**COURSE 3**

**PREPARE, DATA FOR EXPLORATION**

**MODULE 1 - DATA TYPES AND STRUCTURES**

**DATA EXPLORATION**

Picture this: You're working on a project. You've **asked all the right questions**, **applied structured thinking**, and you're **completely in sync with your stakeholders**.

You're off to a great start. But **there's another step in the process: preparing the data correctly**.

**This is where understanding the different types of data and data structures comes in**. Knowing this **lets you figure out what type of data is right for the question you're answering**. Plus, you'll gain practical skills about **how to extract, use, organize, and protect your data**.

**You've learned how to tell a story using data.**

Now **we'll learn more about the data that you'll need to tell the best story possible**.

Here's a real life example of how I've used data to tell a story. In my job, we analyze Medicare enrollment data over time and make connections to how people research Medicare plans on Google. As people 65 and older become more informed decision makers for their health, I use the data to learn if there's an increase in Medicare enrollments and what part Google searches play if there is an increase in demand. Now it's very important that I make sure the data is relevant and valid. I also have to pay attention to questions around access and equity while maintaining the privacy of those conducting searches. The happy ending of my story is that the data in my findings is useful to medical professionals and their patients.

There's so much useful data out there, and you're building the skills you'll need to find and use the right data in the best way. In this course, you'll continue sharpening those skills. So you've already heard a lot about the **data analysis process steps: Ask, Prepare, Process, Analyze, Share and Act.** Now it's **time to learn how to prepare the data**.

You'll learn to **identify how data is generated and collected**, and you'll **explore different formats**, **types and structures of data**. We'll make sure you **know how to choose and use data** that'll help you understand and respond to a business problem.

And because **not all data fits each need**, you'll **learn how to analyze data for bias and credibility**. We'll also explore what clean data means. But wait, there's more. You'll also get up close and personal with databases. We'll cover what they are and how analysts use them. You'll even get to extract your own data from a database using a couple of tools that you're already familiar with: spreadsheets and SQL.

The key here is patience. Like anything worth doing, this will take time and practice. And I'll be with you every step of the way. Still with me? Great. The last few things we'll cover are the basics of data organization and the process of protecting your data. **Data works best when it's organized**. And if you're organizing your data, you'll want to protect it too. I'll show you how to do both and apply it to your own analysis. I'm so excited to help you write your own personal story as you continue exploring the world of data analytics. So let's do it.

**COLLECT DATA**

[**DATA COLLECTION IN OUR WORLD**](https://www.coursera.org/learn/data-preparation/lecture/QCPVt/data-collection-in-our-world)

Every minute of every day millions of texts and hundreds of millions of emails are sent. On top of that, millions of online searches are made and videos viewed and those numbers are only growing. **That's a lot of data**.

We'll talk about the ways that data can be generated and how industries collect data themselves. **Every piece of information is data.** All that data is usually generated as a result of our activity in the world.

With social media and mobile devices, millions and millions of people are adding to the huge amount of data out there, each and every day. Think about it like this. Every digital photo online is one piece of data. Every photo itself holds even more data, from the number of pixels to the colors contained in each of those pixels. But that's not the only way data is made. **We can also generate data by collecting information.**

This **data generation and collection** comes with a few more things to think about. It **needs to be done with consideration to ethics so that we maintain people's rights and privacy**.

**Real world example:**

The United States Census Bureau uses forms to collect data about the country's population. This data is used for a number of reasons, like funding for schools, hospitals, and fire departments. The Bureau also collects information about things like U.S. businesses, creating their own data in the process. The great thing about this is that others can then use the data for their own needs, including analysis. The annual business survey is used to figure out the needs of businesses and how to provide them with resources to help them succeed.

**Survey data is just one example**. There's all kinds of data being generated all the time, and there's lots of different ways to collect it. Even something as simple as an interview can help someone collect data. Imagine you're in a job interview. To impress the hiring manager, you want to share information about yourself. The hiring manager collects that data and analyzes it to help them decide whether to hire you or not. But it goes both ways. You could also collect your own data about the company to help you decide if the company is a good fit for you. Or you can use the data you collect to come up with thoughtful questions to ask the interviewer.

**Scientists also generate data.** They use a lot of observations in their work. For example, they might collect data by studying animal behavior or looking at bacteria under a microscope.**Forms, questionnaires and surveys** are commonly used **ways to collect and generate data**.

One thing to note: data that's generated online doesn't always happen directly. Have you ever wondered why some online ads seem to make really accurate suggestions or how some websites remember your preferences? This is done using cookies, which are small files stored on computers that contain information about users. Cookies can help inform advertisers about your personal interests and habits based on your online surfing, without personally identifying you. As a real world analyst, you'll have all kinds of data right at your fingertips and lots of it too. **Knowing how it's been generated can help add context to the data**, and **knowing how to collect it can make the data analysis process more efficient**.

[**DETERMINE WHAT DATA TO COLLECT**](https://www.coursera.org/learn/data-preparation/lecture/XWEc0/determine-what-data-to-collect)

As a **data analyst**, you'll need to **decide what kind of data to collect and use for every project.** With a nearly endless amount of data out there, this can be quite a bit of a data dilemma, but there's good news. You'll learn which **factors to consider when collecting data**. Usually, you'll have a head start in figuring out the **right data for the job**, because the data you need will be given to you, or your **business task or problem will narrow down your choices**.

Let's start with a question like, what's causing increased rush hour traffic in your city? First, you need to **know how the data will be collected**. You might use observations of traffic patterns to count the number of cars on city streets during particular times. You notice that cars are getting backed up on a specific street. That brings us to **data sources.**

In our traffic example, your observations would be **first-party data**. This **is data collected by an individual or group using their own resources**. Collecting first-party data is **typically the preferred method because you know exactly where it came from**.

You might also have **second-party data**, which is **data collected by a group directly from its audience and then sold**.

In our example, **if you aren't able to collect your own data**, you might **buy it from an organization that's led traffic pattern studies** in your city. **This data** didn't start with you, but **it's still reliable because it came from a source that has experience with traffic analysis.** **The same can't always be said about third-party data or data collected from outside sources who did not collect it directly**. This data might have come from a number of different sources before you investigated it. It might not be as reliable, but that doesn't mean it can't be useful. You'll just want to make sure you check it for accuracy, bias, and credibility.

Actually, **no matter what kind of data you use**, **it needs to be inspected for accuracy and trustworthiness**.

**Remember that the data you choose should apply to your needs, and it must be approved for use.** As a data analyst, it's your job to decide what data to use, and that means choosing the data that can help you find answers and solve problems and not getting distracted by other data. In our traffic example, financial data probably wouldn't be that helpful, but existing data about high volume traffic times would be.

Let's talk about **how much data to collect**. In data analytics, **a population** refers to **all possible data values in a certain data set**. If you're analyzing data about car traffic in a city, your population would be all the cars in that area. But **collecting data from the entire population can be pretty challenging**. **That's why a sample can be useful.** A sample is a **part of a population that is representative of the population**.

You might collect a data sample about one spot in the city and analyze the traffic there, or you might pull a random sample from all existing data in the population. **How you choose your sample will depend on your project.**

As you collect data, **you'll also want to make sure you select the right data type**. For traffic data, an appropriate data type could be the dates of traffic records stored in a date format. The dates could help you figure what days of the week there is likely to be a high volume of traffic in the future.

Finally, you need to **determine the time frame for data collection**. In our example, if you needed an answer immediately, you'd have to use **historical data, which is data that already exists**. But let's say you needed to track traffic patterns over a long period of time. That might affect the other decisions you make during data collection. Now you know more about the different data collection considerations you'll use as a data analyst. Because of that, **you'll be able to find the right data when you start collecting it yourself**.

