**Google Data Analytics**

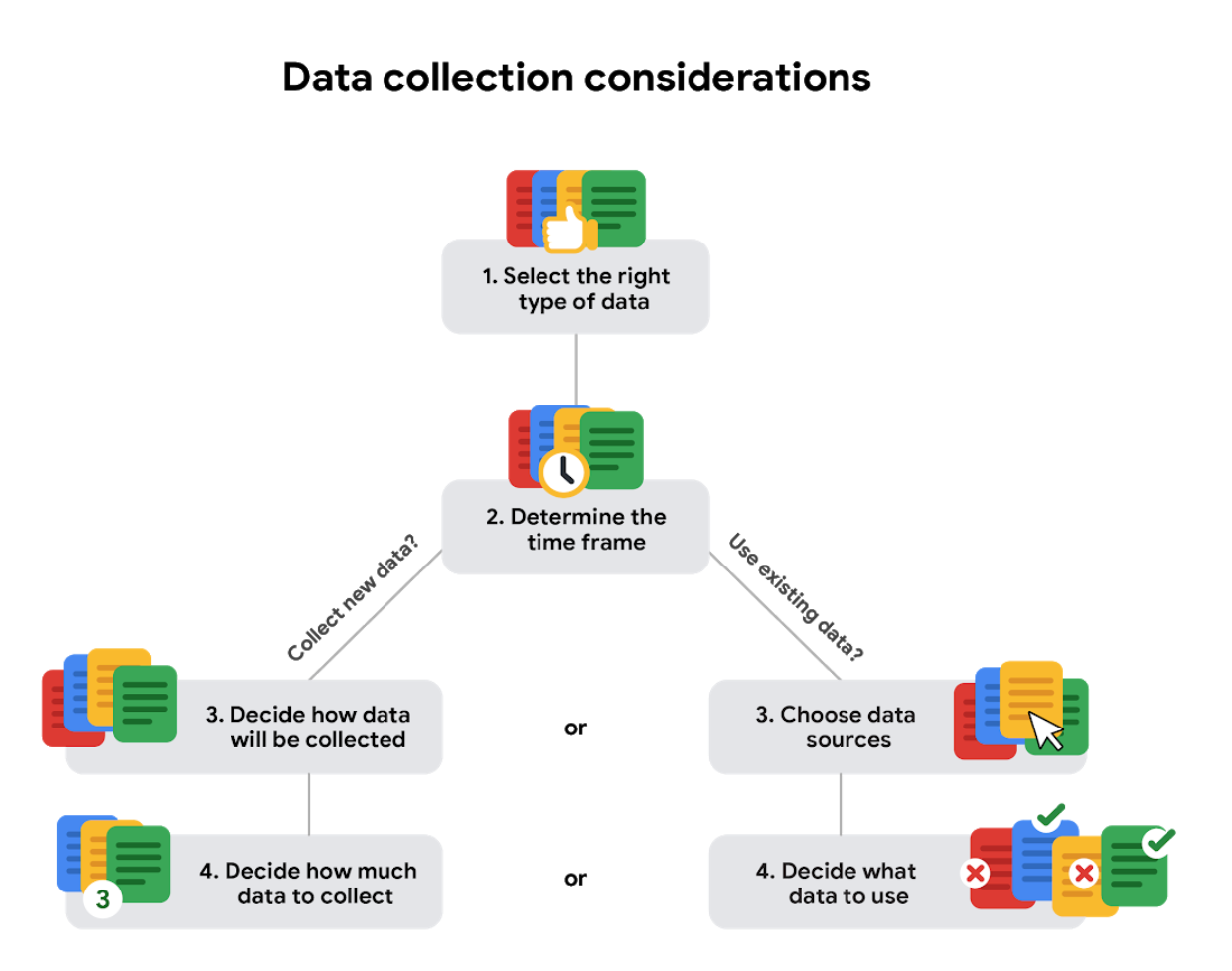
**Prepare for Data Exploration**

**How data is collected**

