Masterschool

June 2024

Fundamentals of Machine Learning

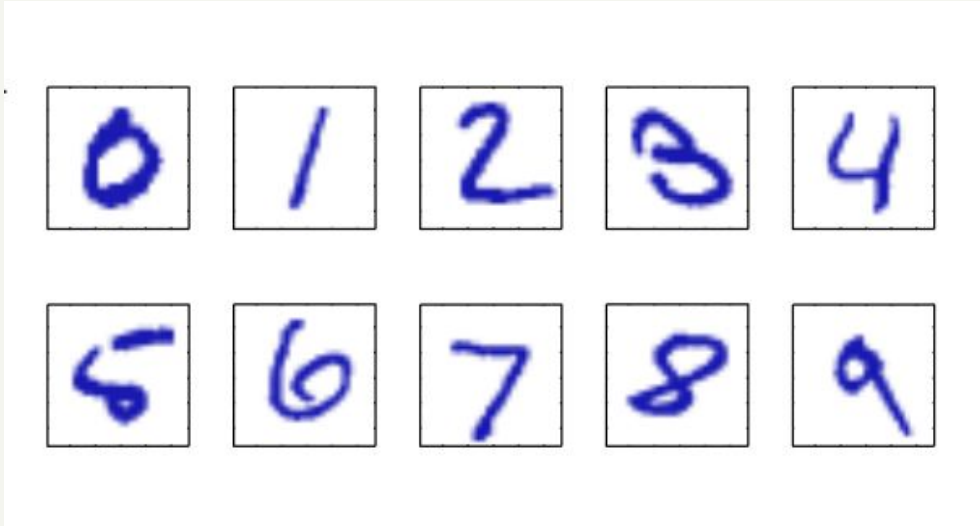
What is Machine Learning?

Machine Learning is a

branch of artificial intelligence focused on building systems that learn from data, identify patterns, and make decisions. As far as we are concerned in this course, it is statistical learning.

Motivating Examples

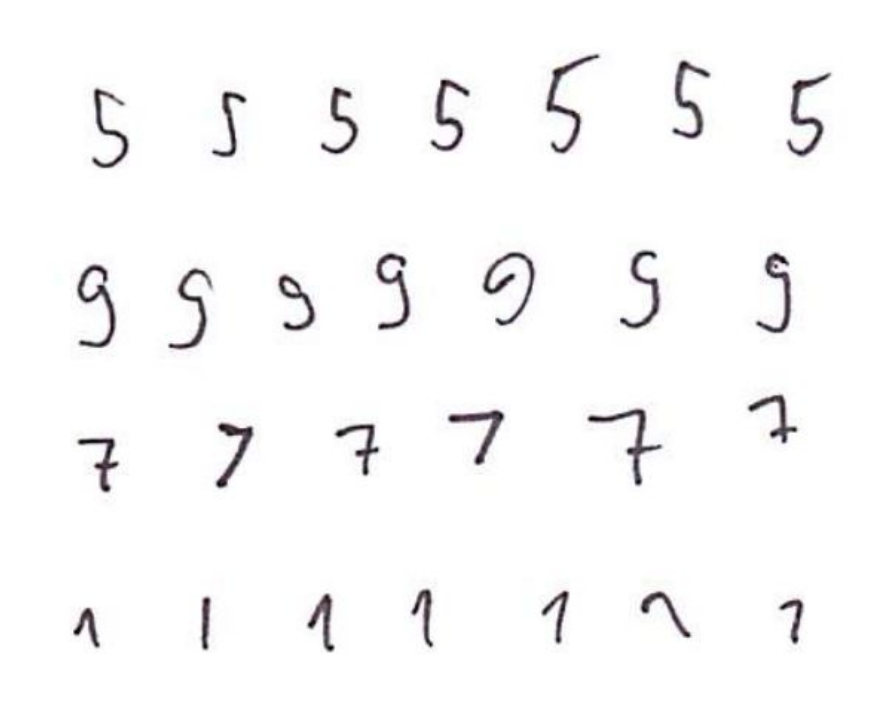
Digit recognition of a Zip Code



Question:

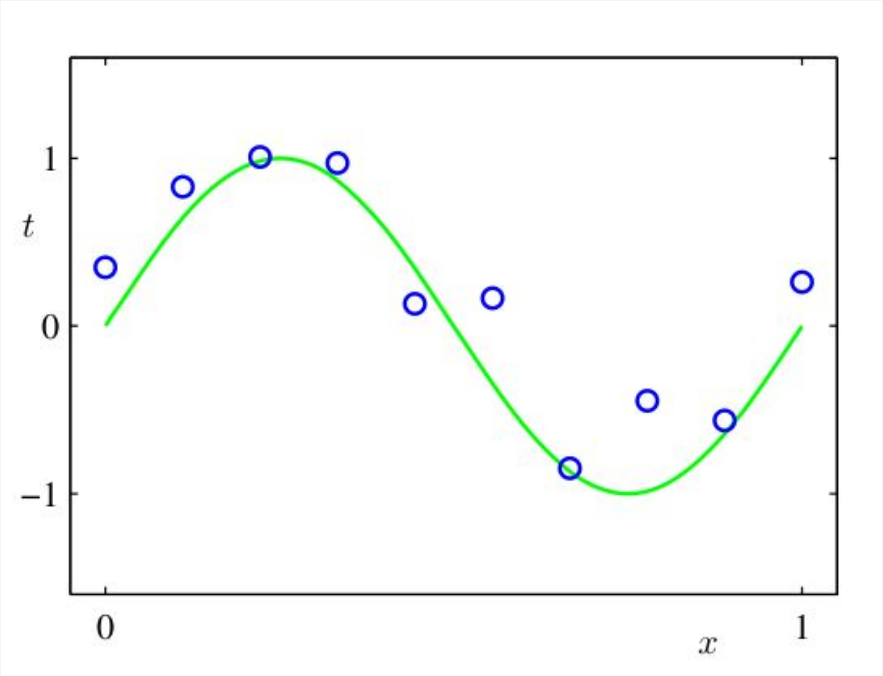
Is it possible to hand-design a rule to map each picture of a digit to its true value?

Digit recognition of a Zip Code



As you can see, the number of cases where we encounter a 5 or 9 is countless. It is impossible to establish a fixed rule to capture all possible scenarios, such as mapping all possible shapes of the digit 5 to the value 5.

