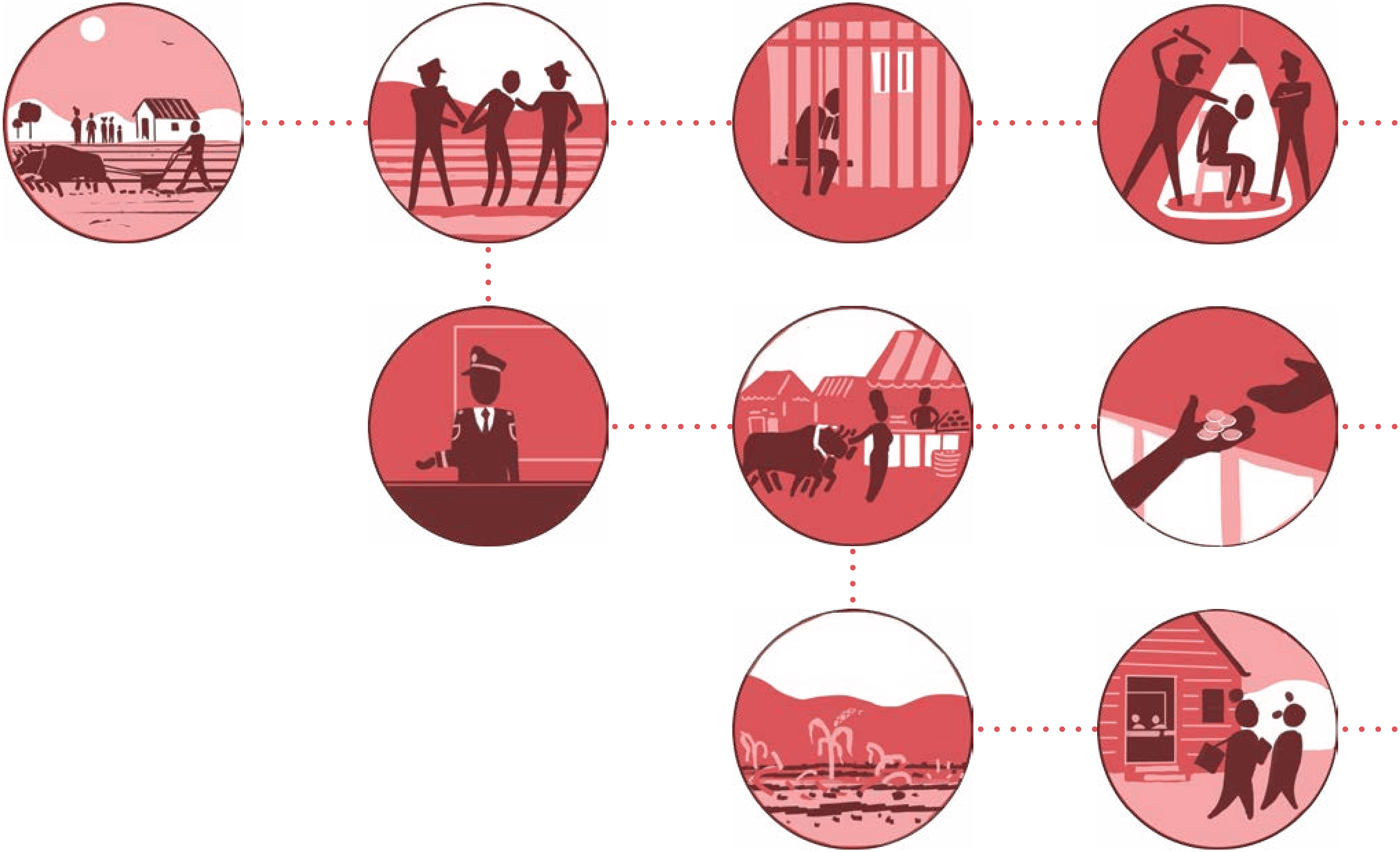
5 Hazards



How to Use This Booklet

Data visualization can make information more memorable, more persuasive, facilitate understanding and ultimately motivate action. And within human rights research, it can help investigators and researchers draw a bigger picture from individual human rights abuses by allowing them to identify patterns that may suggest the existence of abusive policies, unlawful orders, negligence, or other forms of culpable action or inaction by decision-makers.

