DGTIC UNAM
DIRECCIÓN GENERAL DE CÓMPUTO Y DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

GOBIERNO DE LA CIUDAD DE MÉXICO
SECRETARÍA DE EDUCACIÓN, CIENCIA, TECNOLOGÍA E INNOVACIÓN

Best practices in data visualization

Guillermo Aguilar & Carlos Cernuda

With material from Aina Frau-Pascual & Nicolas P. Rougier

Mexico City, ASPP LATAM 2023

Plan

18:00 Principles of data visualization

Hands-on Exercise 1: **mastering matplotlib**

19:00

Tomorrow

09:00 (cont). Exercise 1

Types of visualizations - Use of color - Common pitfalls

Hands-on Exercise 2: which visualization should I use?

\+ **Review** of your solutions as PR

10:30 **END**

(*) Hands-on Exercise 3: images

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

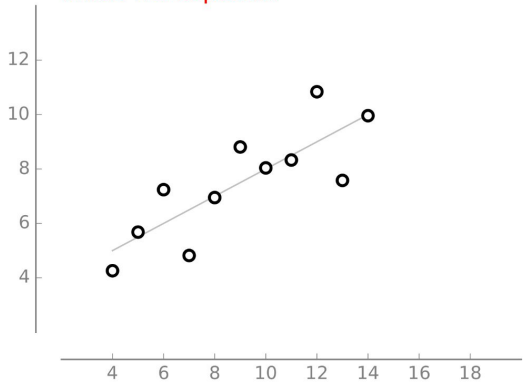
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

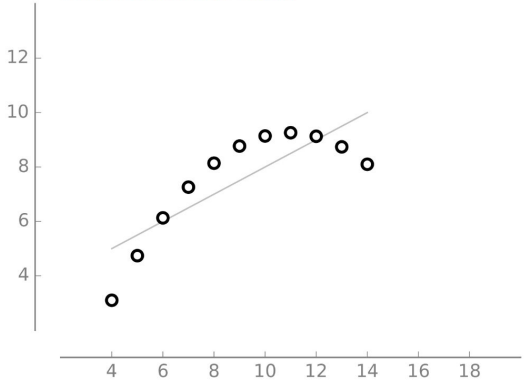
Anscombe (1973)

Classical example: Anscombe's quartet

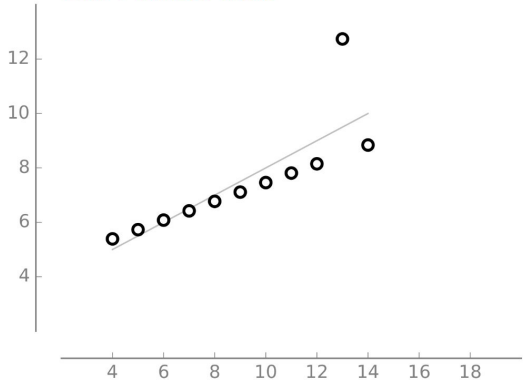
What we expect...



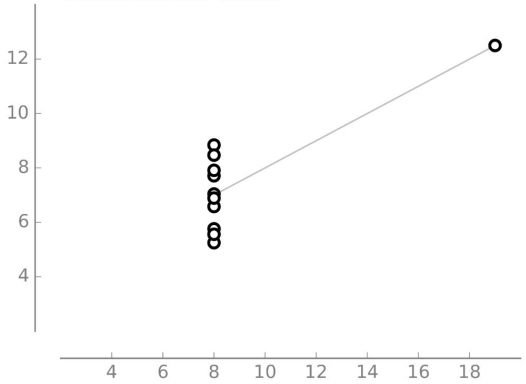
The non-linear case



The Y outlier case



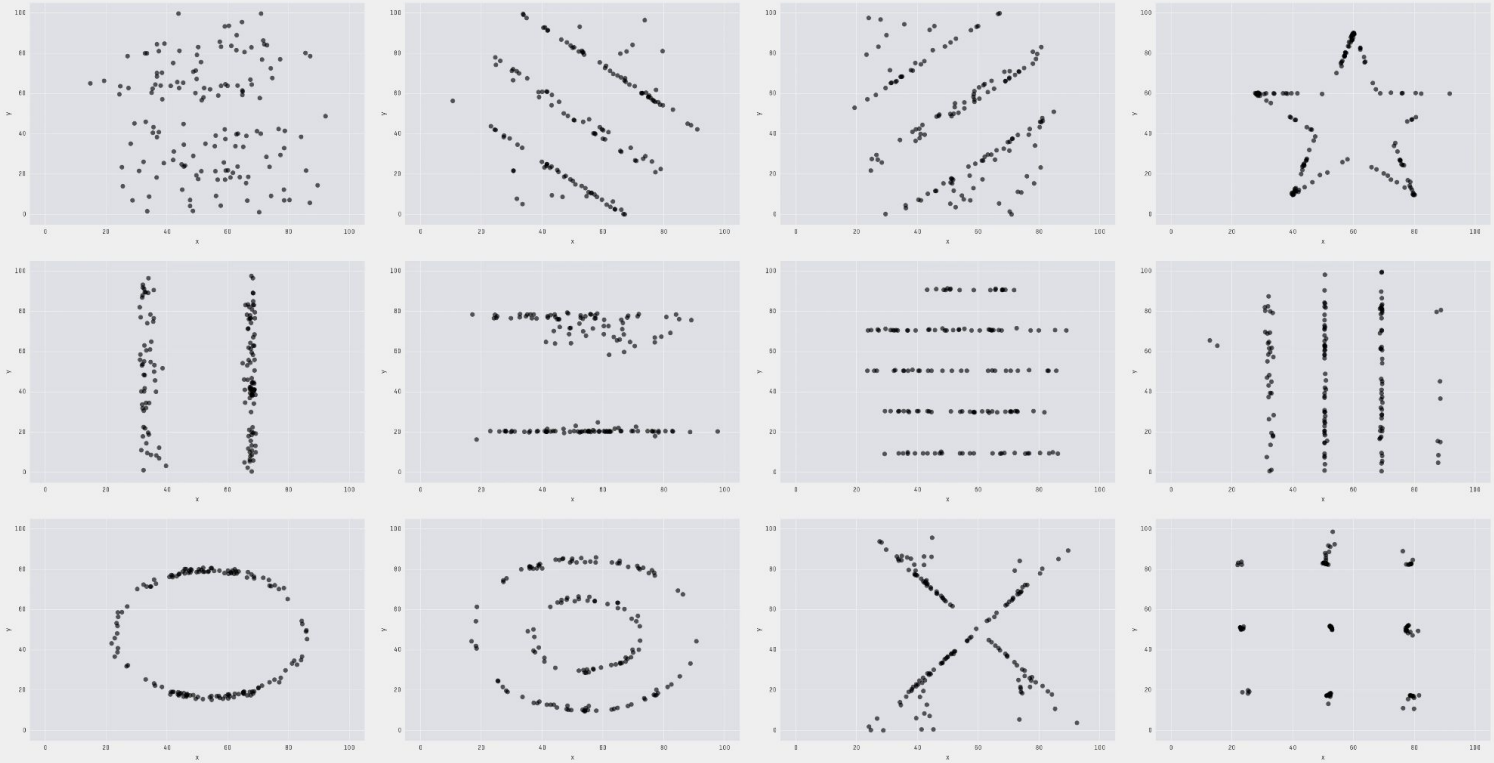
The X outlier case



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

Datasaurus

X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06



And you will read this last

You will read this first

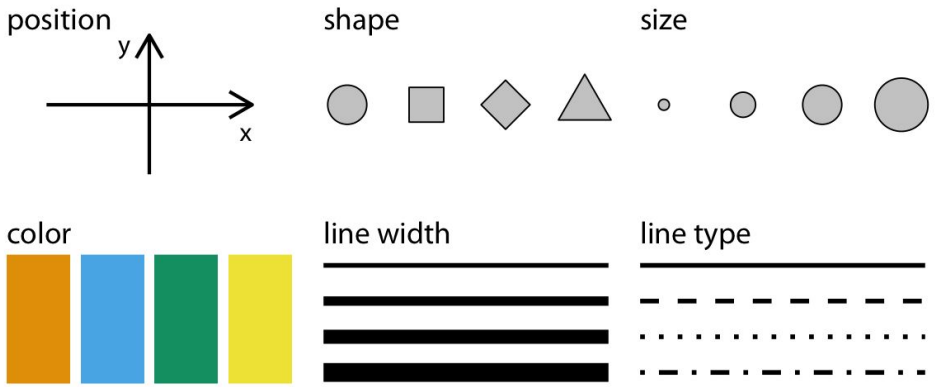
And you will read this

Then this one

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)

