Data Collection

Data, in its simplest form (sometimes called raw data), is just a collection of words, numbers, pictures etc. **Processed data** is data that has been saved into a format, such as a data table, spreadsheet, or comma separated values file. **Clean data** is data that does not suffer from being inaccurate, incomplete, or inconsistent (the 3 "I"s). Finally, **information** is data that has been analyzed to yield insights.

Vocabulary

- Data collection
- Data (raw)
- Processed Data
- Clean Data
- Information
- Big Data
- Outliers
- Data curation
- Data governance

Data collection

There are several steps in the data collection process. All of these steps may not apply in your situation, and you may change the order in which they are done to better suit your circumstances.

A. What is the data you are collecting? While all of the above descriptions of data may seem similar, they describe different states of the data collection process. For example, raw data can be the output of a temperature sensor or user input from an Internet survey form. You must start with a good description of what you are collecting.

B. Anonymous data or data tied to specific people? This is a huge issue in data collection. Some institutions require express permission to collect data with personally identifiable information. How would you anonymize data that contained personal information? One way is to create a unique identifier for each individual and then keep two separate databases: one with only the unique identifier as an index for each individual and another which ties the unique identifier to that individual's personal data. The latter would need protection (encryption, private storage, etc.) but allows you to use the former without risk of exposing data of individuals that should be private.

Your organization should have a process for de-identifying personal information and vetting data collection procedures. See the module on Ethics and Privacy Considerations for more information.

C. The Five Cs of data: clean, consistent, conformed, current, and comprehensive.

1. **Clean**
   Clean data means data that has no missing values, no inaccurate data, no out of range data, no typos, etc. Having 100% clean data can be difficult to achieve. See the Data cleaning step below. During data collection, pre-emptive cleaning can be done by limiting values that may be assigned or calibrating your equipment.
2. **Consistent**
   Consistent data means that the data is the same no matter where it appears in your organization and the definition(s) associated with your data are consistent across your organization. This is sometimes referred to as a "single source of truth". A sales figure should be the same no matter where it is seen. For data collection, this means your collection parameters are always the same no matter where or how the data is collected.
3. **Conformed**
   Conformed data means the values fit the purpose of the data collection. For example, if you are collecting temperature data, then using degrees (whether Fahrenheit or Celsius) is appropriate, while using an arbitrary index is not. In addition, the parameters should be clearly defined. For collecting data in space where "year" is the category, does it mean terrestrial year or Martian year? For accounting, does it mean calendar year or fiscal year?
4. **Current**
   Collecting data means you are collecting current values. However, since today's values are tomorrow's old data, you should timestamp your data as it is collected so users know just how current it is. Current can also mean a single value. When trading stocks there are many price quotes tied to a specific time. However, there is only one "closing price" for a stock. It remains current until the next closing price.
5. **Comprehensive**
   The data should have width as well as depth. When collecting names, enable the collecting of prefixes (Mr., Miss., Doctor, Reverend, etc.); first names; middle names (perhaps more than one); last names (again perhaps more than one and/or hyphenated); honorifics; and titles where appropriate. Routinely you would include first, middle, and last names for things like customer and employee data. But for official records, government functions, etc. then prefixes, titles, honorifics, etc. can be crucial data.

D. The Five Vs of Big Data: volume, velocity, veracity, variety, and value.

1. **Volume**
   Volume of data is an important consideration. Contemporary social media post can run into the millions or billions even for a short collection period. Other data, like weather values, may have small sizes per collection point, but amass to large scales over periods of time and among large numbers of collection points.

2. **Velocity**
   Data can be collected at varying speeds. Financial transactions and social media posts usually require very high-speed collection methods for real-time analysis. Weather information, on the other hand, is usually collected at less frequent intervals. Other data may come in spurts, requiring high speed data collection followed by periods of little or no data collection.

3. **Veracity**
   How true to your purposes is the data? Faulty equipment, incorrect calibrations, sloppy collection methods, etc. can ruin your data collection process by producing invalid data. It may look okay, but it is going to result in inaccurate analysis. Another aspect of veracity is how well do you trust the source of data? If you are using third-party sources, they should be vetted and noted in your data curation process. Blockchain technology can aid in that process.

4. **Variety**
   Some data collection requires a variety of sources. For example, biographies may include data from archives, media posts, government records, audio/video samples, etc. Collecting (and storing) data from a large variety of sources poses unique challenges. If done in real-time, then synchronization of data collection is a vital concern.

5. **Value**
   When collecting data, make sure it is of value to the state purpose of collecting the data. Twitter feeds are popular data collection examples, but if you are not using social media for your project that data is of no value to you. If you are collecting health information about newborns in Asia, then data from other parts of the world or older people is of no value for your specific needs.

E. What are the parameters of the data collection? Is it a limited time collection or ongoing? Do you have a set of constraints for the data, such as a valid temperature range? You must be able to distinguish between outliers and invalid data as well. For example, if you are analyzing temperature data but only want to consider data collected with modern sensors, you would remove hand-collected temperature data before analysis.

F. Data cleaning. Will you clean data as it is collected (with custom forms, for example) or will you do it afterwards? Cleaning is an essential part of data analytics and is often the most time-consuming step. Anything you can do to cut down on data errors during collection will

reduce the amount of time spent cleaning the data. For example, if you are collecting data on cars, having a lookup table with preset car colors will prevent you from having to sort through entries like “cream”, “opal”, “eggshell”, “beige”, etc. all of which are describing white cars. See the Data Cleaning module for more information on this process.

G. Data storage. How will you store the data that is collected? Local, cloud, hybrid, etc.? A pristine copy of your data should be stored in safe places in case your work gets lost and to aid in verification of the analysis. See the module on Data Storage for more information.

