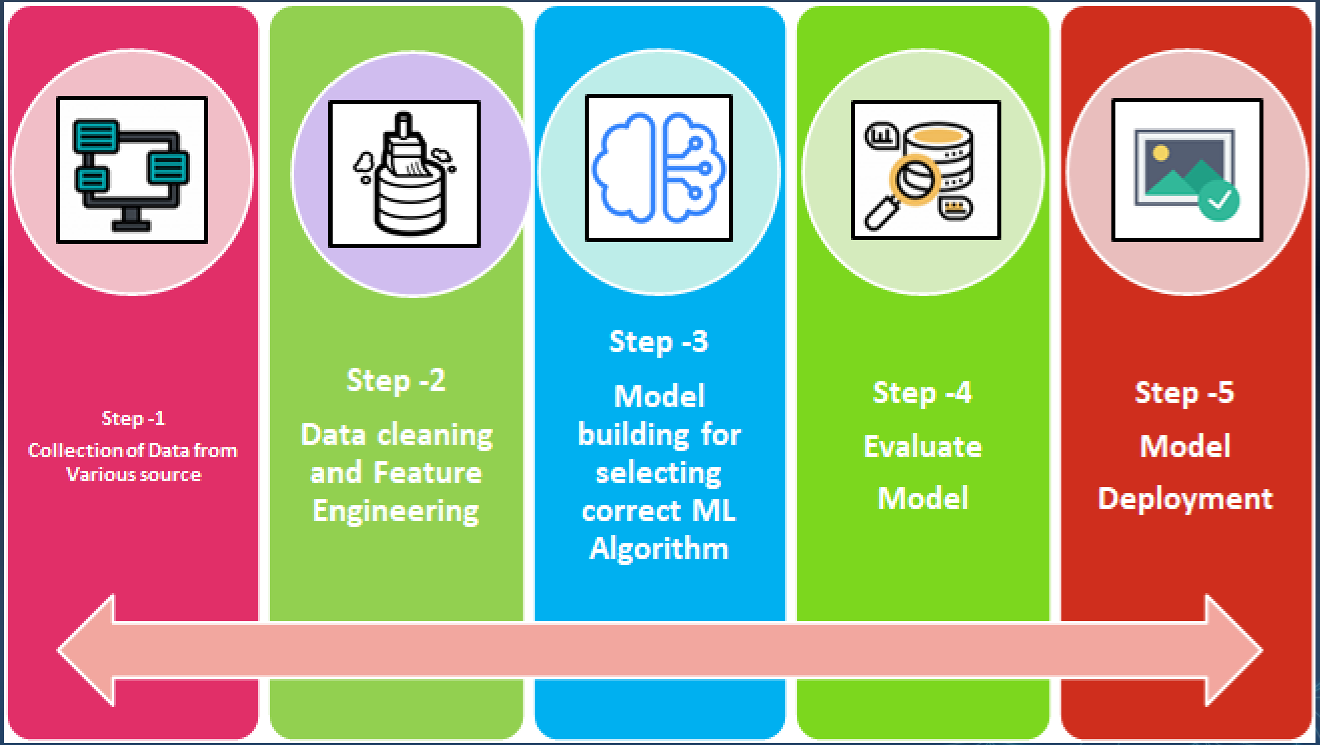
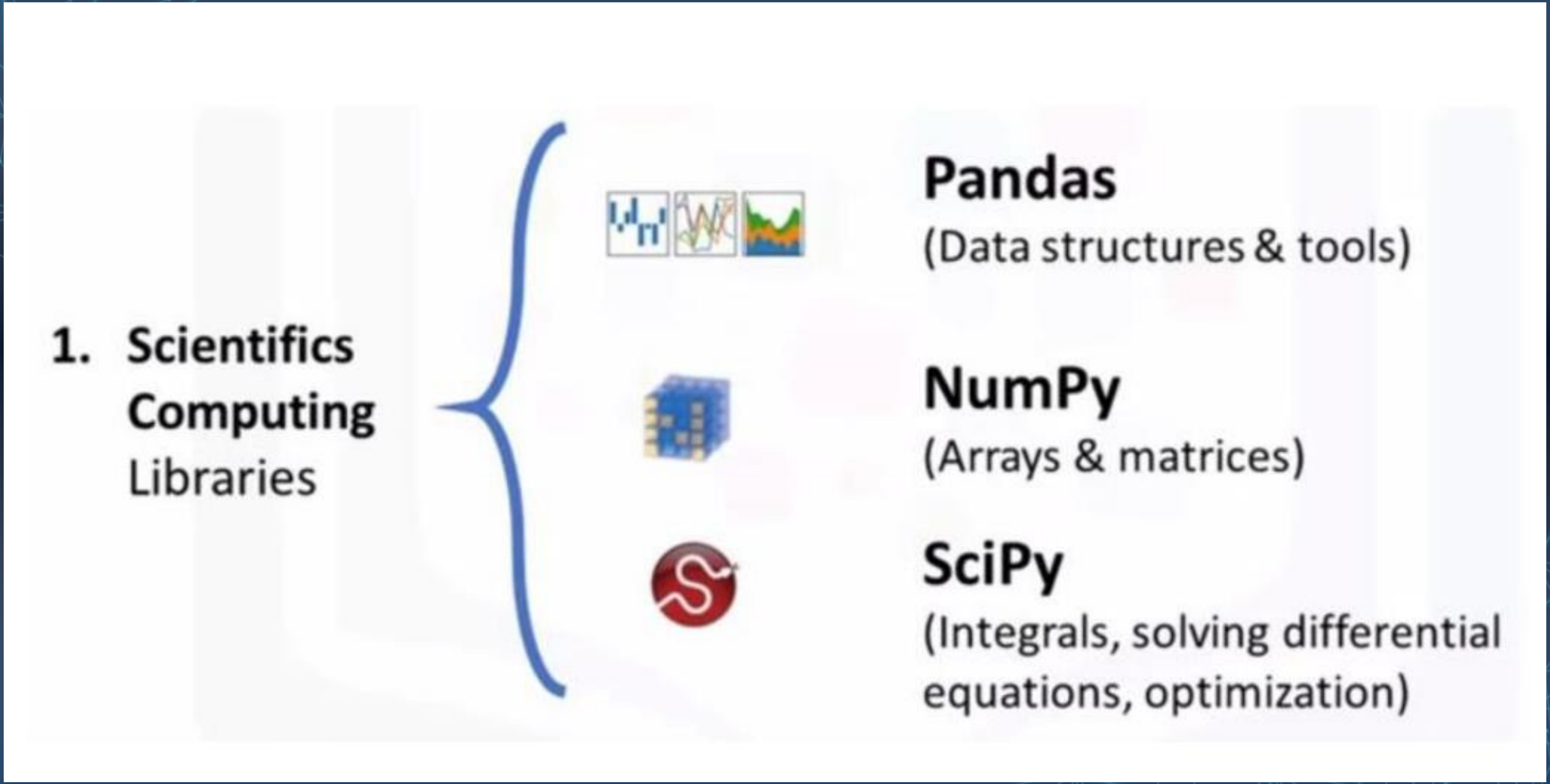
Data Management & Visualization

