DATA SCIENCE INTERNSHIP

Session-1

HDLC TECHNOLOGIES

Agenda

- Introduction
- Python basics
- Jupyter notebook, Google colab
- Pandas
- Numpy
- Matplotlib
- Scikit learn
- Mini Project
- Main Project
- Final Exam

INTRODUCTION

- Data Science is a combination of multiple disciplines that uses statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it.

What is Data Science?

- Data Science is about data gathering, analysis and decision-making.
- Data Science is about finding patterns in data, through analysis, and make future predictions.
- By using Data Science, companies are able to make:
- Better decisions (should we choose A or B)
- Predictive analysis (what will happen next?)
- Pattern discoveries (find pattern, or maybe hidden information in the data)

Where is Data Science Needed?

- Data Science is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.
- Examples of where Data Science is needed:
- For route planning: To discover the best routes to ship
- To foresee delays for flight/ship/train etc. (through predictive analysis)
- To create promotional offers
- To find the best suited time to deliver goods
- To forecast the next years revenue for a company
- To analyze health benefit of training
- To predict who will win elections

How Does a Data Scientist Work?

- A Data Scientist requires expertise in several backgrounds:
- Machine Learning
- Statistics
- Programming (Python or R)
- Mathematics
- Databases
- A Data Scientist must find patterns within the data. Before he/she can find the patterns, he/she must organize the data in a standard format.

Here is how a Data Scientist works:

1. **Ask the right questions** - To understand the business problem.
2. **Explore and collect data** - From database, web logs, customer feedback, etc.
3. **Extract the data** - Transform the data to a standardized format.
4. **Clean the data** - Remove erroneous values from the data.
5. **Find and replace missing values** - Check for missing values and replace them with a suitable value (e.g. an average value).
6. **Normalize data** - Scale the values in a practical range (e.g. 140 cm is smaller than 1,8 m. However, the number 140 is larger than 1,8. - so scaling is important).
7. **Analyze data, find patterns and make future predictions**.
8. **Represent the result** - Present the result with useful insights in a way the "company" can understand.

•

What is Data?

- Data is a collection of information.
- One purpose of Data Science is to structure data, making it interpretable and easy to work with.
- Data can be categorized into two groups:
- Structured data
- Unstructured data

Unstructured Data

- Unstructured data is not organized. We must organize the data for analysis purposes.

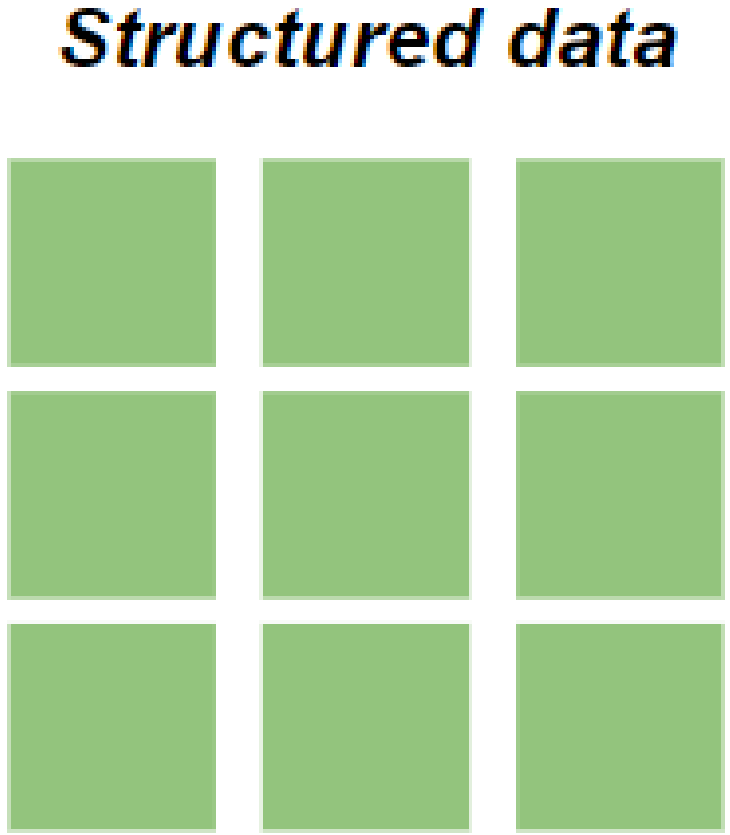
Health Status 09082020 Inbox ×

Date 09082020
Average pulse 70,
Max pulse 80,
Steps 10500

Unstructured data

Example of Unstructured data

Structured Data



How to Structure Data?

- We can use an array or a database table to structure or present data.
- Example of an array:
- `[80, 85, 90, 95, 100, 105, 110, 115, 120, 125]`
- The following example shows how to create an array in Python:
- ```
  Array =[80, 85, 90, 95, 100, 105, 110, 115, 120, 125]
  print(Array)
  ```


  ```

Data Science & Python

- Python is a programming language widely used by Data Scientists.
- Python has in-built mathematical libraries and functions, making it easier to calculate mathematical problems and to perform data analysis.

Python Libraries

- Python has libraries with large collections of mathematical functions and analytical tools.
- In this course, we will use the following libraries:
- **Pandas** - This library is used for structured data operations, like import CSV files, create dataframes, and data preparation
- **Numpy** - This is a mathematical library. Has a powerful N-dimensional array object, linear algebra, Fourier transform, etc.
- **Matplotlib** - This library is used for visualization of data.
- **SciPy** - This library has linear algebra modules
- We will use these libraries throughout the course to create examples.