DSS5201 Data Visualization

Week 1 Course Overview

Yuting Huang

NUS DSDS

2024-08-12

Instructor: Dr. Yuting Huang (yhuang@nus.edu.sg).

- We meet **in person** once a week.
- **Mondays 7-9:30pm at LT9**.

Topics:

- Introduction to Python programming.
- Importing data into Python.
- Data manipulation.
- Principles of data visualization.
- Advanced topics on data visualization.

Evaluation components

	Component	**Group (G) /Individual (I)**
Assessments	5%	I
DataCamp assignments	10%	I
Group project	15%	G
Midterm test	30%	I
Final exam	40%	I

Activate your account to join DataCamp Classroom.

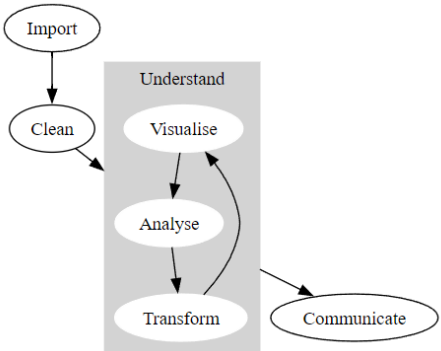
- Via the link on Canvas.
- Sign in using your NUS email account (with domain @u.nus.edu) is required.

Grades are binary. You are graded based on completion by the due date.

- Don't wait till the last minute to submit assignment.
- You will receive a certificate of completion from DataCamp for each course you complete.

What is data science?

Data science is an interdisciplinary field that allows us to turn **raw data** into **understanding**, **insight**, and **knowledge**.



Data science skills empower us to participate in and drive conversations that shape our lives and society as a whole.

- Fundamentally human-centered.
- Facilitates decision making by **quantitatively balancing trade-offs**.

To quantify things reliably, we must

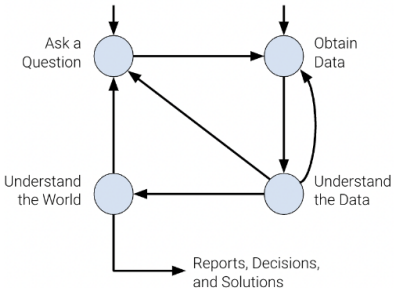
- Find **relevant** data and recognize its limitations.
- Ask the right **questions** and make reasonable assumptions.
- Conduct appropriate **analyses** and explain the insights.
- ... and apply **critical thinking and skepticism** at every step of the way.

This course aims to give you the skills you need about data science and programming.

- Beginner: Fret not! We will help you get started from scratch.
- Intermediate or sophisticated coder: There are materials relevant for you too.

We will give recommendations based on what will serve you best **in the long run**.

The **data science lifecycle** is an iterative process that encompasses the various statistical and computational building blocks of data science.



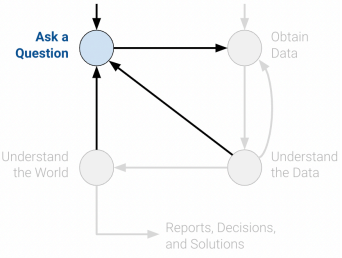
Source: Data100 at UC Berkeley.

Ask a question

Whether by curiosity or necessity, data scientists constantly ask questions.

Here are somethings we should ask ourselves before framing a question:

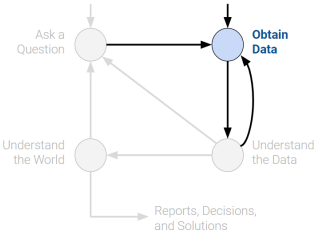
- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are the metrics for success?
  - Establish a clear point to know when to conclude the project.



A careful analysis of any problem requires the use of data.

May be readily available to us, or we may need to collect them ourselves. When doing so, it is important to ask:

- What data do we have? What data do we need?
- How will we sample more data?
  - Web scrapping, collect data manually, run experiments, etc.
- Are our data representative of the population we want to study?

