Module 3: Understanding Data

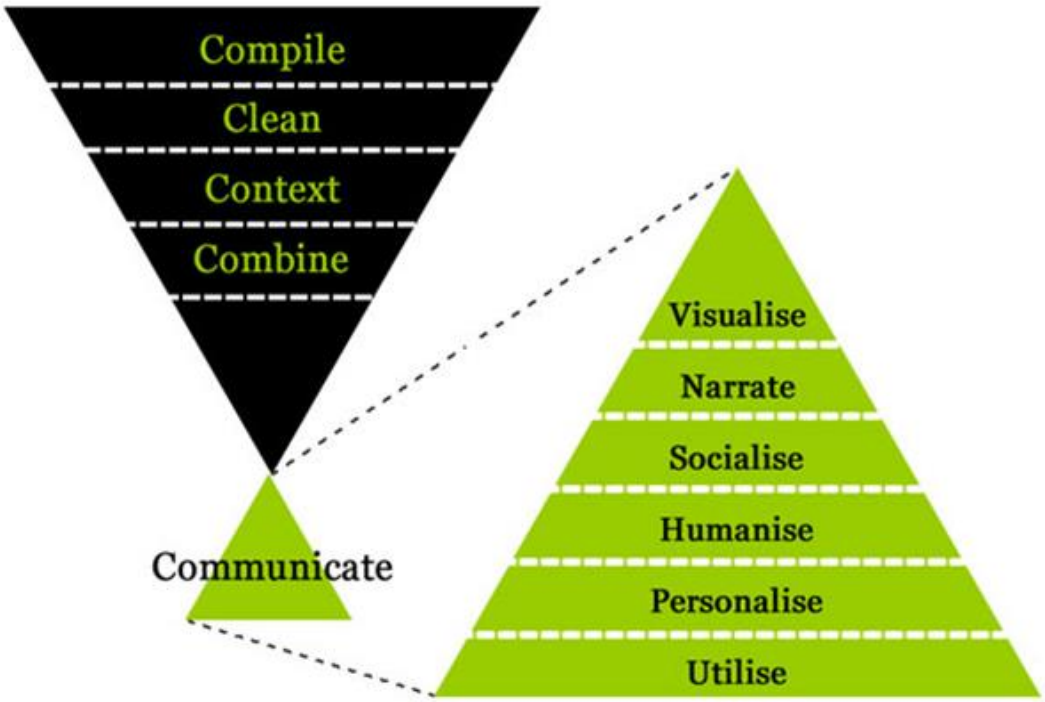
STUDENT WORKBOOK

Key to producing solid data driven analysis is the ability to evaluate data quality and apply basic statistical principles for accurate interpretation. This module will introduce basic concepts of data organization and cleaning as well as questions to help us evaluate the source of the data. Next we will look at basic calculations that can transform numbers into ratios, comparisons and rounded figures that audiences can more easily understand. Next, we will cover essential statistics to ensure data is interpreted correctly and that we recognize data manipulation. Finally we will review principles of data privacy.

At the end of this module, you will be able to:

- Organize a data set for analysis
- Determine whether the source of a data set is trustworthy
- Simplify data into a way audiences can relate to
- Explain basic statistical principles
- Evaluate manipulated data
- Explain the challenges of data privacy

Contents



Lesson 1: Organizing Data

Before beginning data analysis to help answer our hypothesis and questions, we have to be able to understand the information we have. Data is organized under a specific set of standardized rules to make it easier for us to see our data. In our work, we will mostly be working with data tables in spreadsheets, not databases, but many of the same organizational principles apply.

A data table is a spreadsheet that's organized in such a way that the human eye can understand. To reach a conclusion, you can analyze the data table as a whole instead of going row by row.

A database is a dataset organized into columns with each discrete data record in a different row. This organization system allows a computer to analyze the data and recognize similarities, allowing you to reach general conclusions about the data.

Each column head is labeled by the category of data it contains and each row is a separate record. Each column indicates the type of data in that row, whether it be names, ages, gender, organization, etc.

Top Common Mistakes in Data Reporting[1]

Mistake No. 1 – Don't overestimate the meaning of your data
Before even opening the file, data reporters should think carefully about the potential limitations of a data set, and what the data can and cannot tell you about a topic.
"I think the tendency that a lot of people who are beginning to do this sort of work have is [to think] that humans are fallible but numbers are ironclad," Mussenden said. "The data is only as good as how it's collected."

Mistake No. 2 – Not checking the file type
Knowing both the type and size of a data file will help you decide which programs to use to work with it.
Mistake No. 3 – Not cleaning the data first
You've likely waited so long for the data set that it's tempting to jump right in and get to work, but the first few hours of most data projects should involve cleaning up the data to make sure it's usable.

Mistake No. 4 – Not indexing your fields
Most data sets will come organized in a meaningful way, whether it be alphabetically, by date or something else. But while usually not intentional, the way the data set is organized when you get it is rarely the best way to spot the trends you're looking for. The sort feature in Excel is a powerful way to reorder and analyze your data, but can mess it up beyond repair if you don't "index" your fields before sorting. To avoid the headache, create a new column to the far left or far right of your data and label it "index." Then, fill in numbers starting at 1 and counting up through the end of the rows. To undo your sort, just sort by this new column from smallest to largest and Excel will put your data back the way it was.
Mistake No. 5 – Assuming you know what the field names mean
Regardless of how simple or complex your data set seems, always request the "data dictionary," a list of all of the fields in your data set, their names and what type of information is in there, such as dates, numbers or phrases. If there isn't a data dictionary, call the agency or office it came from and ask to talk about the fields with whoever maintains the database or file.

Mistake No. 6 – Not saving each major change as a new copy
Data analysis is one of the only types of reporting in which you can lose all of your hard-fought victories if you hit save after making a mistake. To avoid data catastrophes, make a folder for the project and label each subsequent version of the data with a number and the date (for example CrimeStatsOriginal, CrimeStats1_June20).

Mistake No. 7 – Doing too much at a time
Data analysis can be extremely difficult to double check. With that in mind, it's important to work slowly and take frequent breaks. Mussenden prefers to break for at least 10 minutes per hour, but you'll find your own pace and rhythm as you go.

Mistake No. 8 – Not involving your editor
Your editor probably doesn't sit in on your interviews and most likely doesn't want to watch you shift columns in Excel for an hour, but they do need to have an active role in any data-driven project you do.

[1] https://www.americanpressinstitute.org/publications/data-reporting-common-mistakes/?utm_content=bufferf5c44&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Successful data reporters should take detailed notes on what they do each day, both for their own benefit and in case an editor wants to review their progress. It also helps tremendously to "logic check" your workflow with a trusted colleague to make sure you aren't missing anything obvious or jumping to conclusions.

Mistake No. 9 – Treating visualizations like an end goal
Most data projects start off with a standard list of calculations such as finding means, medians, ranges and minimums and maximums in the data set. But after crunching the easy numbers, it can be tricky to tell which direction to explore next. Create some easy visualizations, like graphs and charts within the Excel program, to help spot patterns that can lead to story ideas or more questions.

Mistake No. 10 – Not knowing when to ask for help

Data Standardization

When working with a database, information can come from different sources, have incomplete fields, come in different structures, and include errors such a double entries or misspellings. This complicates the analysis process and although we recognize the mistakes, the computer does not know how to handle them.

Data standardization or cleaning is the process of cleaning the data and is an important step for data journalism.

One of the first steps in cleaning data is to ensure that all the column headers are correct and complete and that the data type in each row matches the column header.

Many data cleaning processes allow us to clean the entire database with the same set of tools.

If the database has addresses, dates, ages, measures, the first step is to decide a standard way to enter these fields into the database.

For example, this is a non-standardized date column

Date
12 February 2012
12/2/2012
2/12/2012
12/2/12
12/feb/2012

There is not a correct format as long as all of the dates are formatted the same and the computer understands the date format. It's important to choose a format that is most convenient for the whole database. In this case we decided on DD/MM/YY. So now our data after being cleaned looks like this:

Date
12/2/2012
12/2/2012
12/2/2012
12/2/2012
12/2/2012

Each date record is now in the identical format: DD/MM/YY.

The basic principle is to ensure that all the data is entered in the same format, often in all capital letters, without any extra spaces and has an index number.

For example, here is a data set:

Non-Standardized Data:

Name	**Date of Birth**	**Address**	**Salary**
Pancho Juárez López	16 April 1975	41 12th ave. zone 23	$1250
Luis Pérez Almódovar	31/5/1980	6 Fifth avenue Zone 12	US$1000

Standardized Data:

No	**NAME**	**DATE OF BIRTH**	**ADDRESS**	**SALARY (USD)**
1	PANCHO JUAREZ LOPEZ	16/4/1975	41 12TH AVENUE ZONE 23	1250
2	LUIS PEREZ ALMODOVAR	31/5/1980	6 5TH AVENUE ZONE 12	1000

There are Excel functions and special programs specialized in data cleaning. Excel filters, search and replace, trim and other functions are sufficient for basic data cleaning. For more advanced cleaning, Open Refine has features that can handle even the messiest data set. We will review these tools in the labs.

Lesson 2: Verifying Data

J13

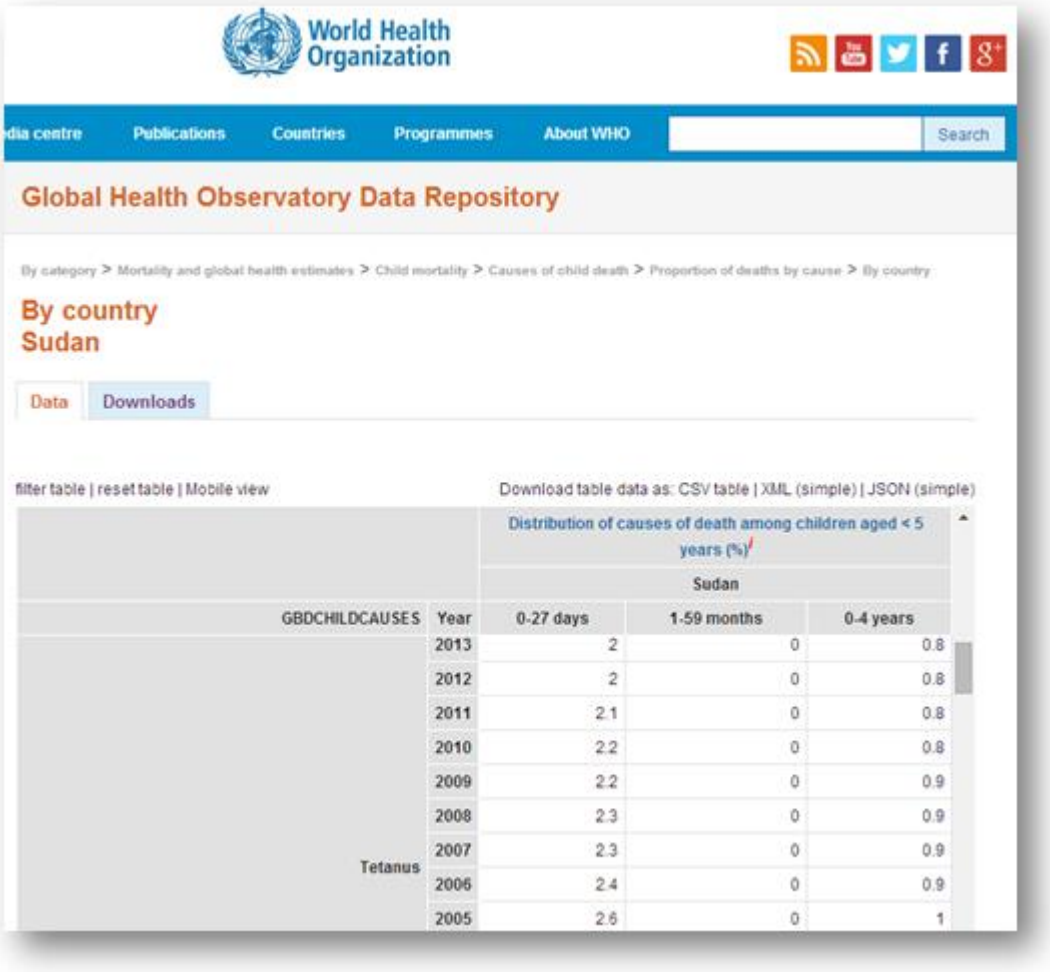
	A	B	C	D	E
			Distribution of causes of death among children aged < 5 years (%)	Distribution of causes of death among children aged < 5 years (%)	Distribution of causes of death among children aged < 5 years (%)
		Sudan	Sudan	Sudan	
	GBDCHILDCAUSES	Year	0-27 days	1-59 months	0-4 years
	HIV/AIDS	2013	0	0.8	0.5
	HIV/AIDS	2012	0	0.7	0.5
	HIV/AIDS	2011	0	0.8	0.5
	HIV/AIDS	2010	0	0.8	0.5
	HIV/AIDS	2009	0	0.8	0.5
	HIV/AIDS	2008	0	0.8	0.5
	HIV/AIDS	2007	0	0.8	0.5
	HIV/AIDS	2006	0	0.8	0.5
	HIV/AIDS	2005	0	0.8	0.5
	HIV/AIDS	2004	0	0.8	0.5
	HIV/AIDS	2003	0	0.7	0.5
	HIV/AIDS	2002	0	0.7	0.4
	HIV/AIDS	2001	0	0.7	0.4
	HIV/AIDS	2000	0	0.6	0.4

- Sources of data?
- Indicators?
- Units of measure?

