On the Design of Scientific Visualisation

Nicolas P. Rougier

September 2024

Visual processing in Primates (Felleman &Van Essen, 1991)

About 27% of the cerebral cortex is devoted to visual processing in humans, vs 52% in macaques (Van Essen, 2004). We are still mostly visual beings. **This draws consequences.**

Anscombe's Quartet (Anscombe, 1973)

What is common to these data sets?

Mean of x	9
Sample variance of x	11
Mean of y	7.5
Sample variance of y	4.12
Linear regression	y = 3.0 + 0.50*x
R squared	0.666
p value	0.0021

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe's Quartet (Anscombe, 1973)

What is common to these data sets?

Mean of x	9
Sample variance of x	11
Mean of y	7.5
Sample variance of y	4.12
Linear regression	y = 3.0 + 0.50*x
R squared	0.666
p value	0.0021

A computer should make both calculations and graphs

Francis Anscombe (1918-2001)

Same stats, Different Graphs (Matejka & Fitzmaurice, 2017)

X Mean: 54.02462 01
Y Mean: 48.09596 99
X SD : 14.52074 09
Y SD : 24.79430 24
Corr. : +0.32809 45

Same stats, Different Graphs (Matejka & Fitzmaurice, 2017)

Cholera outbreak, London (Snow, 1854)

John Snow (1813-1858) is considered one of the fathers of modern epidemiology, in part because of his work in tracing the source of a cholera outbreak in Soho, London, in 1854.

Snow used a dot map to illustrate the cluster of cholera cases around the pump.

Cholera outbreak, London (Snow, 1854)

John Snow (1813-1858) is considered one of the fathers of modern epidemiology, in part because of his work in tracing the source of a cholera outbreak in Soho, London, in 1854.

Snow used a dot map to illustrate the cluster of cholera cases around the pump.

Analyzing 1.1 Billion NYC Taxi and Uber Trips (Schneider, 2015)

*"The New York City Taxi & Limousine Commission has released a staggeringly detailed historical dataset covering over 1.1 billion individual taxi trips in the city from January 2009 through June 2015. Taken as a whole, the detailed trip-level data is more than just a vast list of taxi pickup and drop off coordinates: **it's a story of New York**."*

Todd W. Schneider
toddwschneider.com

We only see what we're watching (Merleaux-Ponty, 1945)

Über Gestaltqualitäten (von Ehrenfels, 1890)

PROXIMITY

Objects that are close together wil be grouped together visually.

SIMILARITY

Two items that share attributes will be visually grouped together.

CLOSURE

The Brain is good at filling in gaps to create a whole.

CONTINUATION

A line will always appear to continue travelling in the same way.

FIGURE & GROUND

Sometimes, the blank space is just as important as the filled space.

Everyone knows what attention is (James, 1890)

And you will read this last

You will read this first

And you will read this

Then this one

What is Visualisation?

Visualisation is a method of computing. *It transforms the symbolic into the geometric, enabling researchers to observe their simulations and computations.* ***Visualisation offers a method for seeing the unseen.*** *It enriches the process of scientific discovery and fosters profound and unexpected insights.*

Visualisation in Scientific Computing
NSF report, 1987

Fundamentals of Data Visualization (Wilke, 2019)

Type of variable	Examples	Appropriate scale	Description
quantitative/numerical continuous	1.3, 5.7, 83, 1.5×10^{-2}	continuous	Arbitrary numerical values. These can be integers, rational numbers, or real numbers.
quantitative/numerical discrete	1, 2, 3, 4	discrete	Numbers in discrete units. These are most commonly but not necessarily integers. For example, the numbers 0.5, 1.0, 1.5 could also be treated as discrete if intermediate values cannot exist in the given dataset.
qualitative/categorical unordered	dog, cat, fish	discrete	Categories without order. These are discrete and unique categories that have no inherent order. These variables are also called *factors*.
qualitative/categorical ordered	good, fair, poor	discrete	Categories with order. These are discrete and unique categories with an order. For example, “fair” always lies between “good” and “poor”. These variables are also called *ordered factors*.
date or time	Jan. 5 2018, 8:03am	continuous or discrete	Specific days and/or times. Also generic dates, such as July 4 or Dec. 25 (without year).
text	The quick brown fox jumps over the lazy dog.	none, or discrete	Free-form text. Can be treated as categorical if needed.