The purpose of this activity is to explore some of the ideas and principles around designing effective data visualization for human rights advocacy.

This activity is broken into a series of six topics each with its own PDF guide.

In practice not every visualization process follows these exact steps in this same order. However, for a workshop setting, we present these as a way to walk through the topics. Each topic has a corresponding list of options and choices. Read through each topic and follow the instructions. Explore the options for each step as you progress.

The six steps are:

Step 1: Choose a human rights issue

Step 2: Discuss some kinds of data you might acquire

Step 3: Consider what question are you trying to answer with your data and visualization

Step 4: Choose a chart type for your visualization

Step 5: Consider some data and visualization hazards

Step 6: Consider some ways your charts can be improved

Step 5

What are some hazards?

Here we will look at some common hazards for data collection and processing, as well as visualization.

Some human rights issue can be difficult to quantify, and different types of bias can affect a data visualization at various steps of the process: data collection, processing, analysis, and even visualization.

Listed here are some hazards and tips on collecting data in a way that is accurate and free from bias, and also respects the privacy and dignity of subjects.

5 | VIZ HAZARDS

Color Association

Colors often have cultural associations, and this can be an advantage or a disadvantage depending on how colors are used. For instance, election maps in the U.S. use a visual convention of representing red for Republicans and blue for Democrats.

In many charts, the abstract use of color may not evoke cultural associations.

However, when using colors in combination with symbols, or to represent specific countries, populations, or places, one should be mindful about strong cultural color associations, particularly when targeting an international audience.

For example, green may represent luck or “go” in the U.S., while in west Asia it is associated with Islam. The color orange is linked with nationalist, conservative, or liberal political parties in different countries around the world.

For a fascinating glimpse at political color association see https://en.wikipedia.org/wiki/Political_colour.

5 | VIZ HAZARDS

Color Value

Color combinations should look good together, but should also be easy to differentiate. Using value (the lightness or darkness of a color) in addition to hue can help make your colors more distinct from each other.

It is easier to perceive approximate differences in length than it is to perceive differences in color—say, when something is twice as long versus twice as light. As such, when using colors for categories it is best to group them into visually distinct bins.

Using distinct color values can also help make the chart legible to persons with color blindness. Some software packages include color palettes designed with color blindness in mind and these are helpful to look for.

5 | DATA HAZARDS

Data Protection

Because of the sensitive nature of working with vulnerable populations and sensitive issues, organizations should take care to minimize risk. Data minimization is the practice of limiting the collection of personal data to only that which is directly relevant to the analysis being undertaken. Data anonymization is the process of removing personally identifiable information from data sets. Note that it may be possible to re-identify data using publicly available data, contextual information, or other methods of data linkage, so you should follow an anonymization framework wherever possible, such as this one created by the UK Anonymisation Network.

Critical data is also at risk from a range of protection issues: malware, staff turnover, theft, confiscation, and even hardware failure. Care should be taken to encrypt data, to limit access, and to maintain encrypted back-ups in more than one physical location.

For more information about practical steps to take to protect data, see the resources curated by the Responsible Data Forum.

5 | DATA HAZARDS

Data Selection

How do you know you're selecting the right data samples to investigate? Is your metric the right one? How does your data relate to the relevant human rights law?

It may be tempting to use data that is cleaner or easier to access, but this may also be misleading, or more easily taken out of context, or may miss the bigger picture.

For instance: focusing on a list of killed human rights defenders may overlook or even downplay tortured individuals, or individuals that have been detained, displaced, or silenced by repressive situations.