Data spreadsheets often hold a wealth of information in a very compact format. Before analyzing data, it is important to understand what the data is measuring and what all descriptions, labels and other contextual information means to ensure a correct interpretation of the data.

Often sector-specific data is produced by specialized professionals who use a lot of jargon and abbreviations to save space in data files. By doing a bit of research and following best practices when it comes to citing the source of the data, it is much easier to understand the data in context. Many datasets come with a codebook or a glossary that explain all the data labels and measurements.

Example: Under Five Causes of Death



World Health Organization

dia centre | Publications | Countries | Programmes | About WHO | Search

Global Health Observatory Data Repository

By category > Mortality and global health estimates > Child mortality > Causes of child death > Proportion of deaths by cause > By country

By country
Sudan

Data | Downloads

filter table | reset table | Mobile view

Download table data as: CSV table | XML (simple) | JSON (simple)

		Distribution of causes of death among children aged < 5 years (%)		
		Sudan		
GBDCHILDCAUSES	Year	0-27 days	1-59 months	0-4 years
Tetanus	2013	2	0	0.8
	2012	2	0	0.8
	2011	2.1	0	0.8
	2010	2.2	0	0.8
	2009	2.2	0	0.9
	2008	2.3	0	0.9
	2007	2.3	0	0.9
	2006	2.4	0	0.9
	2005	2.6	0	1

Take a look at this dataset, which documents the causes of death of children under the age of five in Pakistan: http://apps.who.int/gho/data/node.main.COCD?lang=en

We have three different ways to measure child mortality;

a. Number of deaths by cause

b. Rate of deaths by cause

c. Proportion of deaths by cause

Please indicate which data set would be appropriate to answer each of these questions?

1. Over the last 10 years, is each of the top five diseases becoming a bigger or smaller threat to child health?

2. What is killing the largest percentage of children?

3. Does the budget have enough money to treat each of the children who die of these top five diseases?

This is a commonly used dataset to help evaluate the health of children in a country. When looking at a raw dataset like this it is important to thoroughly review all the information before doing any analysis. This section walks through some questions to ask about the data before using it.

Click on **Rate of deaths by cause** and click on the **Download** tab. Select the CSV file with the **list containing text, codes, and values.**

Sources of Data	Understanding the Indicators	Units of Measure
Data Questions: • What organization produced this data? • Where did the organization source the data from or are they the original source? • Can I find an explanation of the data? • Is there a link to data on the spreadsheet? • How old is the data?	**Data Questions:** • What do the indicators mean? • Can I look up definitions of indicators I don't understand • What are the differences between the age categories? • What indicators are not included in this data that would provide more context?	**Data Questions:** • What do the numbers mean? What is the unit of measure? • What is the difference between a rate or a percentage? • Is this data available using other measures from another source?
Public Service Questions • Does the data come from a trusted source? • Is the data current enough to be relevant? • Can I find more information about the data source?	**Public Service Questions:** • What would the public want to know about this data? • Do the indicators answer the questions I want to ask? • What other information would explain the data	**Public Service Questions:** • Does the unit of measure accurately put the data into context? • Does the unit of measure help the public gauge the risk? • What text do I need to explain the units to my audience?

Responsible data use hinges on the author verifying the validity of the data before reporting on it. Without being an expert in data science, there are a list of questions that can help casual data users identify signs of suspicious or untrustworthy data. Using this list to evaluate data each time will minimize mistakes. Being aware of common data reporting mistakes by others will also ensure responsible data use.

Essential Questions to Ask a Dataset [2]

- Where do these numbers come from?
    - What institution published this data?
    - Does this institution have a record for reliable data collection?
    - Is the report available on their website?
- Who recorded them?
    - Did the institution gather the data itself or did it outsource to another company?
    - What training did the employees undergo?
- How?
    - Was data collected by going to the primary source or was it gathered from a report?
    - Were these the results of a survey in which some people were documented or a census in which almost everyone is documented?
- For what purpose was this data collected?
    - Was this data collected to report to a funder to show that targets have been met?
    - Was the data collected by an outside auditor?
- How do we know it is complete?
    - Can we interview the data collectors?
    - Is there an explanation of the limitations of the data?
- What are the demographics?

[2] Jonathan Stray, Source https://source.opennews.org/en-US/learning/statistically-sound-data-journalism/

    - Who was data collected about and who was left out?
    - Were rural and urban areas represented? Men and women? Abled and disabled?
- Is this the right way to quantify this issue?
    - What exactly is the data measuring and does your story match?
- Who is not included in these figures?
    - Was any group left out because of difficult of access? (such as disabled, people living in areas of violence)
- Who is going to look bad or lose money as a result of these numbers?
    - Was the study commissioned by an organization that is trying to prove that their projects are effective?
    - Was the study commissioned by an outspoken critic on the topic?
- Is the data consistent from day to day, or when collected by different people?
    - If collected over years, was the data collected by the same group following the same methodology?
- What arbitrary choices had to be made to generate the data?
    - How were decisions on things such as sample size made?
- Is the data consistent with other sources? Who has already analyzed it?
    - Are there other data sets on the same topic and to the results align?
- Does it have known flaws? Are there multiple versions?
    - Does the methodology explain potential errors in the data? Are there different copies in different places?

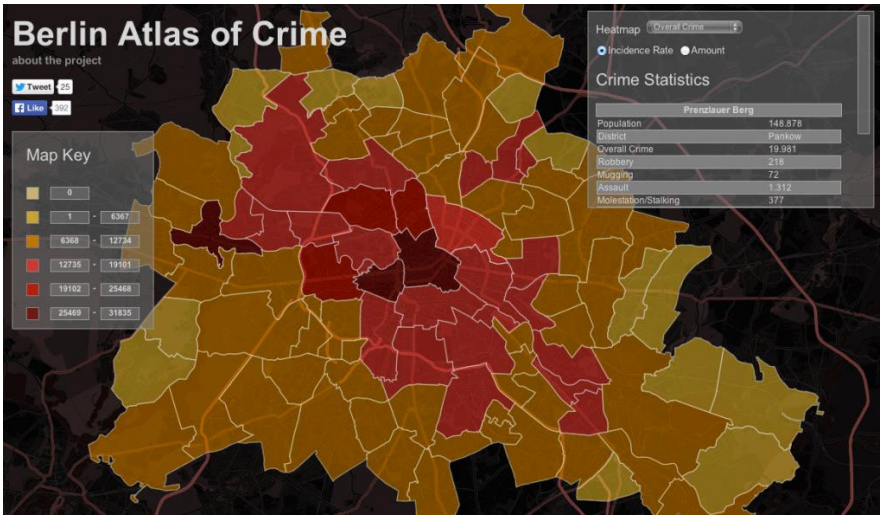
Exercise: Essential Questions to Ask a Dataset

1. **Read** The madrassa and the state of Pakistan
   - Who recorded the data?
   - How?
2. **Read** Afghan refugees will have to leave Pakistan come what may, says Safron minister
   - For what purpose was this data collected?
   - How do we know it is complete?
3. **Read** Aid Minister and the scandal school tsar: We give Pakistan £700m each year towards education - yet the Punjab schools minister is being investigated amid corruption allegations. Here he is... with Justine Greening
   - What are the demographics?
   - Is this the right way to quantify this issue?
   - Who is not included in these figures?
4. **Read:** Agencies Seek Better Data on Violence Against Women in Asia-Pacific
   - What data sources are cited?
   - Is the data consistent from day to day, or when collected by different people?
   - What arbitrary choices had to be made to generate the data?
   - Is the data consistent with other sources? Who has already analyzed it?
   - Does it have known flaws? Are there multiple versions?
5. Read Health indicators project bleak future for Pakistani children
   - Is this the right way to quantify this issue?
   - What exactly is the data measuring and does your story match?

Exercise: Evaluate Data in The Media

Find a story from mainstream Pakistani media and evaluate the source of the data for the story using *Essential Questions to Ask a Dataset*

Example: Understanding Crime Data



Sometimes, we don't look at enough data to come to meaningful insight about data. This often happens because we don't understand the topic and statistics enough or we are so eager to report that we don't look at the data in context or other data sets to compare it to. In effect, we don't have enough data to verify it properly. Crime data is often misreported or exaggerated because of these issues. These are some common mistakes made when reporting on crime.

Is a crime map just a population density map? In this example from Berlin, the highest crime rate is shown in the center of the city, which also has the highest population density: http://www.economicworldmap.net/berlincrime/

- Since a crime rate per population was not calculated, naturally crime rates appear higher where there are more people, but that doesn't mean that an individual's chances of being a victim of crime are higher, despite the dark red impression.
- Is it statistically significant when compared to the rest of the country or past years? (For example, a 5% increase in local crime when there was a 7% national increase, or a 6% increase in 2012)
- Is there a factor skewing results? (For example, a terrorist attack that caused a results spike or someone who under-reports crimes against vulnerable groups?
- Is it a factor of higher police presence? (For example, drug offences can soar if there are police around to catch them)
- Is this a crime category that is notoriously under-reported, like sexual assault? Or over reported? Like speeding in places where police have a quota of speeding tickets per month?

Exercise: Evaluating health care spending data

You have been asked to review basic health care spending data and identify the most important information for citizens to make more informed decisions about how the government should address health care costs and lobby politicians to spend wisely on health to address pressing needs.

- **Step 1:** Open Health Care Spending WB
- **Step 2:** Ensure you understand the context of the data
    - What is the source of the data?
    - What do each of the indicators mean?
        - Which indicators are complementary, or should add up to 100%
    - What are the units? Are they the best units for my analysis?
- **Step 3:** Identify what you think is most important for the public to know
- **Step 4:** Think about:
    - What story angles do you see in the data?
    - Which indicators go together to explain a problem?
    - What experts would you interview for your story?
- **Step 5:** Write a hypothesis and then a 200-300 word summary of your findings

Lesson 3: Summarizing and Simplifying Data Insights

Let's explore basic data related to health data in Pakistan. Child health is a complex issue that often varies widely by geography, income, and other factors. Comparing basic data from a variety of sources can help develop a more complete picture of the major challenges facing child health in the country.

Simplifying Percentages

When people read a series of percentages, they have a hard time connecting with the subject of the data. So always try to simplify a percentage to a fraction or a population rate so that the audience can imagine how many people are affected by an issue. A common denominator, or the biggest number that fits evenly into both the percentage and the total (100%) can help simplify the numbers.

Understanding percentages

To understand how to convert percentages, take a look at these:

- 33% = 33/100 = 3/10 (divide top and bottom by 3)= ⅓
- 75% = 75/100 (divide top and bottom by 25) = ¾

Example

Here are a few examples to further illustrate this point, a fraction or a population rate can help simplify the following statements about maternal health, which use a percentage.

At a national level almost 40% of these children are underweight. Over half the children are affected by stunting and about 9% by wasting. A positive relationship exists between the age of the child and the prevalence rates of stunting and underweight. There are significant provincial variations in malnutrition rates in Pakistan, whereas no differences in malnutrition rates are apparent between sexes. The prevalence of stunting appears to be associated with the overall level of development of the provinces, being lowest in Punjab and highest in Balochistan, the least developed province. -- http://www.fao.org/ag/agn/nutrition/pak_en.stm

Percentage	**Fraction**	**Population rate**
40% of children under five are under weight	Two fifths of children under five are under weight.	Two in five children are under weight for their age.

50% of children are stunted.	Half of children are stunted.	One in two children are too short for their age.
9% of children under five also wasted.	One tenth of children are wasted.	One out of 10 children are far too thin for their height.

Comparison across all three groups: always make comparisons out of the same TOTAL.

Two in five children are under weight for their age = Four in 10

One in two children are too short for their age = Five in 10

One out of 10 children are far too thin for their height = One in 10

What is the difference between underweight, stunted and wasted? Which are relevant or newsworthy?

Try these:

- Worldwide, women earn 30% less than men for the same work.
- Worldwide, 50% of homicides of women were committed by someone close to the woman.
- Worldwide, 80% of women report experiencing some kind of verbal sexual assault in her lifetime.

Comparing Numbers

Now let's try comparing two statements that use percentages, and re-write them in a clear and comprehensible manner for readers.

For example, let's simplify these two statements:

- 24.2% of births are registered in rural areas
- 32% of births are registered in urban areas

To do this convert the percentages in the two statements to fractions:

- 24.2% = 24/100 =6/25 = 1/5
- 32% = 32/100 = **8/25**

So now we can say:

- One out of five births are registered in rural areas
- 8 out of 25 births are registered in urban areas.

But this can we want to use the same denominator in both statements:

- **6 out of 25 births are registered in rural areas**
- **8 out of 25 births are registered in urban areas**

Rounding Off Numbers

Large, complex numbers can cause your audience to stop paying attention to your story. Use rounded, easy to understand numbers can ensure people understand the magnitude of the number without getting lost in all the digits:

Examples	**Rounded Off**
Pakistan's population is 193,569,848 million	Nearly 200 million people live in Pakistan.
Pakistan's fertility rate is 3.26	Women in Pakistan have on average just over three children.

