هکاتون هوش مصنوعی ایرانسل لبز

در چارچوب طرح شهید بابایی بنیاد ملی نخبگان

Quera

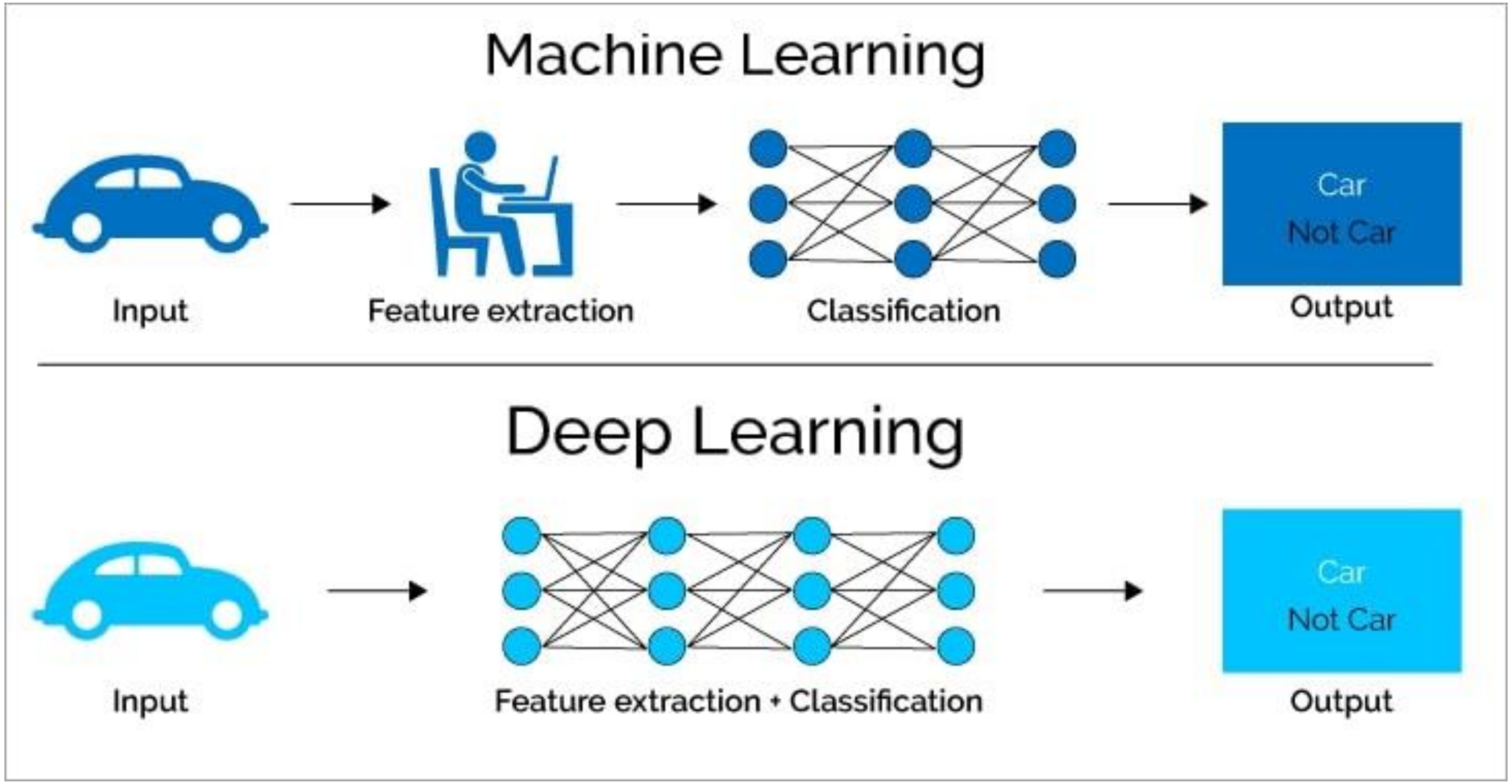
What is Machine Learning?

- Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.
- Machine learning involves the development of algorithms that enable computers to identify patterns and relationships within data, ultimately enabling them to make predictions or decisions based on new, unseen data.
- Machine learning is a subfield of artificial intelligence that focuses on developing algorithms and models that allow computers to perform tasks by learning from data, rather than relying on explicit programming.

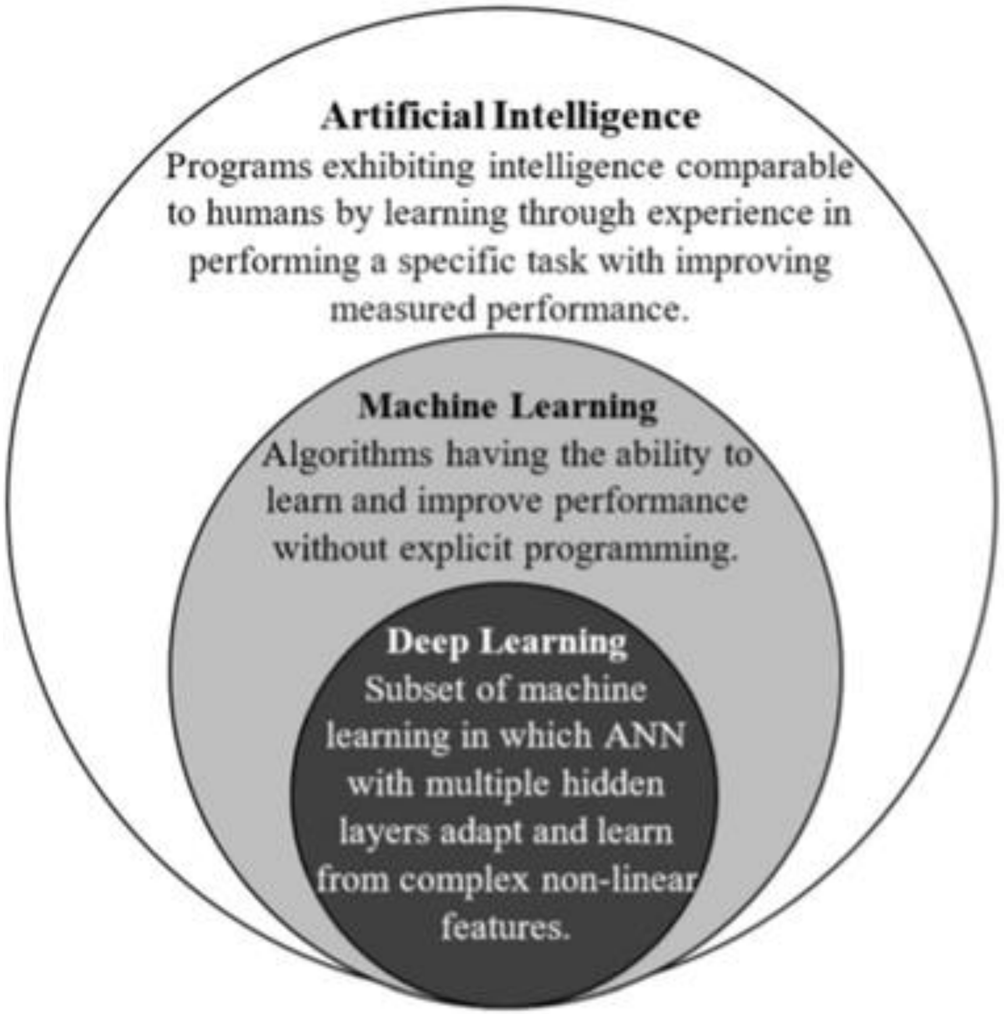
What is Artificial Intelligence

- **Narrow/Weak AI:**
  Image Recognition
  Recommender Systems
  Speech Recognition
  Fraud Detection
  Image Analysis
- **Strong AI**
  Human-Level Conversations
  Creative Arts
  Autonomous Decision-Making
  Adaptation and Learning
  Problem Solving
  Scientific Discovery

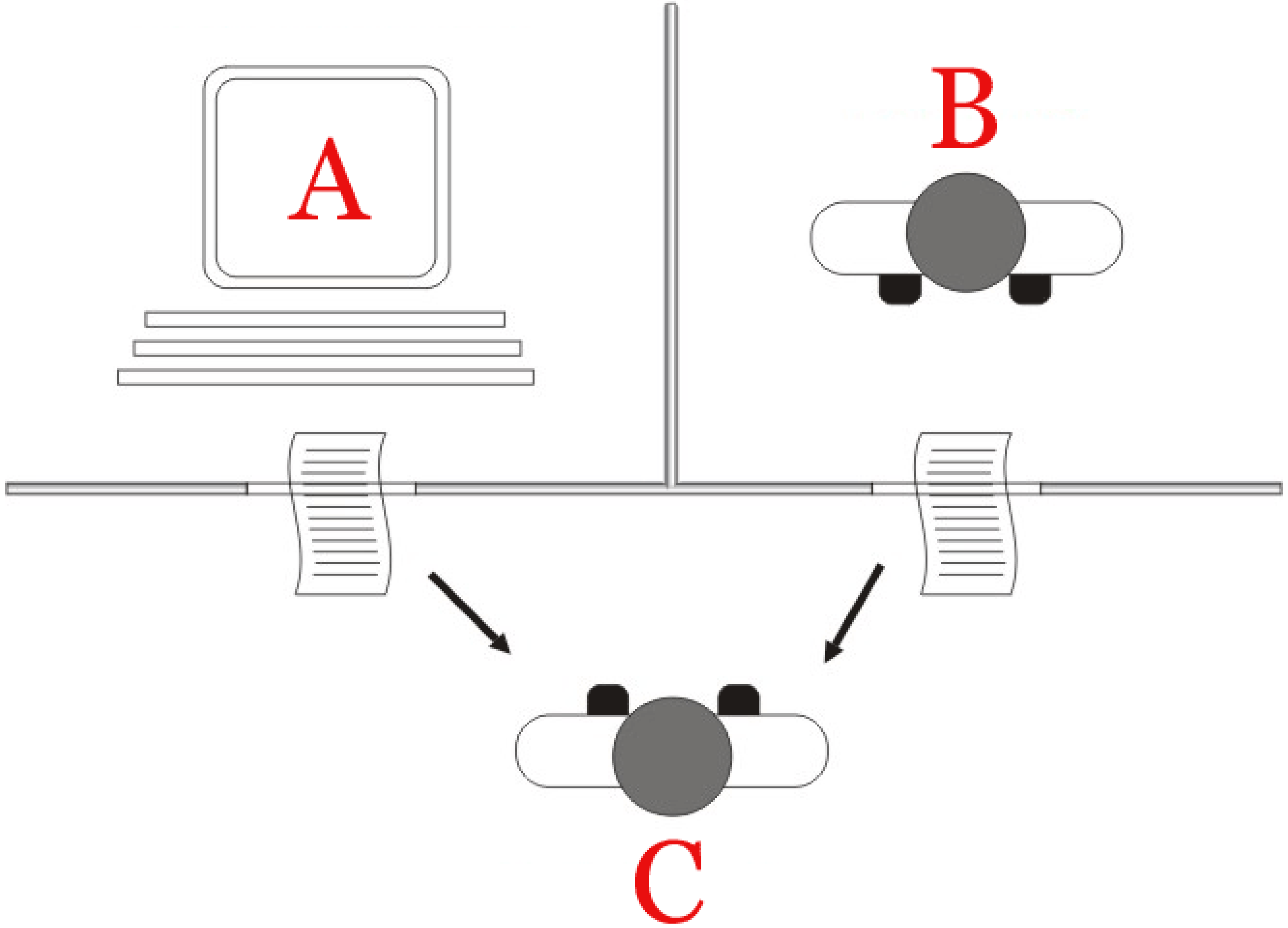
What is Deep Learning?



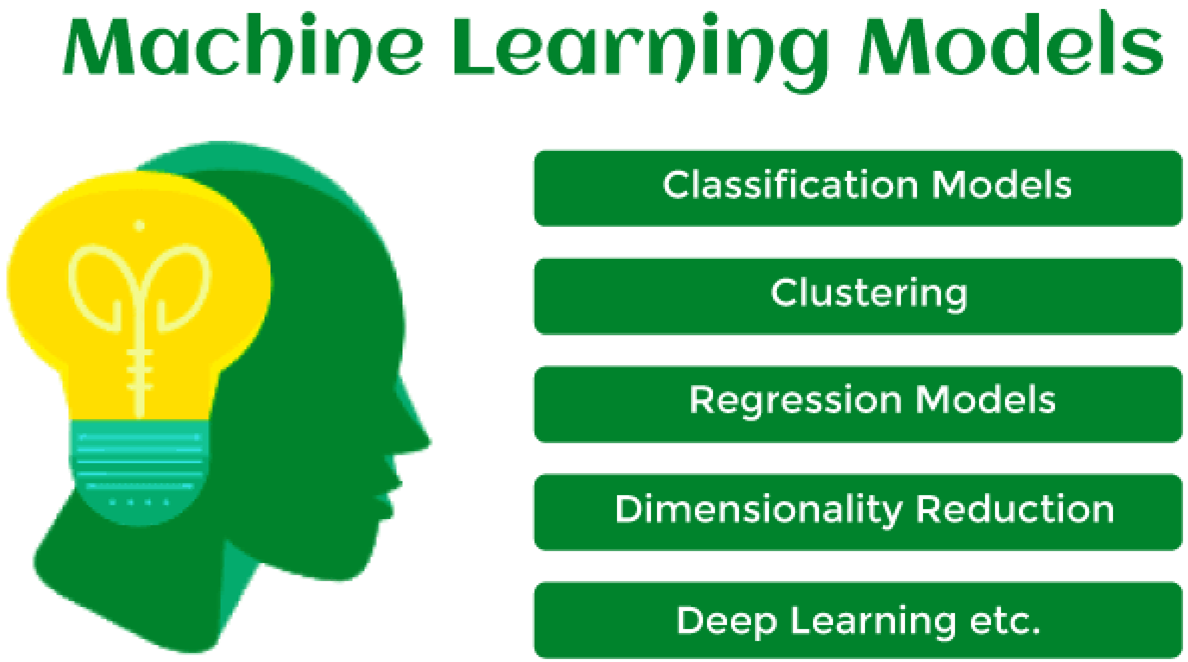
AI vs ML vs DL


Artificial Intelligence
Programs exhibiting intelligence comparable to humans by learning through experience in performing a specific task with improving measured performance.
Machine Learning
Algorithms having the ability to learn and improve performance without explicit programming.
Deep Learning
Subset of machine learning in which ANN with multiple hidden layers adapt and learn from complex non-linear features.

Turing Test



What is a MODEL?



How we learn?

