The grammar of graphs

Best practices on data visualization with examples

Maurício Moreira-Soares - November 3rd 2022
Oslo Centre for Biostatistics and Epidemiology, University of Oslo

The goal of this presentation

- Provide basic guidelines and discuss some pitfalls on data visualization
- Introduce data visualization with **R** and **ggplot2**
- This presentation is based on the paper:

  Ten simple rules for better figures, 10.1371/journal.pcbi.1003833

Think before you plot

- We use graphs to communicate **messages**

Think before you plot

- We use graphs to communicate **messages**
- If your message is not clearly defined it is even difficult to choose a plot

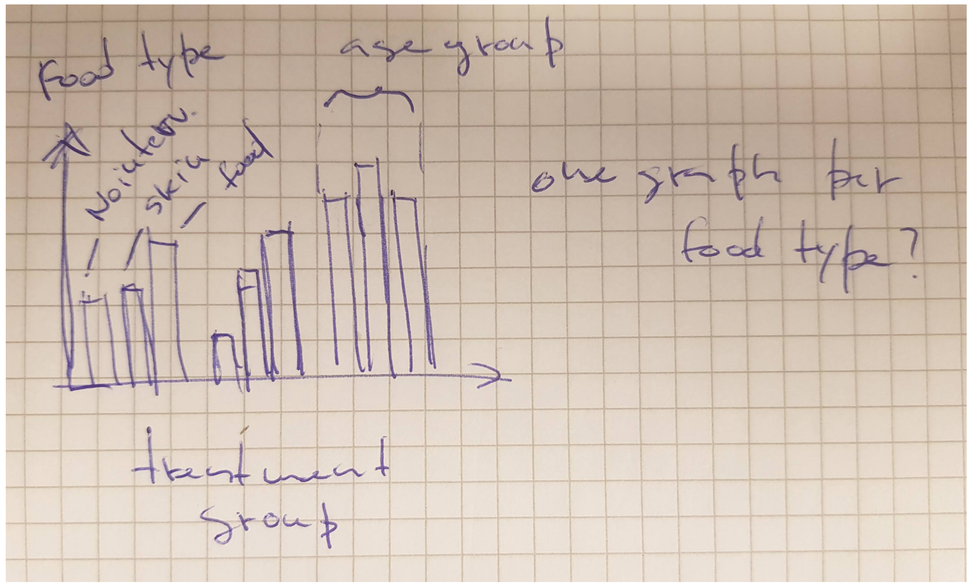
Think before you plot

- We use graphs to communicate **messages**
- If your message is not clearly defined it is even difficult to choose a plot

1. What is the underlying message you wish to deliver?
2. How can a figure best express this message?

Sketch

- Making good graphs can take a lot of time of tweaking in a software
- Pen and paper is key to prevent losing time on graphs that will not be used
- It also helps to identify your message

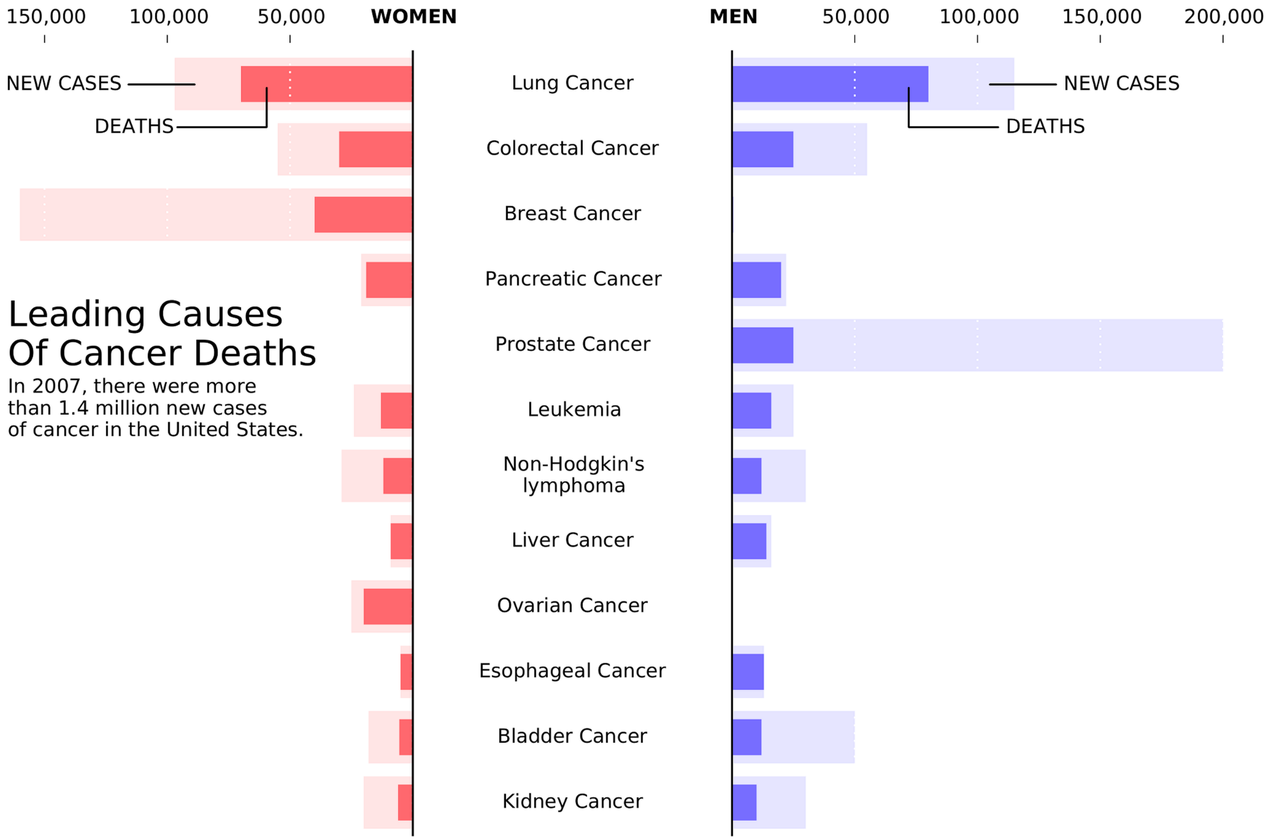


Know your audience

- To deliver an effective message you should know your **audience**
  - What is important for them? What is their motivation?
  - What you need them to do?
  - What data will help you make your point?

