# Principles of Data Visualization

Decoding
Handouts
March 12, 2025

Ronald Geskus; rgeskus@oucru.org,
Oxford University Clinical Research Unit
Centre for Tropical Medicine
Ho Chi Minh City, Viet Nam

.1

**Abstract.** Data visualisation is a key component in different phases of the research process. In the exploratory phase, a clever visualisation can provide important new insights. In the analysis phase, it can help you to understand the results. In the reporting phase, a visual presentation of results is often the most informative way to convey your message.

Over the last 10 to 20 years, many high quality newspapers and magazines have invested in setting up highly skilled graphics teams. Examples are the Financial Times and the New York Times. The scientific community is somewhat lacking behind, but the importance of visualisation is more and more recognized. One example is the creation of the STRATOS (STRengthening Analytical Thinking for Observational Studies) visualisation panel.

In the first part we focused on the process of graph creation. We introduced the "grammar of graphics", which defines the structure and components of a graph. The most popular graphics package in R, ggplot2, is based on this grammar.

The grammar of graphics provides a framework for graph construction, but it doesn't tell you what constitutes an informative graph of high quality. In this second part, we cover the basic principles of graph perception, largely based on the ideas from William Cleveland and Edward Tufte. By knowing these principles, in combination with a language to define graph components, you are better able to discuss the quality of a graph and pinpoint where it can be improved.

# Contents


**1 Intro** **3**

**2 Principles** **4**
2.1 Scale and range . . . . . 4
2.2 Perception . . . . . 9

**3 Decoding: William Cleveland** **11**
3.1 A hierarchy of visual decoding . . . . . 12
3.1.1 Size (numeric) . . . . . 13
3.1.2 Group (categorical) . . . . . 19

**4 Colour** **19**

**5 Exercises part 2** **22**

**6 Aesthetics: Tufte** **27**

**7 Final words** **31**


.2

# 1 Introduction

In the morning class we discussed points for consideration when creating a graph. We introduced the Grammar of Graphics, which provides a handle on discussing graph creation in relation to purpose and quality.

## Contents of this course

How to make an informative visual representation of data and/or results

- Part I: process of graph construction (encoding)
  - How to represent ("map") data/results in graph
  - The structure of a graph; the grammar of graphics
- *Part II: principles of graph perception and interpretation (decoding)*
  - *What happens when you see and read a graph*
  - *Principles for creating truthful and informative graphs*

.3

We mentioned the following points to consider when creating a graph.

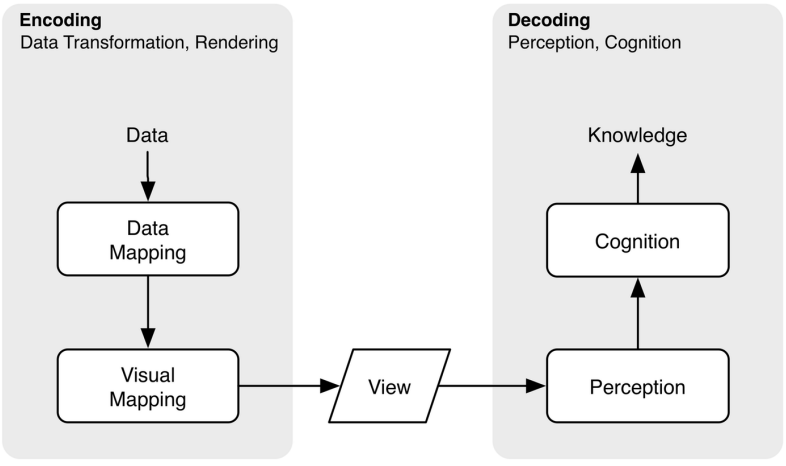
## Points to consider

- *Why* do we make a graph? (purpose)
  1. We want to explore and inspect our data or results
  2. Present data or results, often with a message, to others
- *Where* do we show the graph: research paper, conference, website
- *Who* is the audience. Level of background knowledge
- *What* do we want to show (data)
  - Graph: visual representation of information ("encoding")
    mapping from data/results to figure
  - Show values in isolation and/or focus on patterns

.4

Today we discuss some principles that help in improving the quality of your graphs. The quality of a graph is determined by how we perceive the graph and what we learn from it. It should have a clear and honest message.

## Creating (mapping) and perceiving (seeing and reading)




.5

# 2 Principles of Data Visualisation

## Criteria for graph quality

- Truthful
  - based on data collected in honest research
  - objective display of information
  - use appropriate scale and range
- Intuitive: easy to understand, concise without oversimplifying (see Section "Aesthetics: Tufte")
- Informative (see Section "Decoding: William Cleveland")
  - use appropriate graph type
  - use proximity and alignment to facilitate comparison
  - use labels and annotations to add clarity to the message
- Insightful: reveals evidence hard to observe otherwise (see Section "Aesthetics: Tufte")
- Visually appealing; but honesty and clarity come first

.6

## 2.1 Scale and range

## Scales on the axes, some principles

- Linear scale: values as measured
  - Most easy interpretation of values

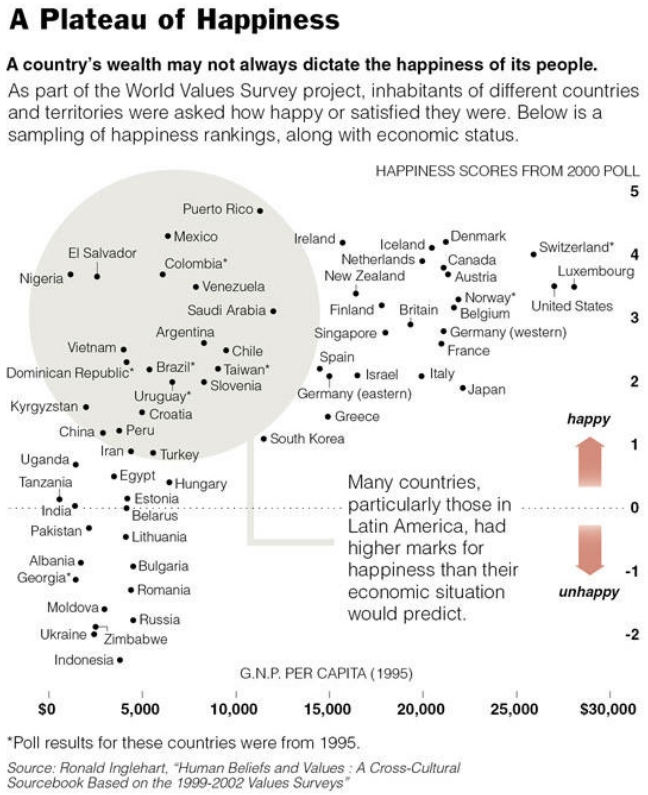
  But:
  - Distribution may be very skewed
  - May give wrong impression of structure
- Log scale
  - Distance reflects percentage change: $\log(x_2) - \log(x_1) = \log(x_2/x_1)$ e.g. $\log_2(2500) - \log_2(1250) = \log_2(2) = 1$ unit on scale
  - Can still use original values as labels
  - Always use log scale for ratios
- Range: is zero included? Does zero have a meaning?
- Categorical data: ordering by i) alphabet? ii) value? iii) subgroup? (e.g. disease type, region)

.7

The use of transformed scales on the axes can completely change the impression of the pattern in the data. An interesting example is the graph that appeared in the New York Times[1].

[1] See https://archive.nytimes.com/www.nytimes.com/imagepages/2005/10/03/science/20051004_HAPP_GRAPHIC.html

## Choice of scale influences interpretation

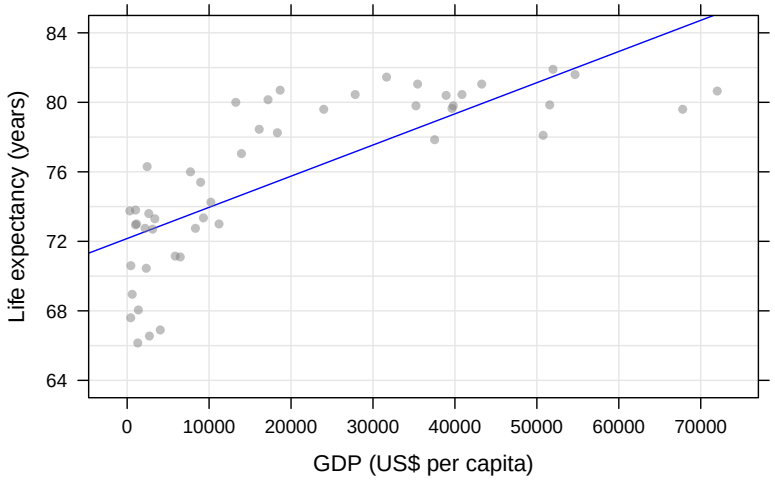



.8

The claim was that "Many countries, particularly those in Latin America, had higher marks for happiness than their economic situation would predict." The author uses a linear scale and tries to distill an interpretation that probably doesn't make sense. If a logarithmic scale had been used for the x-axis, the picture would have been quite different and probably the relation had become quite linear.

A similar example is a WHO graph of life expectancy against Gross Domestic Product (GDP) to Life Expectancy in 2007.