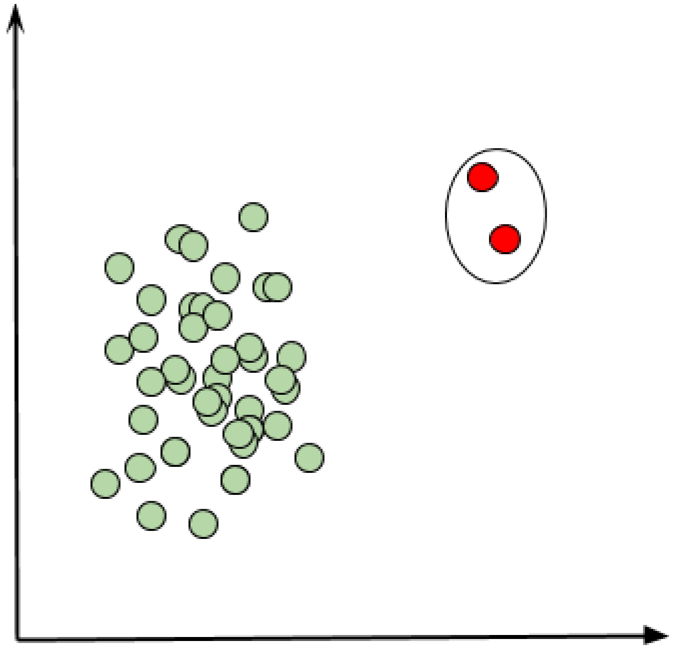
When we are two years old - Clustering

Clustering Example

- Customer Segmentation

Clustering Example

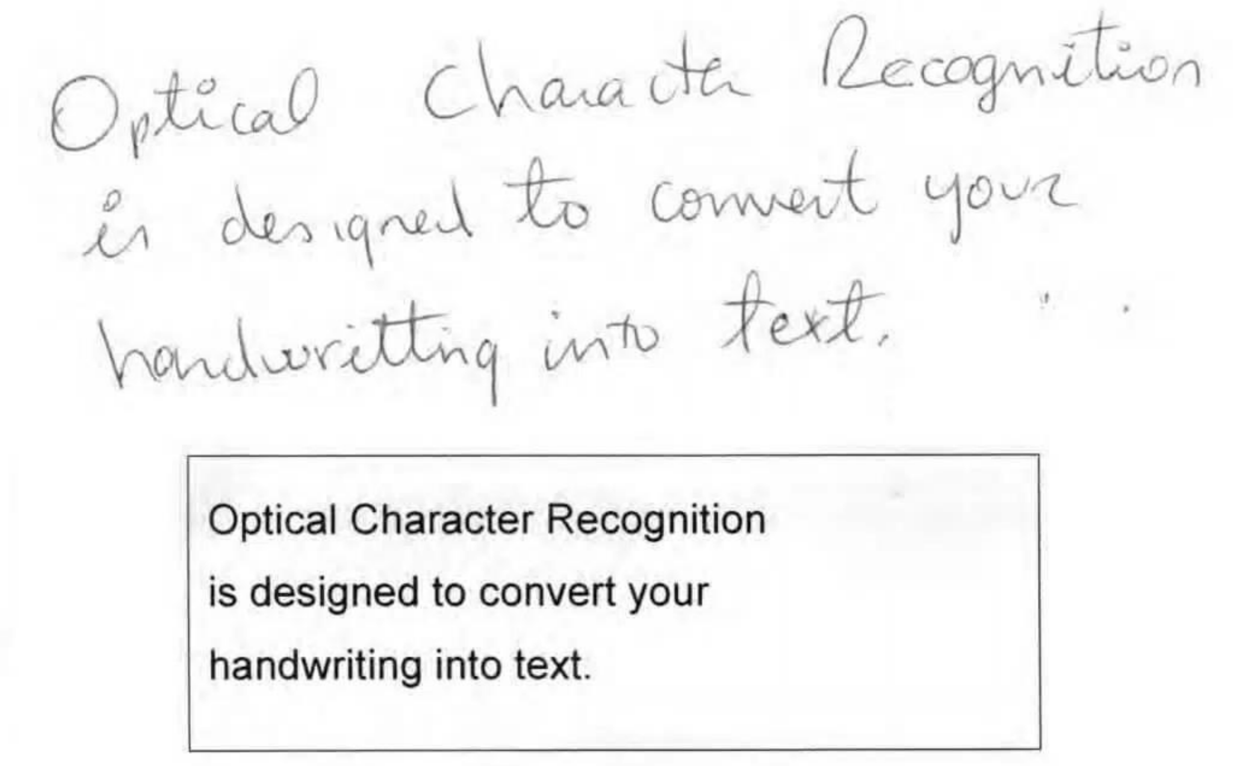
- Anomaly Detection



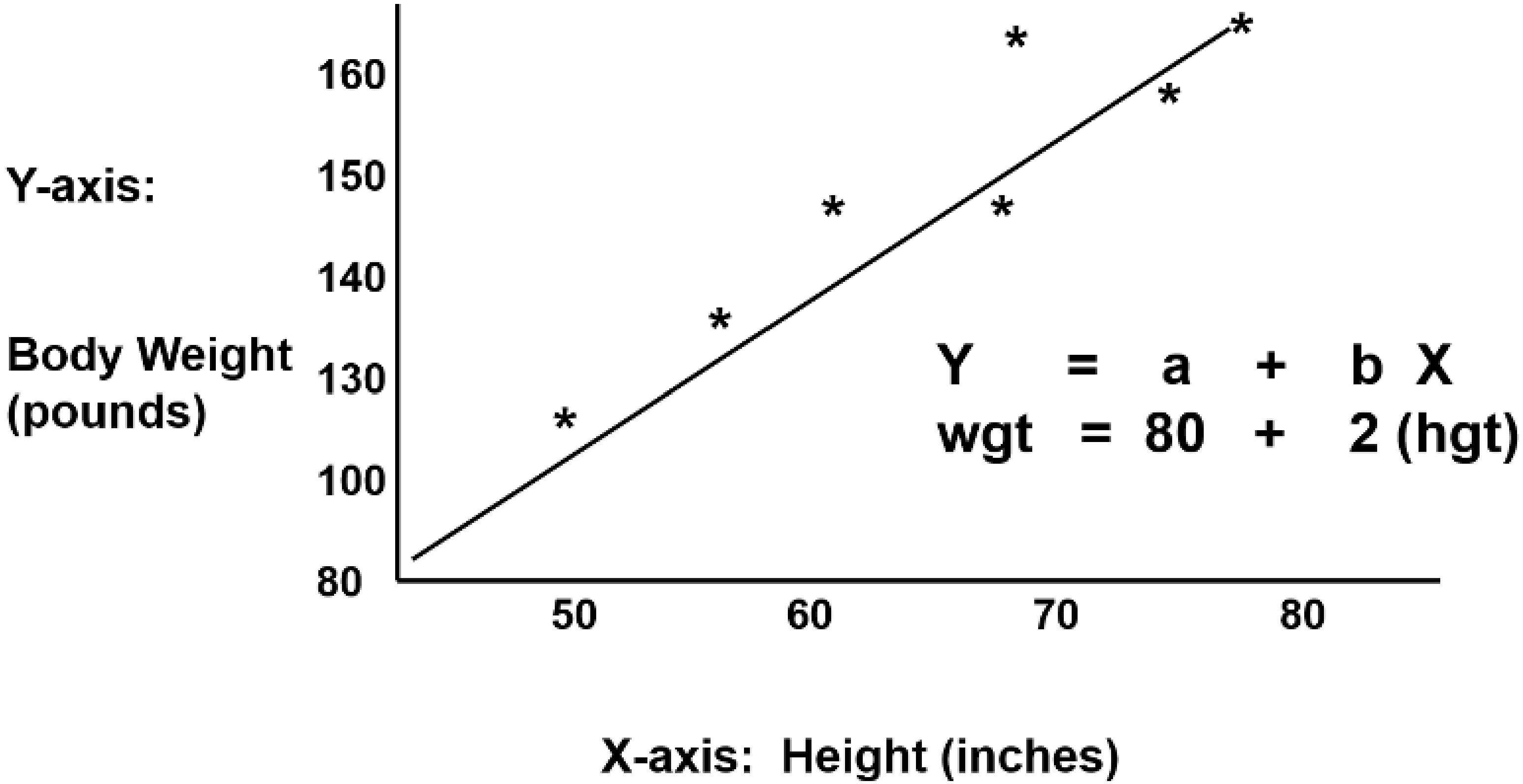
When we are 4 years old - Classification

Classification Example

- Optical Character Recognition



When we are 16 years old - Regression



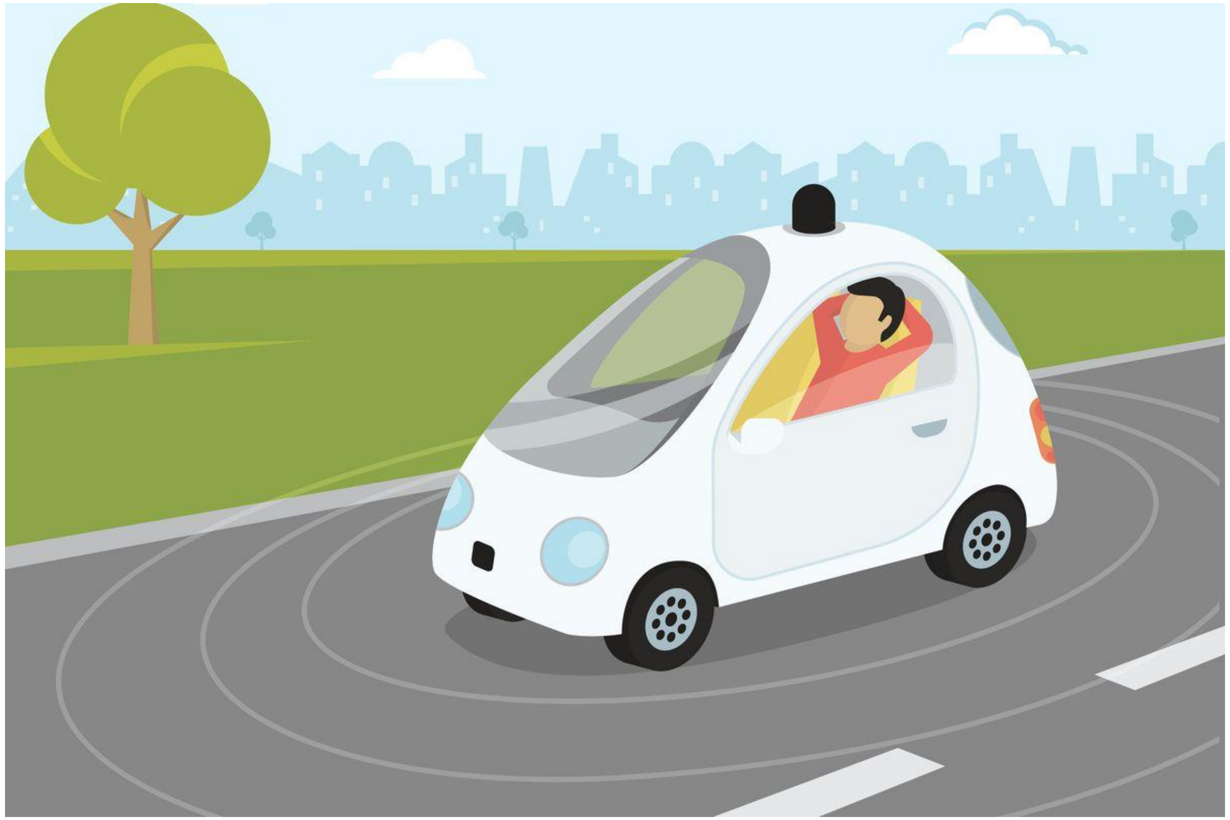
Regression Example vs Time-Series

Regression	Time Series
1. In Regression it does not matter if we reshuffle the data.	1. Time Series consists time from before to the latest and they cannot be shuffled.
2. In Regression data points are independent.	2. In Time Series there is strong correlation between successive values.
3. Regression model predicts only on the basis of the values given.	3. Time Series not only depends on the values given, but also on depends on the sequence in which the values are given.

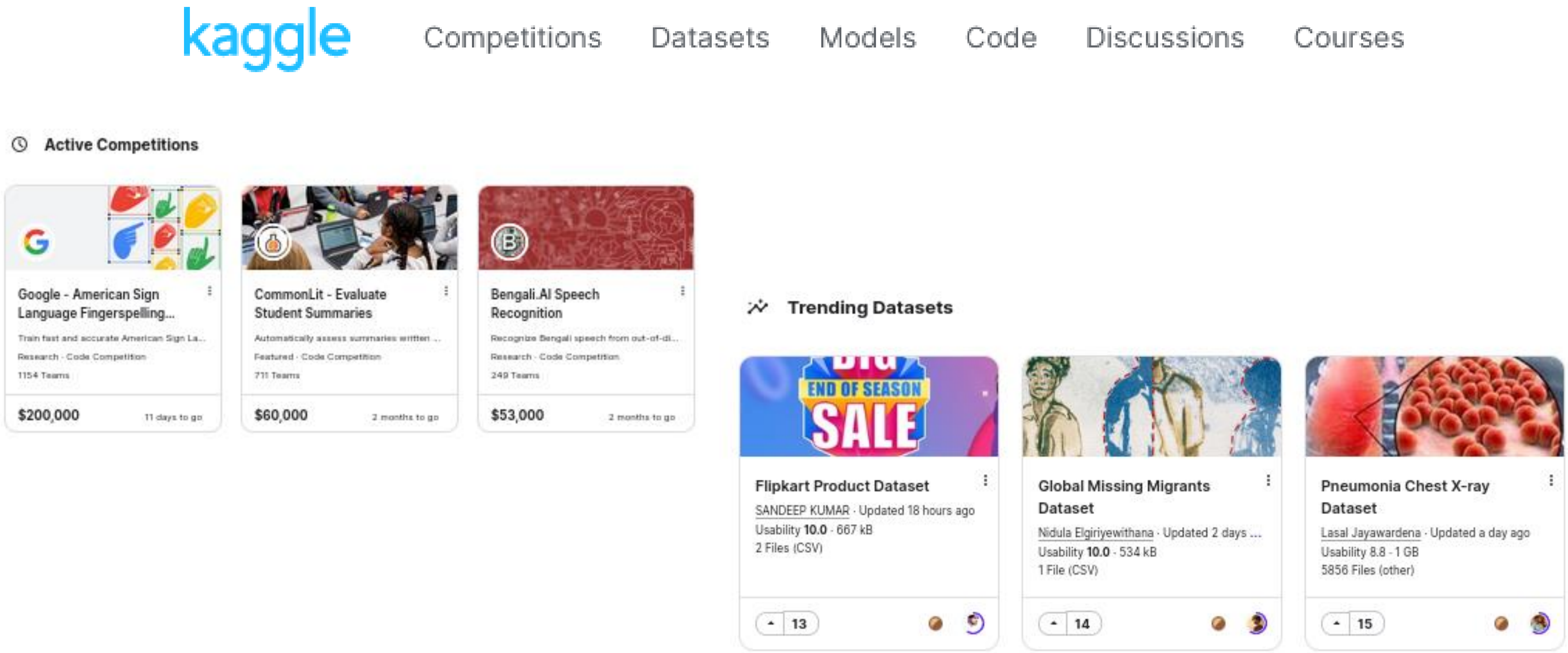
What about Basketball? – Reinforcement Learning

Reinforcement Learning Example

- Self-driving cars



Kaggle.com


kaggle
Competitions
Datasets
Models
Code
Discussions
Courses
Active Competitions
Google - American Sign Language Fingerspelling...
Train fast and accurate American Sign La...
Research · Code Competition
1154 Teams
$200,000
11 days to go
CommonLit - Evaluate Student Summaries
Automatically assess summaries written ...
Featured · Code Competition
711 Teams
$60,000
2 months to go
Bengali.AI Speech Recognition
Recognize Bengali speech from out-of-di...
Research · Code Competition
249 Teams
$53,000
2 months to go
Trending Datasets
END OF SEASON
SALE
Flipkart Product Dataset
SANDEEP KUMAR · Updated 18 hours ago
Usability 10.0 · 667 kB
2 Files (CSV)
13
Global Missing Migrants Dataset
Nidula Elgiriyewithana · Updated 2 days ...
Usability 10.0 · 534 kB
1 File (CSV)
14
Pneumonia Chest X-ray Dataset
Lasal Jayawardena · Updated a day ago
Usability 8.8 · 1 GB
5856 Files (other)
15

