

Why we visualise data

Motivation

Data visualisations can be a very efficient means of identifying patterns in data and conveying a message. The scientific aim of any visualisation is to allow the reader to understand data and extract information:

- intuitively;
- efficiently; and
- accurately.

It is important, when creating a visualisation, to consider the background of the reader or intended audience (Krause 2013). Interpretation is in the eye of the beholder, and a visualisation will only succeed at conveying its message if designed with its audience in mind.

A successful data visualisation will:

- **Grab attention** In a sea of text, a visualisation will stand out. If a reader is short on time or uncertain about whether a document is of interest, an attention-grabbing visualisation may entice them to start reading.
- **Improve access to information** Textual descriptions can be lengthy and hard to read, while skilfully created visualisations permit the extraction of key information more efficiently, making information extraction a fun task.
- **Increase precision** Textual descriptions are frequently less precise than a visual depiction showing data points and corresponding axes, while a text with too many precise numbers can make it hard to follow a line of argument.
- **Bolster credibility** While a textual summary provides a story, a visualisation of the data can add credibility to otherwise unsubstantiated claims: readers can see the numbers for themselves and arrive at the (same?) conclusions.
- **Summarise content** Visual displays allow for summarising complex textual content, aiding the reader in memorising key points.

For these reasons, data visualisations are key elements in almost any kind of publication – scientific papers, media reports, conference presentations, social media posts, video summaries, etc.

Tables, too, are a way to visualise data or statistics, and can be similarly important components of a publication. A table may in some cases visualise data better than a graphic. For example, five numbers are probably better displayed in a table than in a complex pie chart that uses colours, angles, and possibly shading and more than two dimensions.

A brief history of data visualisation

Data visualisation has been, for a long time, both a topic of scientific research and an evolving art form with a variety of high-impact applications.

In 1859, [Florence Nightingale](#), the founder of modern nursing, published her findings on the sanitary status of the British army during the war with Russia. She showed raw data as well as summary statistics in tables and charts (Nightingale, 1859). One chart in particular continues to be celebrated today: a polar area chart on the “causes of mortality in the army in the East”.

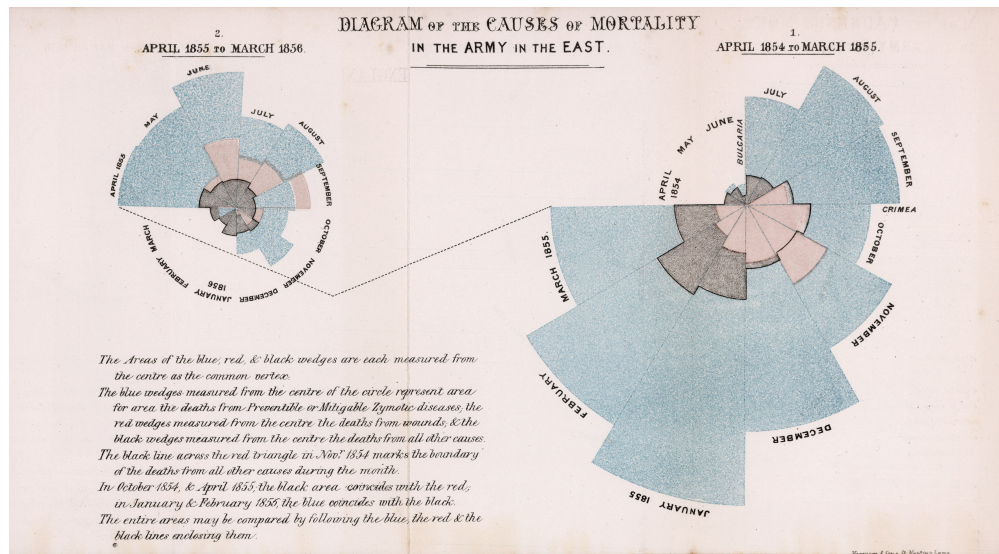


Figure 1: Florence Nightingale's polar area chart, “Diagram of the causes of mortality in the army in the East”. Source: Wikimedia Commons.

What made Nightingale’s graphs “particularly iconic was their powerful use of visual rhetoric to make an argument about data” (Hedley 2020). This quality is also evident in other visualisations produced by Nightingale’s contemporaries.

