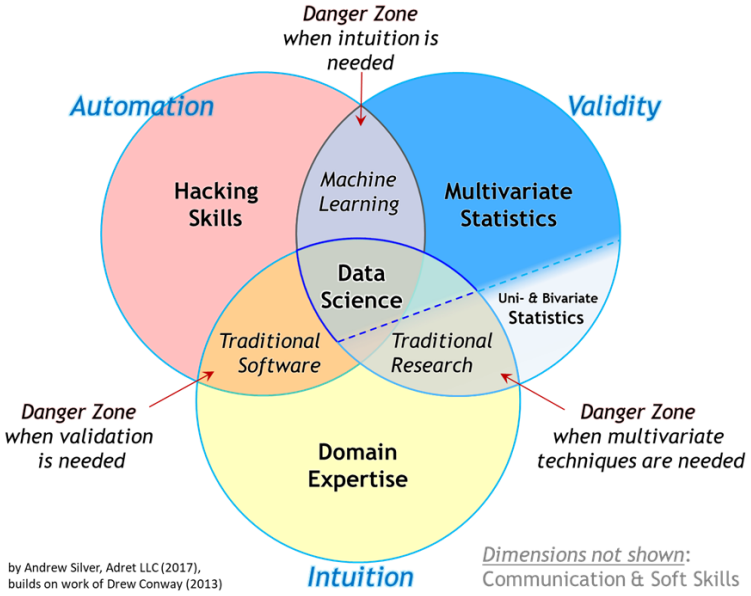
# CodeX: Machine Learning

## Data Science and Practicality

### Definition

💡 Data science is a **dynamic** and **interdisciplinary field** that involves **extracting valuable insights** and knowledge **from data** through a combination of **statistics, mathematics, and computer science**. It's like being a detective for the digital age, using data to solve real-world problems, make informed decisions, and uncover hidden patterns




# Why?

| Reason | Explanation | Example |
| --- | --- | --- |
| **Unleash Your Creativity** | is not just about numbers; it's a canvas for your creativity. Design stunning visualizations, tell compelling stories with data, and explore the exciting world of technology-driven creativity | Spotify Dashboard: Imagine creating your own interactive dashboard that **visualizes your music trends**. Your creativity determines how engaging and visually stunning the map becomes. |
| **Solve Real-World Mysteries** | Become a digital detective! Data science lets you solve real-world puzzles, from predicting trends in social media to uncovering patterns in healthcare data | Dive into healthcare data to **predict disease outbreaks**. By analyzing patterns in patient records, you could help identify potential hotspots, contributing to early intervention and saving lives. |
| **Tech Superpowers Unleashed** | Ever dreamed of having tech superpowers? Learn to wield machine learning, artificial intelligence, and data visualization tools. Data science is | Train a machine learning model to **recognize handwritten digits**. Your code becomes the wizardry that enables a computer to 'see' and interpret |

| Reason | Explanation | Example |
| --- | --- | --- |
| | where technology meets magic, and you get to be the wizard. | what you write, just like magic! |
| **Connect with Your Passion** | Whether you're into fashion, gaming, or the environment, data science connects with your passions. It's not just a skill; it's a ticket to apply your interests in the real world and make a difference. | If you love fashion, use data to analyze trends and predict the next big style. **Your insights could influence designers** and fashionistas, making you a trendsetter in the industry. |
| | | |

### Kickstart your journey now.

You could determine whether your passion is in this field by going through this syllabus. Designed to expose young minds over the current technological field, I have made simplification of useful resources while retaining the information input. I have more complex resources that will give you kickstart for your tertiary studies if you decided to delve more into this magical field that fuse logic and creativity to solve real-world issues. **CodeX** is hoping that this program will be beneficial in both short and long term perspective, to promote sustainability of talent in this competitive and interesting field of studies.

## Practical Applications in the Modern World

https://www.youtube.com/watch?v=uGYJuOyIvzs&list=PLQY2H8rRoyvyLQDomfBj-ptdBGTxAHwV_&index=5&ab_channel=TensorFlow

https://www.youtube.com/watch?v=AveBSbOuOOI&list=PLQY2H8rRoyvyLQDomfBj-ptdBGTxAHwV &index=15&ab channel=Google

https://www.youtube.com/watch?v=L 4BPjLBF4E&ab channel=AIWarehouse

# Introduction to Machine Learning

In this short session, we will focus on supervised machine learning model, the life-cycle of machine learning projects and much more activities to hone our understanding on the solid truth of ML.

