Feeding Our Machine Learning Pipeline

Mohammed Osman
SENIOR SOFTWARE DEVELOPER

@cognitiveosman www.cognitiveosman.com

Overview

Positioning ourselves in the ML pipeline

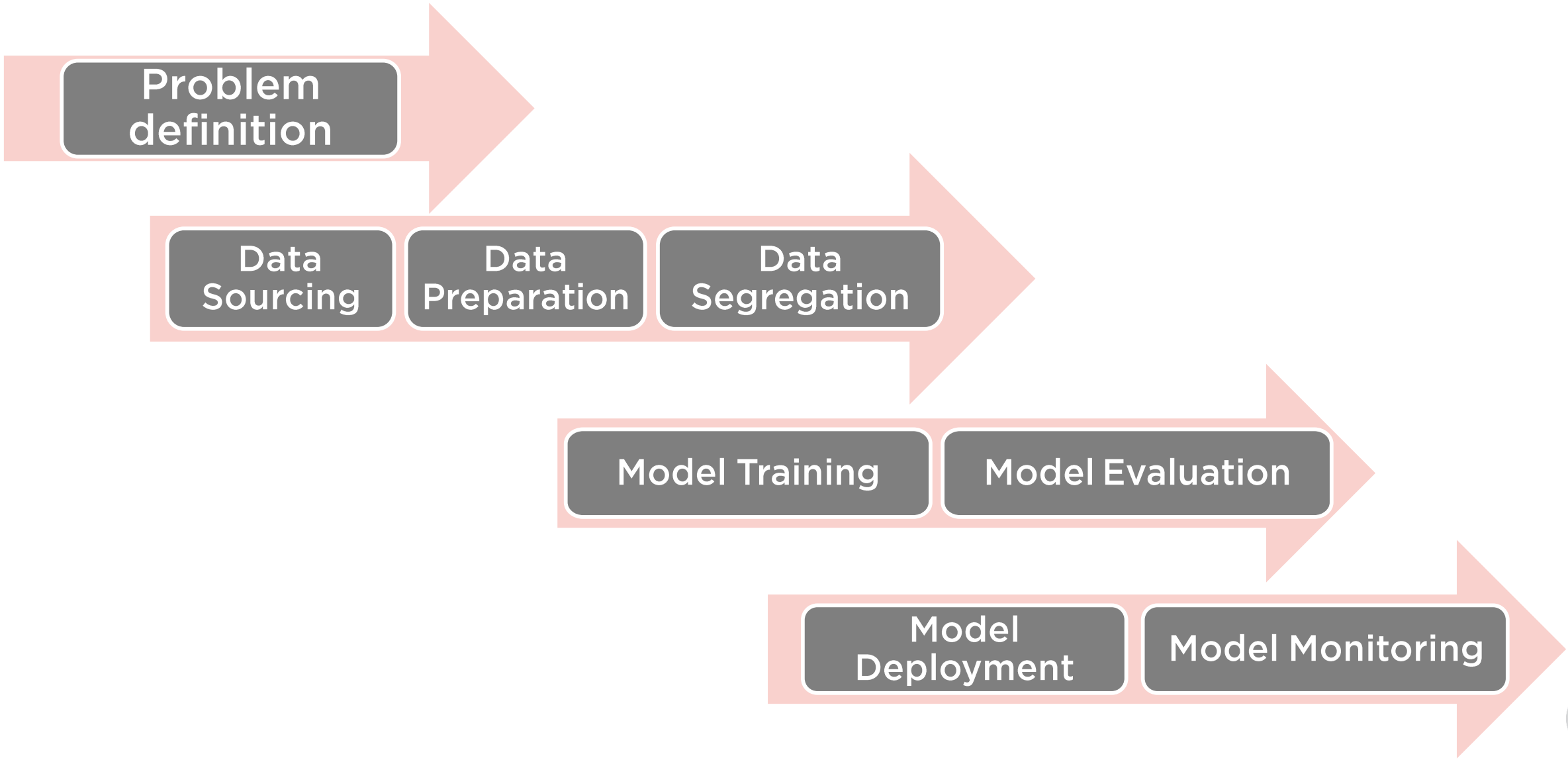
How and why we collect data?

