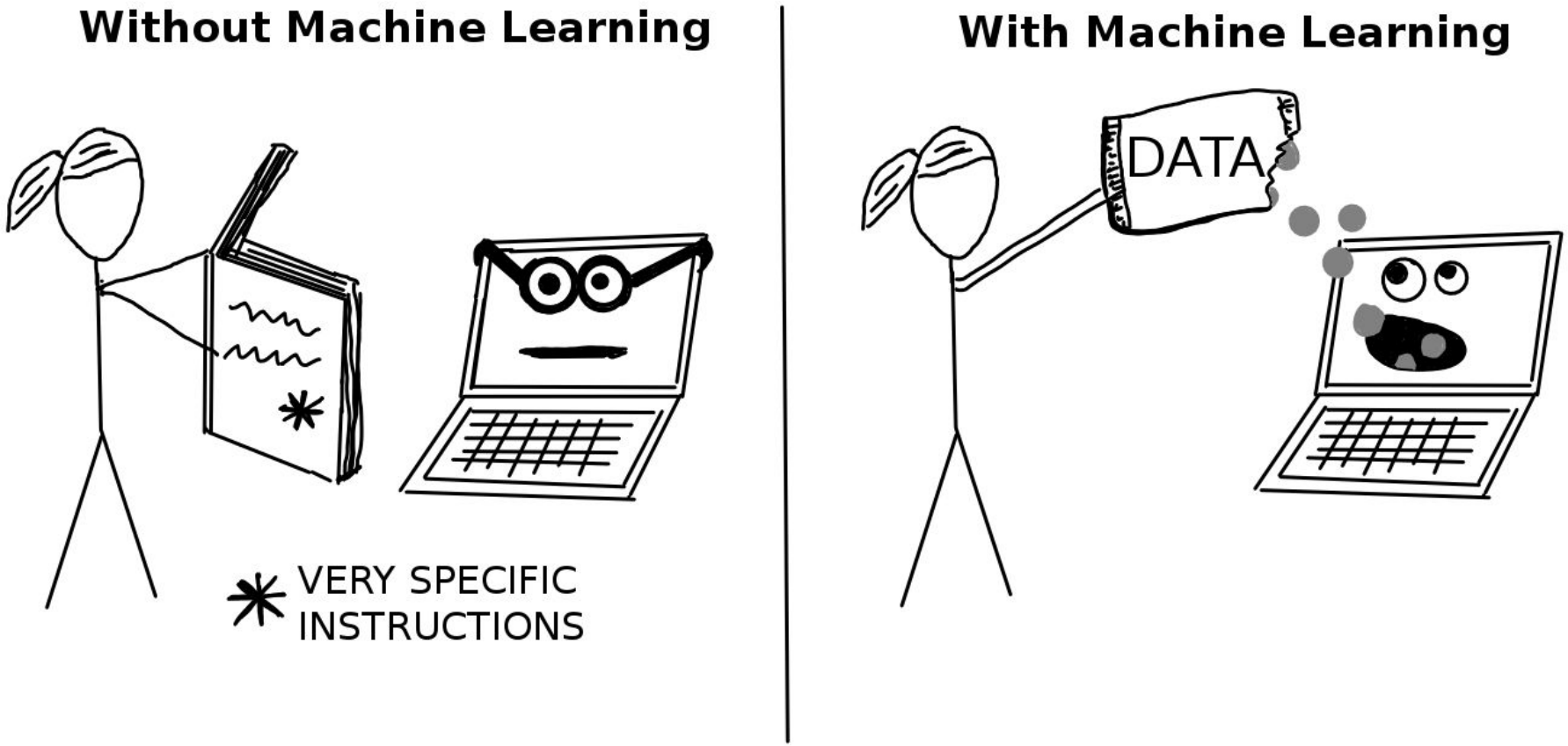
Introduction to Machine Learning

Introduction

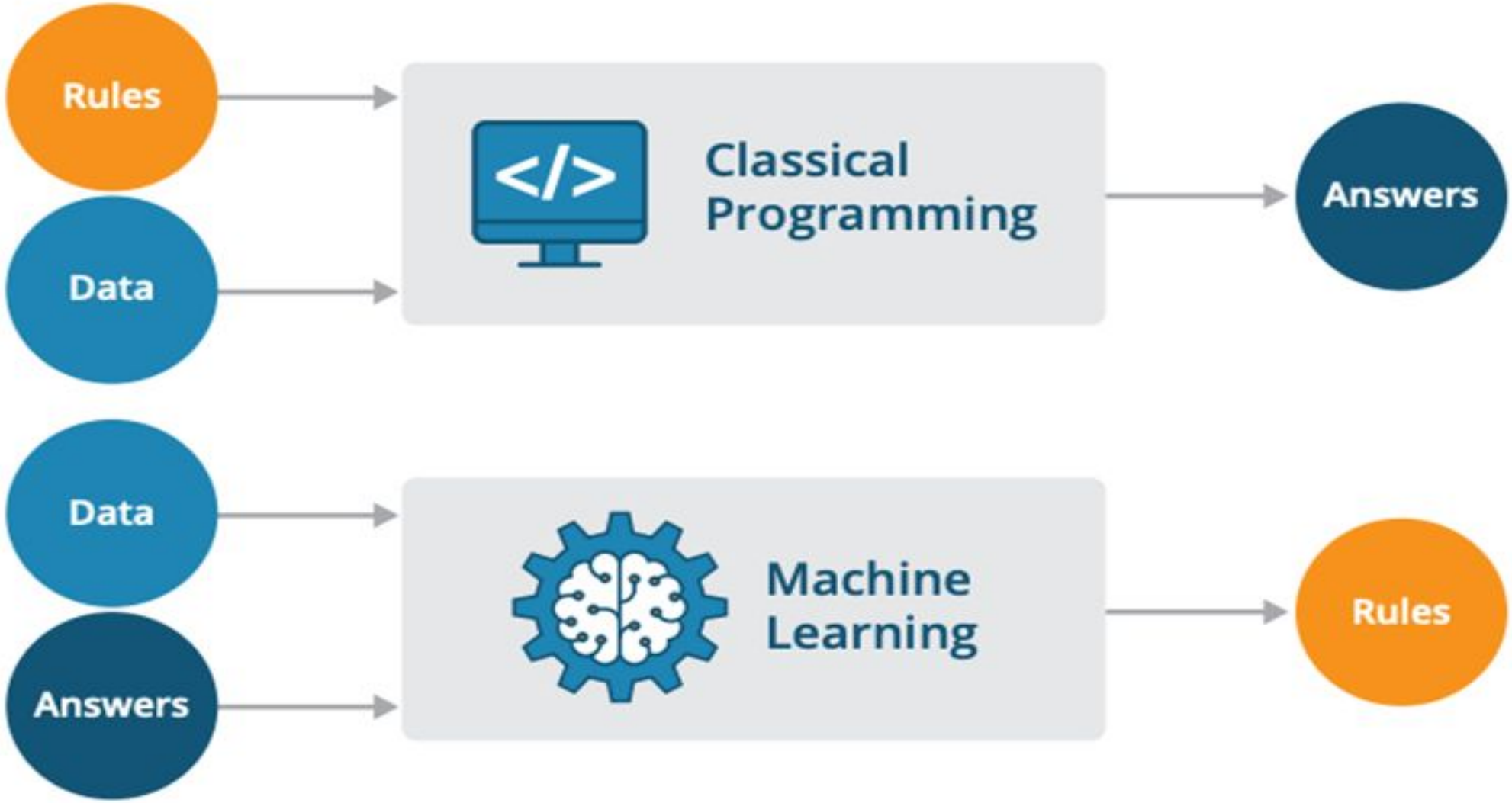
- To solve a problem on a computer, we need an algorithm.
- An algorithm is a sequence of instructions that should be carried out to transform the input to output.
- For example, one can devise an algorithm for sorting.