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

1) Know your audience

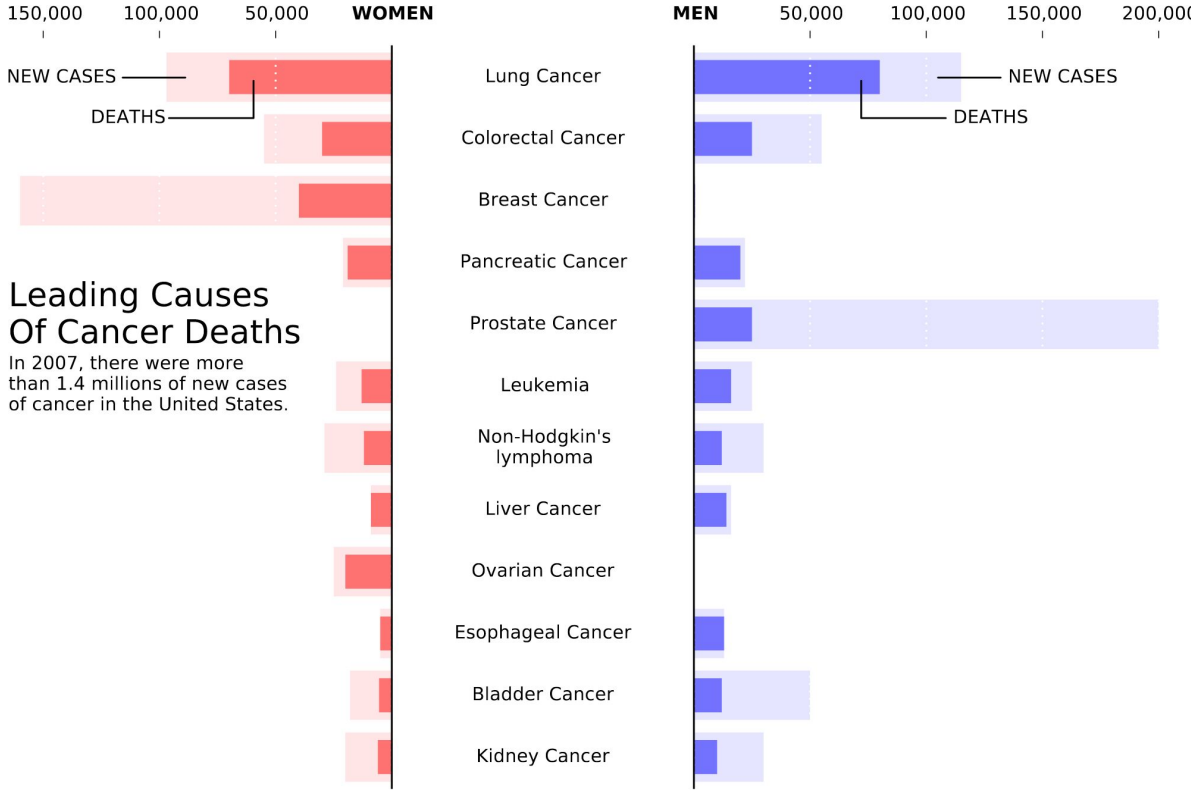
- Complexity +

My colleagues

Scientific community

Student audience

General public



Audience: general public

Main message: cancer

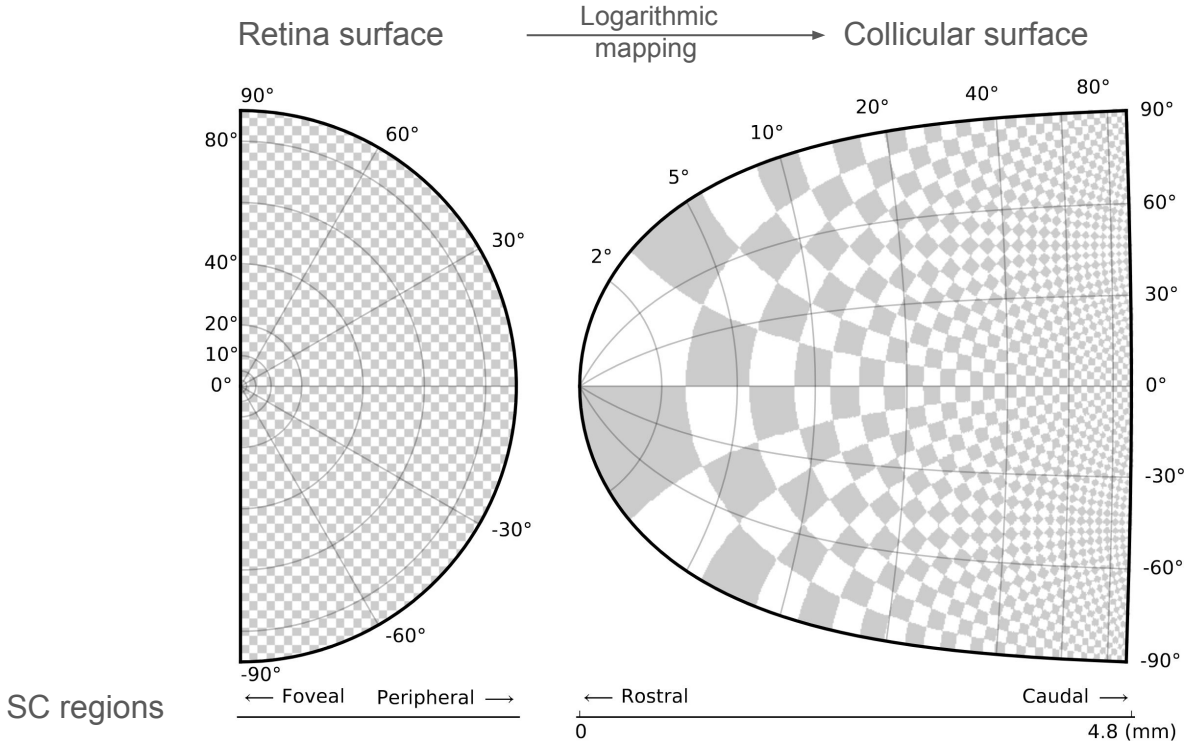
Separated in sex groups: Women / Men

2) Identify your message

Audience:

neuroscience scientific community

Superior Colliculus



SC regions

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.

3) Adapt the figure to support the medium

Figure for a Paper

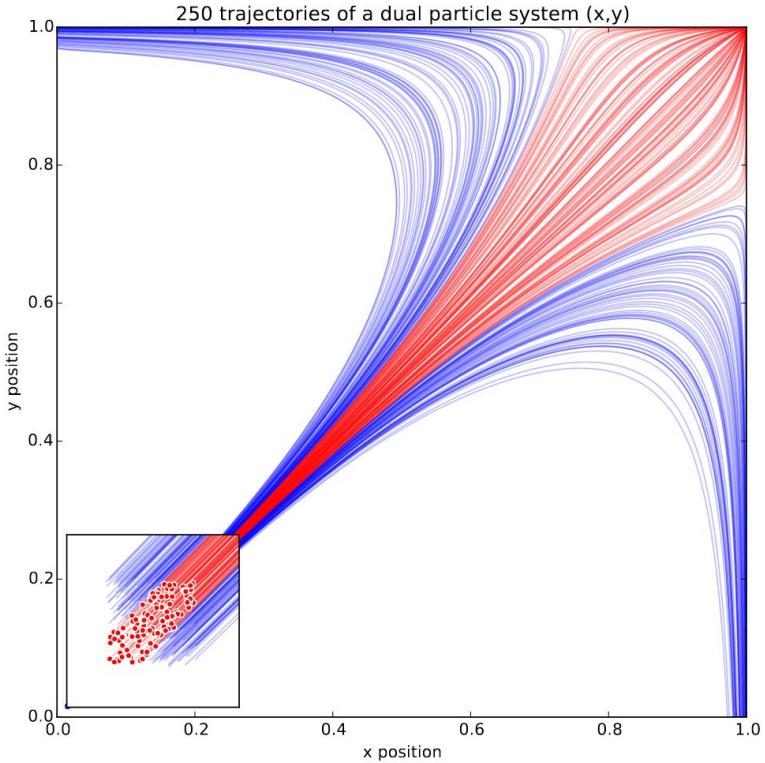
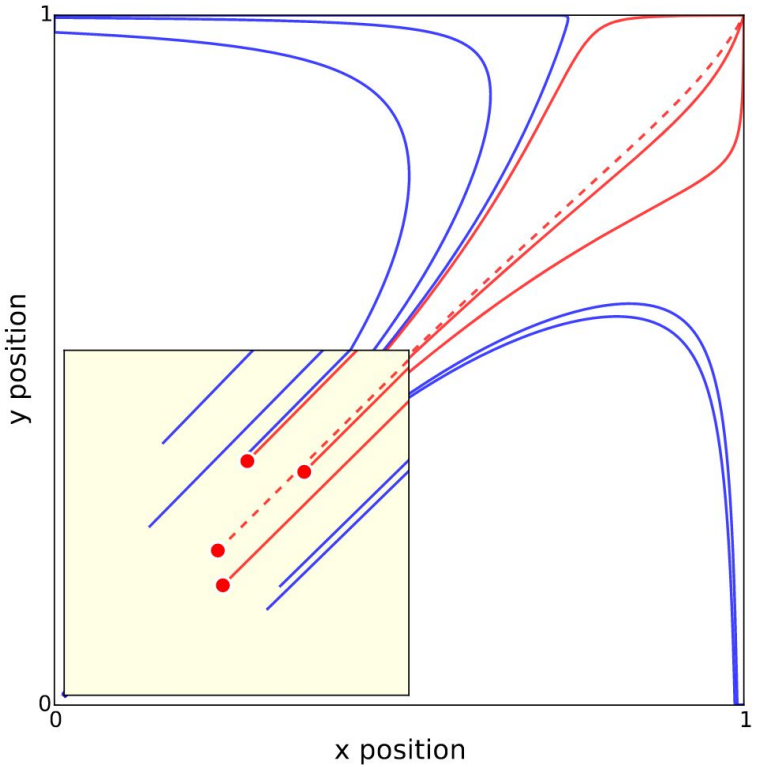


