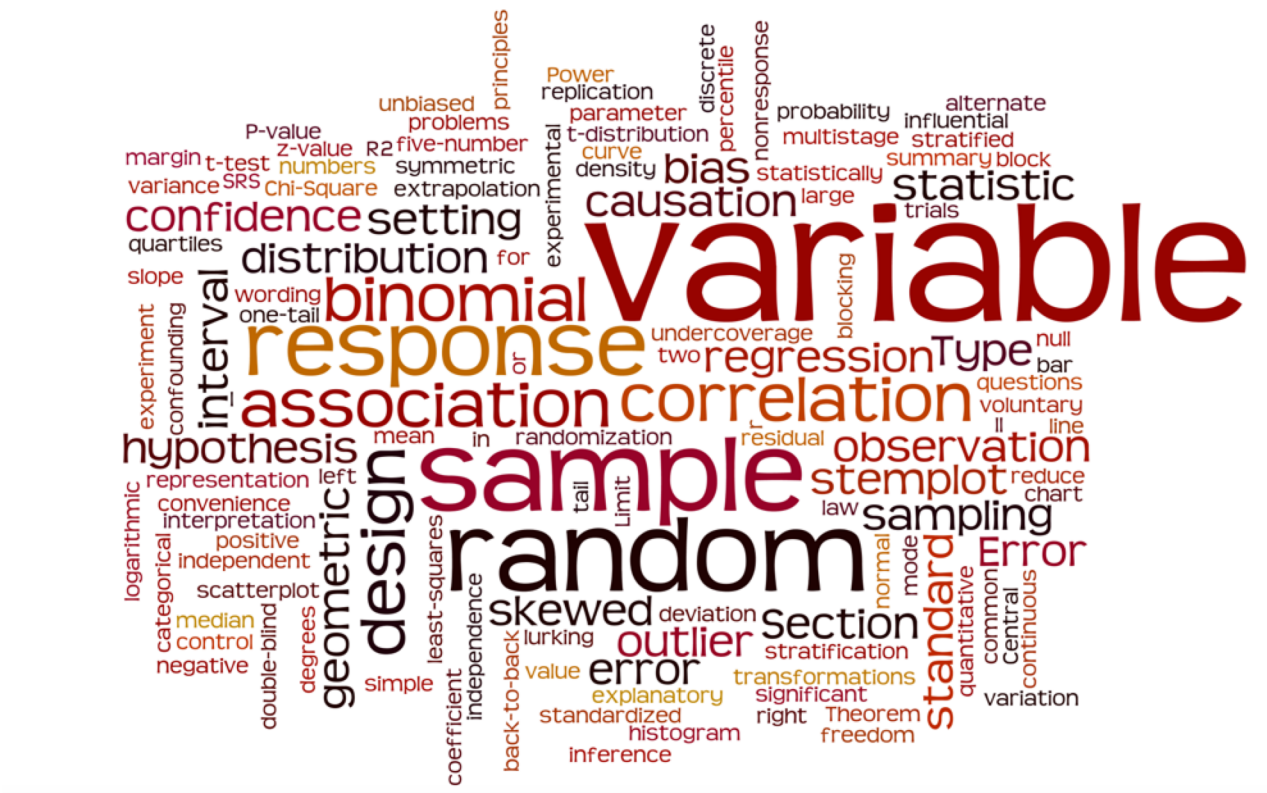
Introduction to Quantitative Methods



Kristin Eccles

Outline

- **Data and data sets**
  - Primary vs. Secondary Data
  - Variables, Individuals, Table/Matrix
- **Data types and levels of measurement**
  - Continuous vs. Discrete
  - Nominal, Ordinal, Interval, Ratio
  - Data standardization / aggregation
- **Uncertainty/Error**
  - Random/systematic
  - Accuracy, precision, resolution
  - Measurement
  - Significant figures
- **Communicating data**
  - Reporting numbers
  - Tables, Graphs, Histograms

Data

- Data collection is important for statistics

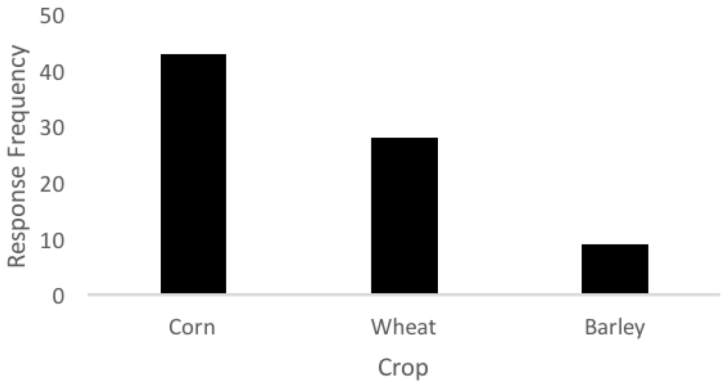
Primary Data	Secondary Data

• Collected from the original source
• Typically time consuming
• Costly (Time= Money)
• Examples: field data, surveys, experiments
• Collected for a specific purpose
  • (We will get into sampling later)
• Fewer potential issues
	• Data collected by another person or agency
	• Less expensive and less time consuming
	• Typically comes in the form of a database
	  • Examples: Census,
	• Potential Problems
	  • Improper collection
	  • Measurement error

Data

- **Qualitative Data**
  - Data about descriptions that can be observed but cannot be computed
  - Classified into two or more categories
    - E.g. Soil types, land use classification, sex
  - Textbook example: 80 farmers are asked to identify their primary cash crop

Crop	Responses
Corn	43
Wheat	28
Barley	9



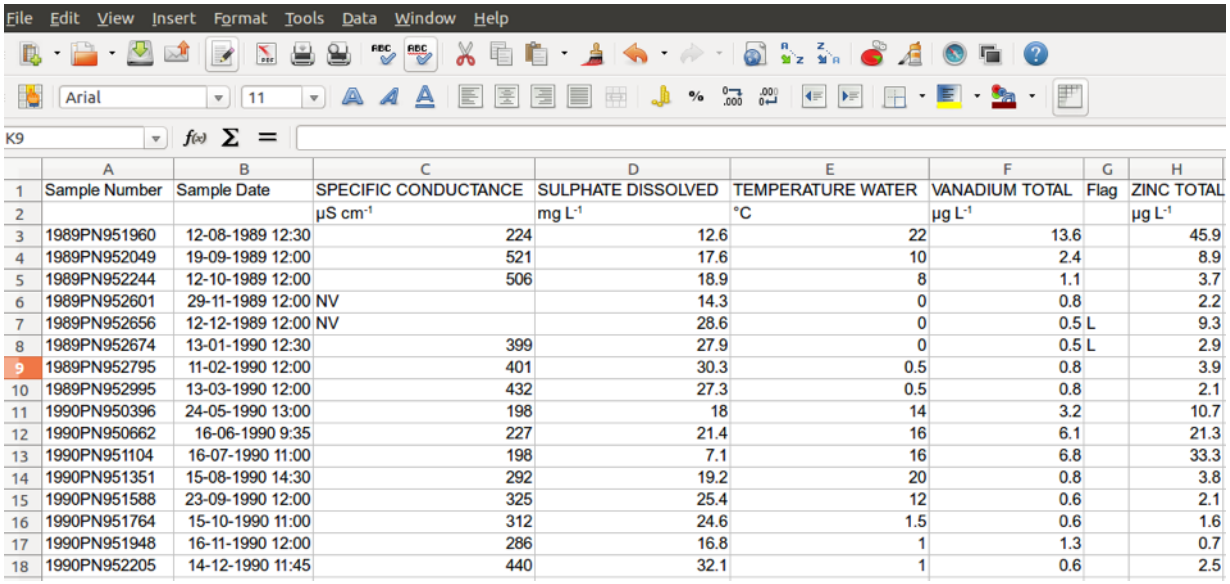
- **Quantitative Data**
  - Data where the observations that are expressed numerically
    - E.g. Temperature, concentration, height, precipitation

Data

- **Variable** – an attribute that is measured
  - Univariate (1), Bivariate (2), Multivariate (>2)
- **Cases, individuals, subjects that are measured**
  - Typically has a unique identifier
- **Values/Data/Observations** – the numbers from each measurements
- **Data Set**
  - Several [related] variables
  - Individuals are in common
  - Forms a matrix (a table)