CSV format

Important ML libraries

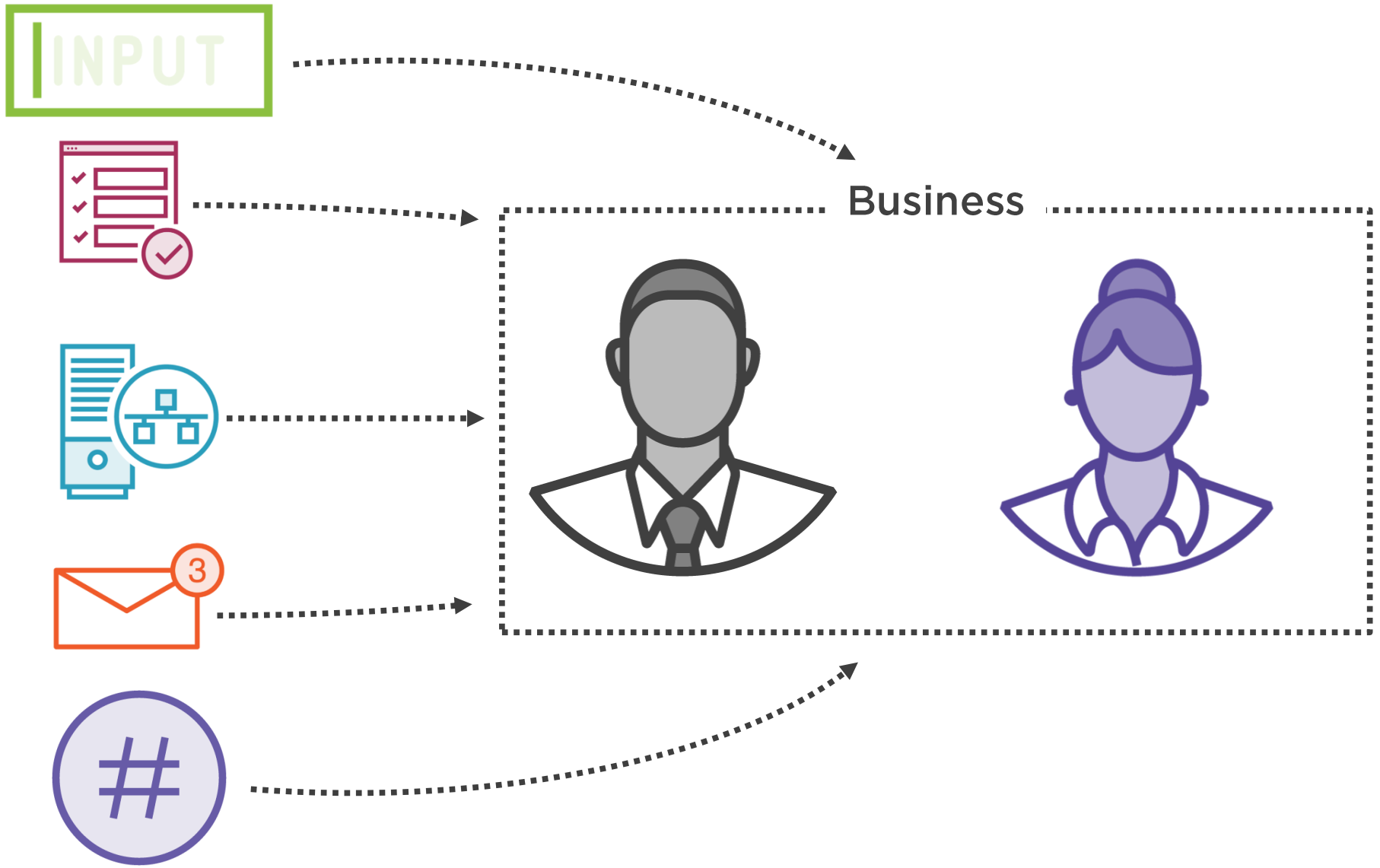
Demo: Loading our data to Python

Data Sourcing



Why Data Sourcing?

