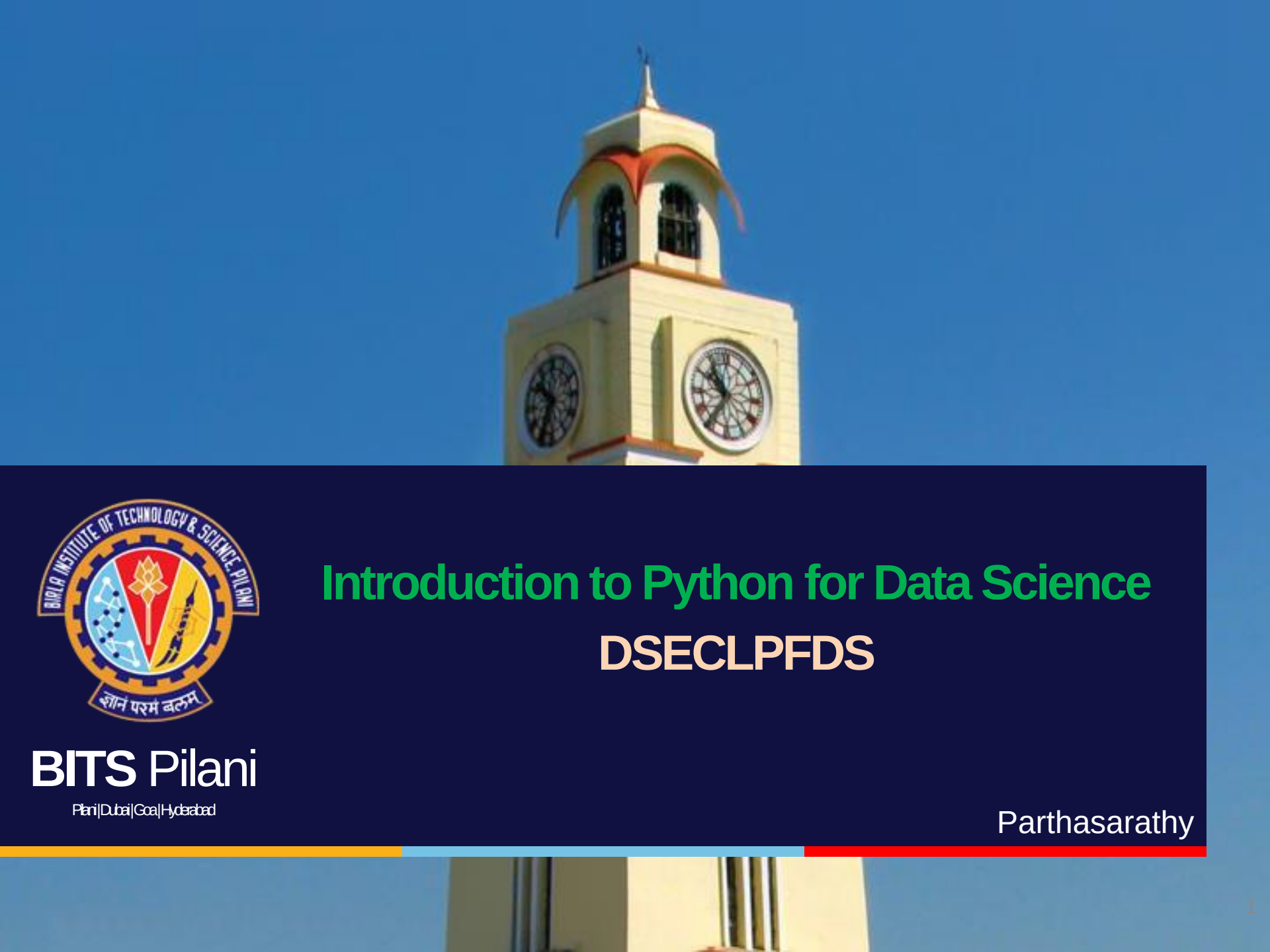


Introduction to Python for Data Science

DSECLPFDS

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Parthasarathy

Agenda for CS #4,5,6

1) Introduction to SciPy Ecosystem
2) Basics of NumPy
3) Basics of Pandas
4) Data Exploration for a Bike Dataset
5) Visualizations using Matplotlib
6) Visualizations using Seaborn
7) Introduction to Scikit-learn

Introduction to SciPy Ecosystem

SciPy (pronounced "Sigh Pie") is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the core packages:

NumPy

Base N-dimensional array package

SciPy library

Fundamental library for scientific computing

Matplotlib

Comprehensive 2-D plotting

IP[y]: IPython

IPython

Enhanced interactive console

SymPy

Symbolic mathematics

pandas

Data structures & analysis

https://www.scipy.org/

https://www.scipy.org/about.html

NumPy

- NumPy stands for Numerical Python.
- NumPy is a Python library used for working with arrays.
- It also has functions for working in domain of linear algebra, fourier transform, and matrices.

Why use NumPy ?

- In Python we have lists that serve the purpose of arrays, but they are slow to process.
- NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.
- The array object in NumPy is called *ndarray*, it provides a lot of supporting functions.

