DATA PREPARATION WITH PANDAS

CONTENTS

- Introduction
- Why Should We Prepare Our Data
- Python
- Python Libraries
- Pandas
- Features of Pandas
- Core Components Of Pandas
- Pandas Operations
- Typical Pipeline For Data Preparation
- Common Tasks Involved In Data Preparation
- Applications Of Pandas
- Companies using Pandas
- Summary

INTRODUCTION

Data preparation is the first step after you get your hands on any kind of dataset. This is the step when you pre-process raw data into a form that can be easily and accurately analyzed. Proper data preparation allows for efficient analysis - it can eliminate errors and inaccuracies that could have occurred during the data gathering process and can thus help in removing some bias resulting from poor data quality. Therefore a lot of an analyst's time is spent on this vital step.

WHY SHOULD WE PREPARE OUR DATA

- ✓ Garbage in, garbage out
- ✓ Reduce errors
- ✓ Remove duplicate records
- ✓ Fix missing values
- ✓ Correct range values
- ✓ Fix formatting (i.e. date, text, number)

PYTHON

- Object-oriented, high-level programming language
- Used as a scripting language to connect existing components together
- Simple, easy to learn syntax emphasizes readability
- Supports modules and packages

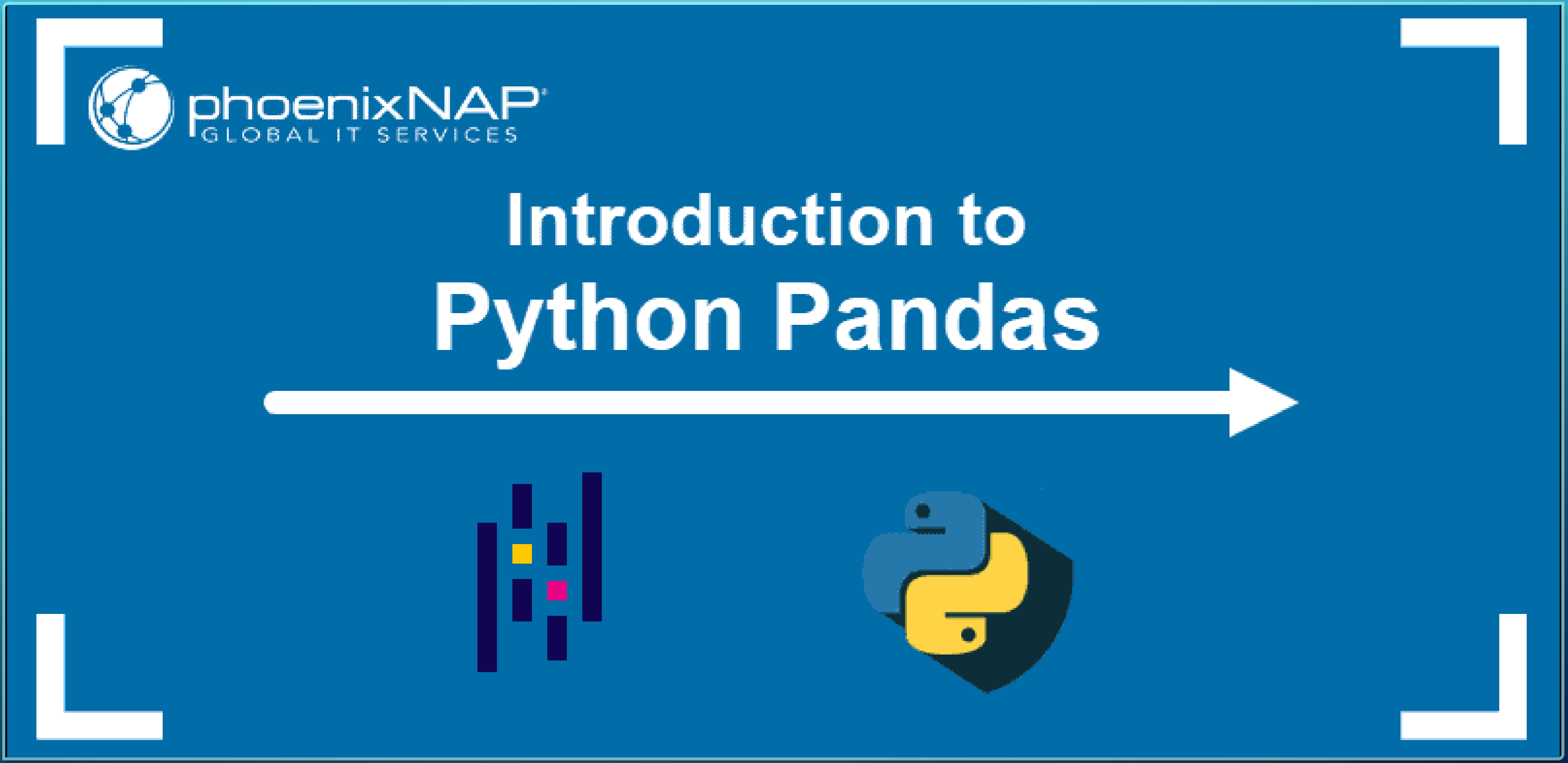
PYTHON LIBRARIES

Many popular Python toolboxes/libraries:-

- NumPy
- SciPy
- Pandas
- SciKit-Learn

Visualization libraries:-

- matplotlib
- Seaborn



Introduction to Python Pandas

PANDAS

- Pandas is a software library written for Python
- Pandas has so many uses that it might make sense to list the things it can't do instead of what it can do
- This tool is essentially your data’s home. Through pandas, you get acquainted with your data by cleaning, transforming, and analyzing it
- Pandas is well suited for different kinds of data, such as:

✓ Tabular data with heterogeneously-typed columns

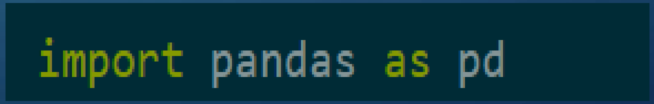
✓ Ordered and unordered time series data