- For some tasks, however, we do not have an algorithm—for example, to tell spam emails from legitimate emails.
- What we lack in knowledge, we make up for in data.
- We can easily compile thousands of example messages some of which we know to be spam and what we want is to "learn" what constitutes spam from them.

Machine Learning



Classical Approach Vs Machine Learning



Definition

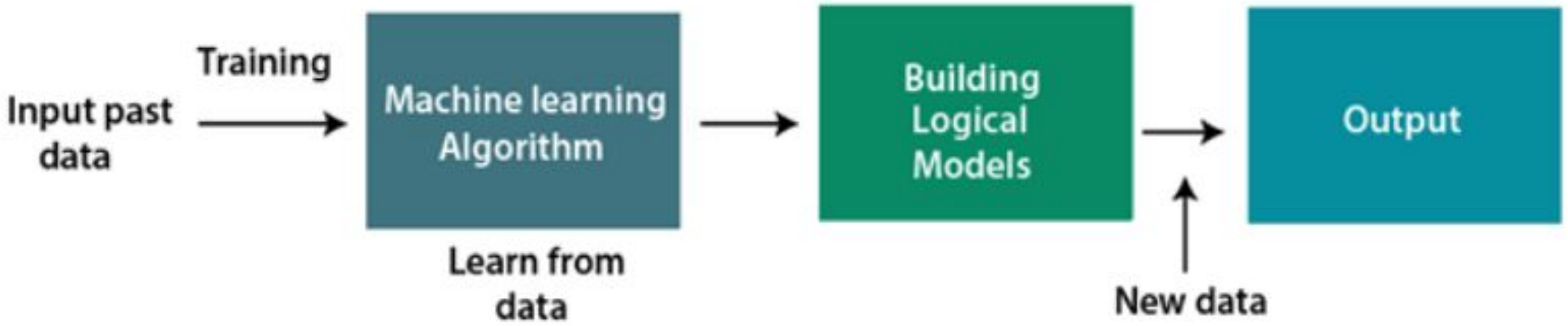
- Machine learning is programming computers to optimize a performance criterion using example data or past experience.

- We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience.

- Machine learning is turning data into information.

How does Machine Learning work ????

- A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it.
- The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.



Why is ML required?

- Rapid increment in the production of data
- Solving complex problems, which are difficult for a human
- Decision making in various sector including finance
- Finding hidden patterns and extracting useful information from data

Key Terminologies

Input Attributes

Target Attribute

Number of new Recipients	Email Length (K)	Country (IP)	Customer Type	Email Type
0	2	Germany	Gold	Ham
1	4	Germany	Silver	Ham
5	2	Nigeria	Bronze	Spam
2	4	Russia	Bronze	Spam
3	4	Germany	Bronze	Ham
0	1	USA	Silver	Ham
4	2	USA	Silver	Spam