Age of Data Abundance



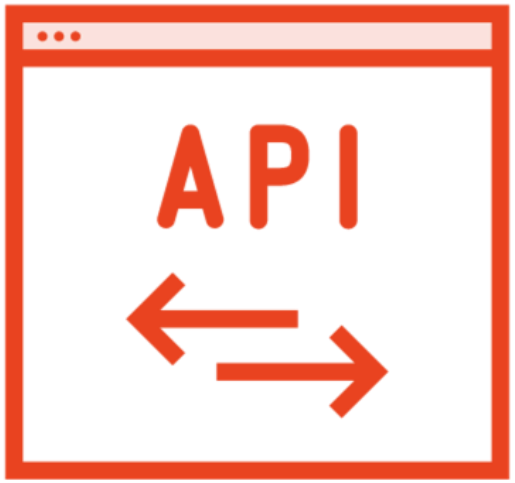
Insane!

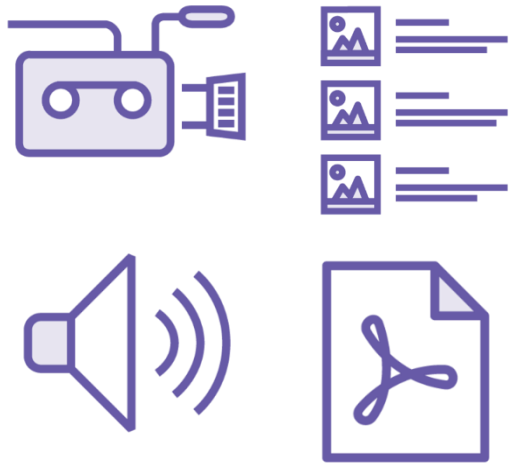
Collecting Mosaic Pieces

Challenges with Heterogeneous Data Sources

Structured Data

Unstructured Data





Proper Integration
Retry
Serialization

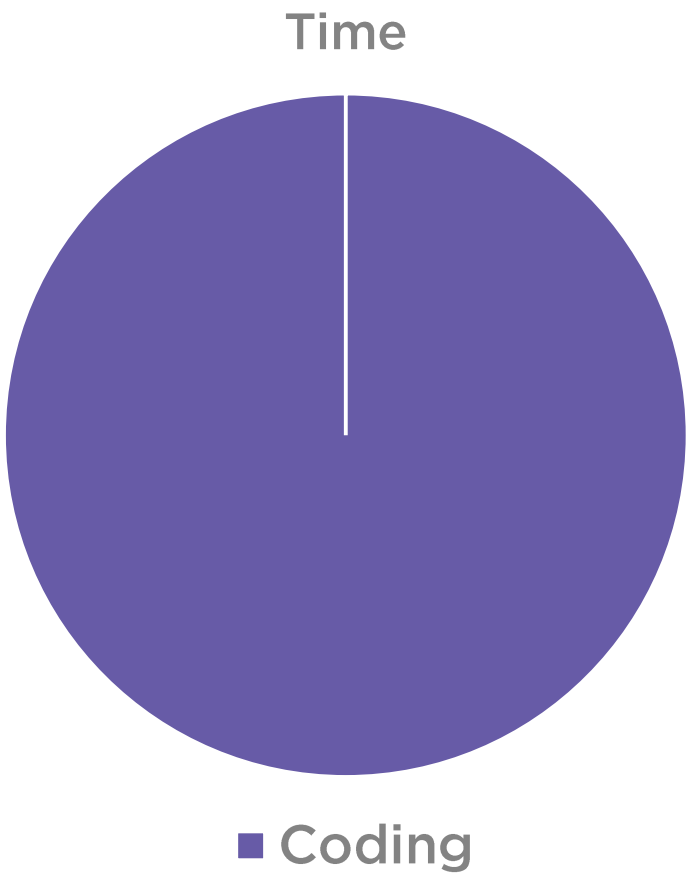
File Size
R/W Locks
Format

Performance
Queries

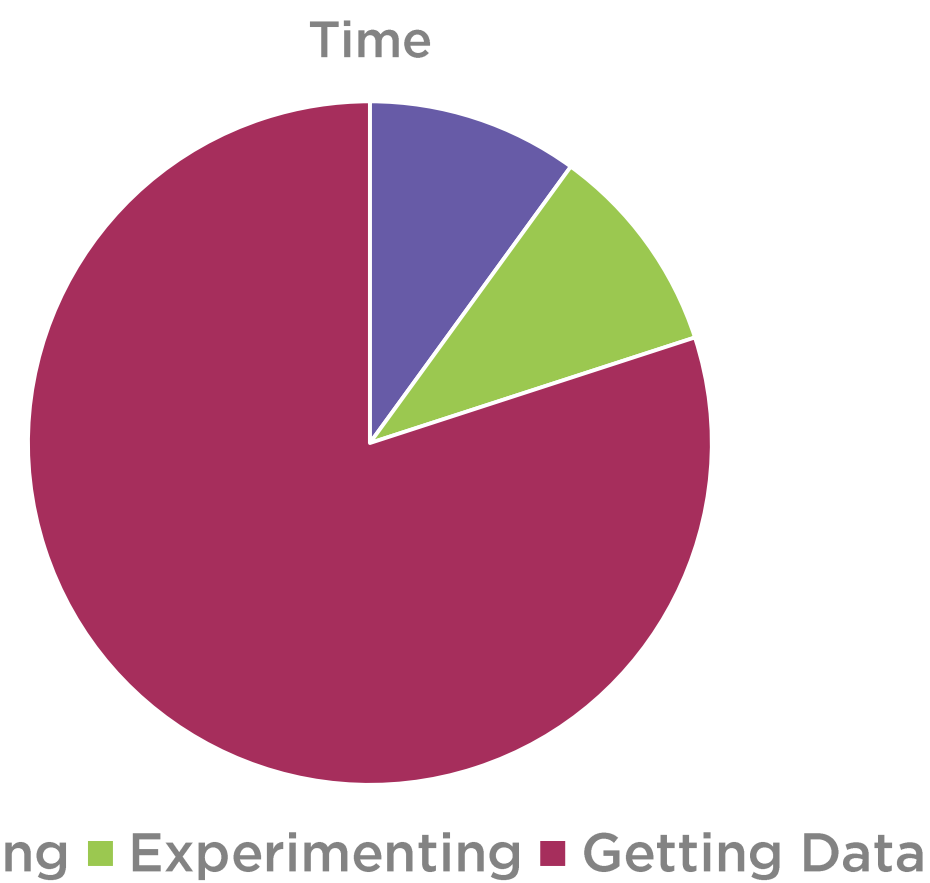
Representation
Processing
Overhead

Reality of Machine Learning Projects

What People Think!



Reality :)



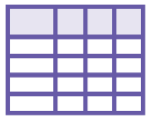
Source: https://tinyurl.com/y49ueurj

This course is not about data sourcing tools, therefore, we will use simple format “CSV” and focus on actual ML instead!

Optional: Understanding CSV Format

CSV (Comma-separated Values) Basics

Text file



Used to represent tabular format

Each line is a record

Each record has multiple columns separated by comma (delimiter)

May contain very first record for column names

CSV Example

1	Name	Age	Gender	Field
2	Peter	22	Male	Engineering
3	Sara	23	Female	Engineering
4	Ali	40	Male	Medicine
5	John	12	Male	Astronomy

```
1 Name,Age,Gender,Field
2 Peter,22,Male,Engineering
3 Sara,23,Female,Engineering
4 Ali,40,Male,Medicine
5 John,12,Male,Astronomy
```

Understanding SciPy

SciPy

Collection of open source libraries for mathematics, science and engineering

Umbrella for NumPy, Matplotlib and Pandas

Understanding NumPy

NumPy

Python library used for scientific computing

Also used for Numerical Analysis, Linear Algebra and Matrix Computations

```
import numpy

pythonArray = [[1, 2, 3],[9,8,7]]

npArray = numpy.array(pythonArray)

print(npArray.shape)

print(npArray[0])

print(npArray[:,1])
```

- Importing NumPy library
- NumPy array from Python array
- NumPy array shape (2,3)
- Print first row [1 2 3]
- Print whole first column [1 9]
- And general array operations in Python