## Example: WHO European Health Report, 2007

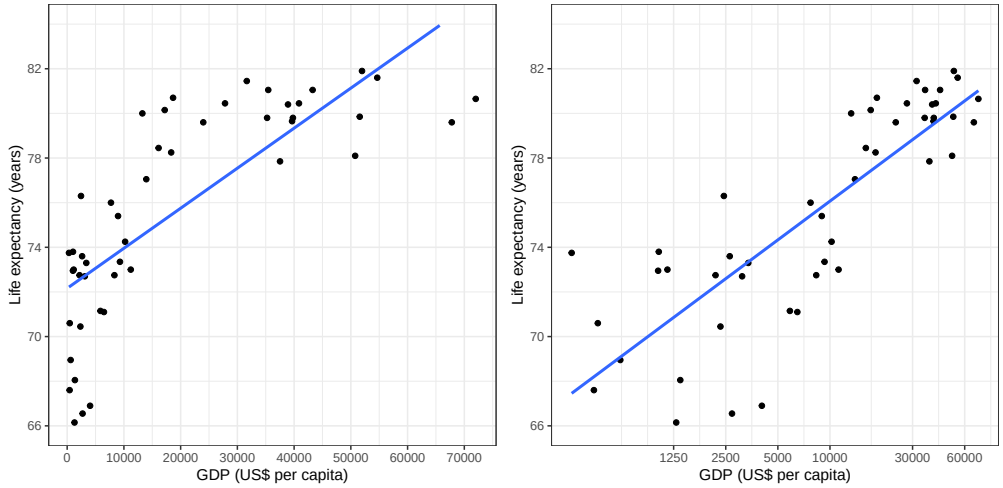



.9

We repeat the graph in slide 9, but show the plot with a log-transformed x-axis alongside. Now the trend looks quite linear; each doubling of GDP leads

to the same increase in happiness. The equidistant values 1250, 2500, 5000 and 10,000 each relate to a doubling of GDP compared to the previous one. Hence, according to the regression line, each doubling of GDP gives a similar number of years increase in life expectancy. Note that the position of the points and the trend do not depend on the base of the logarithm that is chosen $\log_2$ or $\log_{10}$ or natural log.

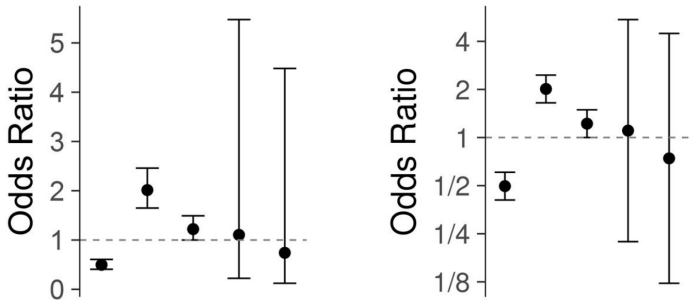
## WHO data 2007, two scales




.10

*Always use a log scale for ratios.* What's wrong with the linear scale for ratios? The problem is that a ratio (such as the odds ratio or relative risk) compares the probability $P_A$ of an outcome in group A with the probability $P_B$ in group B as a ratio $P_A/P_B$. Which group is allocated to A and which to B is our decision. The impression that is given in our visualisation should not depend on this decision. If we swap both groups, an OR of 2 turns into an OR of $1/2 = 0.5$, but both should be as far from an OR of 1. Also, the uncertainty in this value, as quantified by the confidence interval, shouldn't change. In the figure of slide 11 we see that this is the case if we use the logarithmic scale, but not if we use the linear scale.

## Always use a log scale for ratios




$\big(P$: probability outcome; odds $= \frac{P}{1-P}$;

odds ratio $\text{odds}(A)/\text{odds}(B) = \frac{P_A}{1-P_A} / \frac{P_B}{1-P_B}\big)$

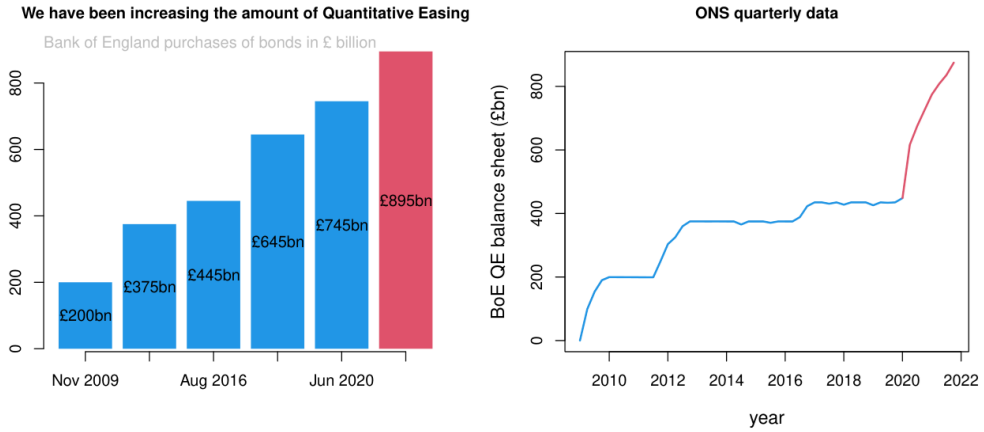
If $\text{OR(F/M)} = \frac{\text{odds(F)}}{\text{odds(M)}} = 2$, then $\text{OR(M/F)} = \frac{\text{odds(M)}}{\text{odds(F)}} = 0.5$

.11

Using an incorrect scale can also be done by purpose, in order to suggest patterns that do not exist. One example is the downplaying of the quantitative easing

measures taken to contain the economic consequences of the pandemic[2]. The left panel of the figure below was shown on the Bank of England website (but it was removed in late 2002). The scale chosen is not linear, not logarithmic, but serves to suggest that the measures were part of an ongoing trend.

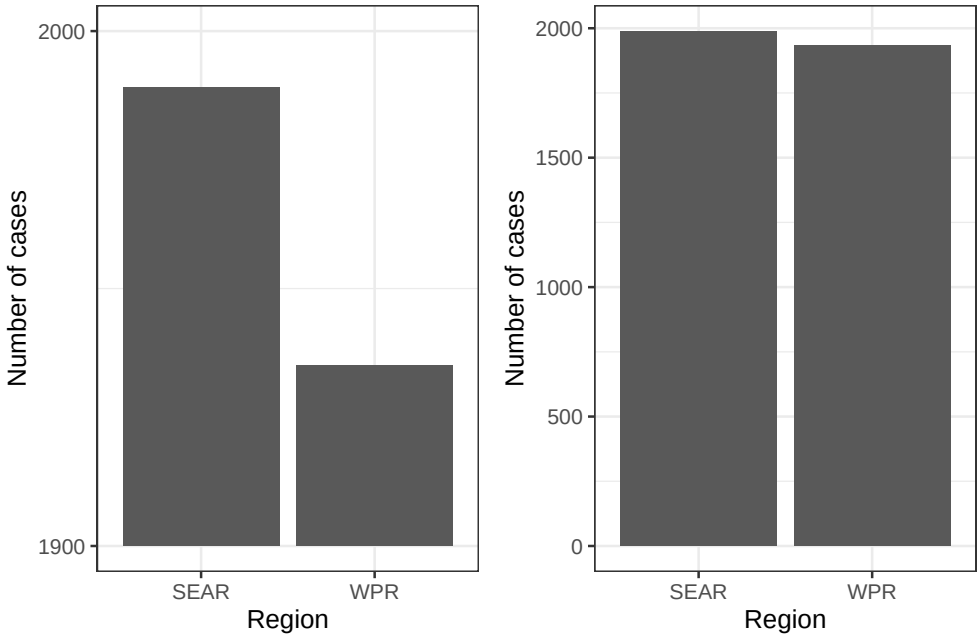
## Change in Quantitative Easing




.12

Another important characteristic is the range that is shown. The graph can suggest large differences by narrowing down the range. In slide 13 (we will come back to this graph later) we plot the number of measles cases in one month for two WHO regions. We should have the y-axis start at zero, otherwise the small difference between these numbers will be overemphasized.