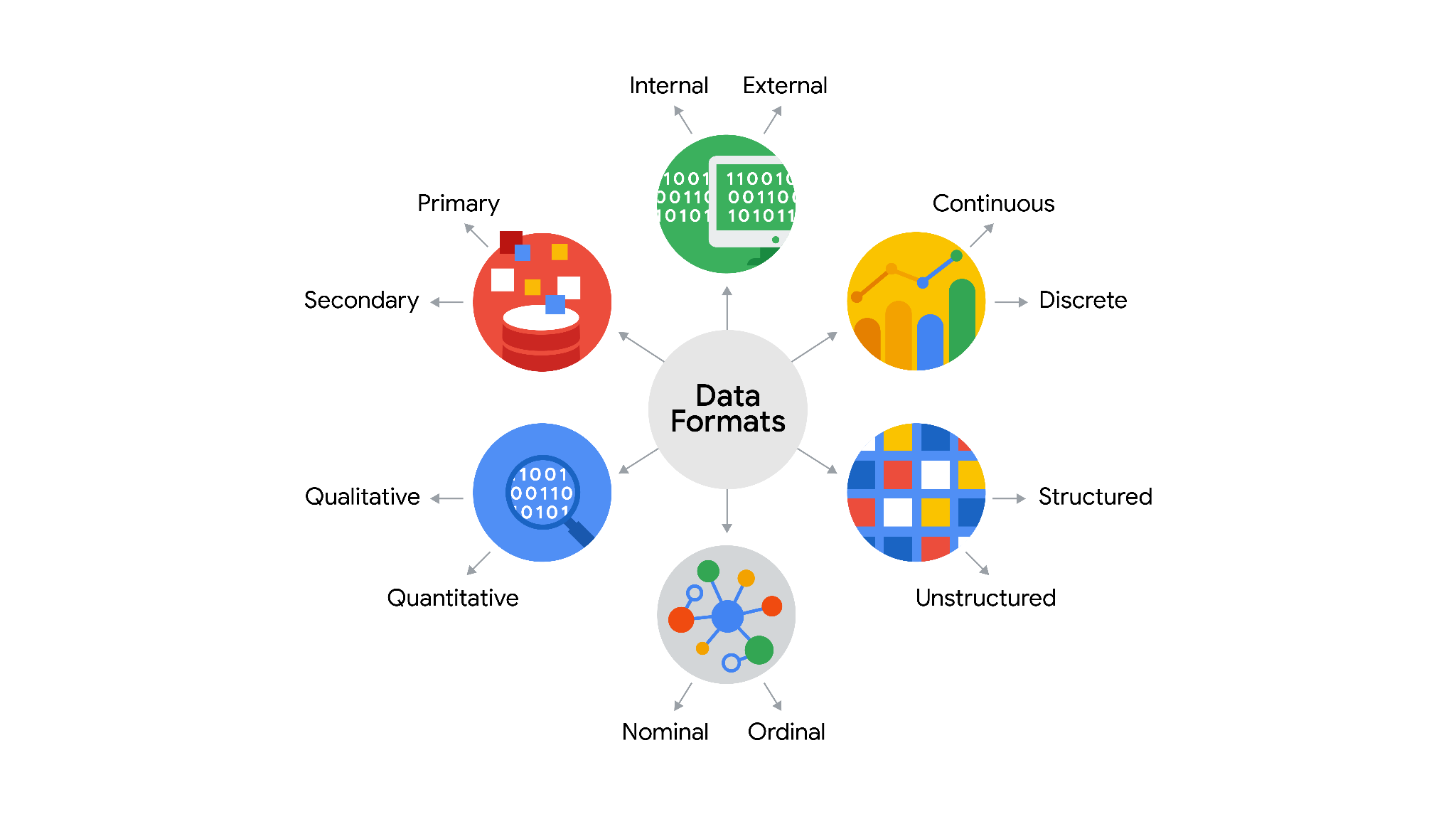
* ﻿﻿Interviews
* ﻿﻿Observations
* ﻿﻿Forms
* ﻿﻿Questionnaires
* ﻿﻿Surveys
* ﻿﻿Cookies