## Definition

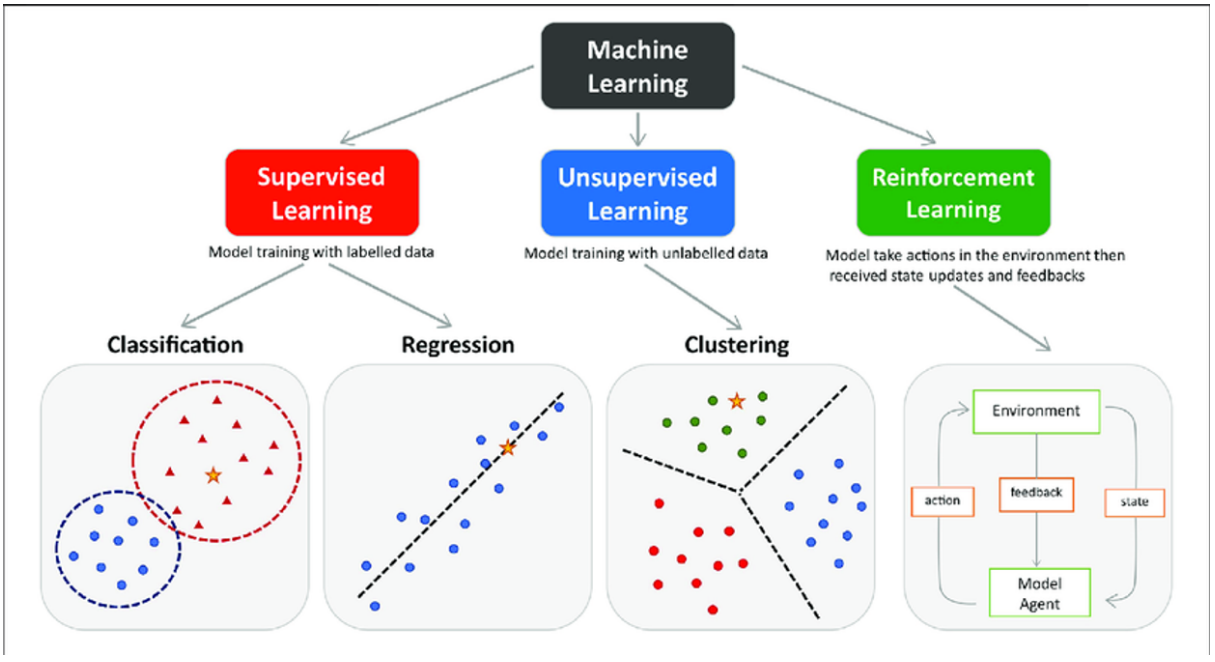
Machine Learning (ML) is a really cool branch of artificial intelligence that focuses on **teaching computers to learn from data** and **make smart decisions** without needing to be told exactly what to do. It's like having a super smart friend who can analyze information and predict things without any explicit instructions. ML is used in a bunch of different areas like finance, healthcare, marketing, and even robotics to solve complex problems and help us make better decisions based on patterns and trends in data.

https://www.youtube.com/watch?v=f uwKZIAeM0&ab channel=OxfordSparks

You won’t believe the amazing things machine learning is doing right now. For starters, smart cars are becoming super cool

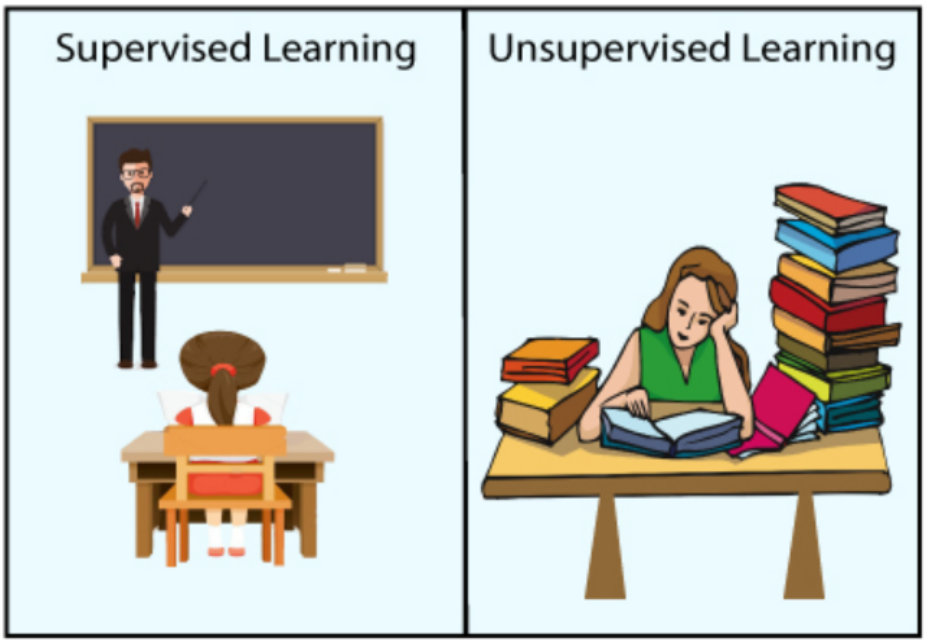
with auto-detection of objects and automatic braking. Imagine cars that can avoid accidents on their own, keeping us safer. And have you ever wondered how web searches work? Machine learning algorithms are behind the scenes, sifting through billions of search results to bring us the most relevant information. It’s like having a super-smart search buddy who knows exactly what we’re looking for and help us find it.

