



Andhra Pradesh State Skill Development Corporation (APSSDC)



Data Analysis Using Python



NumPy

pandas

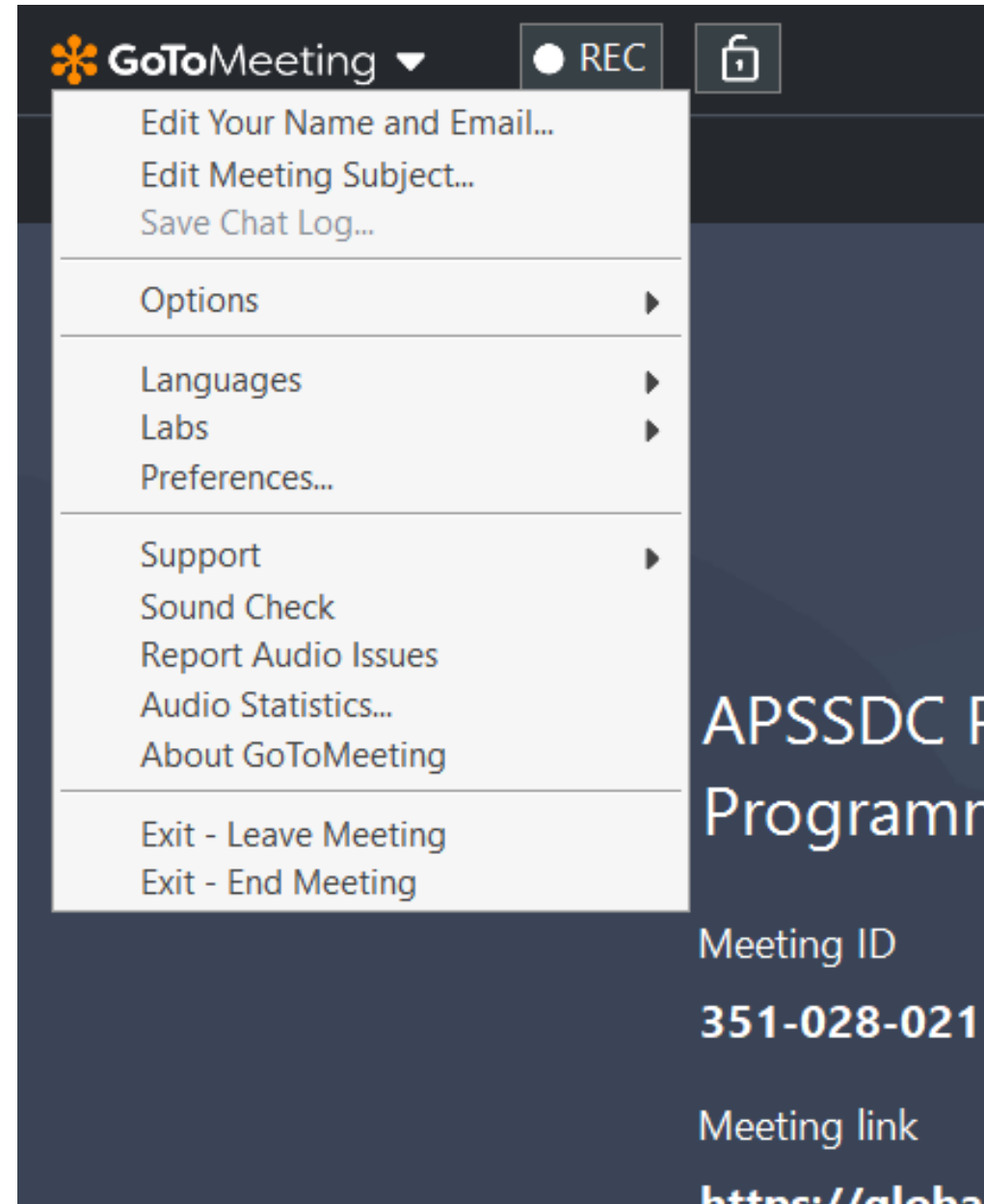


matplotlib



For Attendance and
Verification Purpose

**RollNo-Name-
CollegeCode/
CollegeName
And Registered
Email ID**



The screenshot displays the GoToMeeting application interface. At the top, the GoToMeeting logo is on the left, and 'REC' and a lock icon are on the right. A dropdown menu is open, listing options: 'Edit Your Name and Email...', 'Edit Meeting Subject...', 'Save Chat Log...', 'Options', 'Languages', 'Labs', 'Preferences...', 'Support', 'Sound Check', 'Report Audio Issues', 'Audio Statistics...', 'About GoToMeeting', 'Exit - Leave Meeting', and 'Exit - End Meeting'. On the right side of the interface, the text 'APSSDC P' and 'Program' is visible. Below this, the 'Meeting ID' is shown as '351-028-021', and the 'Meeting link' is partially visible as 'https://globe'.