[**SELECT THE RIGHT DATA**](https://www.coursera.org/learn/data-preparation/supplement/7iFqv/select-the-right-data)

Following are some data-collection considerations to keep in mind for your analysis:

## **How the data will be collected**

Decide if you will collect the data using your own resources or receive (and possibly purchase it) from another party. Data that you collect yourself is called first-party data.

## **Data sources**

If you don’t collect the data using your own resources, you might get data from second-party or third-party data providers. **Second-party data** is collected directly by another group and then sold. **Third-party data** is sold by a provider that didn’t collect the data themselves. Third-party data might come from a number of different sources.

## 

## **Solving your business problem**

Datasets can show a lot of interesting information. But **be sure to choose data that can actually help solve your problem’s question**. For example, if you are **analyzing trends over time**, make sure you **use time series data** — in other words, **data that includes dates.**

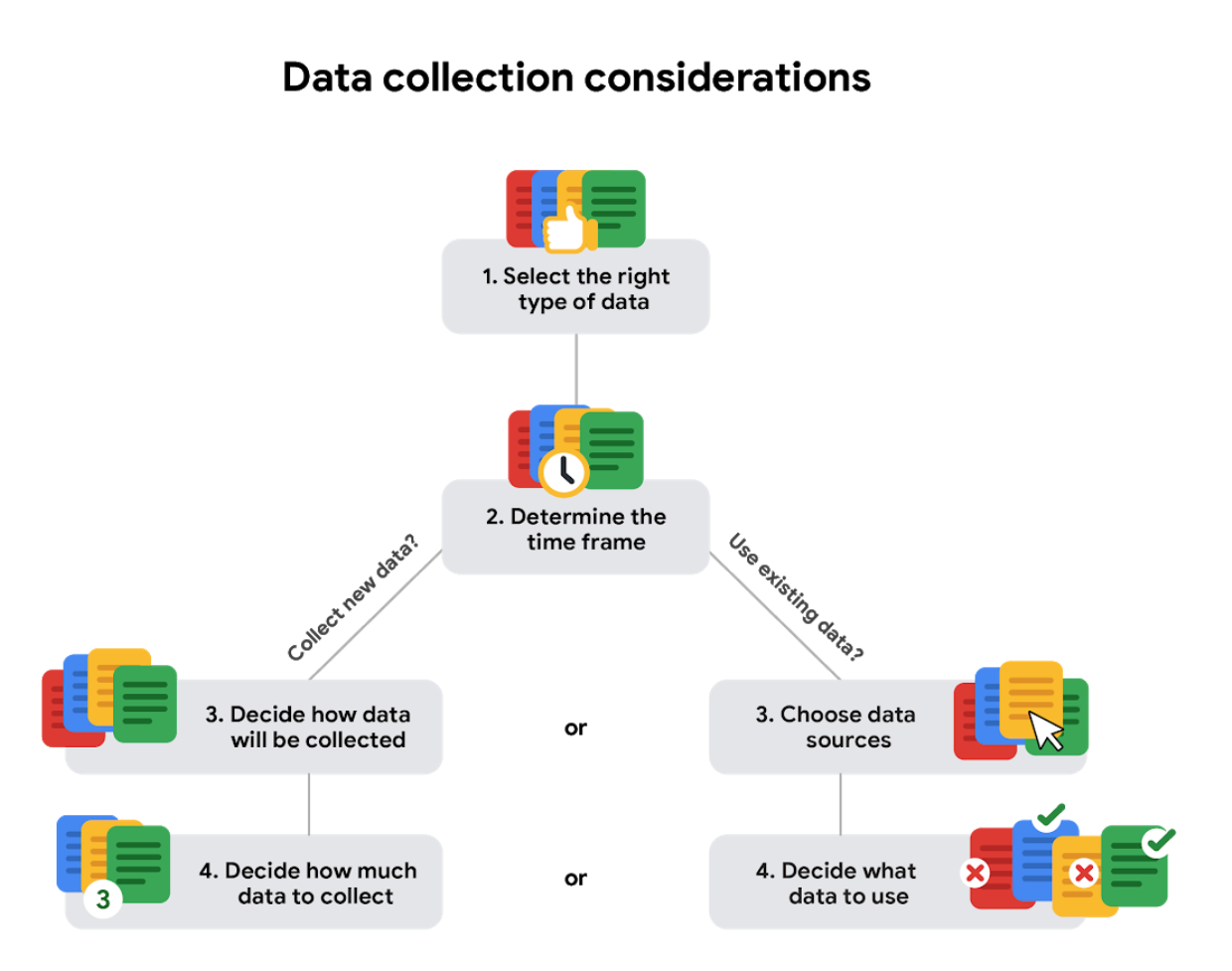
## **How much data to collect**

If you are **collecting your own data**, make **reasonable decisions about sample size**. A random sample from existing data might be fine for some projects. Other projects might need more strategic data collection to focus on certain criteria. **Each project has its own needs**.

## **Time frame**

If you are collecting your own data, decide how long you will need to collect it, especially if you are tracking trends over a long period of time. If you need an immediate answer, you might not have time to collect new data. In this case, you would need to use historical data that already exists.

Use the flowchart below if data collection relies heavily on how much time you have:



**DIFFERENTIATE DATA FORMATS AND STRUCTURES**

[**DISCOVER DATA FORMATS**](https://www.coursera.org/learn/data-preparation/lecture/8rP33/discover-data-formats)

We know we can compare different movies and movie genres. Turns out, you can do the same with data and data formats. Let's use our movie data spreadsheet to understand how that works.

We'll **start with quantitative and qualitative data**.

If we check out column A, we'll find titles of the movies. This is **qualitative data** becauseit **can't be counted, measured, or easily expressed using numbers**. Qualitative data is **usually listed as a name, category, or description**. In our spreadsheet, the movie titles and cast members are qualitative data.

**Quantitative data can be measured or counted and then expressed as a number**. This is **data with a certain quantity, amount, or range**. In our spreadsheet here, the last two columns show the movies's budget and box office revenue. The data in these columns is listed in dollars, which **can be counted**, so we know **that data is quantitative**.

We can go even deeper into **quantitative data** and **break it down into discrete or continuous data**.

**Discrete data is data that's counted and has a limited number of values**. Going back to our spreadsheet, we'll find each movie's budget and box office returns in columns M and N. These are both examples of discrete data that can be counted and have a limited number of values. For example, the amount of money a movie makes can only be represented with exactly two digits after the decimal to represent cents. There can't be anything between one and two cents.

**Continuous data can be measured using a timer**, **and its value can be shown as a decimal with several places**. Let's imagine a movie about data analysts that I'm definitely going to star in someday. You could express that movie's run time as 110.0356 minutes. You could even add fractional data after the decimal point if you needed to.

**Qualitative data** is **usually listed as a name, category, or description**. In our spreadsheet, the movie titles and cast members are qualitative data.

**There's also nominal and ordinal data:**

**Nominal data** is a **type of qualitative data** that's **categorized without a set order**. In other words, this data **doesn't have a sequence**. Here's a quick example. Let's say you're collecting data about movies. You ask people if they've watched a given movie. Their responses would be in the form of nominal data. They could respond "Yes," "No," or "Not sure." These choices don't have a particular order.

**Ordinal data**, on the other hand, is a type of **qualitative data with a set order or scale**. If you asked a group of people to rank a movie from 1 to 5, some might rank it as a 2, others a 4, and so on. These rankings are in order of how much each person liked the movie.

**Internal data** **is data that lives within a company's own systems**. For example, if a movie studio had compiled all of the data in the spreadsheet using only their own collection methods, then it would be their internal data. The great thing about **internal data** is that **it's usually more reliable and easier to collect**, but in this spreadsheet, it's more likely that the movie studio had to use data owned or shared by other studios and sources because it includes movies they didn't make. That means they'd be collecting external data.