IDEs

PyCharm

Visual Studio Code

Sublime Text

Vim

GNU Emacs

Spyder

Atom

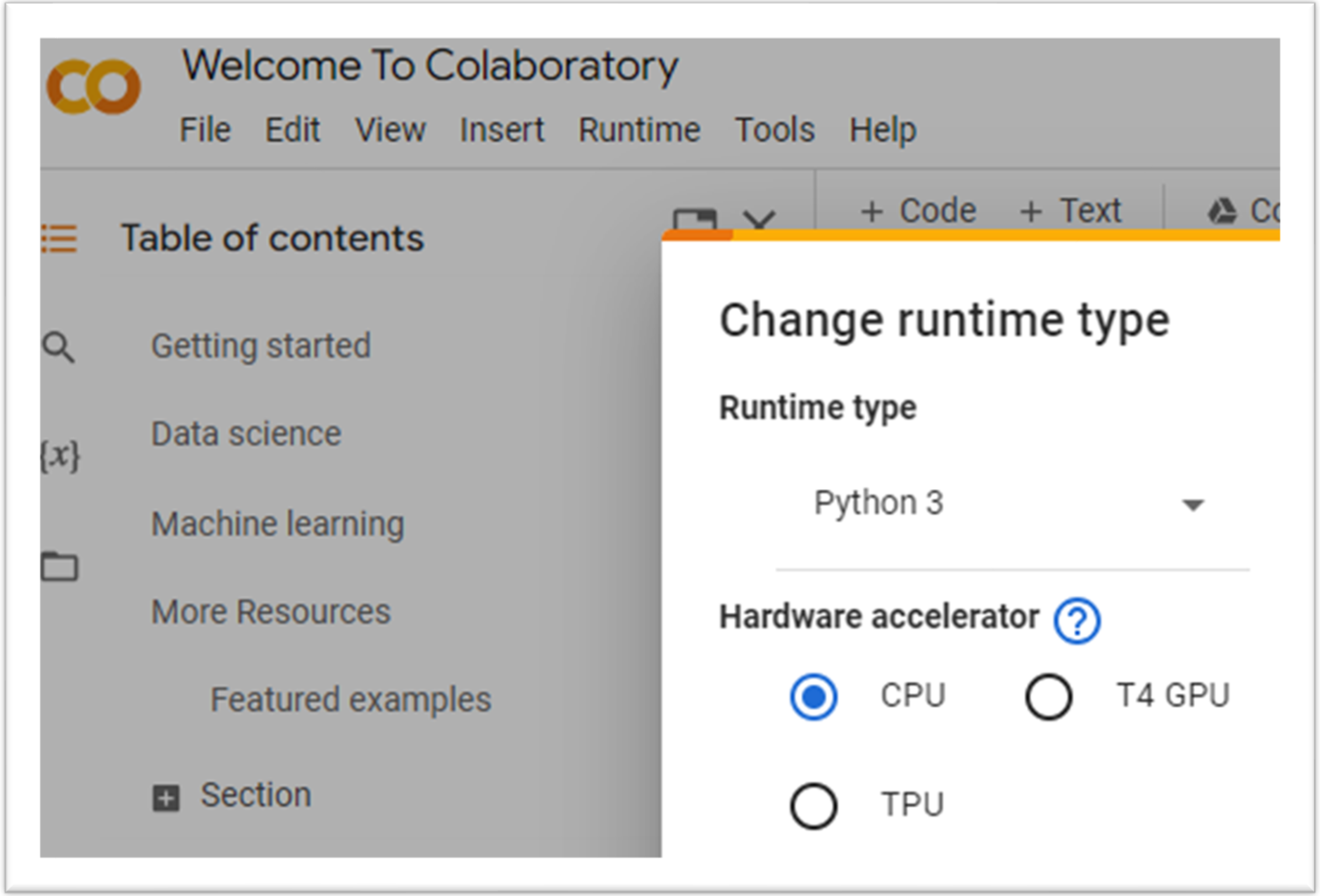
Jupyter

Eclipse

IntelliJ IDEA

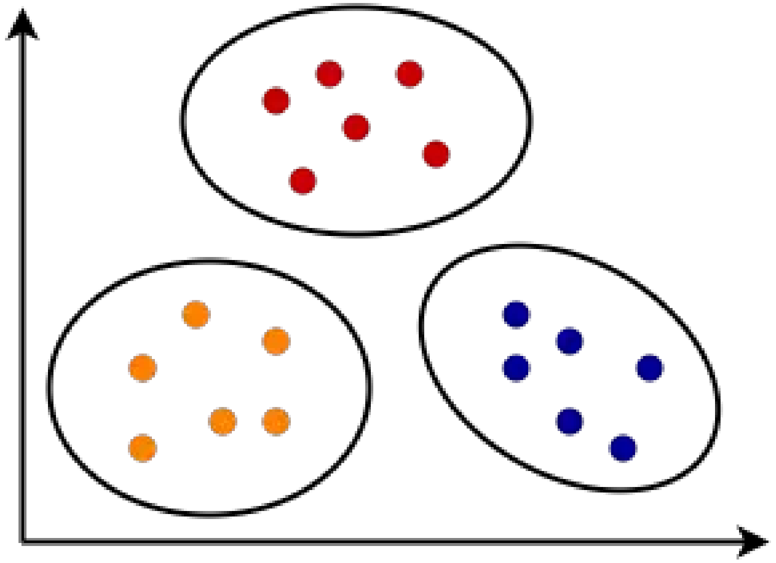
Notepad++

Colab.research.google.com

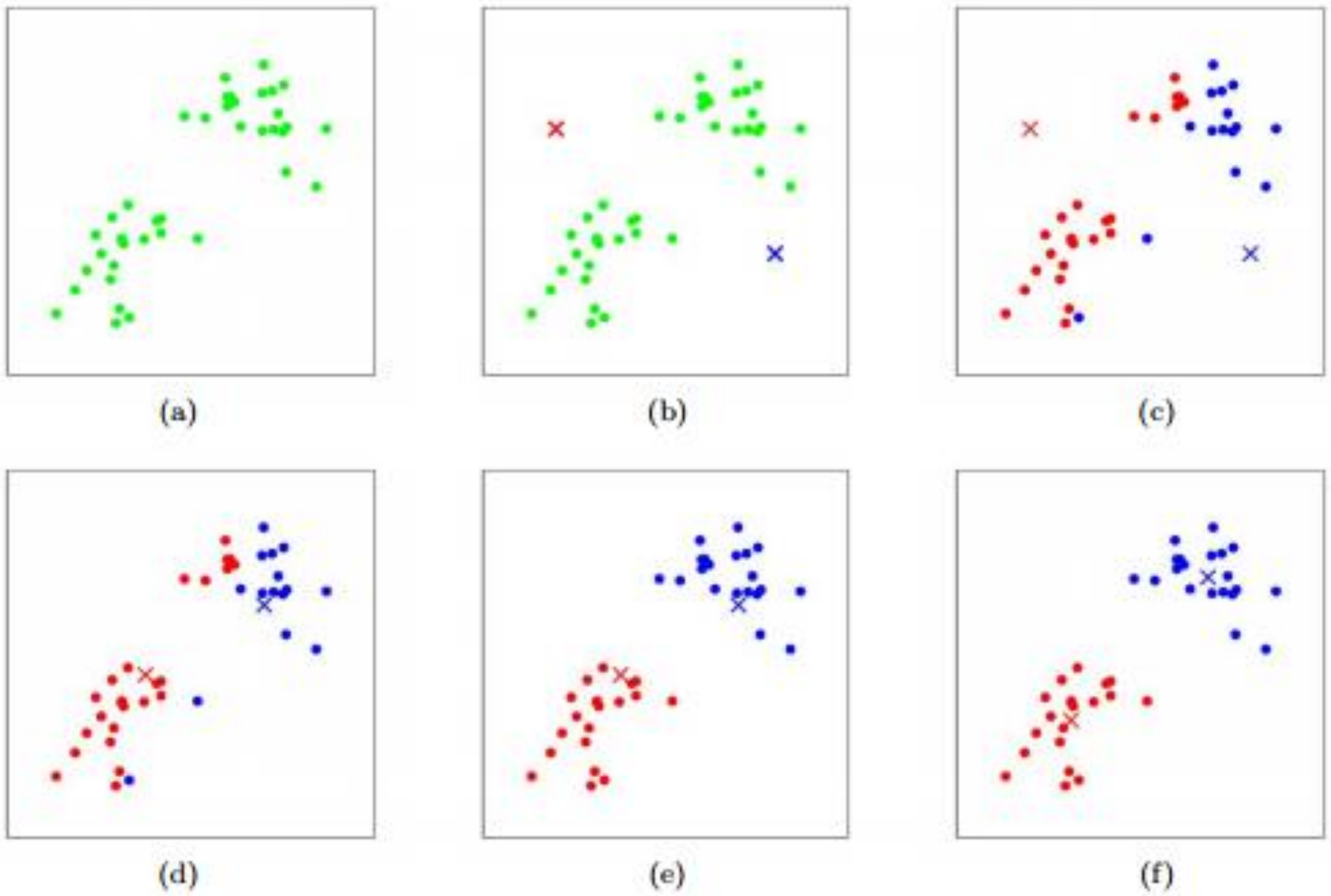


Clustering Algorithms - KMeans

1. Choose the number of clusters K.
2. Randomly select any K data points as cluster centers.
3. Calculate the distance between each data point and each cluster center.
4. Assign each data point to that cluster whose center is nearest to that data point.
5. Re-compute the center of newly formed clusters.
6. Keep repeating the procedure from Step-03 to Step-05 until any of the following stopping criteria is met:
   - Center of newly formed clusters do not change
   - Data points remain present in the same cluster
   - Maximum number of iterations are reached

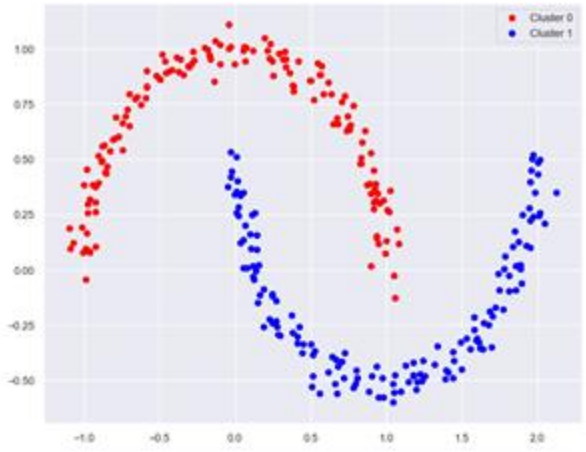


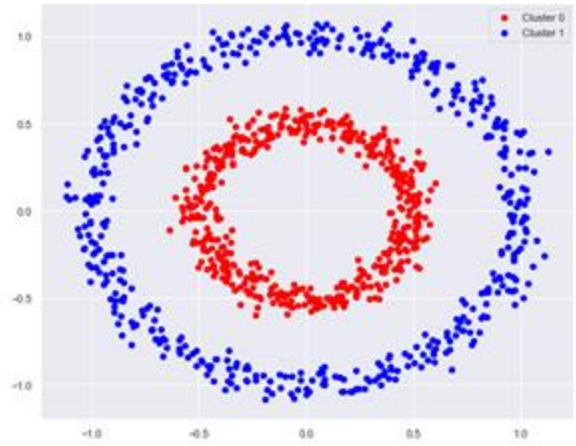
KMeans



KMeans - Limitations

- Setting a value for K
- Numerical variables only
- Sensitive to initial conditions
- Data has no noises or outliers
- Data has symmetric distribution of variables
- Good in a spherical-like shapes
- Variables on the same scale
- There is no collinearity



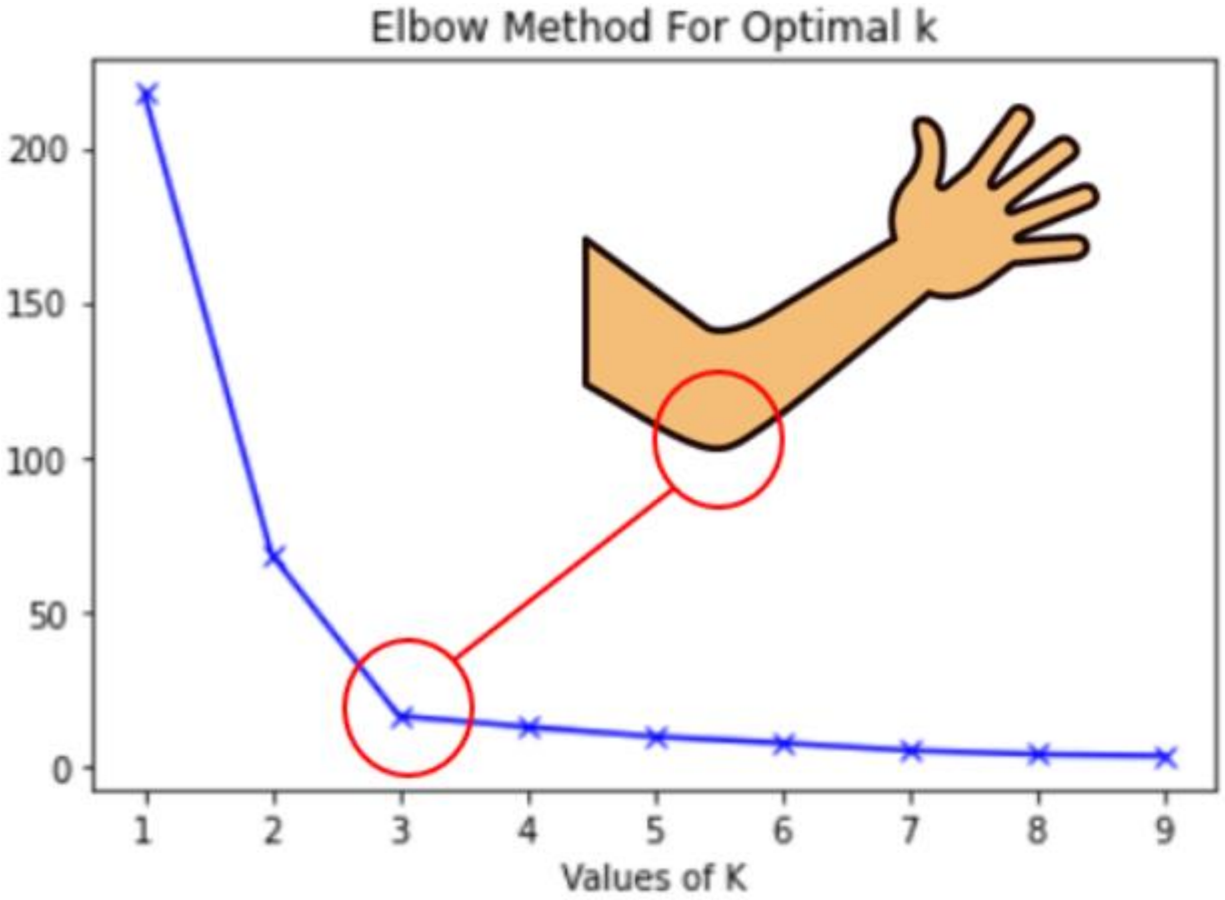


Evaluation Metrics & Elbow Method

https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/
https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam

number of clusters   number of cases   case i centroid for cluster j

$$\text{Inertia} = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$



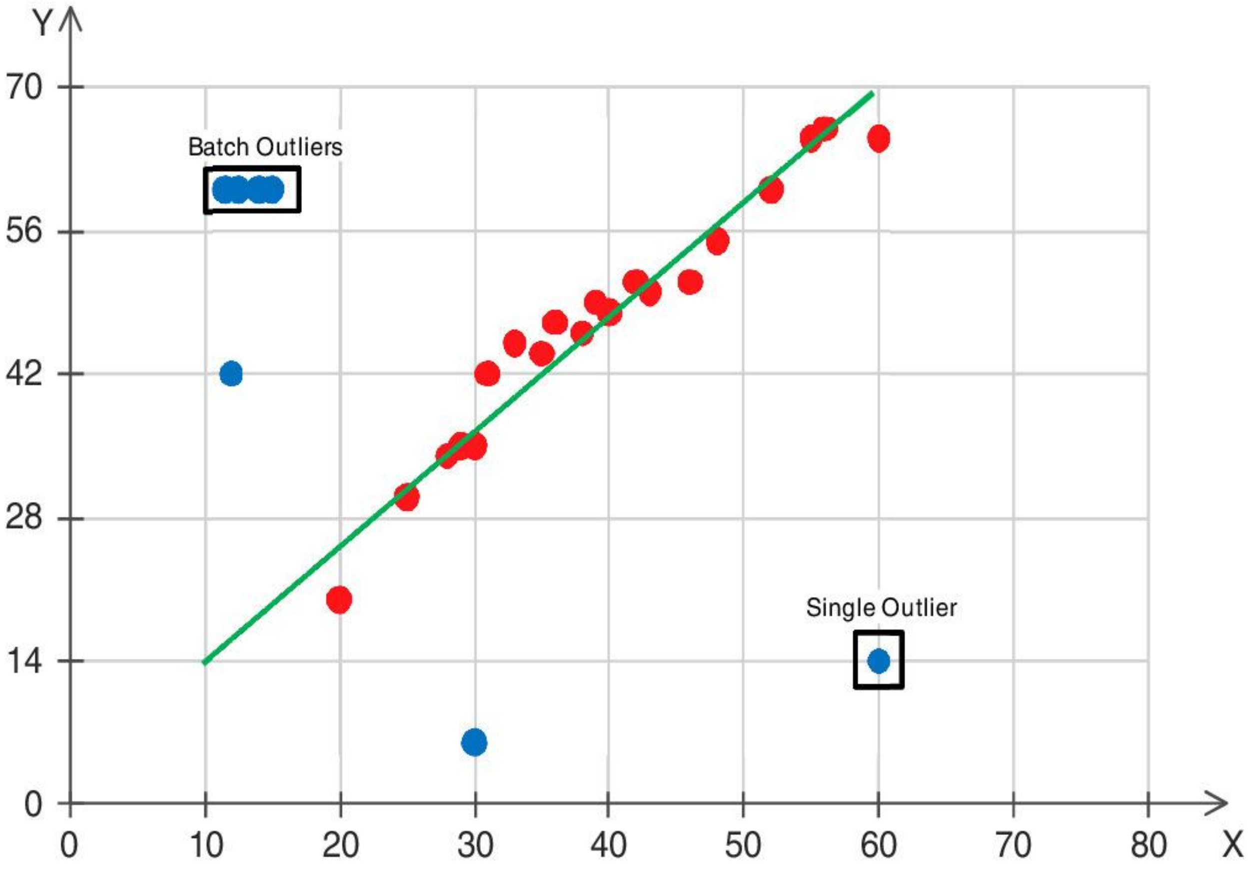
Outliers’ and Noise

Outlier

A data point that deviates significantly from the majority of other data points in a dataset