But that’s not all – machine learning also improves our entertainment experiences. Platforms like YouTube use it to recommend personalised content based on what we like to watch. It’s like having our virtual DJ, always playing the perfect tunes or videos just for us. And in healthcare, machine learning is making waves too. It’s helping doctors diagnose diseases more accurately by analysing medical data like patient records and imaging studies.




Based on this diagram, we will focus on classification superviesd learning.

## Understanding supervised and unsupervised learning

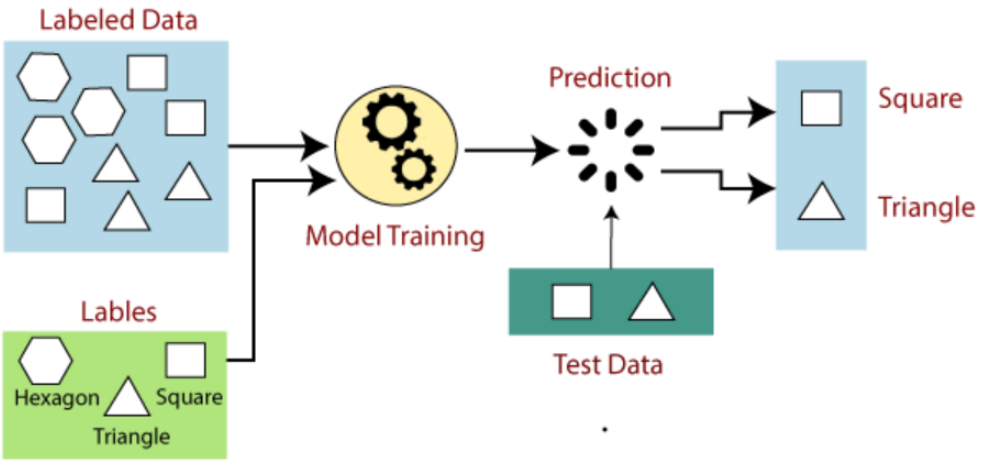



## Supervised Learning

🔑 Supervised learning is a machine learning method in which **models are trained using labeled data**. The labelled data means some input data is already tagged with the correct output. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

In the real-world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc.

The working of Supervised learning can be easily understood by the below example and diagram:




Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.

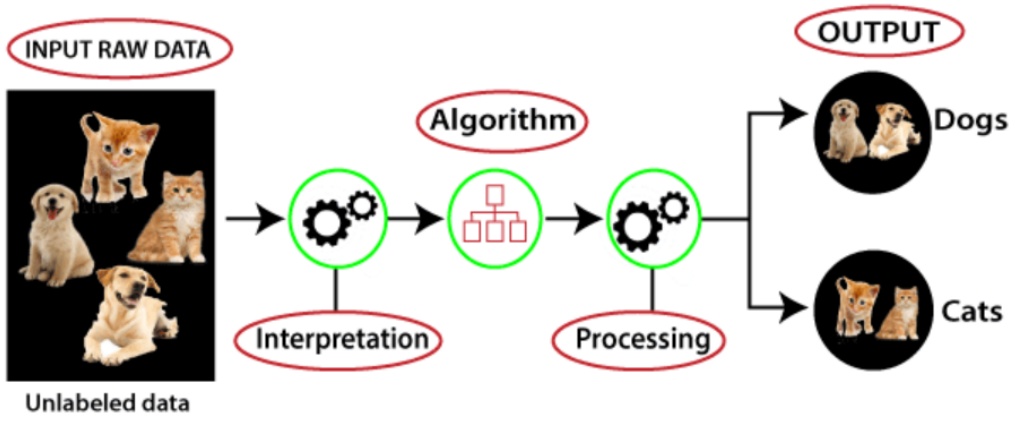
- If the given shape has four sides, and all the sides are equal, then it will be labelled as a **Square**.
- If the given shape has three sides, then it will be labelled as a **triangle**.
- If the given shape has six equal sides then it will be labelled as **hexagon**.

Now, after training, we test our model using the test set, and the task of the model is to identify the given shape. The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.

## Unsupervised Learning

🔑 Unsupervised learning is another machine learning method in which **patterns inferred from the unlabeled input data**. The goal of unsupervised learning is to find the structure and patterns from the input data. Unsupervised learning **does not need any supervision**. Instead, it **finds patterns from the data by its own**.

