WELCOME

Introduction to Machine Learning with Python

SUDIP SHRESTHA, PHD

ABOUT ME

Data Science / AI Lead
Advanced Analytics, Modeling

Instructor: Maven.com

Applied Data Science with Python

CLASS MATERIALS

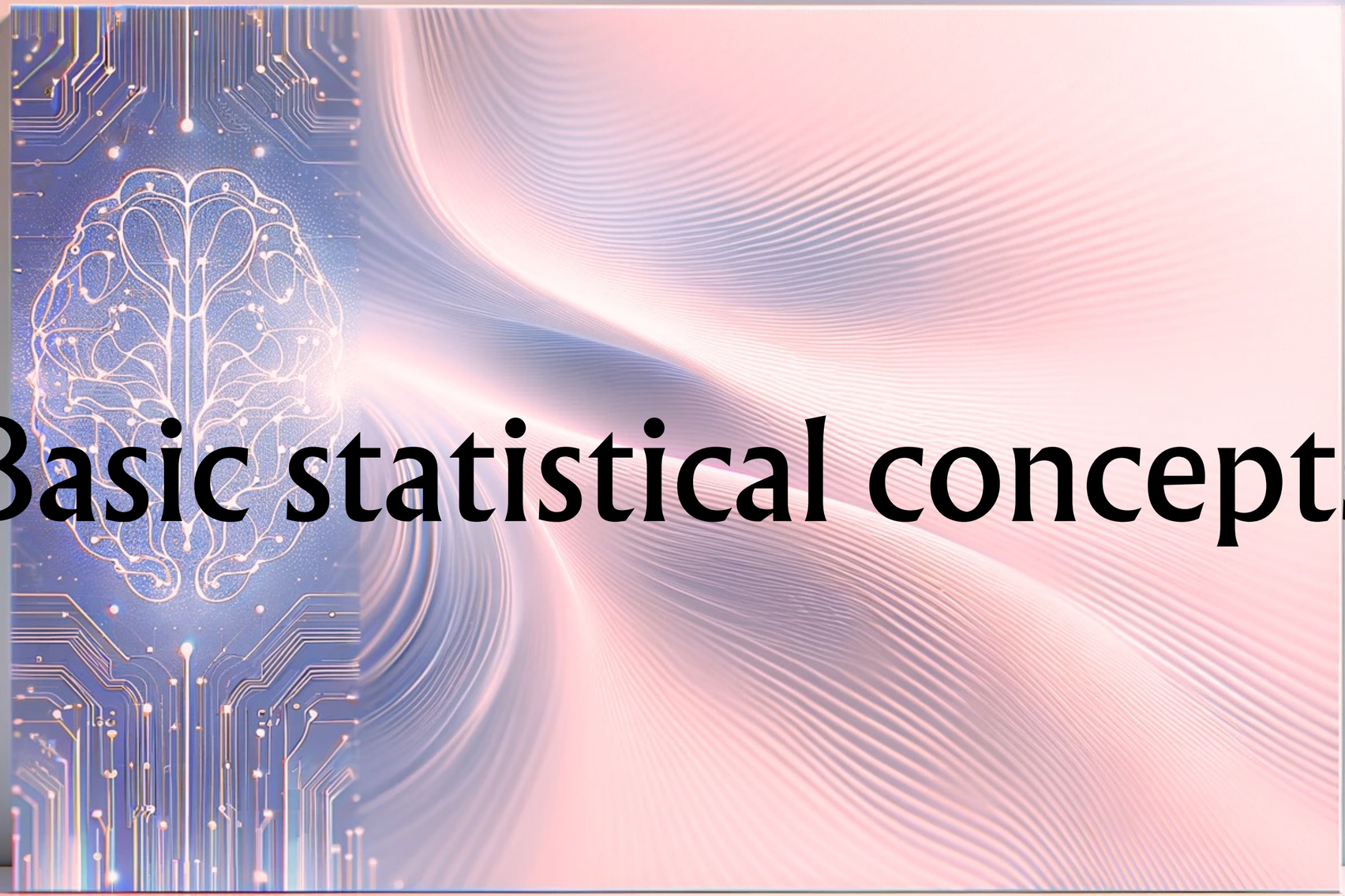
Class Presentation

Notebooks - Python file

Recording

LEARNING OBJECTIVE

- Basic statistical concepts
- Overview of machine learning
- Machine learning lifecycle
- Types of machine learning
- Hands-on Coding



Basic statistical concepts

BASIC STATISTICAL CONCEPT

$$\text{Mean} \quad \mu = \frac{\sum x_i}{n}$$

- The mean is calculated by summing all the values in the dataset and then dividing by the number of values (n). It represents the average value.
- Provides a quick and simple measure of the central tendency of a dataset, helping to summarize and understand large amounts of data efficiently.

Table: Ages of Participants

Participant ID	Age
1	25
2	30
3	22
4	28
5	26

Sum of Ages: 25+30+22+28+26 = 131

No. of participants = 5

Mean Age = 131/5 = 26.2

BASIC STATISTICAL CONCEPT

Median

- The median is the middle value when the data is sorted in ascending order. If there is an even number of values, it is the average of the two middle values.
- Provides a robust measure of central tendency, particularly useful in skewed distributions where it better represents the dataset's central value than the mean.

Table: Test Scores of Students

Student ID	Test Score
1	70
2	85
3	90
4	75
5	88

Sort test score in ascending order

Student ID	Test Score (Sorted)
1	70
4	75
2	85
5	88
3	90

Since there are 5 test scores (an odd number), the median will be the middle value, which is the 3rd score when sorted. Hence The median test score is 85.

BASIC STATISTICAL CONCEPT

Mode

- The value that occurs most frequently in the dataset.
- Identifies the most frequently occurring value in a dataset, offering insights into common trends and patterns.

In this survey data, respondents are asked about their favorite color.

The mode is Blue, as it is the color mentioned most frequently (4 times) by the respondents.

