Attributes

1. Attribute's Use
   (a) Quantitative – Numeric values where mathematical operations make sense
      - Discrete – Distinct countable values
        – Number of people in a room
      - Continuous – Any value within a range is possible
        – Temperature

   (b) Qualitative (Categorical) – Value represents a category / group
      - Favorite pet

2. Attribute's Type – Type of data the attribute contains determines how the attribute is stored in a computer
   (a) Text – Words or characters without order or numerical utility
      - `name: "Milton", favorite_pet: "cat"`

   (b) Integer – Numbers with no decimal points
      - `age: 38, weight: 230`

   (c) Float – Numbers with decimal points
      - `height: 180.5cm`

   (d) Boolean – Two possible values (True / False)

- We can further consider properties of quantitative / qualitative attributes
- Qualitative
  – Nominal – Categorical without an imposed order. Allows distinctness between attributes (same or not same)
    ∗ Color (Red, blue, green)
  – Ordinal – Categorical with an imposed order. Allows comparing if one is greater than the order
    ∗ Shirt size (Small, Medium, Large)
- Quantitative
  – Interval – No meaningful ratio / multiplication among values, but has meaningful differences
    ∗ Temperature in Celsius
  – Ratio – Meaningful ratio / multiplication and difference among values
    ∗ Temperature in Kelvin
    ∗ Counting items
    ∗ Based off some absolute point (0)
- Attribute Names

  - Attributes with uninformative names can make it difficult to know what they are capturing
    - `attr_A, attr_B, ...`
  - Better to use informative naming systems
    - `favorite_pet, name, ...`

Cleaning Data

- Structured data can sometimes require additional work to get it into a structure that you want to work with
- Messy data – Data is not organized and not structured in a well-defined way
- Tidy data – Data is structured where each column represents a distinct attribute, each row corresponds to a unique instance, and each cell contains only one value
- Data structuring – Process of organizing data, often into tables

Before:

Student_ID	Math_Score	Math_Hours_Studied	English_Score	English_Hours_Studied	History_Score	History_Hours_Studied
1	85	10	90	12	88	9
2	90	11	88	11	90	10

After:

Student_ID	Subject	Score	Hours_Studied
1	Math	85	10
1	English	90	12
1	History	88	9
2	Math	90	11
2	English	88	11
2	History	90	10

Storing Data

- There are many different ways to store a dataset and each one has their advantages and disadvantages
  - CSV – Comma Separated Value
  - JSON – JavaScript Object Notation
  - XML – Extensible Markup Language
  - Database
  - One file vs multiple files
  - One directory vs multiple directories
- Need to consider size of dataset when determining the best way to store it

Metadata

- Metadata – Information that describes and provides context about a dataset
  - Data about data
- Code book / data key / data statement / data card is a common place to store metadata about the dataset
- Metadata of the dataset can include:
  - Names and descriptions of attributes
  - Type of data stored in each attribute
  - What values mean (coding scheme) (What does a 1 indicate in an attribute)
  - Units of measurement
  - Date of dataset creation
  - Author and source of data
  - Number of missing values
  - etc.
- Some metadata might make more sense with some data and not others
  - Language of text
  - Resolution of images

Sampling

- We often can't feasibly acquire / process all data from an entire population that we are interested in
  - We will make conclusions based on a more manageable sample of the population

Let's try a sampling exercise:

- I want to understand the relationship between an animal's size and their lifespan. Since I really like bats, I measure the size of the bats (height, width). Does this give me an accurate understanding of:

1. The height of animals?
2. The width of animals?
3. The expected lifespan of animals?
4. The expected lifespan of animals related to their size?

The answer to all of these questions is no!