

Developing Data Products

SC4125

Module 4

Visualization

Disclaimer & Acknowledgement

▶ Images and content from 3rd party sources have been used in this slide collection

In this deck of slides images from many sources have been used per what I believe fits under *fair use* doctrine. Nevertheless, if a copyright owner of any figure would like them to be removed, kindly contact me.

I have attributed the sources as best as I could. However, if there is any misattribution that ought to be rectified, please contact me.

Contact: Anwitaman@ntu.edu.sg

Data Visualization

- ▶ Various purposes of data visualization
 - Exploration of the data while cleaning it and creating your analytics pipeline
 - Communicate information and results
 - Dashboards & UIs (potentially incorporating interactivity)



“Learn the rules like a pro,
so you can break them like
an artist.”

Pablo Picasso

Some prominent influencers

- ▶ Edward Tufte: Considered as a pioneer in the field of data visualization
Reference book: **The visual display of quantitative information**



- ▶ Alberto Cairo: Popularized the idea that data visualization should be seen as “functional art” vs. fine art
Reference book: **The functional art**



The sketch of E. Tufte was illustrated by Merchant for the Brunswick Review, Image source: <https://washingtonmonthly.com/>
The sketch of A. Cairo is created using MS Office's artistic effect using image from: <https://com.miami.edu/profile/alberto-cairo/>

Conveying Information

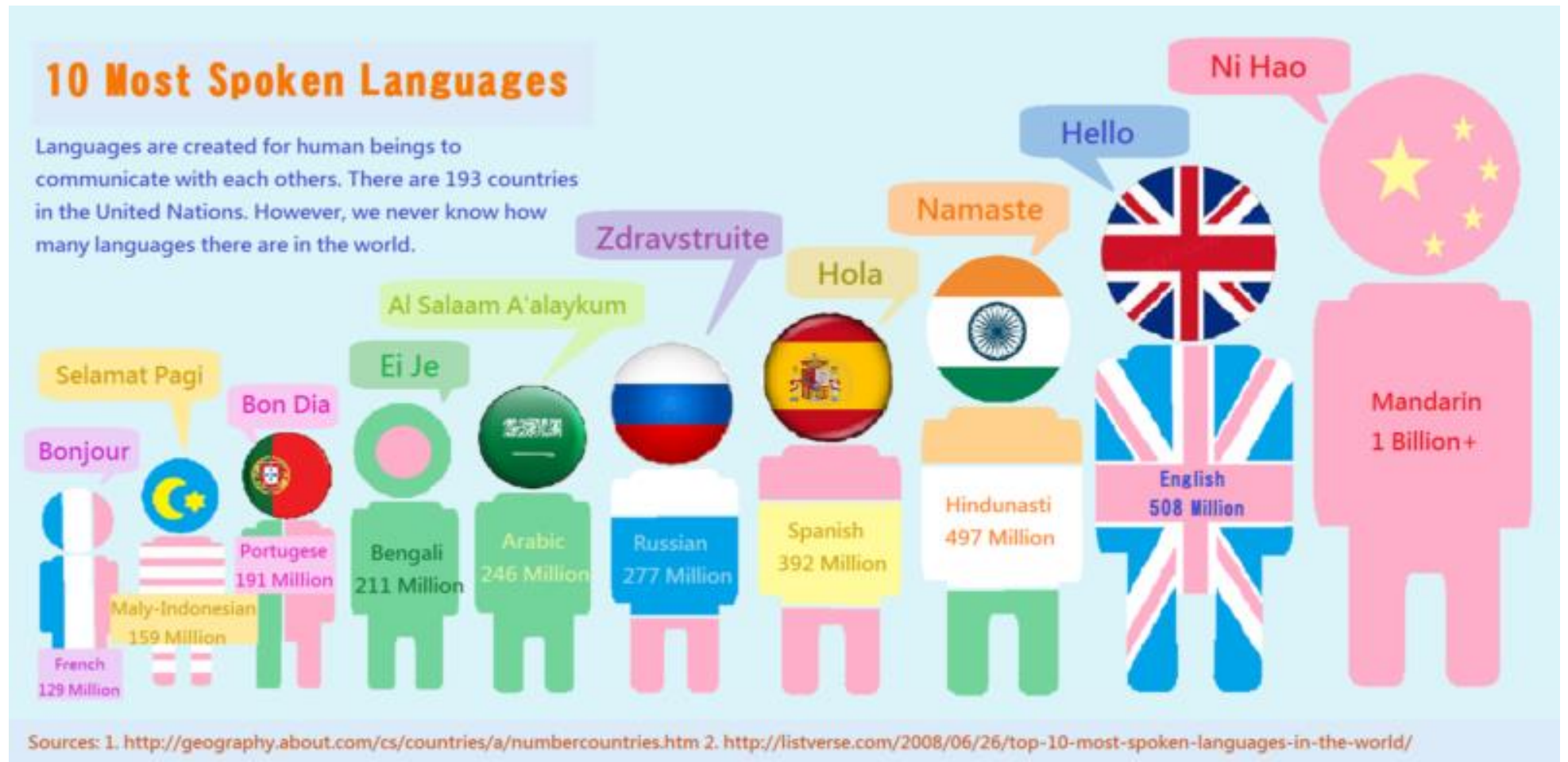
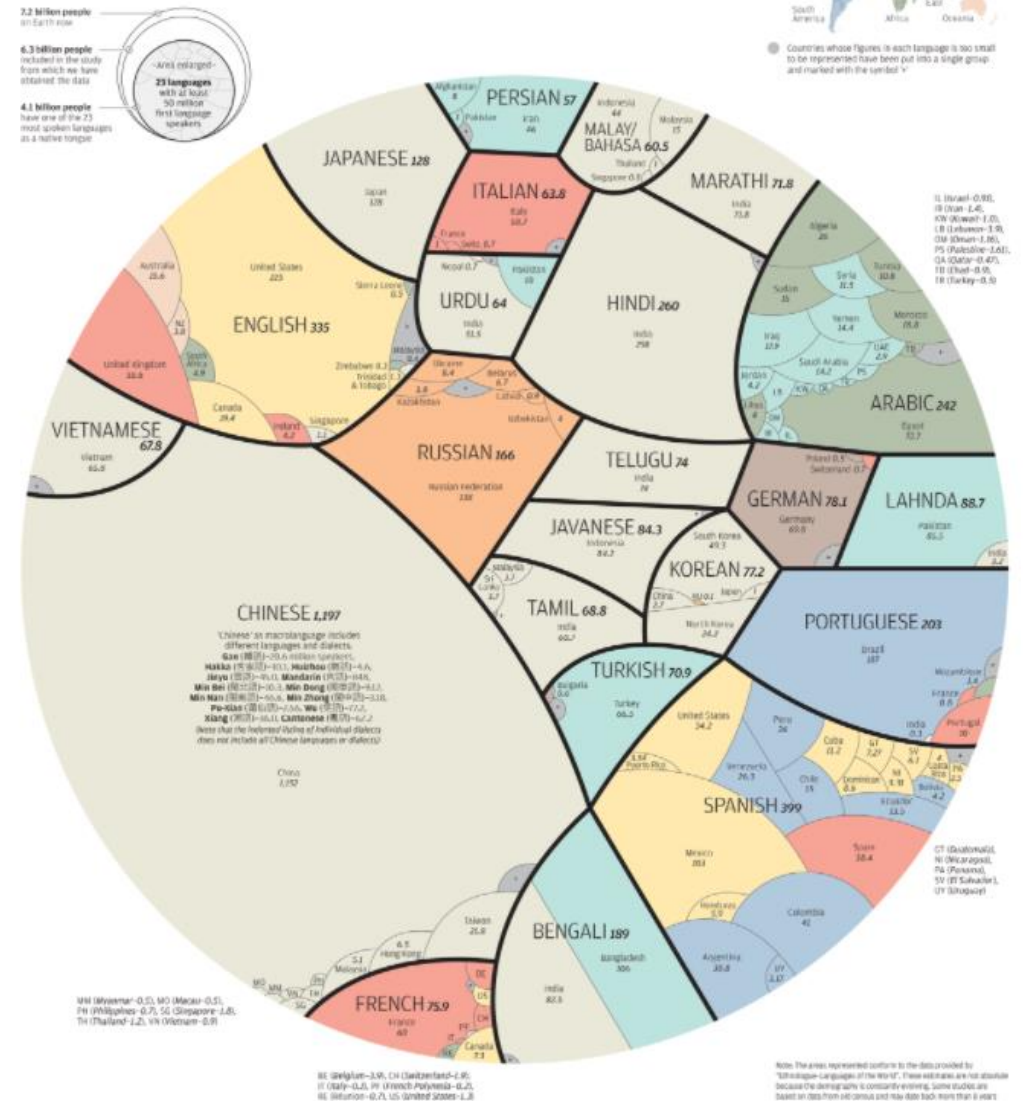


Image source: <https://www.plato-edu.com/>

Conveying Information

A world of languages

There are at least 7,102 known languages alive in the world today. Twenty-three of these languages are a mother tongue for more than 50 million people. The 23 languages make up the native tongue of 4.1 billion people. We represent each language within black borders and then provide the numbers of native speakers (in millions) by country. The colour of these countries shows how languages have taken root in many different regions.



Conveying Information

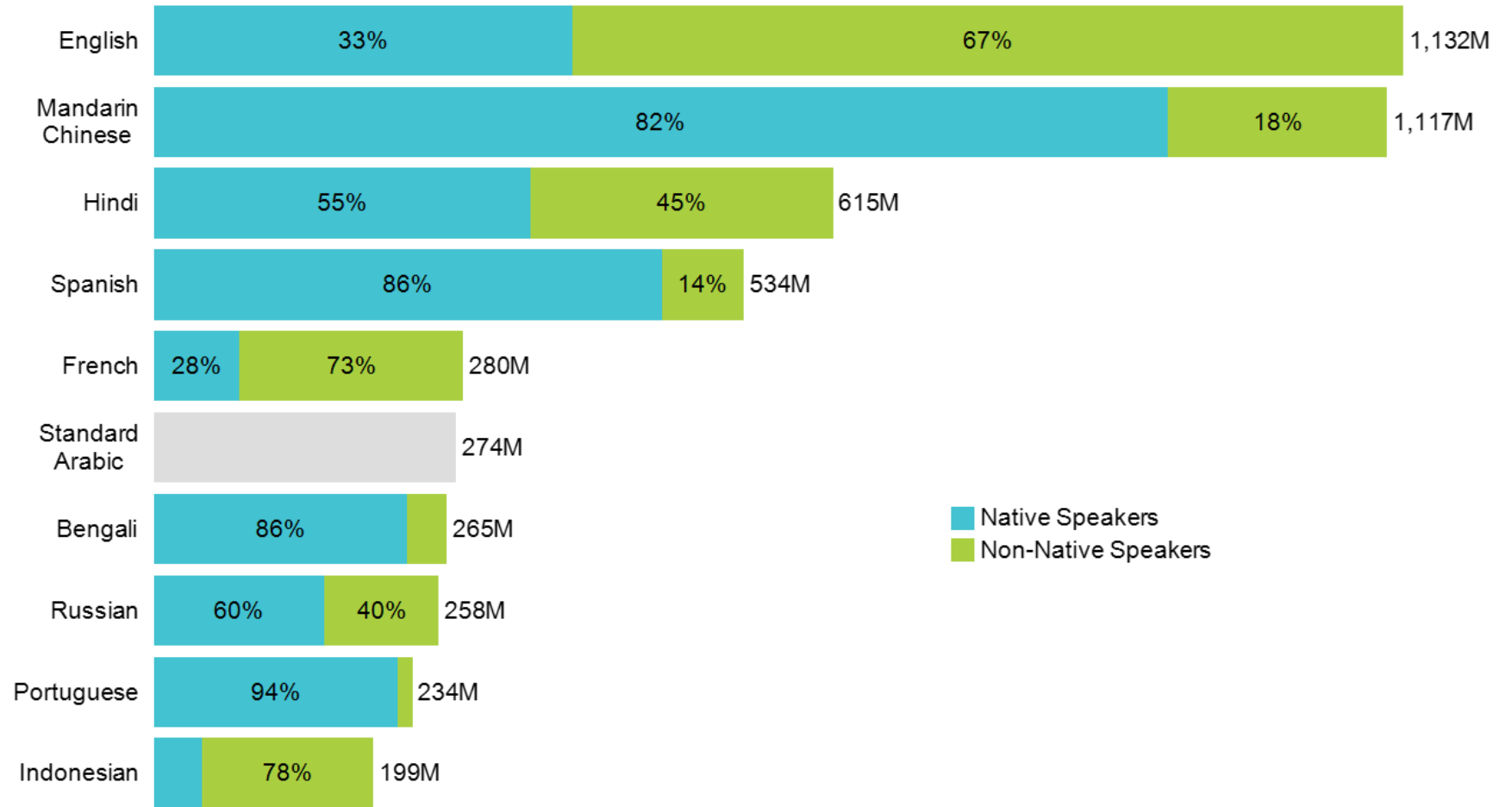


Image source: <https://www.mekkographics.com/10-most-spoken-languages/>

Conveying Information

▶ Chartjunk

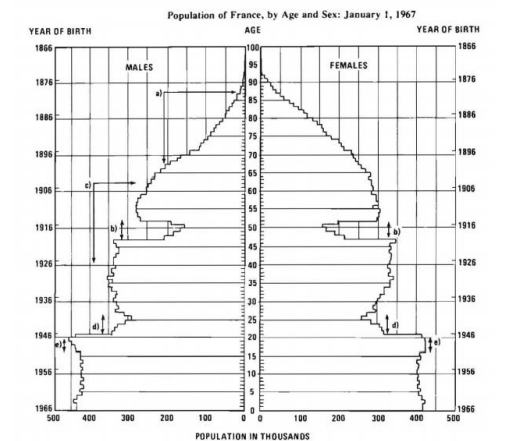
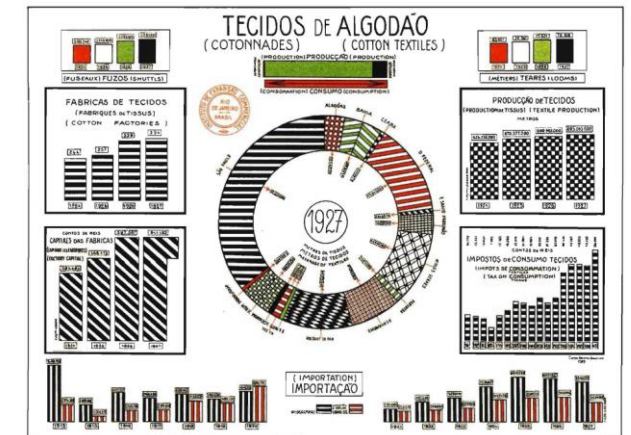
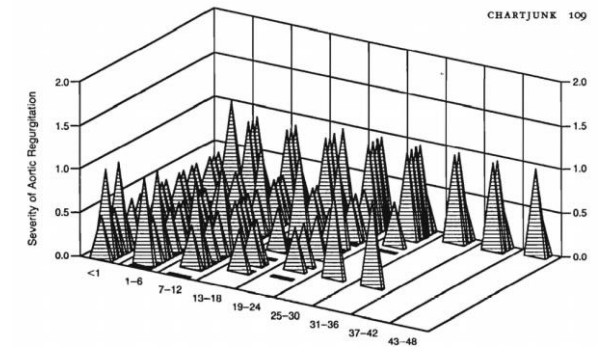
The interior decoration of graphics generates **a lot of ink** that does not tell the viewer anything new. The purpose of decoration varies—to make the graphic appear more scientific and precise, to enliven the display, to give the designer an opportunity to exercise artistic skills. Regardless of its cause, it is all non-data-ink or redundant data-ink, and it is often chartjunk. – **Edward Tufte**, The visual display of quantitative information

A plot with higher data-ink ratio

Chartjunk examples

Vibrations (e.g., Moiré effect), Grids, Ducks

Graphics do not become attractive and interesting through the addition of ornamental hatching and false perspective to a few bars. Chartjunk can turn bores into disasters, but it can never rescue a thin data set.



Source: The visual display of quantitative information by Edward Tufte

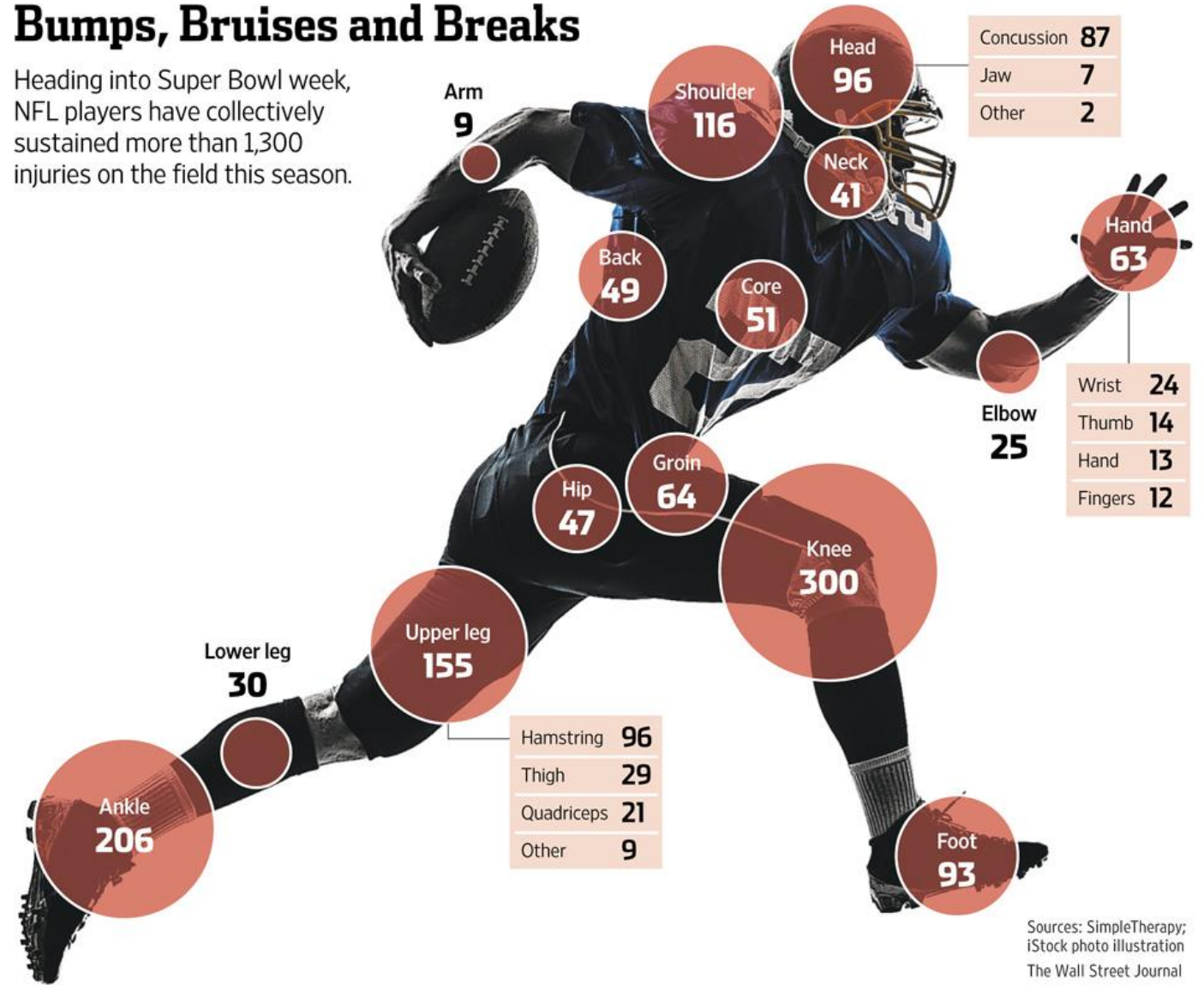
Conveying Information

▶ Chartjunk

- may sometimes work well if **done right?**

Bumps, Bruises and Breaks

Heading into Super Bowl week, NFL players have collectively sustained more than 1,300 injuries on the field this season.



