Many of us have heard the phrase “knowledge is power” or a variant of it. In the world of data science, I believe we can apply a similar phrase, “information is power,” as data leads to information which leads to knowledge. Data is collected every day, whether it be experimental observations, interviews with individuals or groups, or market surveys. However, this data is usually collected in a raw format; that is, the raw data quality must be improved to use the data for a desired purpose. I will be discussing some of the different methods to collect data, as well as how the quality of said data can be improved.

Data can be collected using four different methods: Observations, interviews, document analysis, and surveys. The first method, observation, is how everyone collects data about the world around them. Qualities and quantities are observed and recorded physically or digitally, from controlled research studies down to simple weather forecasts. The next method, interviewing, is arguably the easiest method to collect data on a person. After all, what better way to find out about someone than asking them a question? Document analysis is similar to observation but differs in that documents and records *already written* are analyzed. Lastly, surveys are a “mass interview” method for collecting data from a large sample of a population.

Before exploring how data quality can be improved, we must first understand what the phrase *data quality* refers to. Data is generally considered of high quality if it is “fit” for the desired use, but there are multiple definitions for data quality depending on the application. For example, from a business perspective, data quality is how well data can be used by a business for everyday operation, decision-making, and planning. From a consumer perspective, data quality is how well a consumer can understand the data and use it in a given setting (e.g., a consumer won’t understand data on an online marketplace if it is a table of products’ inventory numbers; the data would be low quality in that case). High quality data is free of redundancies, incomplete data points, and outliers.

Now that we understand what constitutes high quality data, we can begin to brainstorm how to improve our data quality to meet our desired needs. For simplicity, let’s look at the business definition of data quality. For example, at my job, automated probes detect bacterial temperature, pressure, acidity, and oxygenation, then compiles a very messy CSV file of those variables. We cannot read the CSV file in its default format, not just because it is a long list, but because all the units are non-metric, causing machines to read the data differently and making it difficult to compare results. To increase the quality, we *normalize* the data, or transform it to meet one standard. Normalization, or standardization, ensures that machines don’t read multiple data points for the same value (e.g., 20 C and 20 Celsius would be two unique points). Once our data has been normalized (to the International System of Units in our case), it is then considered high quality, as we can then use it to view trendlines and compare across multiple experiments.

Normalization is just one of many techniques to improve data quality across multiple fields. Two other common methods include fixing duplicate records and incomplete entries. Duplicate records are most commonly encountered in businesses that ask customers to input their data. If, say, names and e-mails are input twice, this can lead to repeat entries which can mess up a business’s customer statistics, to name one problem. Fortunately, machine learning can be applied to help remedy the issue. An algorithm can be developed that will detect when a data entry is a duplicate and remove it if it is.

I believe that incomplete data entries have greater consequences than duplicate entries. They reduce the statistical power of a study and lead to invalid conclusions. Fortunately, there are several methods to remedy them. If there are enough data points, the blanks can be omitted entirely with little effect on trends (listwise deletion). This is what I frequently do, as my probes record multiple points per second. If the data is quantitative, the mean value can be substituted as a rough estimate (mean substitution). In addition, standard deviation can be input to the missing point to determine how varying the point(s) will affect the overall trend (sensitivity analysis). The best solution is simply to prevent incomplete data from arising in the first place.

Data quality is highly subjective; these methods are just some that are used across the disciplines that use data. If we collect data by any means without improving its quality, we could run into several issues such as overreporting or underreporting quantities (not normalizing) or false-positives/false-negatives (keeping duplicates or incomplete entries).

References:

1. Ainsworth, Q. (2020). *Data Collection Methods.* JotForm Education. Retrieved September 7, 2020 from https://www.jotform.com/data-collection-methods/.
2. Brigham, A. (2018). *Critical steps for improving your data quality.* Experian Data Quality. Retrieved September 7, 2020 from https://www.edq.com/blog/critical-steps-for-improving-data-quality/.
3. Fleckenstein, M. & Fellows, L. (2018). *Modern Data Strategy.* Springer Science.
4. Joshi, N. (2016). *Four ways to improve your data quality.* Allerin. Retrieved September 7, 2020 from <https://www.allerin.com/blog/4-ways-to-improve-your-data-quality>.
5. Jovancic, N. (2019). *Data Collection Methods for Obtaining Quantitative and Qualitative Data. LeadQuizzes.* Retrieved September 7, 2020 from[*https://www.leadquizzes.com/blog/data-collection-methods/*](https://www.leadquizzes.com/blog/data-collection-methods/).
6. Kelleher, J. D. & Tierney, B. (2018). *Data Science.* The MIT Press.