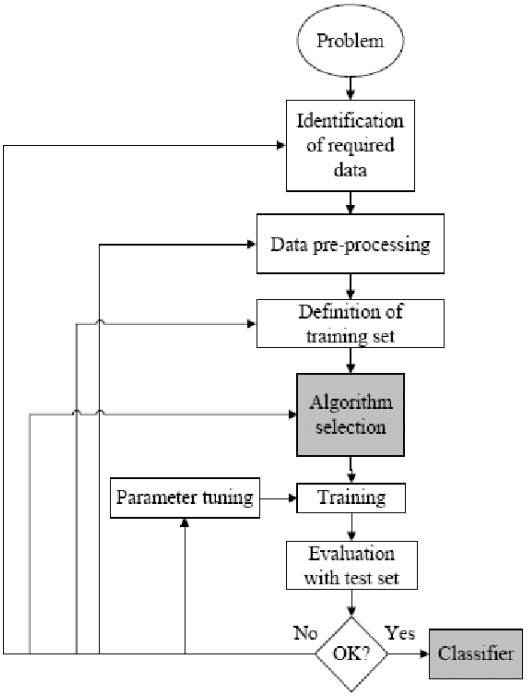
Working of unsupervised learning can be understood by the below diagram:

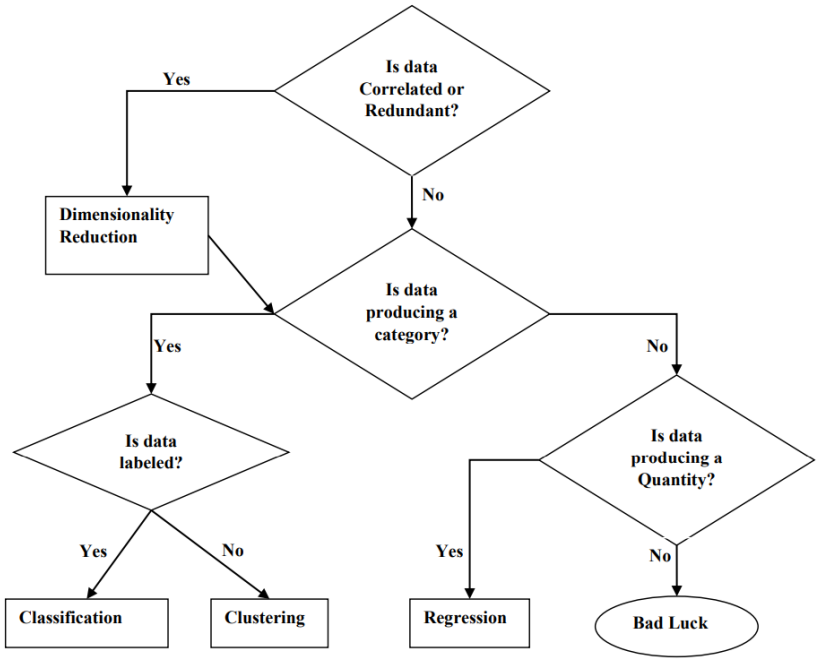



Here, we have taken an unlabeled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabeled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.

Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.

GO to Example first then cross refer with note

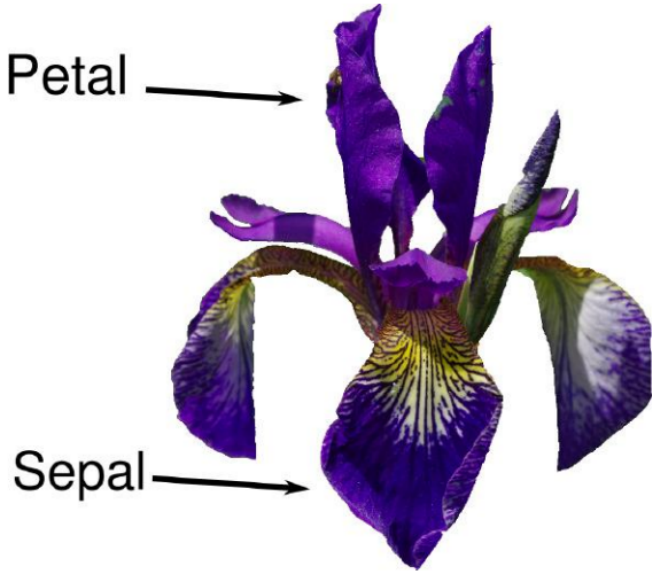



## ‘Hello world’ to ML: Iris (Classification ML)

Let’s assume that a hobby botanist is interested in distinguishing the species of some iris flowers that she has found.

- She has collected some measurements associated with each iris: the length and width of the petals and the length and width of the sepals, all measured in centimeters.
- She also has the measurements of some irises that have been previously identified by an expert botanist as belonging to the species setosa, versicolor, or virginica.
- For these measurements, she can be certain of which species each iris belongs to. Let’s assume that these are the only species our hobby botanist will encounter in the wild.
- Our goal is to build a machine learning model that can learn from the measurements of these irises whose species is known, so that we can predict the species for a new iris.