Instances

Numeric

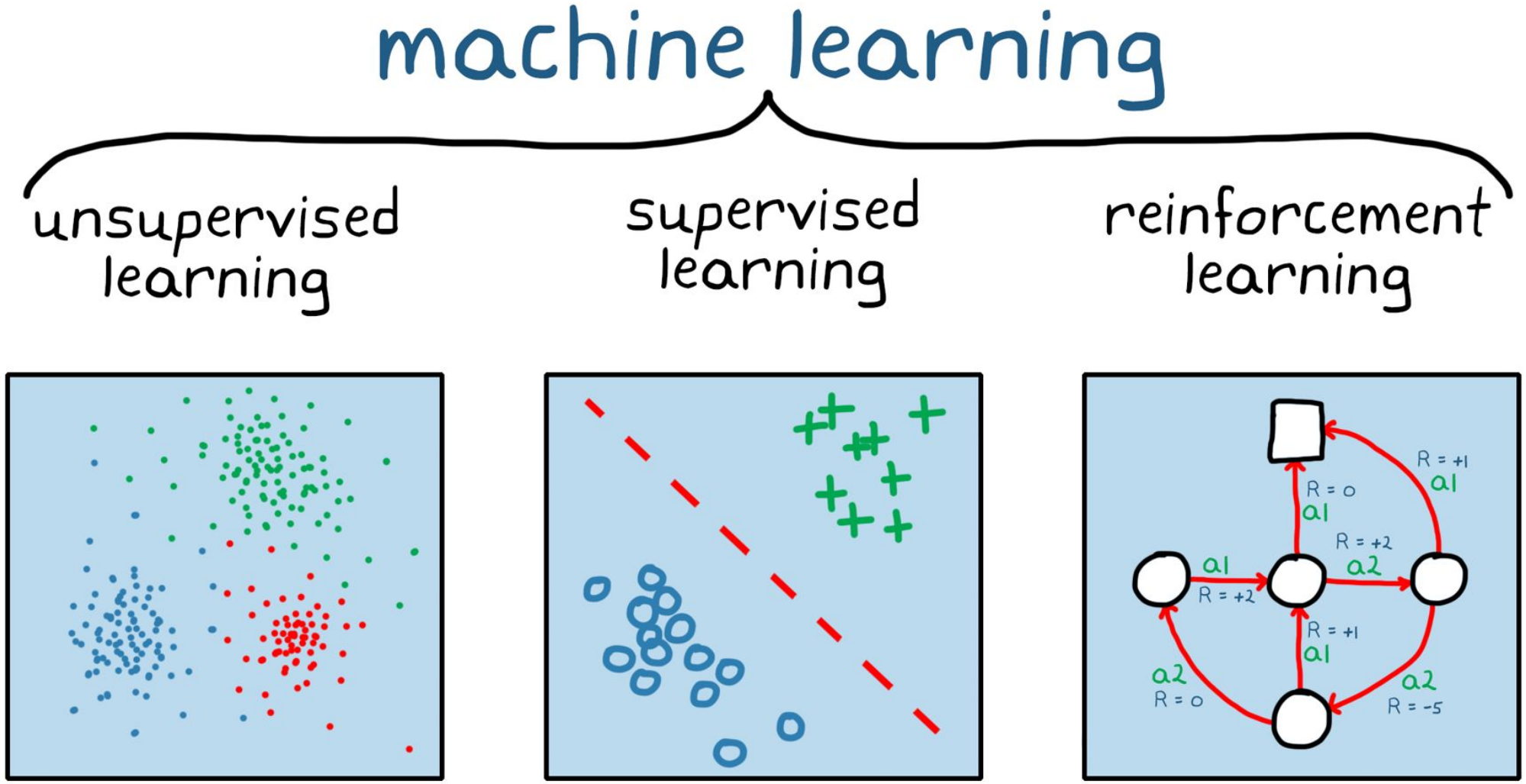
Nominal

Ordinal

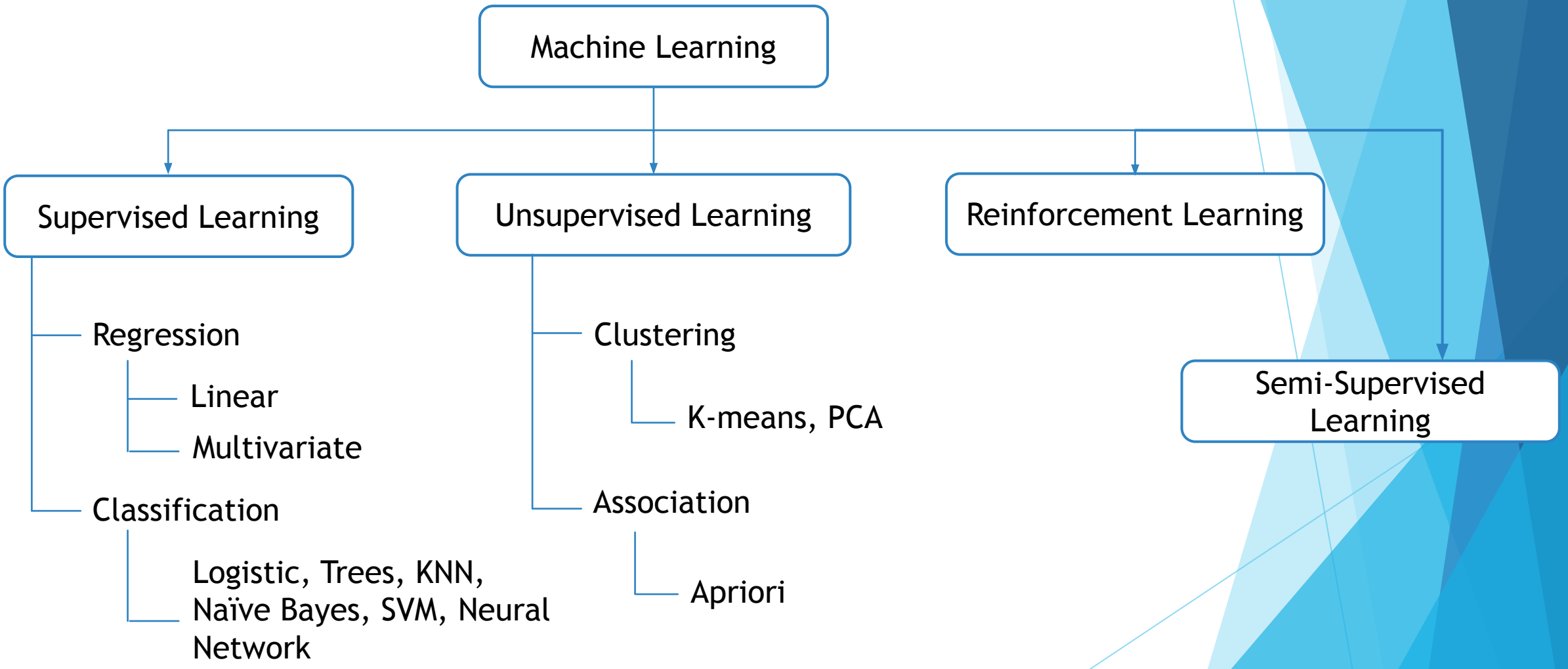
Key Task of Machine Learning

- Classification
  - In classification, our job is to predict what class an instance of data should fall into.
- Regression
  - Regression is the prediction of a numeric value.
- Clustering
  - In unsupervised learning, there's no label or target value given for the data.
  - A task where we group similar items together is known as *clustering*.