**External data** is **data that lives and is generated outside of an organization**. External data becomes **particularly valuable when your analysis depends on as many sources as possible**. A great thing about this data is that it's structured.

**Structured data is data that's organized in a certain format**, such as **rows and columns**. **Spreadsheets** and **relational databases** are two examples of **software that can store data in a structured way**.

You might remember our earlier exploration of **structured thinking**, which **helps you add a framework to a problem so that you can solve it in an organized and logical manner**. You can **think of structured data in the same way**.

**Having a framework** **for the data** makes the **data easily searchable and more analysis-ready**. As a data analyst, **you'll work with a lot of structured data**, which will **usually** be in the form of a **table**, **spreadsheet** or **relational database**.  
  
Sometimes you'll come across **unstructured data**. This is **data that is not organized in any easily identifiable manner**. **Audio** and **video** files are **examples** of **unstructured data** because there's no clear way to identify or organize their content. Unstructured data might have internal structure, but **the data doesn't fit neatly in rows and columns like structured data**.

[**DATA FORMATS IN PRACTICE**](https://www.coursera.org/learn/data-preparation/supplement/mBSNa/data-formats-in-practice)

When you think about the word "format," a lot of things might come to mind. Think of an advertisement for your favorite store. You might find it in the form of a print ad, a billboard, or even a commercial. The information is presented in the format that works best for you to take it in. The format of a dataset is a lot like that, and **choosing the right format will help you manage and use your data in the best way possible.**

## **Data format examples**

As with most things, it is easier for definitions to click when you can pair them with examples you might encounter on a daily basis. Review each data format’s definition first and then use the examples to lock in your understanding.

## **Primary vs Secondary data**

The following table highlights the differences between primary and secondary data and presents examples of each.

| **Data format classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Primary data | Collected by a researcher from first-hand sources | * Data from an interview you conducted - Data from a survey returned from 20 participants * Data from questionnaires you got back from a group of workers |
| Secondary data | Gathered by other people or from other research | * Data you bought from a local data analytics firm’s customer profiles * Demographic data collected by a university * Census data gathered by the federal government |

## 

## 

## 

## 

## **Internal vs External data**

The following table highlights the differences between internal and external data and presents examples of each.

| **Data format classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Internal data | Data that is **stored inside** a company’s own systems | * Wages of employees across different business units tracked by HR * Sales data by store location * Product inventory levels across distribution centers |
| External data | Data that is **stored outside** of a company or organization | * National average wages for the various positions throughout your organization * Credit reports for customers of an auto dealership |

## **Qualitative vs Quantitative data**

The following table highlights the differences between qualitative and quantitative data and presents examples of each.

| **Data format classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Qualitative | A **subjective** and **explanatory** measure of a **quality** or **characteristic** | * Favorite exercise activity * Brand with best customer service * Fashion preferences of young adults |
| Quantitative | A **specific and objective measure**, such as a **number**, **quantity**, or **range** | * Percentage of board certified doctors who are women * Population size of elephants in Africa * Distance from Earth to Mars at a particular time |

## **Continuous vs Discrete data**

The following table highlights the differences between continuous and discrete data and presents examples of each.

| **Data format classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Continuous data | A **type of quantitative data** that is **measured** and can have almost **any numeric value** | * Height of kids in third grade classes (52.5 inches, 65.7 inches) * Runtime markers in a video * Temperature |
| Discrete data | A **type of quantitative data** that is **counted** and has a **limited number of values** | * Number of people who visit a hospital on a daily basis (10, 20, 200) * Maximum capacity allowed in a room * Tickets sold in the current month |

## **Nominal vs Ordinal data**

The following table highlights the differences between nominal and ordinal data and presents examples of each.

| **Data format classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Nominal | A **type of qualitative data** that is categorized **without a set order** | * First time customer, returning customer, regular customer * New job applicant, existing applicant, internal applicant * New listing, reduced price listing, foreclosure |
| Ordinal | A **type of qualitative data** **with** a **set order or scale** | * Movie ratings (number of stars: 1 star, 2 stars, 3 stars) * Ranked-choice voting selections (1st, 2nd, 3rd) * Satisfaction level measured in a survey (satisfied, neutral, dissatisfied) |

## **Structured vs Unstructured data**

The following table highlights the differences between structured and unstructured data and presents examples of each.

| **Data format classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Structured data | Data **organized in a certain format**, like **rows** and **columns** | * Expense reports * Tax returns * Store inventory |
| Unstructured data | Data that **cannot be stored as columns and rows** **in a relational database.** | * Social media posts * Emails * Videos |

[**OPTIONAL SELF-REFLECTION: UNSTRUCTURED DATA**](https://www.coursera.org/learn/data-preparation/quiz/zX8DT/optional-self-reflection-unstructured-data)

[**CONTINUE EXPLORING STRUCTURED DATA**](https://www.coursera.org/learn/data-preparation/lecture/PWTqC/continue-exploring-structured-data)

As a quick refresher, **structured data is data organized in a format like rows and columns**. But there's definitely more to it than that. Structured data works **nicely within a data model, which is a model that is used for organizing data elements and how they relate to one another**.

What are **data elements**? They're pieces of information, such as **people's names, account numbers, and addresses**.

**Data models help to keep data consistent and provide a map of how data is organized.** This **makes it easier for analysts and other stakeholders to make sense of their data and use it for business purposes**.

**In addition to working well within data models, structured data is also useful for databases.** This makes it easy for analysts to enter, query, and analyze the data whenever they need to. This also **helps make data visualization pretty easy** because **structured data can be applied directly to charts, graphs, heat maps, dashboards and most other visual representations of data**. Spreadsheets and databases that store data sets are widely used sources of structured data.

