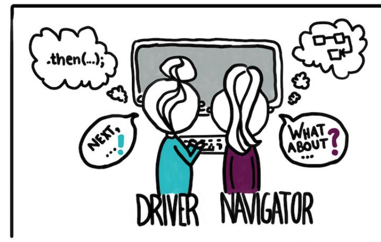
Best practices in data visualization

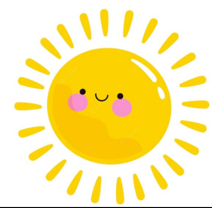
Guillermo Aguilar, Aina Frau-Pascual, Nicolas P. Rougier

Bilbao, ASPP 2022

Plan for this session

16:30	Principles of data visualization
	Hands-on Exercise 1: **mastering matplotlib**
	Review
~ 17:40	Break 5-10 min
~ 17:50	Visualizations and use of color
	Hands-on Exercise 2: **which visualization should I use?**
	Review of your solutions as PR
	(*) Hands-on Exercise 3: images + **Review** of your solutions as PR
19:00	The End!





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

Visualization in Scientific Computing, NSF report, 1987

Classical example: Anscombe’s quartet

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

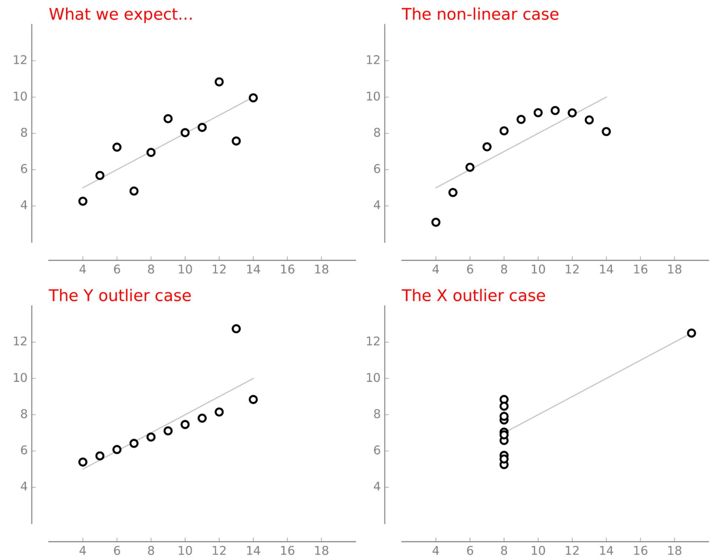
statistic	value
mean of x	9
sample variance of x	11
mean of y	7.50
sample variance of y	4.125
correlation coefficient	0.816
linear regression line	y = 3.00 + 0.500x
coefficient of determination	0.67

Anscombe (1973)

Classical example of why visualization is important

Anscombe, an statistician, created these four datasets, pairs of x and y points. They all share the ‘summary statistics’: mean and variance of each variable, and the correlation coefficient between x and y. So for a computer or an algorithm that only looks at these numbers, these four datasets would be identical.

Classical example: Anscombe's quartet

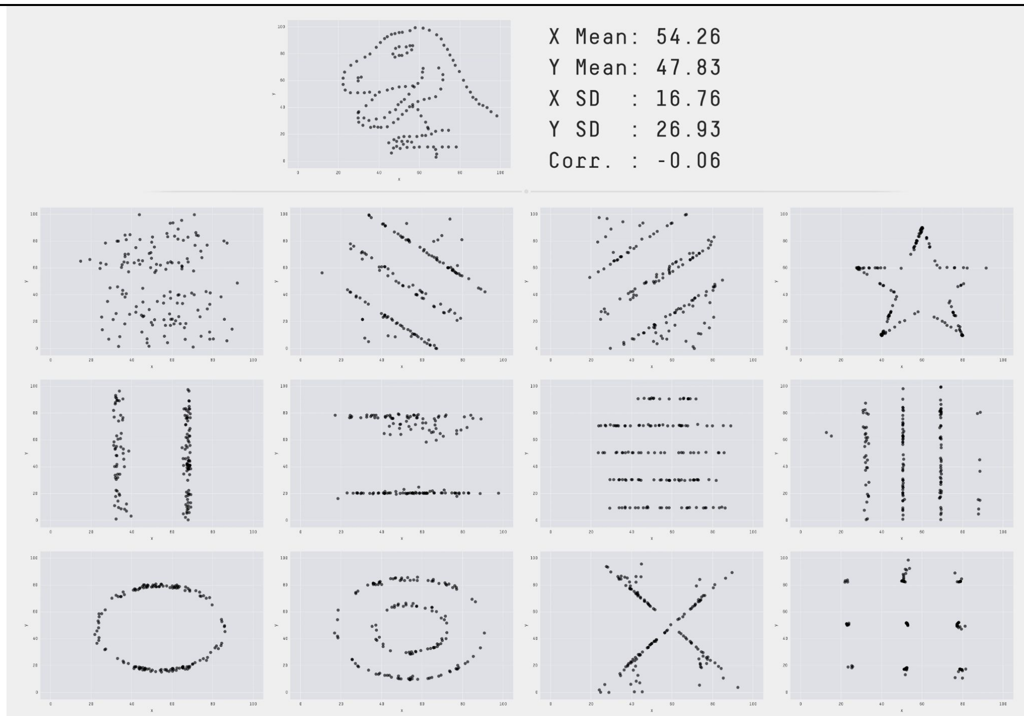


statistic	value
mean of x	9
sample variance of x	11
mean of y	7.50
sample variance of y	4.125
correlation coefficient	0.816
linear regression line	y = 3.00 + 0.500x
coefficient of determination	0.67

However when we visualize them, we see that actually these datasets are completely different!

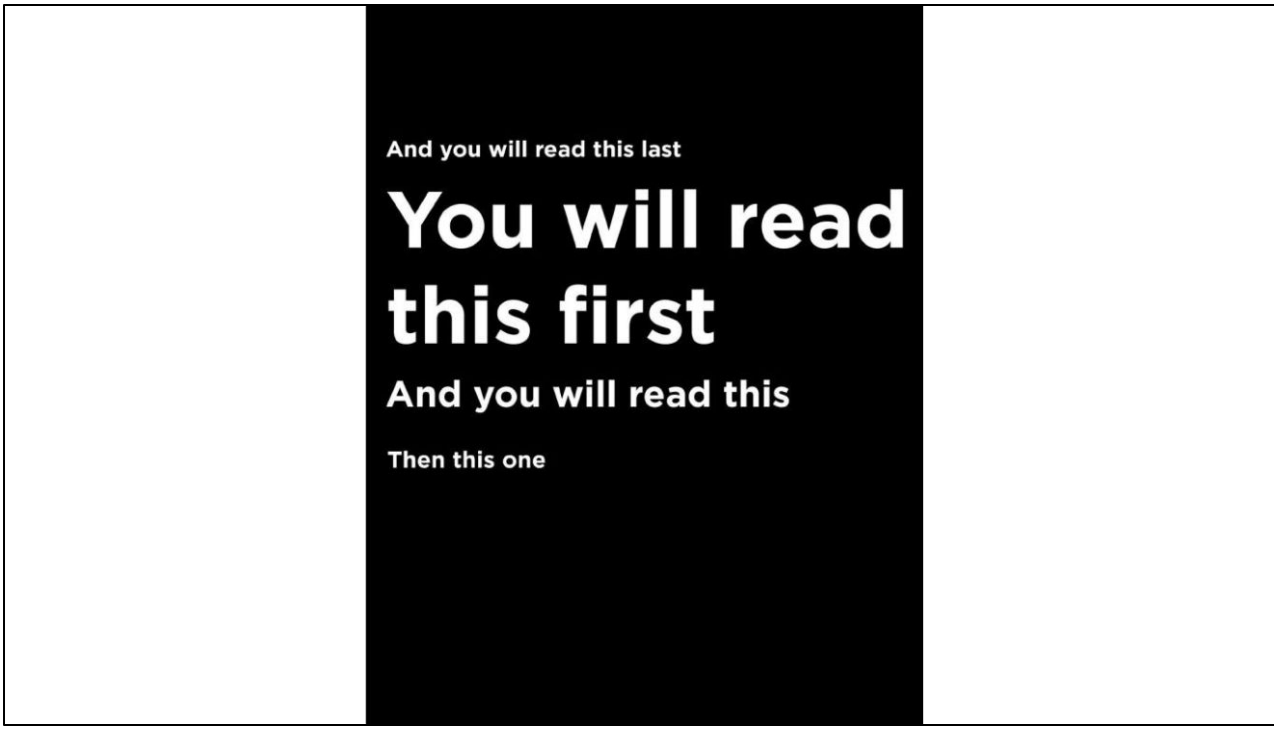
We need of vision and our cognitive interpretation to comprehend these differences.

Datasaurus



In the extreme, from a given set of points - here a cute dinosaur - we can create arbitrarily many different datasets with the same statistics - same mean, standard deviation etc, but looking completely different.
The R package Datasaurus implements this idea.