Calculating Rates

Rates are also useful to simplify statements that quote percentages. For instance, to calculate what part of a population is affected by a condition, divide the total number of people by the number of people affected – the resulting number is your rate.

Example

Here is a statement:

"11% of child deaths are caused by injury"

Let's convert the percentage to a rate:

- You understand that 114% = 11/100.
- Now let's divide the total number here (i.e. 100) by the number of impacted people (i.e. 11), which gives us 100/11 = 9.09
- Based on this calculation we can say that 1 in 9 child deaths are caused by injury

"1 in every 9 children who die, die of injury"

Calculating Rates per population

Rates are used in relation to populations instead of specific numbers. They are necessary because it allows you to make comparisons across regions, demographic groups, etc. even if one group is very big and another is very small.

For example, incidents of violence are always calculated as a rate because it's not the same if two murders occur a year in a town of 1,000 as in a town of 50,0000.

The formula to calculate a rate is:

Rate= Number of incidents/(Population/100,000)

Sometimes rates are calculated by 1,000, 10,000 or 100,000 depending on the prevalence of what you are measuring. For example, if you are calculating a common occurrence, like cancer rates, you may calculate per 10,000 people. The formula would change accordingly:

Rate= Number of incidents/(Population/10,000)

Basically, the second half of the formula seeks to calculate how many groups of your standard measure, be it 10,000 or 100,000, exist in your population.

If we wanted to figure out which province has the highest reporting rates of domestic violence over a certain period of time, we can create a data table with towns, population, number of cases and rate and apply our formula.

Town	**Cases**	**Population**	**Formula**	**Rate per 100,000**
Town 1	500	150,000	500/(150,000/100,000)	333.33
Town 2	200	140,000	200/(140,000/100,000)	142.85
Town 3	100	130,000	100/(130,000/100,000)	76.92
Town 4	80	120,000	80/(120,000/100,000)	66.66
Town 5	50	110,000	50/(110,000/100,000)	45.45
Town 6	20	100,000	20)/(100,000/100,000)	20

Now we can see that, in fact, the bigger the town, the higher the rate of reporting of domestic violence. This of course does not necessarily mean that women in bigger towns experience higher rates of domestic violence. It could mean that the system to report cases to the police is easier or more visible in larger towns or it could be for some other reason. We will explore possible relationships in the next lesson.

Simplify the following statements that contain percentages by using fractions or population rates:

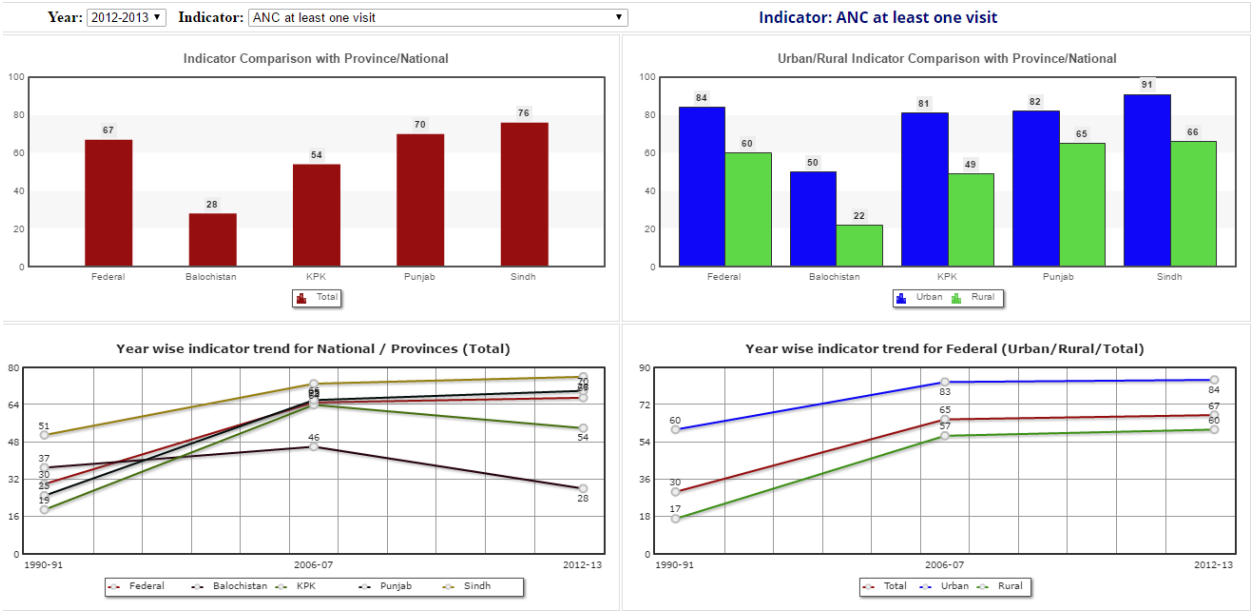
- In Pakistan, 29% of under five deaths are caused by pneumonia.

- In Pakistan, 6% of neonatal deaths are caused by congenital anomalies.
- In Pakistan, health makes up 4.7% of the national budget

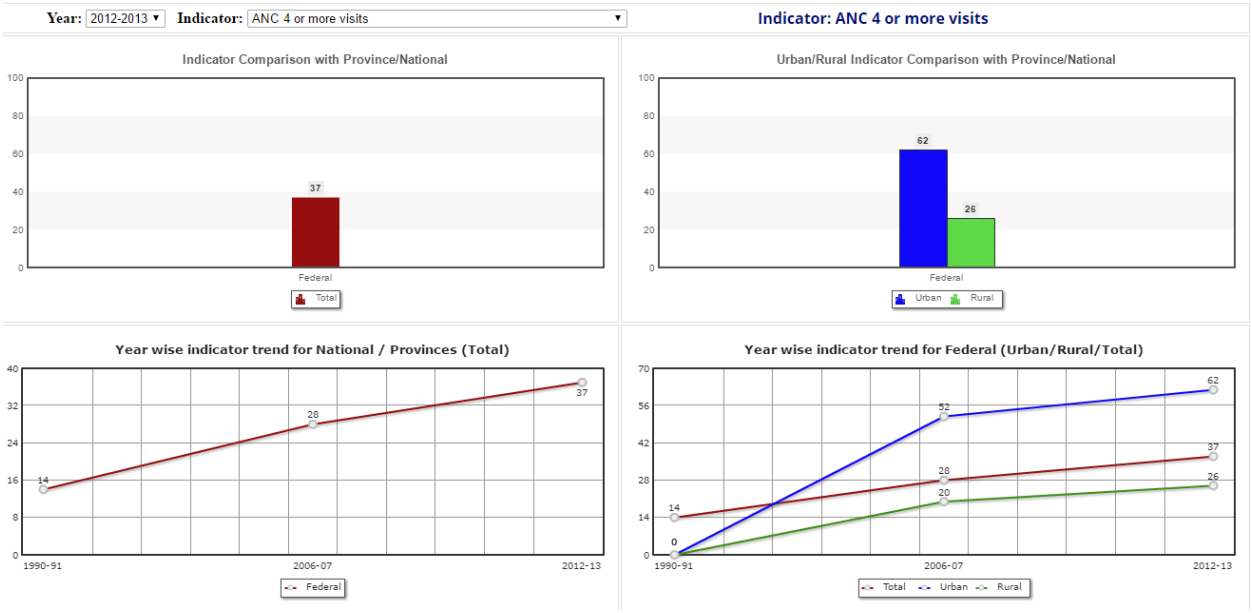
Use simplification to transform data into accessible information:

1. “Six per cent of the world’s chronically malnourished children live in Pakistan. Almost 10 million children suffer from chronic malnutrition (44%), 3.3 million suffer from acute malnutrition (15%), and 1.3 million (6%) are severely malnourished requiring therapeutic care. Up to 60 per cent mothers and children suffer from micro-nutrient deficiencies,” Dr Nizamani said. --Saving lives: MPs pledge to address malnutrition

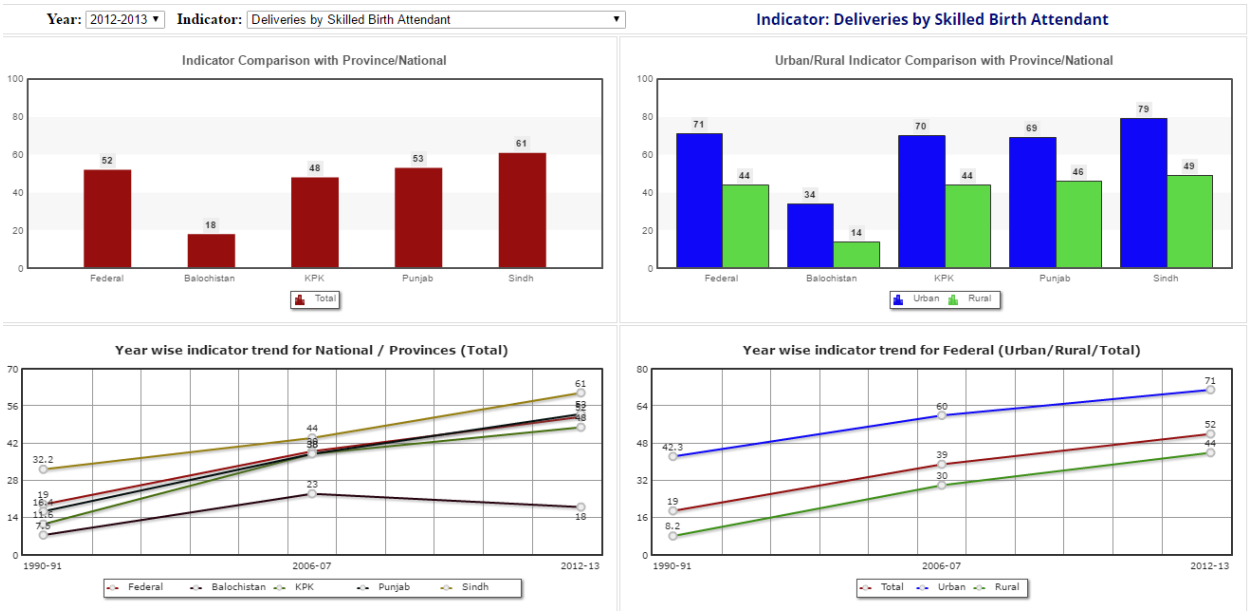
2.



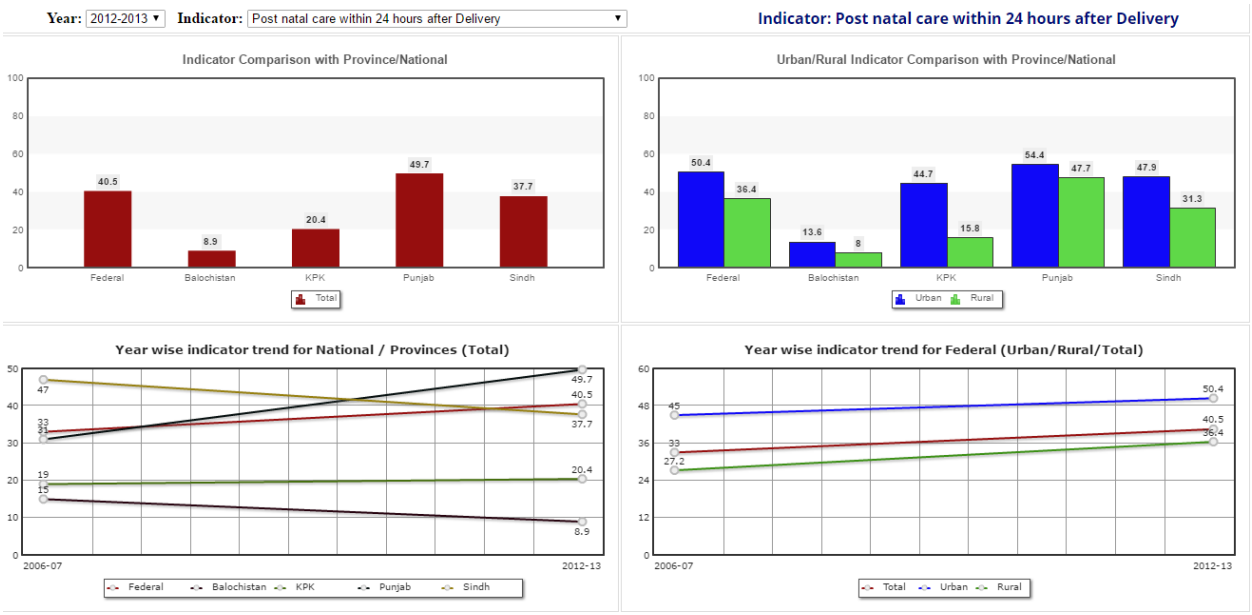
3.