✓ Arbitrary matrix data with row & column labels

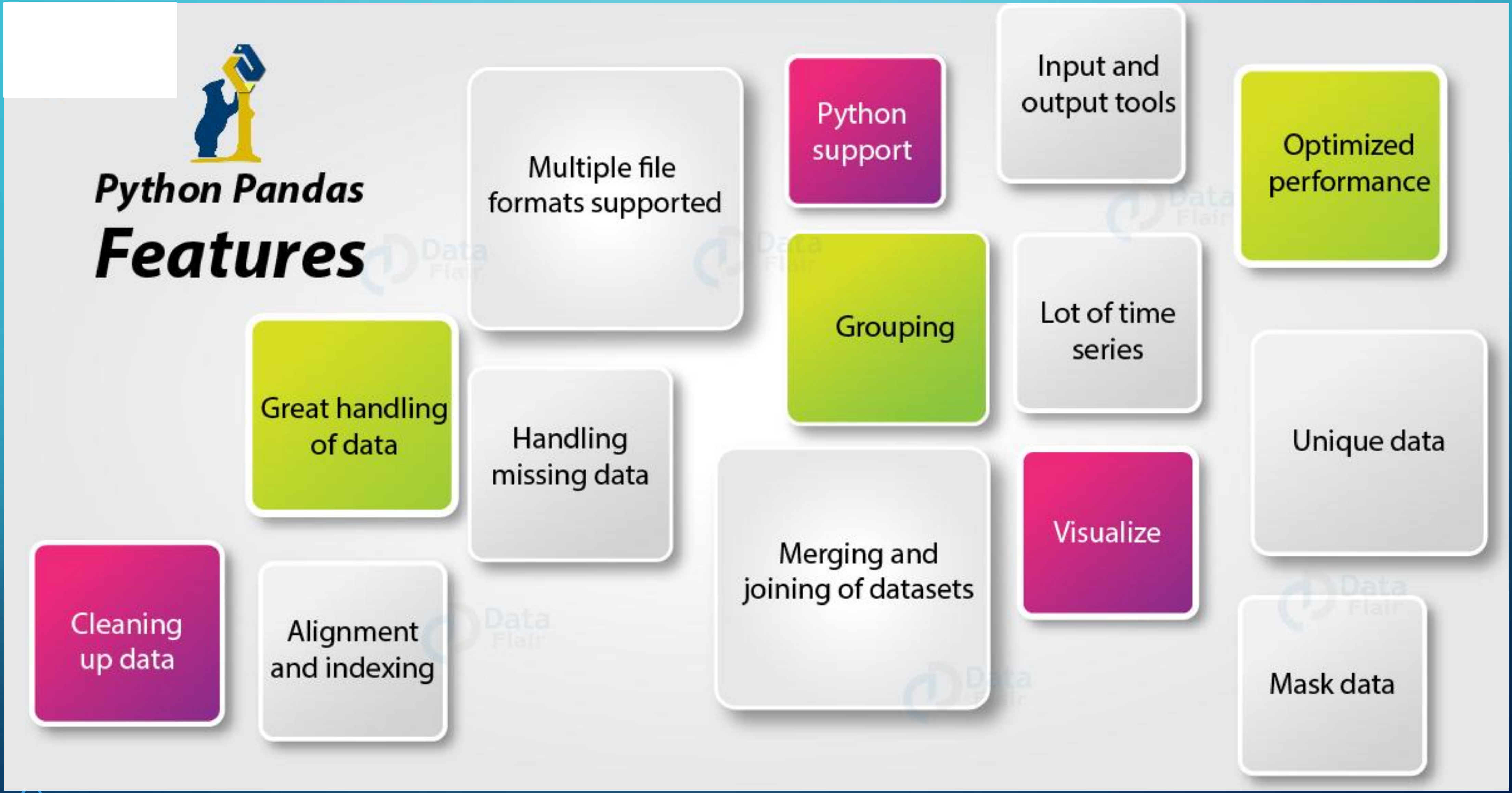
✓ Unlabelled data

✓ Any other form of observational or statistical data sets

To use the pandas library, you need to first import it. Just type this in your python console:



```
import pandas as pd
```



Python Pandas Features

- Multiple file formats supported
- Python support
- Input and output tools
- Optimized performance
- Grouping
- Lot of time series
- Great handling of data
- Handling missing data
- Unique data
- Merging and joining of datasets
- Visualize
- Cleaning up data
- Alignment and indexing
- Mask data

