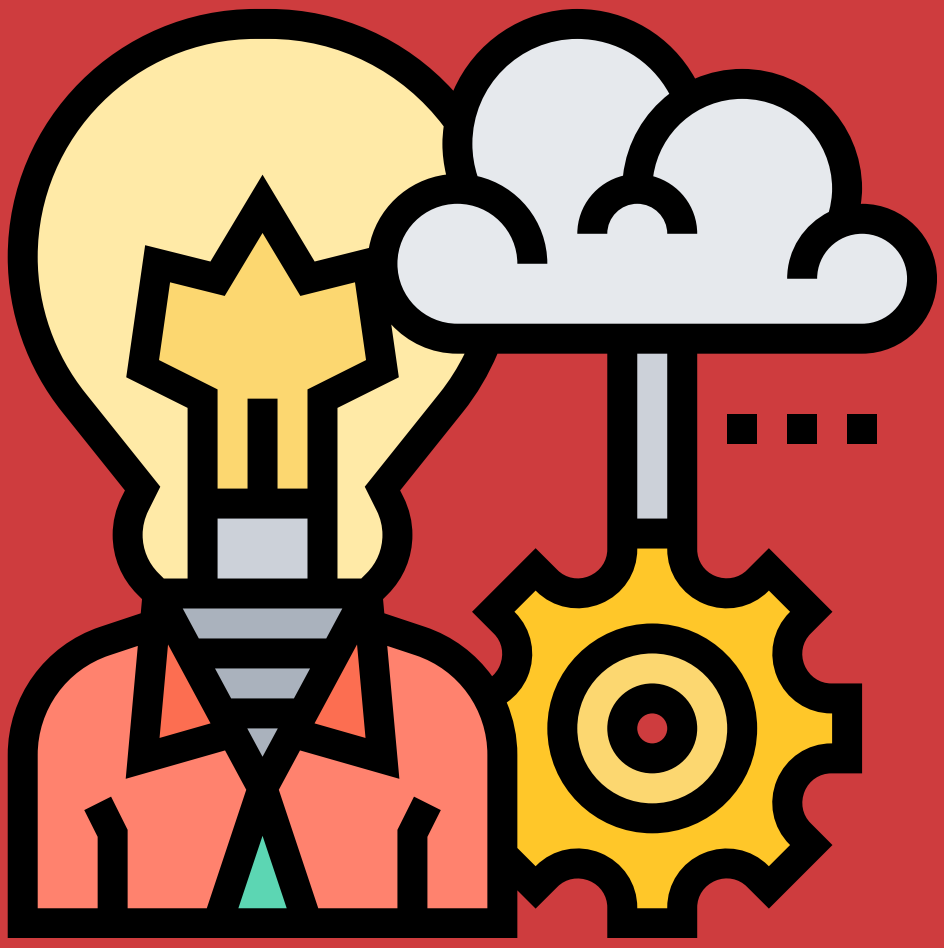
A BRIEF GUIDE TO
DATA UNDERSTANDING



CURATED BY HARPREET SAHOTA

THE ARTISTS OF DATA SCIENCE

WHAT IS DATA UNDERSTANDING?

Data understanding is the process of determing what is in your data, in an effort to more fully understand the conditions under which the problem happens and to explore potential theories using the data.

Data Understanding is focued on the following key processes:

- Data identification and prioritization
- Data collection and preparation
- Data profiling and characterization

WHY DO IT?

We perform data understanding...to understand.

We are trying to make sense of the world by exposing our theories to the realities of our data.

In essence, we use data along with our perceptions, intuition and experience to help us make data-driven decisions.

But, in order to support a data-driven, fact-based problem-solving approach with analytics, we will need to accurately define the problem in its simplest terms.

By doing this, we can frame the problem as a question that can be answered with data.

Once we understand the problem, we use data understanding to begin to more fully explore the theories outlined during our problem framing exercise.

This is where we begin our data exploration in earnest, by linking the problem and the processes that we worked to understand with the reality of the data we see as a result of some real-world processes.

DATA IDENTIFICATION AND PRIORITIZATION

01

First, **articulate the data needed** to help solve the problem. One of the traps that I see data scientists often fall into is thinking they need some specific data—spending countless hours going after it and transforming it into something that can be used, only to find out that it had marginal value.

02

Second, you need to **confirm that the data truly reflects the realities** of the data generating process you are interested in.

This is necessary so that you can make certain that what you think has happened jives with what really happened and what you see in the data.

