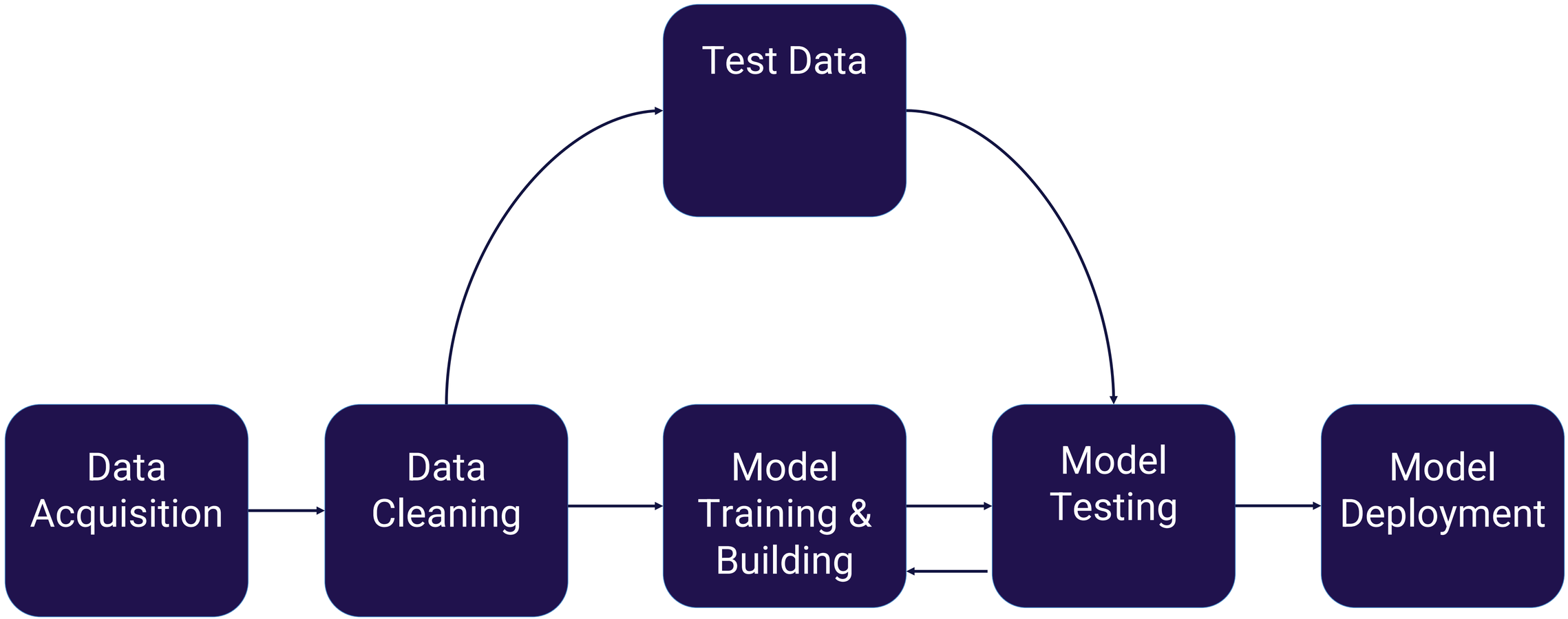
Iterative Approach

Algorithm learns from the data and compare the predicted to the actual.

Learning

The algorithm determines the error and modifies the model accordingly.




Test Data
Data Acquisition
Data Cleaning
Model Training & Building
Model Testing
Model Deployment

INTRODUCTION TO MACHINE LEARNING

IN PRACTICE

TRAINING DATA

For training model parameters.

VALIDATION DATA

To determine which model parameter to adjust.

TESTING DATA

Measures the final performance metrics of the model.

INTRODUCTION TO MACHINE LEARNING

OVERFITTING

Undesirable machine learning behavior that occurs when the machine learning model gives accurate predictions for training data but not for new data.