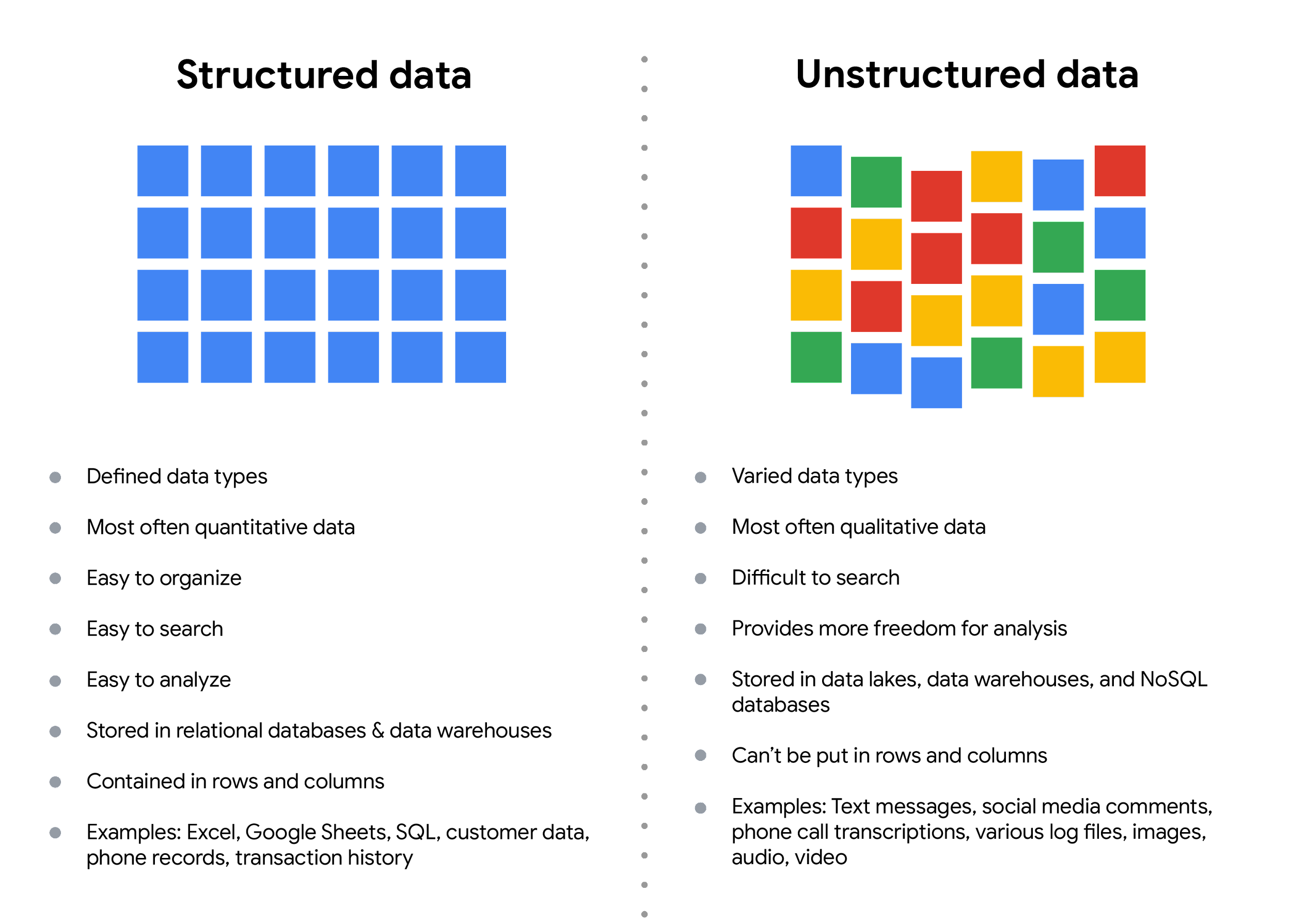
[**THE EFFECTS OF DIFFERENT STRUCTURES**](https://www.coursera.org/learn/data-preparation/supplement/tkt9D/the-effects-of-different-structures)

Data is everywhere and it can be stored in lots of ways. Two general categories of data are:

* **Structured data:** Organized in a certain format, such as rows and columns.This makes it easier to store and query for business needs. If the data is exported, the structure goes along with the data.
* **Unstructured data:** Not organized in any easy-to-identify way. And there is much more unstructured than structured data in the world. Video and audio files, text files, social media content, satellite imagery, presentations, PDF files, open-ended survey responses, and websites all qualify as types of unstructured data.

For example, when you rate your favorite restaurant online, you're creating structured data. But when you use Google Earth to check out a satellite image of a restaurant location, you're using unstructured data.

Here's a refresher on the characteristics of structured and unstructured data:

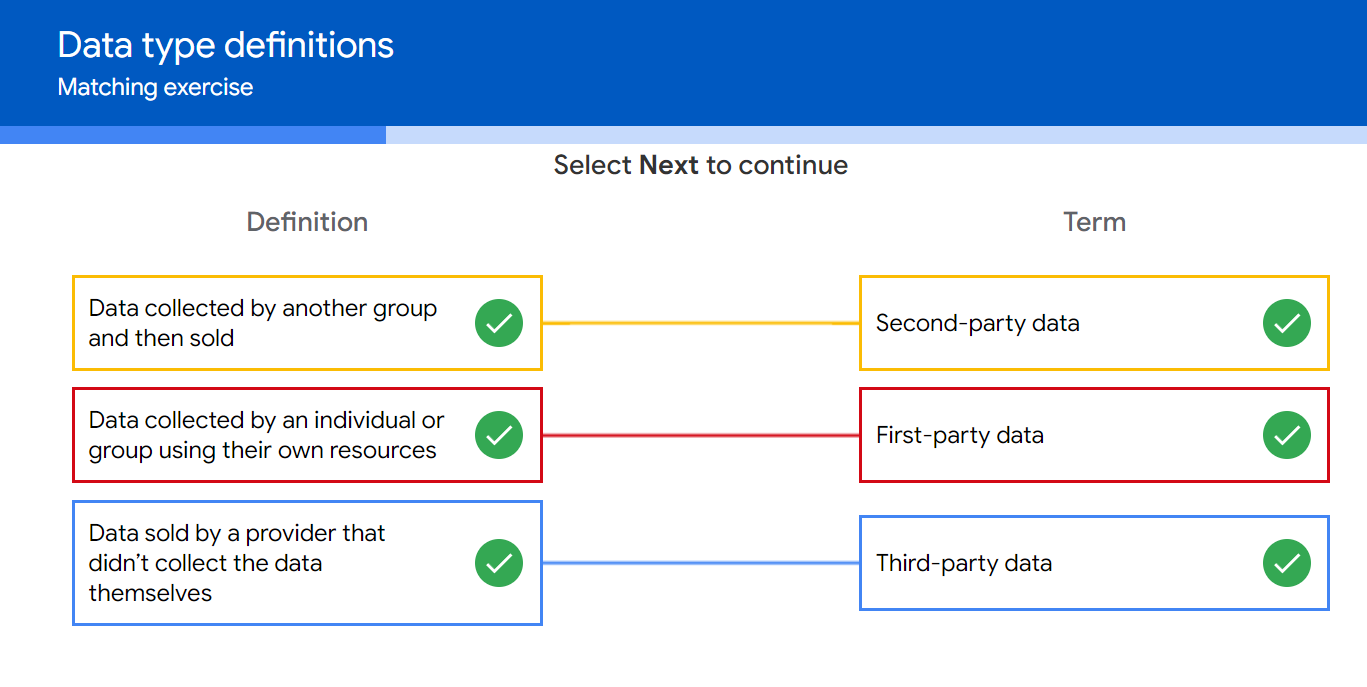


## **The fairness issue**

**The** **lack of structure makes unstructured data difficult to search, manage, and analyze**. But r**ecent advancements in artificial intelligence and machine learning algorithms are beginning to change that**.

Now, the **new challenge facing data scientists** is **making sure these tools are inclusive and unbiased**. Otherwise, certain elements of a dataset will be more heavily weighted and/or represented than others. And as you're learning, **an unfair dataset does not accurately represent the population, causing skewed outcomes, low accuracy levels, and unreliable analysis**.

[**DIFFERENTIATE DATA TYPES**](https://www.coursera.org/learn/data-preparation/ungradedWidget/uw7l4/differentiate-data-types)



[**DATA MODELING LEVELS AND TECHNIQUES**](https://www.coursera.org/learn/data-preparation/supplement/vtp7L/data-modeling-levels-and-techniques)

**Data models help keep data consistent and enable people to map out how data is organized.** A basic understanding makes it easier for analysts and other stakeholders to make sense of their data and use it in the right ways.

**Important note:** As a junior data analyst, you won't be asked to design a data model. But you might come across existing data models your organization already has in place.