H. Data presentation. What is the intended audience and what data presentation best suits that audience? Are they other developers/analysts who need raw data files or are they executives in your organization who need the condensed “executive summary” version? Always make sure your data presentations come from your curated data sets and can be reproduced without differences (aka “single source of truth”) no matter who generates them.

I. Qualitative versus quantitative data. Qualitative data is often found in small sample-size projects where the data is subjective in nature. Quantitative data is often found in scientific data where objective, free-form data is needed. For example, “tall” is a qualitative data observation while “6ft, 2in” is a quantitative observation. They may, in fact, be describing the same person!

J. Structured versus unstructured data. The biggest difference between structured and unstructured data is when structure is applied to the data. Structured data is usually filled into an existing template or structure as it is collected and before it is saved. Unstructured data is usually saved first, then read into a data structure when it is used.

K. Finally, you should have a data curation plan that handles data issues from start to finish. Even when you are done with the data, it should be made available to other users and to ensure repeatability of your analysis.

Sources and methods

- Real-life datasets

  Real-life datasets abound in data science. Sports, census, weather, *et al* all generate volumes of real data. These are often the best datasets to use for beginning data science projects as people can readily relate to the topics and ideas for visualization are easier to generate.

  See the resources section for a few examples.

- API

An API (Application Programming Interface) is a collection of tools that allows for automatic connection with a data service that enables uploading and downloading of data. Far more efficient than web scraping, it is the preferred way to obtain data from a source.

- Synthetically generated

  Data that is synthetically generated can be used to test systems and models. For example, if you want to test your model's ability to detect and eliminate outliers, you can generate data with many outliers and outliers that are at the edge of detection.

  Synthetically generated data is also good for generating large amounts of data to test models. You could create millions of rows of fake customer data to test your system's ability to process Big Data and data that comes in large streams.

- Web scraping

  Web scraping is the process of farming data from a web page by searching the document structure for data elements. It is resource intensive and usually requires a specialized library to enable. Companies that provide data (and those that don't) discourage web scraping as it competes for live use of their website with real customers and skews their website use statistics.

- Open datasets

  Open datasets are free to use. They are often created by governments with tax dollars, or by industry with a vested interest in allowing public analysis of data. While the data itself is free, you may find such datasets being made available by third-party providers who charge a fee for their services, which can range form data cleaning to curation of historical data. Examples include the Open Datasets at AWS (Amazon Web Services) (https://registry.opendata.aws/) and IMDb (Internet Movie Database) (https://www.imdb.com/interfaces/).

- Closed datasets

  Closed datasets are usable by designated partners only. They include legal and medical services, government data that includes secret or sensitive information, business strategies and customer data, etc. If you are using a closed dataset, make

sure you are taking the proper precautions to secure the data and analysis. Typically you would need to "depersonalize" data by removing anything that could personally identify an individual, such as Social Security numbers, driver's license numbers, etc.

Classroom Projects

A. Have members of the class use a roster and ask other students whether they consider that student to be "short, medium, or tall" in height. Do the same with a tape measure or height measurement stand. How do the datasets relate? Is there a way to make the qualitative and quantitative data match? (Use data "buckets".)

B. Make audio recordings and then digitize them. How do the storage requirements compare? Which data collection method is better for archival use? Should you compress the data?

C. Find a school presentation such as an art show or academic contest and create survey forms to collect data from attendees.

D. Create a survey form to conduct a census of the school population. Collect both quantitative as well as qualitative data. Be sure to anonymize the results!

E. Create a collection of visual media items such as artwork, seating charts, pictures, etc. How would you store the data? Is it structured or unstructured? Does it need to be anonymized?

F. Create a science project (growing plants from seeds, etc.) and collect data from it. What are your options for data collection? Can you include data of different types such as numbers, text, and images? What sensors can you use and do they have to be digital? (Thermometers, light meters, etc.) Compare the results among different groups and form hypotheses to explain differences.

G. Have students collect data about something they do every day or week, such as brushing their teeth or doing yard work. Students should create a template for data entry that contains at least five observations. (Day, time of day, duration, weather if outside, etc.) Compare results weekly and for the entire term.

H. Use Raspberry Pi devices to enable data collection from sensors. (See resources below.)

Resources

1. Raspberry Pi projects. https://projects-raspberry.com/raspberry-pi-projects-to-facilitate-research/

2. Open source Raspberry Pi classroom projects. https://opensource.com/education/15/12/5-great-raspberry-pi-projects-classroom
3. Data collection with Raspberry Pi. https://www.instructables.com/Data-Collection-With-Raspberry-Pi/
4. Web search results for classroom projects. https://duckduckgo.com/?q=raspberry+pi+data+collection+classroom+project&atb=v1-1&ia=web
5. Data measurement plan (Six Sigma). https://blog.masterofproject.com/data-collection-plan-six-sigma/
6. Data collection projects using Arduino. https://create.arduino.cc/projecthub/projects/tags/data+collection
7. TI data collection equipment. https://education.ti.com/en/products?category=data-collection
8. TI classroom projects. https://education.ti.com/en/activities/innovator
9. Kaggle datasets. https://www.kaggle.com/datasets
10. Census bureau datasets. https://www.census.gov/data/datasets.html
11. data.world collection of MLB datasets. https://data.world/datasets/baseball
12. Bureau of Economic Analysis data. https://www.bea.gov/data
13. Data.Gov data collection. https://catalog.data.gov/dataset
14. 50 Free Datasets (collection): https://blog.journeyofanalytics.com/50-free-datasets-for-data-science-projects/
15. AWS Open Datasets: https://registry.opendata.aws/
16. IMDb Open Datasets: https://www.imdb.com/interfaces/

This material is based upon work supported by the National Science Foundation under Grant DUE 2055411.Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.