A simplistic but rather impactful visualisation of the water pumps in London associated with transmission of cholera paved the way for root cause identification. Medical doctor John Snow collected data on cholera deaths and created a visualisation where the number of deaths was represented by the height of a bar at the corresponding address in London. This visualisation showed that the deaths clustered around Broad Street, which helped identify the cause of the cholera transmission, the Broad Street water pump (Snow 1854; Wikipedia contributors 2023).

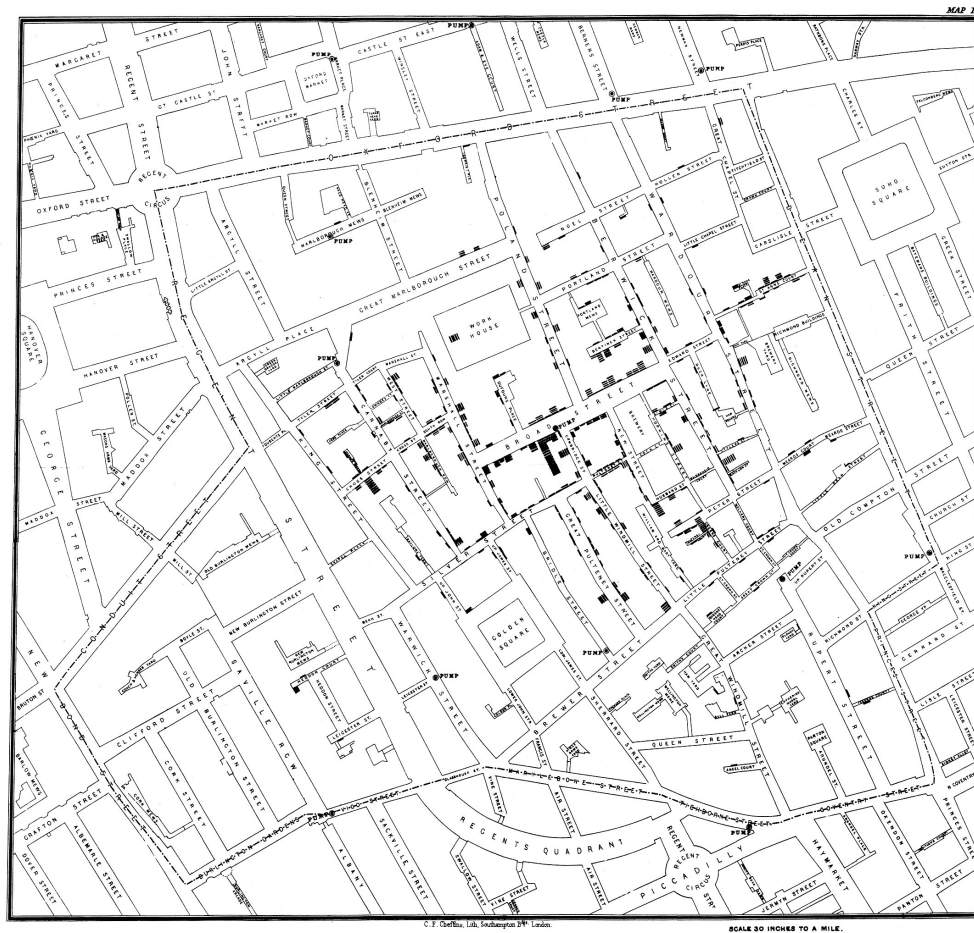


Figure 2: Map by John Snow showing clusters of cholera cases in the London epidemic of 1854.
Source: Wikimedia Commons.

An early complex visualisation was created by Minard in 1861, depicting [data from Napoleon's march on Moscow](#) in 1812/13 and his subsequent retreat.

