Data Visualization

Inspriration

- https://color.adobe.com/nl/create/color-contrast-analyzer
- https://www.datavis.ca/gallery/delights.php
- https://www.tableau.com/blog/stephen-few-data-visualization
- http://dataphys.org/list/
- https://www.printmag.com/article/data-humanism-future-of-data-visualization/
- https://courses.washington.edu/info424/2007/readings/Show_Me_the_Numbers_v2.pdf
- https://dariorodighiero.com/From-Wisdom-to-Data
- https://datavizcatalogue.com/
- https://upgrader.gapminder.org/
- https://hdlab.stanford.edu/palladio/ https://selection.datavisualization.ch/ https://guides.nyu.edu/digital-humanities/tools-and-software/visualization https://datavizcatalogue.com/ https://en.wikipedia.org/wiki/File:Minard.png https://www.datawrapper.de/ https://app.flourish.studio/templates Physical data visualizations: http://dataphys.org/list/ https://www.printmag.com/article/data-humanism-future-of-data-visualization/ https://dariorodighiero.com/From-Wisdom-to-Data

Basics of visualization

Information visualizations are a part of everyday communications and scholarship. These graphics have **powerful rhetorical** force. The visualizations are often more easily consumed than the complex research data on which they depend.

Today, we will focus on understanding the **process** by which visualizations are made helps bring into focus

- what they show
- and what they conceal

All information visualizations are **metrics expressed as graphics**.

The implications of this simple statement are far ranging. Data can be very difficult to interpret in tabular form. Very few individuals are skilled at reading spread sheets, let alone relational databases, to make sense of information. A query might produce thousands of data points. Information visualizations are used to make this quantitative data legible.

They are particularly useful for seeing patterns in large amounts of information, making these apparent in a condensed form.

Anything that can be quantified (given a numerical value) **can be turned into a graph**, chart, diagram, or other visualization.

Points, lines, and areas can be plotted using analog tools: paper and colored pencils. And many of the formats used in digitally produced visualizations are centuries old. The process of making graphs by hand is slow and deliberate. Each point has to be marked, each line created by connecting dots or using mathematical formulae, and each area calculated. At each step of hand-drawing a graph or chart, we reflect on how it is made.

But the ease of production afforded by computational means makes it possible to create polished and sophisticated graphics without critical reflection.

We can easily overlook the fact that all parts of the process—from creating quantified information to producing visualizations—are acts of interpretation.

In addition, the ability to read a visualization requires understanding the semantics of graphic formats. **Visual forms create meaning, they don't just display it.**

A bar chart makes a different statement than a pie chart, for instance, and such insights are crucial to the critical engagement with information visualization (Lengler and Eppler 2007).

To begin, consider the two components of a visualization separately—the metrics and the graphics. Here are two versions of the same information, a table and a bar chart:

[table]

The **table is not very complicated**, it puts dates in one column and number of pages output by an author into a second one.

All of the information in it makes good sense but trying to read columns of numbers to see a pattern in them is difficult.

[barchart]

The **bar chart** makes clear that a steady output of pages occurred in 1972, matched by one spike in 1971, and followed by low output in 1973. The comparison of values is easily done in the visual format, and if we imagine extending the table to include hundreds or thousands of data points, this fact would be even more dramatically clear.

What is the relationship of the data to the visualization?

In this situation, a line of dates is charted on the x-axis and a set of values is indicated by the y-axis. The conventions of charts make this easy to read and even intuitive in layout.

But is there an inherent visual form in the data?

One interesting exercise is to put the same data into other graphical formats to see what happens. Here are two examples of the same data but in a **line chart** and a **pie chart**.

[linechart]

We are immediately confronted with the question of what features of the graphical display are meaningful.

For instance, the **continuous line** on the left graphing the dates **suggests that the rate of change** in the data about pages is a significant factor. **But the "number of pages" data is actually a discrete value.**

While **the bar chart compares the values of each segment to each other, the line chart makes these part of a continuous process**, though this is not the case.

[piechart]

By contrast, **the pie chart suggests that each entry is part of a whole**—that the sum total of pages is significant, not the difference in their value.

Also, in the pie chart, he **values are hard to compare**, the **dates are lost** entirely, and the **concept of the "whole" of the author's output has no meaning**.

--> Neither of these charts makes the correlation of date and page output as clear as the initial bar chart. These are both "bad" graphics (and possibly bad data as well).

The point is that nothing in the data dictates the form of the visualization. These and a host of other charts can be generated from the same data.

Any data set can be put into a pie chart, a continuous graph, a scatter plot, a tree map, and so on. **The challenge is to understand how the information visualization creates an argument and then make use of the graphical format whose features serve your purpose.**

Any sense that data have an inherent visual form is an illusion.

```
--> Go to the Google sheet, look at some of the suggestions.
```

Data creation, as we noted in earlier, **depends on parameterization**. This means that anything that can be measured, counted, or given a metric or numerical value can be turned into data. The concept of parameterization is crucial to visualization because the ways in which we assign value to the data will have a direct impact on the ways it can be displayed. Visualizations are convincing by virtue of their graphic qualities and can easily distort the data. **While all visualizations are interpretations, some are more suited to the structure of a given data set than others.**

Visualization basics

In many cases, **the graphic image is an artifact of the way the decisions about the design were made, not about the data.** Understanding some basics of the relation between graphics and metrics is essential.