Figure for a Talk



4) Captions are not optional. Neither x and y-labels

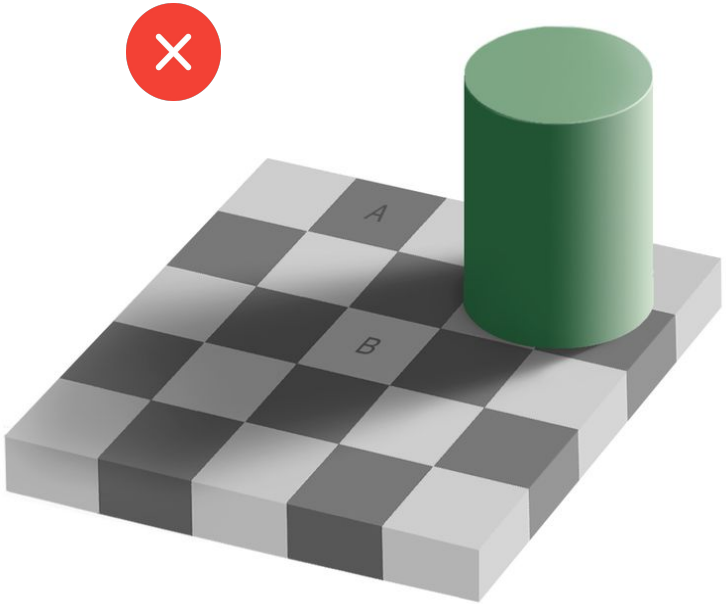


Figure 1. Optical illusion

Caption helps you understand the image

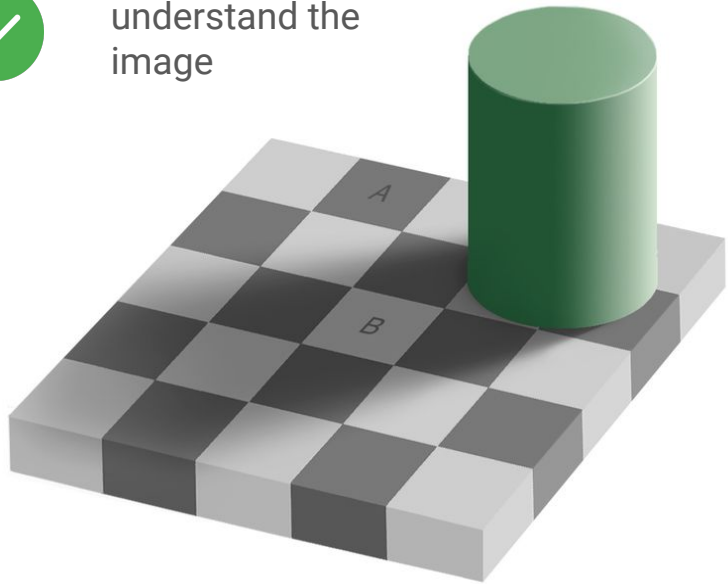
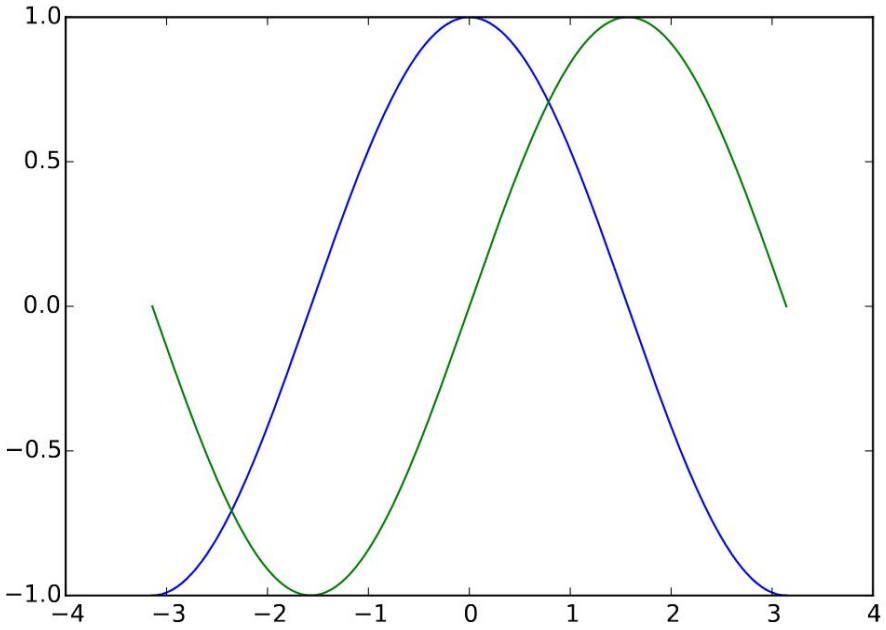


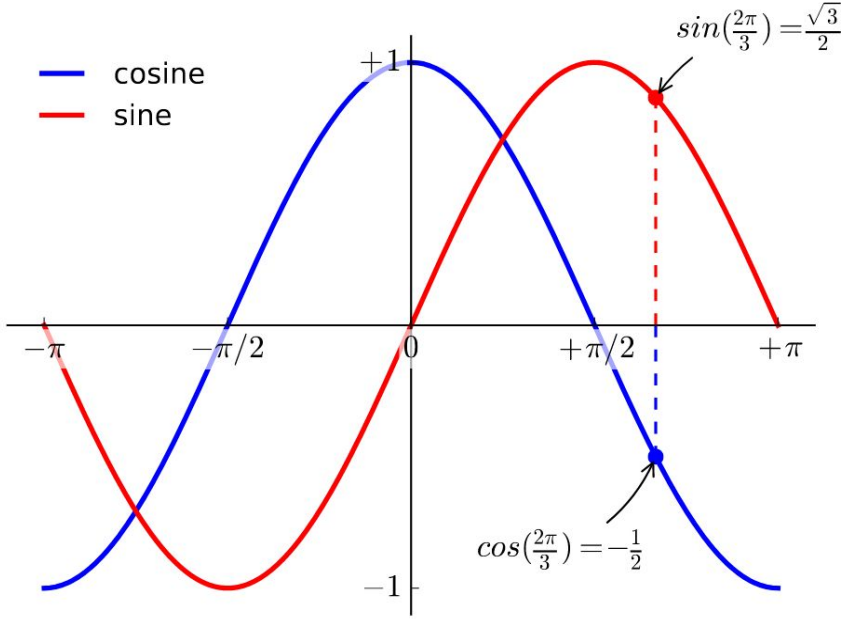
Figure 1. A and B patches are actually the same color even though we perceive them at being different color

5) Do not trust the defaults

Matplotlib defaults



With a bit of work….
(Exercise 1 today)



6) Use color effectively → more on this later

7) Do not mislead the reader

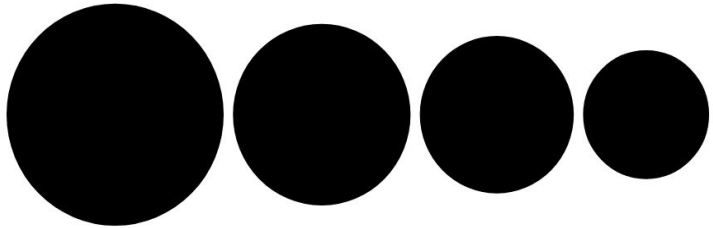
Using full range bars shows a more realistic comparison among them

Relative size using full range

Relative size using partial range

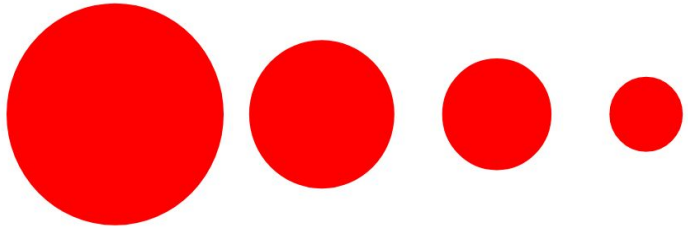
Using **partial range bars misleads the reader** to think the difference is bigger