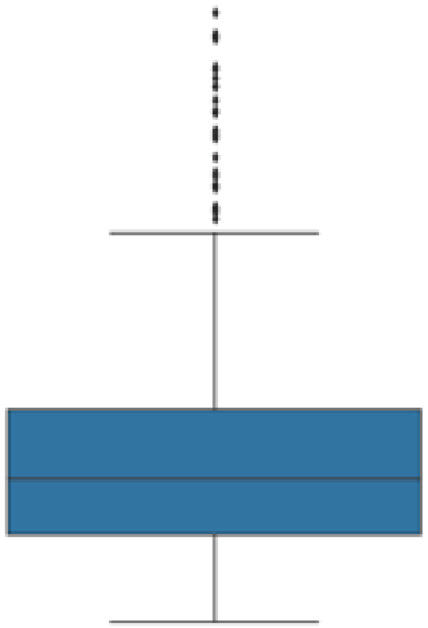
Noise

Refers to random or irrelevant variations present in data that can obscure meaningful patterns or relationships.



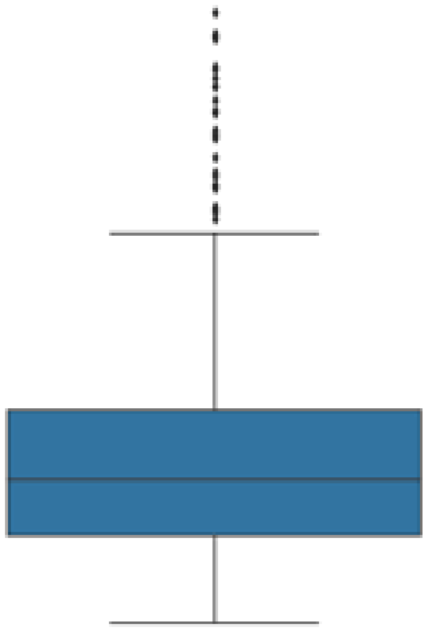
Outlier Detection

- IQR (Interquartile Range)
- Z-Score
- Local Outlier Factor (LOF)
- Isolation Forest
- DBSCAN!
- ...

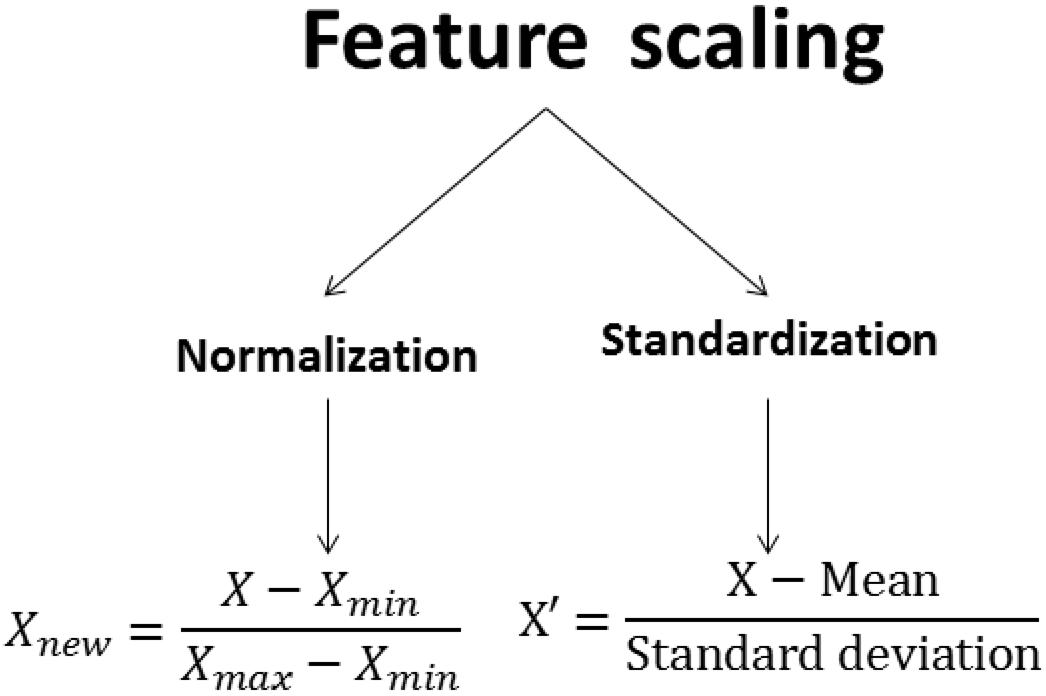


Outlier Handling

- Reporting
- Removing
- Imputation
- Transformation
- Capping and Flooring
- Data Segmentation



Feature Scaling



Other k-Clustering Algorithms

KMeans++

A method for initializing cluster centroids in K-means clustering that selects centroids with a higher probability of being distant from each other.

KMedoids

A clustering algorithm that uses actual data points as cluster representatives (medoids) instead of centroids, making it more robust to outliers.

KModes

A clustering algorithm designed for categorical data that identifies clusters based on the most frequent categorical values.

KPrototype

A hybrid clustering algorithm that combines K-means for numerical data and K modes for categorical data to handle mixed data types within a dataset.

Covariance & Correlation

Covariance is used to understand the relationship between two variables and how they might move in relation to each other.

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables. It not only shows the kind of relation (in terms of direction) but also how strong the relationship is.

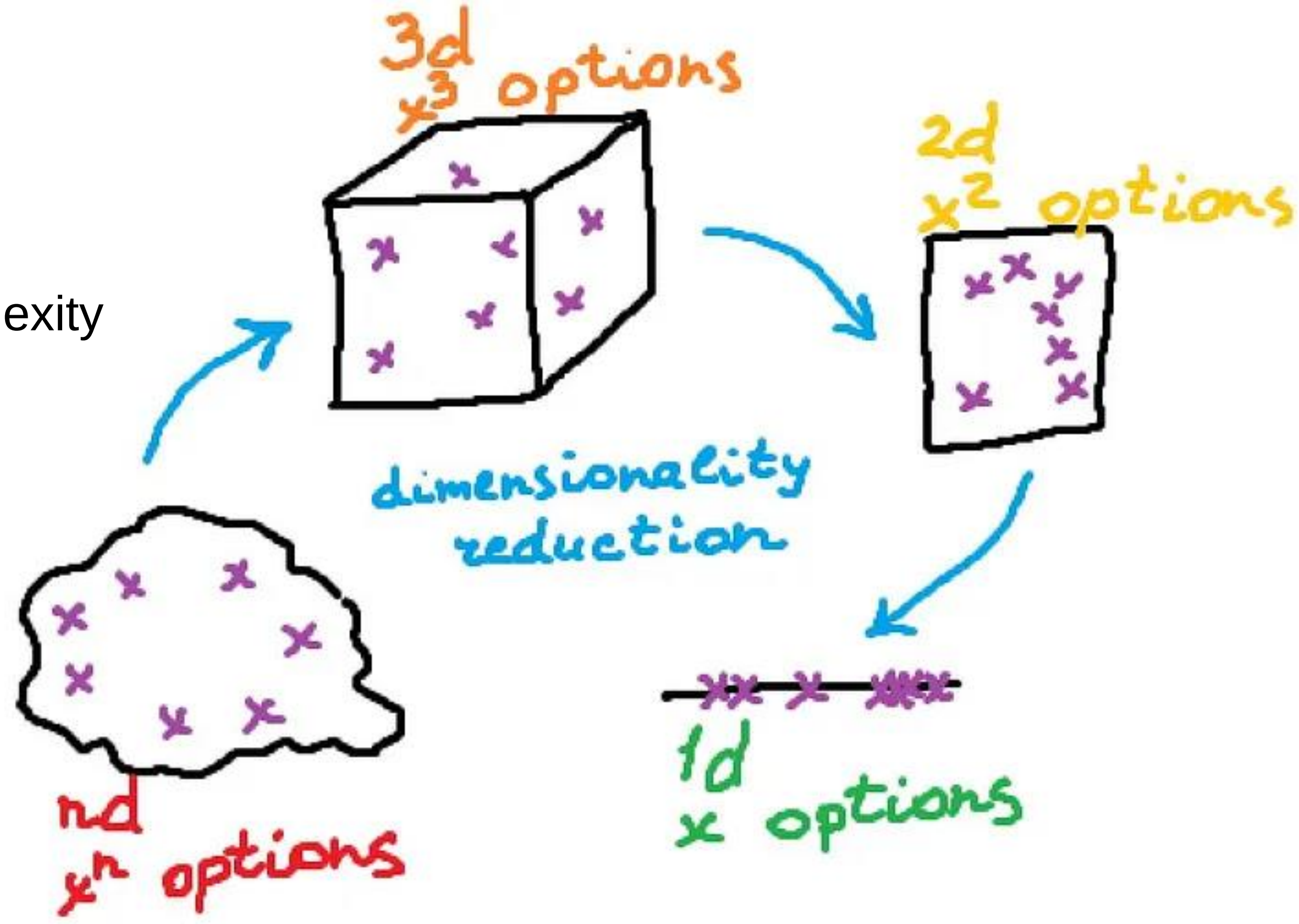
$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

Missing Values Handling

- Rows Deletion
- Columns Deletion
- Mean/Median/Mode Imputation
- Regression
- KNN
- Interpolation
- Predictive Models
- Grouping Techniques

Dimensionality Reduction

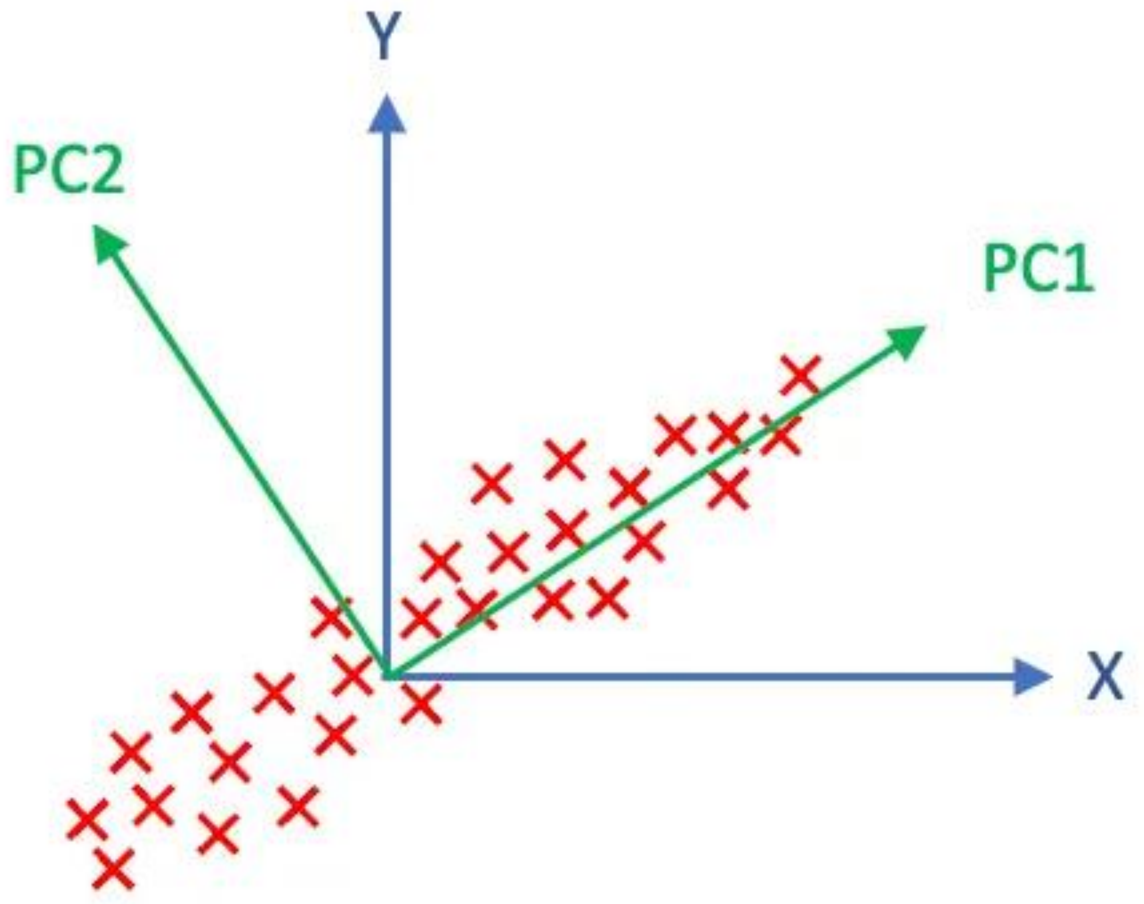
- Improved Visualization
- Feature Selection
- Reduced Computational Complexity
- Noise Reduction
- Overfitting Prevention
- Better Model Performance
- Anomaly Detection



Principal Component Analysis (PCA)

STEPS

1. Data Standardization
2. Calculate Covariance Matrix
3. Compute Eigenvalues and Eigenvectors
4. Sort Eigenvalues
5. Select Principal Components
6. Projection & Dimensionality Reduction