Datasaurus
https://cran.r-project.org/web/packages/datasauRus/vignettes/Datasaurus.html,
https://www.autodesk.com/research/publications/same-stats-different-graphs

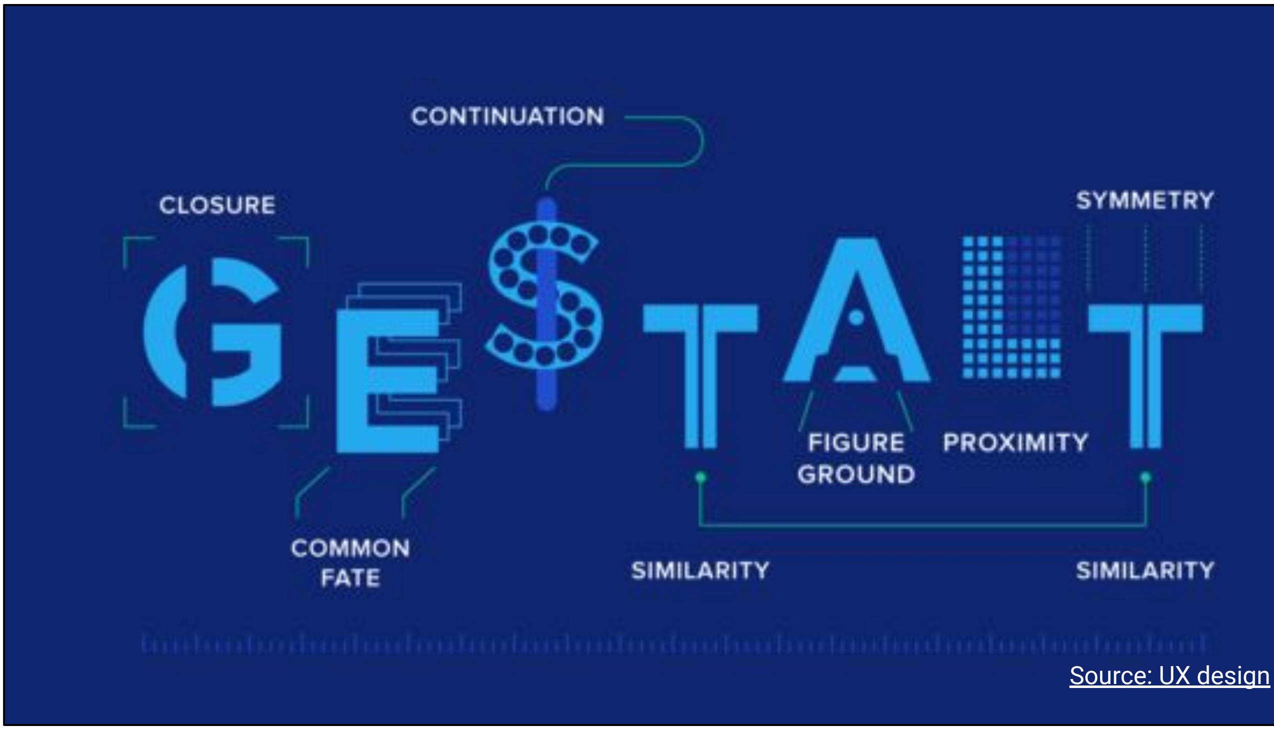


So although you might think that number are just numbers and you just need statistics, that’s not the whole story.

Because we are humans that see, and we communicate our findings (of science) to other humans, we do need to consider these ‘human factors’...
and one of the most important ones in scientific communication is how we visualize things.

We put his example as a demo of how we do not read always from top to bottom, instead we prioritize size.

Thus we need to know a little bit of how we see in order to better present our results and communicate.



Very generally speaking, psychologists have discovered that our vision follow certain rules: Gestalt rules.
For example we know that we tend to see things that have some kind of continuation as one thing -> closure. On the left we don't see 3 pieces separately, we do see a G, as we ‘close’ the figure.
Same with proximity, we tend to see things that are similar and close to each other as one thing → proximity.

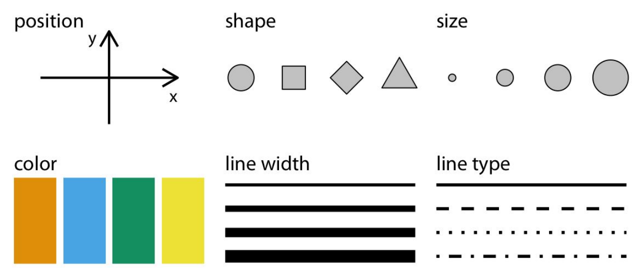
Not into the details here, you can read more about it. Graphic designers know a lot about it.

Main message here is that you can rely on your own vision to know if a figure looks good or not in terms of design, font sizes etc. That doesn’t mean you’re being less scientific or less objective.

Main challenge on data visualization: the **mapping from..**

Data → **Graphical elements**

Type of variable	Examples	Appropriate scale
quantitative/numerical continuous	1.3, 5.7, 83, 1.5x10^{-2}	continuous
quantitative/numerical discrete	1, 2, 3, 4	discrete
qualitative/categorical unordered	dog, cat, fish	discrete
qualitative/categorical ordered	good, fair, poor	discrete
date or time	Jan. 5 2018, 8:03am	continuous or discrete



Claus O. Wilke (2019)

Thus the main challenge in data visualization is the mapping between pure numbers and variables, which a table or dataset in a computer stores, to graphical elements, visual features which a human understands.

We have different kinds of data, you probably know this classification: quantitative or numerical variables, either continuous or discrete, for example like things that are measured with real numbers.
And then qualitative or categorical variables, for example things without an order, dog, cat, fish… or with an order: good, fair poor.

Some variables can be assigned to some graphical elements and not others. You probably know that a continuous variable can be assigned to x and y axis.
But for example a categorical variable, dog, cat, fish, cannot be assigned to say, the line width or the size of a dot.. That doesn't make sense. As it is categorical it should be mapped to something categorical, like line type, or a categorical color palette.

Rules for this mapping are numerous and we will not go into the details here, as also different fields have different internal ‘conventions’ or ‘standards’ on how they do this mapping. BUT in the 2nd exercise you will see and practice this mapping and the decisions that you need to take.

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Better Figures

Nicolas P. Rougier[1,2,3]*, Michael Droettboom[4], Philip E. Bourne[5]

1 INRIA Bordeaux Sud-Ouest, Talence, France, 2 LaBRI, UMR 5800 CNRS, Talence, France, 3 Institute of Neurodegenerative Diseases, UMR 5293 CNRS, Bordeaux, France, 4 Space Telescope Science Institute, Baltimore, Maryland, United States of America, 5 Office of the Director, The National Institutes of Health, Bethesda, Maryland, United States of America

So instead of doing a full lecture of how to do plotting, we go through some simple rules that you can follow.
If you follow these, your visualizations could be very much improved.

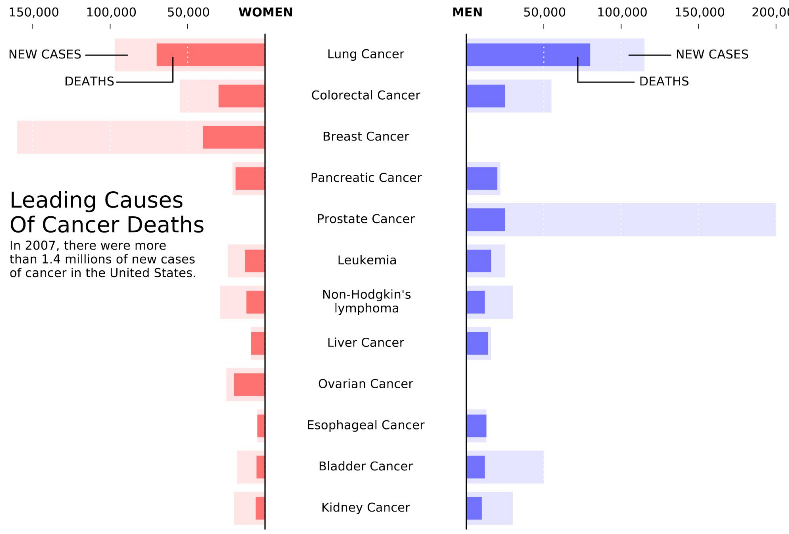
1) Know your audience

My colleagues

Scientific community

Student audience

General public



Audience: general public

Main message: cancer

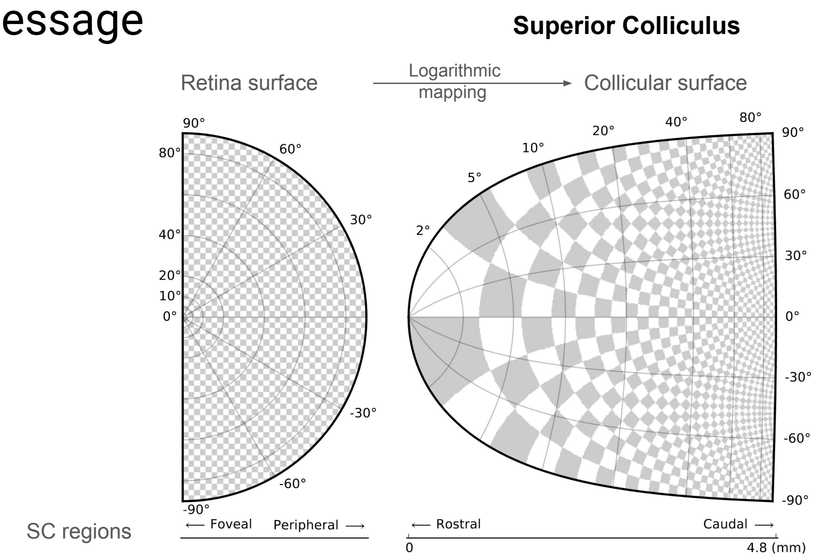
Separated in sex groups: Women / Men

The graphical design of the visual should be **informed by the audience and the message** the visual is to convey.

Focus of this figure:

- Audience: general public
- Qualitative (you can't tell the exact numbers)
- Main message in the center: cancer
- New cases always higher than deaths and deaths part of cases, so we visually have them together

2) Identify your message



Audience:
neuroscience
scientific
community

Main message: Artificial checkerboard pattern demonstrates the magnification of the foveal region in the superior colliculus (brainstem structure). This has to do with the induction of saccadic eye movement that the SC plays a role in.

