Data Collection

Learn about where we get data and what we need to consider ethically.

Most people are aware that data visualizations, machine learning algorithms, and analyses require data. But where does that data come from? Does everyone have access to data for analysis?

In this article, we'll cover:

- The importance of ethics in data collection.
- How data may be collected.
- Common sources of freely available datasets.

Data ethics and privacy

Whenever we talk about data collection, we need to discuss data ethics. Much of the data available to us comes from individuals and would be considered **personally identifiable**, meaning we could use it to identify someone. Many people use the acronym PII (pronounced "pie" — yum!) for the term "personally identifiable information". Examples of PII are address, email, phone number, social security numbers, credit card numbers, and medical records. We all have an obligation to protect personally identifiable information.

Ethical issues regarding data collection may be divided into the following categories:

- **Consent**: Individuals must be informed and give their consent for information to be collected.
- **Ownership**: Anyone collecting data must be aware that individuals have ownership over their information.
- **Intention**: Individuals must be informed about what information will be taken, how it will be stored, and how it will be used.
- **Privacy:** Information about individuals must be kept secure. This is especially important for any and all personally identifiable information.

Data collection

We collect, process, and analyze data to better understand our world and make more informed decisions. The first step in any data work is to collect the data itself. Data can come from a lot of places, including research, governments, technology, observation, or directly from individuals — the list is endless!

We collect this data in many different ways. One way is to seek out information that doesn't yet exist and measure it directly. This can include activities like surveys, observational studies, or recording the results of an experiment. This kind of data might be considered *static*, meaning the information is collected once and does not change. Think about conducting a survey by mail: the survey results are collected and recorded only once.

Data can also be live and ever-changing based on the most up-to-date information. For example, apps and websites can track clicks and time spent on pages across multiple users at the same time without a human actively recording all the data points. Unlike the static data of more traditional methods, sensors and trackers can also continuously update data to include new information in a live feed. Think about weather predictions: the data that goes into weather predictions are updated continuously to get the most accurate predictions.

Finally, rather than collect measurements directly, we can also use existing data that was collected by others or for some other purpose. There are lots of databases that are freely available for public use. We can even compile data from a variety of sources and join them together before an analysis.

Data sources

Many organizations house all kinds of data. Datasets are often kept private or can only be accessed for a fee. This may be done for reasons like protecting the identity of individuals, keeping valuable information from competitors, or making a profit from data collection.

The following list has links and descriptions for websites that provide free access to some interesting datasets. The companies and organizations on this list provide public access to data, allowing anyone with internet access to view this information. The websites vary in how they provide data access: some may have a CSV or Excel file of data that can easily be downloaded to a computer, while others allow access to a database via an API.

- World Health Organization (WHO): Data available on the WHO's site cover a variety of health-related topics, such as COVID-19, air pollution, and even brain health. There are fact sheets and direct access to various datasets, including:
    - Mental health
    - Road traffic mortality

- FiveThirtyEight: This is a very popular analysis website that provides direct access to some of their datasets. Topics include sports, politics, science & health, culture, and economics. Check out some of these interesting finds:
    - Political polls
    - The best NBA players
- Data.gov: The U.S. government has its own open data collection. The site includes information on agriculture, climate, energy, and many other topics. Here are a few unique datasets:
    - Maritime limits and boundaries
    - Storm Events Database
    - Census data for the United States of America
    - Census data for EU countries
- Data Unicef: UNICEF's Data and Analytics team provides global access to data on children. This organization believes that the right data in the right hands can help us make informed and equitable decisions. You can view a variety of topics and data from various countries.

Looking for something specific? Google Dataset Search works like a google search bar for datasets. We think the following datasets look really interesting!

- Orchids — Did you know the total value of trees, plants, and flowers exported from the Netherlands in 2020 was nearly 9.8 billion euros?
- Biodiversity at U.S. national parks — Did you know that Haliaeetus leucocephalus (also know as a Bald Eagle) can be found in just about every U.S. National Park? Check out this data file to explore animal and plant species that have been identified and verified by evidence in national parks.

- Revenue of the cosmetic & beauty industry in the U.S. — Talk about big money: the revenue of the U.S. cosmetic industry was estimated to amount to about 49.2 billion U.S. dollars in 2019.