Using the **disc area** shows a more proportional sizes



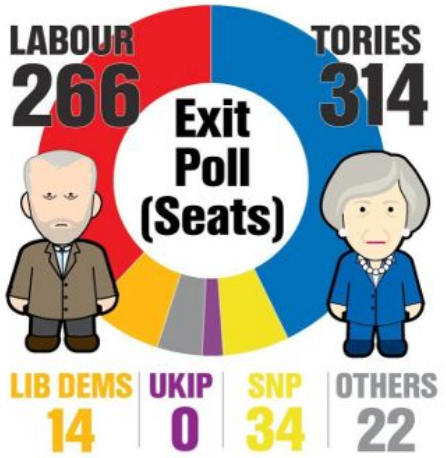
Relative size using disc area

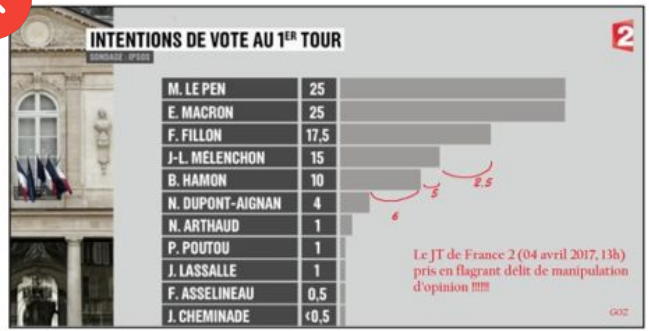
Relative size using disc radius

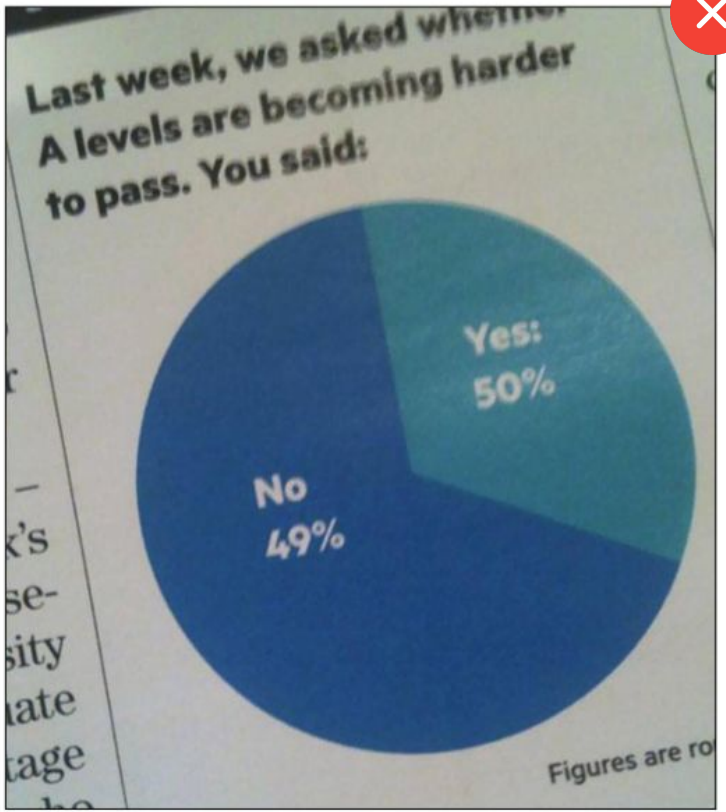


Using the disc radius misleads the reader to think the difference is bigger

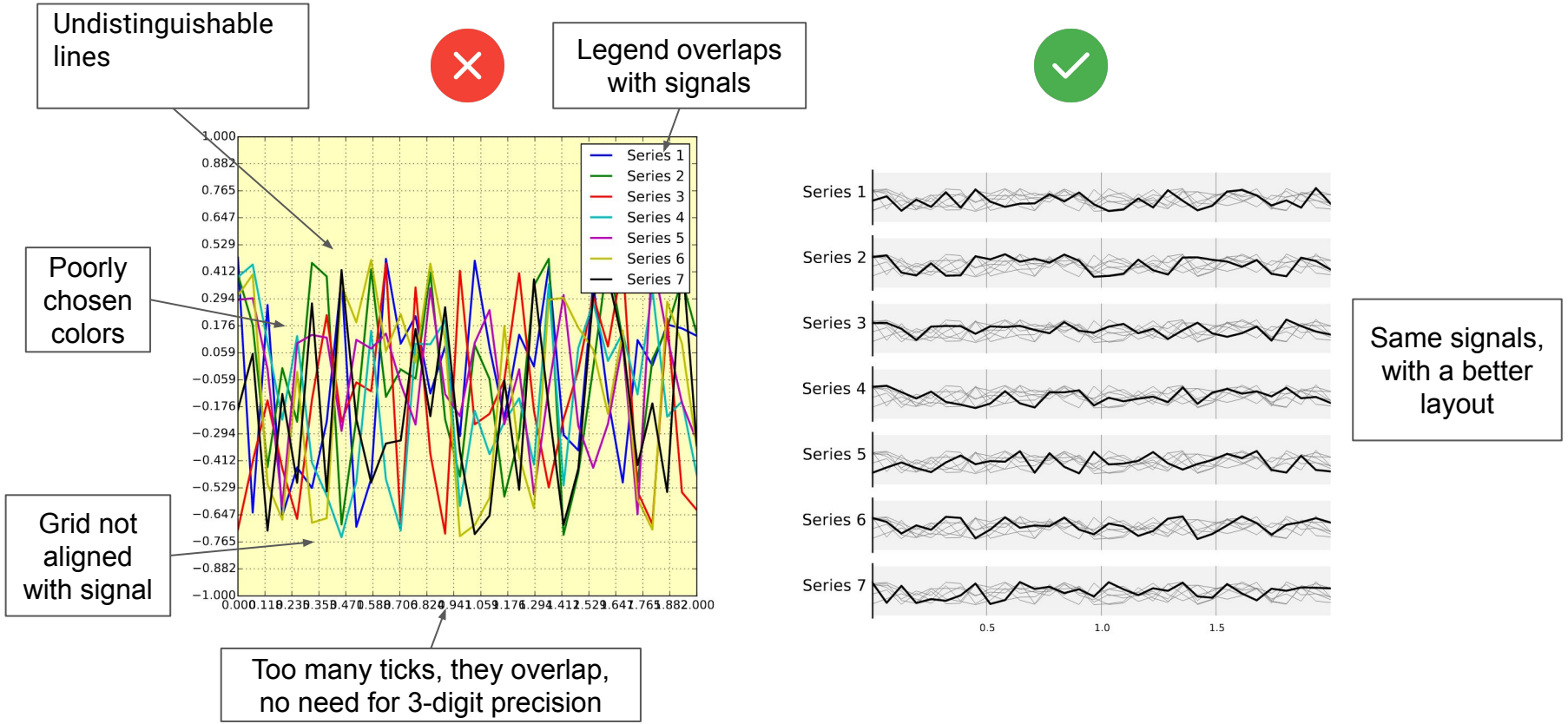
7) Do not mislead the reader. Really.





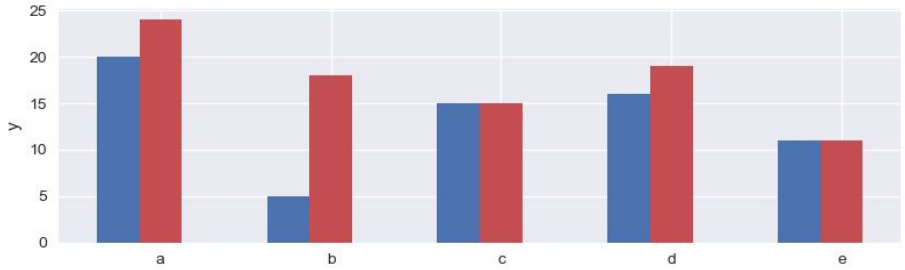


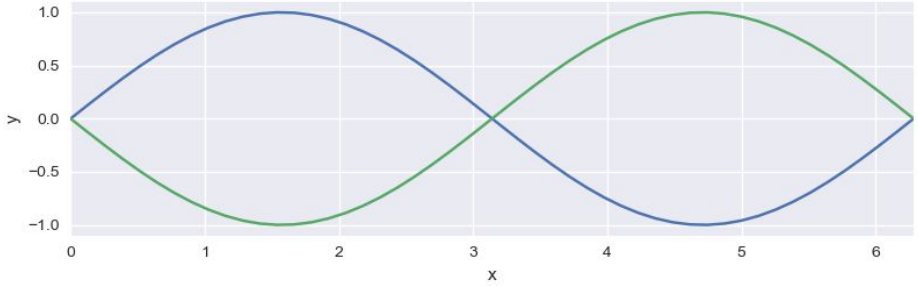
8) Avoid **chartjunk**

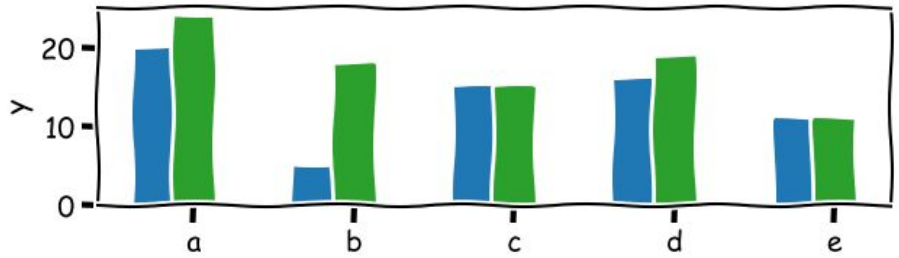


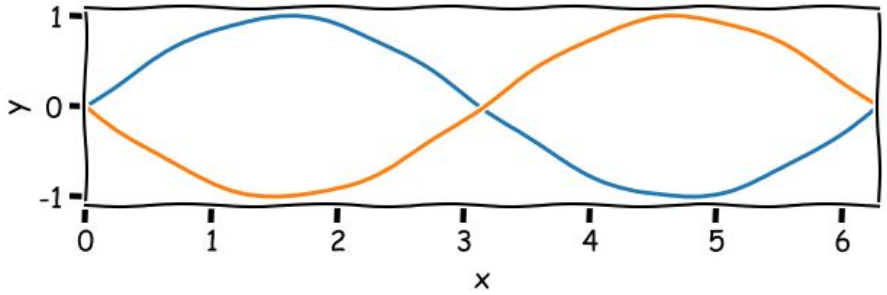
9) Message trumps beauty:

To convey an idea, sometimes an sketch suffices









10) Get the right tool

PDFCrop to remove white borders

GraphViz for creating easy graphs

ImageMagick for scripted image processing

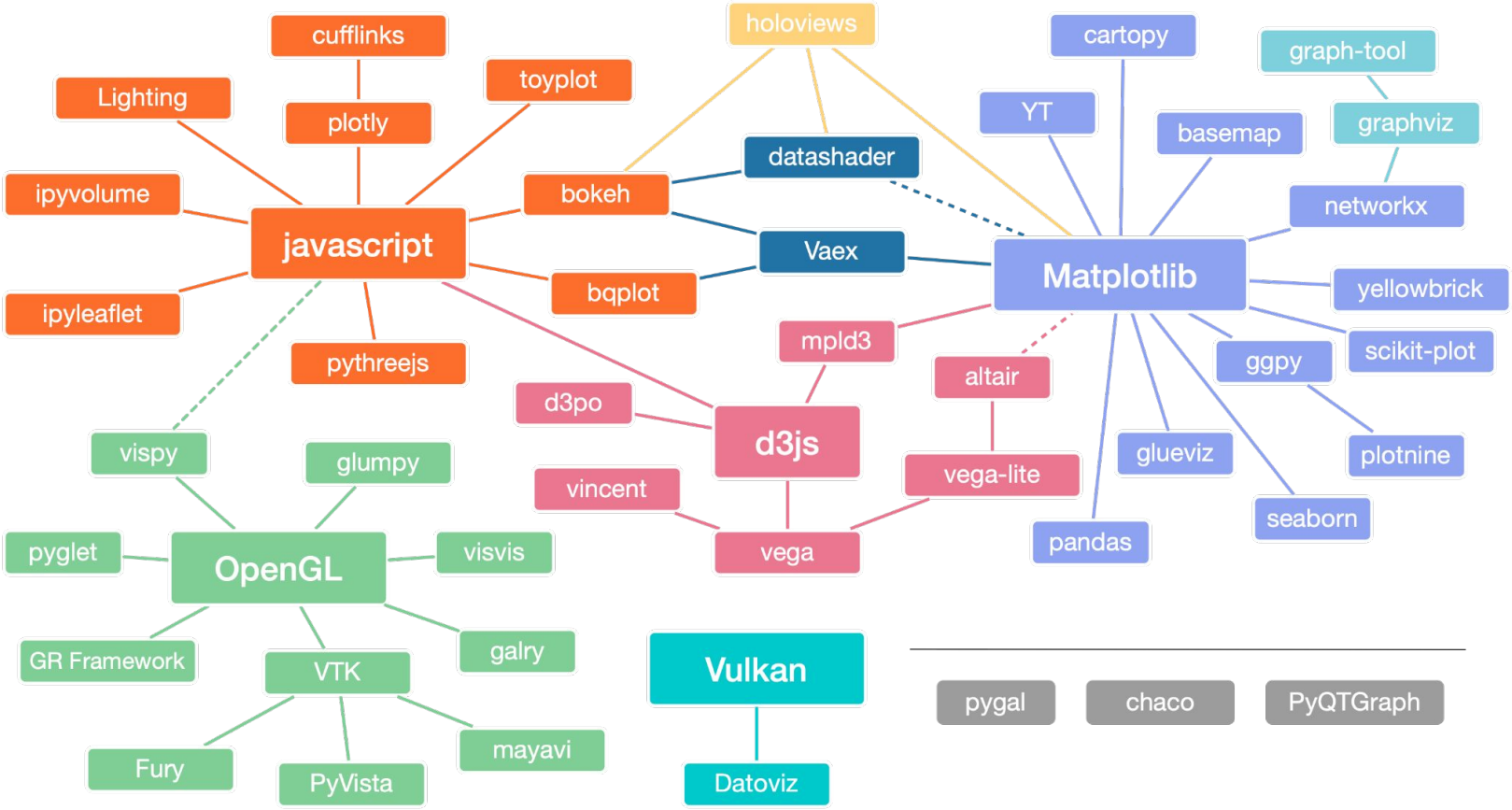
Gimp for bitmap image manipulation

Inkscape for vector image manipulation

Tikz for scripted vector art

And many, many, many others…

Overview of visualization libraries

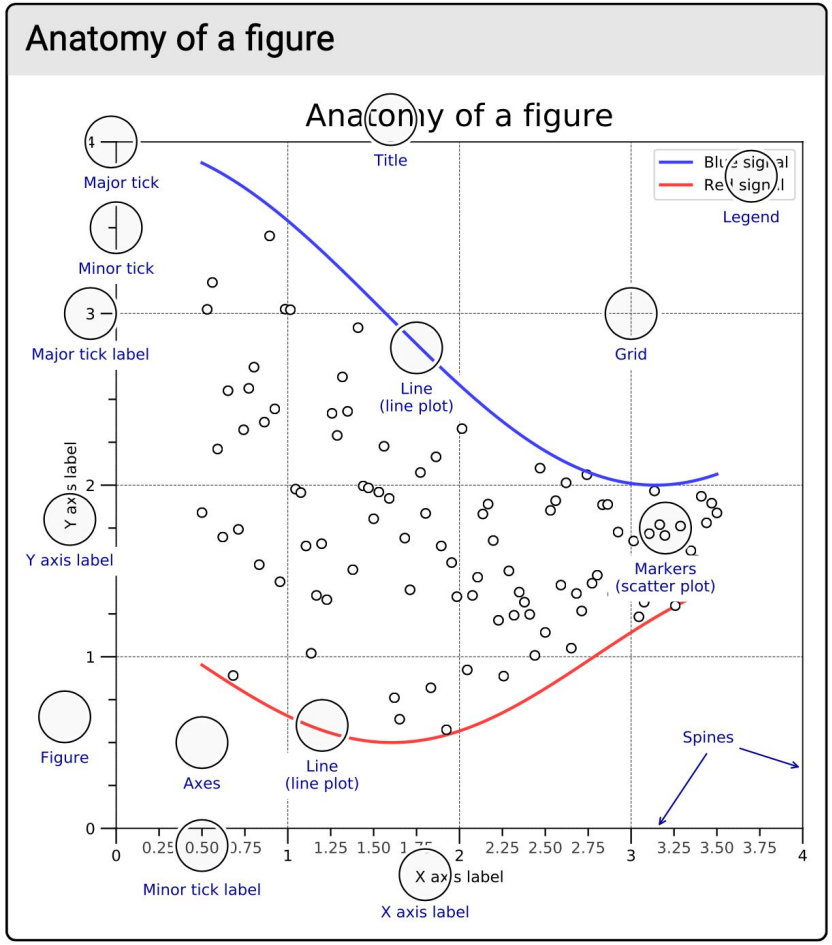


Time for hands-on exercises!

Exercise 1: Mastering matplotlib

Have your cheatsheet at hand!:

https://matplotlib.org/cheatsheets/

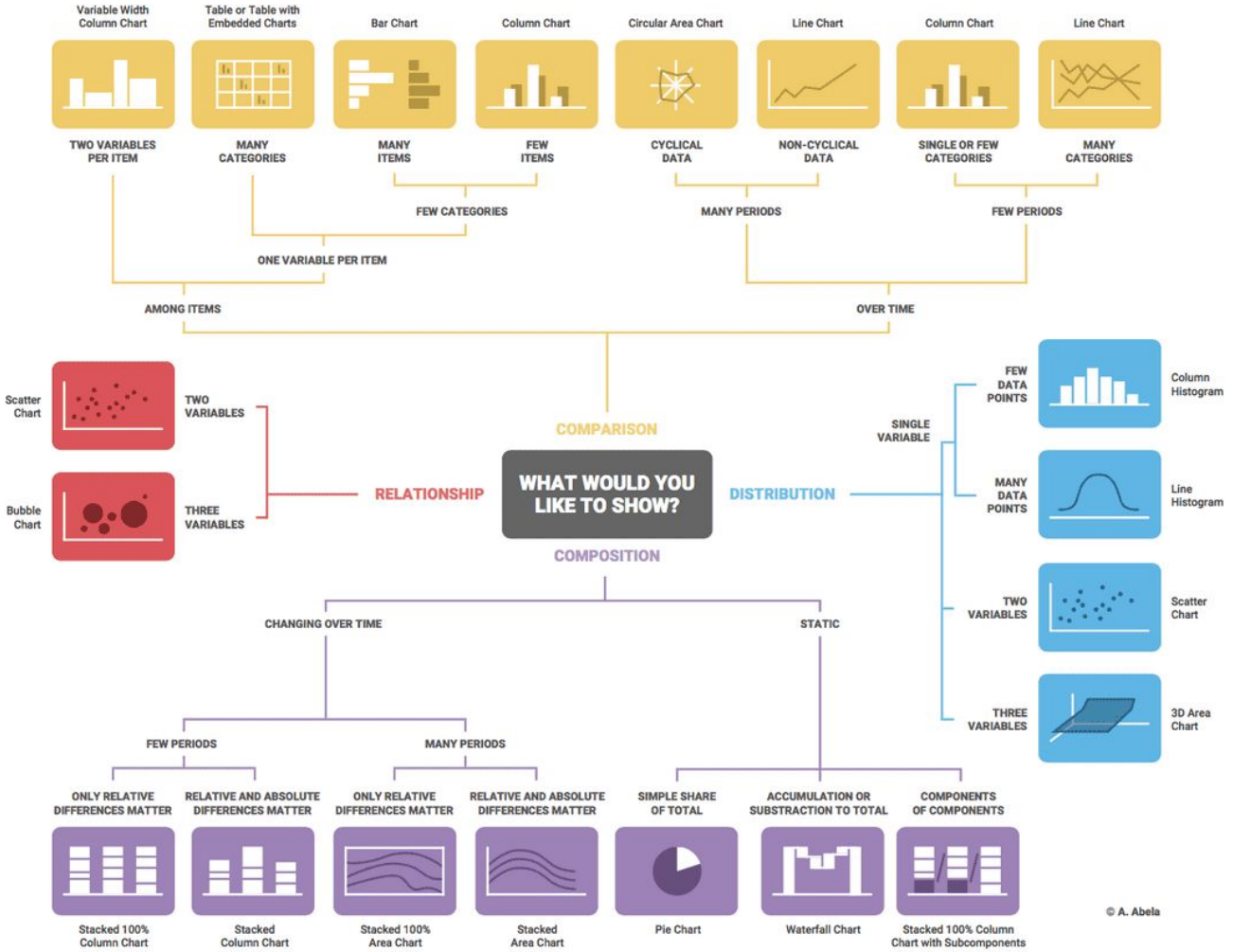


An useful tool:

datavizcatalogue.com

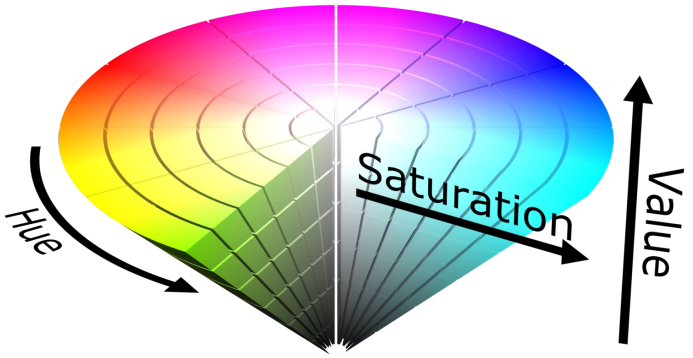
Another,
older tool

Chart
suggestions
by Abela



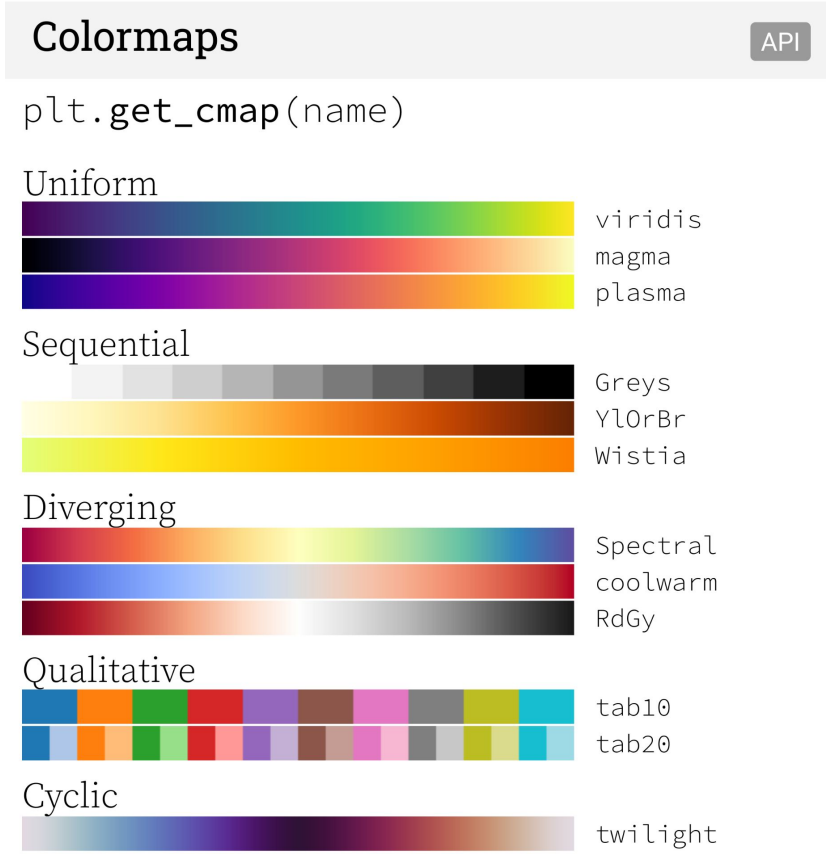
6) Use color effectively

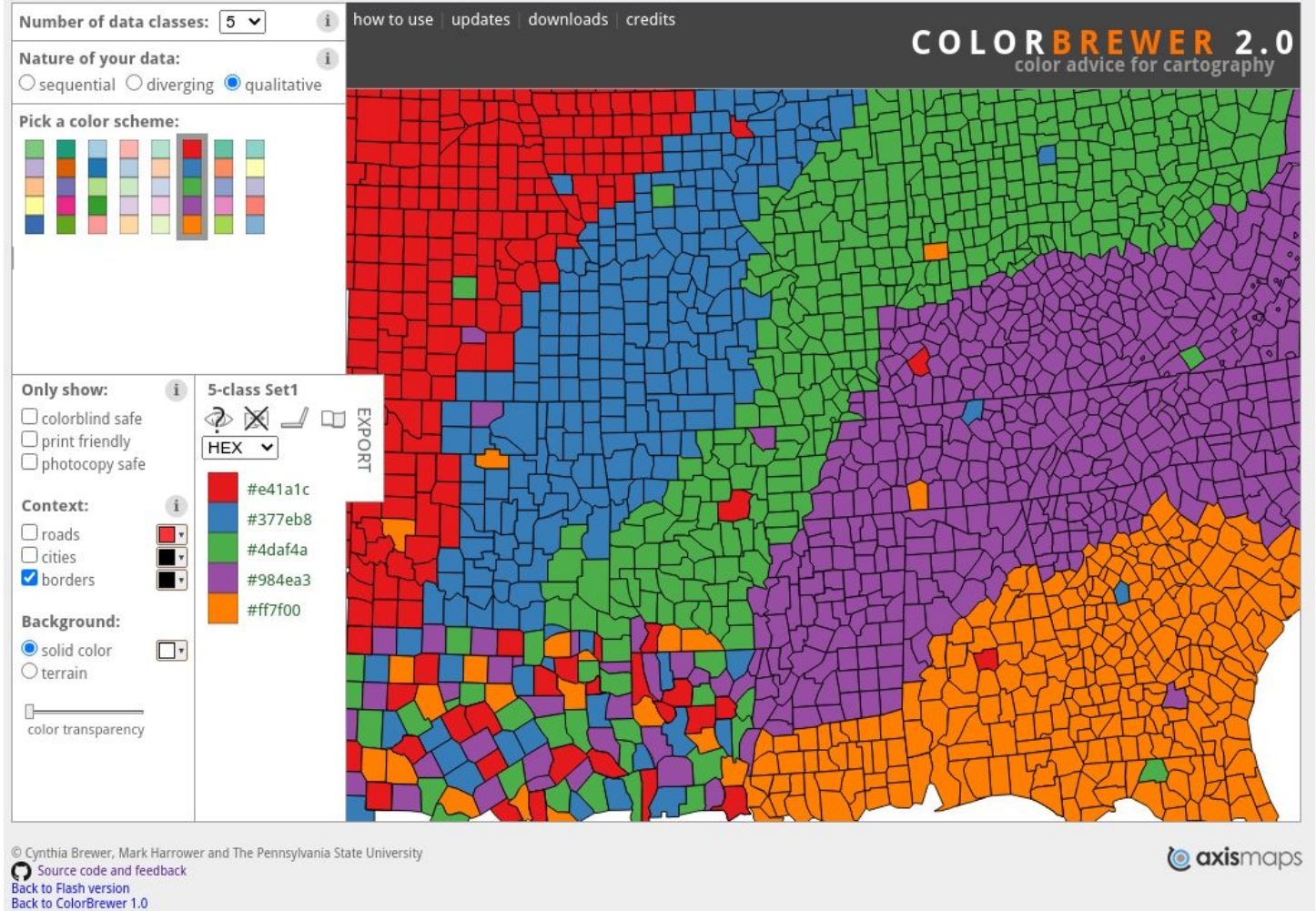
Three dimensions of color: Hue, saturation and brightness



