**Introduction**

Table of Contents

[Problem Types 4](#_Toc100149308)

[Feature Sets 4](#_Toc100149309)

[Types of Learning 6](#_Toc100149310)

[Supervised Learning 7](#_Toc100149311)

[Unsupervised Learning 7](#_Toc100149312)

[Semi-Supervised Learning 9](#_Toc100149313)

[Reinforcement Learning 10](#_Toc100149314)

[Steps of Machine Learning 11](#_Toc100149315)

The world currently revolves around Big Data. Every moment of our lives, we produce a huge amount of information. For example, the temperature changes of a room can be detected, which will produce a large amount of data within a few hours. We also use this data in different ways. For example, the temperature changes can be used to determine at what point to automatically turn on or turn off an air conditioner.

The data that we produce has certain patterns. These patterns are important as well, since we can do things like recognize and predict certain events. Wouldn’t a system that automatically recognizes that the occupants of a house leave by 8 AM every morning and turns of the air conditioner accordingly be amazing? Unfortunately, in real world scenarios, it is not always possible to analyse these patterns manually, especially given the amount of data that we have. This is where Machine Learning comes in.

Essentially, we will be learning how to train computers to be able to detect patterns in data. In **Computer Science**, we aim to develop **automated machinery** to accomplish **non-trivial tasks**. This is essentially what programming is. However, there are several situations in which we cannot **explicitly write a program**, such as hand-writing recognition. We cannot identify the steps required to finish the task. The Machine Learning model we create will instead be used to create the required program.

Consider that we have two countries, and , each with a population of . For each country, we create a height distribution. We then notice that the heights of people form country tend to be higher than the heights of people from country . From this, we can find a certain threshold point, the dividing point between the two height distributions. A very simple program could then be written which simply checks a provided height against the threshold to determine if the person is from country or country .

However, the height distribution and the threshold we determined apply to this specific pair of countries. We are unable to write a general program which can be used to determine which country a given person belongs to given that the pair of countries are random. This is what we mean when we say that we cannot explicitly write a program.

Instead, we need to develop a model that will first determine the threshold for a pair of countries and then generate the required program using that threshold. The way we do this is by providing the computer with a large amount of data along with the output for that data so that it can use the patterns in the data to determine what thresholds are required for specific outputs. This is essentially the learning stage.

This is where Machine Learning differs from conventional programming. In conventional programming, we provide a program and some data and the computer uses the program to determine the output for that data. In Machine Learning, in the learning stage, we provide the data and the output and the computer creates a program. Later on, we can test the program by giving it new data and seeing if the program it generates for that data gives an accurate output. This is essentially the **supervised** learning model.

Formally, Machine Learning is programming computers to optimize a performance criterion using example data or past experience. In the example we saw above, the performance criterion would be how accurately the program generated is able to place a particular person in one of the two countries, which is in turn dependent on how accurate the threshold is. We must decrease the number of misclassifications.

The model we create will be defined up to some parameters, the threshold for example, and the learning part is the program being optimized to those parameters based on the provided data. This can later on be used to make predictions, which will be more accurate the more we train our program.

## Problem Types

Machine Learning problems typically fall into one of two categories, **Classification** and **Regression**.