Demo on NumPy

Lets have a hands-on session on :

- Creation of NumPy array
- Properties of NumPy array
- Prepopulated arrays
- 2D arrays
- Ways to access array elements
- Reshaping
- Computations
  - Arithmetic
  - Comparison
  - Aggregation
  - Boolean
- Note on Linear Algebra and NumPy arrays

Pandas

- Pandas is a Python library used for working with data sets.
- It has functions for analyzing, cleaning, exploring, and manipulating data.
- The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis“
- 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.
- Pandas gives us answers about the data. Like:
  - Is there a correlation between two or more columns?
  - What is average value?
  - Max value?
  - Min value?
- Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called *cleaning* the data.

Pandas

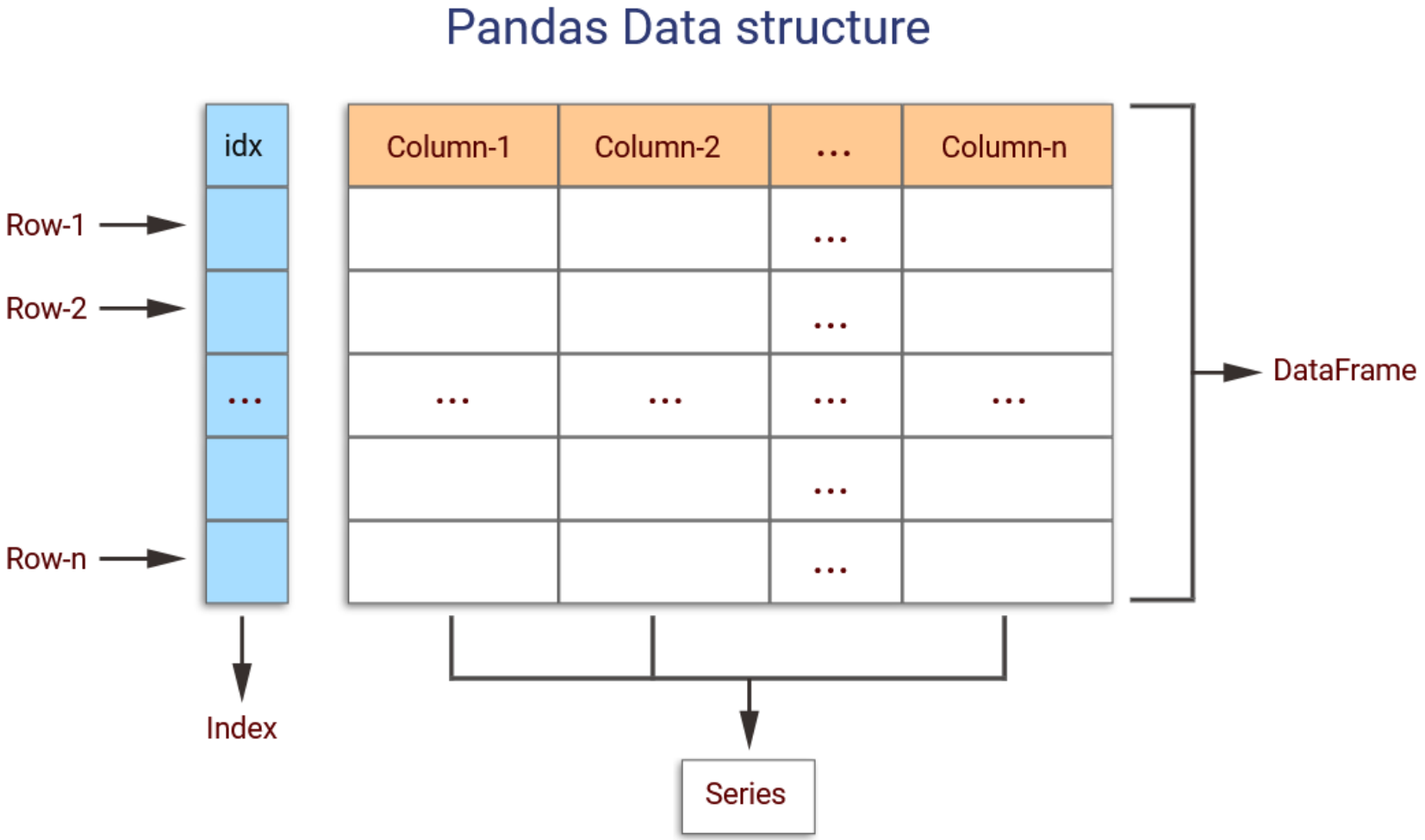
- DataFrame:

A Dataframe is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. In dataframe datasets arrange in rows and columns, we can store any number of datasets in a dataframe. We can perform many operations on these datasets like arithmetic operation, columns/rows selection, columns/rows addition etc.

In the real world, a Pandas DataFrame will be created by loading the datasets from existing storage, storage can be SQL Database, CSV file, and Excel file. Pandas DataFrame can be created from the lists, dictionary, and from a list of dictionary etc.

In simple, a dataframe is a collection of Series and a Series is a collection of scalar values.

Pandas DataFrame



Demo on Basics of Pandas

- Importing Pandas
- Series Creation
- Data Frame Creation
- Reading Files using Pandas
- Writing to a File using Pandas
- Some functions offered by Pandas

Post your queries in the Discussion Forum!!

Feedback

 : 5

 : 3

 : 1

Thank You for your time & attention !

Contact : parthasarathypd@wilp.bits-pilani.ac.in