Eng- Mohamed Khaled Idris
Eng- Mayar Swilam

Review:


Step -1
Collection of Data from Various source
Step -2
Data cleaning and Feature Engineering
Step -3
Model building for selecting correct ML Algorithm
Step -4
Evaluate Model
Step -5
Model Deployment

Python Libraries



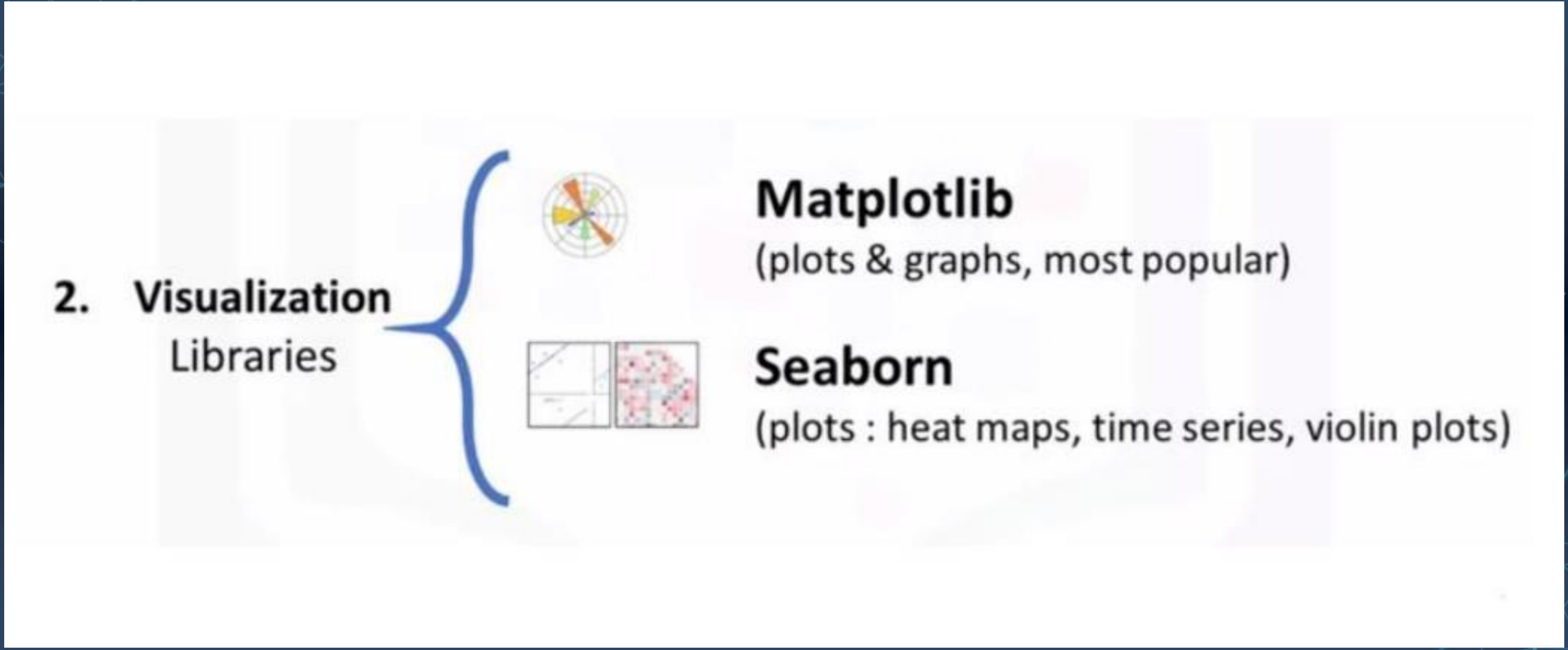
1. **Scientifics Computing** Libraries

Pandas
(Data structures & tools)