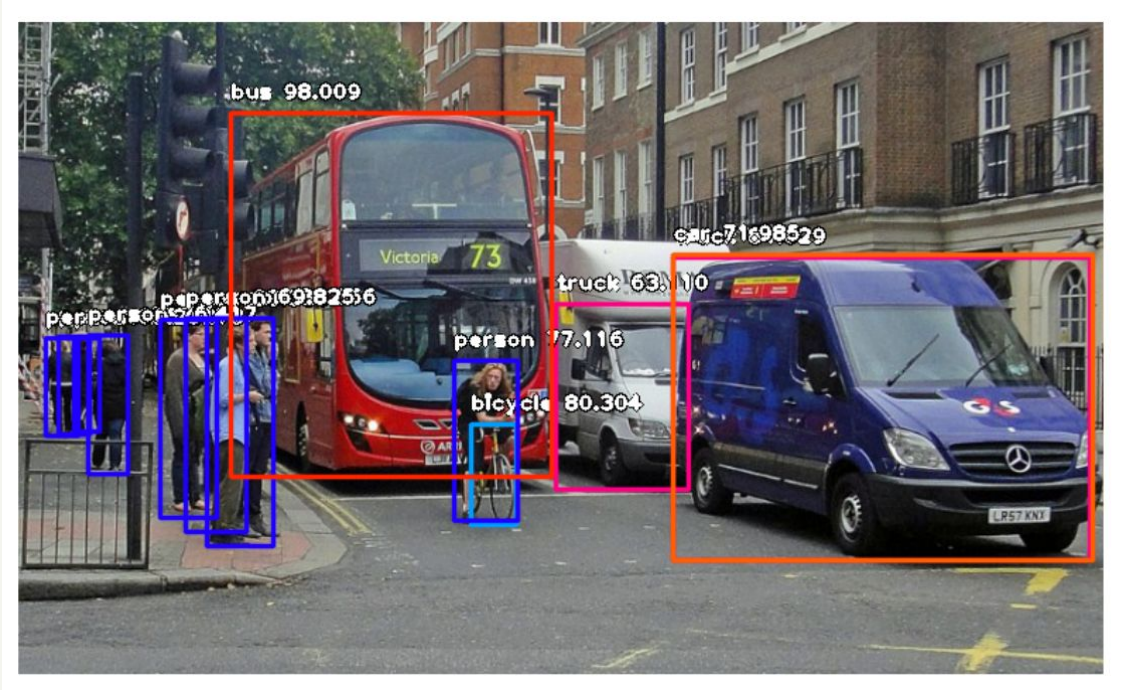
Curve Matching



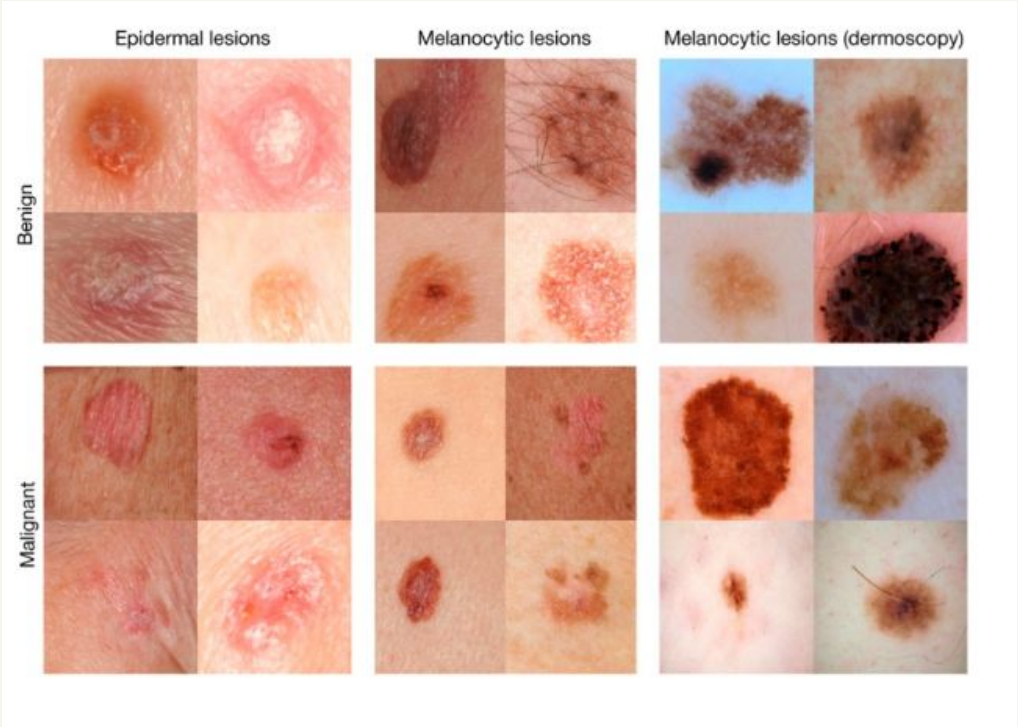
In this graph, we have collected information about a real situation involving two things, x and t.

- Our goal is to draw a line or curve that best matches this information.
- We also want to use this line or curve to guess the value of t when we have new x values.

