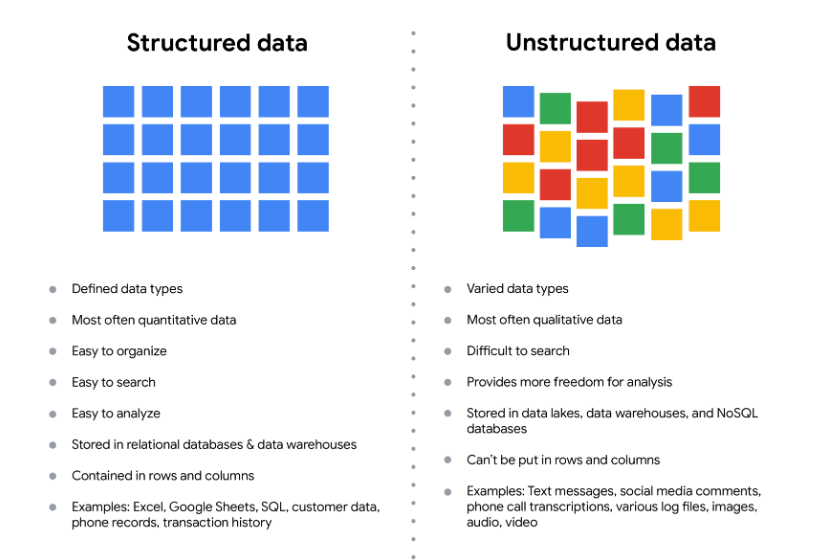
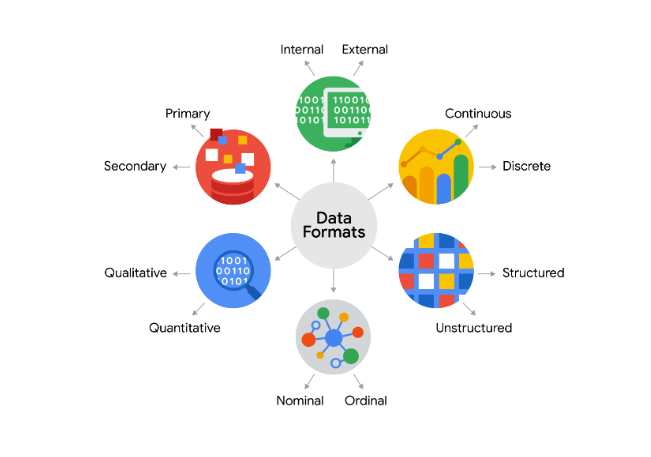
**Introduction to Data Exploration**

* Focus on the prepare stage of the data analysis journey
  + Understanding data types and structures
  + Understanding privacy, bias, credibility, ethics, access
  + Databases
  + Organizing and protecting your data
  + Engaging with the data community (networking)
* **Subtopic: Data Collection in Our World** 
  + Data Collection Methods:
    - Interviews
    - Observations
    - Forms
    - Questionnaires
    - Surveys
    - Cookies – small files that contain information about users
  + Determining what data to collect:
    - How will the data be collected
    - Choose data sources
      * First party data – collecting data yourself
      * Second party data – buying from an outside source who collected it themselves
      * Third party data – collected from different sources who didn’t collect it
        + Least reliable
    - Decide what data to use, don’t get distracted by other data
    - How much data to collect
      * Consider the population
      * Consider collecting a sample – a part of a population that is representative of the population
    - Select the correct data type
    - Determine the time frame for data collection
* **Subtopic: Differentiate between data formats and structures** 
  + Discrete data – counted, limited number of values
  + Continuous data – data that is measured and can have any numeric value
  + Qualitative:
  + Nominal data – data categorized without a set order
  + Ordinal data – data categorized with a set order or scale
    - Eg. rank a movie from 1-5
  + Internal data – data that lives and generated within an organization
  + External data – data that lives and is generated outside of the organization
  + Structured data – data organized in a certain format
    - Tables, spreadsheets, SQL databases
  + Unstructured data – data that is not organized
    - Audio, video files, social media, emails
  + Structured vs Unstructured comparison



* + Data Formats



* + Data Modelling
    - The process of creating diagrams that visually represent how data is organized and structured – think of it as a blueprint
    - Three levels of data modelling, each with a different level of detail
      * Conceptual
        + A high level view of the data structure
        + Eg. How data interacts across an organization
        + Eg. Define business requirements
        + Does not contain technical details
      * Logical
        + Focus on the technical details of a database
        + Doesn’t actually spell out the actual names of database tables
      * Physical
        + Depicts how a database operates
        + Defines all entities and attributes used
    - Data modelling techniques:
      * Entity relationship diagrams – ERD
        + A visual way to understand the relationship between entities in the data model
      * Unified modelling language diagrams – UML
        + Detailed diagrams that describe the structure of a system by showing the systems entities, attributes, operations, and their relationships
* **Subtopic: Know the type of data you’re working with** 
  + Data Types in spreadsheets
    - Number, strings, Booleans
  + Wide and long data
    - Wide – lots of attributes
      * Easy to identify and compare different rows of data
      * Preferred when creating tables and charts with a few variables about each subject
      * Preferred when comparing straight forward line graphs
    - Long – lots of rows (observations)
      * Each row is usually one time point of the same subject
      * Keeps things nice and compact
      * Preferred when storing lots of variables about one subject
      * Preferred when performing advanced statistical analysis or graphing
  + Data Transformation:
    - Adding, copying, replicating data
    - Deleting fields or records
    - Standardizing the names of variables
    - Renaming, moving, combining columns in a database
    - Joining one set of data with another
    - Saving a file in a different format
  + Goals for data transformation:
    - Organization, compatibility, migration, merging, enhancement, comparison