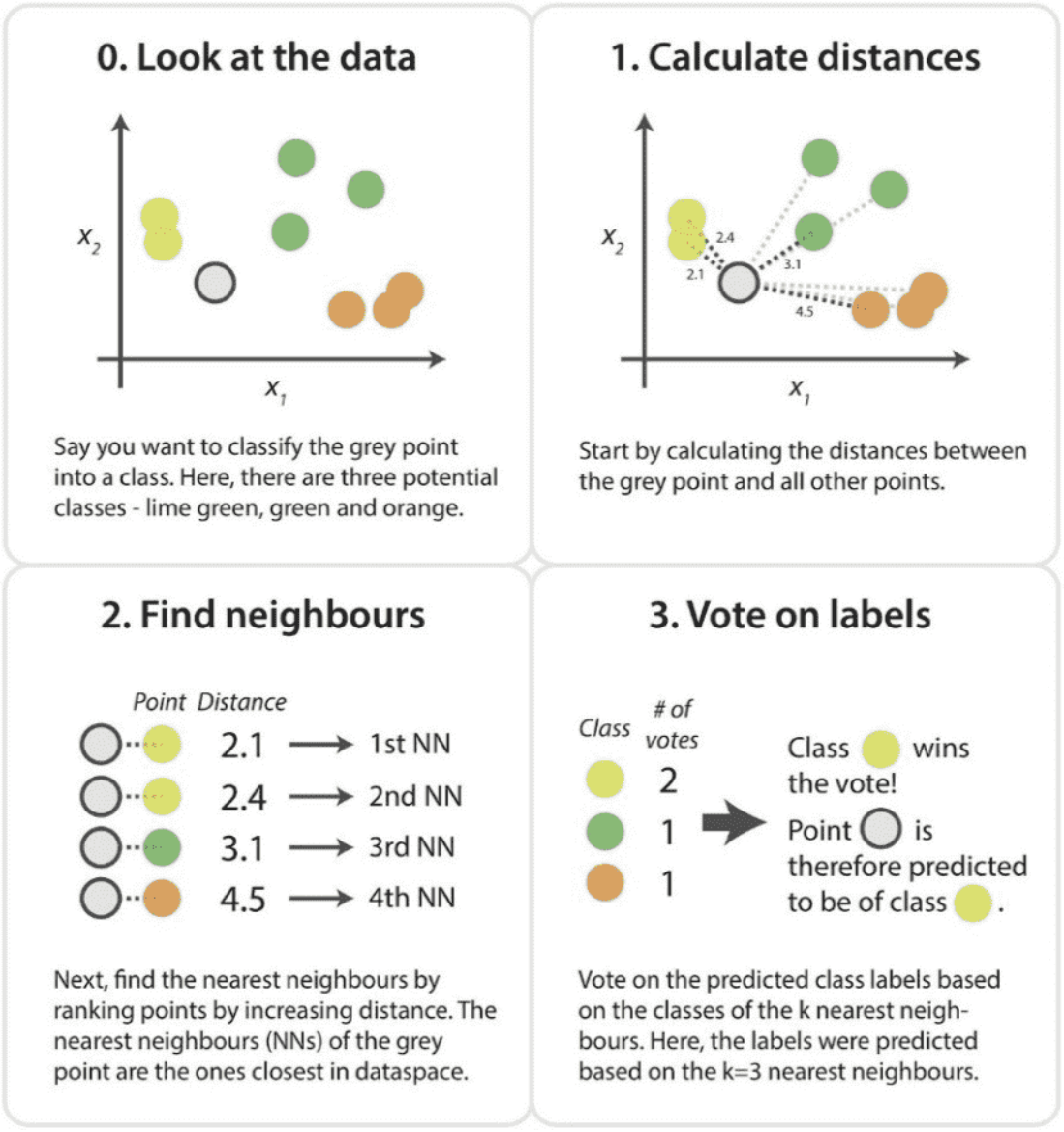

VIEW THE PROCESS HERE:
https://colab.research.google.com/drive/1lkiA-pLayzo8a-KoDAsQ8_kZiIqADxCZ?usp=sharing

# Example of Machine Learning Algorithm : K-nearest neighbor (KNN)

https://www.youtube.com/watch?v=0p0o5cmgLdE&ab_channel=IntuitiveMachineLearning

## K-Nearest Neighbors (KNN) Classification:

1. **Training Phase:**
    - During the training phase, the algorithm "learns" from the labeled examples in the training dataset. Each example consists of features (attributes) and their corresponding class labels.
2. **Distance Metric:**
    - The algorithm uses a distance metric (commonly Euclidean distance) to measure the distance between the new, unlabeled data point and each point in the training dataset.
3. **Neighborhood Identification:**
    - KNN identifies the 'k' training examples that are closest to the new data point in terms of distance. These examples are the "neighbors."
4. **Majority Voting:**
    - For classification, KNN assigns the class label to the new data point based on the majority class among its k-nearest neighbors. It's like asking the neighbors for their opinion, and the new point adopts the most popular class.




