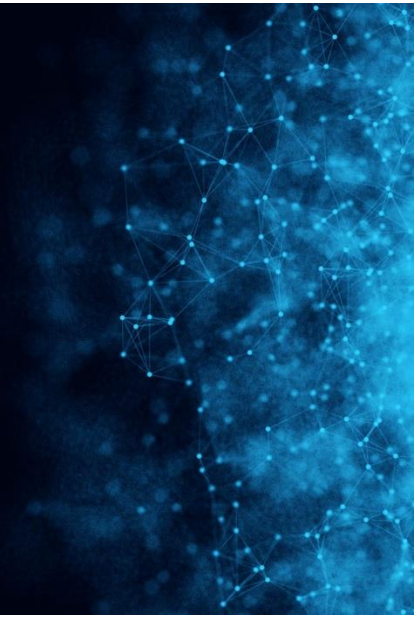


# DA2011
# Machine Learning I

## Lecture 1

Dr. Deshanee Wickramarachchi
Ms. Naethree Premnath
21st September 2025

# Course Aim and Intended Learning Outcomes

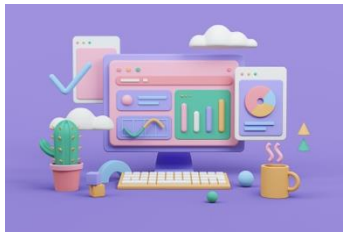
## Course Aim

To provide critical awareness of fundamental concepts in machine learning and hands on experience in problem solving using basic tools and techniques in machine learning.

## Intended Learning Outcomes:

Upon successful completion of the course the student will be able to,

***Demonstrate*** awareness of fundamental concepts in machine learning related prediction problems



***Identify*** and ***apply*** suitable machine learning tools and concepts to solve prediction problems

***validate***, ***interpret*** and ***communicate*** the findings effectively

***demonstrate*** independent learning skills, teamwork skills, and other social skills

# Course Structure

- Credit value – 2C Core
- Course materials - All course materials will be uploaded to LMS.
- Lectures
  - Sundays 8:00am-10:30am over Zoom
  - Once you join the Zoom session, please rename yourself as '24ada<your number>_<first name>'
- Software
  - This course will use Python, accessed via Google Colab for hands-on exercises.
- Recommended Reading
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning (2nd ed). New York: springer
  - Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning. New York: Springer series in statistics.
- Acknowledgement
  - The content of this course draws inspiration from the concepts presented in the lectures by Dr. Pemantha Lakraj at UOC and Prof. Rafael Irizarry at HarvardX.

# Evaluation Criteria

- 100% continuous assessments

| Lab Assignments | Group Project | Individual Reports | Quizzes |
|---|---|---|---|
| 40% | 30% | 20% | 10% |

–Lab Assignments (40%)

| | Date | Assignment on |
|---|---|---|
| Lab Assignment 1 | 19.10.2025 | Regression model building and assessing model accuracy |
| Lab Assignment 2 | 07.12.2025 | Decision Tress and ensemble learning with practical issues |

–Group Project (30%)

| | | |
|---|---|---|
| Interim Discussion | 10% | End of October |
| Final Report | 5% | Early December |
| Final Presentation | 5% | |
| Viva | 10% | |

# Individual Reports (20%)

- After each lecture, you are required to write a **short individual report** not greater than 300 words.
  - Page size: A4, Font size: 12pt, Margin: Normal, Line Spacing: 1.5
  - Font: Times New Roman or Calibri
  - Include your name and Index No at the top
- Your report should include:
  - **Summary:** A concise overview of the key points discussed in the lecture.
  - **Personal Understanding:** What you found most important or interesting.
  - **Limitations:** Aspects that were unclear or areas where you would like further explanation.
- Submission guidelines:
  - Reports should be submitted through the LMS within **48 hours** of the lecture.
  - Deadline: 10:00am on Tuesday
  - Format: PDF file
- This task must reflect **your own understanding** of the lecture content. The use of AI tools is not permitted.