GoToMeeting ▼ ● REC 🔒

- Edit Your Name and Email...
- Edit Meeting Subject...
- Save Chat Log...
- Options ▶
- Languages ▶
- Labs ▶
- Preferences...
- Support ▶
- Sound Check
- Report Audio Issues
- Audio Statistics...
- About GoToMeeting
- Exit - Leave Meeting
- Exit - End Meeting

APSSDC P
Program

Meeting ID
351-028-021

Meeting link
<https://globe>

Session Resources

<http://bit.ly/apssdc-da-mb1>

Agenda

Day1

Intro to Data
and Data
Manipulation
with NumPy

Day2

Data Analysis
with pandas

Day3

Data
Preprocessing
with Scikit-
Learn

Day4

Cleaning Data
in Python

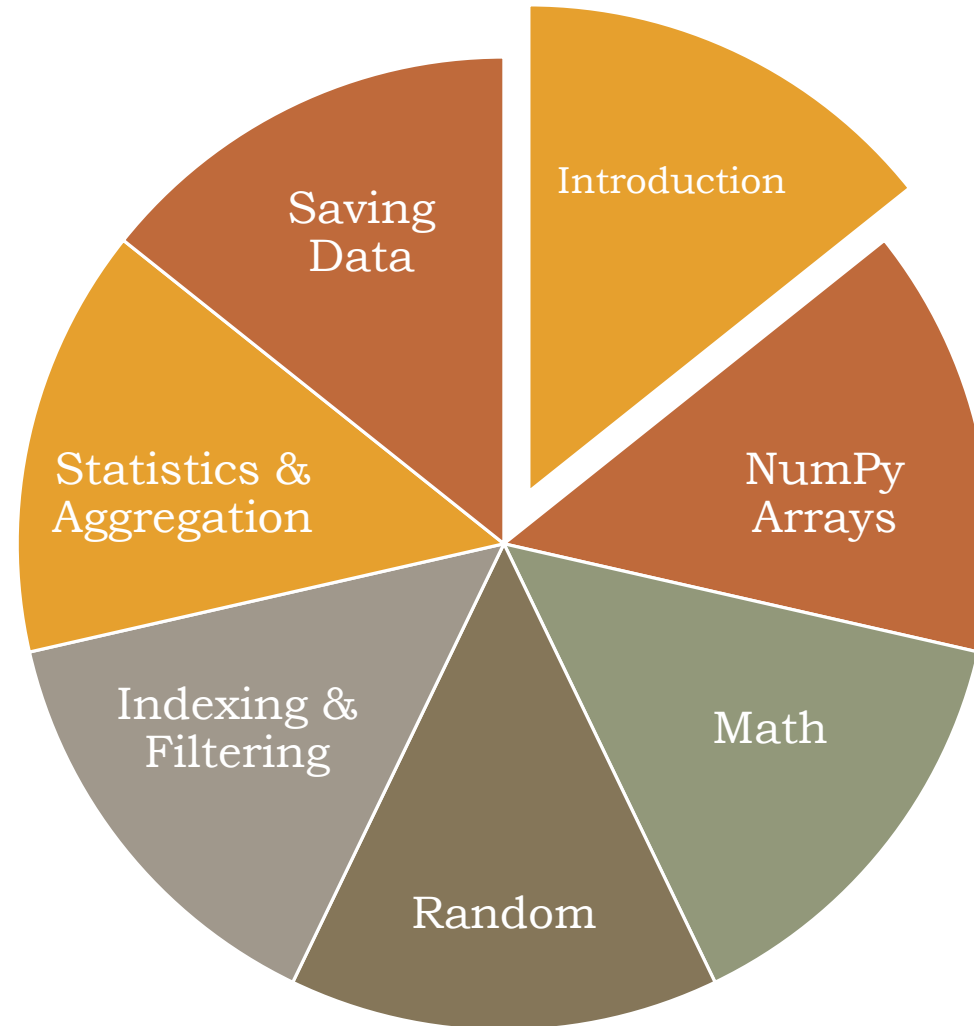
Day5

Introduction
to Data
Visualization
& Matplotlib

Day6

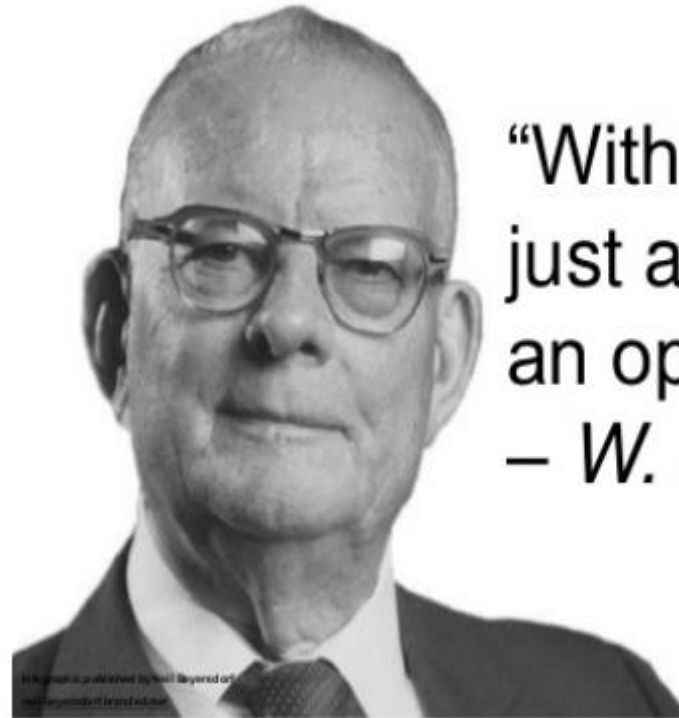
Data
Visualization
using Seaborn

Introduction Data and Data Manipulation with NumPy