**Data collection considerations**

* ﻿﻿How the data will be collected
* ﻿﻿Choose data sources
* ﻿﻿Decide what data to use
* ﻿﻿How much data to collect
* ﻿﻿Select the right data type
* ﻿﻿Determine the time frame

****

**Note:** When analysis depends on as many data sources as possible **external data** is useful.

**Data Formats**

| **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 |

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Internal data | Data that lives 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 centres |
| External data | Data that lives outside of a company or organisation | - National average wages for the various positions throughout your organisation  - Credit reports for customers of an auto dealership |

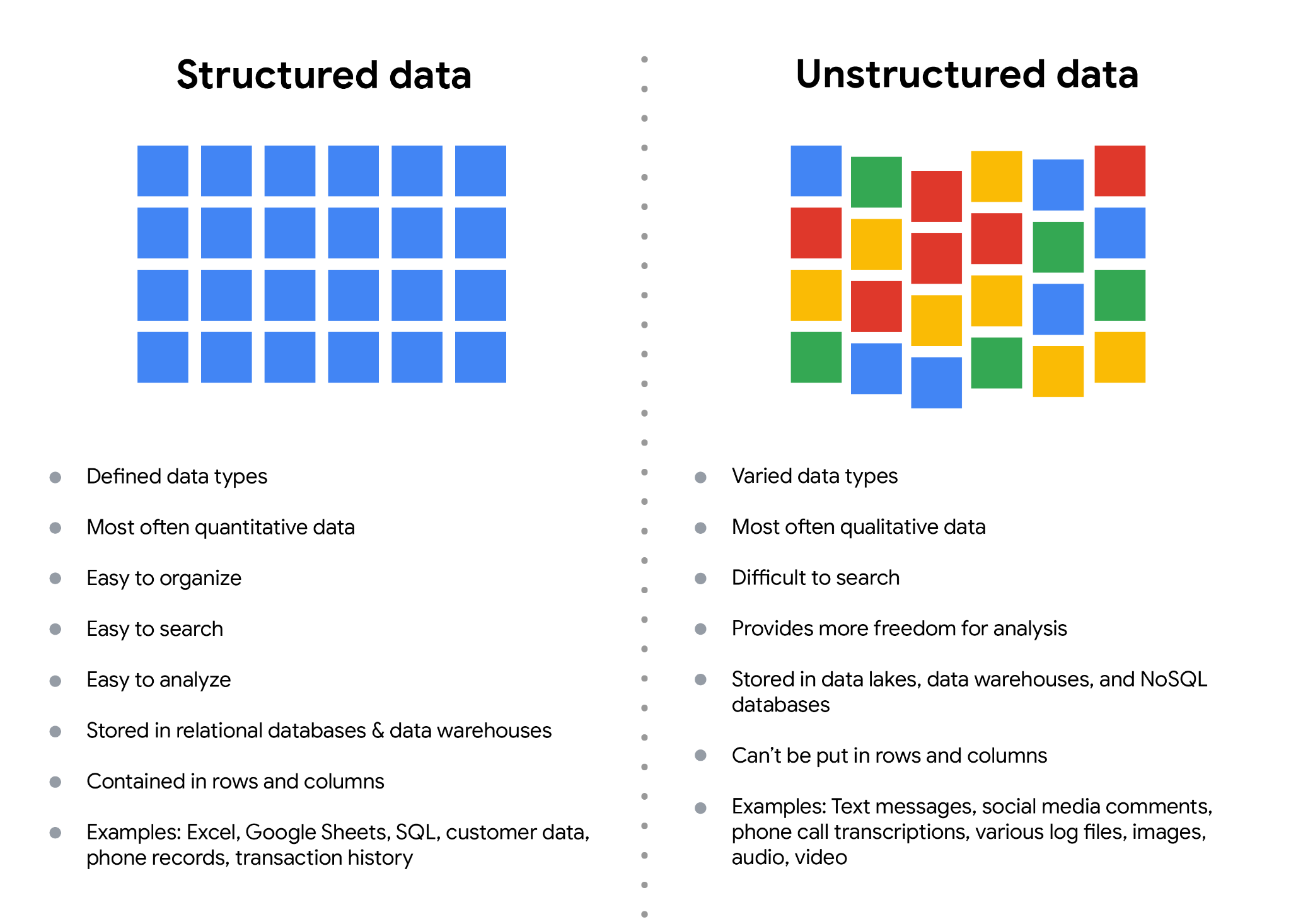
| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Continuous data | Data that is measured and can have almost any numeric value | - Height of kids in third grade classes (52.5 inches, 65.7 inches)  - Runtime markers in a video  - Temperature |
| Discrete data | 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)  - Room’s maximum capacity allowed  - Tickets sold in the current month |

