GENDER 101 COURSE

Module 2: Gender Data Collection and Processing

Overview

This module aims to delve into gender data collection, biases, and processing. There's lots of interesting content for this week, so we've listed below our suggested learning path.

Take this poll

What is your organization's largest data challenge?

- ☐ Too much data
- ☐ Not enough data
- ☐ Political and/or organizational hurdles
- ☐ Lack of capacity and/or training
- ☐ Data biases
- ☐ Other

Gender Data Collection

In this section, we will learn about data collection with emphasis give to methods and tools for conducting effective data collection. We will also go through some considerations for collection.Examining Biases

In this section, we will cover the potential biases in the gender data lifecycle. We will also highlight the importance an intersectional approach is to gender data collection.

Gender Data Processing

In this section, we will learn about data processing, with emphasis given to the methods and tools for conducting effective data processing.

Case Study

Before we delve into this week's activities, in this section, we will revisit our case study. Together, we will walk through both Dr. Lopez and Rumy's goals in data collection and processing.

Activities

Spot the Biases

In this activity, we'll closely examine Dr. Lopez and Rumy and their datasets looking for potential biases that may be contained within.

Spot the Processing Concerns

In this activity, we'll take another look at Dr. Lopez's and Rumy's datasets, but this time examining with the purpose of identifying any potential processing concerns. We will also help them standardize their processing methods.

The Bukavu Series Deep Dive

Review the Bukavu Series and add your thoughts to the discussion board.

Sourced Participant Resources

This section provides a place for you to share resources that you have used or created to help you when it comes to gender, data, and gender data. Click here to add resources you would like to share with other participants! There is no limit to how many you can add. You can also review those submitted by others.

Gender Data Collection

Recall that collection is our first phase of the Gender Data Lifecycle. Collection is the process of gathering and measuring information in a systematic fashion. Gender data collection ensures both sex and gender factors are captured. Example methods of collection include:

- Questionnaires and surveys
- Observations
- Interviews
- Review of existing documents and records

Take this poll

What method of collection do you primarily work with?

- ☐ Surveys
- ☐ Observations
- ☐ Interviews
- ☐ Reviewing existing documents
- ☐ All of these
- ☐ None of these
- ☐ Other

Methods of Data Collection

We learned there are two types of data: quantitative and qualitative.

- Qualitative data is descriptive information that cannot be measured with numerical values. For example, this can include items like emotions.
- Quantitative data is data that can be measured by numerical values. For example, this can include items like height, weight, volume, price, temperature, etc.

Analyze the chart below to help you compare and contrast the advantages and disadvantages of various collection methods. Feel free to refer back to this chart when planning for the collection of your next project! While many of these methods support qualitative data collection, some have opportunities for quantitative data collection.

METHOD	DESCRIPTION	ADVANTAGES	DIS-ADVANTAGES
STRUCTURED INTERVIEWS	1-on-1 using predetermined questions	Easy to administer, takes up less time than other interviews	Does not allow for follow up or variation; may lack depth
UNSTRUCTURED INTERVIEWS	Questions not predetermined and more open-ended	More in depth, allows interviewer to follow up; less rigid, more open responses	More time consuming, less consistency in data collected
FOCUS GROUPS	Group interview that involves group interactions	More time effective, interview several people at once, provides social context	Group dynamics can interfere with data accuracy, must limit number of questions
DIRECT OBSERVATION	On site observation and notation of processes & behaviors	Can gain holistic perspective & observe unexpected outcomes; occurs in natural setting.	Time-consuming, observer may affect behavior of participants. observer bias, may be intrusive
PARTICIPANT OBSERVATION	Researcher participates in group activities while observing others	Active participation provides more complete understanding & context; more natural/less intrusive	Observer may become too close to the topic or participants, can lead to bias
WRITTEN DOCUMENTS	Researcher uses existing documents to study topic (letters, diaries, memos, emails)	May provide factual information otherwise not attainable if writers are deceased; inexpensive	May be tainted by writer's POV or difficult to verify validity, may find conflicting information
ARTIFACTS	Researcher studies items used by different societies or cultures to gain an understanding of the past	Provides insight into how people lived, what they valued, & their knowledge & opinions	May be difficult to interpret meaning, must analyze in appropriate context & use alongside other methods