Here are some fundamental guidelines for thinking about which chart to use:

- The **distinction between discrete and continuous data** is one of the most significant decisions in choosing a design.

Example: in visualizing the height of students in a class, making a continuous graph that connects the dots makes no sense at all. There is no continuity between the height of one student and another. Individual height is a discrete value.

- If you are showing change over time or any other variable, then a continuous graph is the right choice. Example: Change in height for individual students over a five-year period.

- If a graph **shows quantities with area**, use it for percentages of a whole, like a pie chart, not comparative value. If you increase the area of a circle by length of the radius, or a square from the length of the side, you are introducing distortion into the relation of the elements. This is a common error. Example: The population in the town doubled from ten thousand to twenty thousand in five years. The data is visualized with two squares on a map, with the second having its sides twice the length of the first (10,000 to 20,000). But the area of the second square four times that of the first, not double.

From Tufte (1986, p. 55-56):

One satisfactory answer to these questions is to use a table to show the numbers. **Tables usually outperform graphics in reporting on small data sets of 20 numbers or less**.

The special power of graphics comes in the display of large data sets.

At any rate, given the perceptual difficulties, the best we can hope for is some uniformity in graphics (if not in the perceivers) and some assurance that perceivers have a fair chance of getting the numbers right. Two principles lead toward these goals and, in consequence, enhance graphical integrity:

1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.

2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.

```
-> Violations of the first principle constitute one form of graphic
misrepresentation, **measured by the size of the lie**.


The NYT reported in 1978  that the U.S. Congress and the Department of
Transportation had set a series of fuel economy standards to be met by
automobile manufacturers, beginning with 18 miles per gallon in 1978 and
moving in steps up to 27.5 by 1985, an increase of 53 percent:

If the Lie Factor is equal to one, then the graphic might be doing a
reasonable job of accurately representing the underlying numbers. Lie
Factors greater than 1.05 or less than .95 indicate substantial
distortion, far beyond minor inaccuracies in plotting.

- magnitude of the change = (27.5-18)/18 = 0.528 (times hundred to get
percentage)
- magnitude of the change = (5.3-0.6)/0.6 = 7.833 (times hundred to get
percentage).

Thus the numerical change of 53 percent is presented by some lines that
changed 783 percent, yielding…

Other things wrong with this graph:

- On most roads the future is in front of us, toward the horizon, and the
present is at our feet. This display reverses the convention so as to
exaggerate the severity of the mileage standards.
- Oddly enough, the dates on the left remain a constant size on the page
```

```
even as they move along with the road toward the horizon.
- The numbers on the right, as wel as the width of the road itself, are
shrinking because of two simultaneous effects: the change in the values
portrayed and the change due to perspective. Viewers have no chance of
separating the two!

Many published efforts using areas to show magnitudes make the elementary
mistake of varying both dimensions simultaneously in response to changes
in one-dimensional data.

Typical is the shrinking dollar fallacy. To depict the rate of inflation,
graphs show currency shrinking on two dimensions, even though the value of
money is one-dimensional.
```

- The way in which you label and order the elements in a chart will make some arguments more immediately evident. **If you want to compare quantities, be sure they are displayed in proximity**. Example: when comparing the population size of states should you put the states in alphabetical order or put the data in size order? Which is going to make the information more legible?

- The **use of labels is crucial** and their design can either aid or hinder legibility. Where are the labels? How much work are you adding to your reader's experience?

- Another consideration and challenge is the **choice of a scale**. When values are relatively close, the scale of the chart can be kept consistent. But imagine the chart contians an outlier... that eats up the scale... To show this value, the scale would need to extend to forty times its current height. The result would be that the difference between 20 pages and 50 pages would barely register. The legibility of the graph and patterns would be altered. To deal with such anomalies, charts are drawn with "broken" or modified scales, leaving a gap between lower and upper values. These gaps need to be noted and taken into account in some kind of legend, labeling, or documentation.

Checklist for visualizations

- Assess your data: Is it composed of discrete or continuous information?

- Choose the appropriate scale: too small a scale may make the important differences in value hard to spot and too large may exaggerate it.

- If outliers stretch the scale for a few data points, consider a gap in the scale and an explanation.

- Is the labeling efficient for use? What order should the information take to be meaningful and usable (alphabetical order of country names makes them easy to find but might separate values and make them hard to compare visually)?

- Use graphic variables carefully: shapes carry information readily, tonal values should be used for data that has a gradient, texture has little "meaning" in itself, and color can carry symbolic value or simply be used for differentiation.

- Proximity of labels to values is optimal for reducing cognitive load; make it easy for the viewer to correlate information. • Never use changes in area to show a simple arithmetic increase in value. • Review the graphic to see if it contains elements that are "incidental" artifacts of production rather than semantically meaningful ones. • While illustrations, images, or exaggerated forms may be considered “junk,” they can also help set a theme or tone when used effectively.