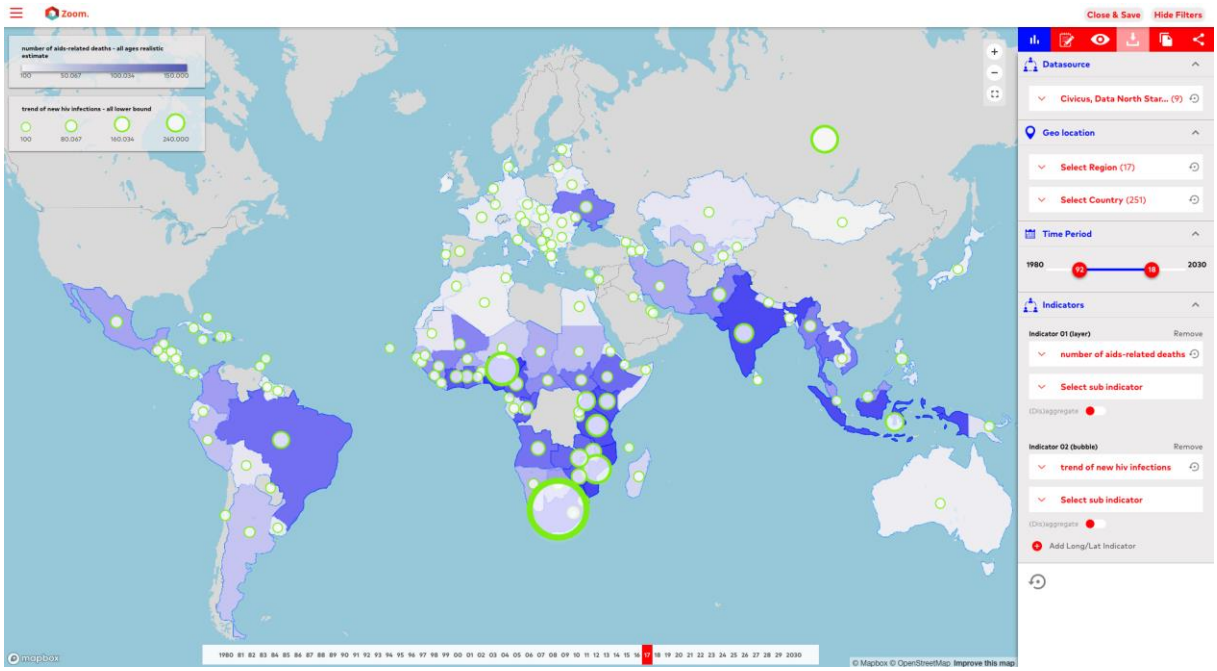
Data guidelines

June 2019



aidsfonds

aidsfonds

1 Welcome to ZOOM

1.1 Welcome

Welcome, we are delighted that you are using Zoom!

Welcome, we are delighted that you are using Zoom!
These guidelines provide you with all the know-how on the practical use of Zoom, and equip you with some key skills to understand, manage and work with data.

This document explains key concepts when working with data like different data types, spurious correlations and visualization types. There is also a separate Zoom manual which provides more detailed step by step guidance on the use of Zoom.

Disclaimer
Any visualisations, correlations and analysis created in Zoom can be inaccurate. Content created in Zoom should only be used when a tested methodology was applied and context has been taken into account. Zoom does not warrant or make any representation regarding use or the results of use of the content in terms of accuracy, reliability or otherwise.

1.2 Introduction to ZOOM

Zoom is a platform for data-informed programming in ending the Aids epidemic.
The first version was developed by Aidsfonds in cooperation with Leiden University and Zimmerman & Zimmerman in 2017. Zoom allows Aidsfonds' staff members and partners to make use of relevant open and private datasets to retrieve up-to-date insights into (sub) locations and (sub) groups, to enable data-informed choices about where and how to work, and to better show the impact of project spending.

The Zoom platform allows authorized users to upload, clean, analyze and visualize datasets. The data and created visualizations can be shared publicly or privately.

If you have a question about Zoom please contact: data@aidsfonds.nl

1.3 Data dictionary

Data: information in a form suitable for storing and processing by a computer. This can be in alphabet, numbers or symbols. Data refers to, or represents, conditions, ideas, or objects.

Data analysis: With data analysis you focus on understanding the data, which can have different goals like to inform decisions or discover useful information (like trends and patterns).

Data visualisation: Data visualization is the process of transforming data into a visual to help people understand patterns, trends, and insights. A core in data visualisation is that the visual is created for a target audience.

Indicator: An indicator is something that points to, measures or otherwise provides a summary overview of a specific concept. For examples indicators are used to measure outputs against program goals. For example the goal could be: “By 2030, end the epidemics of AIDS” and an indicator to measure this is: “Number of new HIV infections per 1,000 uninfected population, by sex, age and key populations”.

Variable: A variable is any characteristics, number, or quantity that can be measured or counted. Age, gender, income and expenses, country of birth, capital expenditure, class grades, eye colour and vehicle type are examples of variables.

2 Why are you using ZOOM?

2.1 Start with the why

Before using Zoom, ask yourself, why do I want to use Zoom?
Most people don't just play with data for fun. They have a story to tell or a problem to solve. These objectives can be very different, examples are: I want to improve my programming by make use of data, I want to communicate about our work, I want to show impact of our work with data, I want to discover patterns and trends, I want to increase my data literacy by playing around with data.

Now think of: why are you currently using Zoom? What is your objective?

When you think the story you want to tell also think of who your target audience is: Who do you want to reach with the analysis and visualisation you are creating in Zoom?

Chapter 5. Building your story, explains in more detail how you can create a story around your data.

2.2 Defining your question

You have thought of why you are using Zoom, now it is time to specify this further by formulating the question you want to answer using data.

Examples are: What is the trend line of new HIV infections in Kenya between 2000 and 2017?
Is there a correlation between Aids related deaths and ART coverage worldwide?