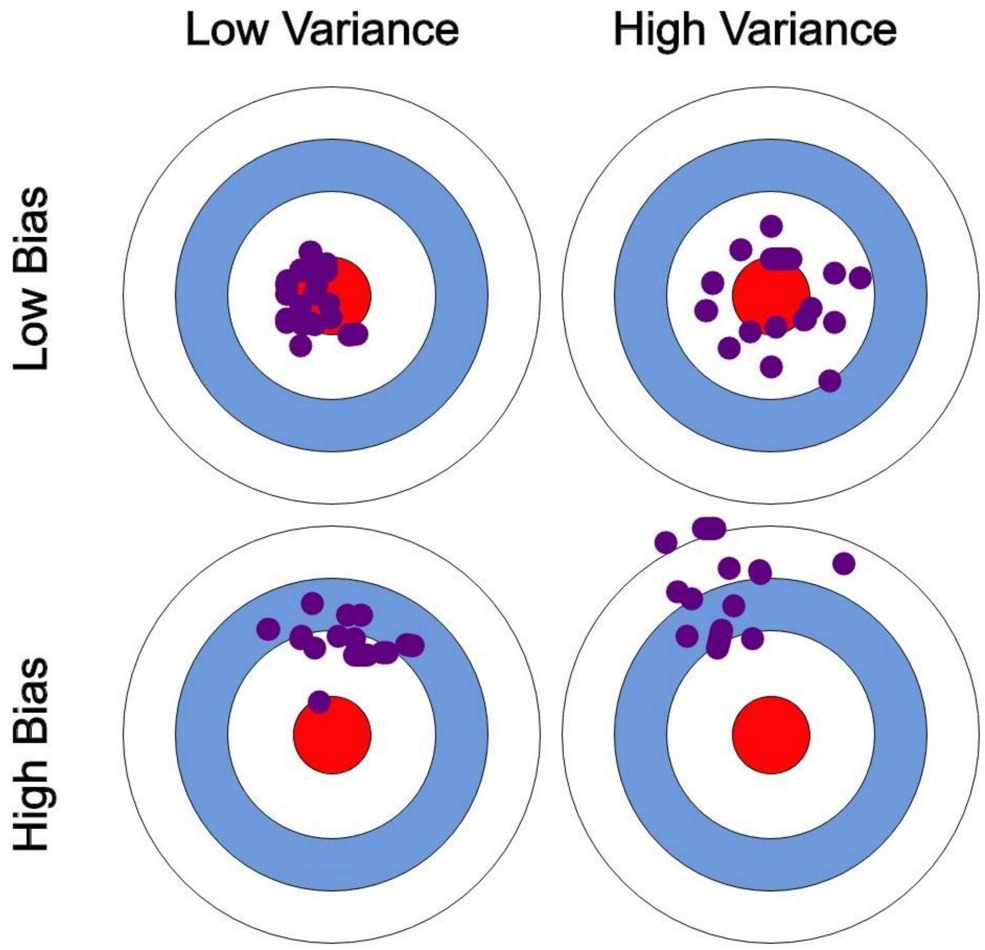
Bias-Variance

Bias

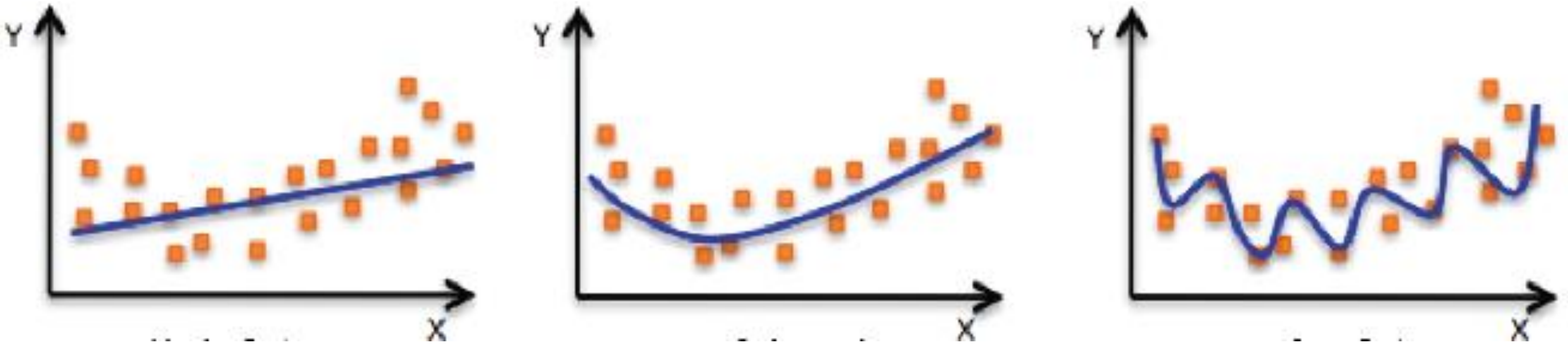
Difference between the prediction of the values by the Machine Learning model and the correct value

Variance

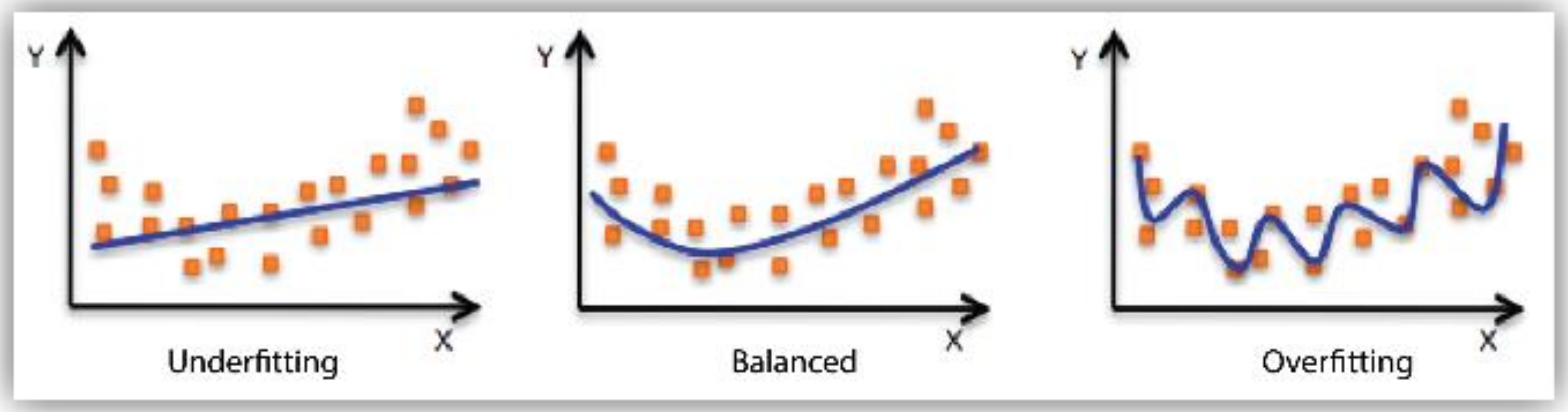
The variability of model prediction for a given data point which tells us the spread of our data