Object Detection



Skin Cancer Detection



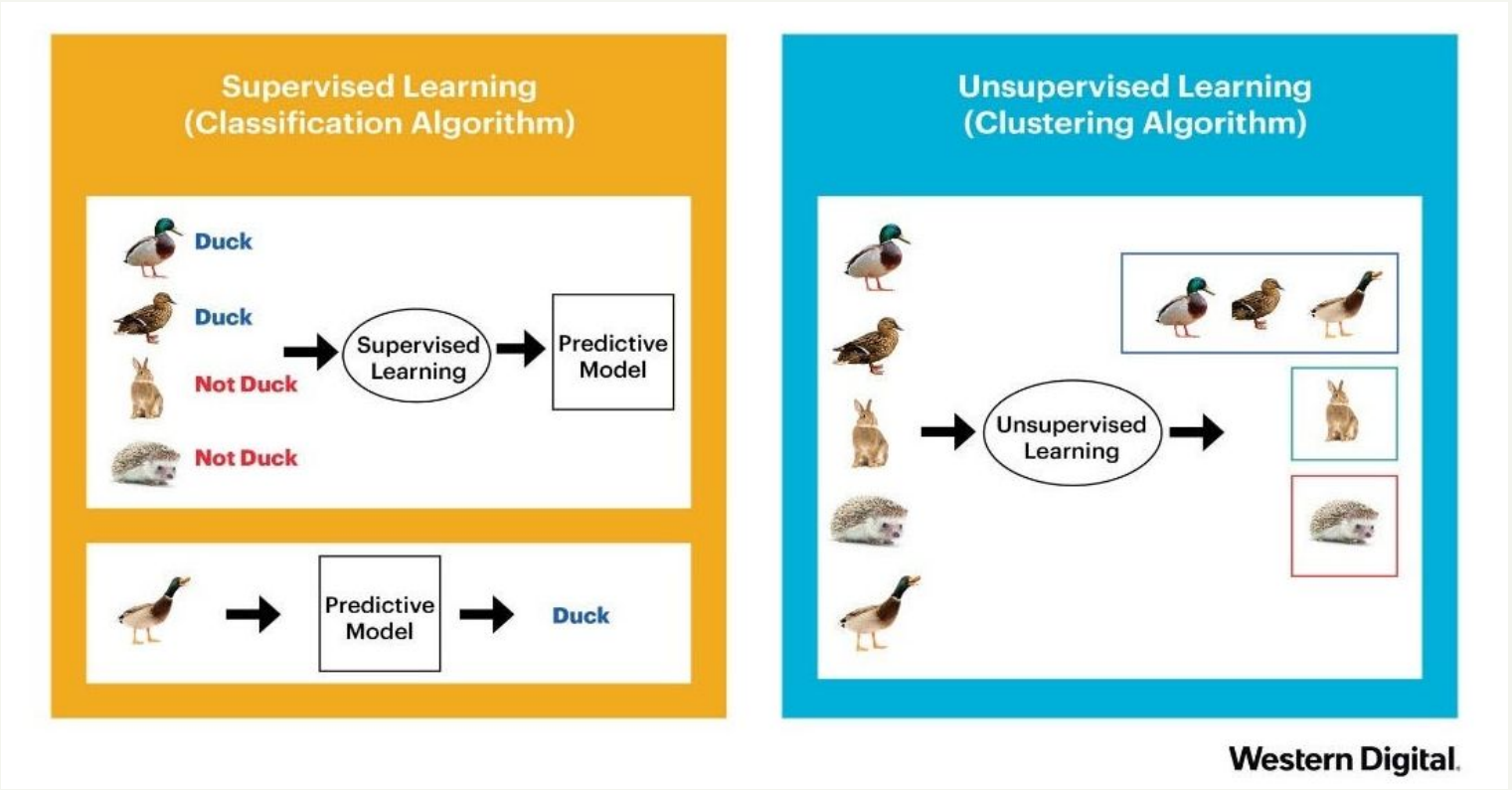
Summary

- Development of algorithms which allow a computer to “learn” specific tasks from training examples.
- Learning means that the computer can not only memorize the seen examples, but can generalize to previously unseen instances.
- Ideally, the computer should use the examples to extract a general “rule” how the specific task has to be performed correctly.

Supervised

vs

Unsupervised



https://miro.medium.com/v2/resize:fit:1400/0*4q_X_xQxevOQY_0u

- **Supervised Learning**

  This type of machine learning involves training a model on a **labeled** dataset, where the input data is paired with the correct output. The goal is to learn a **mapping from inputs to outputs** to make predictions on new, unseen data. Common algorithms include linear regression, decision trees, and neural networks

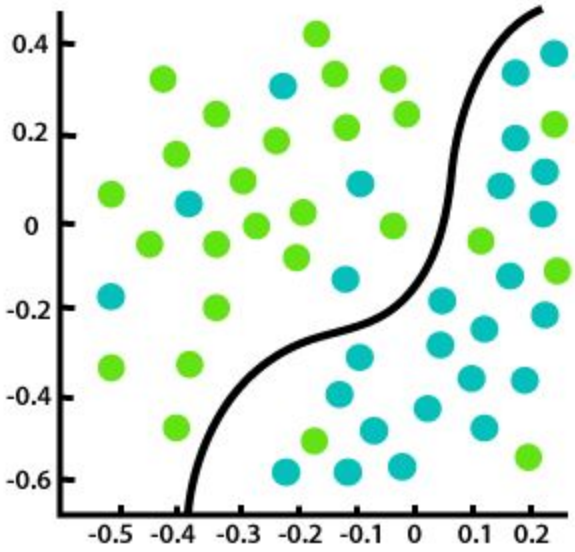
- **Unsupervised Learning**

  In unsupervised learning, the model is trained on **unlabeled** data without explicit instructions on what to predict. The goal is to **identify patterns**, structures, or relationships within the data. Common techniques include clustering (e.g., k-means, hierarchical clustering) and dimensionality reduction (e.g., PCA, t-SNE)

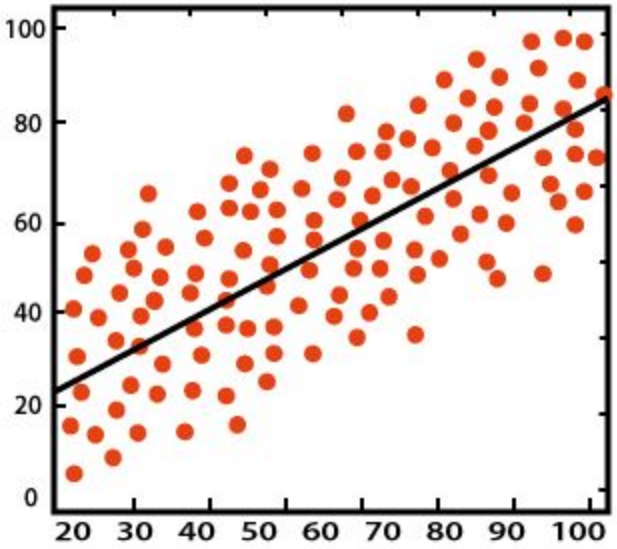
Regression

vs

Classification



Classification



Regression

As you have seen so far from the previous examples, there are two types of variables we want to predict:

- Variables with discrete values
- Variables with continuous values

The first type involves predicting discrete values, known as Classification problems. In these problems, we aim to classify an input into one of several predefined categories.

The second type involves predicting continuous values, known as Regression problems. Here, we aim to predict a continuous value within a range. The expected values may span from $-\infty$ to $+\infty$ or be constrained within a specific interval, such as $[a,b]$.

Deduction
vs
Induction

Who knows what deduction and induction mean?

Deductive Inference

Deductive inference is the process of reasoning from one or more general statements (premises) to reach a logically certain conclusion

Example:

> Premise 1: every person in this room is a student.
> Premise 2: every student is older than 10 years.
> Conclusion: every person in this room is older than 10 years.