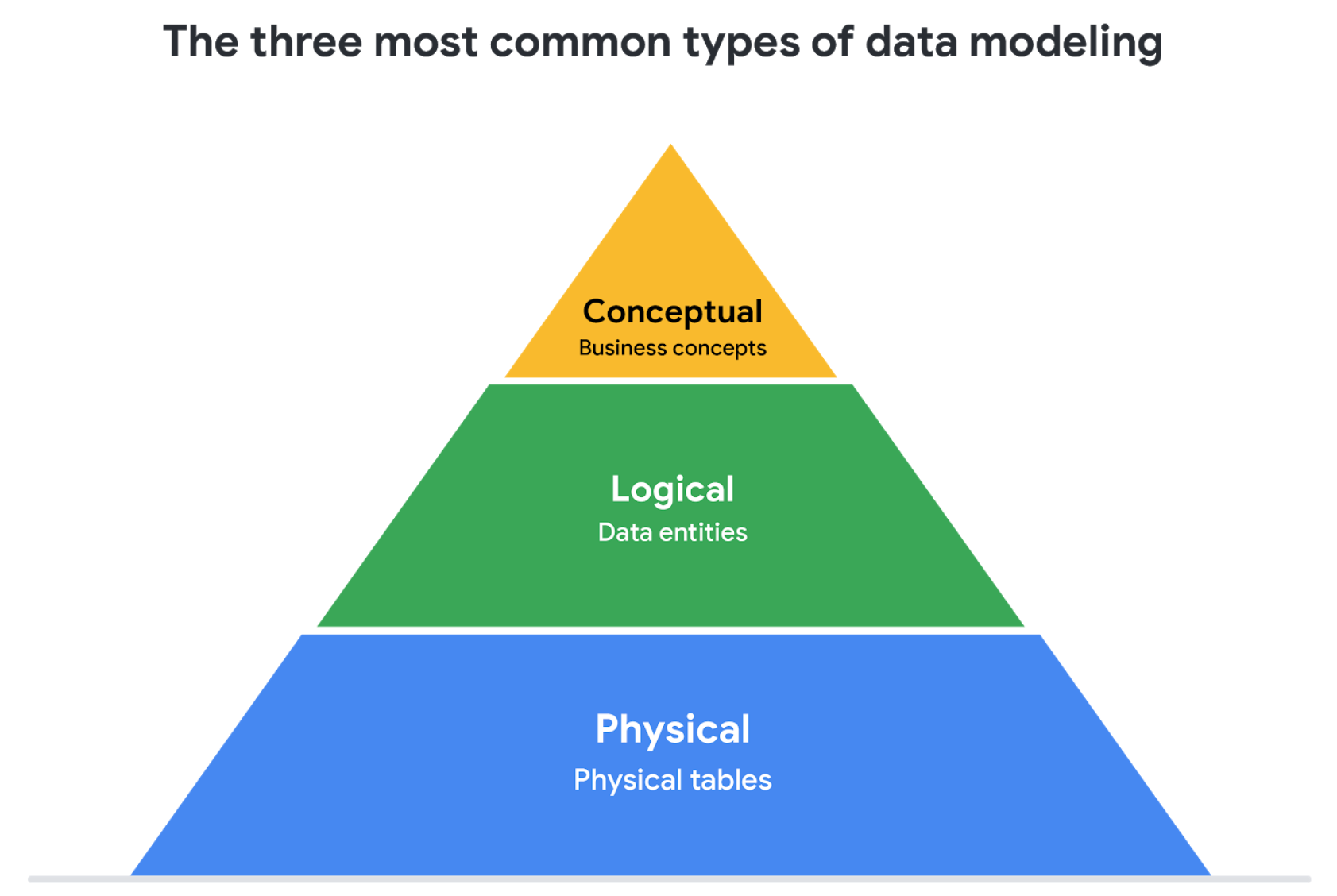
## 

## **What is data modeling?**

**Data modeling** is the process of creating diagrams that visually represent how data is organized and structured. These visual representations are called **data models**. You can **think of** **data modeling as a blueprint of a house**. At any point, there might be electricians, carpenters, and plumbers using that blueprint. Each one of these builders has a different relationship to the blueprint, but they all need it to understand the overall structure of the house. **Data models are similar**; **different users might have different data needs, but the data model gives them an understanding of the structure as a whole.**

## **Levels of data modeling**

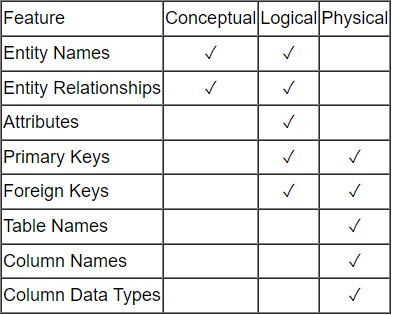
Each level of data modeling has a different level of detail.



1. **Conceptual data modeling** gives a **high-level view of the data structure**, such as **how data interacts across an organization**. For example, a conceptual data model may be used to define the business requirements for a new database. A conceptual data model doesn't contain technical details.
2. **Logical data modeling** focuses on the **technical details of a database such as relationships, attributes, and entities**. For example, a logical data model defines how individual records are uniquely identified in a database. But it doesn't spell out actual names of database tables. That's the job of a physical data model.
3. **Physical data modeling** depicts **how a database operates**. A physical data model defines all entities and attributes used; for example, it includes table names, column names, and data types for the database.

More information can be found in this [comparison of data models.](https://www.1keydata.com/datawarehousing/data-modeling-levels.html)

**THE THREE TYPES OF DATA MODELS:**



We can see that the complexity increases from conceptual to logical to physical. This is why we always first start with the conceptual data model (so we understand at high level what are the different entities in our data and how they relate to one another), then move on to the logical data model (so we understand the details of our data without worrying about how they will actually implemented), and finally the physical data model (so we know exactly how to implement our data model in the database of choice). In a data warehousing project, sometimes the conceptual data model and the logical data model are considered as a single deliverable.

## 

## 

## 

## **Data-modeling techniques**

There are a lot of approaches when it comes to developing data models, but two common methods are the **Entity Relationship Diagram (ERD)** and the **Unified Modeling Language (UML)** diagram. ERDs are a visual way to understand the relationship between entities in the data model. UML diagrams are very detailed diagrams that describe the structure of a system by showing the system's entities, attributes, operations, and their relationships. As a junior data analyst, you will need to understand that there are different data modeling techniques, but in practice, you will probably be using your organization’s existing technique.

You can read more about ERD, UML, and data dictionaries in this [data modeling techniques article](https://dataedo.com/blog/basic-data-modeling-techniques).

## **Data analysis and data modeling**

**Data modeling can help you explore the high-level details of your data and how it is related across the organization’s information systems**. Data modeling **sometimes requires data analysis to understand how the data is put together**; that way, you **know how to map the data**. And finally, data models **make it easier for everyone in your organization to understand and collaborate with you on your data**.

This is important for you and everyone on your team!

[**TEST YOUR KNOWLEDGE ON DATA FORMATS AND STRUCTURES**](https://www.coursera.org/learn/data-preparation/quiz/vn3Ui/test-your-knowledge-on-data-formats-and-structures)

**EXPLORE DATA TYPES,FIELDS, AND VALUES**

[**KNOW THE TYPE OF DATA YOU'RE WORKING WITH**](https://www.coursera.org/learn/data-preparation/lecture/eYxtI/know-the-type-of-data-you-re-working-with)

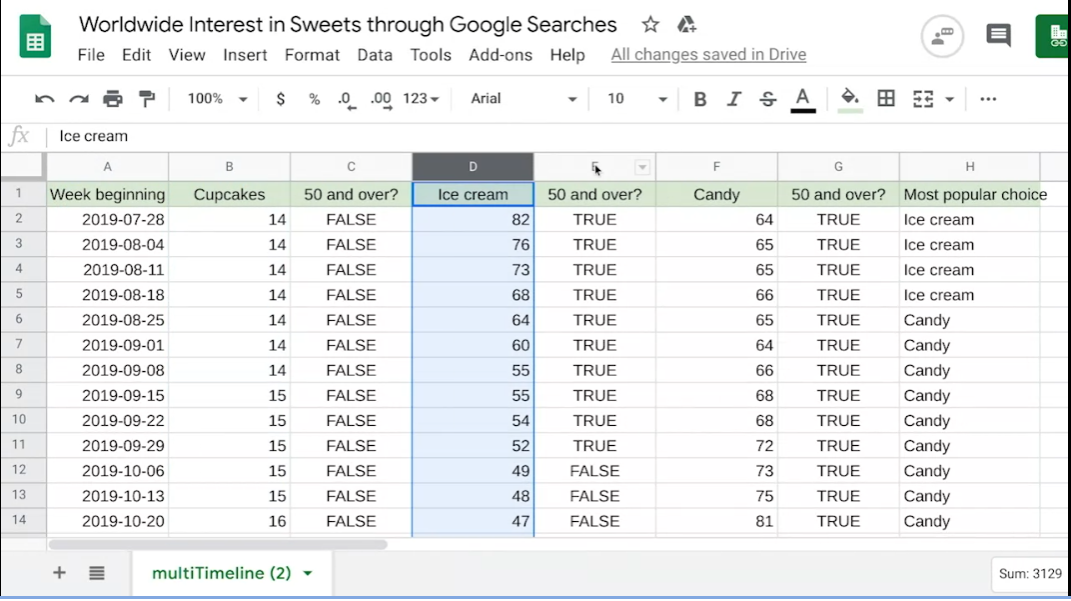
**Another way** you can **describe data**: **the data type**.

