CAU 중앙대학교 CHUNG-ANG UNIVERSITY

PYTHON FOR DATA SCIENCE AND MACHINE LEARNING

Name: **Chanda Simfukwe Ph.D. Candidate**
Supervisor: **Dr. Young Chul Youn**
Date: **10.05.2023**

DATE	TASK TITLE	START TIME	END TIME	TASK COMPLETE
5.04.2023	Course Introduction	6pm	8pm	

1. Introduction to the Course

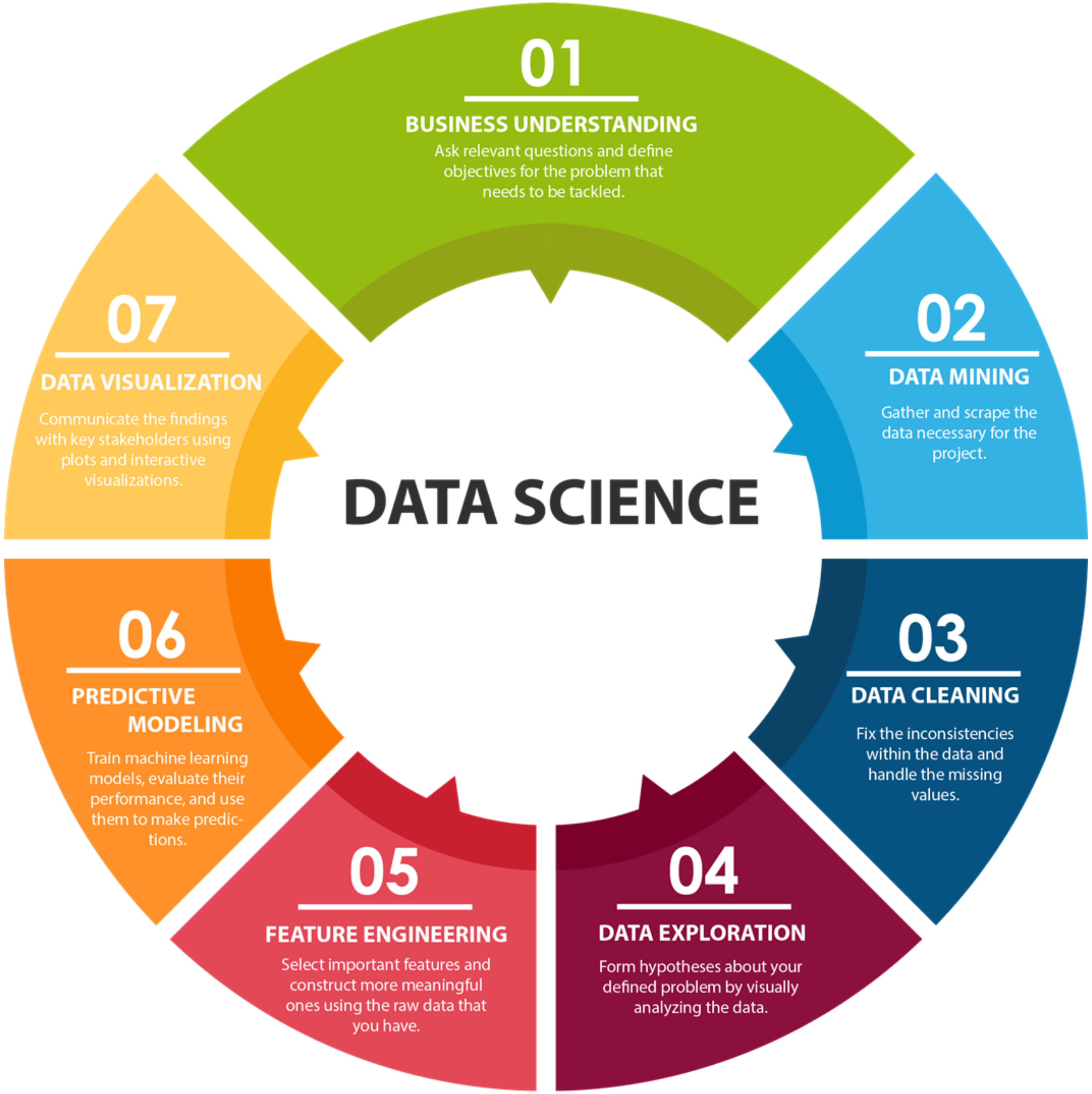
2. Environment Set-up

3. Jupyter Overview

4. Python Basics

5. Python Basics Exercise Overview

1. Introduction to the Course



DATA SCIENCE

01 BUSINESS UNDERSTANDING

Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING

Gather and scrape the data necessary for the project.

03 DATA CLEANING

Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING

Select important features and construct more meaningful ones using the raw data that you have.

06 PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION

Communicate the findings with key stakeholders using plots and interactive visualizations.

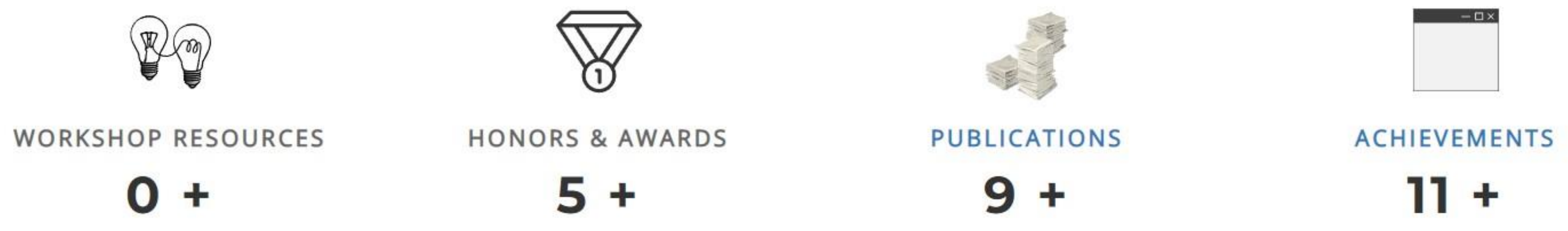
Most Popular Python Data Science Libraries

NumPy
SciPy
Pandas
Seaborn

Scikit-Learn
MatplotLib
Plotly
PySpark

and much more!