If the premises are correct, then all conclusions are correct as well.

Nice in theory! Mathematics is based on this principle.

Inductive Inference

Reasoning that constructs or evaluates general propositions that are derived from specific examples.

Example:

> We drop lots of things, very often.
> In all our experiments, the things fall downwards, not upwards.
> So we conclude that likely, things always fall downwards when we drop them.

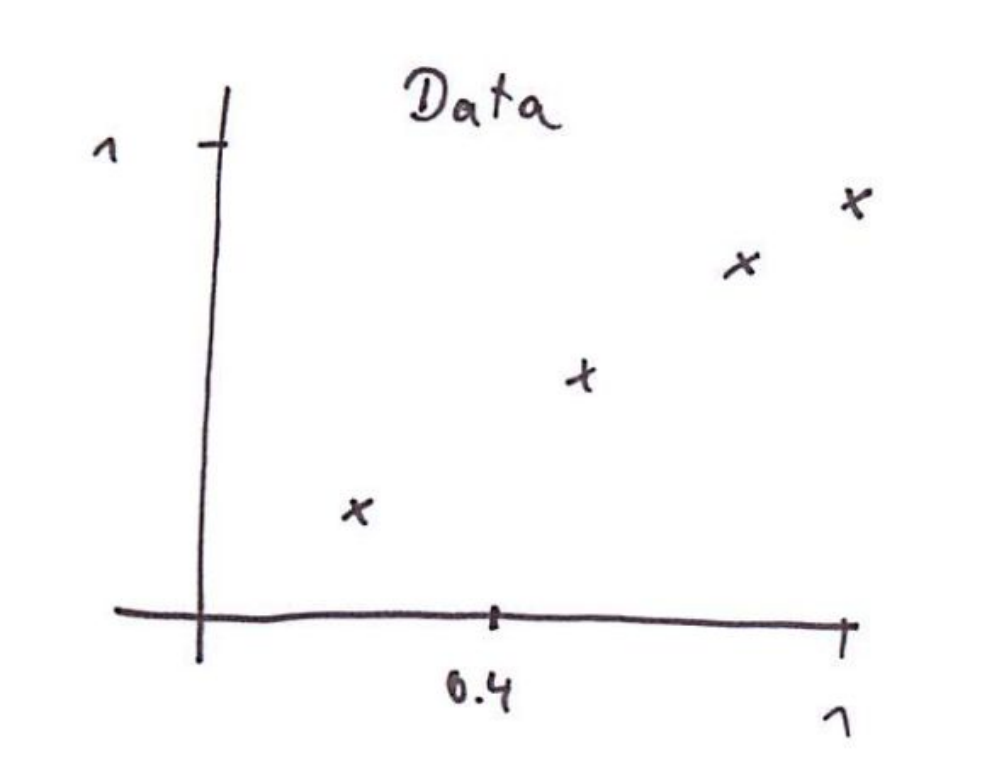
Very important: we can never be sure, our conclusion can be wrong!

Humans do inductive reasoning all the time: we draw uncertain conclusions from our relatively limited experiences.

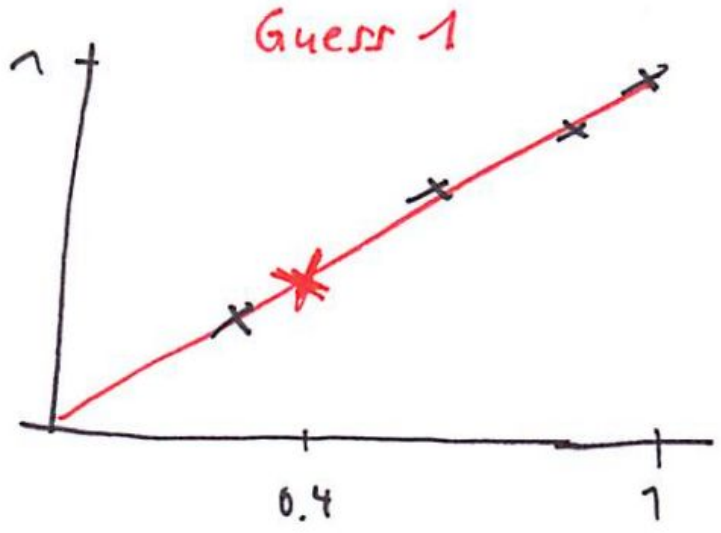
Here comes now our second, more abstract description of what machine learning is:

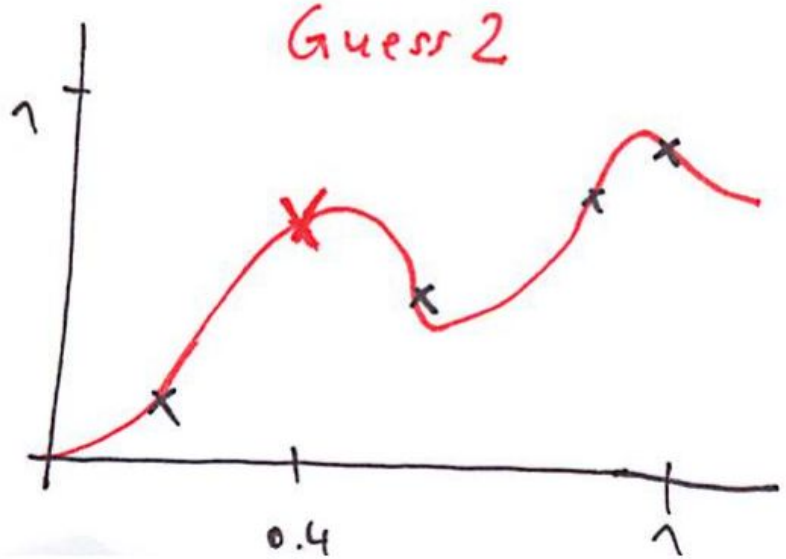
Machine learning tries to automate the process of inductive inference

Why Should Machine Learning Work at All?