03

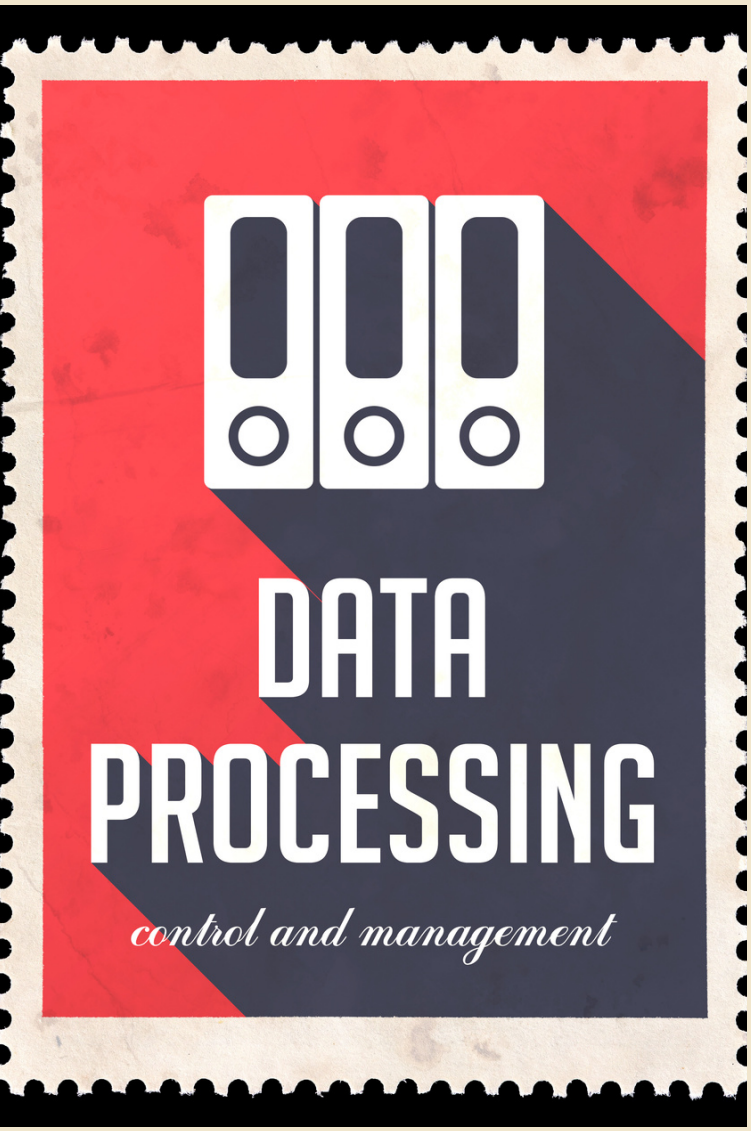
Finally, it is important to verify the veracity of the data to p**revent issues with interpretability**, believability, and trust in your findings.

HOW DO I KNOW WHAT DATA I NEED?

The choice of data will ultimately depend on the methods employed in your analysis. However, at this stage, remember that you are still trying to make sense of the data through exploration. *You are trying to understand at this point, not test your theories, so there is no one right answer.* The goal is to narrow down the things that you don't yet know, including your "known unknowns" and "unknown unknowns"—or as is commonly called, *unconscious ignorance*.

Ask yourself the following questions: What data do I think I need?, How will this be useful to my exploration of the phenomenon?, What assumptions must be true for this to work? On a scale of 1 to 10, how important is getting this data to my success? How feasible is getting this data?

STRUCTURING DATA FOR EXPLORATION



A common question is, "How should I structure my data for analysis and exploration?"

Unfortunately, **there is no one perfect answer**; it depends on several factors, including what tool you will use for data exploration, what questions you want to ask of the data, and what inferences you wish to make.

Most often, you will want as much detail in the data and at the lowest level possible (individual records), as *it is much easier to summarize than it is to disaggregate*.

There are several potential data structures, and each have their own benefits and drawbacks based on the types of analysis that can be performed on this type of data structure.

The process of structuring data for analysis is referred to as **data modeling**, which is a critical competency for those who want to be data scientists or data analysts.

At this point you may need to think through how you want to structure the data.

Data stored in long, skinny tables may be appropriate for time-series or event data such as web log data or laboratory values for patients whereas other types of data may be stored in wide tables.

Consider what questions you have and whether we are seeking to understand:

- Differences between groups,
- Associations between variables,
- Predictions of future events,
- Distribution of a single sample,
- Time series, etc.

DATA PROFILING

PART OF THE DATA SENSEMAKING BEST PRACTICE IS TO ACTUALLY EXPLORE THE DATA. THE DECISIONS MADE IN EXPLORING, INTERPRETING, AND EXPLAINING DATA SPEAKS TO THE TRUE ART OF ANALYSIS. DATA PROFILING IS PERHAPS THE MOST FUN PART OF ANALYTICS—PLAYING WITH DATA!