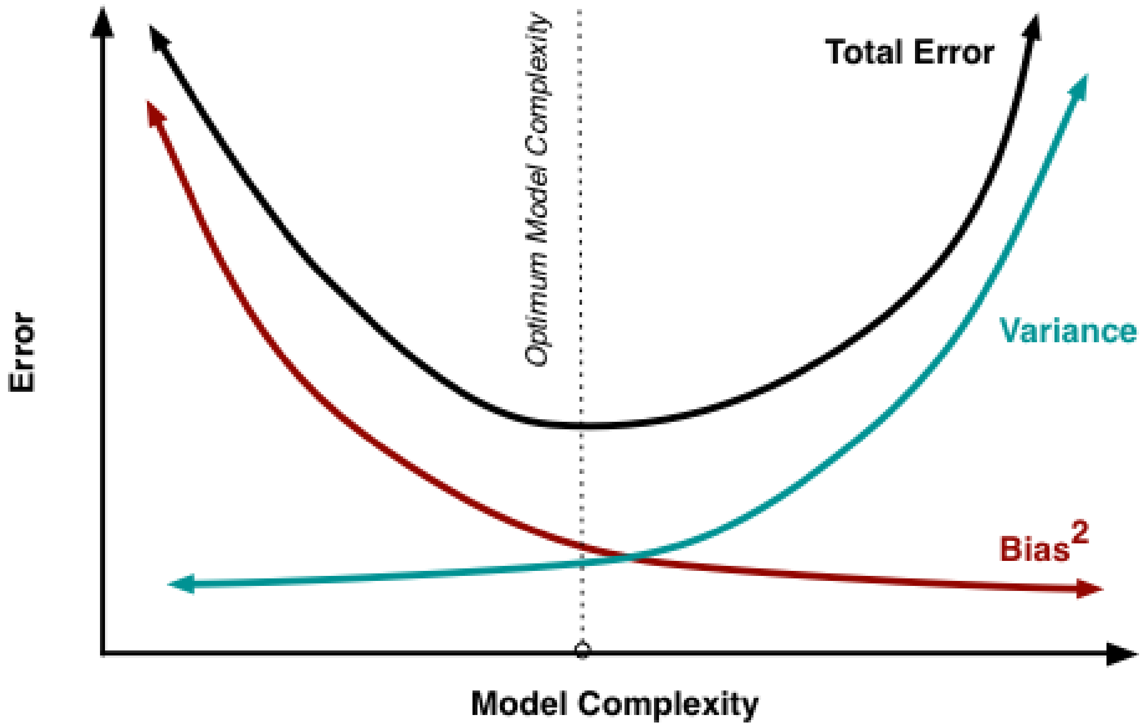
Bias-Variance



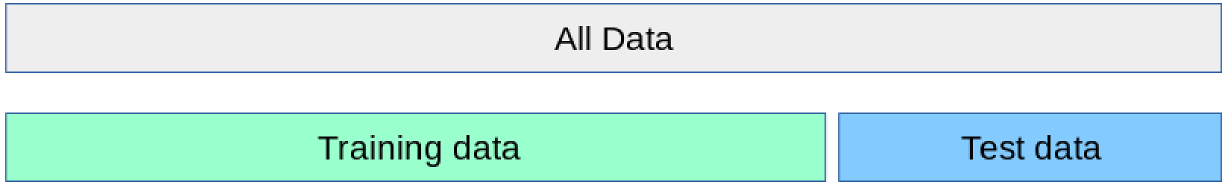
Overfitting & Underfitting



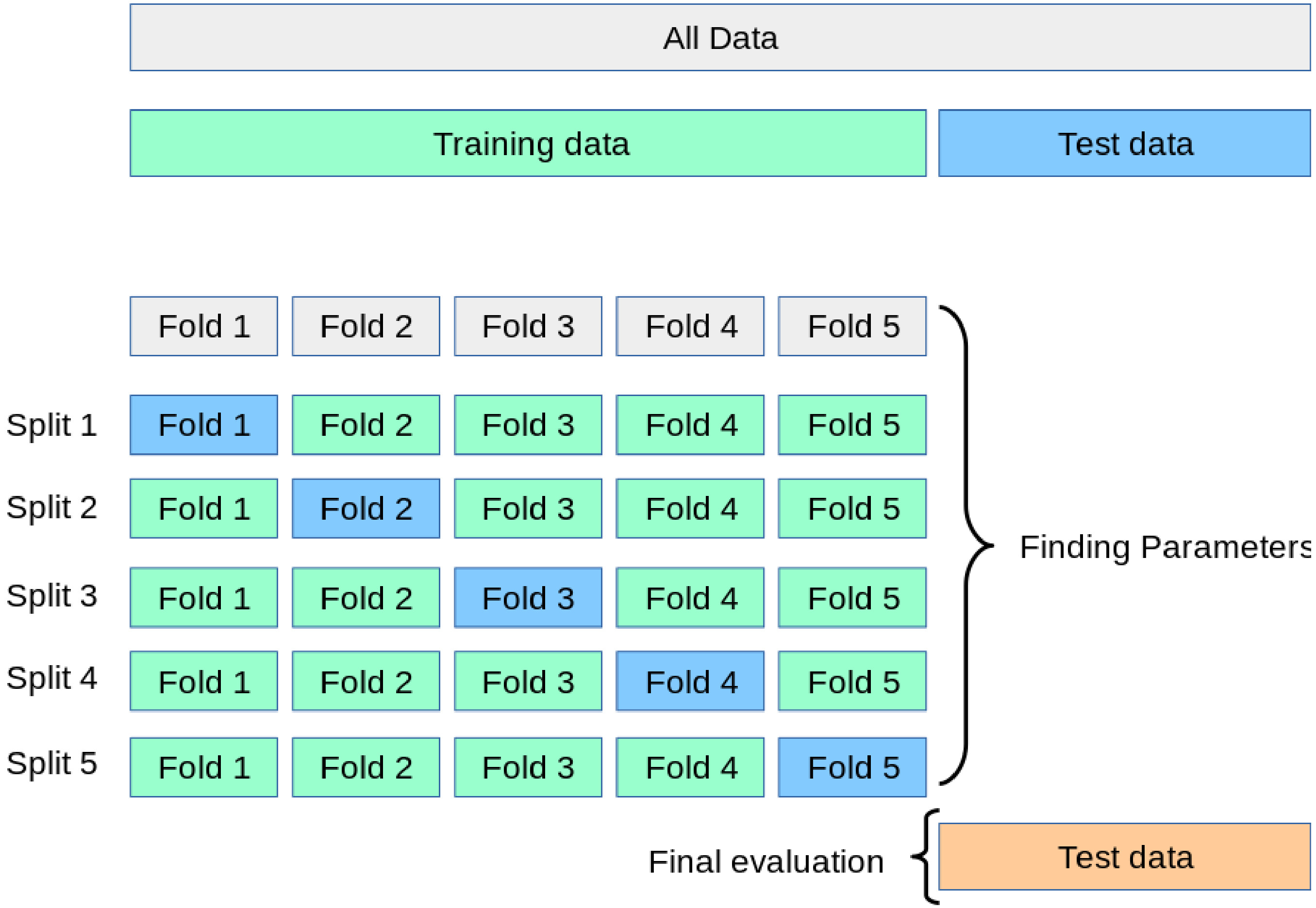
Bias-Variance Trade-off



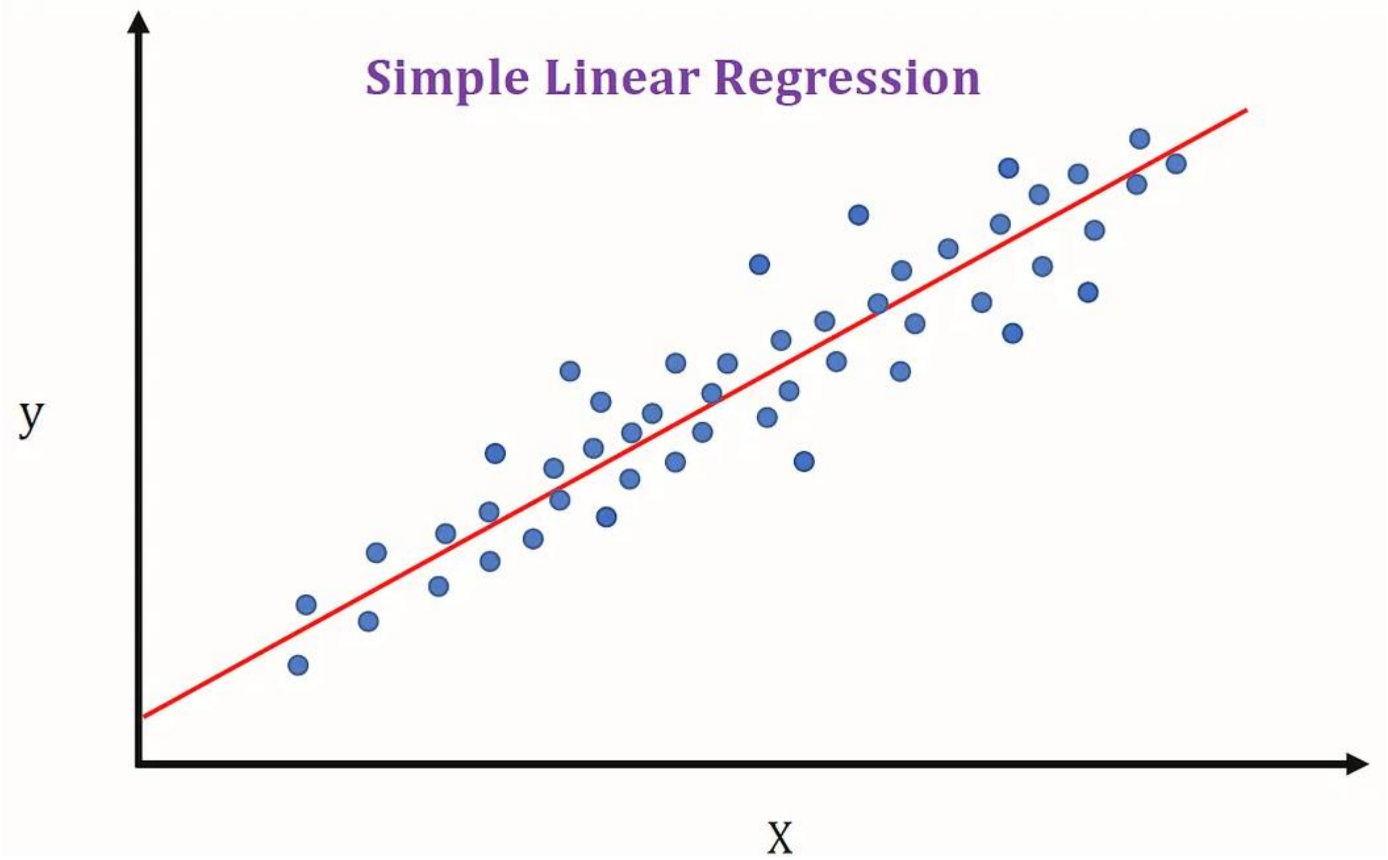
Train-Test Split



Cross Validation



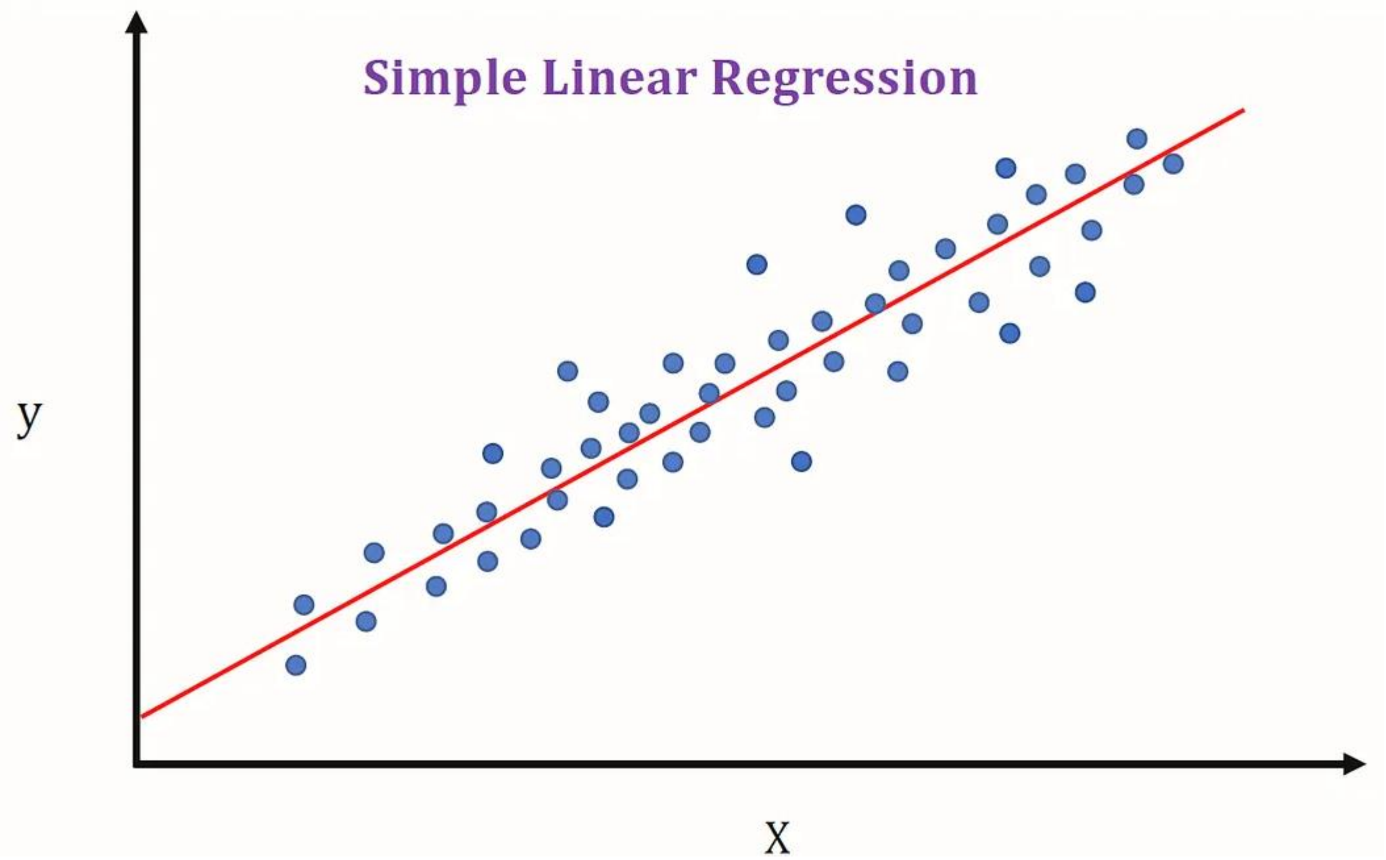
Simple Linear Regression



Simple Linear Regression

$$slope = \frac{covariance(x, y)}{variance(x)}$$

$$intercept = \bar{y} - slope \times \bar{x}$$

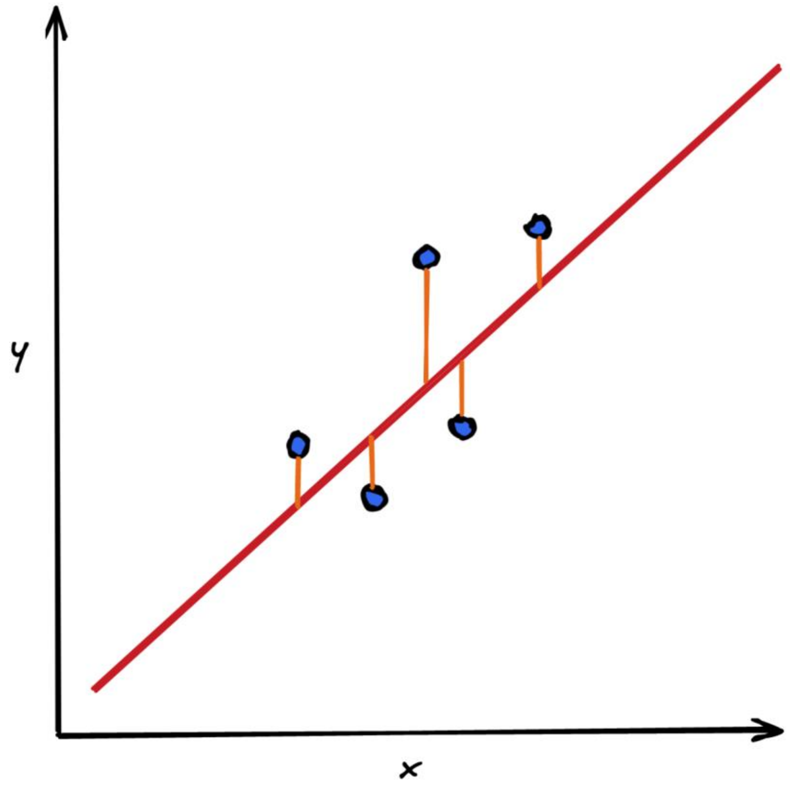


Linear Regression Metrics

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}}$$



Linear Regression Metrics

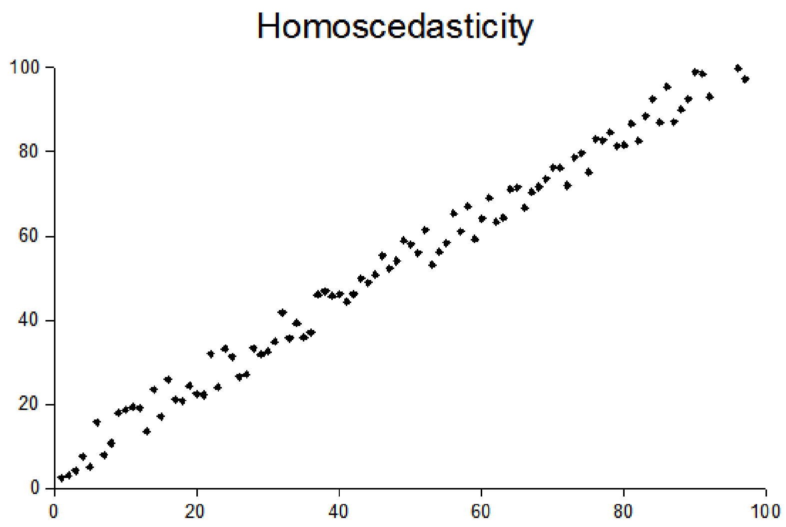
$$SS_{\text{res}} = \sum_i e_i^2 \; , \quad SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

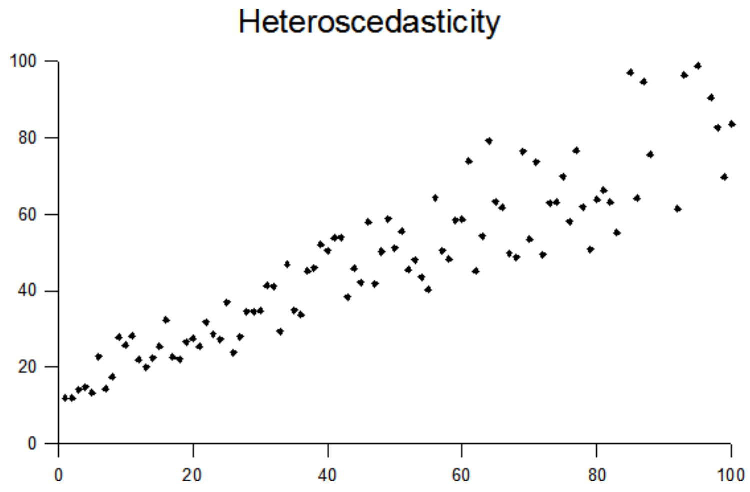
$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

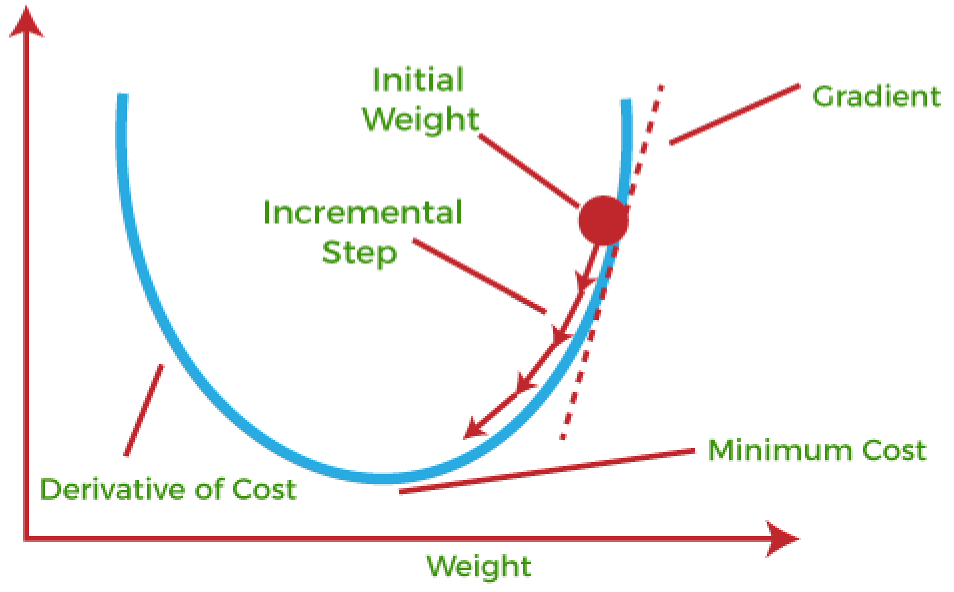
Linear Regression Limitations

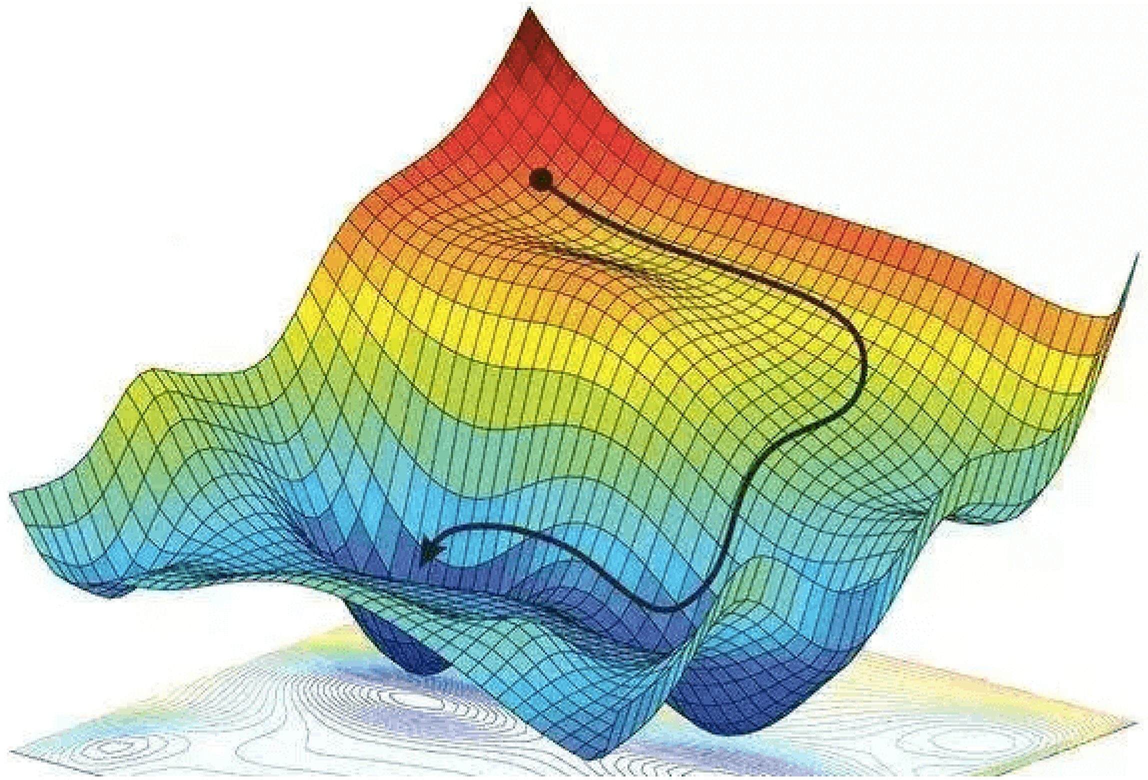
- Assumption of Linearity
- Sensitive to Outliers
- Assumption of Homoscedasticity
- Sensitive to Multicollinearity