- Course resources
  - Go to : www.chandasimfukwe.com



WORKSHOP RESOURCES
0 +

HONORS & AWARDS
5 +

PUBLICATIONS
9 +

ACHIEVEMENTS
11 +

2. Environment Set-up

- Set Up and Installation
  - Objectives
    - Install Python with Anaconda
    - Download zip file of notebooks from resources
    - Open Jupyter and explore notebooks

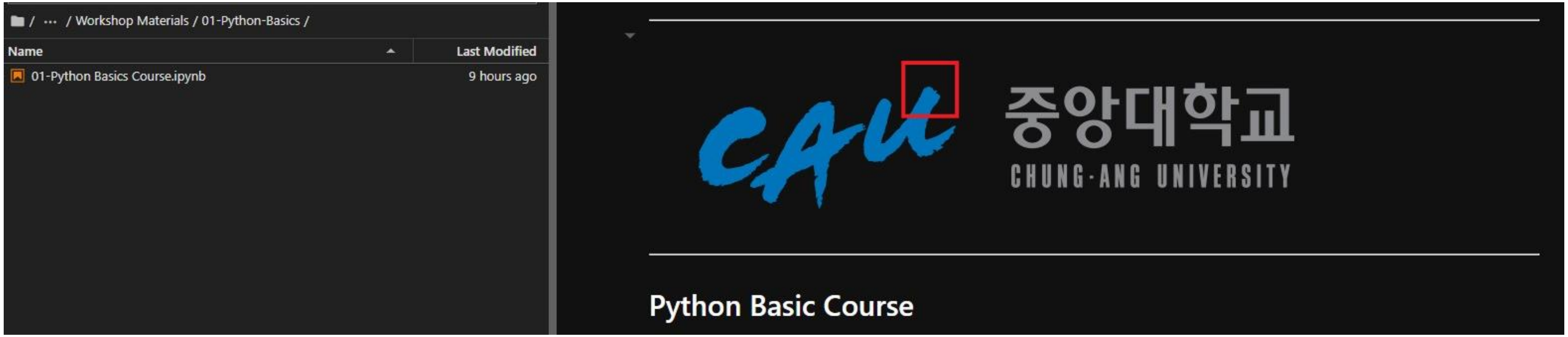
- Anaconda is a distribution of Python
- This mean it includes not only Python, but many libraries that we use in the workshop, as well as its own virtual environment systems.
- Its an “all-in-one” install that is extremely popular in data science and machine learning!

- Jupyter is a development environment where we can write code, display images, and write down markdown notes.
- It is the most popular IDE in data science for exploring and analyzing data!
- It is also a great learning tool.

- Let’s download Anaconda
- Go to: https://www.anaconda.com/
- Or simply Google Search:
- “Anaconda Python Download”

3. Jupyter Overview

- Jupyter Notebooks
  - Check the resources for this lecture and download the zip file.
  - It contains all the .ipynb files and notebooks for the course.
  - Make sure you remember where you saved and unzipped it.



/ … / Workshop Materials / 01-Python-Basics /

Name	Last Modified
01-Python Basics Course.ipynb	9 hours ago

Python Basic Course

Anaconda Virtual Environments

- Virtual Environments allow you to set up virtual installations of Python and libraries on your computer.
- You can have multiple versions of Python or libraries and easily activate or deactivate these environments.
- Let’s see some examples of why you may want to do this.
- Anaconda has a built-in virtual environment manager that makes the whole process easy.
- Check out the resource link for the official documentation that we will go over now.

4. Python Basics

- Topics to cover

  - Data Types
    - Numbers
    - Strings
    - Print Formatting
    - Lists
    - Dictionaries
    - Booleans
    - Tuples and Sets

  - Python Operators
    - Comparison Operators
    - If, elif, and else Statements
    - For Loops
    - While Loops
    - range()
    - List Comprehension
    - Functions

5. Python Basics Exercise Overview

- Exercise resources
  - Download the exercise folder named “Exercise-05.04.2023” from “Workshop Resources” unzip and upload to Jupyterlab to run.

Python Basic Course Exercises

6. NumPy Arrays

Section Goals

- Understanding NumPy
- Create arrays with NumPy
- Retrieve information from a NumPy array through slicing and indexing
- Learn basic NumPy operations
- Test NumPy skills with exercise questions

What is NumPy (Numeric Python)?

- Python library for creating N-dimensional arrays
- Ability to quickly broadcast functions
- Built-in linear algebra, statistical distributions, trigonometric, and random number capabilities

Why use NumPy?

- While NumPy structures look like standard Python lists, they are much more efficient
- The broadcasting capabilities are also extremely useful for quickly applying functions to data sets
- NumPy-based algorithms are generally 10 to 100 times faster (or more) than their pure Python counterparts and use significantly less memory.

```
import numpy as np
my_arr = np.arange(1000000)
my_list = list(range(1000000))
```

Topics to cover

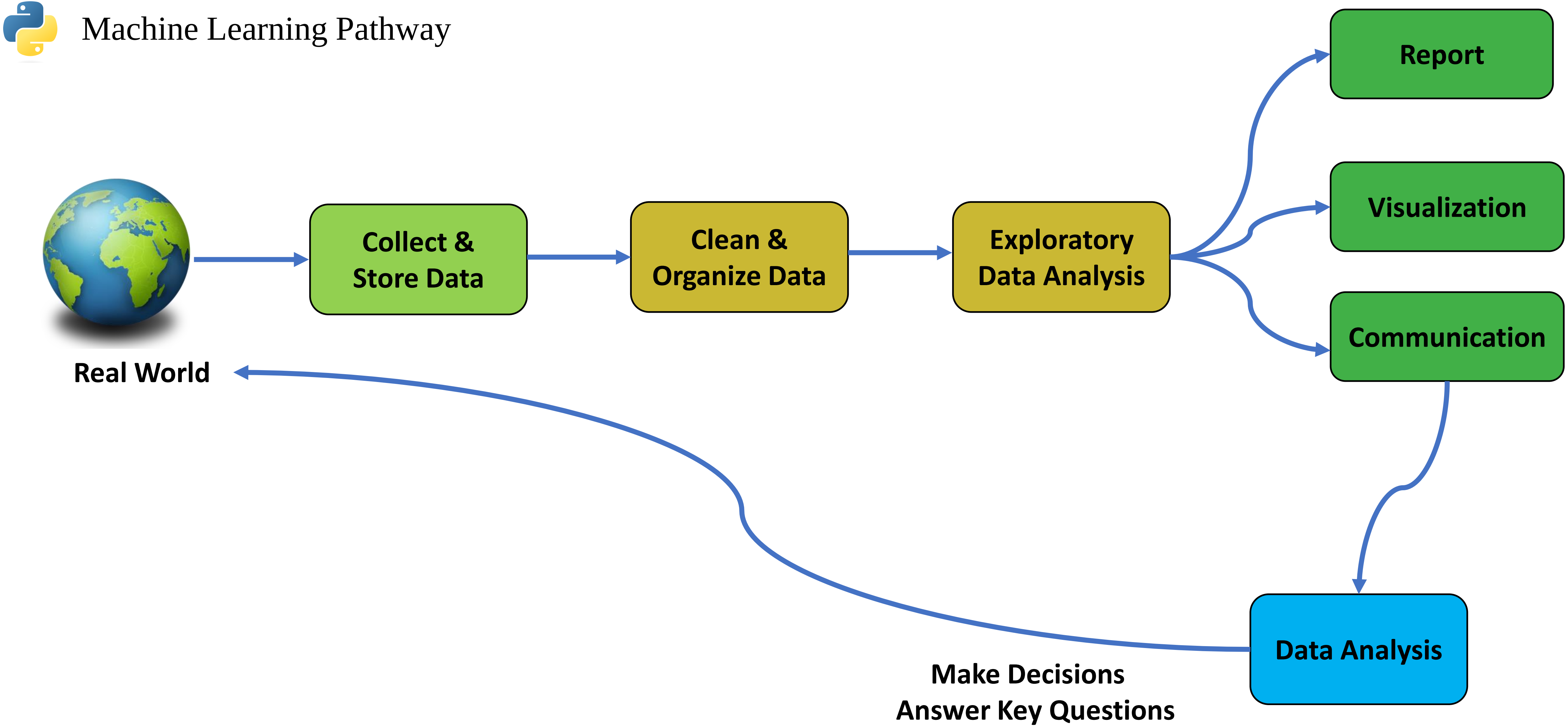
- NumPy Arrays
- Creating NumPy
- NumPy vs. Lists
- Built-in Methods
- Random
- Array Attributes and Methods
- Reshape
- Shape
- dtypes
- Numpy Indexing and Selection
- Broadcasting
- Indexing a 2D Array (Matrices)
- Fancy Indexing
- Selection
- Arithmetic
- Universal Array Functions