Method Spotlight: Ethnography

Ethnography is a qualitative research method that is often used in anthropology. It is an in-depth study of a culture or a singular part of a culture through real-time observation. Proper ethnography often takes long periods of time. It's common practice for ethnographers to spend years living among members of the studied culture. While we might not engage in ethnography for our data collection, there are key points we can draw from the practice. It is important when collecting data to also collect:

- people in their cultural setting;
- the way in which respondents interact with one another and with their social and cultural environment;
- respondent's deeds as well as their words;
- the implicit as well as the explicit;
- what is not said as much as what is said;
- their language, and the symbols, rituals, and shared meanings that populate their world, with the object of producing a narrative account of that particular culture, against a theoretical backdrop.
- (Source: Emerald Group Publishing)

Gender Data Collection Tools

Observational Studies and Interviews

- Devices for interviews and transcriptions, such as phones, digital voice recorders, or video cameras
- Frameworks for Ethnography. Examples: Speech Code Theory

Surveys and Questionnaires

- Google Forms is a survey administration application.
- UNICEF's RapidPro is a platform that allows anyone to build interactive SMS messaging systems and questionnaires.
- SurveyMonkey is an online survey development company.

Holistic Data Planning

- ADAPT is an innovative tool to help plan according to changing data demands. ADAPT was developed by Paris 21. It recently added a gender module.

Further Resources

- National Science Foundation's Common Qualitative Methods
- Dr. Chloe Sharp's list of Top 7 Tools for Online Qualitative Research

Considerations for Gender Data Collection

The History of Colonialism in Data Collection

Colonialism is defined as "the policy or practice of acquiring full or partial political control over another country, occupying it with settlers, and exploiting it economically" (CRRF). Colonialism can also pervade practices, such as data collection. It is no secret that data is powerful. Just as data can be acquired to do good, data can also be acquired and then exploited by bad actors. Therefore, we pay attention not only to what data we collect but also how and why we collect it.

Collection is often the first phase of an organization's data lifecycle. Therefore, the design of the collection method(s) sets the stage for the entire gender data process.

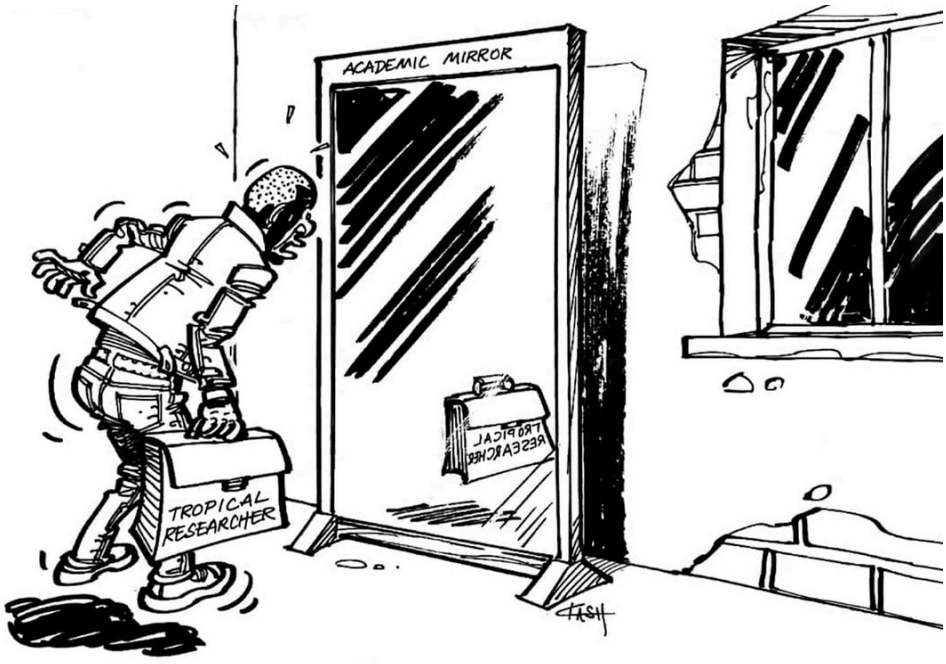
Questions to consider include:

- Who is consulted during the design and planning phase of data collection?
- Whose voices are centered?
- How are they treated?
- How are they compensated?
- How are they involved in the execution of the data collection?
- How are they involved throughout the gender data lifecycle?