# Quizzes (10%)

- There will be 4 quizzes in total.
- Duration: 15 minutes. Will be open for a 48-hour window on LMS.
- You should start the quiz with sufficient time to finish before the cut-off time.

| Quiz | Starts at 00:00am on | Ends at 11:59pm on | Topics tested |
|---|---|---|---|
| Quiz 1 | 27.09.2025 | 28.09.2025 | Machine Learning Basics |
| Quiz 2 | 11.10.2025 | 12.10.2025 | Simple and Multiple Linear Regression |
| Quiz 3 | 01.11.2025 | 02.11.2025 | Decision Trees, Shrinkage Methods |
| Quiz 4 | 15.11.2025 | 16.11.2025 | Ensemble Learning, Handling Practical Issues |

- Any changes to the assessment plan will be notified to you in advance.

# Suggestions for successful completion of this course

- Attend lectures regularly.
- Read suggested textbooks and any other resources available for examples and exercises.
- Seek help early if you struggle, don’t wait until the end of the module. A forum named ‘Clarify Your Doubts’ will be available on LMS for you to ask any clarifications.

# Topics Covered

1. Introduction to Machine Learning
2. Assessing Model Accuracy
3. Simple Linear Regression
4. Multiple Linear regression
5. Decision Tree Methods
6. Regularization and Shrinkage Methods
7. Ensemble Learning
8. Handling Practical Issues
9. K-Nearest Neighbors (KNN) Regression

# Start-of-course Feedback

https://forms.office.com/r/yZjaidaW4r

# Introduction to Machine Learning

# Today you will learn…

- Introduction to Machine Learning
  - What is Statistical Learning and Machine Learning
  - Outcome vs features
  - Supervised vs unsupervised learning
  - Regression vs Classification
- Assessing Model Accuracy
  - Mean Squared Error (MSE)
  - Training and testing data
  - Overfitting
  - Bias-variance trade-off
- Basics of Python for Machine Learning

# What is Statistical Learning?

- Although the term Statistical Learning is new, most of the concepts under this topic were well developed long time ago.
  - At the beginning of the 19th century, Legendre and Gauss published papers on least squares method.
  - Fisher proposed linear discriminant analysis in 1936.
  - In the 1940s, logistic regression was introduced.
  - In the early 1970s, Nelder and Wedderburn introduced generalized linear models which contains both linear and logistic regression as special cases.
  - In mid 1980s Breiman, Friedman, Olshen and Stone introduced classification and regression trees.
- In general, statistical learning refers to a set of statistical tools for modelling and understanding complex data.

# What is Machine Learning?

- Closely related to Statistical Learning.
- Statistical learning focuses more on **understanding the relationship** between variables through inferential methods subject to multiple assumptions.
  - –Regression, ANOVA, Hypothesis testing
- Machine learning focuses more on **prediction** and **learning from data**.
  - –Regression, Classification, Clustering
- Machine learning methods are more suitable for large, complex and unstructured data.

For further reading: https://www.coursera.org/articles/statistical-learning-vs-machine-learning

