Data Analysis with Pandas

Data Analysis

- Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

Example:

- **Analyzing Sales Trends:** Finding the month with the highest revenue
- **Tracking Fitness Progress:** Analyzing daily steps and calories

What is Data Manipulation and Analysis?

- **Data Manipulation**
  - **Definition:** Changing, organizing, or preparing data to make it useful and easier to understand.
  - **Goal:** To clean, transform, and structure raw data for better usability.
  - **Example:**
    - **Organizing a Grocery List:** Sorting random items into categories like "Fruits" or "Dairy".
    - **Fixing Errors in a Student Record:** Correcting missing or wrong grades.

Key Differences Between Data Manipulation and Analysis

Aspect	Data Manipulation	Data Analysis
Focus	Preparing and cleaning data	Extracting insights from prepared data
Goal	Organize and structure raw data	Find patterns, trends, and solve problems
Example	Fixing errors in a student's grade sheet	Analyzing which student scored the highest

What is Pandas?

- **Pandas** is a powerful and popular Python library designed for **data manipulation** (cleaning, transforming, and structuring data) and **data analysis** (finding patterns, trends, and insights).
- It simplifies working with structured datasets like tables, spreadsheets, or time-series data.

What Makes Pandas Unique?

- **Key Features:**
  a. Works seamlessly with structured data formats like CSV and Excel.
  b. Handles missing values easily.
  c. Built on NumPy for fast computations.

Why Use Pandas?

a. **Performance:** Handles millions of rows efficiently.
b. **Ease of Use:** Beginner-friendly syntax for cleaning and transforming data
c. **Integration:** Works with libraries like Matplotlib (visualizations) and Scikit-

Real-Life Examples of Pandas in Action

Finance:

- Analyzing time-series data like stock prices to identify market trends.

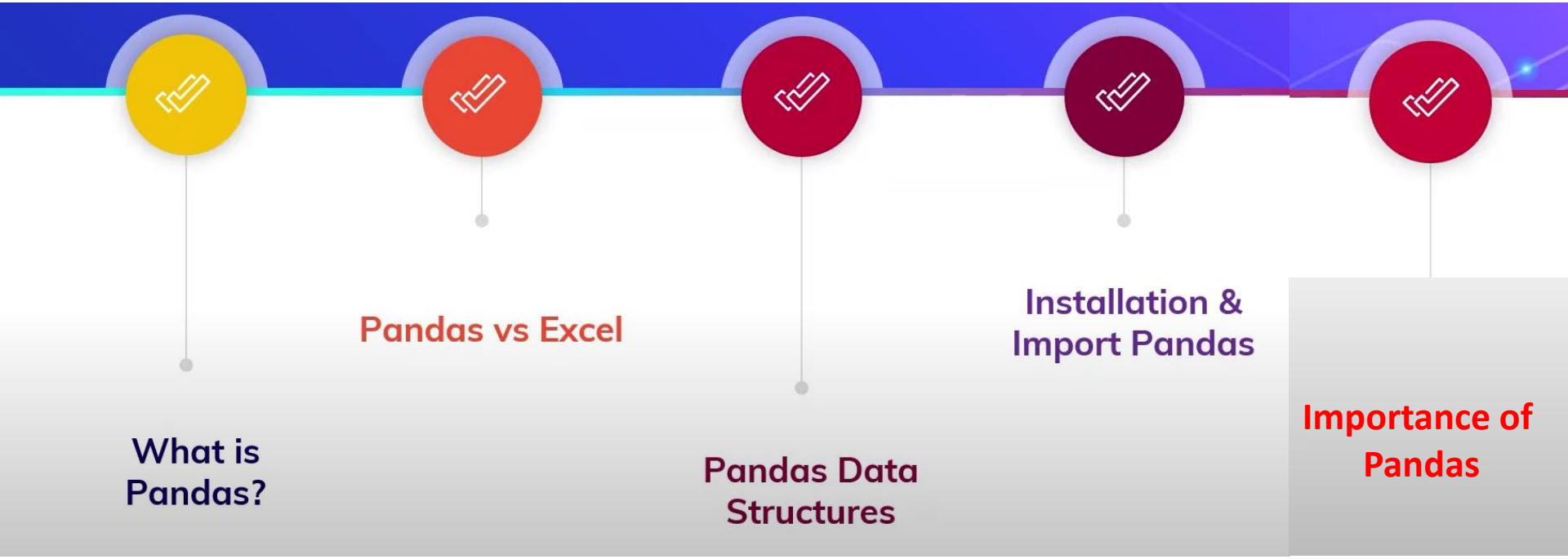
Retail:

- Tracking inventory and finding the most sold products in a store.

Healthcare:

- Analyzing patient records and outcomes from clinical trials.