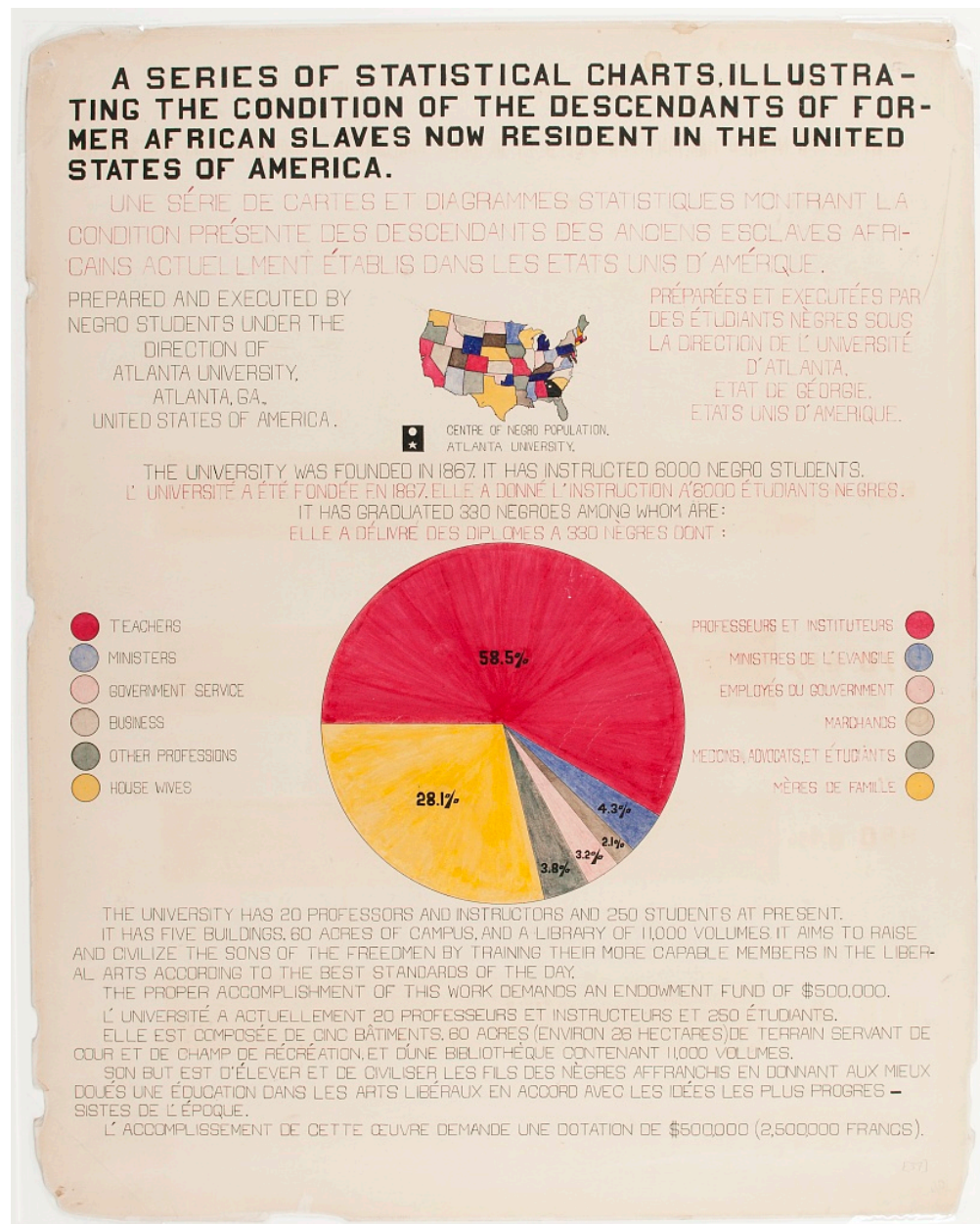


Figure 4: A series of statistical charts illustrating the condition of the descendants of former African slaves now in residence in the United States of America. Drawing, ca. 1900. [//hdl.loc.gov/loc.pnp/ppmsca.33899](http://hdl.loc.gov/loc.pnp/ppmsca.33899). Source: Library of Congress.

Groundbreaking work in modern visualisation was provided by Tukey with his book *Exploratory Data Analysis* (Tukey 1977), and Edward Tufte (Tufte 1990, 2004, 2006).

The historical development of visualisations is provided in various publications by Michael Friendly (Friendly and Denis 2005; Friendly 2018, 2022).

Data were visualised by hand until computers came along. The first monitors and printers worked in text mode only, with a resolution of 25 rows and 80 columns or similar, which did not permit much detail or precision. Graphics terminals and dot matrix printers followed, and resolutions kept increasing with the development of laser printers.

Statistical systems such as SAS enabled creation of data visualisations early on (“Documentation” n.d.).

Arguably the most consistent implementation of a graphics system was realised with ggplot (Wickham 2011, 2016) based on Wilkinson’s *The Grammar of Graphics* (Wilkinson 2005).