Table: Survey of Favorite Colors

Respondent ID	Favorite Color
1	Blue
2	Red
3	Blue
4	Green
5	Blue
6	Red
7	Green
8	Blue
9	Yellow
10	Red

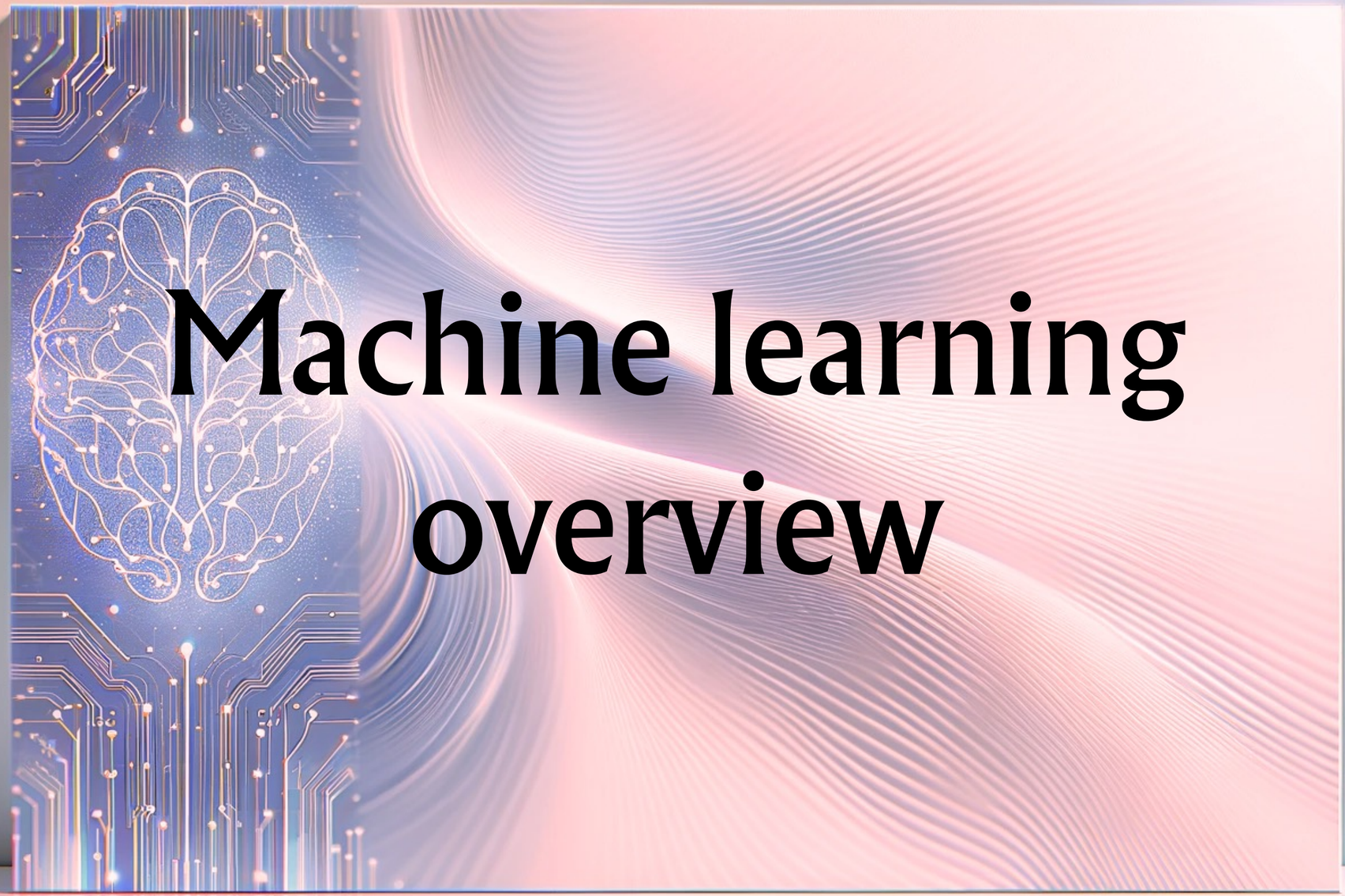
BASIC STATISTICAL CONCEPT

Standard Deviation

- The standard deviation measures the amount of variation or dispersion of a set of values.
- It's calculated as the square root of the variance, where the variance is the average of the squared differences from the Mean.

standard deviation = $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$

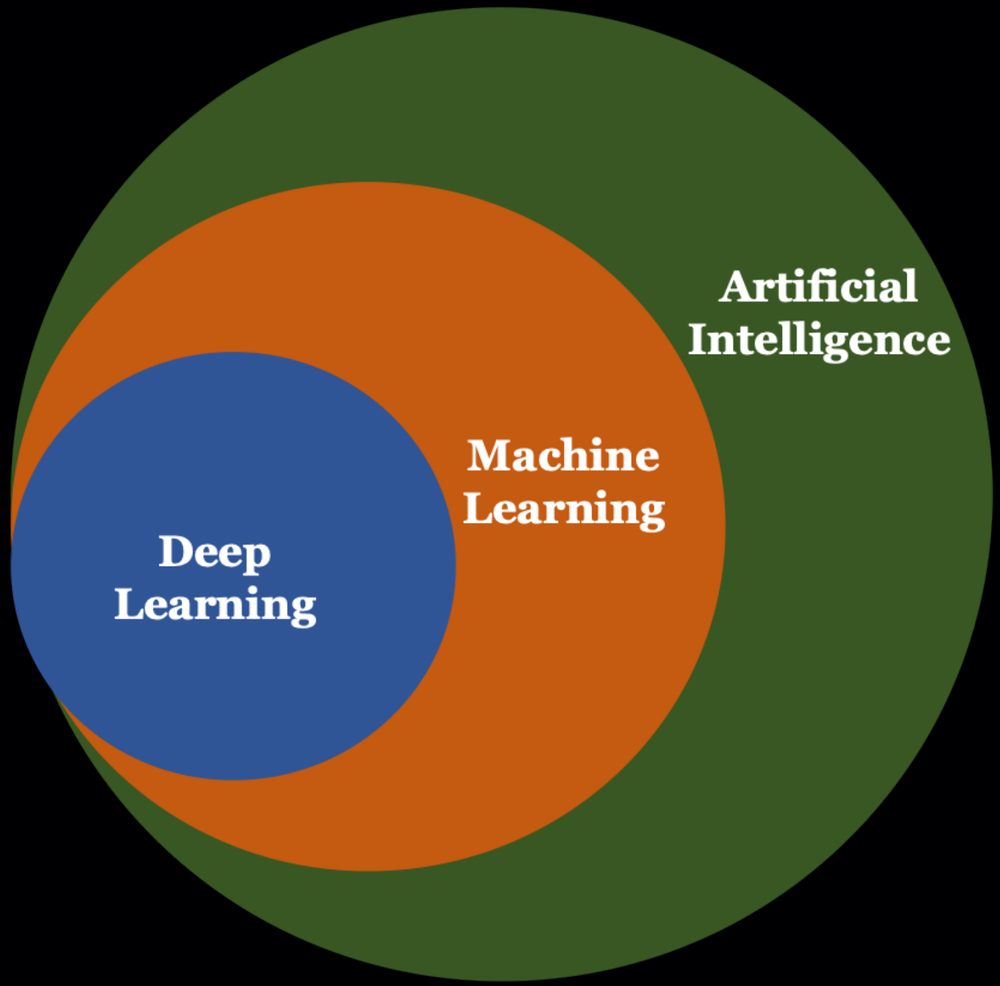
Variance is important in data science as it quantifies the spread of a dataset's values, providing insight into its variability and helping to understand the degree of dispersion around the mean.



