**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 question. For example, if you are analyzing trends over time, make sure you use time series data — in other words, data that includes dates.

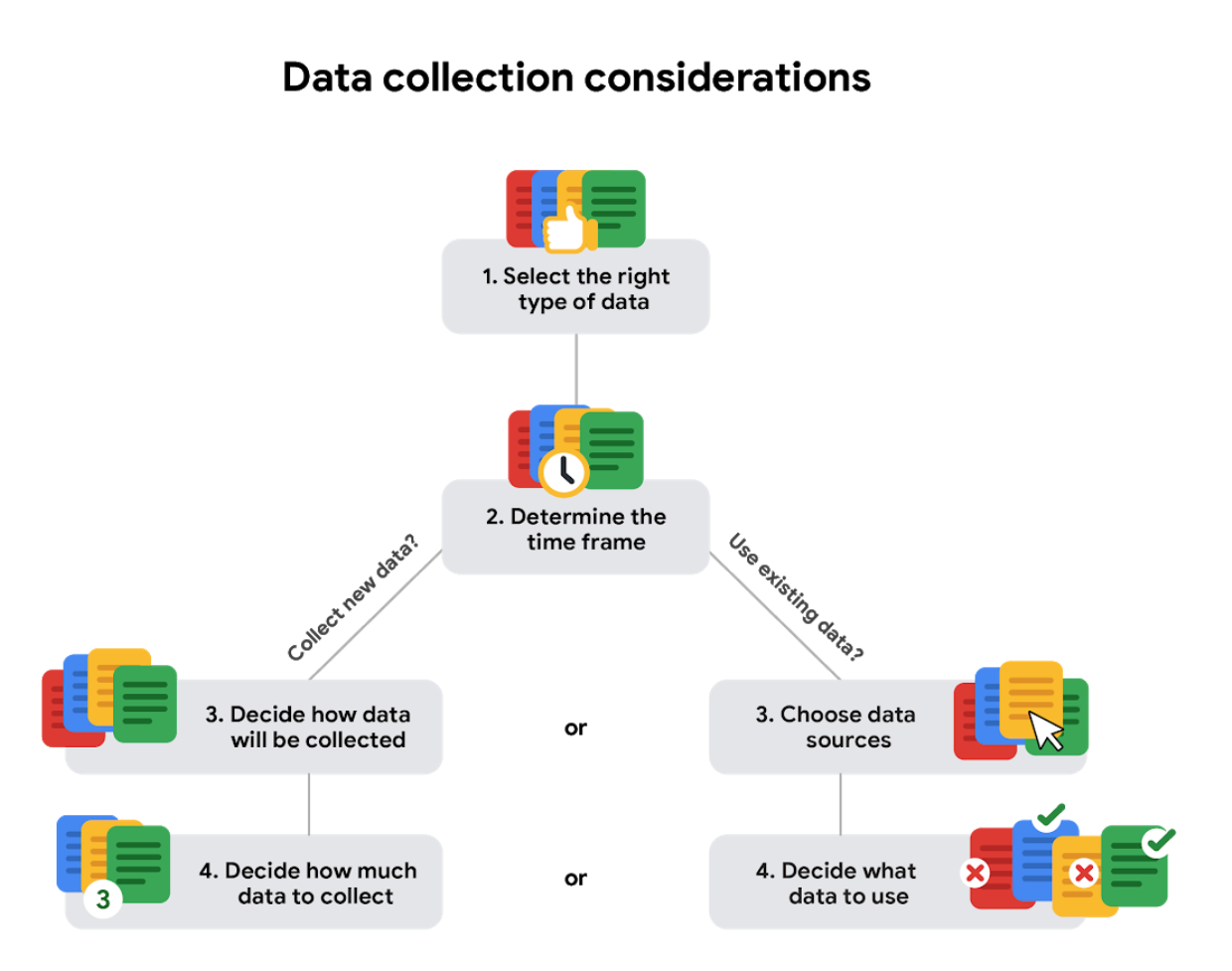
**How much data to collect**

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



**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 versus 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 versus 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 |

**Continuous versus 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 | 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 | 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 |

**Qualitative versus 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 |

**Nominal versus 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 versus 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 |