Types of Machine Learning

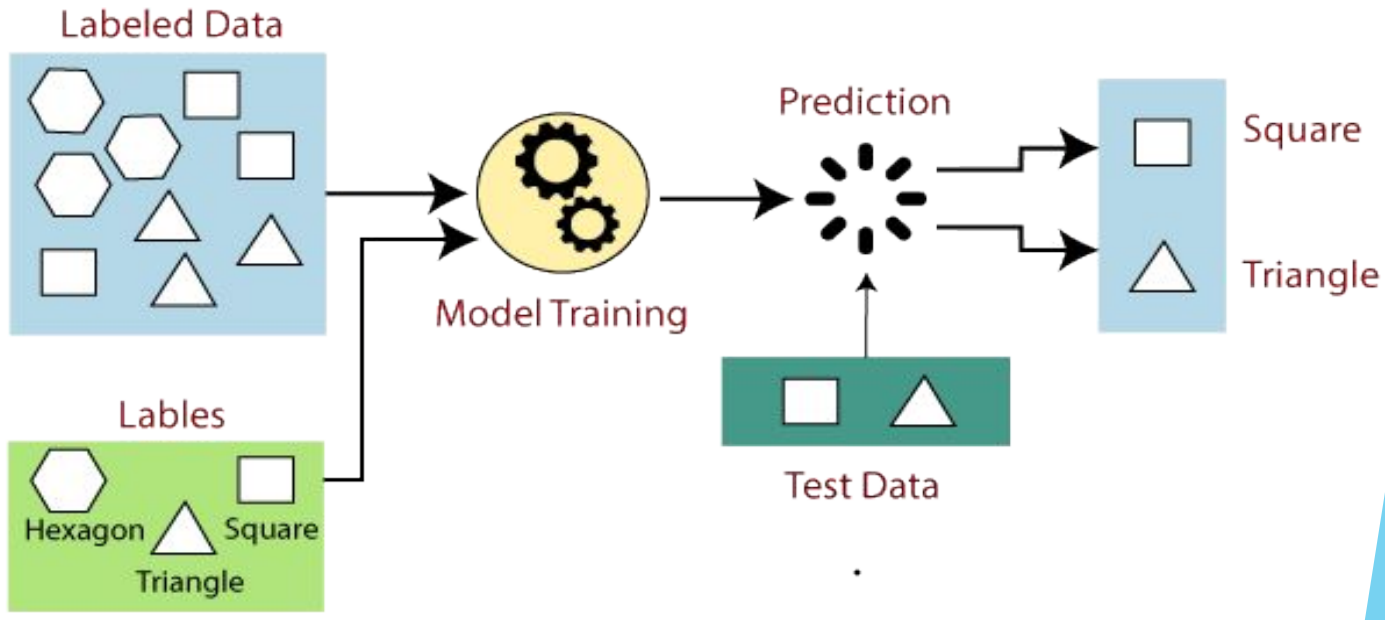


Machine Learning Classification



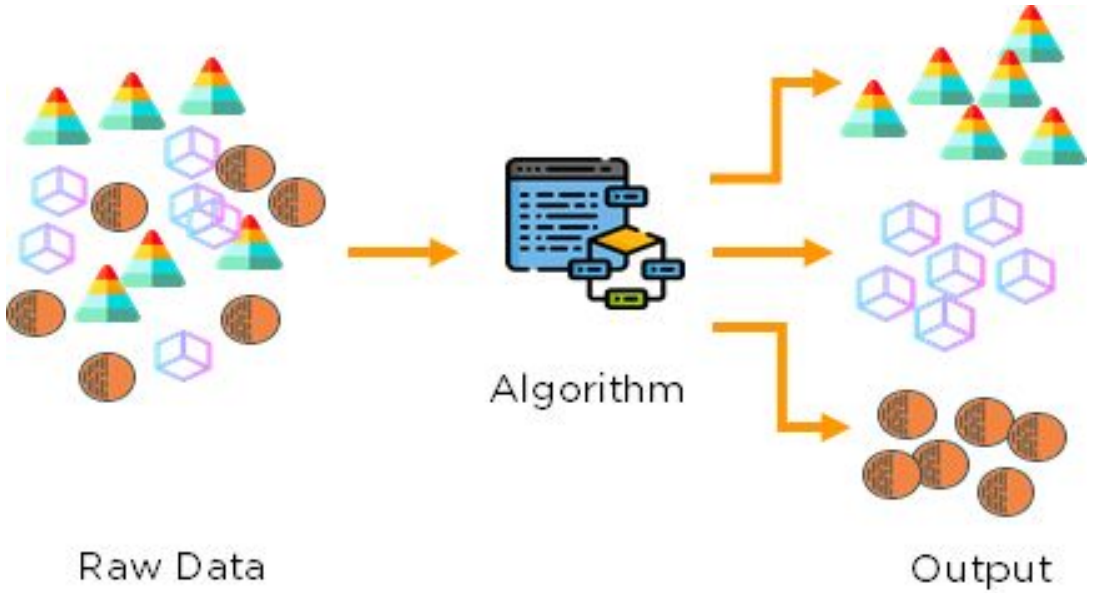
Supervised Learning

- In supervised learning, models are trained using labelled dataset, where the model learns about each type of data.
- Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

