Data Science Fundamentals

Learning Objectives

Outline the data science cycle and machine learning process

Explain the commonly used feature selection and feature engineering methods

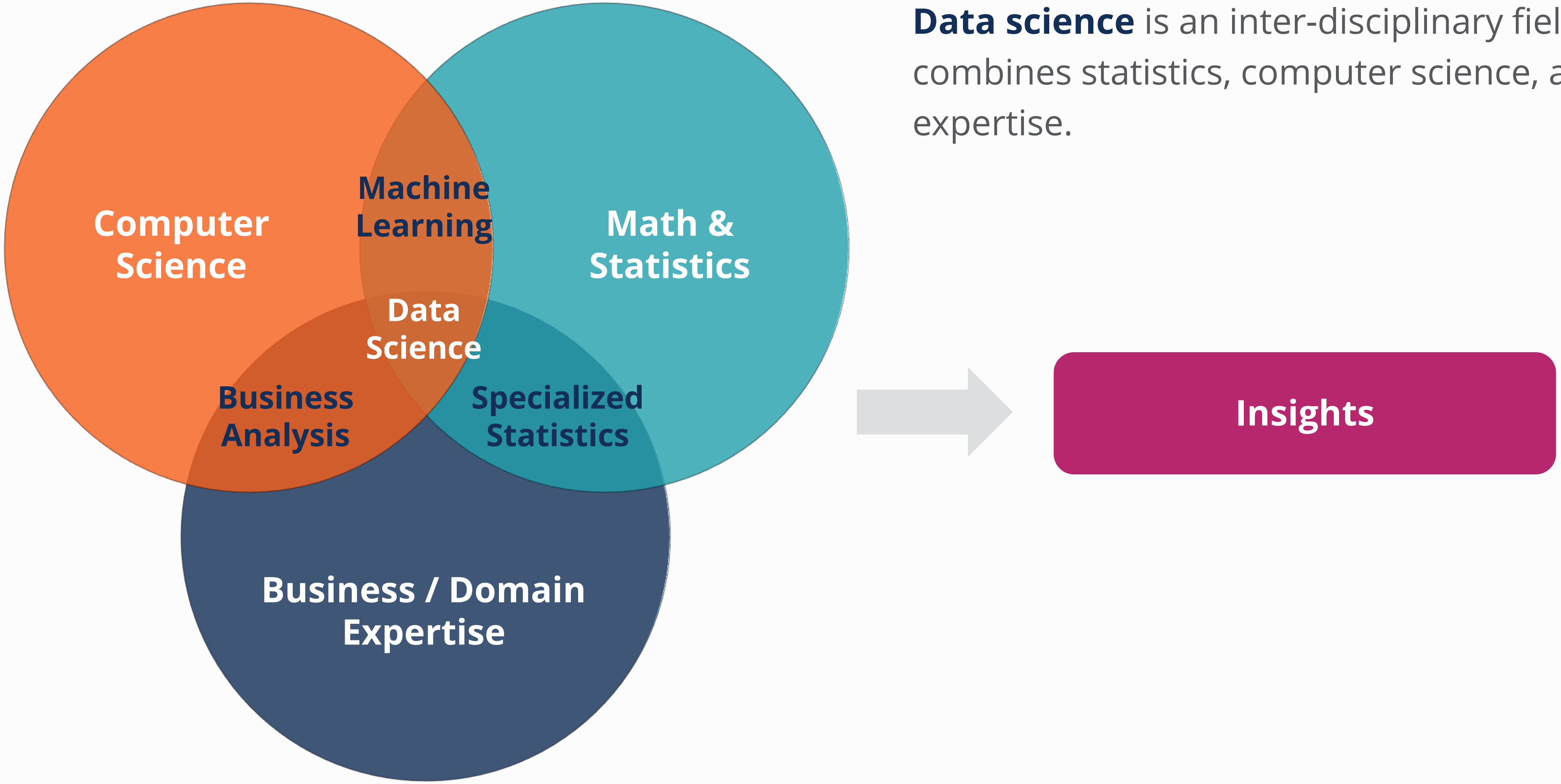
List the algorithms mostly used in supervised and unsupervised learning

Read the key metrics used to evaluate a machine learning model

Explain the techniques used to improve an underfitting or overfitting model

Data Science Introduction

What Is Data Science



Data science is an inter-disciplinary field that combines statistics, computer science, and domain expertise.

How Is Data Science Used in Business

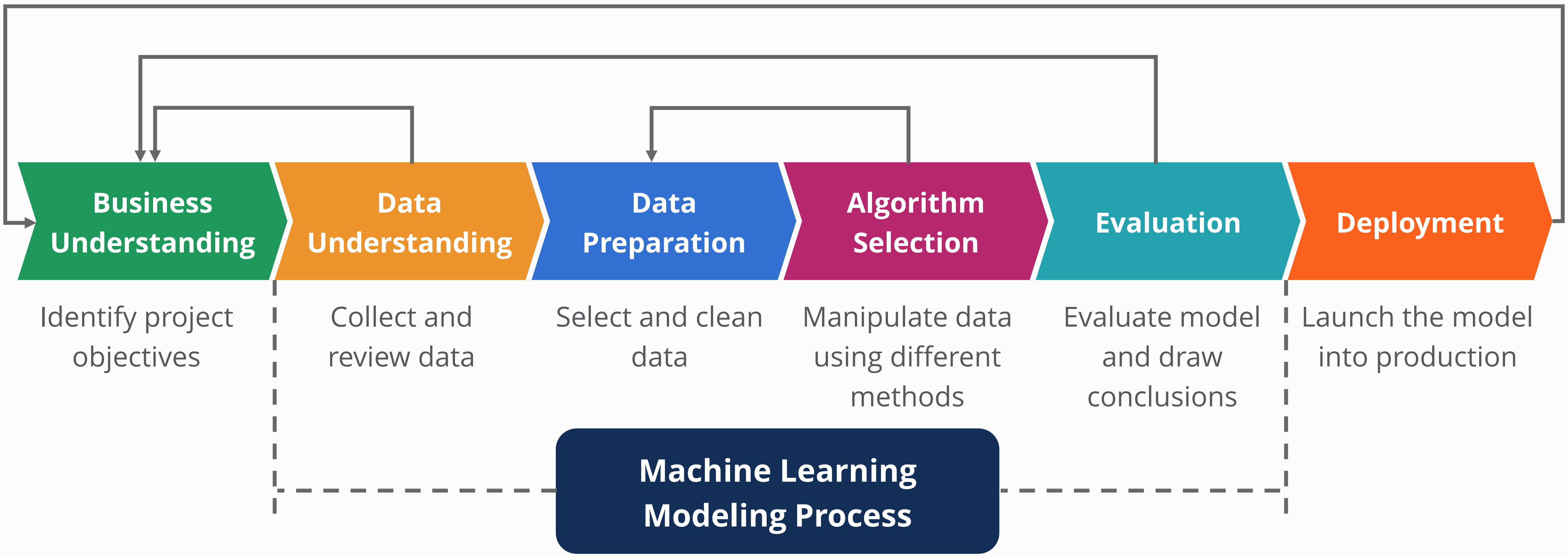
Data science can be used to answer any business question and drive business decisions.

Prevent fraudulent financial transactions and enhance risk mitigation.

Understand user behavior and design better products.

Predict which customers will abandon a product or service.

Data Science Cycle



There is constant feedback going on throughout this cycle.

Machine Learning Overview

Machine learning uses computer algorithms to make predictions from input data.

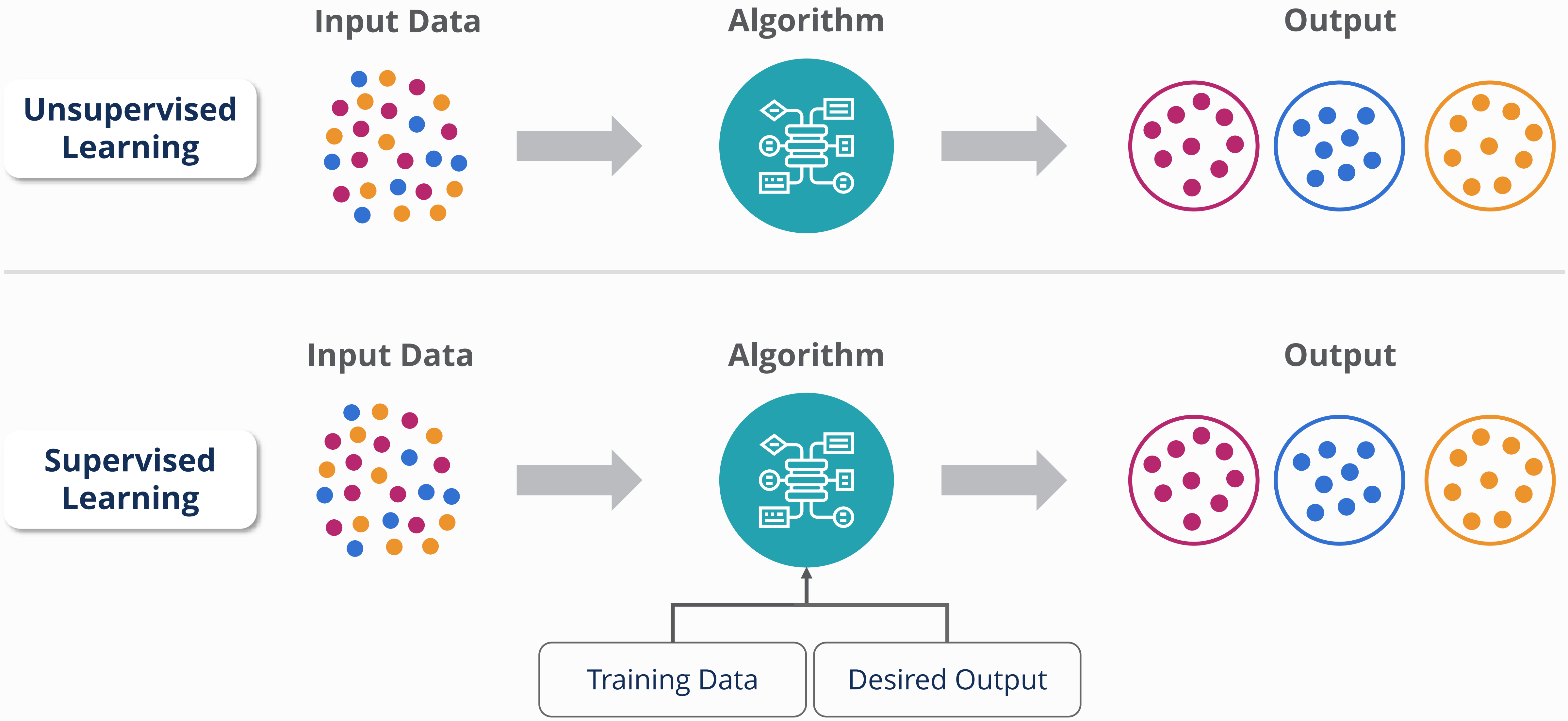
Traditional Programing

The programmer writes codes to define logic

Machine Learning

The computer creates logic from data

Machine Learning Overview



Machine Learning Overview

Machine Learning

Unsupervised Learning

- Group and interpret data based only on input data
- Often used when you have limited understanding of the data and want to explore similarities