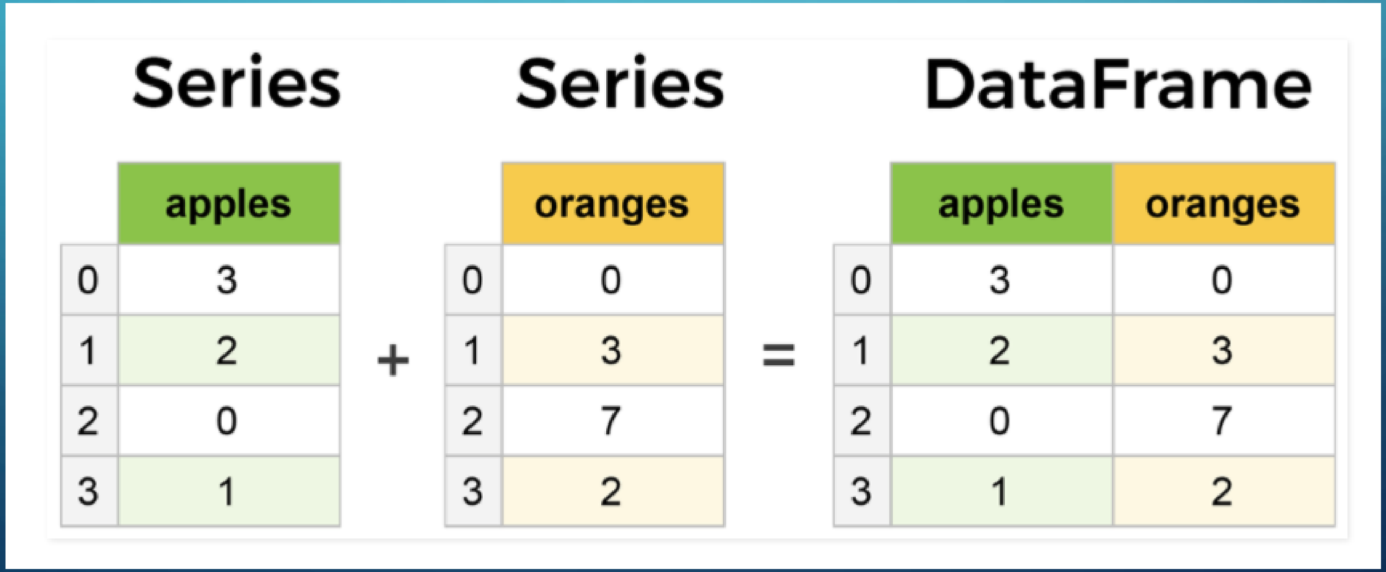
CORE COMPONENTS OF PANDAS

The primary two components of Pandas are:-

➢Dataframe

➢ Series

A Series is essentially column and a Dataframe is a multidimensional Table made up of a collection of Series.