So how do you formulate a question for the story you want to tell?

When formulating your question try to incorporate the following elements: the where, when, who and what.

Where – is your question location based?
- One location: Worldwide, one country or one sub national location.
- Multiple locations: Countries or sub-national locations.

When – does your question have a time dependency?
- One specific time period : for example a certain year.
- Multiple time periods : for example multiple consecutive years or a baseline and endline time.

Who - does you question focus on a specific target group?
- All people
- Specific people - for example certain age group, gender, community, profession etc. (you can also have a combination of variables, for example women between 15-30).

What - What type of analysis do you want to do with the data?
- Compare data
- Over time (for example: a trend line)
- Among indicators/variables

- Among different groups of people
- From different programmes
- Discover a relationship between multiple indicators or variables.
- Show how your data is distributed.
- Visualize a composition to show how individual parts make up the whole of something (for example: answers to a certain question split up by gender).

If we take our previous example question: What is the trend line of new HIV infections in Kenya between 2000 and 2017?

We can divide it up in the following, where, when and what:
- Where: Kenya
- When: Between 2000 and 2017
- Who: not defined, so we assume all people.
- What: A trend line (which is comparing data over time).

Thinking of the where, when and what of the question you want to answer will give you guidance in finding the datasets you will need to answer your question. This is even more important when you want to use multiple data sets. As then you need to make sure that the datasets contain the same where, when and who so you are able to compare the data.

For example if you have the question: What is the difference of new HIV infections between women age 15-30 and men age 15-30 in 2016 in Kenya?

And you have two datasets:
1. One dataset with information on HIV infections of men in Kenya
2. One dataset on HIV infections of women in Kenya.

In order to say something relevant you need to find the common variables within the datasets so you are able to use both datasets. Specifically for the where, when and who this is important:

- *Where* – If – as in the example – you are interested in data from Kenya you will need to check the completeness of the datasets. For example: if one dataset is missing data on a county you won't be able to say anything on the whole of Kenya (because you miss values). To say anything relevant you need to find the common 'where' between the two datasets. Maybe both datasets have data on 3 counties (let's say Mombasa, Nakuru and Kisumu), then you can edit your question to 'What is the difference of new HIV infections between women age 15-30 and men age 15-30 in 2016 in Mombasa, Nakuru and Kisumu?

- *When*: Your data needs to be of the same year to be able to compare it. It would not make sense to compare new infections for women in 2012 with new infections for men in 2016. Comparing datasets from different years can be relevant when investigating a trend line for the same target group, for example: "What is the difference of new HIV infections between 2012 and 2016 of women age 15-30 in Kenya?"

- *Who*: Your target group needs to be the same, for example if you have one dataset containing data on new HIV infections of all women (age 0 - 100), and the other dataset containing data on new HIV infections is of men age 15 - 30, the 'who' is different and can't be compared.

Sometimes the specific data you need is not available. You can then decide to change your question, or not use the data. If you do choose to use incomplete data for your analysis, make sure to communicate the limitations of the data to your audience.

After you formulated your question also check within our organisation or online if it might already be answered. So you don't duplicate efforts and learn from experiences of other people.

3 The data

3.1 Data types

There are different types of data. The two major categories are qualitative and quantitative data.

Qualitative data is information about qualities of something; information that can't be measured in numbers. For example information collected in in-depth interviews, direct observations or written documents.

Quantitative data is information about quantities. That is, information that can be measured and written down with numbers.

All quantitative data can be visualized in Zoom, as an addition it also recognizes latitude and longitude which makes visualisation of points on a map possible. Not all qualitative data can be visualized in Zoom. There are two types of qualitative data that can be visualised in Zoom:

Categorical data: is data that can be put into categories. For example Yes, No, Maybe, Don't know. But also different knowledge levels, gender and age categories, etc.

Location data (names): with location data we mean written country names. Zoom recognizes countries when the country name is spelled correctly or official ISO country codes are used. Every country has an unique 2 and 3 letter ISO country code. For example for the Netherlands the official ISO country codes are: NLD (official 3 letter country code) and NL (official 2 letter country code). For some selected countries sub national data is also recognized when spelled correctly.

3.2 Data access

Which data you can see and use in Zoom depends on the access you have to the data based on your user role and permission (To read more about user roles and permission see *Chapter 7. User management*)
There are three data access types: open, shared and closed. All three can be found in Zoom.

Open data is data that can be used, reused and redistributed freely by anyone. Open data meets the following three criteria:

1. Accessible (ideally via the internet) at no more than the cost of reproduction, without limitations based on user identity or intent.
2. In a digital, machine readable format for interoperation with other data (so for example not in a PDF).
3. Free of restriction on use or redistribution in its licensing conditions.

Shared data is data that is shared only with named people or organisations.

Closed data is data that can only be accessed by its subject, owner or holder.

Zoom consists of open, shared and closed datasets. All users can access the open datasets. User management enables that authorisation can be given and restrictions set to access and use of shared or closed data sets (To read more about user management see *Chapter 7. User management*).

3.3 Data collection methods

There are broadly two sorts of data collection categories: primary data and secondary data.

Primary data are those that you or your team has collected yourself.
Secondary data come from other sources - perhaps administrative data of a government partner, survey data from another organization, study, or statistics bureau. An example is the UN AIDS or World Bank data.

Before you decide to collect data yourself, check what data is already available in Zoom, within your organisation or publicly. You might find out that the data (and maybe even analysis) you need has already been done.

There are different methods in which data can be collected, here are six of the most used:

Questionnaires - List of a research or survey questions asked to respondents, and designed to extract specific information. It serves four basic purposes: to collect the appropriate data, make data comparable and amenable to analysis, minimize bias in formulating and asking question, and to make questions engaging and varied. This data is often quantitative or categorical, but it can also maintain qualitative and location data. *Examples are:* HIV/AIDS knowledge survey, Human rights databases, stigma index.

