Lesson 11

Machine Learning – General Concepts

Kush Kulshrestha

Example – Dr. Delta

Dr. Delta works to diagnose patients with malaria.
However, it takes a long time for her to see everyone.

Luckily, Dr. Delta has **historical patient data about what factors predict malaria**, such as body temperature, travel history, age, medical history.

Dr. Delta can use historical data as an input in a **machine learning algorithm** to help her predict whether a new patient will have malaria.

Example – Dr. Delta

Patient No	Patient Name	Temperature	Travel History	Age	Medical History	Other features…

Patient Has Malaria

Features (x)

Labels (y)

Example – Dr. Delta

Based on our experience of the world, we have an understanding of relationships between features

Computers acquire human intuition and quantify it, by extracting patterns from raw data

Training / Test Split

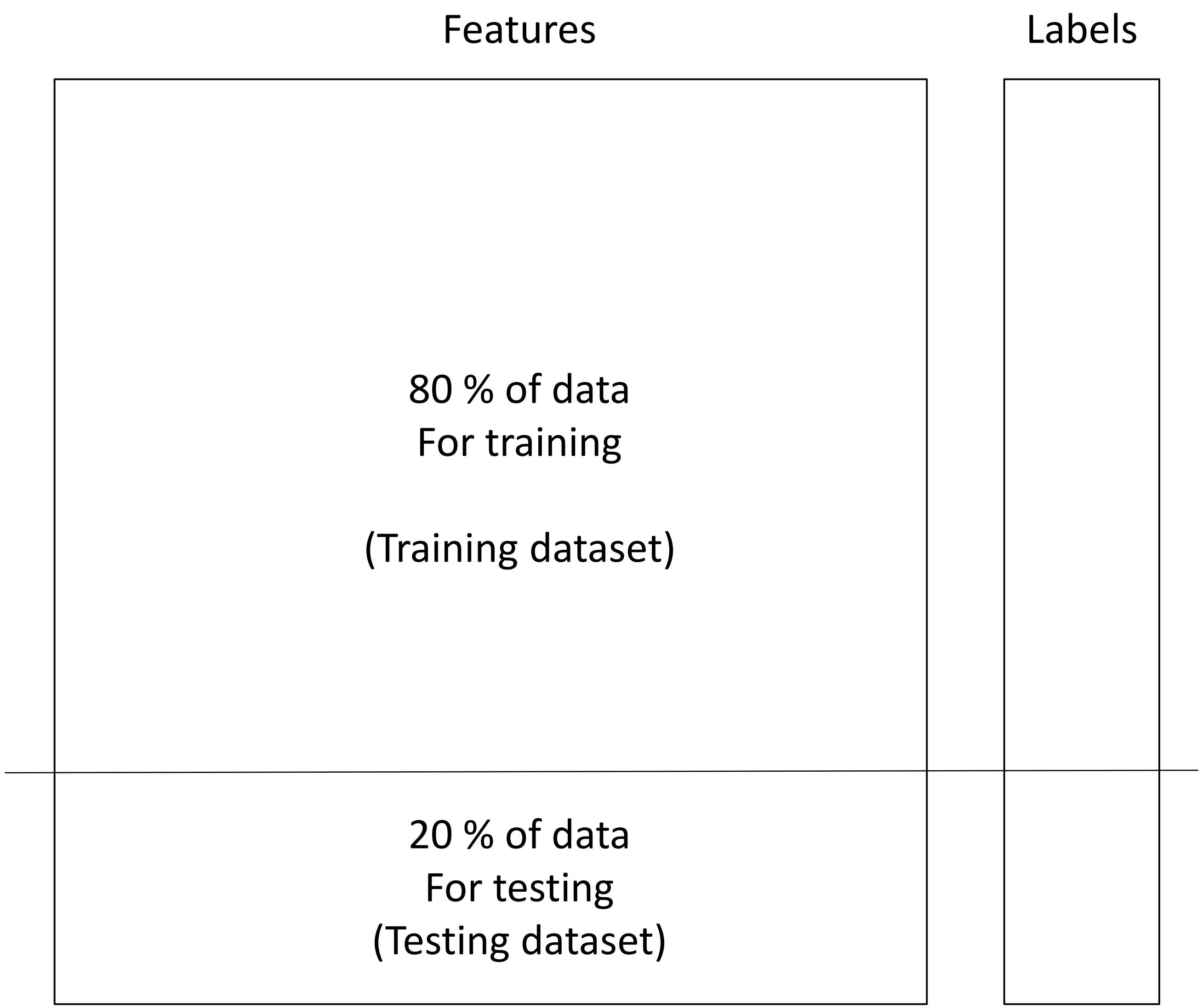
Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
			⋮			

labels
⋮

Training / Test Split

Step I:

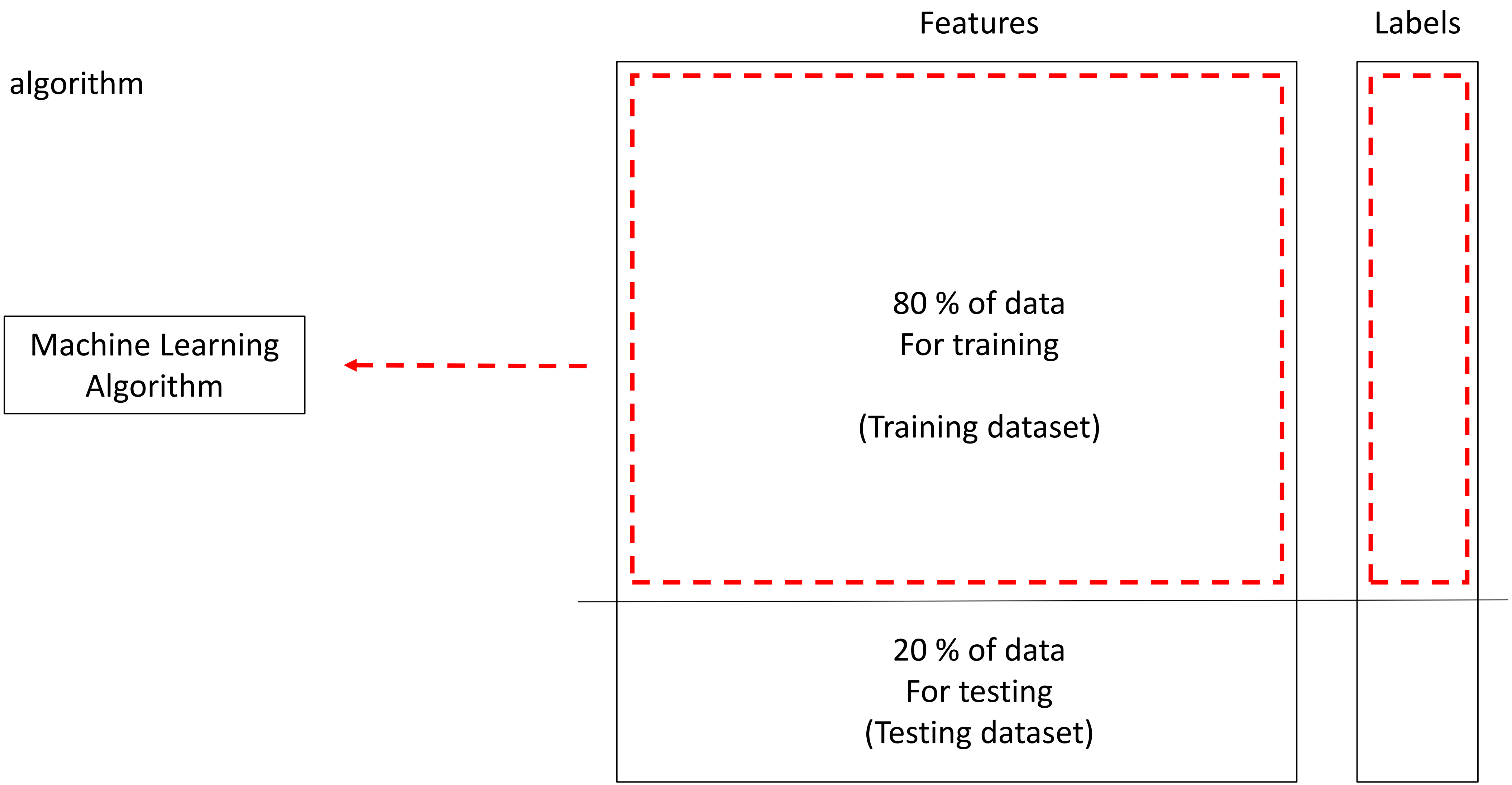
Splitting the dataset



Training / Test Split

Step II:

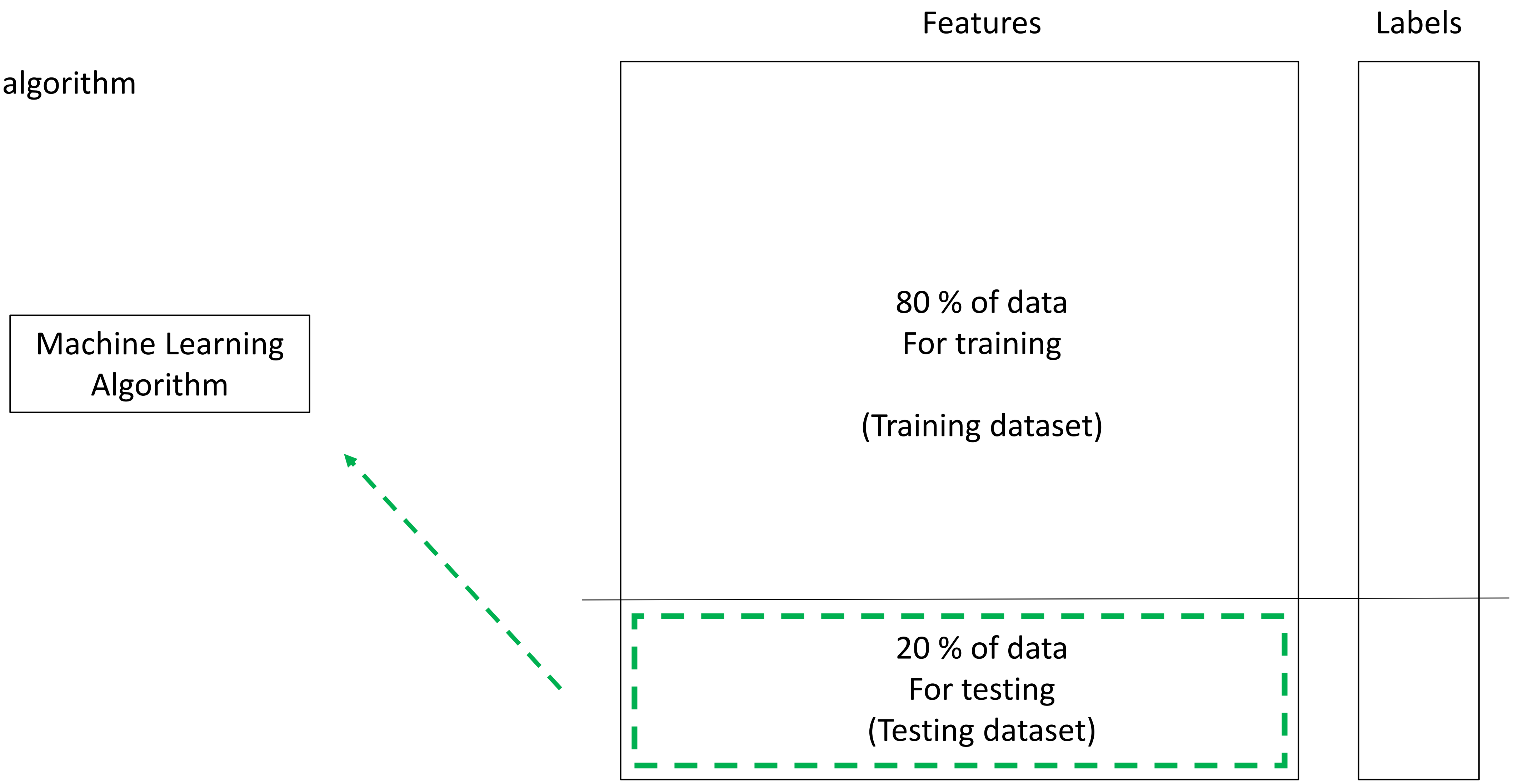
Training the algorithm



Training / Test Split

Step III:

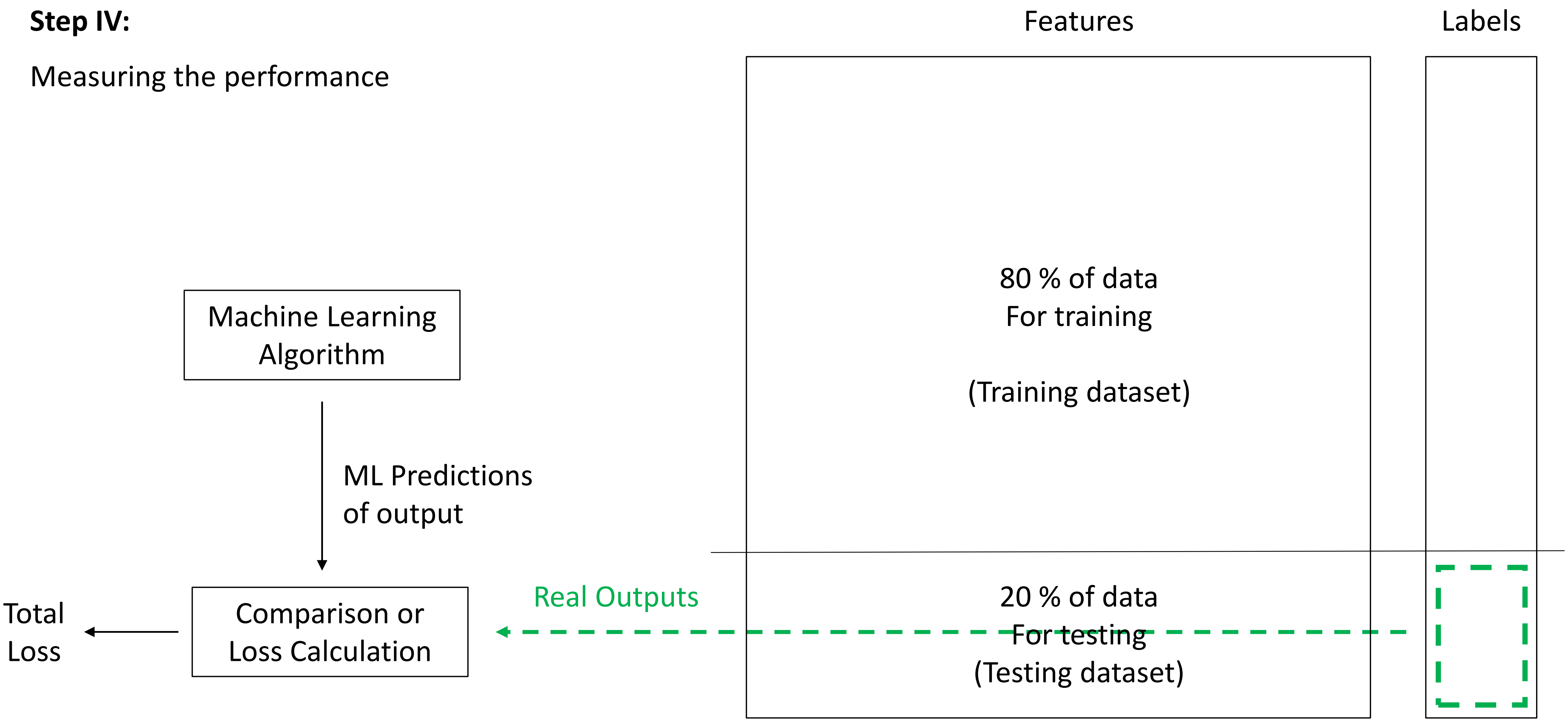
Testing the algorithm



Training / Test Split

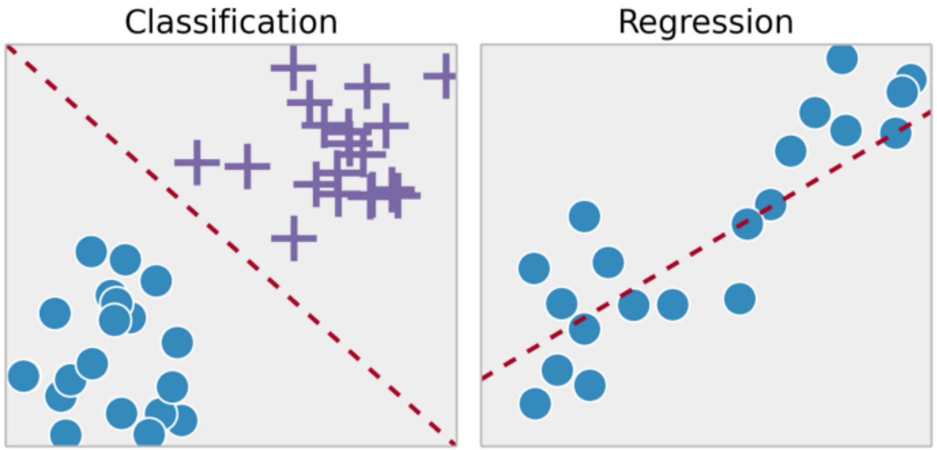
Step IV:

Measuring the performance

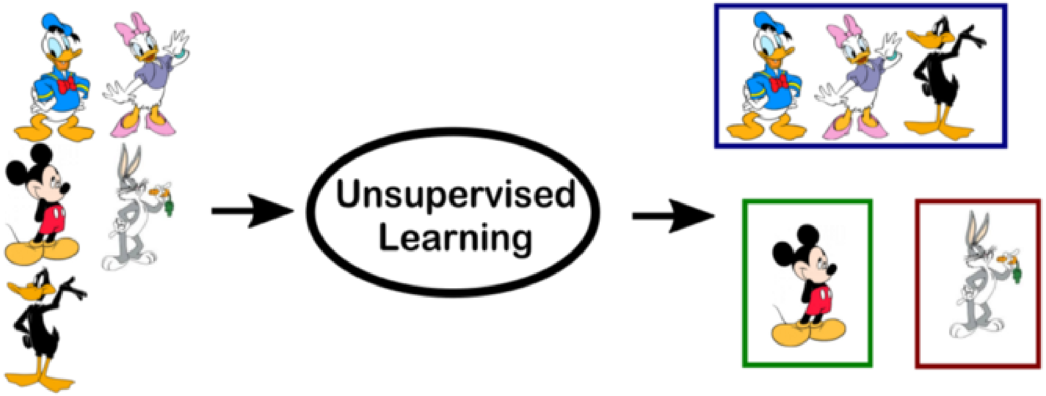


Types of ML Algorithms

Supervised Machine Learning Algorithms



Unsupervised Machine Learning Algorithms



Types of ML Algorithms

Supervised Machine Learning Algorithms

- The goal is to find specific relationships or structure in the input data that allow us to effectively produce correct output data.

- When conducting supervised learning, the main considerations are model complexity, and the bias-variance trade-off. Note that both of these are interrelated.

- Model complexity refers to the complexity of the function you are attempting to learn — similar to the degree of a polynomial.

- This is because a high-complexity model will **overfit** if used on a small number of data points. Overfitting refers to learning a function that fits your training data very well, but does not **generalize** to other data points, in other words, you are strictly learning to produce your training data without learning the actual trend or structure in the data that leads to this output.

Types of ML Algorithms

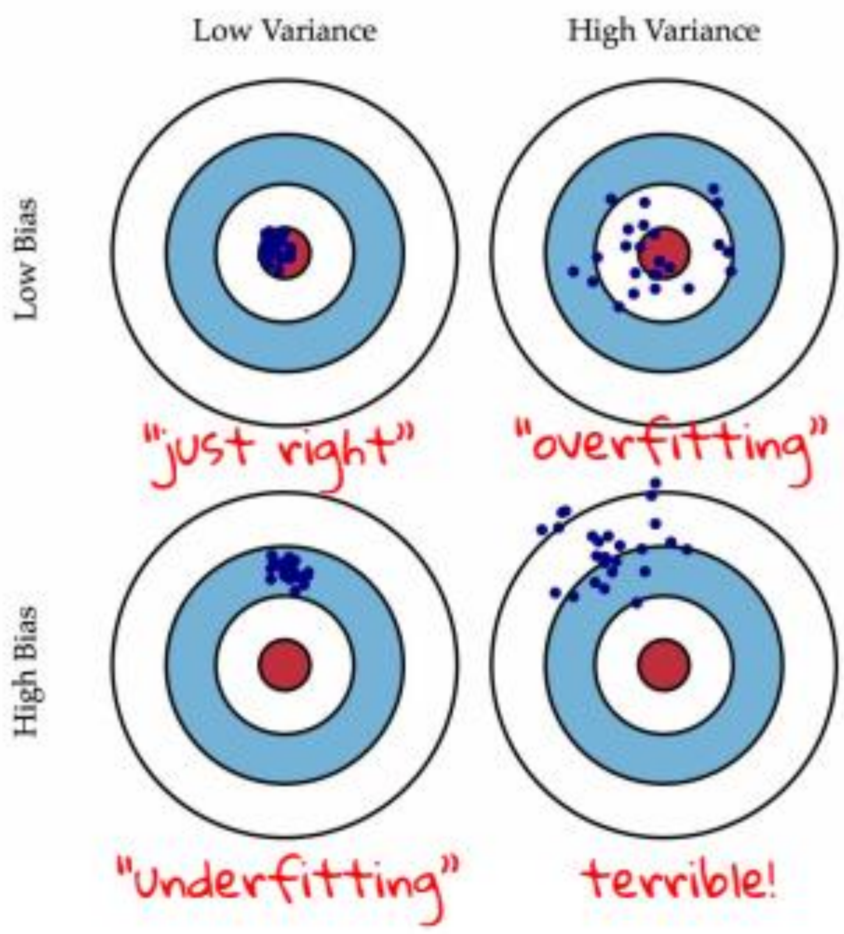
Supervised Machine Learning Algorithms and Bias-Variance trade-off

- The bias-variance tradeoff also relates to model generalization.