Sources: SimpleTherapy;
iStock photo illustration
The Wall Street Journal

Image source:

<https://www.wsj.com/articles/SB10001424052702303277704579344753369526502>

Displaying Information

Chartjunk

- may sometimes be useful if **done right?**

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. Here, the point was to **convey is the idea itself, rather than any precise data.**

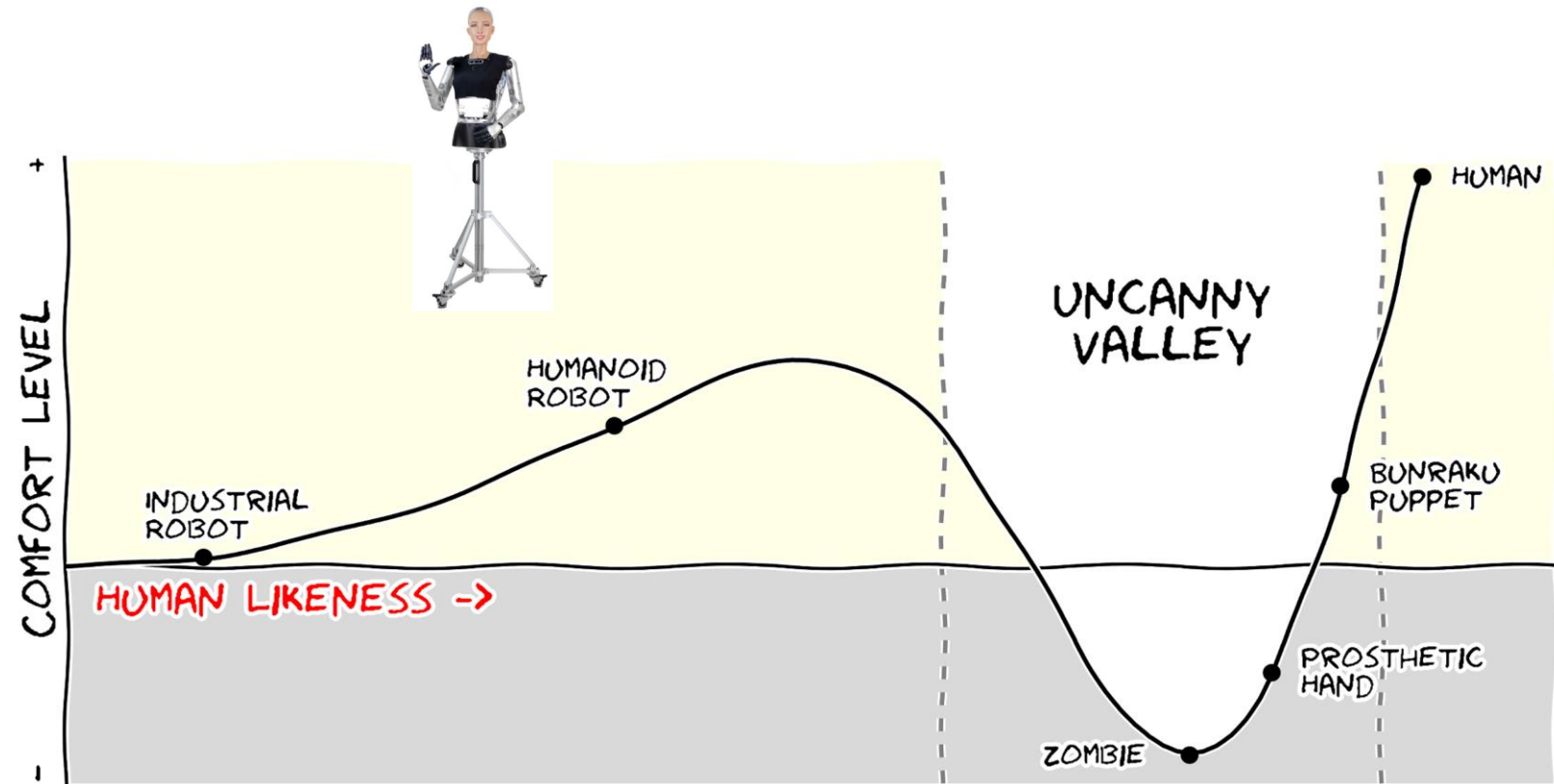


Image sources:

<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833>

<https://www.robotshop.com/de/en/hanson-robotics-sophia-2020-rd-version.html>

Conveying Information

Chartjunk

- may sometimes work well if **done right**?
- may work in **non-scientific**, e.g. mass-media, settings?

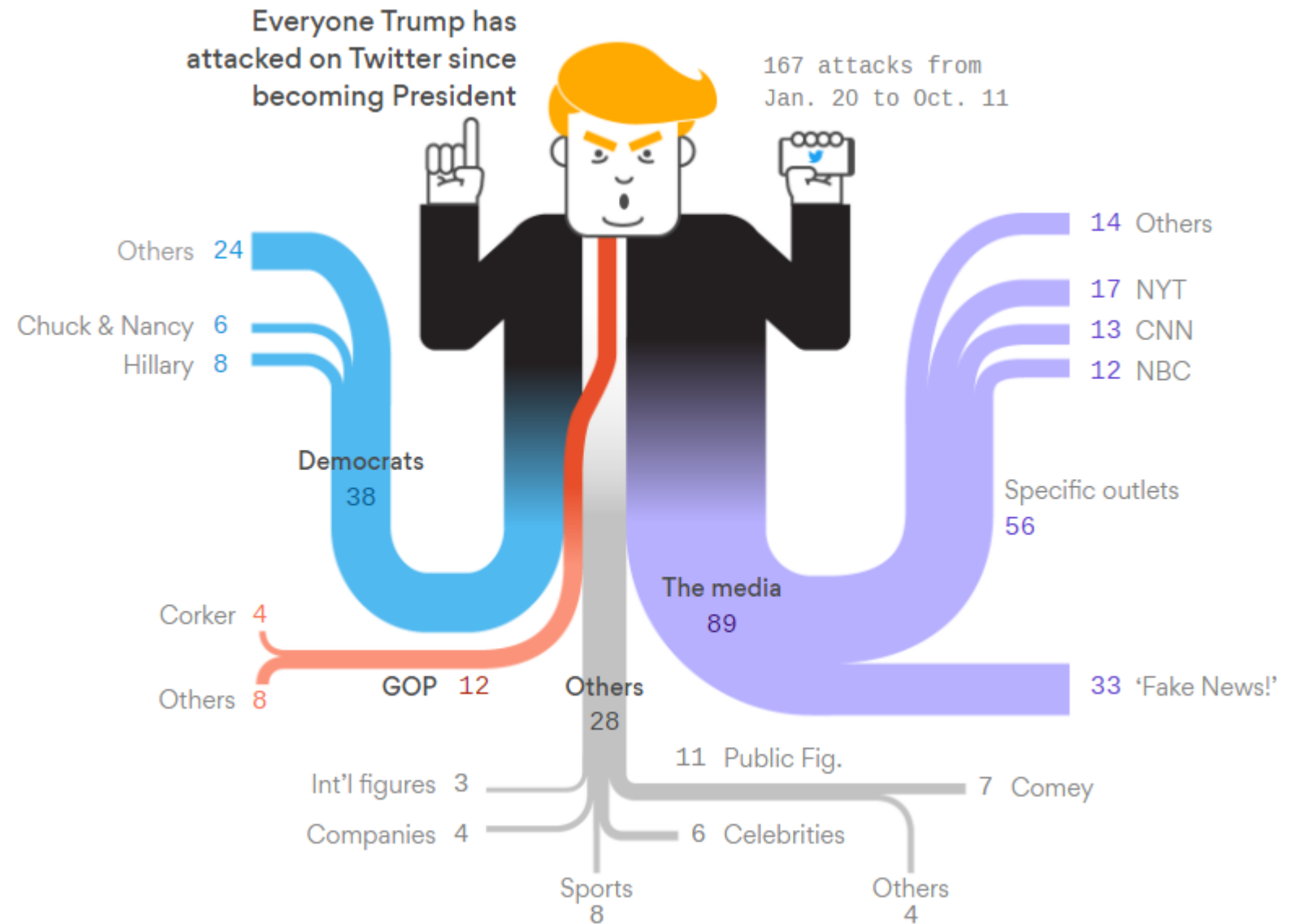


Image source:

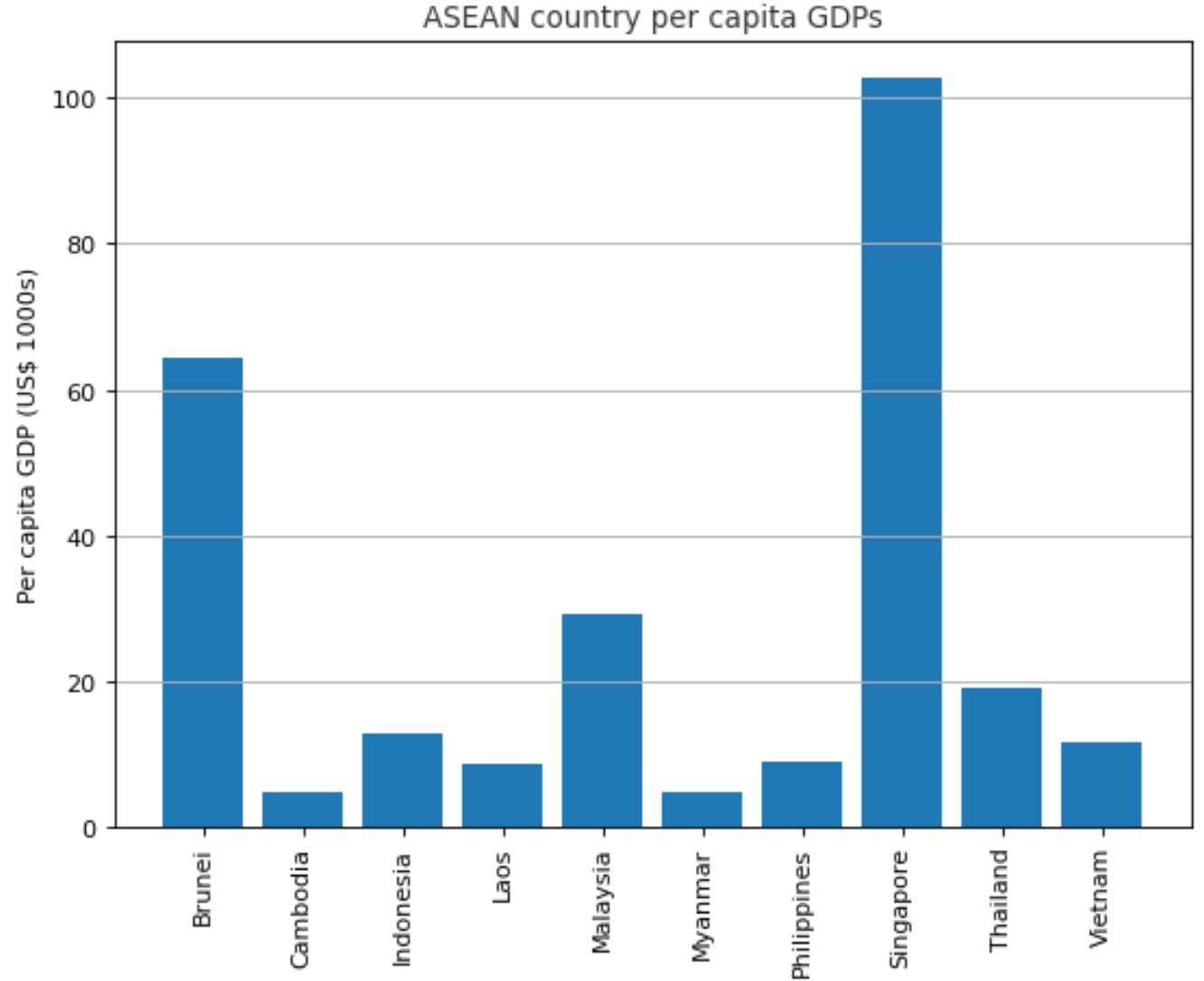
<https://www.axios.com/who-trump-attacks-the-most-on-twitter-1513305449-f084c32e-fcdf-43a3-8c55-2da84d45db34.html>

Example: A bar plot

▶ Data-ink

- Edward Tufte's maxim:

1. **Above all else show the data.**
2. Maximize the data-ink ratio
3. Erase non-data-ink.
4. Erase redundant data-ink.
5. Revise and edit.



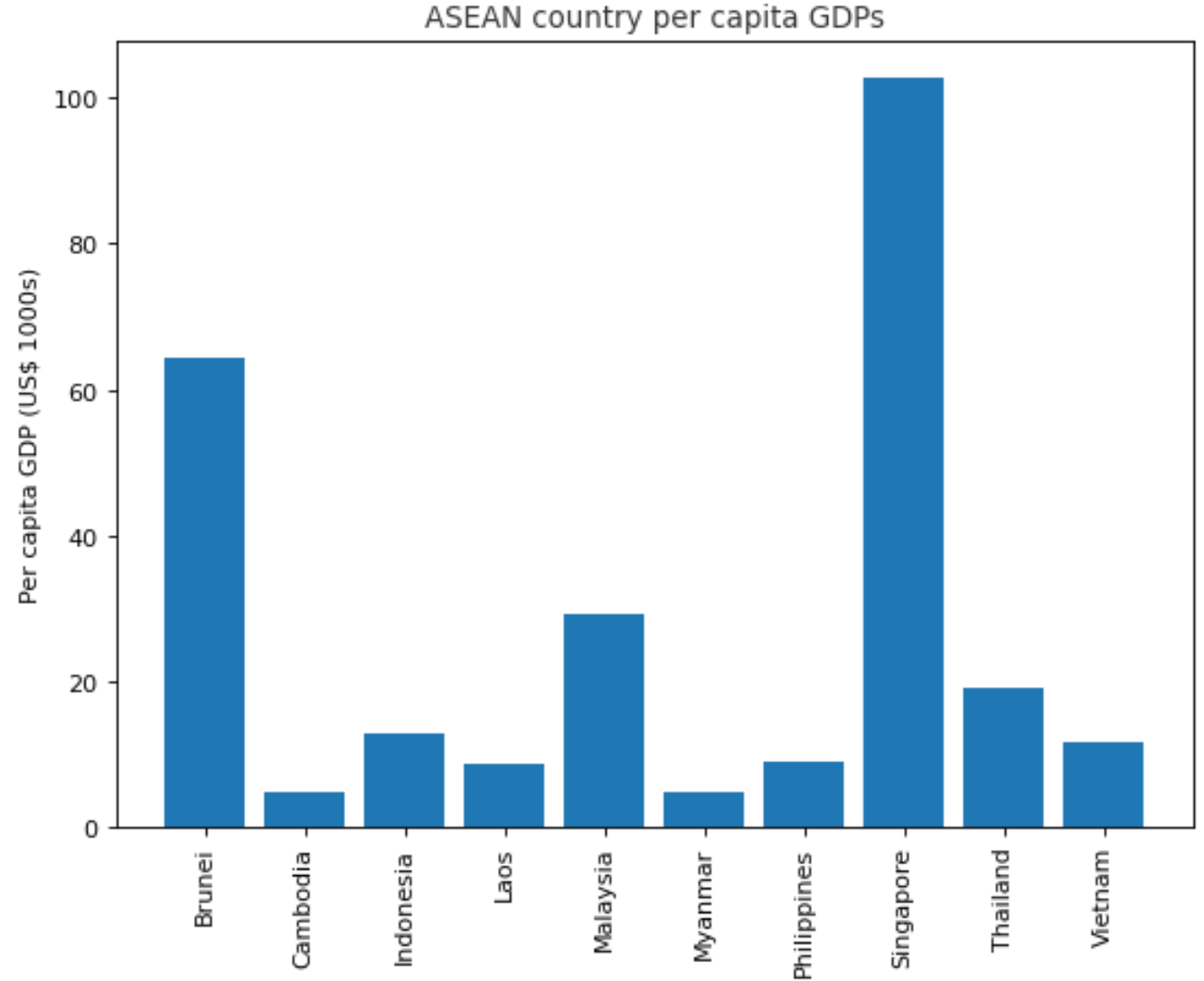
Data source: https://en.wikipedia.org/wiki/List_of_ASEAN_countries_by_GDP (accessed/as on 27 August 2021)

Step by step ...

- Edward Tufte's maxim:

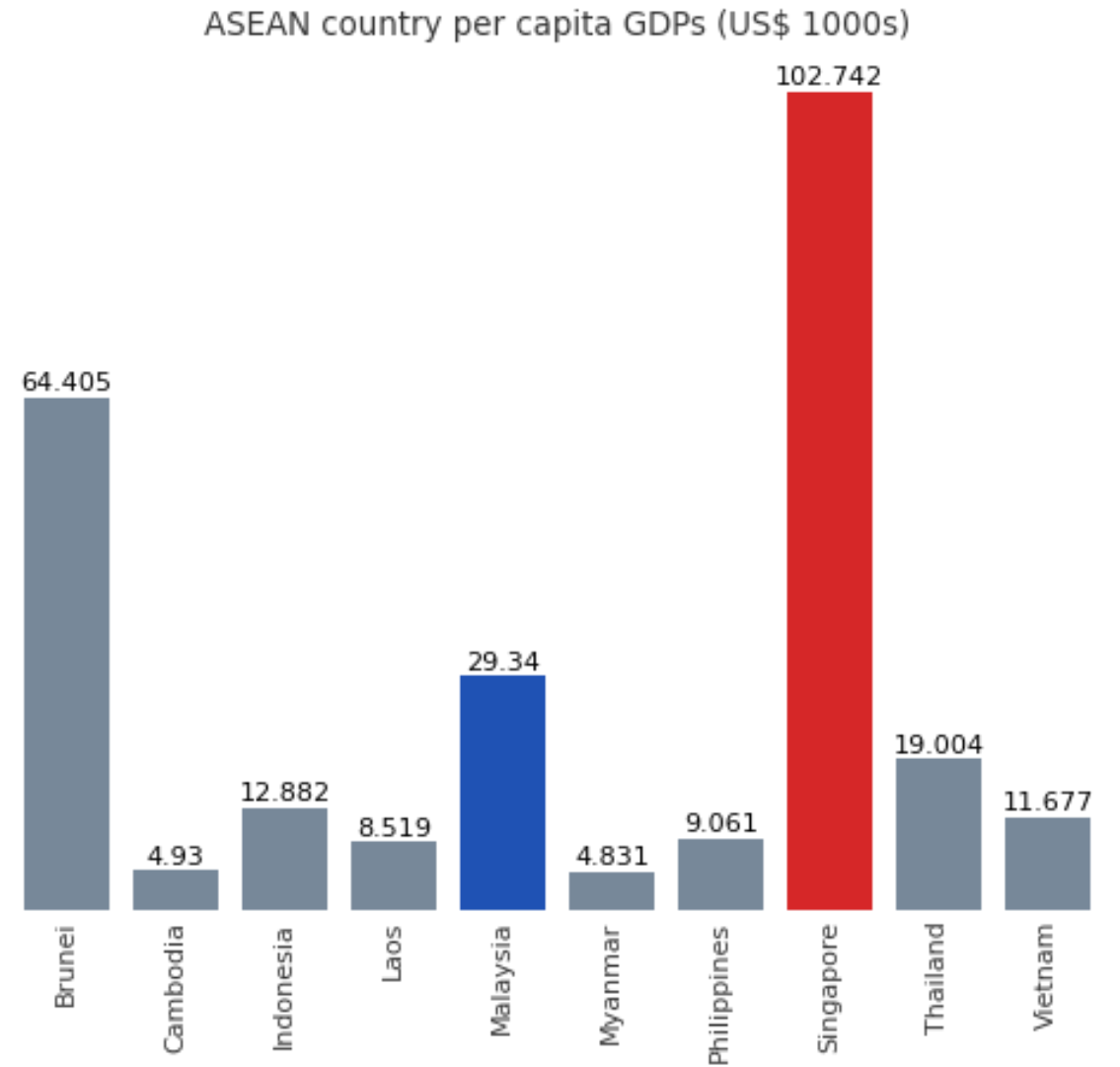
1. Above all else show the data.
2. Maximize the data-ink ratio
3. Erase non-data-ink.
4. Erase redundant data-ink.
5. Revise and edit.