Know your audience

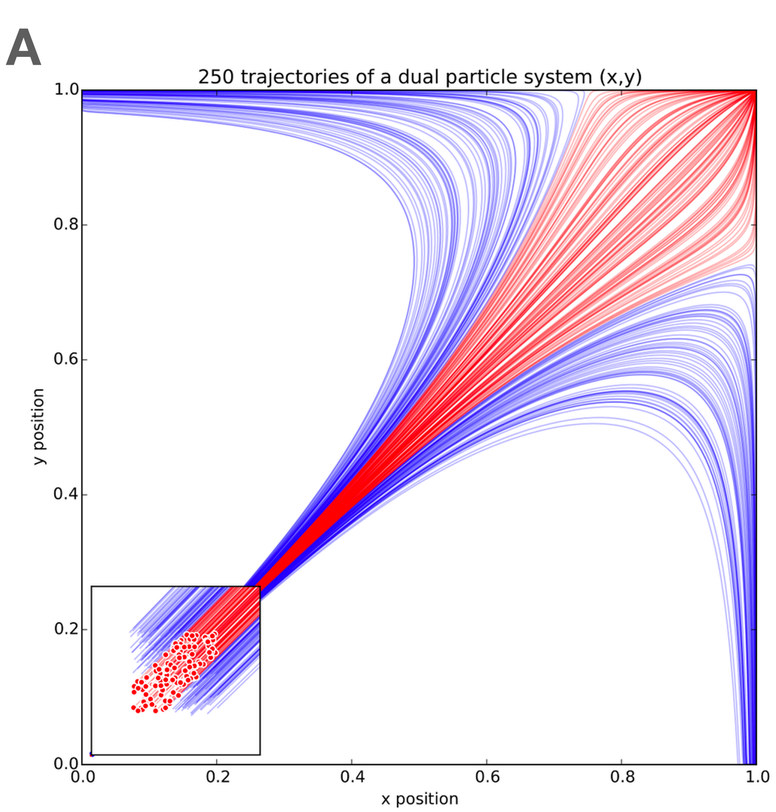
- Figure published in NY Times
- Why not classical double column?
- Could it be used in a sci. journal?

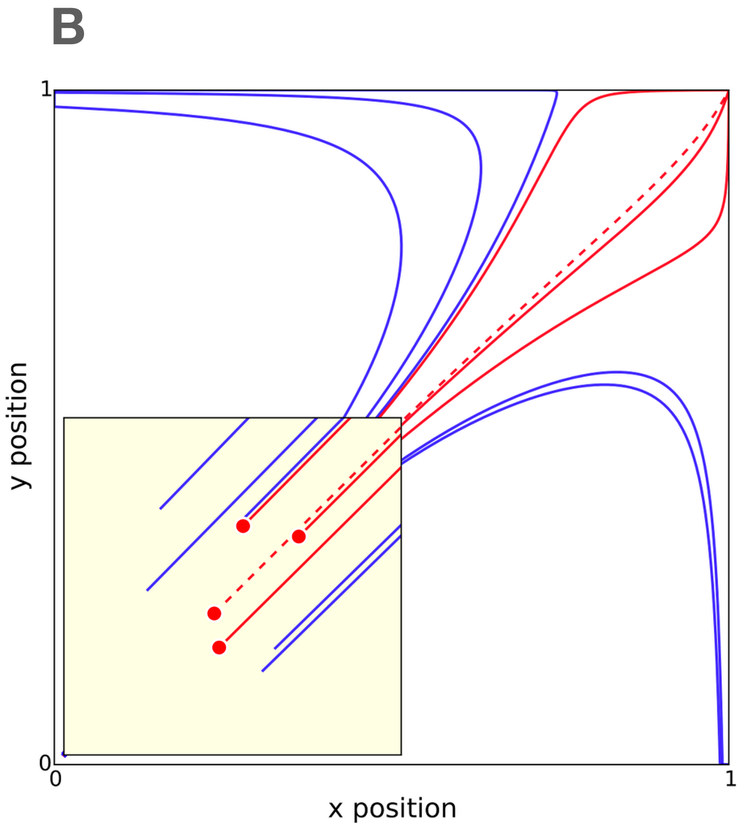


10.1371/journal.pcbi.1003833

Adapt the figure to the support medium

- Which figure would you use in a journal? And for a presentation?





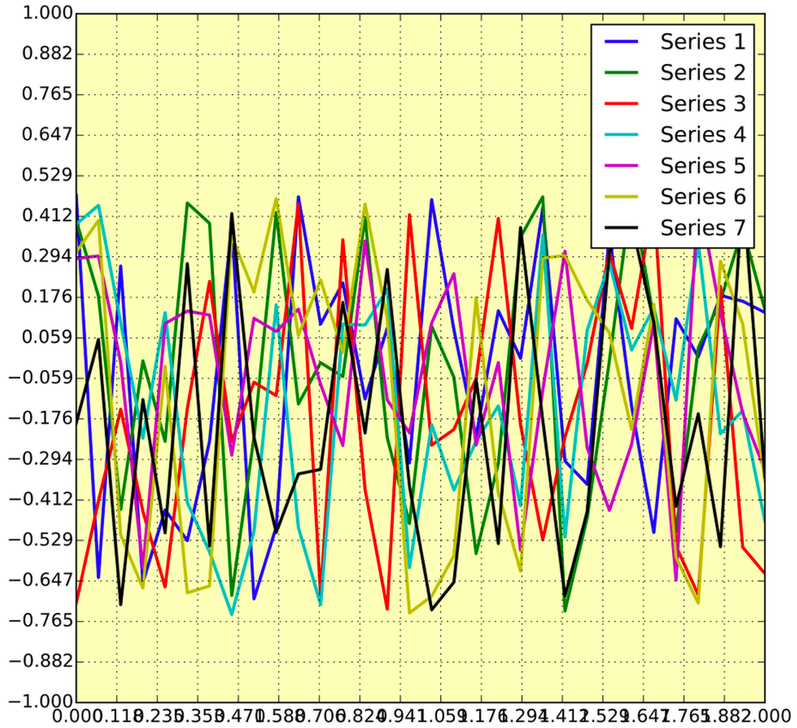
10.1371/journal.pcbi.1003833

Avoid chart junk

- Everything in your figure should be used to improve the message
- Some visual elements are unnecessary and might add confusion

