Study Unit 4

Introduction to Data and Data Science

Introduction to Data and Data Science Outline

- Definition, Types and Sources of data
- Who is a data scientist?
- How to become a data scientist?
- Programing languages for data science
- Lifecycle of data science
- Data science use cases

Study Unit Duration

This Study Session requires a minimum of 3 hours' formal study time.

You may spend an additional 2-3 hours on revision.

Preamble

Data is present everywhere and is collected every day. We make calls and send messages on the phone every minute. We tweet and retweet messages on Twitter, post pictures and videos on Instagram, countries like Kenya, Somalia, South Sudan, Uganda, and Ethiopia count their citizens and foreigners at a well-defined point of time. Hospitals take clinical records of patients, and teachers count the number of students present in school every day. With these available huge amounts of data, organizations focus more and more on using the insights from data to evaluate progress, build solutions and make an informed decision. The need to extract useful insight is a must for a business in today's world.

Learning Outcomes of Study Unit 1

Upon completion of this study unit, you should be able to:

1.1 Describe the various forms of data

1.2 Distinguish between primary and secondary sources of data

1.3 Describe Data Science and its applications in solving real world problems

Terminologies, Acronyms and their Meaning

AI	Artificial Intelligence
ML	Machine Learning
RL	Reinforcement Learning
DL	Deep learning
EDA	Exploratory Data Analysis
np	Numpy
plt	Matplotlib

NaN	Not a Number
NULL	Missing value
Viz	Visualization
TF	TensorFlow
Os	Operating system
Pd	Pandas
Sns	Seaborn

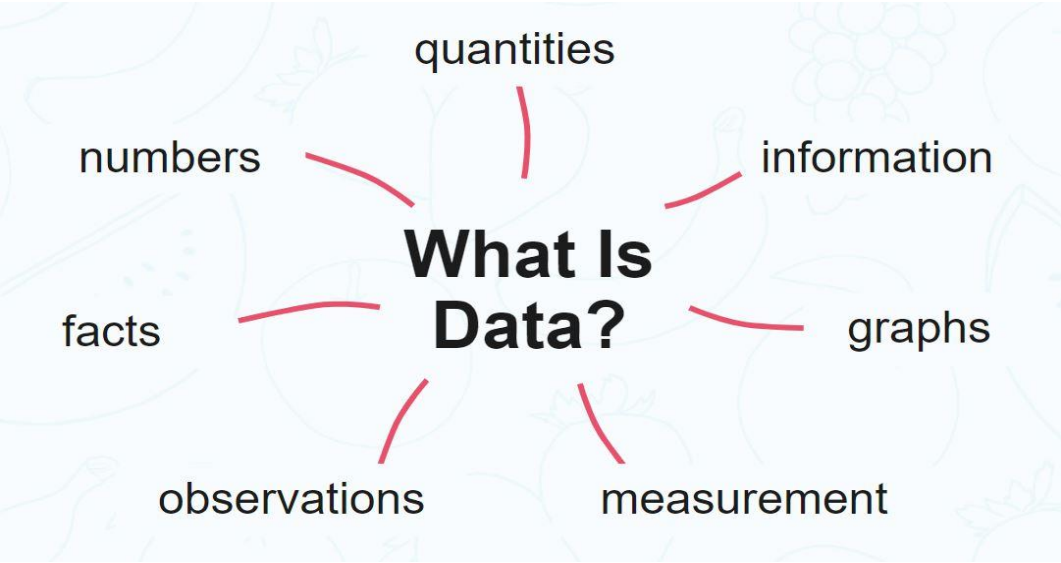
Prerequisites

A basic understanding of Python programming in CS14 (Programming in Python) is required.

4.1 Overview of Data

4.1.1 Definition of Data

Data refers to unorganized and unprocessed facts, which do not hold complete meaning unless processed to drive meaningful insights. In other words, data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things. That is, data can be words (texts), sounds, images, or numbers written on papers, stored on a computer, and in fact, it could be a fact that is stored inside your mind right now.



Source: https://www.twinkl.ae/teaching-wiki/data

Data is fundamentally inert and has no real meaning or value until we analyze it. Data is a raw material for information, and the result of data processing is called information.