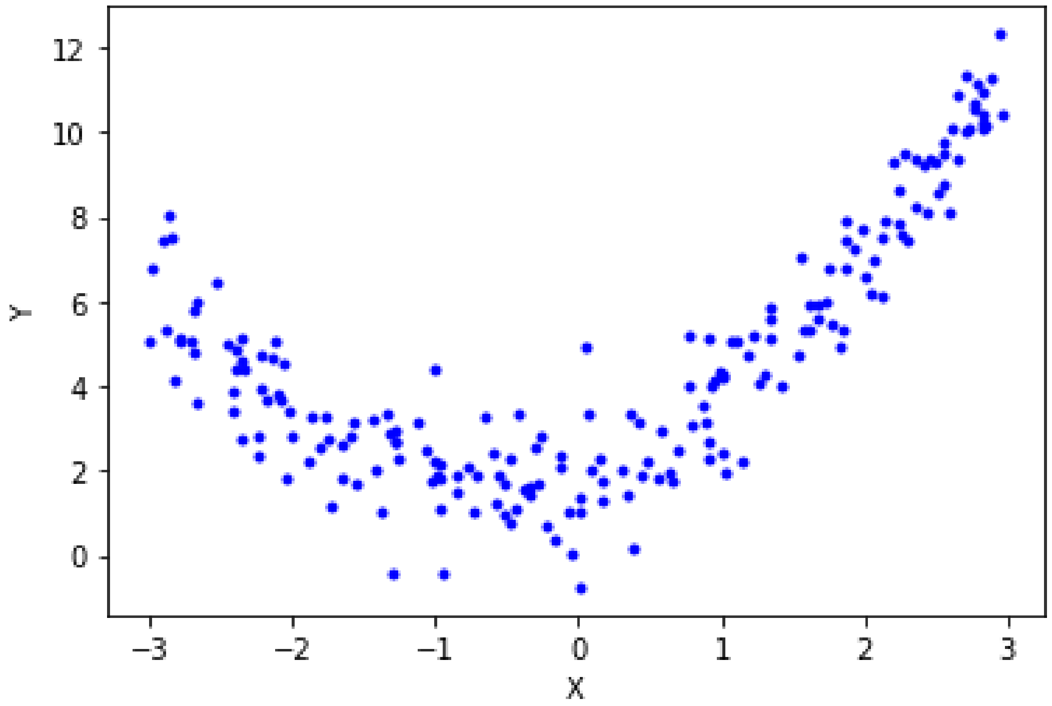
Gradient Descent





https://www.geeksforgeeks.org/gradient-descent-in-linear-regression/

Simple Polynomial Regression



$$y = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

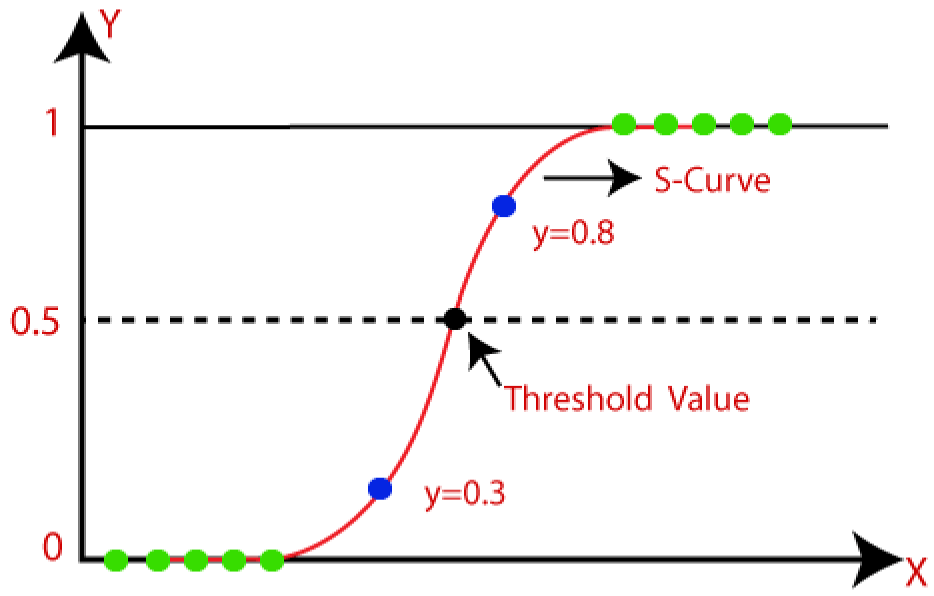
Logistic Regression

Similar to linear regression, logistic regression is also used to estimate the relationship between a dependent variable and one or more independent variables, but it is used to make a prediction about a categorical variable versus a continuous one. A categorical variable can be true or false, yes or no, 1 or 0, et cetera. The unit of measure also differs from linear regression as it produces a probability, but the logit function transforms the S-curve into straight line.

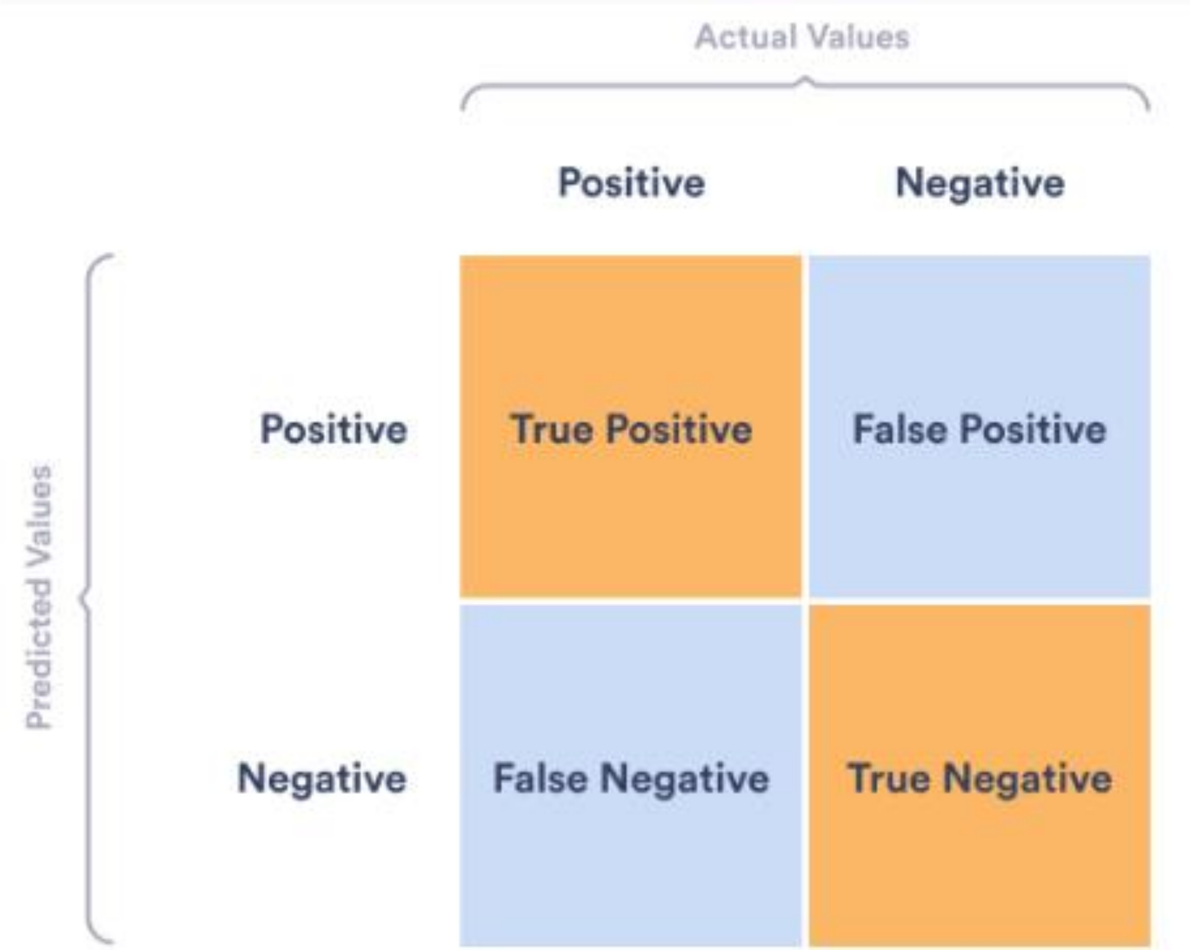
- **Fraud Detection**
- **Disease Prediction**
- **Churn prediction**
- **...**

Logistic Regression

$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

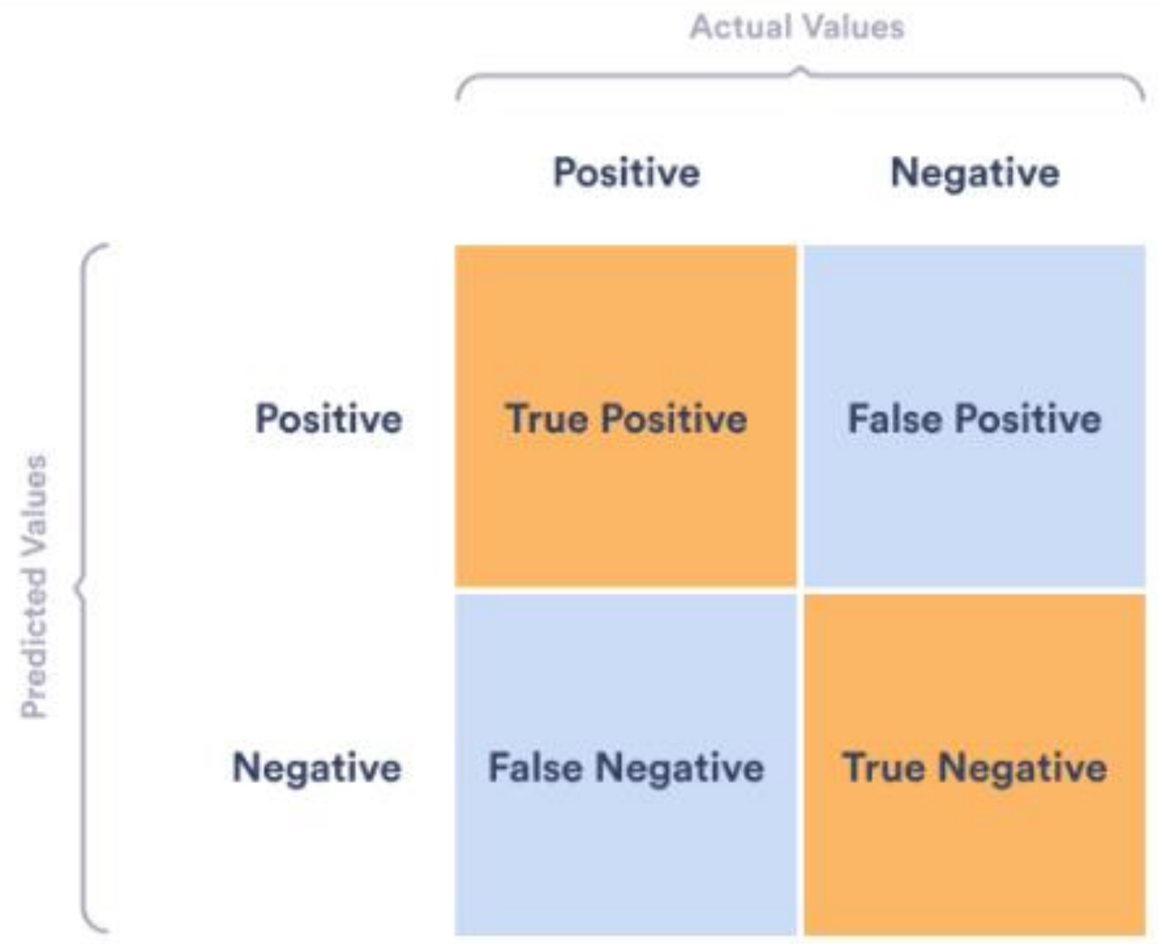


Confusion Matrix (TP, TN, FP, FN)



Metrics

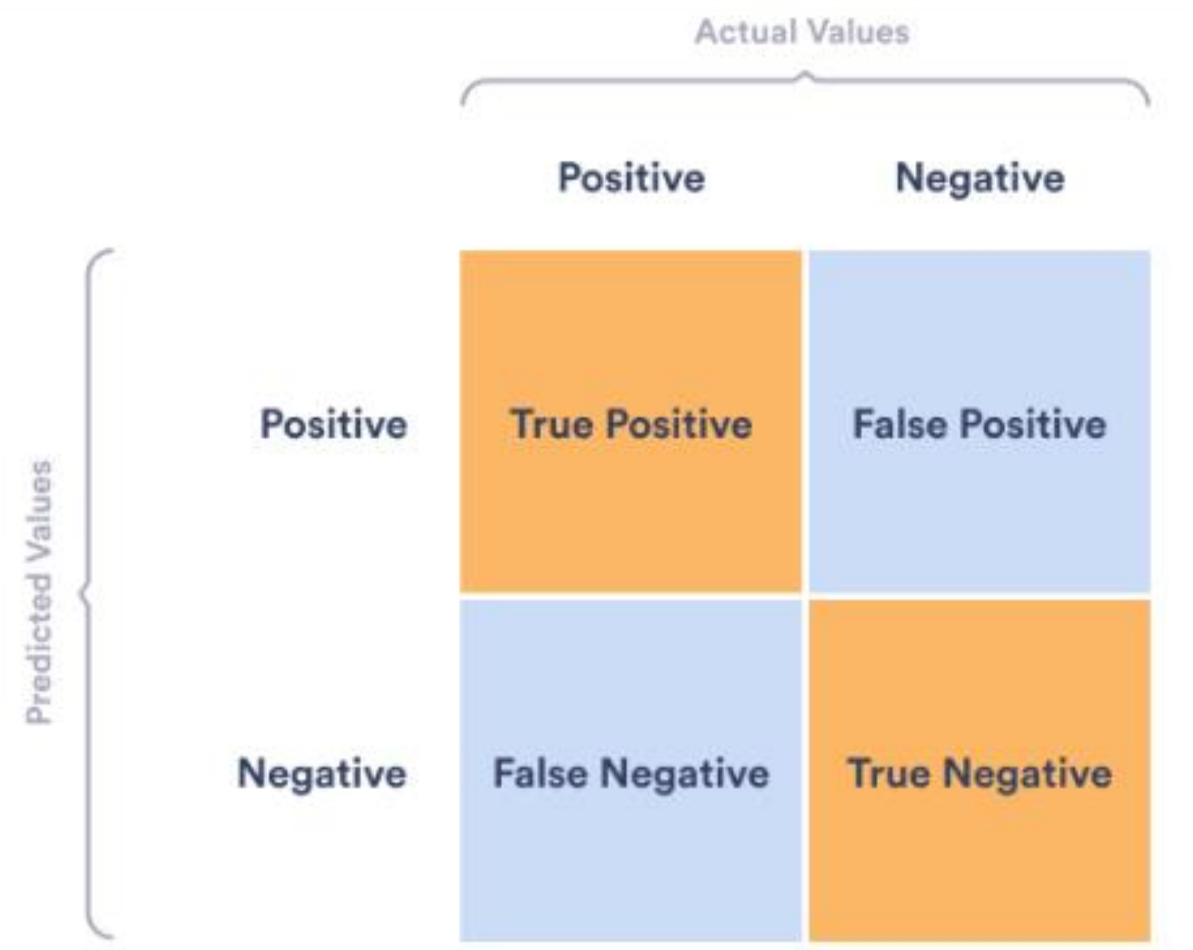
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$