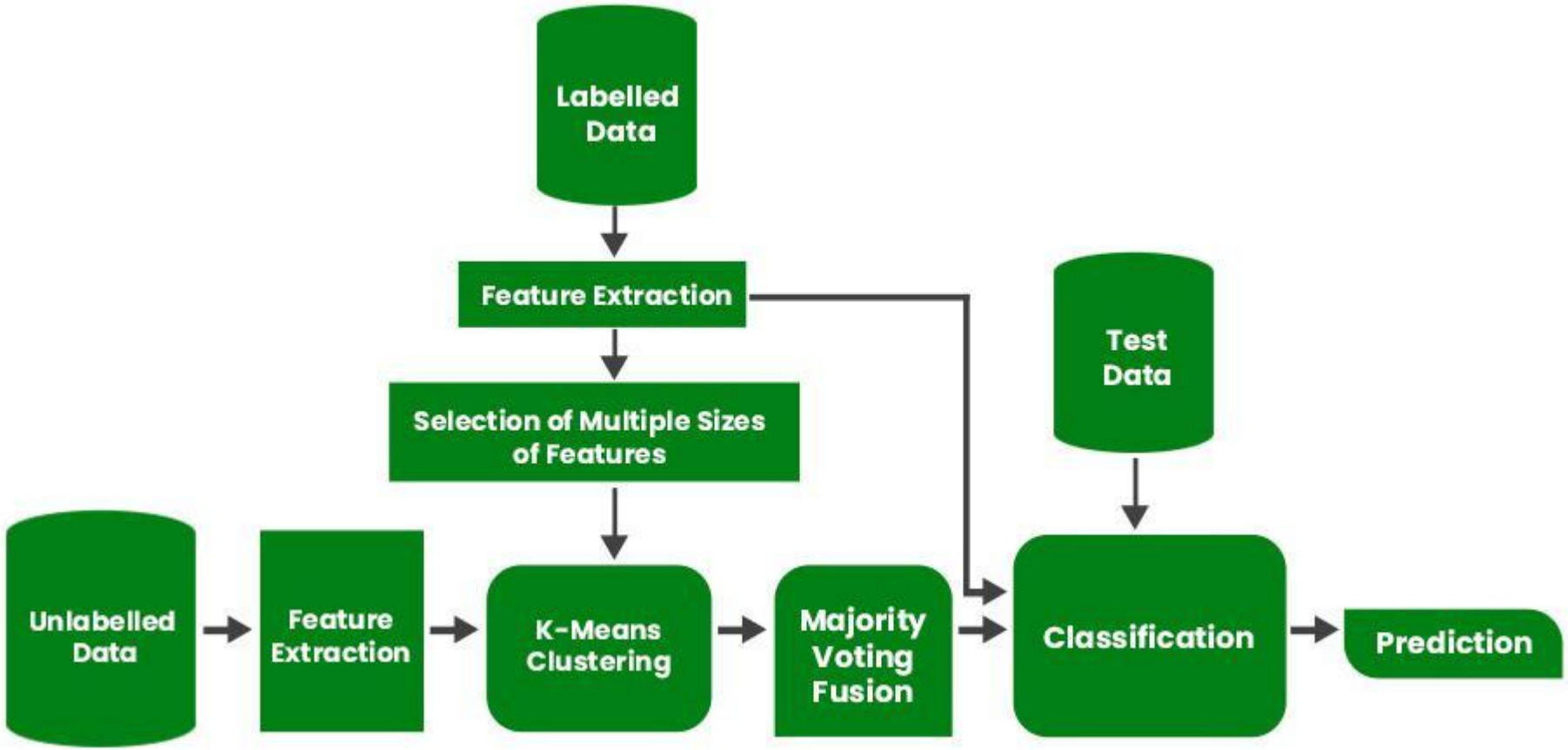


Unsupervised Learning

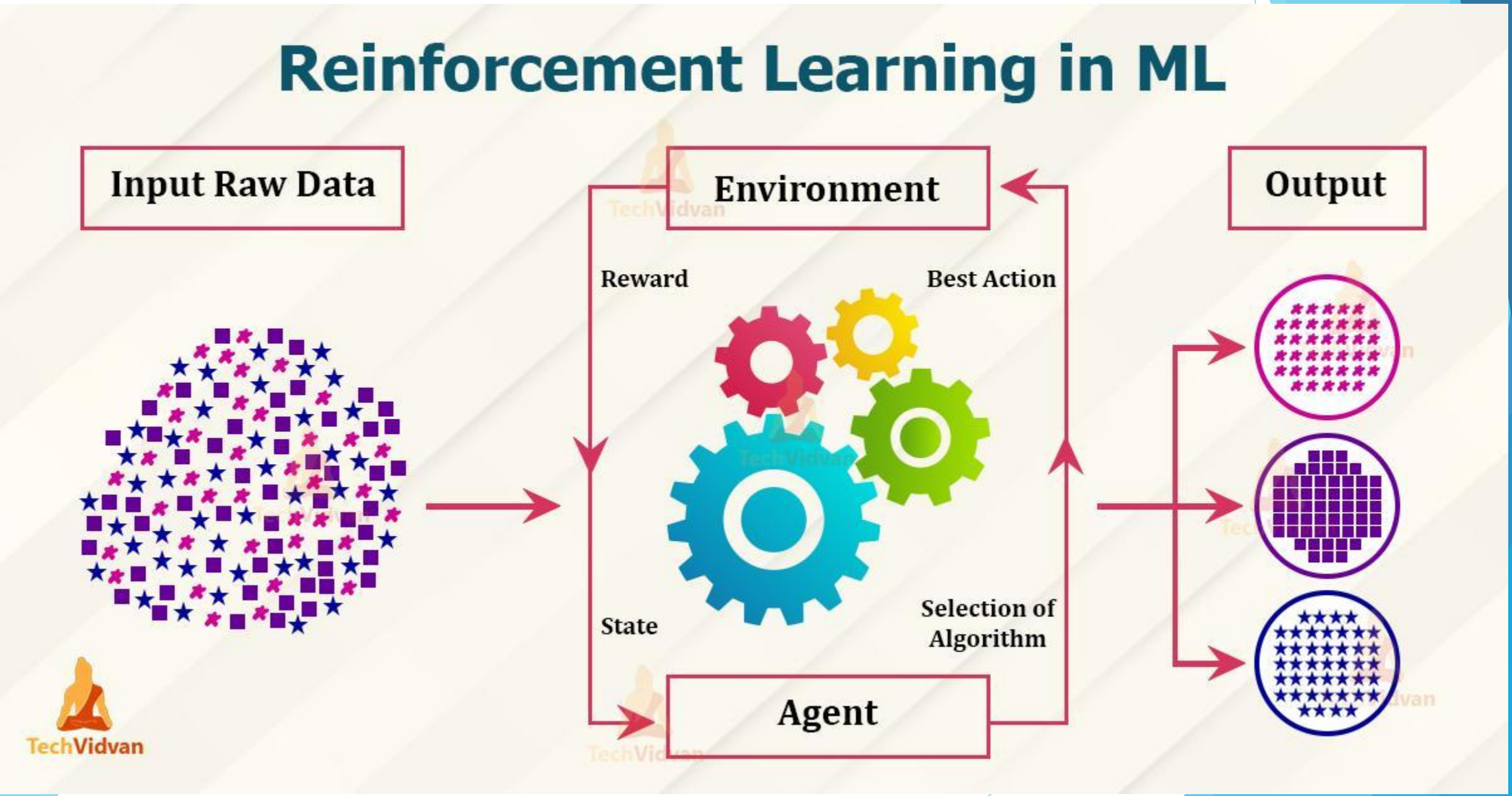
- “Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.”
- Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data.
- The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.