Series

	apples
0	3
1	2
2	0
3	1

+

Series

	oranges
0	0
1	3
2	7
3	2

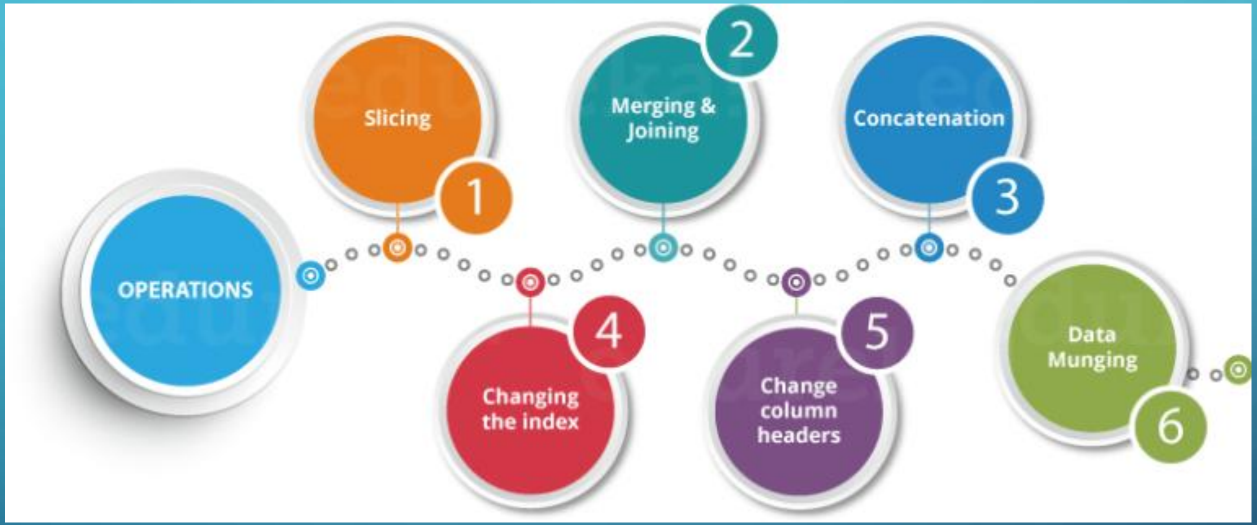
=

DataFrame

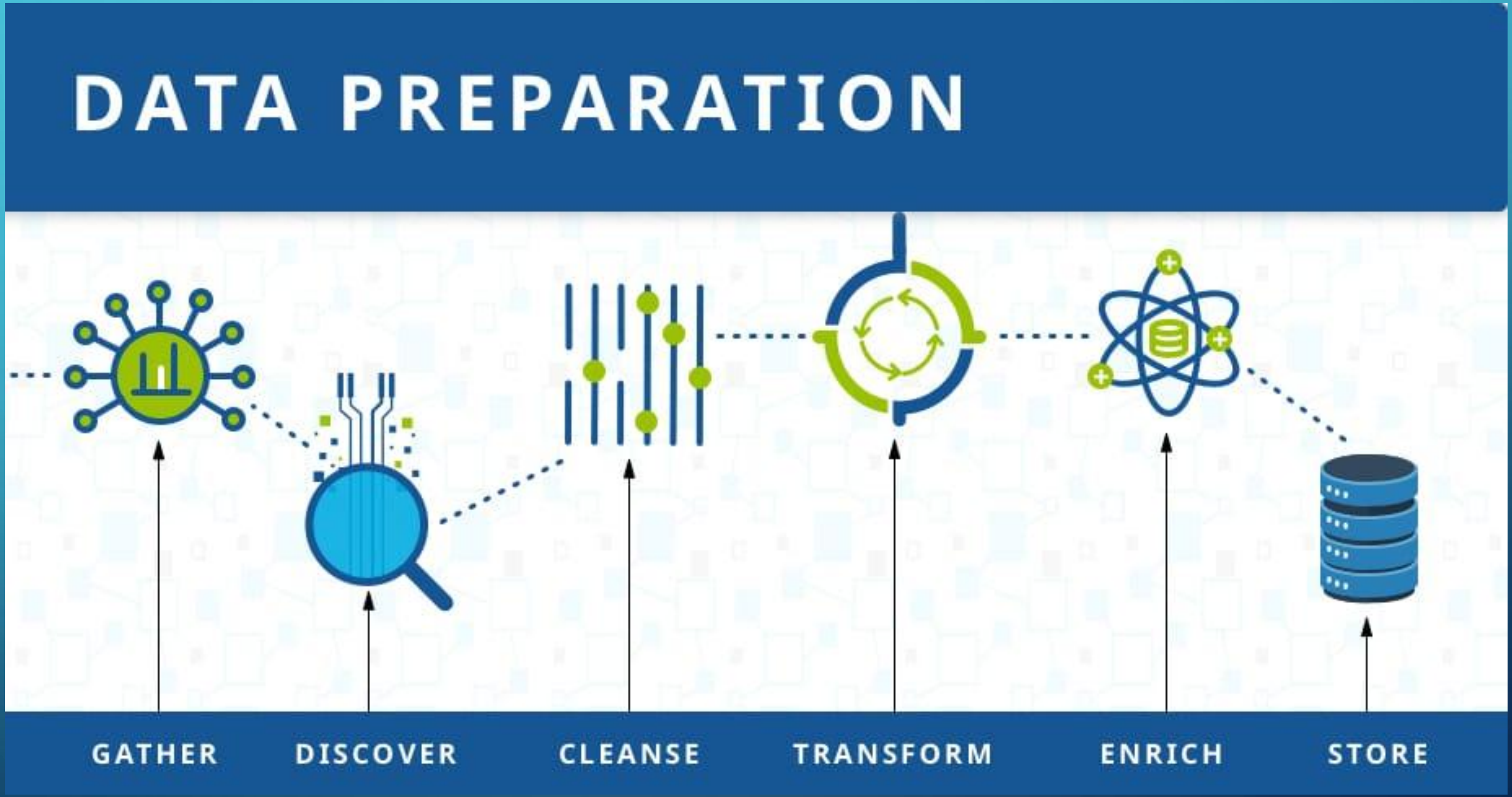
	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

PANDAS OPERATIONS

Using Python pandas, you can perform a lot of operations with series, data frames, missing data, group by etc. Some of the common operations for data manipulation are listed below:



TYPICAL PIPELINE FOR DATA PREPARATION



- The first step of a data preparation pipeline is to **gather** data from various sources and locations
- Before any processing is done, we wish to **discover** what the data is about. At this stage, we understand the data within the context of business goals and Visualization of the data is also helpful here
- The next stage is to **cleanse** the data of missing values and invalid values. We also reformat data to standard forms
- Next we **transform** the data for a specific outcome or audience
- We can **enrich** data by merging different datasets to enable richer insights
- Finally, we **store** the data or directly send it out for analytics

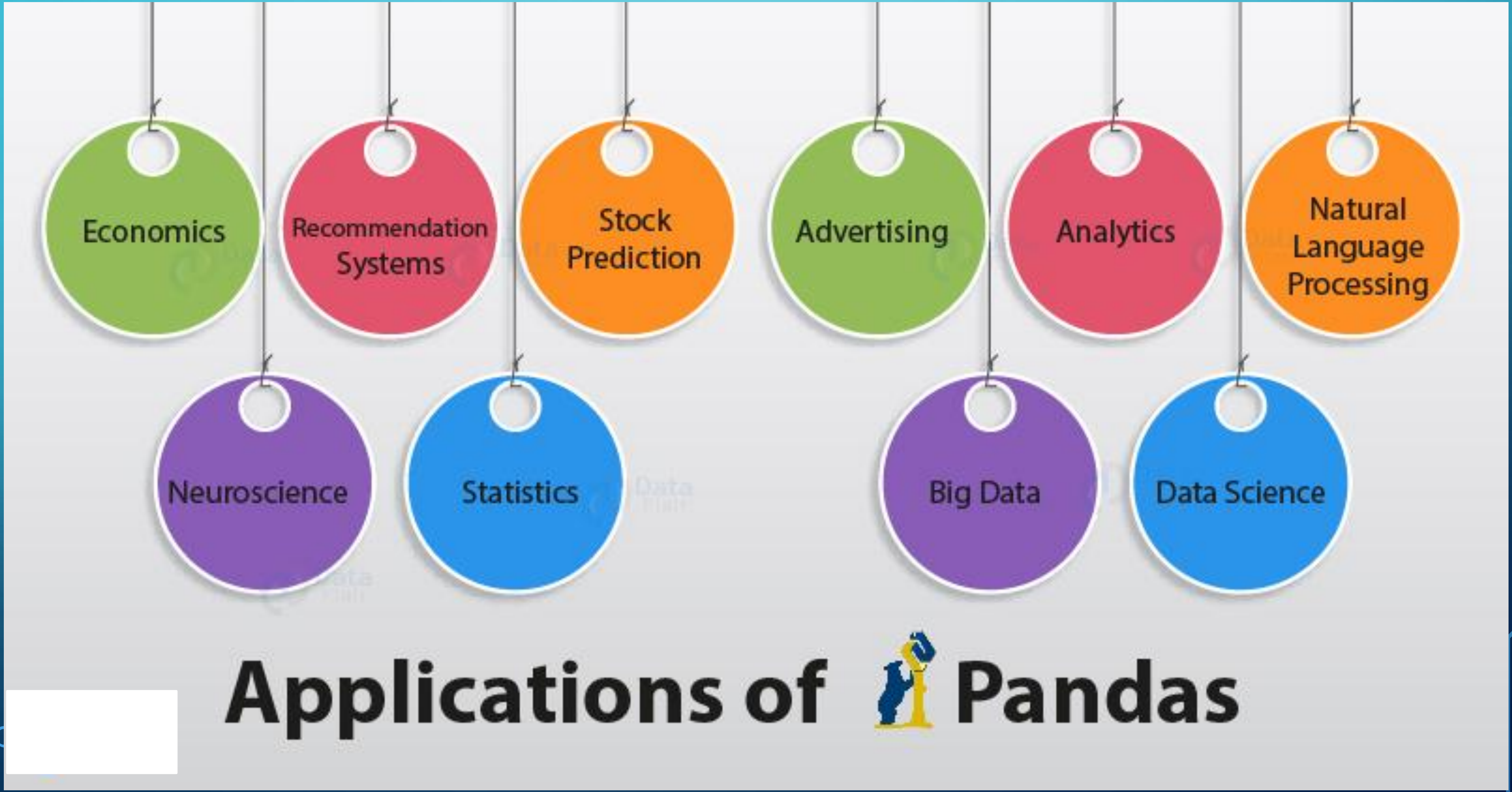
COMMON TASKS INVOLVED IN DATA PREPARATION