A key component of ggplot2 is the feature of creating conditional displays. These allow for subsetting the data by the values of one or more variables. The concept was introduced by Cleveland (Cleveland 1993, 1994), showcasing a widely known barley data set that was analysed in textbooks for decades until a visualisation strongly suggested an error in the data set – two years of crop yield of one variety at one of six farms were accidentally swapped. It took a visualisation to reveal what numerous numerical analyses had missed!

Cleveland named the display type “trellis”, being inspired by a trellis in his garden (Cleveland 1993). The trellis concept was first implemented in S (Becker and Chambers 1984) and S-PLUS (Becker and Cleveland 1996). When R (R Core Team 2021) was developed by Robert Gentleman and Ross Ihaka in Auckland NZ in the 1980s, a package named `lattice` was developed (Sarkar 2008) since “trellis” carried a trademark (by Lucent and later MathSoft). With the advent of ggplot, the term faceting (with functions `facet_grid` and `facet_wrap`) replaced the “lattice” with its “panels”.

Battle-Baptiste, W., and B. Rusert. 2018. *W.e.b. Du Bois’s Data Portraits: Visualizing Black America: The Color Line at the Turn of the Twentieth Century*. The W.E.B. Du Bois Center at the University of Massachusetts.

Becker, Richard A., and John M. Chambers. 1984. *S: An Interactive Environment for Data Analysis and Graphics*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Becker, Richard A., and William S. Cleveland. 1996. *S-PLUS Trellis Graphics User’s Manual*. Seattle, WA: MathSoft.

Cleveland, William S. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.

———. 1994. *The Elements of Graphing Data*. Summit, NJ: Hobart Press.

Corbett, J. 2001. “Charles Joseph Minard, Mapping Napoleon’s March, 1861.” CSISS Class 2001. 2001. <https://escholarship.org/uc/item/4qj8h064>.

“Documentation.” n.d. Statistical Analysis System. Accessed June 19, 2023. <https://support.sas.com/en/documentation.html>.

Du Bois, W. E. B. 1900. *The Exhibit of American Negroes*. Paris.

Friendly, M. 2018. “A Very Brief History of Visualization: Visions, Stories and Pictures.” Chicago, IL. 2018. <http://datavis.ca/papers/CHF-2x2.pdf>.

- . 2022. “Remembrances of Things EDA.” 2022. https://www.researchgate.net/publication/361191335_Remembrances_of_Things_EDA.
- Friendly, M., and D. Denis. 2005. “The Early Origins and Development of the Scatterplot.” *J. Hist. Behav. Sci.* 41 (2): 103–30. <https://doi.org/10.1002/jhbs.20078>.
- Hedley, Alison. 2020. “Florence Nightingale and Victorian Data Visualisation.” *Significance* 17 (2): 26–30. <https://doi.org/10.1111/1740-9713.01376>.
- Krause, Andreas. 2013. “Concepts and Principles of Clinical Data Graphics.” In *A Picture Is Worth a Thousand Tables: Graphics in Life Sciences*, 3–21. Springer. https://doi.org/10.1007/978-1-4614-5329-1_1.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Robinson, A. H. 1967. “The Thematic Maps of Charles Joseph Minard.” *Imago Mundi* 21: 95–108.
- Sarkar, Deepan. 2008. *Lattice: Multivariate Data Visualization with r*. New York, NY: Springer.
- Snow, John. 1854. “Mode of Communication of Cholera.” Piccadilly (London), UK: John Churchill. 1854. <https://archive.org/details/b28985266/page/52/mode/2up?view=theater>.
- Tufte, Edward R. 1990. *Envisioning Information*. Graphics Press.
- . 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press.
- . 2004. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.
- . 2006. *Beautiful Evidence*. Cheshire, CT: Graphics Press.
- Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
- Wickham, Hadley. 2011. “Ggplot2.” *Wiley Interdisciplinary Reviews: Computational Statistics* 3: 180–85. <https://onlinelibrary.wiley.com/doi/10.1002/wics.147>.
- . 2016. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. Vol. 2. Use r! Springer International Publishing. <https://doi.org/10.1007/978-3-319-24277-4>.
- Wikipedia contributors. 2023. “1854 Broad Street Cholera Outbreak.” 2023. https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak.
- Wilkinson, Leland. 2005. *The Grammar of Graphics*. Statistics and Computing. New York: Springer-Verlag. <https://doi.org/10.1007/0-387-28695-0>.