A

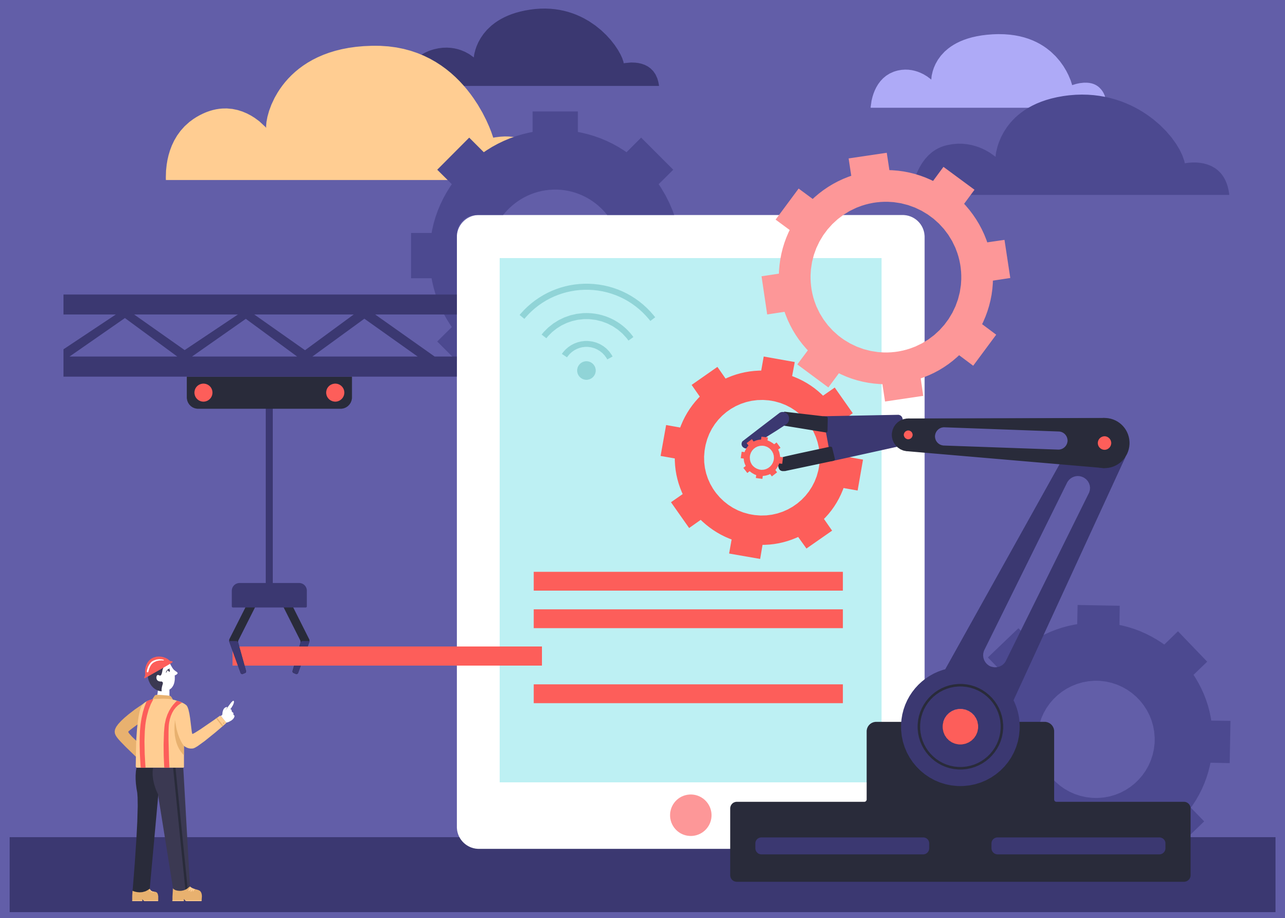
Noise

The model fits the noise in the dataset

B

Performance

Low error in training dataset but high error in the testing and validation dataset.



INTRODUCTION TO MACHINE LEARNING

UNDERFITTING

Undesirable machine learning when the model is unable to capture and learn from the pattern.