7. NumPy Basics Exercise Overview

- Exercise resources
  - Download the exercise folder named “Exercise-19.04.2023” from “Workshop Resources” Unzip and upload to Jupyterlab to run.
  - Note: Solutions provided

NumPy Basics Exercises

Machine Learning Pathway



Real World

Collect & Store Data

Clean & Organize Data

Exploratory Data Analysis

Report

Visualization

Communication

Data Analysis

Make Decisions
Answer Key Questions

8. Pandas

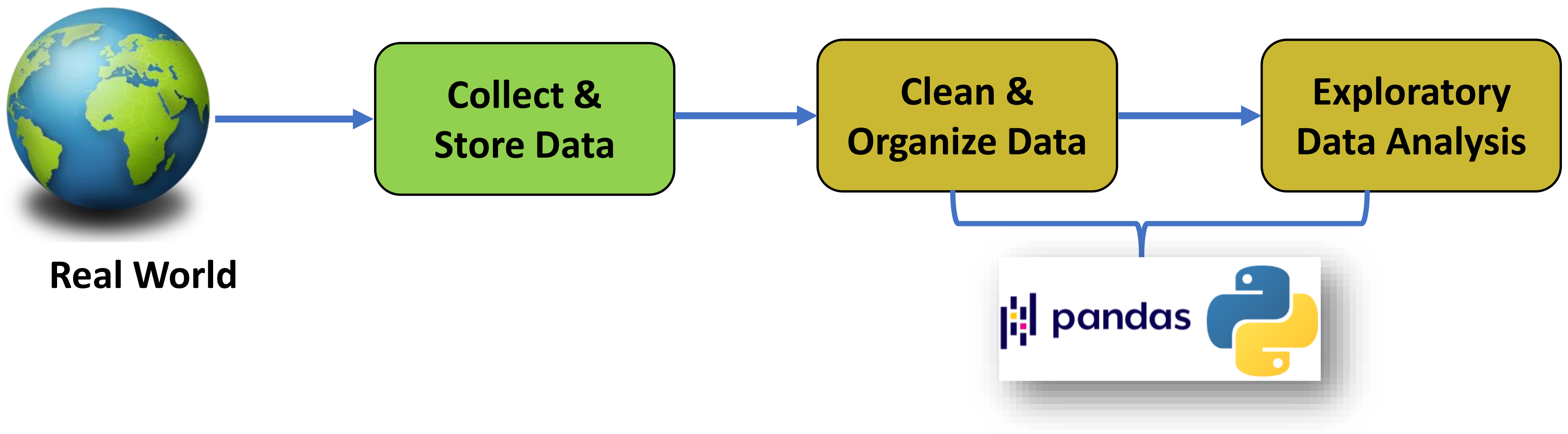
- Pandas is an open-source library built on top of Numpy
- It allows for fast analysis and data cleaning and preparation
- It excels in performance and productivity
- It also has built-in visualization features
- It can work with data from a wide variety of sources and formats
  - https://pandas.pydata.org/docs/index.html

Install Pandas

- You’ll need to install pandas by going to your command line or terminal and using either