But a plan of attack is needed. It is often useful to both new analysts as well as experienced data scientists to have a set of questions when beginning to explore the data. As you will recall, during the process of data identification and prioritization, we did just that. It's time to pull those questions back out and use that to remind yourselves why you thought the data was a good idea in the first place. Data understanidng gives you the space to do exactly that.

One of the first things I like to do when exploring a new dataset is to get to know its personality, sometimes called the **culture** or **gestalt** of a dataset.

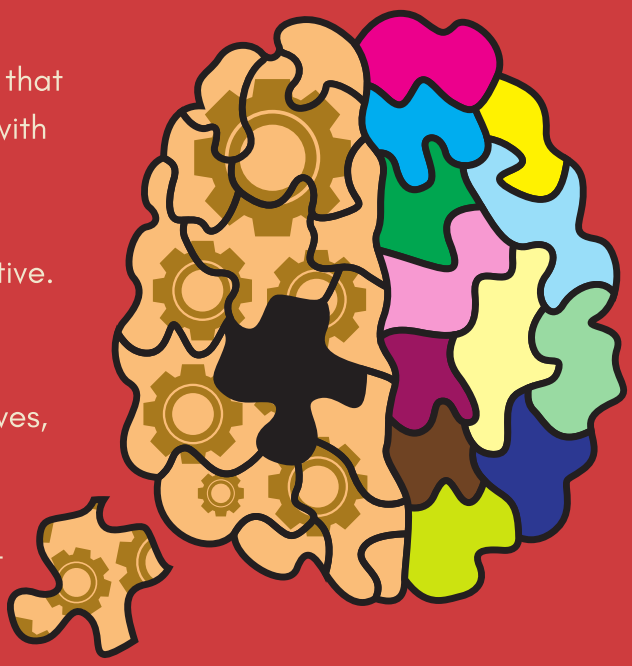
What does it look like? Is it tall and lanky, short and chubby?

Similarly, when we talk about a dataset, we might say that it is "*left-brained*" (i.e., full of big numbers or precise with lots of numbers to the right of the decimal).

The data might be more logical, analytical, and objective.

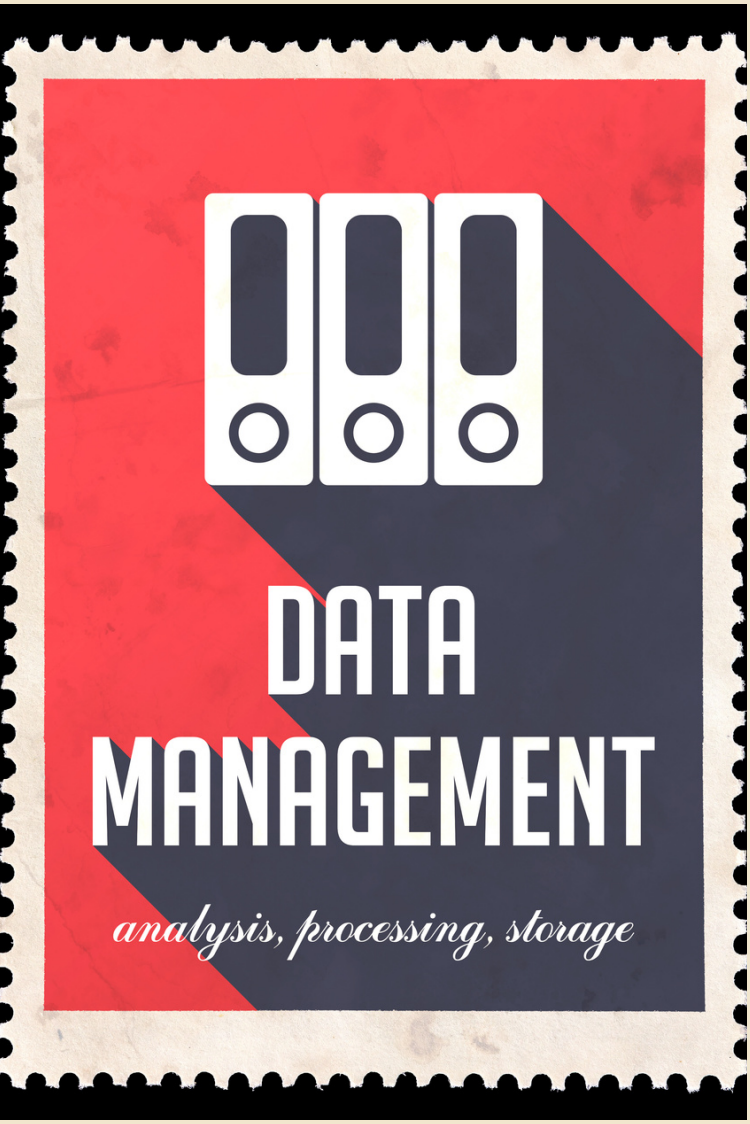
However, a dataset that is "*right-brained*" might have unstructured data like categorical values, long narratives, or open-ended responses to survey questions.