Year: 2012-2013
Indicator: ANC 4 or more visits
Indicator: ANC 4 or more visits
Indicator Comparison with Province/National
Federal
Total
Urban/Rural Indicator Comparison with Province/National
Federal
Urban
Rural
Year wise indicator trend for National / Provinces (Total)
1990-91
2006-07
2012-13
Federal
Year wise indicator trend for Federal (Urban/Rural/Total)
1990-91
2006-07
2012-13
Total
Urban
Rural

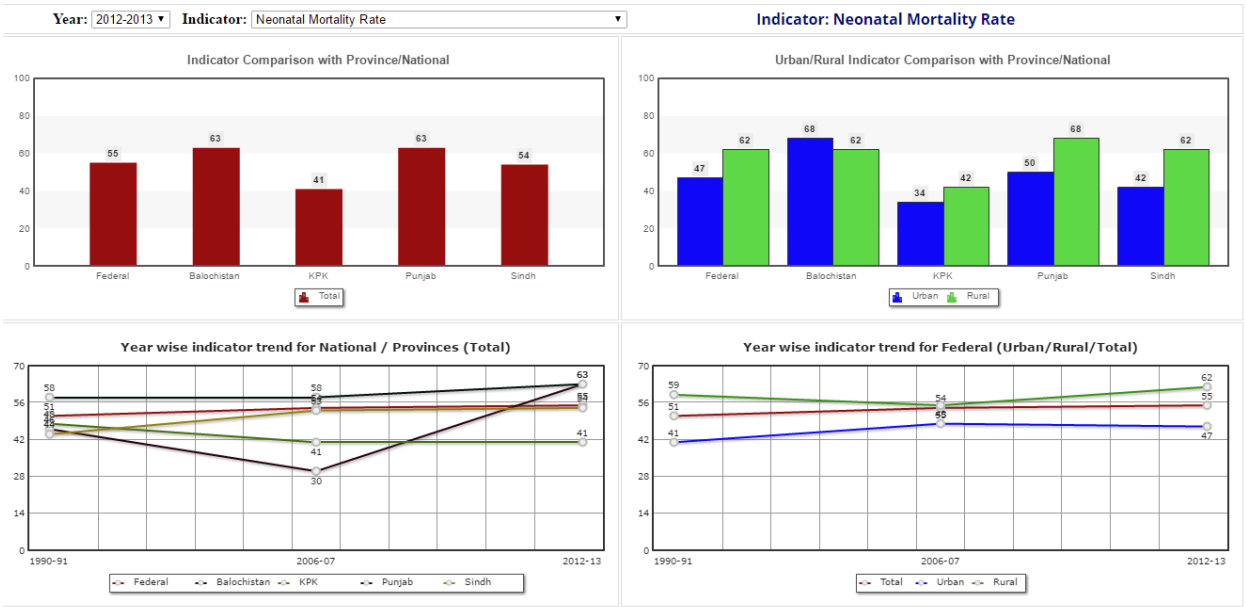
4.


Year: 2012-2013
Indicator: Deliveries by Skilled Birth Attendant
Indicator: Deliveries by Skilled Birth Attendant
Indicator Comparison with Province/National
Federal
Balochistan
KPK
Punjab
Sindh
Total
Urban/Rural Indicator Comparison with Province/National
Federal
Balochistan
KPK
Punjab
Sindh
Urban
Rural
Year wise indicator trend for National / Provinces (Total)
1990-91
2006-07
2012-13
Federal
Balochistan
KPK
Punjab
Sindh
Year wise indicator trend for Federal (Urban/Rural/Total)
1990-91
2006-07
2012-13
Total
Urban
Rural

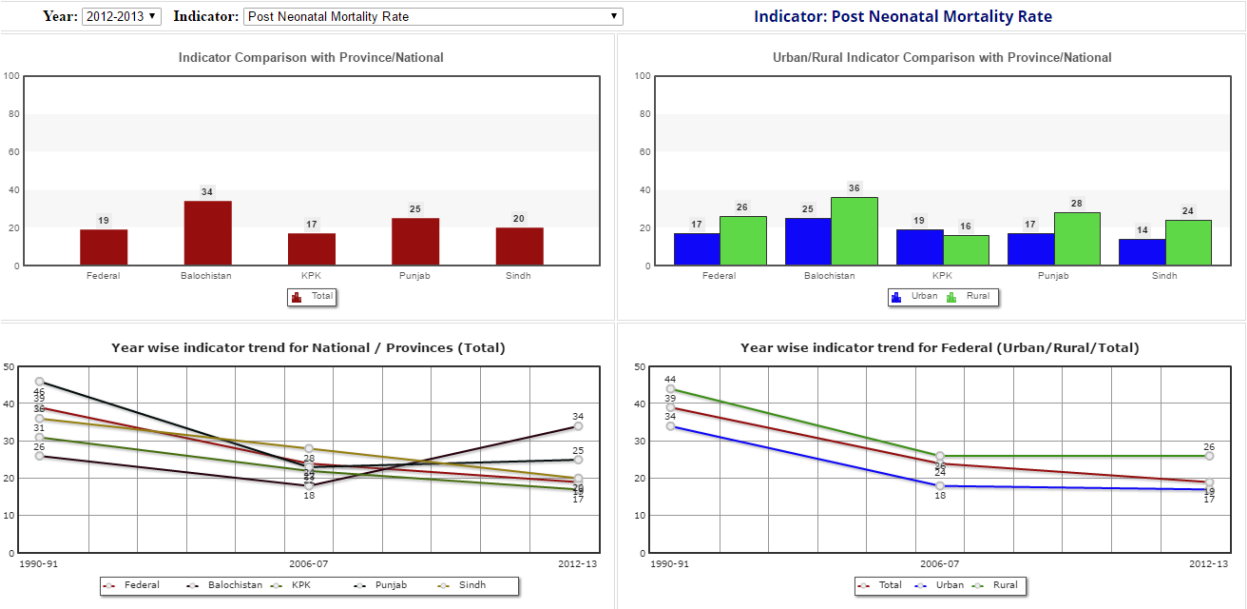
5.


Year: 2012-2013
Indicator: Post natal care within 24 hours after Delivery
Indicator: Post natal care within 24 hours after Delivery
Indicator Comparison with Province/National
Federal
Balochistan
KPK
Punjab
Sindh
Total
Urban/Rural Indicator Comparison with Province/National
Urban
Rural
Year wise indicator trend for National / Provinces (Total)
2006-07
2012-13
Year wise indicator trend for Federal (Urban/Rural/Total)

6.


Year: 2012-2013
Indicator: Neonatal Mortality Rate
Indicator: Neonatal Mortality Rate
Indicator Comparison with Province/National
Federal
Balochistan
KPK
Punjab
Sindh
Total
Urban/Rural Indicator Comparison with Province/National
Urban
Rural
Year wise indicator trend for National / Provinces (Total)
1990-91
2006-07
2012-13
Year wise indicator trend for Federal (Urban/Rural/Total)

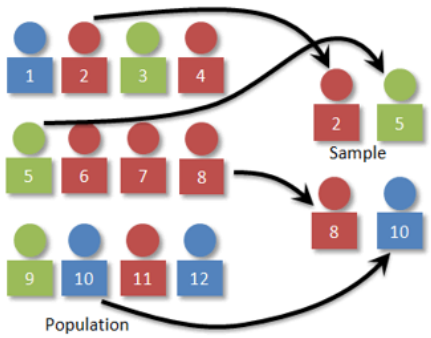
7.



8. Choose another survey from http://nhsrc.gov.pk/ (Dashboards, Click Here to View Pakistan Health

Lesson 5: Essential Statistics

Sampling



Sampling Considerations

- What's the universe?
- How will you draw the sample?
- How far will you want to break it down by demographics?
- What sort of accuracy do you need?

As data becomes a more plentiful source of information, it is increasingly important to be able to evaluate how the data was collected and whether it meets research standards for reliability. Understanding sampling and margin of error help determine whether polls or surveys are representative and what conclusions can be drawn.

What is sampling?

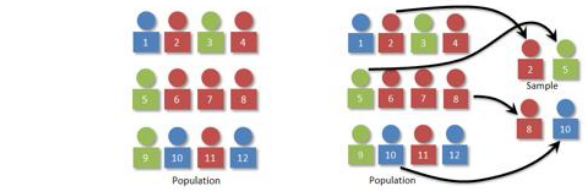
From Evidence to Stories: Thinking Like a Data Journalist

Since it is impossible or very time consuming and very expensive to gather data from every single person in a population, researchers usually employ 'sampling'. The goal of sampling is to choose a sample of people that accurately represents the entire population.

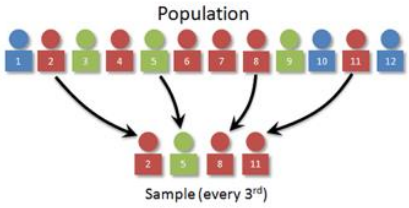
Sampling Considerations

- What's the universe?
- How will you draw the sample?
- How will you get the items, docs or data?
- How far will you want to break it down by demographics?
- What sort of accuracy do you need?

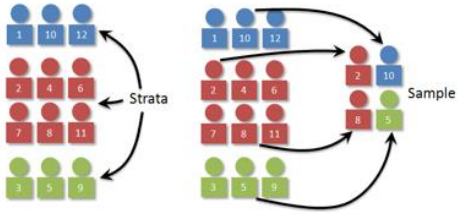
Common Sampling Methods



Random



Systematic



Stratified

There are several methods to select samples from a population. Here are some examples of methods used for sampling:

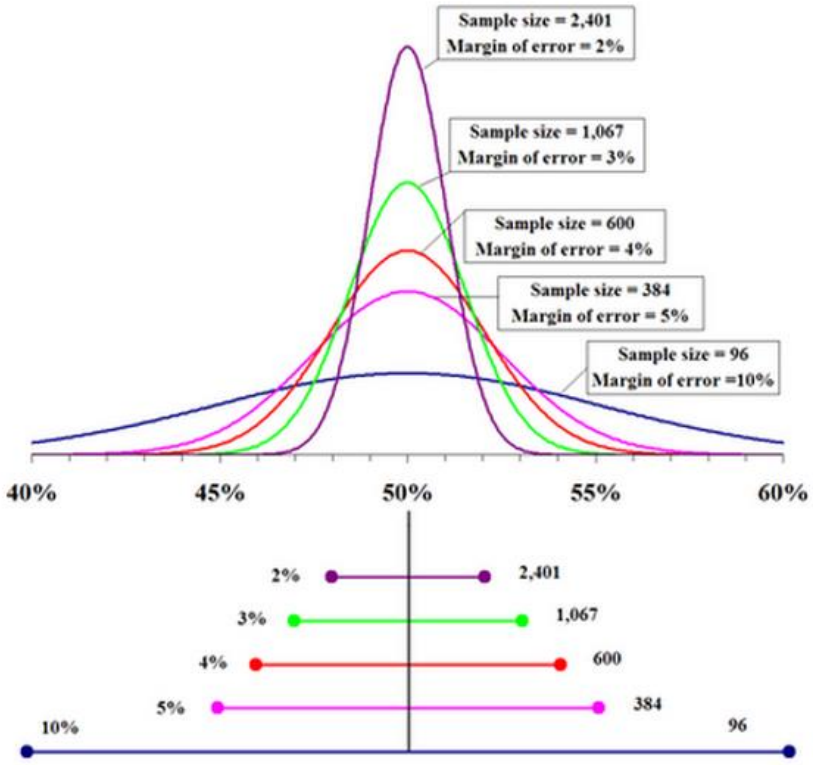
Random Sampling	Every item has an equal chance of being included.
Systematic sampling	Every Nth record is selected.
Stratified sampling	Pulling your sample based on another underlying number – such as population. Rather than pouring all the records for four counties in a pot and pulling randomly – you pull a random sample from each county.
Oversampling	Pulling more of a particular group in order to do further research with that group. For example, you notice an interesting trend among disease rates from a specific gender and age group, so you pull extra samples from that demographic to study the trend further.

Bad Sampling: Unscientific Polls

- Web polls
- Radio or TV call-in polls
- Man / woman on the street polls
- Twitter polls

- Web polls: You only get people who have access to the internet and read your website
- Radio or TV call-in polls: You only reach members of your audience, who already have their own political demographic, with the time and inclination to participate
- Man / woman on the street polls: This is a tiny sample and can't represent a population
- Twitter polls: Again, respondents are only Twitter users who follow you

Understanding Margin of Error



[3]

Polls generally involve taking a sample from a certain population. Because it is impractical to poll everyone in the population, researchers take smaller samples that are intended to be representative, that is, a random sample of the population. The margin of error is a measure of confidence in the polling results.

The more people that are interviewed, the more likely it is that the sample is representative of a population. If a poll has a margin of error of 2.5 percent, that means that if you ran that poll 100 times — asking a different sample of people each time — the overall percentage of people who responded the same way would remain within 2.5 percent of your original result in at least 95 of those 100 polls.

[3] https://en.wikipedia.org/wiki/Margin_of_error

Example: Small sample size and the vaccine and autism debate

Read: Sticking with the truth

In 1998, Andrew Wakefield and 12 of his colleagues published a case series in the *Lancet*, which suggested that the measles, mumps, and rubella (MMR) vaccine may predispose to behavioral regression and pervasive developmental disorder in children. Despite the small sample size (*n*=12), the uncontrolled design, and the speculative nature of the conclusions, the paper received wide publicity, and MMR vaccination rates began to drop because parents were concerned about the risk of autism after vaccination.