    - ○ conda install pandas
    - ○ pip install pandas

Machine Learning Pathway



Real World

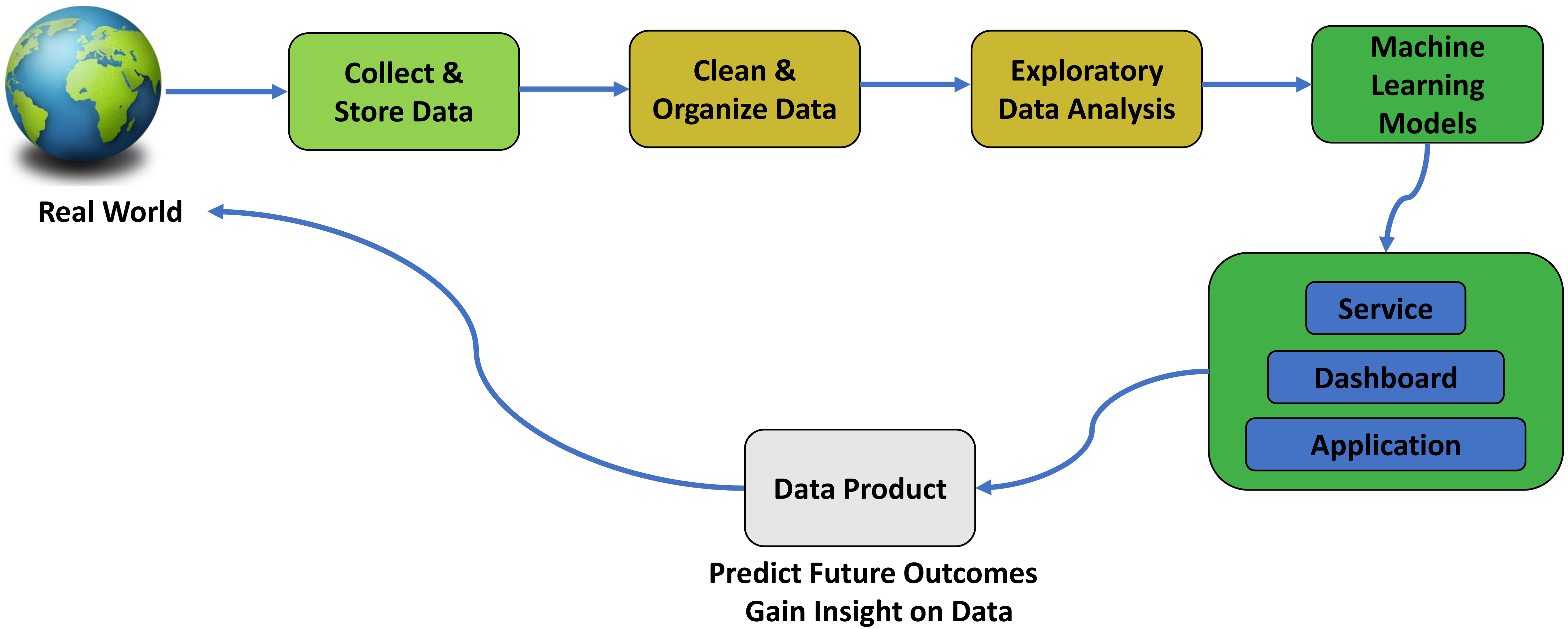
Collect & Store Data

Clean & Organize Data

Exploratory Data Analysis

pandas

Machine Learning Pathway



Real World

Collect & Store Data

Clean & Organize Data

Exploratory Data Analysis

Machine Learning Models

Service

Dashboard

Application

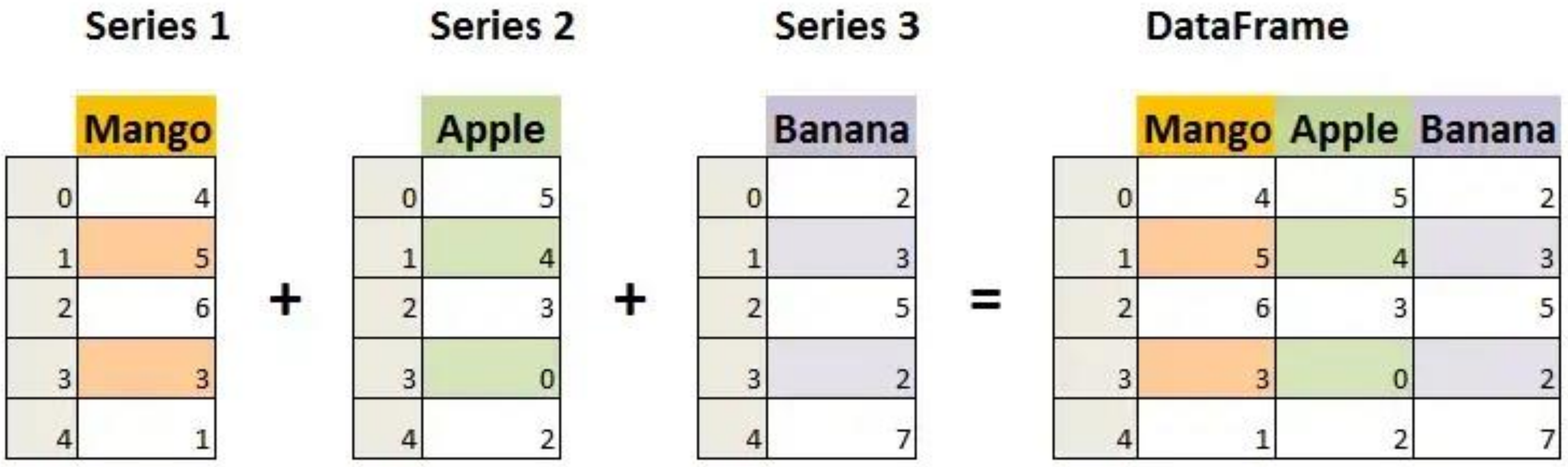
Data Product

Predict Future Outcomes
Gain Insight on Data

Topics to cover

- Series
- DataFrames
- Conditional Filtering
- Missing Data
- Group By Operations
- Merging Joining and Concatenating
- Operations