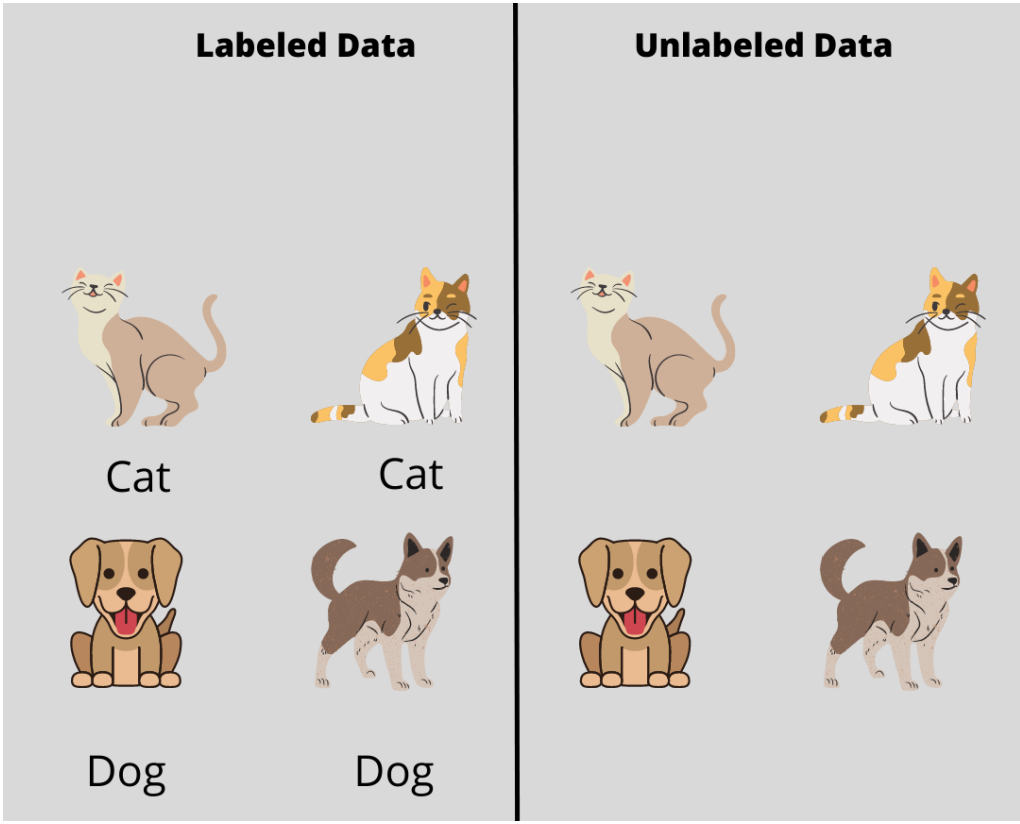
# What is Machine Learning used for?

- Predictive Analytics
  - –regression models, decision trees, neural networks etc.
- Image Recognition
  - –identify defects in product images, classify images and computer vision tasks
- Natural Language Processing (NLP)
  - –speech recognition, content generation, document processing
- Recommendation Systems
  - –to determine the best products and services for individuals
- Autonomous Systems
  - –reinforcement learning: self-driving cars

For further reading: https://www.coursera.org/articles/statistical-learning-vs-machine-learning

# Supervised vs unsupervised learning

- Machine learning falls into two main categories: supervised and unsupervised
- Supervised Learning:
  - Model learns from labelled data
  - The actual outcome is known for the data used to train the model
  - Eg: Suppose you want to train an image classifier to identify cats and dogs. You would provide labelled images where each image is labelled as a cat or a dog.
- Unsupervised Learning:
  - The model is not given labelled data.
  - It should discover patterns or structures within the data without any guidance.



# Outcome vs features

- In supervised learning, data comes in the form of outcomes and features.
  - –Outcome: what we want to predict
  - –Features: what will be used to predict the outcome
- Notation:

| $Y$ | $X_1, X_2, \ldots, X_p$ |
|---|---|
| **Outcome/Response**<br>**Dependent variable** | **Features/Predictors/Covariates**<br>**Independent variables** |

- We develop algorithms that take in the features and predict the outcome.
- Machine learning techniques are used to train these algorithms using observed data, so that they can be used to predict the outcome when it is unknown.