Machine learning overview

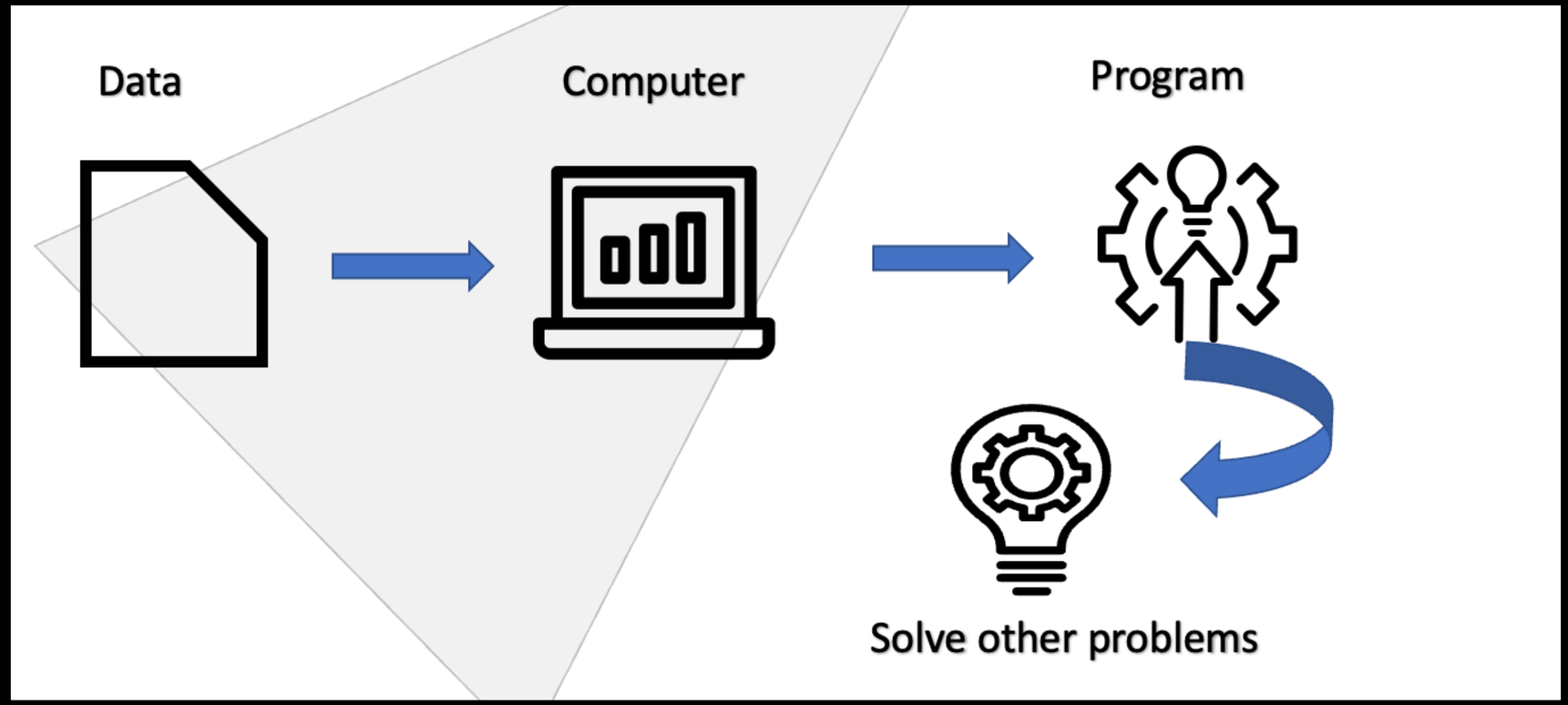
MACHINE LEARNING OVERVIEW

- Machine learning is considered as a subfield of Artificial Intelligence. The idea of Machine learning is that a system learns from the given data, identifies pattern in the data and make decisions with minimal human intervention.