A

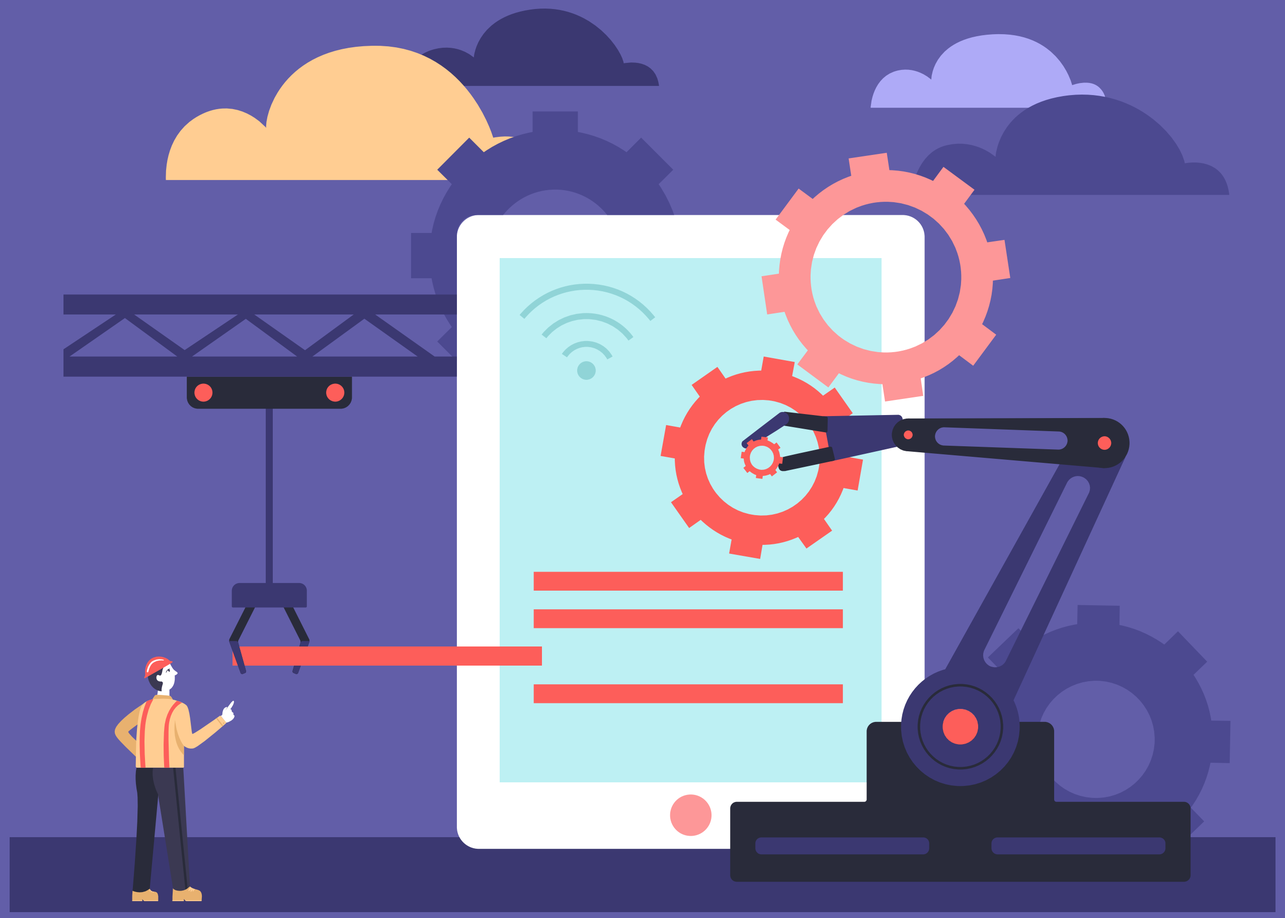
Lacking

Model did not learn the underlying pattern of the data.