Data set - Example

- **Variables**
  - Water chemistry
  - Columns (letters)
- **Cases/Subjects**
  - Samples (on different dates)
  - Rows (numbers)
- **Values/Data**
  - Numbers (399...)
  - Or 'NV'
  - Cells (B3:H11)



	A	B	C	D	E	F	G	H
1	Sample Number	Sample Date	SPECIFIC CONDUCTANCE	SULPHATE DISSOLVED	TEMPERATURE WATER	VANADIUM TOTAL	Flag	ZINC TOTAL
2			µS cm^{-1}	mg L^{-1}	°C	µg L^{-1}		µg L^{-1}
3	1989PN951960	12-08-1989 12:30	224	12.6	22	13.6		45.9
4	1989PN952049	19-09-1989 12:00	521	17.6	10	2.4		8.9
5	1989PN952244	12-10-1989 12:00	506	18.9	8	1.1		3.7
6	1989PN952601	29-11-1989 12:00	NV	14.3	0	0.8		2.2
7	1989PN952656	12-12-1989 12:00	NV	28.6	0	0.5	L	9.3
8	1989PN952674	13-01-1990 12:30	399	27.9	0	0.5	L	2.9
9	1989PN952795	11-02-1990 12:00	401	30.3	0.5	0.8		3.9
10	1989PN952995	13-03-1990 12:00	432	27.3	0.5	0.8		2.1
11	1990PN950396	24-05-1990 13:00	198	18	14	3.2		10.7
12	1990PN950662	16-06-1990 9:35	227	21.4	16	6.1		21.3
13	1990PN951104	16-07-1990 11:00	198	7.1	16	6.8		33.3
14	1990PN951351	15-08-1990 14:30	292	19.2	20	0.8		3.8
15	1990PN951588	23-09-1990 12:00	325	25.4	12	0.6		2.1
16	1990PN951764	15-10-1990 11:00	312	24.6	1.5	0.6		1.6
17	1990PN951948	16-11-1990 12:00	286	16.8	1	1.3		0.7
18	1990PN952205	14-12-1990 11:45	440	32.1	1	0.6		2.5

FYI – Data are plural (technically), but often used as a 'mass noun' (singular)

A single data point is a datum

Refs:http://www.ec.gc.ca/data_donnees/SSB-QualEau_OSM-WaterQual/Surface_Water_Quality/Athabasca_River_at_27_Baseline_1989_2010_units_corrected.xls

Data types

Discrete

- Integers
- Restricted set of possible values
- Categories
- Counts
  - e.g., number of fish (count), deformed? (y/n), species of fish (trout, pike, etc)

Continuous

- Real number / Floating point
- Any value possible
  - e.g., temperature, pollution levels, conductivity

Levels of measurement

- How 'detailed' is your information (scale of measurement)
- Levels of measurement:
  - Nominal (lowest)
  - Ordinal
  - Interval
  - Ratio (Highest)
- NOIR (handy mnemonic)
- Conversion between levels possible but **ONLY** higher to lower
- **Important:** – level of measurement determines what analysis methods are available!

Discrete

Nominal

- Lowest level (discrete and qualitative)
- 2 or more categories
- Categories are arbitrary/ unordered
- Typically mutually exclusive
- Yes not > than no

Ordinal

- Discrete categories
- Rank or order to the categories
- Distance between categories is unknown
- E.g. Likert Scale

Like	Like Somewhat	Neutral	Dislike Somewhat	Dislike
1	2	3	4	5

Continuous

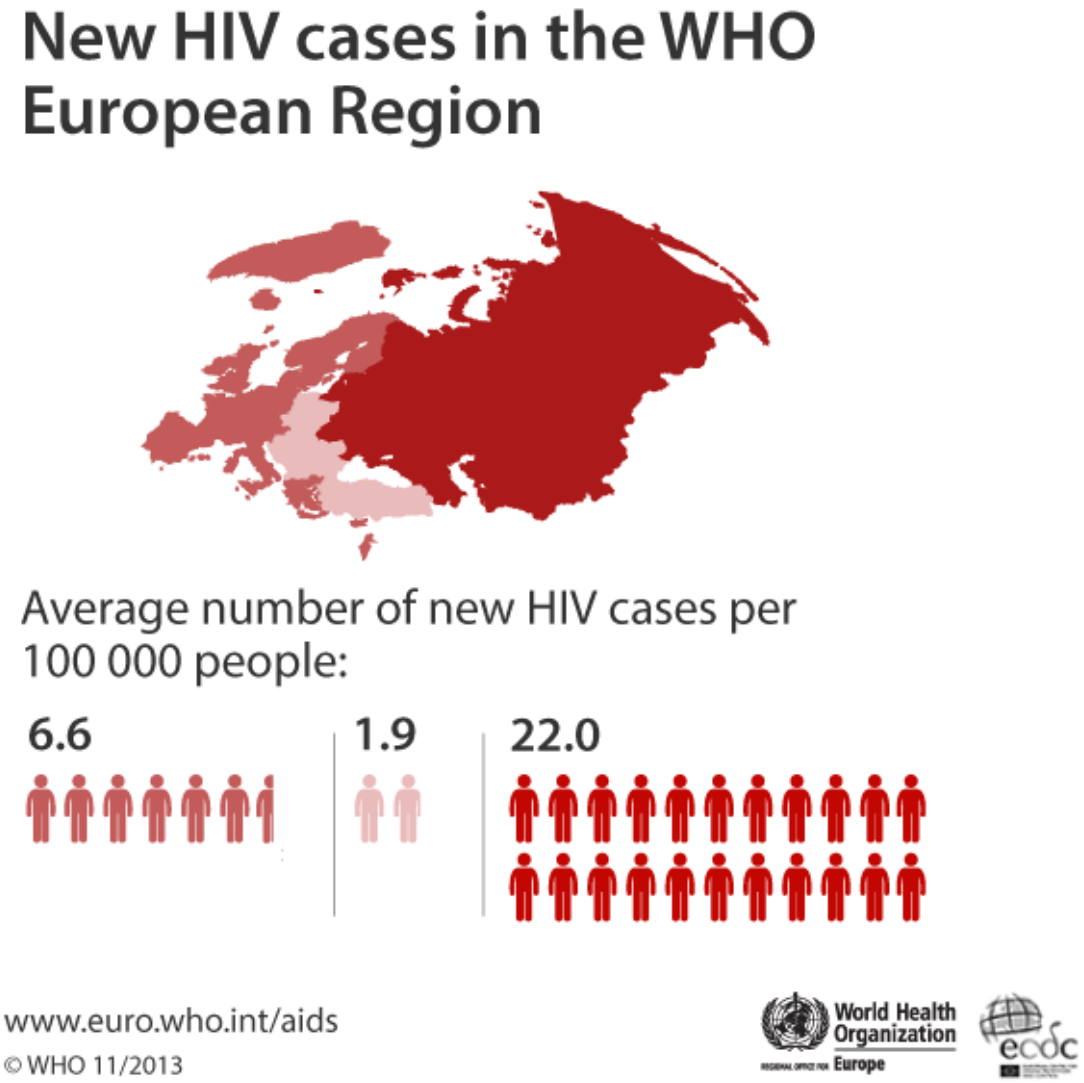
Interval

- Along a scale of measurement
- Zero is arbitrary
- Distance between attributes does have meaning
- Difference between 100˚C - 90˚C = 90˚C - 80˚C
- But what does 0˚C mean?