Series and DataFrame



Series 1

	Mango
0	4
1	5
2	6
3	3
4	1

+

Series 2

	Apple
0	5
1	4
2	3
3	0
4	2

+

Series 3

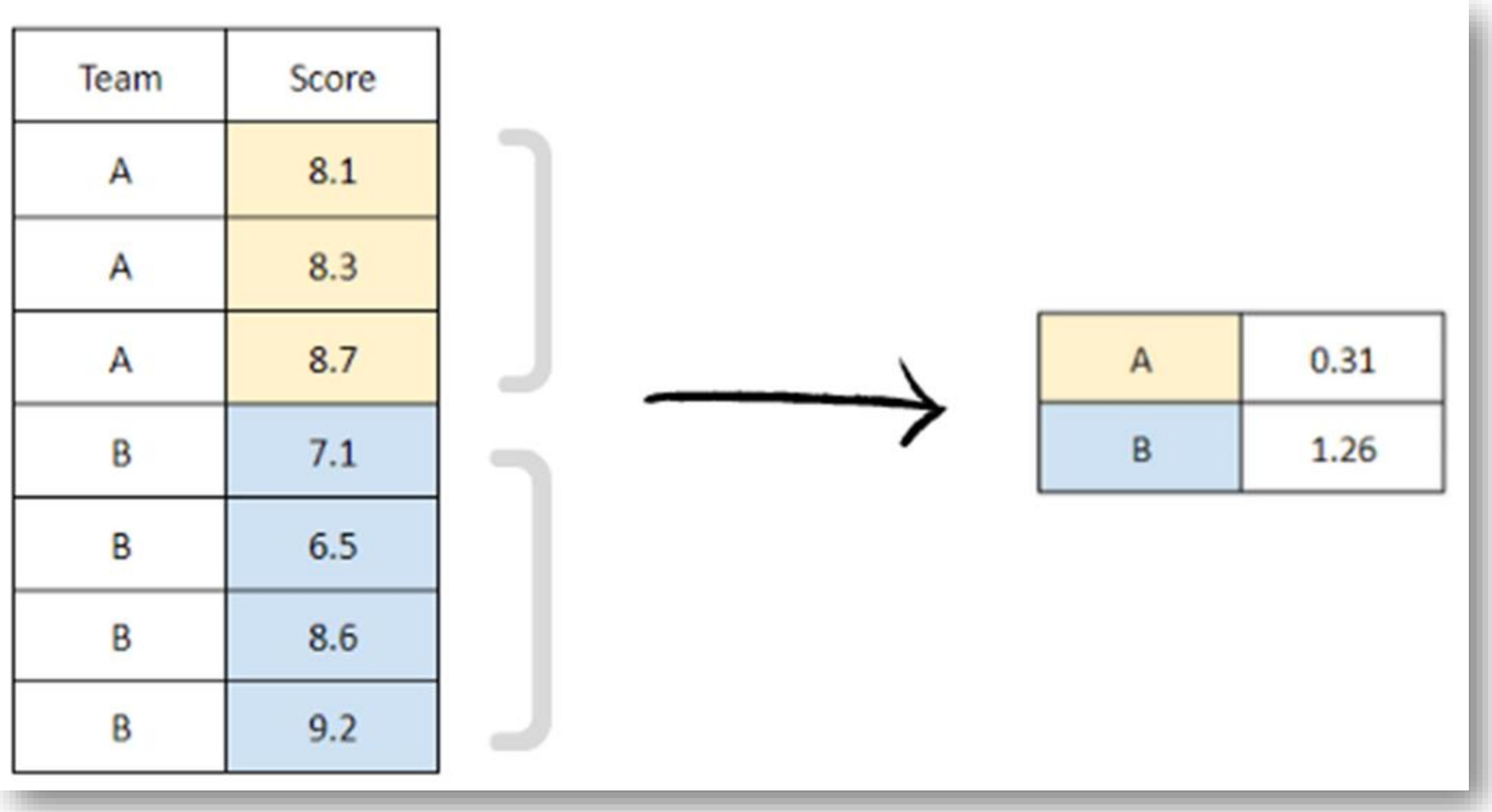
	Banana
0	2
1	3
2	5
3	2
4	7

=

DataFrame

	Mango	Apple	Banana
0	4	5	2
1	5	4	3
2	6	3	5
3	3	0	2
4	1	2	7

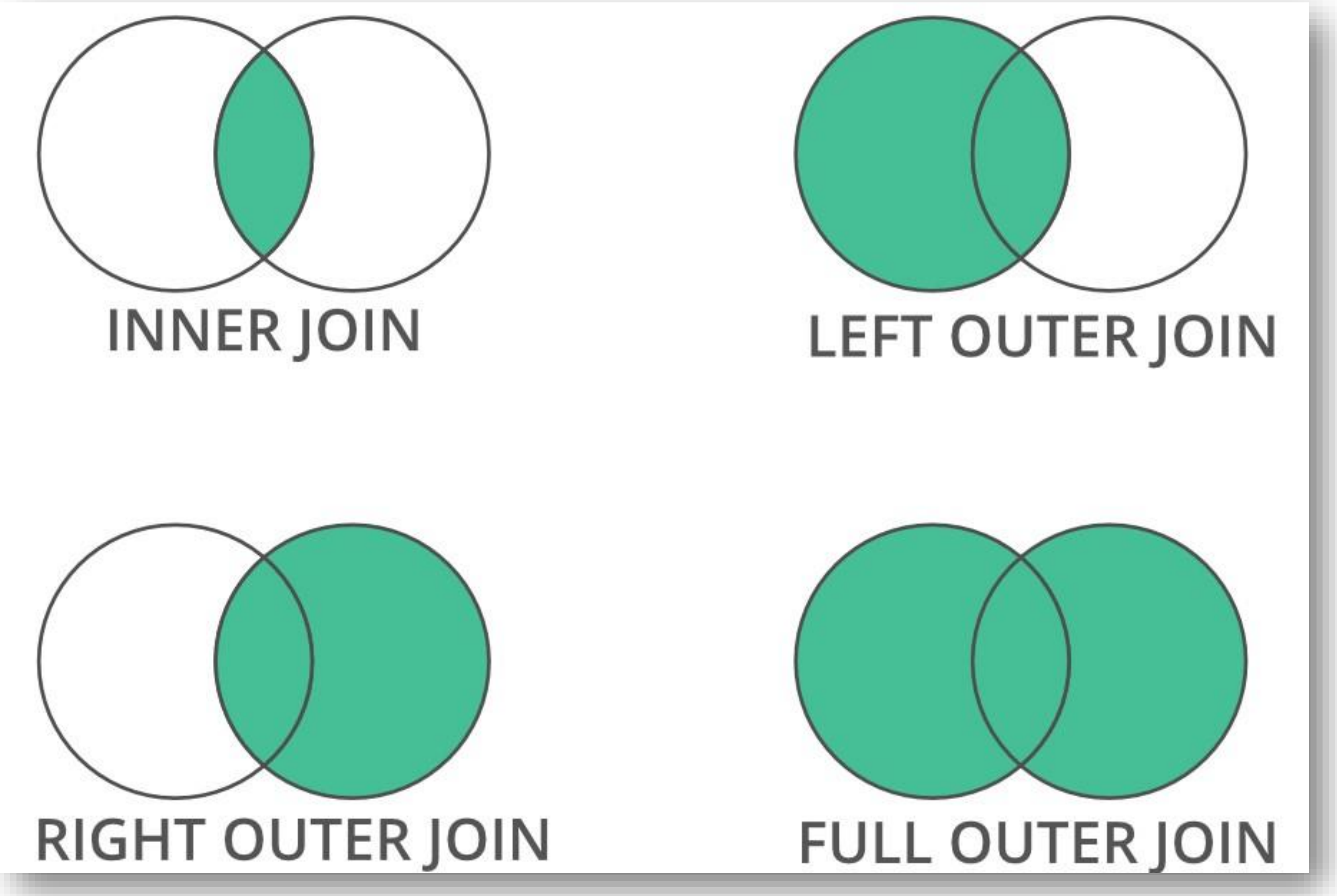
GroupBy



Team	Score
A	8.1
A	8.3
A	8.7
B	7.1
B	6.5
B	8.6
B	9.2

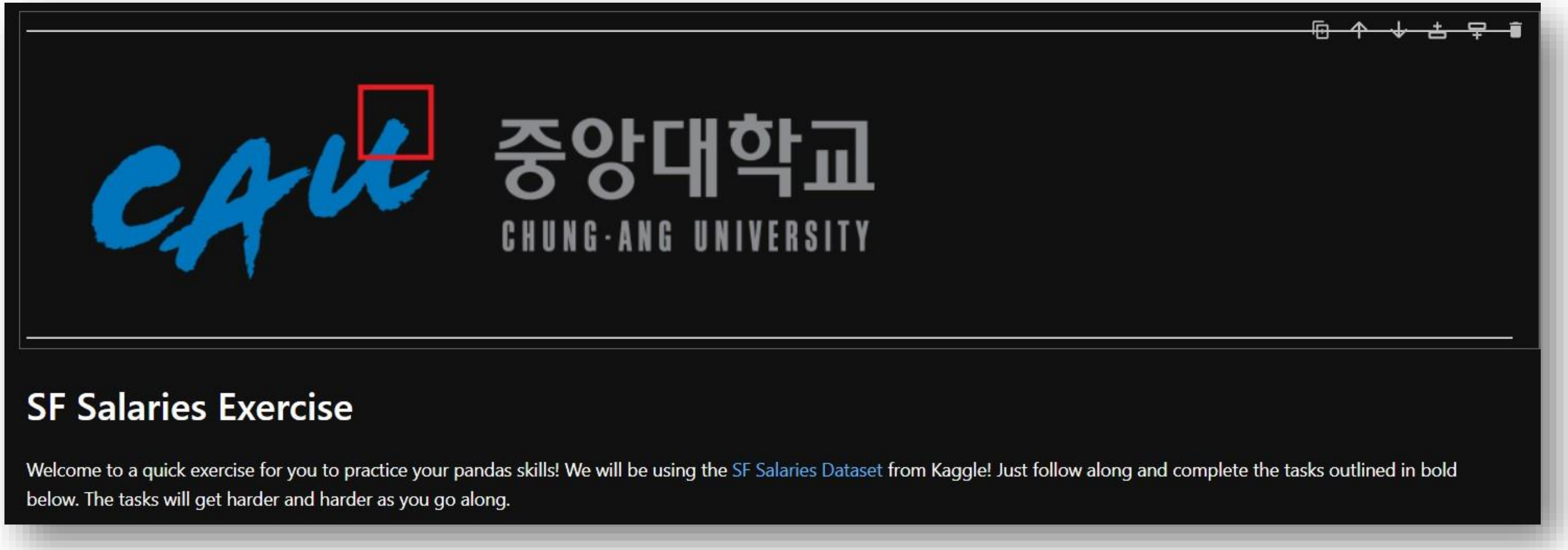
A	0.31
B	1.26

Merge, Join



8. Pandas

- Exercise resources
  - Download the exercise folder named “Exercise-26.04.2023” from “Workshop Resources” Unzip and upload to Jupyterlab to run.
