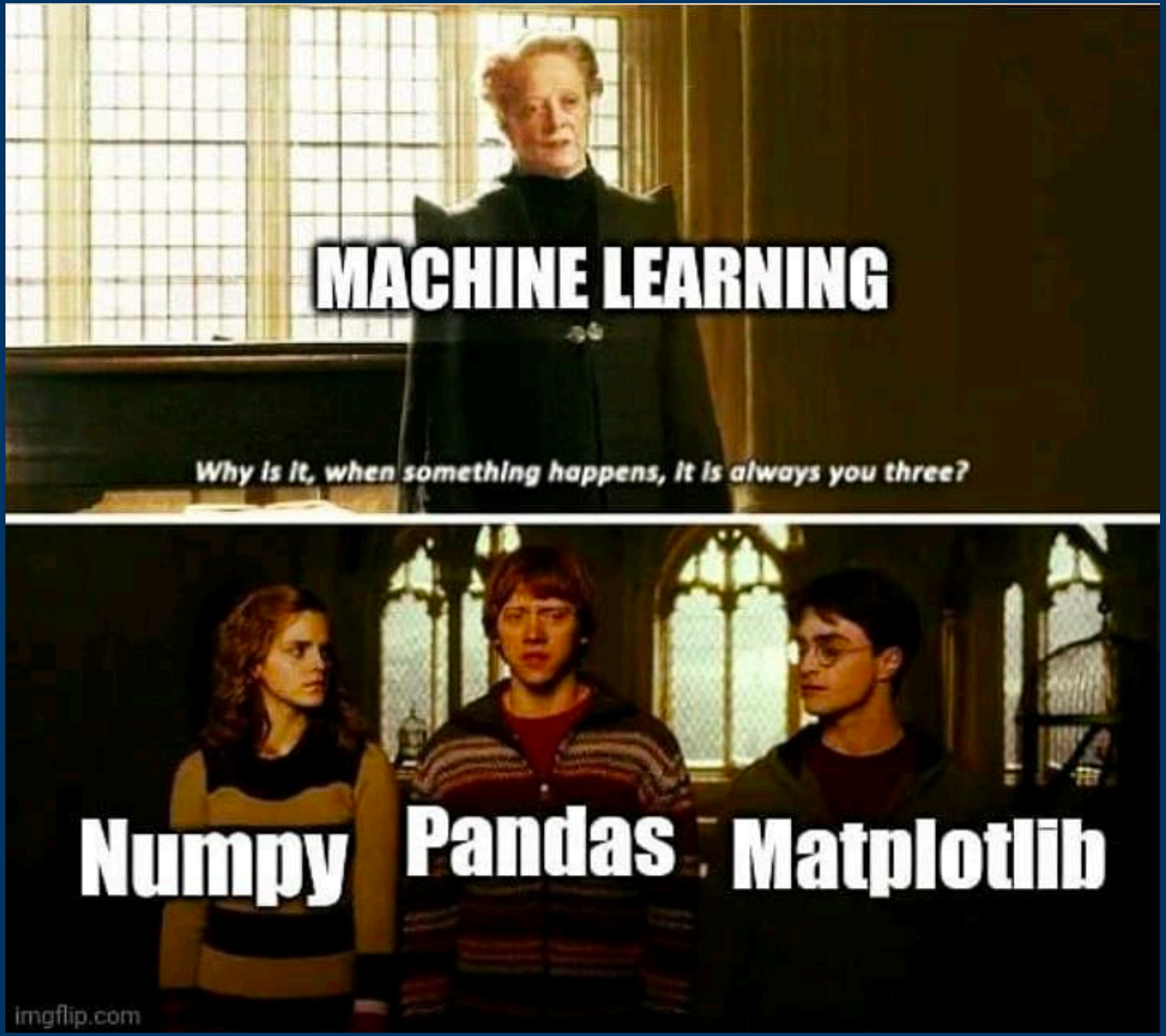
Recitation 5

CS - 210



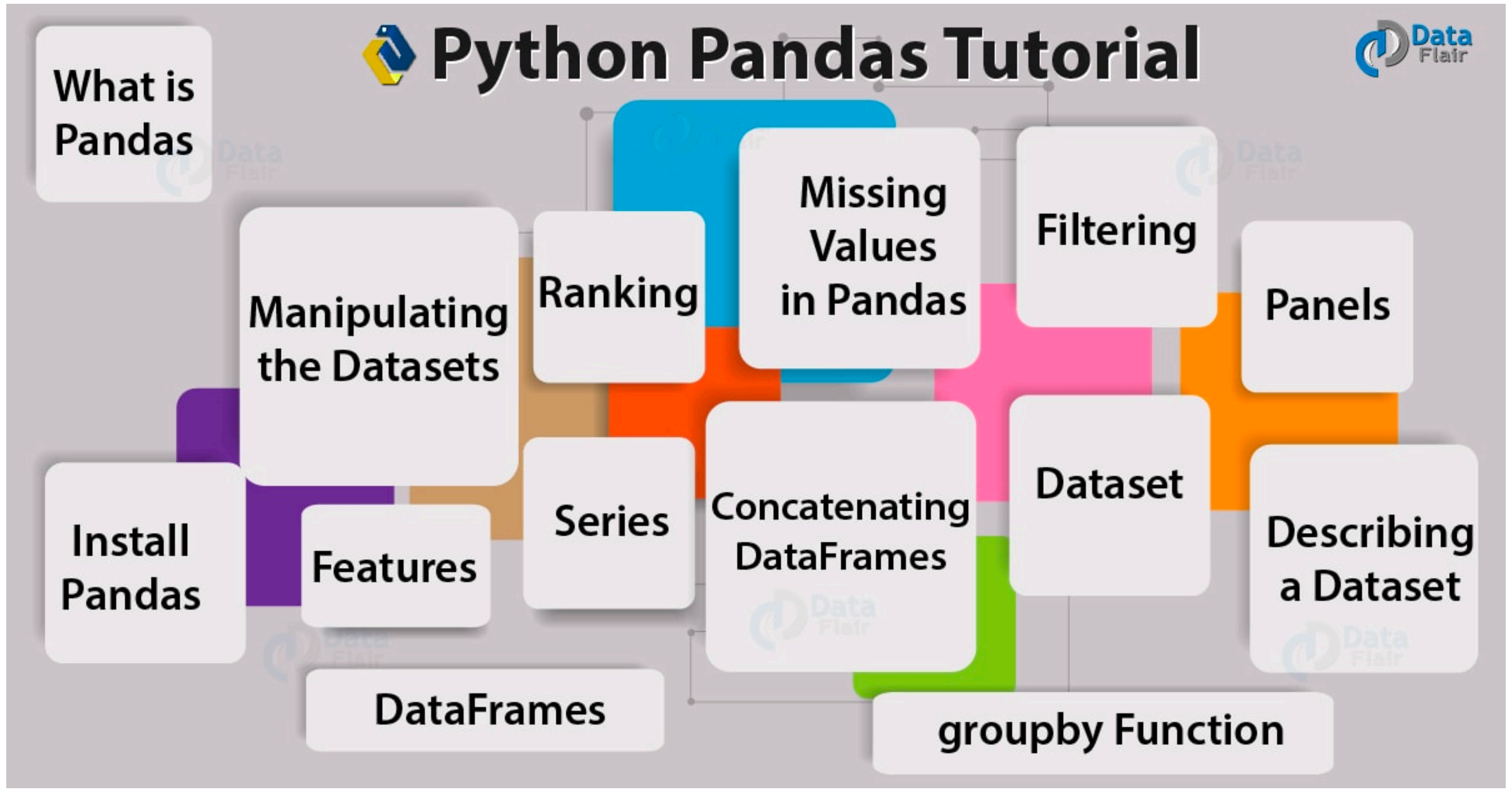
Anamika Lochab

Pandas

A data analysis tool for Python

- "pandas is an open-source library providing high-performance, easy-to-use data structures, and data analysis tools for Python."

- Name derived from the term "panel data", an econometrics term for datasets that include observations over multiple periods of time for the same individuals.



Flexibility and Versatility

It can handle Any Data

- Reads data from varied formats: CSV, Excel, SQL databases, and more.
- Wide array of functions to filter, slice, and dice data.
- Easily handles missing data.