Semi Supervised Learning



Reinforcement Learning



How to choose the right Algorithm

- First, you need to consider your goal. What are you trying to get out of this? (Do you want a probability that it might rain tomorrow, or do you want to find groups of voters with similar interests?) What data do you have or can you collect?
- If you're trying to predict or forecast a target value, then you need to look into supervised learning. If not, then unsupervised learning is the place you want to be.
- If you've chosen supervised learning, what's your target value?
- Is it a discrete value like Yes/No, 1/2/3, A/B/C, or Red/Yellow/Black? If so, then you want to look into classification.
- If the target value can take on a number of values, say any value from 0.00 to 100.00, or -999 to 999, or +_ to -_, then you need to look into regression

Contd…

- If you're not trying to predict a target value, then you need to look into unsupervised learning.
- Are you trying to fit your data into some discrete groups? If so and that's all you need, you should look into clustering.
- Do you need to have some numerical estimate of how strong the fit is into each group? If you answer yes, then you probably should look into a density estimation algorithm.
- You should spend some time getting to know your data, and the more you know about it, the better you'll be able to build a successful application.

Steps in developing a machine learning application

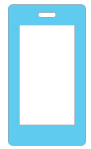
Collect data

Prepare the input data

Analyse the input data

Train the algorithm

Test the algorithm



Use it

Reading a dataset

- Continuous Data

Id	Sepal-Length	Sepal-Width	Petal-Length	Petal-Width	Species
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3	1.4	0.2	Setosa
3	4.7	3.2	1.3	0.2	Setosa
4	7	3.2	4.7	1.4	Versicolor
5	6.4	3.2	4.5	1.5	Versicolor
6	6.9	3.1	4.9	1.5	Versicolor
7	6.3	3.3	6	2.5	Virginica
8	5.8	2.7	5.1	1.9	Virginica
9	7.1	3	5.9	2.1	Virginica

- Categorical Data

ID	Marital Status	Race	Sex	Income
1	Widowed	White	Female	<=50K
2	Widowed	White	Female	<=50K
3	Widowed	Black	Female	<=50K
4	Divorced	White	Female	<=50K
5	Separated	White	Female	<=50K
6	Divorced	White	Female	<=50K
7	Separated	White	Male	<=50K
8	Never-married	White	Female	>50K
9	Divorced	White	Female	<=50K
10	Never-married	White	Male	>50K

- Continuous Data
- Categorical Data
- Targets (Categorical / Continuous)

- Features / Attributes
- Classes
- Training Set / Testing Set

Training, Testing and Validation Dataset

TRAINING SET

The subset of data used to train a machine learning model

TEST SET

The subset of data used to evaluate the performance of a trained machine learning model on unseen examples, simulating real-world data