## Include zero: measles cases in October 2019, two WHO regions



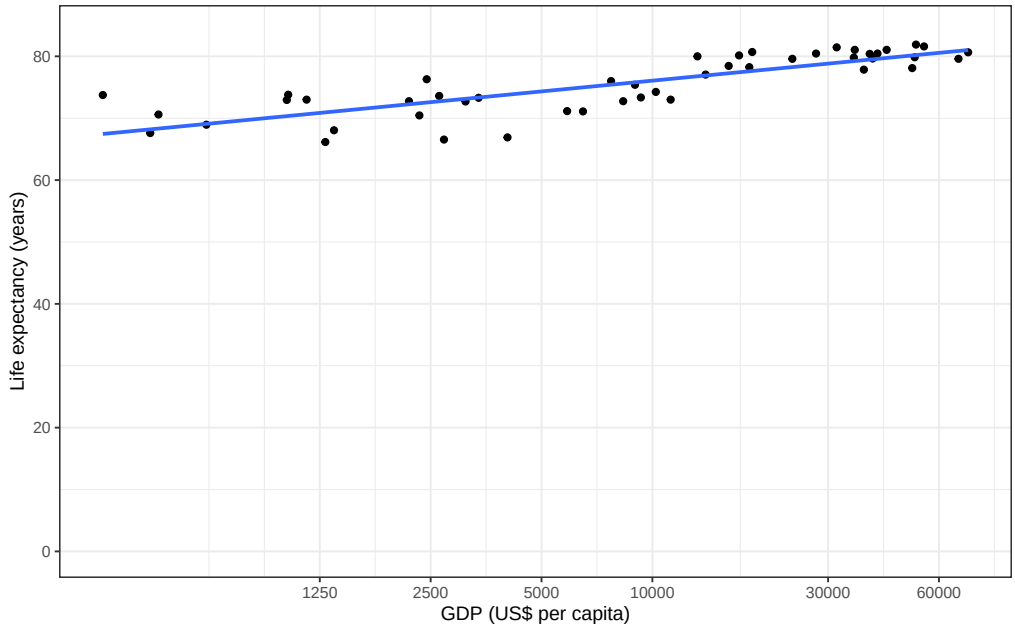

.13

However, in slide 14 there is no reason to include zero, because all life expectancies in the European region are above 60 years. Including zero will downplay

[2]See https://arxiv.org/abs/2409.06473

the trend.

## Don’t include zero: no country has zero life expectancy

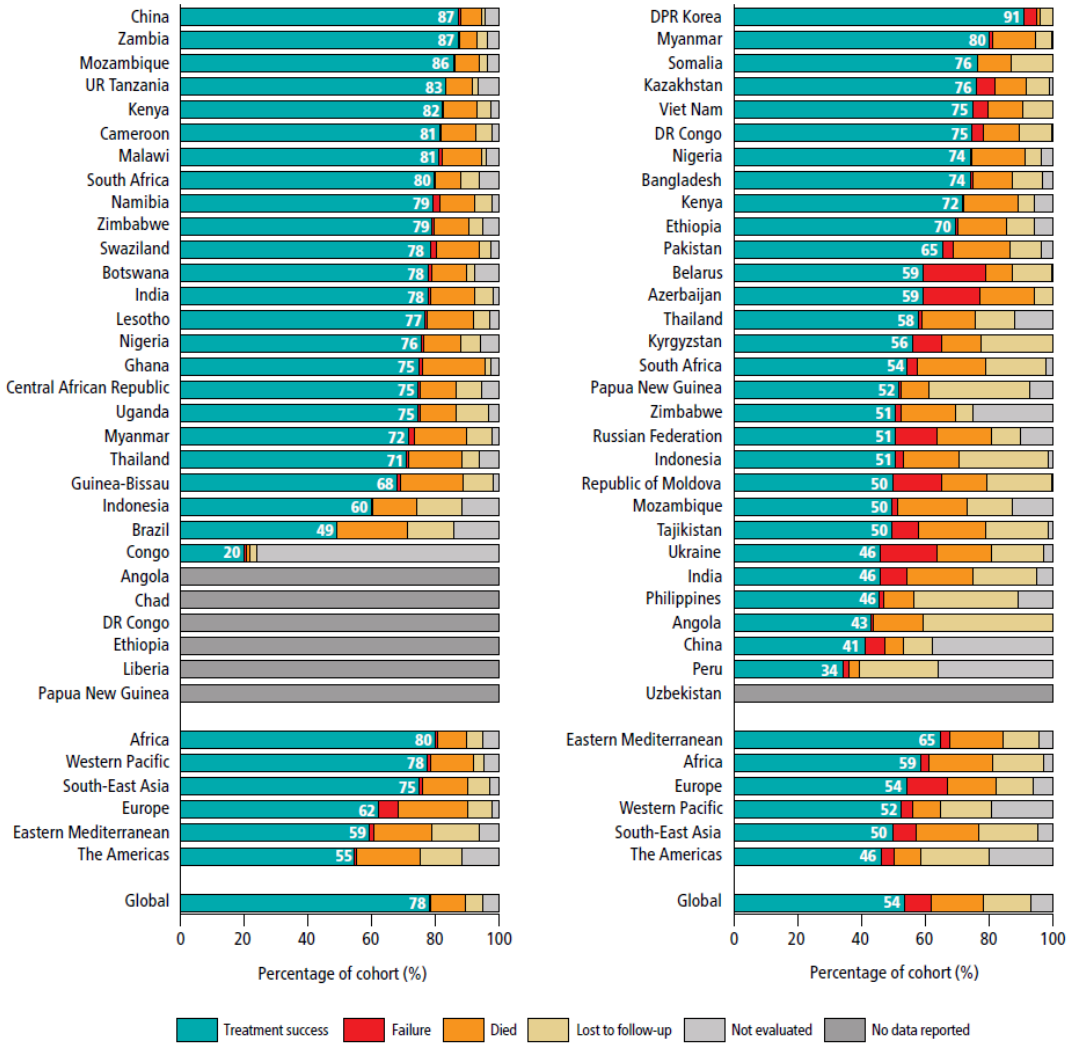



.14

Until now we talked about numbers along the scales. We can also have dates or categorical data along one of the axes. In slide 15, we see that the countries and regions have been ordered according to the percentage that was successfully treated. Later we’ll discuss whether this stacked bar chart is the best way to present the data.

**Treatment outcomes for new and relapse HIV-positive TB cases in 2015, 30 high TB/HIV burden countries, WHO regions and globally**

**Treatment outcomes for rifampicin-resistant TB cases started on treatment in 2014, 30 high MDR-TB burden countries, WHO regions and globally**




.15

## 2.2 Theories of perception

Daniel Kahneman describes two modes of thought in his book Thinking, Fast and Slow. System 1 is fast and intuitive, while system 2 is slow and logical. Those systems often give different results. In visual perception, there is a similar difference between "seeing" (intuitive) and "reading" (logical). A well designed graph makes first impression it gives correspond with the information that can be learned from the graph. An example of the difference between seeing and reading is in slide 27 with respect to the peak in measles infection.

## Two modes of perception: Seeing and Reading

- Daniel Kahneman: Thinking, Fast and Slow
  Fast: intuitive, "seeing"
  Slow: logical, "reading"
- What happens when you interpret a graph
  1. Graphical perception: first impression, *see*.
     "There was a peak in measles cases in the first half of 2019"
  2. Graphical cognition: draw conclusions, *read*.
     "The outbreak was mainly in the African region"
     "The number of cases in the South-East Asian region has declined over time"

.16

There is a lot of theory on human perception. The gestalt principles of design have relevance for the design of statistical graphs. In chapter 4 of his recent book "Being You", Anil Seth gives an interesting account of his recent insights on the nature of human perception.

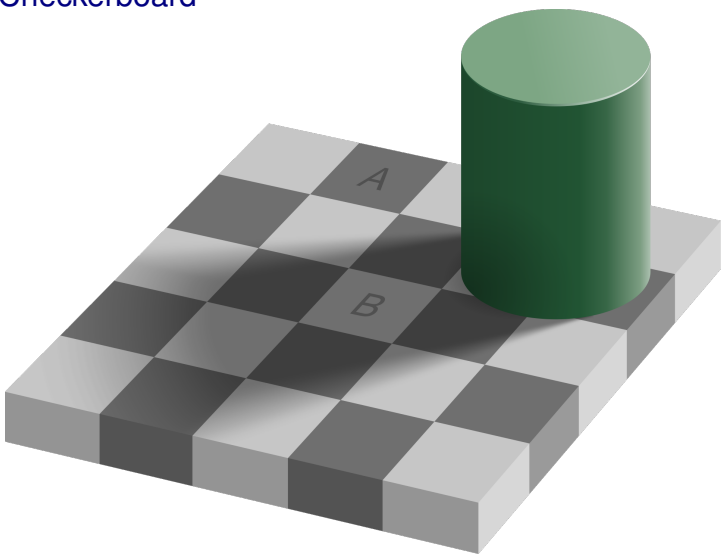
## Human perception

- Gestalt principles. See e.g. https://www.toptal.com/designers/ui/gestalt-principles-of-design
- Anil Seth: Being You, A New Science of Consciousness (2021)
  - Perception: sensory signals that are interpreted based on the brain's expectations or beliefs about their causes.
  - We never experience sensory signals themselves, we only experience interpretations of them, which are constructions of the brain
  - Example: white paper remains white when you move from indoors to outdoors, because your brain knows its whiteness cannot change

.17

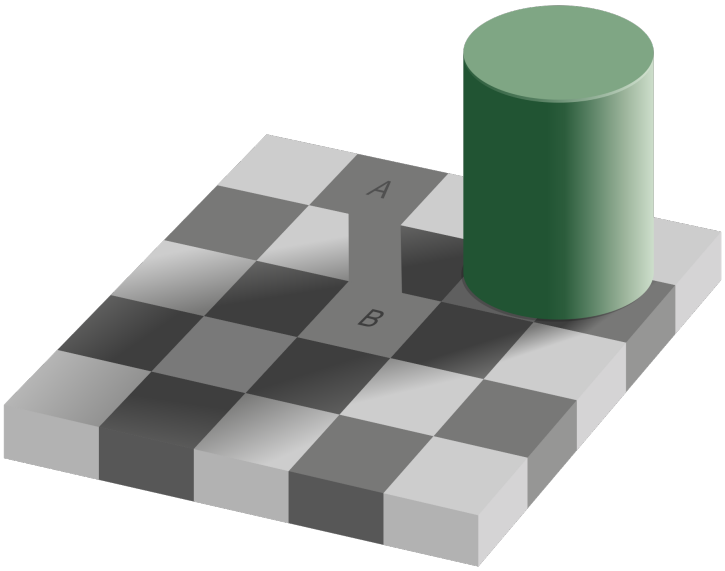
An example in which your interpretation in influenced by the surrounding information is Adelson's checkerboard. Which square is darker in slide 18, A or B? Does your answer change if you look at the checkerboard in slide 19? What happens if you cover the green cylinder in slide 19, so that it is no longer visible? We see that the perception of darkness is influenced by the brain's interpretation of the shadow that is cast by the cylinder.