Direct observations - In direct observation, a researcher watches and records while an action is happening. This can be noted as qualitative data (when it is described) or categorical data (when is what is seen is put in categories) quantitative information (for example when services, people or the environment are scored on knowledge, attitudes or behaviour).
Examples are: scorecards, mystery client tool, observations.

Self reporting - A self-report study is a type of survey, questionnaire, or poll in which respondents read the question and select a response by themselves without researcher interference.
Examples are: SMS surveys, use of a chatbot.

The three methods mentioned above often provide quantitative data that can be added and used in Zoom. The three methods below generally provide qualitative data, this data is usually not in Zoom, but it can be used to build your story and provide context to your data visualisations. To learn more about that see *Chapter 5. Building your story.*

Focus Group Discussions - At a focus group discussion a group of people is brought together to discuss a specific topic of interest. A moderator or facilitator introduces the topics, guides the discussion and stimulates group participation. The outcome of the discussion usually is a qualitative report.

Qualitative Interviews - In qualitative interviews the interviewees are given space to expand their answers and accounts of their experiences and feelings. Moreover, their answers are not pre-categorised in the interview schedule. The aim is often interpretation and understanding of how and why. The data from qualitative interviews is qualitative data. Some information can be transferred into categorical or quantitative data for example when gender, age, location of the interviewee is collected.

Case-study - a case study is a research method involving an up-close, in-depth, and detailed examination of a 'case' which can be a person, group or situation. It is often over a period of time and takes into account contextual conditions.

All methods can be used to get information on one moment in time, or be used to monitor over time (for example a baseline questionnaire at the beginning of the project and endline questionnaire at the end of the project to measure the possible impact of the project). If you have a one off survey or a monitoring survey this has effect on the analysis or visualisation you can do, you can learn more about that in *Chapter 4.2. Data visualisations*.

3.4 Data quality

When you check the data quality, you check if the data is fit for its intended uses in operations, decision making and planning. As when your data quality is bad, you should not use it in your operations. As noted before there are different data sets in Zoom, here we focus on how you can check data quality of all datasets (primary and secondary) in Zoom and dive more in specifics for datasets that you have collected yourself (primary data sets).
It is difficult (especially for secondary datasets – as you often don't know how they have been collected) to analyse the quality of the data. It is important to understand that all datasets will have their own limitation. Awareness of that can support you in deciding when you should or should not use the data.

3.4.1. All data sets in Zoom (primary and secondary data)

Check metadata
Metadata is 'data about the data set' it gives you information on: who the owner of the data is, when the data has been collected, how it has been collected, source, type of access, when the data set was last updated etc. This information is critical as it provides you with some basic data quality checks. For example when you are mapping health facilities on a map, but you find out that the dataset you are using is from 2000, this might not be relevant anymore. Appendix A shows all metadata fields collected in Zoom.

Check the completeness of your dataset
Knowing what data is in the dataset you are using is key to understanding your dataset and knowing what kind of analysis you can do with that data set.
When you are using a dataset, take some time to critically review it to find out what data it actually contains.

Imagine that you want to see the average of the Aids related deaths worldwide and you plot that in a trend line. But then you take a better look at the dataset and find out that of 195 countries that exists the dataset you have used contains data of only 90 countries. Should you then create a worldwide trend line? The answer is no, as it only represents <50% of the countries in the world.

Another example - imagine you want to map the results of a HIV knowledge survey on municipalities in the Netherlands. When you take a look at the data you find out that there are three municipalities that only have 1 respondents. Can you then say something relevant of the knowledge per municipality? No, as one person cannot be found representative for the knowledge level of a whole municipality
To solve this problem, you could broaden the scope. When you look at province level find all the provinces have more than 100 respondents, which can then allow you to say something relevant of the knowledge per province.

3.4.2. Data sets where you are involved with the collection (primary data)

Data collection

When you are collecting data you need to be aware of the following seven elements before, during and after the data collection:

1. Reliability
2. Representativity
3. Validity
4. Confidentiality
5. Objectivity
6. Comparability
7. Participatory

These elements are part of Aidsfonds' Planning Monitoring, Evaluation and Learning (PMEL) guidance and are key in assuring the quality of your data.

Data Cleaning

As you are interested in using Zoom to upload your data, we assume that you have data collected and have it in a spreadsheet. To get to this point you have spent time and energy designing a questionnaire, the surveyors have worked with respondents to ensure quality data collection. Now, it is time to turn all this hard work and preparation into meaningful insights.

What do you do first?

Before you begin analyzing your data, it is crucial to ensure that your data set is complete and correct. The data cleaning process helps you achieve this by detecting and removing errors and inconsistencies in your data. After all, you don't want to be making important decisions based on wrong insights simply because your data had errors!

If you like to learn more about data cleaning, SocialCops has developed a comprehensive e-book on basic data cleaning skills with step by step guidance. Or check out the school of data course on data cleaning. It is required to clean your dataset before uploading it to Zoom.

3.5 Personal identifiable information

We want to ask you not to load any personal identifiable information into Zoom. Personal identifiable information is information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context. The reason for not storing this type of data is that when data is in the wrong hands this could violate privacy or even form a risks for the people named in the dataset. If you have any dataset you want to upload we want to ask you to take out any personal identifiable information like, name, phone number, address. If you have any concerns if a dataset might contain personal identifiable information, please check with the Zoom team first before uploading the dataset.

4 Data analysis and visualization

Now you have defined the question you want to answer, the story you want to tell and the data you want to use, it is time for the next step, actually working with the data. Here you have two options:

1. Data Analysis
2. Data visualisation

Data analysis: With data analysis you focus on understanding the data and identifying or discovering the trends and patterns of the data.

Data visualisation: Data visualization is the process of transforming data into a visual to help people understand patterns, trends, and insights. A core in data visualisation is that the visual is created for a target audience.