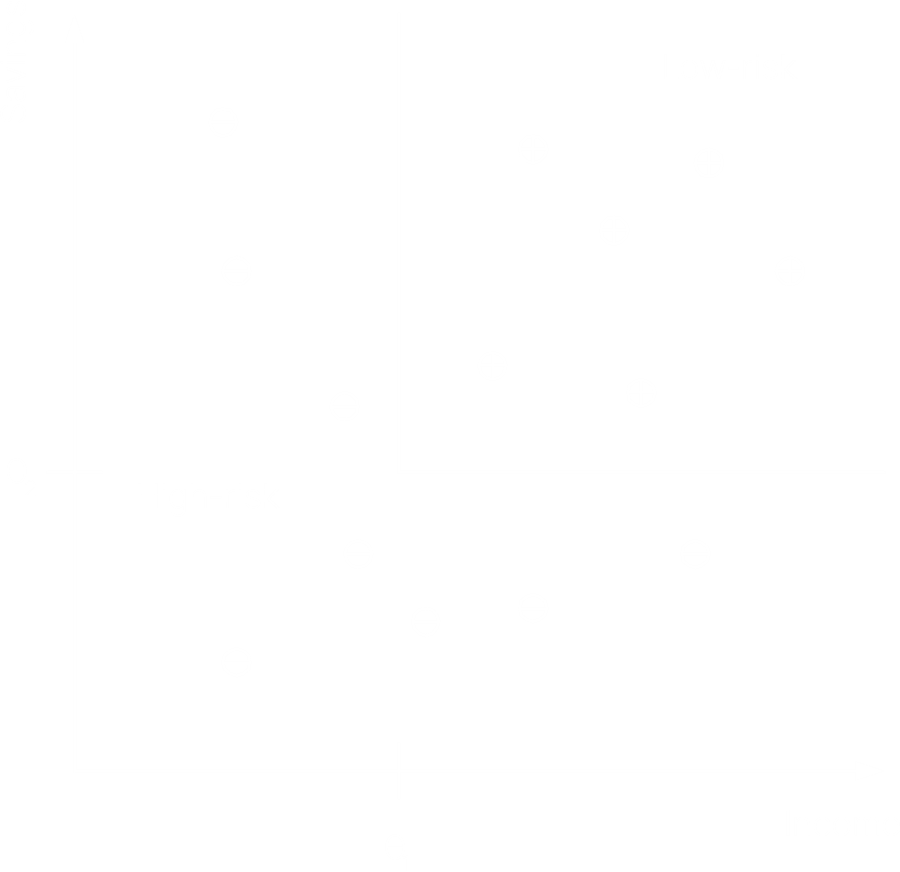
Classification problems deal with **discrete outputs**. We literally place data into classes. For example, if we want to decide whether a particular car is a good car or a bad car based on some metrics, this would be a classification problem.

Regression problems deal with **continuous outputs**. We predict a value for a piece of data and compare it with the actual value to find the amount of error in our model. For example, predicting the price of a car given its features is a regression problem.

## Feature Sets

Suppose we have **data points**, , , , , and they each have a corresponding **output**, , , , .

The value of a particular output is determined by **features**, , , , . Thus, a particular data point, , has some values for these features, from which we can determined some output . Thus, the features make up the **feature set**.



Consider the example above. Here, each data point has 2 features, Income and Savings. Based on these values, we are providing some output, namely whether the data-point is low-risk or high-risk.

Continuing the above example, a potential machine learning model would try to determine some threshold values for each feature, and above, using which the output can be determined. Thus, if a data point has a higher income that and a higher savings than , it is a low-risk data point.

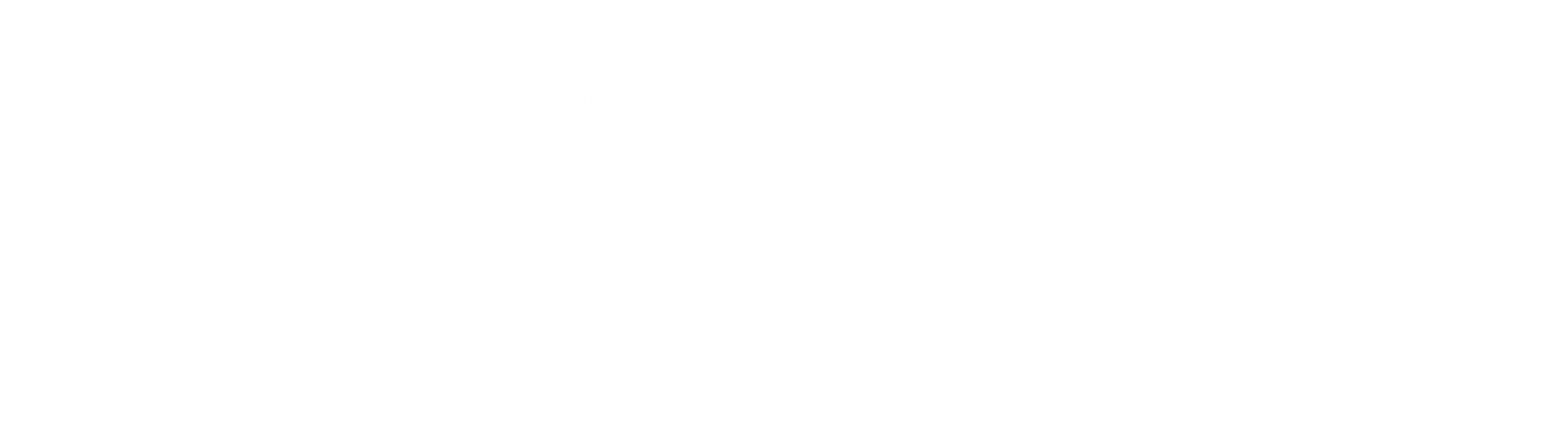
The model or function that is being used to determine the outputs can be considered . In a regression problem, we would need to minimize the loss function, i.e. .