### Adelson's Checkerboard




.18

### Adelson's Checkerboard (II)




.19

# 3 Decoding: William Cleveland

The book "The Elements of Graphing Data" by William Cleveland gives a list of principles that help in creating an informative graph, i.e. a graph that efficiently transfers information. These are principles to take into account when making the graph ("encoding" information), because they help when reading the graph ("decoding" information)[3].

[3] See also http://www.stat.auckland.ac.nz/~ihaka/120/Notes/ch05.pdf and http://hbiostat.org/doc/graphscourse.pdf for an overview of the principles of visual perception

### William Cleveland: The Elements of Graphing Data

- Which graph is better at accurately revealing relevant information
- Study on most efficient ways to transfer information from data to a graph. Guidelines on creating informative graphs
- Pattern recognition; assessment of relative magnitude of values *detection*, *assembly* and *estimation*

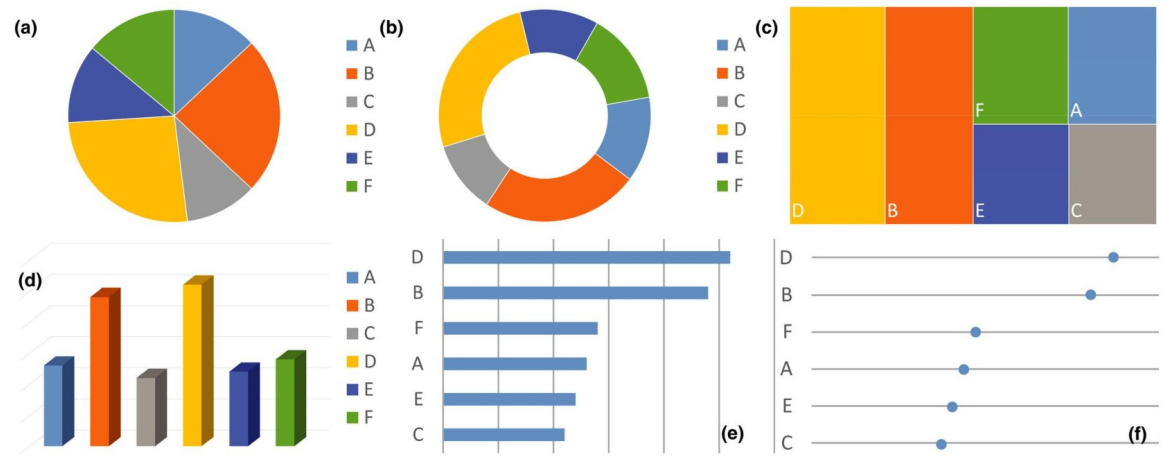
**detection** recognition of geometry (encoding of values via scales)

**assembly** discerning patterns; grouping

**estimation** assessment of relative magnitude of values ("how much more?")

.20

We will give a ranking of visual representations with respect to their ability to compare numbers or groups. Look at the figure in slide 21. Which of the six graph types (a) to (f) do you think is suited to compare the six numbers.




.21

## 3.1 A hierarchy of visual decoding

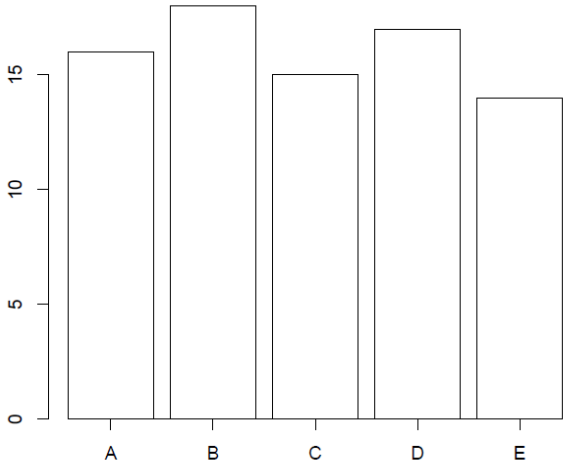
### Decoding

- Table look-up: attach value to what is shown; pattern recognition
- Size (numeric)
  In decreasing level of efficiency:
  1. Position along common scale
  2. Position along identical, nonaligned scales
  3. Length; grid lines are useful
  4. Angle, slope; don't use pie charts
  5. Area
  6. Volume; don't use unneccesary third dimension
  7. Colour
- Group (categorical)
  - Colour
  - Characters/symbols; "+" and "o" preferred for two groups
  - Linetypes (e.g. one thin black and one thicker gray line)

.22

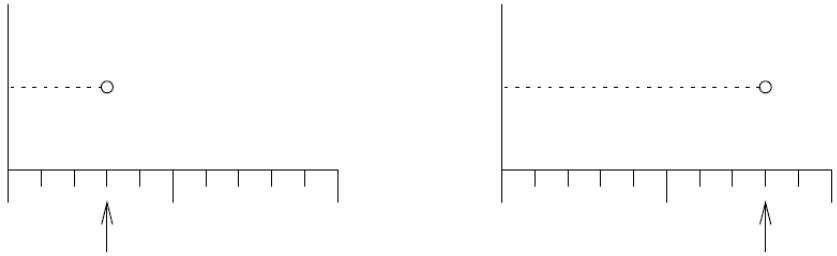
### 3.1.1 Size (numeric)

1. Position along common scale




.23

2. Position along identical, nonaligned scales

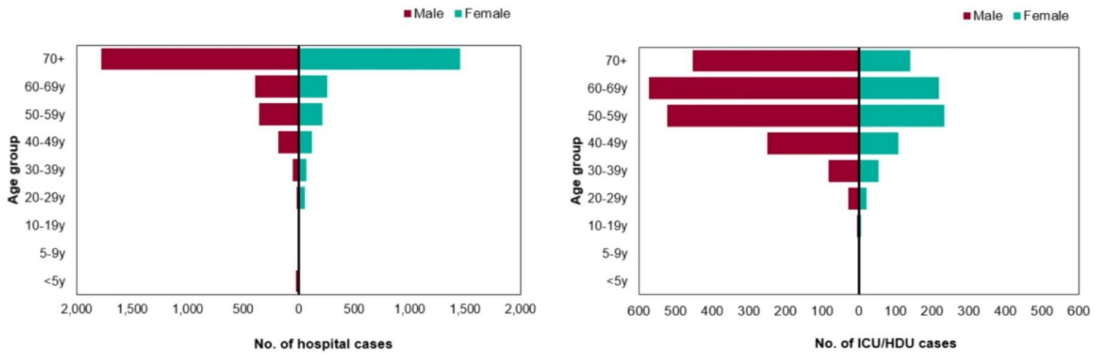


.24

An example of a non-aligned scale is a so-called back-to-back histogram.

ICU mortality in covid-19

Hospitalised cases (L) and ICU/HDU cases (R).
PHE, data from England.