4.1 Data Analysis

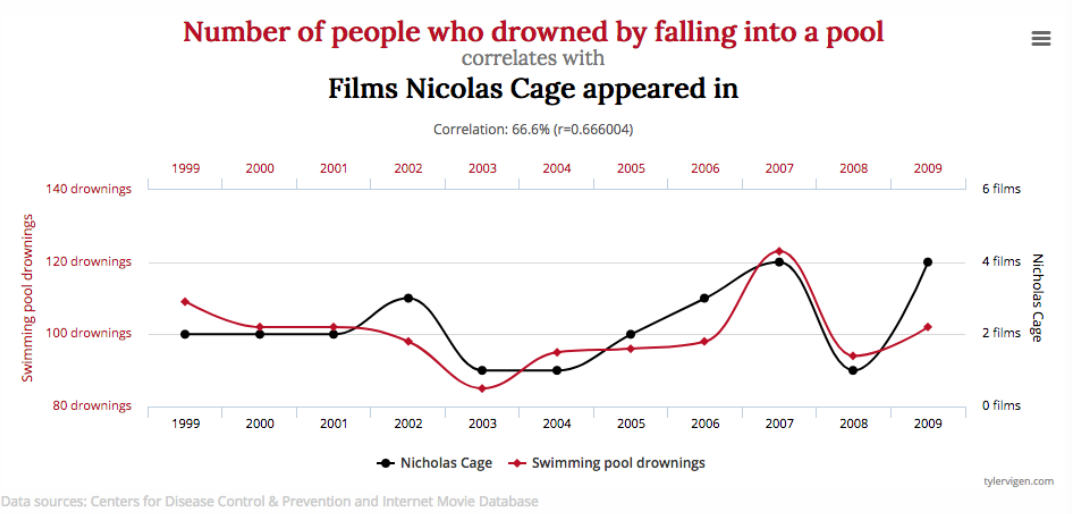
When you want to use Zoom to identify or discover trends and patterns in the data, be aware that Zoom can plot two variables but is not a statistical method to discover correlations between data. To discover correlation between data we advise you to use tested statistical correlation measurement methods like the Pearson correlation coefficient. When you are trying to discover a relationship between two indicators it is very important to be aware of the possibility of spurious correlations. If you like to learn more about data analysis, the school of data has a nice online course called introduction into exploring data.

4.1.1 Spurious correlation

What if you plot two indicators in one chart and you see a relationship?
First thing to note is that correlation doesn't imply causation - which means that if two indicators correlate this does not mean one thing was *causing* the other.

In his book "spurious correlation" Tyler Vigen (2019)[1] shows multiple examples of spurious correlations. The example below seems to show a clear correlation between number of people drowned by falling into a pool and films Nicolas Cage appeared in.

[1] Vigen, T. (2019). *15 Insane Things That Correlate With Each Other.* [online] Tylervigen.com. Available at: http://www.tylervigen.com/spurious-correlations [Accessed 1 May 2019].



This is a obvious case of a spurious correlation as we all know that Nicolas Cage movies aren't actually causing drownings. But why does it matter so much?

It is important not to take a correlation as a causation as you could be misleading your audience.

4.1.2 Difference between causation and correlation

What is the difference between causation and correlation?
In a causal relationships, one event prompts the other. Which means that one factor causes the other factor. For example there is a causal relationship between the weather and icecream sales, when the temperature is higher sales go up.
Correlation, on the other hand, refers to any relationship between two factors — they move together, but it's not necessarily that one causes the other.

4.1.3 Explaining to your audience

If you find a correlation and want to share it, be aware of the wording you use. Data storytelling is like any other kind of narrative: your words matter. You have to choose them carefully. Heather Krause (2018)[2] gives a nice overview of words you can use for correlation and which imply causation.

Words That Are Fine for Correlation	Words That Imply Causation
Get	Cause
Have	Increase
Linked	Decrease
More	Benefits
Less	Impacts
Tied to	Enhances
Connected	Affect
Related	Improves
Tend	Worsens

[2] Krause, H. (2018). *Are You Falling Victim to Spurious Correlations? (Here's How to Stop) - Datassist.* [online] Datassist. Available at: https://idatassist.com/falling-victim-spurious-correlations-heres-stop/ [Accessed 1 May 2019].

4.1.4 Consult a data analyst or scientist

If you think you possibly found a relevant causation or correlation, consult a data analyst or scientist to go through your data to test if there is a actual correlation. As you don't want to mislead your audience with spurious correlations.

4.2 Data Visualisation

Two reasons why data visualisations are powerful to show your data:

1. **Visuals are more intuitive than text alone** - Text has a hard time competing with visuals. This is evident when it comes to delivering any piece of communication that is meant to be easily digestible, like an executive summary or report. We are hard-wired to find emotional cues within visuals. Photos, drawings, shapes, and colors can be quickly interpreted, which is why data visualizations are so effective.

2. **Data visualized properly leads to information and knowledge** - When you process data, analyze it, and visualize it, you are adding value to it. That data now has more value because it has been transformed into information. Once that information is consumed (received, understood, remembered, and accessible) it becomes knowledge.

