The Shape of Data

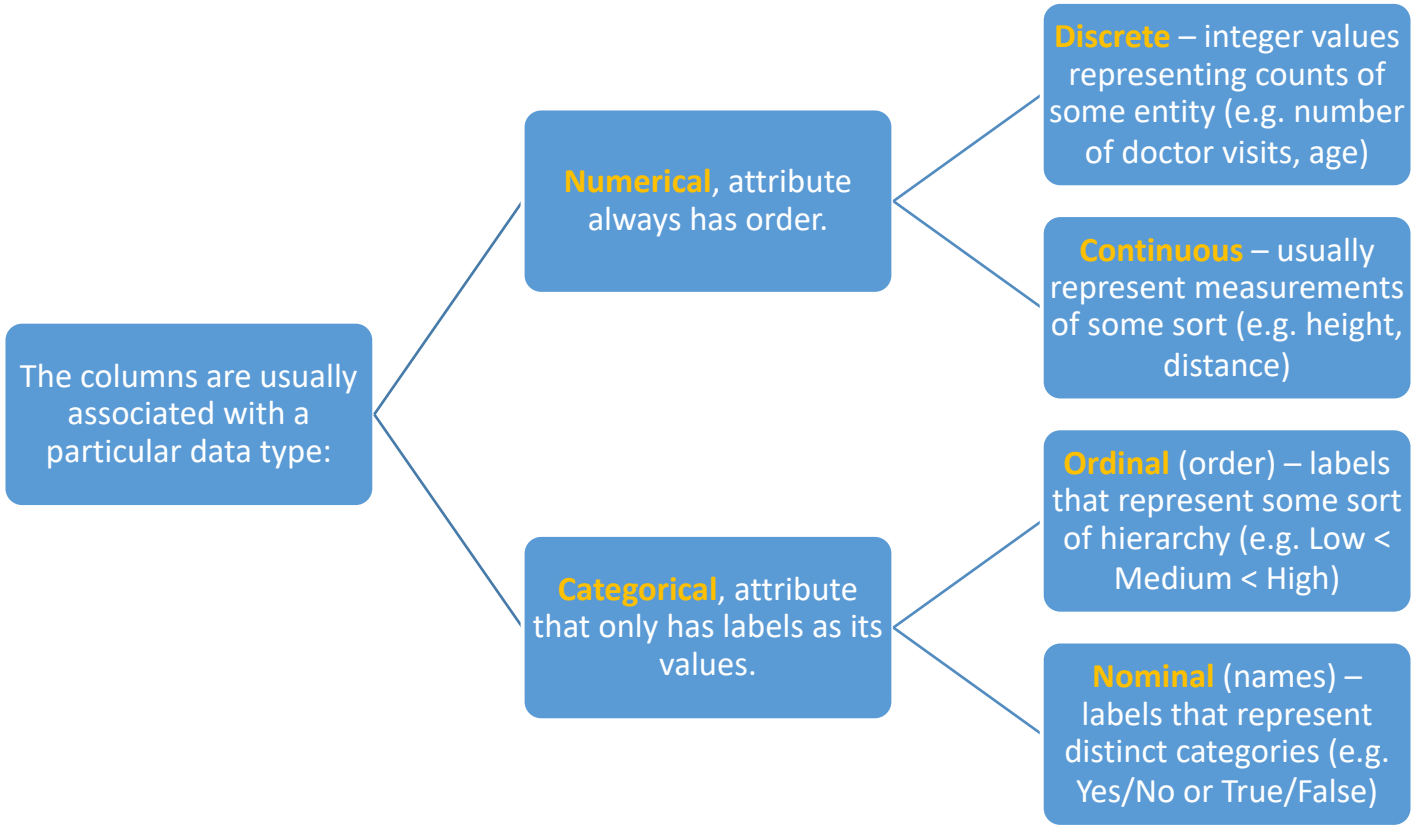
- In data science most of the data encountered is in tabular format (like our tennis data set).
- We call this kind of data **structured data.**
- As opposed to **unstructured data** which usually appears in the form of text.
- In this course we will take a look at both. We will start with structured data.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

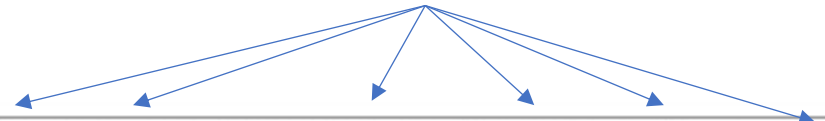
Structured Data

- Structured data consists of tables where
  - each column describes an attribute of the data objects in question. We often call the columns **variables** or **attributes**.
  - Each row describes a single **observation** or data object.
- For example, in our tennis data set each row describes a day in terms of its attributes (columns).