Avoid chart junk

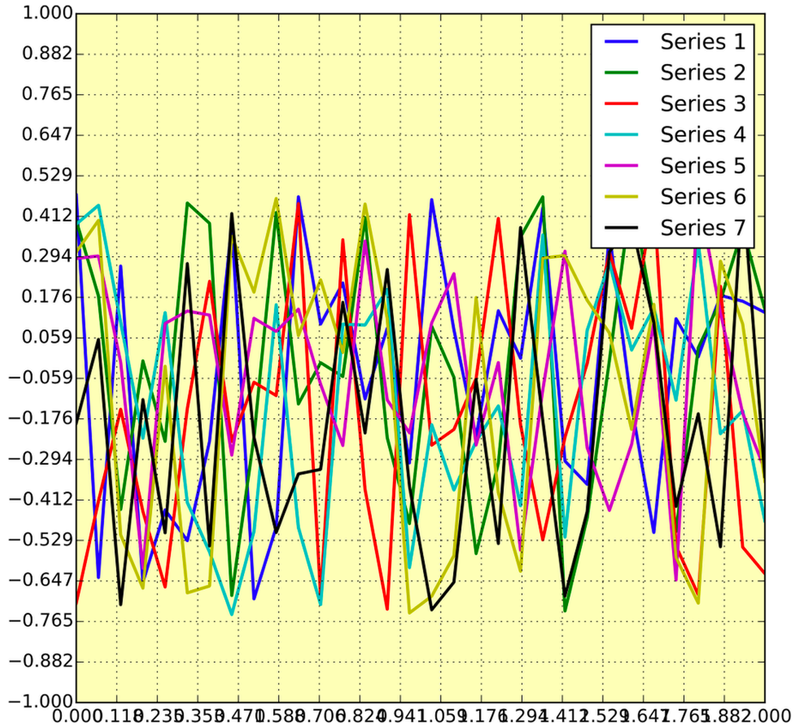
- Can you identify the message in this graph?



10.1371/journal.pcbi.1003833

Avoid chart junk

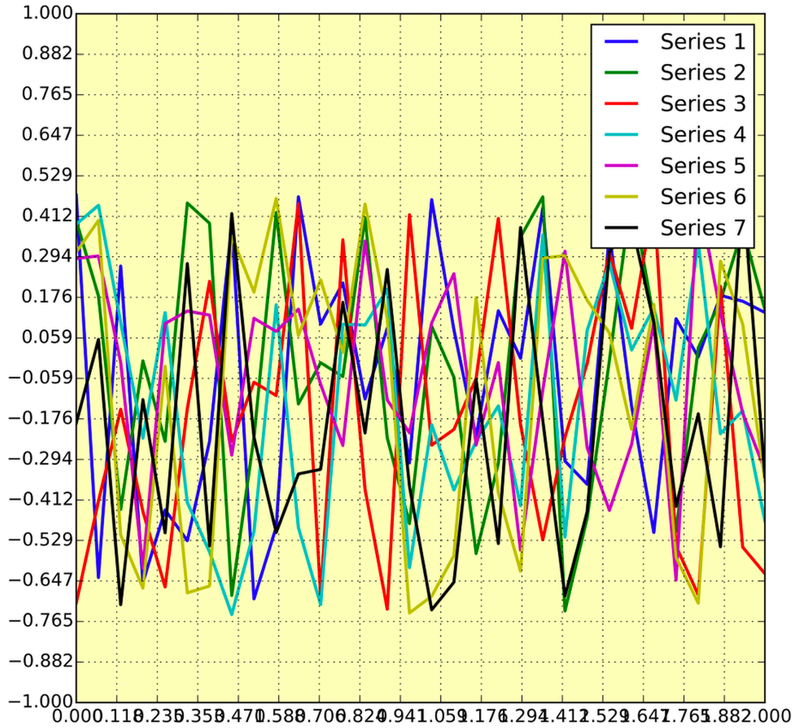
- How it can be improved?

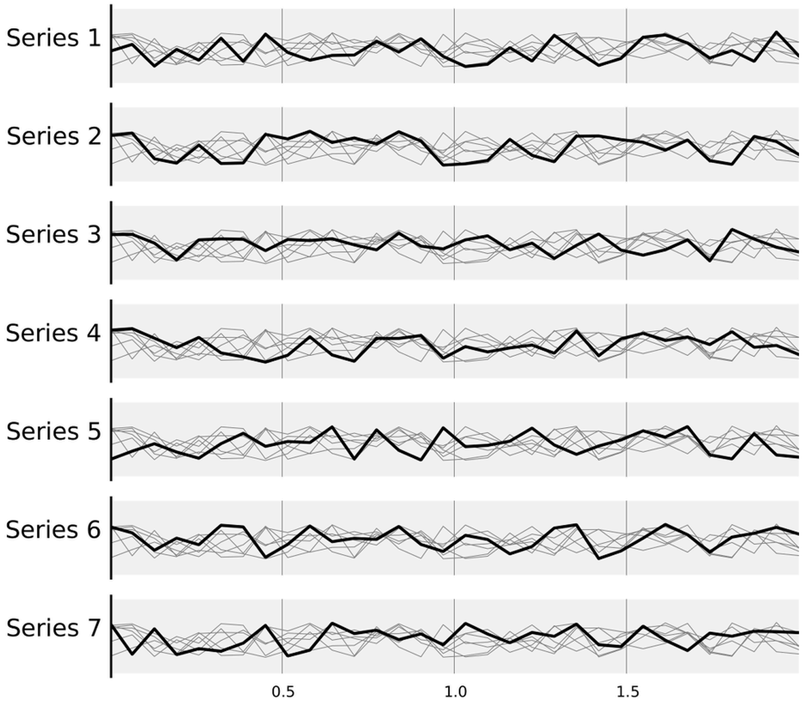


10.1371/journal.pcbi.1003833

Avoid chart junk

- Better?