Ratio

- Highest level (continuous and quantitative)
- E.g. distance, area, elevation, temperature in Kelvin

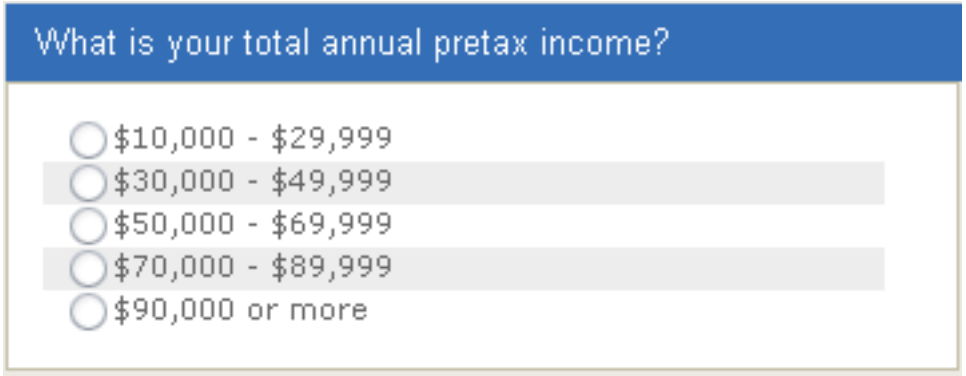
Data Blunders



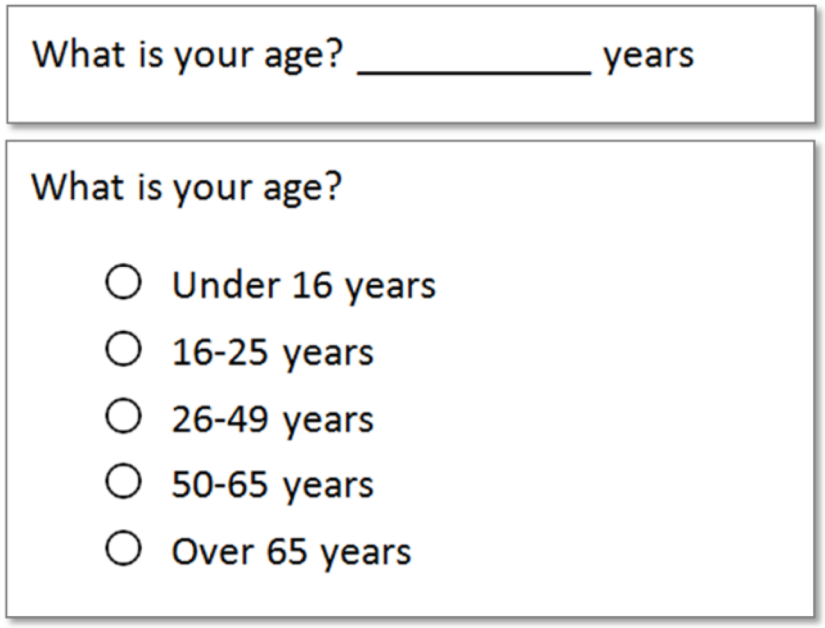
http://www.euro.who.int/en/health-topics/communicable-diseases/hivaids/data-and-statistics/infographic-average-number-of-new-hiv-cases-per-100-000-people-in-the-european-region-2013

“Down grading data”

- Income is a _____ variable?

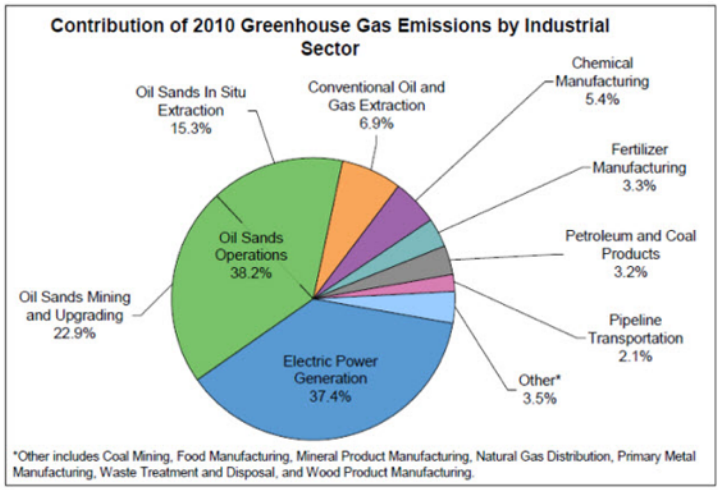


- Age is a _____ variable?



Other data types

- Data can be 'standardized' to give context
- Avoid issues in interpretation
- **Rates**
  - Per [divide by] time, person, country, etc.
- **Proportions**
  - An amount ÷ total
  - Between 0 and 1
- **Percentages**
  - A proportion expressed out of 100
  - Find proportion and multiply



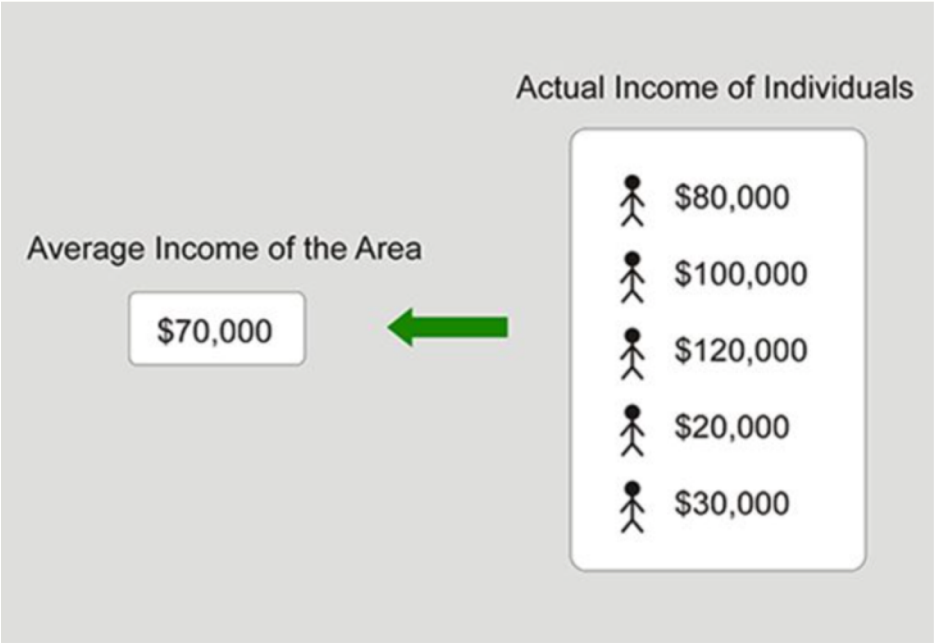
Imgs http://www.prism-magazine.org/mar12/feature_01.cfm http://www.energy.alberta.ca/oilsands/791.asp

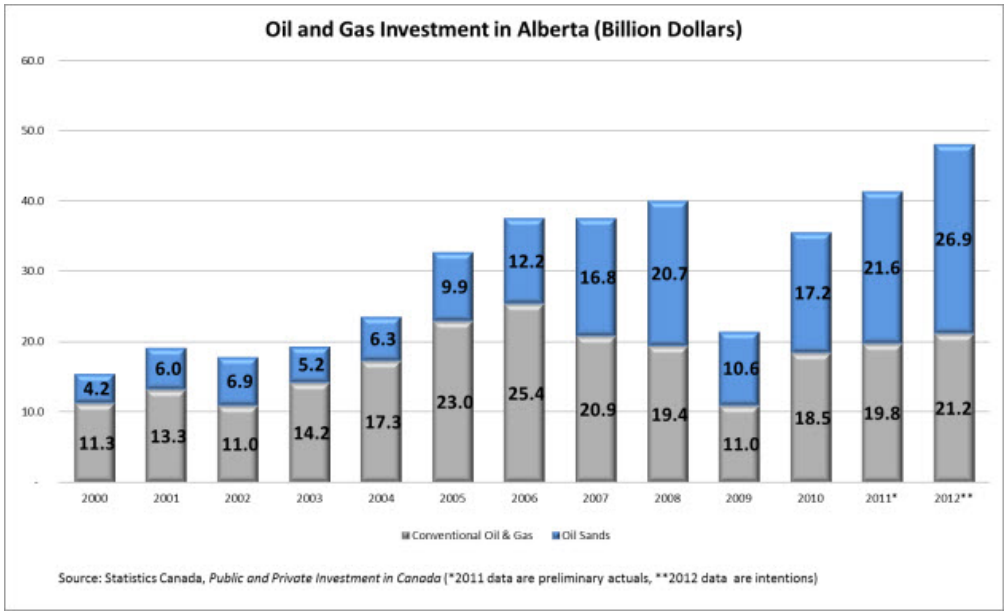
Spatial vs. Aspatial Data

- Spatial data (also termed geospatial data) references something happening on the surface of the earth.
  - Comprised of location information attribute
  - Different ways to encode space:
    - Latitude and Longitude coordinates
    - Postal codes
    - Place names
- Data that cannot be related to the surface of the earth the data is termed aspatial or non- spatial.