# Training Dataset View

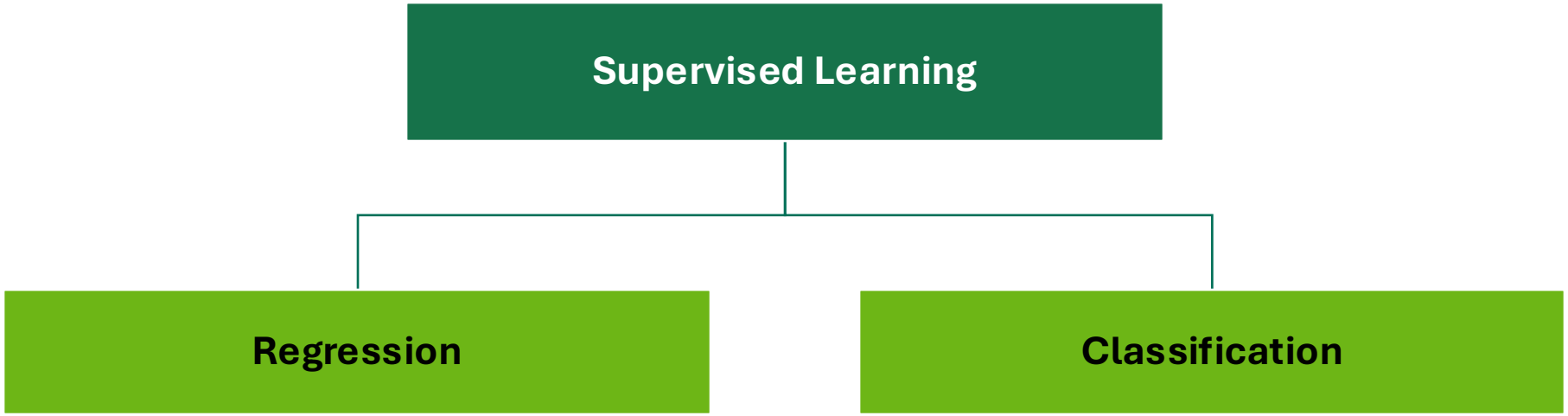
- A typical dataset would look like:

| Unit | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|---|
| 1 | 24 | 1 | 1.2 | 100 | 0.2 |
| 2 | 30 | 0 | 2.0 | 120 | 0.4 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $n$ | 22 | 1 | 1.5 | 80 | 0.1 |

- In notations:

| Unit | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|---|
| 1 | $y_1$ | $x_{11}$ | $x_{12}$ | $x_{31}$ | $x_{41}$ |
| 2 | $y_2$ | $x_{21}$ | $x_{22}$ | $x_{32}$ | $x_{42}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $n$ | $y_n$ | $x_{n1}$ | $x_{n2}$ | $x_{n3}$ | $x_{n4}$ |

# Regression vs Classification




- Outcome is **continous**
- Eg:
  - –Predicting stock prices
  - –Estimating house prices

- Outcome is **categorical**
- Eg:
  - –Classify image as dog or cat
  - –Detect handwritten digits

# Group Activitiy

- First work on this activity individually. (2 minutes)
- Then, you will be assigned to breakout rooms. Discuss your answers with your peers and agree to one set of solutions. Apoint one team leader to present. (5-6 minutes)
- Once we are back in the main class, be ready to present your answers.

- State whether each of the following scenarios falls into regression or classification.

1. Estimating a student's final exam score from their assignment marks and attendance.
2. Predicting whether a patient has diabetes (Yes/No) based on medical test results.
3. Predicting if a customer will churn (leave) or stay with a company.
4. Forecasting sales revenue for the next quarter.
5. Determining if an email is spam or not spam.
6. Forecasting sales revenue for the next quarter.
7. Categorize customer reviews as positive, neutral, or negative.
8. Predicting house prices based on location, size, and features.

# Assessing Model Accuracy

# Prediction Models

- Suppose we have $n$ observations in the dataset. $Y_i$ denotes the outcome and $X_i = (X_{i1}, X_{i2}, \dots, X_{in})$ denotes the $p$ predictors for the $i^{\text{th}}$ unit.
- The observed values are usually denoted with simple letters.
- We believe that there is some relationship between the outcome and the predictors.

$$Y_i = f(X_i) + \varepsilon_i \quad \text{for } i = 1, \dots, n$$

  - $f$ is fixed but unknown
  - $\varepsilon_i$ is random error that captures other discrepencies
- We need to estimate the unknown function $f$ using the observed data.

# Mean Squared Error

- The estimated function is denoted by $\hat{f}(x)$.
- The predicted outcome for the $i^{\text{th}}$ unit based on $\hat{f}(x)$ is:

$$\widehat{y_i} = \hat{f}(x_{i1}, \ldots, x_{ip})$$

- We want to predict the outcome as close as possible to the actual value $y$. In other words, we want to minimize the error:

$$y - \hat{y}$$

- In regression, the most commonly used  measure to assess the model accuracy is the Mean Squared Error (MSE).