4.1.2 Types of data

Data can be qualitative or quantitative.

Qualitative data

This refers to data that can be observed and recorded. It is non-numerical in nature. This type of data is collected through methods of observations, survey, opinion of people on a particular topic, and similar methods. Examples of qualitative data include gender (male or female), opinion (agree,

neutral, disagree), blood type (A, B, AB, and O), country of origin (Ethiopia, Somalia, South Sudan, Sudan, Uganda, Kenya, etc.), and so on. Qualitative data is often known as categorical data.

Quantitative data

Quantitative data is the type of data that can be measured in the form of numbers or counts. For example, distance from Nairobi to Kigali, number of hours to complete introduction to data science 1, length of a table, revenue realised by Somalian government, speed covered by a car, age of a student, weight of a goat, etc.

Quantitative data has quantifiable information that can be used for mathematical computations and statistical analysis which informs real-life decisions. For example, a manufacturing company in Uganda will need an answer to the question, “How much is the production cost?”. Quantitative data can be used to answer questions such as “How many?”, “How often?”, “How much?”.

Quantitative data can be divided into two types, namely; discrete and continuous data.

1. Discrete data

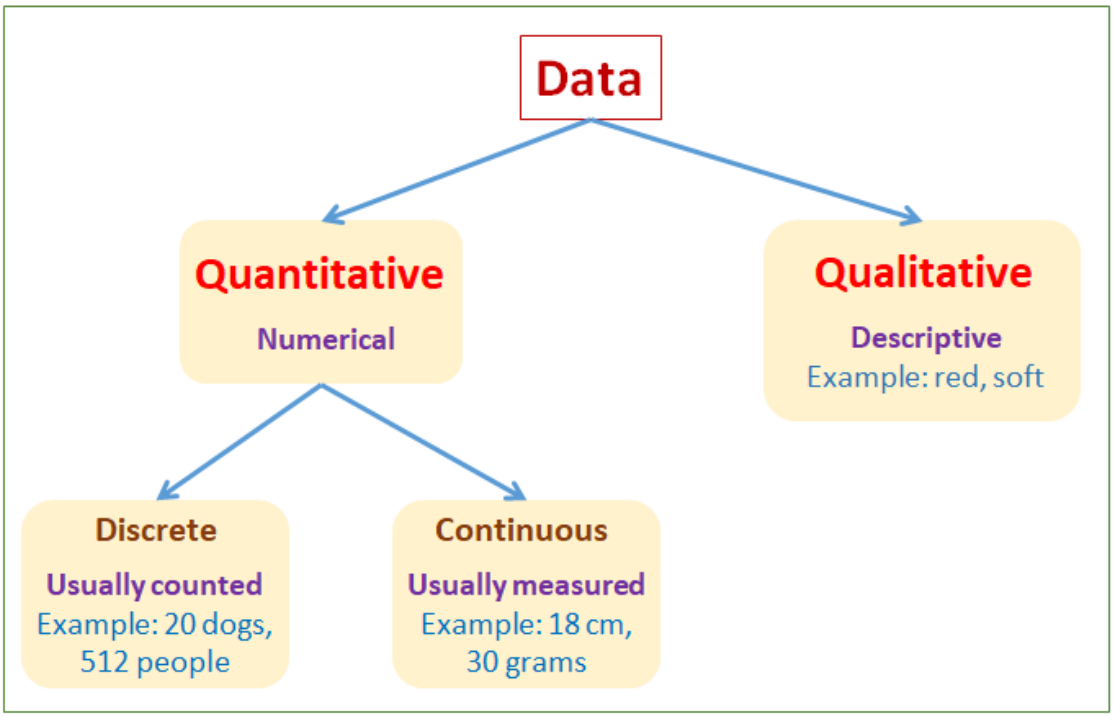
Discrete data is a type of data that consists of counting numbers only. That is, it can be counted and has a finite number of possible values. Examples of discrete data includes the number of students taking introduction to data science 1, the number of days in a year, number of females in South Sudan, etc. You can see that these data take on only certain numerical values. Also, if you count the number of phone calls you receive for each day of the week, you might get values such as zero (no call), one, two, or five.