VALIDATION SET

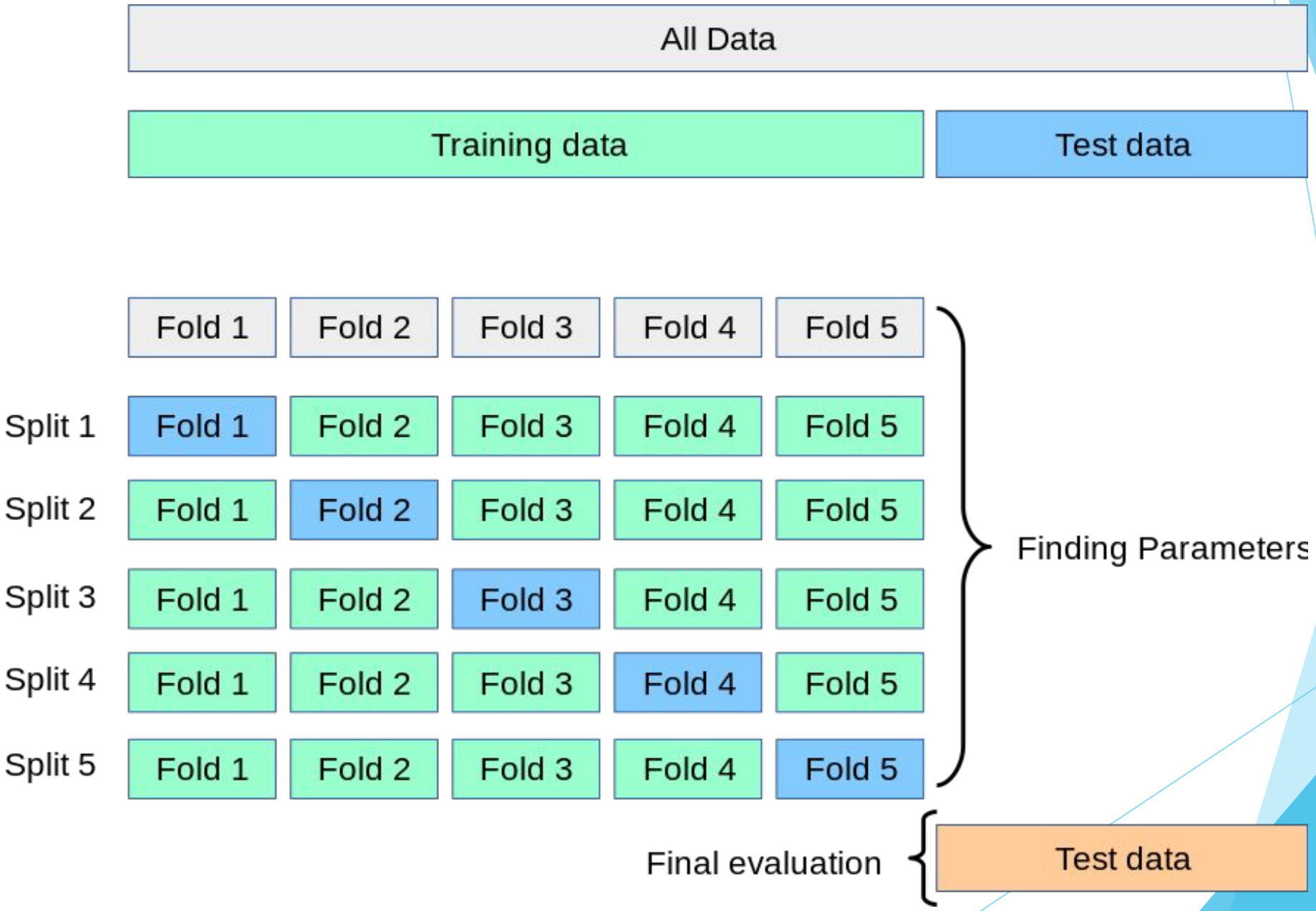
The intermediary subset of data used during the model development process to fine-tune hyperparameters

Cross Validation

- In machine learning, we couldn't fit the model on the training data and can't say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. For this purpose, we use the cross-validation technique.

- Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data. It involves dividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds. This process is repeated multiple times, each time using a different fold as the validation set. Finally, the results from each validation step are averaged to produce a more robust estimate of the model's performance.