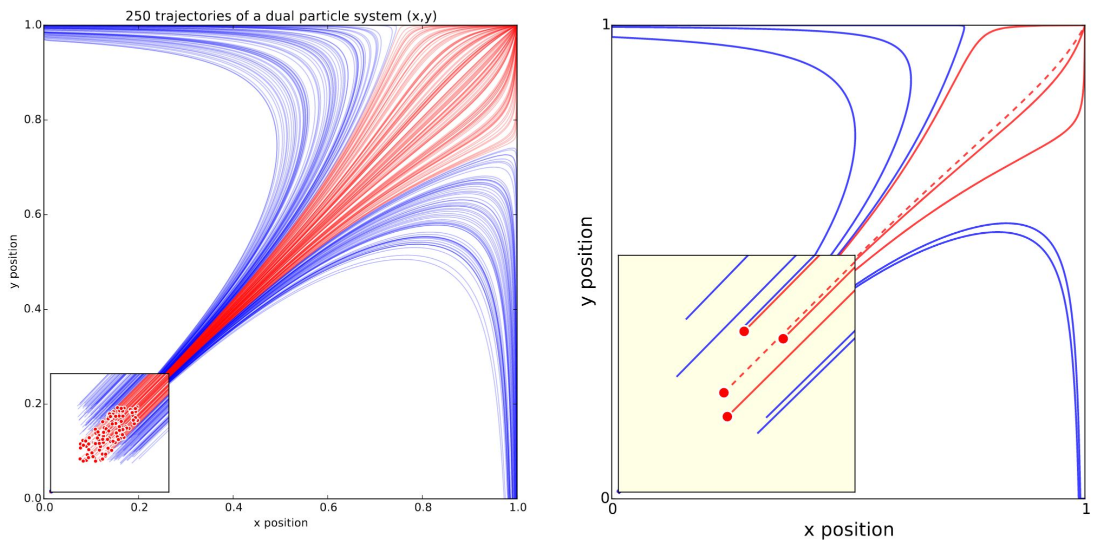
Depending on the audience the complexity of your message has to be different.
Once you identify which is the right level of complexity for your message, then you visualize your message.
For the audience the main message should be easily graspable at a first sight, with minimal reading of captions and text.
That means that as more general the audience is, the simplest the visualization has to be.

One message -> one figure

On this example:
Audience → neuroscientific community.
Main message: artificial checkerboard pattern log mapping from a brainstem substructure to another

3) Adapt the figure to support medium

Simulation of trajectories of a dual-particle system



Each media is different. Seeing a figure in a printed poster on a conference is different than having it digitally in a computer, and different that a figure projected in an oral presentation. All these media have different physical sizes, but more importantly, each of them also implies **different ways of viewing and interacting with the figure**.

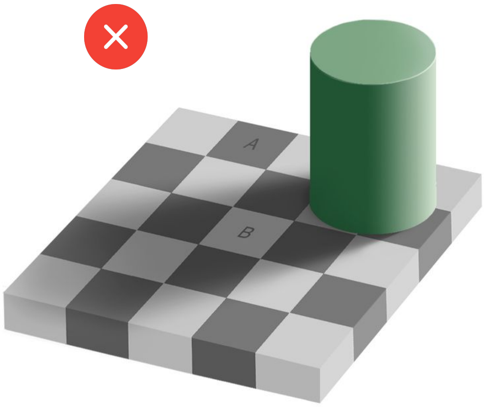
On the left:

- image for a journal article where the reader is free to look at every detail. Red lines: initial conditions Line transparency used to show density of lines

On the right:

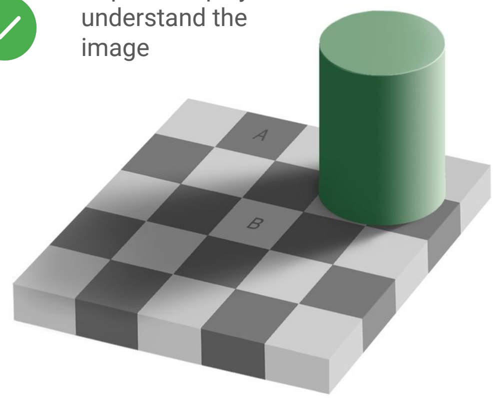
- Image for an oral presentation. Time-limited display. Many details have been removed. Some parts modified so that it is easier to reference.

4) Captions are not optional



Optical illusion

Caption helps you understand the image



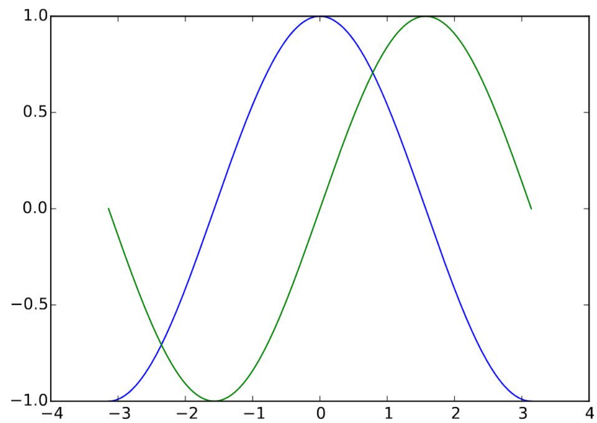
A and B patches are actually the same color even though we perceive them at being different color

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

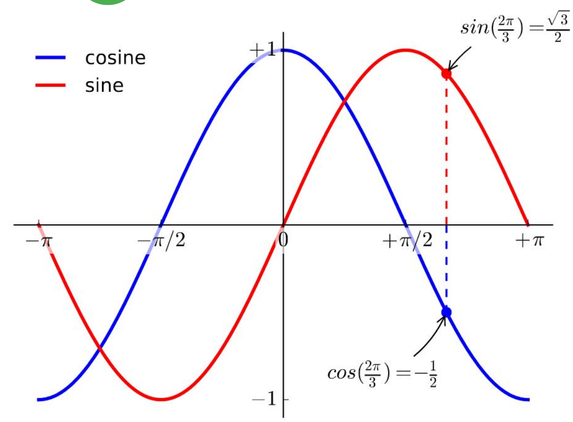
Be concrete and explicit in the caption. Describe all what is going on, if there are different panels, colors, line types, you should mention what they all mean.

5) Do not trust the defaults

Defaults for a matplotlib plot



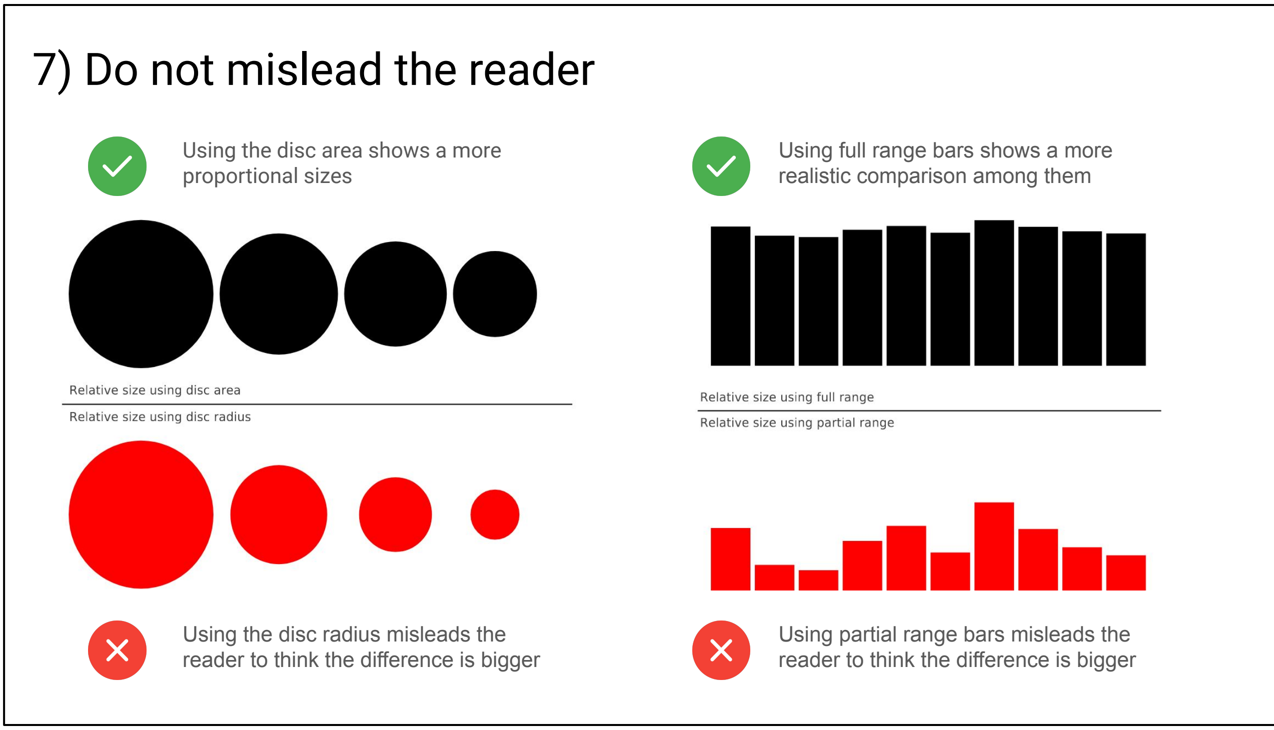
With a bit of work…



Defaults are good enough for any plot but best for none. Fine-tuning the plot will allow you to better express the message.

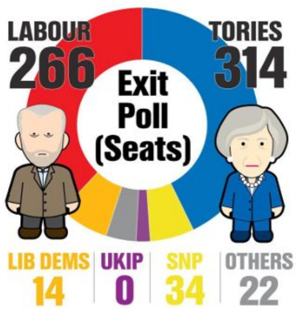
All plots require at least some manual tuning of the different settings to better express the message, be it for making a precise plot more salient to a broad audience, or to choose the best colormap for the nature of the data.