When trying to identify discrete data, we ask the following questions; Can it be counted? Can it be divided into smaller parts?

2. Continuous data

Continuous data is a type of data that arise as a result of measurement. It has an infinite number of possible values within a given range. Example of continuous data includes height, weight, temperature and length.

Quick Summary



4.2 Sources of Data

The following are the two sources of data:

1. Primary data
2. Secondary data

4.2.1 Primary Data

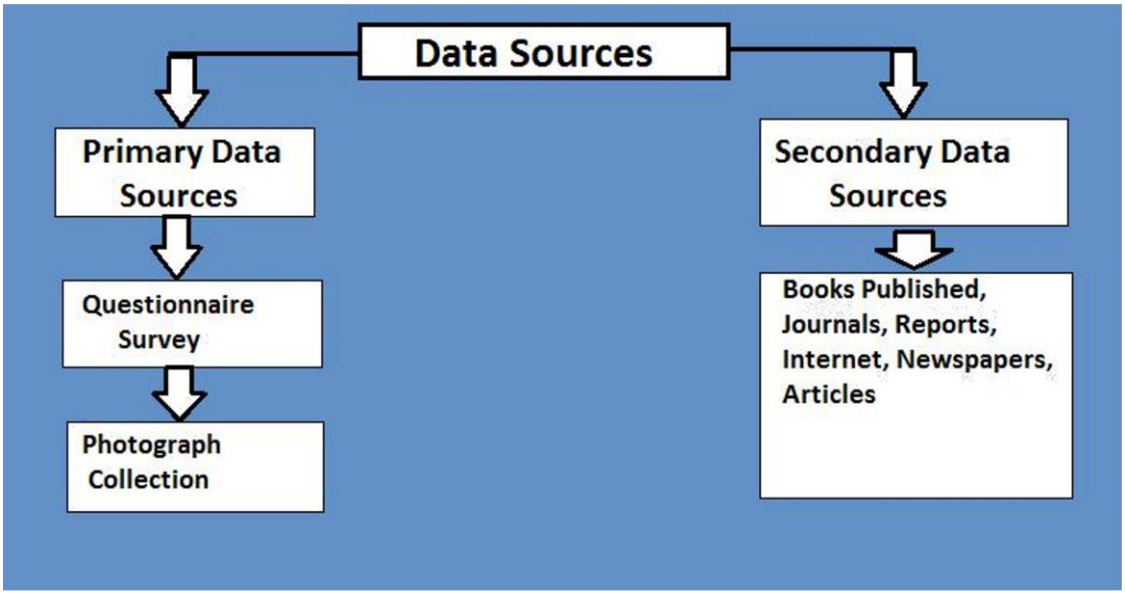
These are first-hand information collected by an investigator. The data collected are pure and original and collected for a specific purpose. This type of data has never undergone any data preprocessing before. For example, population census conducted by the government of Kenya after every 10 years.

4.2.2 Secondary Data

Secondary data refers to second-hand information. They are data acquired from optional sources like magazines, books, documents, journals, reports, the web and more. That is, they are not

originally collected rather obtained from already published or unpublished sources. Secondary data are impure in the sense that they have undergone data preprocessing at least once.

Quick Summary



Additional resources

For more resources in this section please consider the following:

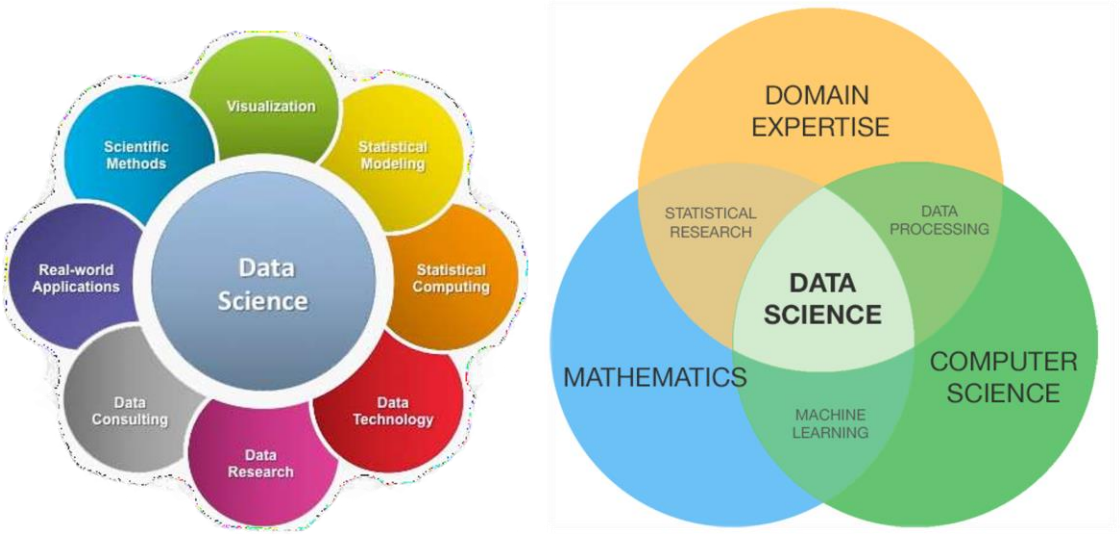
- https://www.mathsisfun.com/data/data.html
- https://www.twinkl.ae/teaching-wiki/data
- https://www.questionpro.com/blog/quantitative-data
- https://courses.lumenlearning.com/odessa-introstats1-1/chapter/sampling-and-data
- https://www.mymarketresearchmethods.com/data-types-in-statistics
- https://studiousguy.com/sources-of-data-collection
- https://byjus.com/commerce/what-are-the-sources-of-data
- http://bit.ly/30-data-science-terms

4.3 Introduction to Data Science

In our previous section, you learnt that data is fundamentally inert, and has no real meaning or value until we give it. In this section you will learn how to use data science to give powerful meaning to your data.

4.3.1 What is Data Science?

Data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data. The field of data science enables us to turn raw data into understanding, insight, and knowledge.



4.3.2 Who is a Data Scientist?

A data scientist is a professional who deals with a massive explosion of data and uses his/her skills in mathematics, statistics, computing, business domain and the scientific method to give data a shape so that it can better express itself. Data scientist makes sense out of this tsunami of information, identify hidden patterns and draw conclusions and insights. In other words, data scientist focuses on analysing the past and current data, predicting the outcomes with the sole aim of making information.

Harvard Business Review (HBD) in 2012 named data scientist the sexiest job of the 21st century. Data science helps us build a strong foundation for the data-driven world, access the power of

artificial intelligent (AI) related technology and developing an operational model to derive business insights from raw data to support decision making.

4.3.3 Data Scientists' Skills

Data scientists must acquire skills like data cleaning, data analysis, and data visualization to be able to effectively communicate information or findings to inform high-level decisions in an organization. As such, it incorporates skills from computer science, mathematics, statistics, communication and business.

4.3.4 How to Become a Data Scientist?

The following are general steps to becoming a data scientist:

- Practice every class activity in this course
- Be active in the group discussion
- Question everything about data
- Ask questions when you don't understand a concept
- Visit other learning resources provided in this course
- Learn from your course mate (other students doing this course with you)
- Know how to programme (Introduction to programming language and this course will help you)
- Build projects using the skills learn in this course. Check here for some of my projects.