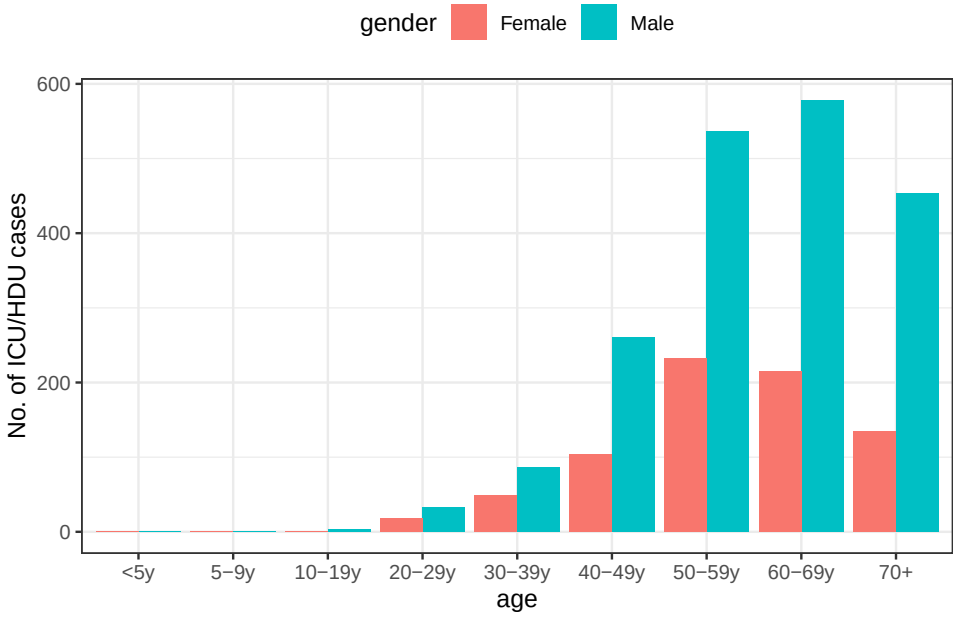

where men are much more likely to have severe disease

GRESHAM COLLEGE

.25

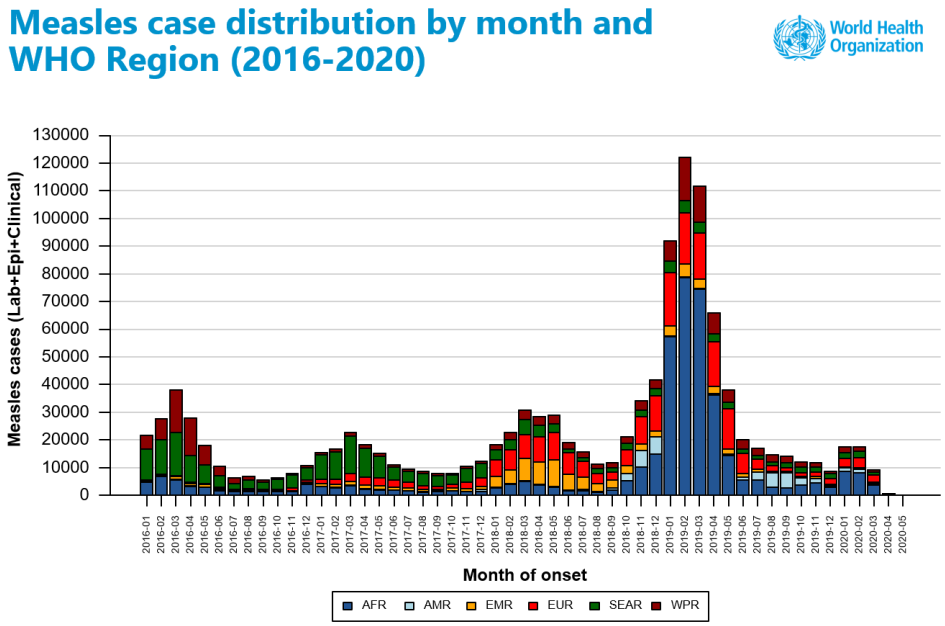
Slide 25 shows mortality from covid-19 is shown by age, separately for males and females[4]. This type of histogram is quite popular, especially in demographics when stratifying numbers by gender. There's one drawback however: it is somewhat difficult to compare the values of males and females. For this, a dodged barchart would be more suitable (slide 26).

## Alternative: dodged barchart




.26

## 3. Length



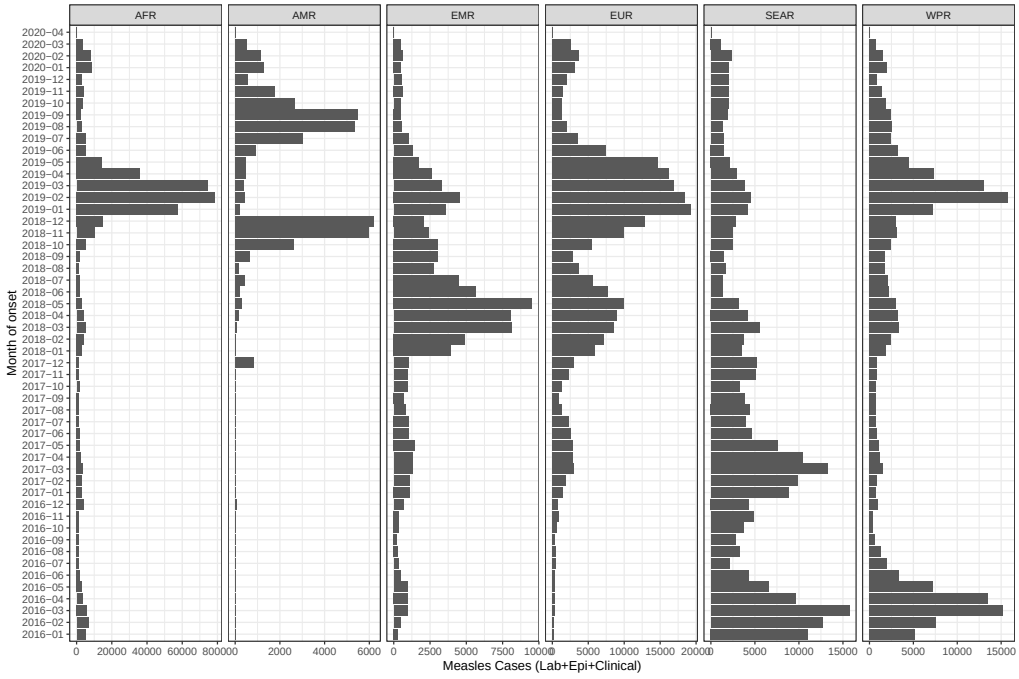

.27

The next one in the hierarchy is length. A common graph type in which length is the mode of presentation is the stacked bar chart. The figure in slide 27 shows the number of measles cases per month from January 2016 until April 2020 over

[4]Copied from the spring 2020 overview lecture of the SARS-CoV-2 pandemic by Chris Whitty https://www.gresham.ac.uk/lectures-and-events/covid-19

the six WHO regions[5]. What is easy to read from the figure are the total numbers over all regions as well as the numbers for the African region. Since the other bars are stacked, we can only use the length to compare regions in a month or to see trends over time within a region. For example, there were two peaks in the Western Pacific Region, in the beginning of 2016 and in the beginning of 2019. Which peak was higher? The same holds in the barchart of slide 15: it is difficult to tell which country or region has the highest percentage that died.

Patterns are much easier to observe if we position the numbers per region along identical, possibly non-aligned scales. For example, the figure in slide 28 is much better for observing trends over time within a region, and we observe that the peaks in the Western Pacific Region were about equal. Note that the range of the scale differs per region, which is preferred if we primarily want to compare trends within a region and not numbers over the regions.

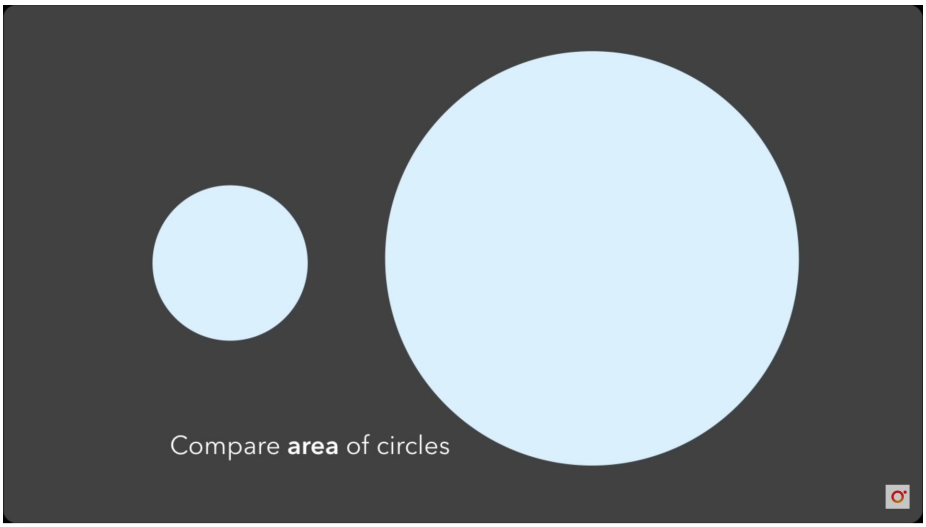



.28

4/5. Angle and area The next two figures are copied from the excellent presentation at https://www.youtube.com/watch?v=vc1bq0qIKoA. Compare the two areas in slides 29 and 30. How many times does the larger area fit into the smaller one? This is more difficult to answer for slide 29. The reason is that we can only compare areas in the first comparison, while in the second comparison we can also compare lengths. In fact, the answer to both is the same: one is seven times larger than the other.

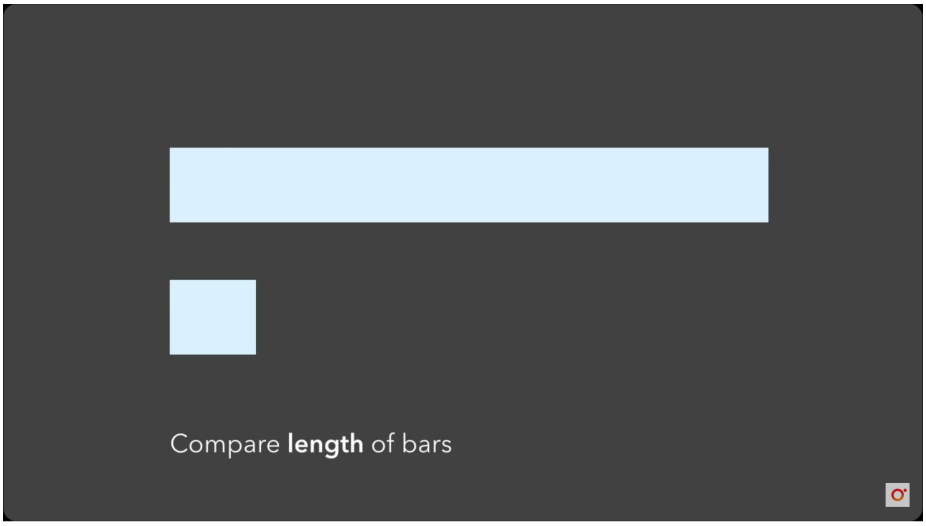
[5]For a definition of the regions see https://www.who.int

How many times larger is the second circle?