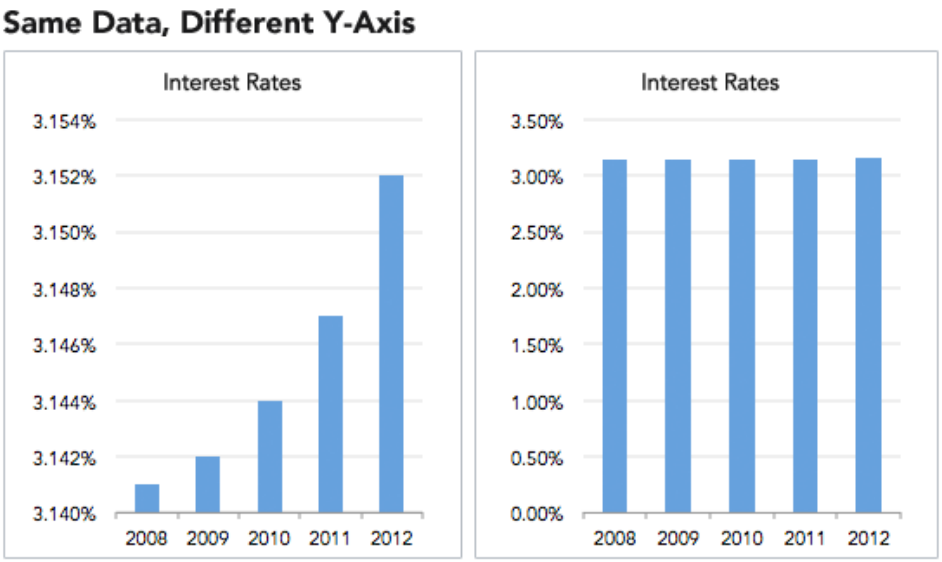
4.3 Lying with data visualisations

Although data visualisations are one of the most important and powerful tools we have to we have to show and share data. But it's just as easy to mislead as it is to educate using charts and graphs. Ravi Parikh (2014)[3] show three common ways to lie with data visualizations, two of the examples he notes are changing the Y-axis and use of cumulative graphs.

4.3.1 Changing the Y-axis

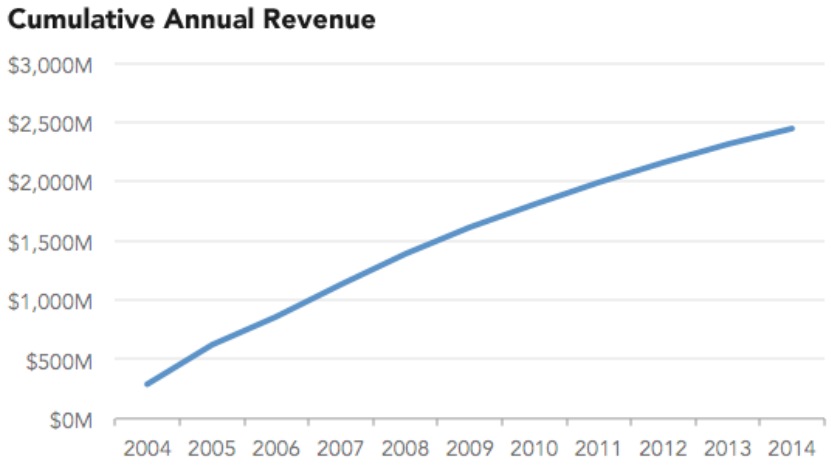
One of the easiest ways to misrepresent your data is by messing with the y-axis of a bar graph, line graph, or scatter plot. In most cases, the y-axis ranges from 0 to a maximum value that encompasses the range of the data. However, sometimes we change the range to better highlight the differences. Taken to an extreme, this technique can make differences in data seem much larger than they are. Let's see how this works in practice. The two graphs below show the exact same data, but use different scales for the y-axis:

[3] Parikh, R. (2014). *Three Common Ways to Lie with Data Visualization*. [online] Heap | Mobile and Web Analytics. Available at: https://heap.io/blog/data-stories/how-to-lie-with-data-visualization [Accessed 1 May 2019].

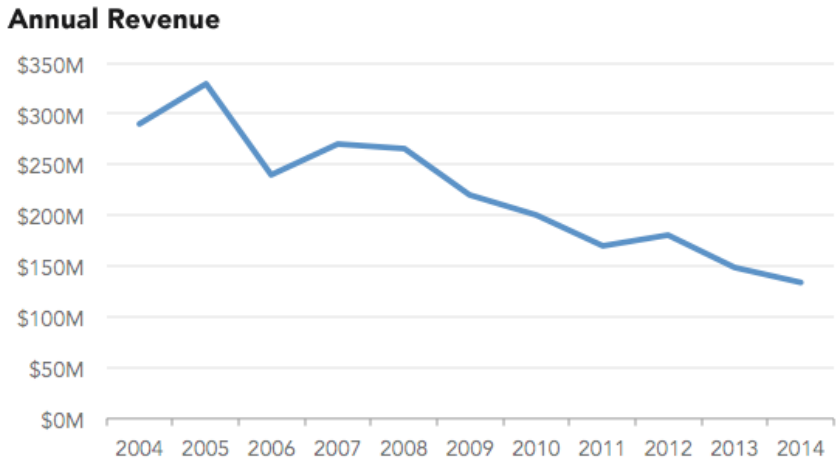


4.3.2 Use of a cumulative graphs

Many people opt to create cumulative graphs of things like number of users, revenue, downloads, or other important metrics. For example, instead of showing a graph of our quarterly revenue, we might choose to display a running total of revenue earned to date. Let's see how this might look:



We can't tell much from this graph. It's moving up and to the right, so things must be going well. But the non-cumulative graph paints a different picture:



Now things are a lot clearer, it shows that revenues have been declining for the past ten years. If we scrutinize the cumulative graph, it's possible to tell that the slope is decreasing as time goes on, indicating shrinking revenue. However, it's not immediately obvious, and the graph is incredibly misleading.[4]

4.4 Design with your audience in mind

Thinking about the intended audience, or what the purpose of your visualisation is, before you start making it will help you to create the most appropriate visualisation for your needs. It can also help you save time when visualising your data, as you'll only include elements that'll be useful for your intended audience. Kyle Maloney (2019)[5] of the University of Sydney defines three audiences groups:

Yourself - Visualising your data can be an important part of data analysis during research, allowing you to discover trends in your data, find relationships, or even plan out future data collection. In making this kind of visualisation, you don't need to be producing a polished end product. Instead, you should develop a workflow that allows you to visualise your data quickly, and in a repeatable manner. It can be wise to document your visualisation workflow to allow you to re-use it in future analyses, and to ensure that you know how you produced a particular visualisation. Documentation may seem dull, but in a few months, you'll be very glad you did it!