A **data type** is a specific kind of data attribute that tells what kind of value the data is. In other words, a **data type tells you what kind of data you're working with**.

Data types can be different depending on the query language you're using.

For example, SQL allows for different data types depending on which database you're using.

Now **a data type in a spreadsheet can be one of three things: a number, a text or string, or a Boolean**. You might find spreadsheet programs that classify them a bit differently or include other types, but these value types cover just about any data you'll find in spreadsheets.



Looking at columns B, D, and F, we find number data types. Each number represents the search interest for the terms "cupcakes," "ice cream," and "candy" for a specific week. The closer a number is to 100, the more popular that search term was during that week. One hundred represents peak popularity. Keep in mind that in this case, 100 is a relative value, not the actual number of searches. It represents the maximum number of searches during a certain time. Think of it like a percentage on a test. All other searches are then also valued out of 100. You might notice this in other data sets as well. Gold star for 100! If you needed to, you could change the numbers into percentages or other formats, like currency.

These are all examples of number data types. In column H, the data shows the most popular treat for each week, based on the search data. So as we'll find in cell H4 for the week beginning July 28th, 2019, the most popular treat was ice cream. This is an example of a **text data type**, **or a string data type**, which is a **sequence of characters and punctuation that contains textual information**. In this example, that information would be the treats and people's names. These can also include numbers, like phone numbers or numbers in street addresses. But these numbers wouldn't be used for calculations. In this case they're treated like text, not numbers.

In columns C, E, and G, it seems like we've got some text. **But the text here isn't a text or string data type.** Instead, it's a **Boolean data type**. A Boolean data type is a data type with **only two possible values: true or false.**

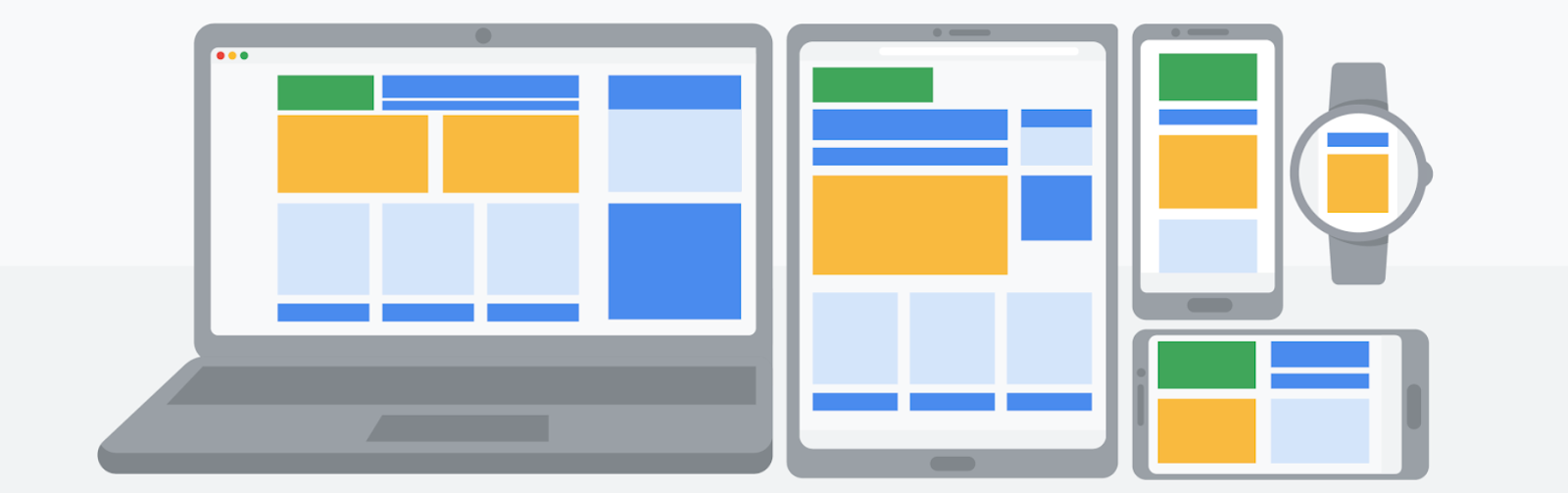
Columns C, E, and G show Boolean data for whether the search interest for each week is at least 50 out of 100. Here's how it works. To get this data, we've created a formula that calculates whether the search interest data in columns B, D, and F is 50 or greater.

In cell B4, the search interest is 14. In cell C4, we find the word false because, for this week of data, the search interest is less than 50. For each cell in columns C, E, and G, the **only two possible values are true or false**. We could change the formula so other words appear in these cells instead, but **it's still Boolean data**.   
  
**Common issue** that people encounter in **spreadsheets**: **mistaking data types with cell values**.

For example, in cell B57, we can create a formula to calculate data in other cells. This will give us the average of the search interests in cupcakes across all weeks in the dataset, which is about 15. The formula works because we calculated using a number data type. But if we tried it with a text or string data type, like the data in column C, we'd get an error. Error values usually happen if a mistake is made in entering the values in the cells. The more you know your data types and which ones to use, the less errors you'll run into.

[**USE BOOLEAN LOGIC**](https://www.coursera.org/learn/data-preparation/supplement/GgZMN/use-boolean-logic)

In this reading, you will explore the basics of Boolean logic and learn how to use single and multiple conditions in a Boolean statement. These conditions are created with Boolean operators, including **AND**, **OR**, and **NOT**. These operators are similar to mathematical operators and can be used to create logical statements that filter your results. **Data analysts** use **Boolean statements to do a wide range of data analysis tasks**, such as **writing queries for searches** and **checking for conditions** when writing programming code.