NumPy
(Arrays & matrices)

SciPy
(Integrals, solving differential equations, optimization)

Python Libraries

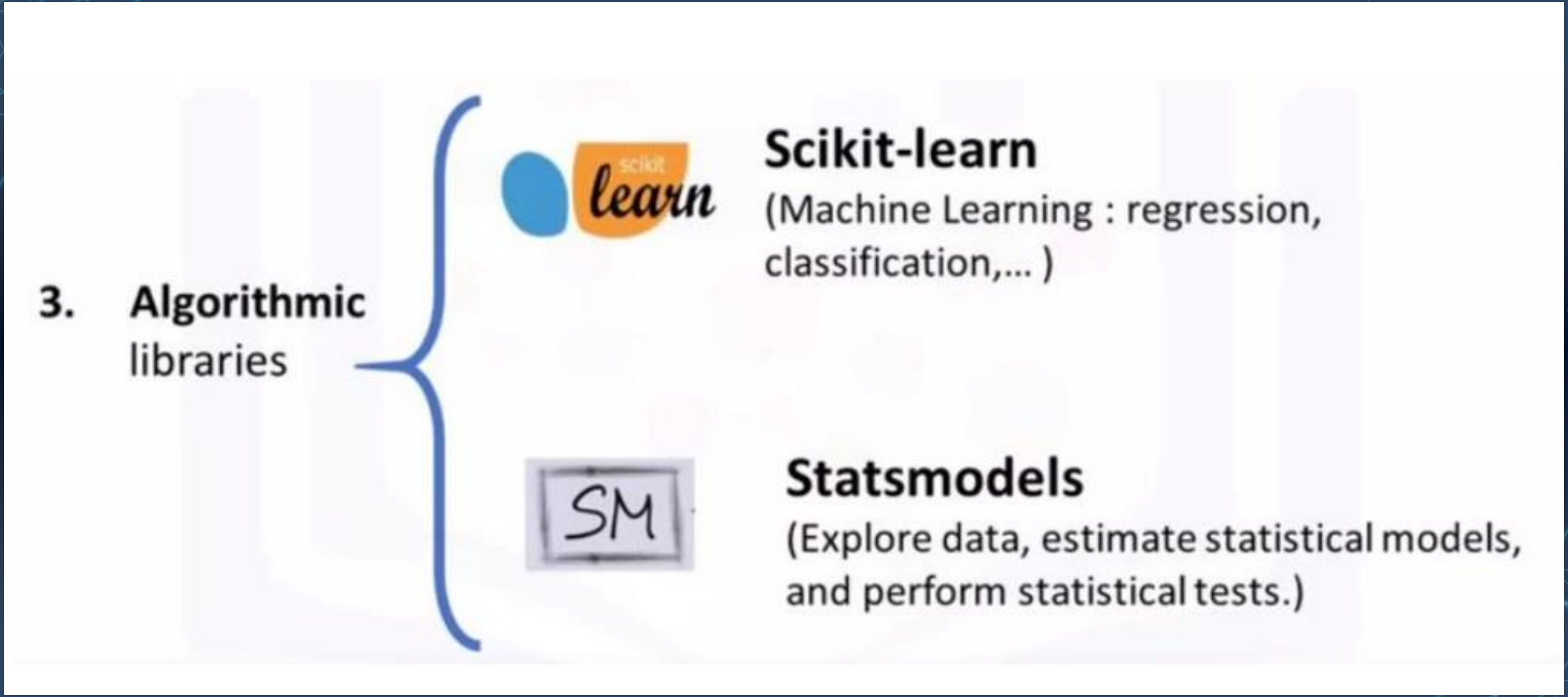


2. Visualization Libraries

Matplotlib
(plots & graphs, most popular)

Seaborn
(plots : heat maps, time series, violin plots)

Python Libraries



3. **Algorithmic** libraries

Scikit-learn
(Machine Learning : regression, classification,...)

Statsmodels
(Explore data, estimate statistical models, and perform statistical tests.)

Python Libraries

TensorFlow

PyTorch

TENSORFLOW & PYTORCH

Data Management

- In order to manage data through python, there are many ways to do it !
- Also, you may need to do some mathematical operations to matrices, lists or any structure
- You can simply manage your data as the same as SQL

HOW THEN ?

NumPy

What is NumPy ?

- NumPy is an open-source Python library that facilitates efficient numerical operations on large quantities of data.