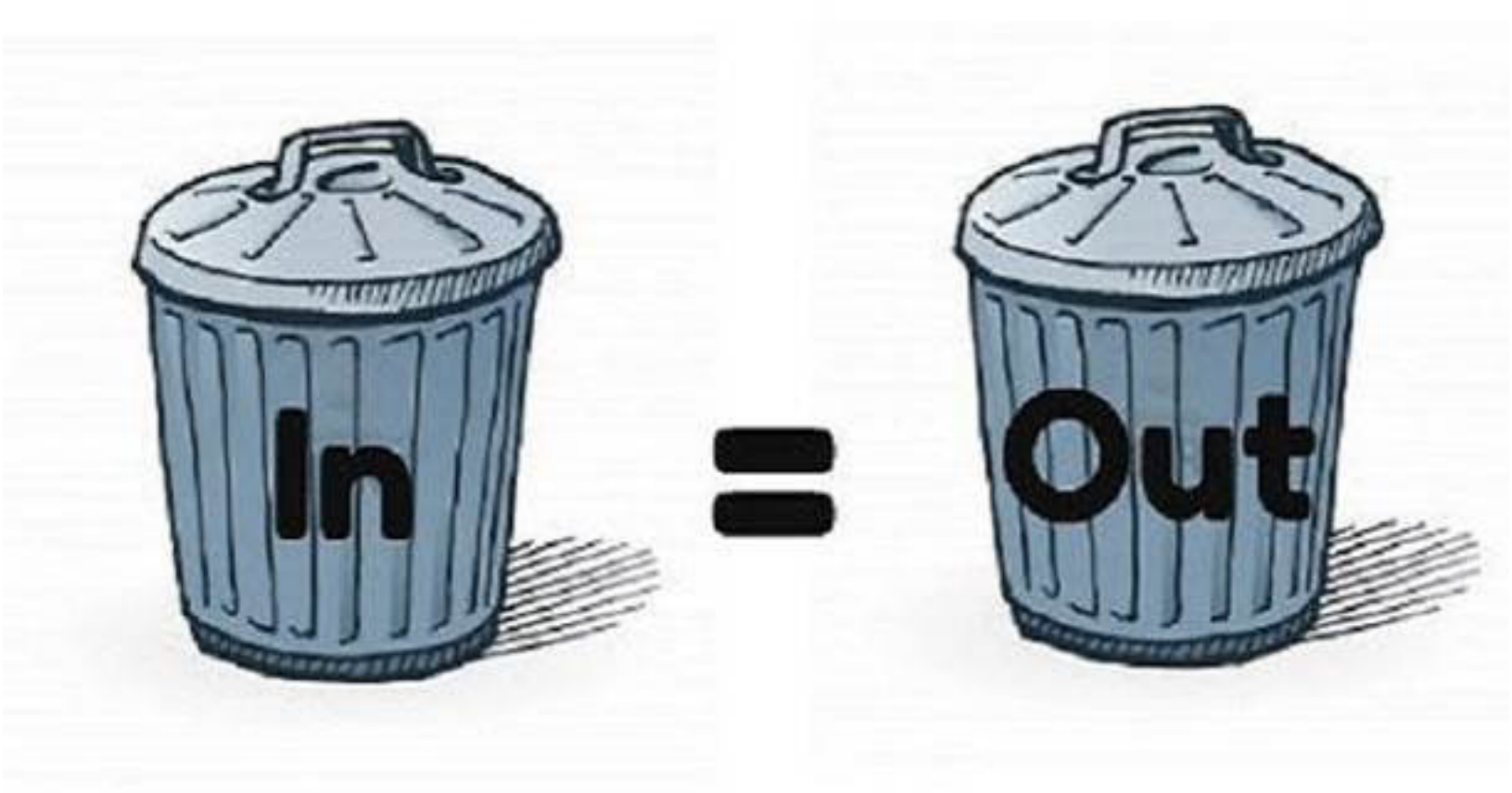
Ecological Fallacy

- Be aware that data may be **aggregated**:
  - Collected from **individual** observations/subjects
  - Then **averaged** (or other descriptive statistic) across space, time, age, etc.
  - Very common: avoids confidentiality issues
    - E.g. Census is collected at individual level but reported by unit of aggregation (dissemination block, census tract)
- This makes for a great summary but you cannot conclude anything about the individual cases!

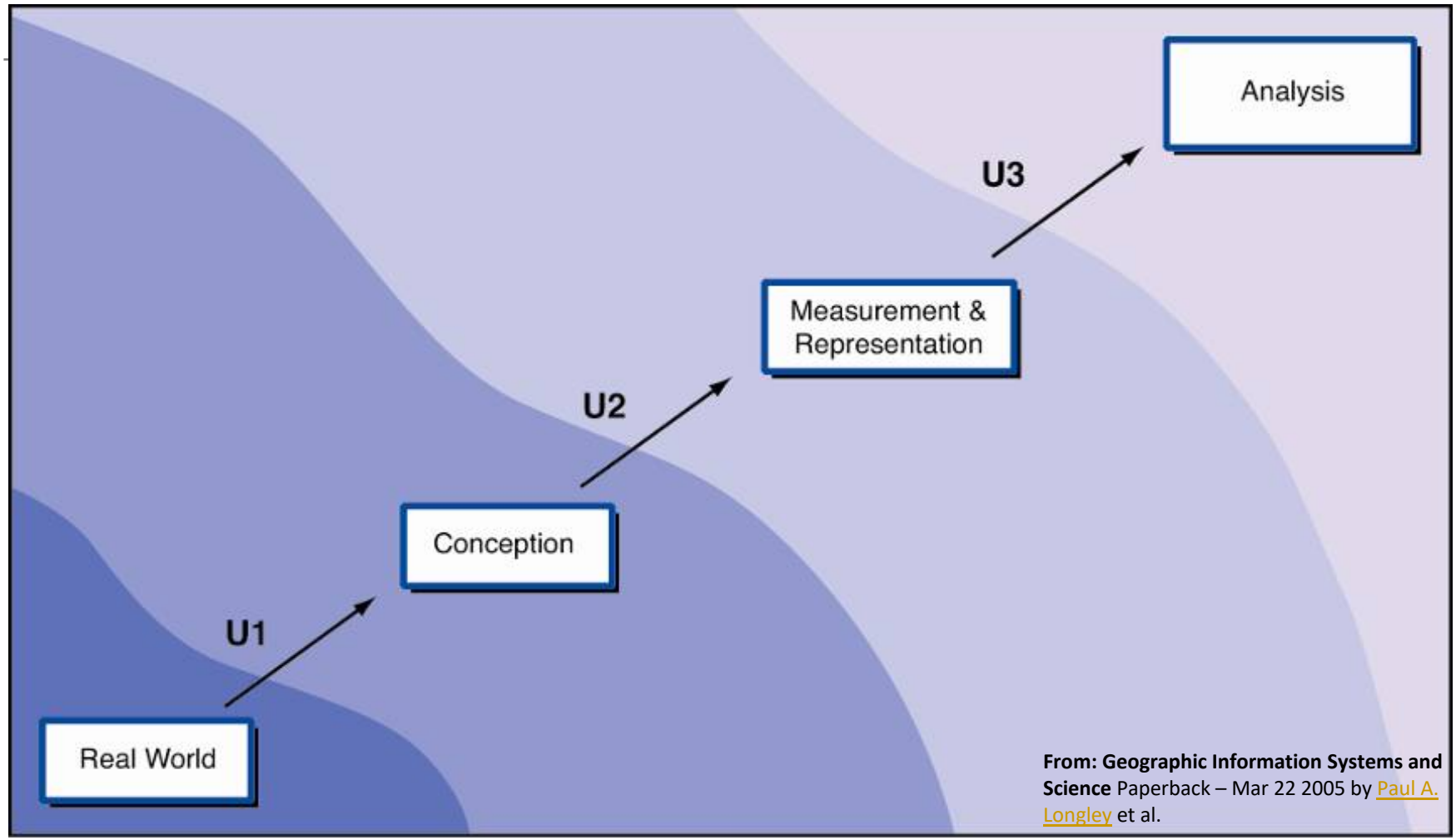




How good is the data?