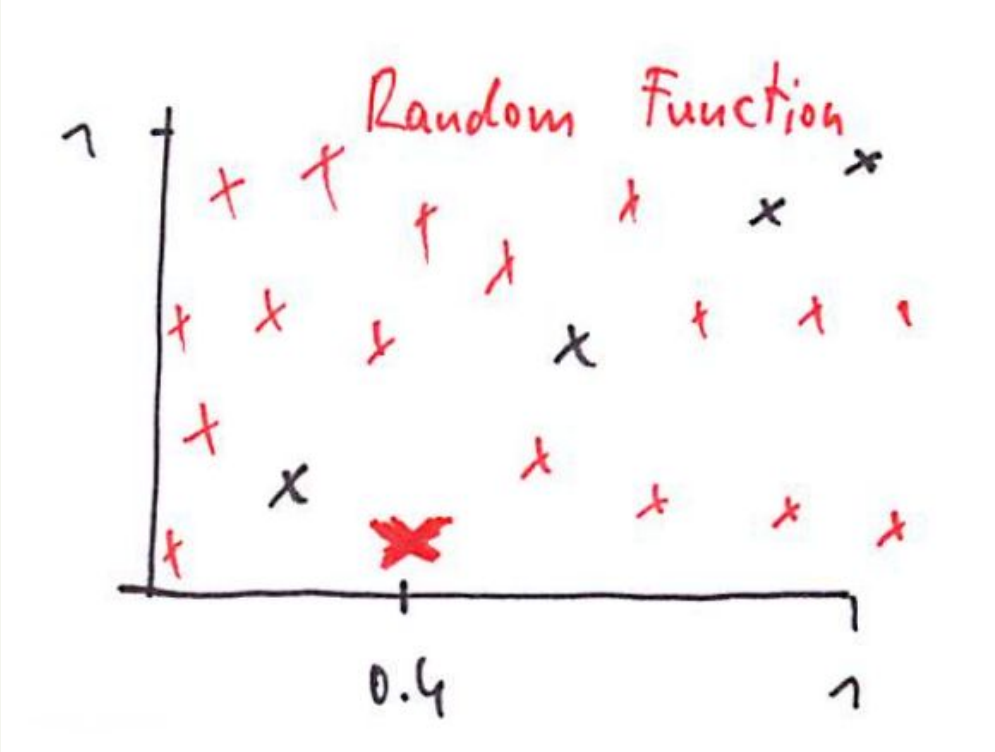


- This image displays the relationship between two variables on the x and y axes.
- The collected data is represented by crosses.
- What do you think is the value of y if x=0.4?





Which one is better?



- Now I tell you that the value of y was generated by a random function.
- What do you think that y is now, if x=0.4?

Consequences:

- We will never know the underlying process which generated data
- We always have to make assumptions about how data was generated. This assumption is called:

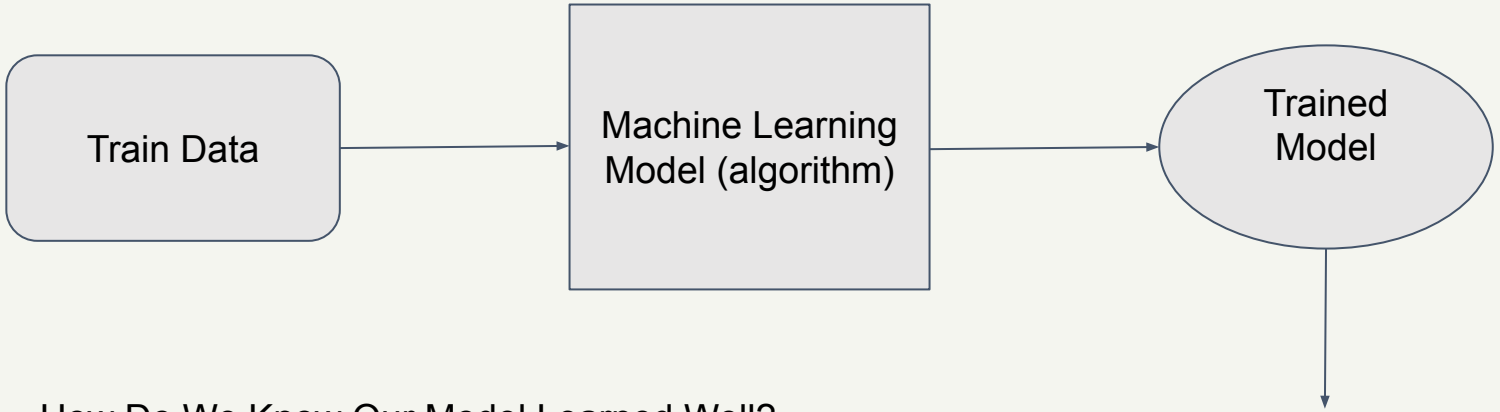
“Inductive bias”

Inductive Bias:

- Output Y has something to do with input X
- Similar inputs lead to similar outputs
- There is a simple relationship or simple rule to generate the output for a given input.
- The function f is simple (but caution, this is not the end of the story!)

Note: Sometimes the inductive bias is clear, but other times it is not. However, rest assured that every machine learning algorithm must have an inductive bias.