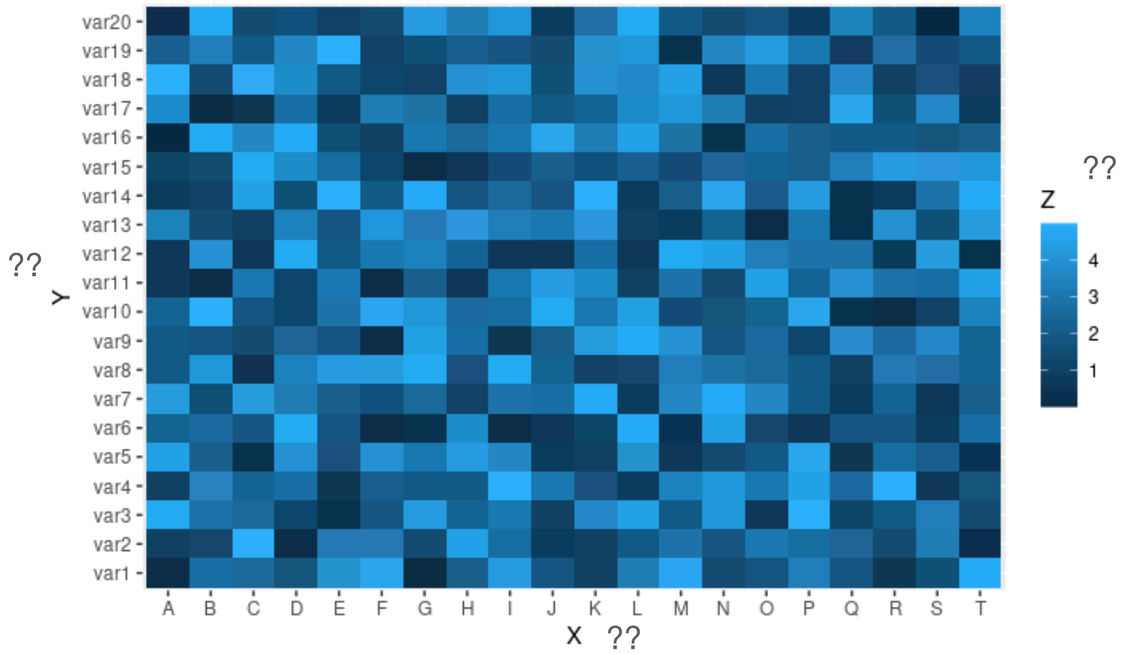




10.1371/journal.pcbi.1003833

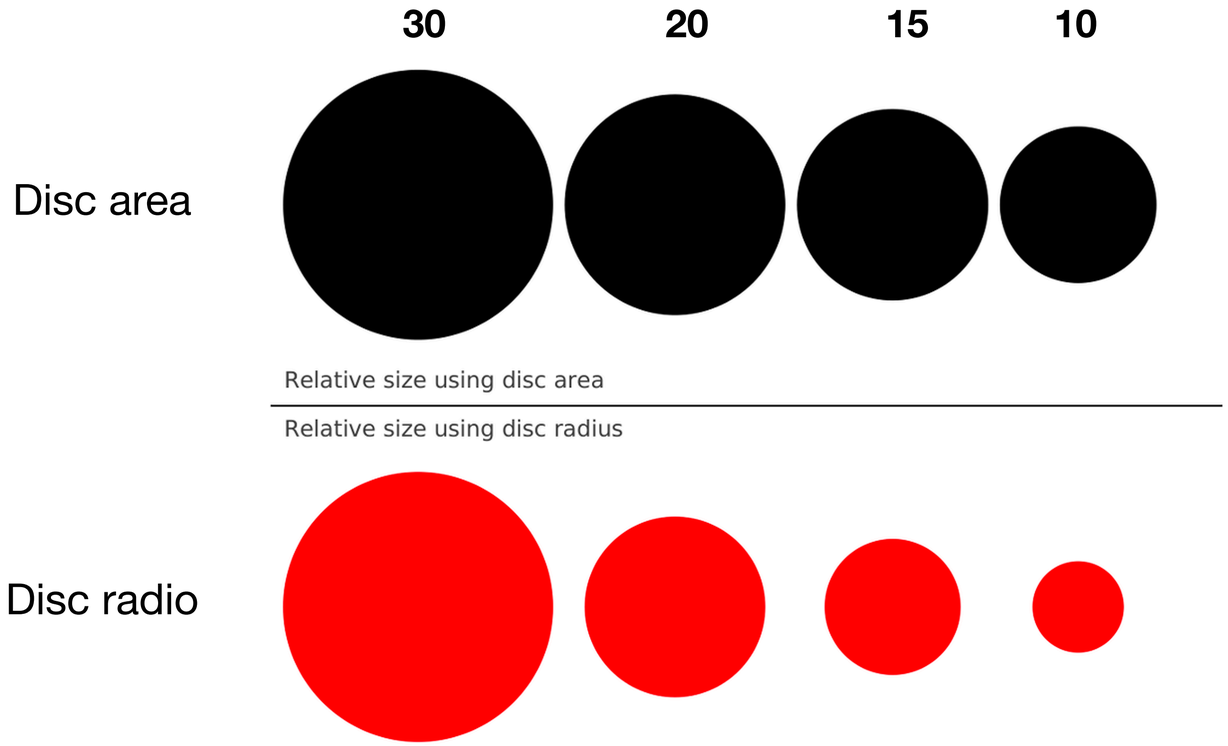
Captions and labels are not optional

- It is not possible to explain everything in the figure
- Label should be meaningful



Do not mislead your audience

- Scientific communication must be honest and transparent



Do not mislead your audience

- Do you think that there is a significant difference between the bars?

10.1371/journal.pcbi.1003833

Do not mislead your audience

- And now?