As the Columbia journalism review points out, even though the subsequent science was clear, definitively disproving any link between vaccines and autism, in the interest of "balance" journalists quoted people on both sides of the debate. ‘‘While it's somewhat reassuring that almost half the US stories (41 percent) tried, to varying degrees, to rebut the vaccine-autism connection, the study raises the problem of "objectivity" in stories for which a preponderance of evidence is on one side of a "debate." In such cases, "balanced" coverage can be irresponsible, because it suggests a controversy where none really exists."

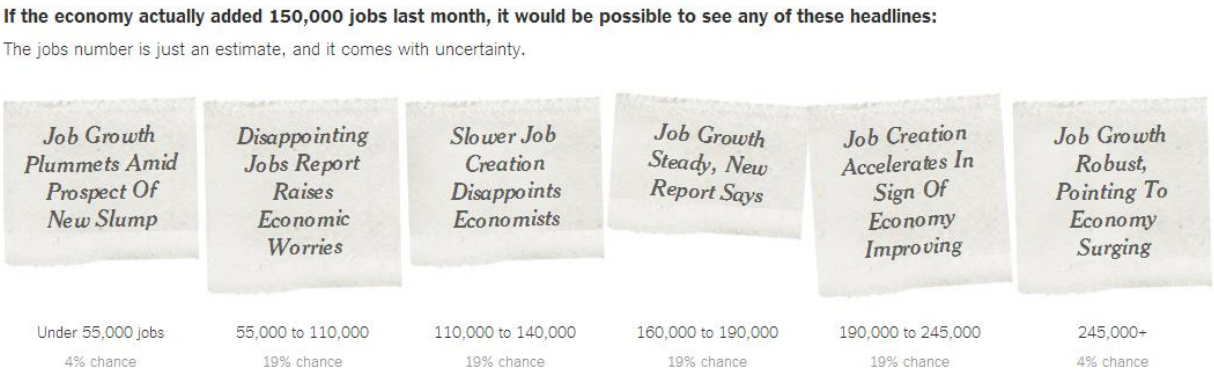
With a little knowledge of statistics, we know that a sample size of only 12 people is too small to suggest any findings at all. It was chance, and other mistakes in the study, that suggested that vaccines and autism had any causal link.

Example: Small samples hidden in big samples

Read: The hidden dangers of ethnic minority data in big surveys

In this case, the British Labor Force Survey sample size of the about 100,000 is very large. But the article cited here, Migrants 'milking' benefits system: Foreigners more likely to claim handouts focuses in on only on housing benefit claimants born in Pakistan or Bangladesh aged 40-44. The survey only captured 27 individuals who belong to that demographic. So while the sample of the entire study is large, underrepresented minorities continue to be underrepresented. This presents a particular challenge for journalists trying to produce public interest reporting on groups that are underrepresented in the media. Before using large data studies that may include some minorities, look for other surveys targeted particularly at those groups.

Example: How not to be misled by the jobs report



This article seeks to explain how even with a very large sample size, data based on a sample is inherently imprecise and instead each number represents a possible range of numbers that describe employment growth: http://www.nytimes.com/2014/05/02/upshot/how-not-to-be-misled-by-the-jobs-report.html?_r=1

Read the article, and answer the following questions:

- What is the sample size for any given month?
- How many jobs are there actually in the economy?
- What do the moving bars represent?
- Why are all the headlines possibly correct?

Exercise: Margin of Error

“National poll: candidate A lead over candidate B grows by 2 points in August to 56% of votes”

The margin of error is 2.5%

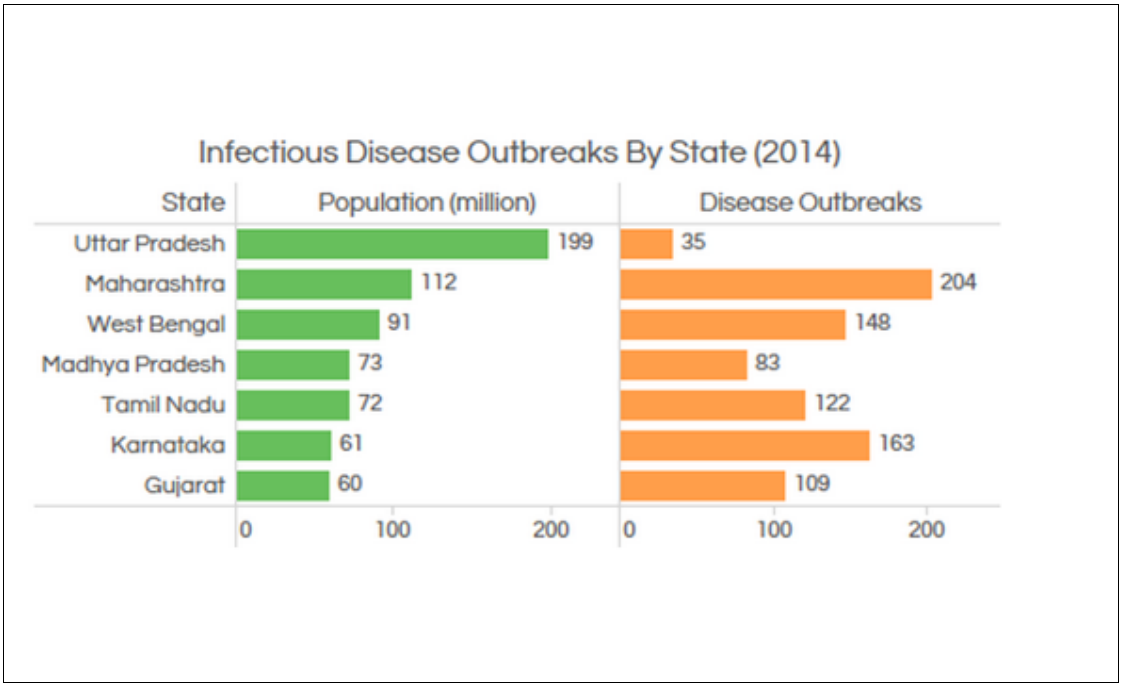
Headline: “National poll: candidate A lead over candidate B grows by 2 points in August to 56% of votes”.

The margin of error is 2.5%.

Based on this headline and the associated margin of error, answer the following questions:

- What is the possible range of the population intending to vote for Candidate A in August?
- What is the possible range of the population intending to vote for Candidate A in July?
- Is there any overlap?
- What does the overlap mean?
- What is a more accurate headline?

Assessing Data Reliability (1 of 3)



In this example, we will try to assess how reliable is the data that's represented in data visualizations.

In 2014, the government of India released data on outbreaks of epidemic diseases for various states in India. The data relate to diseases, such as diarrhea, cholera, and malaria.

Here is a visualization based on the data[4]. Let's first interpret this visualization.

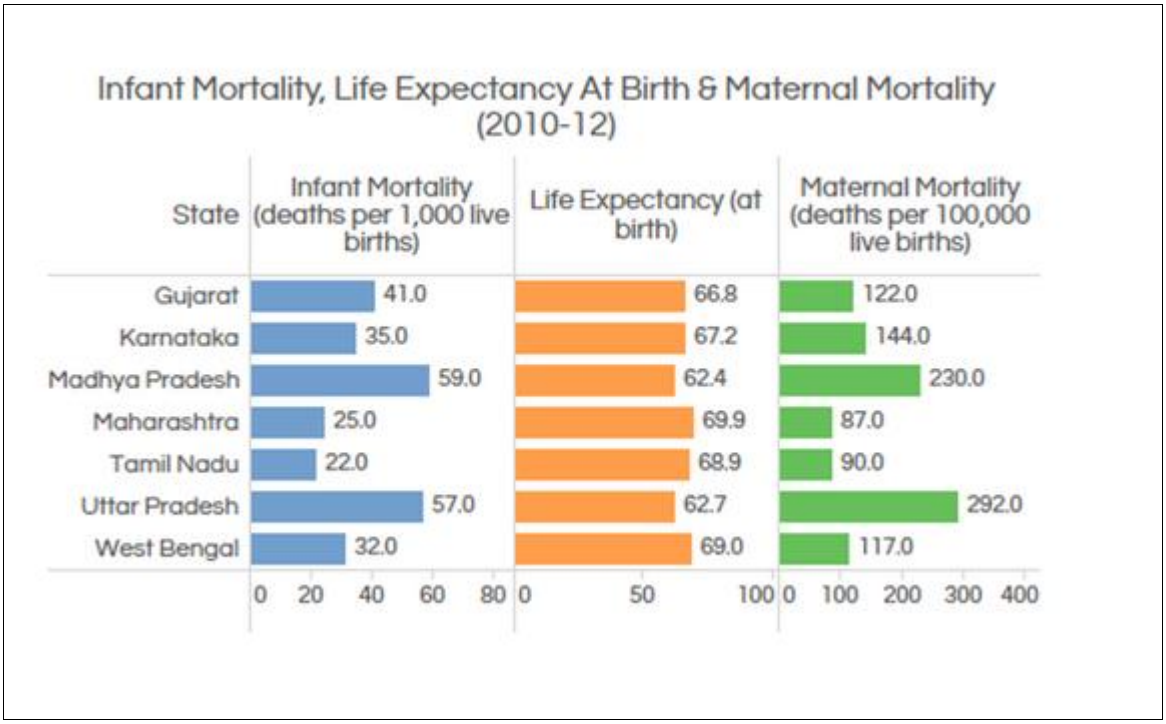
Questions:

- Which state had the largest number of disease outbreaks?
- Which state had the highest rate[5] of disease outbreaks?
- Which state has both the highest population and lowest number of disease outbreaks?
- What is the difference between disease outbreaks and number of cases of the disease and number of fatalities from disease?

[4] http://www.indiaspend.com/cover-story/lies-and-statistics-how-indias-most-populous-state-fudges-crime-data-11091

[5] The rate is the number of disease outbreaks divided by the population

Assessing Data Reliability (2 of 3)



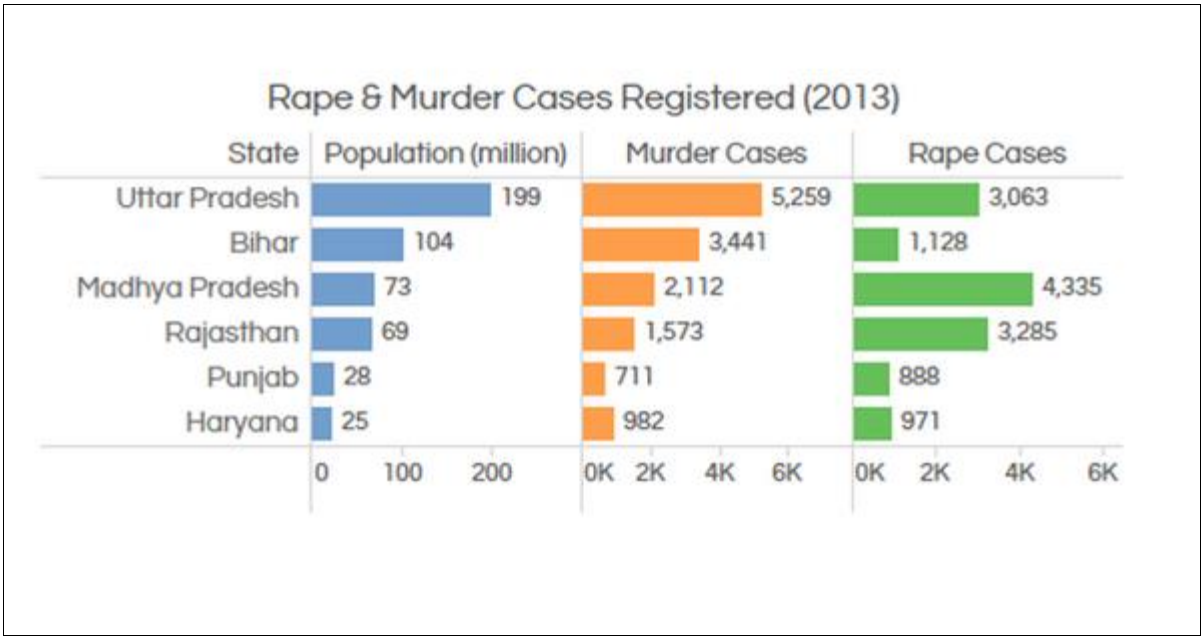
Looking at the previous visualization we can make this claim – "The state of Uttar Pradesh in India has the highest population however it also has the lowest number of disease outbreaks. "

To evaluate this claim, let's also consider this visualization about other health indicators from India.

Questions:

- Which state has the highest rate of infant mortality?
- Which state has the longest life expectancy? And shortest?
- Which state has the highest maternal mortality?
- How does this data change our confidence in the disease outbreak data for Uttar Pradesh?