## Types of Learning

There are five categories of machine learning algorithms:

1. **Supervised Learning** – The machine learns using training data which includes the correct outputs.
2. **Unsupervised Learning** – The machine learns using training data which does not include the correct outputs.
3. **Semi-Supervised Learning** – The machine learns using training data a few of which include the correct outputs while others do not.
4. **Reinforcement Learning** – The machine learns by using rewards and punishments. For a specific series of actions, the machine is given a reward while for others, it is punished. Using this, the machine understands which series of actions it needs to take.
5. **Active Learning** – The machine learns while also taking feedback from the user. For specific cases, it might ask the user to step in and correct it.

### Supervised Learning



The first step in **Supervised Learning** is to divide the data into two parts, the **training set** and the **test set**. For example, say of the data is set aside as the training set and the other is set aside as the test set.

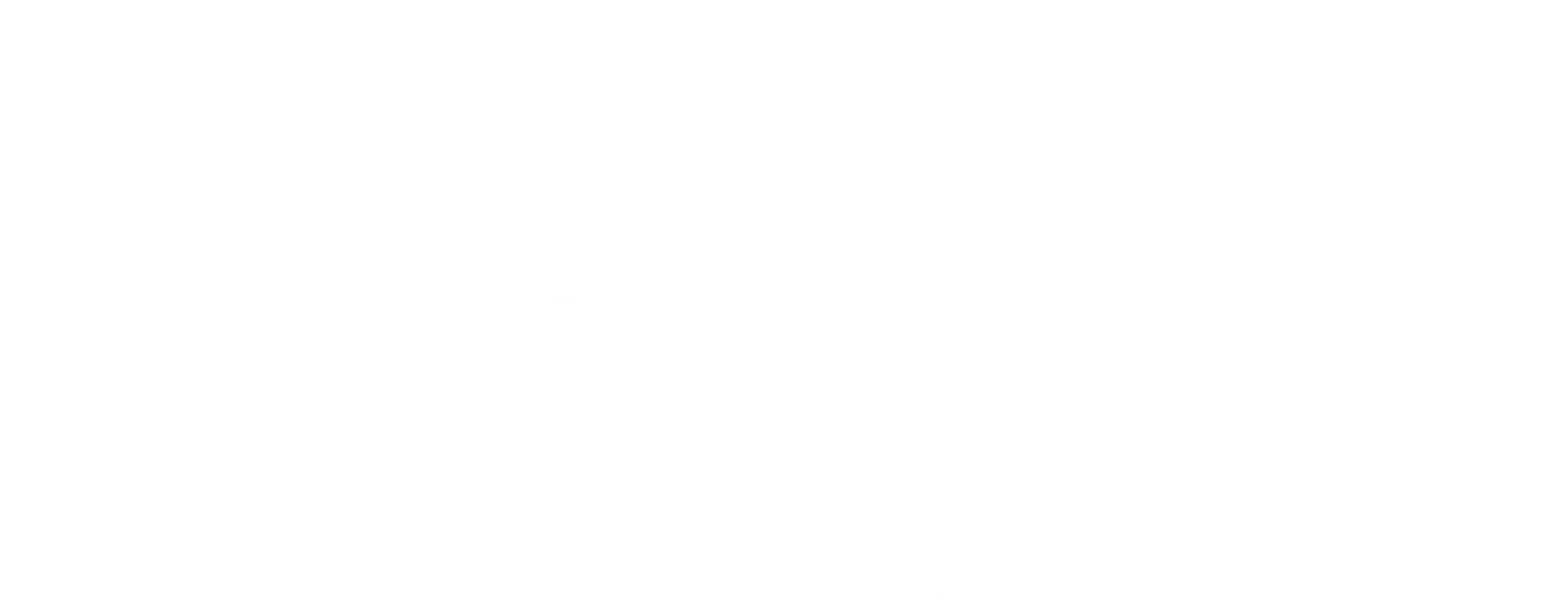
Next, using the training set, the machine creates a **model**. Once this is done, the test set is used to verify how accurate the model is. The outputs from the test set are not given to the model, only the input. The model gives a set of predictions, and these predictions are cross-checked with the actual outputs to find the percentage of cases in which the model gave the correct output.

### Unsupervised Learning

In **unsupervised learning**, we essentially derive a structure from the data without knowing the effect of the variables. We do this by **clustering** the data based on the relationships between the variables. Since we do not have the correct results, we do not have any feedback for our predictions.

