# Object Oriented Programming (Using Python)

## UNIT- III

## Python Libraries:

- Basics of Open Source Libraries for
  - Data pre-processing
  - Modeling and
  - Visualization

Prof. R. MADANA MOHANA
Professor, Artificial Intelligence & Data Science
http://rmadanamohana.com/

# Python Libraries

- **Python** is one of the most popular languages used by data scientists and software developers alike for data science tasks.
- It can be used to predict outcomes, automate tasks, streamline processes, and offer business intelligence insights.
- There are quite a few open-source libraries that make Python data tasks much easier.

# Python Libraries

The following are the most important libraries for data science tasks available in the Python ecosystem covering areas such as data processing, modeling, and visualization:

**Data Mining**

1. Scrapy
2. BeautifulSoup

# Python Libraries

## Data Processing and Modeling

3. NumPy (Numerical Python)
4. SciPy (Scientific Python)
5. Pandas (Python Data Analysis)
6. Keras
7. SciKit-Learn
8. PyTorch
9. TensorFlow
10. XGBoost

# Python Libraries

## Data Visualization

11. Matplotlib
12. Seaborn
13. Bokeh
14. Plotly
15. Pydot

# Python Libraries

## Scrapy

- One of the most popular Python data science libraries, Scrapy helps to build crawling programs (spider bots) that can retrieve structured data from the web – for example, URLs or contact info. It's a great tool for scraping data used in, for example, Python machine learning models.
- Developers use it for gathering data from APIs. This full-fledged framework follows the Don't Repeat Yourself principle in the design of its interface. As a result, the tool inspires users to write universal code that can be reused for building and scaling large crawlers.

# Python Libraries

## BeautifulSoup

- BeautifulSoup is another really popular library for web crawling and data scraping.
- If you want to collect data that's available on some website but not via a proper CSV or API, BeautifulSoup can help you scrape it and arrange it into the format you need.

# Python Libraries

## NumPy

- NumPy (Numerical Python) is a perfect tool for scientific computing and performing basic and advanced array operations.
- This library offers many handy features performing operations on n-arrays and matrices in Python.
- It helps to process arrays that store values of the same data type and makes performing math operations on arrays (and their vectorization) easier.
- In fact, the vectorization of mathematical operations on the NumPy array type increases performance and accelerates the execution time.

# Python Libraries

## SciPy

- This useful library includes modules for linear algebra, integration, optimization, and statistics.
- Its main functionality was built upon NumPy, so its arrays make use of this library.
- SciPy works great for all kinds of scientific programming projects (science, mathematics, and engineering).
- It offers efficient numerical routines such as numerical optimization, integration, and others in submodules.
- The extensive documentation makes working with this library really easy.

# Python Libraries

## Pandas

- Pandas is a library created to help developers work with "labeled" and "relational" data intuitively.
- It's based on two main data structures: "Series" (one-dimensional, like a list of items) and "Data Frames" (two-dimensional, like a table with multiple columns).
- Pandas allows converting data structures to DataFrame objects, handling missing data, and adding/deleting columns from DataFrame, assigning missing files, and plotting data with histogram or plot box. It’s a must-have for data wrangling, manipulation, and visualization.

# Python Libraries

## Keras

- Keras is a great library for building neural networks and modeling.
- It's very straightforward to use and provides developers with a good degree of extensibility.
- The library takes advantage of other packages, (Theano or TensorFlow) as its backends.
- Moreover, Microsoft integrated CNTK (Microsoft Cognitive Toolkit) to serve as another backend.
- It's a great pick if you want to experiment quickly using compact systems – the minimalist approach to design really pays off!

# Python Libraries

## SciKit-Learn

- This is an industry-standard for data science projects based in Python.
- Scikits is a group of packages in the SciPy Stack that were created for specific functionalities – for example, image processing.
- Scikit-learn uses the math operations of SciPy to expose a concise interface to the most common machine learning algorithms.
- Data scientists use it for handling standard machine learning and data mining tasks such as clustering, regression, model selection, dimensionality reduction, and classification.
- Another advantage: It comes with quality documentation and offers high performance.

# Python Libraries

## PyTorch

- PyTorch is a framework that is perfect for data scientists who want to perform deep learning tasks easily.
- The tool allows performing tensor computations with GPU (Graphics Processing Unit ) acceleration.
- It's also used for other tasks – for example, for creating dynamic computational graphs and calculating gradients automatically.
- PyTorch is based on Torch, which is an open-source deep learning library implemented in C, with a wrapper in Lua (Lua is a lightweight, high-level, multi-paradigm programming language designed primarily for embedded use in applications).

# Python Libraries

## TensorFlow

- TensorFlow is a popular Python framework for machine learning and deep learning, which was developed at Google Brain.
- It's the best tool for tasks like object identification, speech recognition, and many others.
- It helps in working with artificial neural networks that need to handle multiple data sets.
- The library includes various layer-helpers (tflearn, tf-slim, skflow), which make it even more functional.
- TensorFlow is constantly expanded with its new releases – including fixes in potential security vulnerabilities or improvements in the integration of TensorFlow and GPU.

# Python Libraries

## XGBoost

- Use this library to implement machine learning algorithms under the Gradient Boosting framework.
- XGBoost is portable, flexible, and efficient.
- It offers parallel tree boosting that helps teams to resolve many data science problems.
- Another advantage is that developers can run the same code on major distributed environments such as Hadoop, SGE (Sun Grid Engine), and MPI (Message Passing Interface).

# Python Libraries

## Matplotlib

- This is a standard data science library that helps to generate data visualizations such as two-dimensional diagrams and graphs (histograms, scatterplots, non-Cartesian coordinates graphs).
- Matplotlib is one of those plotting libraries that are really useful in data science projects - it provides an object-oriented API for embedding plots into applications.
- Python can compete with scientific tools like MatLab or Mathematica.
- However, developers need to write more code than usual while using this library for generating advanced visualizations.
- Note that popular plotting libraries work seamlessly with Matplotlib.

# Python Libraries

## Seaborn

- Seaborn is based on Matplotlib and serves as a useful Python machine learning tool for visualizing statistical models – heatmaps and other types of visualizations that summarize data and depict the overall distributions.
- When using this library, you get to benefit from an extensive gallery of visualizations (including complex ones like time series, joint plots, and violin diagrams).

# Python Libraries

## Bokeh

- This library is a great tool for creating interactive and scalable visualizations inside browsers using JavaScript widgets.
- Bokeh is fully independent of Matplotlib.
- It focuses on interactivity and presents visualizations through modern browsers – similarly to Data-Driven Documents (d3.js).
- D3.js is a JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of Scalable Vector Graphics, HTML5, and Cascading Style Sheets standards.
- It offers a set of graphs, interaction abilities (like linking plots or adding JavaScript widgets), and styling.

# Python Libraries

## Plotly

- This web-based tool for data visualization that offers many useful out-of-box graphics – you can find them on the Plot.ly website (https://plotly.com/).
- The library works very well in interactive web applications.
- Its creators are busy expanding the library with new graphics and features for supporting multiple linked views, animation, and crosstalk integration.

# Python Libraries

## pydot

- This library helps to generate oriented and non-oriented graphs.
- It serves as an interface to Graphviz (written in pure Python).
- You can easily show the structure of graphs with the help of this library.
- That comes in handy when you're developing algorithms based on neural networks and decision trees.