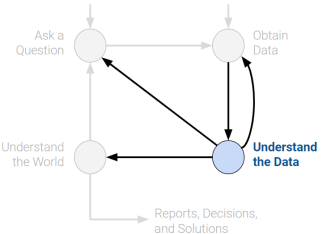


Understand the data

Raw data are not inherently useful. Therefore, translating raw data into actionable insights is a key job of a data scientist.

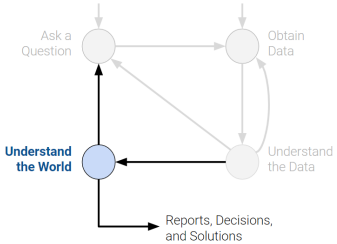
We may choose to ask:

- How are the data organized, and what do they contain?
- Do we have relevant data?
- Are there biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?



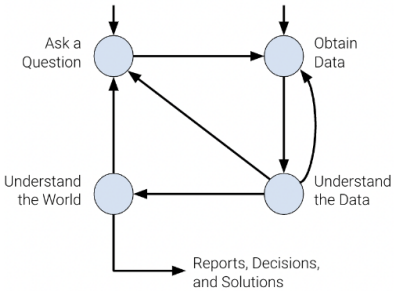
After observing the patterns in our data, we begin to answer our questions.

- What do the data say about the world?
- Do the data answer our questions or solve the problem we had?
- How robust are our conclusions and can we trust the results?



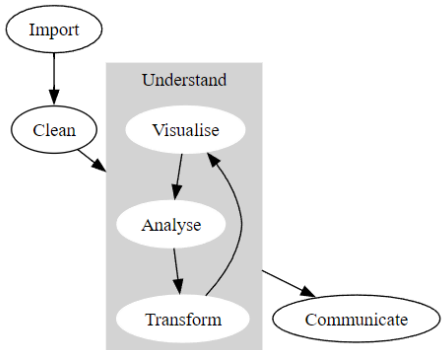
It is a set of general guidelines, rather than a hard set of requirements.

- In our journey this semester, we will cover the techniques used in data exploration.



Source: Data100 at UC Berkeley.

The goal for our class is to help you learn the most important tools that allow you to do data science.



- The workflow is roughly organized according to the order in which you will use them in a data science project.
- Of course, you will iterate through them multiple times.

Why do we visualize?

1. Discover patterns that may not be obvious from numerical summaries:
   - Anscombe's Quartet.

Table 1: Anscombe's Quartet

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50

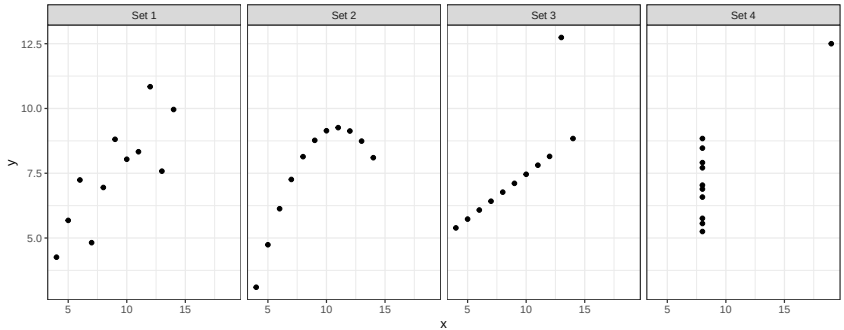
How does the (x_1, y_1) pair differ from the (x_2, y_2) pair?

- If we compare a variety of **descriptive statistics**, the four data sets appear to be identical.

TABLE 2: Descriptive Statistics

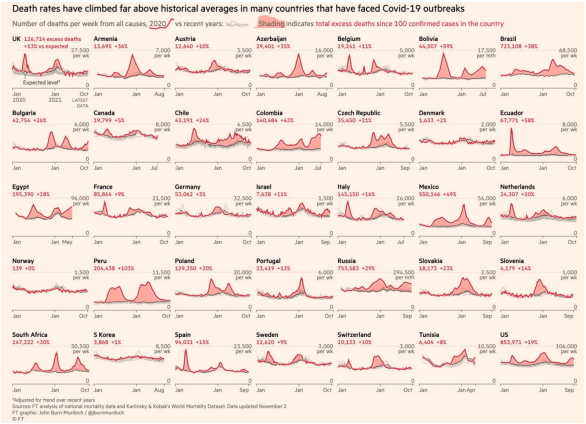
	x1	x2	x3	x4	y1	y2	y3	y4
N	11	11	11	11	11.0	11.0	11.0	11.0
Mean	9	9	9	9	7.5	7.5	7.5	7.5
Median	9	9	9	8	8.0	8.0	7.0	7.0
SD	3	3	3	3	2.0	2.0	2.0	2.0