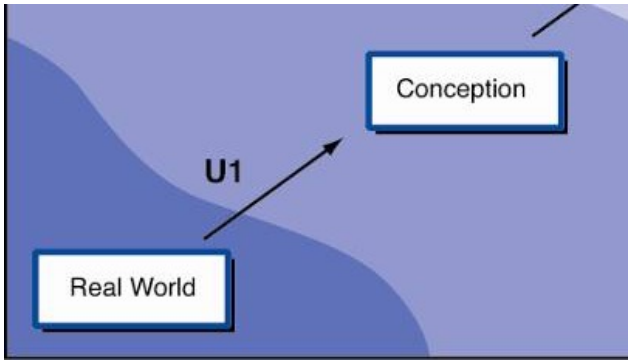
Sources of uncertainty



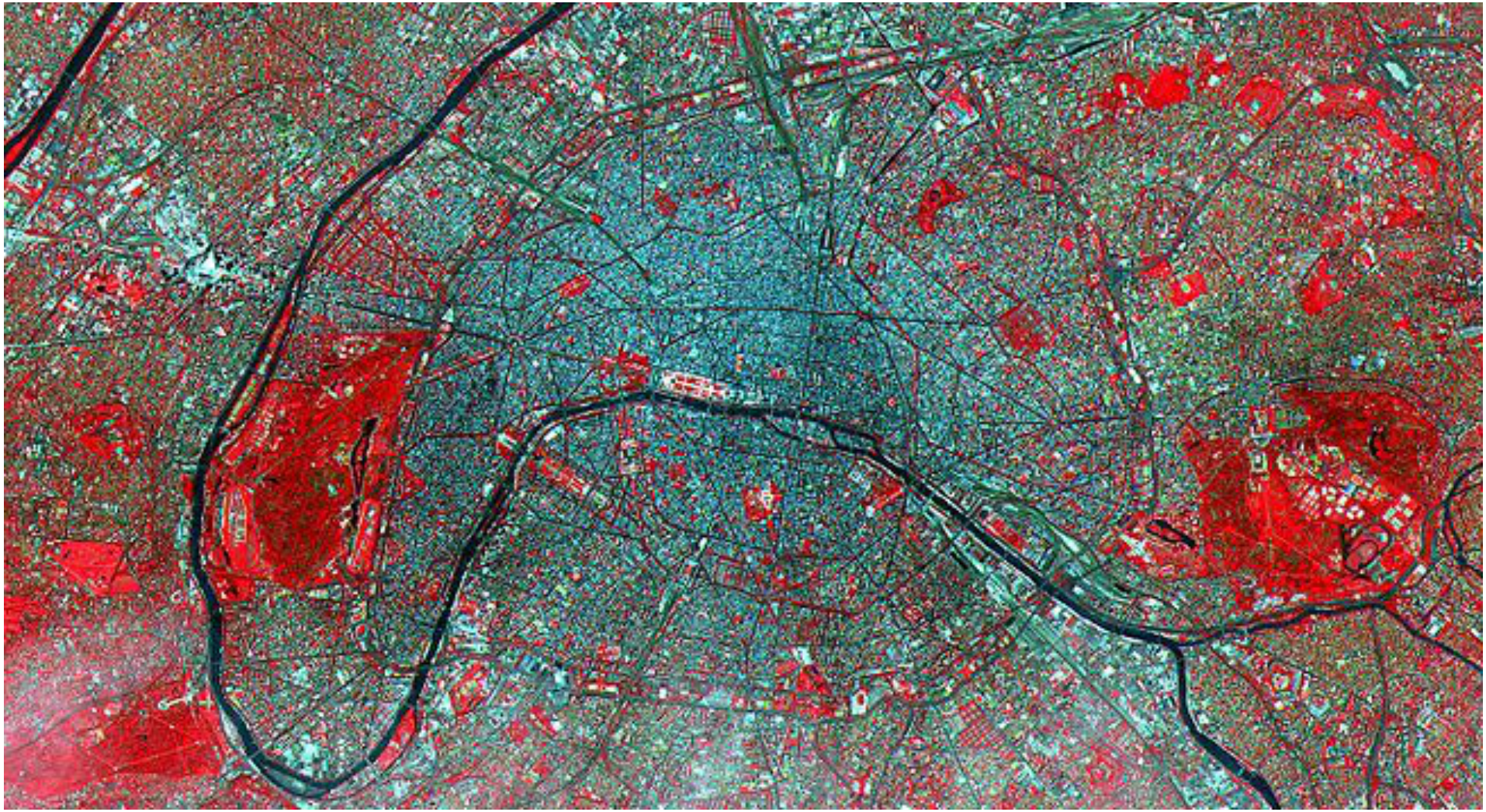
From: Geographic Information Systems and Science Paperback – Mar 22 2005 by Paul A. Longley et al.

A conceptual view of uncertainty- how good is the data we are using?

Uncertainty in Conception



- Conception- the way in which something is perceived or regarded
- How do we make classifications?
  - Human Geography
  - Physical Geograpy
- How does this impact uncertainty?



https://horizon-magazine.eu/article/disused-factories-and-satellites-helping-thwart-urban-sprawl_en.html

Standardized Classifications

TABLE 2.2

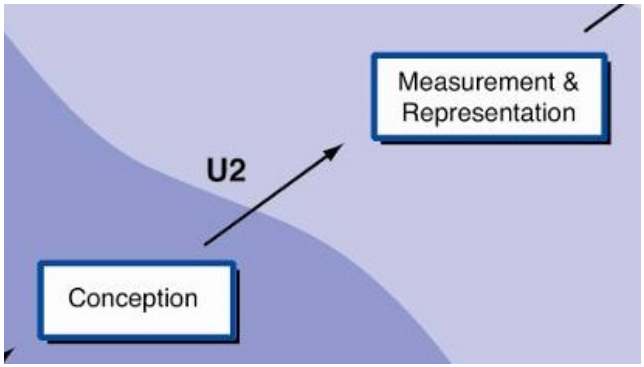
A Portion of the National Land Cover Database System

1. Water		2. Developed		3. Barren		4. Forested Upland		5. Shrubland	
11	Open Water	21	Low Intensity Residential	31	Bare Rock/Sand/Clay	41	Deciduous Forest	51	Shrubland
12	Perennial Ice/Snow	22	High Intensity Residential	32	Quarries/Strip Mines/Gravel Pits	42	Evergreen Forest		
		23	Commercial/Industrial/ Transportation	33	Transitional	43	Mixed Forest		

6. Non-Natural Woody		7. Herbaceous Upland Natural/Semi-Natural Vegetation		8. Herbaceous Planted/Cultivated		9. Wetlands	
61	Orchards/Vineyards/ Other	71	Grasslands/Herbaceous	81	Pasture/Hay	91	Woody Wetlands
				82	Row Crops	92	Emergent Herbaceous Wetlands
				83	Small Grains		
				84	Fallow		
				85	Urban/Recreational Grasses		

Source: United States Geological Survey (USGS), USGS Land Cover Institute

Uncertainty in Measurement



- There is always error:
  - **Systematic** (aka a **bias**)
  - **Random** (sometimes referred to as noise or random variability)
- Causes?
  - Faulty instrument
  - Operator mistake
  - Rounding error
  - Recording error
  - Settling time
  - Interference
  - Etc.