MACHINE LEARNING OVERVIEW

A program that we can use to solve some other problem



APPLICATION OF MACHINE LEARNING

Spam email detection

Movie recommendation

Home value prediction

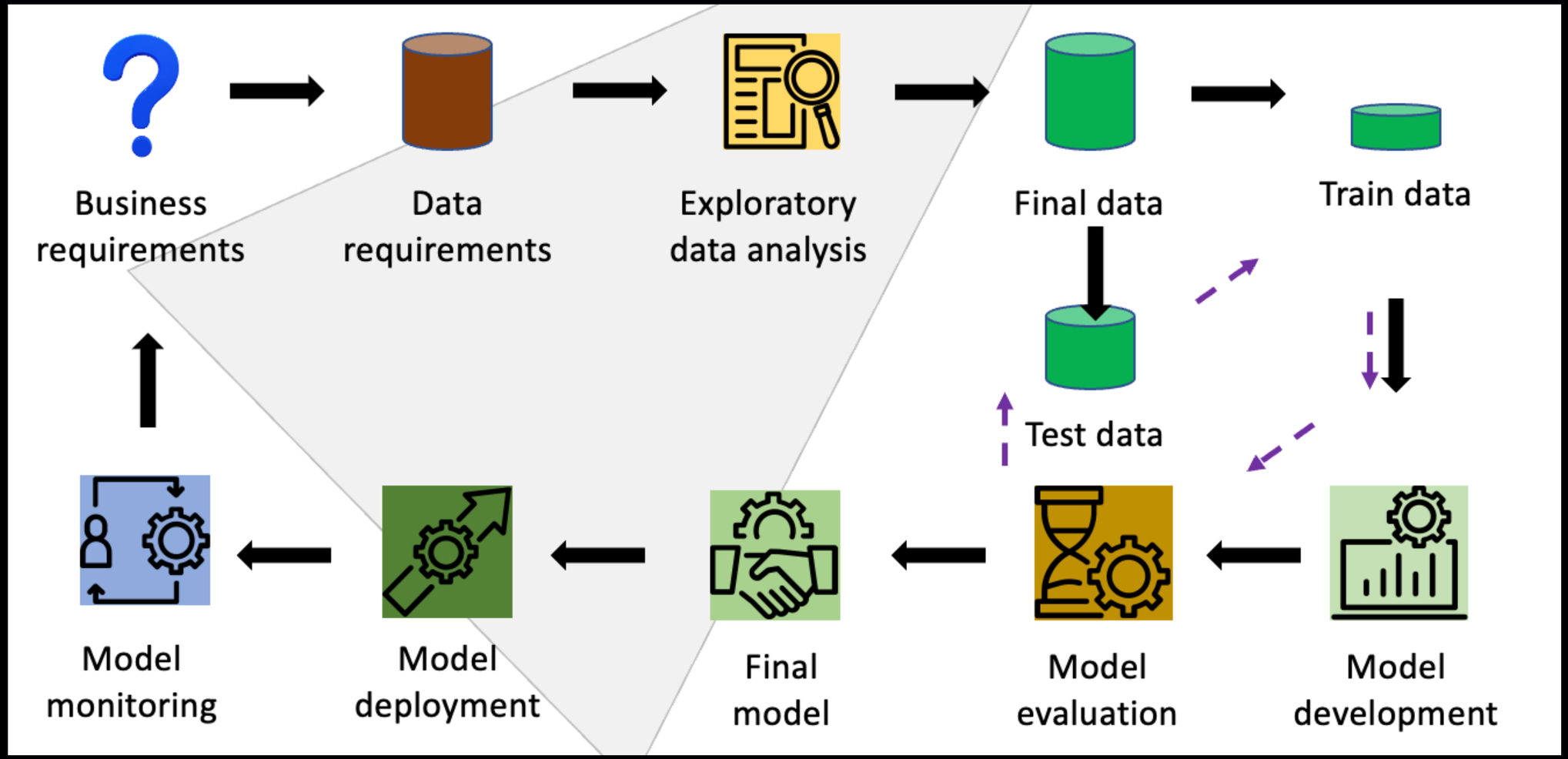
Medical diagnosis

Fraud detection

Speech recognition