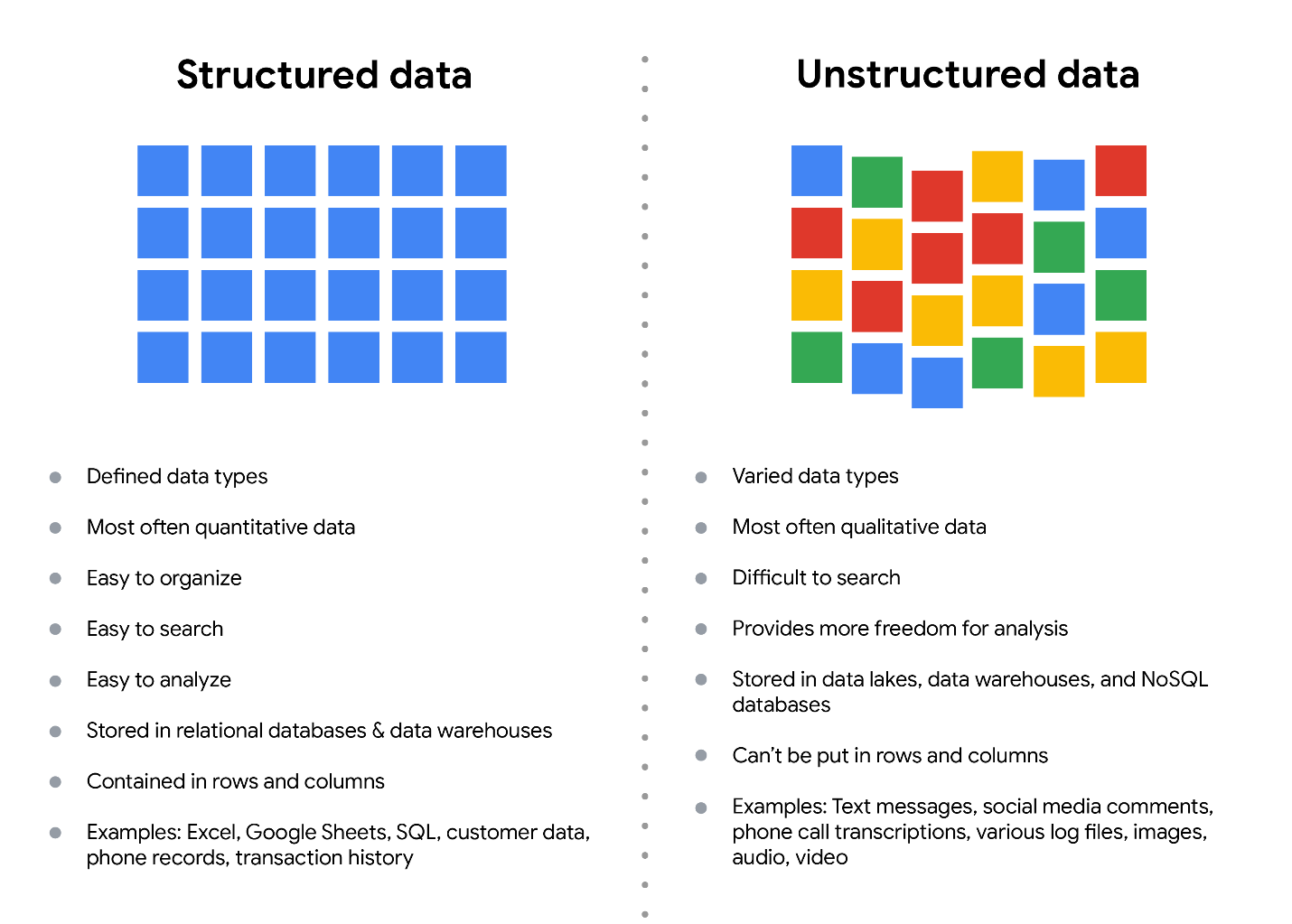
**The effects of different structures**

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.
* **Unstructured data:** Not organized in any easy-to-identify way.

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:



**Structured data**

As we described earlier, **structured data** is organized in a certain format. 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**

**Unstructured data** can’t be organized in any easily identifiable manner. 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.

**The fairness issue**

The lack of structure makes unstructured data difficult to search, manage, and analyze. But recent 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.

**Wide data** is a dataset in which every data subject has a single row with multiple columns to hold the values of various attributes of the subject. It is helpful for comparing specific attributes across different subjects.