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

# Training and Testing Data

- After fitting a model to the observed data, we must evaluate whether it truly works by testing its ability to generalize to unseen data.
- We cannot use the same data that we used to train the model, for testing. We need new data, that the model hasn't seen earlier.
- This is usually achieved by splitting the observed data into two parts: training and testing data.
  - We may use 80% of the observed data to train the model and the remaining 20% to test the performance of the model.
  - Other splits, such as 75%–25% or 70%–30%, are also commonly used.
- Python libraries have inbuilt functions to randomly split the dataset into train and test sets. Eg: `scikit-learn` -> `train_test_split` function

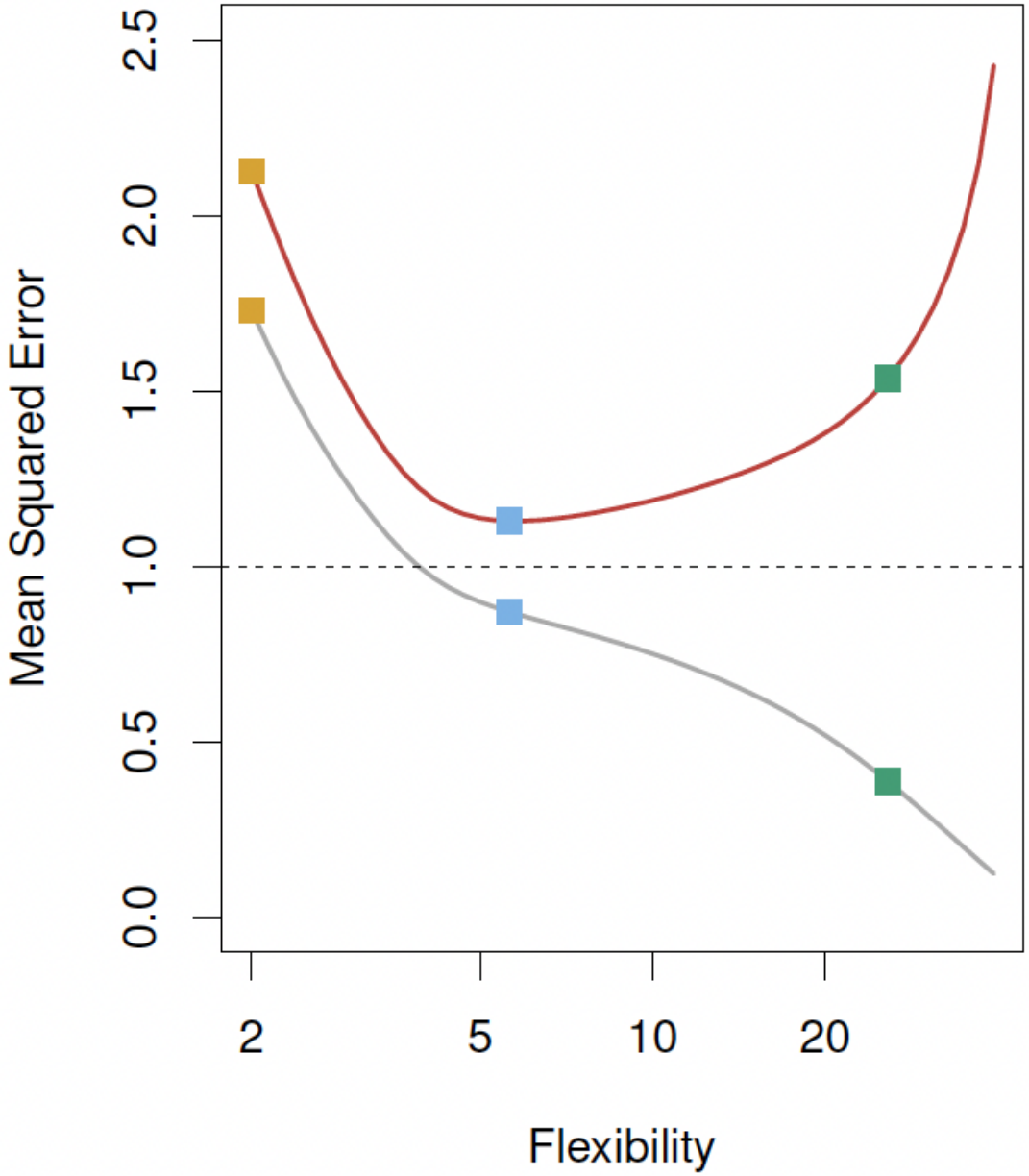
# Overfitting

- Sometimes a model will perform well on the training data with small MSE, but perform badly on the test data with large MSE.
- This is called the overfitting problem.
- Reasons:
  - Too complex models
  - Insufficient data or poor quality of data
- Ways to overcome this common problem of overfitting will be addressed later in this course.