Practical Definition of Machine Learning

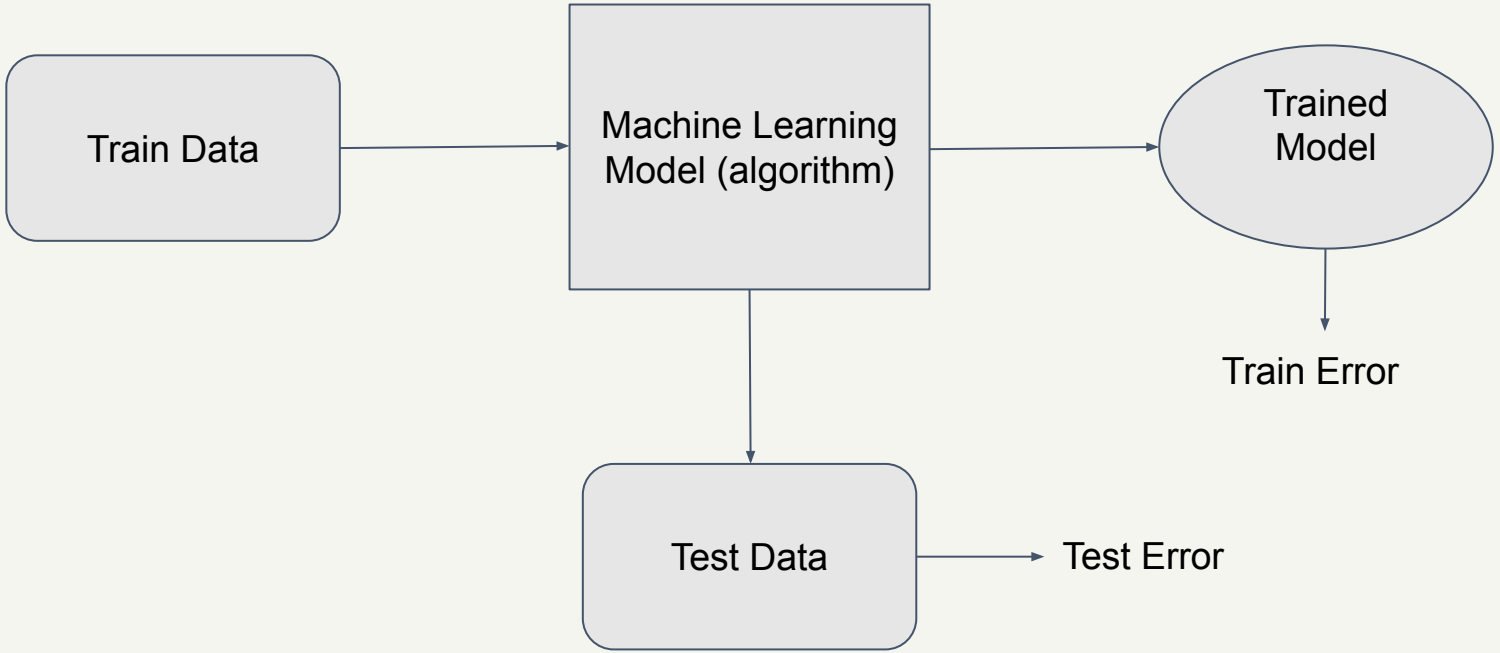


How Do We Know Our Model Learned Well?

- Ensure the model accurately learned the training data
- After optimizing the model, calculate the training error
- The training error indicates how well the model performs on your data

Is That Enough? Does Ensuring My Model Learned the Training Data Well Mean I've Built a Good Model?

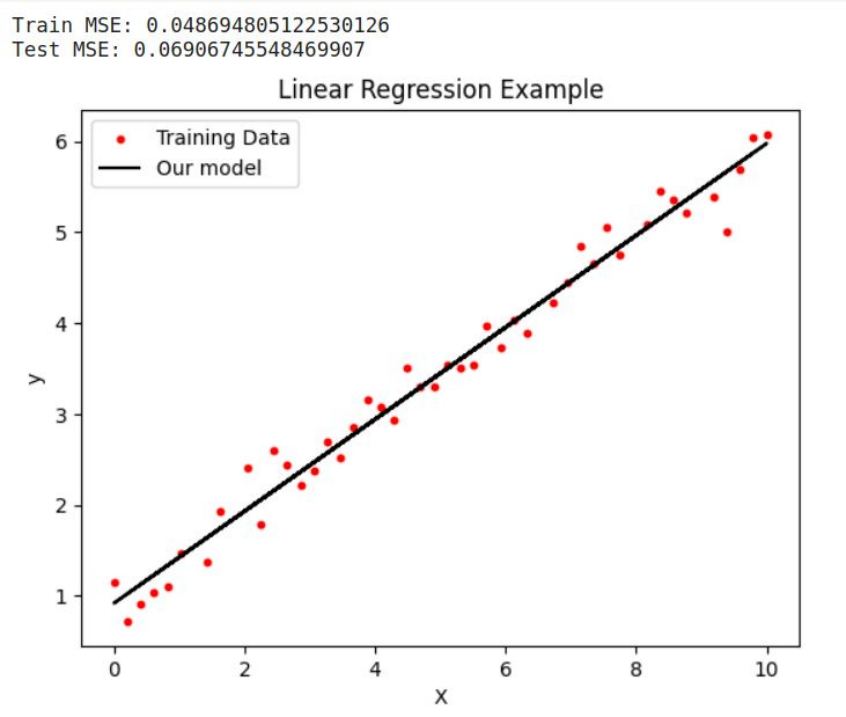
After all, we are showing all the samples (x, y) to the model.



I should test my model on the test data and calculate the test error. Only then can I determine if my model has performed well. **The model must perform well on unseen data points.**

Hands-on Exercise
Bias & Variance

Linear Regression



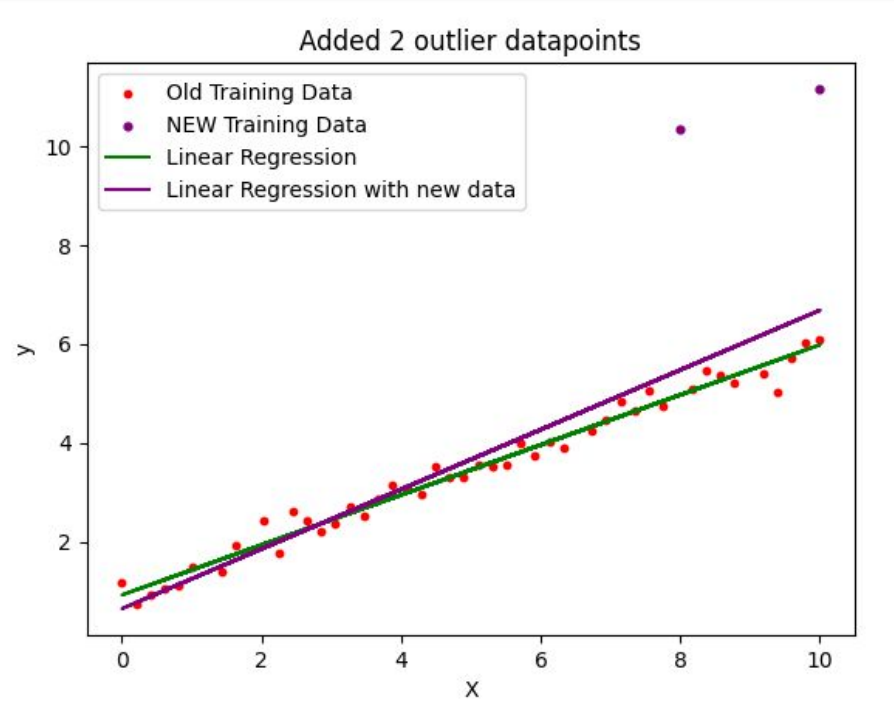
- In the hands-on exercise, we demonstrated how easily a linear regression model can be fitted to a data sample
- We split our data into training and testing datasets
- Next, we calculated the training and testing errors, known as train-MSE and test-MSE, respectively
- In the context of regression, this error is referred to as MSE, which stands for Mean Squared Error

A few remarks

- This model assumes a very simple rule about the nature of the data
- However, in spite of this strong assumption, it **fits data very well** simply because the data is in fact more or less linear in nature
    - Whenever a model fits data very well, we say that the model has a **low Bias**
    - A model's bias indicates how far, on average, the model's predictions are from the actual data points

What is variance?