Let's start by removing the grid lines



A plot with higher data-ink ratio

► With emphasis on some data records

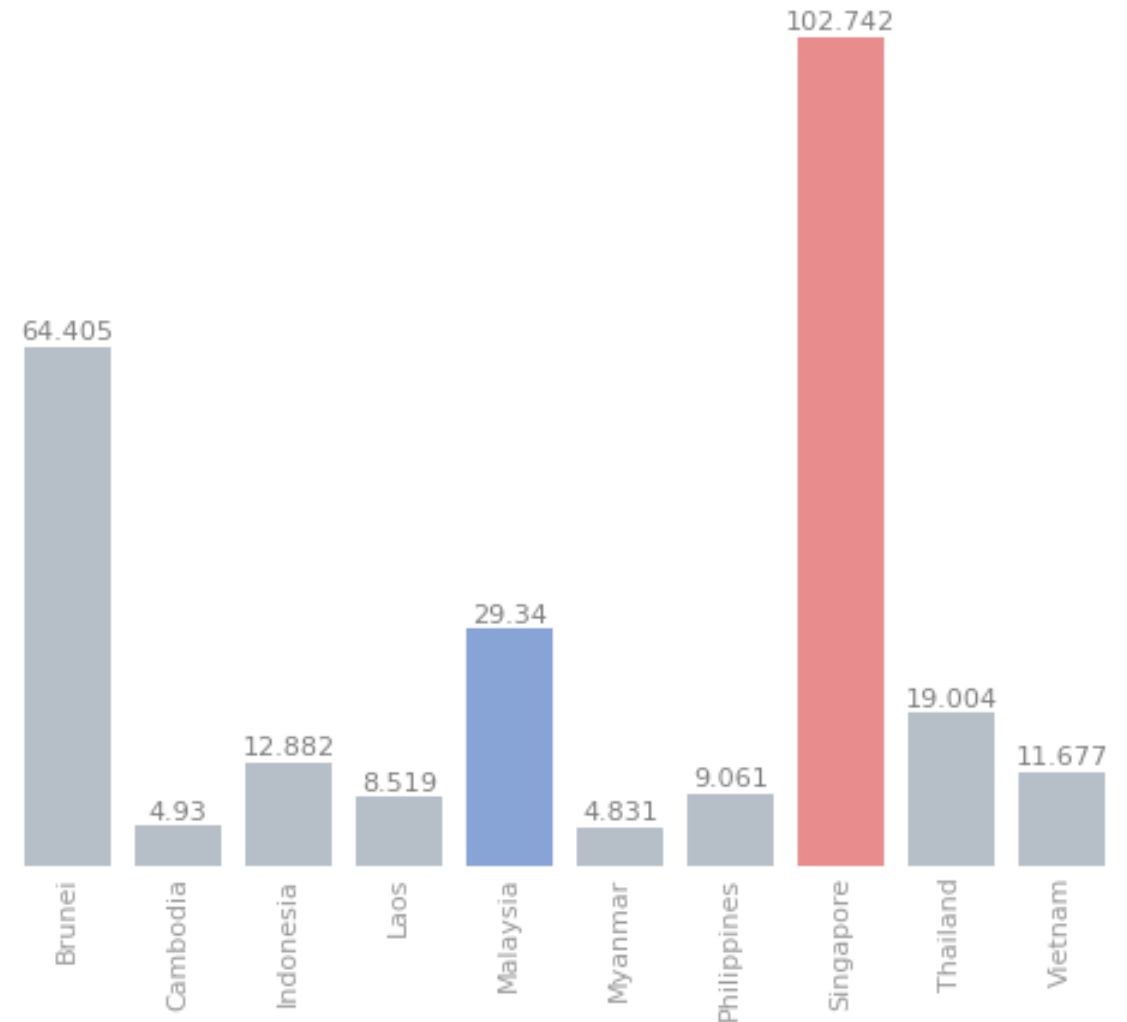


A plot with higher data-ink ratio

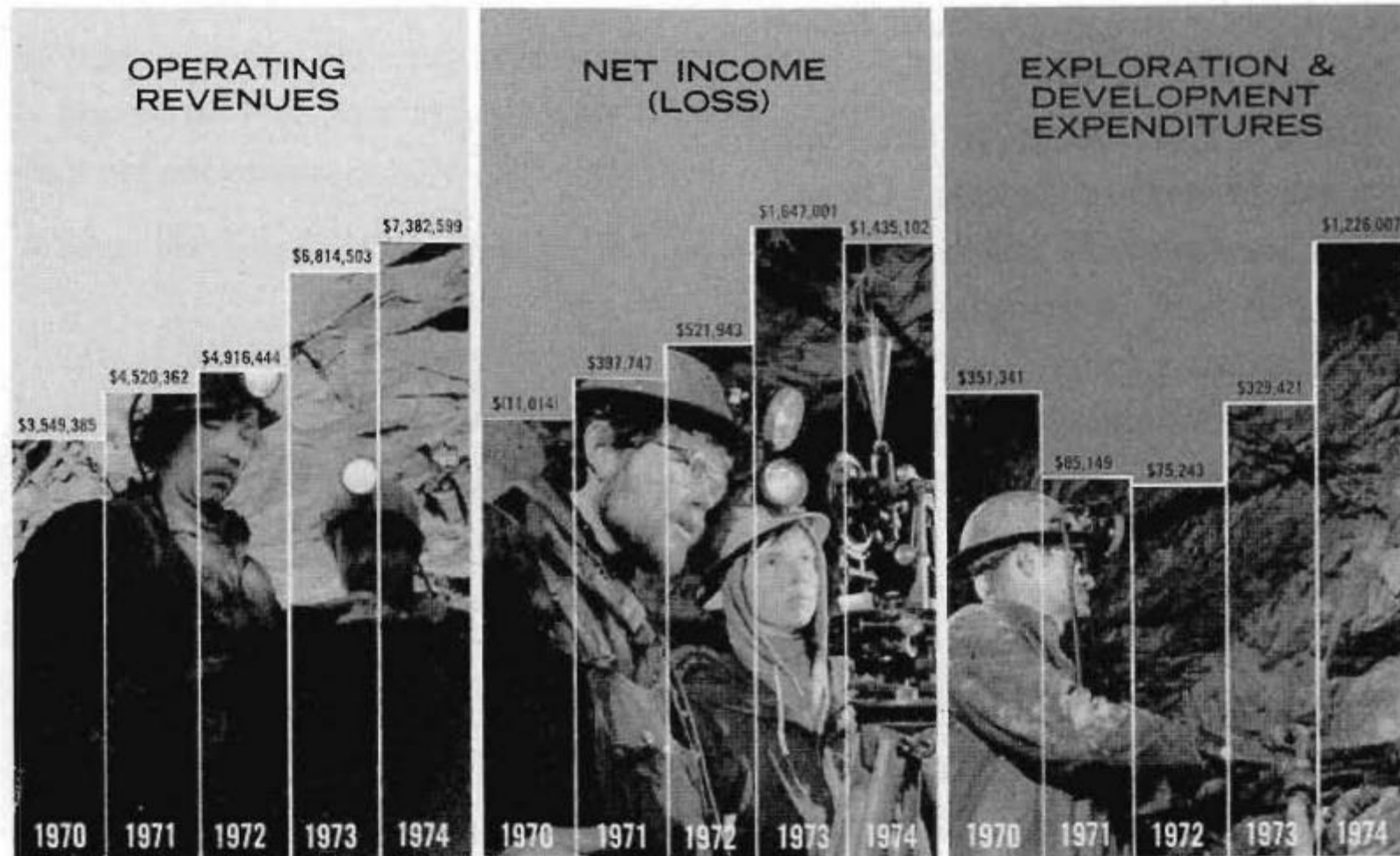
Ungraded task 4.1: Create a similar plot using your favourite tools (using same, or similar data). If you have ideas to further improve data-ink ratio, share the ideas and your final results too!

Ungraded task 4.2: Identify some other plots (which is not a bar plot) from a public source, which do not have a good data-ink ratio in your judgement. Consider how you may improve them, and share your improved results along with the 'original' you improved upon.

ASEAN country per capita GDPs (US\$ 1000s)



What is wrong with this graphic? (beside the chartjunk)

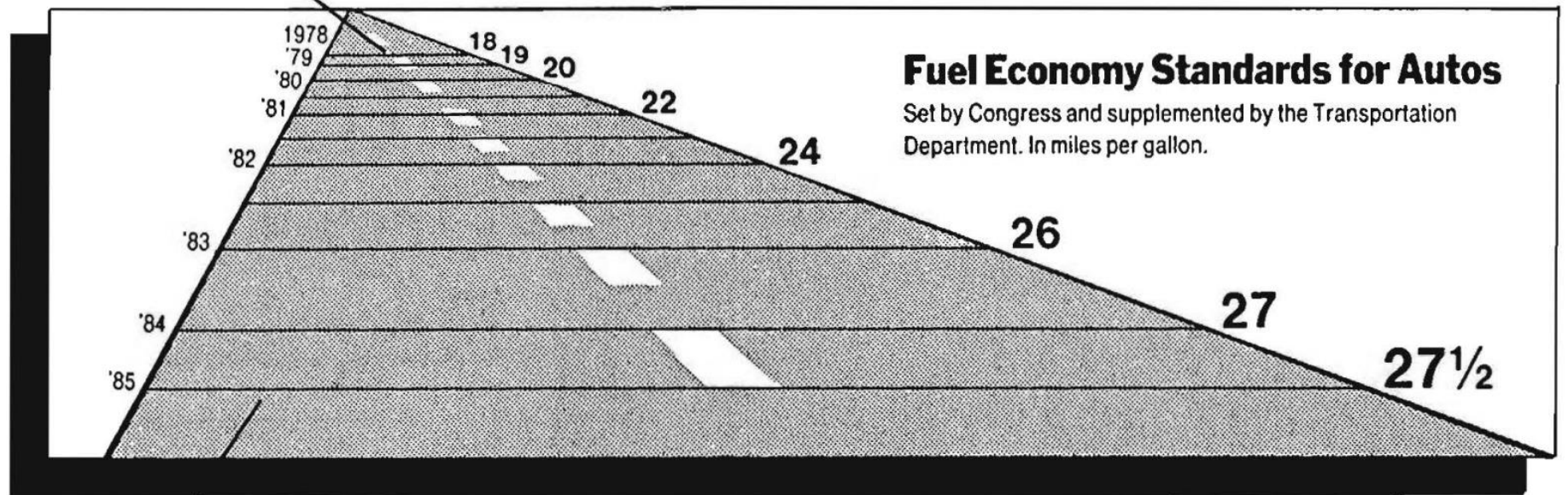


Source: The visual display of quantitative information by Edward Tufte

Graphical integrity and the lie-factor

$$\text{lie-factor} = \frac{\text{Size of effect shown in graphic}}{\text{Size of effect in data}}$$

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



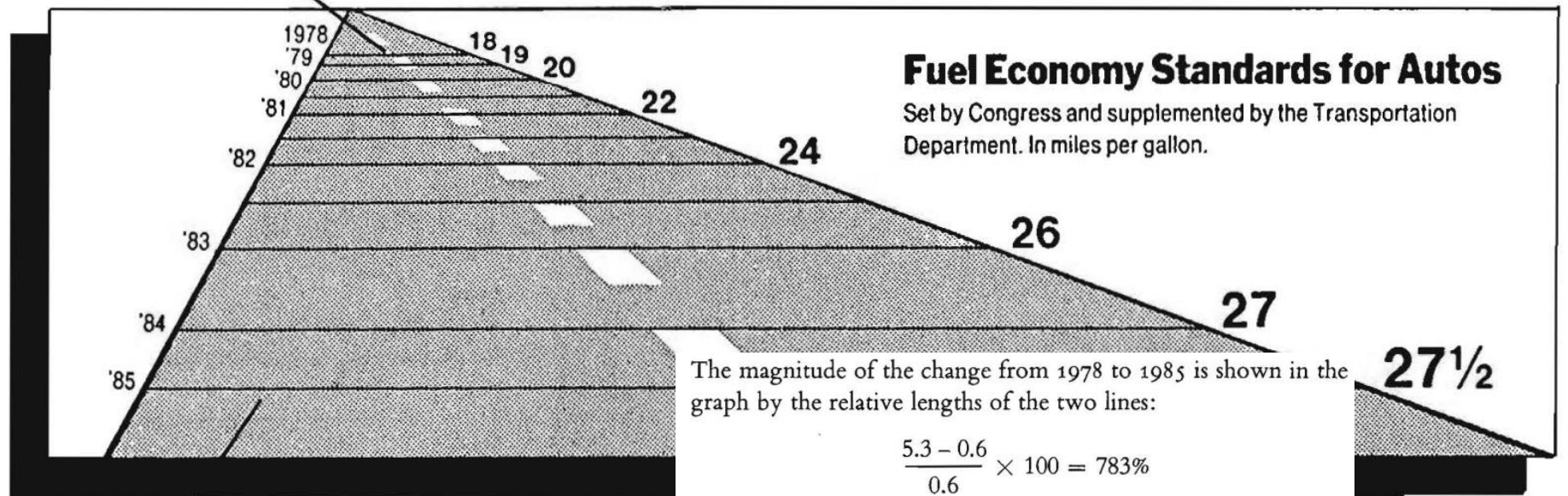
This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Source: The visual display of quantitative information by Edward Tufte

Graphical integrity and the lie-factor

$$\text{lie-factor} = \frac{\text{Size of effect shown in graphic}}{\text{Size of effect in data}}$$

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



The magnitude of the change from 1978 to 1985 is shown in the graph by the relative lengths of the two lines:

$$\frac{5.3 - 0.6}{0.6} \times 100 = 783\%$$

Thus the numerical change of 53 percent is presented by some lines that changed 783 percent, yielding

$$\text{Lie Factor} = \frac{783}{53} = 14.8$$

This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

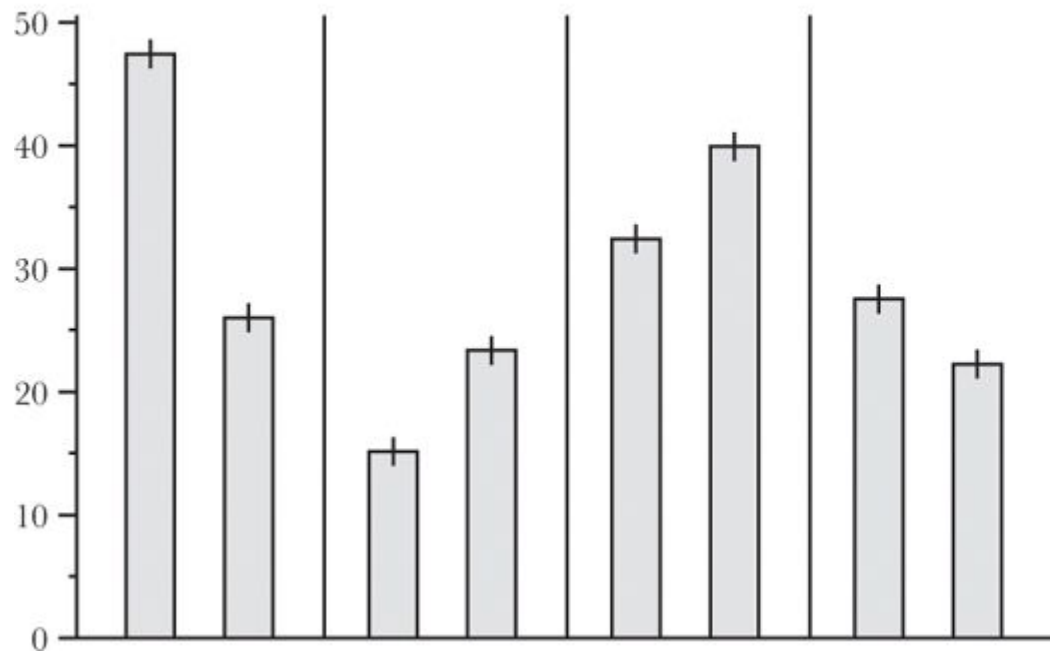
A few key take-aways on visualization by Edward Tufte

- Maximize **data-ink ratio**
- Get rid of **chartjunk**
 - **Caveat:** Chartjunk done right, may work!
- Aim for **no/low lie-factor**

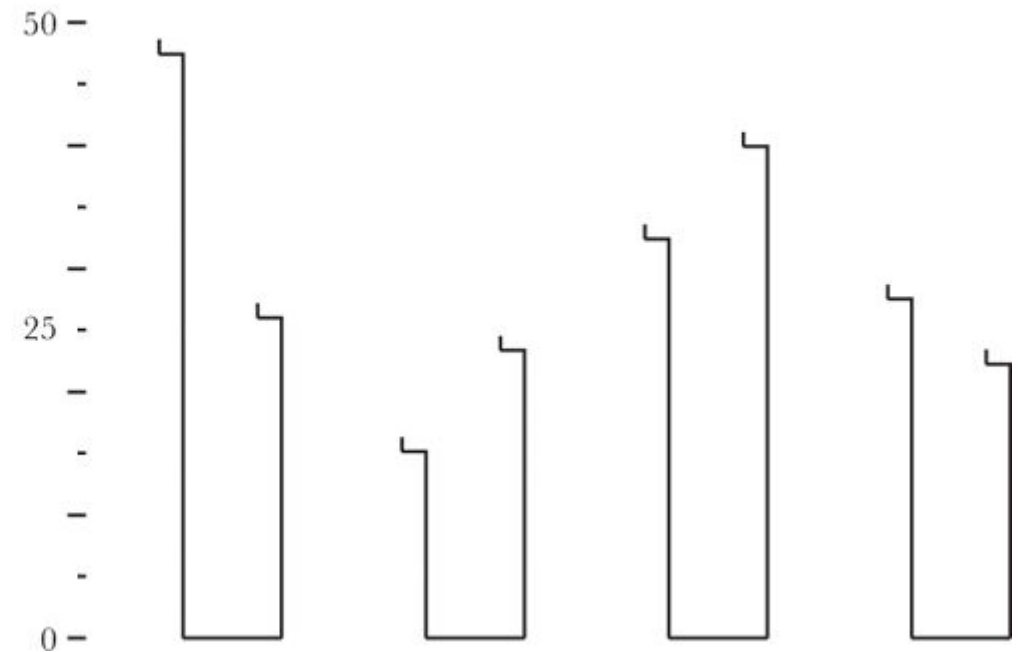
A word of caution: Extremes are easy, strive for balance

► Which do you like better?

Traditional Bar Chart



Maximized Data-Ink Ratio



Source: The functional art by Alberto Cairo

More on graphical integrity: Misleading graphics

▶ Spurious correlations!

Worldwide non-commercial space launches correlates with Sociology doctorates awarded (US)

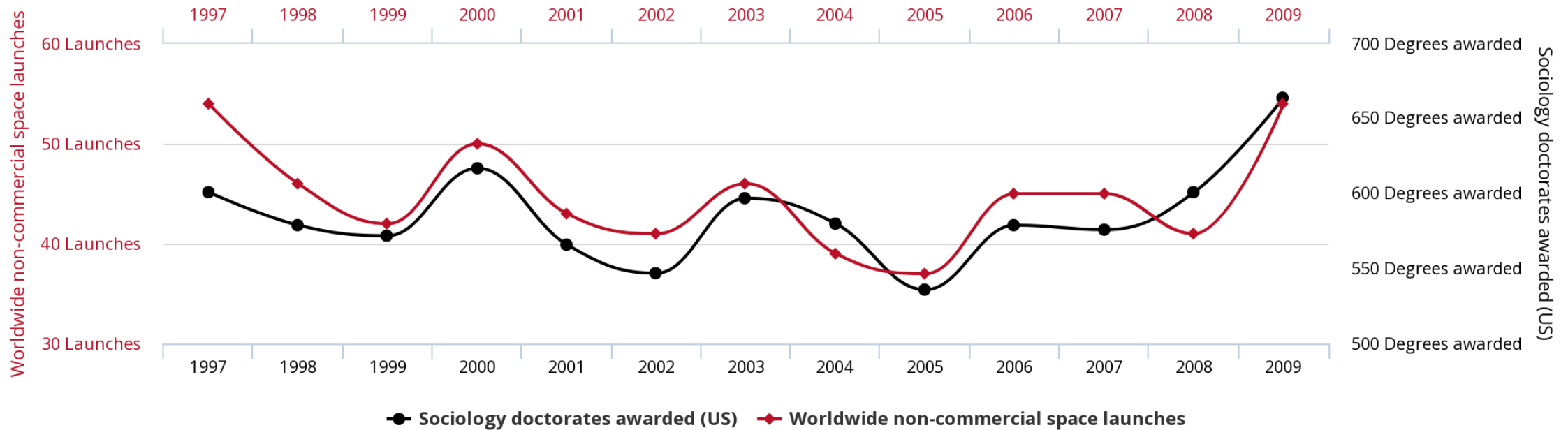
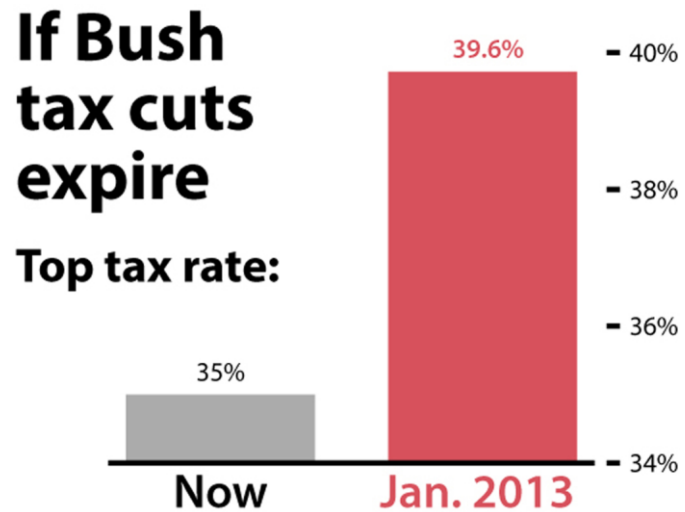


Image source: <http://www.tylervigen.com/spurious-correlations>

tylervigen.com

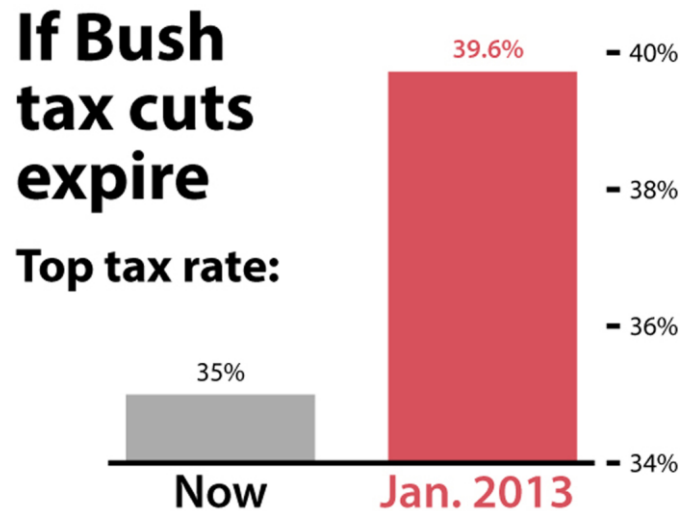
More on graphical integrity: Misleading graphics



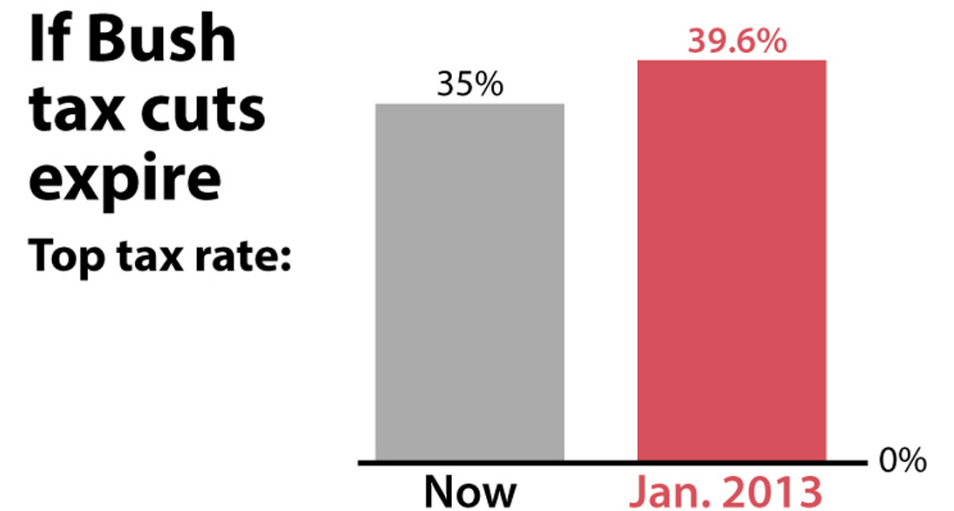
Source: How charts lie by Alberto Cairo

More on graphical integrity: Misleading graphics

▶ Chart baseline is not at 0!