Additionally, if you are using data that has already been collected - then you should strive to answer these questions. Thoughtful data practices must be vigilantly applied throughout all phases of the data lifecycle.

The Bukavu Series

To shed light on how colonialist data practices still exist, the Bukavu Series shows "social scientists from the Global North who work on the African continent [who] have long relied on the knowledge and skills of local researchers and researcher assistants. Without this assistance, navigating the subtleties and complexities on the ground would be impossible for many Northern scholars – particularly in insecure settings and conditions of conflict. Yet this dynamic grows out of colonial power relations, and even today researchers from the Global South are far too often characterized as mere “fixers,” “guides,” or “data collectors” for their Northern counterparts – when they should in fact be recognized as crucial collaborators in academic research."



Anonymous Word Cloud: How did the Bukavu Series make you feel?

Data Sovereignty

What is Data Sovereignty? For the sake of this course, there are two definitions to balance.

1. "In the corporate technology sector data sovereignty often refers to the understanding that data which are stored outside of an organization's host country and still subject to the laws in the country where the data are stored." (Cedric Jeannot, 2016)

   An example of this definition would be The General Data Protection Regulation 2016/679, also known as GDPR. GDPR is a regulation in EU Law that sets guidelines to protect the data of EU citizens. It must be followed by all applications or websites that include European members. (Investopedia).

2. More broadly, another definition of data sovereignty is the idea of a group or individual's right to control and maintain their own data.

   An example of the second definition would be the efforts towards indigenous data sovereignty. "Indigenous data sovereignty is the right of each Native nation to govern the collection, ownership, and application of the tribe's data. Native nations exercise Indigenous data sovereignty through the interrelated processes of Indigenous data governance and decolonizing data."

In this case, indigenous communities are defined as, "...peoples and nations...having a historical continuity with pre-invasion and pre-colonial societies that developed on their territories, consider themselves distinct from other sectors of the societies now prevailing on those territories, or parts of them. They form at present non-dominant sectors of society and are determined to preserve, develop and transmit to future generations their ancestral territories, and their ethnic identity, as the basis of their continued existence as peoples, in accordance with their own cultural patterns, social institutions and legal system"(Indigenous Data Governance: Strategies from United States Native Nations).

Why does indigenous data sovereignty matter? Missing and Murdered Indigenous Women, Girls, and 2-Spirit People (MMIWG2S) provides one example of the importance of data sovereignty.

2-Spirit or "Two-spirit" refers to a person who identifies as having both a masculine and a feminine spirit, it is a term most often used in indigenous communities.

MMIW, MMIWG, MMIWG2S

The Missing and Murdered Indigenous Women human-rights crisis disproportionately affects Indigenous peoples in Canada and the United States, notably those in the First Nations, Métis, and Inuit (FNMI) and Native American communities (Wikipedia). Due to racism, sexism, and other forms of systemic oppression, such as continued colonialist practices, indigenous women, girls, and 2-spirit people are more likely to be missing and/or murdered. However, the response to these cases by law enforcement has oftentimes been negligent due to those very same factors.

"The Urban Indian Health Institute, for example, found in 2016 there were 5,712 reported cases of missing and murdered Indigenous women (MMIW) [in the United States], but only 116 of those were logged in the Department of Justice database." (Tracking data of missing and murdered Indigenous people: A conversation with Annita Lucchesi)

Statistics like these reveal why indigenous data sovereignty matters.

"Why is data-keeping so important in understanding the MMIW crisis?

For us, our database is an opportunity to build on the knowledge and expertise of Indigenous communities and express the needs and priorities of families. It's an opportunity for us to take some leadership in changing the power dynamic, in saying, "We know ourselves. We know our experiences of violence. We know the dynamics of our community, and we know how to fix them." It steps away from a colonial government that's complicit. Indigenous data sovereignty is the right Indigenous people have to govern data about themselves and their homelands." - Annita Lucchesi

More resources:

- Reclaiming Power and Place: The Final Report of the National Inquiry into Missing and Murdered Indigenous Women and Girls
  "The National Inquiry must look into and report on the systemic causes of all forms of violence against Indigenous women and girls, including sexual violence. We must examine the underlying social, economic, cultural, institutional, and historical causes that contribute to the ongoing violence and particular vulnerabilities of Indigenous women and girls in Canada. The mandate also directs us to look into and report on existing institutional policies and practices to address violence, including those that are effective in reducing violence and increasing safety."
- They Disappear “Not Once, but Three Times: In Life, In the Media, and In the Data"
- Indigenous Data Sovereignty: Towards an Agenda (Book)
  Premised on the United Nations Declaration on the Rights of Indigenous Peoples, this book argues that indigenous peoples have inherent and inalienable rights relating to the collection, ownership and application of data about them, and about their lifeways and territories.