Img: Anna Crawford & Allison Neil – Field course 2013

Accuracy

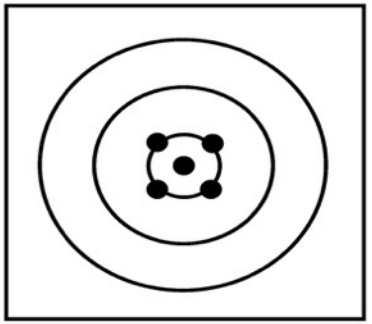
- How close is the measurement to the actual/true value?
- Low accuracy is caused by *systematic* error (**bias**)
  - typically an instrument issue
- Accuracy improves with **calibration** against a standard
  - Offsetting the bias
  - Can be hard to detect (so re-calibrate often)

Precision

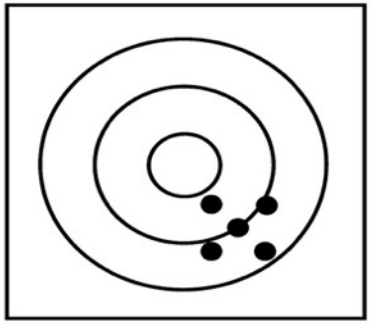
- How well do measurements agree with each other? (repeatable/reproducible)
  - Low precision caused by *random* error
  - Operator issues (typically)
- If you have low precision, better take more measurements (increase the sample size)
- Can also mean:
  - how many **significant digits** a given number has – the **numerical precision**
  - level of refinement in a measurement – **measurement resolution**

Precision vs. Accuracy

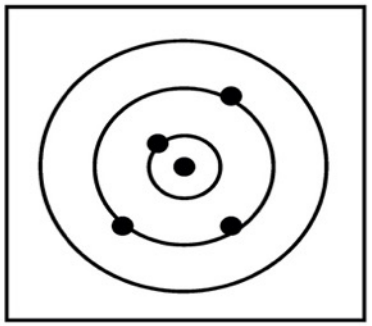
Case 1:
Precise,
accurate



Case 2:
Precise,
inaccurate



Case 3:
Imprecise,
accurate



Case 4:
Imprecise,
inaccurate

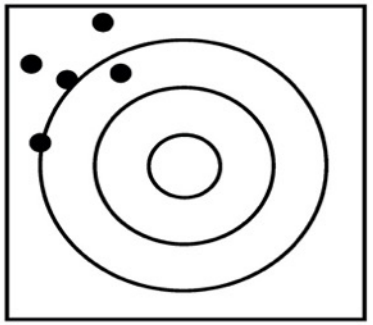
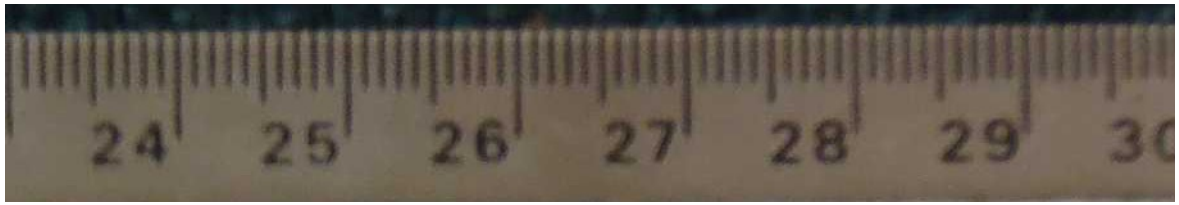


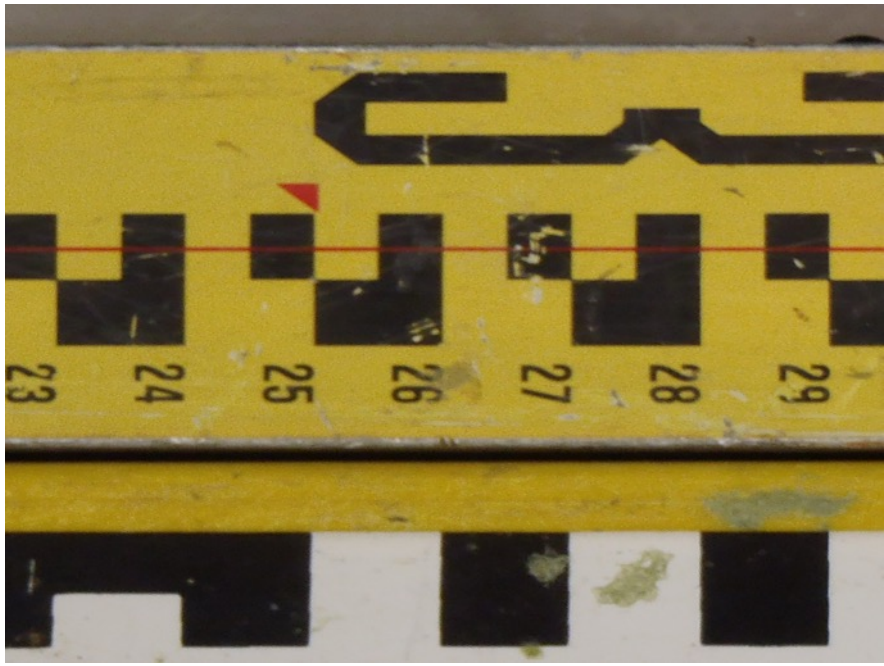
FIGURE 2.1
The Measurement Concepts of Precision and Accuracy:
The Target Analogy

Measurement resolution

- How finely can you make a measurement?
  - Confusingly: also known as **precision**
- The precision of the instrument affords greater numerical precision
- How fine are gradations on a scale?

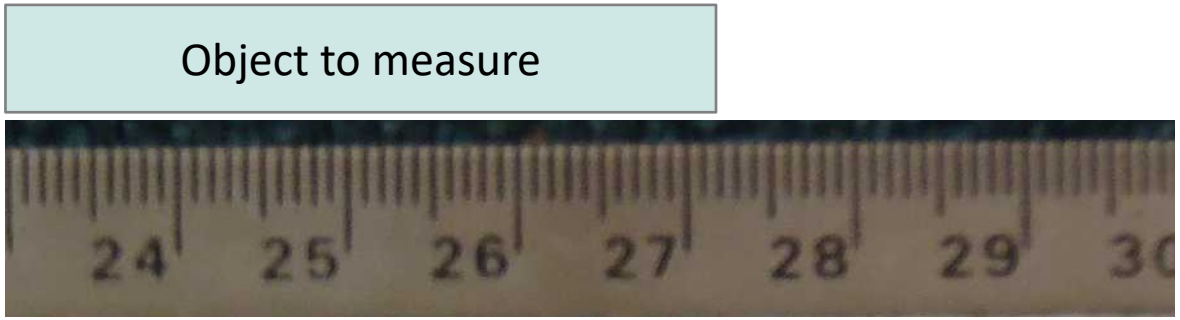
How closely can you determine that 2 measurements are not the same?