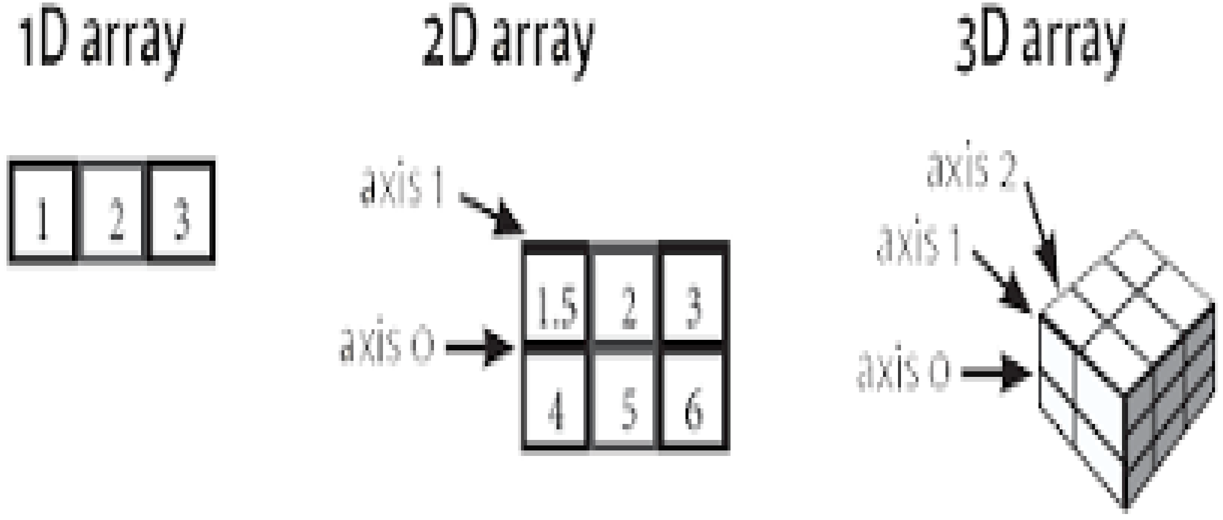
- There are a few functions that exist in NumPy that we use on pandas DataFrames.

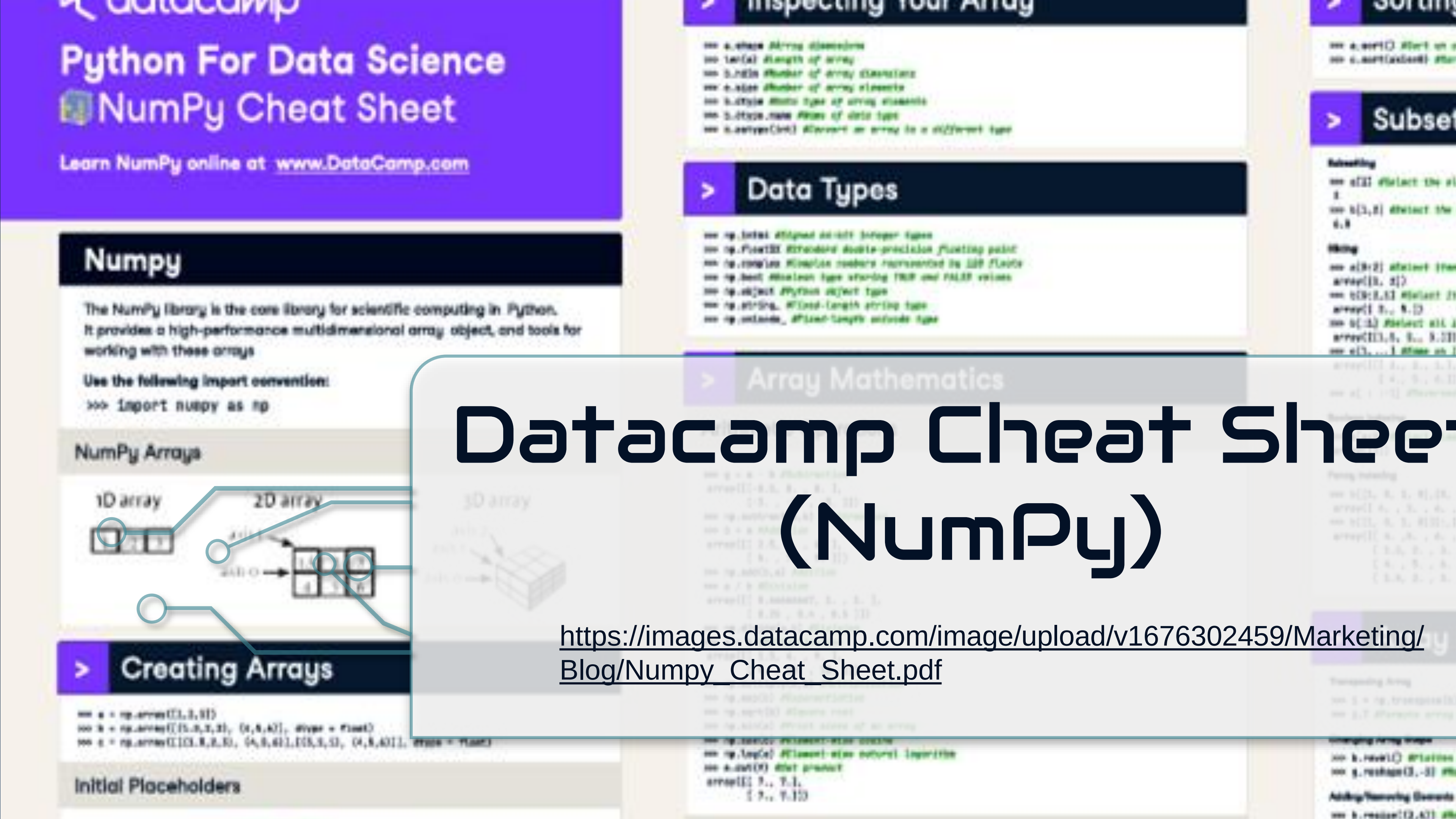
- For us, the most important part about NumPy is that pandas is built on top of it. So, NumPy is a dependency of Pandas.

