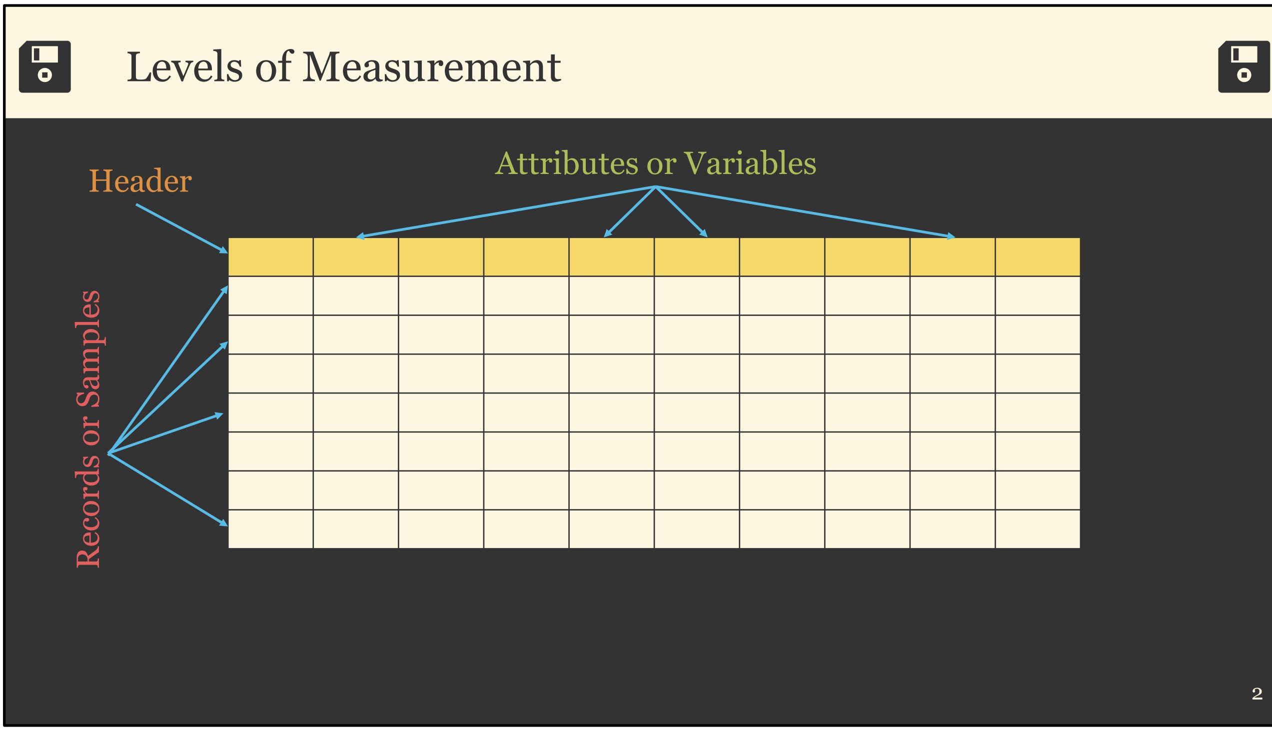
Methods in Open Science

Data

Image from NASA:
https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

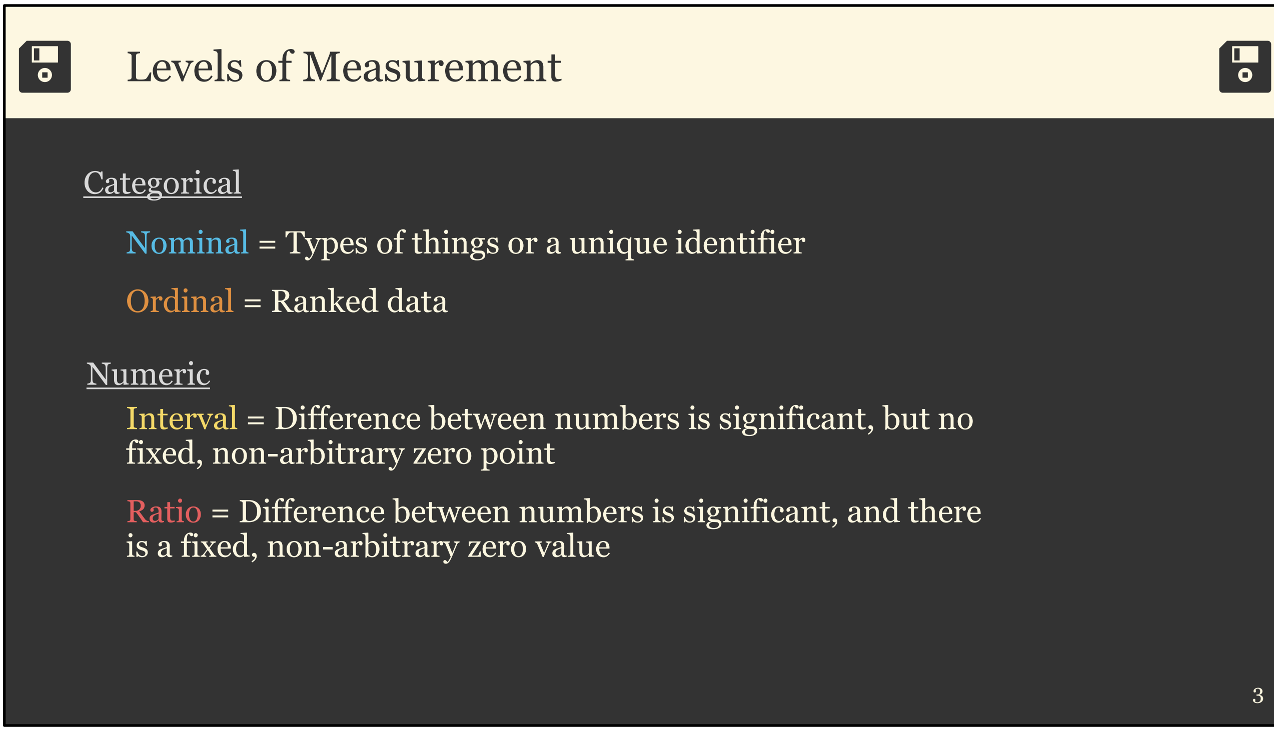
Now that I have broadly introduced data science and its uses and value, we will now focus on some of the key characteristics of data. As the name implies, data science is all about working with data to extract useful and actionable information, so all data scientists must feel comfortable working with a variety of data types. Further, data are often messy, complex, or incomplete in the real world.



Although data can be stored or presented in different formats, one common means is as a two-dimensional table. In a data table, each row is generally an individual record or sample while each column is some information about that specific instance. Columns are commonly termed attributes or variables. The first row is often reserved as descriptors of the columns and is known as a header.

For example, a table could be generated to store information about students at a high school or college. Each row or record could represent information about a specific student while each column would hold a specific piece of information about that student, such as his or her name, address, age, birth date, etc. The header could describe or provide a title for each of the columns or specific attributes.

Although different data structures exists, tables are a very common means to store data, so you will work with many two-dimensional tables in this course. In fact, you have probably already worked with tables in the form of Excel Spreadsheets.



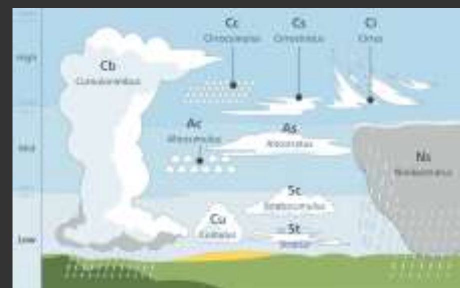
One common means to categorize or better understand data is by grouping them by level of measurement. This is important since the level of measurement will determine what methods are appropriate to analyze the data and what visual representations are most effective.

Data can generally be grouped into two broad types: categorical and numeric. Categories generally represent types of things that are unordered (i.e., nominal) or ordered (i.e., ordinal). Numeric data, in contrast, is some type of continuous measure and can be further partitioned into interval and ratio types.

Note that the word data is technically plural. So, it is correct to say “data are” as opposed to “data is.” However, this can seem a bit awkward or incorrect even though it is grammatically correct.