  - Note: Solutions provided

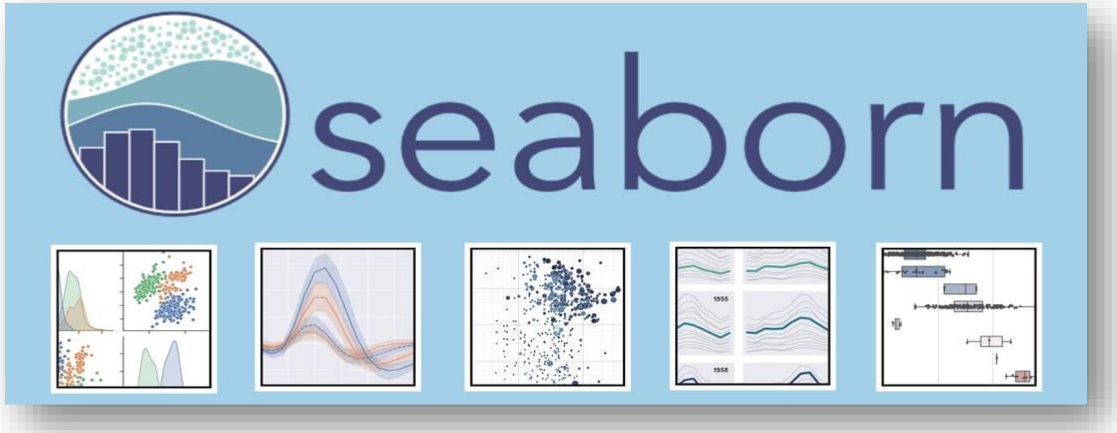


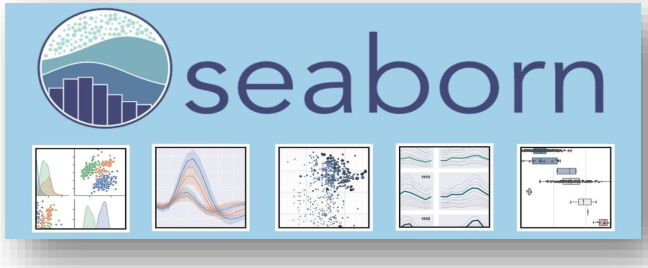
SF Salaries Exercise

Welcome to a quick exercise for you to practice your pandas skills! We will be using the SF Salaries Dataset from Kaggle! Just follow along and complete the tasks outlined in bold below. The tasks will get harder and harder as you go along.

9. Seaborn

- Seaborn is a statistical plotting library for data visualization
- It has beautiful default styles
- It also is designed to work very well with pandas dataframe objects