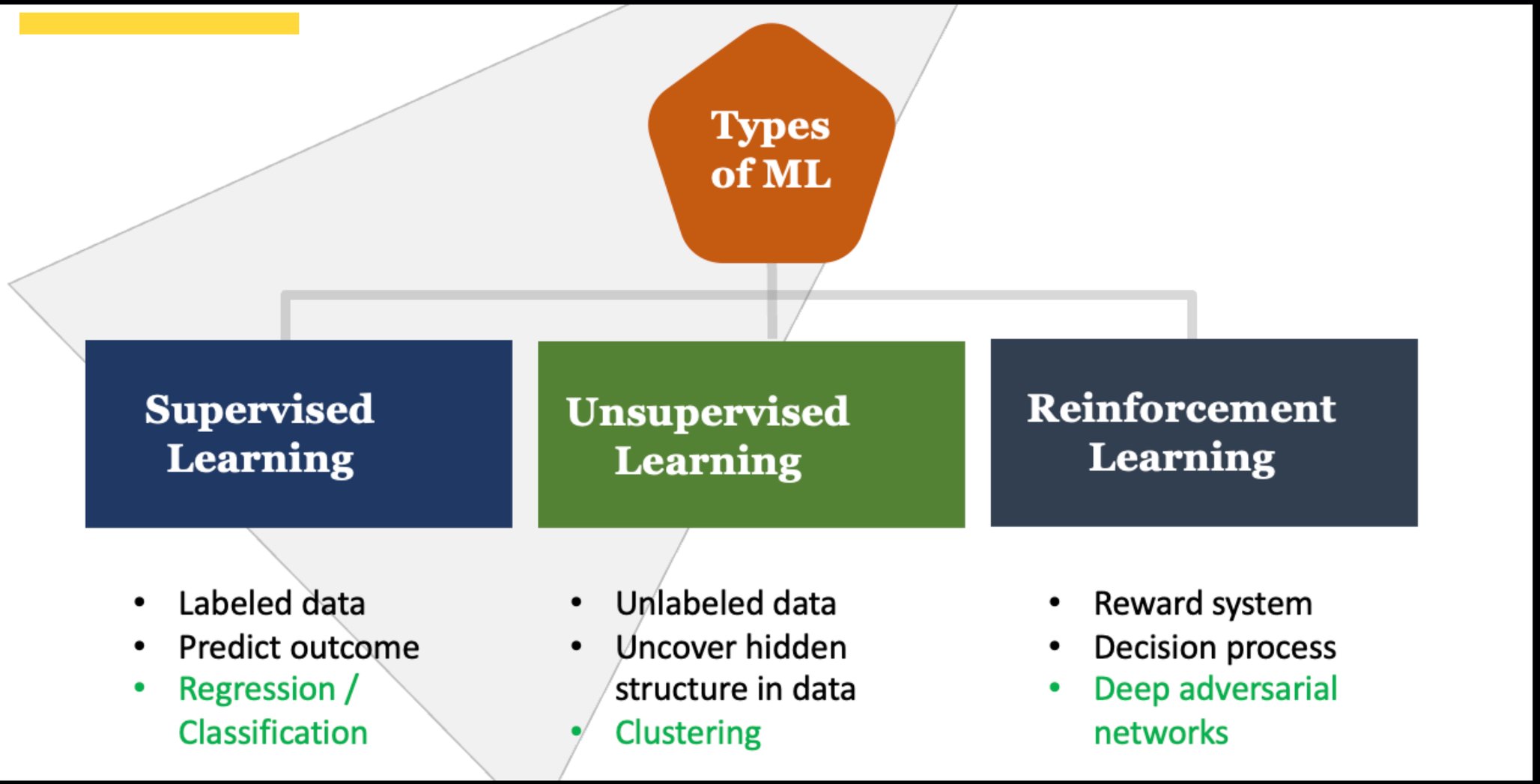
Machine learning lifecycle

ML MODEL LIFECYCLE



Types of ML

TYPES OF ML MODELS



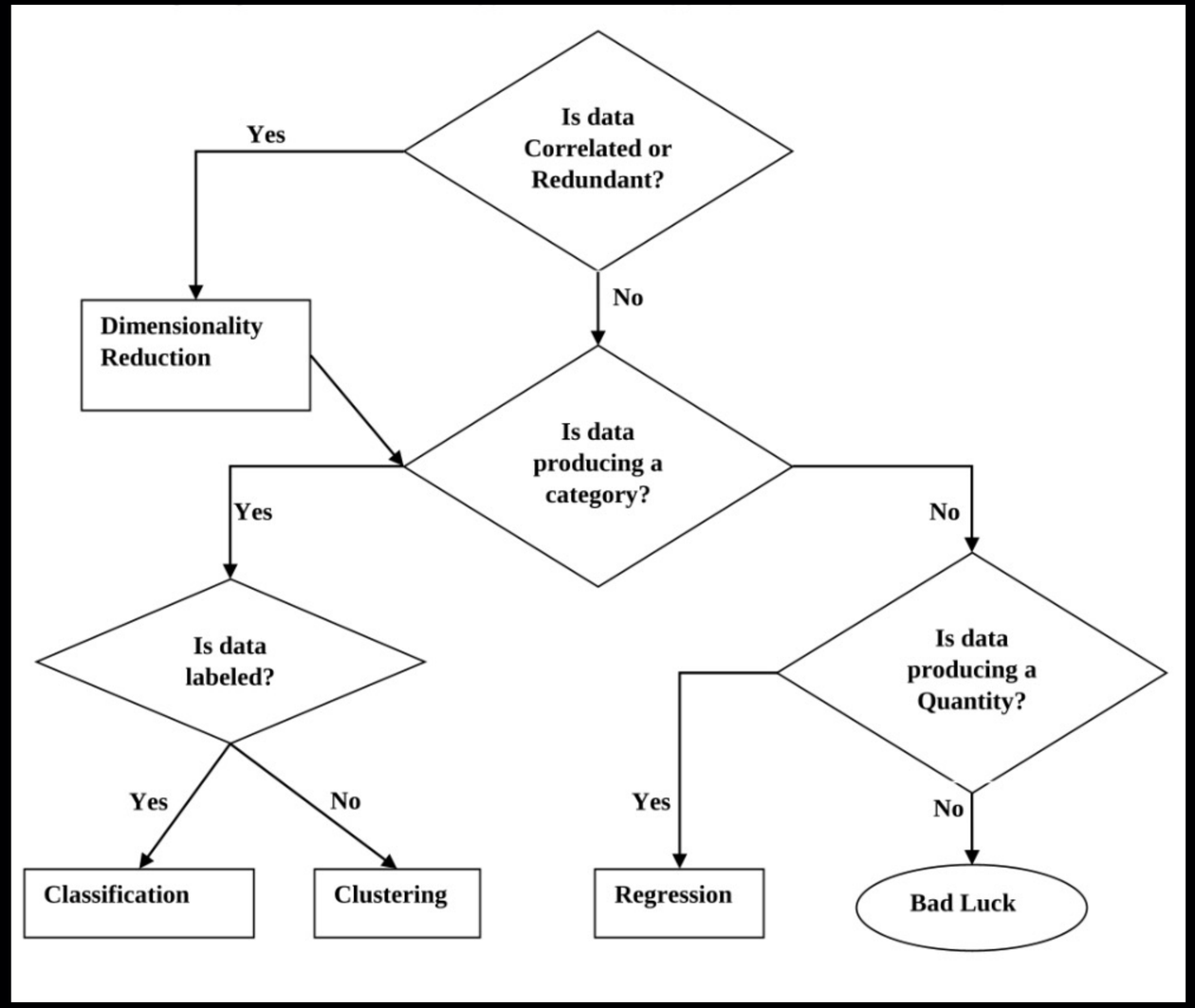
SUPERVISED LEARNING

- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- Support Vector Machine
- Naive Bayes Classifiers