Assessing Data Reliability (3 of 3)

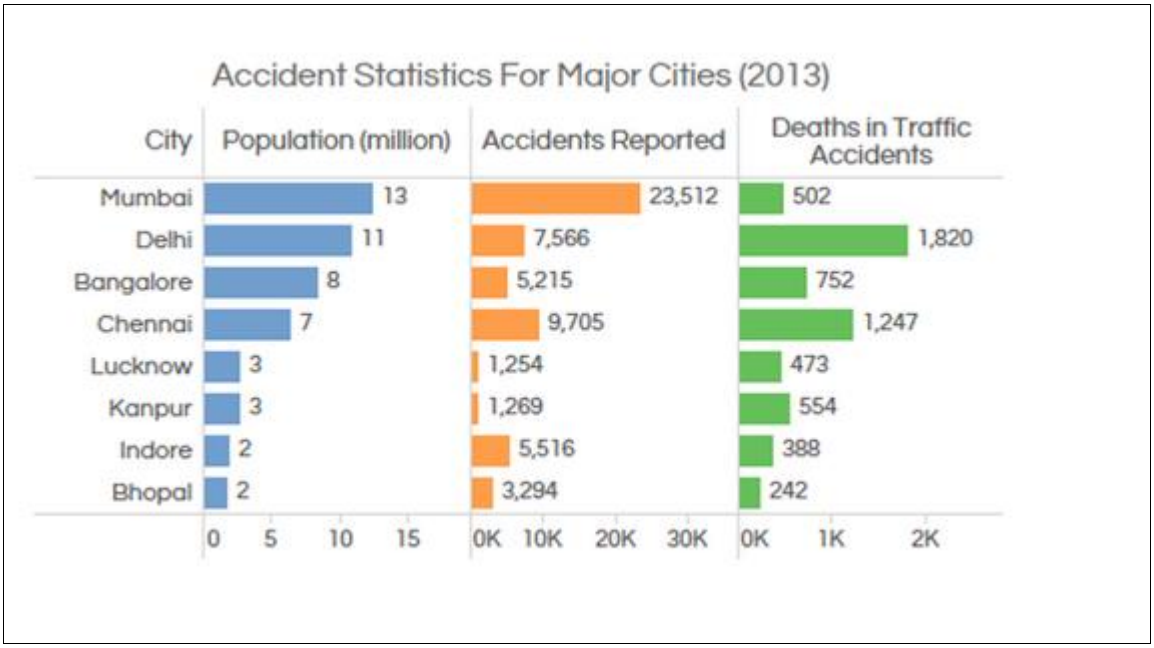


Let's continue from the previous example, and compare crime data from the some of the states in North India. Notice that the state we focused on in the last two visualizations – Uttar Pradesh – is present in this data.

Questions:

- Which state records more murders than rapes?
- Which is more difficult to cover up or under-report: a murder or a rape?
- Look back at our findings about Uttar Pradesh from previous questions. Does it seem credible that women are safer in Uttar Pradesh as compared to other states in North India?
- Based on the data, which state's data should we be skeptical of?

Practice: Assessing Data Reliability



Based on what we learned in the previous example, let's look at another scenario. Here accidents data from major Indian cities is compared. Try answering these questions to assess the reliability of the underlying data.

Questions:

- Which city records the most accidents?
- Mumbai and Delhi have similar populations, 13 million and 11 million, respectively. Delhi has far more vehicles, but Mumbai recorded thrice as many accidents. Does this mean Mumbai drivers are more reckless?
- In Mumbai, there was one death for every 50 accidents whereas in Delhi there was one death for every 4 accidents. Do you think accidents in Delhi are more deadly?
- Which number is easier to under report, accidents or deaths in accidents?

Lesson 5: Evaluating Data Interpretation

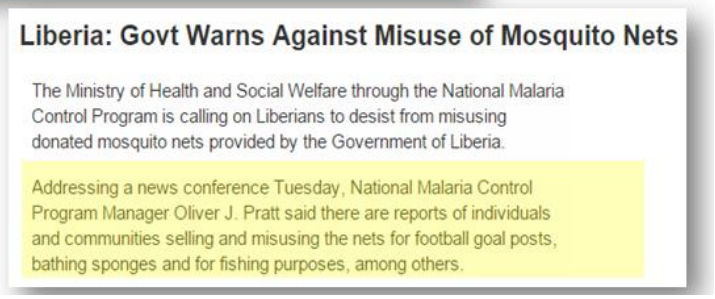
- Choosing proxy data to answer a question
- Comparing things that are not alike
- Extrapolating a pattern from coincidence
- Confusing correlation and causation
- Finding trends in too little data
- Producing aggregated data by aggregating individual cases

Data is a fantastic basis for beginning to produce information that can guide public policy. However, it is important to understand the limitations of a dataset, what conclusions can be reasonably drawn and to recognize claims have been made that can't be supported by the data at hand. All data journalism is based on interpretation, even in the most concrete headlines:

- "Crime rates fall"
- "Humans are causing climate change"
- "Countries with more guns have more deaths by firearms"

In this lesson, we will evaluate claims made by others about data for accuracy. Errors fall into several major categories that we watch out for when determining whether data analysis is correct.

Choosing Proxy Data



Liberia: Govt Warns Against Misuse of Mosquito Nets

The Ministry of Health and Social Welfare through the National Malaria Control Program is calling on Liberians to desist from misusing donated mosquito nets provided by the Government of Liberia.

Addressing a news conference Tuesday, National Malaria Control Program Manager Oliver J. Pratt said there are reports of individuals and communities selling and misusing the nets for football goal posts, bathing sponges and for fishing purposes, among others.

In many cases, 'what we want to measure' and 'what we are able to measure' are not the same - leading to the selection of a proxy indicator. A proxy is a stand-in indicator of what we want to measure. For example, often news articles use Per Capita Gross Domestic Product, or the total amount of money that a country makes, as a proxy for measuring quality of life in that country. That means instead of measuring quality of life by looking at all the factors influencing well-being, it substitutes average individual wealth to measure quality of life or wellbeing in a country.

Here's another example[6] – about the use of insect-treated mosquito nets to prevent malaria.

Questions:

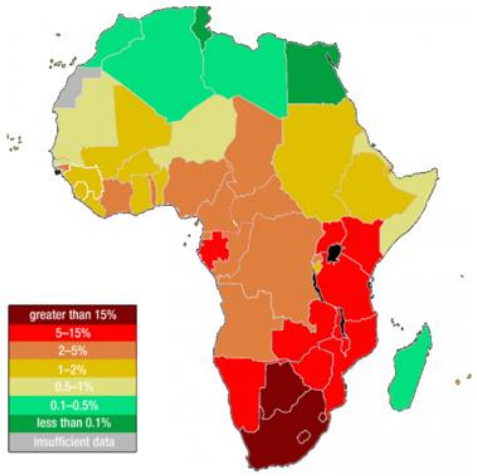
- Is the number of mosquito nets per person an accurate representation of how many people are sleeping under mosquito nets?

- To understand how much of the population has access to mosquito nets, what would have to be measured?

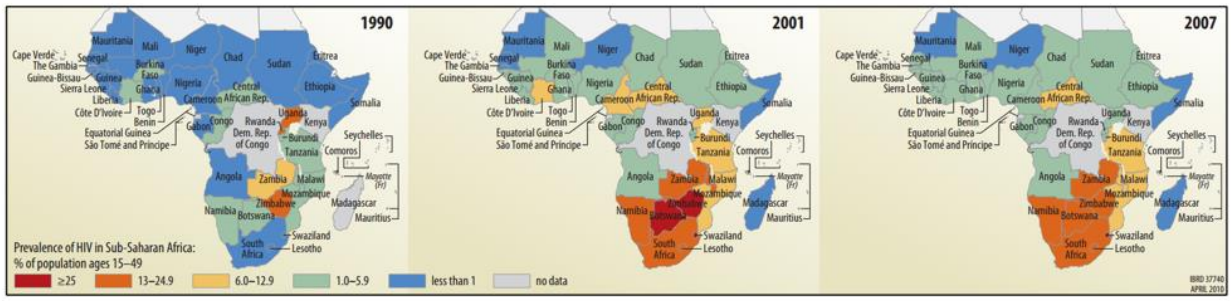
[6] http://www.internewskenya.org/dataportal/assets/img/data_visualisations/Preventingthebite.png http://allafrica.com/stories/201506031681.html

Comparing Things that are Not Alike (1 of 2)

Map 1



Map 2



Often, we don't get all the data needed to tell a story - either from year to year, place to place, or in the detail. We have to make decisions about how much information we can derive with a limited dataset. For example, let's imagine we get teacher salary data from all over a country. It may be tempting to hone in on where teachers are paid the least but there are other factors: cost of living may be lower, education requirements for teachers across the country may vary, incentive pay may mean teachers migrate to certain areas for higher pay or there may be a rotation system in place.

Let's consider an example - suppose you are studying these two maps to find out how is Sudan doing in the fight against HIV and AIDS compared to other African countries over the last 10 years?

- In Map 1[7], we see the HIV rates for 2015, but only for 2015.
- In Map 2[8], the data is older, but we can see HIV rates over three periods: 1990, 2001, and 2007.

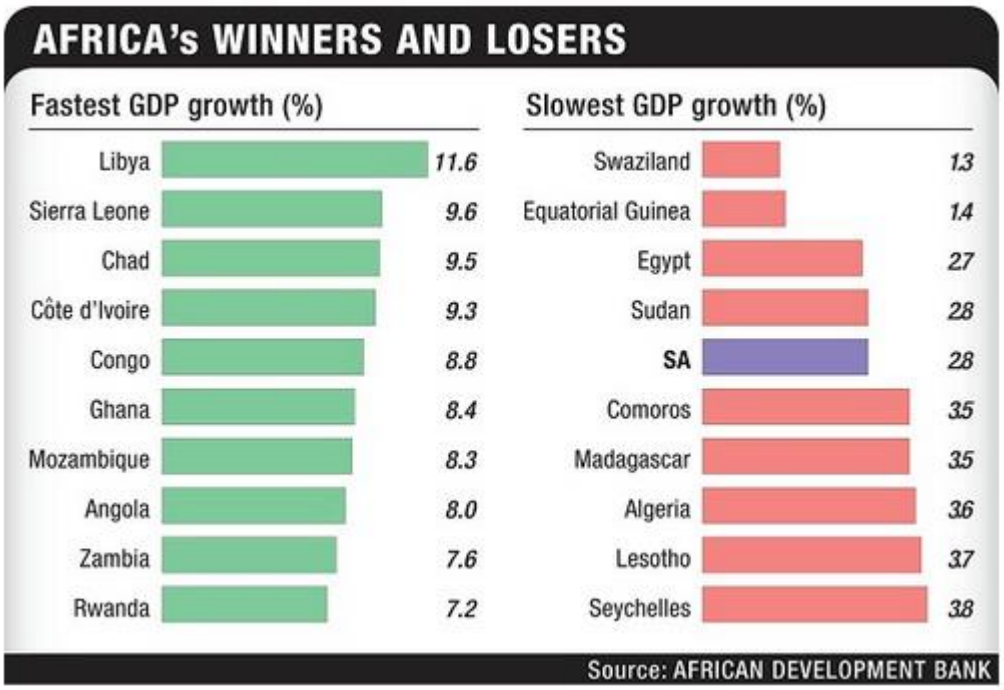
[7]http://blogs-images.forbes.com/judystone/files/2015/02/Map-of-HIV-Prevaance-in-Africa1-e1423067828157.png

[8]http://siteresources.worldbank.org/INTPROSPECTS/Images/334933-1271876733261/6992744-1328626949160/8422535-1328627766358/Africa_&_HIV.pdf

Questions:

- What is the advantage of the first map?
- Based on the first map, which countries should we compare Sudan to?
- Which countries had similar rates of HIV 15 years ago?

Comparing Things that are Not Alike (2 of 2)

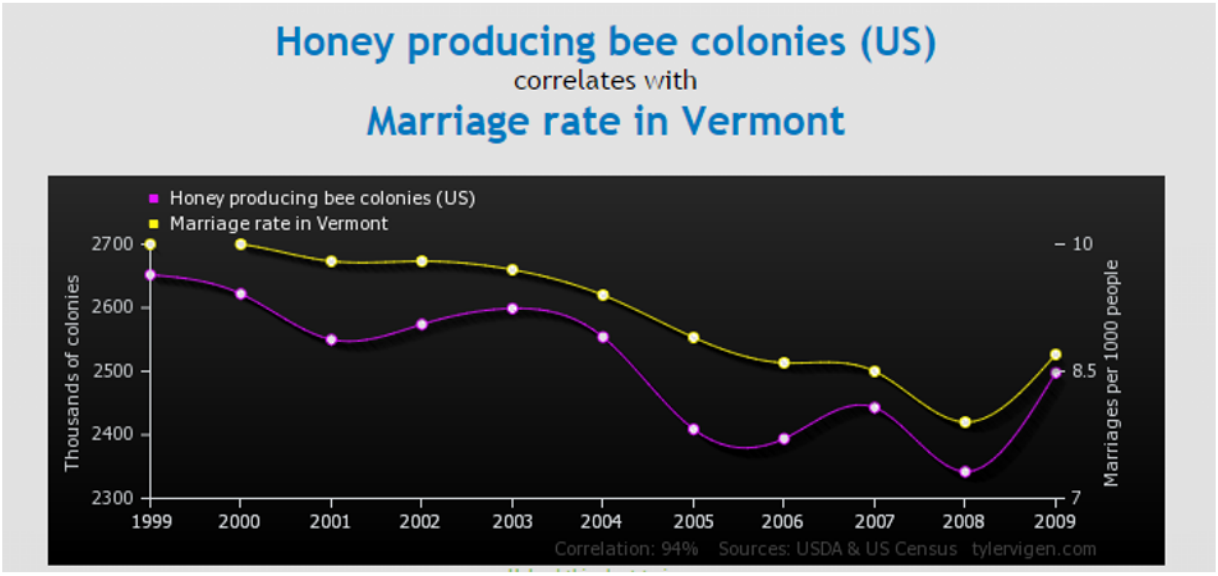


Now let's consider another example – about Sudan's economic growth: http://www.bdlive.co.za/economy/2013/08/08/south-africa-slumps-into-the-bottom-10-on-africa-growth-chart

Questions

- If you want to understand Sudan's economic growth over the last 10 years, which countries would you compare it to?
- Does it make sense to compare Sudan to neighboring countries?
- Does it make sense to compare it to countries that had a similar GDP per capita 10 years ago?