Additional resources

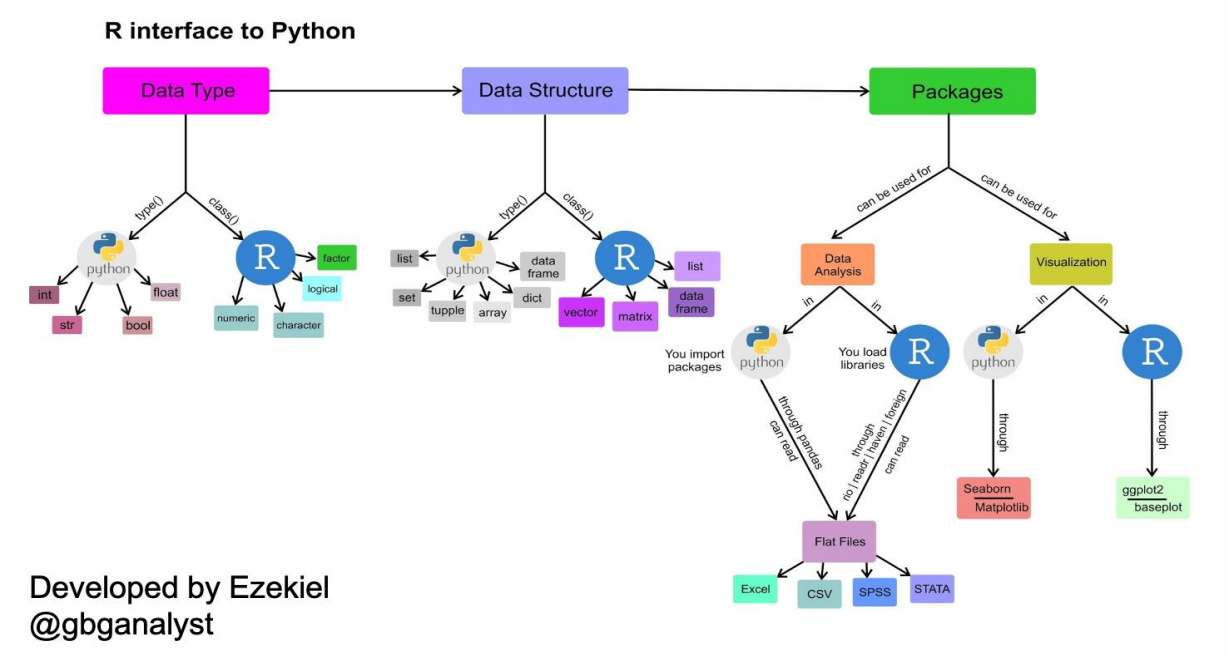
For more additional steps, check out details via

https://www.dataquest.io/blog/how-to-become-a-data-scientist/

4.3.5 Programing Languages for Data Science

A programming language is a formal language comprising a set of instructions that produce various kinds of output. There are several programming languages for data science and you as a data scientist should learn and master at least one language for data science project.

Python is one of the most widely used data science programming language in the world today. It is an open-source, easy-to-use language that has been around since the year 1991. This course will teach you how to use Python for data cleaning, analysis and visualization.



Additional resources

For more information about R, Scala, Julia, and other programming languages please visit this source.

4.3.6 Data Science with Python - Libraries

Packages are a collection of related modules that aim to achieve a common goal. Finally, the Python standard library is a collection of packages and modules that can be used to access built-in functionality. In an ideal world, you would import any necessary modules into your Python scripts without any issues.

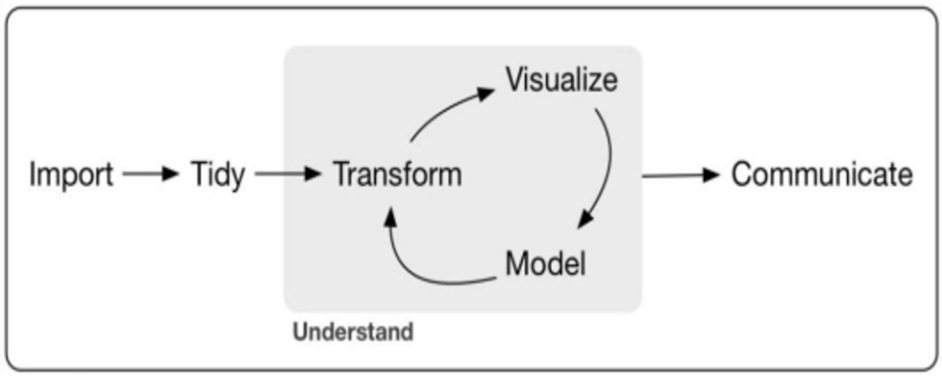
Top 6 most important Python libraries and packages for data science are:

1. Numpy
2. Pandas
3. Matplotlib
4. Scikit-Learn
5. TensorFlow
6. Keras

In this course, you will learn how to use Numpy, Pandas, Matplotlib and or Seaborn for various data science activities.