NumPy Arrays

- NumPy arrays are unique in that they are more flexible than normal Python lists.
- They are called ndarrays since they can have any number (n) of dimensions (d).
- They hold a collection of items of any one data type and can be either a vector (one-dimensional) or a matrix (multi-dimensional).
- NumPy arrays allow for fast element access and efficient data manipulation.

NumPy Arrays





Python For Data Science
NumPy Cheat Sheet

Learn NumPy online at www.DataCamp.com

Numpy

The NumPy library is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays

Use the following import convention:

```
>>> import numpy as np
```

NumPy Arrays

Creating Arrays

```
>>> a = np.array([1,2,3])
>>> b = np.array([(1.5,2,3), (4,5,6)], dtype = float)
>>> c = np.array([[(1.5,2,3), (4,5,6)],[(3,2,1), (4,5,6)]], dtype = float)
```

Initial Placeholders

Inspecting Your Array

```
>>> a.shape #Array dimensions
>>> len(a) #Length of array
>>> b.ndim #Number of array dimensions
>>> e.size #Number of array elements
>>> b.dtype #Data type of array elements
>>> b.dtype.name #Name of data type
>>> b.astype(int) #Convert an array to a different type
```

Data Types

```
>>> np.int64 #Signed 64-bit integer types
>>> np.float32 #Standard double-precision floating point
>>> np.complex #Complex numbers represented by 128 floats
>>> np.bool #Boolean type storing TRUE and FALSE values
>>> np.object #Python object type
>>> np.string_ #Fixed-length string type
>>> np.unicode_ #Fixed-length unicode type
```

Array Mathematics

```
>>> np.log(a) #Element-wise natural logarithm
>>> e.dot(f) #Dot product
array([[ 7., 7.],
       [ 7., 7.]])
```

Subset

Datacamp Cheat Sheet (NumPy)

https://images.datacamp.com/image/upload/v1676302459/Marketing/Blog/Numpy_Cheat_Sheet.pdf

Pandas

What is Pandas ?

- *Pandas* is a very popular library for working with data (its goal is to be the most powerful and flexible open-source tool, and in our opinion, it has reached that goal).

- DataFrames are at the center of pandas. A DataFrame is structured like a table or spreadsheet. The rows and the columns both have indexes, and you can perform operations on rows or columns separately.

DataFrame

- A pandas DataFrame can be easily changed and manipulated. Pandas has helpful functions for handling missing data, performing operations on columns and rows, and transforming data.

- If that wasn’t enough, a lot of SQL functions have counterparts in pandas, such as join, merge, filter by, and group by.

- With all of these powerful tools, it should come as no surprise that pandas is very popular among data scientists.

Datacamp Cheat Sheet (Pandas)

https://images.datacamp.com/image/upload/v1676302204/Marketing/Blog/Pandas_Cheat_Sheet.pdf

Learn Pandas Basics online at www.DataCamp.com

Pandas

The Pandas library is built on NumPy and provides easy-to-use data structures and data analysis tools for the Python programming language.

Use the following import convention:

```
>>> import pandas as pd
```

Pandas Data Structures

Series

A one-dimensional labeled array capable of holding any data type

Dataframe

A two-dimensional labeled data structure with columns of potentially different types