VS



Source: How charts lie by Alberto Cairo

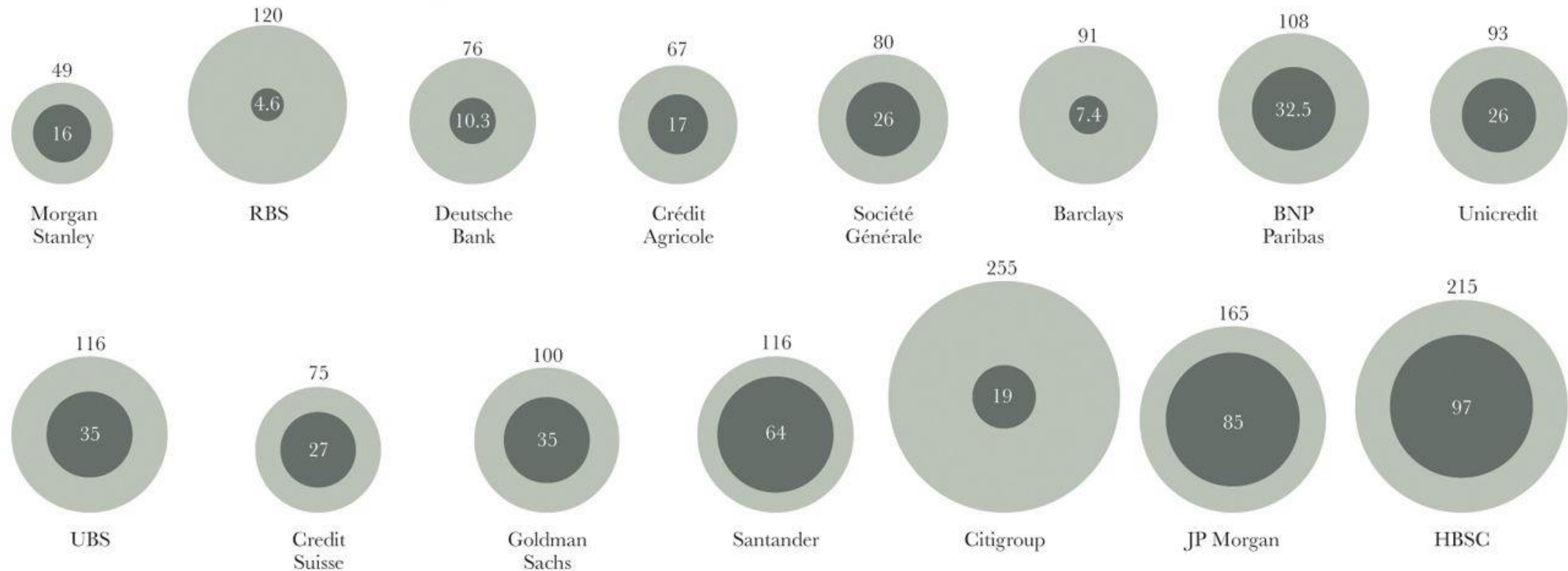
More confusing graphics

Market Capitalization of the World's Biggest Banks

In billions of dollars

● January 2007 ● January 2009

Source: Bloomberg



Source: The functional art by Alberto Cairo

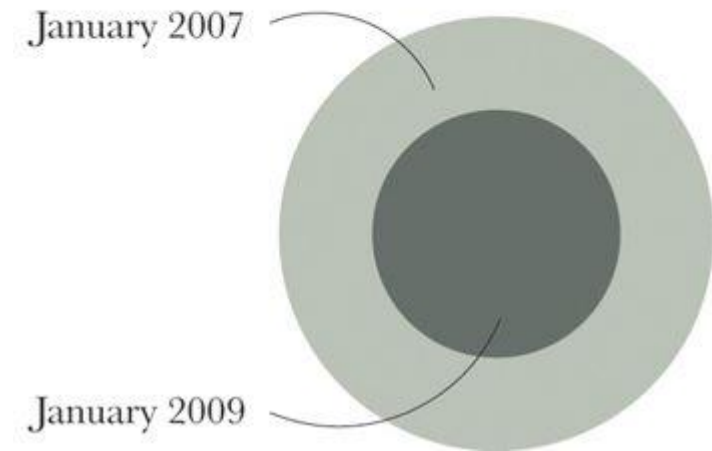
More confusing graphics

- ▶ The Outer bubble represents \$80 billion.
What does the second bubble represents?

Market Capitalization of Société Générale

Billions of dollars

Source: Bloomberg



Source: The functional art by Alberto Cairo

More confusing graphics

- ▶ The Outer bubble represents \$80 billion.
What does the second bubble represents?

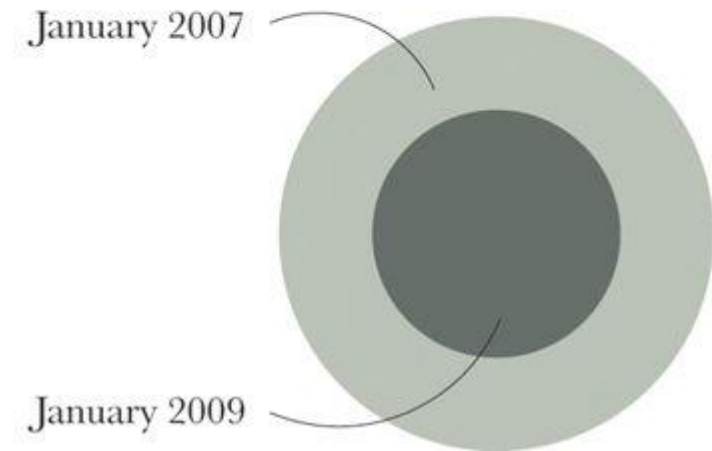


Want readers to **compare areas**.
But many tend to compare heights.

Market Capitalization of Société Générale

Billions of dollars

Source: Bloomberg



Source: The functional art by Alberto Cairo

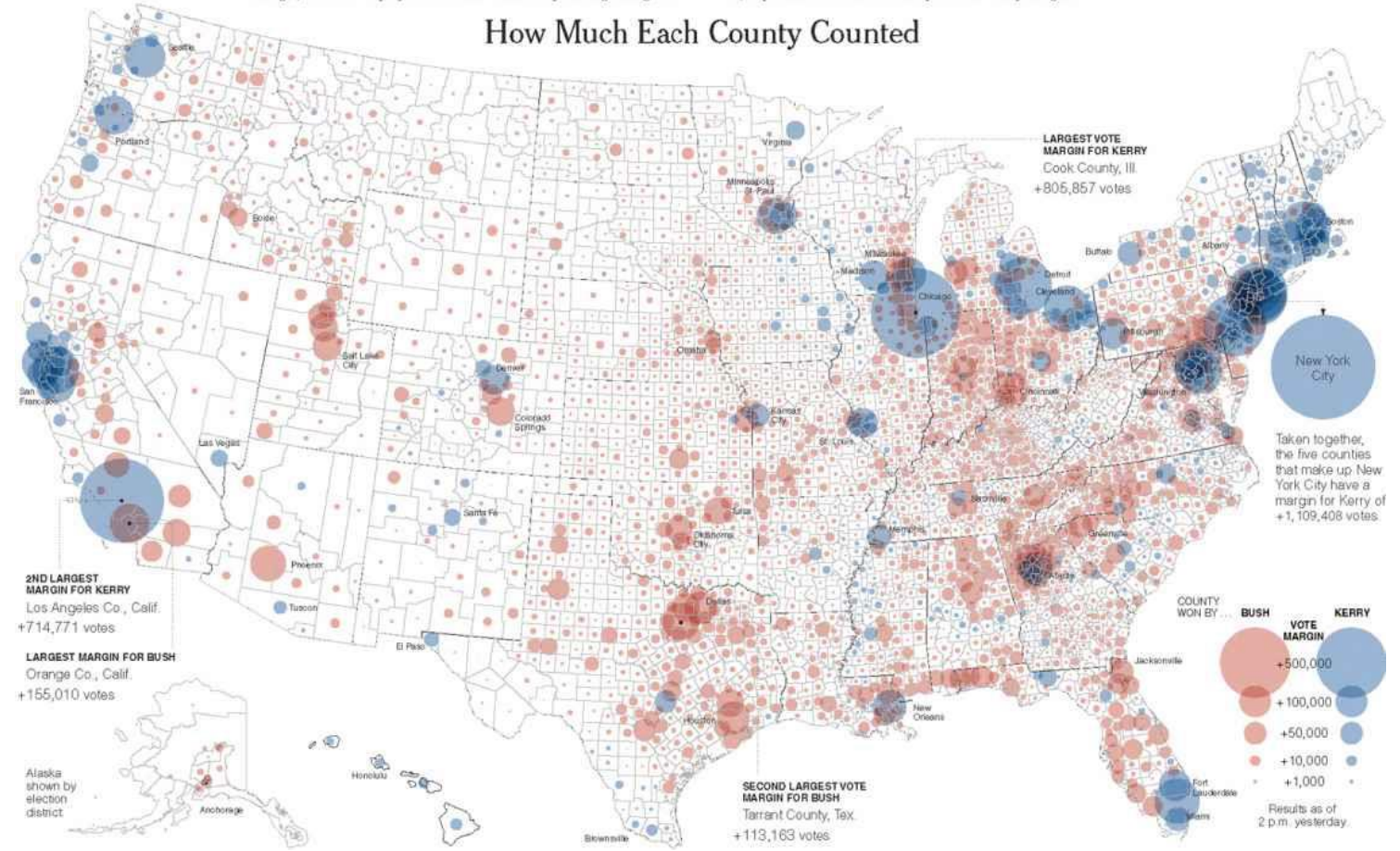
Bubbles are generally ok for the Big Picture



Red and Blue, the Divided Electorate, in All Its Shades

The simple formula for winning an election is to get more votes than your opponent in as many counties as possible. It worked for President Bush. The map below, which uses the size of circles to indicate a candidate's winning margin, shows how it played out. Senator John Kerry had huge margins in

many counties with large cities, and those margins were enough for him to win some of those states. However, Mr. Bush's relatively smaller but consistent margins in suburban and rural counties, in much of the South and West, helped him overcome Mr. Kerry's urban-county margins.

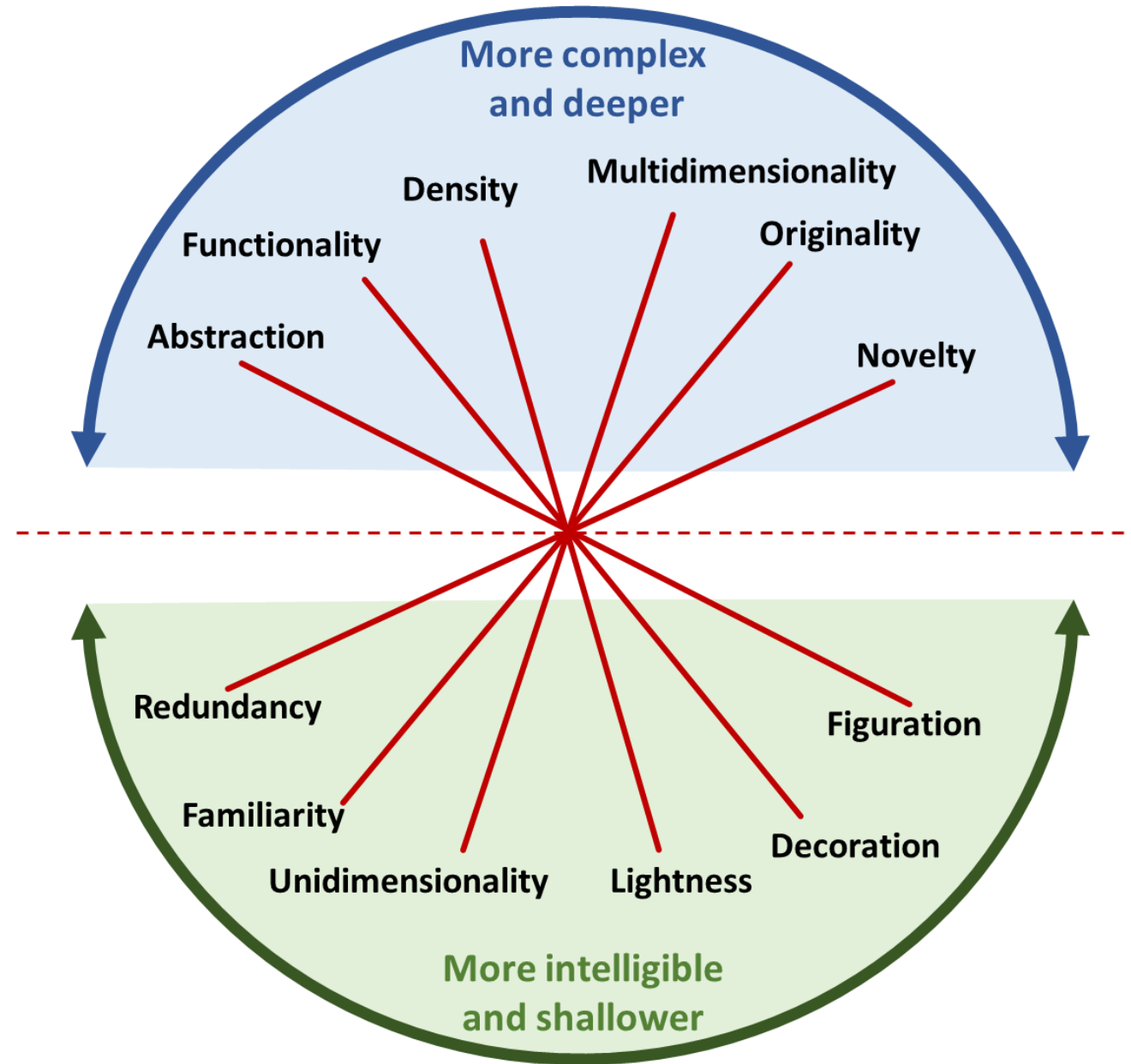


Source: The functional art by Alberto Cairo

Planning your visualization

▶ Visualization Wheel – Alberto Cairo

An exercise in meta-visualization:
a *visualization for planning visualizations*

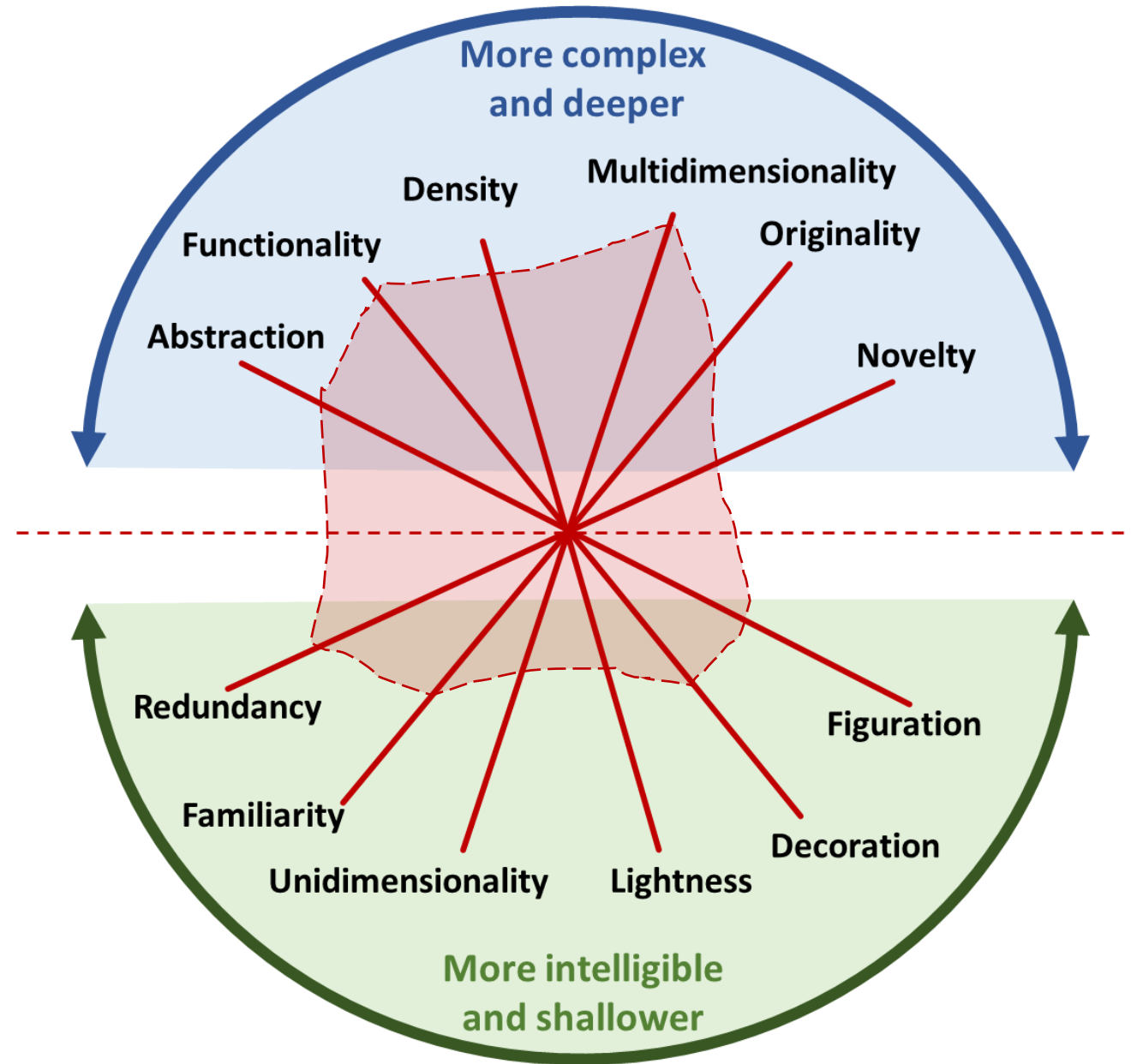


Redrawn. Original concept of visualization wheel is from The functional art by Alberto Cairo

A hypothetical visual ...

► Visualization Wheel – Alberto Cairo

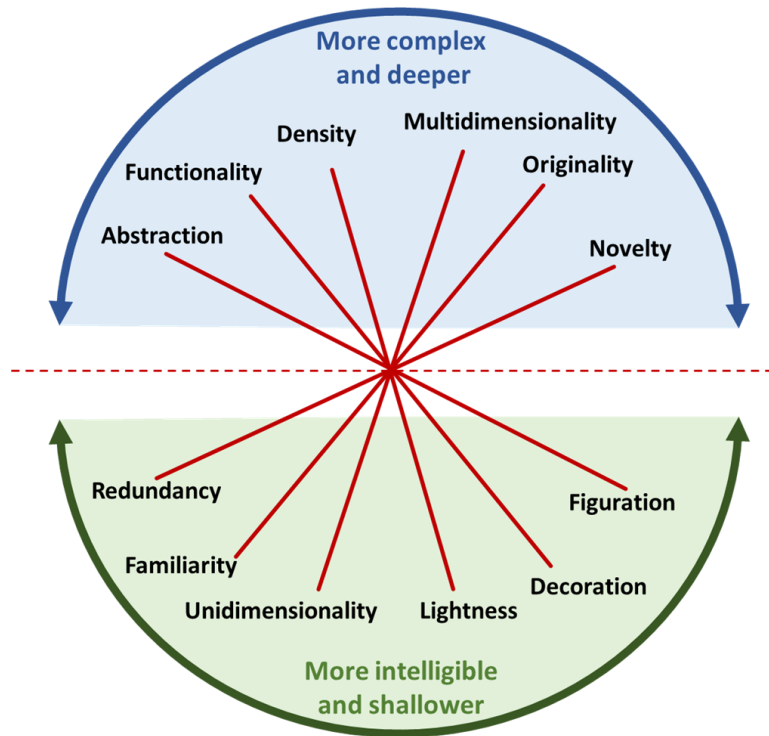
An individual visualization may have a specific combination of characteristics.
Caution: This is highly subjective.