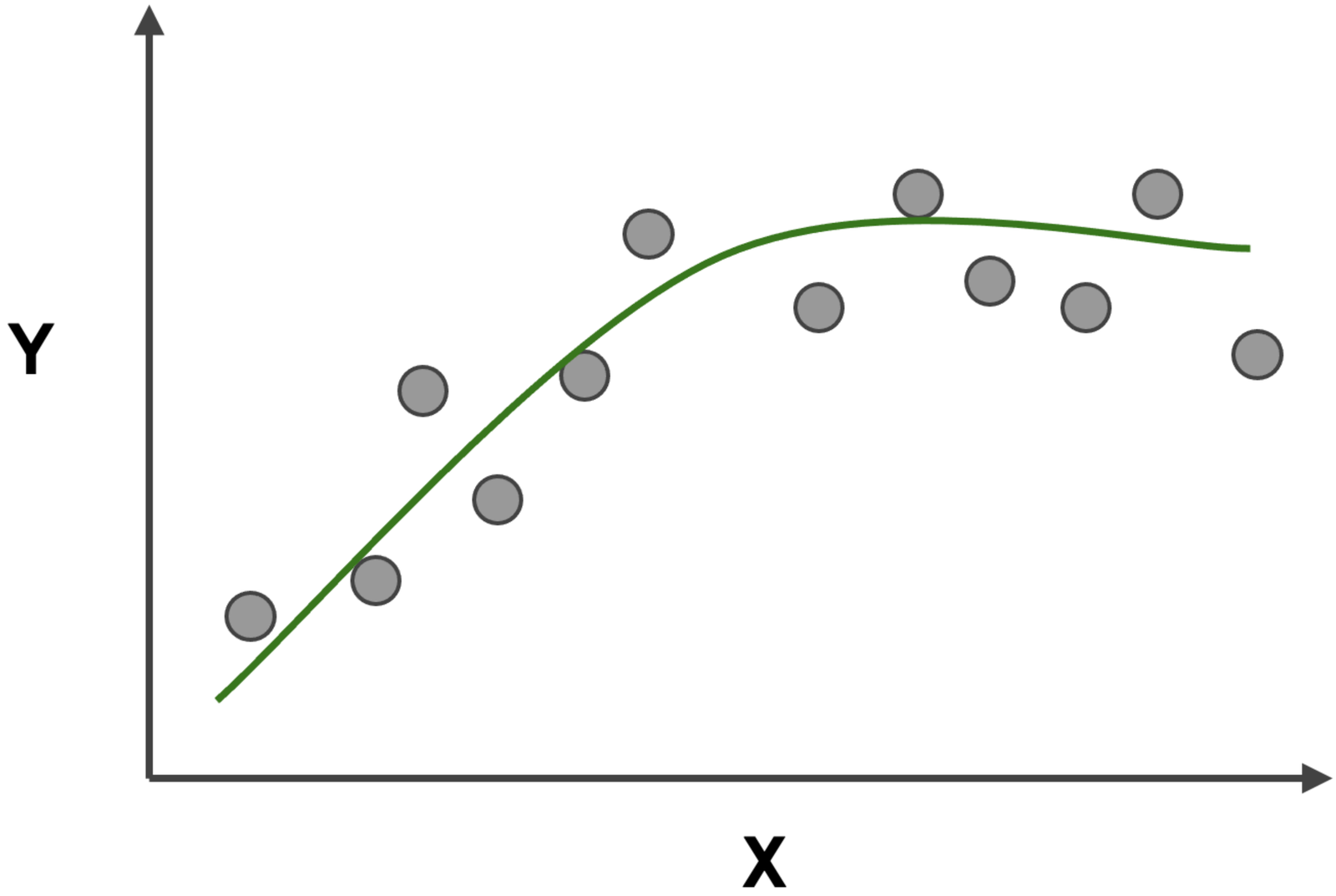
B

Performance

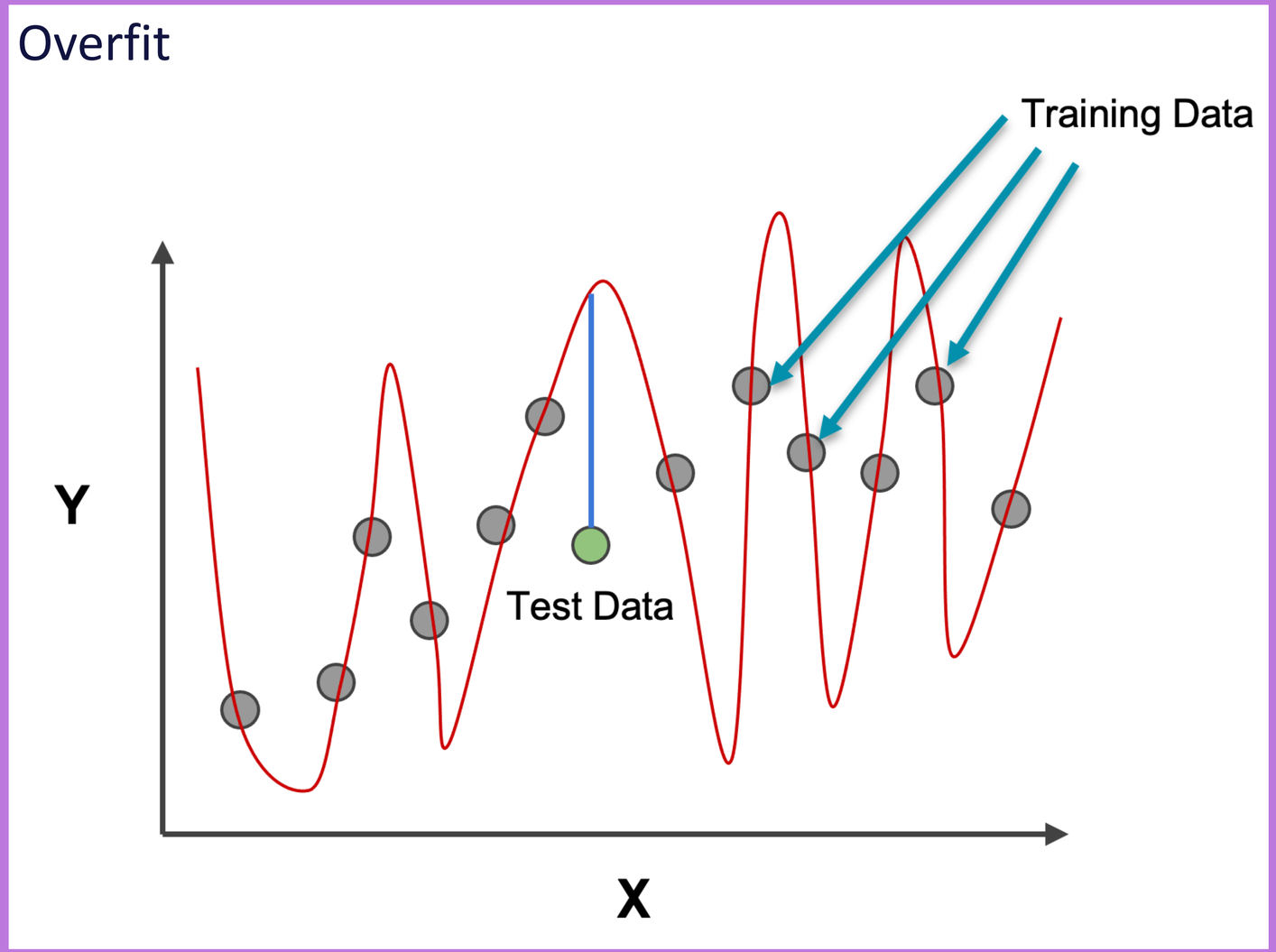
Low variance but high bias.