- In any model, there is a balance between bias, which is the constant error term, and variance, which is the amount by which the error may vary between different training sets.

- So, high bias and low variance would be a model that is consistently wrong 20% of the time, whereas a low bias and high variance model would be a model that can be wrong anywhere from 5%-50% of the time, depending on the data used to train it.

(Note that bias and variance typically move in opposite directions of each other; increasing bias will usually lead to lower variance, and vice versa.)

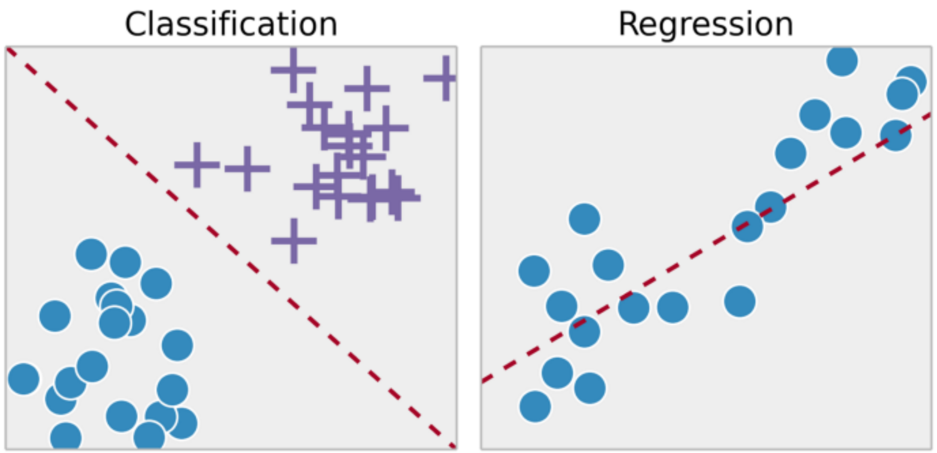


Types of ML Algorithms

Supervised Machine Learning Algorithms and its types

- In classification algorithms, task is to separate the dataset into different classes. Eg, Is a given email message spam or not spam? Is this an image of a dog, a cat, or a hamster?

- In regression, the goal is to predict any continuous variable depending the input data. Eg, What is the value of a house in California? Or What is the probability that a user will click on this ad?



Types of ML Algorithms

How Supervised Algorithms learn

Imagine you are a teacher and you asked a question to your students

The labels Y provide the correct answer to the question your students are trying to solve. Since you know the answer you can reward a good student performance and punish a bad one.

So every time your student (Mr. model) estimates the number of people with malaria (Y*), you compare it with the true answer and check how well Mr. model is doing.

Since model doesn’t know anything in starting, it will randomly guess the number of people with Malaria. They are going to be wrong. One way of checking how wrong the results are is to measure (Y* - Y)

More is this value, farther is Mr. model’s guess from the true value and more is the loss. This is called loss function, which quantifies the difference of predictions from the real values.

Mr. models goal is to minimize the loss function.

A loss function quantifies how unhappy you would be if you use f(x) to predict Y* when the correct value is y.

In other words, a loss function how well our model fits the data.

Types of ML Algorithms

Unsupervised Machine Learning Algorithms

- In unsupervised Machine learning, we wish to learn the inherent structure of our data without using explicitly-provided labels.

- The most common tasks within unsupervised learning are clustering, representation learning, and density estimation. Since no labels are provided, there is no specific way to compare model performance in most unsupervised learning methods.
- Two common use-cases for unsupervised learning are exploratory analysis and dimensionality reduction.

- Unsupervised learning is very useful in exploratory analysis because it can automatically identify structure in data. For example, if an analyst were trying to segment consumers, unsupervised clustering methods would be a great starting point for their analysis. In situations where it is either impossible or impractical for a human to propose trends in the data, unsupervised learning can provide initial insights that can then be used to test individual hypotheses.