Just as we say about "right-brained" people, we might consider this to be more intuitive, thoughtful, and subjective.



Often, this data is left on the cutting room floor in our analytic process, but as tools for working with unstructured data become commonplace, we see increasing interest in these types of data.

DATA QUALITY



Data quality issues are all around us, but sometimes it is hard to truly understand the importance of data quality for data science.

Viewing data quality through two lenses:

- **Technical quality**– the data fails the basic tests for valid values or has missing data. Specifically, we look for:
  - *Invalid data*—Are there incorrect values or overloaded fields?
  - *Missing data*—Is there an absence of content or of valid content?
- **Business quality**– often more critical and much harder to evaluate; some things to consider to determine data quality from a business perspective include:
  - *Relevance*—Is it meaningful?
  - *Accuracy*—Is the information accurate, or are errors introduced by "fat fingers" or laziness (duplicate records)?
  - *Consistency*—Do we see different business rules being applied?
  - *Timeliness*—Can we access it in the time frame that is useful?
  - *Comparable*—Can we compare values from different sources?
  - *Completeness*—Does the data tell the whole story?

DATA QUALITY

WHAT TO LOOK FOR

IN TERMS OF DATA QUALITY, THE THINGS YOU MIGHT TYPICALLY LOOK FOR INCLUDE:

ACCURACY

Can you determine from the dataset whether the data was recorded correctly? Are there anomalies or nonsensical values? Examples might include a “pregnant male” or values that are out of the expected range. Do you see any typos, multiple formats representing the same data, missing/default values?

COMPLETENESS

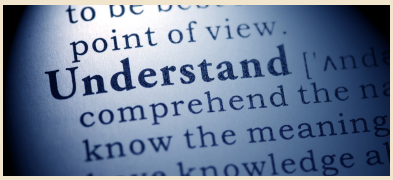
Can you determine whether all relevant data was recorded? What else would have been useful to tell the complete story that you are trying to convey? Are there duplicate values?

TIMELINESS

Do any measures of time present challenges or illustrate significant gaps in time? For example, lab results that don’t seem to relate to an encounter?

KEY DELIVERABLES

WHAT SHOULD YOU ACCOMPLISH AT THE END OF THIS PROCESS?

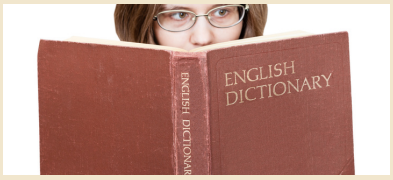


UNDERSTANDING OF THE DATA

You should have a thorough understanding of what values each column contain, what range the take on, the cardinality, the number of unique values, missing values, etc

DATA PROFILE REPORT

Use a tool like pandas profiling to make it easier to do this and save the output in the report section of your repository.



DATA DICTIONARY

Create a table with each row being a variable name, and columns such as description, percent missing, number of unique values, range of values, etc. Something that will give anyone reviewing your project enough information to get a sense of the data. Use the profile report to summarize what you see

CLEAN DATASET

At the end of this process you should have a cleaned dataset that is ready for exploratory data analysis, feature engineering, and further downstream analysis.

SUMMARY

Data understanding includes a number of processes that help us identify and prioritize data, acquire and transform the data, and understand its content.

Let's pause and summarize what we have learned so far:

- Profiling a dataset helps you understand its personality and whether this data is going to be useful for us in solving your problem.
- The type of variable determines what methods you can use to profile the data; for example, you cannot calculate measures such as mean or standard deviation on a categorical variable.

By profiling and characterizing data, you can assess their suitability for helping to provide clarity in solving a problem or understanding a phenomenon.

The activities included in the data understanding best include:

- Inspecting data to understand what we have and how it is organized
- Evaluating the current quality of the data and identify possible issues
- Determining the overall utility of the data for addressing the problem
- Assessing other data that may be useful when combined

Questions that you should be asking yourself at this point include:

- What have I learned?
- Does this help me understand the phenomenon better than before?
- Do my theories about what might be impacting, influencing, or causing my problem have more clarity?