Where on the Visualization Wheel is The Big Mac Index?

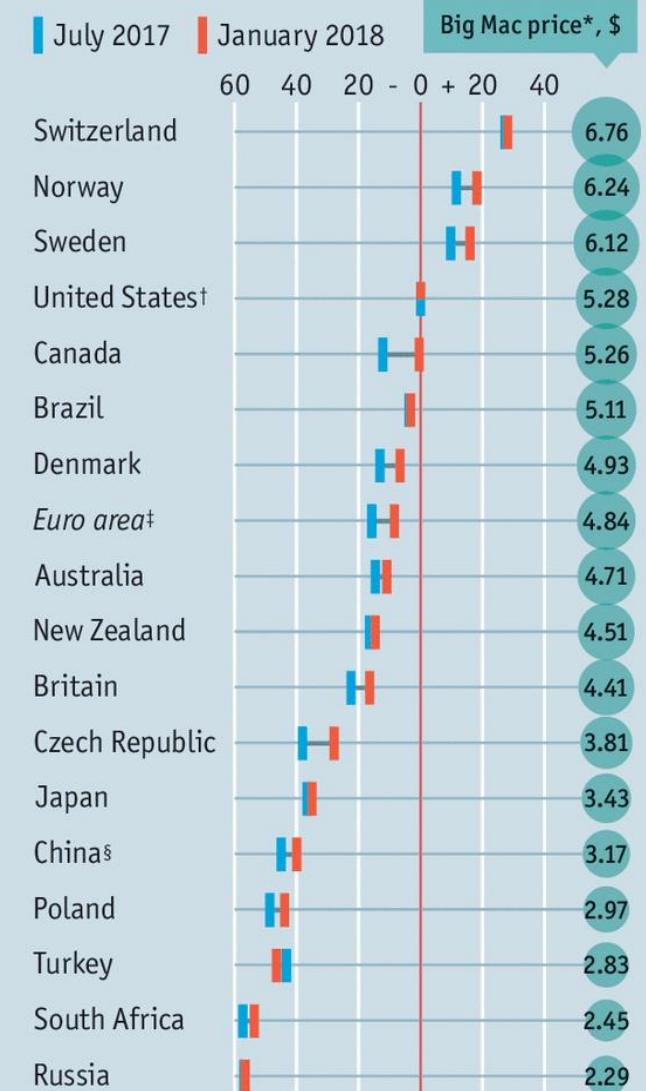
► Visualization Wheel – Alberto Cairo

Where does the Big Mac Index figure from **The Economist** fit in?



The Big Mac index

Local currency under(-)/over(+) valuation against the dollar, %



*At market exchange rates (Jan 17th 2018)
 †Average of four cities ‡Weighted average of member countries
 §Average of five cities

Sources: McDonald's; The Economist

Try it yourself

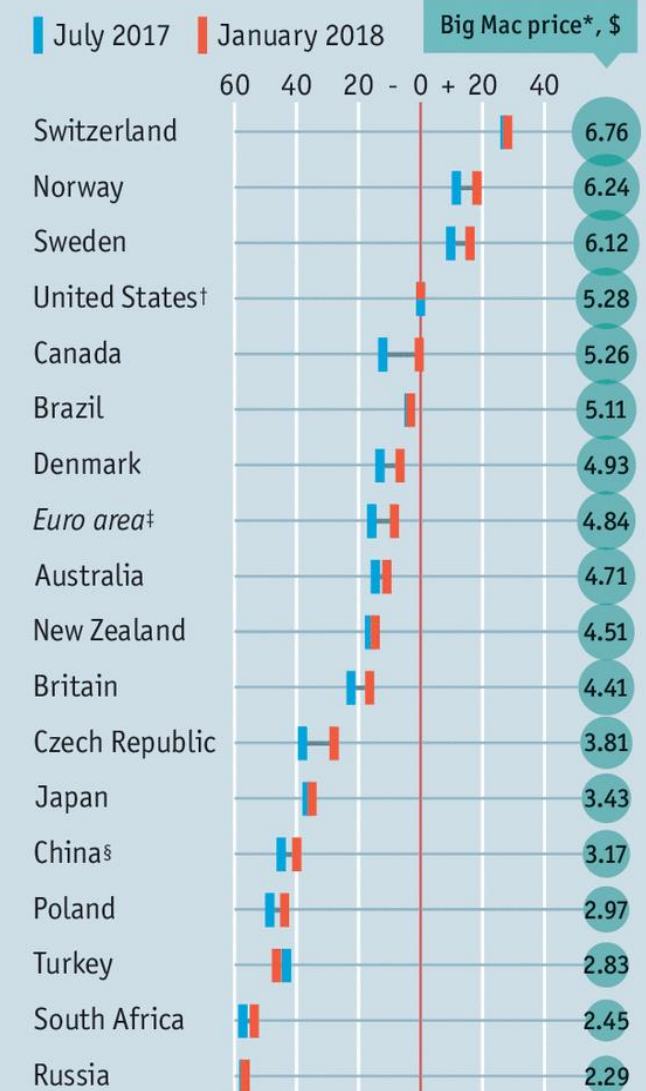
Here's the Big Mac Index data

<https://github.com/TheEconomist/big-mac-data>

Ungraded task 4.3: Create a similar plot with the data, using framework of your choice. Share your code & plot result.

The Big Mac index

Local currency under(-)/over(+) valuation against the dollar, %



*At market exchange rates (Jan 17th 2018)

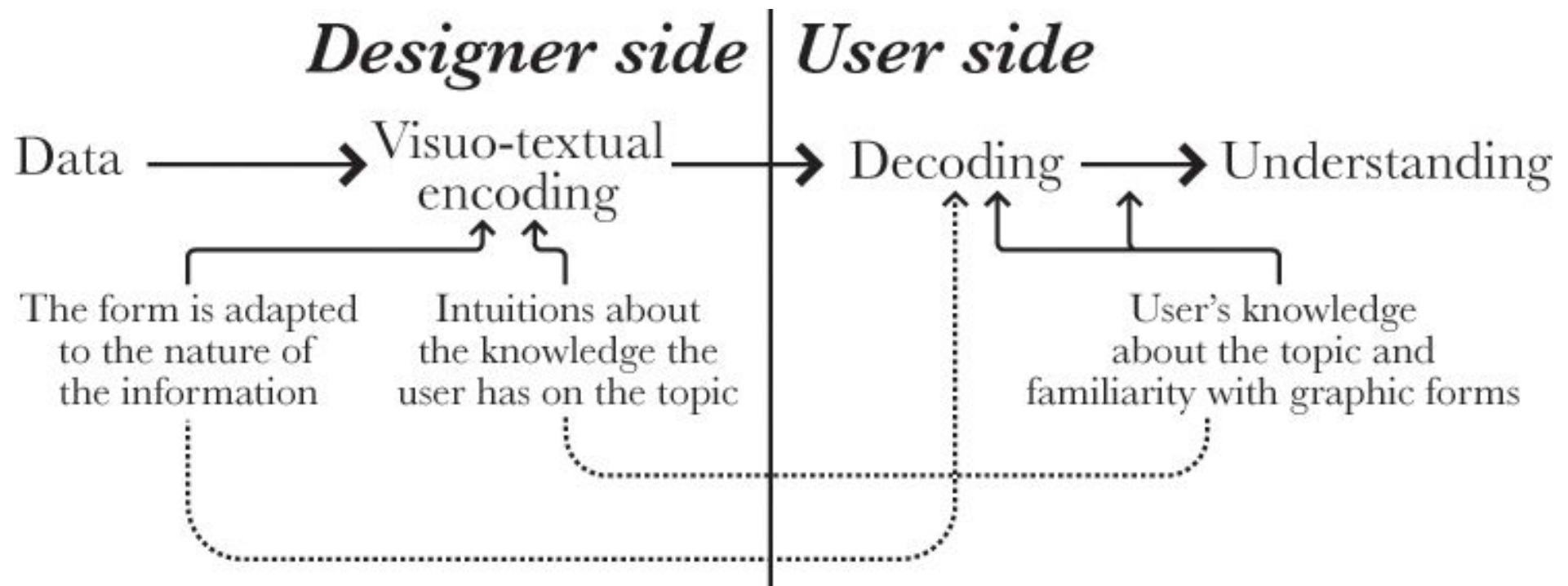
†Average of four cities ‡Weighted average of member countries

§Average of five cities

Sources: McDonald's; *The Economist*

Complexity adapted to the audience

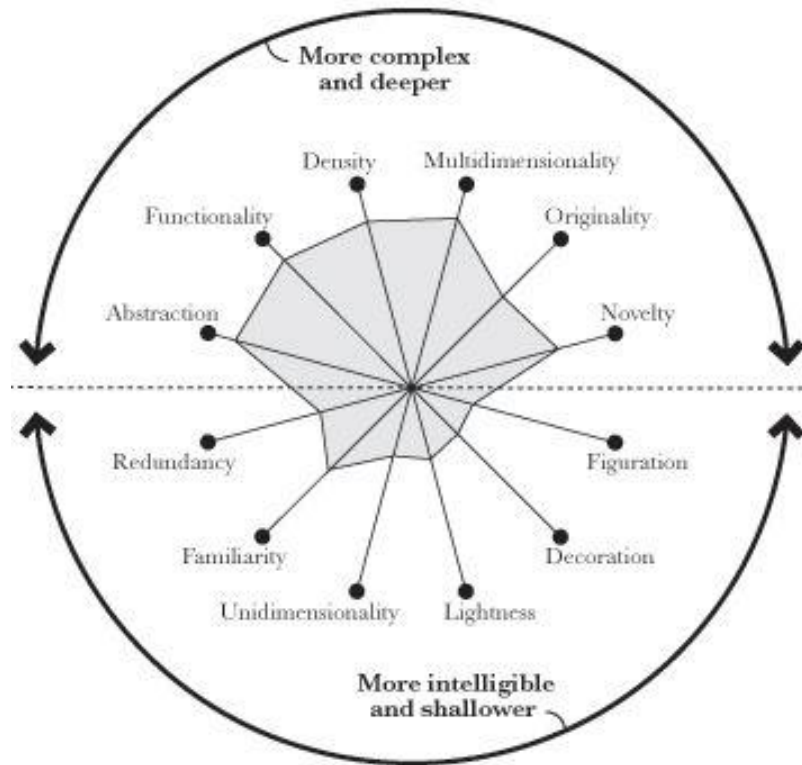
- ▶ If the subject is complex
Aim should be to clarify, rather than simplify/dumb it down



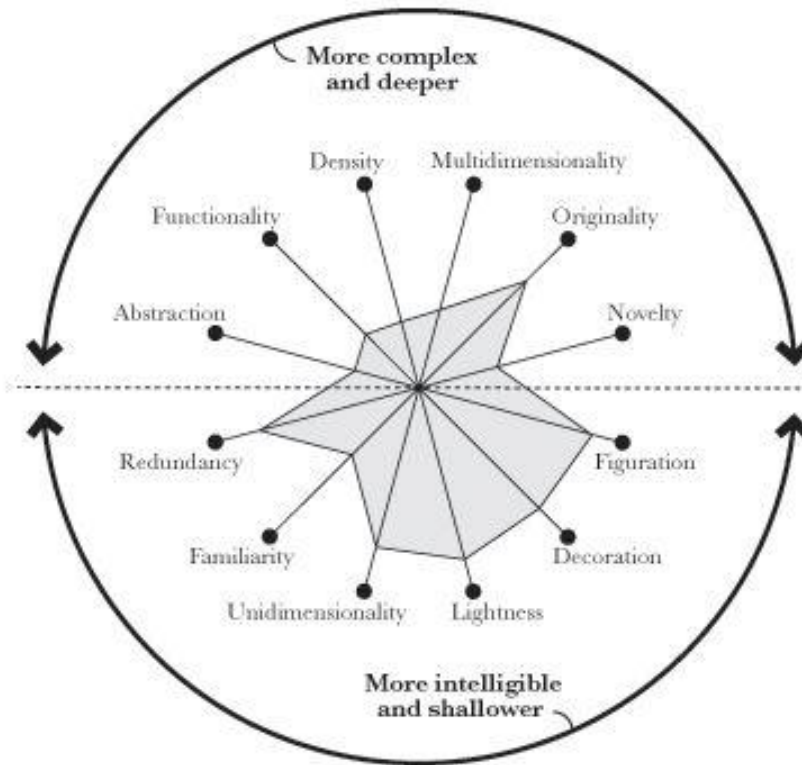
Source: The functional art by Alberto Cairo

Complexity adapted to the audience & purpose

The wheel preferred by scientists and engineers

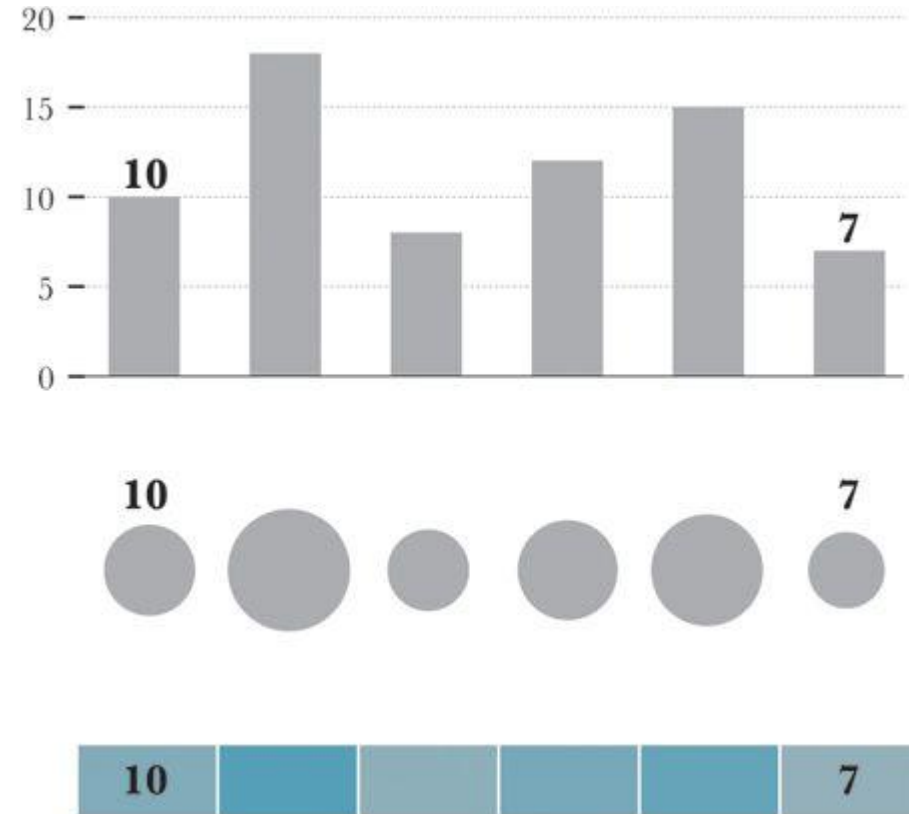


The wheel favored by artists, graphic designers, and journalists



Source: The functional art by Alberto Cairo

Same information, different ways to display



Source: The functional art by Alberto Cairo

Accurate

Generic

Length (aligned)



Length



Slope



Angle



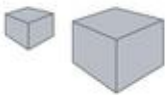
Area



Color intensity



Volume

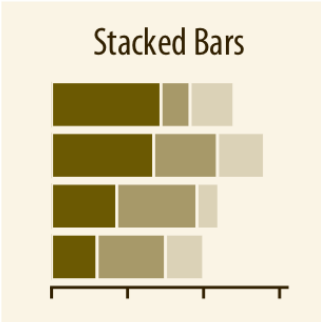
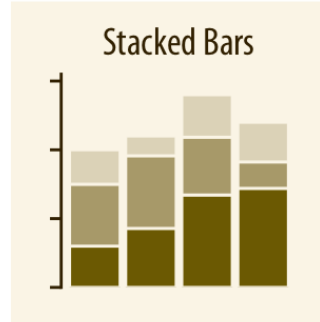
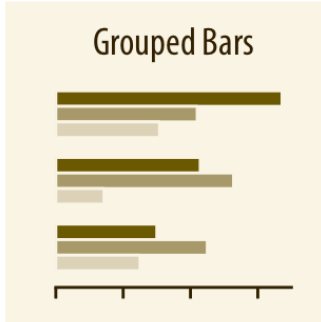
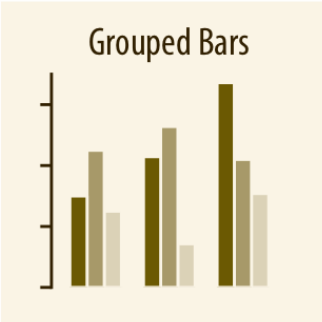
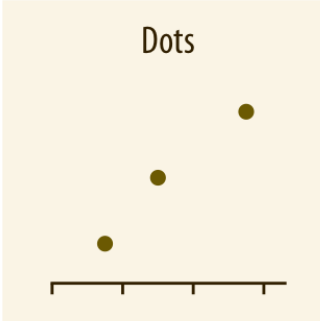
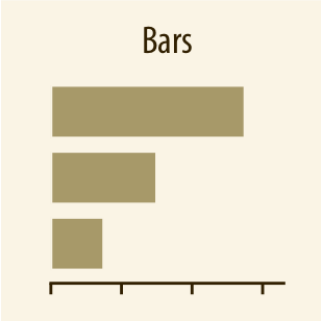
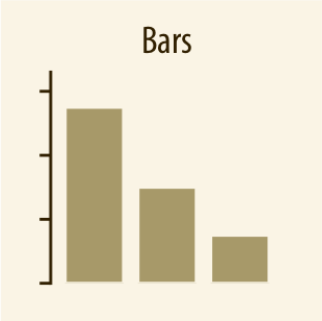


Color hue



A directory of visualizations

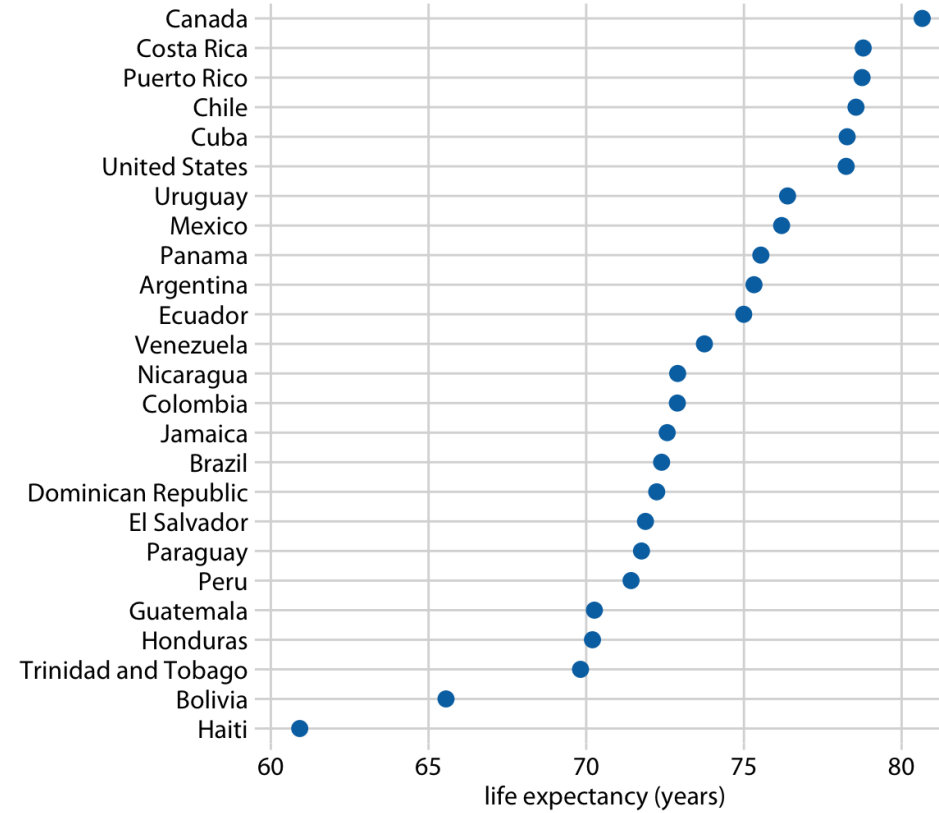
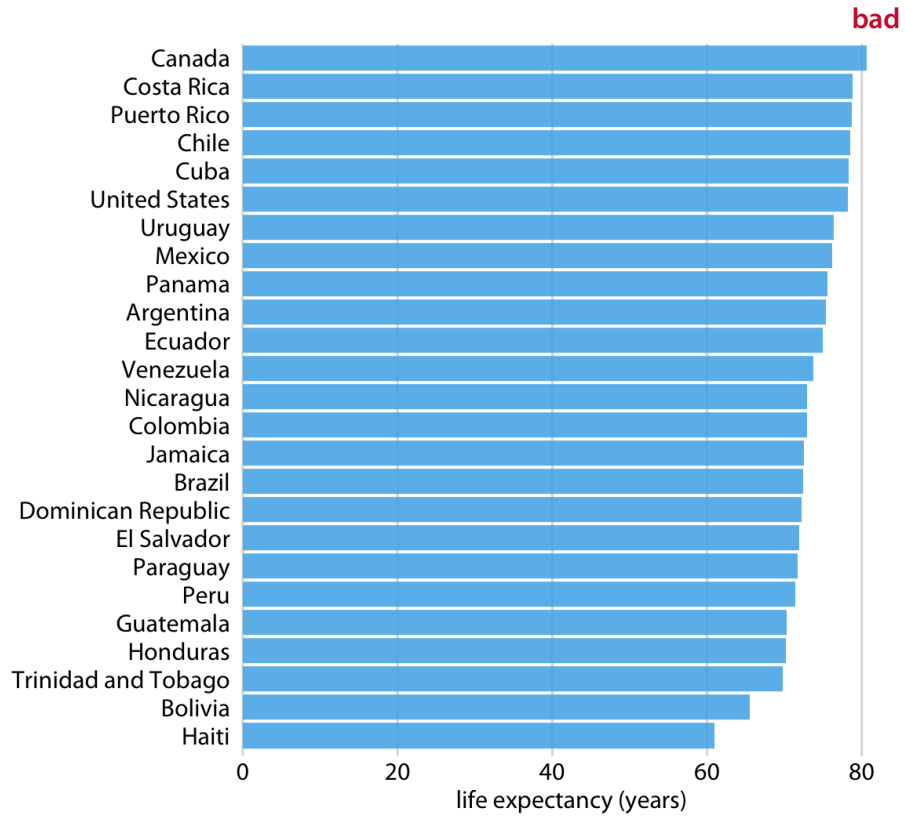
► Amounts