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Qualitative | Subjective and explanatory measures of qualities and characteristics | - Exercise activity most enjoyed  - Favourite brands of most loyal customers  - Fashion preferences of young adults |
| Quantitative | Specific and objective measures of numerical facts | - Percentage of board certified doctors who are women  - Population of elephants in Africa |

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Nominal | A type of qualitative data that isn’t categorised with 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)  - Income level (low income, middle income, high income) |

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Structured data | Data organised in a certain format, like rows and columns | - Expense reports  - Tax returns  - Store inventory |
| Unstructured data | Data that isn’t organised in any easily identifiable manner | - Social media posts  - Emails, Videos |

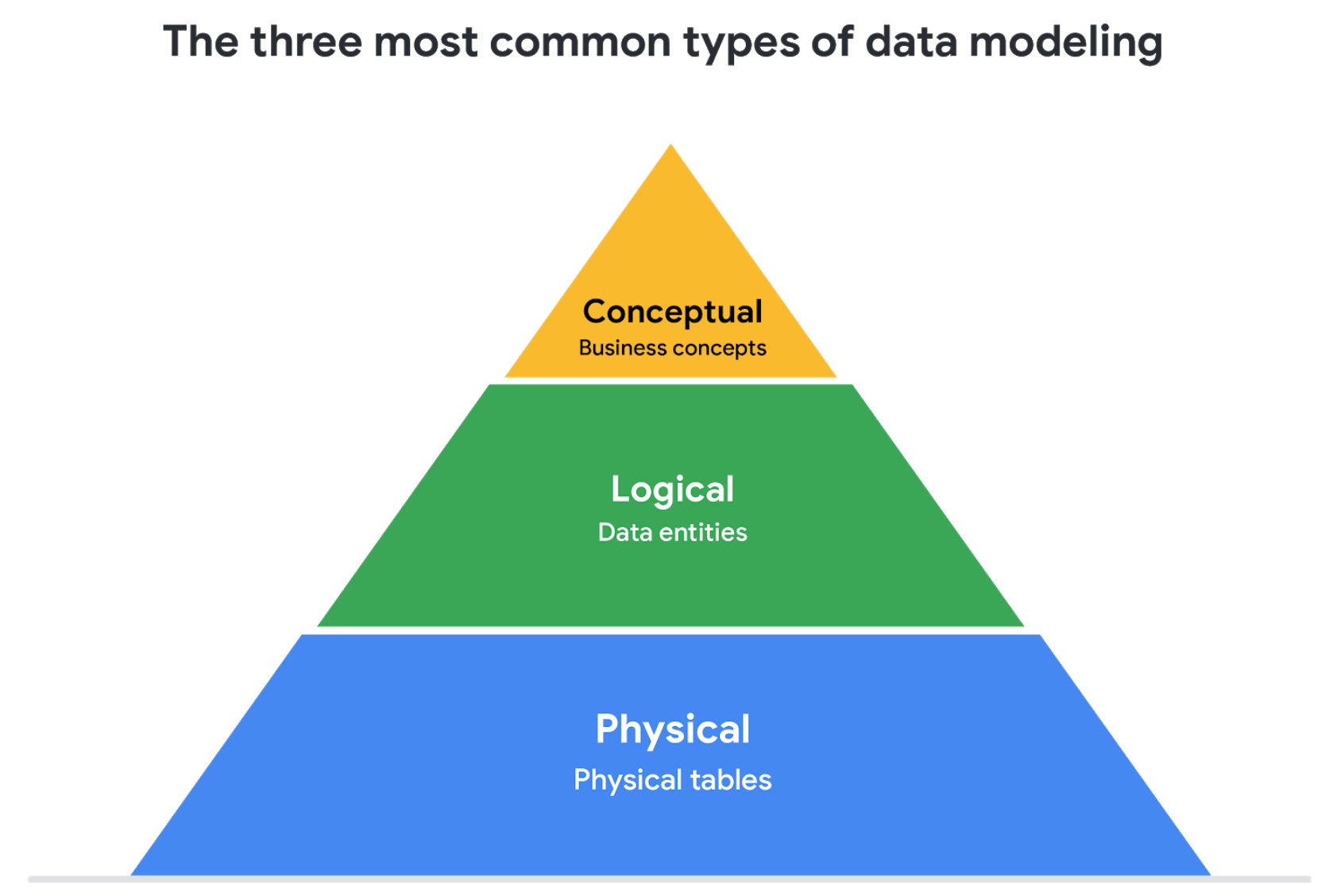
**Structured and Unstructured Data**



**Data Model:**

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

You can think of data modelling 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.



1. **Conceptual data modelling** gives a high-level view of the data structure, such as how data interacts across an organisation. 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 modelling** 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 modelling** 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.

**Reference Link**: [comparison of data models.](https://www.1keydata.com/datawarehousing/data-modeling-levels.html)

## **Data-modelling techniques:**

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

**Reference Link**: [data modelling techniques article](https://dataedo.com/blog/basic-data-modeling-techniques)

**Data Transformations:**

Data transformation is the process of changing the data’s format, structure, or values.

Data transformation usually involves:

* Adding, copying, or replicating data
* Deleting fields or records
* Standardising 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

Goals for data transformation might be:

* Data **organisation**: better organised 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 organisation 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

**Long and Wide Data:**

**Long data** is data where **each row contains a single data point** for a particular item.

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

| **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 |