- It is only through visualizations that differences across these data sets emerge.

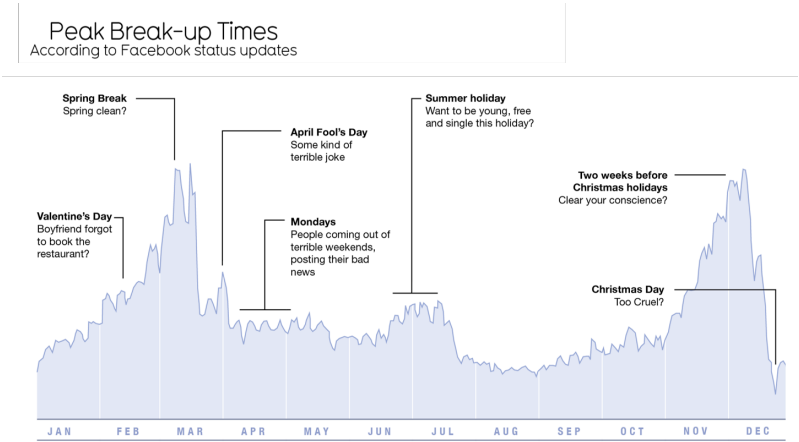


Why do we visualize?

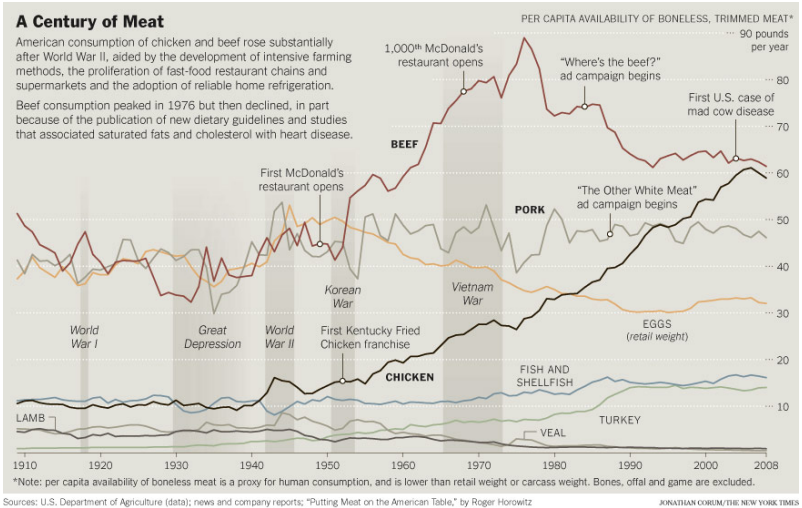
2. Convey information in a way that is otherwise difficult or impossible to convey.



Source: Financial Times, Dec 20, 2021.



Source: Knowledge is Beautiful.



Source: The New York Times, Mar 15, 2011.

By the end of this course, you will be able to...

- Get insights from data.
- **Be resourceful** in collecting data and finding solutions.
- ... **reproducibly and collaboratively**, using modern programming tools and techniques.

Resourcefulness is one of the most important skills to succeed in data science.

- Ability to find and use available resources to solve problems.
- Be proactive and persistent.
- Be adaptable.
  - Always be ready to learn something new.
  - Keep up with the latest tools and best practices.

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

- Near-term goals:
  - Does the code actually do what you think it does?
  - Are the tables and figures reproducible from the code and data?
- Long-term goals:
  - Can the code be used for other (newer) data?
  - Can you extend the code to do other things?

Source: The Turing Way Project.