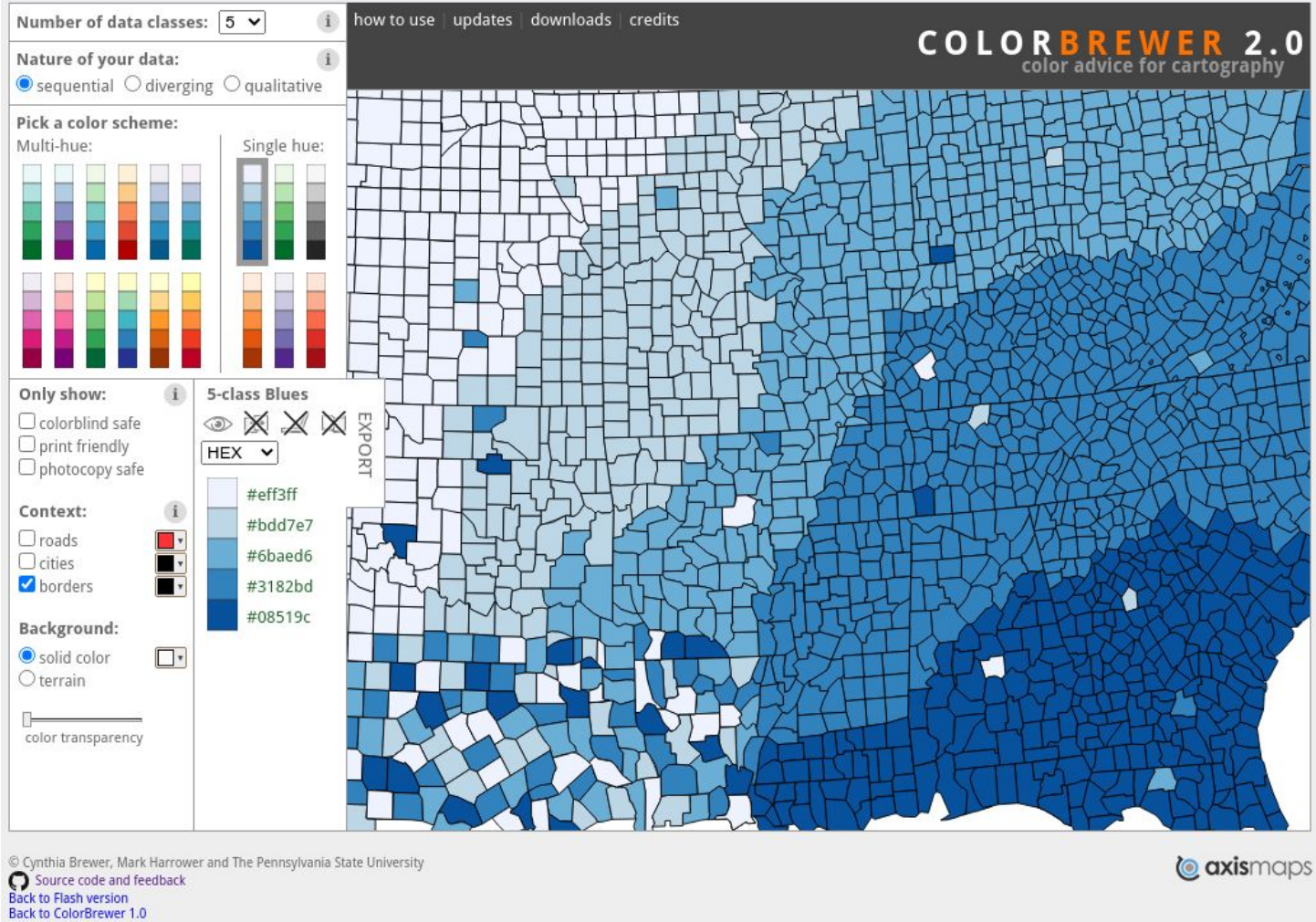
Types of color scales

- **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



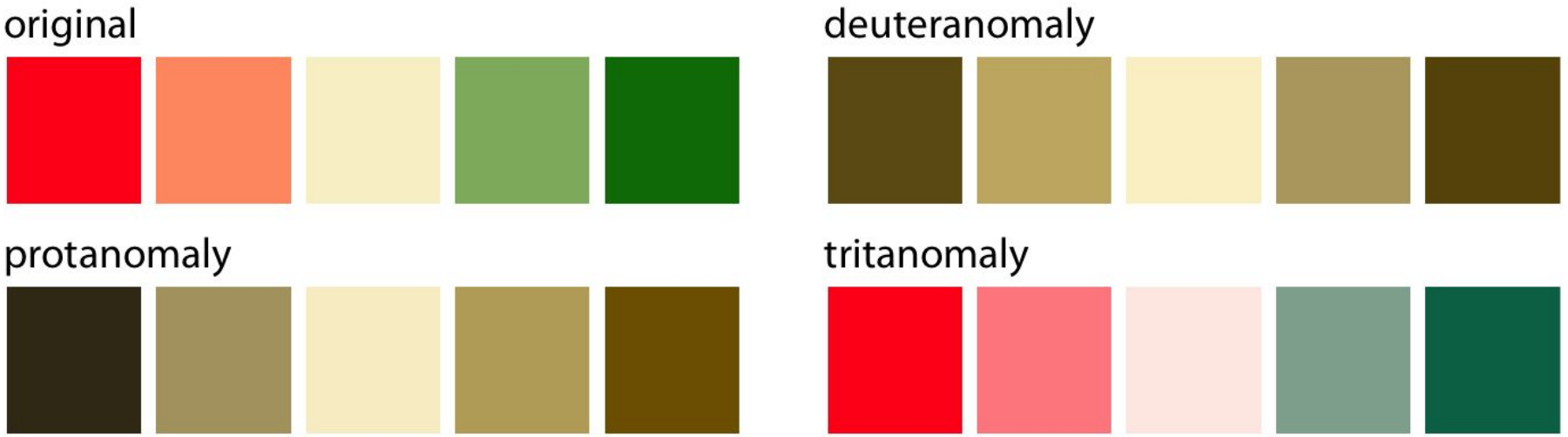


https://colorbrewer2.org



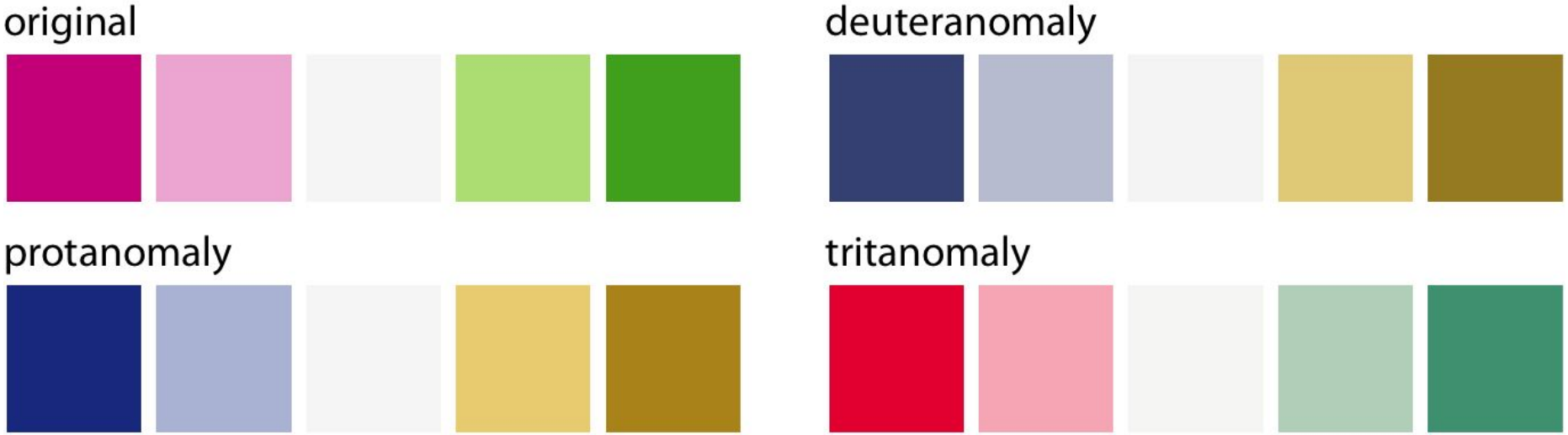
https://colorbrewer2.org

Consider colorblindness



A red–green contrast becomes indistinguishable under red–green color vision deficiency (deuteranomaly or protanomaly) From Wilke (2019)

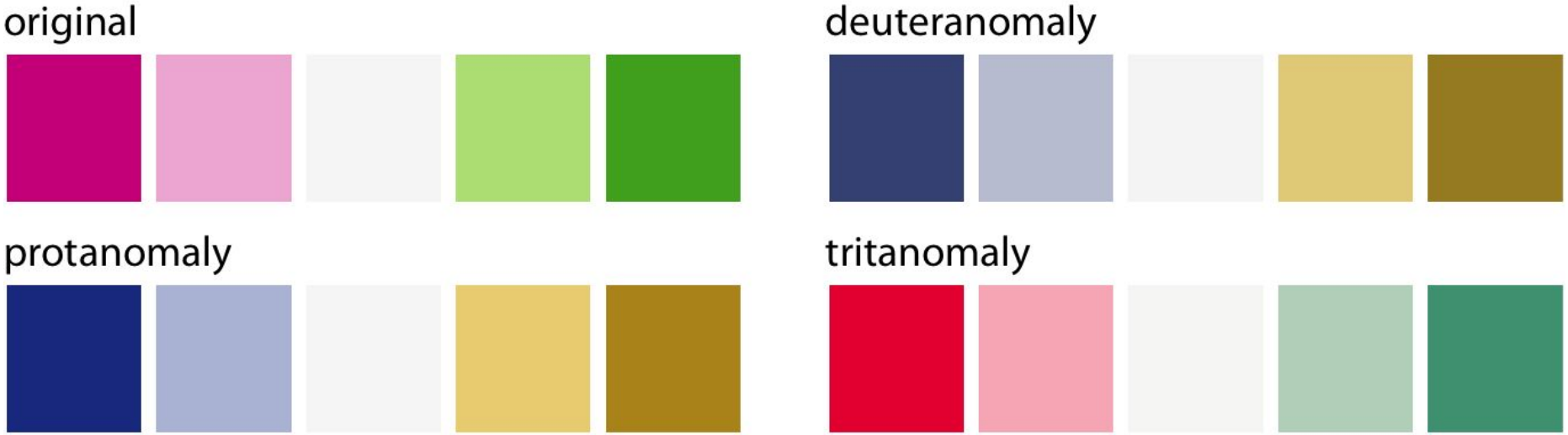
Consider colorblindness



The ColorBrewer PiYG (pink to yellow-green) scale looks like a red–green contrast to people with regular color vision but works for all forms of color-vision deficiency.

From Wilke (2019)

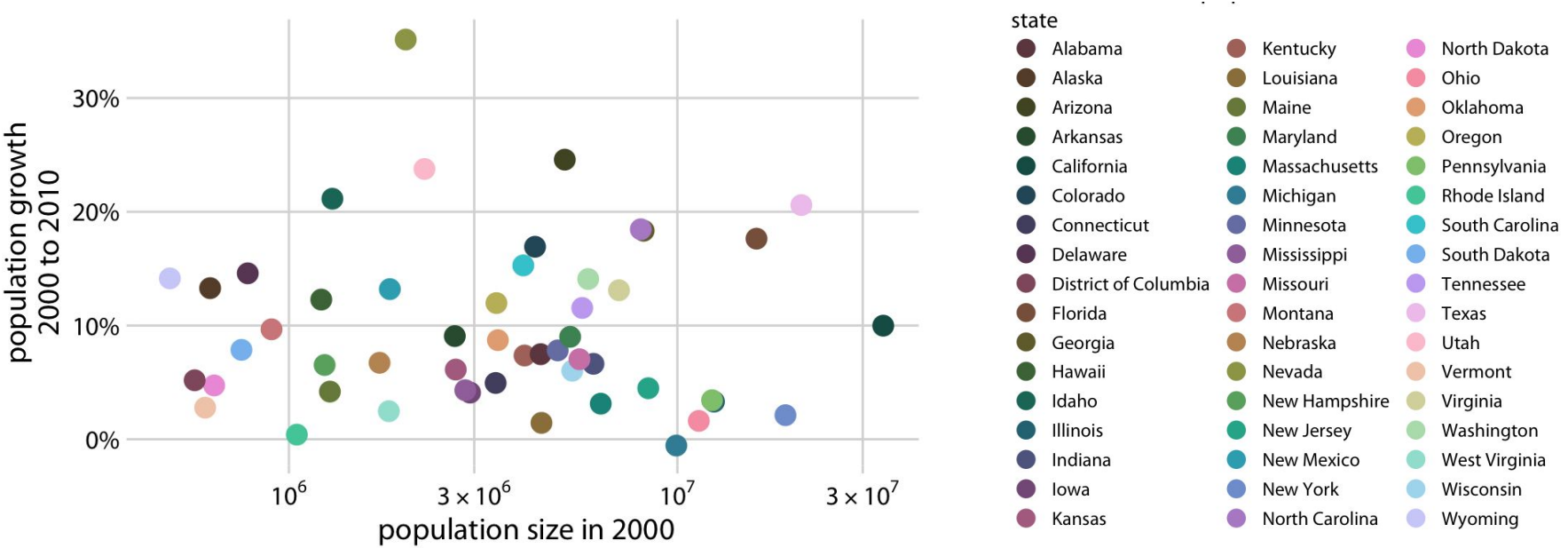
Consider colorblindness



The ColorBrewer PiYG (pink to yellow-green) scale looks like a red–green contrast to people with regular color vision but works for all forms of color-vision deficiency.

From Wilke (2019)

Common pitfall: encoding too much information



Claus O. Wilke (2019)

Common pitfall: using the wrong color scale

rainbow scale

rainbow converted to grayscale

The jet/rainbow color scale is **NOT a sequential** colormap,

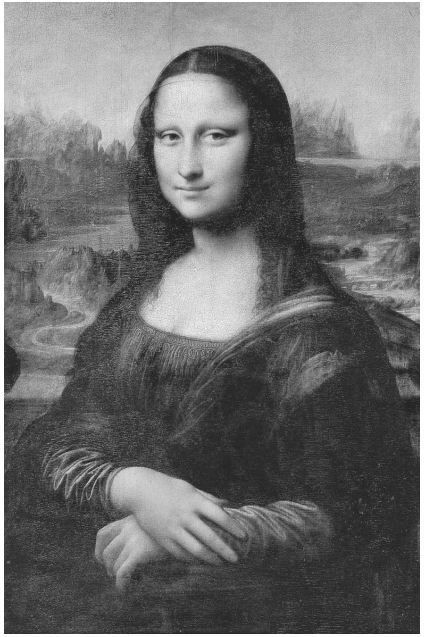
→ our perception of it is NOT linear but **circular!**

So do not use it for data that is not circular.