Python Libraries for Data Analysis


What is Pandas?
Pandas vs Excel
Pandas Data Structures
Installation & Import Pandas
Importance of Pandas

What is Pandas?

- The name "pandas" has a reference to both "panel data", and "python data analysis" and was created by wes mckinney in 2008.
- Pandas is a python library used for working with data sets.
- It has functions for analyzing, cleaning, exploring, and manipulating
- Read and write data structures and different formats : csv, XML, JSON,ZIP etc.

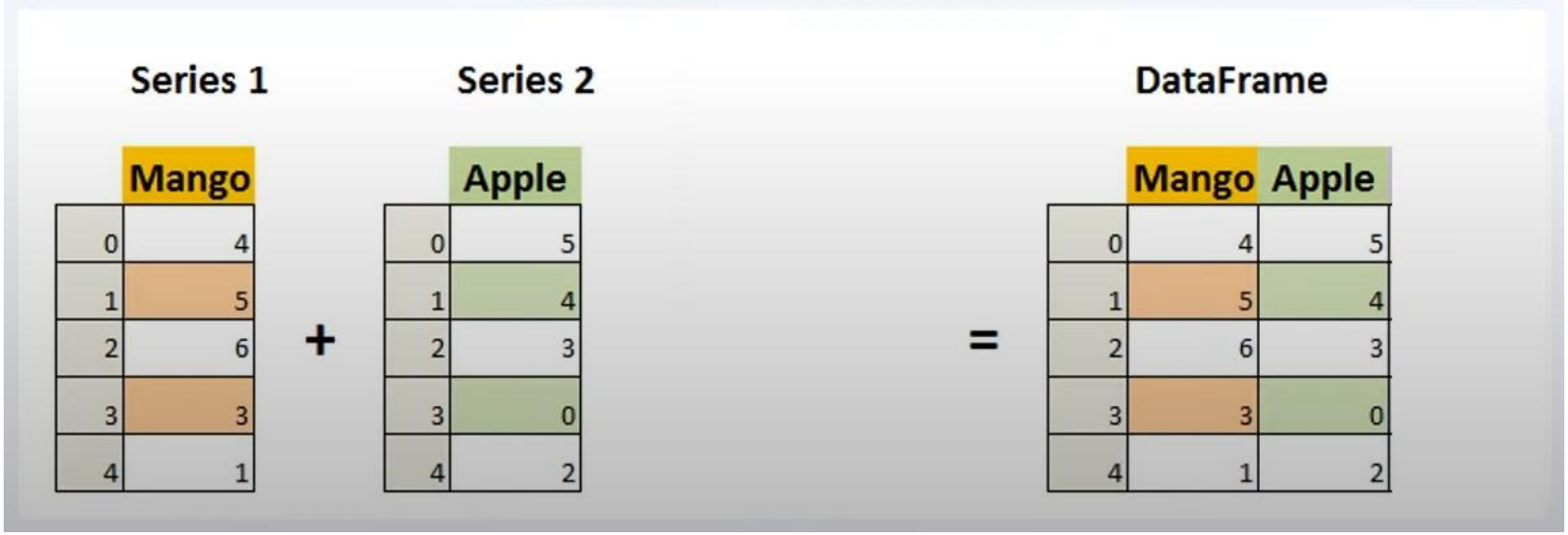
Pandas Data Structures

Three Data Structures:

Series: One-Dimensional labeled arrays pd.Series(data)

DataFrames: Two-Dimensional data structure with columns, much like a table.

Panel : A panel is a 3D container of data.



Importance of Pandas in Python

- Pandas allows us to analyze big data and make conclusions based on statistical theories.
- Pandas can clean messy data sets, and make them readable and relevant.
- Easy handling of missing data (represented as NaN) in floating point as well as non-floating point data
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects
- Data set merging and joining. Flexible reshaping and pivoting of data sets Provides time-series functionality.

Data Structures in Python Pandas

- The Pandas provides two data structures for processing the data.

Series & DataFrame and Panel

Series

- **Series :** It is defined as a one-dimensional array that is capable of storing various data types.

```
import pandas as pd
a = pd.Series( )
print(a)
```

Key Pandas Concepts

DataFrame:

- A **DataFrame** is a **two-dimensional labeled data structure** in Pandas, similar to a table in a database, an Excel spreadsheet, or a SQL table.
- It consists of **rows** and **columns**, where:
  a. Rows have indices (labels).
  b. Columns have names (labels).

Installation & Import Pandas

INSTALLATION

pip install pandas

IMPORT

Import pandas as pd

- Creating Series using other objects like dictionary or tuple:

Works with Missing data also

- NumPy gives broad-casting error whereas Pandas handles missing data

DataFrame in Python Pandas