Lecture 0-1 Data Analysis: an Overview

Yue Ma, July 2024

Contents

Lecture 0	1.Self Introduction	Let's know each other :)
	2.Discussion: What is Data Analysis?	What is your current understanding? Introduce your relevant experience.
	3.Pipeline and Useful Tools	Learn the basic concepts and tools
Lecture 1	4.Algorithm I: Statistics	Probability Distribution; Hypothesis Test
	5.Algorithm II: Mathematical Modelling	Build Models with Pre-knowledge
	6.Algorithm III: Machine Learning	The Modern Technique

Self Introduction

01

Name

02

Major/Interested Majors

03

Hobbies

04

Any questions?

Discussion

What is your current understanding of Data Analysis?

- What does data analysis do?
- How do we do it?

Do you have any relevant experience?

- What was the purpose?
- How did you do it?

What is Data Analysis?

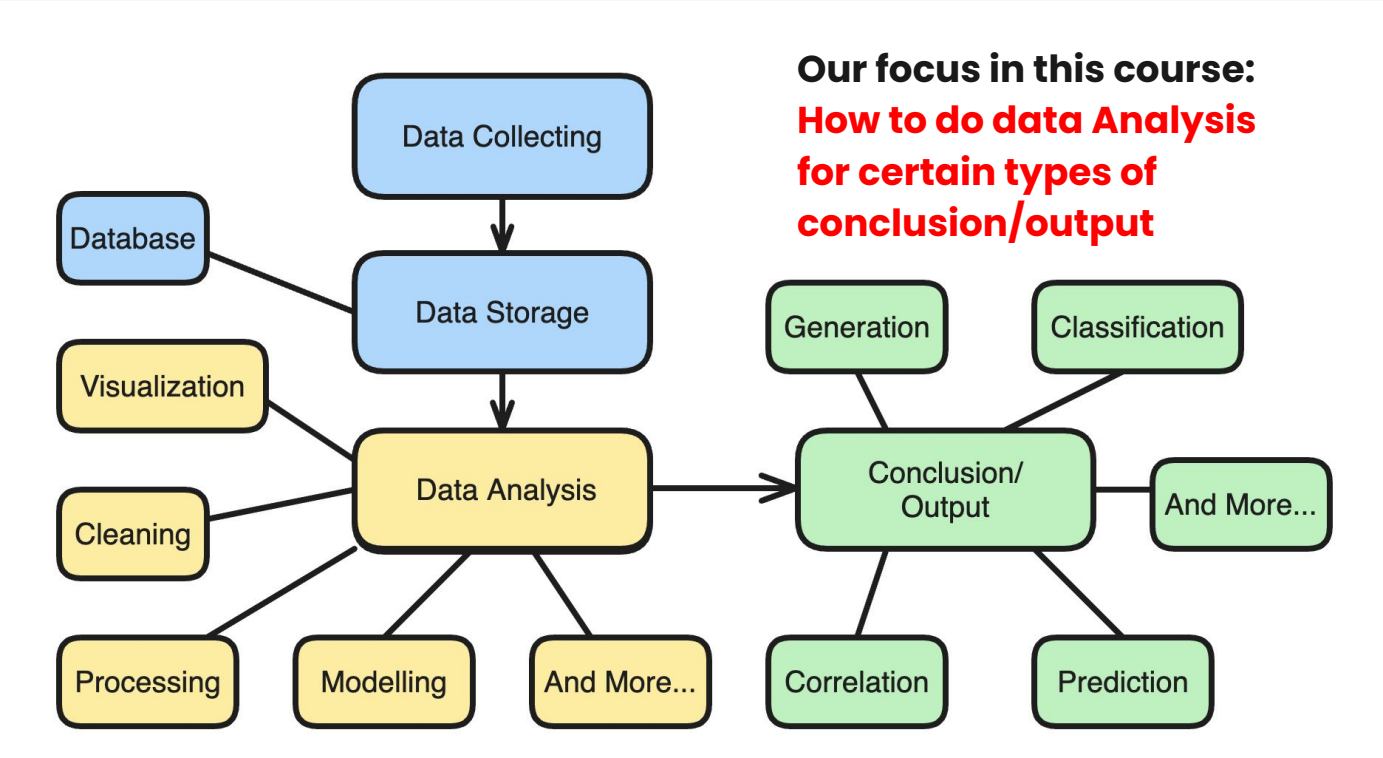
Definition on Wikipedia:

Data analysis is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

Methods

Goals

Pipeline and Useful Tools



Pipeline and Useful Tools

What tools can we use for data analysis?

- **Human brain + Calculators:** Good for modelling, but not realistic for massive analysis :/
- **Excel; SPSS:** Commercial software with pre-defined operations and statistical functions. Easy to use but limited functions.
- **R (Programming Language)**: Also with pre-defined statistical functions, and could be used for very basic machine learnings.
- **Python (Programming Language)**: An increasingly popular and powerful tool. Exceptionally good for machine learning.

Pipeline and Useful Tools

Demo: find the average value of a 1-D dataset with:

- **Human brain + Calculators**
- **Excel**
- **Python (Programming Language)**

Problem Set 1 - P1

What are the definitions of the following quantities for a dataset?
You can use google, ChatGPT, etc. but you need to write the formulas and explanations down in your own words.

1. **Mean/Average**
2. **Median**
3. **Standard Deviation**
4. **Quartile**

Problem Set 1 - P2

What do the following task mean in data analysis? Explain and provide an example for each one.
You can use google, ChatGPT, etc. but you need to write the formulas and explanations down in your own words.

1. **Classification**
2. **Regression**