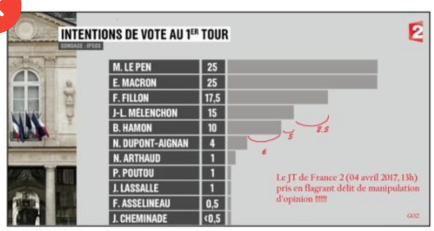
6) Use color effectively → more on this later

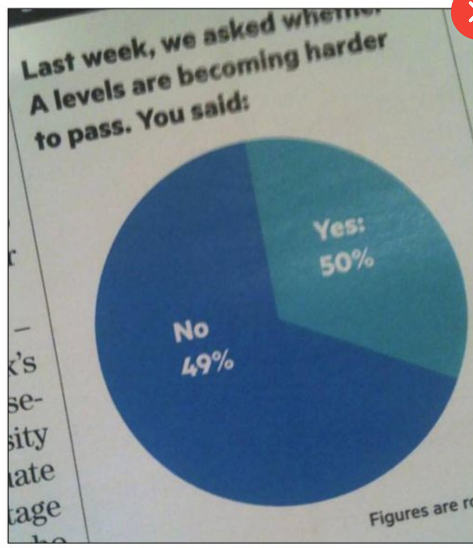


What distinguishes a scientific figure from other graphical artwork is the presence of data that needs to be shown as objectively as possible. A scientific figure is, by definition, tied to the data (be it an experimental setup, a model, or some results) and if you loosen this tie, you may unintentionally project a different message than intended. However, representing results objectively is not always straightforward. As a rule of thumb, make sure to always use the simplest type of plots that can convey your message and make sure to use labels, ticks, title, and the full range of values when relevant. Lastly, do not hesitate to ask colleagues about their interpretation of your figures.

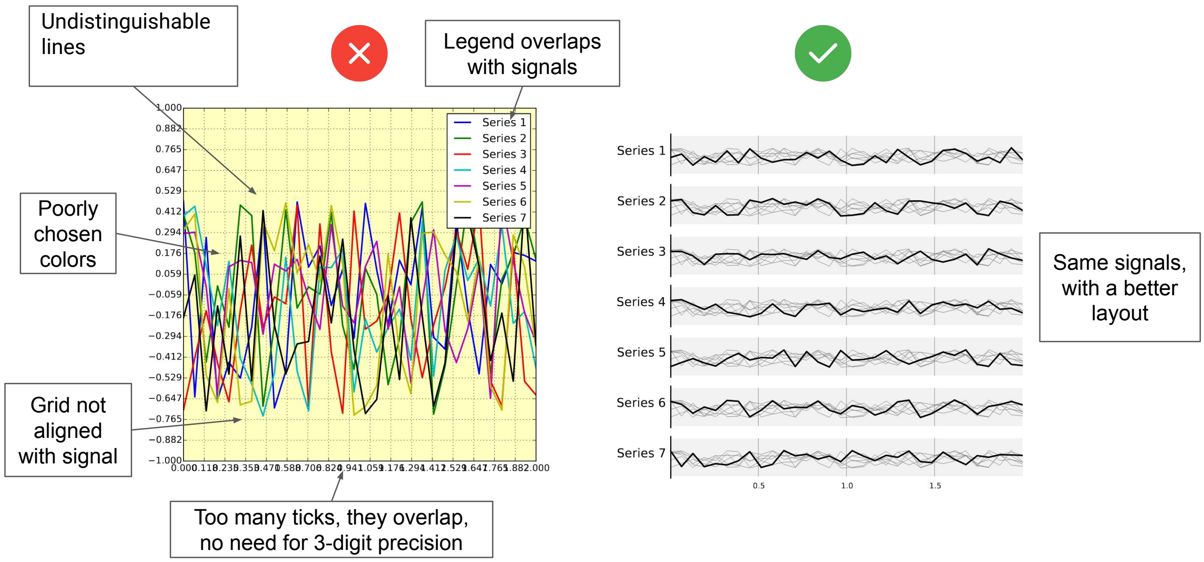
7) Do not mislead the reader. Really.





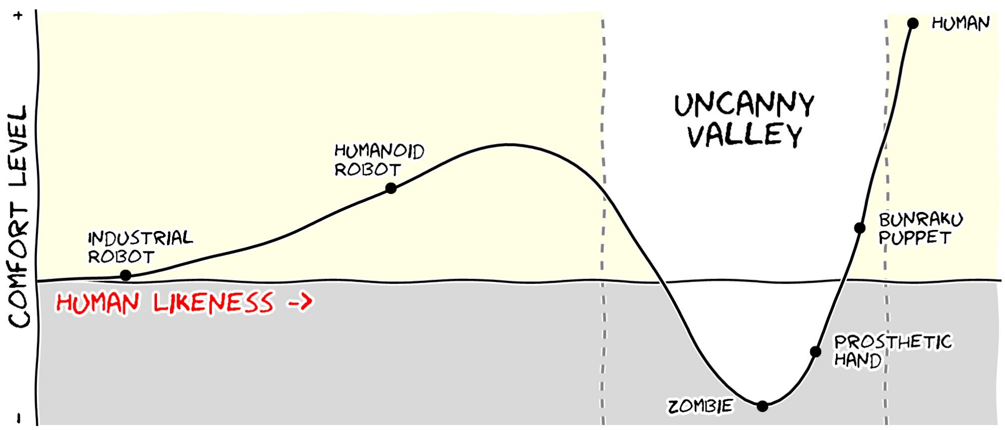


8) Avoid chartjunk



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). For example, chartjunk may include the use of too many colors, too many labels, gratuitously colored backgrounds, useless grid lines, etc.

9) Message trumps beauty



To deliver an idea or sometimes you don't have to be super precise and accurate. You can give a rough idea, with a sketch. This is OK!

(The uncanny valley is a well-known hypothesis in the field of robotics that correlates our comfort level with the human-likeness of a robot.)

10) Get the right tool

PDFCrop to remove white borders

GraphViz for creating easy graphs

ImageMagick for scripted image processing

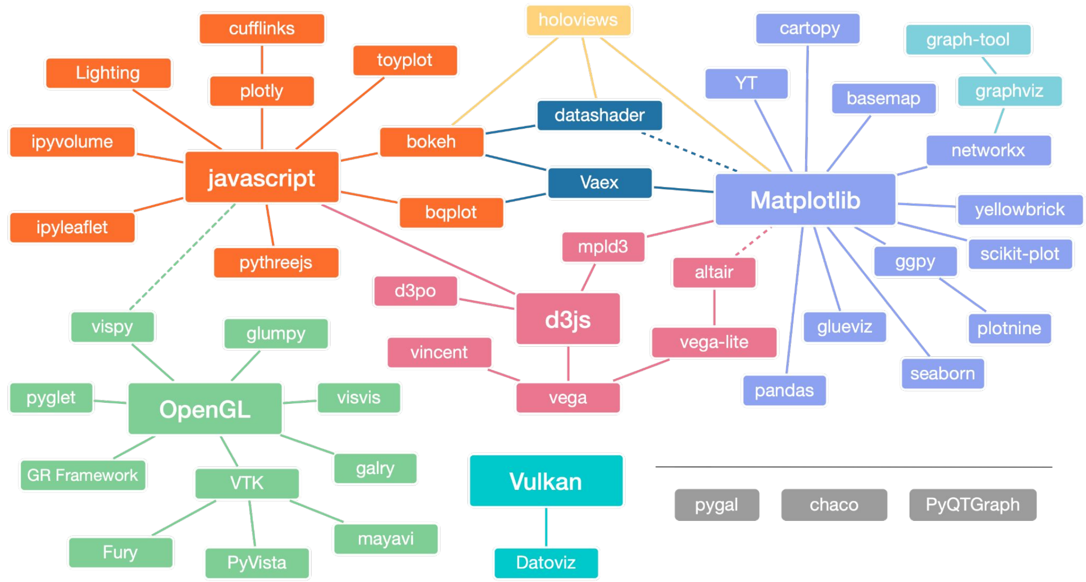
Gimp for bitmap image manipulation

Inkscape for vector image manipulation

Tikz for scripted vector art

And many, many, many others…

Overview of visualization libraries



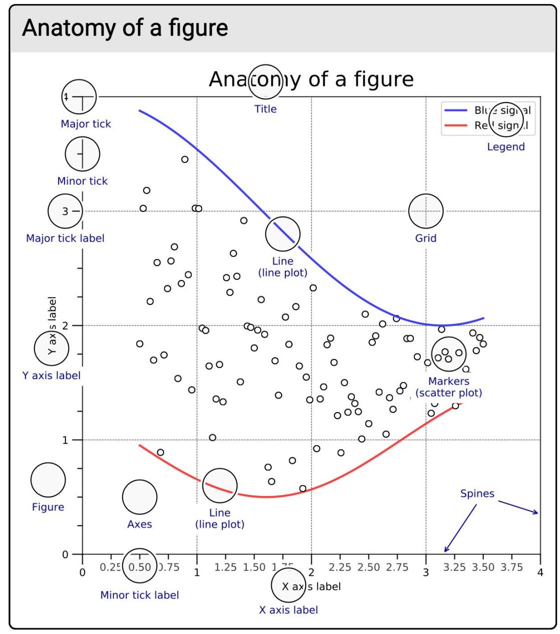
A jungle of visualization libraries. Here we will focus on the most used one: **matplotlib.**
After today's session feel free to expand your knowledge of other libraries, seaborn, plotnine, etc. Check out the Repo's README for links to further resources.

Time for hands-on exercises!