Visualization Analysis and Design (Munzner, 2014)

[VAD Fig 5.1]

Ten Simple Rules for Better Figures (Rougier et al, 2014)

BROWSE PUBLISH ABOUT

SEARCH

advanced search

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

EDITORIAL

Ten Simple Rules for Better Figures

Nicolas P. Rougier, Michael Droettboom, Philip E. Bourne

Published: September 11, 2014 • https://doi.org/10.1371/journal.pcbi.1003833

3,040 Save | 91 Citation | 638,626 View | 2,512 Share

Article | Authors | Metrics | Comments | Media Coverage

Download PDF

Print | Share

Check for updates

Rule 1: Know Your Audience

Rule 2: Identify Your Message

Figures

Included in the Following Collection

1. Know Your Audience

It is important to **identify**, as early as possible in the design process, the **audience** and the message the visual is to convey.

- COMPLEXITY +

- Scientific Community
- Colleagues
- Students
- General Public

2. Identify your message

A figure is meant **to express an idea** or introduce some facts or a result that would be too long (or nearly impossible) to explain only with words, be it for an article or during a time-limited oral presentation.

Here, the visual pattern is fake but conveys the main message about the retinal transformation.

3. Adapt the Figure to the Support Medium

A figure can be **displayed on a variety of media**, such as a poster, a computer monitor, a projection screen (as in an oral presentation), or a simple sheet of paper (as in a printed article).

Each of these media represents different physical sizes for the figure, but more importantly, each of them also implies **different ways of viewing and interacting** with the figure.

✘ Presentation

✔ Article or Poster

✔ Presentation

✘ Article or Poster

4. Captions are not Optional

Optical Illusion

A and B patches are actually the same color even though we perceive them at being different.

Whether describing an experimental setup, introducing a new model, or presenting new results, you cannot explain everything within the figure itself, **a figure should be accompanied by a caption**.

The caption explains how to read the figure and provides additional precision for what cannot be graphically represented.

5. Do Not Trust the Defaults

Basic rendering using the default style.

Better rendring with a bit of work and experence.

Any plotting library or software comes with a set of default settings. These settings are to be used for virtually any type of plot, they are not fine-tuned for a specific type of plot.

In other words, they are good enough for any plot but **they are best for none**.

6. Use Color Effectively

Color is an important dimension in human vision and is consequently equally important in the design of a scientific figure. However, as explained by Edward Tufte, **color can be either your greatest ally or your worst enemy** if not used properly.

Do not use jet/rainbow colormap. Ever!

Qualitative: rapid variation of colors, used mainly for discrete/categorical data.

Diverging: variation between colors used to highlight deviation from a median value

Sequential: variation of a unique color, used for quantitative data varying low to high.

6. Use Color Effectively (bis)

Color is an important dimension in human vision and is consequently equally important in the design of a scientific figure. However, as explained by Edward Tufte, **color can be either your greatest ally or your worst enemy** if not used properly.

Do not use jet/rainbow colormap. Ever!

Colour

✔ Original colors

Jet

✘ Qualitative Colormap

Viridis

✔ Sequential colormap

7. Do not Mislead the Reader

Using the disc area shows a more proportional sizes

Relative size using disc area

Relative size using disc radius

Using the disc radius misleads the reader (difference looks bigger)

Using full range bars shows a more realistic comparison among them

Relative size using full range

Relative size using partial range

Using partial range bars misleads the reader (difference looks bigger)

What distinguishes a scientific figure from other graphical artwork is the **presence of data** that needs to be shown as objectively as possible. A scientific figure is, by definition, tied to the data (be it an experimental setup, a model, or some results).

If you loosen this tie, you may unintentionally project a different message than intended.

7. Do not Mislead the Reader (bis)

The Sun

The Herald Sun

8. Avoid "Chartjunk"

Chartjunk refers to all the unnecessary or confusing visual elements found in a figure that do not improve the message (in the best case) or add confusion (in the worst case).