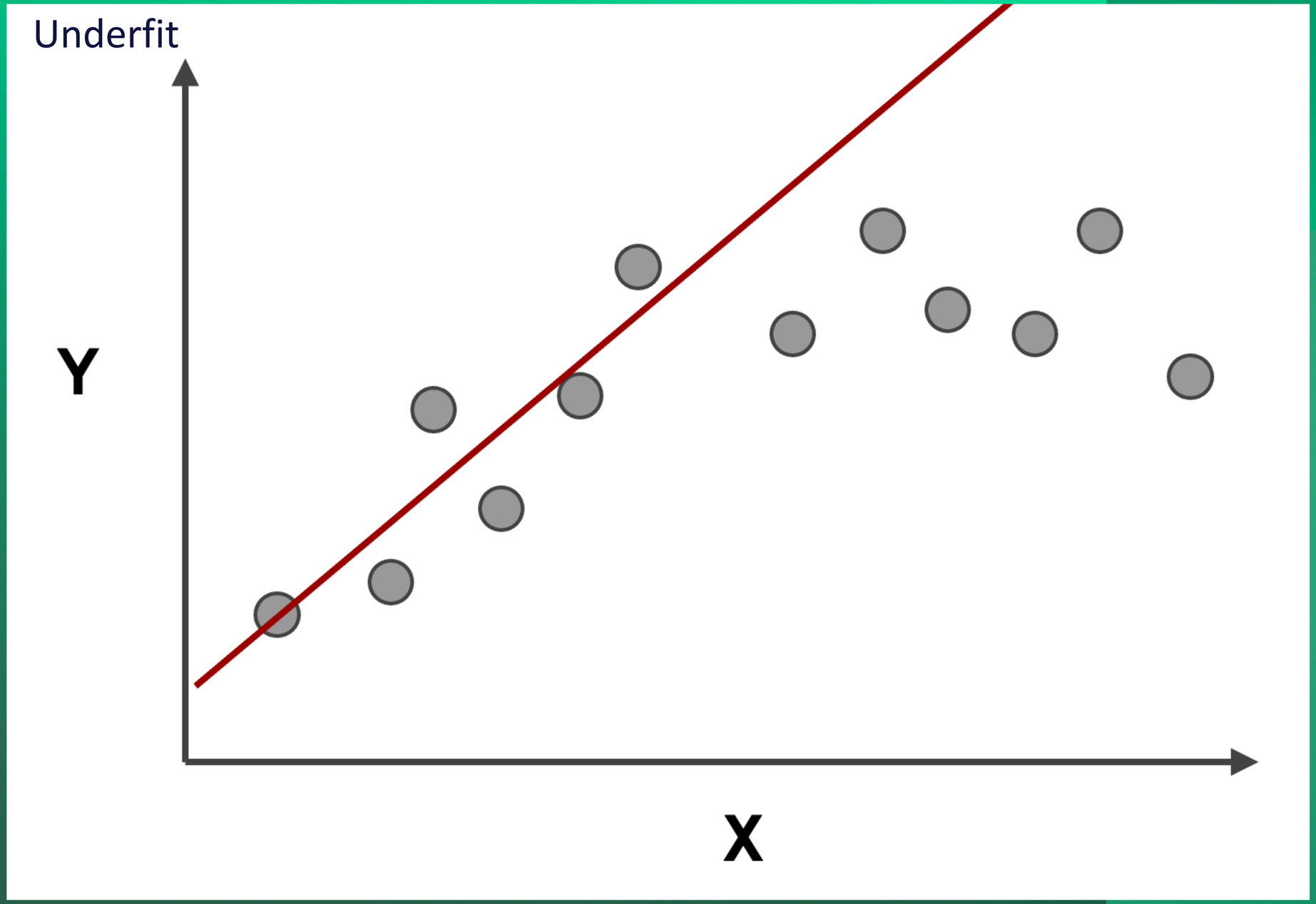


Good Fit



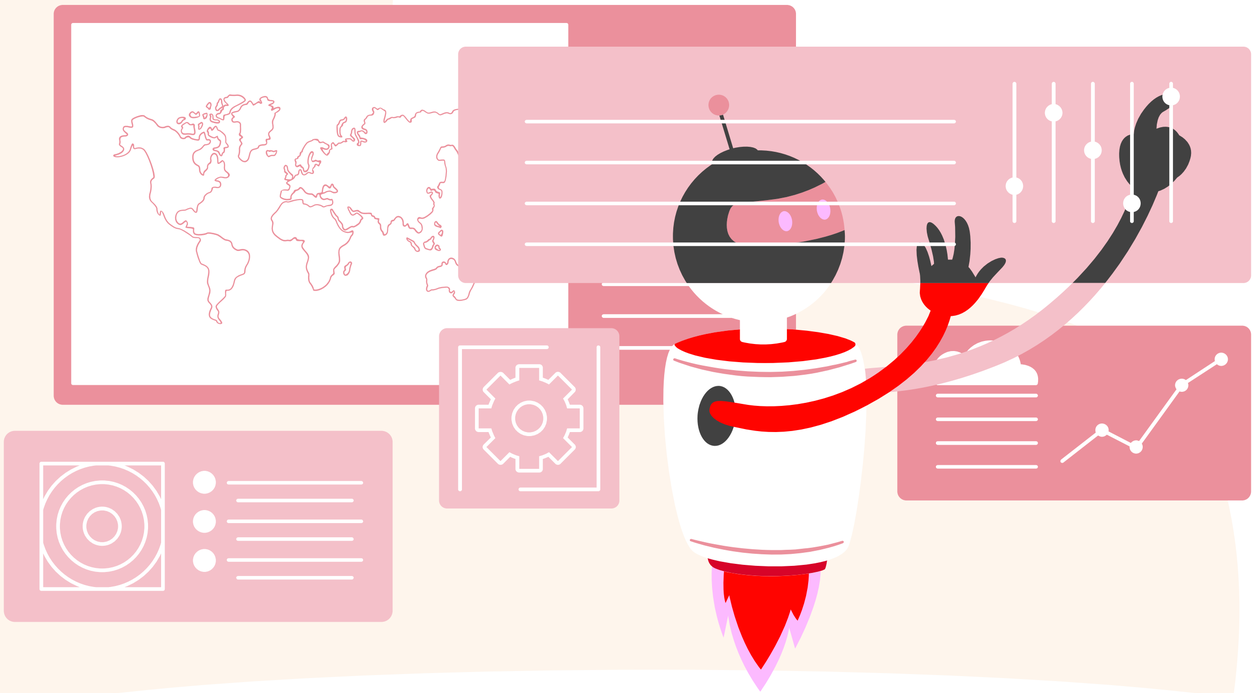
Overfit




Underfit
Y
X

INTRODUCTION TO MACHINE LEARNING

PERFORMANCE METRICS



CLASSIFICATION

- Accuracy
- Sensitivity
- Specificity
- Metrics for Multiclassification

Prediction

- Mean Absolute Error
- Mean Absolute Percentage Error
- Root Mean Squared Error
- Median Error