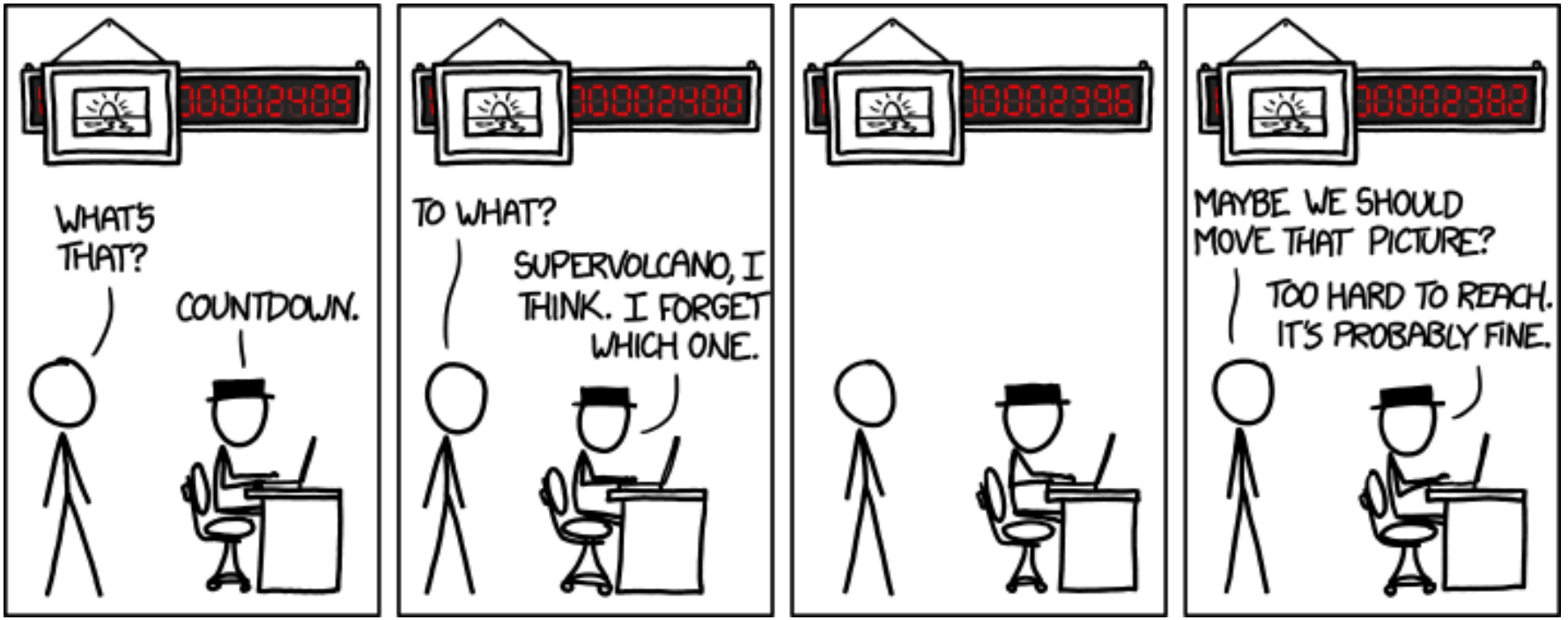
Measurements

- Measuring against a scale (continuous data)
- Pick the closest gradation (i.e., *to the nearest mm*)
- Report this with a unit: length = 272 mm
  - Uncertainty is *implied*:
  - Can be reported **explicitly** as 272 $\pm$ 0.5 mm
  - 271.5 mm ≤ length ≤ 272.5 mm
  - This represents the possible range
- This is an uncertainty!
- Note that uncertainty doesn't mean a "*mistake*"



Significant figures

Important, since they convey meaning...



http://imgs.xkcd.com/comics/countdown.png

Significant Figures/Digits

- 'Sig Figs': The digits that are *known to be correct* in a number *+* ***one more***
  - Example: the object length = 272 $\pm$ 0.5 mm
  - 2 and 7 are absolutely correct, but the last 2 is a bit iffy...
- When you look at a number, which digits/figures are:
  - Important because they convey meaning?
  - Act as a place holder only?
    - Are simply there because of over-precision?
      - i.e., they convey a false precision...
      - 15/35 = 0.428571428

Which digits are significant?

- **Rules**
- **YES**: Zeros between non-zeros: 400,003 or 400.008
- **YES**: Trailing zeros on the right of the decimal point: 2.33000 or 45.0
- **NO**: Zeros that set-off a decimal point on their left: 0.00034 or 0.023 (3.4×10^{-4} or 2.3×10^{-2})
- ***Likely*** **NO (but possibly YES):** Trailing zeros to the left of the decimal point: 450 000 or 7 000 000 (4.5×10^{5} or 7×10^{6})
- BUT – sometimes the last zero **IS** significant; in which case you need to use common sense (or be told) Example: “I used a metre stick and the whale measured 10 m long”

Reporting with correct sig. figs

- GENERAL PRINCIPLE: *Don't introduce spurious precision*
- Calculate with extra digits and *then* round to report
- **x & ÷** Can have no more **sig. figs** than carried by the least precise *measured* number
- **+ & -** Can have no more **decimal places** than are in the *measured* number with the fewest decimal places
- What is a *measured* number?
  - What was measured (using an instrument likely)
  - It is not:
    - a conversion factor
    - a mathematical constant
    - an integer count (these are 'exact' numbers)

Rounding

- **Rules:**
- If the first digit you need to remove is 5,6,7,8 or 9, round the digit to the left *up* (round up)
- If the first digit you need to remove is 0,1,2,3 or 4, leave the digit on the left *as is* (round down)
- **Once in a while:**
  - If the first digit you need to remove is a 5 followed by zeros? Then you are exactly half-way between 0 and 1 and you have a tie!
    - Round up if the number on the left is odd
    - Round down if the number on the left is even

Why worry about uncertainty?

- Suppose you measure the temperature in two places with two thermometers with different precision (±1˚C, ±2˚C)
  - River 1 = 21˚C
  - River 2 = 22 ˚C

  - Which is warmer?

Reporting values

- Use appropriate significant figures
- Follow with the correct unit – metric *much preferred*
- Space between number and unit – *except* when the unit is not a letter (e.g., % or º)
- Make sure super- and sub-scripts are correct (preferred over division: m s^{-1} not m/s)
- Examples:
  - The road was 4.5 km long and 3 m wide
  - His temperature was 41ºC and so he missed the final exam
  - The wind was 24.3 m s^{-1} during the storm

Displaying data (Table vs Graph)

Table	Graph

• Hard to absorb info
• Best when data are difficult to graph:
  • Technically
  • Too busy
• Or when a graph would be trivial (2-3 points)
	• Visual and impressive
	• Can readily grasp info
	• Picture = 1000 words
	• Can display large amounts of data (if done properly)
	• Not easy to see exact values

Captions

- Explains the figure or table so that it can '*stand on its own*':
  - Start with a brief title
  - Then an explanation (as required)
  - More detail about what is displayed
  - Explain symbols (if they are not in a legend)
  - What is the 'take home' message of the graph/table in words?
  - Then a source (required, if you didn't measure the values yourself)
- Figure captions **go *below*** a figure
- Table captions **go *above*** the table

Tables

- Basically like a data matrix
- Variable name and **unit** at the top of the column (1st row)
- Observation/Subject/Row name in the 1st column
- Horizontal line – above and below column header and at the end of the table (no grids, fancy backgrounds, etc)
- Black and white!

Table 3. Water quality measurements from 1989 to 2010 in the Athabasca River, downstream of the Oil Sands development. Source: Joint Oil Sands Monitoring, Environment Canada, 2010.

Date	Conductivity (µS cm^{-1})	Temperature (°C)
1989-08-12	312	18
1990-08-15	285	14
1991-01-15	348	1

Different types of graphs

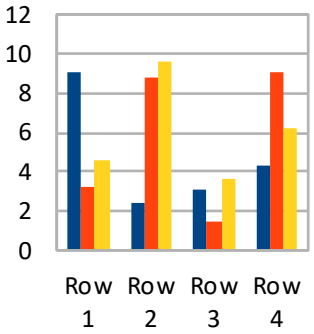
Scatterplot – shows relationships (*interval/ratio*)

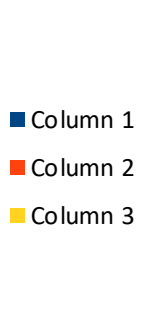
Bar (vertical or horizontal) – compare amounts across categories (*nominal*), display frequencies/histograms (*ordinal* categories)

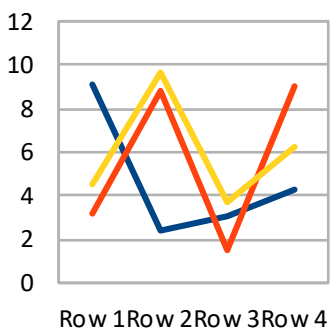
Line – direction of change across categories, time series (*ordinal* categories)

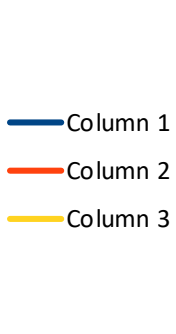
Pie – show proportions (*nominal* categories)

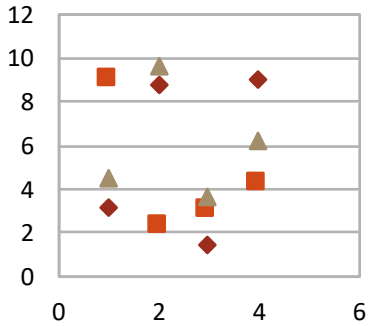
Many others...