What is Data?

Data are facts and statistics collected together for reference or analysis.



“Without data you’re just another person with an opinion.”
– *W. Edwards Deming*

Interesting insights

Bombardier showcased its C Series jetliner that carries Pratt & Whitney's Geared Turbo Fan (GTF) engine, which is fitted with 5,000 sensors that generate up to 10 GB of data per second. A single twin-engine aircraft with an average 12-hr. flight-time can produce up to 844 TB of data.

Saudi Aramco laid 650km of new pipelines across a mountain range of red sand dunes. How do they monitor that?

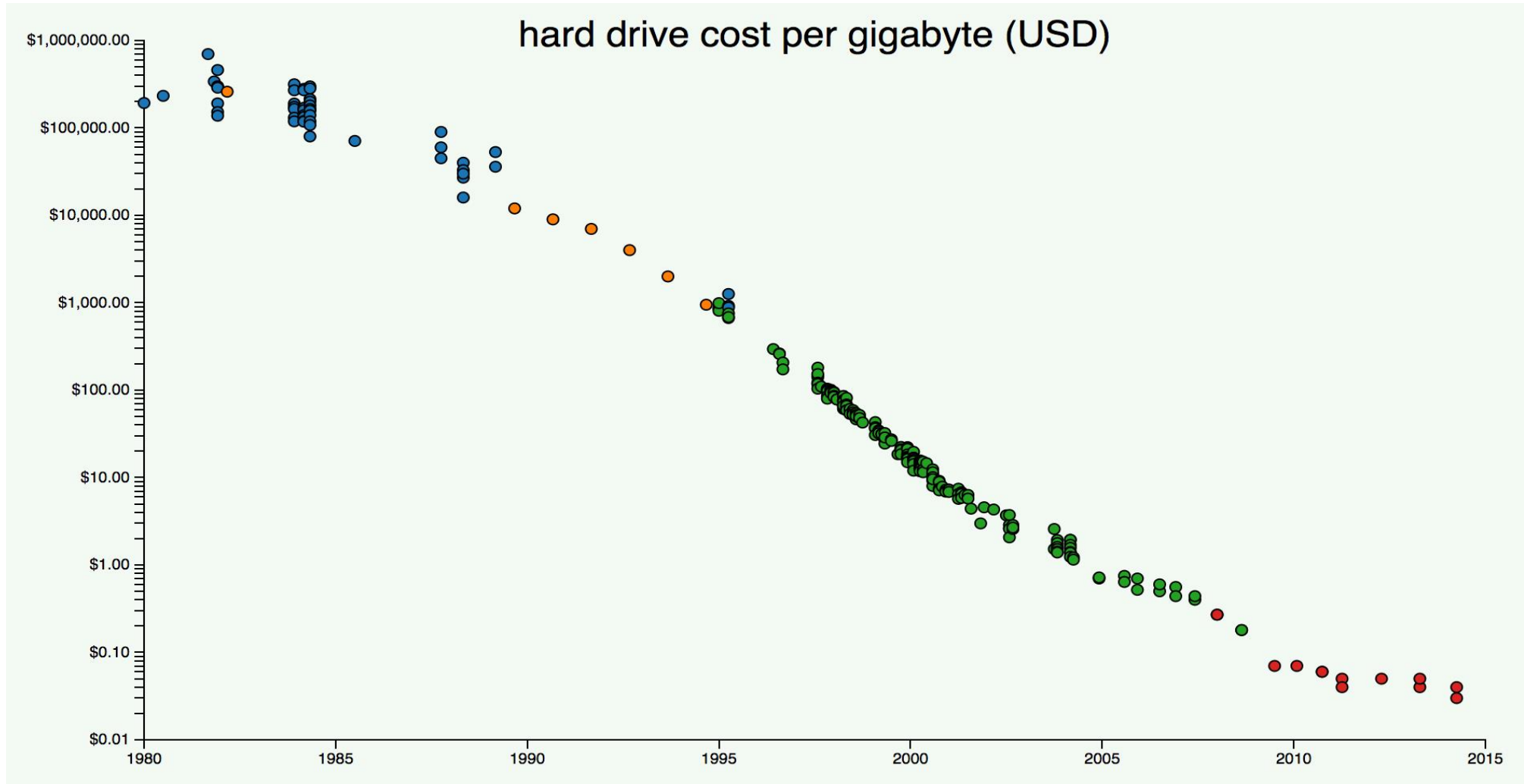
Using 100,000 sensors and data points on wells, pipelines, plants and terminals, it directs every drop of oil and cubic foot of gas that comes out of the kingdom