Supervised Learning

- Develop predictive model based on both input and output data
- Often used when you have desired output to repeat in the future

Data Understanding

Data Understanding

Business Understanding → **Data Understanding** → Data Preparation → Algorithm Selection → Evaluation → Deployment

This step is to collect and review data.

Collect Data

- Internal sources
- Outside sources

Exploratory Data Analysis

Exploratory Data Analysis

Exploratory Data Analysis: a first glance on the data to see any trends or patterns.

Input Data/ Features

Income	Credit Score	Age
$56,000	755	43
$38,000	682	22
$120,000	731	38
$65,000	595	54
$52,00	784	68

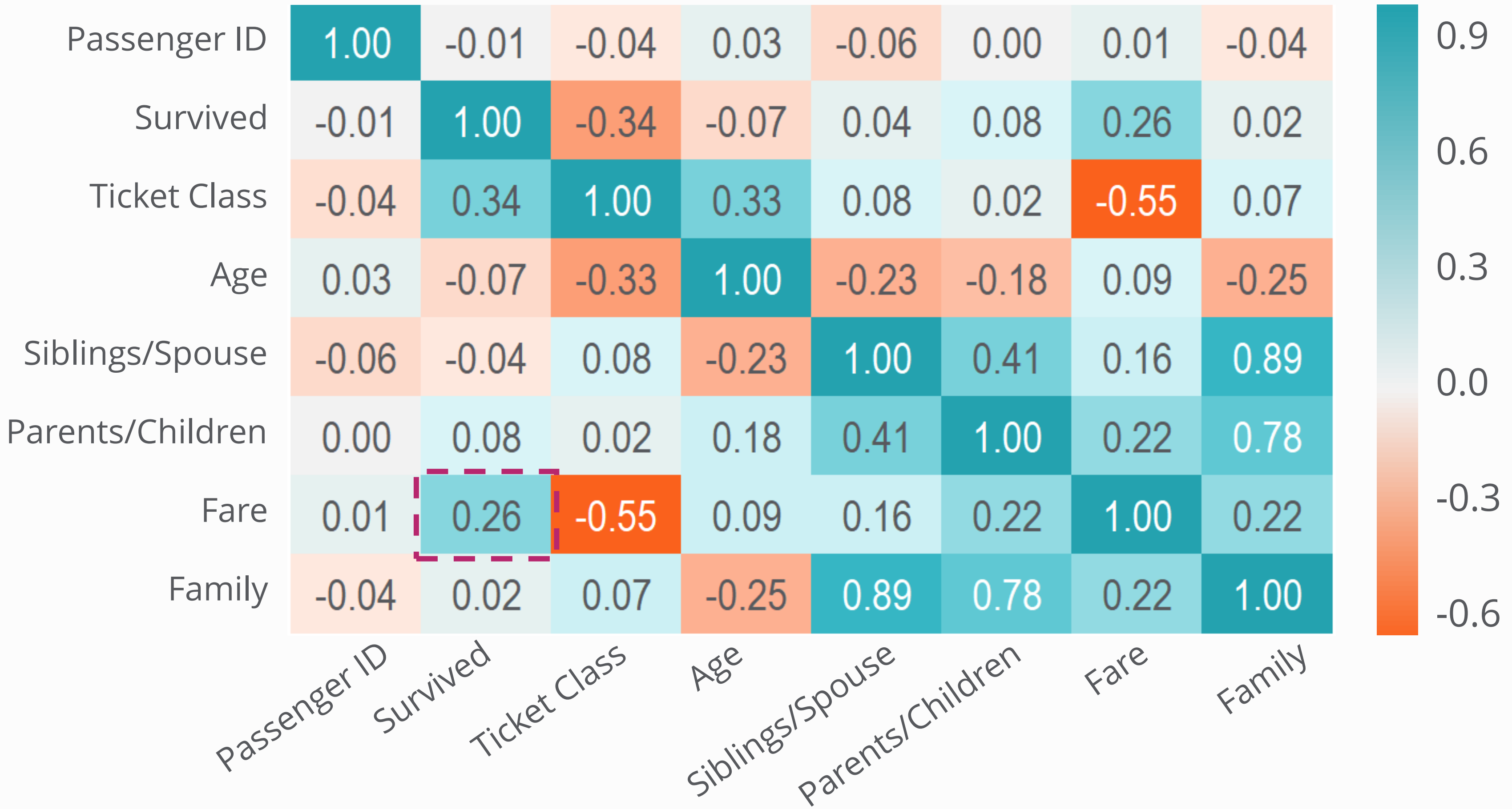
Output Data/ Target Variables

Default Payment
No
Yes
No
Yes
No

Example: Build a model to predict credit card default payment

Exploratory Data Analysis

Basic **descriptive statistics** are used here along with plotting of different variables within the dataset.



Case Demonstration

The purpose of this demonstration is to give **you an overview of the machine learning process** and **data science cycle**.

Case Objective: Predicting the house prices within New Taipei City in Taiwan

Data Source: Market historical dataset of real estate valuation downloaded from UC Irvine Machine Learning Repository

Data Preparation

Data Preparation

Business Understanding → Data Understanding → **Data Preparation** → Algorithm Selection → Evaluation → Deployment

This step is to set up the data and preparing it for machine learning modeling.

Feature Selection

Feature Engineering

Feature Selection

Feature selection: select the related features from the dataset and remove the irrelevant ones.

Irrelevant features can negatively impact the performance of a machine learning model.

- Reduce processing time
- Improve analysis results

Feature Selection

Common feature selection methods:

Principal Component Analysis (PCA)

- Reduce the features that have a high correlation with each other
- Keep only the principal components if there are too many starting features

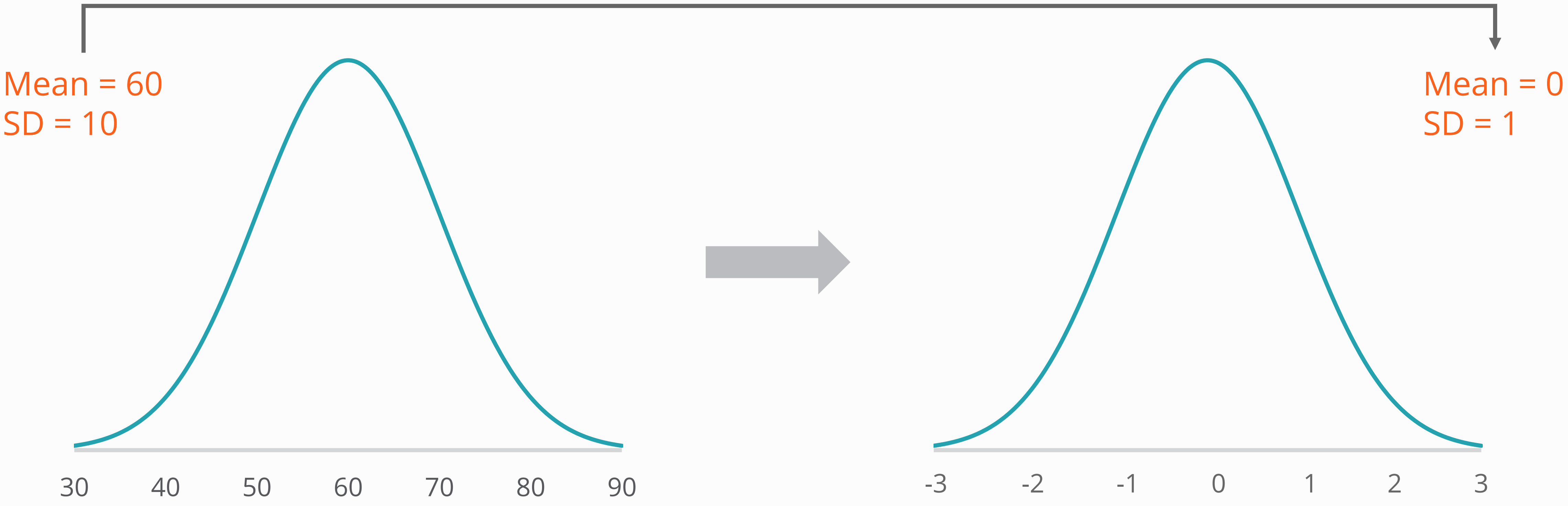
Feature Importance

- Leverage decision tree algorithms to determine which features are more important towards the output
- Remove irrelevant features

Feature Engineering

Feature engineering is the process to set up your data for better model performance.

- **Standardization** transforms the data to have a mean of 0 and a standard deviation of 1.



Standardization helps to rescale the distance of the data for prediction.