Examining Biases

Biases in Gender Data

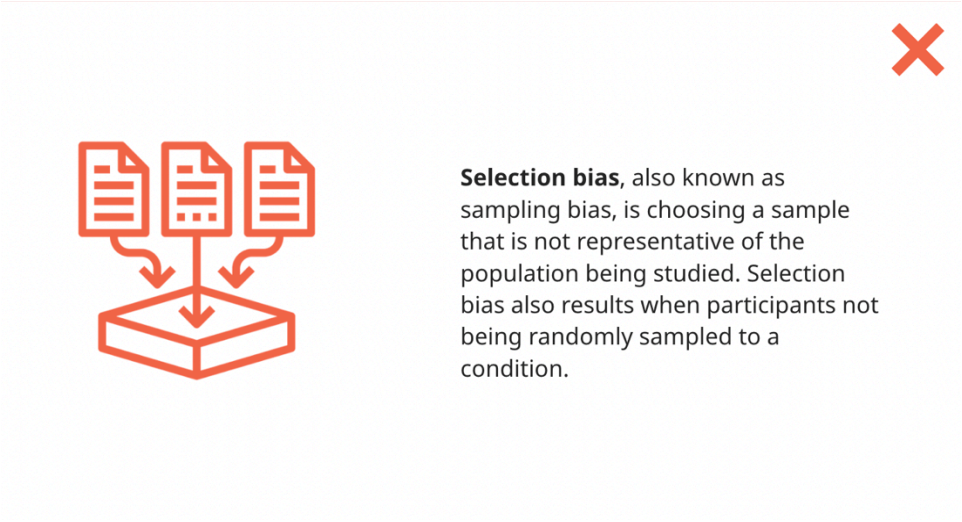
- After reading the considerations for collection, we will move forward with the phase of data collection itself.
- Ultimately, we all must strive for gender equity. Specifically, in gender data, our goal is to collect samples that truly represent the population, paying close attention to marginalized groups.
- We must understand the current barriers throughout the data lifecycle in order to properly address them. Some of these barriers are biases. Biases are personal and/or systemic judgments and pre-conceptions.
- Biases can appear and pervade at any stage of the gender data lifecycle. For now, we will focus on the first phase of collection.

- In the Activity: Spot the Biases, we will practice identifying different examples of biases utilizing the terminology

Biases in Data Collection

- Collection bias is a systemic and/or ingrained prejudice that influences data capture.

Social Desirability Response Bias is a type of response bias that occurs when a respondent is influenced to provide answers that are socially desirable and are not necessarily their true responses.



Selection bias, also known as sampling bias, is choosing a sample that is not representative of the population being studied. Selection bias also results when participants not being randomly sampled to a condition.

Non-response bias means certain members of the population being studied do not participate in the data capture. There are many reasons for non-responses bias. Examples include the medium of study (e.g. online survey) or time study is conducted (e.g. morning times).

Question Wording/Ordering Bias is a type of bias that occurs when conducting surveys or interviews. If the question is worded or ordered in such a way that it favors one response over another, then it is considered biased.

Response bias is a general term for any external factor that can influence a respondent's answers.

Demand characteristics is a term used to describe the ways participants of a study can influence the outcome. For example, if a supervisor is conducting a study of her employees, then employees may be inclined to answer in a certain way.

Outlier/Extreme Response Bias is a type of bias that occurs when there is extreme data values. These extreme values have the potential to skew the study's results.