Exercise 1: Mastering matplotlib

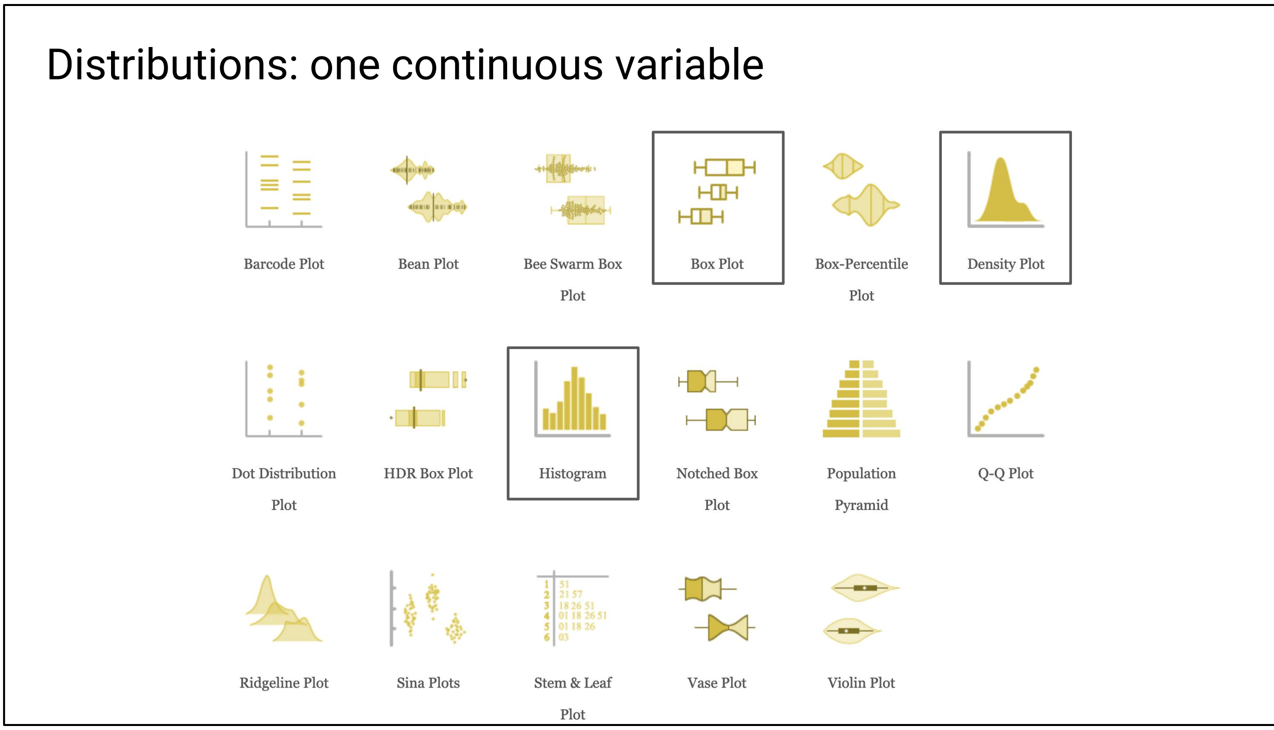
Have your cheatsheet at hand!:

https://matplotlib.org/cheatsheets/



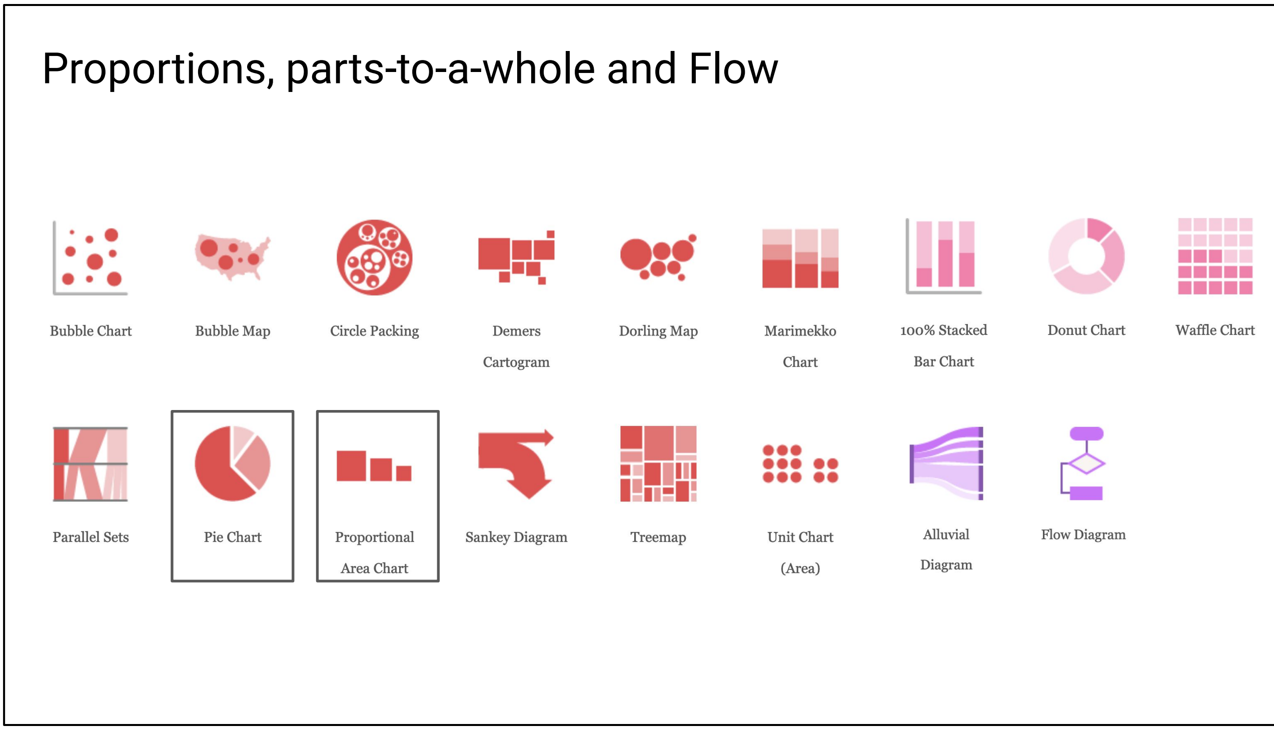
Many types of data visualization tools: datavizcatalogue.com

https://datavizcatalogue.com/search.html
https://datavizcatalogue.com/blog/chart-selection-guide/



Charts that display frequency in the data and/or how spread out over an interval the data values are. Often, these chart types are useful for generating shapes or patterns that provide insight into the kind of distribution present in the dataset.

Marked with a square are the ‘prototypical’ visualizations of each category. Depending on your scientific field these ‘prototypes’ can vary in their use.

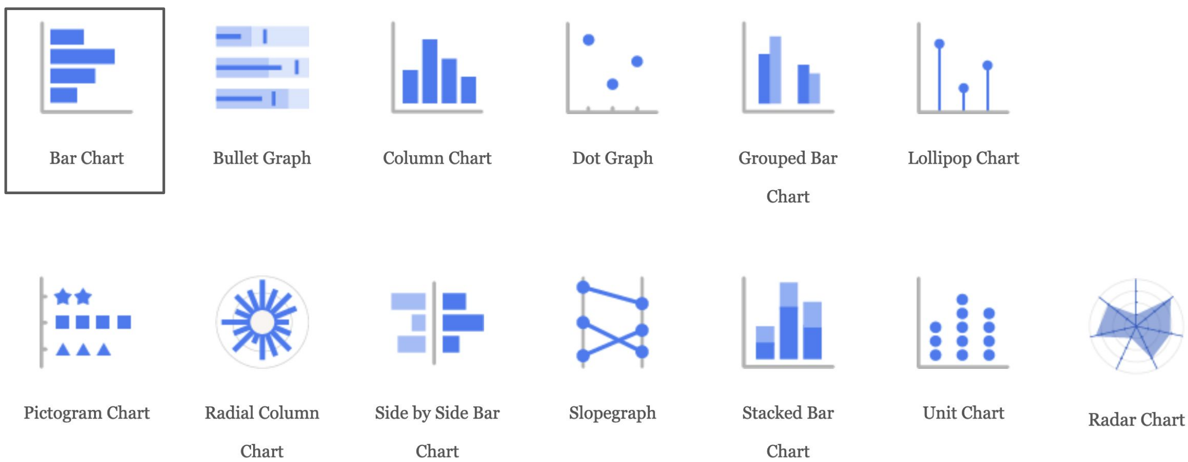


Charts that use area size to communicate differences or similarities. This can be to communicate magnitude, flow, to compare values, or to show a parts-to-a-whole relationship.

Charts used to show proportions but in a parts-to-a-whole relationship. So when the goal is to show how the parts of a variable relate to a total. These charts can also be useful in showing how the data is divided up.

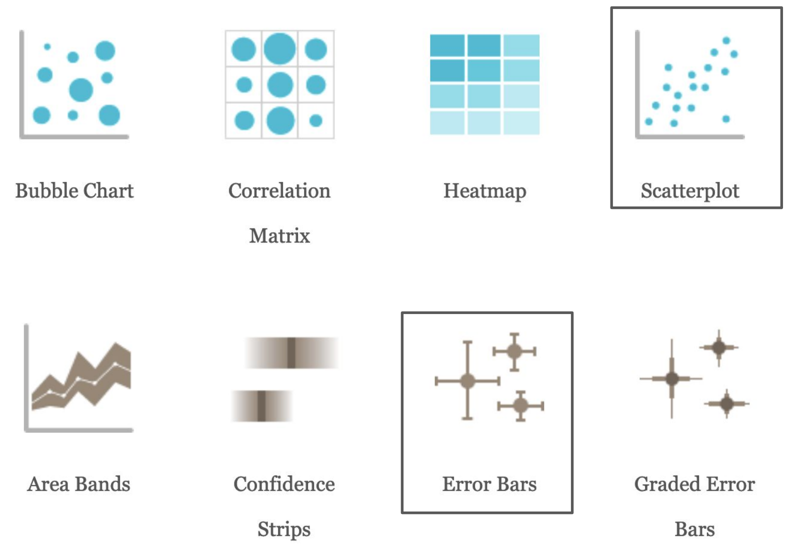
Charts used to communicate the movement or flow of entities or to communicate how a process or system works.

