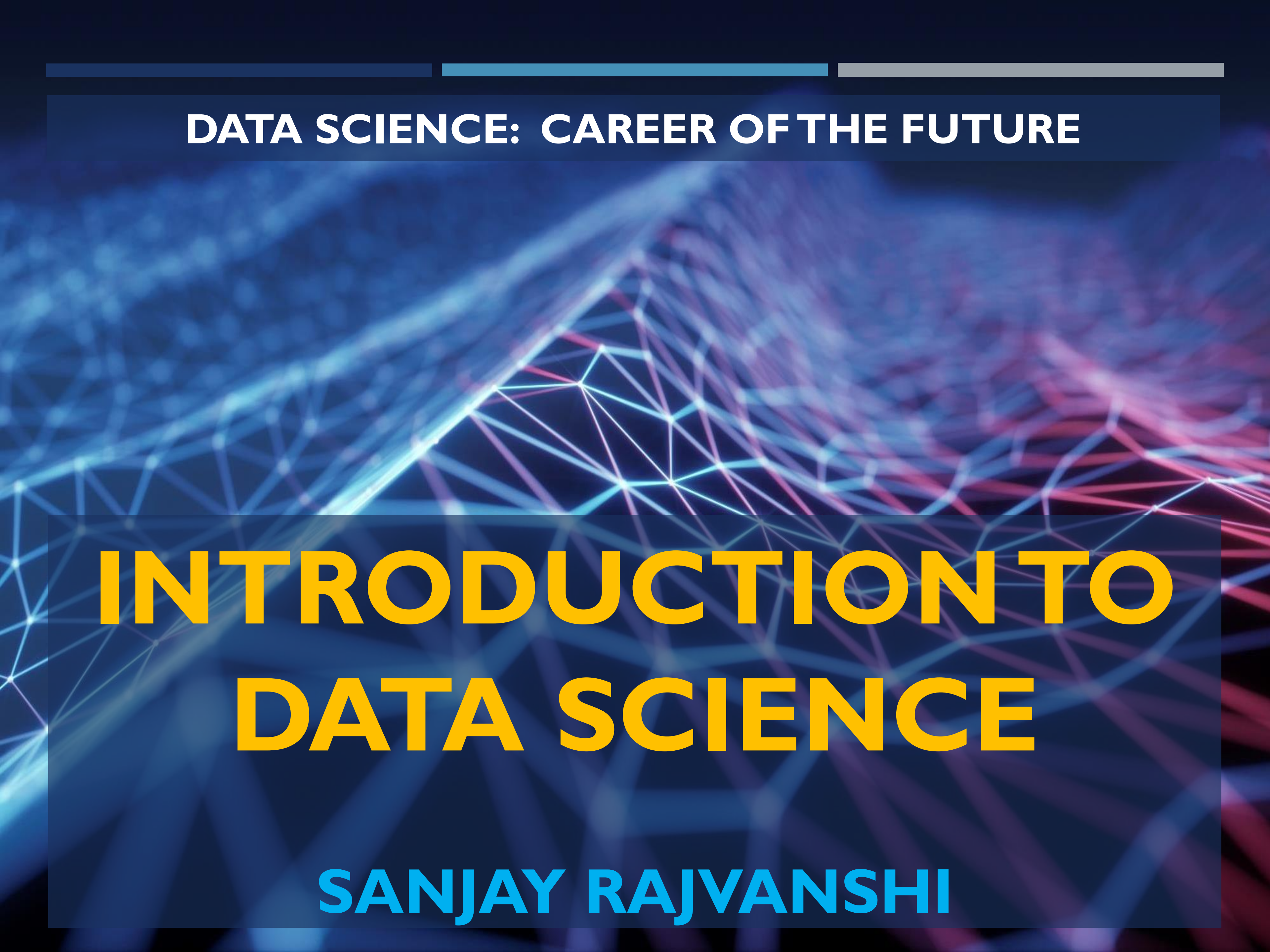
DATA SCIENCE: CAREER OF THE FUTURE



INTRODUCTION TO DATA SCIENCE

SANJAY RAJVANSHI

SCHEDULE

Session	Date	Time	Topic
1	Sep 25	7:00 pm – 8:00 pm	Introduction to data science and associated tools.
2	Oct 2	7:00 pm – 8:00 pm	Introduction to Python. Learn how to use Python for data analysis. Python is simple, yet powerful language that is often used in data science.
3	Oct 9	7:00 pm – 8:00 pm	Data wrangling with Python. Learn how to gather data and make it useful for analysis.
4	Oct 16	7:00 pm – 8:00 pm	Data visualization and analysis with Python. Learn how to create useful visualizations to aid in the analysis of the data.
5	Oct 23	7:00 pm – 8:00 pm	Brief introduction to artificial intelligence and machine learning. Get a peek into how to make data based predictions.

Note: All classes are on Wednesdays.