4.3.7 Lifecycle of Data science.

The general lifecycle of data science involves:



Sources: Hadley Wickham, R for data science

Import: The first step in data science is to import data into Python, R, Spreadsheet (MS Excel), or any other data processing software. Without data, there is no data science.

Tidy: Another step is data cleaning or munging, and it is one of the most time-consuming tasks of a data scientist. Since much of the datasets available are not cleaned, it is, therefore, a data scientist's work to prepare the datasets in an easy to use format.

Principle of tidy data states that:

1. Every column in the data is a variable
2. Every row in the data is an observation
3. Every cell is a single value relating to a specific variable and observation

Tidying data is an aspect of data munging or wrangling.

Transform: The other step of data munging is transforming or creating a new variable from the existing variables

Visualize: Data visualization refers to the graphical representation of data by using visual elements such as charts, Infographics, and maps in understanding the data.

Models: Models are used to answer questions from the data. A good data scientist will implement various machine learning algorithms which require good coding and interpretation skills. Data science uses a branch of Artificial Intelligence (AI) called Machine Learning (ML) to analyze data and predict the future.

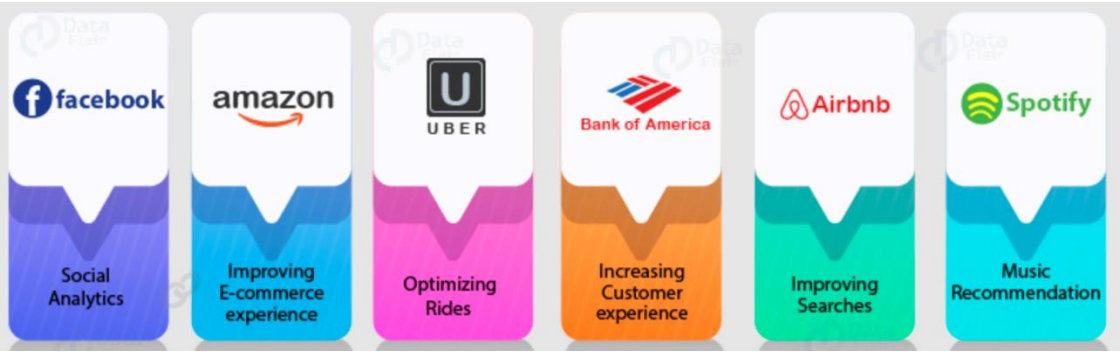
Communicate: The last step of data science is communication, an absolutely critical part of any data analysis project. A data scientist will need to report findings to the stakeholders, and tools like Jupyter notebook, Rmarkdown, or PowerPoint or MS-Word make this communication easier.

4.3.8 Data Science Use Cases

Data science helps us achieve some major goals that either were not possible or required a great deal more time and energy just a few years ago. The following are the most relevant and efficient data science use cases:

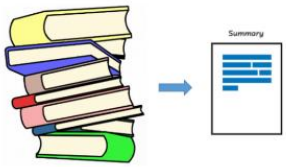
- Forecasting (sales, revenue and customer retention)
- Pattern detection (weather patterns, financial market patterns, etc.)
- Recommendations (movies, restaurants and books you may like)
- Anomaly detection (fraud, disease, crime, etc.)
- Automation and decision-making (background checks, credit worthiness, etc.)
- Classifications (email spam, diseases, phishing website)
- Recognition (facial, voice, text, etc.)

Other use cases by some organizations:



Source: https://data-flair.training/blogs/data-science-use-cases

Summary of Study Unit 4



In this study unit, you have learnt that:

1. Data can come inform of texts, sounds, images, or numbers written on papers or stored on a computer
2. Data can be qualitative or quantitative. It is qualitative if it is non-numerical in nature and quantitative if it can be measured in form of numbers or counts.
3. The field of data science enables us to turn raw data into understanding, insight, and knowledge
4. The lifecycle of data science comprises import, tidy, transform, visualize, models, and communicate.

Additional resources

For more uses cases, please visit here and here by clicking on the link.

In your Introduction to Python programming, you were introduced to Jupyter notebook. Please consider this as an additional resource for Markdown in Jupyter notebook

https://www.datacamp.com/community/tutorials/markdown-in-jupyter-notebook