Dropping

Asking For Help

Sort & Rank

Read and Write to Excel

Read and Write to SQL Query or Database Table

Selection

Also see NumPy Arrays

Getting

Boolean indexing

Setting

Summary

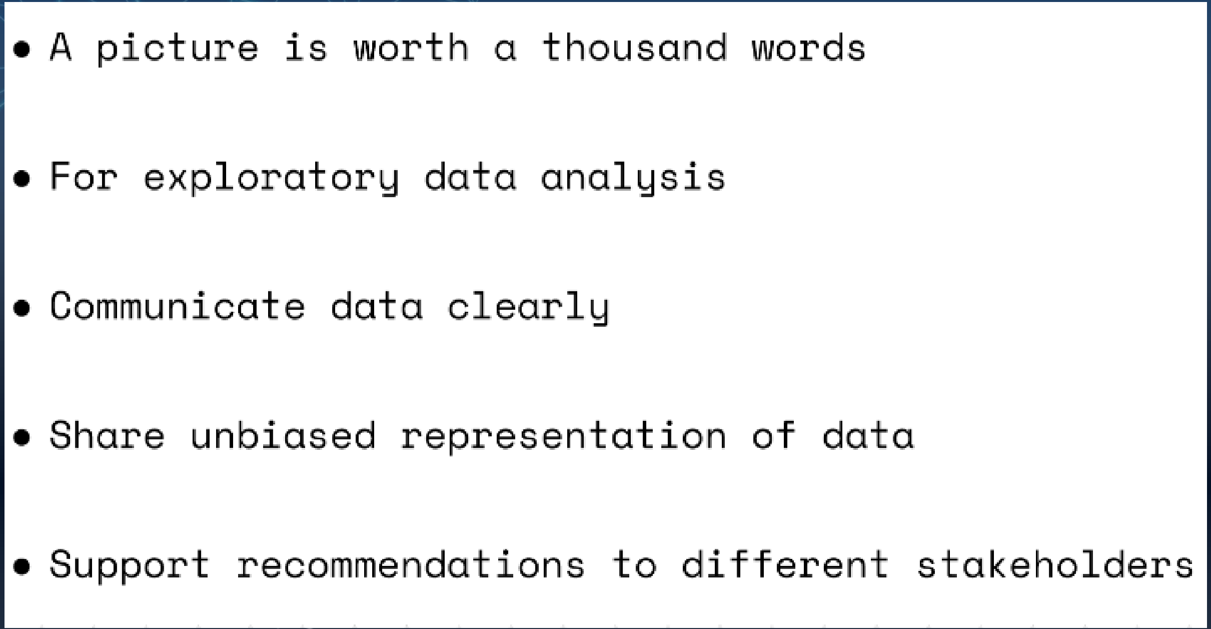
Applying Functions

Data Alignment

Internal Data Alignment

Arithmetic Operations with Fill Methods

Learn Data Skills Online at www.DataCamp.com

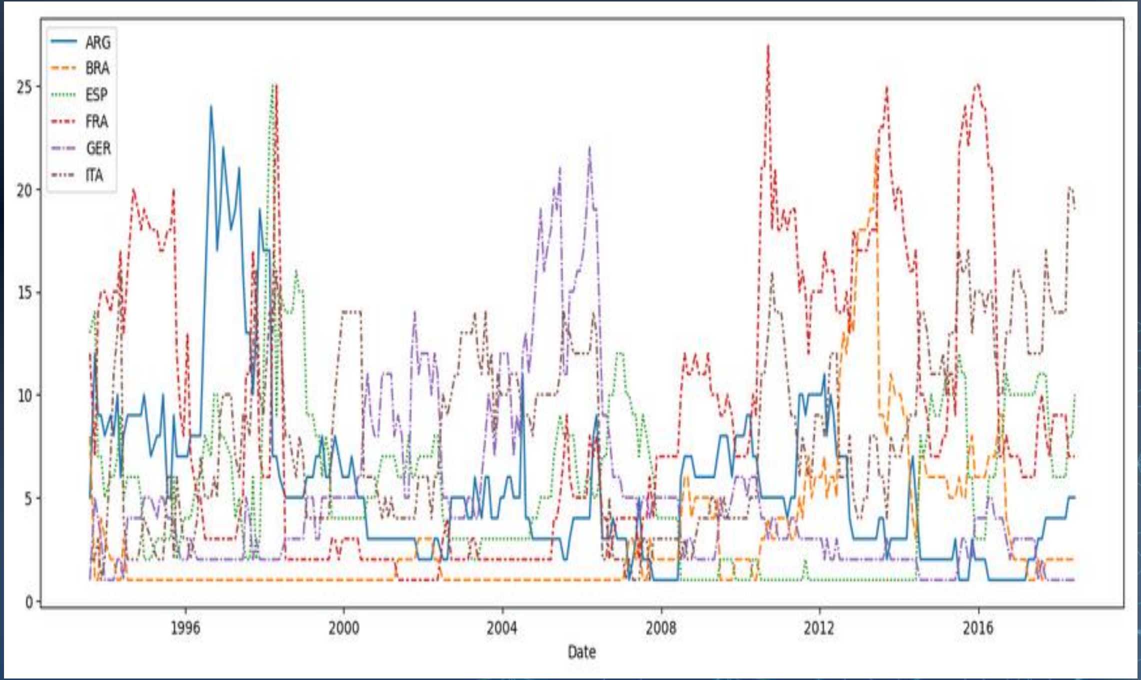


- A picture is worth a thousand words
- For exploratory data analysis
- Communicate data clearly
- Share unbiased representation of data
- Support recommendations to different stakeholders

Why Visualization ?