Others in your field - Visualisations can be an important part of sharing your results with colleagues or organisations you work with, through figures in papers or posters, or as part of presentations. These visualisations are showing off your work, and so need to be polished, clear, and well annotated. You should consider employing figure types that are commonly used in your field, as familiarity with the layout will help your audience to quickly grasp your results.

Public outreach - Visualisations can be an excellent way to engage the interest of non-specialists in the results of a research project. These visualisations should be kept very simple and clear - don't try to display too much competing information on a single visualisation! Clear annotations and descriptions of what's being displayed should be provided, but make sure to avoid using any specialist terminology in your descriptions. It's a good idea to highlight any key points on your visualisation, as a public audience might lack the specialist knowledge to pick out the main messages without prompting.

Next to defining your main audience it is also important to take into account the accessibility and possible cultural sensitivity of the visualisation you make.

Cultural and accessibility - It's a good idea when creating a visualisation to consider whether your audience might include people with visual impairments, such as colour blindness. To keep your visualisation accessible, ensure that colour isn't the sole way of conveying information by using patterns and labels to help differentiate the information. You can also find colour hues and saturation that will still be differentiable to colour blind vision.
Resources like ColorBrewer allow you to come up with colourblind safe (as well as printing in grey scale safe) colour palettes.

[4] Parikh, R. (2014). *Three Common Ways to Lie with Data Visualization*. [online] Heap | Mobile and Web Analytics. Available at: https://heap.io/blog/data-stories/how-to-lie-with-data-visualization [Accessed 1 May 2019].

[5] Maloney, K. (2019). *Subject guides: Data Analysis and Visualisation: Creating a Visualisation*. [online] University of Sydney. Available at: https://libguides.library.usyd.edu.au/c.php?g=508301&p=3476685 [Accessed 1 May 2019].

When using symbols and patterns in your visualisation, be aware of possible cultural sensitivity of some symbols or colours to your audience. In some areas maps can also be sensitive to certain audiences due to contentious borders.

4.5 Use of common conventions

Common conventions for visualisations should be used if possible, as not doing so can introduce unnecessary confusion. For example, maps are normally drawn with north to the top of the page; a map that's oriented in a different direction stands a higher chance of being misunderstood by readers. Also make sure that your numbers add up. For example when you use a pie or donut chart the total percentage should add up to 100%.

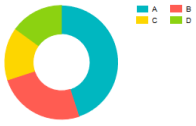
4.6 Choosing your data visualisation

When you have formulated your why, your question, have a clear idea who your audience is and found your data. How do you decide which visualisation type you will use?

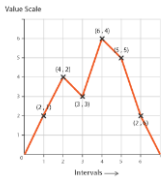
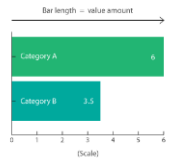
Some types of visualisations are encountered more frequently than others in everyday life. These visualisations will have an advantage in that most viewers will know how to interpret them without needing any additional instructions. More novel visualisations can be eye-catching and help to engage the interest of your audience, but it's important to keep in mind that they'll also take longer for your audience to understand, and may require an explanation.
Make sure to always keep the goal of your visualisation in mind and assess whether the visualisation you've chosen is the most effective way of achieving that goal.

To choose the right visualisation is necessary to know that different visualisations highlight different aspects of the data, and therefore, they fulfil a different function in displaying data. The following visualisations are available in Zoom, we split them up between graphs, maps and table. In the table below you can find which visualisation types can be made in Zoom and when you do and don't use them.

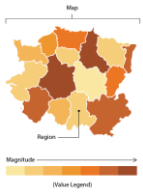
4.6.1 Visualisation types in Zoom[6]

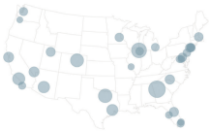
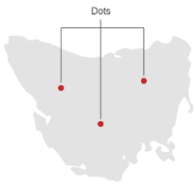
Type	Description	Do use when..	Don't use when..
Donut Chart 	A Donut Chart shows how a single entity is divided into several components. Although It can be seen as a Pie Chart with the center cut out, there is no lack of information in using Donut Charts.	...the full ring must show 100% of the data ...if the point is to compare the part to the whole. ...you want to give a quick idea of the proportional distribution of the data. ...you have categorized data	... you have more than 6 categories. ...the aim is to show a trend over time. ...you want to make comparisons. ...you want to compare values - use Bar Charts instead. ...the category values are too similar or too different. ...you have to display real-time data.

[6] For more detailed information on charts and maps in Zoom see the 'Zoom data visualisation guidelines':
https://docs.google.com/document/d/1WcDwmBBGqkxX3Kpw_jcWFDZ1lQ7W7iLIn0_x_ZTu9H0/edit

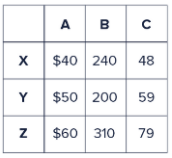
Line chart 	A Line Chart represent quantitative data over a continuous variable, which is usually temporal. The Line Chart (or Line Plot, or Line Graph) displays information as data points connected by straight lines	...you have to represent quantitative data in a timescale; ...changes over the time are small; ...you are looking for a trend over the time; ...you have to represent quantitative data over a ordinal attribute (but carefully).	...you are not going to show axes labels. ...you want to compare the same quantities plotted on different scales. ...you need to represent more than 3-4 lines per graph: the risk is that it is very difficult to read. ...you have to represent categorical data: this may encourage conclusions like: "The more male a person is, the taller he/she is". ... you don't need to compare value over the time (use a Bar Chart instead).
Bar chart 	A Bar Chart use rectangular bars with lengths proportional to the values they represent. Several kinds of Bar Chart can be distinguished depending on the variables plotted. However, in all of them these variables must belong to the same hierarchical level	... you want to compare different values that are hierarchically equivalent; ... it is important to quantify the difference between those values.	... Never use to compare values with different units or hierarchy. ... you are going to plot your data on three dimensions.