Nominal and Ordinal Data

- **Nominal Data**
  - Categories where no order is implied
  - Types of cars, rocks, hand tools, clouds, bugs, soils, forests, movie genres, etc.
  - Unique identifiers for specific features: states (FIPS codes), student numbers, ZIP codes, social security numbers, etc.
- **Ordinal Data**
  - Categories with an implied order
  - Academic year (Freshman, Sophomore, Junior, Senior), 1st Place vs. 2nd Place vs. 3rd Place, Age groups, Letter grades, Education level
  - **Likert Scale:** Very Satisfied, Satisfied, Indifferent, Dissatisfied, Very Dissatisfied



https://commons.wikimedia.org/wiki/File:Cloud_types_en.svg

Very Dissatisfied	Dissatisfied	Indifferent	Satisfied	Very Satisfied

4

Nominal data are categories without an implied order. This slide provides many examples. As a specific example from my field, geologist categorize rocks into three broad types based on how they form: igneous, metamorphic, and sedimentary. Note that there is no implied order in these categories. For example, an igneous rock is not better or worse than a sedimentary rock. These broader categories can be further divided into more specific types. For example, we can differentiate some sedimentary rocks based on the size of the grains that make them up (e.g., shale, sandstone, or conglomerate).

Ordinal data are categories with an implied order. One common example is the Likert scale (Very Dissatisfied to Very Satisfied) or academic year (Freshman, Sophomore, Junior, or Senior).

To highlight the importance of level of measurement, it would not make sense to take the average or mean of nominal or ordinal data. However, this would be meaningful for numeric data.

Interval and Ratio Data

❖ Interval Data
- Difference between numbers is significant
- NO fixed, non-arbitrary zero point
- Temperature (Fahrenheit and Celsius scales), Time, Dates, IQ

❖ Ratio Data
- Difference between numbers is significant
- There IS a fixed, non-arbitrary zero value
- Distance, Elevation, Speed/Velocity, Mass, Energy, Power, Barometric pressure, Income, Weight, Height, Blood pressure, Heart rate, etc.
- All percentages and proportions

5

Numeric data can be further divided into interval and ratio types. For both types, the difference between numbers is significant or meaningful. However, interval data do not have a fixed, non-arbitrary zero value whereas ratio data do.

A common example of interval data is temperature measured using the Celsius or Fahrenheit scale. For both scales, zero does not imply “no temperature”. As a result, a temperature that is twice the value is not twice as hot. For example, 40-degrees Celsius is not twice as hot as 20-degrees Celsius. Other examples of interval data include time, dates, and IQ scores.

Ratio data do have a fixed or non-arbitrary zero value. For example, a length of zero meters means no length or a volume of 0 cubic meters means no volume. With a non-arbitrary zero value, twice the quantity will be twice the magnitude. For example, a 100-kilometer highway is twice as long as a 50-kilometer highway. All proportions and percentages are also ratio data, such as percent unemployment.

Although the difference between interval and ratio data can be a bit confusing, it often does not matter in terms of how the data are analyzed or visualized. So, we will generally just discuss nominal, ordinal, and numeric data and not differentiate interval and ratio. However, it is worth discussing the difference.

When is a number not treated as a number?

6

It should be noted that some numbers represent nominal or ordinal data as opposed to numeric data. For example, a ZIP code is simply a unique identifier for a geographic area (as shown in the provided map). So, this numeric value is simply a code, it is not meant to be used mathematically. Other examples of numeric values that are treated as categories include social security numbers or student ID numbers.