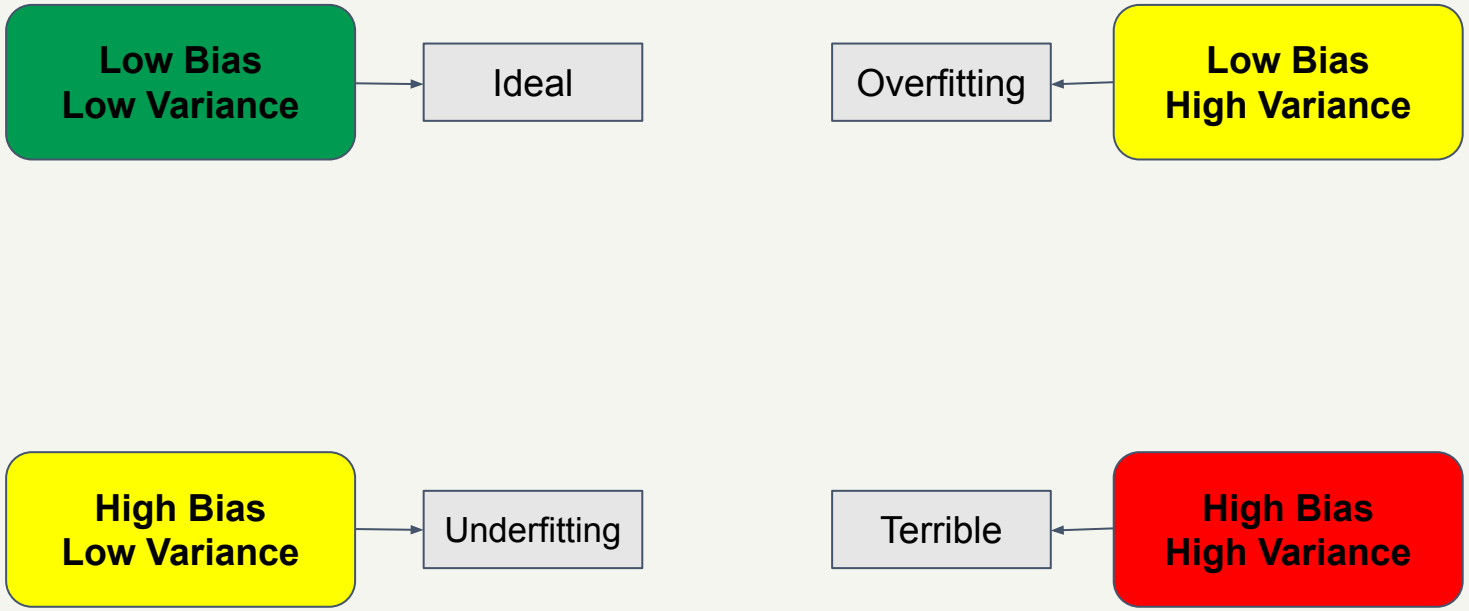
To understand variance, we'll conduct an experiment. We'll add two outlier data points to our existing data sample and observe the effect



Our observations:

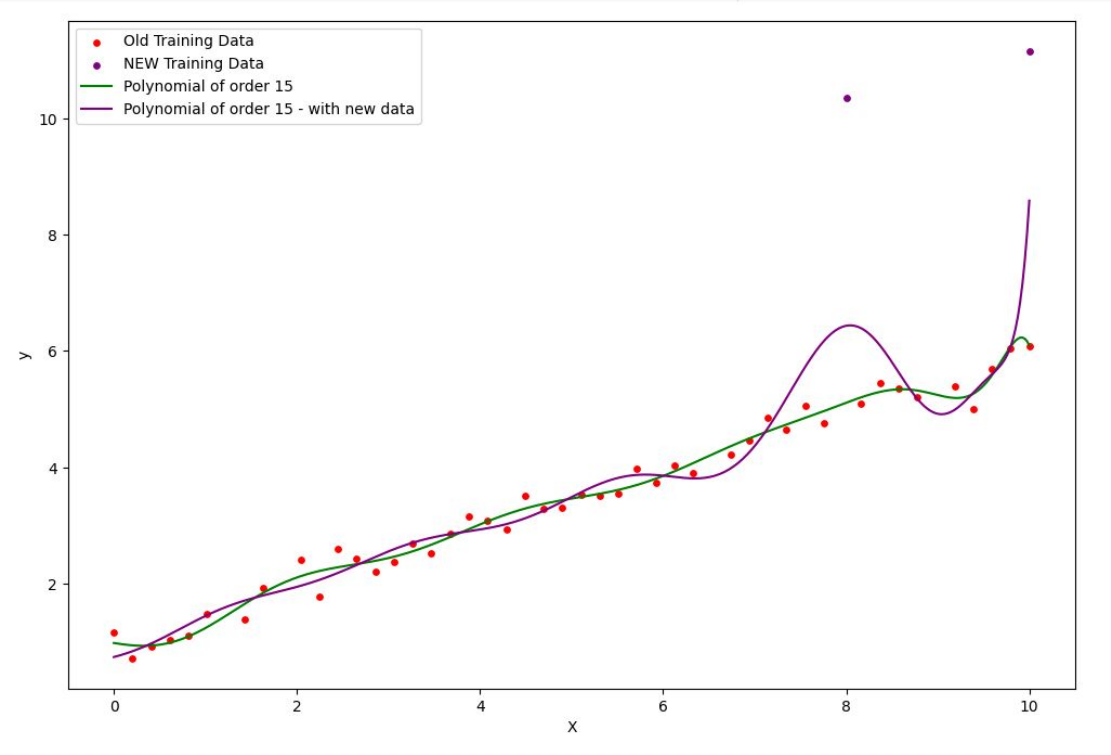
- The slope of the model changed, but despite the significant difference between the outliers and other samples, the variability was small
- Therefore, we say this model has low variance
- Even with the introduction of new, significantly different data points, the model did not change much
- You can imagine that the slope would have remained more or less the same if the outliers were closer to the other data points. This extreme case highlights the model's robustness to outliers

So far, we've developed an understanding of what bias and variance mean. These are two crucial characteristics of every machine learning model we train. Generally, we can encounter four scenarios:



Low Bias, High Variance

- How can we find a model with high variance and low bias?
- Such models are typically sophisticated in nature
- From the category of linear models, a polynomial model with an order 15 for example can be considered as a very sophisticated model

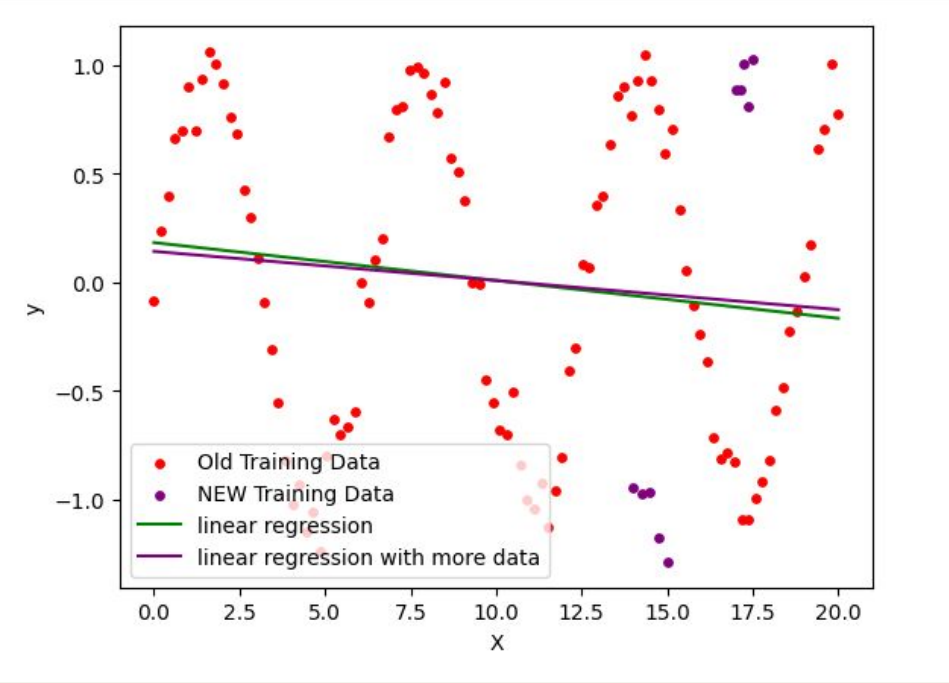


Our observations:

- The model has a low bias
- However, it tries very hard to fit those two outlier points, indicating high variance
- Models with high variance overfit the data
- As a result, they tend to exhibit volatile behavior, especially around noise (outliers)
- This makes them unreliable when tested on unseen data

High Bias, Low Variance

- How can we find a model with high bias and low variance?
- Such models are typically simple in nature
- From the category of linear models, a linear model such as “ax+b” or perhaps a polynomial of order 2 such as “ax^2 + bx + c” can be good candidates

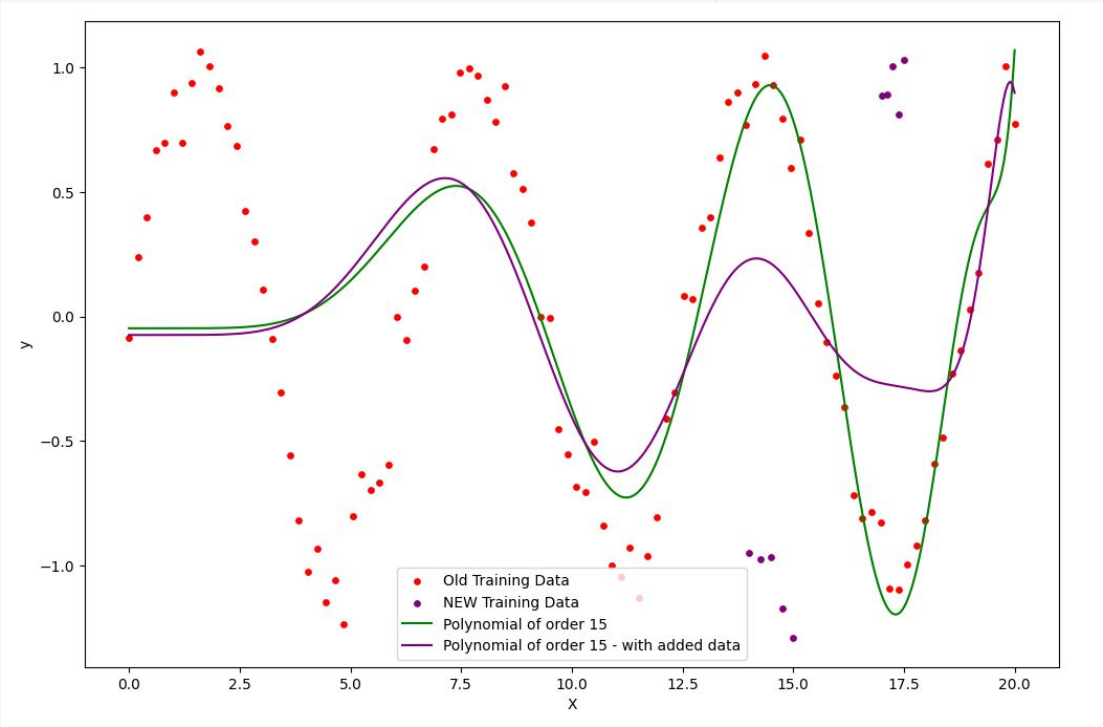


Our observations:

- The model has a high bias
- Despite adding new outliers, the slope of the model slightly changes, indicating a low variance

High Bias, High Variance

- This is the worst scenario
- This rarely happens in ML, however it is still possible
- Typically we have models which overfit the data, thus having a high variance, and low bias



Our observations:

- Nature of data is sinusoidal
- Nevertheless, we wrongly tried to fit a polynomial of order 15 to such a data (bad decision)
- It goes without saying, polynomial family is a bad choice for such a data
- As a result, we not only have a high bias, but also a high variance because the model behavior drastically changes by adding outliers

Parametric
VS
Non-Parametric Models

Parametric models

These models assume a specific form for the underlying data distribution and are defined by a finite set of parameters. Examples include **linear regression** and **logistic regression**. Parametric models are typically easier to interpret and computationally efficient, but they may not perform well if the chosen form does not accurately represent the data

$$\hat{y} = f(x) = ax + b$$

Non-Parametric models

Non-parametric models make fewer assumptions about the data distribution and can adapt to more complex relationships but may require more data and computational resources. Examples are **Decision-Trees** and **KNN algorithm,** which we'll discuss in the future lessons