Full range: 0-100

Relative size using full range

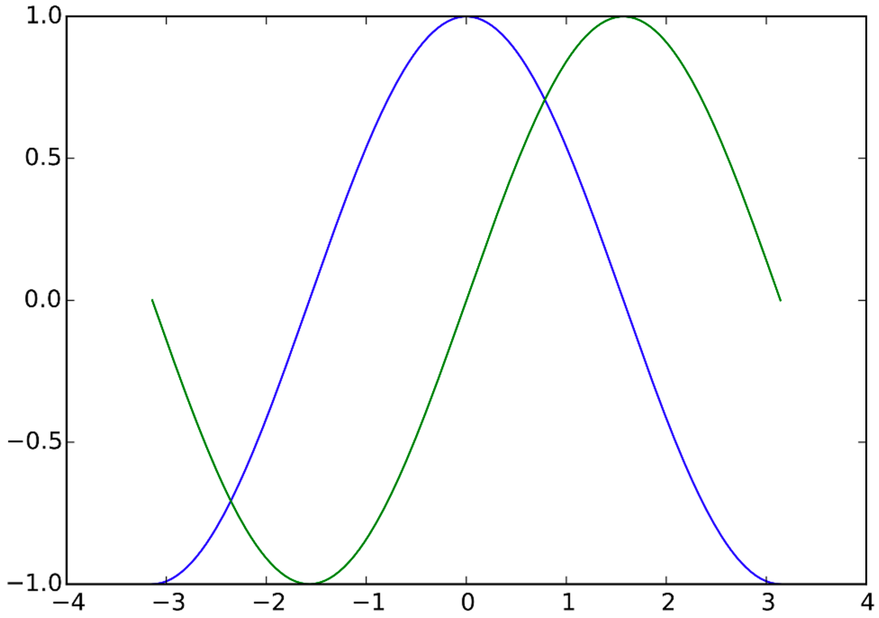
Relative size using partial range

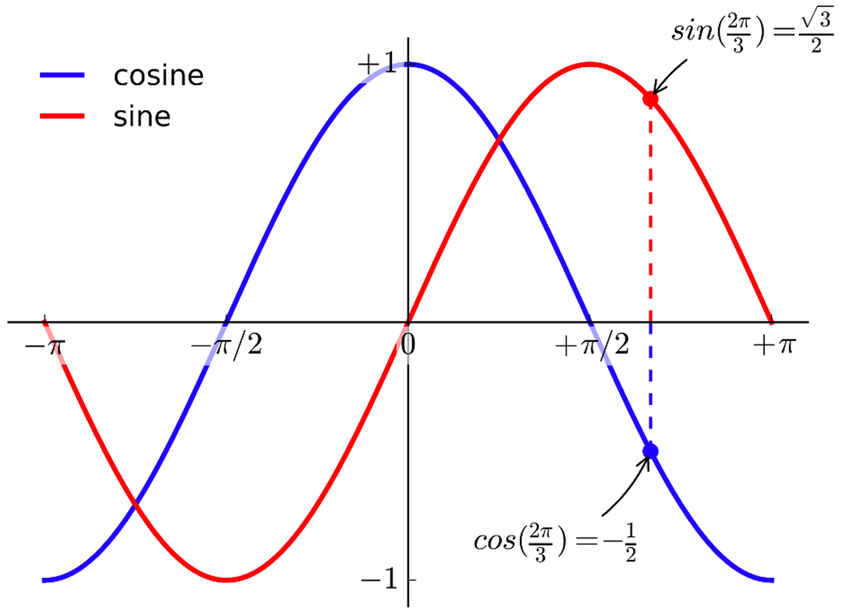
Partial range: 80-100

10.1371/journal.pcbi.1003833

Do not trust the defaults

- The previous misleading settings are often offered as defaults





Color matters

- 300 Million people around the world are color blind

Green-type colorblindness (common)

Color matters

- 300 Million people around the world are color blind
- Data visualization libraries embraced the challenge of more inclusive palettes

Green-type colorblindness (common)

Color matters

- 300 Million people around the world are color blind
- Data visualization libraries embraced the challenge of more inclusive palettes
- Color independent figures are ideal

Green-type colorblindness (common)

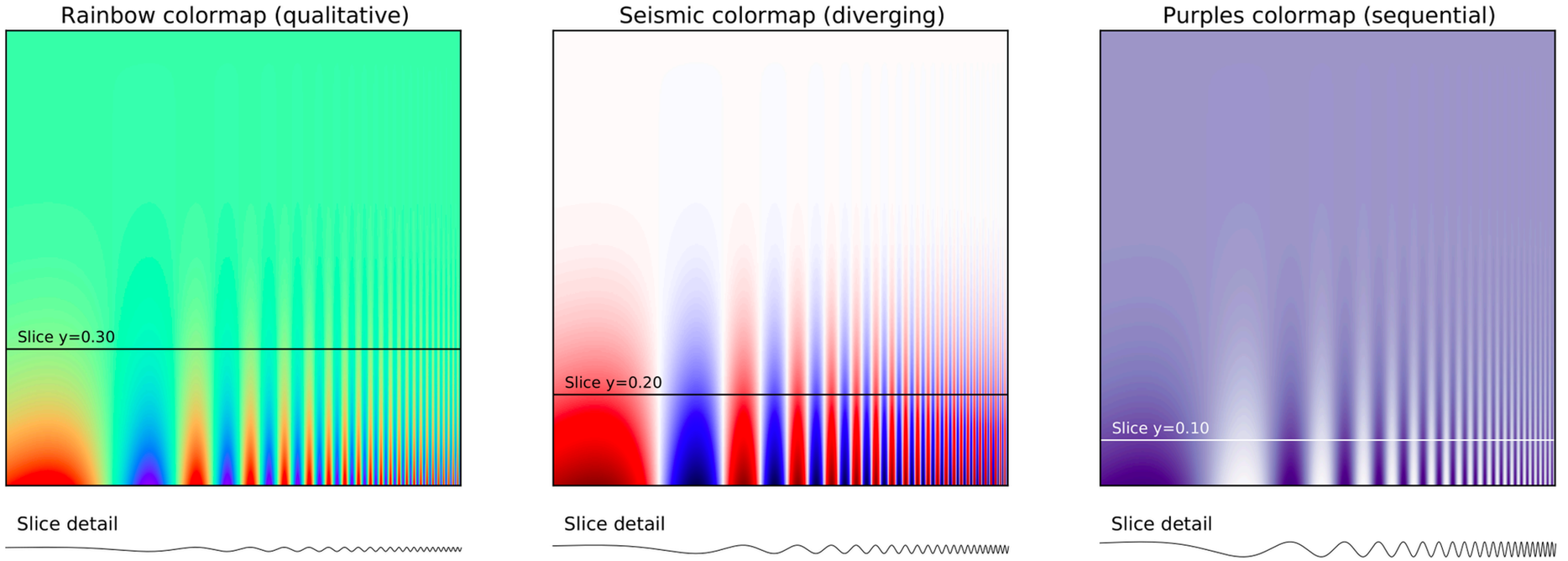
Color matters

- 300 Million people around the world are color blind
- Data visualization libraries embraced the challenge of more inclusive palettes
- Color independent figures are ideal
- You should have a reason to use colors