Units of Measurement

Quantity	Unit of Measure	Abbreviation	Quantity	Unit of Measure	Abbreviation	Quantity	Unit of Measure	Abbreviation
Mass	Kilograms	kg	Force	Newtons (kg*m/s^2)	N	Magnetic Flux	Weber (V*s)	Wb
Length	Meters	m	Power	Watts (Joules/s)	W	Magnetic Flux Density	Tesla (Wb/m^2)	T
Time	Seconds	s	Pressure	Pascal (N/m^2)	Pa	Inductance	Henry (Wb/A)	H
Amount of Substance	mole	mole	Energy	Joules (N*m)	J	Frequency	Hertz (1/s)	Hz
Electrical Current	Ampere	A	Electrical Charge	Coulombs (A*s)	C	Solid Angle	Steradian	sr
Temperature	Kelvin	K	Electrical Voltage	Volt (W/A)	V	Illuminance	Lux (lm/m^2)	lx
Luminous Intensity	Candela	cd	Electrical Resistance	Ohm (V/A)	Ω	Luminous Flux	Lumen (cd*sr)	lm

https://www.nist.gov/pml/weights-and-measures/metric-si/si-units

Index = a unitless measure

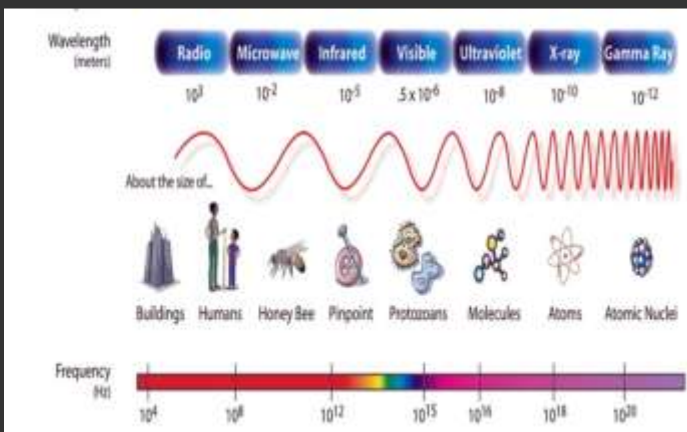
Many physical measurements have different units in which they can be measured. For example, temperature could be measured in Fahrenheit, Celsius, or Kelvin. Length could be measured in millimeters, centimeters, meters, kilometers, feet, yards, or miles. Atmospheric or barometric pressure could be measured in Pascals, atmospheres, pounds per square inch, or millimeters of mercury.

In this class, we will rely on the units defined by the International System of Units where appropriate and/or possible. For example, the standard unit for length is the meter, and the standard unit for energy is the Joule. Unfortunately, for those of us that grew up in the United States, these units have not been universally adopted, so they can take some getting used to. For example, I generally know if I need to where a jacket from a temperature provided in degrees Fahrenheit. However, I do not have this intuition for temperatures reported in Celsius.

The chart on this page lists the standard units for a variety of physical measurements.

Note that some measures are unitless. Examples include proportions and percentages. An index is a common generic term for a unitless measure, and you will hear this term used in this context or with this meaning throughout this course.

Difference in Magnitude



https://commons.wikimedia.org/wiki/File:EM.png

Unit	Distance
Kilometers (km)	1,000 or 1×10^{3} m
Meters (m)	1 or 1×10^{0} m
Centimeters (cm)	0.01 or 1×10^{-2} m
Millimeters (mm)	0.001 or 1×10^{-3} m
Micrometers (or microns) (μm)	0.000001 or 1×10^{-6} m
Nanometers (nm)	0.000000001 or 1×10^{-9} m
Ångstrom (Å)	0.0000000001 or 1×10^{-10} m

8

Another complexity is that phenomenon can be measured over a wide range of magnitudes. One good example of this is the wavelength of electromagnetic radiation. Electromagnetic radiation, such as radio waves or the wavelengths that we perceive as visible light, are conceptualized as an oscillating wave where each oscillation has a length. Different types of radiation are categorized based on ranges of wavelengths.

Electromagnetic radiation covers a wide range of wavelengths. For example, an X-ray has a wavelength that is similar in size to a single atom whereas a single oscillation in a radio wave can be meters in length. As a result, we must be able to measure wavelength across a wide range of magnitudes. This can be accomplished using units that differ by powers of 10. For example, a meter is equivalent to 100 centimeters. We can also represent the value using scientific notation.