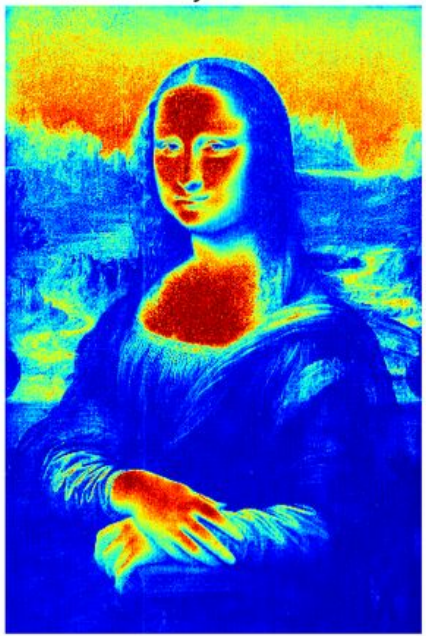
Common pitfall: using the wrong color scale

Colour

Grey



Jet



Viridis

Exercise 2: which visualization should I use?

- Work in paris, develop in your fork
- Do only 1 exercise (A - F)
- Goal: a visualization that is *publication-ready*
- Do a Merge request when ready.
- We'll review your visualizations together and comment them together

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

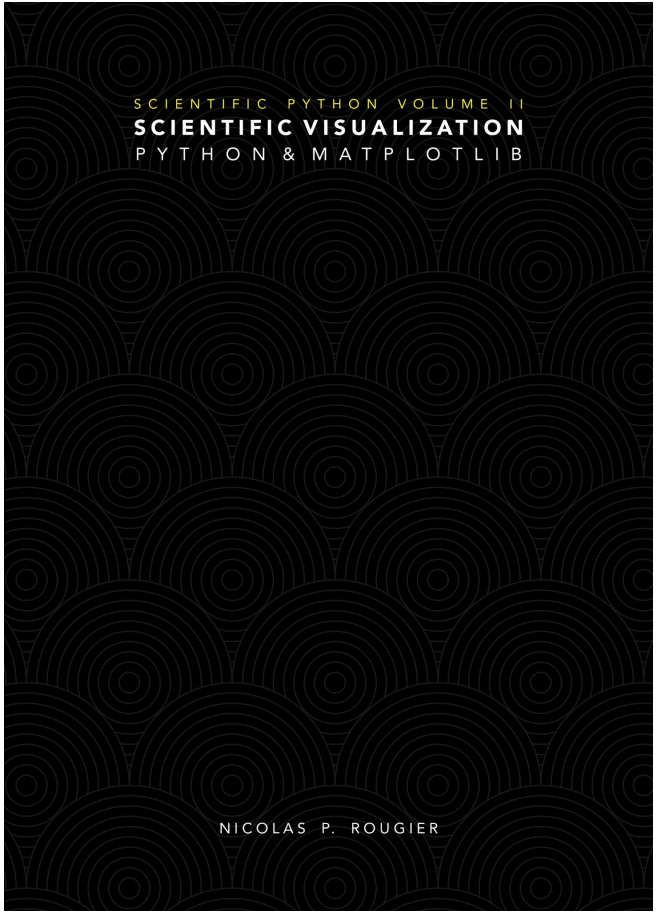
At the conceptual level :

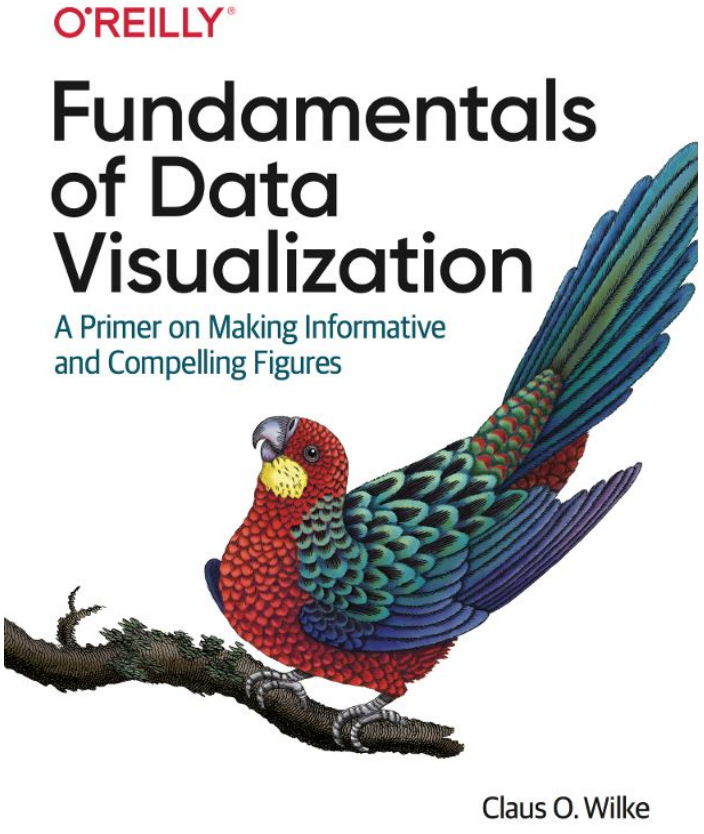
- **Ten simple rules for better figures**, Nicolas P. Rougier, Michael Droettboom, Philip E. Bourne (2014)
- **Fundamentals of Data Visualization**, book by Claus O. Wilke (2019)
- **Chart Suggestions - a though-starter** by A. Abelas.
- **Data Visualization Catalogue**
- **Edward Tufte**'s series of 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

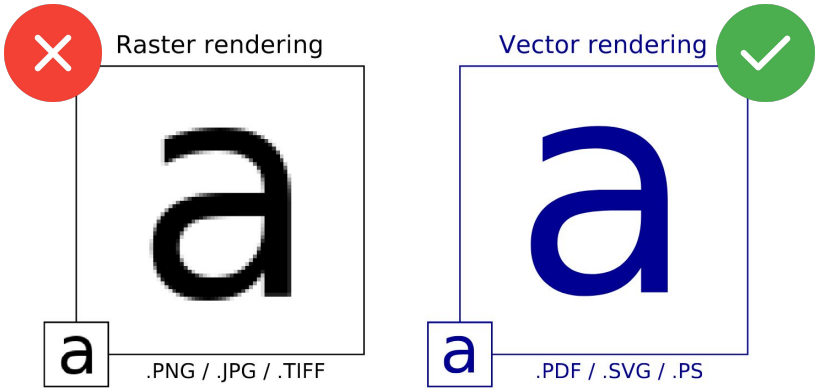


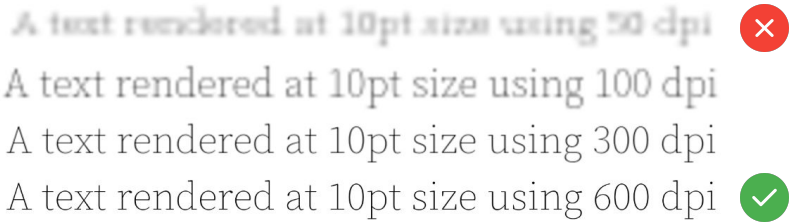


Some extra tips

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)

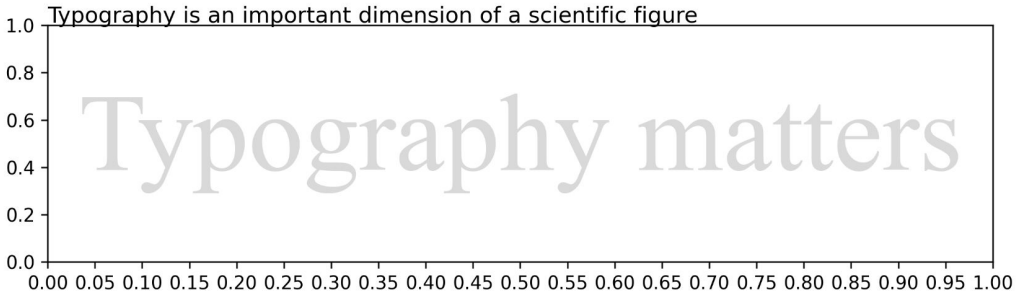
Vector formats

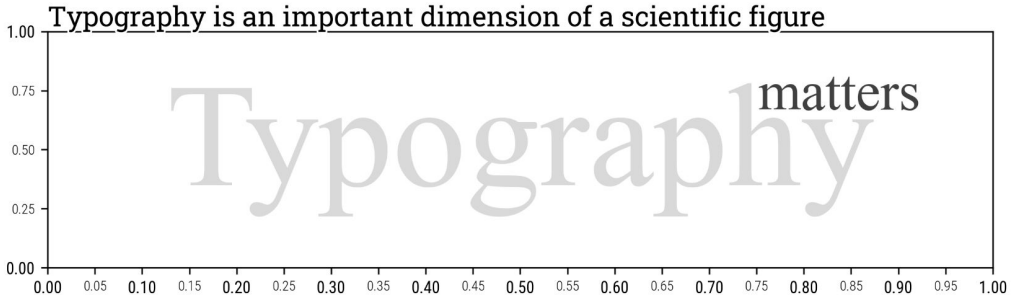
PDF: Portable Document Format
SVG: Scalable Vector Graphics

Text rendered in matplotlib and saved using different dpi

Font stack choice

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





Serif
DejaVuSerif.ttf

Sans
DejaVuSans.ttf

Serif
RobotoSlab-Regular.ttf

Sans
RobotoCondensed-Regular.ttf

Serif
SourceSerifPro-Regular.otf

Sans
SourceSansPro-Regular.ttf

Monospace
DejaVuSansMono.ttf

Cursive
Apple Chancery.ttf

Monospace
RobotoMono-Regular.ttf

Cursive
Merienda-Regular.ttf

Monospace
SourceCodePro-Regular.ttf

Cursive
ITC Zapf Chancery.ttf