Green-type colorblindness (common)

Use color effectively

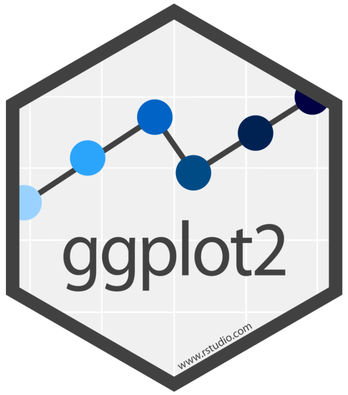
- The rainbow and seismic colormaps are equally bad
- They hide details in higher frequencies (and they are not color blind friendly)



10.1371/journal.pcbi.1003833

Choose the right tool

- The right tool is the one you know best and it works
- But sometimes you need to learn something new
- We will give a brief introduction to ggplot2



Quick start on ggplot2

- All plots start with the function **ggplot()** which receives the data
  - ggplot(data = df)
- Then we **add** components to the graph using the sign `+`
  - ggplot(data = df) + geom_line(aes(x = time, y = v))

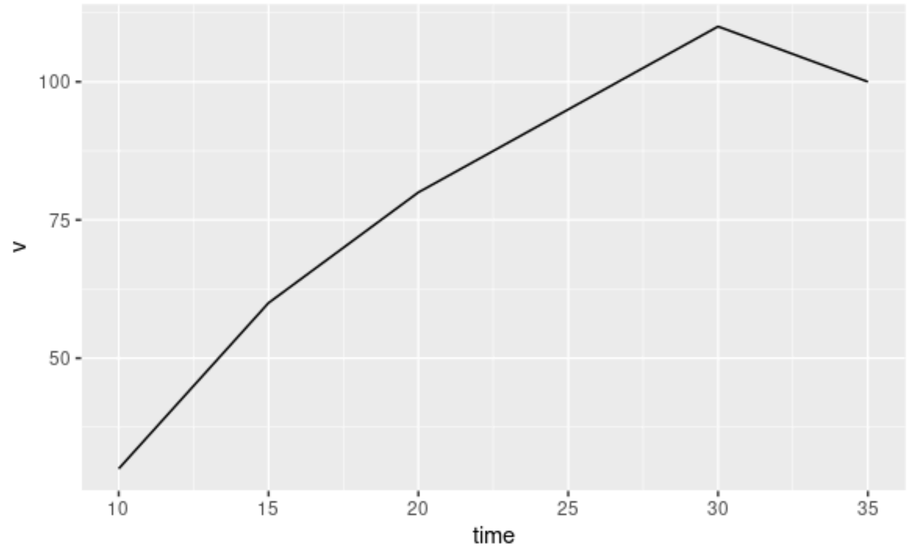
Quick start on ggplot2

- All graph types start with **geom_**
  - geom_line(), geom_bar(), geom_hist(), geom_smooth(), …
- They receive aesthetic mappings with the function **aes()**
  - **geom_line(aes(x = time, y = v))**

Line plots

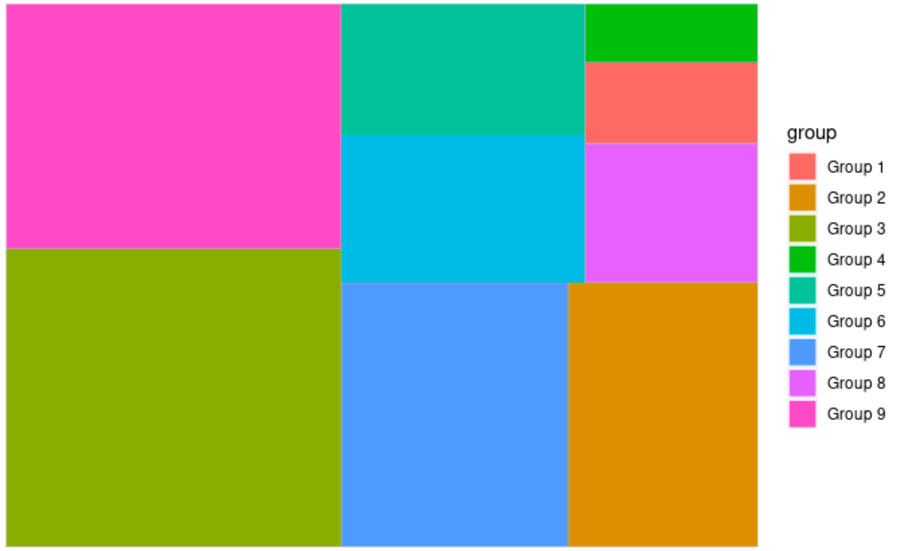
- Useful for time dependent variables

  - ggplot(data = df) + geom_line(aes(x = time, y = v))