Extrapolating a Pattern from a Coincidence (1 of 2)



Sometimes, datasets match up quite closely, mirroring a non-existent trend. For example, there is a close relationship between rates of marriage and rates of honey production in the US state of Vermont[9]. Most likely, these two variables are unrelated.

Often, if you go in looking for a pattern without an open mind, you will find trends that may not be real. Always look for other data sets to confirm or disprove a trend that looks too convenient.

[9]http://www.tylervigen.com/page?page=1

Extrapolating a Pattern from a Coincidence (2 of 2)

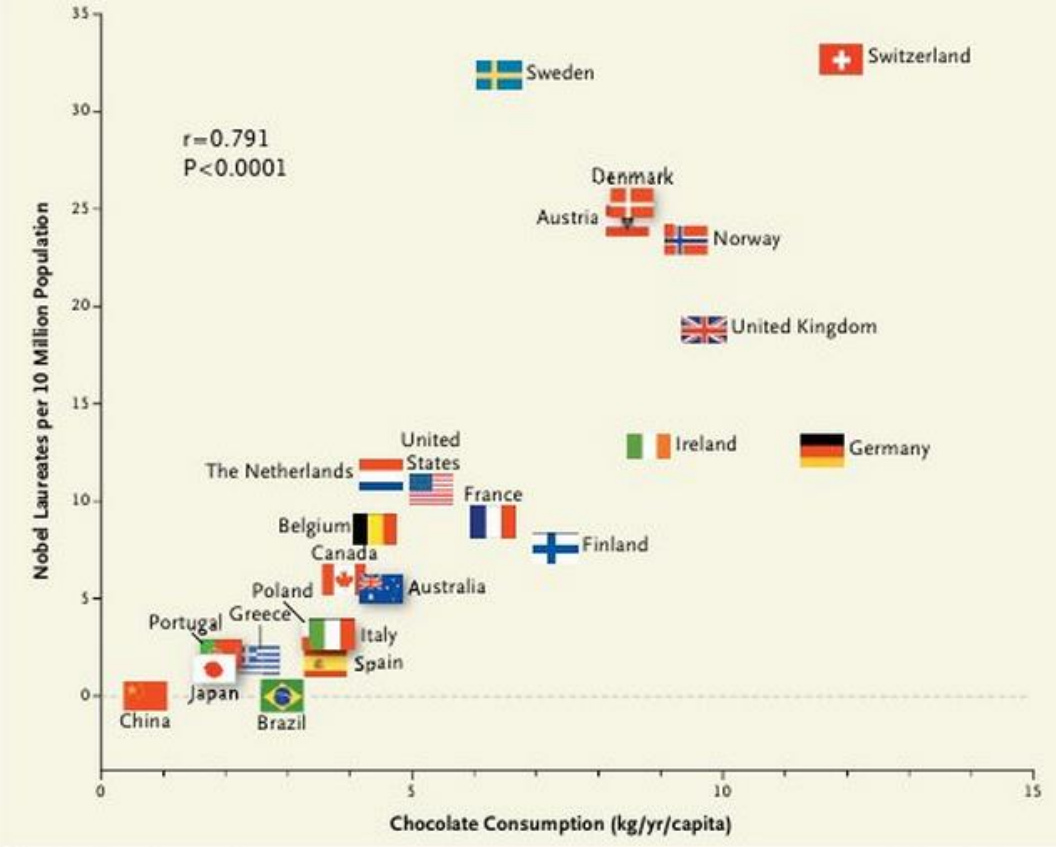


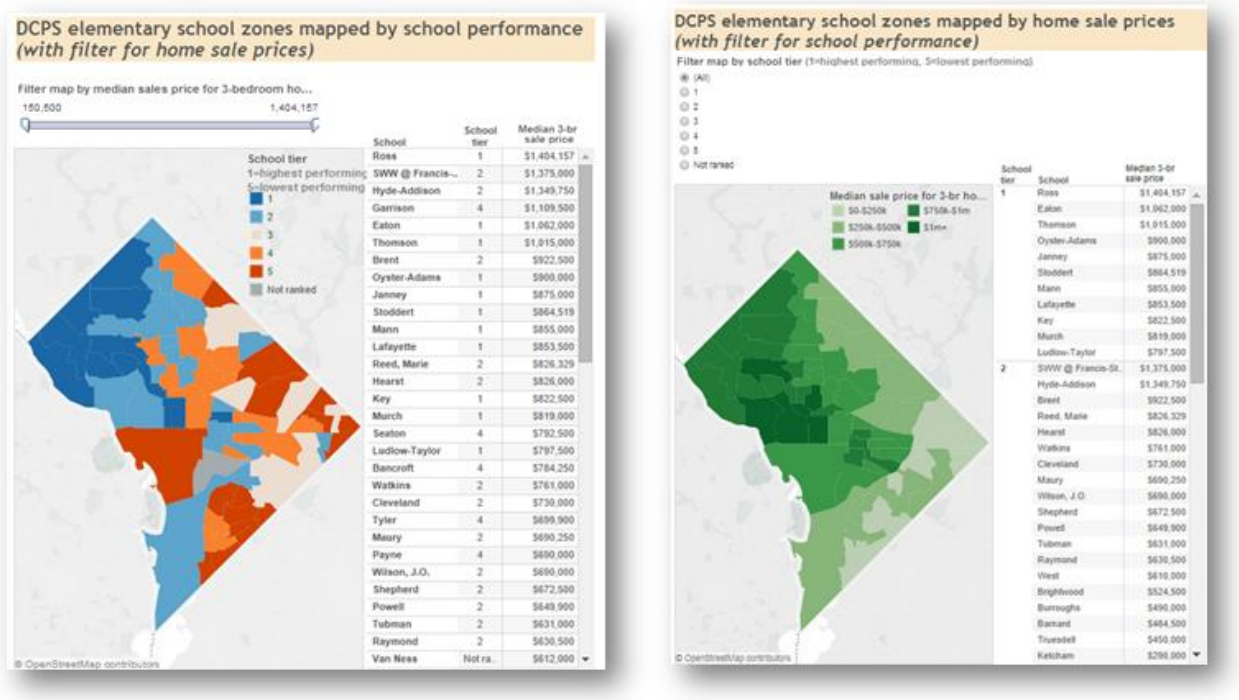
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Here is another example: The number of Nobel prizes won by a country (adjusting for population) correlates well with per capita chocolate consumption.

Questions:

1. What could the possible relationships between the two variables be?
2. Do these variables seem likely to be related?

Confusing Correlation and Causation (2 of 3)



Let's consider an example: https://www.washingtonpost.com/blogs/all-opinions-are-local/wp/2015/07/22/the-correlation-between-test-scores-and-home-prices/

It seems there is a correlation between housing prices and test scores. However is one variable causing the other?

- Housing prices increasing could cause wealthy, educated people to move into the neighborhood and test scores of their children go up.
- Or, schools with good test scores attract wealthier families willing to pay more to live in that neighborhood.
- It may be another hidden variable. Maybe green space attracts good teachers and cause housing prices to go up.
- Or it could be a factor we don't know about, or it could be a coincidence.

Confusing Correlation and Causation (3 of 3)

Chart A

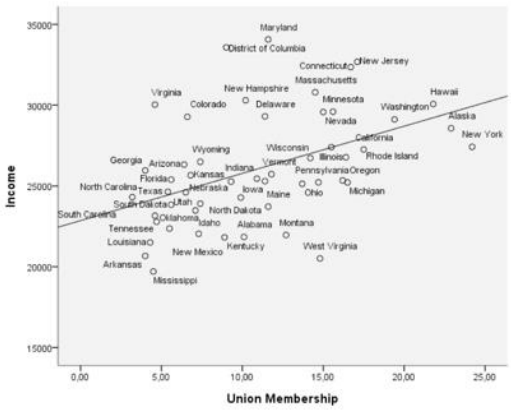


Chart B

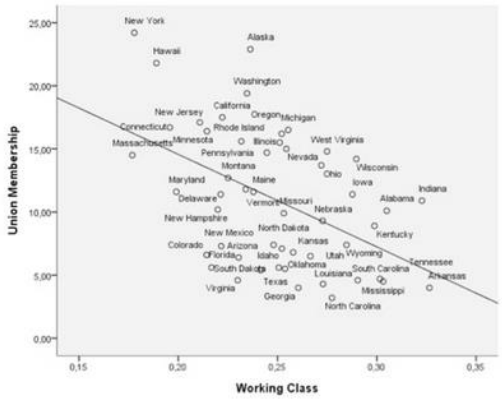
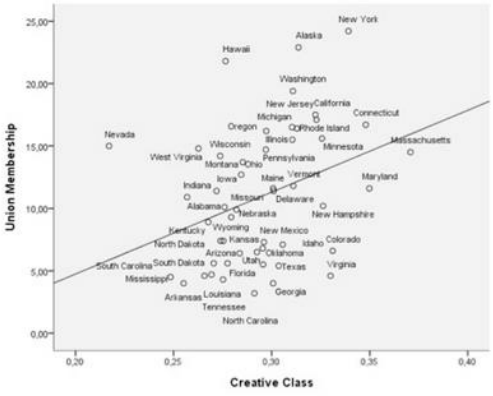


Chart C



Now try this. Observe these three charts.

These charts show how union memberships correlates with three difference variables - Income (Chart A[10]), working class (Chart B[11]), and creative class (Chart C[12]) - across states in the US.

Questions:

- What are the possible relationships between unionization and higher income?
- What are the possible relationships between the working class and unionization?
- Between unionization and working class? Creative class and working membership?
- How would you determine which are causing which?

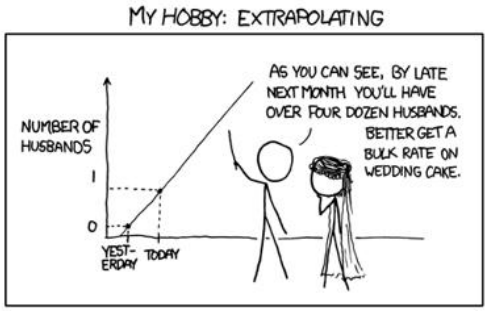
[10]http://www.creativeclass.com/creative_class/_wordpress/wp-content/uploads/2011/03/union4.png

[11]http://www.creativeclass.com/creative_class/_wordpress/wp-content/uploads/2011/03/union5.jpg

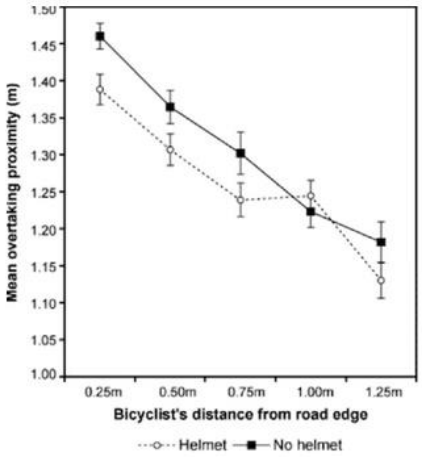
[12]http://www.creativeclass.com/creative_class/_wordpress/wp-content/uploads/2011/03/union6.jpg

Finding Trends in too Little Data and Making Generalizations

A lighthearted example



A real-life example



The larger the sample size, the more data collected, and the longer the duration of the study, the more likely it is to be able to draw conclusions from data findings.

In this humorous example[13], the author has only one data point: one day passes and the woman gained one husband. The author could extrapolate that on day 2 the woman will have 2 husbands; on day 3 she will have 3 husbands, etc. This is a silly example but often people use too little data to jump to broad conclusions.

Let's also consider a real-life example where a one man study "proved" that people drive closer to bicycle riders with helmets. In this example, a British researcher collated data on how close cars drive to him when he is and isn't wearing a helmet[14]. He rode 200 miles, and found that when he was wearing helmet, cars came about 3.35 inches closer while passing him.