Feature Engineering

Feature engineering is the process to set up your data for better model performance.

- **Normalization (min-max scaling)** rescales the data to values between 0 and 1.

Income	Credit Score	Age
$56,000	755	43
$38,000	682	22
$120,000	731	38
$65,000	595	54
$52,00	784	68

Income	Credit Score	Age
0.2195	1.0000	0.4565
0.0000	0.5438	0.0000
1.0000	0.8500	0.3478
0.3293	0.0000	0.6957
0.1707	0.9563	1.0000

Feature Engineering

Feature engineering is the process to set up your data for better model performance.

- **One Hot Encoding** turns categorical data into number columns.

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Algorithm Selection

Algorithm Selection

Business Understanding → Data Understanding → Data Preparation → **Algorithm Selection** → Evaluation → Deployment

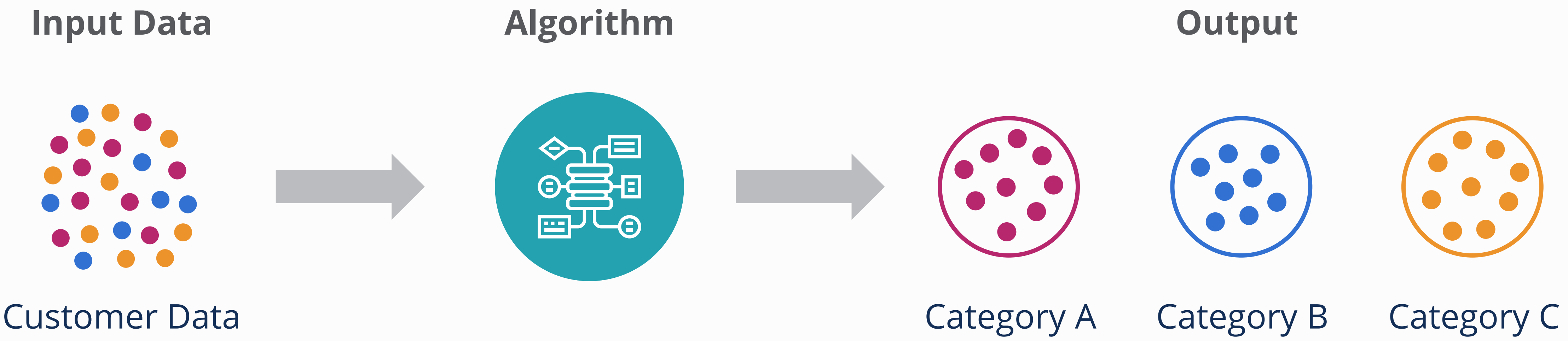
This step is to select machine learning algorithms that will contribute to the prediction of the results.

The algorithms are the key pieces that allow the machine to **learn from input data** and **improve from experience**.

Unsupervised Learning Algorithms

Supervised Learning Algorithms

Algorithms for Unsupervised Learning

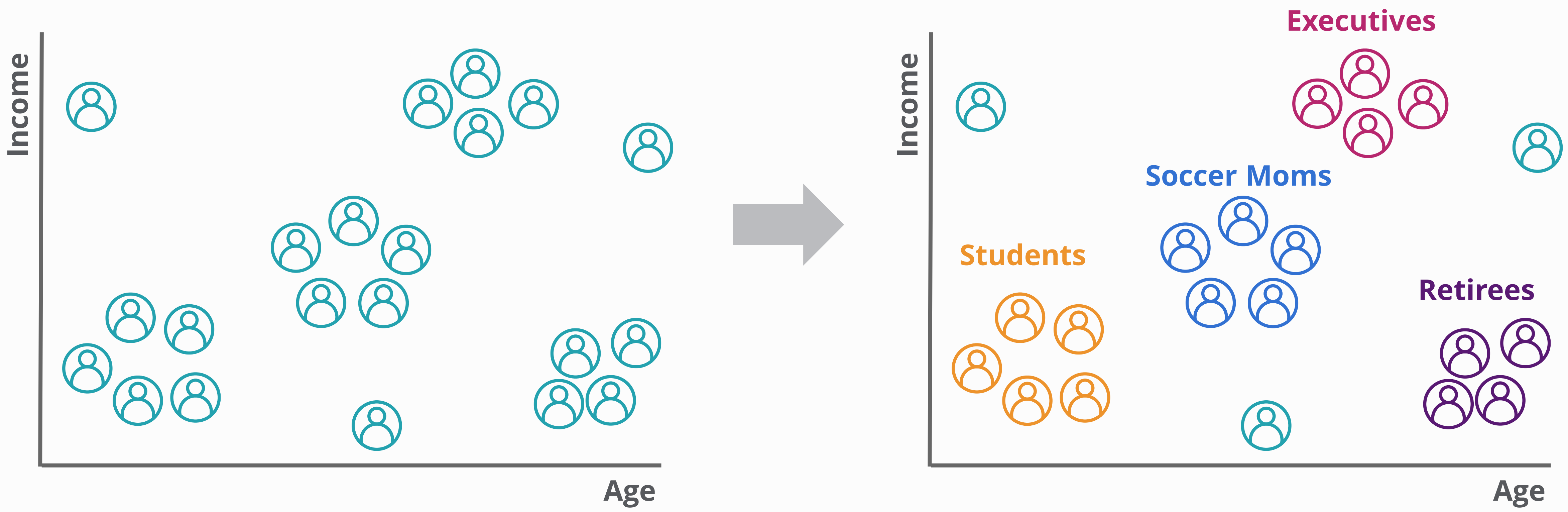


Algorithms used for unsupervised machine learning

- K-means clustering
- Hierarchical clustering

K-Means Clustering

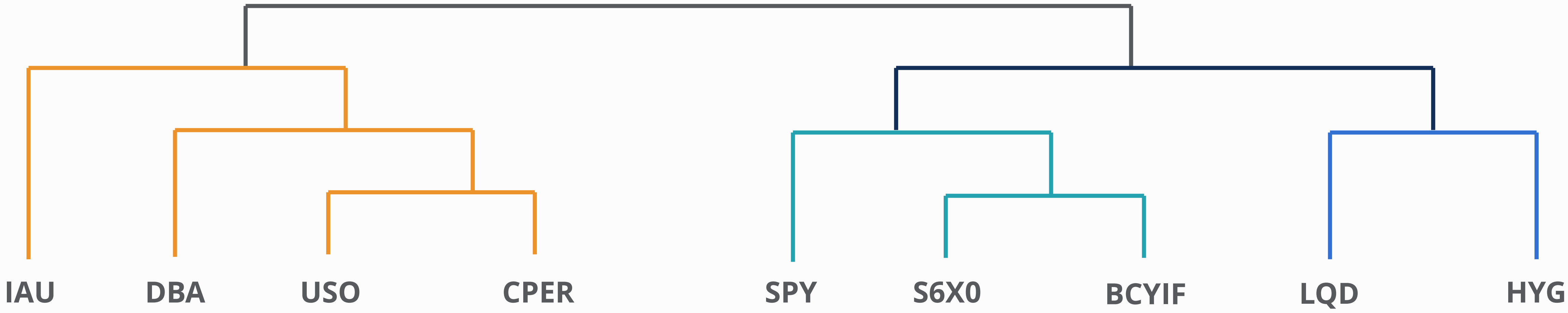
K-means clustering: a popular type of clustering algorithm to identify groups and trends.



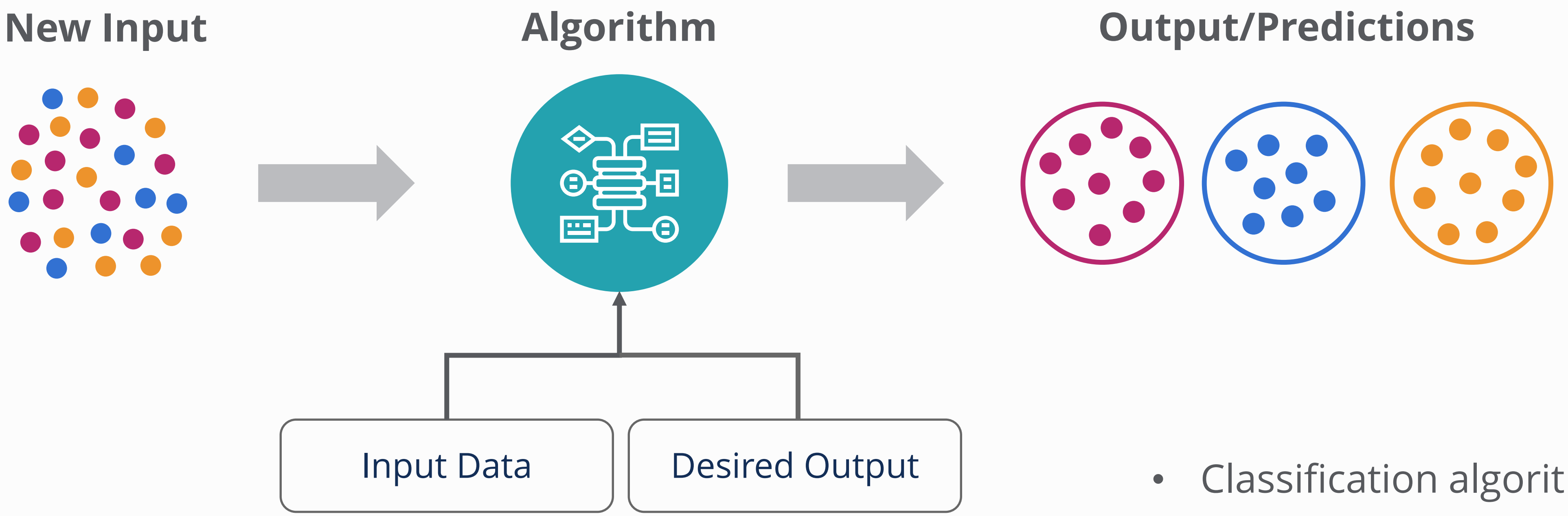
Hierarchical Clustering

Hierarchical clustering is another method of finding similarities within the dataset.