**Bias:**

Subconscious or conscious preference for or against people, object, subject etc.

**Types of bias:**

* **Sampling bias:** when the sample does ot represent the population
* **Observer (experimenter/ research bias) bias:** The tendency for different people to observe things differently
* **Interpretation bias:** The tendency to always interpret ambiguous situations in a positive or negative way
* **Confirmation bias:** The tendency to search for or interpret information in a way that confirms pre-existing beliefs

**Identifying and finding new data:**

Acronym: **ROCCC** (rock) or **CROCC** (croc)

**R -** Reliable

**O -** Original

**C -** Comprehensive

**C -** Current

**C -** Cited

If you have **original** data from a **reliable** organisation and it's **comprehensive**, **current**, and **cited**, it **ROCCCs**!

**Data ethics**

Well-founded standards of right and wrong that dictate how data is collected, shared, and used

**Aspects of data ethics**

* ﻿﻿Ownership
* ﻿﻿Transaction transparency
* ﻿﻿Consent
* ﻿﻿Currency
* ﻿﻿Privacy
* ﻿﻿Openness

# **Data anonymization:**

Data anonymization is the process of protecting people's private or sensitive data by eliminating that kind of information. Typically, data anonymization involves blanking, hashing, or masking personal information, often by using fixed-length codes to represent data columns, or hiding data with altered values

**de-identification**, which is **a process used to wipe data clean of all personally identifying information**.

**Open Data:**

In data analytics, **open data** is part of **data ethics,** which has to do with using data ethically. **Openness** refers to free access, usage, and sharing of data. But for data to be considered open, it has to:

* Be available and accessible to the public as a complete dataset
* Be provided under terms that allow it to be reused and redistributed
* Allow universal participation so that anyone can use, reuse, and redistribute the data

Data can only be considered open when it meets all three of these standards

**Data interoperability**

The ability of data systems and services to openly connect and share data

**Databases**

A storage for data

**Primary key:**

* ﻿﻿Used to ensure data in a specific column is unique
* ﻿﻿Uniquely identifies a record in a relational database table
* ﻿﻿Only one primary key is allowed in a table
* ﻿﻿Cannot contain null or blank values
* A primary key may also be constructed using multiple columns of a table. This type of primary key is called a **composite key**