Comparisons: more than one variable



Charts that help communicate the differences or similarities between values in a dataset. Often, these charts are used to make comparisons across categories and/or for communicating the ranking between categories.

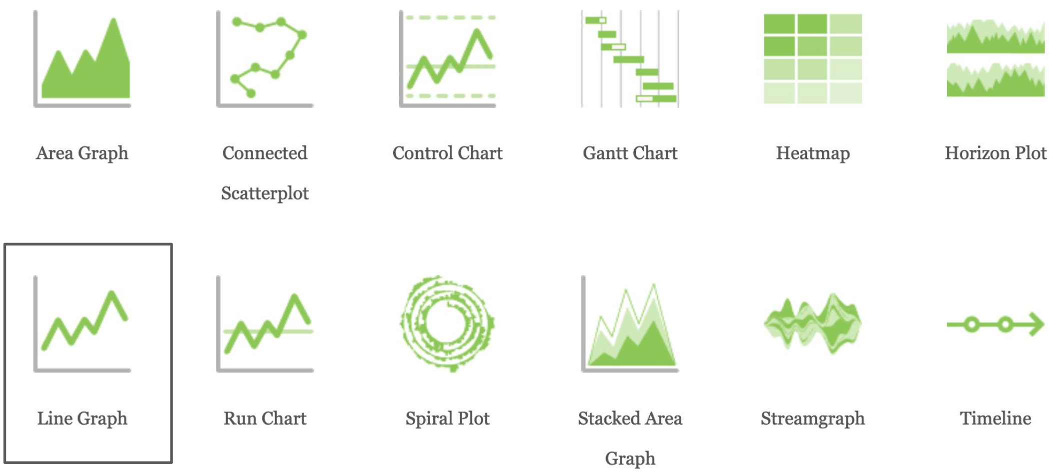
Correlations and Uncertainty/Error



Charts used for finding and/or showing the correlation (when one variable has an effect on another) between two or three variables.

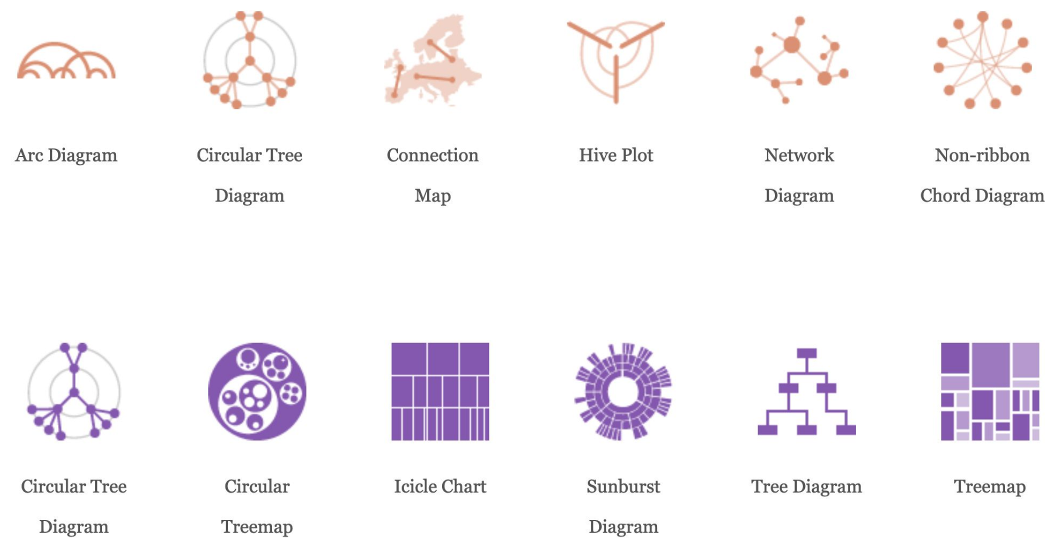
Charts used to communicate and/or analyse the range of error or uncertainty within a dataset.

Data over time: timeseries



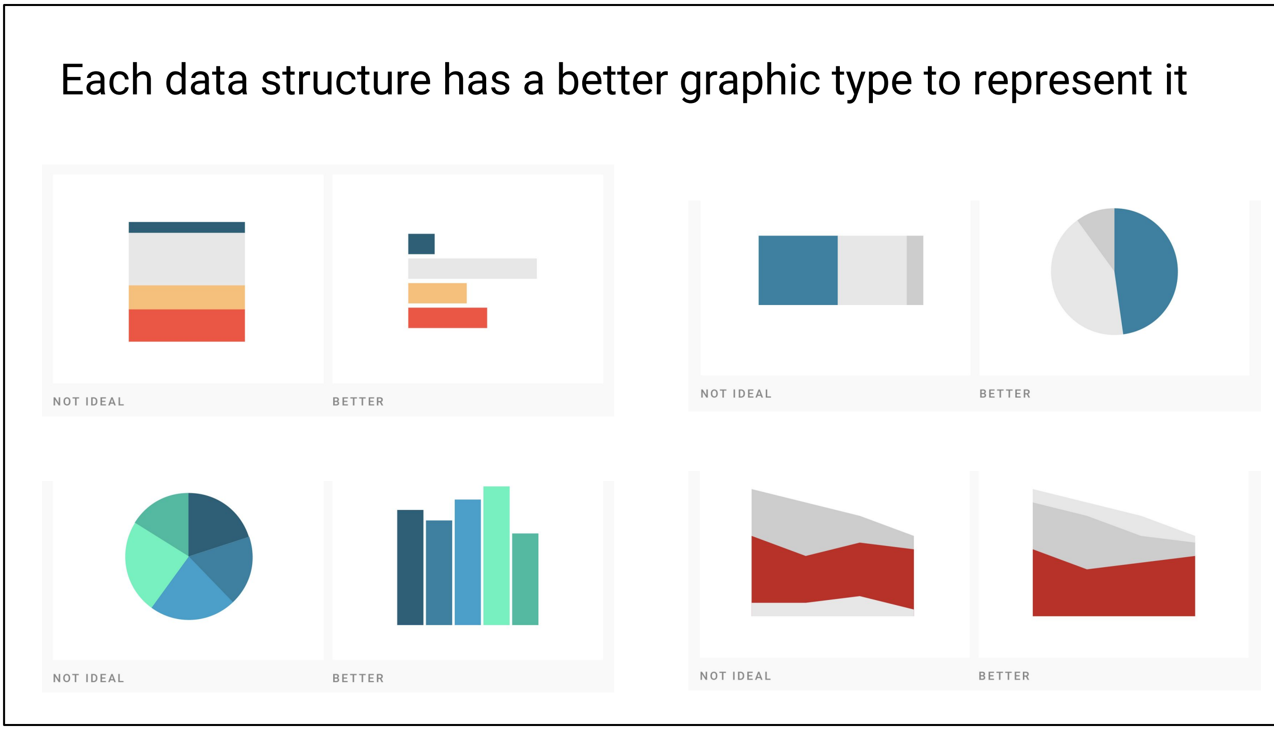
Charts typically used to show data changing over a time period to communicate or analyze trends and patterns in a dataset. The more diagrammatic editions below, such as a **Timeline** or **Gantt Chart** are more used to communicate the sequencing of events.

Connections and Hierarchy



Charts used to communicate or analyse how things are connected, the structure of networks, or the relationships contained within networks.

Charts used to communicate and/or analyse how things are ranked and ordered together in an organisation or system.



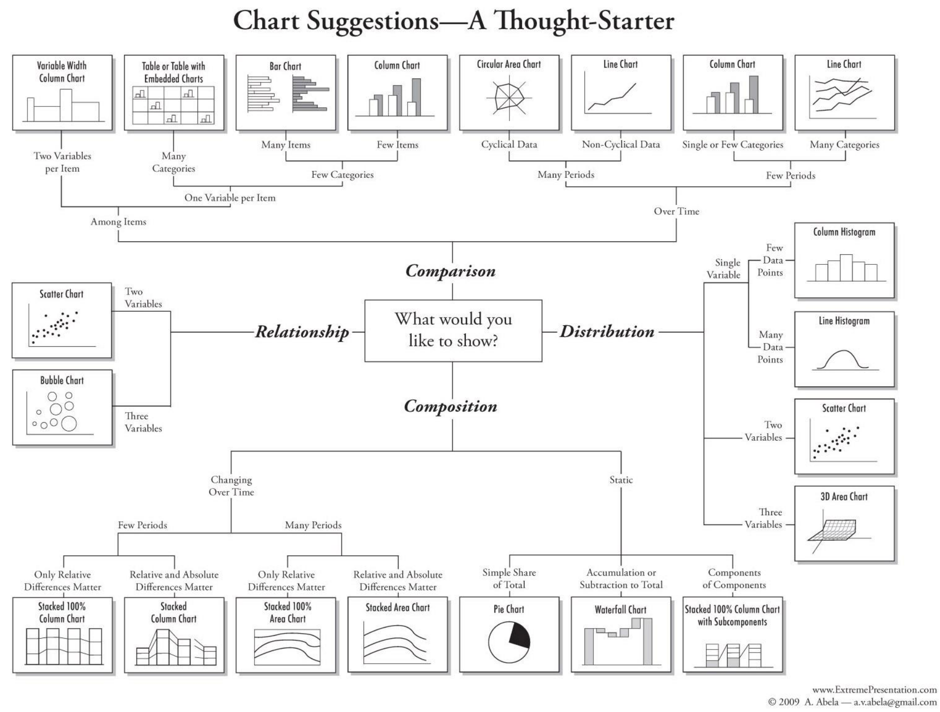
Stacked column charts are great to show the parts of multiple totals. If you only want to show parts of one total, consider a bar chart instead.