To check for data selection issues, ask yourself if the data is *valid* (does it measure the right you are assessing?), *reliable* (was the data collection process dependable?), and *unbiased* (was it collected by a group that respects scientific standards or by a group with a conflict of interest?).

5 | VIZ HAZARDS

Distortion

While manipulation of the facts or deception of the reader is usually unintentional in the human rights realm, accidentally misleading visualizations may be deceptive.

Common distortion techniques include:

- truncated y-axis (such alteration of the axis range leads to exaggeration or understatement of the quantities presented – see Non-Zero Baseline below);
- representing quantity using area (when there is not a one-to-one relationship between the data and the graphical area, this can be distorting);
- stretched aspect ratio (when the scale of one axis in a line chart is widened, the angle of the line changes, suggesting a different rate of increase or decrease than the data suggests); and
- inverted axis (switching axes often leads viewers to interpret the data as supporting a message that is opposite to what is in the data).

5 | VIZ HAZARDS

Double Y-Axis

Displaying two different types of data on the same chart make for easy comparison and can save space. However, while it may be tempting to compare two different trends on the same chart, using two different scales on the *y*-axis can be misleading, particularly if the scales are dramatically different.

Attention is also drawn to the intersection of lines, which may be arbitrary depending on the choice of scales.

If you are convinced that a double *y*-axis is the best way to display the data, use very different colors to illustrate the two different data sets. Color code your axes accordingly, to reinforce the data association. If your audience reads from right to left, put a label on your primary dataset on the left *y*-axis, and your secondary dataset on the right. Avoiding the same chart type for the two data sets can also help clarify the difference. For instance, instead of using two lines, perhaps combine a line for one dataset and a bar chart for the other.

5 | DATA HAZARDS

Incomplete Data

Data used for human rights analysis is almost always incomplete. Minorities and other vulnerable populations may be excluded from official data gathering exercises because of deliberate policy, lack of access or resources, or other reasons. Even when such groups are included, they may make up such a small portion of the sample that it is functionally impossible to use disaggregated data. When human rights groups collect their own data, the scale, timeline, or precariousness of events may confound complete data collection.

When using received data, it is important to understand the sampling and data collection methods and their limitations. Especially crucial here is knowing whether a specific data set is the product of a randomized collection strategy or is a convenience sample.

By using statistical methods like multiple systems estimation, it may still be possible to draw rigorous conclusions from incomplete data. Using such methods requires a high degree of analytical sophistication.

5 | VIZ HAZARDS

Non-Zero Baseline

Starting a y-axis at zero has its advantages and disadvantages. Charts where the y-axis does not start at zero can exaggerate the differences between data point values and the steepness of the slope of connecting lines that might otherwise appear rather slight. In some situations this may be deceptive.

However, in other situations slight variation may have enormous impact (for examples, a single degree difference in global warming). If these small changes are significant to your story, it may make sense to not start the y-axis at zero.

Context should be taken into account, and choosing whether or not to start an axis at zero should be considered with caution.

5 | DATA HAZARDS

Selection Bias

Selection bias is when data for analysis are chosen in a way that it is neither a complete enumeration of all the possible data (like a census) nor a random, scientific sample. Selection bias often affects data related to human rights violations. This is because human rights-related data is frequently collected in situations of limited access, danger, high stigma, or risk of punishment. In addition, survivors and victims may mistrust the organization collecting data, or fear that their information may not be secure against misuse.

Data samples affected by selection bias are referred to as convenience samples. These rely on information “selected by those who provide it or observe it,” rather than on a probability-based selection procedure. For example, testimonies from individuals who choose to tell their stories to NGOs, text messages coming from disaster-stricken areas, and cases captured through crowd-sourcing platforms are all limited to people who elect to come forward with their experiences, or who have access to cellphones and are able to send text messages. In general, convenience samples often favor people who have one or both of consistent interaction with, or access to, online data. By relying on convenience samples for data-based policy decisions and responses to human rights violations, human rights data practitioners run the risk of systematically excluding those people who do not enjoy such access. This is problematic, because people with low access to the internet or online data are already likely to have been marginalized in other ways, and may be more vulnerable to human rights abuses.