- During the collection phase, it is critical to limit as much bias as possible. It is important to avoid making assumptions about the lives of women and men, girls and boys; as these assumptions extend beyond gender. Biases could also include assumptions about ability, socio-economic status, race, ethnicity, and sexuality.
- Some ways to combat biases include:
    - Review questions thoroughly for surveys, focus groups, and interviews. Try to view it from all possible respondent's point of views. If you struggle with this the following bullet may help.
    - Consult a gender expert before and throughout the entirety of the project/while you are conducting research (all phases of the data lifecycle). This is especially important because as stated before biases can appear and pervade at any stage of the gender data lifecycle. That means that while your collection phase may be bias-free, bias could be introduced in the analysis phase thereby corrupting the data and compromising the conclusions.
    - Ensure situational, geographical, and cultural context is considered. Consider ways in which you can thoughtfully include members of the community in your process.

We All Count's Data Equity Framework

We encourage taking into account We All Count's Data Equity Framework. Please click here to read their explanations of each phase of the lifecycle.

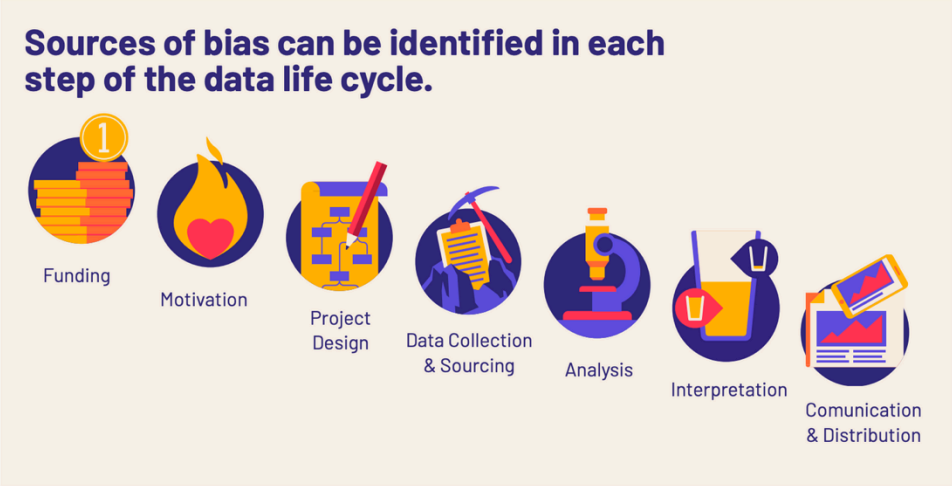


Image Source: We All Count

Connections to our Case Study

As we move forward with Rumy and her organization's efforts to provide COVID-19 relief and organize vaccine awareness campaigns, we need to be cognizant of biases and their effects. Recall, the current gender gaps in the development community, outlined by the Ladysmith Report, are also heavily affected by these biases.

Rumy is creating a community survey to determine their needs during the COVID-19 crisis and vaccine access. The goal is to gather information to provide proper relief to households. Some questions to brainstorm before engaging in this week's activity include:

- What needs to be considered during the design of Rumy's study?
- What are the potential demand characteristics Rumy should be aware of?
- What are the possible areas where bias could be present in the collection phase of the data lifecycle?
- What are tangible ways Rumy can limit biases in data collection?