**Foreign key:**

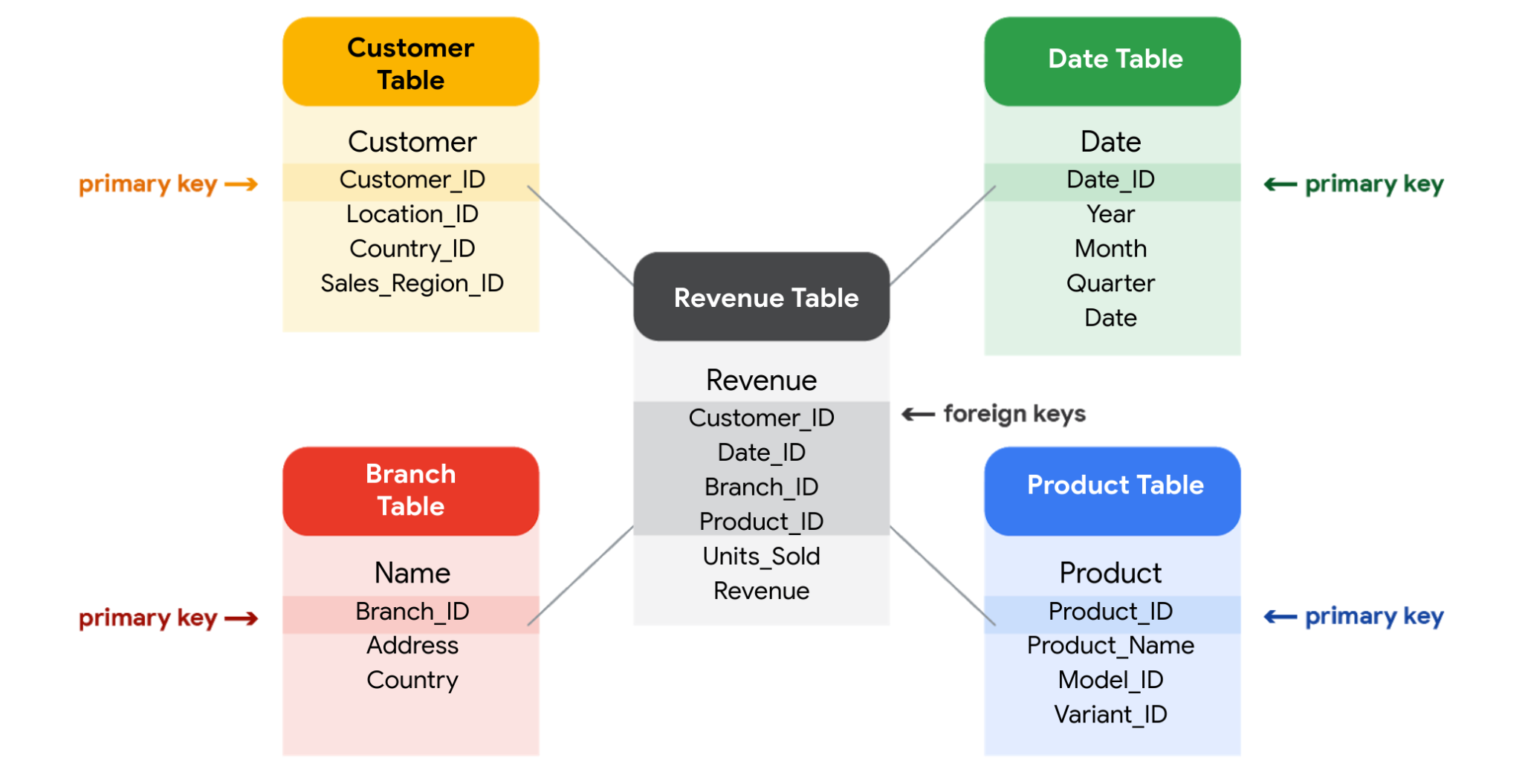
* ﻿﻿A column or group of columns in a relational database table that provides a link between the data in two tables
* ﻿﻿Refers to the field in a table that's the primary key of another table
* ﻿﻿More than one foreign key is allowed to exist in a table

## **Relational databases**

A **relational database** is a database that contains a series of tables that can be connected to show relationships. Basically, they allow data analysts to organise and link data based on what the data has in common.