The splits occur based on where the model thinks the differences should be split, which is done mathematically by calculating the distances of each type.



Algorithms for Supervised Learning



The goal of the algorithm is to **map the relationship between the input and output**. This allows the model to produce predictions when given new inputs.

- Classification algorithms
- Regression algorithms
- Ensemble algorithms
- Validation/resampling technique

Regression and Classification Algorithms

Regression algorithms: predict an output given input in the form of a numeric value

Classification algorithms: predict the output of a given input in the form of categorical value.

Company Profile

↓

Regression

↓

Return on Investment

Security Profile

↓

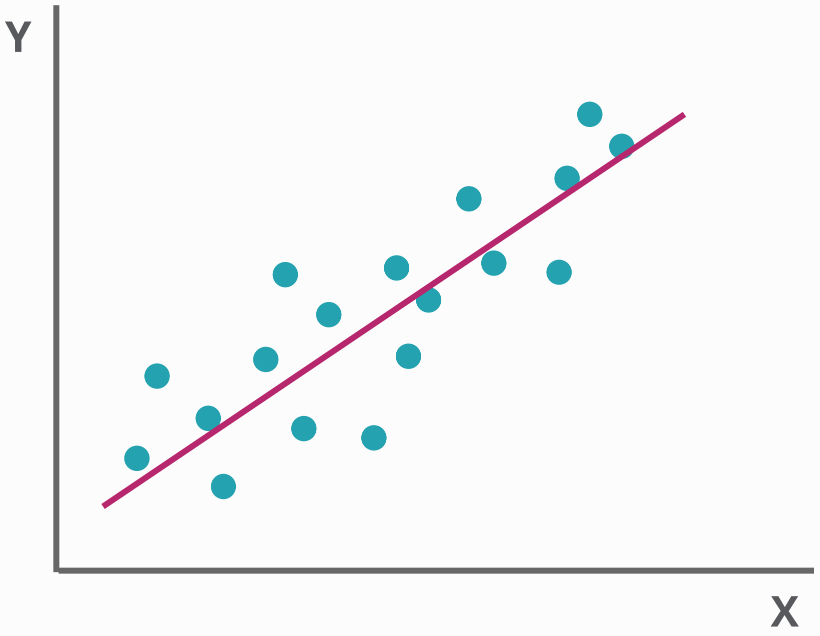
Classification

↓

Sell	Buy	Hold

Linear Regression

Linear regression is a type of regression model.



$$y = mx + b$$

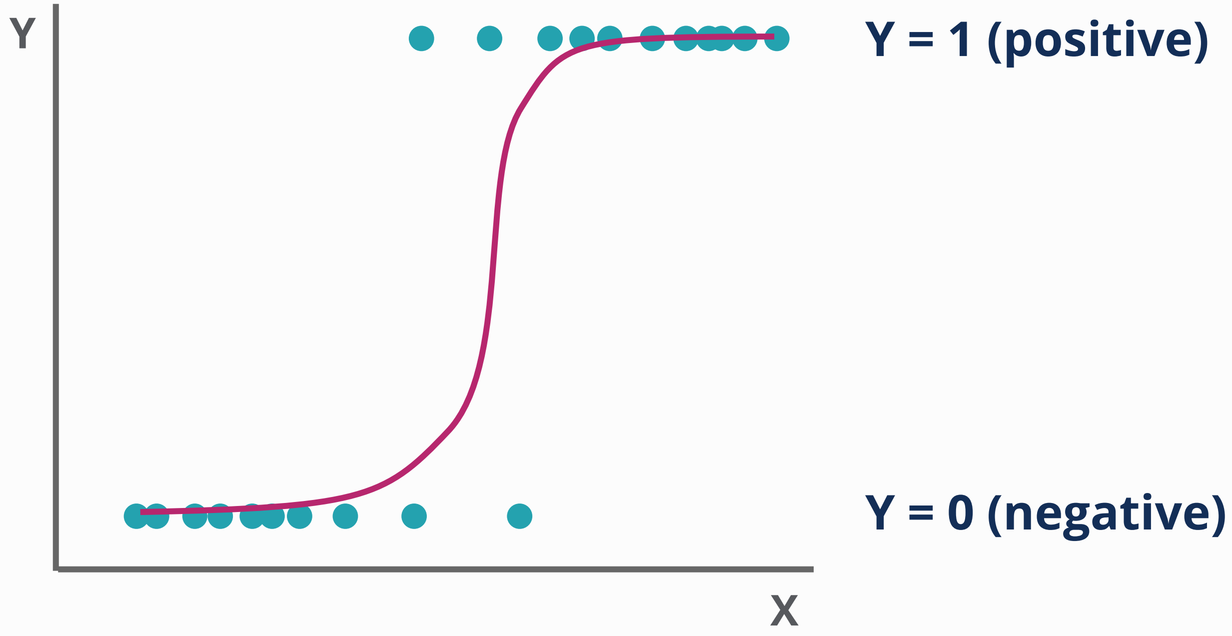
- **x**: Input
- **y**: Output
- **m**: Coefficient value
- **b**: Intercept

The linear regression algorithm helps us figure out the values of m and b so we can make predictions.

Logistic Regression

Logistic regression is a type of binary classification algorithm.

The output variables are sorted into two categories.

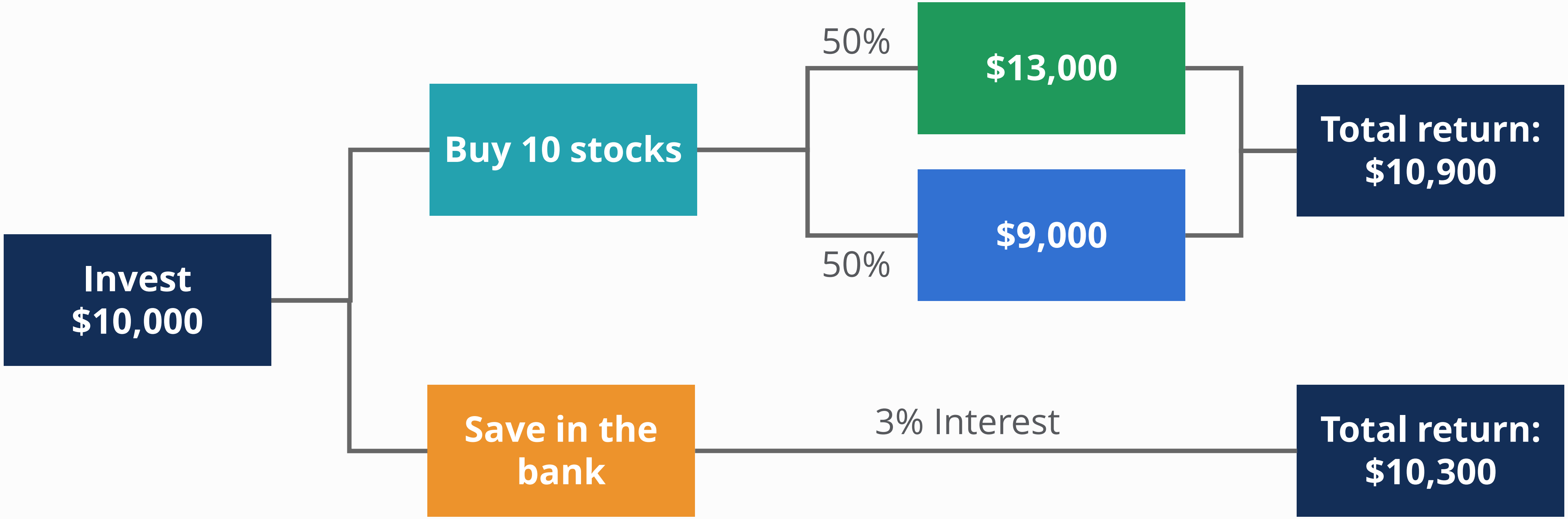


The logistic regression algorithm calculates the probability of output data being positive or negative.

Decision Tree

The decision tree algorithm can be used to predict both categorical or numeric outcomes.

The decision tree algorithm splits the dataset into hierarchical branches until it reaches the results to answer the question.



Other Common Algorithms

Algorithm	Type
Linear Regression	Regression Models
Ridge Regression	Regression Models
Lasso Regression	Regression Models
Logistic Regression	Classification Models
Linear Discriminant Analysis	Classification Models
Naive Bayes	Classification Models
Decision Tree	Regression & Classification
K-Nearest Neighbors (kNN)	Regression & Classification
Support Vector Machines (SVM)	Regression & Classification