Rich Functional Set

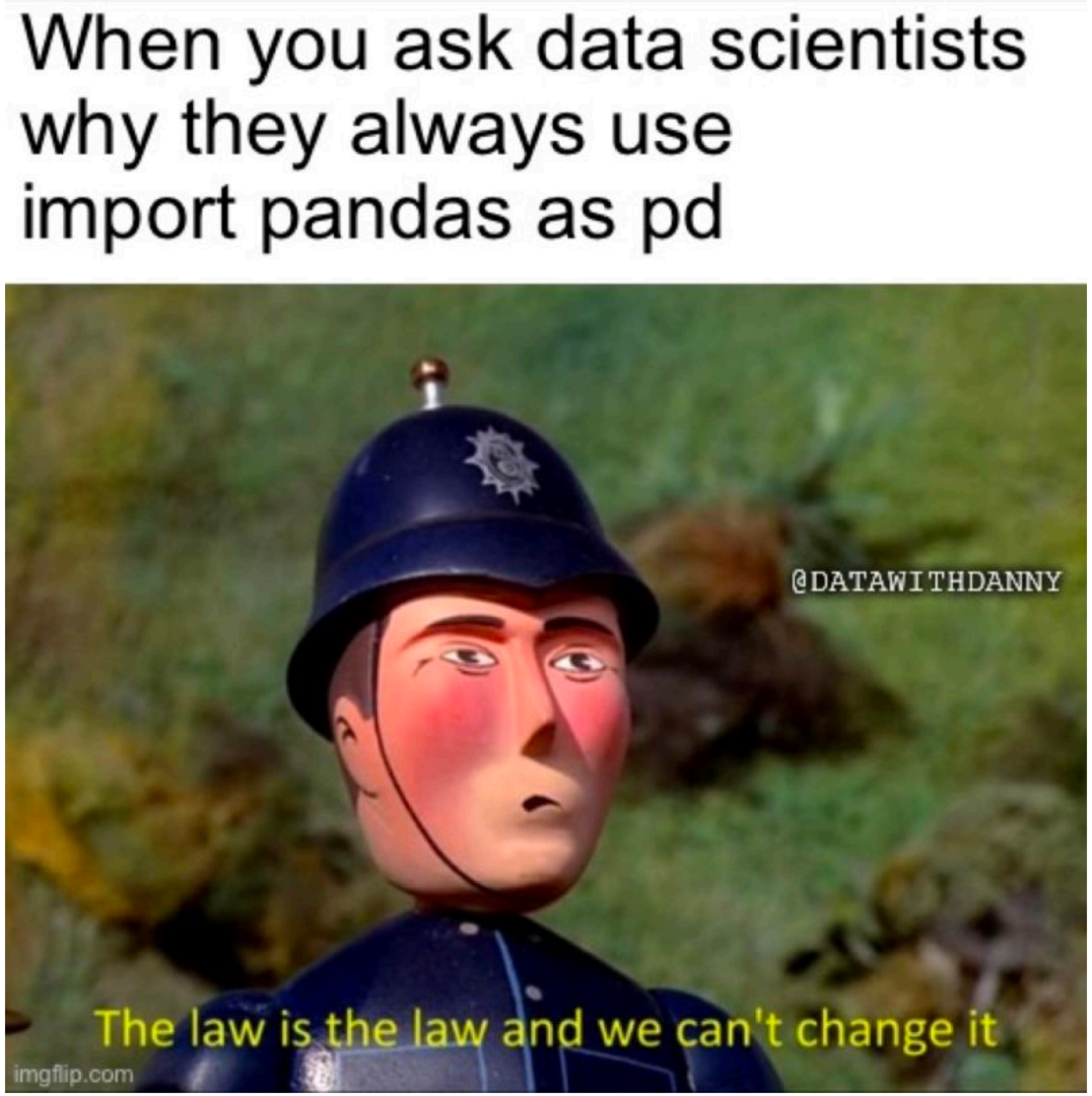
Built for Data Analysis

- Data Cleaning: Handling missing data, outlier detection.
- Data Transformation: Pivot tables, aggregation, and more.
- Data Visualization: Integrated with Matplotlib for basic plotting.

Integration with Other Libraries

Plays Well with Others

- NumPy for numerical operations.
- Matplotlib and Seaborn for visualization.
- SciKit-Learn for machine learning tasks.



When you ask data scientists why they always use import pandas as pd

Performance

Built for Speed

- Built on top of NumPy, enabling efficient array computing.
- Optimized operations for large datasets.
- C-friendly data structures for faster computations.

Series

What is a Series?

- A Series is a one-dimensional labeled array capable of holding data of any type.
- It's essentially a column in an Excel sheet.
- The main difference between a list in Python and a Series in pandas is the presence of the index.
- You can specify custom indices for a Series. This custom index allows for more descriptive data access.

DataFrame

The DataFrame is one of the most powerful data structures in pandas, essentially a two-dimensional table with labeled axes (rows and columns). This is perfect for tabular data.

1. Creating a DataFrame
2. Accessing Data
3. Modifying Data
4. Handling Missing Data
5. Aggregating and Grouping
6. Joining and Merging

Connecting the Dots: NumPy and Pandas

How Pandas Builds Upon NumPy

- The foundation of Pandas lies in NumPy. Every DataFrame or Series is essentially a collection of NumPy arrays
- Additional features like flexible indexing, column alignment, and data manipulation which Pandas offers on top of NumPy.
  - Indexing: Can use labels for indexing.
  - Columns: Have named columns.
  - Data Alignment: Automatically aligns data by index and columns.