Data Types


The columns are usually associated with a particular data type:
Numerical, attribute always has order.
Discrete – integer values representing counts of some entity (e.g. number of doctor visits, age)
Continuous – usually represent measurements of some sort (e.g. height, distance)
Categorical, attribute that only has labels as its values.
Ordinal (order) – labels that represent some sort of hierarchy (e.g. Low < Medium < High)
Nominal (names) – labels that represent distinct categories (e.g. Yes/No or True/False)

Variables



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

All of the variables in this data set are categorical variables – variables whose values only consist of labels/levels.

Outlook and PlayTennis are **nominal categorical** variables, the labels/level cannot be considered ordered, i.e. $Yes \not\prec No$ and $No \not\prec Yes$

The remaining variables are all ordinal categorical variables – the labels/levels can be considered ordered, i.e.
$Cool < Mild < Hot$

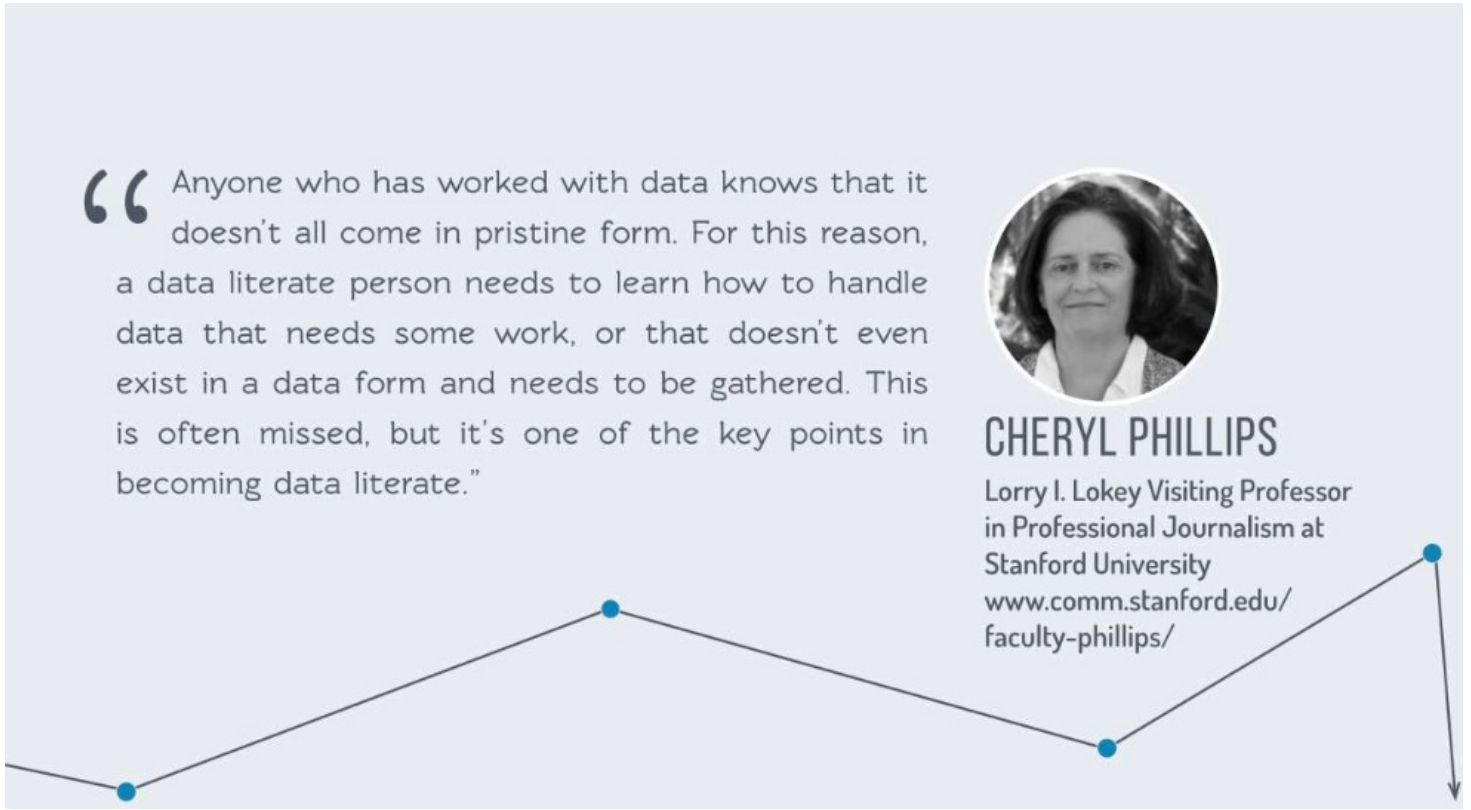
ID	Gender	Age	Income	Rating
1	Male	28	$50,000	4.5
2	Female	35	$65,000	3.8
3	Male	22	$40,000	4.2
4	Female	45	$80,000	4.8
5	Male	31	$55,000	3.5

- **ID**: Discrete numerical variable representing a unique identifier for each individual.
- **Gender**: Nominal categorical variable representing the gender of the individual (Male/Female).
- **Age**: Discrete numerical variable representing the age of the individual.
- **Income**: Continuous numerical variable representing the income of the individual.
- **Rating**: Continuous numerical variable representing a rating given by the individual.