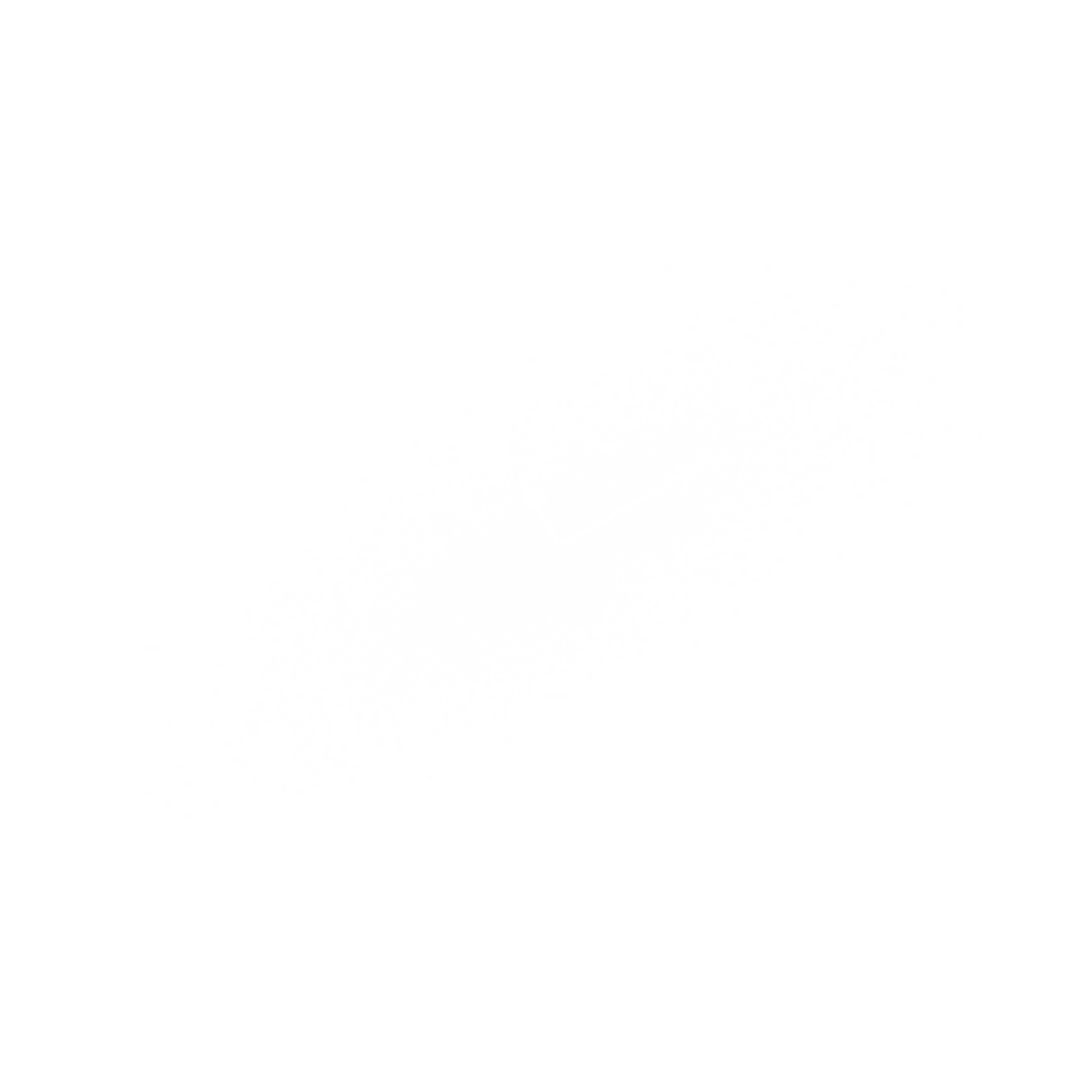
Unsupervised learning can be of several types, such as Agglomerative Clustering, PCA, ICA, etc.

In the **Agglomerative Clustering** algorithm, we attempt to divide data points into groups and determine how many distinct groups there are.



The number of groups is a debatable question. For example, some may say there are three groups of data points in the case above, while others may say there are two, the data point on top and the ones on the bottom.

In **PCA**, we reduce the number of dimensions in the feature set.



For the graph above, the feature set has two dimensions. If we consider the line that is roughly along , and we transpose all the data points onto that line, we can reduce this to just one dimension. Reducing the number of dimensions can be very useful when we are dealing with a very large feature set. For example, if the DNA of each human being is a data point, then the feature set could be every characteristic determined by the DNA sequence, which is in the millions.

Now, the question is, why did we choose the line that is roughly along ? Why not say, ? If we take any line and transpose the data points onto that line, we will find that the line gives us the widest range. This is useful because it becomes difficult to work with a narrow range.