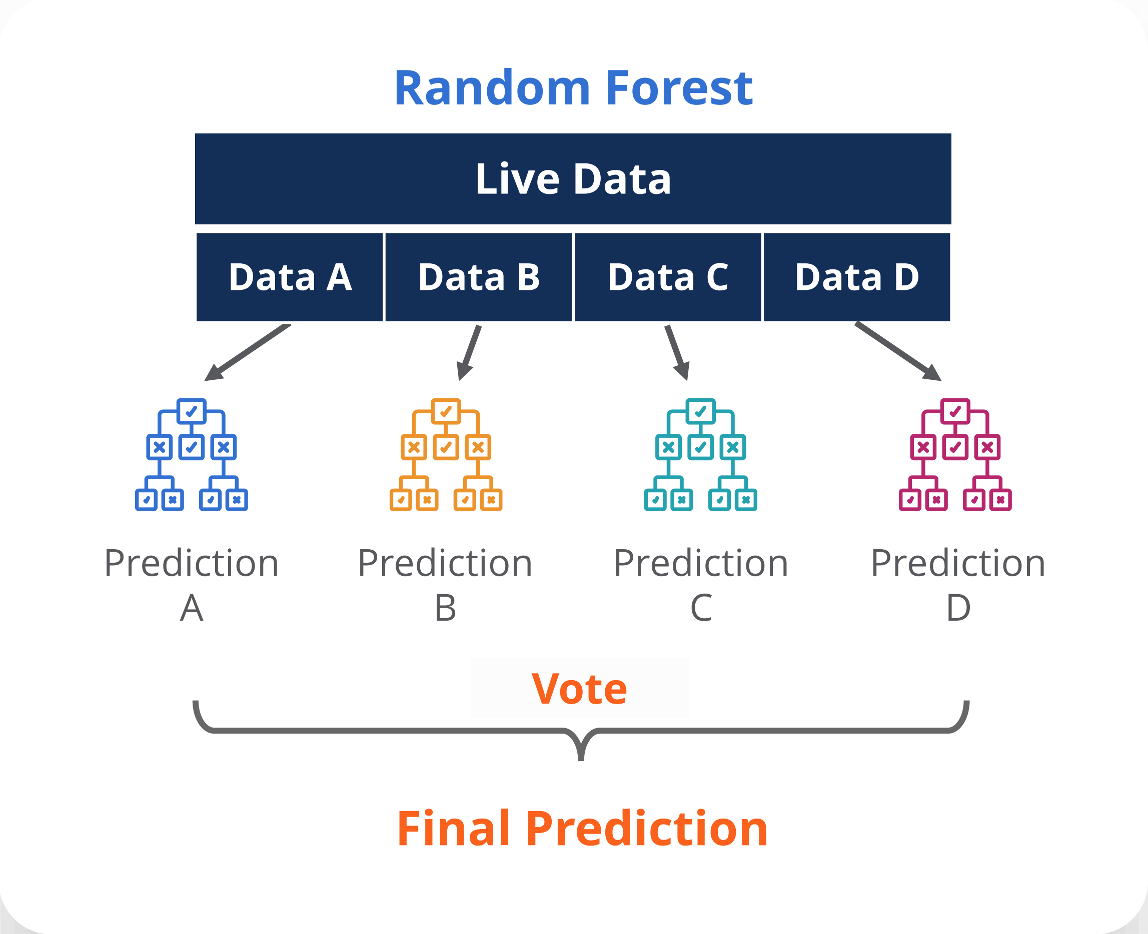
Ensemble Models

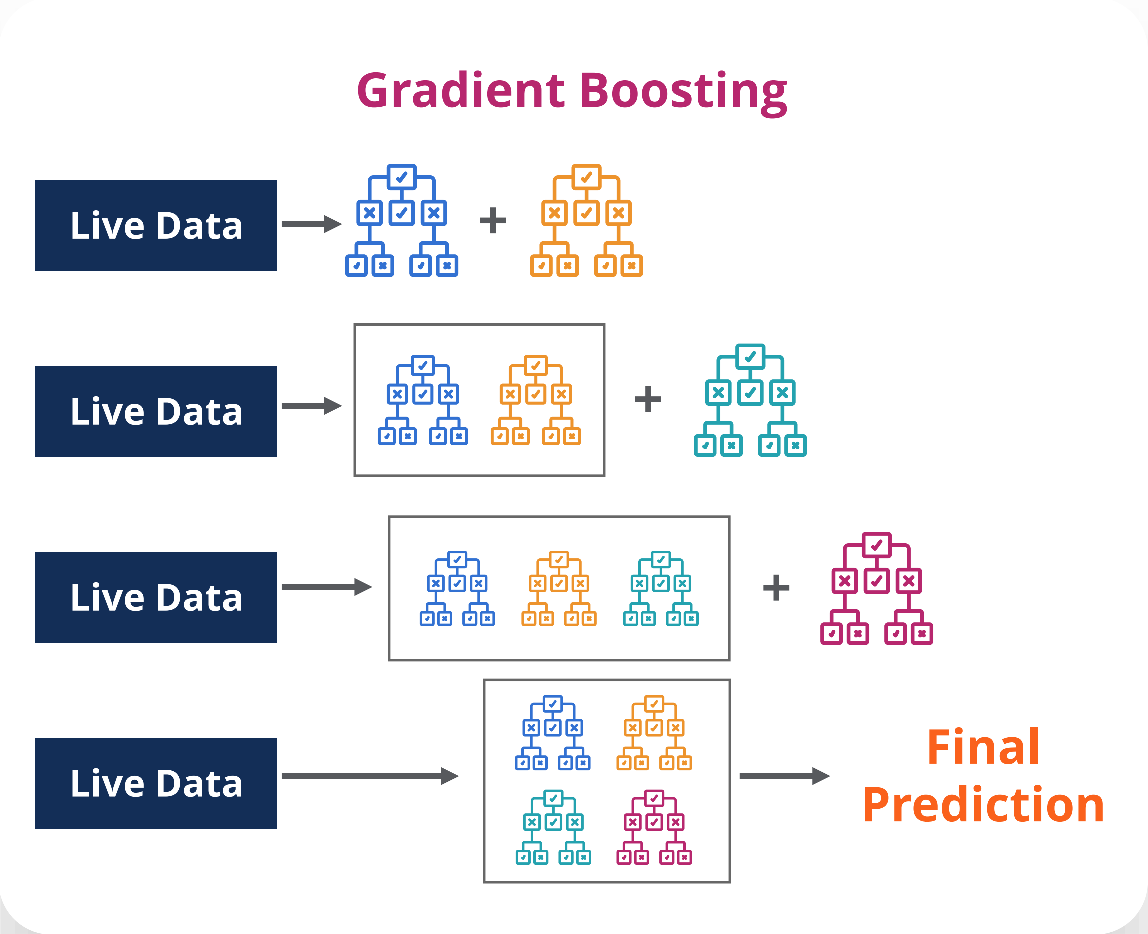
Ensemble models are a creation of different algorithms modeled into one.

Ensemble models can be used for **both classification or regression**.

Empirically, ensemble models tend to add **~5% improved performance** over stand-alone machine learning models.

Ensemble Models





Ensemble models can be any combination of the machine learning algorithms.

How to Choose an Algorithm?

Regression Algorithms

Classification Algorithms

Resemble Models

Find similar examples

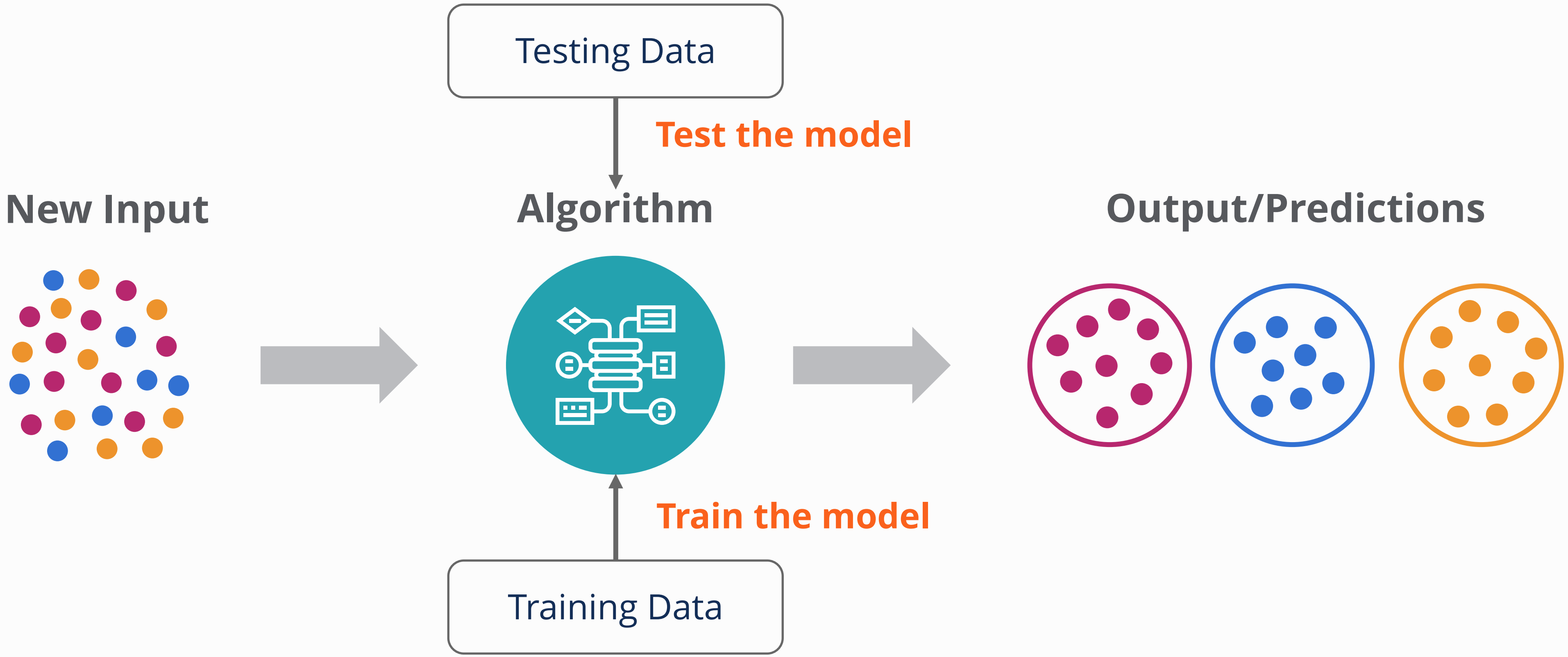
Try with your own data

Go through data science cycle

Validation/Resampling Techniques

Validation or resampling techniques are commonly used in supervised learning.

The goal of validation is to get a better estimate of how the model would perform with data that it has not seen before.