Test Image from X_test

TRAINED MODEL

Test Image from X_test

VENOMOUS

Correct Label from y_test

TRAINED MODEL

Test Image from X_test

VENOMOUS

Correct Label from y_test

TRAINED MODEL

VENOMOUS

Prediction on Test Image

Test Image from X_test

TRAINED MODEL

VENOMOUS

Prediction on Test Image

VENOMOUS

Correct Label from y_test

VENOMOUS == VENOMOUS

Compare Prediction to Correct Label

Test Image from X_test

TRAINED MODEL

SI BESHIE

Prediction on Test Image

VENOMOUS

Correct Label from y_test

VENOMOUS == SI BESHIE

Compare Prediction to Correct Label

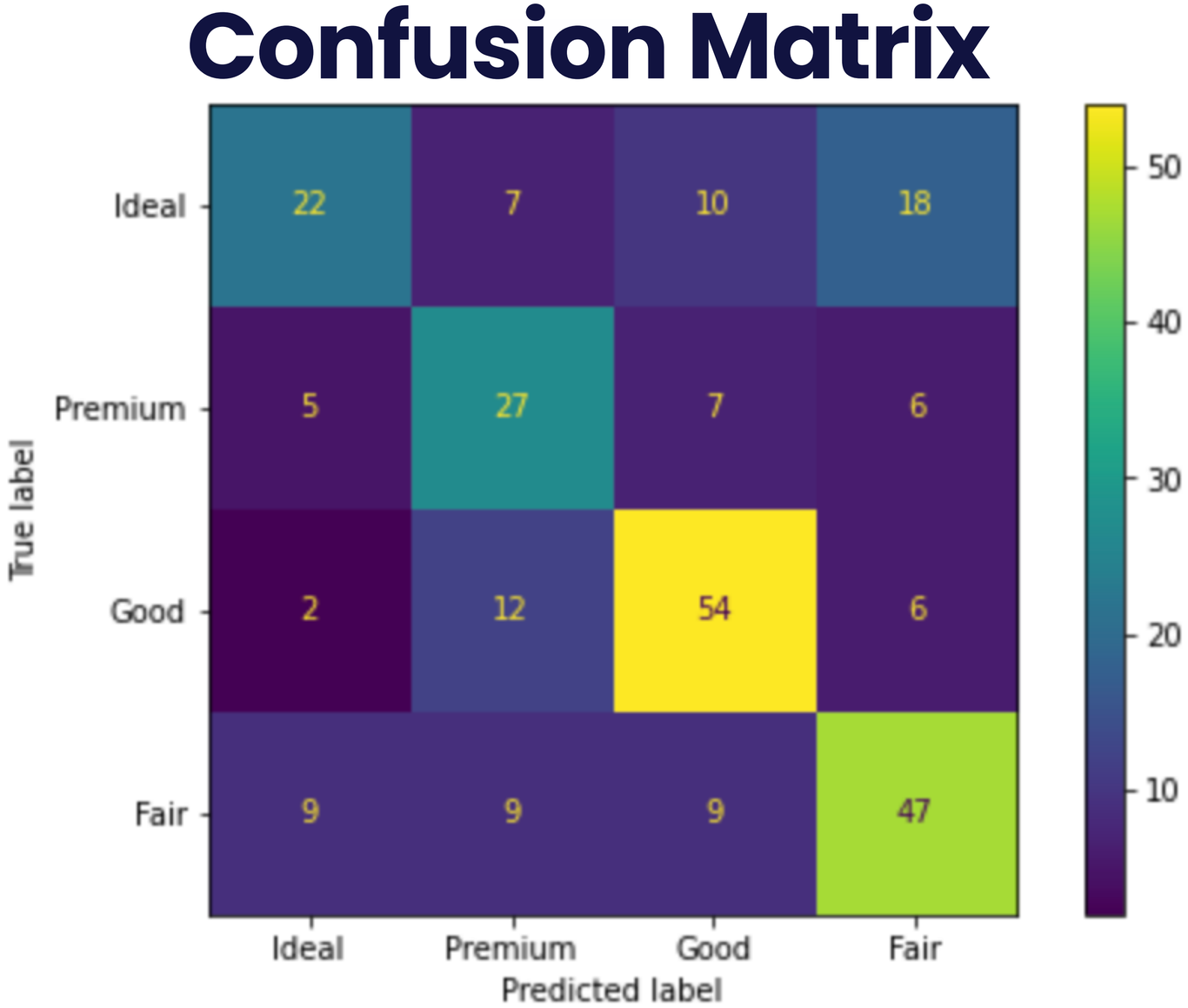
Classification Metrics

- $Accuracy = \frac{Number\ of\ Correctly\ Classified}{Total\ Number\ of\ Samples}$
- $Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negatives}$
- $Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$
- $Critical\ Success\ Index\ (Jaccard's\ Statistics) = \frac{True\ Positive}{True\ Negative + False\ Positive + False\ Negative}$
- $Youden's\ Statistics = Specificity + Sensitivity\ - 1$

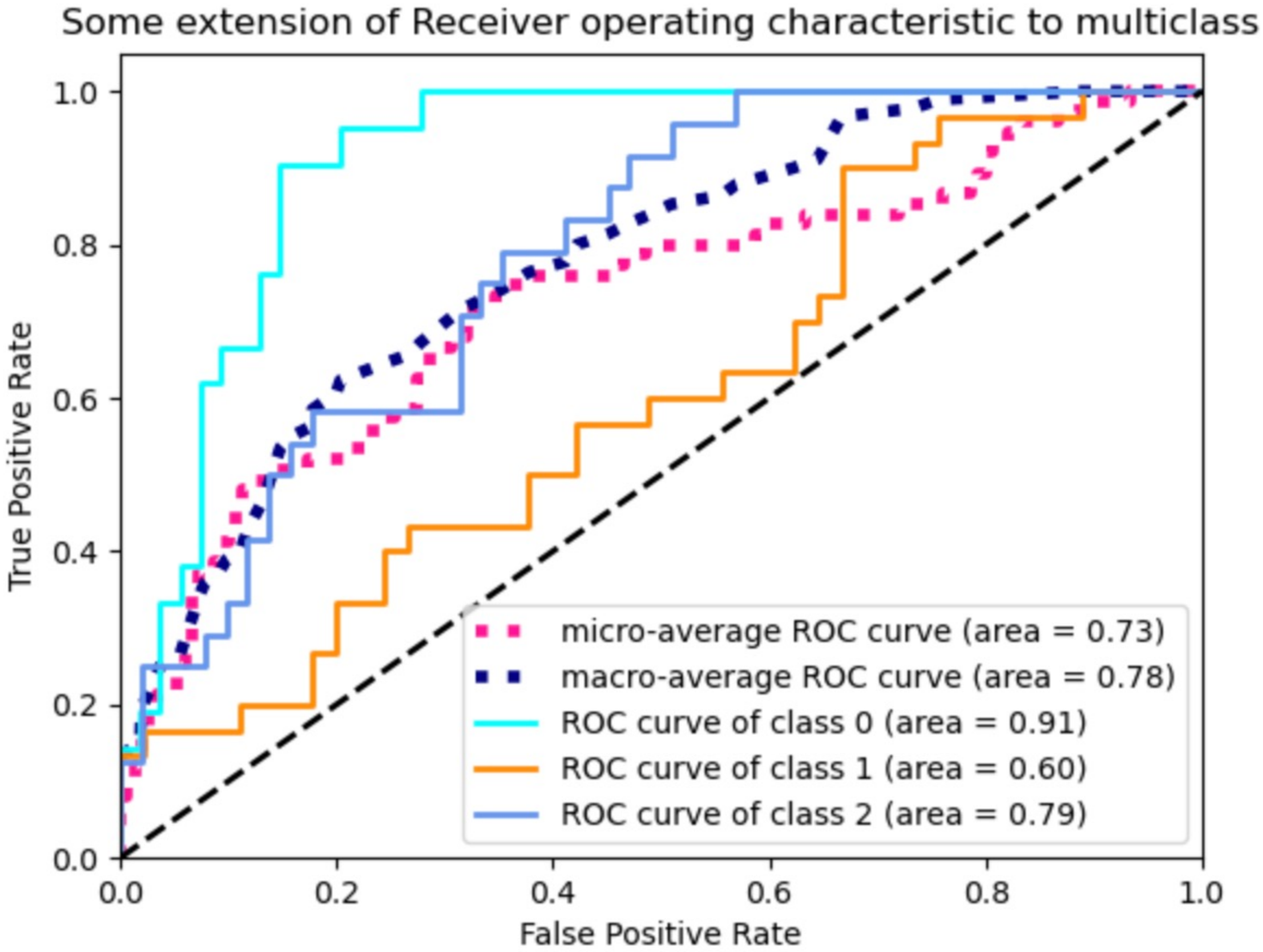
What About Multiclass Data?