.29

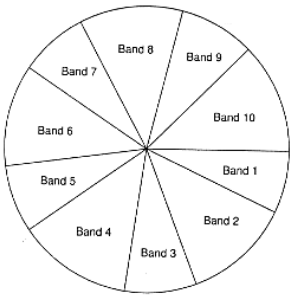
How many times larger is the first bar?



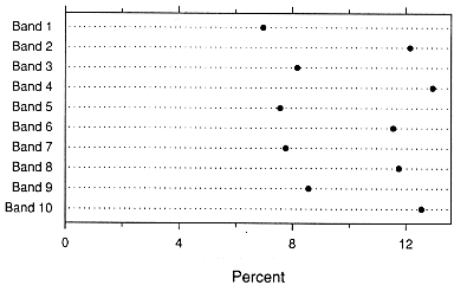

.30

For the same reason, a pie chart is almost never the best choice. The next graph shows that patterns are very difficult to observe in pie charts because we need to rely on angles or areas to make the comparison.

## Angle: pie chart versus dot plot



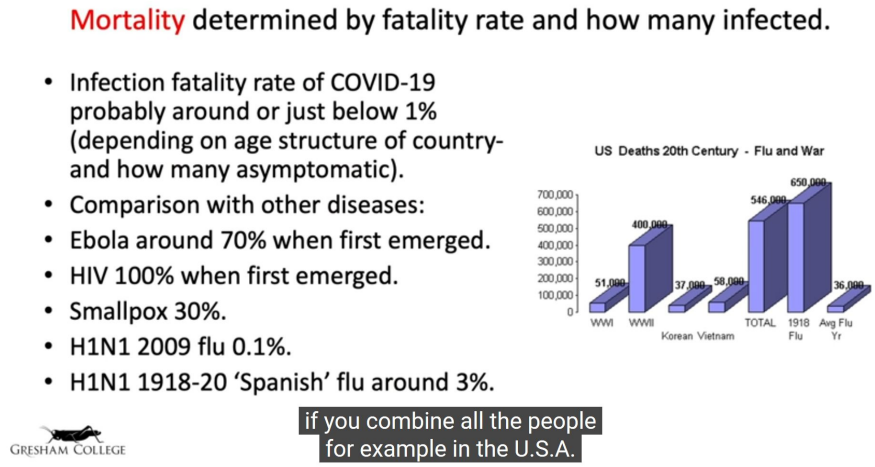

4.19 PIE CHART. The pie chart falls in the category of a pop chart — a graphical method used frequently in the mass media and certain business presentations but far less in science and technology. Both table look-up and pattern perception are less efficient for pie charts than for dot plots.




4.20 DOT PLOT. The data from Figure 4.19 are graphed by a dot plot. Patterns emerge that cannot be decoded from Figure 4.19.

.31

**6. Volume** Probably the most frequently used graphs using volume are the three dimensional barchart and three dimensional pie chart. In the morning we saw an example of a three dimensional bar chart. We show another one in slide 32 (from Chris Whitty's presentation). Not only does the third dimension add nothing, it can even distort the truthful representation of data because the third dimension makes lower bars use relatively more "ink". We have already argued why a pie chart should not be used. The only graph that is worse than a pie chart is a three dimensional pie chart.



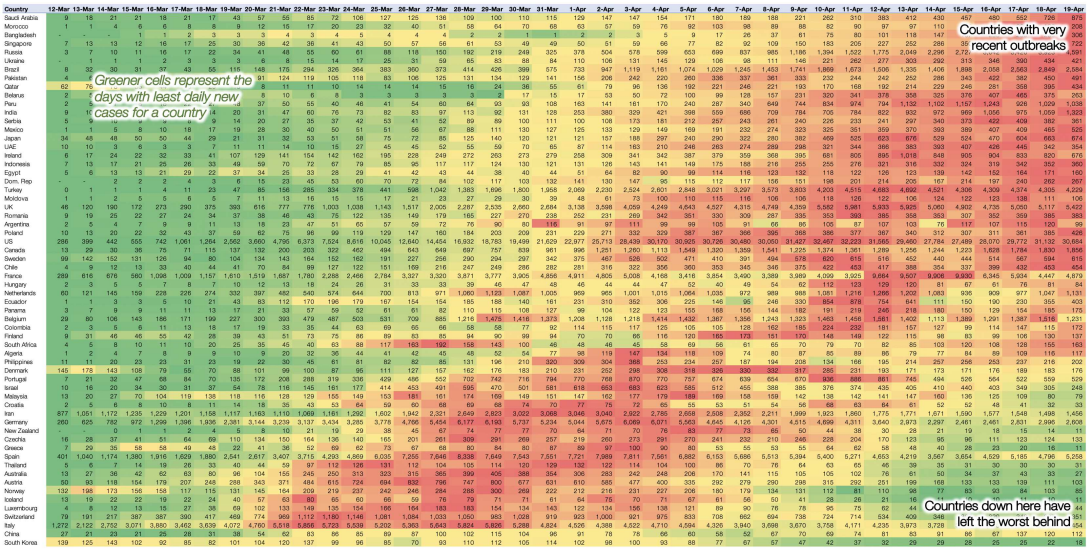

.32

**7. Colour** Colour is last in the hierarchy. The reason is that it cannot be directly related to size. But it may become easier to link colours to numbers by using a legend as a guide. If the actual values are not of primary importance and it's more the general trend that matters, colour can be a powerful concise way to visualize variation in values. This especially holds if the colour has a general interpretation (like green for "good" and red for "bad", or blue for "cold" and

red for “hot” in many cultures). The heatmap below concisely shows the trends in number of new SARS-CoV-2 infections during the start of the pandemic.

## Alternative as heatmap

Chart 3: Daily **New** Coronavirus Cases Worldwide*

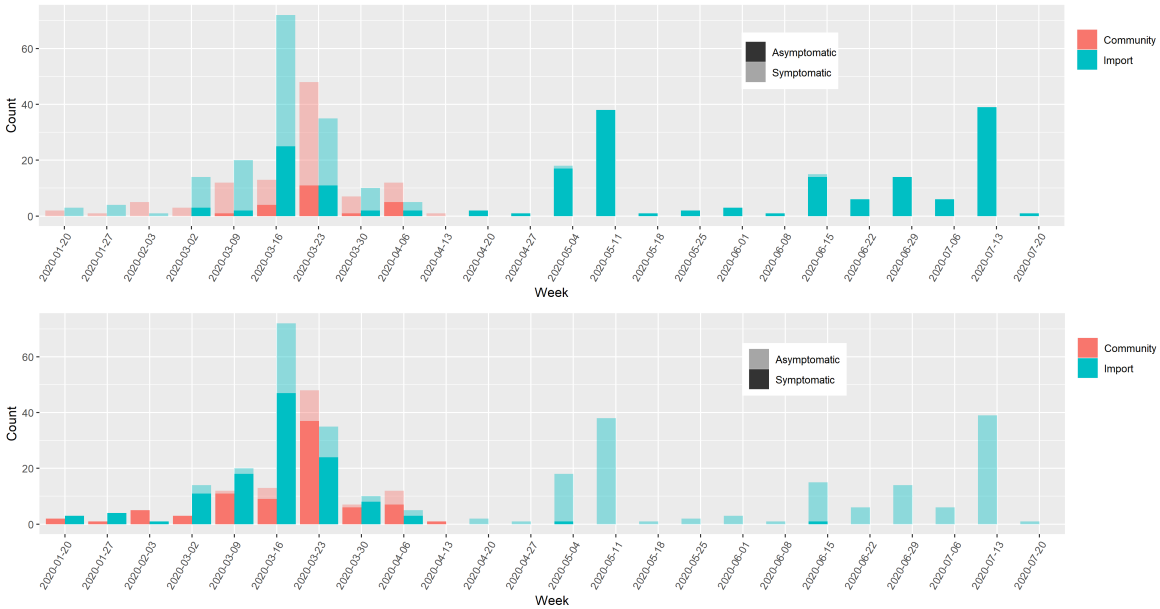



*As of April 19th 2020. Top ~60 countries in total cases, with added countries with special interest. Daily new cases calculated as 4-day averages to smooth outliers and weekend effects. Order based on average of last 5 days over average of last 25 days, adjusted manually to highlight the trend
Source: Tomas Pueyo Analysis, Johns Hopkins data via Github: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv

.33

An example in which for many of us our fast brain interprets a component of colour is shown in slide 34. In general, a deeper colour (more “chroma”) is seen as more serious, certainly when it is red. Therefore, most people will find it easier to link chroma to symptomatic status in the second figure, even though the factual information is the same in both figures.