## Key Concepts:

- **K (Number of Neighbors):**
  - The value of 'k' is a crucial parameter. A larger 'k' considers more neighbors, potentially making the algorithm more robust to noise but may oversmooth the decision boundary. *‘No free lunch theorem’*
- **Decision Boundary:**
  - KNN doesn't explicitly create a model with a predefined structure. The decision boundary is determined by the distribution of data points in the feature space.

## Example Analogy:

Imagine you have a map with houses labeled 'A' and 'B.' You want to determine the label for a new house but don't know it yet. You look at the closest three houses on the map (k=3) and notice two of them are labeled 'A' and one is labeled 'B.' Based on the majority, you predict that the new house is also likely to be labeled 'A.'

## Summary:

The k value in the k-NN algorithm defines how many neighbors will be checked to determine the classification of a specific query point. For example, if **k=1**, the instance will be assigned to the same class as its single nearest neighbor. Defining k can be a balancing act as different values can lead to overfitting or underfitting. **Lower values of k** can have **high variance**, but **low bias**, and *larger values of k* may lead to *high bias* and *lower variance.* The choice of k will largely depend on the input data as data with more outliers or noise will likely perform better with higher values of k. Overall, it is recommended to have an odd number for k to avoid ties in classification, and cross-validation tactics can help you choose the optimal k for your dataset.

# Bias and Variance

# Cheat Sheet – Bias-Variance Tradeoff

**What is Bias?** $bias = \mathbb{E}[f'(x)] - f(x)$

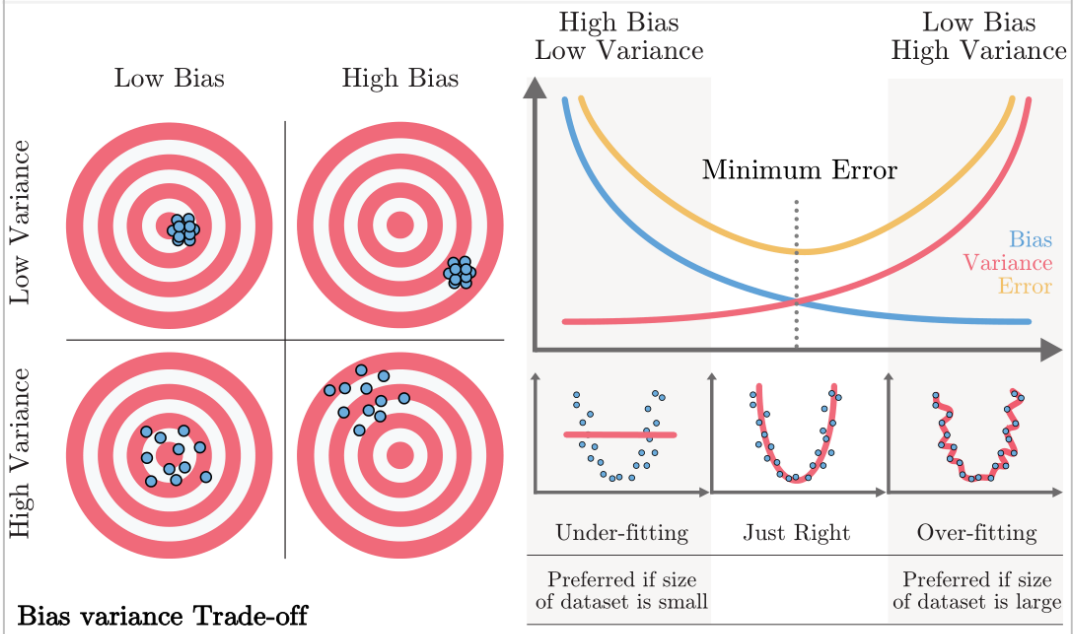
- Error between average model prediction and ground truth
- The bias of the estimated function tells us the capacity of the underlying model to predict the values

**What is Variance?** $variance = \mathbb{E}\left[(f'(x) - \mathbb{E}[f'(x)])^2\right]$

- Average variability in the model prediction for the given dataset
- The variance of the estimated function tells you how much the function can adjust to the change in the dataset

**High Bias**
- → Overly-simplified Model
- → Under-fitting
- → High error on both test and train data

**High Variance**
- → Overly-complex Model
- → Over-fitting
- → Low error on train data and high on test
- → Starts modelling the noise in the input




**Bias variance Trade-off**

- Increasing bias reduces variance and vice-versa
- Error = bias$^2$ + variance +irreducible error
- The best model is where the error is reduced.
- Compromise between bias and variance

Source: https://www.cheatsheets.aqeel-anwar.com



# Imbalanced Data Classification

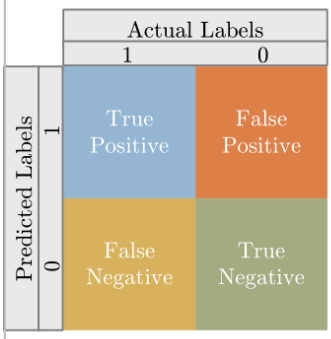
# Cheat Sheet – Imbalanced Data in Classification