In order to avoid chartjunk, **try to save ink** (or electrons in the computing era).

Typical chartjunk

Better layout

8. Avoid "Chartjunk" (bis)

Chartjunk refers to all the unnecessary or confusing visual elements found in a figure that do not improve the message (in the best case) or add confusion (in the worst case).

In order to avoid chartjunk, try to save ink (or electrons in the computing era).

Less is More

A remake of www.darkhorseanalytics.com

Made with matplotlib 2.0

9. Message Trumps Beauty

In Science, message and readability of the figure is the most important aspect while **beauty is only an option.**

This figure is an extreme case where the message is particularly clear even if the aesthetic of the figure is questionable.

10. Use the Right Tool

- **PDFCrop** to remove white borders
- **GraphViz** for creating easy graphs
- **ImageMagick** for image processing
- **Gimp** for bitmap image manipulation
- **Inkscape** for vector image manipulation
- **Tikz** for scripted vector art (Latex)
- And many, many, many others...

Figure by **Izaak Neutelings**
`tikz.net/fourier_series`

10. Use the Right Tool

Mandelbrot Set

The Mandelbrot set is the set of complex numbers c for which the function $f_c(z) = z^2 + c$ does not diverge when iterated from $z = 0$, i.e., for which the sequence $f_c(0)$, $f_c(f_c(0))$, etc., remains bounded in absolute value.

Here is a rendering using shading and normalized recounts. See **On Smooth Fractal Coloring Techniques (Härkönen, 2007)**

Domain Coloring

"Domain coloring is a technique for ***visualizing complex functions by assigning a color to each point of the complex plane****. By assigning points on the complex plane to different colors and brightness, domain coloring allows for a four dimensional complex function to be easily represented and understood."*

Wikipedia

Contours with shadows

Filled contours with dropshadows is a nice effect that allows you to use a sequential colormap while distinguishing positive and negative values.

If you look closely at the figure, you can see that the drop shadow is external for positive values and internal for negative values.

Escher-like Spiral Tilings (Kaplan, 2019)

Escher like projections can be obtained using the complex exponential function and some specific periods that can be computed quite easily.

To do the figure on the left, I mostly followed explanations given by **Craig S. Kaplan**, professor at the University of Waterloo on his blog post **Escher-like Spiral Tilings (2019)**.

Recursive Voronoi cells

This recursive Voronoi (n=5) is based on the precise placement of random points using the **fast Poisson disk sampling** by **Robert Bridson (2017)** that guarantee blue noise distribution (random & approximately evenly spaced).

Randomised Self-Organizing Maps (Rougier & Detorakis, 2021)

Self-organizing map is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction".

We developed a **randomized self-organizing map** based on the blue noise distribution of neurons. The figure shows the self-organisation of the map when fed with random RGB colors.

One million integers embedded into 2D space (Williamson, 2018)

"This is the first million integers, represented as binary vectors indicating their prime factors, and laid out using the UMAP dimensionality reduction algorithm by Leland McInnes.

Each integer is represented in a high-dimensional space, and gets squished down to 2D so that numbers with similar prime factorisations are closer together than those with dissimilar factorisations.

A very pretty structure emerges; this might be spurious in that it captures more about the layout algorithm than any true structure of numbers."

John H Williamson, 2018
johnhw.github.io

Simplex Noise (Perlin, 2001)

"Simplex noise is the result of an n-dimensional noise function comparable to Perlin noise ("classic" noise) but with fewer directional artifacts and, in higher dimensions, a lower computational overhead."

Wikipedia

The animation is a simple scatter plot using 4D Simplex noise to ensure temporal consistency. Original idea by **Necessary Disorder.**

Further Resources (Open Access)

See also (blogs)

- Necessary disorder
- Folds and other math stories
- Paul Bourke
- Processing

Books (not Open Access)

- The Visual Display of Quantitative Information (1983)
- Trees, maps, and theorems (2009)
- Visualization Analysis and Design (2014)

Some scientific posters

Some scientific conference posters

Some public outreach posters