Exercise Sheet: Practical Data Collection

Data, Data Storage, Data Collection

Exercise 1 - Data Collection Methods

For each research question, indicate the most suitable data collection method and justify your choice in one sentence.

(**1**) How do students spend their free time?
(**2**) What is the average temperature in a city?
(**3**) How many people visited a website last month?
(**4**) Does a new drug reduce symptoms?
(**5**) What are the most popular restaurants in town?

(**1**) Survey allows directly asking the students about their activities. Conducting interviews might also provides deeper, subjective insights into how the students choose to spend their free time.
(**2**) Sensors (e.g., weather stations) provide continuous, and high-frequency measurements of temperature. This method is ideal for collecting accurate, real-time data over time. Depending on who is asking the question, the data might also be fetched from already recorded data, e.g., via an API.
(**3**) If the website provides an API (e.g., Google Analytics API), you can directly retrieve visitor logs, page views, and other analytics data. If you own the website, you can log events (like queries), meaning that the data is already and should just be retrieved.
(**4**) Conducting a controlled experiment where one group receives the new drug and another receives a placebo enables causal inference about symptom reduction.
(**5**) Collecting reviews and ratings from restaurant platforms (e.g., Google Maps or TripAdvisor) can identify restaurants with the highest ratings, most reviews, or most check-ins as proxies for popularity. Here again, the data is already available and should just be retrieved, altough it probably needs to be scrapped.

Exercise 2 - Identifying and Mitigating Bias in Data Collection

For each scenario, identify the type of bias, explain why it is problematic, and propose a mitigation strategy.

(**1**) A survey about student stress is sent to students who attended lectures.

The survey only targets students who attend lectures, excluding those who skip or are less engaged, who may experience different levels of stress. This results in a non-representative sample of the entire student population. The bias might be classified as selection bias (overrepresentation of some groups), or futher refined to undercoverage bias (some groups – the students not attending the lecture) have a null probability of being in the sample: they are already out of the acces frame. This also highlight that undercoverage bias is a specific type of selection bias. A mitigation strategy would be to distribute the survey to all registered students via email, the university portal, or some student organizations.

(**2**) A fitness tracker records steps made during the day but only when the user wears it.

If the fitness tracker is not worn consistently, the resulting data will be incomplete, potentially leading to underestimation or overestimation of the user's activity. For example, the tracker might be removed during high-intensity activities (e.g., for safety reasons), only worn during specific activities (e.g., running or swimming), forgotten at home, removed for charging, or taken off for comfort. Such bias is called *measurement bias*. Mitigation strategies include combining tracker data with complementary sources (e.g., smartphone motion data) or applying algorithms that detect and correct non-wear periods based on user behavior patterns.

(**3**) A company develops a new product and posts a poll about it on its website.

Posting a poll on the company's website primarily reaches existing or loyal customers, who are more likely to express positive opinions, skewing feedback about the new product. This is an example of response bias. To mitigate the bias, the poll could be distributed through neutral third-party platforms (e.g., social media) to reach a broader and more diverse audience. Another solution might be to target non-customers to gather unbiased opinions or to change the collection method entirely, e.g., via to focus group.

(**4**) A study on sleep habits collects data from smartwatches.

The study excludes individuals who do not own or use smartwatches, often correlated with age, income, or lifestyle. Like the survey about student stress, this is an example of undercoverage bias, which limits the generalizability to a broader population. As soon as we want to avoid restricting the findings to smartwatch users, the only plausible mitigation is to use alternative data collection methods to include non-smartwatch users.

(**5**) A recommendation system collects user clicks to train its model, but only shows popular items to new users.

The system reinforces the popularity of already popular items by only showing them to new users. This is a classic example of self-perpetuating cycle also known as feedback loops (there is no standard name for this kind of bias). The loop limits the discovery of less popular but potentially relevant items, reducing diversity and personalization. Randomization can help break the feedback loop, e.g., randomly showing less popular items.

(**6**) A sensor network in a city that has more sensors in wealthy neighborhoods shows good air quality reports.

The sensor distribution overrepresents wealthy neighborhoods, leading to air quality reports that do not reflect the experiences of the entire city. This can mask pollution hotspots in poorer areas. We did not cover specifically this bias in the lecture, but it is known as spatial bias. A solution to mitigate the findings would be deploy more sensors in underrepresented areas. In case where this is not feasible, another solution is to weight sensor data by neighborhood population or density of the sensor in the area.