		predicted condition	
	total population	prediction positive	prediction negative
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**

But what about Multiclass Data?



What About Multiclass Data?



Performance Metrics for Regression

$$\text{Mean Absolute Error} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}|$$

Advantages:

- **Easy to understand and interpret.**
- **Treats errors equally, not giving disproportionate weights to large errors.**

Disadvantage:

- Not penalizing the impacts of outliers.

Performance Metrics for Regression

Mean Absolute Percent Error = $\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\widehat{y_i}}{y_i}\right|$

Advantages:

- **Provides error as a percentage metric (easy interpretation).**

Disadvantage:

- **Can be undefined large values for y_i closer to zero.**
- **Biased towards predictions that are systematically towards the actual values.**

Performance Metrics for Regression

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

Advantages:

- **Penalizes large errors.**

Disadvantage:

- **Disproportionately penalize large errors.**
- **Sensitive to outliers.**
- **Units are not the same as the original data (less intuitive interpretations).**

Performance Metrics for Regression

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

Advantages:

- **Penalizes large errors.**
- **Units are the same as the original data (more intuitive interpretations).**

Disadvantage:

- Sensitive to outliers.

Which Regression Metrics to Use?

1. **Domain knowledge plays a key role, example in predicting machine downtime small fluctuations in error may not be significant but this is not the case if you are trying to predict dosage of a medication.**

2. **Plot the error values to see its distribution and compare the error metric.**

3. **Do summary statistics of your error values and compare your error metric.**

INTRODUCTION TO MACHINE LEARNING

UNSUPERVISED MACHINE LEARNING

A

DATA HAS NO LABEL

- Previously discussed performance metrics cannot be used.
- Uses measures of distances and similarity.

B

PATTERN DISCOVERY

- The model is tasked to learn from the underlying distribution.
- Capture patterns or behavior of the data.

INTRODUCTION TO MACHINE LEARNING

TYPES OF UNSUPERVISED MACHINE LEARNING

DIMENSIONALITY REDUCTION

- A technique used to reduce the number of variables in a high dimensional data.
- Makes the data more manageable.
- Linear Algebra: PCA, NMF, SVD
- Statistical Embedding : T-SNE
- Statistical Physics: UMAP.
- Deep Learning: Autoencoders

ANOMALY DETECTION

- Also known as outlier detection.
- Identification of data points that deviate from normal behavior.
- Tree based: Isolation Forest
- Density Bsed: Local Outlier Factor, DBSCAN
- Statistics: Z-Scores, Interquartile range, Box-Whiskers.

CLUSTERING

- Collecting similar objects into a common group called cluster or segment.
- Distance based: K-Means
- Tree based: Hierarchical Clustering
- Density Based: Density-Based Spatial Clustering of Application with Noise (DBSCAN)
- Linear Algebra: Spectral Clustering

ASSOCIATION RULE

- Also called the Market Basket Analysis.
- Reveals how items or events are associated with each other.