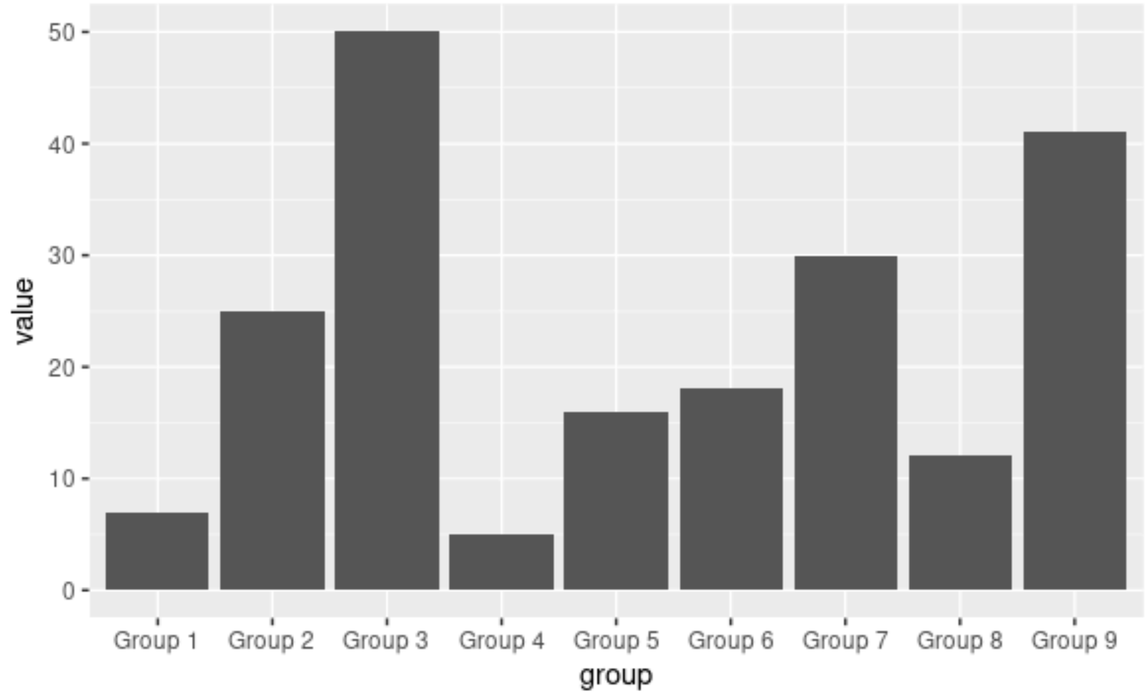
Treemap plot

- I reminder that humans are bad at comparing areas
  - ggplot(data = df) + geom_treemap(aes(area = value, fill = group))



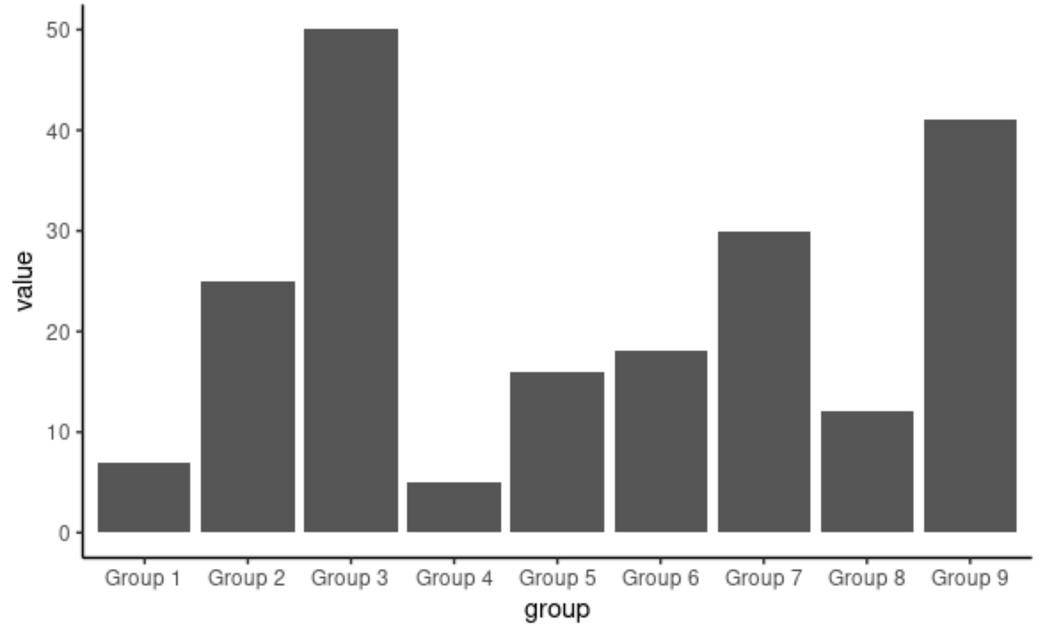
Bar plot

- A better representation to compare groups
  - ggplot(data=df) + geom_bar(aes(x=group, y = value), stat = “identity")



Improve

- With ggplot you can add elements that will improve the visualization
- ggplot(data=df) + geom_bar(aes(x=group, y = value), stat = "identity") + **theme_classic()**