Source: Fundamentals of Data Visualization by Claus O. Wilke

Examples

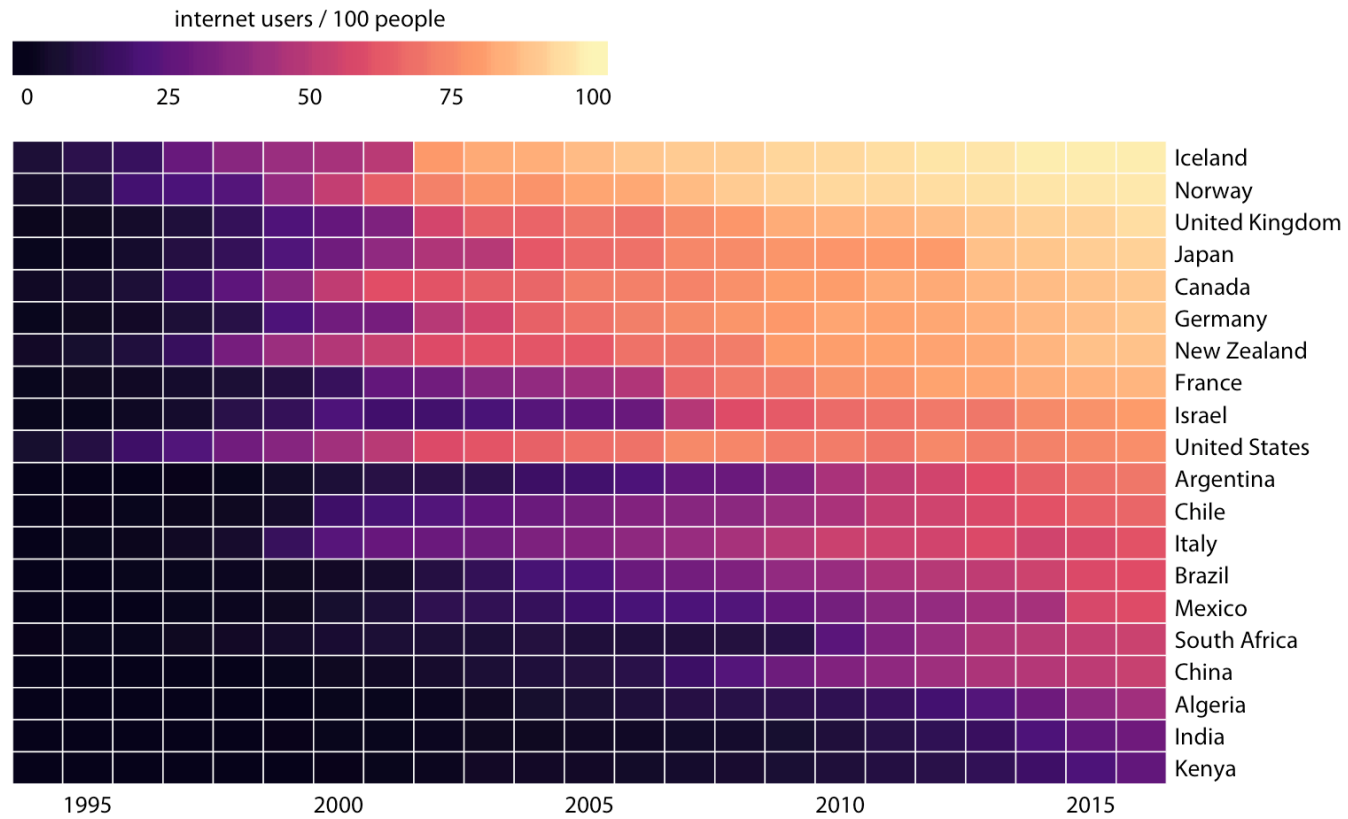
Amounts



Source: Fundamentals of Data Visualization by Claus O. Wilke

A more sophisticated example

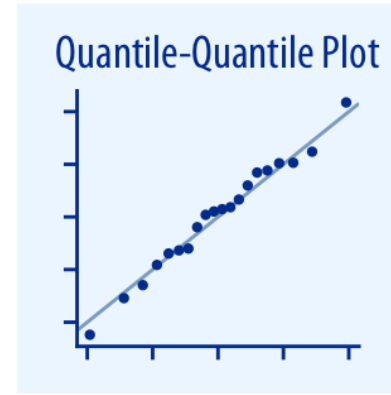
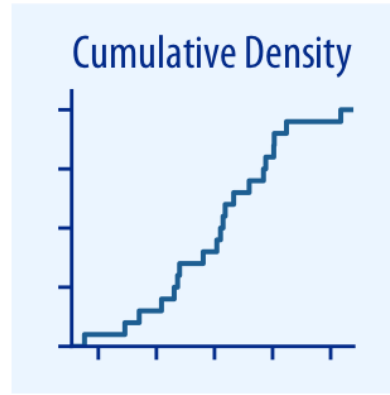
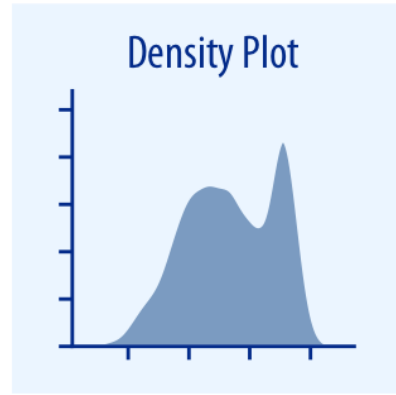
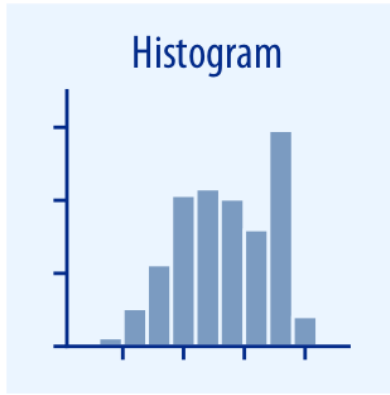
► Amounts: Heatmap to capture temporal dimension



Source: Fundamentals of Data Visualization by Claus O. Wilke

A directory of visualizations

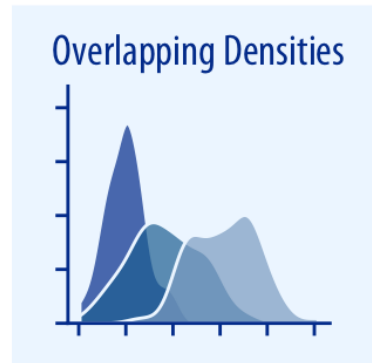
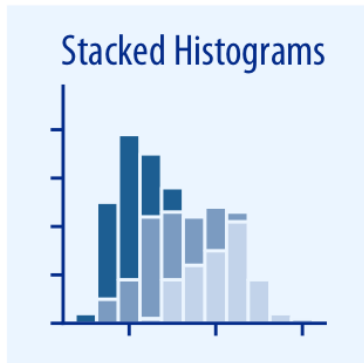
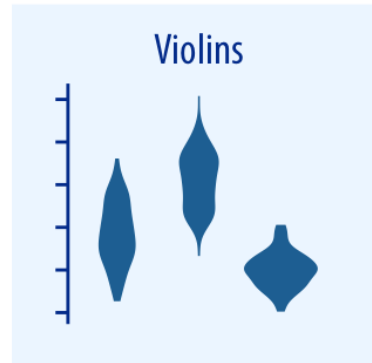
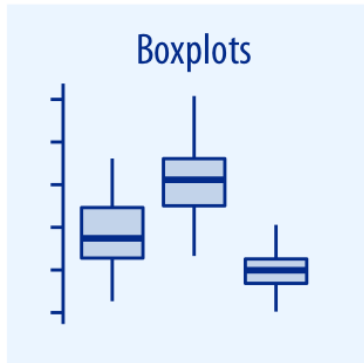
► Distributions



Source: Fundamentals of Data Visualization by Claus O. Wilke

A directory of visualizations

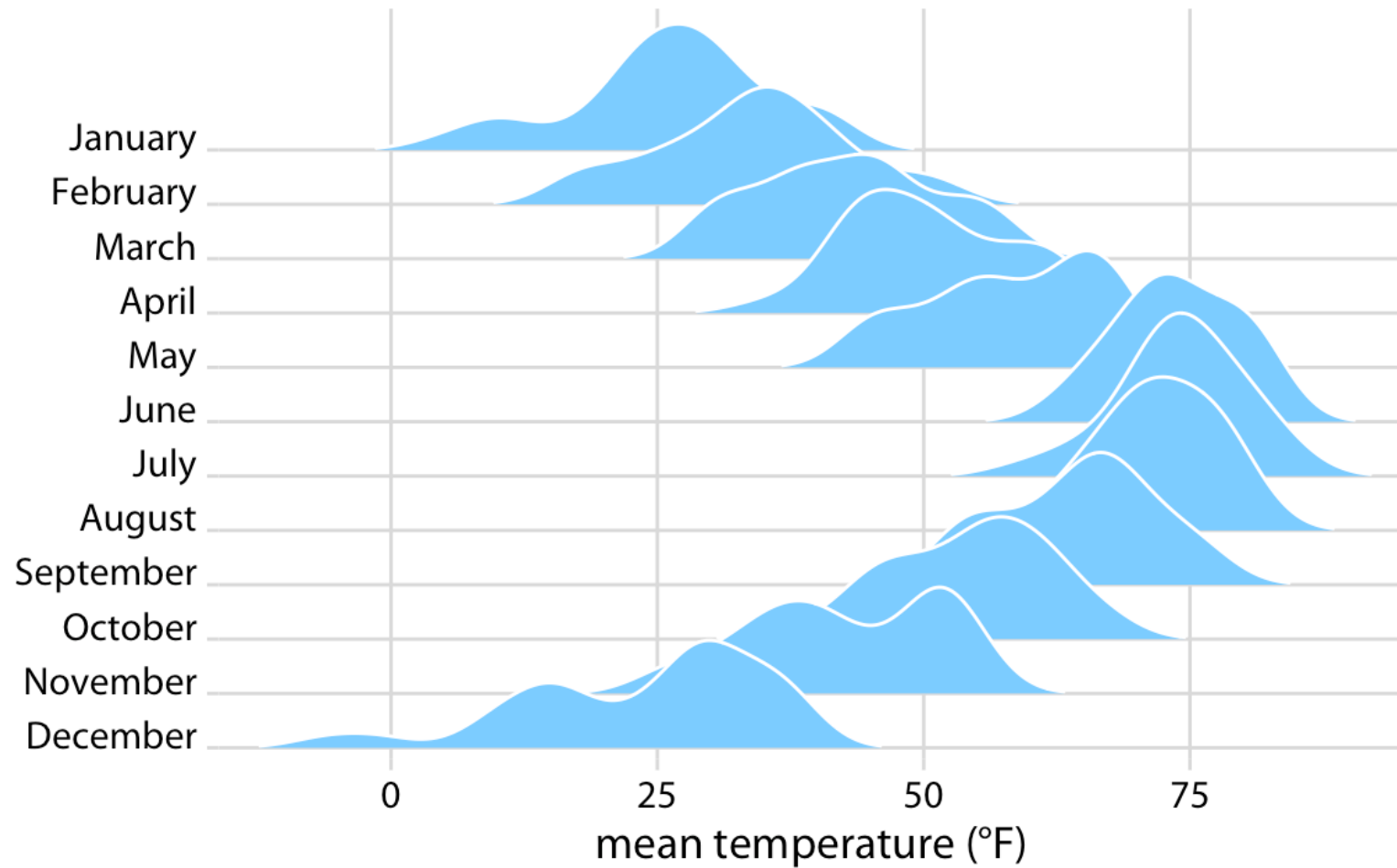
► Distributions



Source: Fundamentals of Data Visualization by Claus O. Wilke

Example

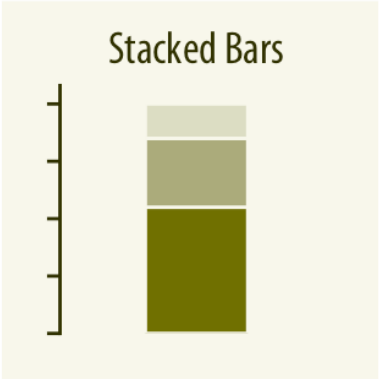
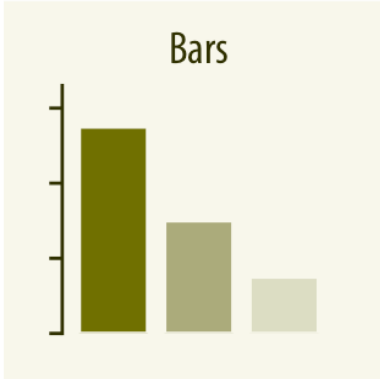
► Distributions



Source: Fundamentals of Data Visualization by Claus O. Wilke

A directory of visualizations

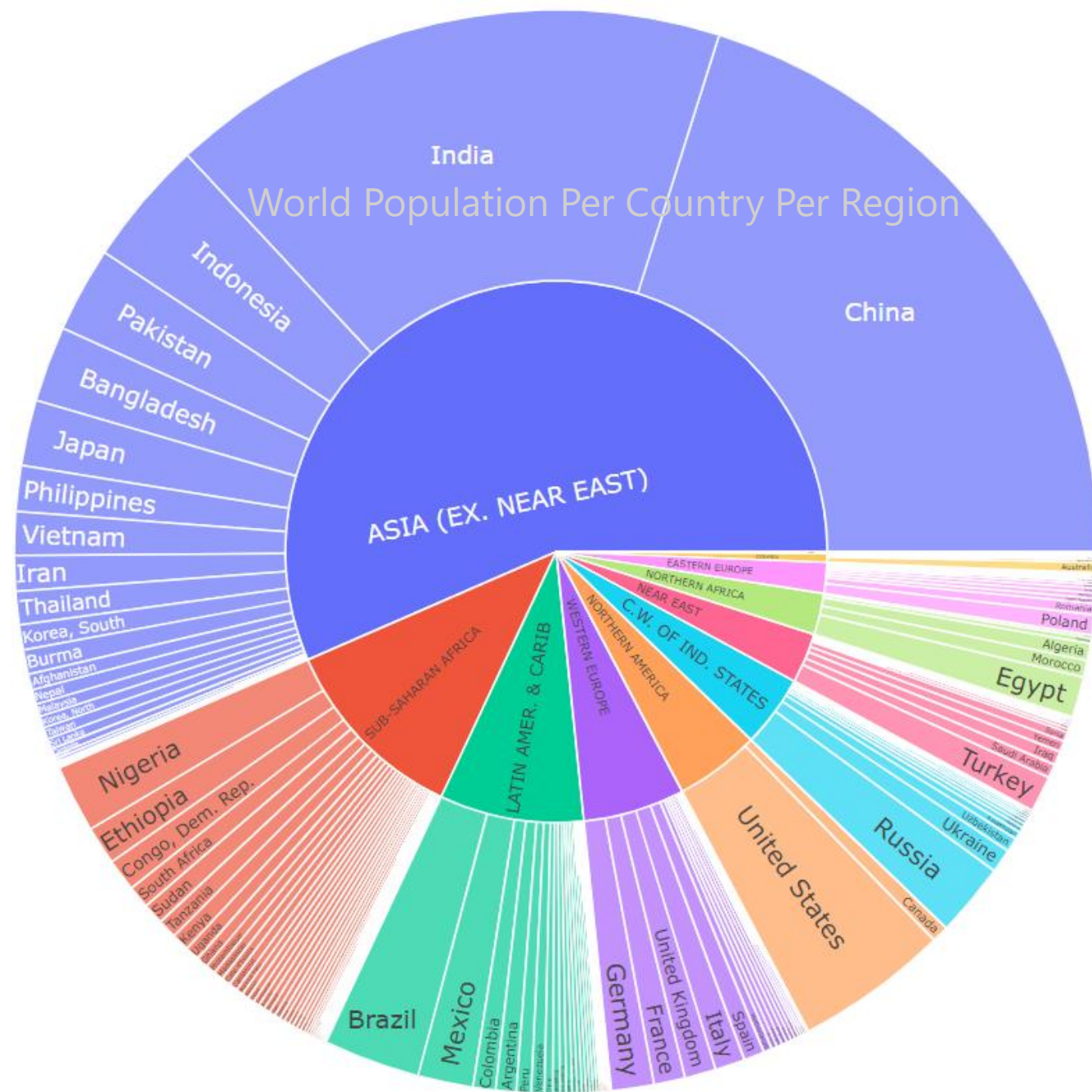
► Proportions



Source: Fundamentals of Data Visualization by Claus O. Wilke

Example

► Proportions: Sunburst chart to capture hierarchical information



Data source: Countries of the world (as used in Module 2)