### **Database Normalisation**

**Normalisation** is a process of organising data in a relational database. For example, creating tables and establishing relationships between those tables. It is applied to eliminate data redundancy, increase data integrity, and reduce complexity in a database.



**Metadata:**

**Metadata** is used in database management to help data analysts interpret the contents of the data within the database

Metadata is stored in a single, central location, and gives the company standardised information about all of its data

* Metadata creates a single source of truth by keeping things consistent and uniform
* Metadata also makes data more reliable by making sure it's accurate, precise, relevant, and timely

3 common types of metadata:

* ﻿﻿**Descriptive**

Metadata that describes a piece of data and can be used to identify it at a later point in time

* ﻿﻿**Structural**

Metadata that indicates how a piece of data is organised and whether it is part of one, or more than one, data collection

* **﻿﻿Administrative**

Metadata that indicates the technical source of a digital asset

**Metadata repository**

A database specifically created to store metadata

* ﻿﻿Describe the state and location of the metadata
* ﻿﻿Describe the structures of the tables inside
* ﻿﻿Describe how the data flows through the repository
* ﻿﻿Keep track of who accesses the metadata and when

**Organising and Protecting Data:**

* **﻿﻿Makes it easier to find and use**
* **﻿﻿Helps you avoid making mistakes during your analysis**
* **﻿﻿Helps to protect your data**

**Best practices when organising data**

* ﻿﻿Naming conventions
* ﻿﻿Foldering
* ﻿﻿Archiving older files
* ﻿﻿Align your naming and storage practices with your team
* ﻿﻿Develop metadata practices (creating metadata for how data would be organised for the project)

## **Best practices for file naming conventions**

Review the following file naming recommendations:

* Work out and agree on file naming conventions early on in a project to avoid renaming files again and again.
* Align your file naming with your team's or company's existing file-naming conventions.
* Ensure that your file names are meaningful; consider including information like project name and anything else that will help you quickly identify (and use) the file for the right purpose.
* Include the date and version number in file names; common formats are YYYYMMDD for dates and v## for versions (or revisions).
* Create a text file as a sample file with content that describes (breaks down) the file naming convention and a file name that applies it.
* Avoid spaces and special characters in file names. Instead, use dashes, underscores, or capital letters. Spaces and special characters can cause errors in some applications.

## **Best practices for keeping files organised**

Remember these tips for staying organised as you work with files:

* Create folders and subfolders in a logical hierarchy so related files are stored together.
* Separate ongoing from completed work so your current project files are easier to find. Archive older files in a separate folder, or in an external storage location.
* If your files aren't automatically backed up, manually back them up often to avoid losing important work.

**Data Security:**

**Encryption** uses a unique algorithm to alter data and make it unusable by users and applications that don’t know the algorithm. This algorithm is saved as a “key” which can be used to reverse the encryption; so if you have the key, you can still use the data in its original form.

**Tokenization** replaces the data elements you want to protect with randomly generated data referred to as a “token.” The original data is stored in a separate location and mapped to the tokens. To access the complete original data, the user or application needs to have permission to use the tokenized data and the token mapping. This means that even if the tokenized data is hacked, the original data is still safe and secure in a separate location.

**Online Presence**

A professional online presence can

* ﻿﻿Help potential employers find you
* ﻿﻿Make connections with other analysts
* ﻿﻿Learn and share data findings
* ﻿﻿Participate in community events

Online Blogs and Communities:

* ﻿﻿O'Reilly
* ﻿﻿Kaggle
* ﻿﻿KDnuggets
* ﻿﻿GitHub
* ﻿﻿Medium
* Try connecting online with communities and sign up for a newsletter to keep up with the latest news on technology.

**Mentors:**

A professional who shares their knowledge, skills, and experience to help you develop and grow

* [SCORE](https://www.score.org/)
* [MicroMentor](https://www.micromentor.org/)

**Sponsor**

A professional advocate who's committed to moving a sponsee's career forward within an organisation