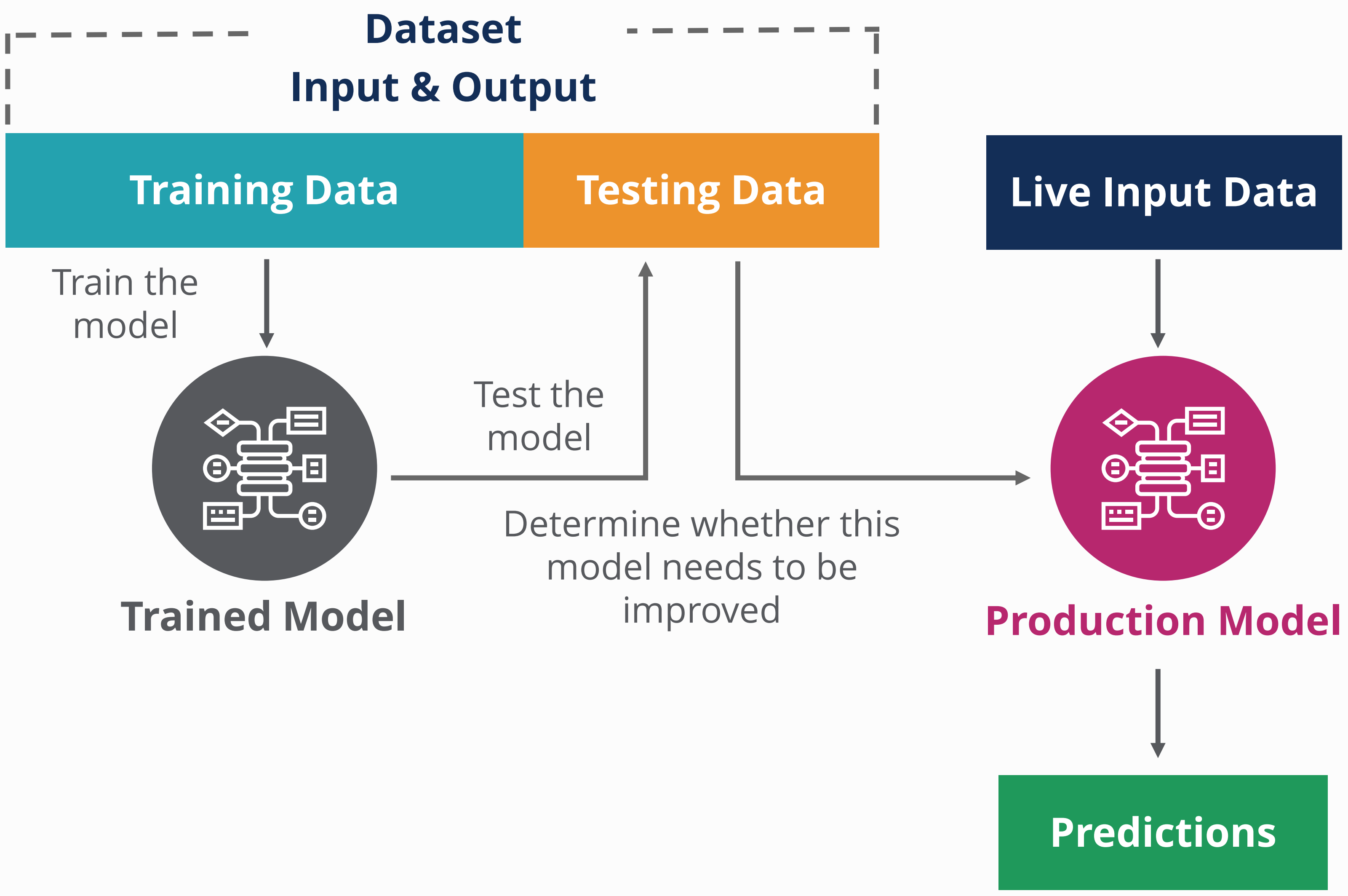
Validation/Resampling Techniques

We want the model to perform well on the training data as well as the testing data.

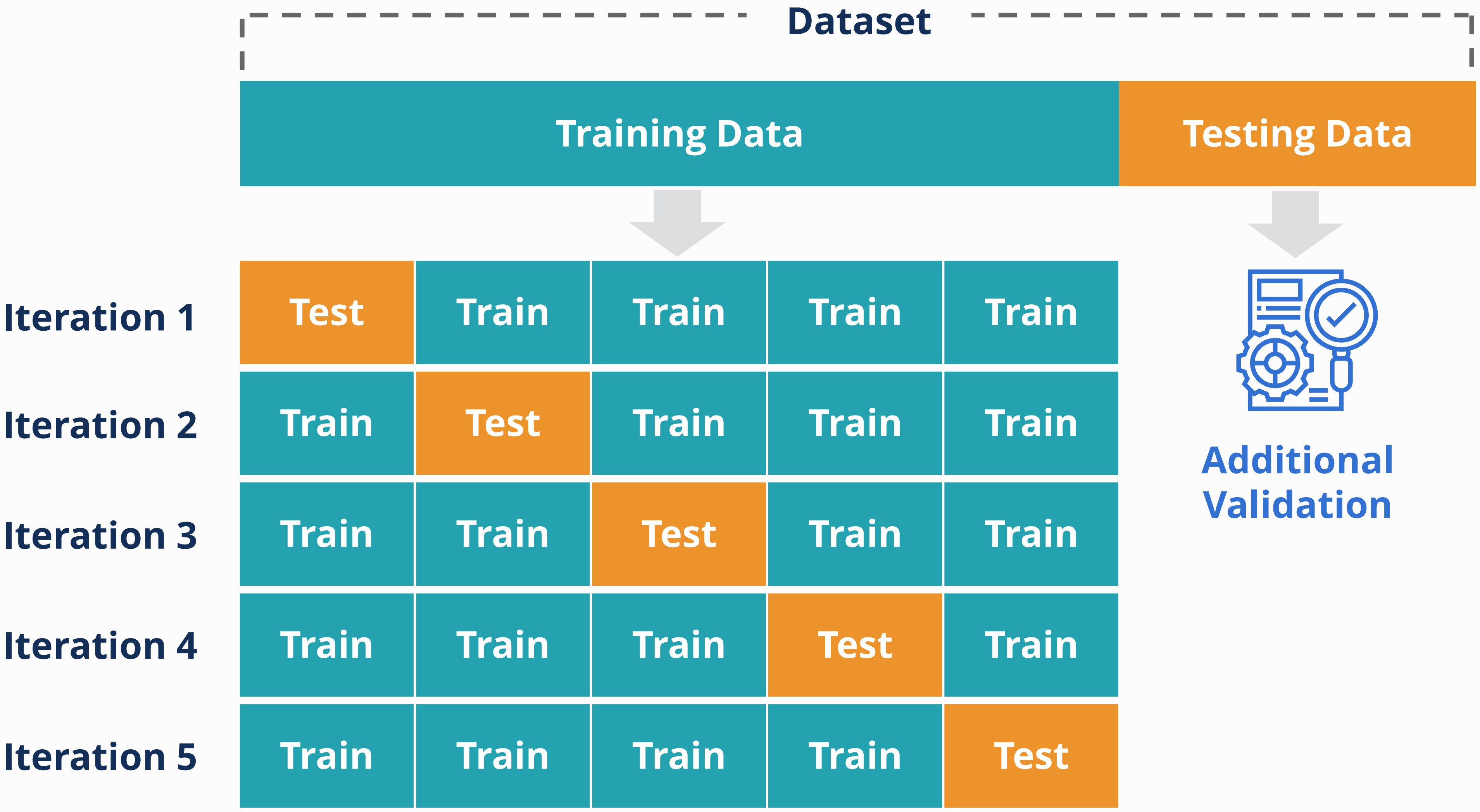
This validation process is **unique to supervised learning**.

- Train and test split
- K-fold cross validation
- Train, Validation, and Test Split

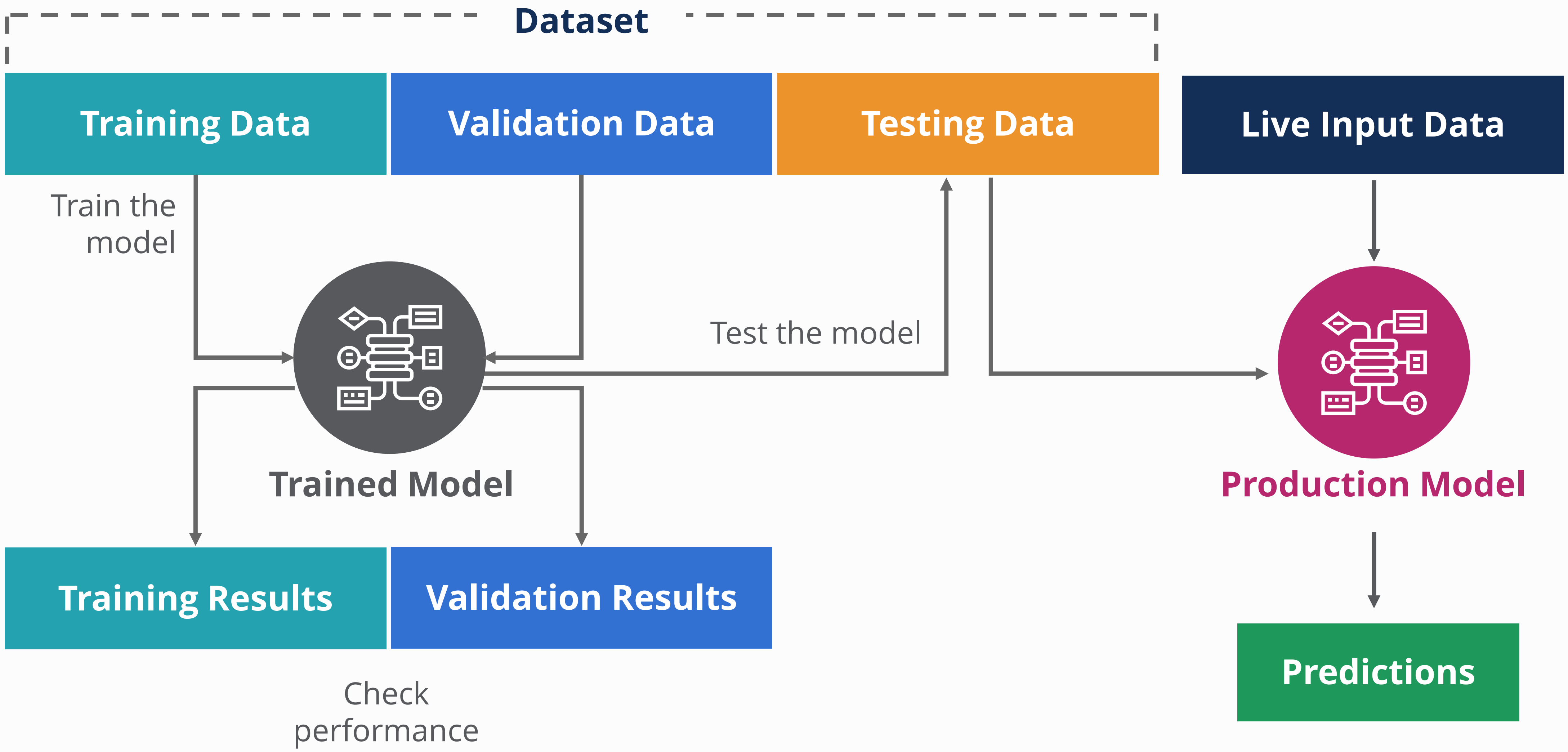
Train and Test Split



K-Fold Cross Validation



Train, Validation, and Test Split



Evaluation

Evaluation

Business Understanding → Data Understanding → Data Preparation → Algorithm Selection → **Evaluation** → Deployment