A lack of probability-based selection procedure also makes statistical inference an unreliable tool for drawing conclusions based on the sample. A random sample – in which every member of the population has the same chance of being selected – becomes more and more representative of the population as a whole as the number of individuals in the sample grows, making statistical analysis a powerful and accurate tool for drawing conclusions. But in a convenience sample with selection bias, this logic no longer holds. Because not every member of the population has the same chance of being selected for the sample, the sample does not become more representative of the general population when it grows larger. Moreover, it is impossible to quantify the probability that some members are less likely than others to be selected, meaning that the type and degree of bias remains unknown.

For more information, see the Human Rights Data Analysis Group's Core Concepts page, particularly their section on Convenience Samples.

5 | VIZ HAZARDS

Spurious Correlation

As the saying goes, "correlation is not causation." Correlation is when two (or more) variables show related trajectories. While this could indicate a causal relationship, there may be other unknown or hidden factors at play.

The careful design of experiments can help rule out some factors, however human rights groups should be extremely cautious and rigorous about making rights claims using correlations.

Similarly, when drawing trend lines, the way the line fits the data points can often depend on the variables included in the equation. Depending on the program used for the visualization, it may be possible to modify the trend line's equation and make sure to isolate only the relevant variables and control for alternative influencing factors.

5 | VIZ HAZARDS

Uncertainty

One of the great powers of working with statistics is the ability to work with and even quantify uncertainty.

This is not always a familiar concept for lay viewers, so visualizing this presents a challenge. Marks on paper imply specificity, so care must be taken when representing a range.

Uncertainty may be expressed as an interval or range of values that reasonably represent the result. Sometimes it is expressed as a range of possible outcomes or paths.

Some ways to represent uncertainty include error bars, fading gradients, violin plots or even dotted lines.

Clear annotation is also key to clarifying what is being represented by the visualization.

5 | DATA HAZARDS

Violative Collection

Projects collecting data from vulnerable populations should consider risks posed to those populations.

While universities require a human subjects review, foundations and NGOs do not have the same infrastructure in place.

Even anonymized data can be re-identified when linked with other data sets. This is particularly dangerous where populations are not only marginalized but also criminalized.

For more on data practices for human rights practitioners, see DatNav, a guide to using digital data for human rights research, as well as this 2016 report on data ethics, and this handbook from the Responsible Data Forum.

Visual Literacy

Visual literacy, the ability to interpret charts and images, varies widely among and within different populations. Visual literacy is a skill that is learned.

While some audiences may be familiar and comfortable with some chart types, others may be less so. In some cases, your audience may be comfortable navigating familiar chart types or novel visualizations. For other audiences, it may be best to walk your audience through the stages or features of a given visualization. While for some audiences, a visual or pictorial representation may be preferable to geometric abstraction.

Finally, in some cases, it may be best not to use a chart at all and simply present a key data point directly.

The best way to determine the visual literacy of your audience is to test your visuals with representatives of your audience before finalizing your publication.

About this Booklet

This activity is the result of a research collaboration between the Center for Human Rights and Global Justice and Tandon School of Engineering at New York University, and was funded with a grant from the John D. and Catherine T. MacArthur Foundation.

The booklet was authored by John Emerson and Margaret Satterthwaite with help from contributors Brianne Cuffe and Sidra Mahfooz.

It was inspired by Shiqing He and Eytan Adar's Vizit cards, The Data Visualisation Catalogue by Severino Ribecca, and Tamara Munzner's Nested Model for Visualization Design and Validation.

For more information about data visualization and human rights along with links to resources, research and tools, visit our project page at http://visualizingrights.org.

Cover illustrations by John R. Holmes for the Open Society Justice Initiative report *Presumption of Guilt: The Global Overuse of Pretrial Detention* (2014.)

Please send suggestions, comments, or feedback to john@backspace.com.