Note: We see later that we will treat numerical ID variables like they appear in this table as nominal categorical variables because it makes no sense to use these identifiers as numerical values, we cannot order them or do mathematical transformations on them.

Real-World Data is Noisy

https://www.slideshare.net/dataremixed/17-key-traits-of-data-literacy

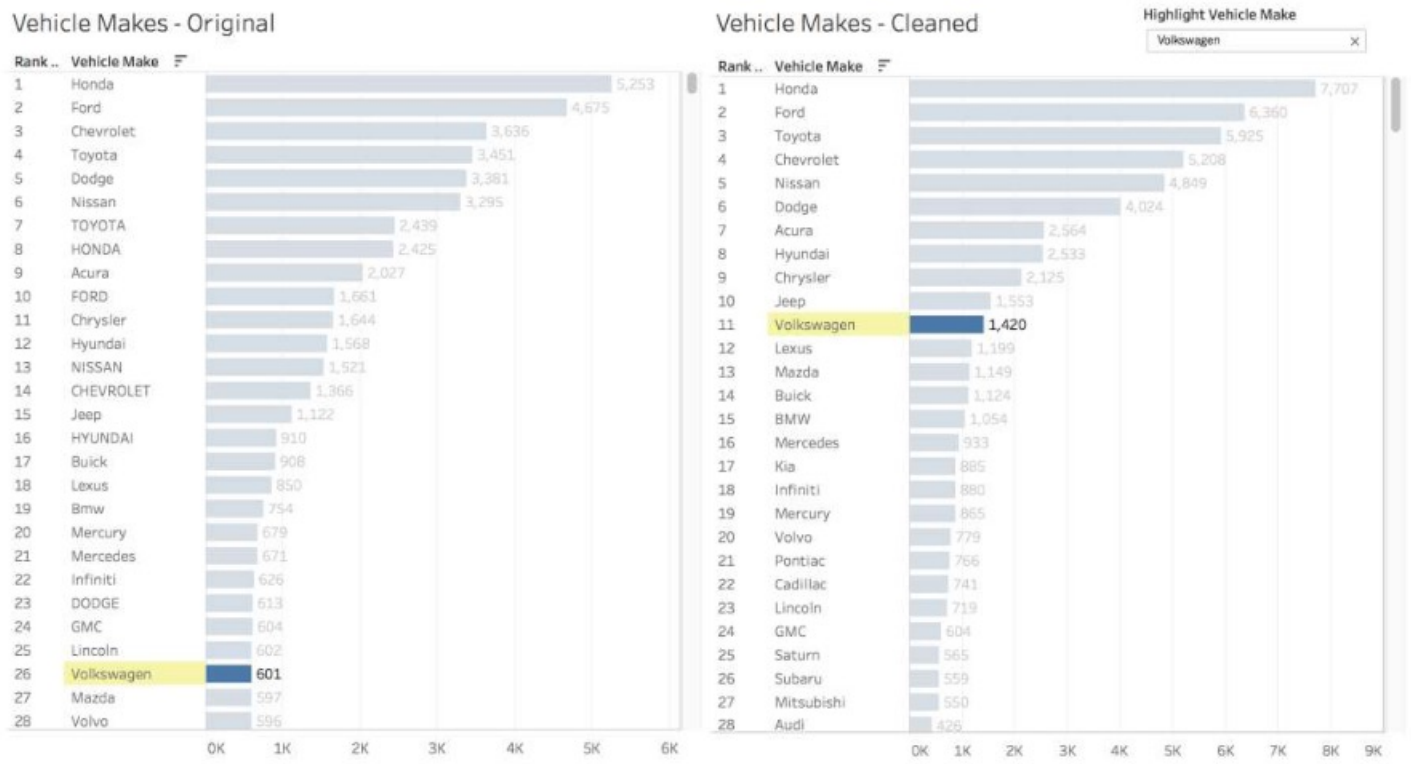


The Effects of Noisy Data

- A marketing firm was tasks to determine how Volkswagen ranks among the top brands
- They found a lot of misspellings of the name Volkswagen

The Effects of Noisy Data

Noisy vs. Clean Data



- With the noisy data the Volkswagen brand ranked at position #26.
- Once cleaned, that is, once the spelling error of the name were removed the brand moved to position #11.
- A huge difference!