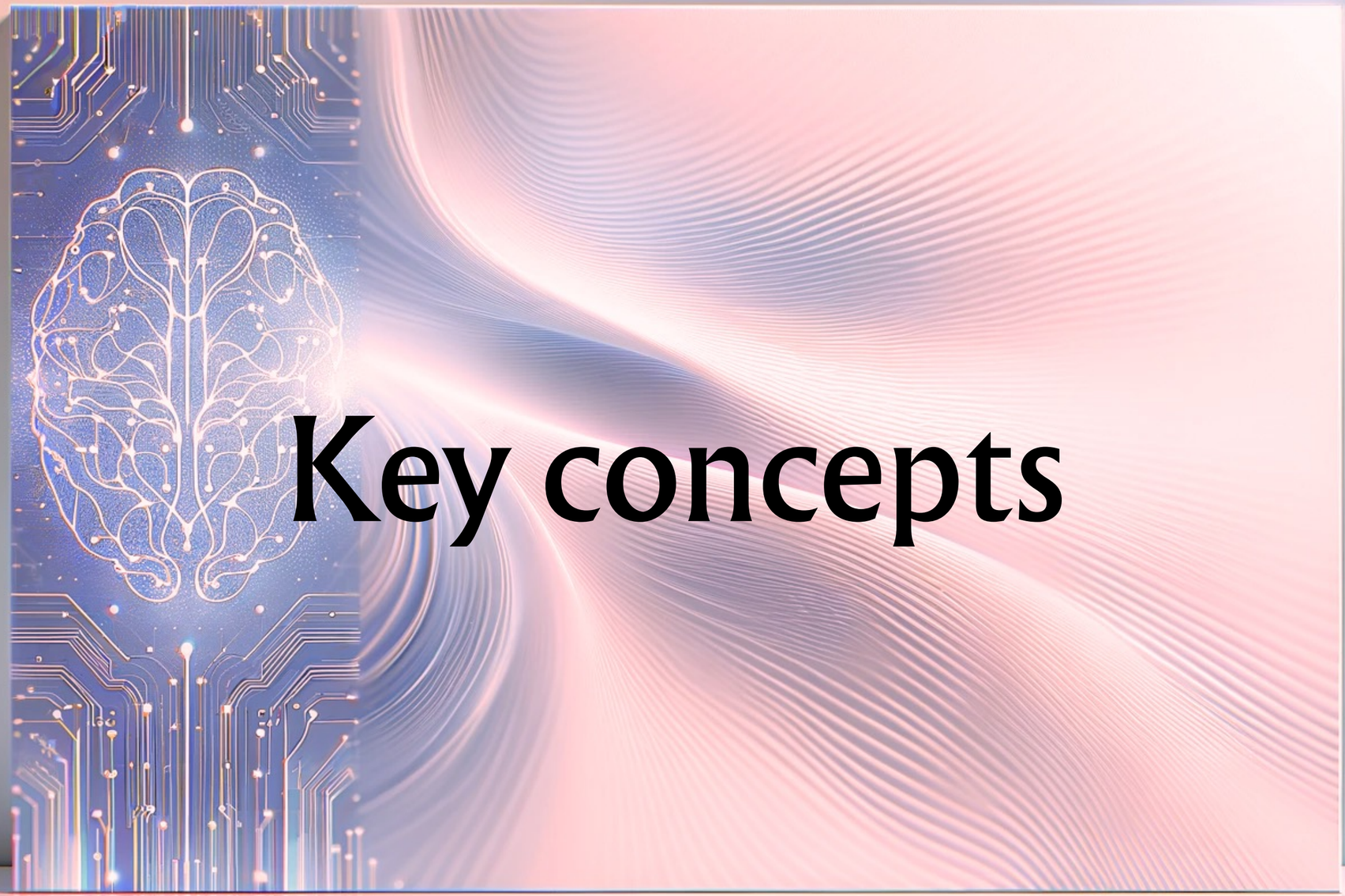
UNSUPERVISED LEARNING

- Clustering
- Principle Component Analysis

TASK SUITED FOR ML



Source: Machine Learning with Python
www.tutorialspoint.com

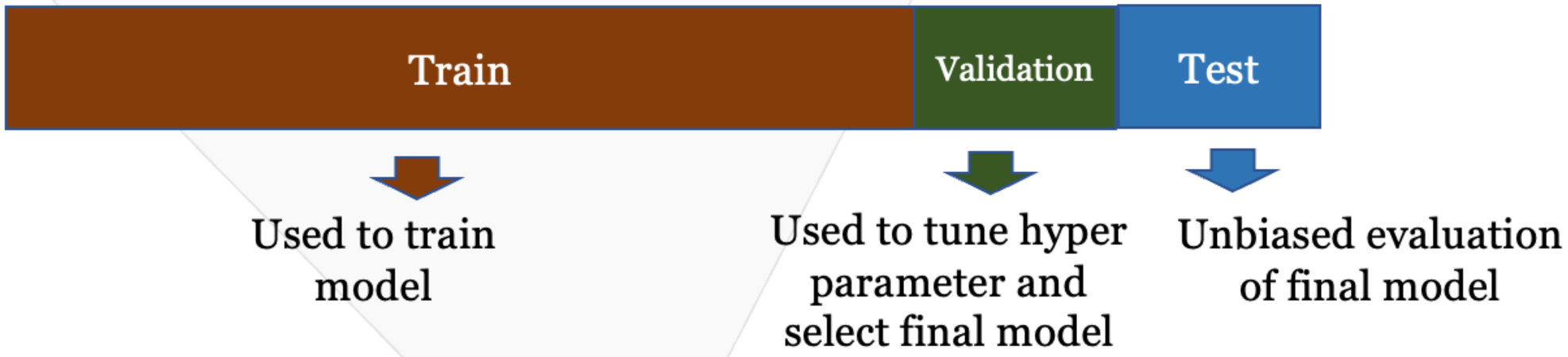

Key concepts

EXPLORATORY DATA ANALYSIS (EDA)

- Data check : count, number of variables, schema/data types
- Outlier
- Missing values
- Summary statistics
- Duplicated records
- Trend
- Sample Vs. Population
- Correlation
- Data clean up (unwanted characters, spaces, formatting..)
- Visualization: boxplot, histogram, bivariate plot

DATA SPLIT

- Before model training, data is split into training and testing sets
- Common split: 80% for training, 10% for validation and 10% for testing




```
from sklearn.model_selection import train_test_split
```

HYPERPARAMETER TUNING

- Some examples
  - Logistic regression classifier : L1 (variable selection) and L2 regularization (overfitting)
  - Neural network: learning rate
  - Decision tree: number of branches

- Two strategies for hyperparameter tunning/search
  - GridSearchCV
  - RandomizedSearchCV

```
from sklearn.model_selection import GridSearchCV, RandomizedSearchCV
```

CROSS VALIDATION (CV)

- Process of splitting the training data into k-folds, then evaluating the model on each fold using the model trained on the remaining folds