Classifier that always predicts label blue yields prediction accuracy of 90%

## Accuracy doesn't always give the correct insight about your trained model

| | | |
|---|---|---|
| **Accuracy:** %age correct prediction | Correct prediction over total predictions | One value for entire network |
| **Precision:** Exactness of model | From the detected cats, how many were actually cats | Each class/label has a value |
| **Recall:** Completeness of model | Correctly detected cats over total cats | Each class/label has a value |
| **F1 Score:** Combines Precision/Recall | Harmonic mean of Precision and Recall | Each class/label has a value |

## Performance metrics associated with Class 1




(Is your prediction correct?) (What did you predict)

True Negative

(Your prediction is correct) (You predicted 0)

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1 score} = 2x\frac{(\text{Prec x Rec})}{(\text{Prec} + \text{Rec})}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

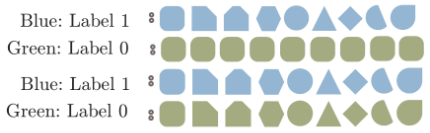
$$\text{False +ve rate} = \frac{FP}{TN + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

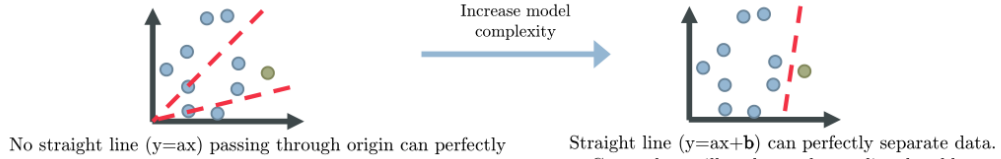
$$\text{Recall, Sensitivity} = \text{True +ve rate} = \frac{TP}{TP + FN}$$

## Possible solutions

1. **Data Replication:** Replicate the available data until the number of samples are comparable
2. **Synthetic Data:** Images: Rotate, dilate, crop, add noise to existing input images and create new data
3. **Modified Loss:** Modify the loss to reflect greater error when misclassifying smaller sample set
4. **Change the algorithm:** Increase the model/algorithm complexity so that the two classes are perfectly separable (Con: Overfitting)




$$loss = a * loss_{green} + b * loss_{blue} \quad a > b$$




No straight line (y=ax) passing through origin can perfectly separate data. **Best solution:** line y=0, predict all labels blue

Straight line (y=ax+**b**) can perfectly separate data. Green class will no longer be predicted as blue

Source: https://www.cheatsheets.aqeel-anwar.com



# Overfitting vs Underfitting

🔥 **Underfitting** occurs when a *model is too simple* to capture the underlying patterns in the data. It fails to learn the training data and performs poorly on both the training and testing sets.

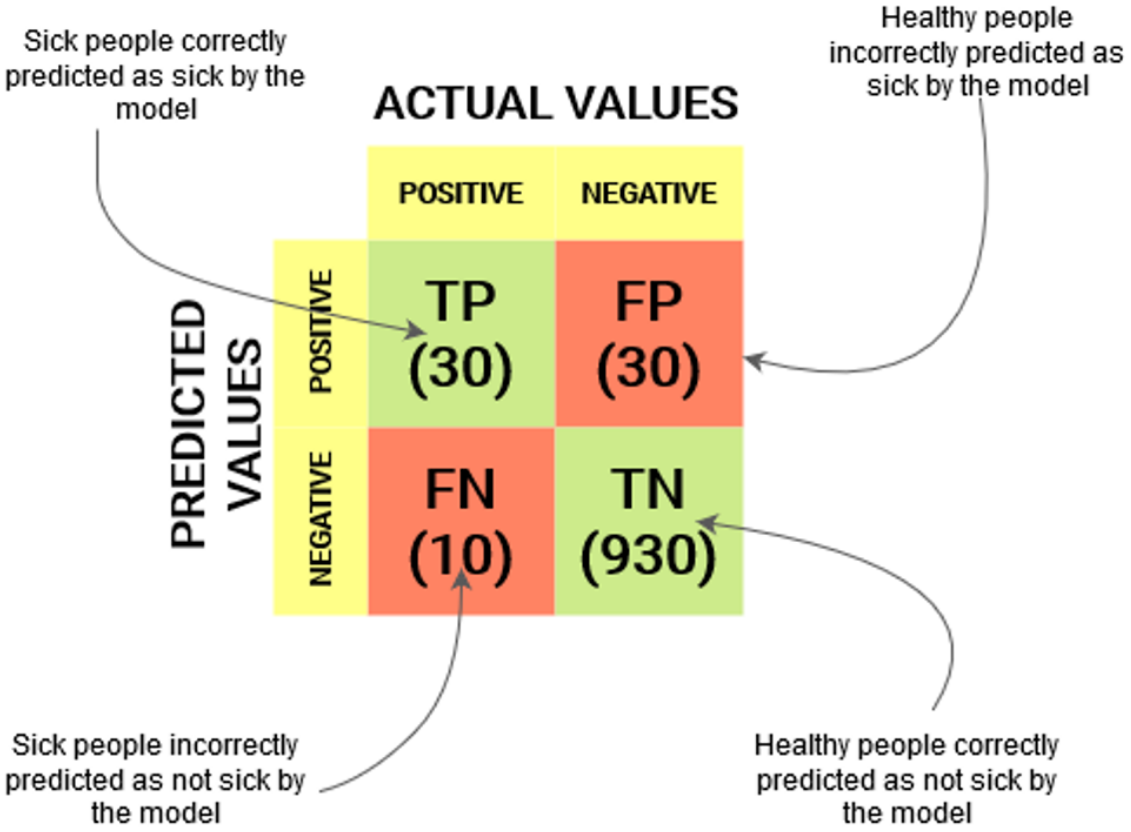
**Overfitting** occurs when a *model is too complex* and learns the training data too well, including its noise. As a result, the model performs well on the training set but poorly on unseen data.