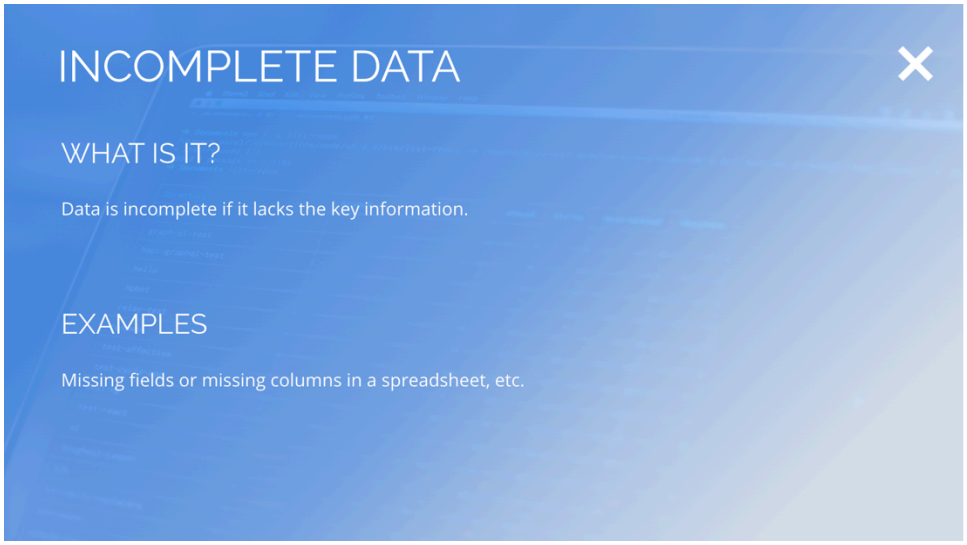
Gender Data Processing

Processing is the subsequent phase in the Gender Data Lifecycle after Collection. Raw data (also called dirty or unclean data) is data directly from the source and is often the state of data as it comes our of the Collection phase, heading into Processing. Raw data is often incomplete, has duplicate, has inaccuracies, and has inconsistencies. Processing is the series of operations needed to prepare raw data for analysis. Thus, gender data ensures that sex and gender factors are processed properly.

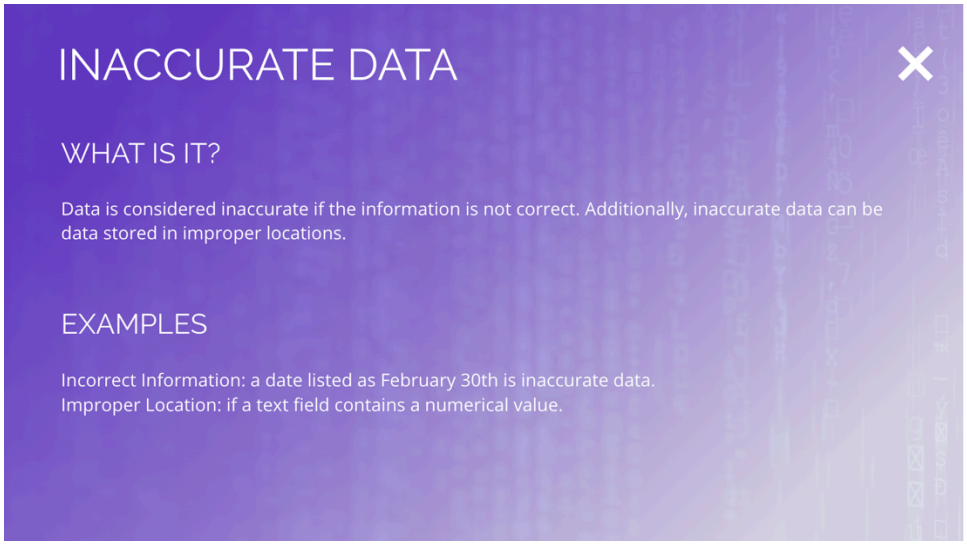
Example methods of processing include:

- Cleaning data
- Aggregating data
- Securing data

Raw Data (aka Dirty or Unclean Data) Terminology




DUPLICATE DATA
WHAT IS IT?
Data is considered a duplicate if it repeats itself.
EXAMPLES
If there are multiple entries from the exact same date, time, and social security number, then these entries may potentially be duplicates.


INACCURATE DATA
WHAT IS IT?
Data is considered inaccurate if the information is not correct. Additionally, inaccurate data can be data stored in improper locations.
EXAMPLES
Incorrect Information: a date listed as February 30th is inaccurate data.
Improper Location: if a text field contains a numerical value.

Connections to our Case Study

As we move forward with Dr. Lopez and her efforts to collect COVID-19 and vaccine data, specifically with a gender focus, we need to be aware of data processing.

Dr. Lopez is collecting large amounts of data she eventually wants to analyze and visualize. A lot of her data is raw and needs processing. Some questions to brainstorm before engaging in this week's activity include:

- What state is Dr. Lopez's data in? What types of raw data does she have?
- What standards and decisions need to be made (for cleaning, aggregation, and security)?
- What are possible areas where bias could be present in the collection and processing phases of the data lifecycle?
- How can Dr. Lopez limit biases during these phases?

Data Cleaning

When you first receive your data, it is likely not in a form that is optimized for analysis or visualization. Therefore, it is important to make sure that you clean your dataset. Clean data is data that is accurate and formatted to a set of rules. These rules can be determined by your organization or by your data analysis needs. The goal of data cleaning is to improve the data's eventual interpretability.

Data cleaning is a sequence of three steps aimed at producing high-quality data and taking into account the above criteria.