4.6.2 Maps available in Zoom

Type	Description	Do use when..	Don't use when..
Choropleth map 	A thematic map divided in different areas, which are colored or patterned depending on a categorical value or in proportion to a quantitative value.	...you want to visualize how a certain attribute varies across geographic areas. ... if you have a significant number of data per area. For example if you map respondents answers per municipality in a choropleth map, but if you have less than 10 respondents in certain areas, this is not significant.	... the data set has geographical data that's not relevant to your use case, or if the geographic areas have too different dimensions. Use a Bar Chart instead. ... if you don't have a significant number of data per area. - Make the administrative areas bigger (for example instead of using municipality, see if you have significant data for province)

Bubble map 	Bubble maps can be viewed as a combination of a map and a Bubble Chart, where the size of the circle represent a numeric value, and it is related to a geographic point or area	... you want to compare proportions over geographic regions. Since the size of the circles is not dependent on the geographical area over which it stands, it can be used when difference between areas are big.	...differences between values are too small
Dot map 	The Dot Map, also known as Point Map shows the distribution of phenomena as dots on a map. Each of these dots can correspond to: - One phenomenon, so there is a one-to-one relation between data and the visual dot - Many phenomena, a dot corresponds to an aggregated value, for example with census data, one point may represent 1,000 people..	Dot Maps highlight variation in pattern, such as clustering, thus they enable the user to visualize intuitively data distribution. Area with many dots correspond to a high concentration of phenomena, while a low concentration is detected in areas there dots are sparse. (note you need a dataset that includes latitude and longitude data to be able to create a dot map)	Dot maps are bad at displaying absolute quantities: don't use them for retrieving exact values

4.6.3 Table available in Zoom

Type	Description	Do use when..	Don't use when..
Table 	A table chart is a means of arranging data in rows and columns.	... you need to provide precision of information/value. This can be more clear in a table then a visual chart.	... you want to show trends and don't need to show detailed information/values. ...you have too many details, your table will be too large and incomprehensive,

	A	B	C
X	$40	240	48
Y	$50	200	59
Z	$60	310	79

5 Building your story

You have now learned how to formulate the question you want to answer, find the data and visualize your selected data in Zoom. Now it is time to build a story around your visualisation. The phrase "data storytelling" has been associated with many things: data visualizations, infographics, dashboards, data presentations, and so on. Too often data storytelling is interpreted as just visualizing data effectively, however, it is much more than just creating visually-appealing data charts. Data storytelling is a structured approach for communicating data insights.

Stories are easier to remember then data or a visualisation alone.
For example: If you ask somebody to remember a grocery list, they usually have trouble repeating the full list back to you. But when you tell it to them in a story, a lot more groceries are remembered. For example I ask you to memorise the following items:

- Bread
- Soap
- Butter
- Wallet
- Cheese
- Peanut butter

It is easier to remember when put in a story:

"The women was washing her hands with soap because she was planning to make a cheese sandwich, but was out of bread, butter and cheese. The only thing she had was peanut butter but she does not like that. She took her wallet and went to the supermarket "

The same goes with data and visualisations, they are easier to remember when it is shared with a narrative. Narrative is — along with visual analytics — an important way to communicate analytical results to non-analytical people. The data gives you the *what*, but people and stories know the *why*.

This is important because data in itself does not put the findings in a context or tell the story about the people behind the data. It should not be forgotten that data and visualisations are about real people with real lives, needs and challenges. Showing spurious correlations or displaying information out of context can have an negative effect on peoples lives and organisations.

For example there is an organisation that receives the same amount of money in 2017 as in 2018. But in 2017 it reached more people. In the given example, it could mean that the organisation has acted less efficient, but other possibilities are that the environment has deteriorated, making it more difficult to reach people or treatment has become more expensive. This is why it is important to provide a narrative to the data to explain the why.

5.1 Type of data stories

Thomas Davenport (2014)[7] defines four key dimensions that determine the type of story you can tell with data and analytics

1. Time
2. Focus
3. Depth
4. Method

Time: Stories can be about the past, present, or future. The most common story is about the *past*, a reporting story which described what happened in s certain period of time (for example, week, month, year). Stories about the *present* focus on what is currently happening. Stories about the future are predictions, they use predictive analytics. They take stories from the past to create a statistical model, which is then used to predict the *future*.

Focus: Are you trying to tell a what story, a why story, or a how to address story? *What* stories are a reporting stories—they tell what happened. *Why* stories go into the underlying factors that caused the outcome. *How to address* stories explore various wa+ys to improve or address the situation identified in the what and the why stories. A complete story often has what, why and how to address in it.

Depth: Lightweight or in-depth. *Lightweight* stories are relatively small, ad hoc investigations to find out why something was happening. *In-depth* stories are long, analytically-driven searches, for example to find a solution to a complex problem.

Method: This is based on the analytical method used. Are you trying to tell, for example, a correlation story—in which the relationships among variables rose or fell at the same time. For example height and weight. Or a causation story, in which you'll argue that one variable caused the other? For example: Too much sun causes a burned skin. In most cases, doing some sort of controlled experiment is really the only way to establish causation. Like also noted in *Chapter 4.1.* you need to be careful with suggesting a causation of correlation.