A directory of visualizations

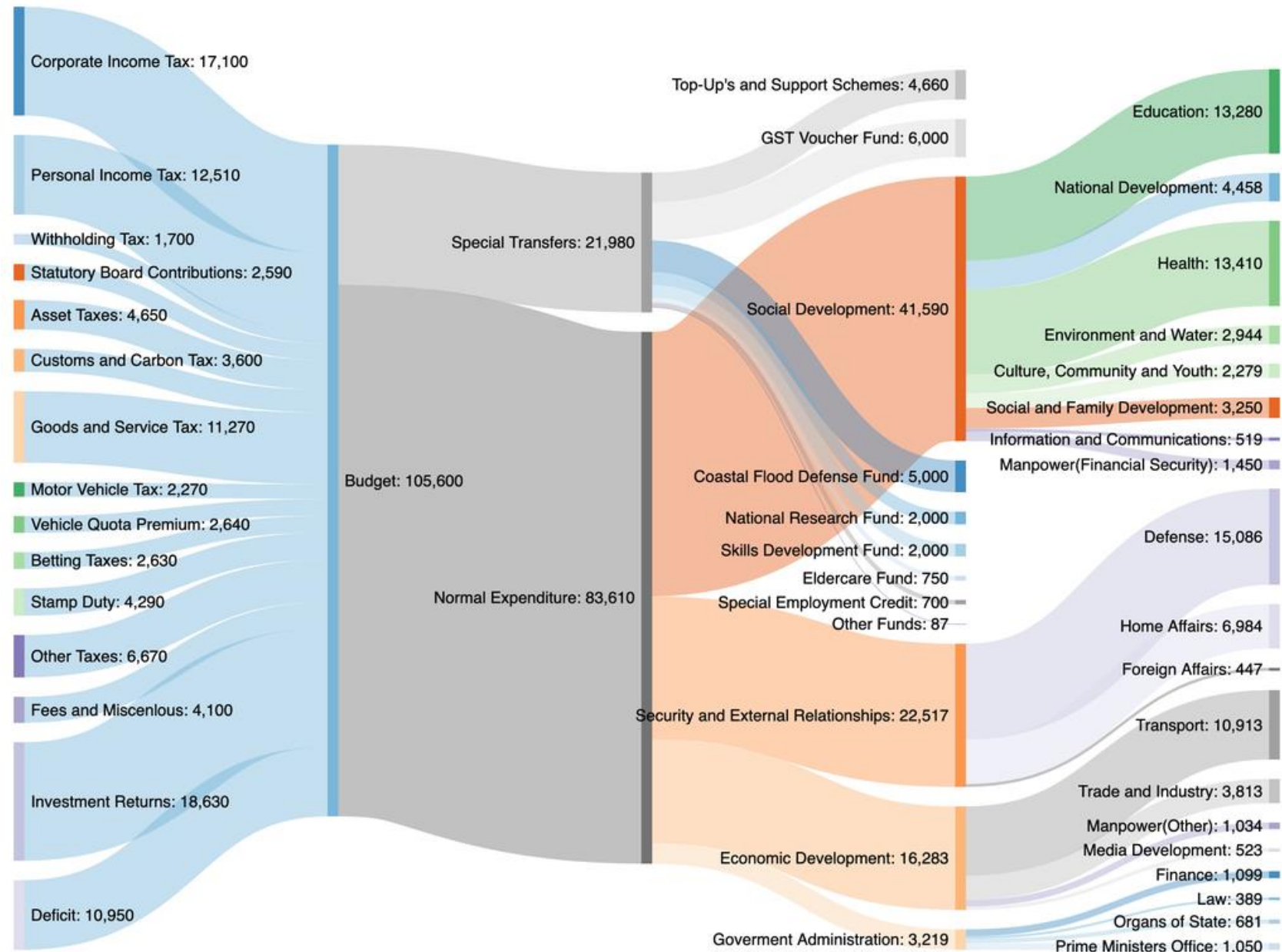
► Proportions



Source: Fundamentals of Data Visualization by Claus O. Wilke

Example

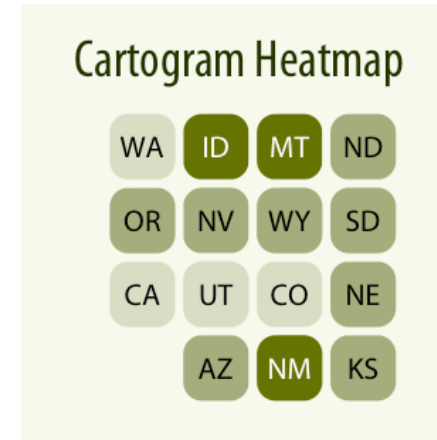
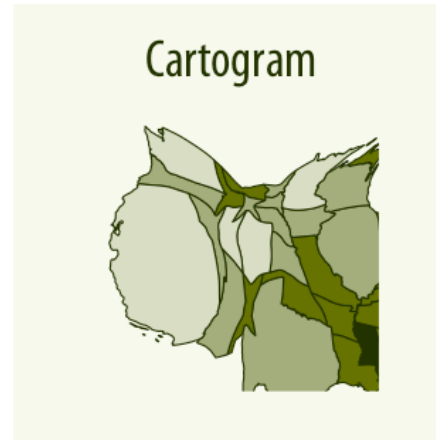
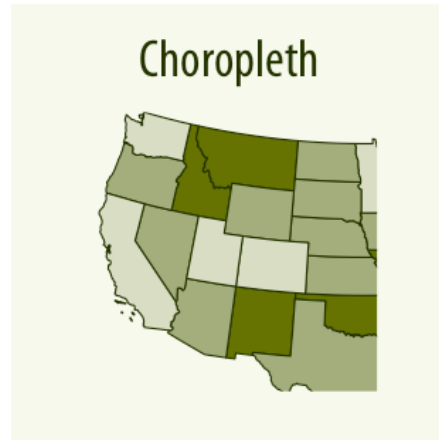
SG Government's Projected Revenue and Expenditure 2020 (In Millions of SGD)



Source: https://www.reddit.com/r/dataisbeautiful/comments/f62y0g/ocsingapore_governments_projected_revenue_and/

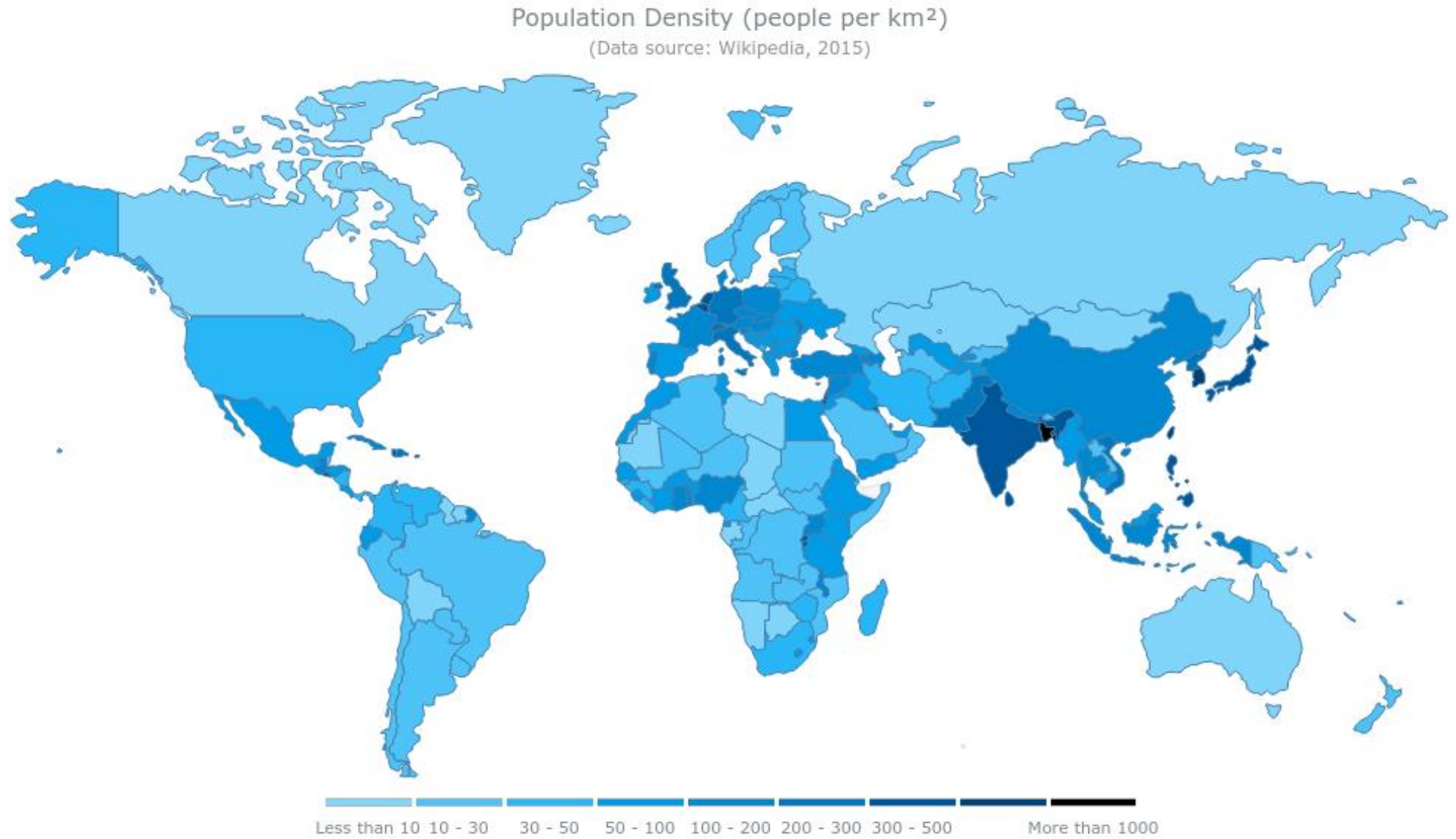
A directory of visualizations

▶ Geospatial data



Source: Fundamentals of Data Visualization by Claus O. Wilke

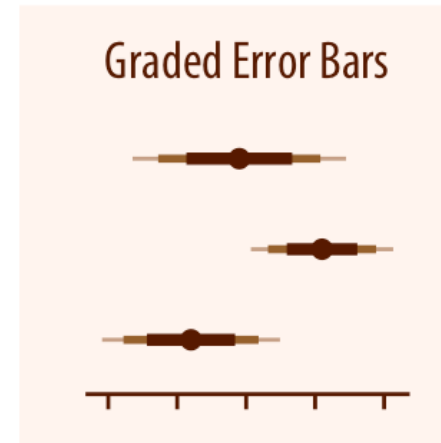
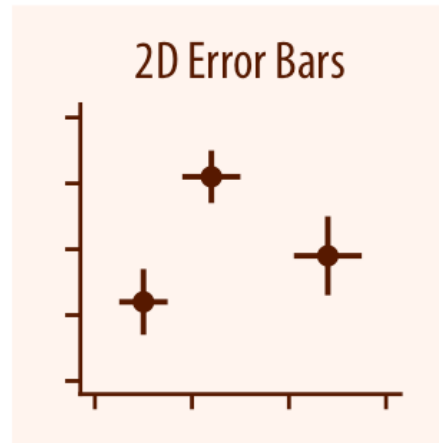
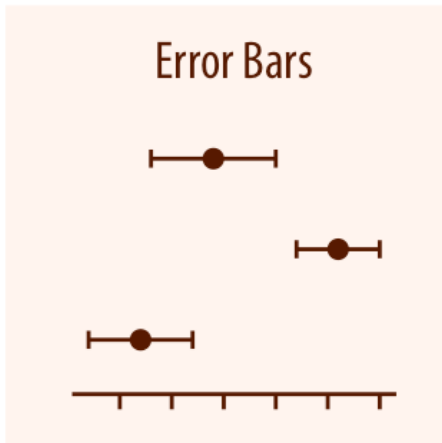
Example



Source: https://www.anychart.com/products/anymap/gallery/Maps_General_Features/World_Choropleth_Map.php

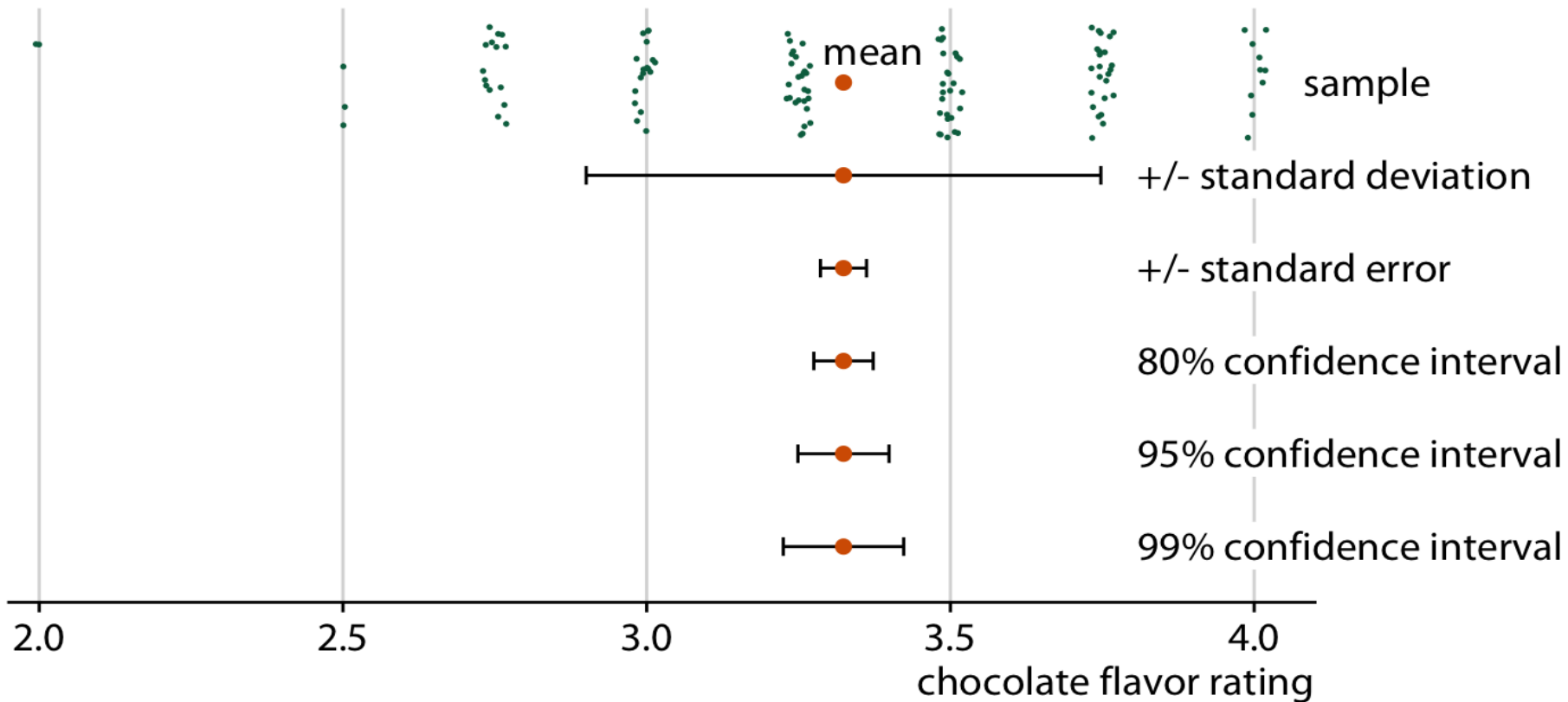
A directory of visualizations

► Uncertainty



Source: Fundamentals of Data Visualization by Claus O. Wilke

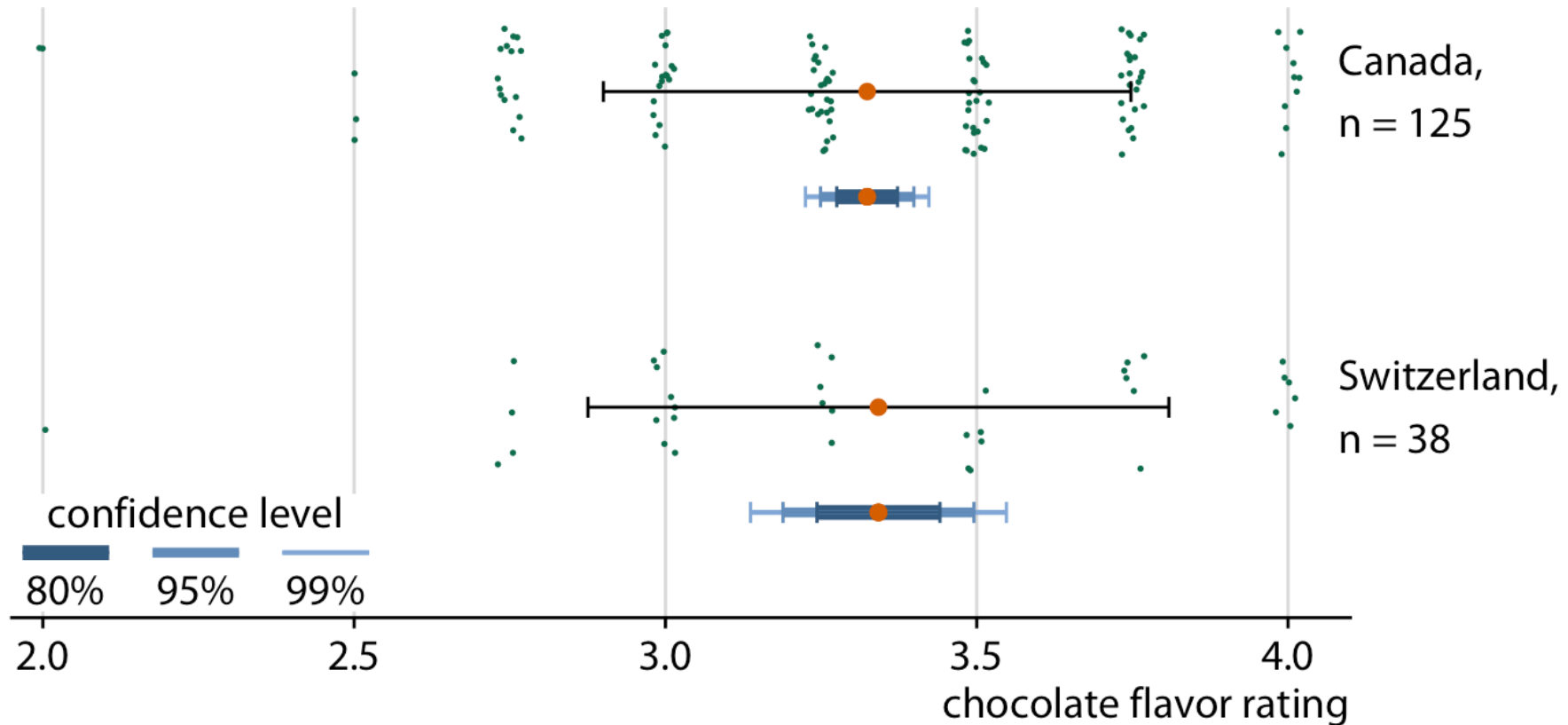
Example



$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$
$$\frac{\sigma}{\sqrt{n}}$$

Source: Fundamentals of Data Visualization by Claus O. Wilke

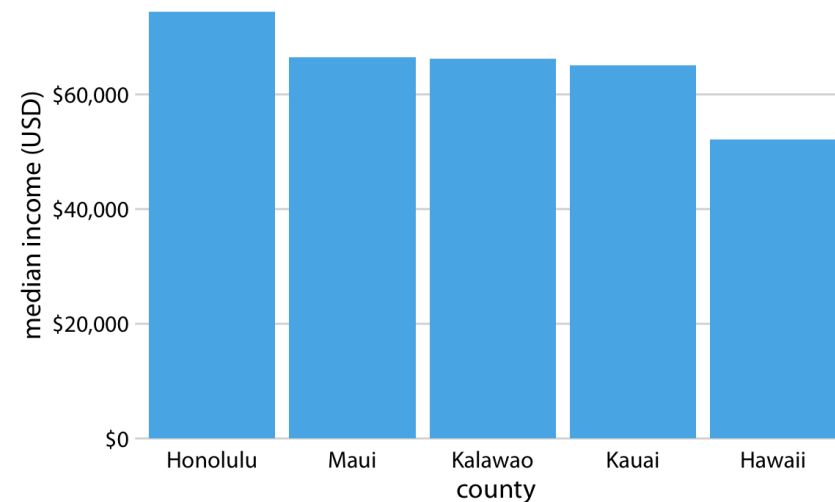
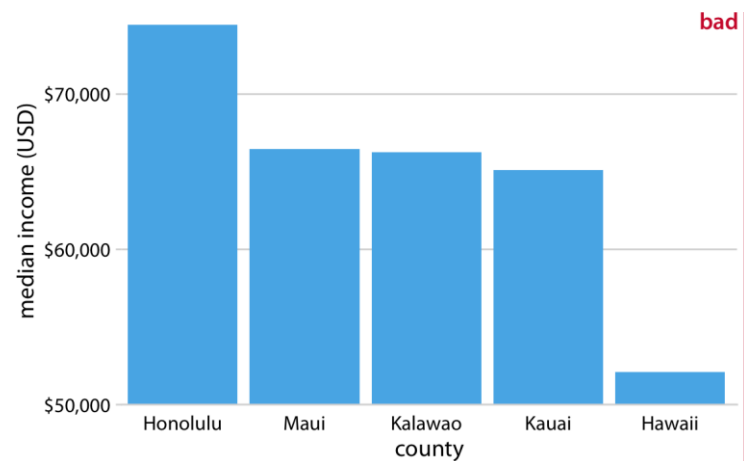
Example



Source: Fundamentals of Data Visualization by Claus O. Wilke

Principle of proportional ink

- ▶ The sizes of shaded areas in a visualization need to be proportional to the data values they represent

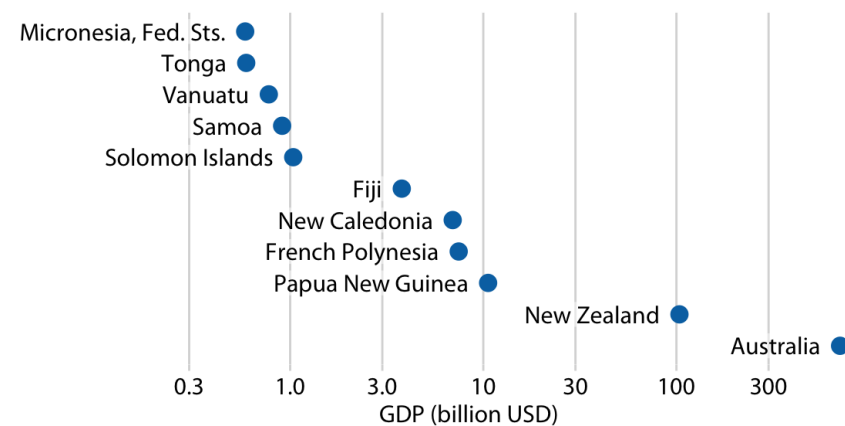
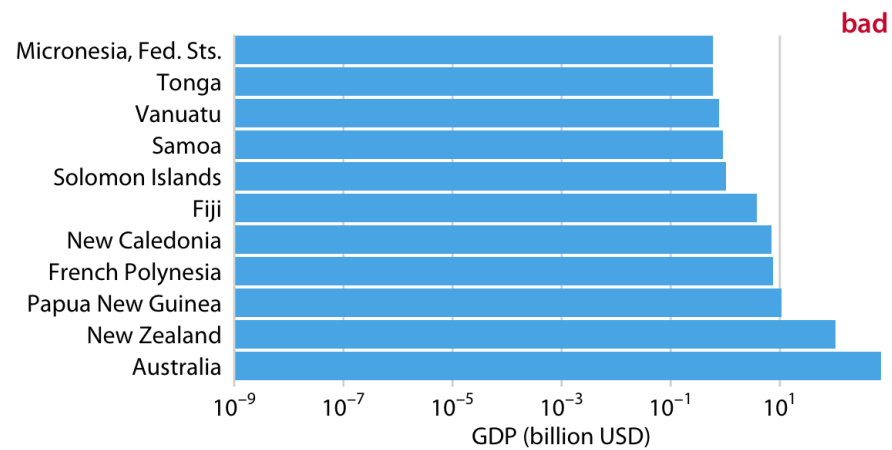


Example with a linear scale

Source: Fundamentals of Data Visualization by Claus O. Wilke

Principle of proportional ink

- ▶ The sizes of shaded areas in a visualization need to be proportional to the data values they represent



Example with a logarithmic scale

Source: Fundamentals of Data Visualization by Claus O. Wilke

Accessibility: Mind the colorblind people



▶ 3(+1) principles of Color Universal Design

- Choose color schemes that can be easily identified by people with all types of color vision
- Use not only different colors but also a combination of different shapes, positions, line types and coloring patterns, to ensure that information is conveyed to all users
- Clearly state color names where users are expected to use color names in communication
- + Moreover, aim for visually friendly and beautiful designs

This material on accessibility for the colorblind people is based on: <https://jfly.uni-koeln.de/color/>

Accessibility: Mind the colorblind people



▶ How colorblind people see colors?



