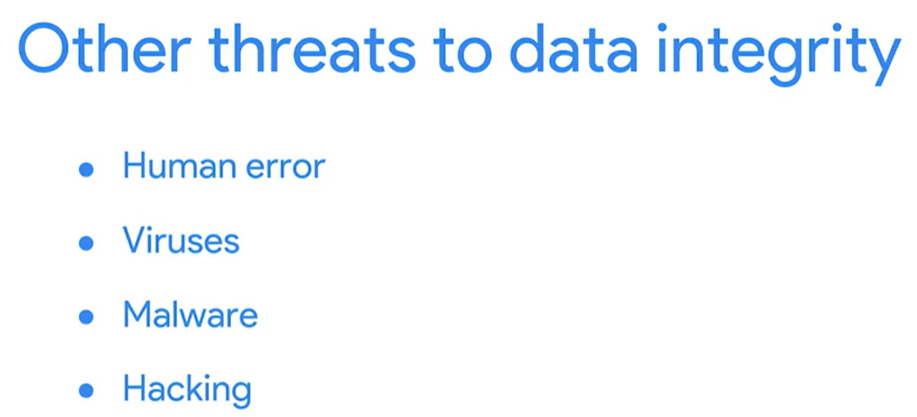
**Tính toàn vẹn dữ liệu (intergrity)**

**Data integrity** is the accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle.

**Data replication** is the process of storing data in multiple locations.

**Data transfer** is the process of copying data from a storage device to memory, or from one computer to another.

**The data manipulation** process involves changing the data to make it more organized and easier to read. Data manipulation is meant to make the data analysis process more efficient, but an error during the process can compromise the efficiency



**Reference: Data constraints and examples**

As you progress in your data journey, you'll come across many types of data constraints (or criteria that determine validity). The table below offers definitions and examples of data constraint terms you might come across.

| **Data constraint** | **Definition** | **Examples** |
| --- | --- | --- |
| **Data type** | Values must be of a certain type: date, number, percentage, Boolean, etc. | If the data type is a date, a single number like 30 would fail the constraint and be invalid |
| **Data range** | Values must fall between predefined maximum and minimum values | If the data range is 10-20, a value of 30 would fail the constraint and be invalid |
| **Mandatory** | Values can’t be left blank or empty | If age is mandatory, that value must be filled in |
| **Unique** | Values can’t have a duplicate | Two people can’t have the same mobile phone number within the same service area |
| **Regular expression (regex) patterns** | Values must match a prescribed pattern | A phone number must match ###-###-#### (no other characters allowed) |
| **Cross-field validation** | Certain conditions for multiple fields must be satisfied | Values are percentages and values from multiple fields must add up to 100% |
| **Primary-key** | (Databases only) value must be unique per column | A database table can’t have two rows with the same primary key value. A primary key is an identifier in a database that references a column in which each value is unique. More information about primary and foreign keys is provided later in the program. |
| **Set-membership** | (Databases only) values for a column must come from a set of discrete values | Value for a column must be set to Yes, No, or Not Applicable |
| **Foreign-key** | (Databases only) values for a column must be unique values coming from a column in another table | In a U.S. taxpayer database, the State column must be a valid state or territory with the set of acceptable values defined in a separate States table |
| **Accuracy** | The degree to which the data conforms to the actual entity being measured or described | If values for zip codes are validated by street location, the accuracy of the data goes up. |
| **Completeness** | The degree to which the data contains all desired components or measures | If data for personal profiles required hair and eye color, and both are collected, the data is complete. |
| **Consistency** | The degree to which the data is repeatable from different points of entry or collection | If a customer has the same address in the sales and repair databases, the data is consistent. |

**Clean data + alignment to business objective = accurate conclusions** VLOOKUP + DATEIF

**Alignment to business objective + additional data cleaning = accurate conclusions**

**Alignment to business objective + newly discovered variables + constraints = accurate conclusions**

If you have insufficient data, you can identify trends with the data that is available and qualify your findings accordingly.

**When you find an issue with your data**

When you are getting ready for data analysis, you might realize you don’t have the data you need or you don’t have enough of it. In some cases, you can use what is known as proxy data in place of the real data. Think of it like substituting oil for butter in a recipe when you don’t have butter. In other cases, there is no reasonable substitute and your only option is to collect more data.

Consider the following data issues and suggestions on how to work around them.

**Data issue 1: no data**

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the analysis after you have collected more data. | If you are surveying employees about what they think about a new performance and bonus plan, use a sample for a preliminary analysis. Then, ask for another 3 weeks to collect the data from all employees. |
| If there isn’t time to collect data, perform the analysis using proxy data from other datasets.  *This is the most common workaround.* | If you are analyzing peak travel times for commuters but don’t have the data for a particular city, use the data from another city with a similar size and demographic. |