Model evaluation is the step where we compare the model's performance on the training data with its performance on the testing data.

Regression Metrics

- R^2
- MAE
- MSE
- RMSE

Classification Metrics

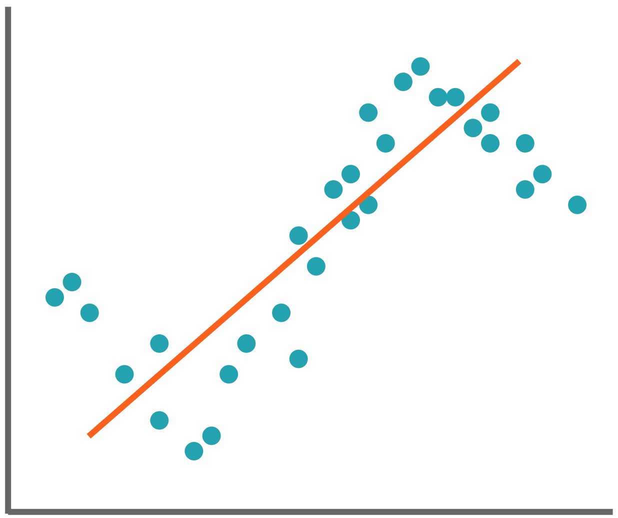
- Accuracy
- AUC

Underfitting vs. Overfitting

The purpose of machine learning is to find the **real relationship** between inputs and outputs.

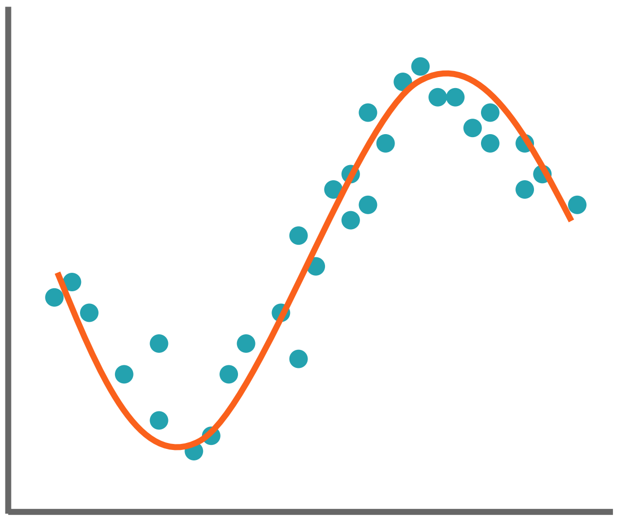
Underfit

Over generalizes the data



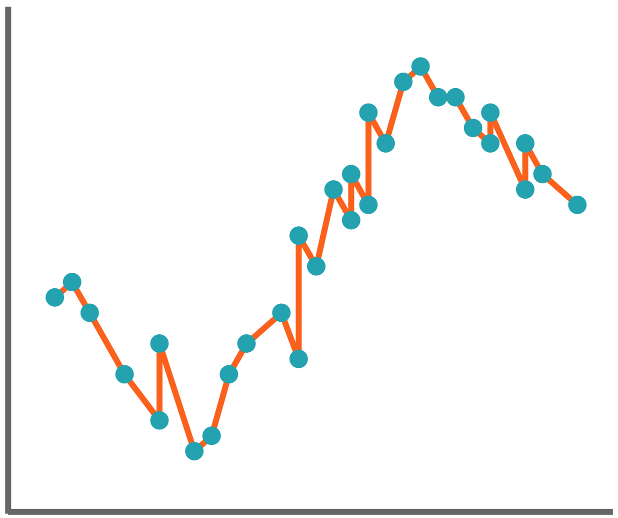
Good Fit

Generalize enough to predict future patterns



Overfit

Does not generalize enough

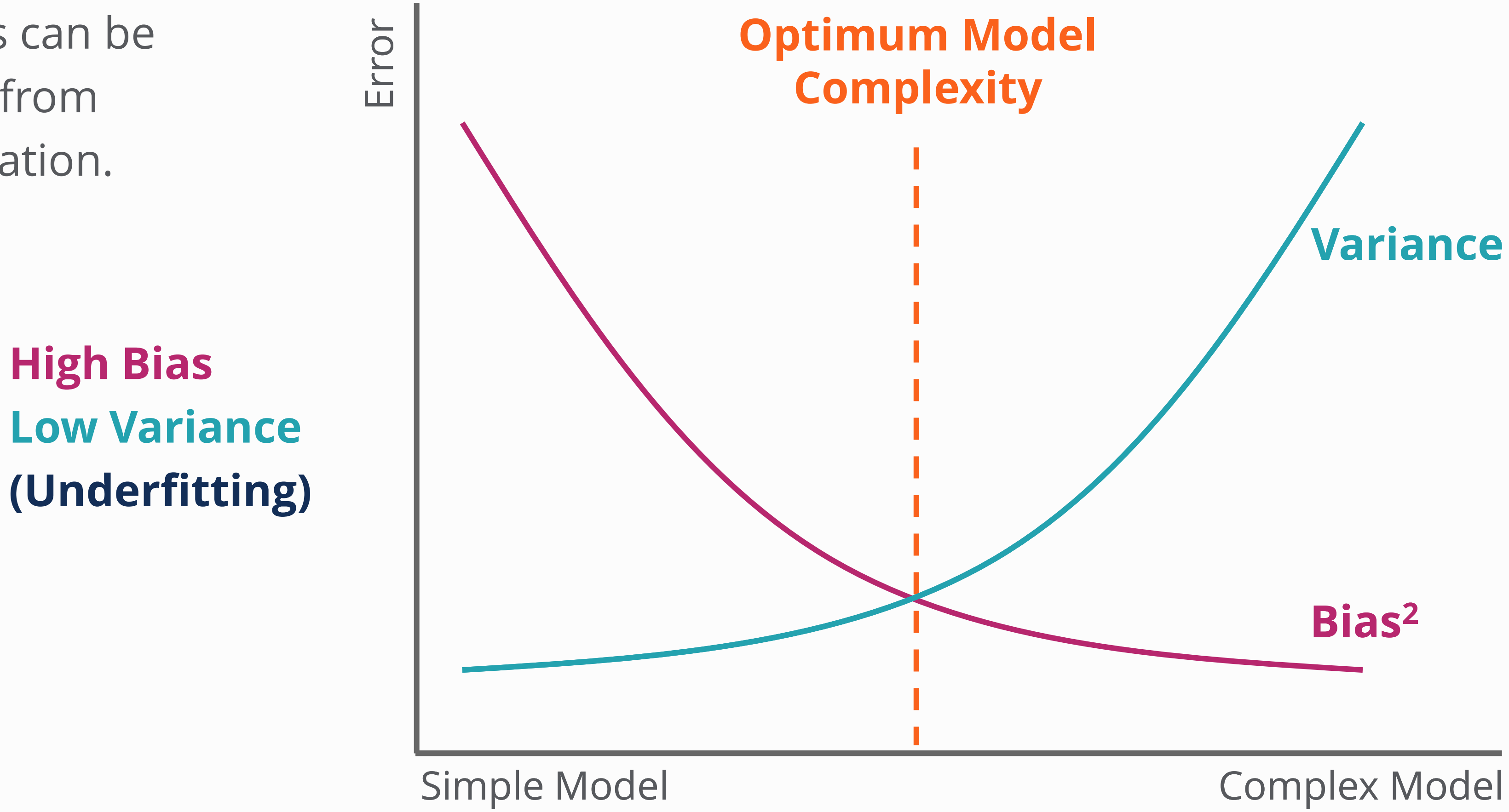


Bias vs. Variance

Bias is the difference between the prediction and the actual value.

High bias can be reduced from regularization.