Metrics

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

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

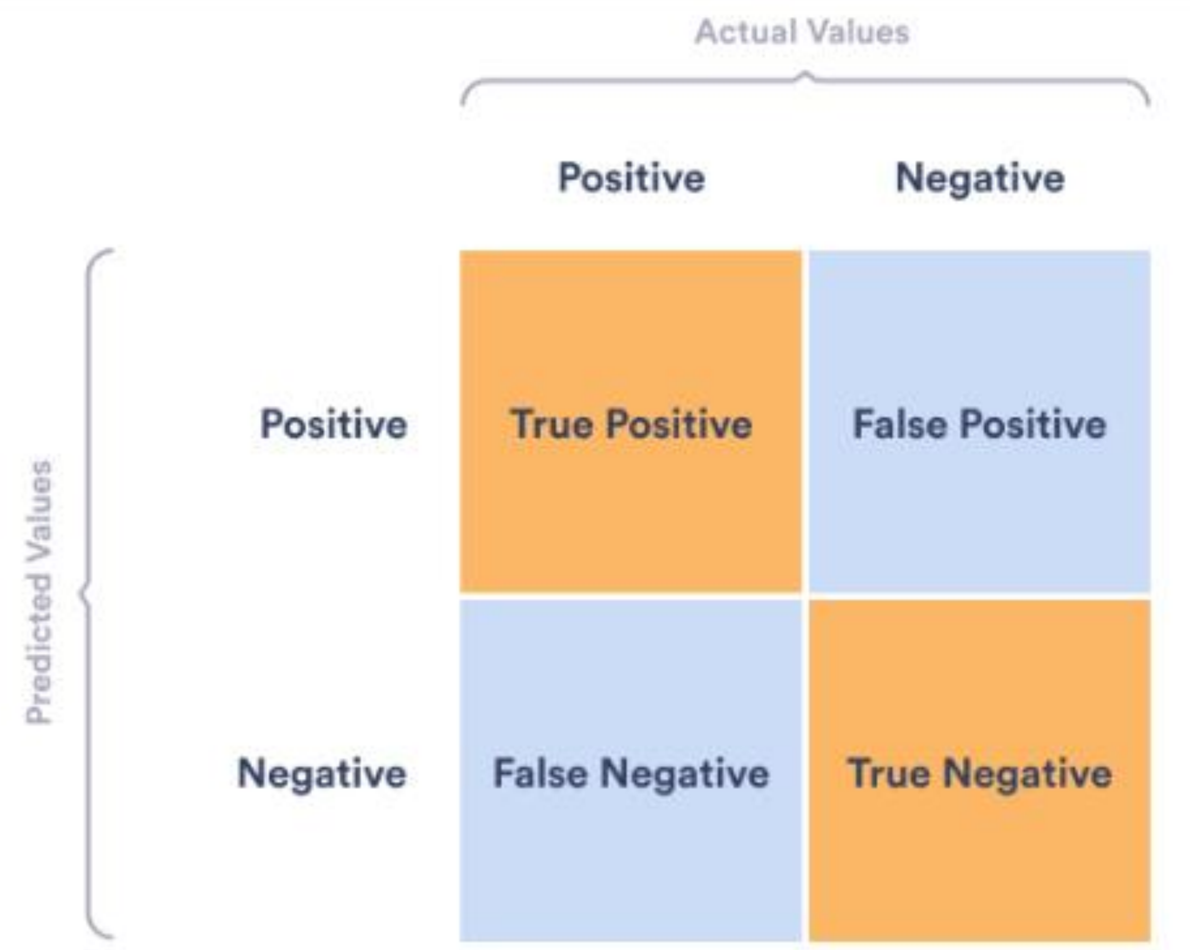


Metrics

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

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

$$Recall = \frac{TP}{TP + FN}$$



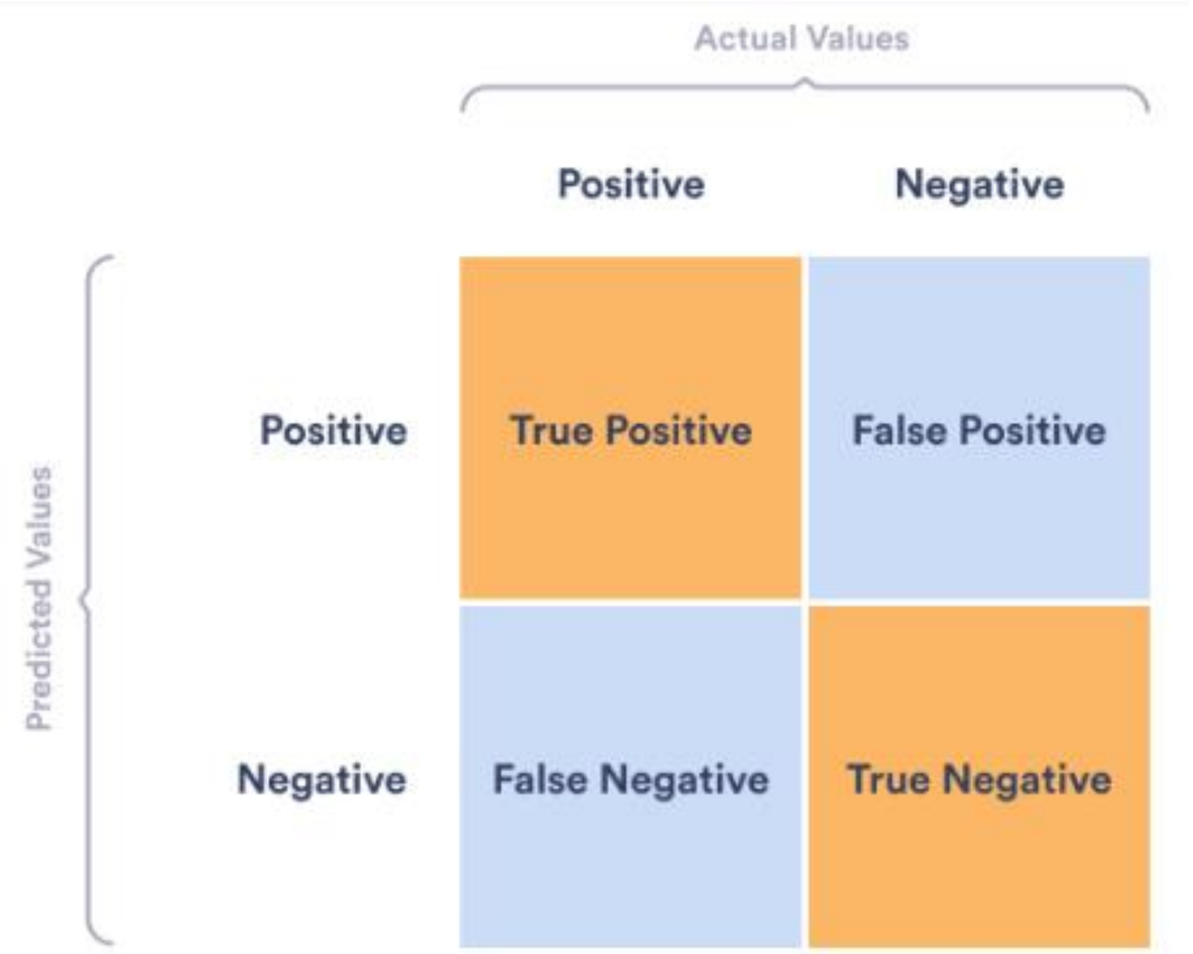
Metrics

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

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

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = \frac{2 \times recall \times precision}{recall + precision}$$



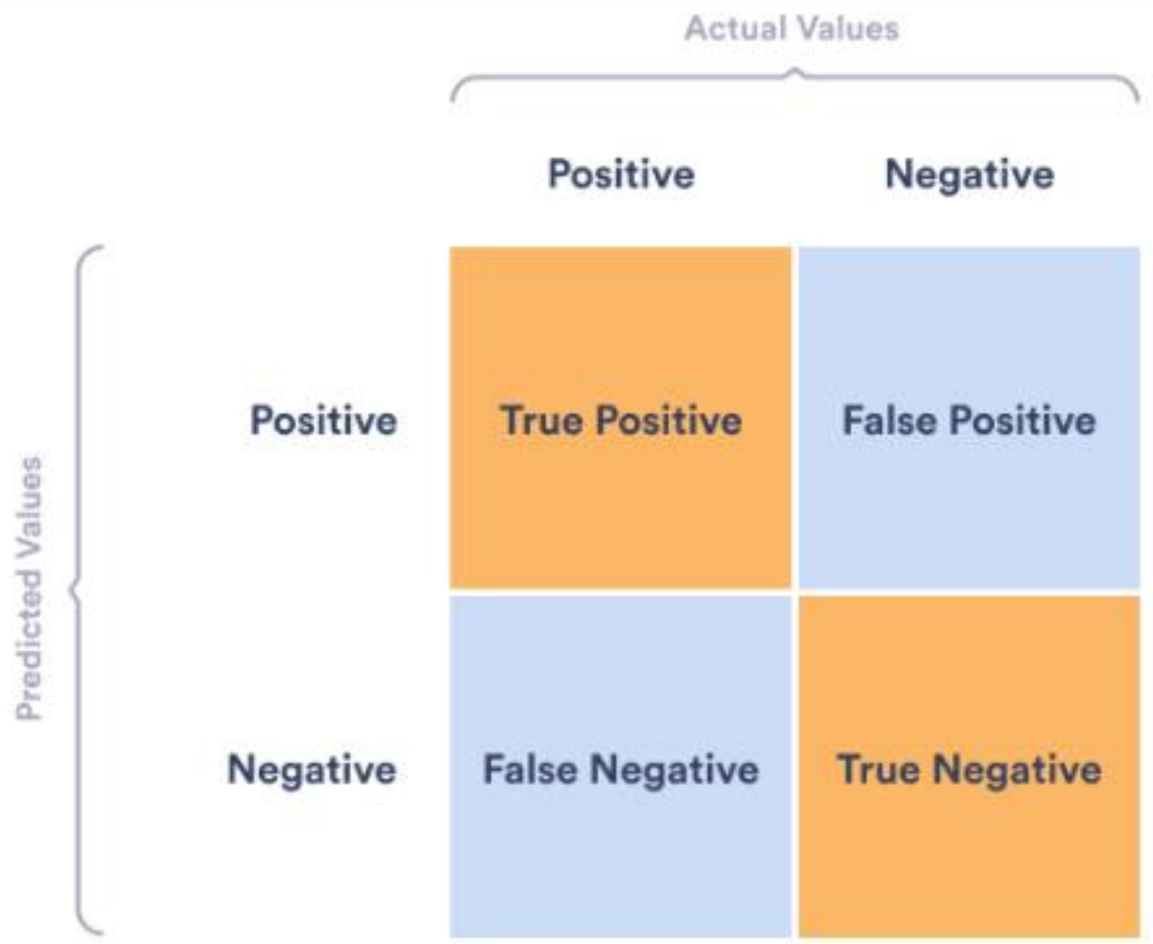
Metrics

$$Sensitivity = TPR = \frac{TP}{TP + FN}$$

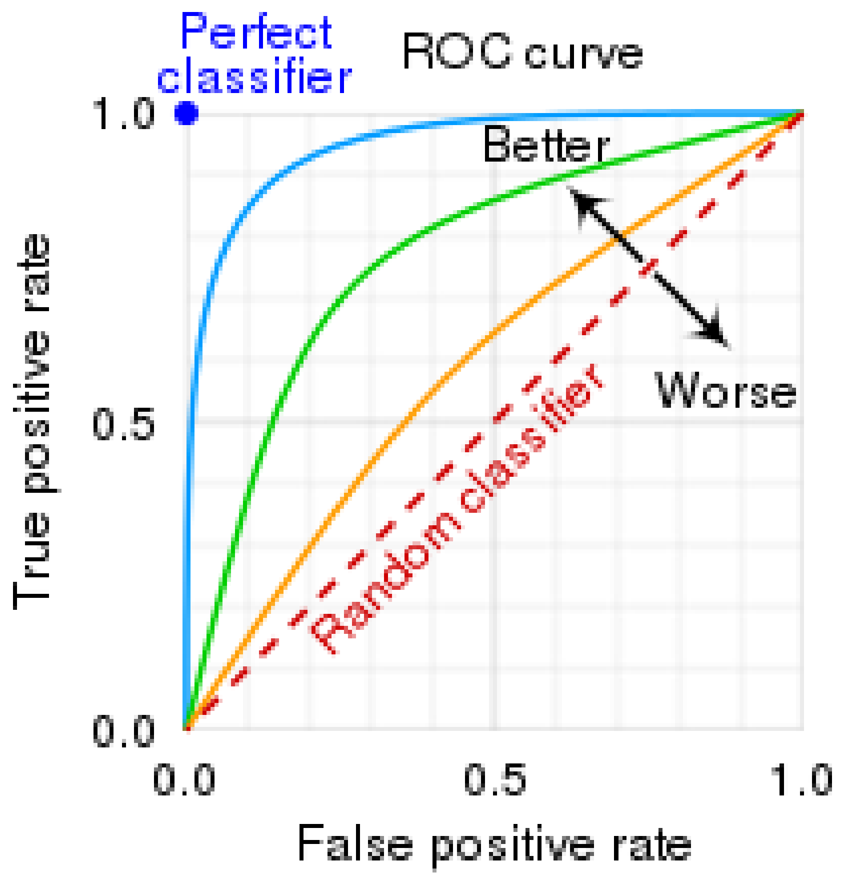
$$Specificity = TNR = \frac{TN}{TN + FP}$$

$$FNR = \frac{FN}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$



AUC-ROC

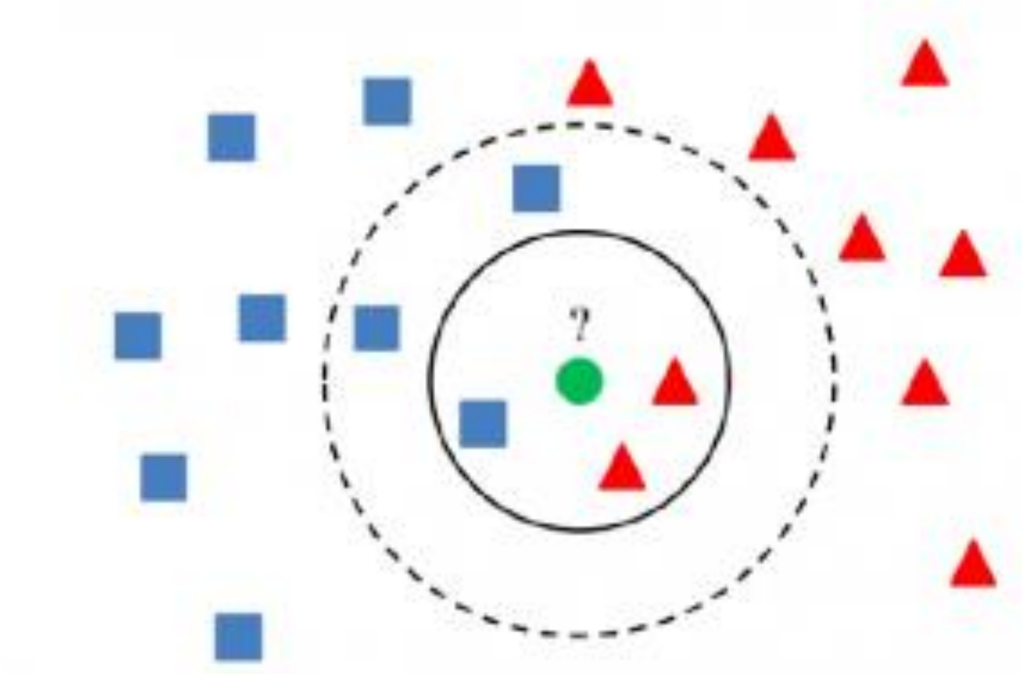


https://dasha.ai/en-us/blog/auc-roc

https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

KNN (K-Nearest Neighbors)

- Classification
- Regression
- Data Imputation
- Anomaly Detection
- Clustering!



Multiclass Averaging

Confusion Matrix

Predicted	Classes	Actual		
		X	Y	Z
	X	10	2	3
	Y	1	8	2
	Z	2	3	9

Multiclass Averaging

Confusion Matrix

	Classes	Actual		
		X	Y	Z
Predicted	X	10	2	3
	Y	1	8	2
	Z	2	3	9

Class	True Positive (TP)	False Positive (FP)	False Negitive (FN)	Precision	Recall
X	10	5	3	0.66	0.76
Y	8	3	5	0.72	0.61
Z	9	5	5	0.64	0.64

Multiclass Averaging

$$\text{PrecisionMicroAvg} = \frac{(TP_1 + TP_2 + \ldots + TP_n)}{(TP_1 + TP_2 + \ldots + TP_n + FP_1 + FP_2 + \ldots + FP_n)} = 0.675$$

$$\text{RecallMicroAvg} = \frac{(TP_1 + TP_2 + \ldots + TP_n)}{(TP_1 + TP_2 + \ldots + TP_n + FN_1 + FN_2 + \ldots + FN_n)} = 0.675$$

$$\text{PrecisionMacroAvg} = \frac{(Prec_1 + Prec_2 + \ldots + Prec_n)}{n} = 0.673$$

$$\text{RecallMacroAvg} = \frac{(Recall_1 + Recall_2 + \ldots + Recall_n)}{n} = 0.67$$

Multiclass Averaging

- **Micro-Averaging**

  Sum up the true positives, false positives, and false negatives across all classes, then calculate the metric to emphasize overall performance across all instances.

  Useful when you want to emphasize overall performance across all instances.

- **Macro-Averaging**

  Calculate the performance metric for each class individually and then average these metrics to assess overall class-agnostic model performance.

  Useful when you want to evaluate the model's overall performance without considering class distribution.