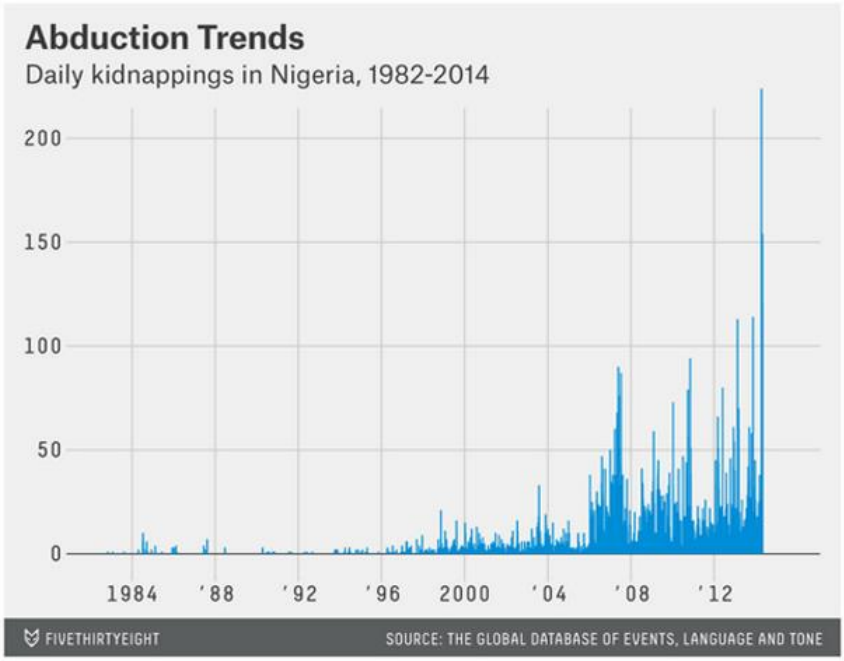
There are a couple of problems in this helmet study. First, this is not enough data. We would have to collect data from many people from many demographics driving at many times of day in many places to identify any trends about proximity to drivers with and without helmets. What else would need to be measured?

[13] http://www.explainxkcd.com/wiki/index.php/File:extrapolating.png
[14]http://www.vox.com/2014/5/16/5720762/stop-forcing-people-to-wear-bike-helmets

Let's look at another example about crime in the United States:
http://www.npr.org/2015/07/01/418555852/nationwide-crime-spike-has-law-enforcement-retooling-their-approach

- Where does the headline indicate crime is spiking?
- Which cities does the story mention have seen a rise in crime rates?
- Does the story compare crime over multiple years?
- What data would you use to determine whether there has been a nationwide crime spike in your country?

Trying to Produce Aggregate Data by Aggregating Individual Cases



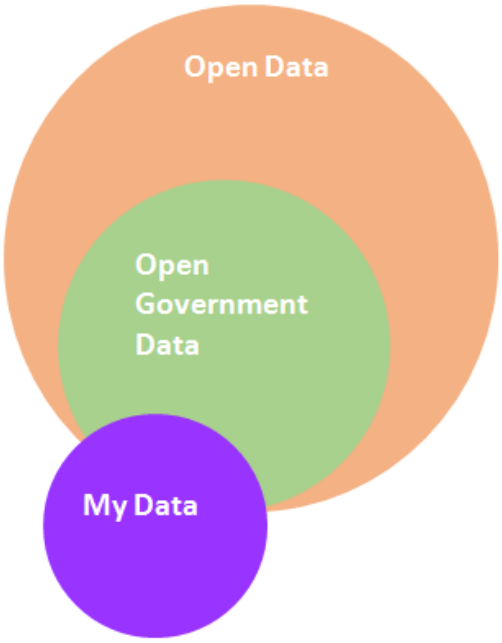
Sometimes, in the absence of official data, journalists and CSOs will try to aggregate data from unofficial sources as a proxy. For example, journalists may aggregate media reports of migrants being lost at sea while crossing the Mediterranean to try to estimate the number of migrants who have been lost. Sometimes, this technique fails because of errors such as double counting or a hole in reporting from a certain region or demographic.

Let's consider the example of reporting of the "Bring Back Our Girls" campaign, sparked by the kidnapping of Nigerian school girls by Boko Haram, media outlets used data from an NGO that sourced the numbers from global media reports on abductions.

Read the following article: https://source.opennews.org/en-US/articles/gdelt-decontextualized-data/, then answer the following questions:

- What was the author trying to accomplish in accessing aggregate data about the number of abductions in Nigeria?
- Why was the data used not an accurate representation of the number of abductions that had actually taken place?
- Where else could the author have gotten this data?
- Why is it important to show the number of abductions over a time period?

Lesson 4: Data Privacy



Despite all the energy behind open data, privacy and surveillance are increasingly prominent concerns when it comes to both governments and companies sharing data. The common practice to "anonymize" data, or to strip out the identifying characters, can sometimes, usually around commercial data, be undone by algorithms. In other circumstances, some data can have unintended effects.

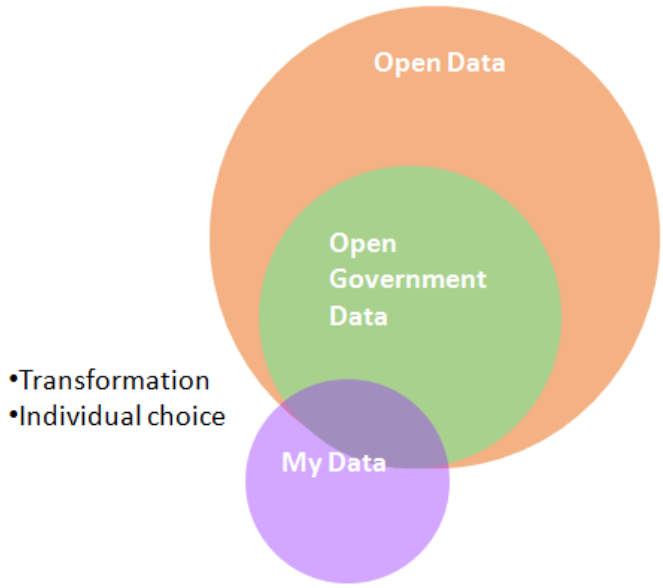
Open Data

"Open data and content can be freely used, modified, and shared by anyone for any purpose"

My Data

Who owns data about me, who controls it, who has access to it? Can I see data about me, can I get a copy of it in a form I could reuse or share, can I get value out of it? Would I even be allowed to publish openly some of the data about me, if I wanted to?

Intersection of Open Data and My Data



My Data becomes Open Data (via transformation)[15]: Important datasets that are (or could be) open come from "my data" via aggregation, anonymization, and so on. Much statistical information ultimately comes from surveys of individuals, but the end results are heavily aggregated (for example, census data).

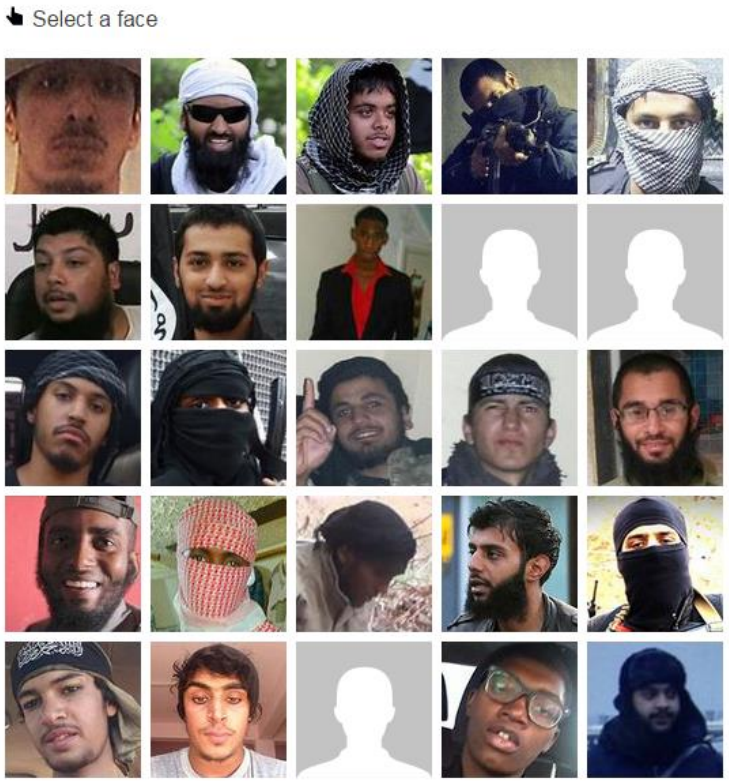
My Data becomes Open Data (by individual choice): There may be people who want to share their individual, personal, data openly to benefit others. A cancer patient could be happy to share their medical information if that could assist with research into treatments and help others like them.

The Right to Choose: if it's my data, just about me, I should be able to choose to access it, reuse it, share it and open it if I wish. Where open data should be freely available for use, reuse and redistribution by anyone, we could think that "my data" should freely available for use, reuse and redistribution by me

This means "my data" is an important source but also that it is essential that the open data community have a good appreciation of the pitfalls and dangers here – e.g. when anonymization or aggregation may fail to provide appropriate privacy. The more important point is that the open government community should consider precisely how such considerations should be made, what resources government data publishers need to make smart decisions, and where the red lines are for protecting privacy and individual well-being in the data's march towards public good and transparency.16

[15] http://blog.okfn.org/2013/02/22/open-data-my-data/#sthash.zo14wGXL.dpuf
[16] http://techpresident.com/news/wegov/24895/what-does-privacy-have-do-open-government

Privacy: Example



At least 700 people from the UK have travelled to support or fight for jihadist organizations in Syria and Iraq, British police say. About half have since returned to Britain. Most of those who went to the conflict zone are thought to have joined the militant group that calls itself Islamic State.

This BBC News database[17] details the stories of over 100 people who have died, been convicted of offences relating to the conflict or are still in the region. The information on these pages has been compiled from open sources and BBC research.

Questions:

- What is the news value of publishing personal details in each case?
- What does the story gain by the personal data?
- What would the story lose if the personal data was removed?
- How could individuals in this story be impacted by the release of their personal details?

[17] http://www.bbc.com/news/uk-32026985

Data Privacy: Exercise

List at least one argument in favor of and one argument against releasing data in the following situations:

- Neighborhood crime data
- List of individuals infected by Ebola with street addresses
- List of families receiving government financial assistance
- Names of locations of hospitals and prisons with the highest fatality rates

Data Denial for Privacy Reasons

Avoiding data denial:

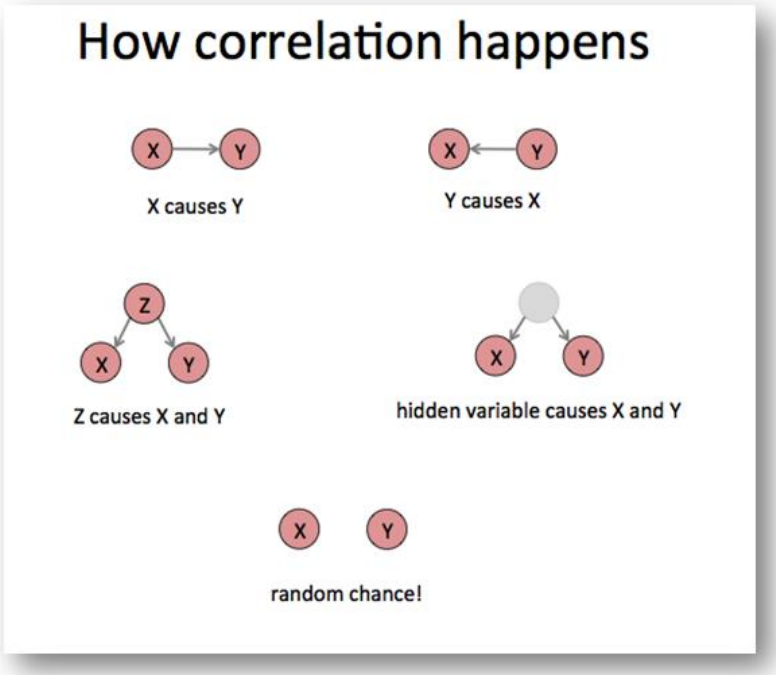
- Be specific about your data request
- Request the data in a .CSV or Excel file
- Ask to see the documentation of allowable fees
- Be persistent about your data requests
- Don’t ask for personal data

To avoid these excuses, be very specific about your data request including the geographical area, time period and institutions you are requesting data from. Also request the data in a .CSV file or Excel that can be emailed or transferred to you on a pen drive and ask to see the documentation of allowable fees. Be persistent about your data requests: call or visit until it becomes a bigger hassle to deal with you than it does to fulfil the data request. Don’t ask for personal data including phone numbers, addresses or national ID numbers about subjects.

Questions

- In your country context, what data is considered particularly sensitive and what are the arguments for and against releasing it?
- Who would benefit from the data remaining private and who would benefit from the data being public?

Confusing Correlation and Causation (1 of 3)



Correlation is a mutual relation of two or more things. Or in other words, correlation is when the value of two variables goes up or down in a similar pattern.

The tricky thing about correlation is it is very difficult to know which variable is influencing the other, if there is any relationship at all[18].

When writing a headline, it is impossible to directly assert a relationship between one variable and another.

[18]https://source.opennews.org/media/img/uploads/article_images/correlation_1.png