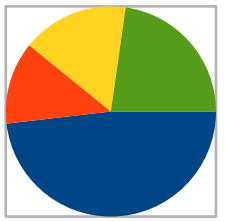


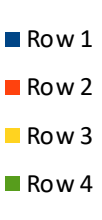












Example graph

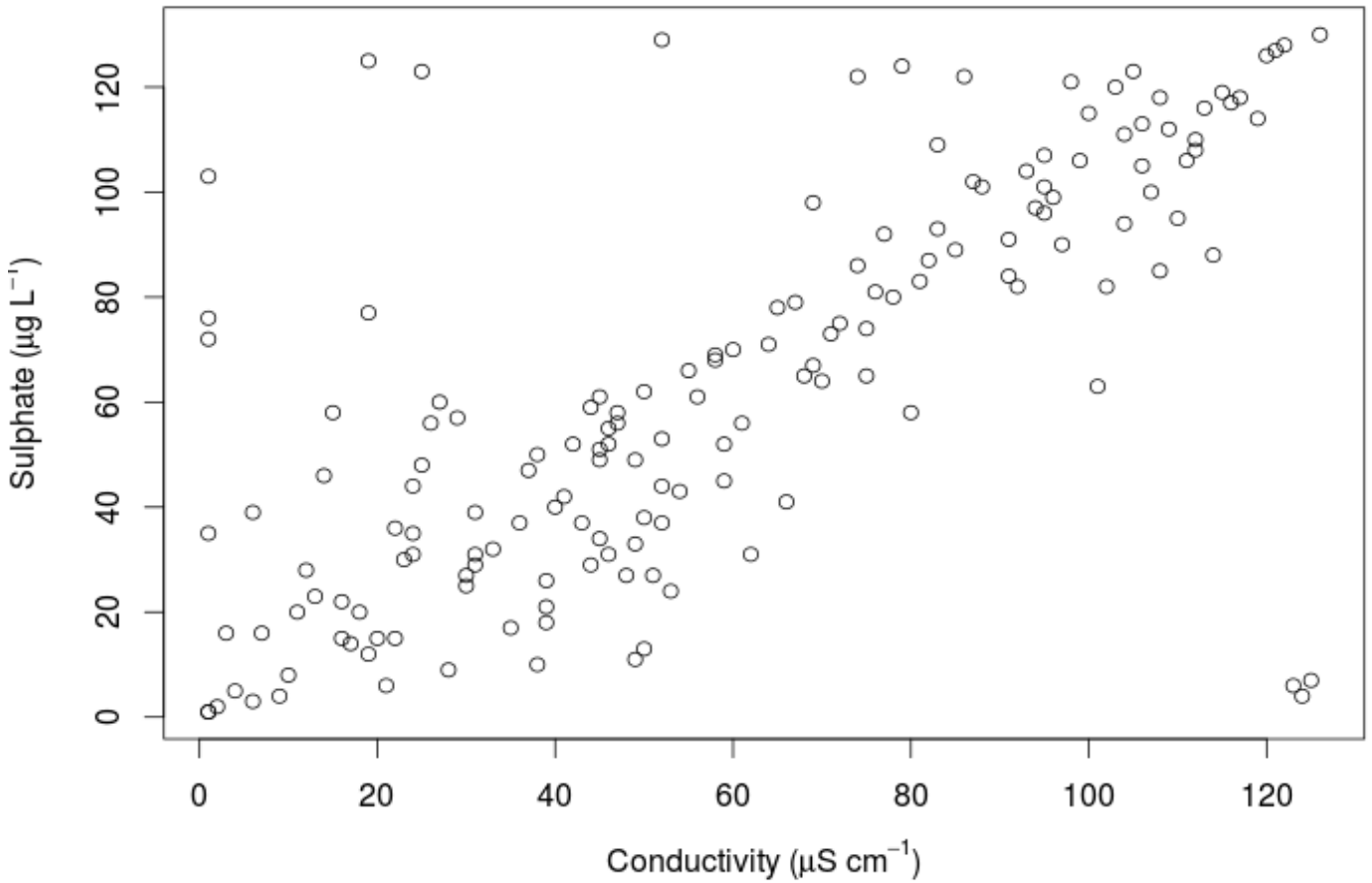


Figure 1. The relationship between conductivity and sulphate concentration from 1989 to 2010 in the Athabasca River, downstream of the Oil Sands development. Source: Joint Oil Sands Monitoring, Environment Canada, 2010.

Parts of a graph

- Axes:
  - X-axis (abscissa)
  - Y-axis (ordinate)
- Scale
  - Tick marks
  - Values
- Symbols (for data points)
- Lines (connect points)
- Labels (see over →)

Labels

- Axis titles/labels = Variable (unit)
- ~~Main title~~ (*No* title! Use a caption)
- Legend (for >1 variable)
- Data labels

Considerations

- Grid lines – almost **never** required
- Second y-axis – only if it helps
- Lines? - ok if you are looking at local rates of change, and there is some rationale for connecting symbols (time series)
- Scale adjustment:
  - This can emphasize certain ideas (*subjective*)
  - Zoom in (restrict scale) – make trends look larger
  - Zoom out (expand scale) – diminish trends and differences
  - In general – try to fit the scale to the data range BUT include 0 (at the origin), if appropriate

Histograms

- **Purpose**: To look at a variable and see how data are distributed
- Define 'bins' or classes of (typically) equal size
  - Note: changing the **number of bins** will change the look of the histogram
- Count the number of occurrences per bin
  - How frequently the data fall into a given bin
- Plot a bar graph with the bins as categories and the frequencies (counts) on the y-axis

Example of a histogram

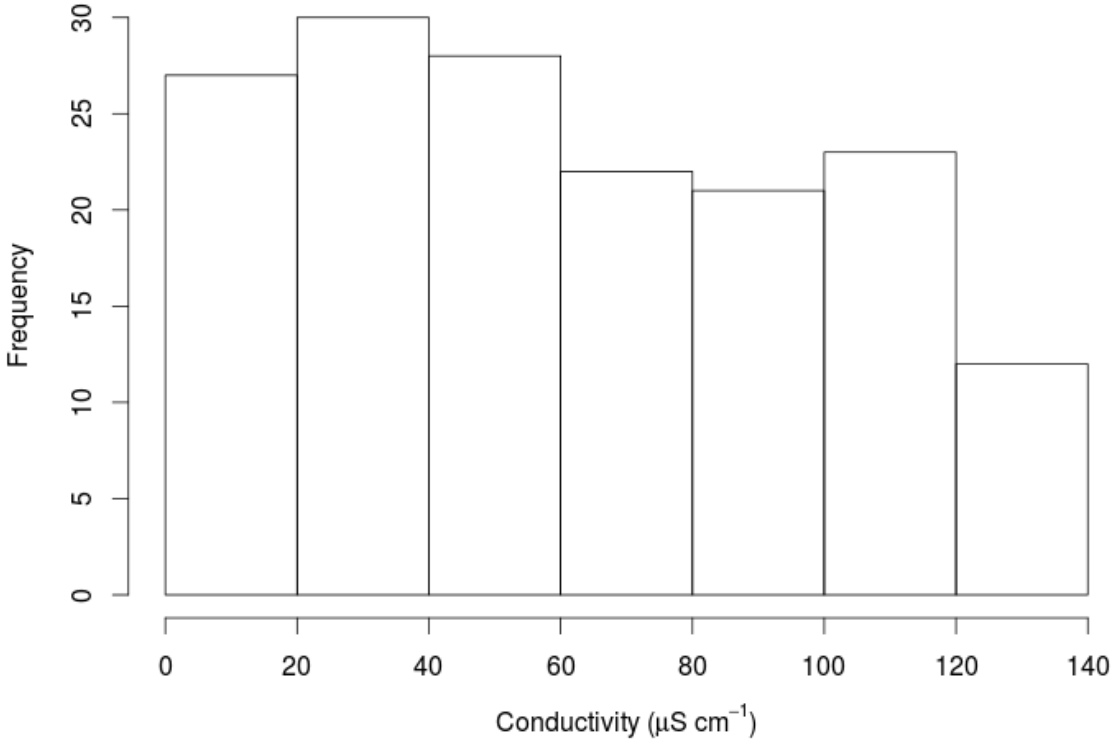


Figure 2. Histogram of conductivity from 1989 to 2010 in the Athabasca River, downstream of the Oil Sands development. Source: Joint Oil Sands Monitoring, Environment Canada, 2010.

Summary

- Type of data (nominal, ordinal, interval/ratio) determines what statistics you can do
- An awareness of the uncertainty in your data will prevent inappropriate conclusions
- Communicate data in the most efficient and unambiguous way possible! Follow conventions!