The table on this slide shows varying units of measurement for lengths and wavelengths. Visible wavelength ranges are often reported in units of micrometers or nanometers since these values are easier to work with. For example, visible light consists of wavelengths from roughly 400 to 700 nanometers or 0.4 to 0.7 micrometers. In units of meters, this would be 4e-7 to 7e-7, which are more difficult values to work with and comprehend. For larger distances or lengths, using small units would not make sense. For example, we generally report driving distances in kilometers or miles as opposed to in meters or feet.

Difference in Magnitude

- **Log** = power that you must raise the base by in order to obtain the desired value

$$Log_{10}(100) = 2$$

$$10^2 = 100$$

Original Value	Log Base 10	Log Base 2	Lab Base e
1	0	0	0
10	1	3.321928	2.302585
100	2	6.643856	4.60517
1,000	3	9.965784	6.907755
10,000	4	13.28771	9.21034
100,000	5	16.60964	11.51293
1,000,000	6	19.93157	13.81551
10,000,000	7	23.2535	16.1181
100,000,000	8	26.57542	18.42068
1,000,000,000	9	29.89735	20.72327
10,000,000,000	10	33.21928	23.02585

9

When dealing with numbers that vary over a wide range of magnitudes, log transformations can be useful. A log is the power that a base value must be raised by in order to obtain the desired value. For example, the value 100 is the base 10 raised to a power of 2.

In the example chart, base 10 logs are shown. Here, we have values that range from one to ten-billion. Using the base 10 log, this range of values could be truncated into a range of 0 to 10. Or, 10 raised to the 0^{th} power is 1 and 10 raised to the 10^{th} power is ten-billion. Note that it is possible to use other bases, such as base 2 or base e. Base e is known as the natural log.

Regardless of the base used, logs allow us to represent values covering a wide range using a more manageable range, and this is one of the common reasons logs are employed. However, once values have been log transformed, the difference between values is no longer uniform. For example, using the base 10 system a log of 0 and 1 represent the values 1 and 10 whereas a log of 1 and 2 represent the values of 10 and 100. So, the same unit change of the log (0 to 1 and 1 to 2) is not the same magnitude (10 vs. 100). This is an important consideration when we use logs. For example, log transforming an axis on a graph could greatly change the perception of the data.

Uncertainty in Measurement

- Uncertainty = the difference between the real world and its representation as data
- Accuracy = agreement with truth
- Precision = agreement with other, repeated measurements

It should be noted that all data have some inherent uncertainty. We generally define uncertainty as the difference between the real world and its representation as data.

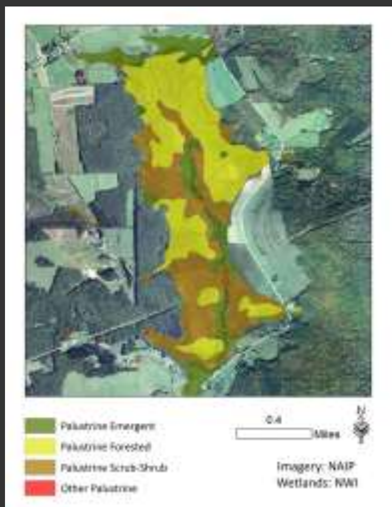
In day-to-day speech, we generally use the terms accuracy and precision interchangeably. However, they have distinct meanings in science, mathematics, and data science.

Accuracy has to do with how well data approximate the correct or actual value in the real world. In contrast, precision has to do with now repeatable measurements are between each other.

The difference here is often explained with a target where the bull's eye represents the correct or real-world value. The example on the left represents an archer that is not accurate but precise. So, he or she is missing the bull's eye, but each arrow hits in nearly the same place. So, he or she is consistent even if the bull's eye is missed. The next example shows a pattern that is less precise but accurate. There is more spread in the pattern, or the archer is less consistent, but the average position would be near the bull's eye. The third example is neither accurate nor precise since the archer is not consistent and the average location would not be near the bull's eye. The last pattern is both precise and accurate.

As a second example, imagine you are collecting GPS measurements at a location. If you get roughly the same coordinates each time you take a measurement, then the measurements are precise but not necessarily accurate, as there could be some systematic error or calibration issue causing the measurements on average to not be close to the correct coordinates.