Aggregation
Anonymization
Augmentation
Tasks involved in Data Preparation
Blending
Decomposing
Deletion

Data preparation involves one or more of the following tasks:

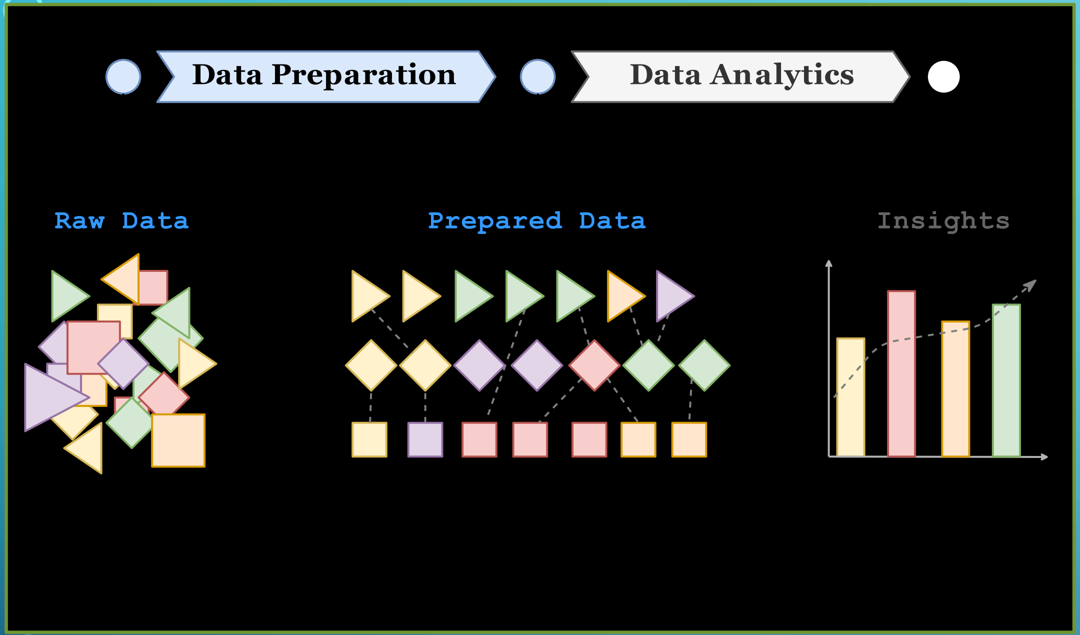
- **Aggregation**: Multiple columns are reduced to fewer columns. Records are summarized
- **Anonymization**: Sensitive values are removed for the sake of privacy
- **Augmentation**: Expand the dataset size without collecting more data. For example, image data is augmented via cropping or rotating
- **Blending**: Combine and link related data from various sources. For example, combine an employee's HR data with payroll data
- **Decomposing**: Decompose a data column that has sub-fields. For example, "6 ounces butter" is decomposed into three columns representing value, unit and ingredient
- **Deletion**: Duplicates and outliers are removed. Exploratory Data Analysis (EDA) may be used to identify outliers

APPLICATIONS OF PANDAS



COMPANIES USING PANDAS

SUMMARY



Raw data is usually not suitable for direct analysis. This is because the data might come from different sources in different formats. Moreover, real-world data is not clean. Some data points might be missing. Some others might be out of range. There could be duplicates. Data preparation is therefore an essential task that transforms or prepares data into a form that's suitable for analysis.

Data preparation assumes that data has already been collected. However, others may consider data collection and data ingestion as part of data preparation. Within data preparation, it's common to identify sub-stages that might include data pre-processing, data wrangling, and data transformation.

THANK YOU

Q
&
A