A column chart is a better choice than a pie chart to show which value is the highest.

But a pie chart is a valid choice when the focus is on the value that's close to or exactly 25%, 50%, 75%.

When stacking values, order is important!

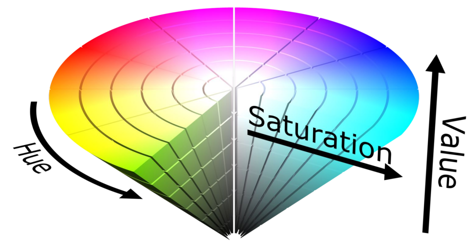
If you’re
completely
overwhelmed by
options,
you can use:



If you’re completely overwhelmed or lost in the amount of options out there… you can use this workflow to decide which visualization to use first. You will get already something very reasonable!

6) Use color effectively

Three dimensions of color: Hue, saturation and brightness

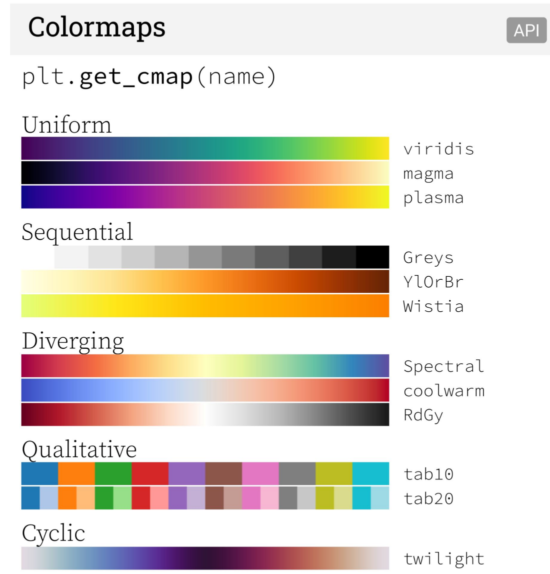


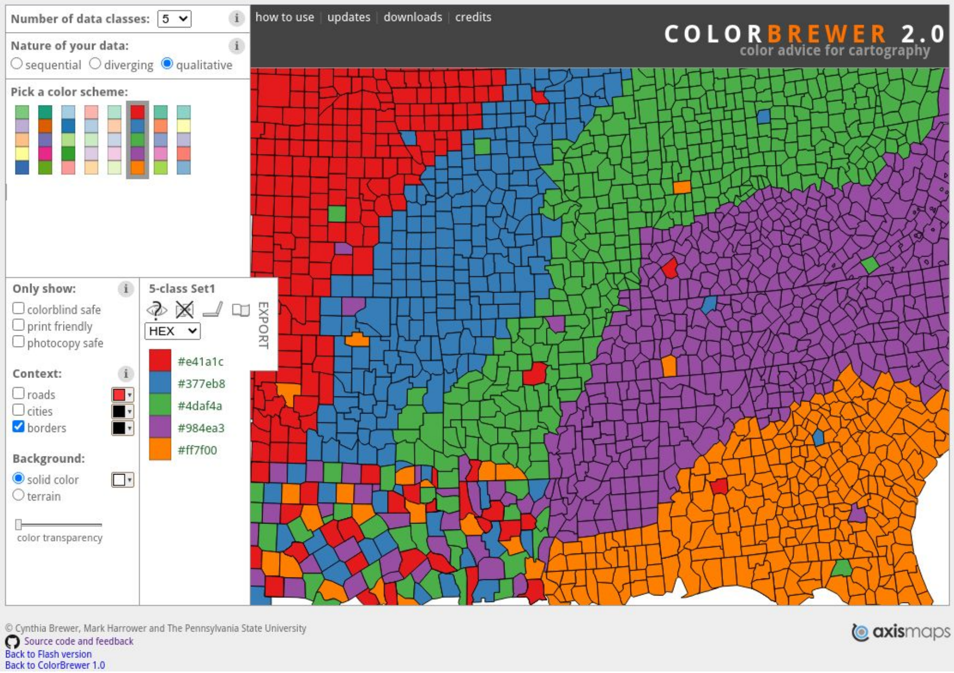
https://lisacharlottemuth.com/
https://blog.datawrapper.de/beautifulcolors/
https://blog.datawrapper.de/colorguide/
https://blog.datawrapper.de/colors/

https://colorizer.org/
https://www.sessions.edu/color-calculator/ -> do not use tetraedric

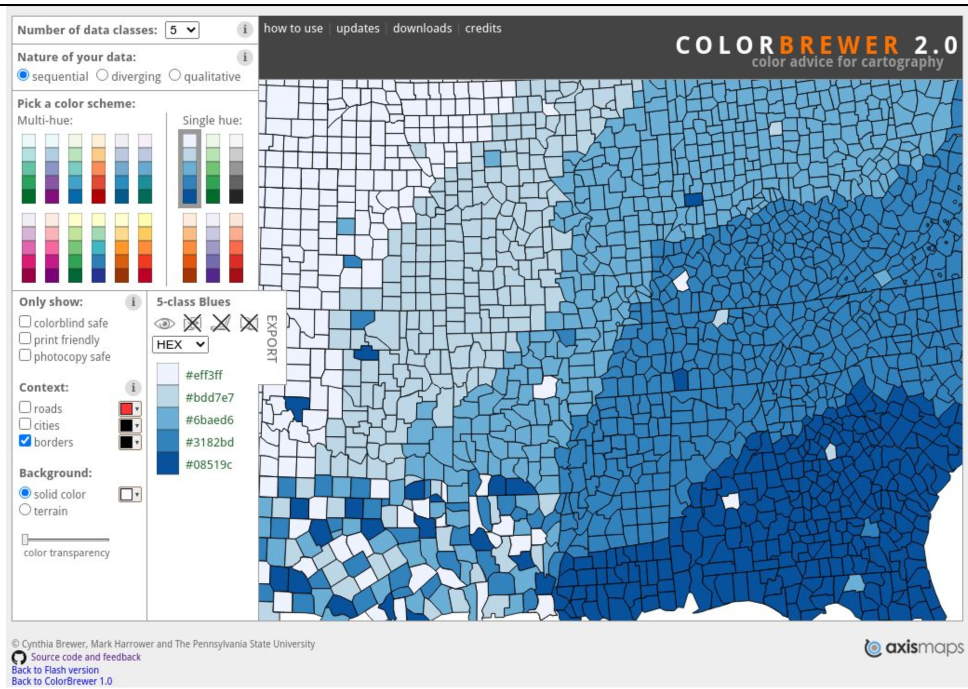
Types of color scales

- **Qualitative/categorical**: data with no order
  - e.g. cities, countries
- **Sequential**: increasing or decreasing data
  - e.g. year
- **Diverging**: data with a natural zero
  - e.g. % change, temperature
- **Circular**
  - e.g. orientation, direction



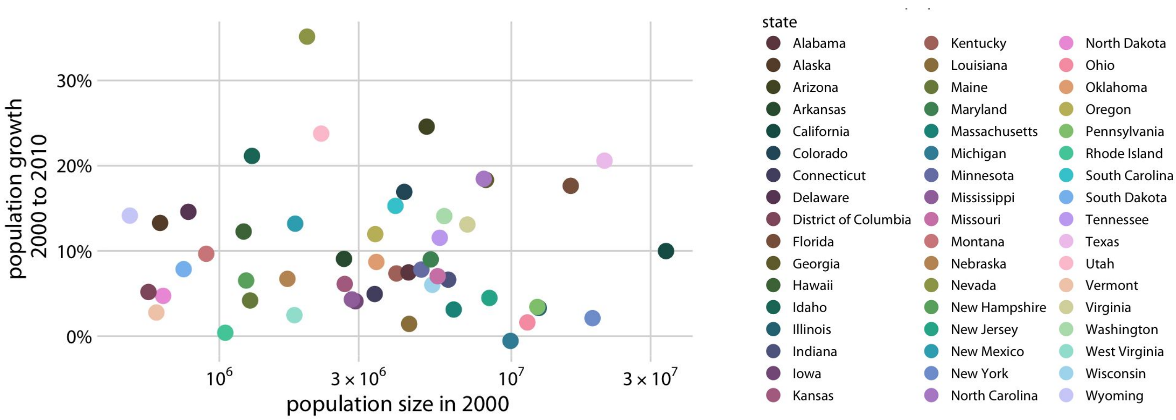


https://colorbrewer2.org



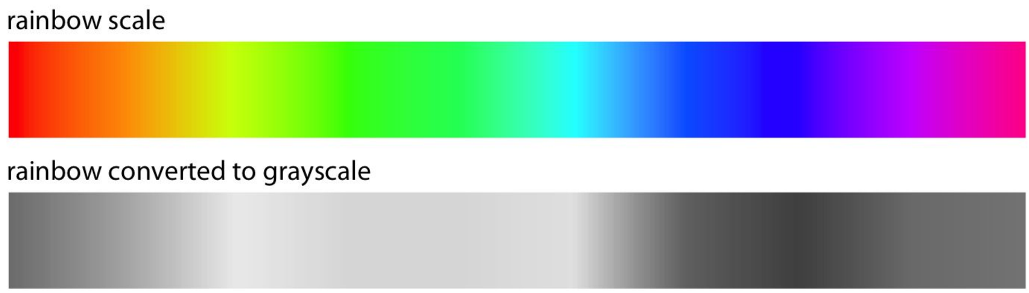
https://colorbrewer2.org

Common pitfall: encoding too much information



Claus O. Wilke (2019)