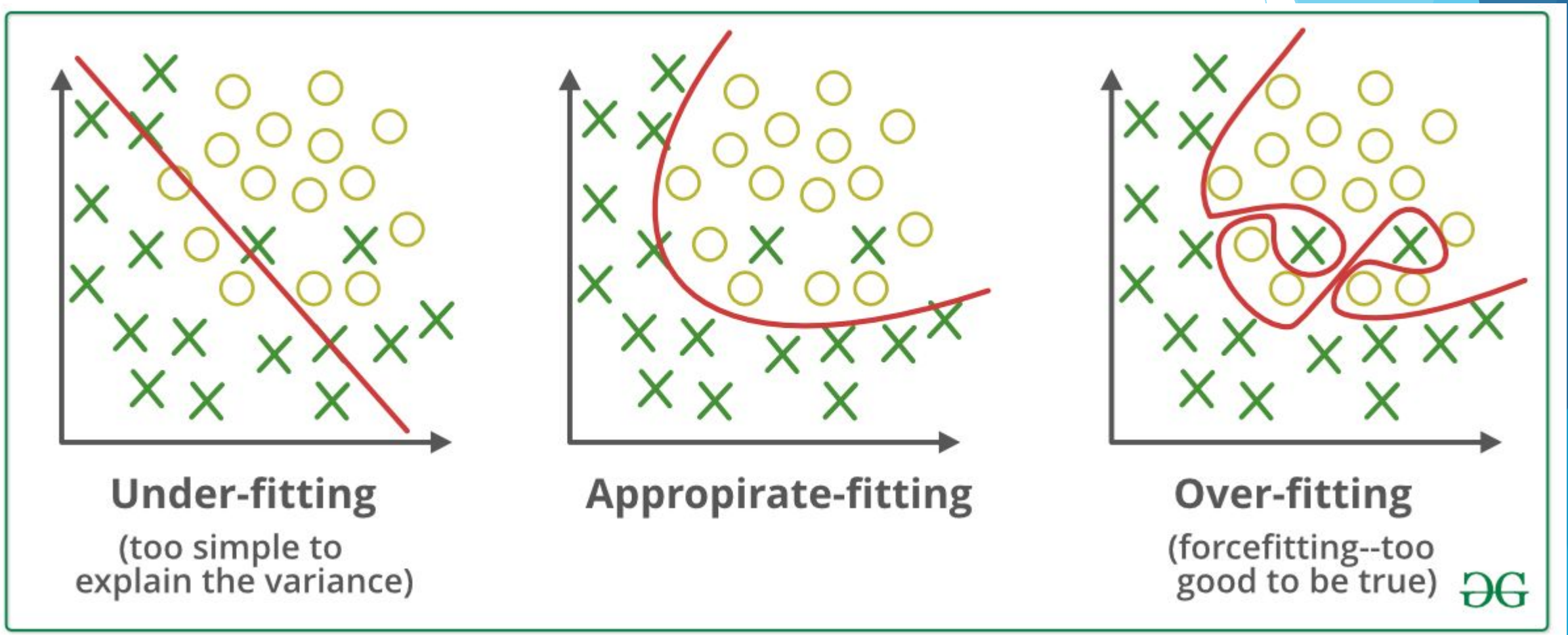
Cross Validation



Overfitting and Underfitting

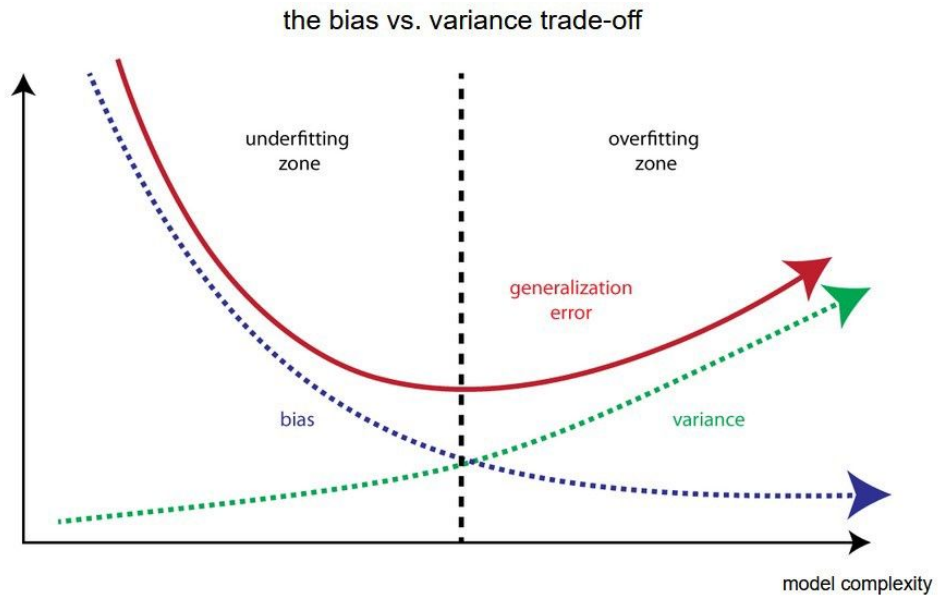
- When we talk about the Machine Learning model, we actually talk about how well it performs and its accuracy which is known as prediction errors.
- Let us consider that we are designing a machine learning model.
- A model is said to be a good machine learning model if it generalizes any new input data from the problem domain in a proper way.
- This helps us to make predictions about the future data, that the data model has never seen.
- Now, suppose we want to check how well our machine learning model learns and generalizes to the new data.
- For that, we have overfitting and underfitting, which are majorly responsible for the poor performances of the machine learning algorithms.

Overfitting and Underfitting



Bias and Variance

- **Bias:** Assumptions made by a model to make a function easier to learn. It is actually the error rate of the training data. When the error rate has a high value, we call it High Bias and when the error rate has a low value, we call it low Bias.
- **Variance:** The error rate of the testing data is called variance. When the error rate has a high value, we call it High variance and when the error rate has a low value, we call it Low variance.



Bias and Variance

- **Low Bias**: Suggests less assumptions about the form of the target function.
- **High-Bias**: Suggests more assumptions about the form of the target function.

- **Low Variance**: Suggests small changes to the estimate of the target function with changes to the training dataset.
- **High Variance**: Suggests large changes to the estimate of the target function with changes to the training dataset.

Underfitting

- A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data.

- Underfitting destroys the accuracy of our machine learning model.
- Its occurrence simply means that our model or the algorithm does not fit the data well enough.

- An underfitted model has **high bias** and **low variance**.

Underfitting

- It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data.
- In such cases, the rules of the machine learning model are too easy and flexible to be applied on such minimal data and therefore the model will probably make a lot of wrong predictions.
- Underfitting can be avoided by using more data and also increasing the features by feature selection.

Underfitting

- **Reasons for Underfitting:**
  - High bias and low variance
  - The size of the training dataset used is not enough.
  - The model is too simple.
  - Training data is not cleaned and also contains noise in it.
- **Techniques to reduce underfitting:**
  - Increase model complexity
  - Increase the number of features, performing feature engineering
  - Remove noise from the data.
  - Increase the number of epochs or increase the duration of training to get better results.

Overfitting

- A statistical model is said to be overfitted when the model does not make accurate predictions on testing data.
- When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set.
- And when testing with test data results in High variance.
- Then the model does not categorize the data correctly, because of too many details and noise.
- The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.

Overfitting

- A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

- Very Good Validation Accuracy and Very Poor Testing Accuracy.

- The overfitted model has **low bias** and **high variance.**

Overfitting

- **Reasons for Overfitting are as follows:**
  - High variance and low bias
  - The model is too complex
  - The size of the training data
- **Techniques to reduce overfitting:**
  - Increase training data.
  - Reduce model complexity.
  - Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
  - Ridge Regularization and Lasso Regularization
  - Use dropout for neural networks to tackle overfitting.

Overfitting and Underfitting

- Use these steps to determine if your machine learning model, deep learning model or neural network is currently **underfit** or **overfit**.

- Ensure that you are using validation loss next to training loss in the training phase.
- When your validation loss is decreasing, the model is still underfit.
- When your validation loss is increasing, the model is overfit.
- When your validation loss is equal, the model is either perfectly fit or in a local minimum.

Issues in Machine Learning

- Which Algorithm to select?
- How much training data is sufficient?
- Prior knowledge held by the learner is used at which time and manner to guide the process of generalization from examples?
- What is the best strategy for choosing a useful next training experience, and how does the choice of this strategy affect the complexity of the learning problem?
- To reduce the task of learning approximation of the problems is carriedout, what will be the best approach?
- To improve the knowledge representation and to learn the target function, how can the learner automatically alter its representation?

Applications of Machine Learning

- Automating Employee Access Control
- Protecting Animals
- Predicting Emergency Room Wait Times
- Identifying Heart Failure
- Predicting Strokes and Seizures
- Predicting Hospital Readmissions
- Stop Malware
- Understand Legalese
- Improve Cybersecurity
- Get Ready For Smart Cars

Applications of Machine Learning