**Long data** is data in which each row represents one observation per subject, so each subject will be represented by multiple rows. This data format is useful for comparing changes over time or making other comparisons across subjects.

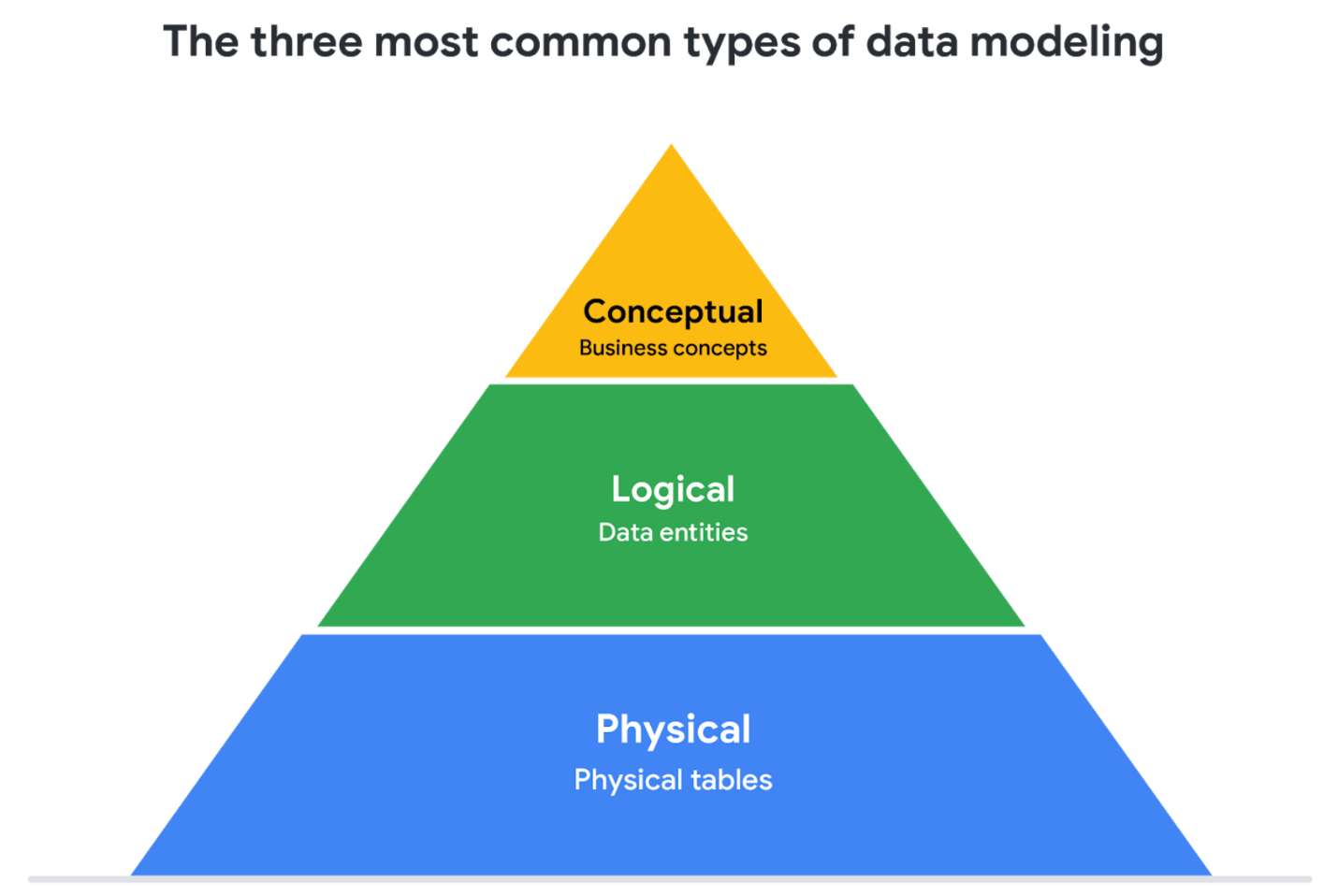
A screenshot of a computer

Description automatically generated

**Data modeling** is the process of creating diagrams that visually represent how data is organized and structured.  These visual representations are called **data models**.

**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.

**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:** Apiece 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