

First, **questions** need to be asked.

Sometimes, a dataset may be given first. Based on that dataset and the data in it, some questions can be formed around the data set. Other times, questions come first, those questions determine what data should be used.

Some good questions can be:

* What am I trying to find?
* Is there a problem I’m trying to solve?
* What are the characteristics of students passing their projects?
* How can I better stock my store with products people want to buy?

In the real world, data that is dealt with is usually data that is contained in multiple varying datasets. The right questions help focus on the relevant parts of the data and can provide meaningful insights.

After the questions have been formed, the data must then be **wrangled**. Which means making sure the data that is needed is there and in great quality.

There are three parts to wrangling:

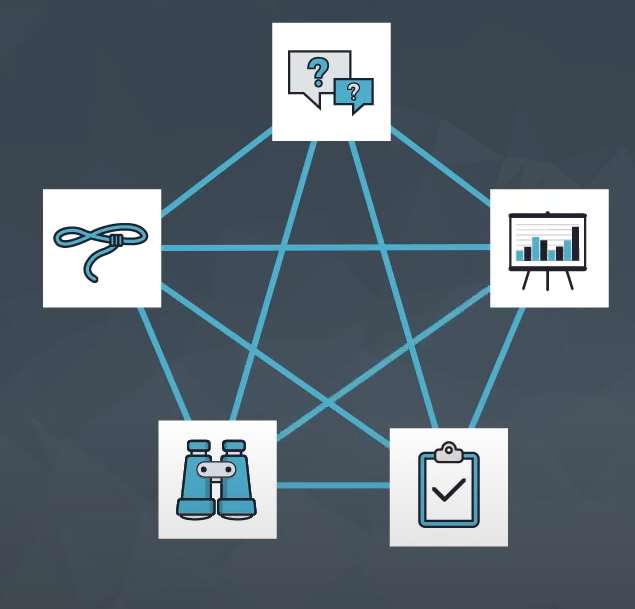
1. **Gathering Data:** If the data is already given, then all that needs to be done is to open it. If data has not been provided, it’s important to think about what data is needed, that answers the previously asked questions, then collect the data from all the sources that are available.
2. **Assessing the data:** It’s important to assess the data to identify any problems in quality or structure.
3. **Clean the data:** Involves modifying, replacing or removing data to ensure the dataset is as high quality and structured as possible.

Wrangling is all about getting the data needed in a form that can be worked with. Sometimes it can be easy other times it can be difficult. Hopefully after wrangling, the data is clean and can be easily worked with.

After wrangling, the data should be **explored**. **Exploratory Data Analysis (EDA)** is the process of exploring and augmenting the data to maximize the potential in analysis, visualization and models. It involves finding patterns in the data, visualizing the relationship in the data and building intuition about the overall data. After exploring, certain operations can be done like removing outliers or creating new features based on other data in the dataset, the latter is called feature engineering. Engineering and modifying the data can often improve the quality of the analysis.

Sometimes, during the EDA process, it might be necessary to go back into the first few steps, like if during exploration, issues have been found with the data. EDA might be where the most about the data is learned (in most cases).

Data analysis is not necessarily linear, each process can be intertwined with the other processes.



After exploring the data, **conclusions can be drawn** or **even predictions can be made**. For example, predicting which students will fail a project and reaching out to those students, or predicting which products are more likely to sell and stocking up on those products. This step is usually approached with machine learning or inferential statistics.

After drawing conclusions (or making predictions), the results should be **communicated**. The analysis is only as valuable as its ability to be communicated. The analysis should be justified and its meaning convened in the insights found. If the goal is to build a system based on the analysis, then what is built, the process of reaching design decisions and reporting on how well it performs need to be shared appropriately.

There are many ways to communicate results:

* Reports
* Slides
* Blog posts
* Emails
* Presentations
* Conversations

Regardless of the method of communication, visualization is always powerful.