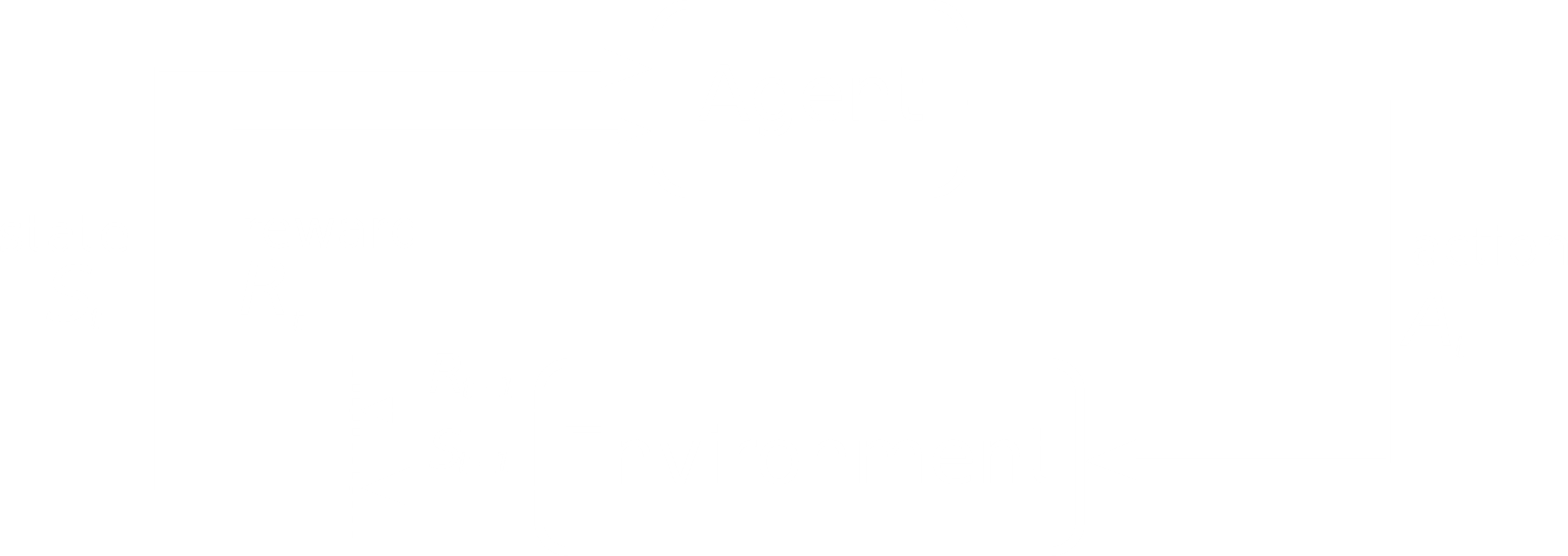
In **ICA**, we deal with non-clustering problems. Consider that there is sound coming from multiple sources that are getting mixed together. Using ICA, we can separate out each of the sound sources from the mixed signal.

### Semi-Supervised Learning

We know that in **Semi-Supervised Learning**, some of the data points have known outputs while others do not. In this case, the machine first works with just the data-points for which we know the outputs. Based on this, it creates a model. Next, it adds the other data points and refines the boundaries of the model to create a final one.

### Reinforcement Learning

In **reinforcement learning**, the machine learns to assess the goodness of policies and learn from past good action sequences to generate a policy.



Consider that we are building a machine that can play chess. After each move, the state of the chess board will change. Thus, at the end of a particular game, the machine knows which moves it made at each state and what the outcome of that sequence was. After playing millions of games, it will be able to figure out what the best possible move in a particular state is that will give the highest chances of winning.

## Steps of Machine Learning

The basic steps of machine learning are:

1. Data Collection
2. Data Preparation – Once the data is collected, we may need to modify it a little. For example, some data points may have erroneous or missing values for some features. We need to deal with these.
3. Choosing a Model – There are multiple types of models we can choose from.
4. Training
5. Evaluation
6. Parameter Tuning – There are specific parameters which can be manually tuned. For example, one parameter is how fast the training stage is completed. We can adjust this to make it faster or slower, which will in turn affect our evaluation results.
7. Prediction

In this course, we will be focusing on:

* **Batch learning**, which means all the training data will be available at the start of the training stage, as opposed to online learning, where data becomes available in real time.
* **Passive learning**, as opposed to active learning, meaning we will not intervene to manually help the machine in the learning phase.
* **Non-structured prediction**, meaning the outcomes of the data points will be simple, such as values or Booleans, and not structures like sequences or trees.