## Number of SARS-CoV-2 infections in Vietnam




.34

### 3.1.2 Group (categorical)

Colours are very suitable to distinguish groups of a categorical variable. Other options are the use of character type for locations ("+" and "o" are the preferred symbols to distinguish two groups in a scatterplot) or line types for lines (dashed, dotted, solid).

# 4 Colour

Several models have been developed to represent colour[6].

## Basics

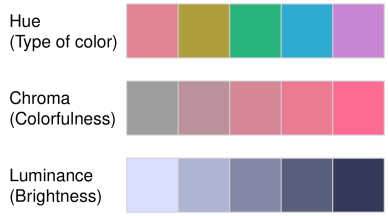
- Three colour models
  - RGB (Red-Green-Blue): how computers represent colour (≈ the 3 eye cones in humans)
  - HSV (Hue-Saturation-Value): ≈ wavelength-colour purity-brightness
  - HCL (Hue-Chroma-Luminance): ≈ wavelength-colourfullness-brightness closest to perceptual properties
- Take colour-blindness into account:
  deuteranopia: green deficient; protanopia: red deficient; tritanopia: blue deficient
- In R colours are internally coded in hexadecimal format

```
     black      orange     skyblue bluishgreen
 "#000000"   "#E69F00"   "#56B4E9"   "#009E73"
```

R has 657 built in color names (see e.g. CheatSheet)

.35

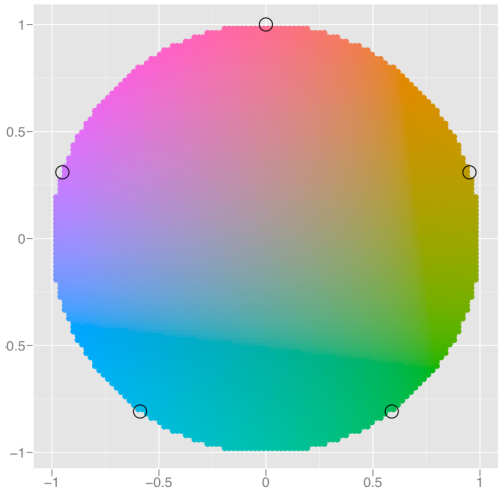
## HCL explained




.36

## HCL colour wheel (fixed luminance)

[6] A fairly techinical introduction is https://www.r-project.org/conferences/DSC-2003/Proceedings/Ihaka.pdf. A definition of the terms can be found on the Wikipedia page https://en.wikipedia.org/wiki/HSL_and_HSV#Formal_derivation




.37

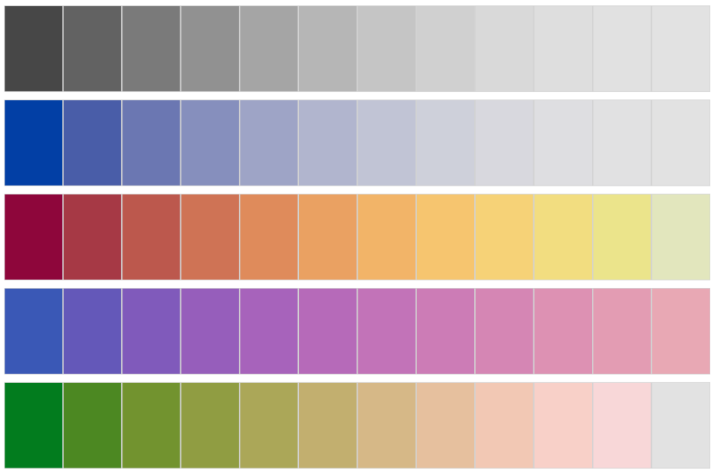
If we want to distinguish groups or we want to represent numeric values by colour, we use a range of colours, called a *palette*. There are three main types of palettes.

## Represent value by colour: palettes

- Qualitative: categorical information
  - same perceptual weight to each category: vary hue, keep chroma and luminance fixed
  - may choose hues based on convention (red and blue for political parties in United States, cold/hot, risky/safe)
- Sequential: for ordered/numeric information
  - Single or two hues; change luminance linearly along with value, may change chroma
- Diverging: for ordered/numeric information in two directions around neutral value
  Combines two sequential palettes with different hues

.38

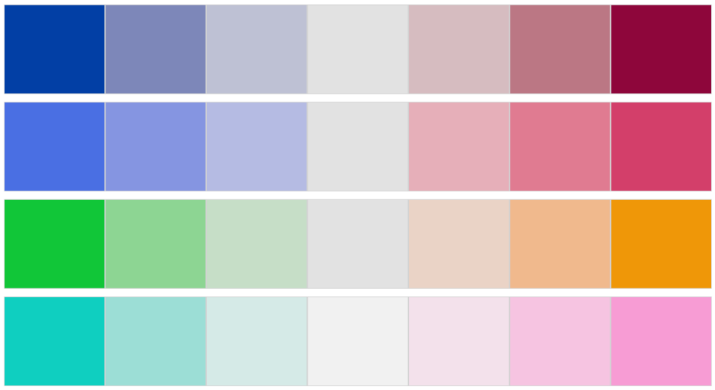
## Sequential palettes



1st: vary luminance only; 2: vary chroma and luminance;
3-5: vary hue, chroma and luminance

.39

## Diverging palettes



.40

## `hcl_palettes()` in `colorspace` package

```
R> hcl_palettes(plot = TRUE)
```


9/27

.41

## Some advice on colour palettes

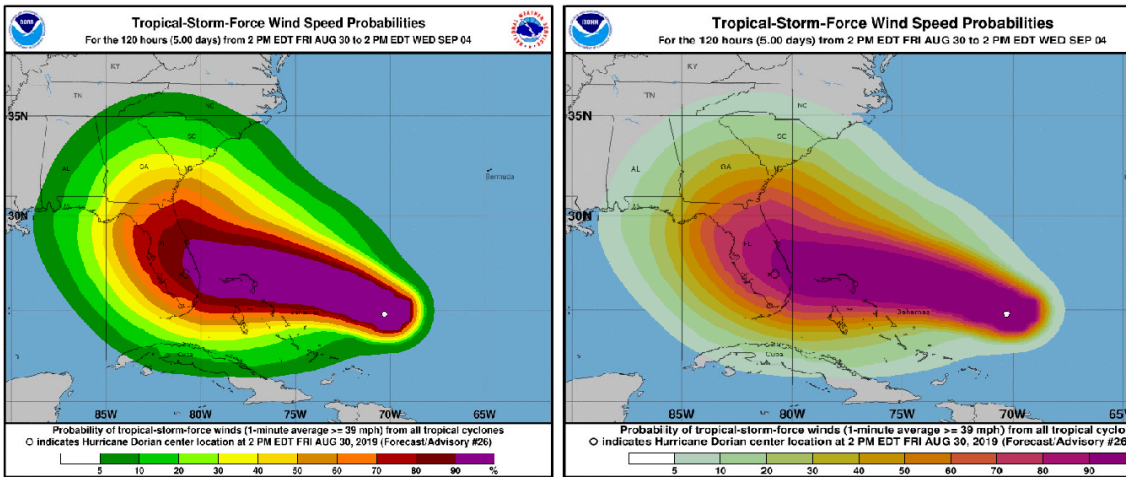
- Avoid large areas of flashy, highly saturated colors; use light colors (higher luminance, lower chroma)
- For points and lines: lower luminance, higher chroma
- If size of coloured areas has a meaning, then use colours of similar luminance. Lighter colours tend to make areas look larger
- `http://hclwizard.org/`: palette creator and colour deficiency emulator
- Top R colour palettes
- All R colour palettes (including interactive colour picker)
- PrettyCols

.42

An interesting example in which use of high-chroma colours may have led to confusion is the example in the next slide[7].

[7] See `https://www.zeileis.org/news/dorian_rainbow/`

### Highly saturated colours may have confused Donald Trump

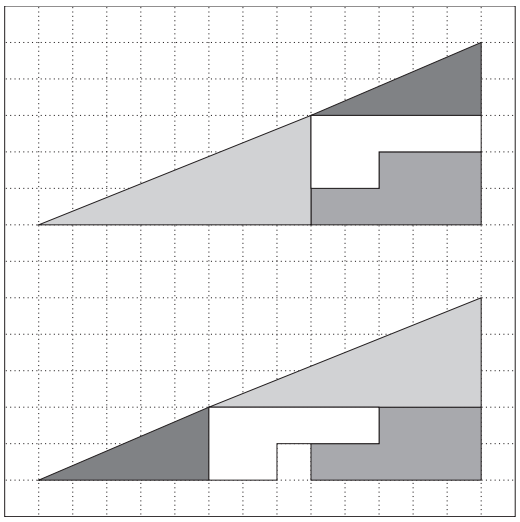



.43

The use of the rainbow palette with very saturated colours may have created the impression to Donald Trump (and others) that the probability of the hurricane hitting Alabama (the green area) was quite high, while it was only 5 to 10%.
For further advice on using colours can be found in the works by Maureen Stone[8] and Stephen Few[9]

# 5 Exercises part 2

1. Look at the following figure. The bottom figure contains the same pieces as the top one, but they have been rearranged.