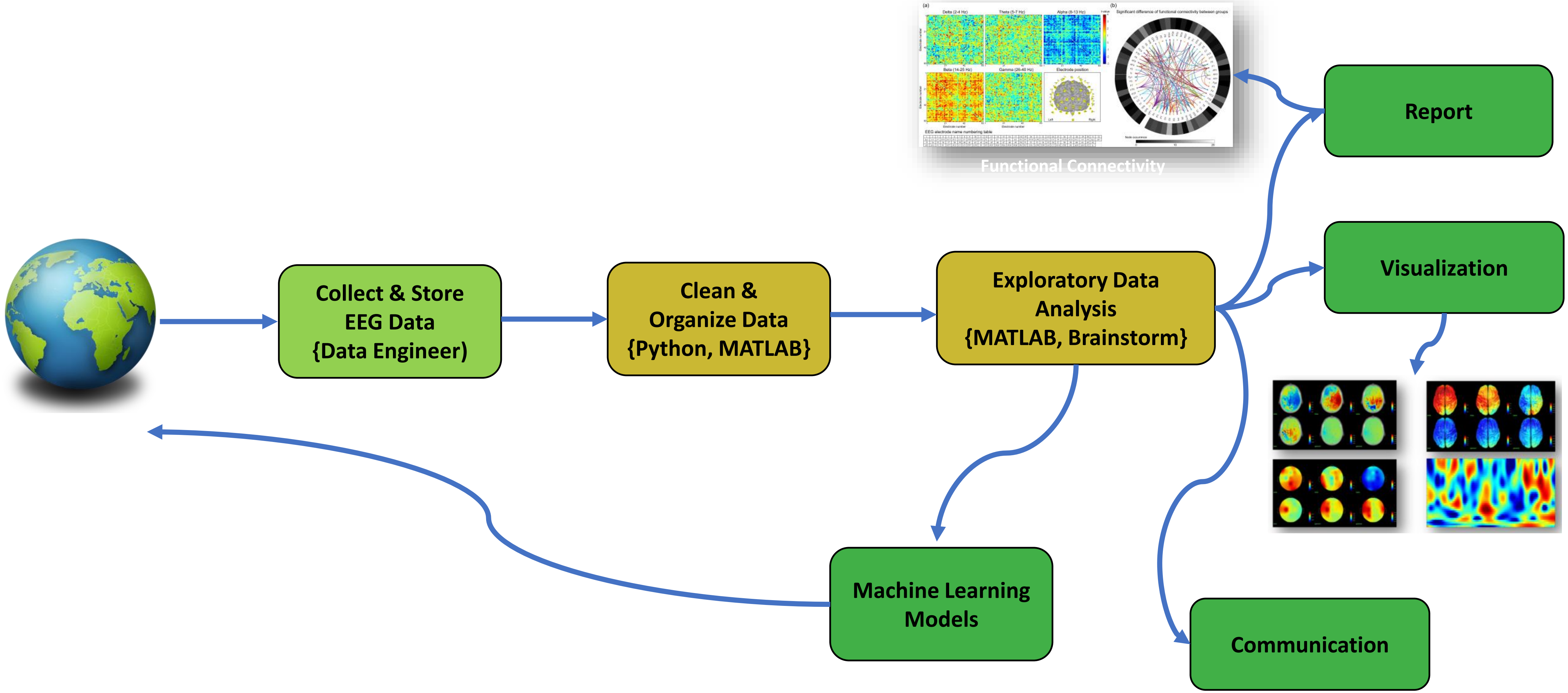


Install Seaborn

- You’ll need to install seaborn by going to your command line or terminal and using either
  - `conda install seaborn`
  - `pip install seaborn`

Documentation

- https://seaborn.pydata.org/



Collect & Store EEG Data {Data Engineer)

Clean & Organize Data {Python, MATLAB}

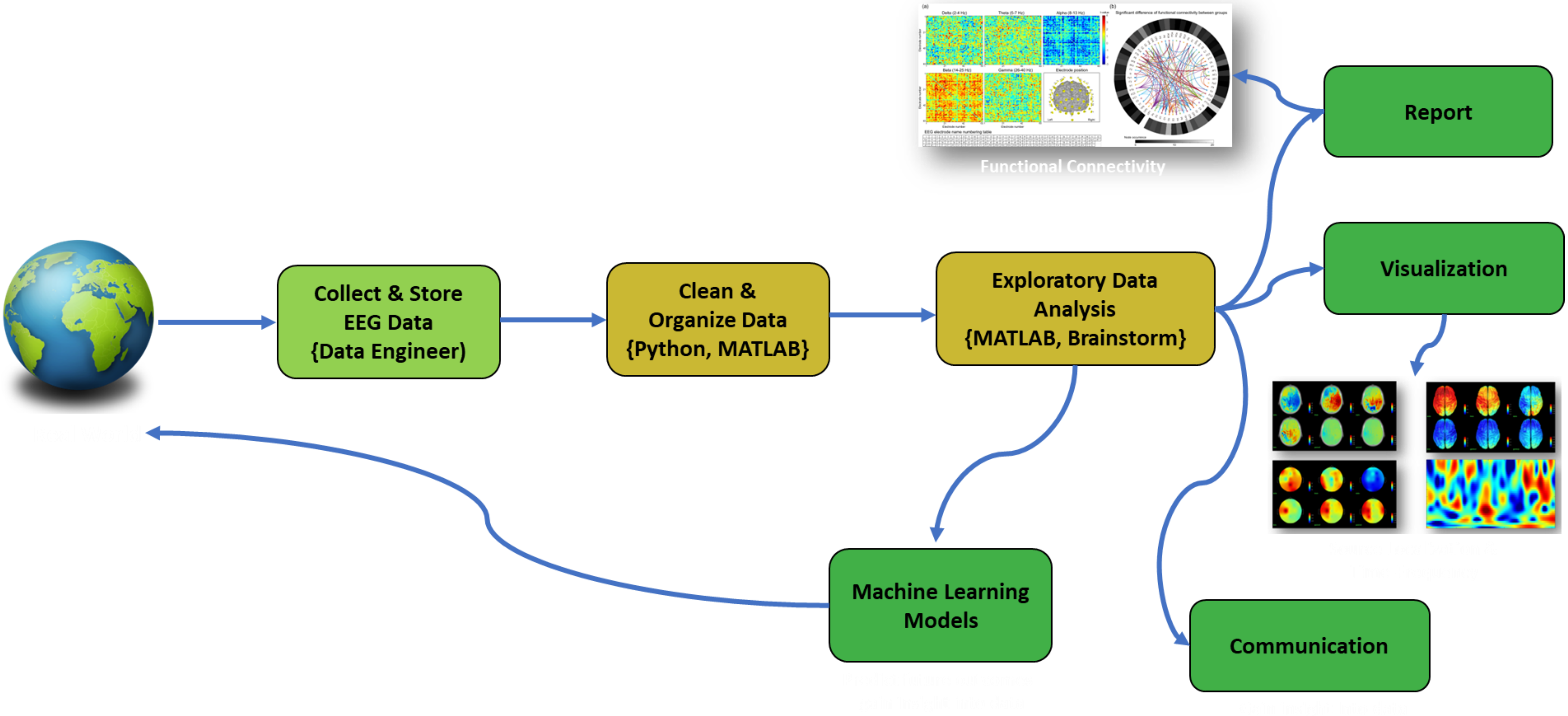
Exploratory Data Analysis {MATLAB, Brainstorm}

Report

Visualization

Machine Learning Models

Communication



Real World

Collect & Store EEG Data {Data Engineer)