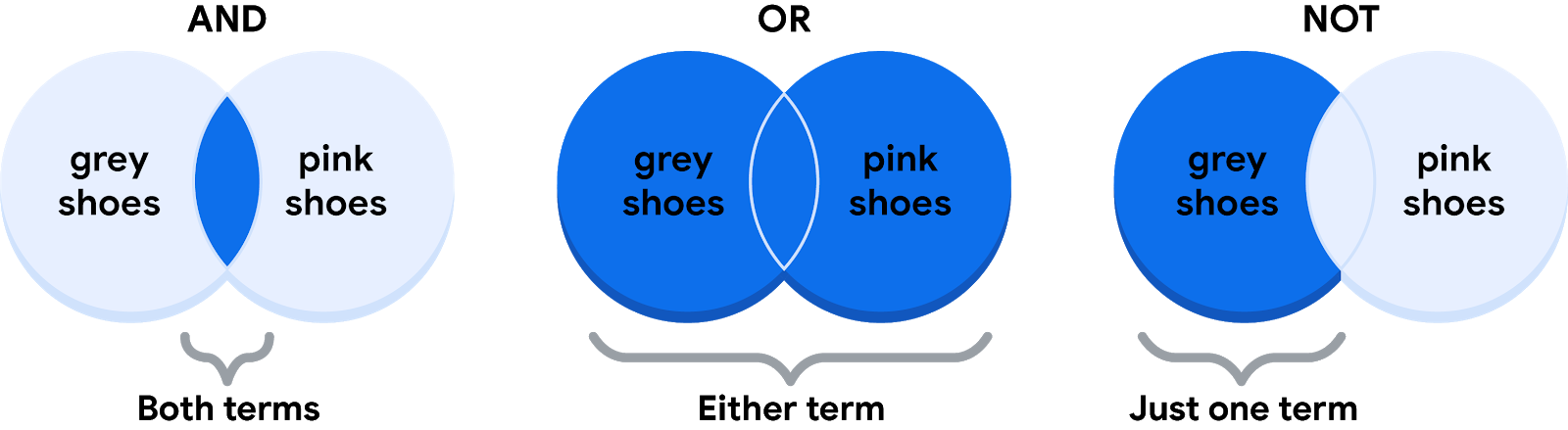


## **Boolean logic example**

Imagine you are shopping for shoes, and are considering certain preferences:

* You will buy the shoes only if they are any combination of pink and gray
* You will buy the shoes if they are entirely pink, entirely gray, or if they are pink and gray
* You will buy the shoes if they are gray, but not if they have any pink

These Venn diagrams illustrate your shoe preferences. **AND** is the center of the Venn diagram, where two conditions overlap. **OR** includes either condition. **NOT** includes only the part of the Venn diagram that doesn't contain the exception.



The intersection of these circles is highlighted to indicate the AND condition requires shoes to be both gray and pink. The Venn diagram that represents OR includes a circle labeled gray shoes overlapping with a circle labeled pink shoes. The entirety of both circles is highlighted to indicate the OR condition means any shoe with gray, pink, or some combination satisfies the requirement. The Venn diagram that represents NOT includes a circle labeled gray shoes overlapping with a circle labeled pink shoes. The portion of the gray shoes circle that does not intersect with the pink shoes circle is highlighted to indicate the NOT condition requires shoes to not include pink.

## **Use Boolean logic in statements**

In queries, Boolean logic is represented in a statement written with Boolean operators. An **operator** is a symbol that names the operation or calculation to be performed.

### **The AND operator**

Your condition is “If the color of the shoe has any combination of gray and pink, you will buy them.” The Boolean statement would break down the logic of that statement to filter your results by both colors. It would say **IF (Color="Gray") AND (Color="Pink") then buy them**

The **AND** operator lets you stack both of your conditions.

Below is a simple truth table that outlines the Boolean logic at work in this statement. In the **Color is Gray** column, there are two pairs of shoes that meet the color condition. And in the **Color is Pink** column, there are two pairs that meet that condition. But in the **If Gray AND Pink** column, only one pair of shoes meets both conditions. So, according to the Boolean logic of the statement, there is only one pair marked true. In other words, there is one pair of shoes that you would buy.

| **Color is Grey** | **Color is Pink** | **If Grey AND Pink, then Buy** | **Boolean Logic** |
| --- | --- | --- | --- |
| Grey/True | Pink/True | True/Buy | True AND True = True |
| Grey/True | Black/False | False/Don't buy | True AND False = False |
| Red/False | Pink/True | False/Don't buy | False AND True = False |
| Red/False | Green/False | False/Don't buy | False AND False = False |

### **The OR operator**

The **OR** operator lets you move forward if either one of your two conditions is met. Your condition is “If the shoes are gray or pink, you will buy them.” The Boolean statement would be **IF (Color="Gray") OR (Color="Pink") then buy them**.

Notice that any shoe that meets either the **Color is Gray** or the **Color is Pink** condition is marked as true by the Boolean logic. According to the truth table below, there are three pairs of shoes that you can buy.

| **Color is Gray** | **Color is Pink** | **If Grey OR Pink, then Buy** | **Boolean Logic** |
| --- | --- | --- | --- |
| Red/False | Black/False | False/Don't buy | False OR False = False |
| Black/False | Pink/True | True/Buy | False OR True = True |
| Gray/True | Green/False | True/Buy | True OR False = True |
| Gray/True | Pink/True | True/Buy | True OR True = True |

### **The NOT operator**

Finally, the **NOT** operator lets you filter by subtracting specific conditions from the results. Your condition is "You will buy any gray shoe except for those with any traces of pink in them." Your Boolean statement would be **IF (Color="Gray") AND (Color=NOT "Pink") then buy them**

Now, all of the gray shoes that aren't pink are marked true by the Boolean logic for the **NOT** Pink condition. The pink shoes are marked false by the Boolean logic for the **NOT** Pink condition. Only one pair of shoes is excluded in the truth table below.

| **Color is Grey** | **Color is Pink** | **Boolean Logic for NOT Pink** | **If Grey AND (NOT Pink), then Buy** | **Boolean Logic** |
| --- | --- | --- | --- | --- |
| Grey/True | Red/False | Not False = True | True/Buy | True AND True = True |
| Grey/True | Black/False | Not False = True | True/Buy | True AND True = True |
| Grey/True | Green/False | Not False = True | True/Buy | True AND True = True |
| Grey/True | Pink/True | Not True = False | False/Don't buy | True AND False = False |

## **The power of multiple conditions**

For data analysts, the real power of Boolean logic comes from being able to combine multiple conditions in a single statement. For example, if you wanted to filter for shoes that were gray or pink, and waterproof, you could construct a Boolean statement such as: “**IF ((Color = "Gray") OR (Color = "Pink")) AND (Waterproof="True")**

Notice that you can use parentheses to group your conditions together.

## **Key takeaways**

Operators are symbols that name the operation or calculation to be performed. The operators **AND**, **OR**, and **NOT** can be used to write Boolean statements in programming languages. Whether you are doing a search for new shoes or applying this logic to queries, Boolean logic lets you create multiple conditions to filter your results. Now that you know a little more about Boolean logic, you can start using it!

## **Resources for more information**

* Learn about who pioneered Boolean logic in this historical article: [Origins of Boolean Algebra in the Logic of Classes](https://www.maa.org/press/periodicals/convergence/origins-of-boolean-algebra-in-the-logic-of-classes-george-boole-john-venn-and-c-s-peirce).
* Find more information about using **AND**, **OR**, and **NOT** from these [tips for searching with Boolean operators](https://libguides.mit.edu/c.php?g=175963&p=1158594).

[**DATA TABLE COMPONENTS**](https://www.coursera.org/learn/data-preparation/lecture/nHw8B/data-table-components)

Here's a riddle for you. What do a **music playlist**, a **calendar agend**a, and an **email inbox** **have in common**?

The answer is **they're all arranged in tables**.

Go ahead and check out your email inbox or a favorite playlist, or look at your calendar agenda. There's tables in every one! **A data table,** or **tabular data,** has a **very simple structure**. It's **arranged in rows and columns**. You can call the **rows "records"** and the **columns "fields".** They **basically** mean the **same thing**, but **records and fields can be used for any kind of data table**, while **rows** and **columns** are usually **reserved** for **spreadsheets**.



When talking about **structured databases**, people in data analytics usually go with "records" and "fields." **Sometimes a field can also refer to a single piece of data, like the value in a cell**. In any case, you'll hear both versions of these terms used throughout this program and your job.

We'll use the new terms we just introduced. So **each song** is a **record**. **Each record** has the **same fields** as the other records in the same order. In other words, **the playlist has the same information about each song.**

Each **song characteristic**, like the **title** and the **artist**, **is a field**.