Common pitfall: using the wrong color scale



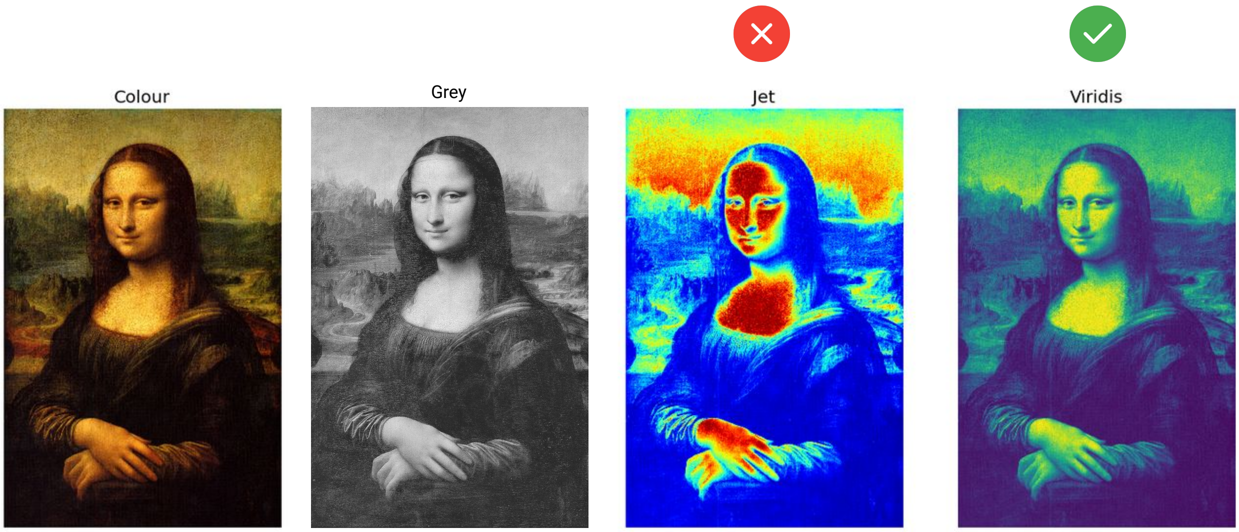
The jet/rainbow color scale is **NOT a sequential** colormap, as our perception of it is NOT linear but **circular!**

We perceived the rainbow / jet color scale as not linear!

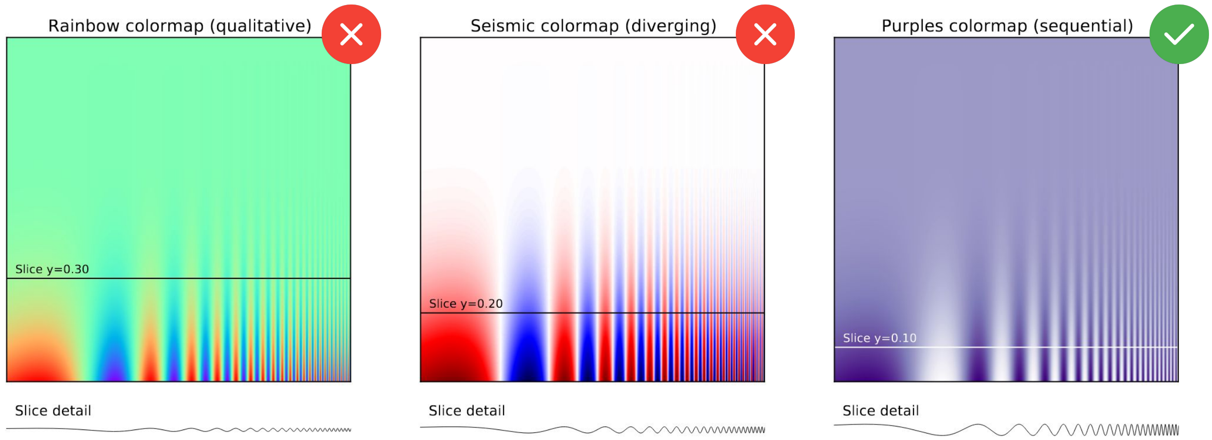
Differences in hue are not related to difference in light intensity. This becomes clear when we render a grayscale version of it. We tend to see the yellow green part as lighter as the blue.

So we cannot map increasing or decreasing data to hue! DO NOT USE the rainbow / jet color scale.
It’s not a default in matplotlib anymore!

Common pitfall: using the wrong color scale



Common pitfall: using the wrong color scale



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.

Use the colormap that is the most relevant to your data. Consider color blindness when choosing colors.

Time for hands-on exercises!

Exercise 2: which visualization should I use?

Have your cheatsheet at hand!:

https://matplotlib.org/cheatsheets/

Extra-Material (from ASPP-2021)

- **Scales & projections** (**notebook**). Tutorial on different type of scales (log scale, symlog scale, logit scale) and projections (polar, 3D, geographic).

- **Animation** (**notebook**). Animation with matplotlib can be created very easily using the animation framework. This notebook shows how to create an animation and save it as a movie.

Further Resources

At the implementation level (code, galleries and how-tos):

- **Seaborn library**, a library for statistical data visualization. Very recommended as a next step in your learning journey.
- **Matplotlib Cheatsheets**, Nicolas P. Rougier (2020)
- **Scientific Visualization – Python & Matplotlib**, open-source book from Nicolas P. Rougier (2021)
- **Python Graph Gallery**, Yan Holtz (2017)
- **Matplotlib Gallery**, Matplotlib team

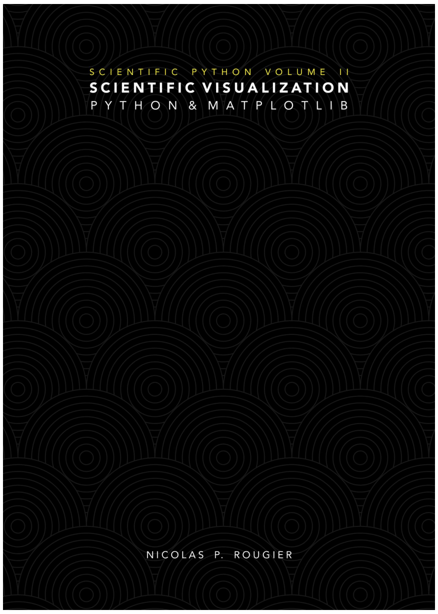
At the conceptual level :

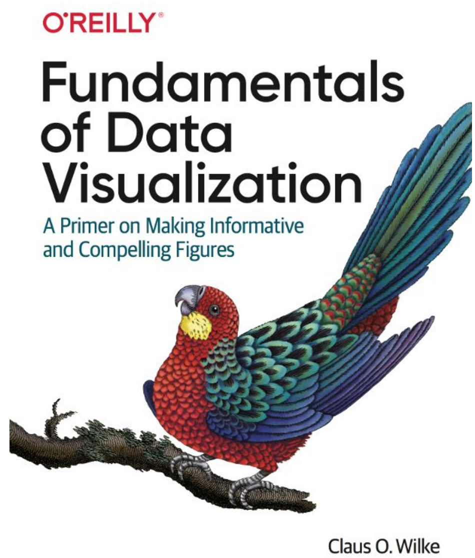
- **Ten simple rules for better figures**, Nicolas P. Rougier, Michael Droettboom, Philip E. Bourne (2014)
- **Fundamentals of Data Visualization**, book by Claus O. Wilke (2019)
- **Chart Suggestions - a though-starter** by A. Abelas.
- **Data Visualization Catalogue**
- **Edward Tufte**'s series of books: The Visual Display of Quantitative Information (1983), Envisioning Information (1990), Beautiful Evidence (2006), etc.

Interactive visualizations:

- **Widgets in Jupyter notebook**
- **Plotly**

Selected further resources



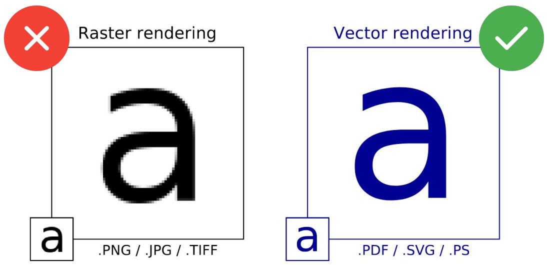


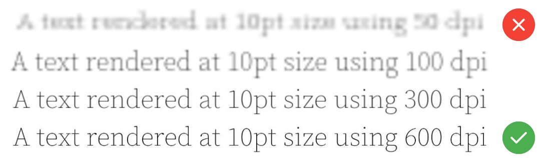
Resources:
https://github.com/rougier/2021-Dataviz
https://github.com/rougier/2021-Dataviz/blob/master/Lesson/dataviz.pdf
https://hal.inria.fr/hal-03427242/document
https://datavizcatalogue.com

Some extra tips

Exporting a figure: vector format!

As a rule of thumb: Save in vector format and with enough DPI (dots per inch)





Bitmap formats

PNG: Portable Network Graphics (lossless)
JPG: Joint Photographic Experts Group (lossy)

Vector formats

PDF: Portable Document Format
SVG: Scalable Vector Graphics

Text rendered in matplotlib and saved using different dpi

Format of a figure: lossy/lossless compression, what is DPI, how to create a figure with text of same size as paper text…
Script to modify a figure that was already done (after review), generic label

Font stack choice

Influence of typography on the perception of a figure. Choose the right font for you.