Clean & Organize Data {Python, MATLAB}

Exploratory Data Analysis {MATLAB, Brainstorm}

Report

Functional Connectivity

Visualization

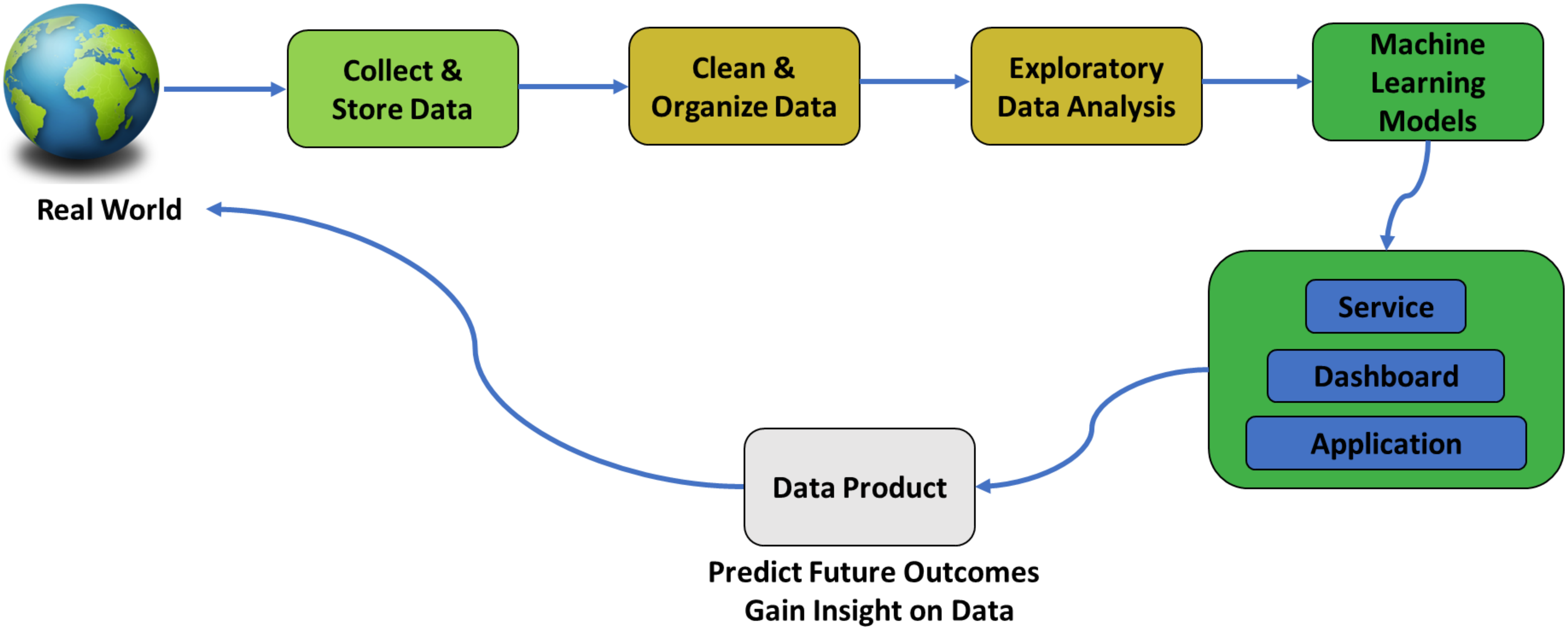
Sources Localization & Time-Frequency

Communication

Gain insight into data

Machine Learning Models

Predict future outcomes gain insight into data



Collect & Store Data

Clean & Organize Data

Exploratory Data Analysis

Machine Learning Models

Service

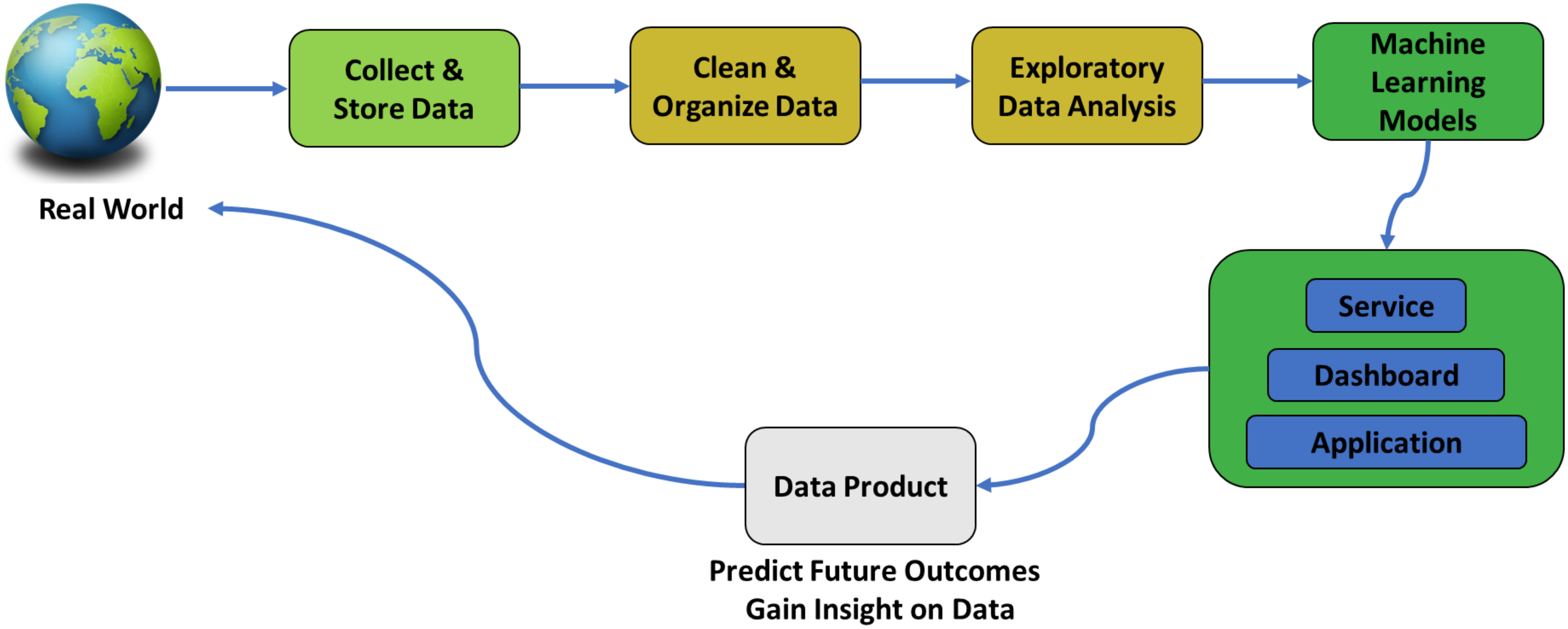
Dashboard

Application

Data Product

Predict Future Outcomes
Gain Insight on Data

Real World



Real World

Collect & Store Data

Clean & Organize Data

Exploratory Data Analysis

Machine Learning Models

Service

Dashboard

Application

Data Product

Predict Future Outcomes
Gain Insight on Data