The important of Visualization

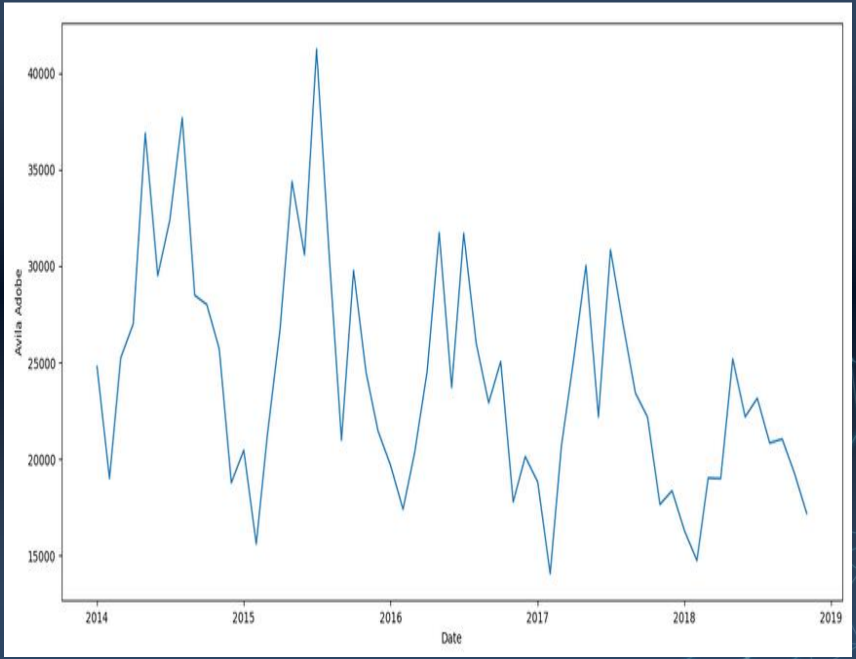
- Considering only the years represented in the figure, which countries spent at least 5 consecutive years in the #1 ranked spot?

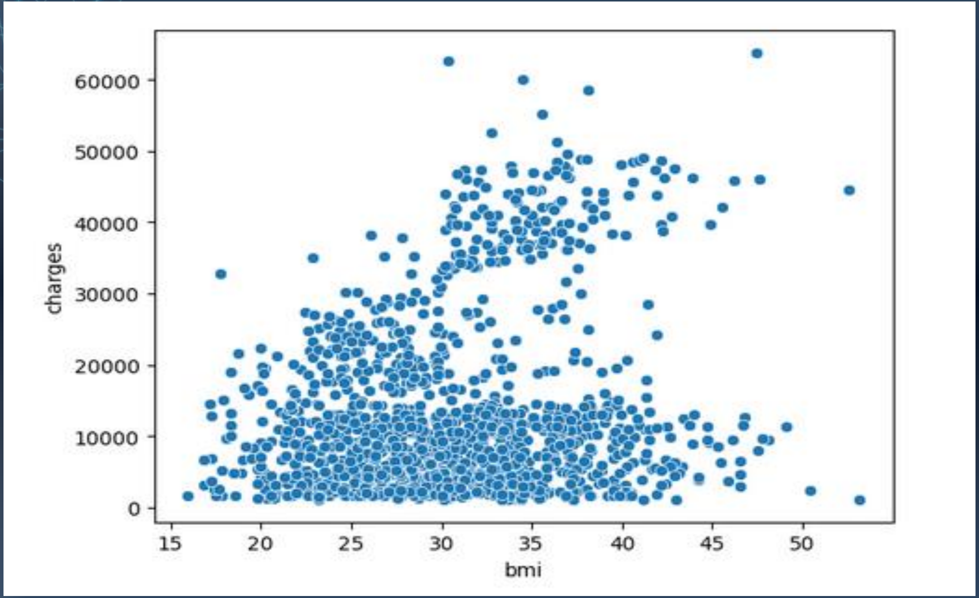


Does Avila Adobe get more visitors:

- **in September-February (in LA, the fall and winter months), or**
- **in March-August (in LA, the spring and summer)?**

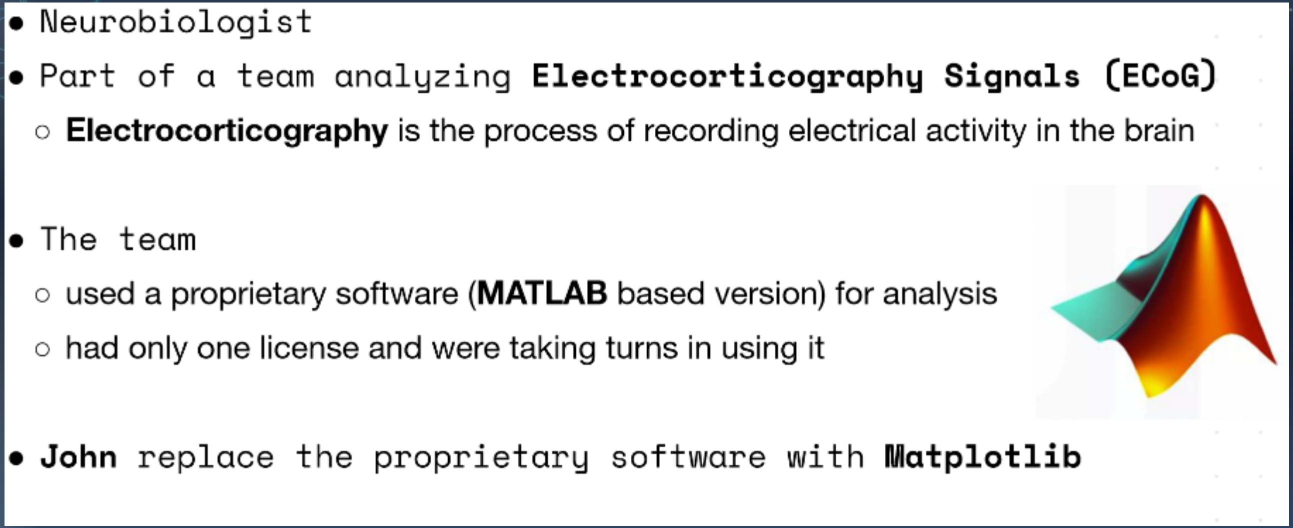
Using this information, when should the museum staff additional seasonal employees?