Non color blind



Protanope
(red cone cells defective)



Deuteranope
(green cone cells defective)



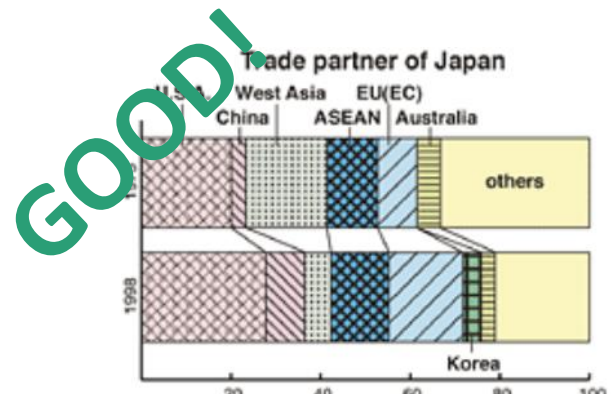
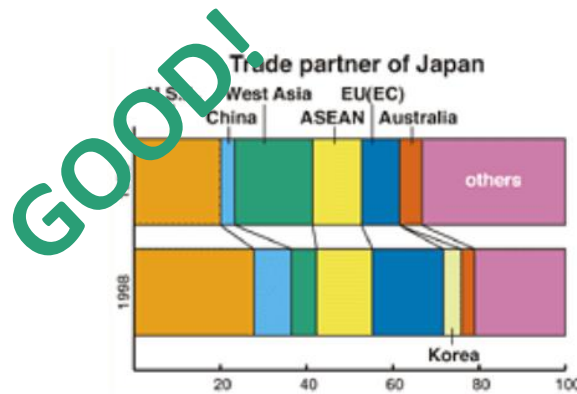
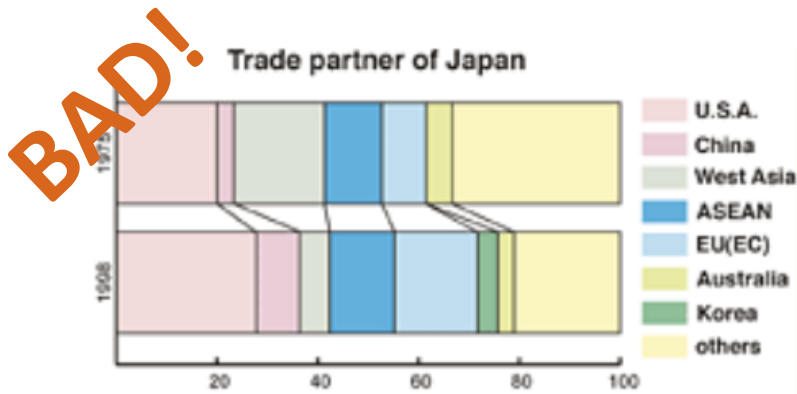
Tritanope
(blue cone cells defective)

This material on accessibility for the colorblind people is based on: <https://jfly.uni-koeln.de/color/>

Accessibility: Mind the colorblind people



- ▶ Depending on the category difficulty differentiating different combinations of color-pairs



Though Tufte discouraged
This sort of fill!

This material on accessibility for the colorblind people is based on: <https://jfly.uni-koeln.de/color/>

Accessibility: Mind the colorblind people



► Set of unambiguous colors for everyone!

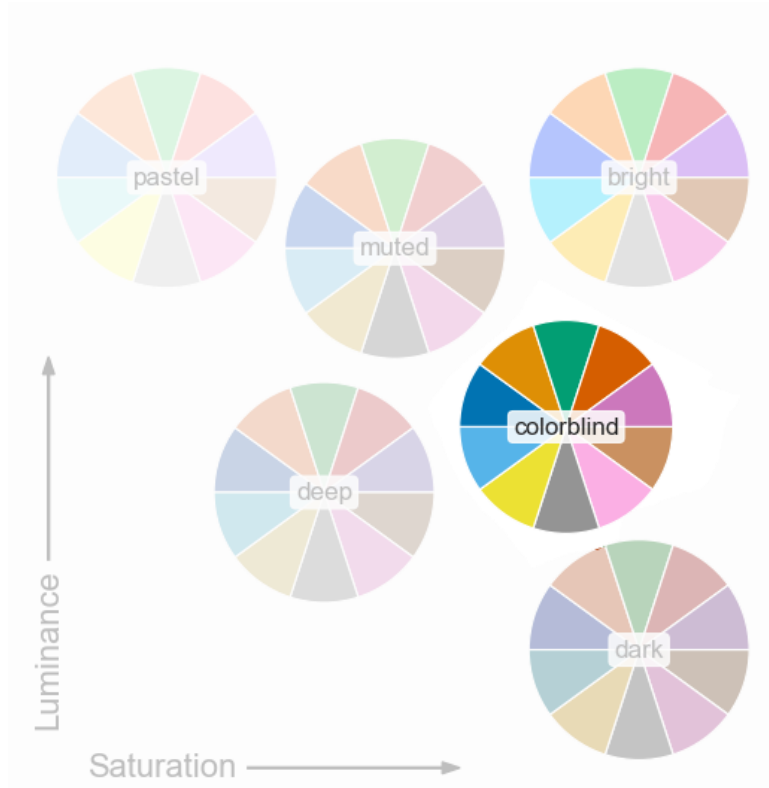
| Original | Simulation | | | Hue | C,M,Y,K (%) | R,G,B (0-255) | R,G,B (%) | |
|----------|------------|--------|--------|----------------|-------------|---------------|---------------|------------|
| | Protan | Deutan | Tritan | | | | | |
| 1 | | | | Black | -° | (0,0,0,100) | (0,0,0) | (0,0,0) |
| 2 | | | | Orange | 41° | (0,50,100,0) | (230,159,0) | (90,60,0) |
| 3 | | | | Sky Blue | 202° | (80,0,0,0) | (86,180,233) | (35,70,90) |
| 4 | | | | bluish Green | 164° | (97,0,75,0) | (0,158,115) | (0,60,50) |
| 5 | | | | Yellow | 56° | (10,5,90,0) | (240,228,66) | (95,90,25) |
| 6 | | | | Blue | 202° | (100,50,0,0) | (0,114,178) | (0,45,70) |
| 7 | | | | Vermillion | 27° | (0,80,100,0) | (213,94,0) | (80,40,0) |
| 8 | | | | reddish Purple | 326° | (10,70,0,0) | (204,121,167) | (80,60,70) |

This material on accessibility for the colorblind people is based on: <https://jfly.uni-koeln.de/color/>

Accessibility: Mind the colorblind people



- ▶ Some plotting tools have in-built suitable color palettes



Source: https://seaborn.pydata.org/tutorial/color_palettes.html

Ungraded task 4.4: Create some visualizations using the `countries of the world` dataset (following the principles we explored in this module), to expose the resulting data you had obtained after cleaning it as per Module 2, ungraded task 2.2.

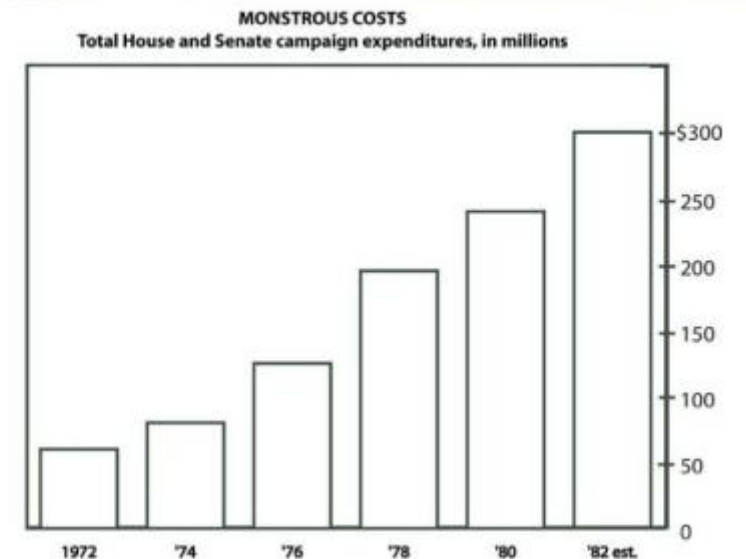
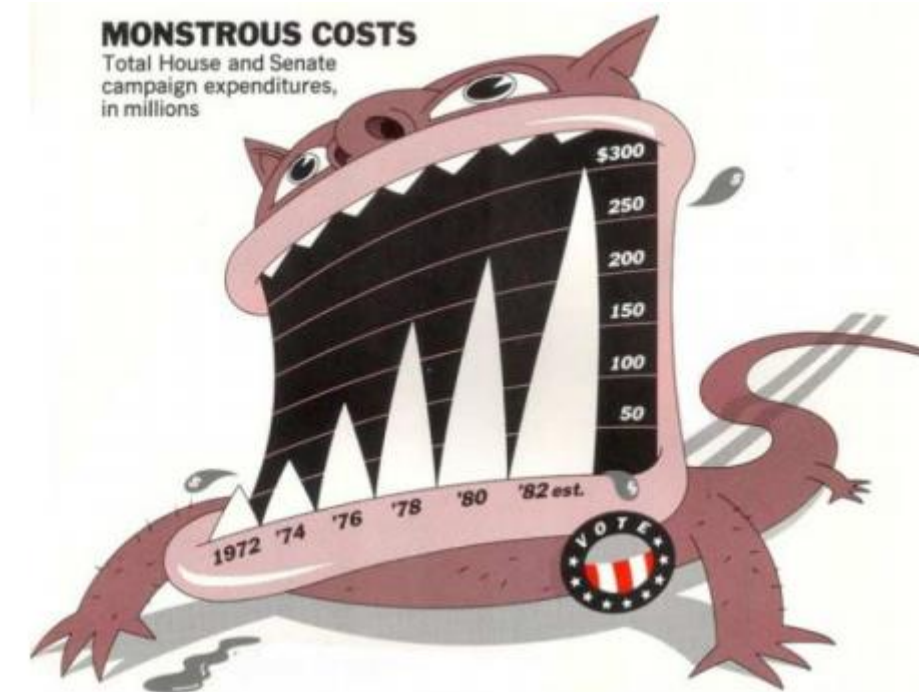
Share your code & results, and discuss which all principles you tried to follow, and how!

Recommended readings

Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts

Scott Bateman, Regan L. Mandryk, Carl Gutwin,
Aaron Genest, David McDine, Christopher Brooks

Department of Computer Science, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
scott.bateman@usask.ca, regan@cs.usask.ca, gutwin@cs.usask.ca,
aaron.genest@usask.ca, dam085@mail.usask.ca, cab938@mail.usask.ca



<http://www.stat.columbia.edu/~gelman/communication/Bateman2010.pdf>

Recommended readings

Color Universal Design (CUD) **- How to make figures and presentations that are friendly to Colorblind people -**

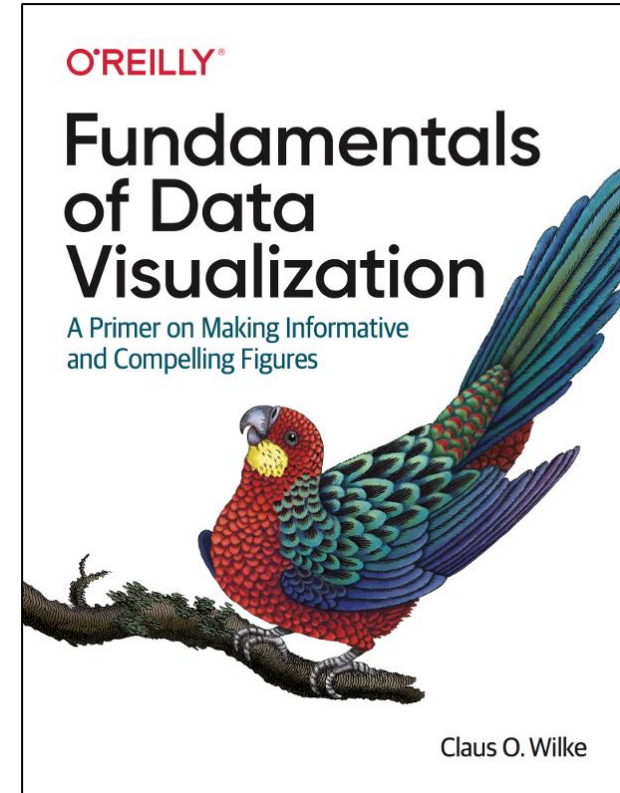
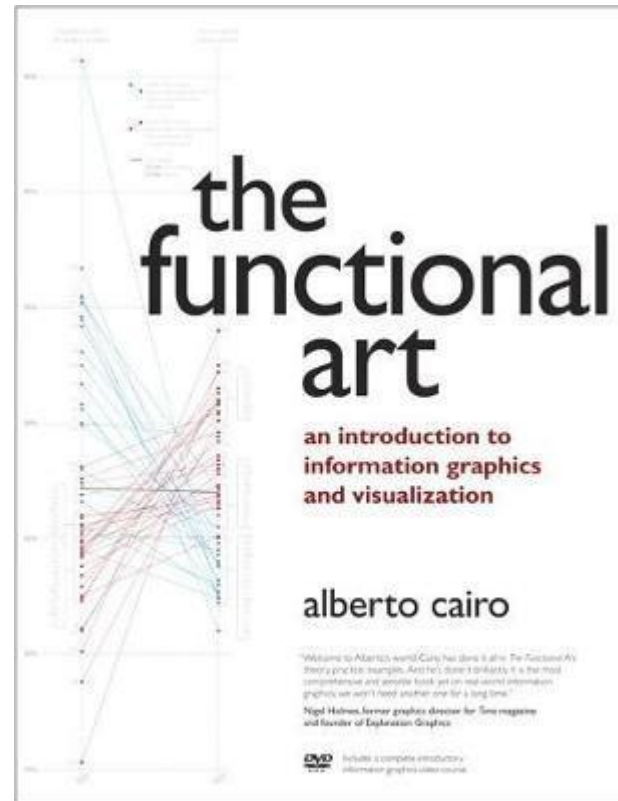
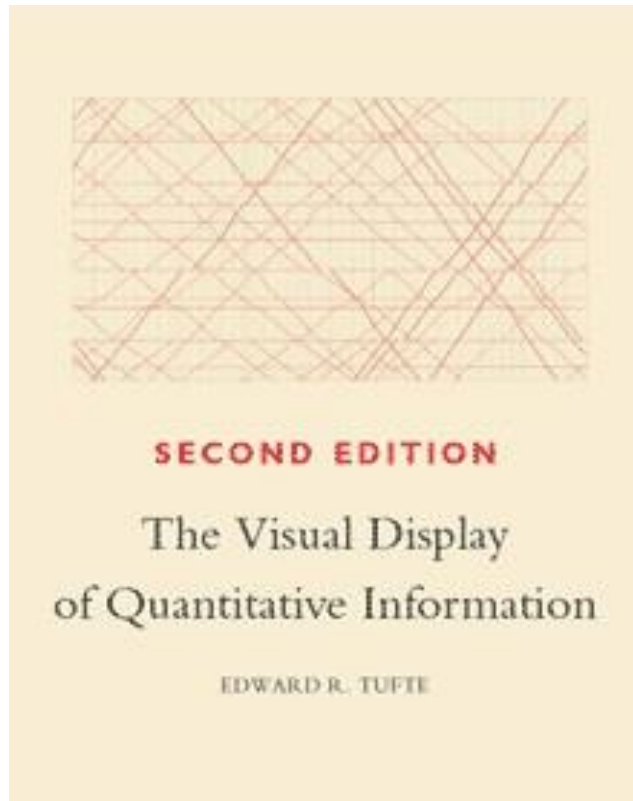
Masataka Okabe
Jikei Medial School (Japan)

Kei Ito
University of Tokyo, Institute for Molecular and Cellular Biosciences (Japan)



<https://jfly.uni-koeln.de/color/>

Selected References



<https://clauswilke.com/dataviz/>

A guide to choose your plot type

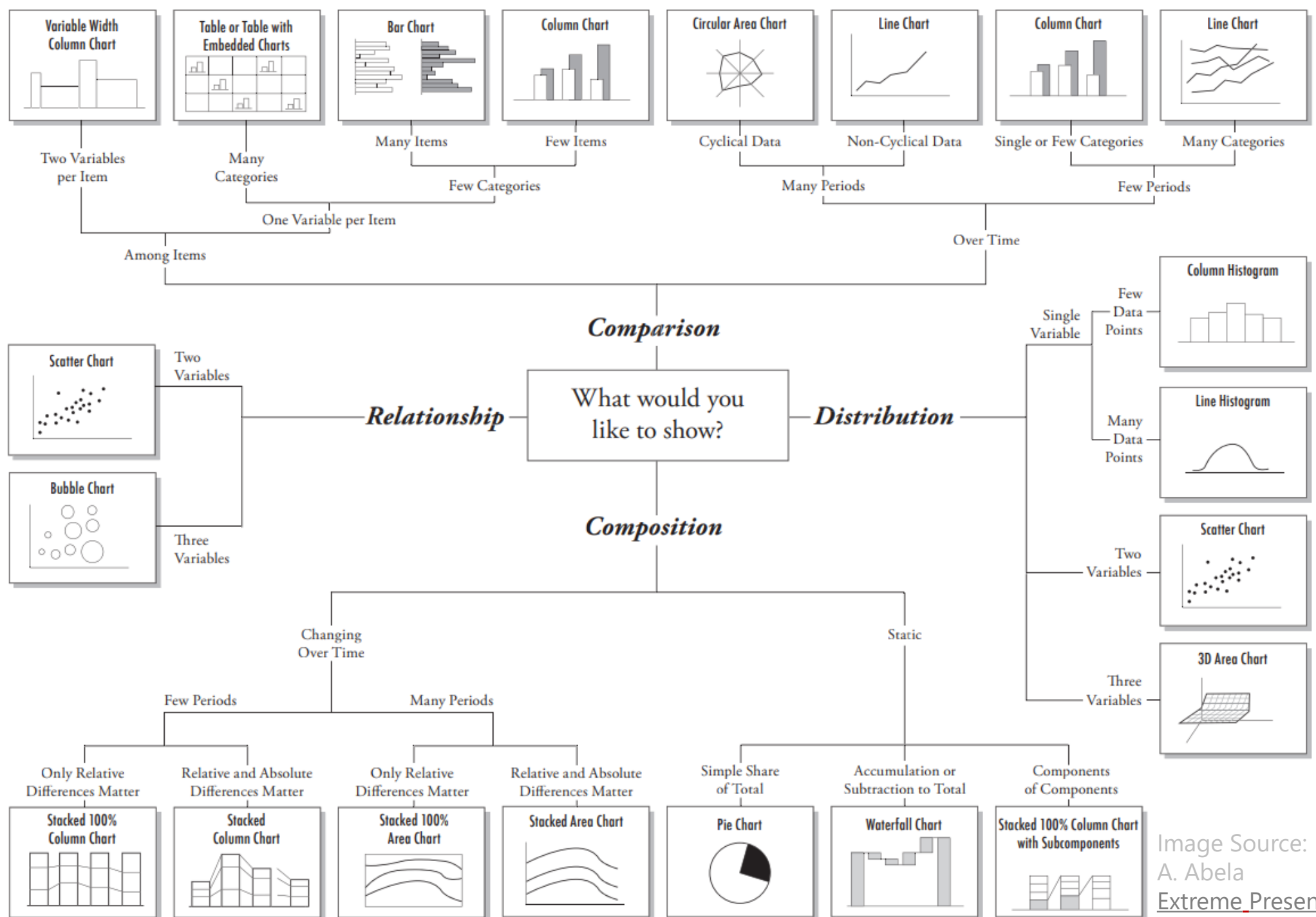


Image Source:
A. Abela
Extreme Presentation