- Inspection: Detect incomplete, duplicated, incorrect, and inconsistent data.
- Cleaning: Fix or remove discovered anomalies.
- Verifying: Results are inspected to verify correctness after cleaning.

Examples of Data Cleaning

Below are examples data cleaning (showcasing the before 'dirty data' and after 'clean data'). Please note that data cleaning depends on you and your organization's needs and standards.

- Examples

Data Cleaning Tutorials in Excel

- Column Formatting
- Table Formatting
- Macros
- Conditional Formatting

Data Aggregation

Data aggregation is systematically combining data from multiple sources together. Processing may or may not involve aggregation.

Some key questions to consider:

- Is there a difference in how the data has been collected? Let's say you conducted interviews with a particular ordering of questions. Later, you decided to switch around the order for the remaining interviews. The change in methodology may be enough to consider that the two datasets may not be best to combine.

- Is there a difference between when the data was collected? Let's say you want to aggregate economic disparity between genders across the globe. However, your United States data is from 2015 and your UK data is from 2002. The difference in years needs to be considered (and possibly addressed) before deciding to continue with combining the datasets.

Connections to our Case Study

Since the outbreak of COVID-19 in early 2020, many organizations have been collecting data around the globe. One such example of data aggregation is the Coronavirus (COVID-19) Vaccinations.

We'll return to this example when we discuss the Visualization phase of the Gender Data Lifecycle.

Data Security

Data security is the commitment to confidentiality and integrity of data. Data security should always be a consideration at every phase of the data lifecycle. Specifically, when dealing with gender data, there needs to be standards in place to ensure the safety of all respondent's privacy. Data security may look different depending on your methodology and organization's standards. However, there are key things all organizations can consider:

- Protect Personally Identifiable Information also known as PII (any data that can be used to distinguish or trace an individual's identity). You can watch our video here describing what constitutes PII.
- Decide on a data encryption standard. Data encryption is the process of converting data into code to prevent unauthorized access.
- Understand the legal policies and procedures surrounding your data. For example, health information in the USA are held accountable to the requirements of the Health Insurance Portability and Accountability Act or HIPAA.

Connections to our Case Study

When Dr. Carolina Lopez engages with health data of women and girls across the globe, she needs to be highly aware of the data security considerations.

- How is Dr. Lopez protecting PII?
- What is her data encryption strategy?

Case Study

Dr. Carolina Lopez and COVID-19 Data

- Dr. Carolina Lopez is currently in the collection and processing phase of the Gender Data Lifecycle. She is in the field, side-by-side with healthcare workers, to ensure the availability of sex-disaggregated data. Ultimately, she wants to collect data that will inform how best to ensure an equal voice for women during decision making in the COVID-19 vaccine roll-out.

- Dr. Lopez is an example of the many women in the global health workforce. Two-thirds of the global health workforce is female (Source: The Interpreter).

Dataset 1: COVID-19 Cases and Vaccinations

- Since the outbreak of COVID-19 in early 2020, many organizations have been collecting data around the globe. One such example of data aggregation is the COVID-19 Dashboard by the Center for Systems Science and Engineering at Johns Hopkins University.

- Dr. Lopez is auditing data like the ones incorporated into the dashboard above. Recently, however, Dr. Lopez has been concerned with local efforts to capturing vaccine roll-out for COVID-19. She was recently provided a dataset from a local clinic in the United States that presented some issues. Consider Dr. Lopez and her goals when engaging in the following activities.

Rumy and COVID-19 Vaccine Awareness

- Rumy believes that women and lower-income earners will be disproportionately economically impacted by COVID-19. Even more concerning, Rumy believes these effects will compound during the vaccination roll-out. She needs to know the where, why, and how to help during this process. She is beginning to conduct a survey for households to determine the gendered effects of COVID-19 and to see where her organization can help with vaccine awareness.

Recent articles that Rumy has been looking at include:

- Women, lower-income earners hit hardest by Social Distancing (Source: Nous)
- 'Vaccine apartheid': Africans tell UN they need vaccines (Source: AP News)
- How Will COVID-19 Affect Women and Girls in Low- and Middle-Income countries? (Source: Center for Global Development)

Dataset 2: Household Surveys

Rumy is in the initial phases of design for her household survey. She piloted a few questions with a sub-group in Chile. She wants to improve her survey and needs to consider collection biases and data processing. Ultimately, her goal is to develop mitigation strategies that specifically target the impact of the outbreak and vaccine roll-out on women and build women's resilience.