One study predicts that by 2020, 1.7 MB of data will be created every second for every person on earth.

The average number of AI projects for a business is expected to increase to 35 by 2022 from four this year, according to a Gartner Inc. survey of about 100 organizations of various sizes, many of them with annual revenue of \$1 billion to \$3 billion. The research and advisory firm also said the number of its clients requesting help in dealing with AI suppliers grew 57% between 2017 and 2018.

As per the report by NASSCOM and Blueocean, India is reigning big data analytics with a value of \$1.2 billion placing it among the top 10 big data analytics markets in the world. They have also anticipated the growth becoming eight-fold by 2025, soaring to \$16 billion. With this vision in mind, every sector is now looking forward to Data analytics for its evolution.

Storage capacity, size & cost



Data Generation





Numerical Python (Numpy)

The library made for scientific and mathematical computations

What is Numpy?



- Numerical Python, popularly known as Numpy has been designed to carry out mathematical computations at a faster and easier rate.
- Further this library enriches the programming language Python by providing powerful data structures like multi dimensional arrays beyond matrices and linear arrays.
- Besides that, Numpy provides a large library of high level mathematical functions to operate on these structures.

How to install Numpy

- In command line

`pip install numpy`

- Anaconda distribution

`conda install numpy`

Python Objects vs Numpy

python objects

1. high-level number objects: integers, floating point
2. containers: lists (costless insertion and append), dictionaries (fast lookup)

Numpy provides

1. extension package to Python for multi-dimensional arrays
2. closer to hardware (efficiency)
3. designed for scientific computation (convenience)
4. Also known as array oriented computing

Why Numpy when we have “Lists” ?

Python has inbuilt data structure “List” which is also technically an array which allows different data types.

The answer to this question comes in following three aspects

1. **Size** – Numpy data structures take less space
2. **Performance** – They are inherently faster than lists.
3. **Functionality** – Scipy and Numpy have optimized functions.
4. **Vectorization** of the operations

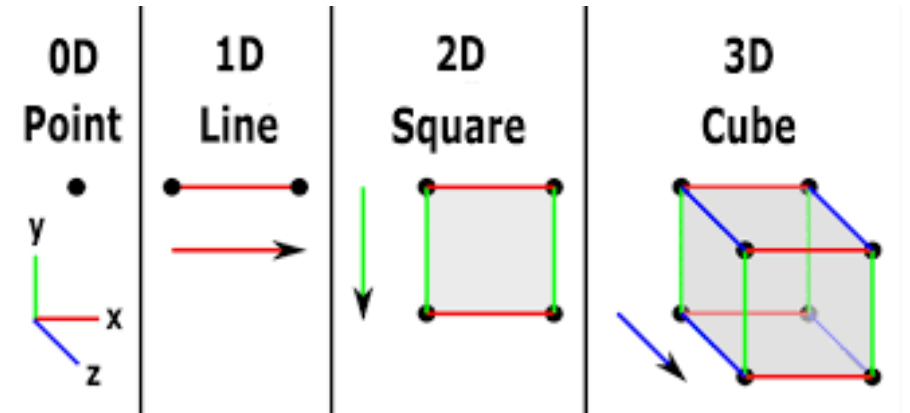
Now let's create some numpy arrays and play around with them!!

Nd-array object

- Nddarray is multidimensional object which can contain only single data type objects.
- It can be a string type or numeric or integer data type.
- If we mixture of strings and numbers are used, all are converted to strings.

Attributes of ND array object

- Dimension – It tells us the number of dimensions of the nd array object. Number of dimensions can range from 1 to 100s and 1000s
- Shape – It gives the shape of the nd array object. That is the length of each dimension.



Attributes of ND array object

- Size - Total number of elements in numpy array
- Dtype – It tells about the type of data being stored in the object.
- Strides – How many steps to be taken to move to next row!!

Some miscellaneous numpy arrays

- `Np.zeros()`
- `Np.arange()`
- `Np.linspace()`
- `Np.full()`