Fundamentals

What is Data Science?

Data Science is an inter-disciplinary academic field that uses:

- Statistics
- Scientific Computing
- Processes
- Algorithms, and
- Systems

to extract and extrapolate knowledge from noisy structured and unstructured data.

Data Science vs Data Analysis

John Turkey, in **1962**, coined a term called “Data Analytics”, which is similar to modern day data science.

But there are major differences between Data Science and Data Analytics:

Data Science	Data Analysis
Works with large, complex datasets	Works with small structured datasets
Used to work with unstructured texts and images using machine learning algorithms to build predictive models	Answers specific questions and constraint to only a specific kind of problem
Involves tasks such as statistical analysis, data preprocessing, feature engineering and model selection	Involves tasks such as Data cleaning, Data visualising, Exploratory Data Anaysis

Data

Data is a collection of discrete or continuous values which can be processed to extract meaningful information.

For example, A file containing the name, age, section, attendance and remarks of students in a class is data which conveys information regarding students. It

can be further processed to obtain average marks, who are low on marks or attendance etc.

Problems related to Data

There are two main problems related to data. One is **longevity** and the other is **accessibility**.

Longevity means *long term* or *long life*.

Accessibility refers to *how much of somthing can be used*.

Data Longevity

Data storage to store data for a long period time has been always a problem. There has been no way to do so for centuries and even eternities.

Data Accessibility

A lot of the scientific data that is used for processing is never published in data repositaries such as databases. In a survey of 516 studies published between 2 and 22 years, only 20% of the studies were willing to provide data.

FAIR

In order to improve reproducability (another researcher working on the same data under the same conditions should get the same results), the FAIR principles are promoted. FAIR stands for:

- Findable
- Accessible
- Interoperable
- Reusable

Data Classification

Data can be classified as **structured**, **unstructured** and **semi-structured**. Here is a comparison between the three:

Structured	Unstructured	Semi Strucutred
Data is arranged and neatly organized, and are addressable for effective analysis	Data which do not follow any organized format or pre-defined structure	Data which does not fit rows and columns, but are not completely unorganizes as well
Table, Database	Text document, Image	JSON file, XML file, HTML file

The main goal of data classification is to arrange data in such a form that it becomes fairly available to the user. It focuses on three iport factors:

- **Homogeneity**: Data present in a particular group should be similar to each other.
- **Clarity**: Data should clearly define its position in a particular group.
- **Stability**: Any sort of investigation on data must not affect the same set of classification.