Consider reading these two articles from We All Count prior to completing Rumy's activity:

- Representation and Visibility in Quantitative Surveys
- Who is the Head of Your Household?

Activities

Activity 1: Spot the Biases

Step 1. Access Dataset 1 (Dr. Lopez's dataset) to download or make a copy. Mark up and take note of any areas of the data that may suffer from biases. Repeat for Dataset 2 (Rumy's dataset).

Step 2. Refer to the Examining Biases section, are there any terms that can be applied to the biases you found?

Step 3. Address these prompts:

- Summarize in a list the types of biases found in the datasets.
- Of the biases you found, which one(s) do you think are the most important to address first? Why?
- What is one method or consideration that Dr. Lopez and/or Rumy can apply during data collection to prevent the biases you found?
- Are either of these datasets and example of gender data? Why or why not?

Step 4. Join the discussion! Post your answers and comment on at least one other participant's answers.

Activity 2: Spot the Processing Concerns

Step 1. Access Dataset 1 (Dr. Lopez's dataset) to download or make a copy. Mark up and take note of any areas of the data that may suffer from biases. Repeat for Dataset 2 (Rumy's dataset).

Step 2. Referring to the Data Cleaning section, are there any terms that can be applied to the raw data issues you found?

Step 3. Address these prompts:

- Summarize in a list the the types of raw data and processing concerns found in the datasets.
- Choose one type of processing concern you found and create a standard for Dr. Lopez and/or Rumy to follow throughout their work. Explain your reasoning. (Example: date format should be MM/DD/YY because each is a U.S. based study)
- For each dataset, what do you think is the most important processing concern to address first ? Why?

Step 4. Join the discussion! Post your answers and comment on at least one other participant's answers.

Activity 3: The Bukavu Series Deep Dive

Take a moment to review the comics and their explanations from The Bukavu Series

Choose one comic to focus on and answer the following questions (you are also welcome to choose more than one):

- What is the title of the comic you chose? (Example: "Undermining female researchers," "Traumatic experiences in the field," "Power imbalances in academic writing")
- The Bukavu Series is not specifically focused on gender data projects. Therefore, recall Week 1's activity, Data Lifecycle Discussion, what consideration(s) did this comic bring up for the gender data lifecycle specifically?
- Do you have additional questions or comments about the Bukavu Series?

Glossary & Resources

- Collection is the process of gathering and measuring information in a systematic fashion.
- Biases are personal and/or systematic judgments and pre-conceptions.
- Selection Bias, also known as sampling bias, is choosing a sample that is not representative of the population being studied. Selection bias also results when participants not being randomly sampled to a condition.
- Response Bias is a general term for any external factor that can influence a respondent's answers.
- Non-response bias means certain members of the population being studied do not participate in the data capture. There are many reasons for non-responses bias. Examples include the medium of study (e.g. online survey) or time study is conducted (e.g. morning times).
- Demand characteristics are the ways participants of a study can influence the outcome.
- Social Desirability Response Bias is a type of response bias is when a respondent is influenced to provide answers that are socially desirable and are not necessarily their true responses.
- Question Wording/Ordering Bias occurs if the question is worded or ordered in such a way that it favors one response over another.
- Outlier/Extreme Response Bias occurs when there are extreme data values.
- Processing is the series of operations needed to prepare data for analysis.
- Raw data (also called dirty or unclean data) is data directly from the source.
- Incomplete data is data that lacks the key information.
- Duplicate data is data that repeats itself.
- Inaccurate data is information that is not correct.
- Inconsistent data is unstandardized data.

- Clean data is data that is accurate and formatted to a set of rules. These rules can be determined by your organization or by your data analysis needs. Data cleaning helps edit or remove incomplete, duplicate, incorrect, and inconsistent data.
- Data aggregation is systematically combining data from multiple sources together. Processing may or may not involve aggregation.
- Data security is the commitment to confidentiality and integrity of data.
- Personally Identifiable Information also known as PII is any data that can be used to distinguish or trace an individual's identity.
- Data encryption is the process of converting data into code to prevent unauthorized access.