10-fold CV

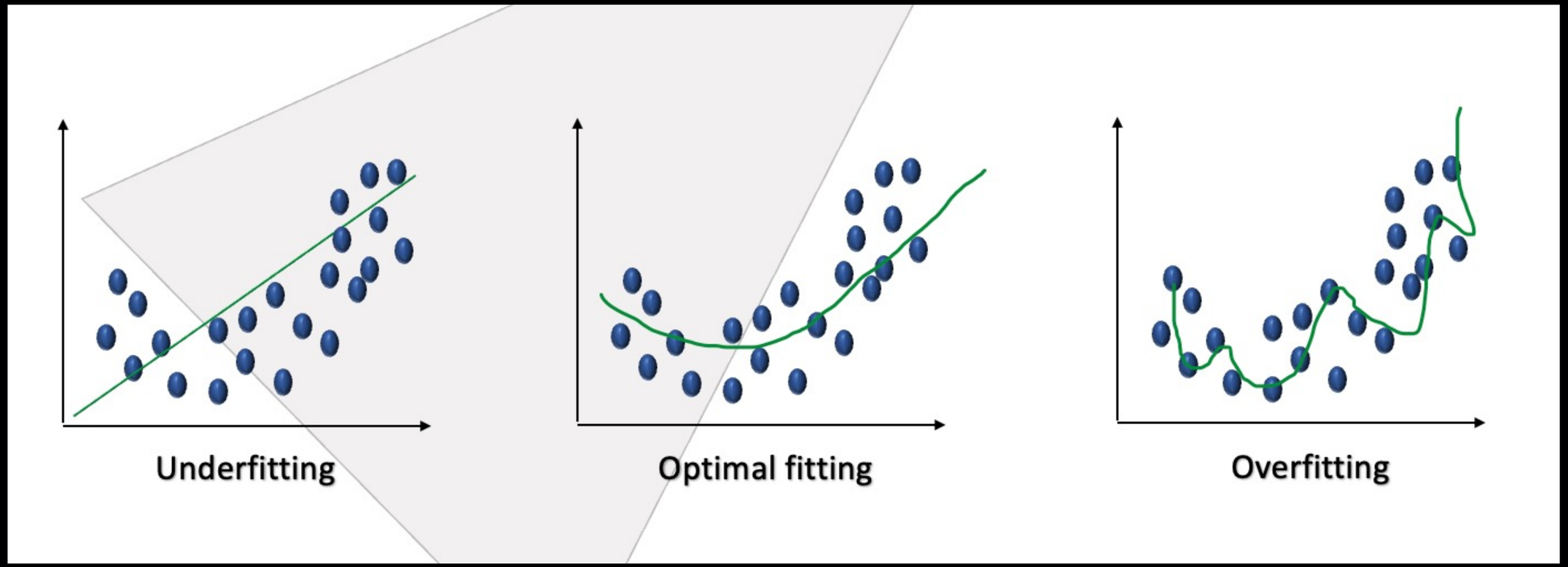
	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Split1	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Split2	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Split3	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Split4	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Split5	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Split6	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Split7	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Split8	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Split9	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Split10	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10

All data

Training

Testing

MODEL UNDER/OVERFITTING



MODEL EVALUATION: CLASSIFICATION

- **Confusion matrix** is a widely used approach for evaluating classification model's performance

		Predicted	
		−	+
Actual	−	TN	FP
	+	FN	TP

TN (True Negative): Correctly classified -ve class

FP (False Positive): Wrongly classified as +ve class

FN (False Negative): Wrongly classified as -ve class

TP (True Positive): Correctly classified +ve class

MODEL EVALUATION: CLASSIFICATION

Spam classificaton

Positive : Spam

Negative: Non-Spam

The goal of the model is to accurately identify spam emails

Cancer detection

Positive : Cancer

Negative: Non-cancer

The goal of the model is to accurately identify the presenece of cancerous cell

MODEL EVALUATION: CLASSIFICATION

Precision

- Precision answers to question:

What proportion of positive prediction was actually correct?

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

		Predicted	
		−	+
Actual	−	TN	FP
	+	FN	TP

MODEL EVALUATION: CLASSIFICATION

Precision

Where to use

We want to minimize FP

MODEL EVALUATION: CLASSIFICATION

Precision

Where to use

We want to minimize FP

FP

Incorrectly predicted spam for non-spam email

MODEL EVALUATION: CLASSIFICATION

Precision

Where to use

We want to minimize FP

FP

Incorrectly predicted spam for non-spam email

Business Impact

For businesses or individuals, missing an important email could mean missing crucial communications or opportunities. Therefore, it's vital that only emails which are very likely to be spam are filtered out.

MODEL EVALUATION: CLASSIFICATION

Recal
l

- Recall answers to question:

What proportion of actual positives was predicted correctly?

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

		Predicted	
		−	+
Actual	−	TN	FP
	+	FN	TP

MODEL EVALUATION: CLASSIFICATION

Recal
l

Where to use

We want to minimize FN

MODEL EVALUATION: CLASSIFICATION

Recal
l

Where to use

We want to minimize FN

FN

Incorrectly predicted non-cancer for cancer cell

MODEL EVALUATION: CLASSIFICATION

Recal
l

Where to use

We want to minimize FN

FN

Incorrectly predicted non-cancer for cancer cell

Health Impact

Missing a diagnosis (a false negative) can be life-threatening in medical contexts. It is crucial to diagnose as many true cases as possible to ensure that patients receive appropriate treatment promptly.

MODEL EVALUATION: CLASSIFICATION

F1- Score

- F1-Score answers to question:

How to take into account for both precision and recall?

$$F1 - Score = 2\,X\,\frac{Precision\ X\ Recall}{Precision + Recall}$$

		Predicted	
		—	+
Actual	—	TN	FP
	+	FN	TP

MODEL EVALUATION: CLASSIFICATION

Accuracy

- Accuracy answers to question:

What is the accuracy/correct prediction of model?

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

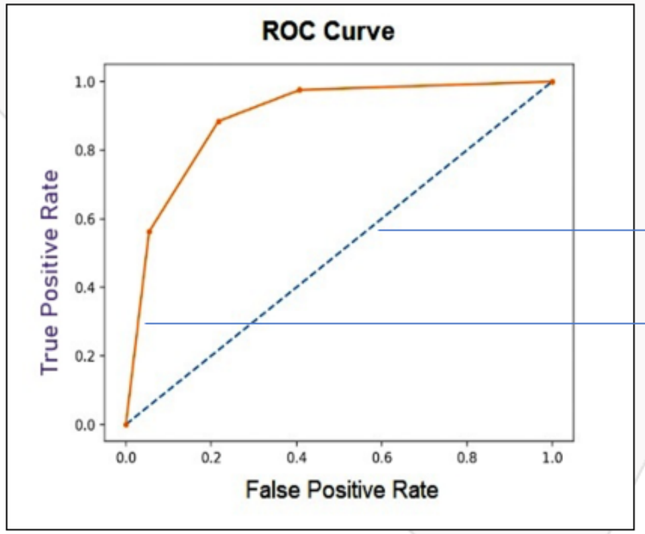
		Predicted	
		−	+
Actual	−	TN	FP
	+	FN	TP

MODEL EVALUATION: CLASSIFICATION

ROC-AUC

- ROC-AUC answers to question:

How well the model is capable of distinguishing between classes?



