

## Data visualisation: Introduction and motivation

By Radovan Bast, UiT The Arctic University of Norway

ORCID: 0000-0002-7658-1847

Date: October 2020



Manuscript:

In this video I will introduce and motivate data visualisation, and define the goals for this course. We will focus on data visualisation for publications and presentations in the academic context, but not only there. Data visualisation courses are sometimes very general, sometimes very specific. We will try a middle ground: We will try to focus on practical recommendations with an emphasis on reproducibility. We will see that for reproducibility, we will probably need to learn some scripting in Python or R. But our ambition will not be to turn this into a programming course and we will not focus on programming languages. Instead I will try to bring you on a good track to show you which tools exist and provide hopefully useful starting points for you to explore.

There are many excellent books on the topic, and here I only list three, and I particularly wish to highlight the first one "Fundamentals of data visualization" by Claus Wilke which I really enjoyed and I recommend to everybody.

I encourage you to also explore more resources which you can find in the accompanying slides: An excellent paper and other presentation slides. Check out also the #tidytuesday hashtag on Twitter where people share their challenges and plots.

I really like this quote from Claus Wilke's book: "One thing I have learned over the years is that automation is your friend." I think figures should be autogenerated as part of the data analysis pipeline (which should also be automated), and they should come out of the pipeline ready to be sent to the printer. There should be no manual post-processing needed. Because all of us have been or will get into a situation where we need to adjust figures a few days before the deadline for a manuscript or the PhD thesis.

And this is one of the two take-home messages: Avoid tools which cannot be scripted. Stay away from tools where plots can only be modified manually by pointing and clicking. It may cause you last minute trouble, and it may cause trouble for the group leader or the next PhD candidate who needs to recreate or adapt the figures of a student who already left. Optimise for comprehension and accessibility. This means font-size, colours, suitable representation, title, and caption. We will discuss these in this data visualisation video series.

So why do we visualise data? A classic example to motivate data visualisation is Anscombe's quartet: Four data sets which look distinctively different (see the left panel), but all four plots have the same statistics. They have the same mean of  $x$  and  $y$ , same variance of  $x$  and  $y$ , same correlation between  $x$  and  $y$  and same linear regression.

This can be taken to the extreme with this nice, animated example: The animation shows different graphs, even a dinosaur, again with the same stats. If we only looked at the numbers (right panel), we would miss the pattern. So we almost always need to do both.

With data visualisation we get more insight into data and it is easier for us to see patterns and problems. Both calculations and graphs should be studied, and each will contribute to understanding. But it's not only about understanding – data visualisation allows us also to communicate insight to others and facilitate their understanding of our work in presentations, in publications, and when communicating research with the public. We also create typically plots because – well – our supervisors tell us to, and in this case we often copy existing style and culture. And that is good, but hopefully after this course you will be able to sometimes suggest a different way to visualise the data in your research group.

In other videos of this series we will discuss figure design and the design process. We will learn the vocabulary of visualisations, discuss which tools to choose, demonstrate reproducible plots, talk about data formats, and finally talk about how to arrange figures and tables for reports and publications.