Exercise Sheet: Introduction to Data and the Data Lifecycle

Data, Data Storage, Data Collection

Exercise 1 – Climate Monitoring

A citizen-science network and weather stations collect data on temperature, rainfall, wind speed, cloud cover, and other environmental factors. Scientists want to use these data to study how local temperatures have changed over the last 20 years and to detect unusual weather events. The end goal is to be able to predict extreme weather events and inform policy decisions.

Data is collected at different resolutions and frequencies, often in multiple formats (CSV files, images, and sensor logs). Data quality may vary because of sensor errors, missing values, or calibration issues, and changes in the measurement equipment over time. Thus, the scientists need to combine heterogeneous sources and want to produce reliable results for policymakers and the public.

(**1**) What is the question that the researchers are trying to answer? What would make the question more precise?
(**2**) Identify several sources of data in this scenario. For each source, indicate who or what produced the data and how it was collected.
(**3**) The scientists receive temperature readings as CSV files, cloud-cover information as satellite images, and equipment maintenance notes as free text from technicians. Classify each of these three formats as structured, semi-structured, or unstructured. What does this classification tell you about how easily they can be combined?
(**4**) What potential issues can you expect regarding quality, trustworthiness of usability or the data?
(**5**) Suppose that two weather stations report temperatures in different units and at different frequencies. Which stage(s) of the lifecycle are involved in making these data usable together?
(**6**) The researchers publish a graph showing a strong increase in average temperature. What information would you want to know before trusting this conclusion?

Exercise 2 – A Broken Scrapper

A small online bookstore runs a nightly job that scrapes competitor prices from their websites to update its own recommended prices. The pipeline is simple: a scraper visits a fixed list of competitor pages, extracts the listed price for each book, stores it in the bookstore's internal price table, and a morning report shows the average competitor price for the day.

One week, the marketing team notices that the reported average competitor price has been exactly \$0.00 for the past ten days. When they check the competitor sites manually, real prices are clearly still shown.

(**1**) The final report contains only \$0.00 prices. Can you identify the stage of the data lifecycle responsible for the problem? Justify your answer.
(**2**) Propose at least one plausible technical explanation for why the scraper kept running (no crash, no error) while returning wrong data.

(**3**) Why might nobody have noticed for ten days? What does that suggest about how the pipeline should have been designed?
(**4**) Propose two concrete check the team could add to catch this kind of problem earlier next time: one close to the source, and one on the final output.

Exercise 3 – The 92% Satisfaction Rate

This exercise previews ideas around sampling and bias that we will see in more details in the follinwg sessions.

A university wants to know whether students are satisfied with their degree programmes. It sends an online satisfaction survey to all 10 000 students.

After one week, 3 000 students have responded. Of these, 2 760 indicate that they are satisfied with their programme. The university publishes the following statement:

"92% of our students are satisfied with their degree programme."

The data pipeline was the following:

1. Students receive a link to an online questionnaire.
2. Responses are collected in a spreadsheet.
3. Incomplete responses are removed.
4. The university computes the proportion of remaining responses indicating satisfaction.
5. The result is published in the university's annual report.

You learn the following additional information:

- First-year students were much more likely to respond than final-year students.
- The questionnaire did not require authentication, so a student could submit it more than once.

(**1**) At which stages of the data lifecycle do you see potential problems?
(**2**) Is the statement "92% of our students are satisfied" justified by the data? Explain.
(**3**) Which additional information would you want before deciding whether the result is trustworthy?
(**4**) Suppose the university cannot obtain more responses. What could it change in the pipeline to make the final result more informative?
(**5**) Identify one thing that could be technically correct at one stage of the pipeline but still contribute to a misleading final conclusion.