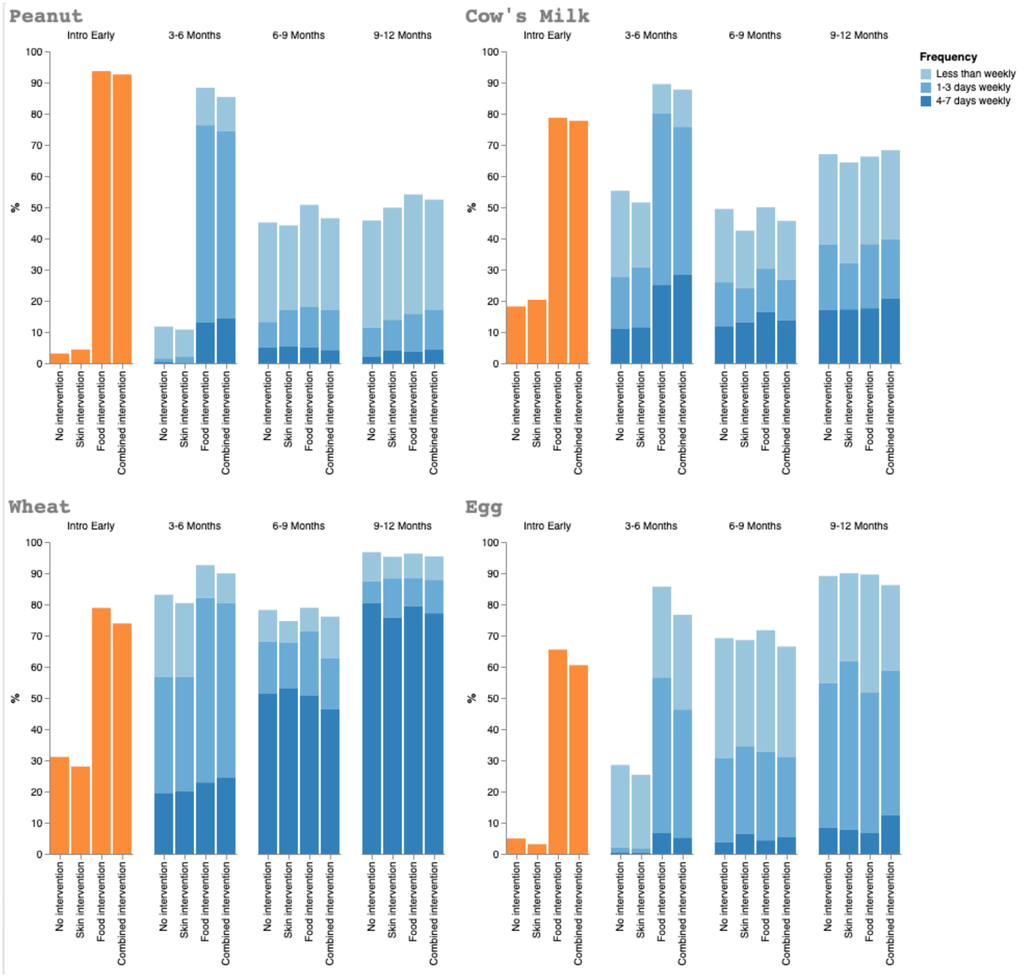
Choose the right format

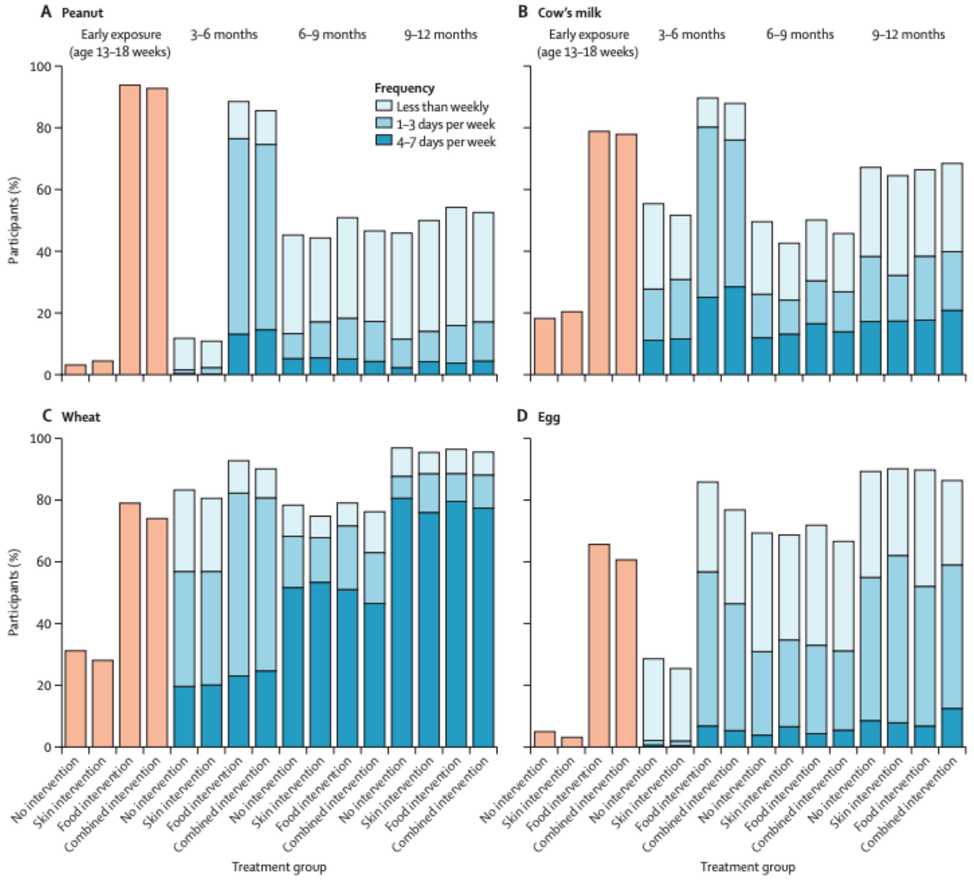
Pixel based vs vectorial

- Vectorial formats are scalable: you can resize figures without loosing quality
  - PDF and SVG are vectorial formats while JPG is pixel based
  - Vectorial formats are preferred by publishers

Choose the right format

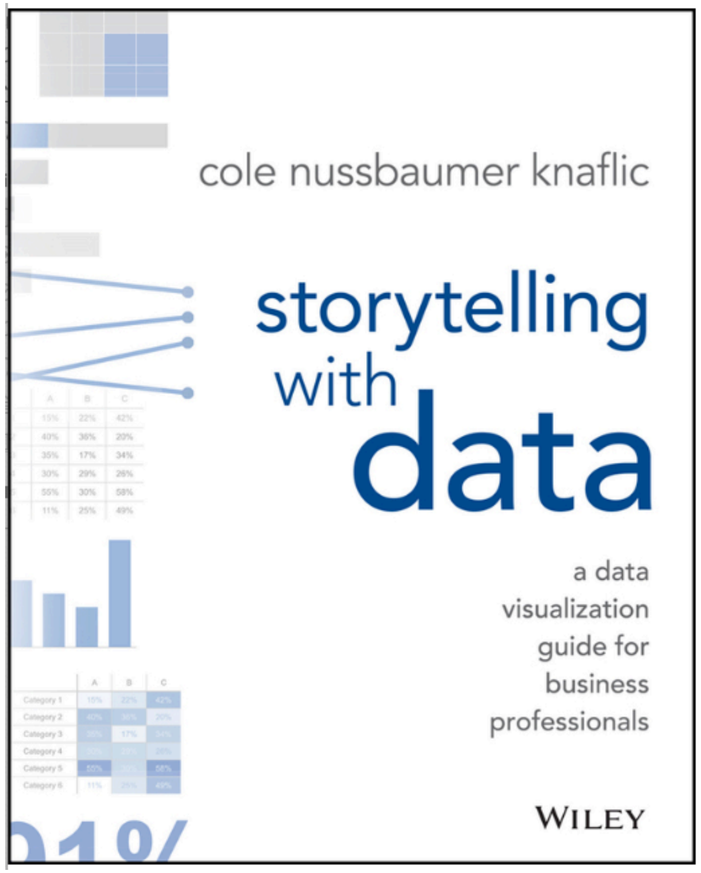
Pixel based vs vectorial





Resources

- For inspiration read “Storytelling with data”
- They provide excellent free resources at https://www.storytellingwithdata.com/



Remember

- The two most important principles are knowing your **audience**
- And defining a clear **message** to convey with data visualization

Thank you!