Source: https://www.askpython.com/python/examples/roc-curves-machine-learning

ROC: Receiver Operating Characteristic Curve
AUC: Area Under Curve i.e. the area that lies under the ROC

AUC = 0.5

Model AUC e.g. 0.8

MODEL EVALUATION: REGRESSION

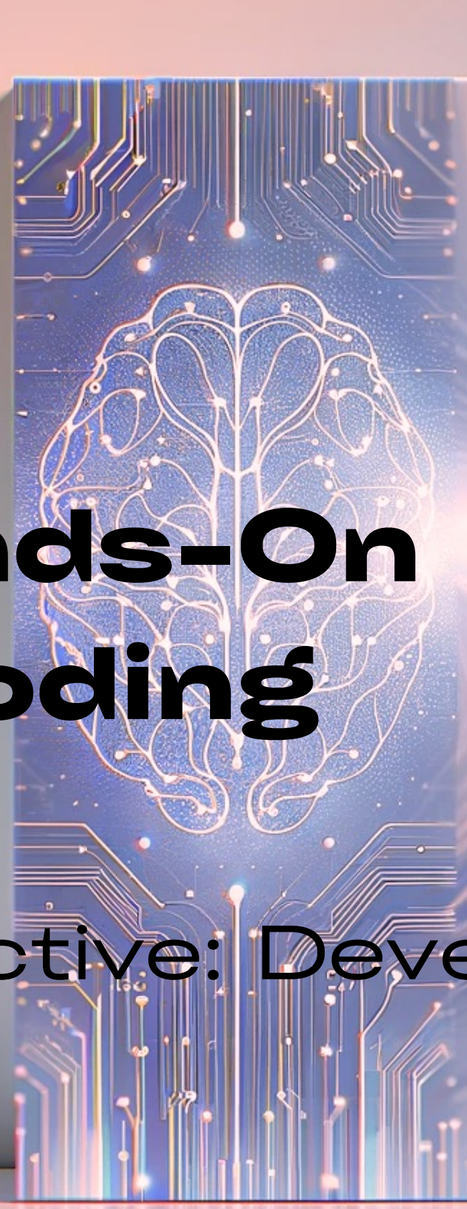
RMSE

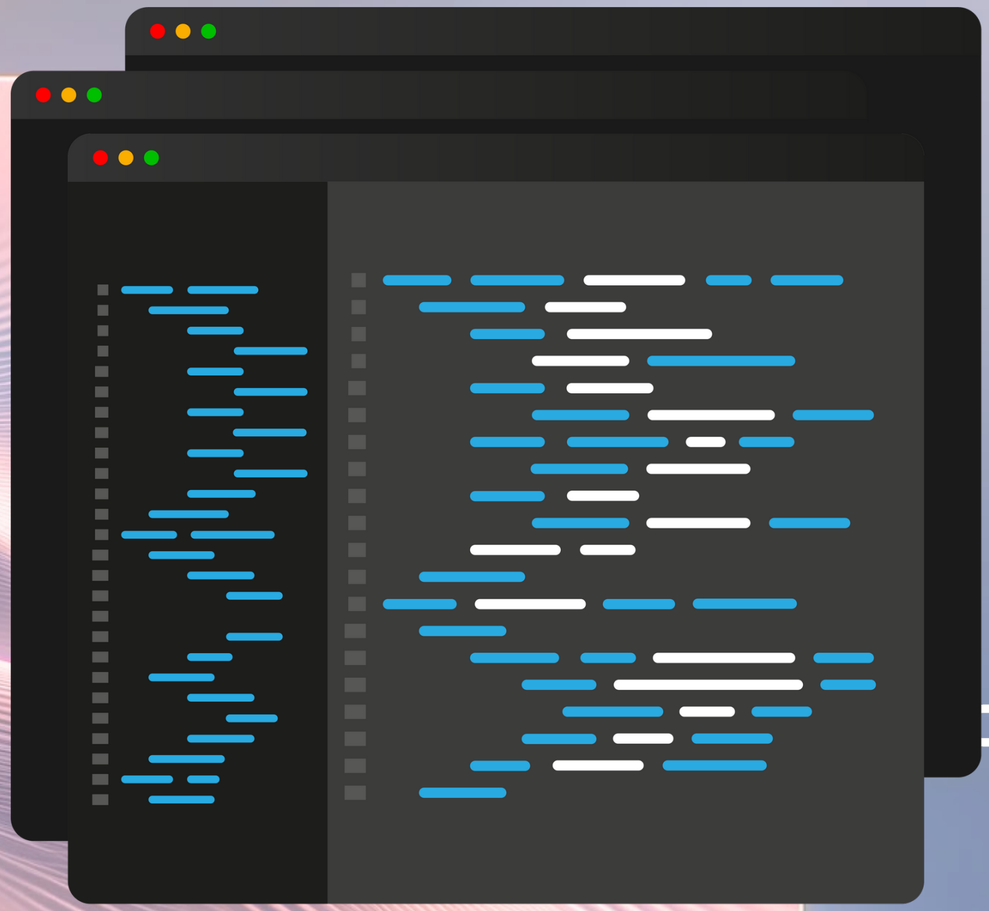
$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(Predicted_i - Actual_i)^2}{N}}$$

Lower the better!

Hands-on coding

Hands-On Coding





Objective: Develop a Machine Learning Model

Health Insurance Cost Prediction Using Regression

INTRODUCTION

This project focuses on applying regression techniques to predict health insurance costs using a real-world dataset. It's an opportunity for you all to delve into the practical applications of data science in the healthcare industry.

OBJECTIVE

The aim is to develop a predictive model that accurately estimates insurance charges based on demographic and health-related factors such as age, BMI, smoking status, and more.

DATASET

Data file name: Insurance.csv

Source: Kaggle

Direct Link:

https://www.kaggle.com/datasets/mirichoi0218/insurance?resource=download.

THANK YOU!

Data Science / AI Lead

Advanced Analytics, Modeling

Instructor: Maven.com

Applied Data Science with Python