Each **separate field** has the **same data type**, but **different fields** can have **different types**. Let me show you what I mean. For the song list, the song titles are a text or string type, while the song's length could be a number type if you're using it for calculations. Or it could be a date and time type.

The **column for favorites** is **Boolean** since it has **two possible values: favorite or not favorite**. We can view spreadsheets in the same way.

The **records** in a **spreadsheet** might be about all sorts of things: **clients, products, invoices, or anything else**. **Each record** has **several fields**, which reveal more about the clients, products, or invoices. The value in every cell contains a specific piece of data, like the address of a client or the dollar amount of an invoice.

As a data analyst, lots of data will come your way, and **records**, **fields**, and **values in data tables** will **help you** **navigate analysis**.

Understanding the **structures of the tables** you're working with is a **part of that**. And hopefully, while you're working hard on your analysis and those tables, you can **have a little fun** with a different data table: the one with your favorite playlist!

[**TRANSFORMING DATA**](https://www.coursera.org/learn/data-preparation/supplement/EOCT4/transforming-data)

## 

## **What is data transformation?**

****

A woman presenting data, a hand holding a medal, two people chatting, a ship's wheel being steered, two people high-fiving each other

In this reading, you will explore how data is transformed and the differences between wide and long data.

**Data transformation** is the **process** of **changing the data’s format, structure, or values**. As a data analyst, there is a good chance you will need to transform data at some point to make it easier for you to analyze it.

Data transformation usually involves:

* Adding, copying, or replicating data
* Deleting fields or records
* Standardizing the names of variables
* Renaming, moving, or combining columns in a database
* Joining one set of data with another
* Saving a file in a different format. For example, saving a spreadsheet as a comma separated values (.csv) file.

## 

## **Why transform data?**

Goals for data transformation might be:

* Data **organization**: better organized data is easier to use
* Data **compatibility**: different applications or systems can then use the same data
* Data **migration**: data with matching formats can be moved from one system to another
* Data **merging**: data with the same organization can be merged together
* Data **enhancement**: data can be displayed with more detailed fields
* Data **comparison**: apples-to-apples comparisons of the data can then be made

## **Data transformation example: data merging**

Mario is a plumber who owns a plumbing company. After years in the business, he buys another plumbing company. Mario wants to merge the customer information from his newly acquired company with his own, but the other company uses a different database. So, Mario needs to make the data compatible. To do this, he has to transform the format of the acquired company’s data. Then, he must remove duplicate rows for customers they had in common. When the data is compatible and together, Mario’s plumbing company will have a complete and merged customer database.

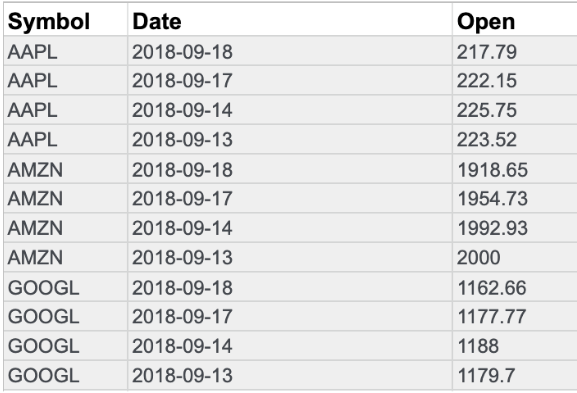
## **Data transformation example: data organization (long to wide)**

To make it easier to create charts, you may also need to transform long data to wide data. Consider the following example of transforming stock prices (collected as long data) to wide data.

**Long data** is data where **each row contains a single data point** for a particular item. In the long data example below, individual stock prices (data points) have been collected for Apple (AAPL), Amazon (AMZN), and Google (GOOGL) (particular items) on the given dates.

**LONG DATA EXAMPLE:**

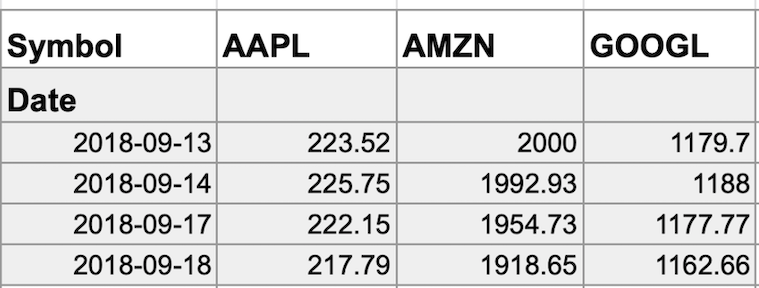
**STOCK PRICES**

****

**Wide data** is data where **each row contains multiple data points** for the particular items identified in the columns.

**WIDE DATA EXAMPLE:**

**STOCK PRICES**

****

With data transformed to wide data, you can create a chart comparing how each company's stock changed over the same period of time.

You might notice that all the data included in the long format is also in the wide format. But **wide data is easier to read and understand**. That is why **data analysts typically transform long data to wide data more often** than they transform wide data to long data.

**THE FOLLOWING TABLE SUMMARIZES WHEN EACH FORMAT IS PREFERRED:**

| **Wide data is preferred when** | **Long data is preferred when** |
| --- | --- |
| Creating tables and charts with a few variables about each subject | Storing a lot of variables about each subject. For example, 60 years worth of interest rates for each bank |
| Comparing straightforward line graphs | Performing advanced statistical analysis or graphing |

# 

# 

# 

# 

# 

# 

# GLOSSARY TERMS FROM MODULE 1

**Terms and definitions for Course 3, Module 1**

**Agenda:** A list of scheduled appointments

**Audio file:** Digitized audio storage usually in an MP3, AAC, or other compressed format

**Boolean data:** A data type with only two possible values, usually true or false

**Continuous data:** Data that is measured and can have almost any numeric value

**Cookie:** A small file stored on a computer that contains information about its users

**Data element:** A piece of information in a dataset

**Data model:** A tool for organizing data elements and how they relate to one another

**Digital photo:** An electronic or computer-based image usually in BMP or JPG format

**Discrete data:** Data that is counted and has a limited number of values

**External data:** Data that lives, and is generated, outside of an organization

**Field:** A single piece of information from a row or column of a spreadsheet; in a data table, typically a column in the table

**First-party data:** Data collected by an individual or group using their own resources

**Long data:** A dataset in which each row is one time point per subject, so each subject has data in multiple rows

**Nominal data:** A type of qualitative data that is categorized without a set order

Ordinal data: Qualitative data with a set order or scale

**Ownership:** The aspect of data ethics that presumes individuals own the raw data they provide and have primary control over its usage, processing, and sharing

**Pixel:** In digital imaging, a small area of illumination on a display screen that, when combined with other adjacent areas, forms a digital image

**Population:** In data analytics, all possible data values in a dataset

**Record:** A collection of related data in a data table, usually synonymous with row

**Sample:** In data analytics, a segment of a population that is representative of the entire population

**Second-party data:** Data collected by a group directly from its audience and then sold

Social media: Websites and applications through which users create and share content or participate in social networking

**String data type:** A sequence of characters and punctuation that contains textual information (Refer to Text data type)

Structured data: Data organized in a certain format such as rows and columns

**Text data type:** A sequence of characters and punctuation that contains textual information (also called string data type)

**United States Census Bureau:** An agency in the U.S. Department of Commerce that serves as the nation’s leading provider of quality data about its people and economy

Unstructured data: Data that is not organized in any easily identifiable manner

**Video file:** A collection of images, audio files, and other data usually encoded in a compressed format such as MP4, MV4, MOV, AVI, or FLV

Wide data: A dataset in which every data subject has a single row with multiple columns to hold the values of various attributes of the subject