Bias vs. Variance Trade-off



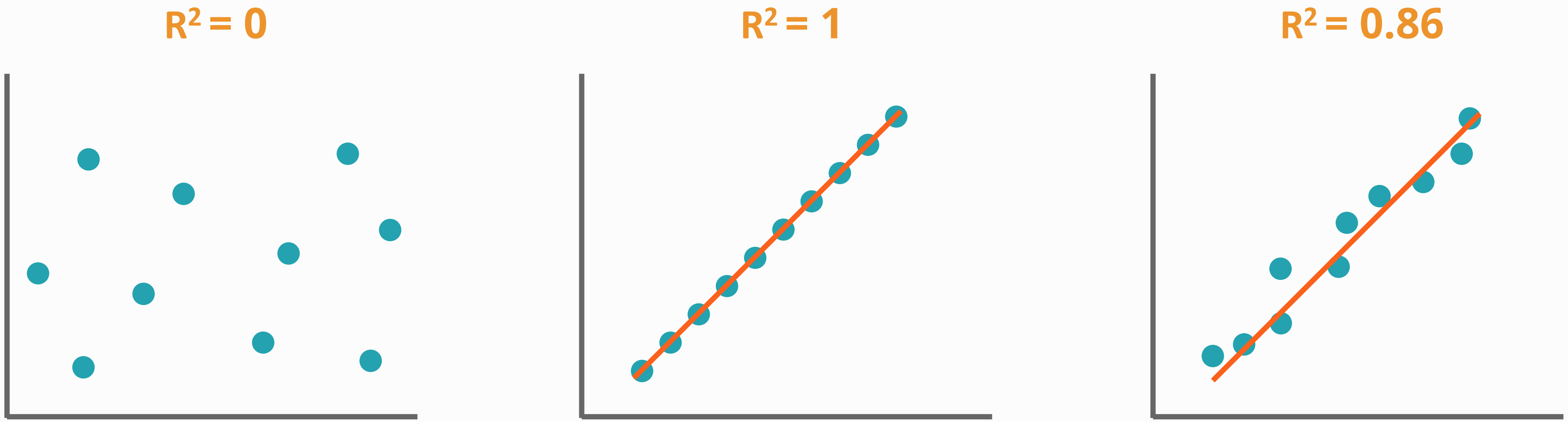
Variance is the difference in fits in between datasets.

High variance can be resolved by reducing complexity.

Regression Metrics – R^2

Coefficient of Determination (R^2) is one of the most used metrics to evaluate regression models.

R^2 measures how close the data are to the fitted regression line.

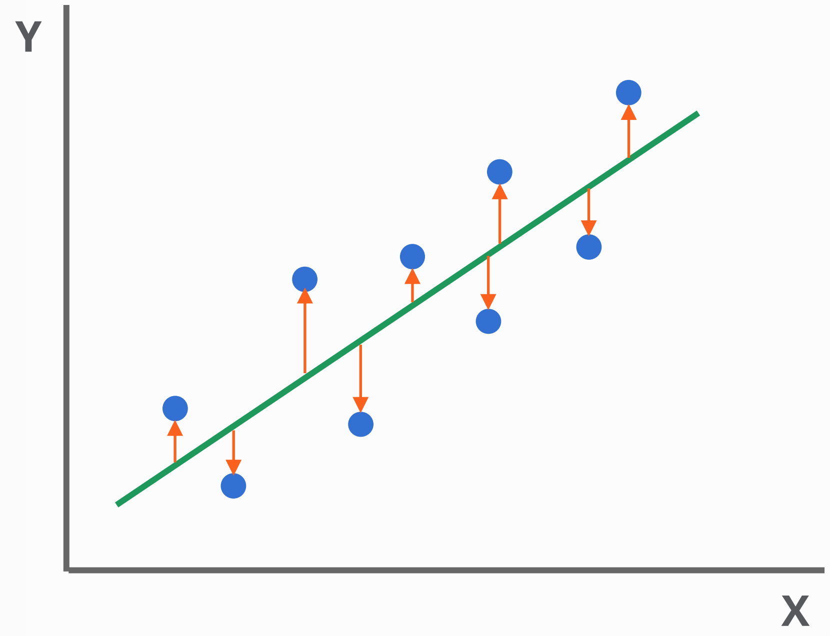


Higher R^2 indicates better fit of the model.

Regression Metrics – MAE, MSE and RMSE

MAE, MSE and RMSE measure how far the predicted values deviate from the actual values.

Metrics	Range	Sensitive to Outlier
Mean Absolute Error (MAE)	0 - ∞	No
Mean Squared Error (MSE)	0 - ∞	Yes
Root Mean Square Error (RMSE)	0 - ∞	Yes



Lower MAE, MSE and RMSE indicate better model performance.

Classification Metrics – Accuracy

Accuracy

Area Under Curve (AUC)

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

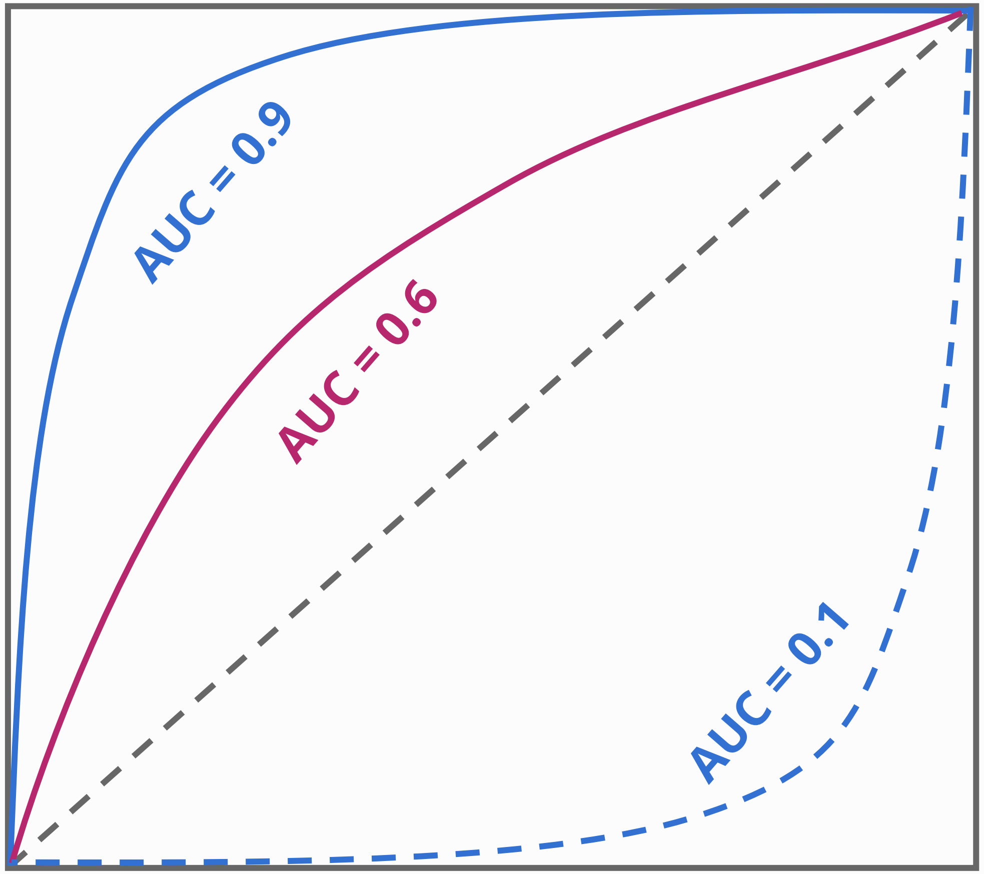
Classification Metrics – AUC

Area Under Curve (AUC) evaluate both true positives and true negatives.

		Prediction	
		Negative (0)	Positive (1)
Actual	Negative (0)	True Negative	False Positive
	Positive (1)	False Negative	True Positive

High AUC means that the model is correctly classifying the output results.

AUC value ranges from 0.5 to 1.



Evaluation Metrics Summary

Look at the evaluation metrics on both training data and testing data.

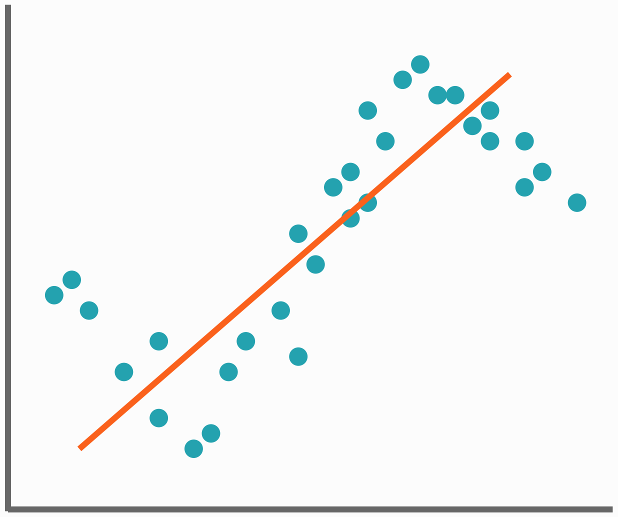
On Training Data	On Testing Data	How Well the Model Fits
High performance	High performance	Good fit
Low performance	Low performance	Underfit
High performance	Low performance	Overfit

Example:

- R^2 is high on the training data and high on the testing data – Good fit
- R^2 is low on the training data and low on the testing data – Underfit
- R^2 is high on the training data but low on the testing data – Overfit

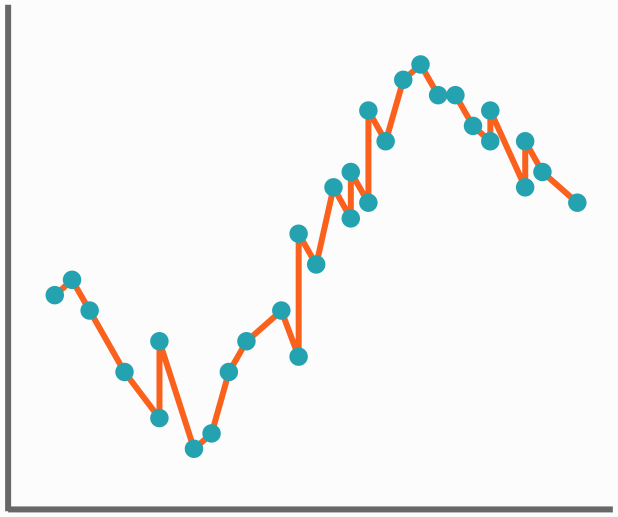
How to Improve the Model

Underfit



- Select more features
- Select a more complex algorithm
- Improve feature engineering
- Add more data

Overfit



- Select fewer features
- Select a simpler algorithm
- Improve feature engineering
- Add more data