**Data issue 2: too little data**

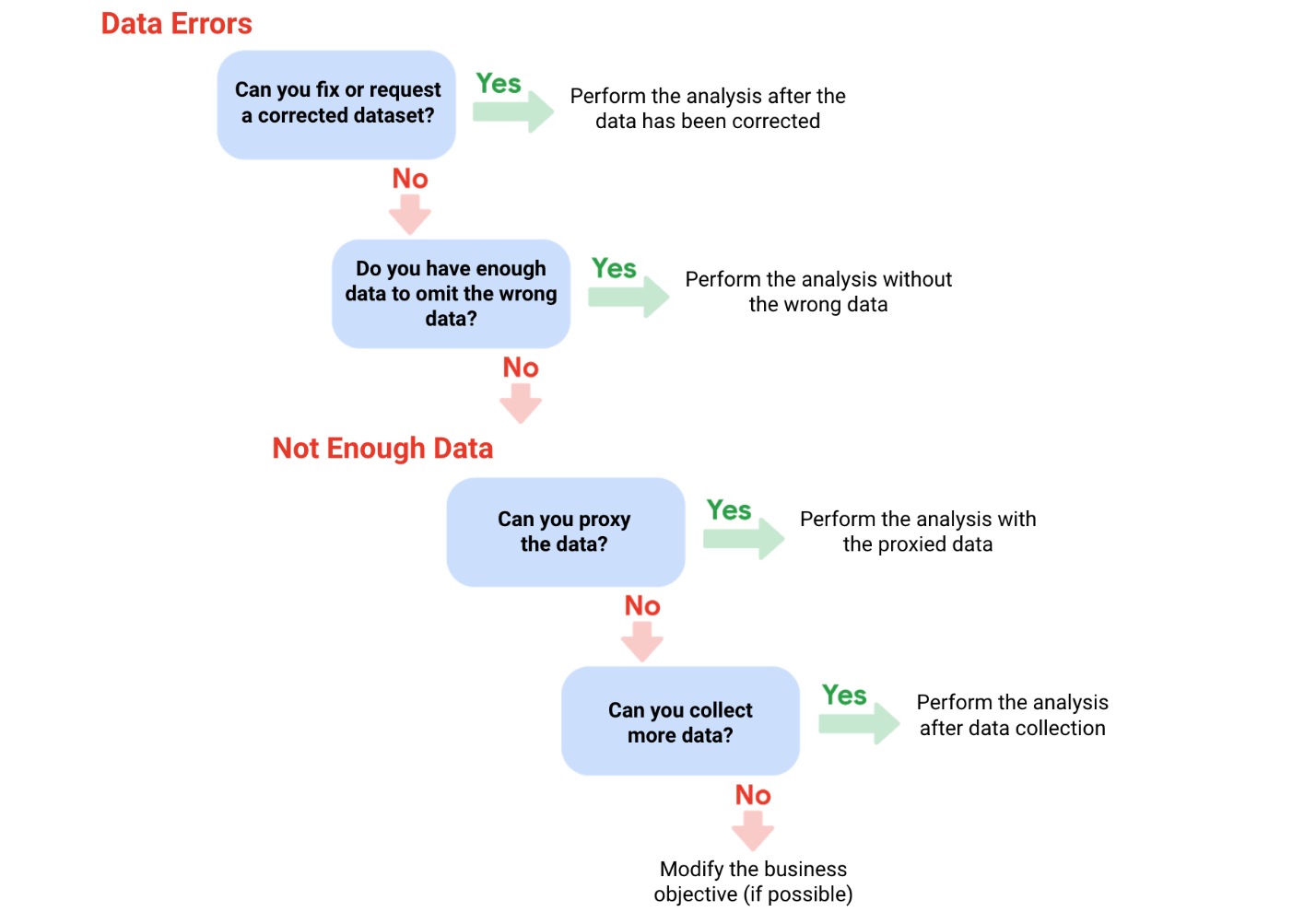
| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Do the analysis using proxy data along with actual data. | If you are analyzing trends for owners of golden retrievers, make your dataset larger by including the data from owners of labradors. |
| Adjust your analysis to align with the data you already have. | If you are missing data for 18- to 24-year-olds, do the analysis but note the following limitation in your report: *this conclusion applies to adults 25 years and older* *only*. |

**Data issue 3: wrong data, including data with errors\***

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| If you have the wrong data because requirements were misunderstood, communicate the requirements again. | If you need the data for female voters and received the data for male voters, restate your needs. |
| Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors. | If your data is in a spreadsheet and there is a conditional statement or boolean causing calculations to be wrong, change the conditional statement instead of just fixing the calculated values. |
| If you can’t correct data errors yourself, you can ignore the wrong data and go ahead with the analysis if your sample size is still large enough and ignoring the data won’t cause systematic bias. | If your dataset was translated from a different language and some of the translations don’t make sense, ignore the data with bad translation and go ahead with the analysis of the other data. |

***\* Important note:*** *Sometimes data with errors can be a warning sign that the data isn’t reliable. Use your best judgment.*

**Use the following decision tree as a reminder of how to deal with data errors or not enough data:**



A white text with black text

Description automatically generated

**Glossary terms from module 1**

**Terms and definitions for Course 4, Module 1**

**Accuracy:** The degree to which the data conforms to the actual entity being measured or described

**Completeness:** The degree to which the data contains all desired components or measures

**Confidence interval:**  A range of values that conveys how likely a statistical estimate reflects the population

**Confidence level:** The probability that a sample size accurately reflects the greater population

**Consistency:** The degree to which data is repeatable from different points of entry or collection

**Cross-field validation:** A process that ensures certain conditions for multiple data fields are satisfied

**Data constraints:** The criteria that determine whether a piece of a data is clean and valid

**Data integrity:** The accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle

**Data manipulation:** The process of changing data to make it more organized and easier to read

**Data range:** Numerical values that fall between predefined maximum and minimum values

**Data replication:** The process of storing data in multiple locations

**DATEDIF:** A spreadsheet function that calculates the number of days, months, or years between two dates

**Estimated response rate**: The average number of people who typically complete a survey

**Hypothesis testing:** A process to determine if a survey or experiment has meaningful results

**Mandatory:** A data value that cannot be left blank or empty

**Margin of error**: The maximum amount that the sample results are expected to differ from those of the actual population

**Random sampling:** A way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen

**Regular expression (RegEx):** A rule that says the values in a table must match a prescribed pattern