# Classification ML Evaluation metrics

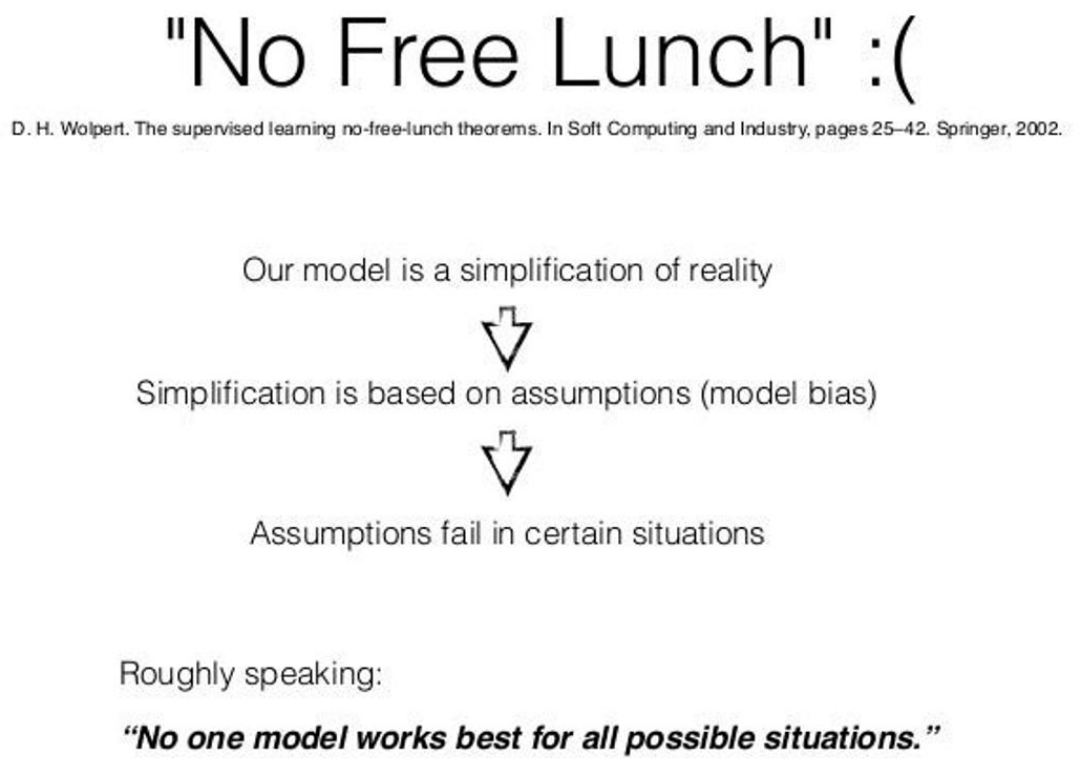
## Accuracy

💡 It represents the percentage of correctly predicted instances out of the total instances. Accuracy is a straightforward measure and is often used for balanced datasets where the classes are approximately equally distributed. However, it might not be the best metric for imbalanced datasets, where one class significantly outnumbers the other

## Confusion Matrix (Classification ML)




## No Free lunch theorem




💡 When it comes to data modeling, the beginner’s question is always, "what is the best machine learning algorithm?" To this the beginner must learn, the No Free Lunch Theorem (NFLT) of Machine Learning. In short, NFLT states, **there is no super algorithm**, that works best in all situations, for all datasets. So the best approach is to try multiple MLAs, tune them, and compare them for your specific scenario

## Tweaking Parameter

💡 Tweaking parameters in machine learning is an essential part of the model development process. The process involves adjusting hyperparameters to improve the performance of a machine learning model. Hyperparameters are configuration settings external to the model itself that cannot be learned from training data but can significantly impact the model's performance

## INTERESTED IN LEARNING MORE???????

Scan here to access the completed version of syllabus