Understanding Matplotlib

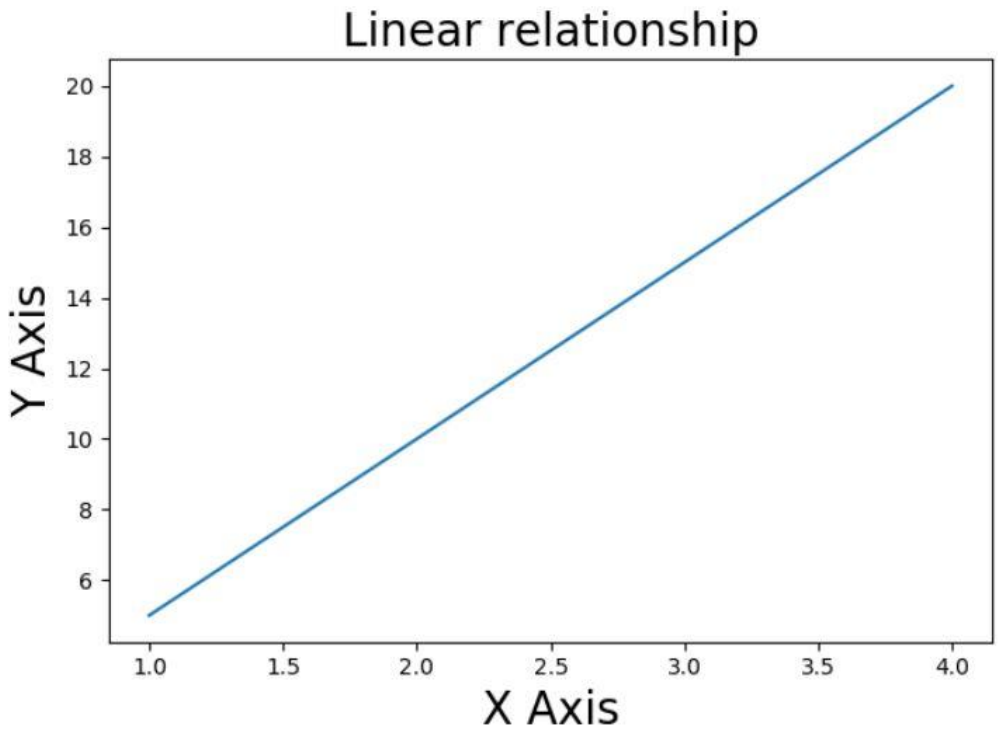
Matplotlib

Python 2D plotting library which produces quality figures

Supports many different types of figures

```
import matplotlib.pyplot as plt

plt.plot([1,2,3,4],[5,10,15,20])

plt.title("Linear relationship",
fontsize=20)

plt.xlabel("X Axis", fontsize=20)

plt.ylabel("Y Axis", fontsize=20)

plt.show()
```

- Importing Matplot library
- Setting up plotting values
- Setting up plot title
- Setting up X and Y axes
- Showing the plot



Understanding Pandas

Pandas

Python library that helps with data analysis

Calculates statistics, cleans data, and persists it

Core Components of Pandas

	Cars
0	5
1	4
2	1
3	7

	Boats
0	2
1	6
2	0
3	2

Series

	Cars	Boats
0	5	2
1	4	6
2	1	0
3	7	2

DataFrame

```
import pandas as pd

data = {

    'cars': [5, 4, 1, 7], #series

    'boats': [2, 6, 0, 2] #series

}

vehicles = pd.DataFrame(data,
index=['Peter', 'Sara', 'Ali',
'John'])

print(vehicles.info())

print(vehicles.loc['Ali’])

print(vehicles.head())
```

- Importing Pandas
- Defining DataFrame from two Series
- Assigning indices to rows specific DataFrame element
- Getting information about the data
- Indexing certain element in the frame
- Get first data elements

Demo

Loading our data using Python

Summary

Machine Learning pipeline revisited

Data sourcing why and how

CSV format

SciPy

- Numpy
- Matplotlib
- Pandas

Demo to load our data