# Example

- We are comparing three ML models: orange, blue and green squares.
- Grey line: Training MSE
- Red curve: Test MSE

- Which model is the best?



# Bias-variance trade-off

- The U-shape that is observed in the test MSE curve is the result of two competing properties of statistical learning models: Bias and Variance
- The expected test MSE can be decomposed into three parts. [Mathematical proof is beyond the scope of this course]
  - Variance of $\hat{f}(x)$
  - Squared bias of $\hat{f}(x)$
  - Variance of random error
- In order to minimize the expected test MSE we should simultaneously minimized the variance and squared bias.
- More flexible and complex models have higher variance and low bias.
- We need to find a balance between these to achieve a low test MSE.

# Python Basics for ML

# Essential Python libraries

- NumPy (**Num**erical **Py**thon )
  - –the core library for numerical computing in Python
  - –provides a high-performance multi-dimensional array object (*np.array*), and tools for working with data structures
  - –*NumPy basics for beginners:* *https://numpy.org/doc/stable/user/absolute_beginners.html*
- SciPy
  - –SciPy builds on NumPy and provides many functions that operate on NumPy arrays
  - –SciPy contains multiple subpackages catering to different scientific computing domains. Eg: scipy.*stats*
  - –*scipy.stats User Guide:* *https://docs.scipy.org/doc/scipy/tutorial/index.html#*

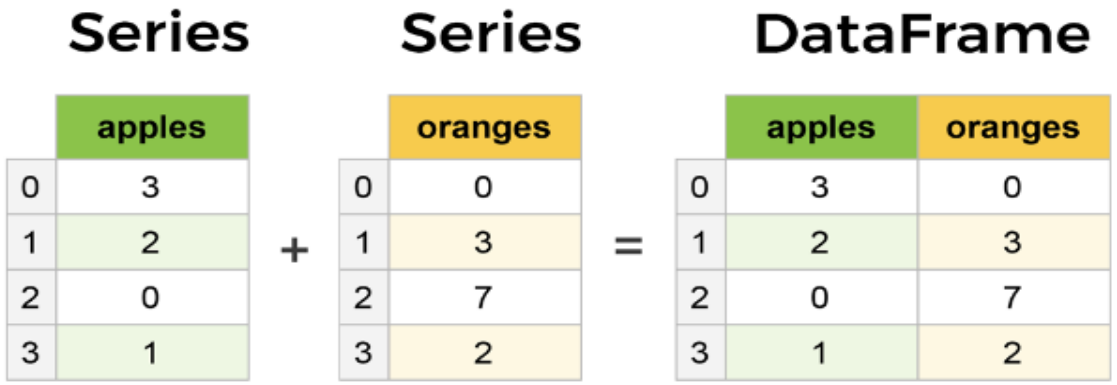
# Essential Python libraries

- matplotlib
  - –most commonly used Python library for data visualization
  - –matplotlib.pyplot module facilitates a plotting system similar to that of MATLAB.
  - –*matplotlib User Guide: https://matplotlib.org/stable/users/index*

- sklearn
  - –Scikit-learn, also known as sklearn, is an open-source, robust Python machine learning library.
  - –enables practitioners to rapidly implement a vast range of supervised and unsupervised machine learning algorithms through a consistent interface.
  - –sklearn User Guide: https://scikit-learn.org/stable/user_guide.html

# pandas

- The pandas library is one of the most important tools at the disposal of Data Scientists and Analysts working with Python.
- Core components of pandas: Series and DataFrames



# Thank you
# See you next week!