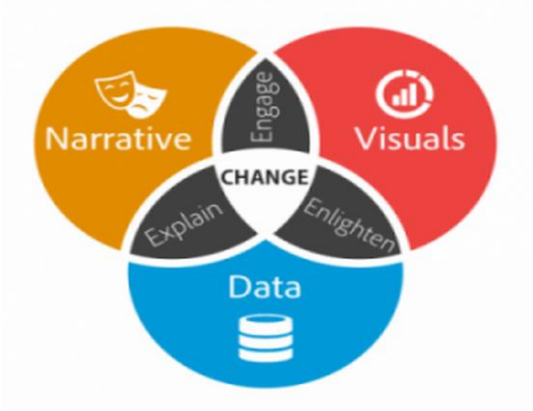
5.2 Elements of data stories

Data storytelling is a structured approach for communicating data insights, Brent Dykes (2016)[8] notes that it involves a combination of three key elements: *data*, *visuals*, and *narrative*. And that it is important to understand how these different elements combine and work together in data storytelling:

When narrative is coupled with data, it helps to ***explain*** to your audience what's happening in the data and why a particular insight is important. Context and commentary is often needed to fully appreciate an insight. When visuals are applied to data, they can ***enlighten*** the audience to insights that they wouldn't see without charts or graphs. Many interesting patterns and outliers in the data would remain hidden in the rows and columns of data tables without the help of data visualizations.

[7] Davenport, T. (2014). *10 Kinds of Stories to Tell with Data*. [online] Harvard Business Review. Available at: https://hbr.org/2014/05/10-kinds-of-stories-to-tell-with-data [Accessed 1 May 2019].
[8] Dykes, B. (2016). *Data Storytelling: The Essential Data Science Skill Everyone Needs*. [online] Forbes.com. Available at: https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/#2931cbea52ad [Accessed 1 May 2019].

Finally, when narrative and visuals are merged together, they can ***engage*** or even entertain an audience . It's no surprise we collectively spend billions of dollars each year at the movies to immerse ourselves in different lives, worlds, and adventures. When you combine the right visuals and narrative with the right data, you have a data story that can influence and drive ***change***.



5.3 Combining quantitative and qualitative information

So what kind of narrative can you add to your data to build a story around your data visualizations? Collected qualitative information (*see Chapter 3*) can provide a narrative for your data or visualisation. This can come from several sources like interviews, case studies or in depth research. You can also use your own point of view as a narrative, for example by reflecting on how a visualisation relates to the theory or change of your organisation, or add a personal story.

6 Sharing

6.1 Colours and symbols

As noted in *Chapter 4.4* when creating you visualisations it is important to be aware of colours and symbols that you use before sharing with your target audience.

6.2 Decide what to share

In Zoom it is possible to share both data and visualisation with selected people or publicly.

Before sharing a visualisation make sure to check if it does not reveal any sensitive information. As already established in *Chapter 3.5* there should not be any personal identifiable information in Zoom. But sometimes information on groups – like ethnicity or sexual preference - can be very sensitive too. And that information in the wrong hands can put people in danger. For example when a government official of a county where homosexuality is illegal finds out where LGBTQ meet up places are, this can put people at risk. So always be aware of what is in your dataset or visualisation before sharing it within your organisation, or publicly.

6.3 Provide context and metadata

In *Chapter 3.4.2* it was established that metadata is important when working with data. The same goes when you share visualisations. Provide the following metadata and context with your visualisation:

1. In the description: explain what is shown - what data are you using, what year, what version, how was it collected, by whom?
2. If you are using public datasets, provide a link to the source.
3. Tell people why they should care about your graphic. Why did you make this graph?
4. Be aware of your wording, don't mislead your audience. When you are displaying a correlation make use of correct wording as suggested in *Chapter 4.1.*
5. When relevant and not sensitive, provide a way for people to contact you if they have questions about the visualisation.

7 User management

7.1 User management and why it is important

Zoom has a user management structure set up to sure that people can access Zoom, and get the correct permissions to ensure they have access to datasets and visualisations important for their work. Most importantly, user management ensures that (sensitive) information in closed or private datasets is only shared with the people that should have access to it. Only a select number of people is able to manage users in Zoom. If you need access to a certain dataset or have a question about user management please contact data@aidsfonds.nl

8 Appendix A - Metadata fields

Type of data field	Mandatory /optional
Title of Dataset	Mandatory
Description	Mandatory
Data Source	Mandatory for external data sets
Organisation (owner of the data)	Mandatory
Year of Data Collection	Mandatory
Accessibility of dataset: Private, Public, On request	Mandatory
Accessibility of data (open - private - request)	Mandatory
Data quality / reliability*	Optional for primary data added by Aidsfonds and partners.
Date uploaded/ last updated	Automatically collected in Zoom

*Data quality / reliability questions

Concept	Possible questions	Type answer
Validity	1. Have you tested the tool (questionnaire/topic list) in a pilot or with a test group before conducting it?	Y/N/Don't know
Objectivity	2. The data contains information which can be considered sensitive (f.e. financial, health, food security information)?	Y/N/Don't know
Representativity	3. How did you select respondents?	open question
	4. How many respondents were interviewed/participated?	Nr field

9 Appendix B - Want to learn more?

Interesting links to increase your data knowledge and skills

Building data skills

- Data skills course: https://schoolofdata.org/courses/
- Video on open closed shared data: https://vimeo.com/125783029
- Data collection techniques: https://cyfar.org/data-collection-techniques
- Metadata https://www.opendatasoft.com/2016/08/25/what-is-metadata-and-why-is-it-important-data/

Data analysis

- Spurious correlation examples: http://www.tylervigen.com/spurious-correlations
- Spurious correlations: https://idatassist.com/falling-victim-spurious-correlations-heres-stop/

Data visualisation

- Creating a visualisation: https://libguides.library.usyd.edu.au/c.php?g=508301&p=3476685
- Data visualisation catalogue: https://datavizcatalogue.com/index.html
- The Data viz project: http://datavizproject.com/
- How to lie with data visualisation https://heapanalytics.com/blog/data-stories/how-to-lie-with-data-visualization

Data protection

- Data protection toolkit: https://www.mdc-toolkit.org/data-protection-starter-kit/

Building your story

- Type of data stories: https://hbr.org/2014/05/10-kinds-of-stories-to-tell-with-data
- Book on storytelling with data: http://www.storytellingwithdata.com/book/
- Data storytelling basics: https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/#6028c24a52ad