IS THERE ANY RELATION BETWEEN BODY MASS INDEX AND INSURANCE CHARGES ?

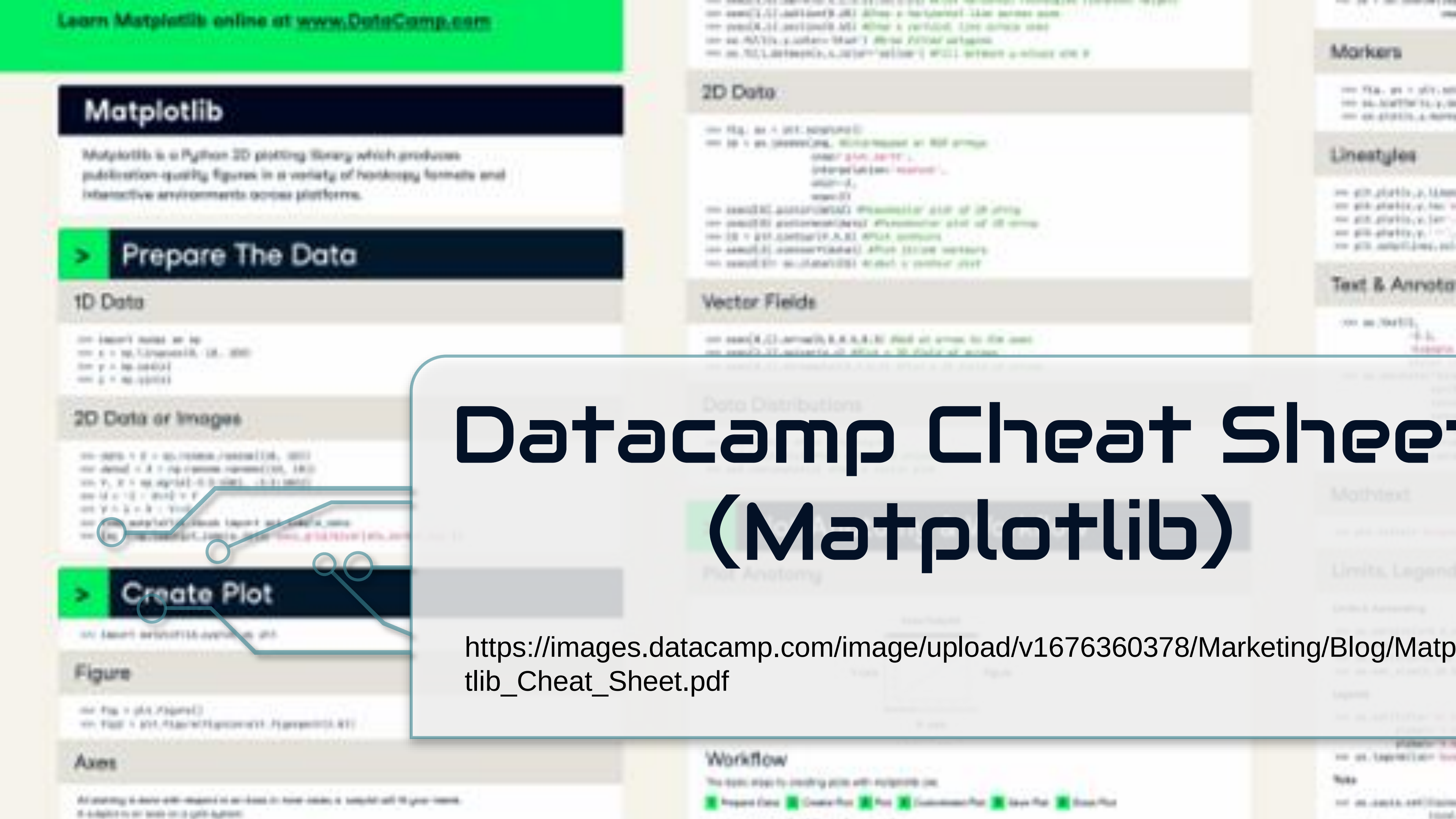
Matplotlib



- Neurobiologist
- Part of a team analyzing **Electrocorticography Signals (ECoG)**
  - **Electrocorticography** is the process of recording electrical activity in the brain

- The team
  - used a proprietary software (**MATLAB** based version) for analysis
  - had only one license and were taking turns in using it

- **John** replace the proprietary software with **Matplotlib**



Learn Matplotlib online at www.DataCamp.com

Matplotlib

Matplotlib is a Python 2D plotting library which produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.

Prepare The Data

1D Data

2D Data or Images

Create Plot

Figure

Axes

2D Data

Vector Fields

Workflow

Markers

Linestyles

Datacamp Cheat Shee (Matplotlib)

https://images.datacamp.com/image/upload/v1676360378/Marketing/Blog/Matp
tlib_Cheat_Sheet.pdf

Questions

Thank You