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?

Possible questions could be

- "How has local temperature changed over the last 20 years?"
- "How frequently have unusual weather events occurred over the last 20 years?"

Let us consider the first question. It can be more precise by specifying where (which location, region?) the evolution of temperature is studied. The exact time should be made explicit since the "last 20 years" will change. The phrasing also makes it ambiguous whether we are comparing 2026 and 2006, or if we consider the evolution over time between the two. In the later case, we should also precise whether we want to understand global trends over the period or the daily/monthly fluctuations. The question should also precise what temperature do we consider, e.g., the average or the maximum over the period.

(**2**) Identify several sources of data in this scenario. For each source, indicate who or what produced the data and how it was collected.

Possible source include weather stations and citizen recordings (plus satellite images and technician logs as per the next questions). For weather station, data is typically collected automatically and would record temperature, rainfall, wind speed, etc., at regular intervals. Citizen recordings might be done manually using personal equipment. Both are essentially sensor logs. One should distinguish between values measured by the weather station and by citizen recordings as the quality may vary. Taking into account the next question, satellite images might also provide maps of temperature, cloud cover, vegetation.

(**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?

Images are unstructured since they have no predefined fields the relevant information (e.g., cloud cover) is not represented explicitly and must be extracted/ Free-text maintenance notes are unstructured as well requiring natural language processing, although they could be partially structured if technicians filled in a form instead. Temperature readings are most likely structured with fixed fields such as date, time, station ID, and temperature. Note that storing them in a CSV is not enough of a justification to conclude that the data is structured. Combining these directly is hard: only the CSV data can be queried or joined as-is; the other two need to be processed into a more structured form first.

(**4**) What potential issues can you expect regarding quality, trustworthiness of usability or the data?

Examples include missing or incomplete measurements, sensor errors or calibration problems, changes in measurement frequency or resolution, changes or replacements of sensors over time, inconsistent units, etc. These issues affect both the quality of the data and the conclusions that can be drawn from it.

(**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?

Primarily cleaning, since the units can be standardized and the measurements can be transformed into a common representation. The choice of how to store and combine the resulting data also involves storage, e.g., if the analysis can deal with different frequencies.

(**6**) The researchers publish a graph showing a strong increase in average temperature. What information would you want to know before trusting this conclusion?

You should consider the provenance of the data, the period and locations covered, missing values, sensor changes, preprocessing, how the average was computed, and how the graph was built: essentially all stages of the lifecycle.

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.

The immediate problem most likely originates in the **Collection** stage: the scraper is failing to correctly extract the price, even though the later stages (storing, reporting) are working correctly on whatever value they receive. The absence of validation means that the **Cleaning** stage did not detect the invalid values. Nothing checks whether \$0.00 is even a plausible price before it is stored and reported. The fact that the later stages continued to process them also illustrates that lifecycle stages are interconnected rather than independent.

(**2**) Propose at least one plausible technical explanation for why the scraper kept running (no crash, no error) while returning wrong data.

A common cause: the competitor's website changed its layout (a redesign), so the rule the scraper uses to locate the price on the page no longer matches anything. Rather than raising an error, the extraction code may quietly default to an empty value or zero, letting the rest of the pipeline continue to run "successfully" on bad data.

(**3**) Why might nobody have noticed for ten days? What does that suggest about how the pipeline should have been designed?

Because the pipeline had no failure signal at all. It kept running without crashing or logging an error, so nothing prompted anyone to look. The bad values flowed silently all the way to the final report. This suggests the pipeline needed some form of validation between collection and storage/reporting; "it ran without crashing" is not the same as "it produced correct data."

(**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.

A check close to the input would be to reject or flag any scraped price that is \$0, negative, or wildly different from the last known price for that item, instead of storing it as-is. This is essentially data validation. A check close to the output would be to alert whenever the reported average changes abruptly or looks suspiciously uniform (e.g., every value being exactly \$0.00). This is essentially a sanity check.

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?

Potential problems occur at several stages. Collection may produce a non-representative set of respondents and duplicate responses. Cleaning removes incomplete responses but does not necessarily solve these problems. Analysis may incorrectly generalize the result from respondents to all students. Communication presents the result as applying to all students without explaining these limitations.

(**2**) Is the statement "92% of our students are satisfied" justified by the data? Explain.

Not necessarily. The data shows that 92% of the *respondents* indicated satisfaction. It does not establish that 92% of all students are satisfied. The low response rate and differences in response rates between student groups may introduce bias.

(**3**) Which additional information would you want before deciding whether the result is trustworthy?

Examples include the response rate by student group, the number of duplicate responses, how incomplete responses were handled, the exact wording of the question, and how respondents compare with the overall student population.

(**4**) Suppose the university cannot obtain more responses. What could it change in the pipeline to make the final result more informative?

Possible improvements include preventing duplicate responses, reporting the response rate, reporting results separately for different student groups, weighting responses if appropriate, improving the questionnaire, and communicating the limitations of the result.

(**5**) Identify one thing that could be technically correct at one stage of the pipeline but still contribute to a misleading final conclusion.

For example, removing incomplete responses may be technically correct cleaning, and computing $2760/3000 = 92\%$ may be mathematically correct analysis. However, the final interpretation can still be misleading because the respondents are not necessarily representative of all students.