Difficulties in Measurement



- Difficult to measure
- Lack of access
- Inconsistency
- Poorly defined
- Fuzzy/Gradational
- Human error
- Aggregation

11

What are the reasons for uncertainty in data?

First, not all phenomenon are easily measured. For example, measuring the length of a board in order to make a cut with a saw can generally be done with limited uncertainty. In contrast, measuring the height of a tree is generally more difficult. Some phenomenon are also not consistent. For example, your height and weight can fluctuate throughout the day depending on how long you have been laying down or whether you have just eaten.

You may also not be able to measure what you need to measure due to access or safety constraints. For example, collecting data during a forest fire may not be possible due to safety concerns.

Sometimes it is even difficult to define what we want to measure or record. For example, some phenomenon can be easily categorized, such as the academic year of a student, whereas others are more difficult to partition into discrete, well-defined categories, such as types of soils or types of forests. Further, boundaries between categories may not be well defined or gradational. For example, the location of a shoreline can vary with the tides. The map on the page provides an example of a mapped wetland. Wetlands are generally difficult to map, partially because their boundaries are generally gradational with the surrounding uplands.

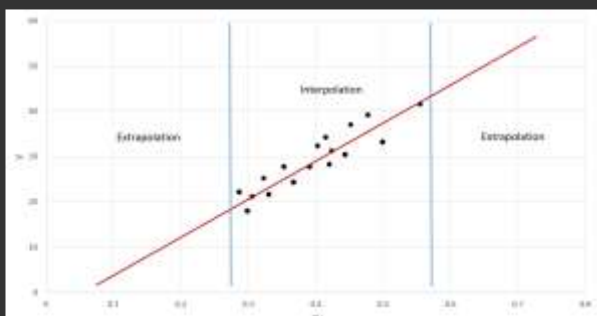
There can also be human error or biases that impact measurements. For example, one technician my overestimate a value while another underestimates it. Another issue is aggregation. For example, disease occurrence rates can be counted at different scales, such as by ZIP code, by county, or by state. This will result in differing levels of generalization and potentially different observed patterns.

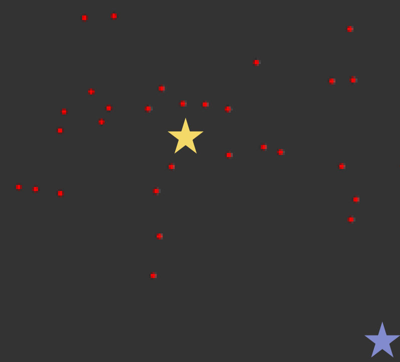
This discussion is not meant to indicate that we cannot trust our data. Instead, it is important to clearly articulate error and uncertainties in our data and our analyses. This is a key component of ethical data analytics. The concept of uncertainty and error will be a continuing theme throughout this course.

Interpolation and Extrapolation

Interpolation = estimate within the bounds of data

Extrapolation = estimate outside of the bounds of the data





12

We can’t measure everything everywhere. For example, it is common to collect measurements at point locations, such as collecting climate or weather data at weather stations. If you would like to estimate values at locations that were not sampled, you will need to apply interpolation or extrapolation as a means to obtain a continuous surface from your discrete measurements. Since these locations were not actually measured but estimated from nearby sample locations, there will be some error or uncertainty in the product.

Extrapolation is generally considered to be more uncertain than interpolation. As an example, if you are trying to predict a variable y using a variable x with a line or linear relationship, estimates within the bounds of your sample data will likely be less uncertain in comparison to estimates outside of the bounds of your data. Maybe outside of the bounds the relationship is no longer linear, begins to increase exponentially, or levels off. Without data points or samples within these ranges of values, this would be hard to determine.

In a spatial context, extrapolation would be making an estimate outside of the geographic extent of your data, as represented by the purple star. In contrast, interpolation would be making an estimate within the geographic bounds of your data, as represented by the yellow star.

Summary of Key Points

- Data scientists work with data.
- Two-dimensional tables are a common means to store data.
- Data can be categorized based on level of measurement.
- Real data can be complex because the world is inherently intricate.
- Data scientists help make sense of data to help create actionable information to solve pressing problems.

This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.

Thanks! Hope you found this useful.