SESSION 2 – RECAP

- Python Basics
- Data Types
- Control Flow Statements
- Packages/Libraries
- Introduced Plots
- Exercises – all solutions in *Intro to Data Science-S2-Solutions-Final.pdf* (also sent via email)

SESSION 2 – RECAP

- Is "like this' a valid string?
- What is the output of the following?

```
count = 5 // 2
if count > 2:
     print ("> 2")
else:
     print ("<= 2")
```

- What do the following do?
  - !=
  - +=
  - ==
  - **

- What is the output of the "print"?

```
weightList = [24, 22, 30]
i = 0
totalWeight= 0
while i < 3:
   totalWeight += weightList [i]
print (totalWeight)
```

- Identify correct vs incorrect:
  - `if = 5`
  - `for i < 10:`
  - `x == 5`
  - `if (y = 5):`
    `print ("ok")`

SESSION 2 – BUBBLE SORT

- Python code:

```
marksList = [30, 50, 11, 7, 57, 88, 75, 89, 69,
29]

lenML = len (marksList)

lenSort = lenML - 1

sortedFlag = False

while (not sortedFlag):

        lenSort = lenML - 1

        sortedFlag = True

for i in range (lenSort):

        if (marksList [i] >

              marksList [i + 1]):

              temp = marksList [i]

              marksList [i] =

                          marksList [i+1]

              marksList [i + 1] = temp

              sortedFlag = False

print (marksList)
```

- Reference:
  - https://en.wikipedia.org/wiki/Bubble_sort

SESSION 3: DATA WRANGLING WITH PYTHON

SESSION 3 – AGENDA

- Data wrangling with Python. Learn how to gather data and make it useful for analysis.
- Learn how to use Python for data analysis. We will start to learn how to make the data suitable for the problem, clean/convert/transform it – sometimes referred to as data wrangling or data munging.
- Specifically we will focus on DataFrames, large amount of data, and how to analyze that.

SESSION 3 – PRE-WORK

- Explore large data sets and pick one per your interest:
  - Montgomery County, MD data sets – https://data.montgomerycountymd.gov/
  - US Govt. open data sets – https://www.data.gov/
  - Non Govt. website with lots of data sets – https://www.kaggle.com/
  - **Pay attention to the licensing terms before downloading**
  - You may contact the library or the instructor for any help in identifying data set(s) you might be looking for or for any other questions related to the data set(s).

SESSION 3 – PRE-WORK

- Familiarize with pandas library (https://pandas/pydata.org)
- It provides two primary data structures:
  - Series (1-dimensional)
  - DataFrame (2-dimensional)
- Review and try examples/code from the following:
  - Intro to data structures (https://pandas.pydata.org/pandas-docs/stable/getting_started/dsintro.html)
  - 10 minutes to pandas (https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html)
  - Try Cookbook on pandas website (https://pandas.pydata.org/pandas-docs/stable/user_guide/cookbook.html#cookbook)

DATAFRAME

- ***DataFrame** is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects.* [3]
- Created in many different ways. We focus on creating a DataFrame from a csv file.
- Operations (selected for this session from a large set possible with DataFrames)
  - Viewing, <TAB> completion
  - Column labels, counts, data types, size,
  - Selection, Addition, Deletion
  - Arithmetic and logical operations at cell, row, column, DataFrame levels
  - Statistical, Transpose, Sorting, Boolean Indexing, Setting, Missing values
  - Append, Grouping, Selective assignment
- Refer to websites on previous slide

DATA SCIENCE SOLUTION LIFECYCLE

- Data Science solution lifecycle (iterative):
  - **Problem identification**
  - **Identify data**
  - **Clean, transform data**
  - Analyze, visualize
  - Identify algorithm(s)
  - Implement
  - Maintain and support

PROBLEM IDENTIFICATION

- For this introductory class, we will work on a simple problem.
- Of course, a problem becomes even simpler if the data is readily available.
- Problem: Find out the number of students attending Montgomery College by campuses.

IDENTIFY DATA

- Montgomery College enrollment data is published by Montgomery County, MD and made available via its Open Data Portal website.
- Download *Montgomery College Enrollment Data* from https://data.montgomerycountymd.gov/Education/Montgomery-College-Enrollment-Data/wmr2-6hn6

DATA WRANGLING

- Making data suitable for analysis
  - Cleaning data
    - Missing values
  - Transforming data
    - String to numbers or vice versa
    - Conversion of coded values
  - Handling outliers
    - Values that are exceptionally out of place
  - Normalize data
    - Technique to adjust the spread of data
- Do pretty much any type of data management that increases the data suitability for the analysis

EXERCISE

- Create a Python file with name "S3-Exx"
- We will cover all the topics in previous slides in this exercise working directly in the Jupyter notebook

SESSION 3 – HOME WORK

- Intro to data structures (https://pandas.pydata.org/pandas-docs/stable/getting_started/dsintro.html)
  - Series, DataFrame arithmetic operations
- 10 minutes to pandas (https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html)
  - Selection (inc. by position), Boolean indexing, Setting values, Missing values, Merging, Grouping
- Try Cookbook on pandas website (https://pandas.pydata.org/pandas-docs/stable/user_guide/cookbook.html#cookbook)
  - if-then, Splitting, Building criteria, Selection, Slicing, Sorting, Grouping, Creating example data

SESSION 4 – AGENDA

- Data visualization and analysis with Python.
- Create useful visualizations to aid in the analysis of the data.
- Create and customize various types of graphs
- Learn some statistical techniques

REFERENCES

Note: you are not required to sign-up for an account on any of the sites to read these articles.

1. *Official website for Python and tutorials –*
   a. https://www.python.org/
   b. https://docs.python.org/3/tutorial
2. *Another good Python reference and tutorials –*
   a. https://www.w3schools.com/python/
   b. https://www.w3schools.com/python/default.asp
3. *pandas (Open source library providing data structure and data analysis tools) –*
   a. https://pandas.pydata.org/
4. *numpy (Fundamental package for scientific computing with Python) –*
   a. https://numpy.org/