   Explain where the extra square comes from in the bottom figure.
2. Maybe the Nightingale Rose inspired Will Burtin, a graphics designer from Germany, when in 1951 he visualized resistance to three types of antibiotics for 16 different bacteria. Figure 1 shows the minimum Inhibitory

[8] https://www.perceptualedge.com/articles/b-eye/choosing_colors.pdf
[9] https://nbisweden.github.io/Rcourse/files/rules_for_using_color.pdf

Concentration (MIC, in $\mu$g/ml) for each antibiotic and bacteria combination. Lower is better, indicating less antibiotic is needed to treat the bacteria. A distinction is made between Gram positive and Gram negative bacteria[10].

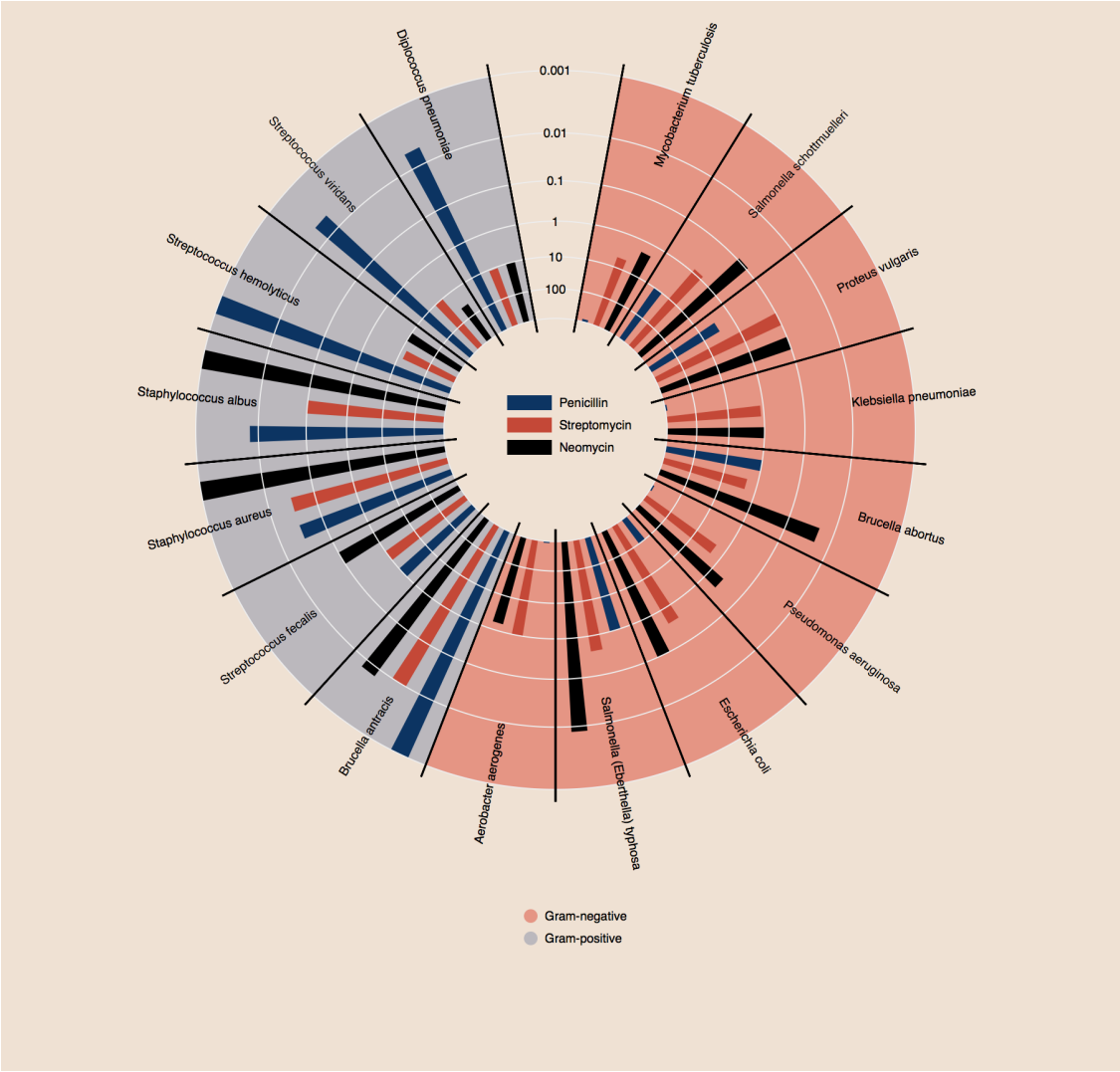



Figure 1: MIC for three anitbiotics by type of bacteria

In 2008, the journal Chance asked its readers to come up with a new visualisation. An adaptation of the graph of one of the winners is shown in Figure 2[11]. The main purpose of the original study was to investigate which bacteria were resistant to what type of antibiotics. Therefore, the bacteria are split along the y-axis according to the number of antibiotics to which the bacteria are resistant. An MIC value above 0.1 was defined as being resistant.

(a) For both figures, what are the geometric objects, the aesthetic attributes and the coordinate system. Define the scales; are there any labels and a legend? Has a statistical transformations been applied? Are there any position adjustments? Are there subplots (faceting)?

[10]The original data is available as `antibiotic` in the `lucid` R package

[11]See Howard Wainer and Shaun Lysen, "That's Funny...", American Scientist, 97: 272-275, 2008

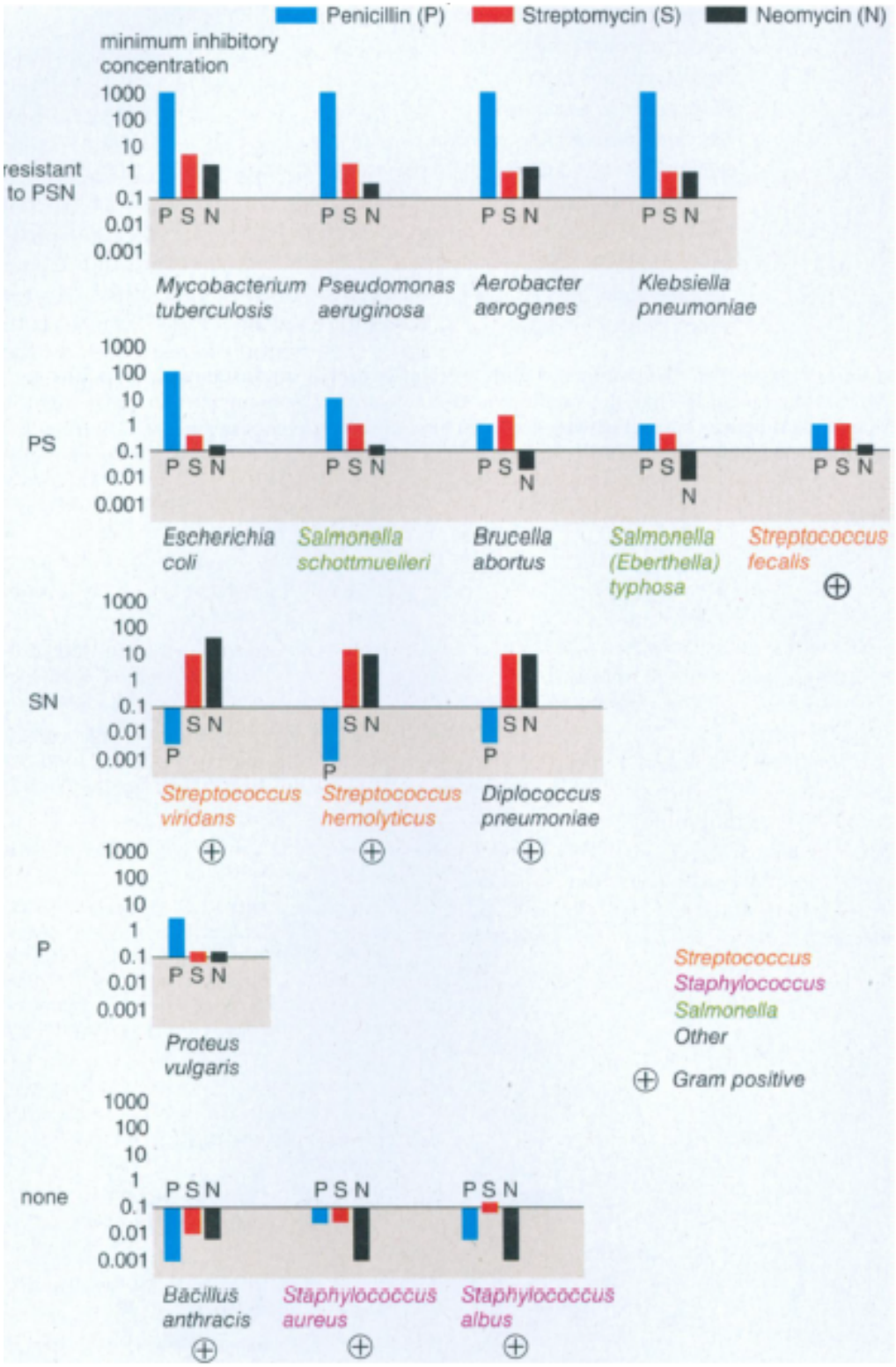



Figure 2: Redesign of the original figure by Will Burtin

(b) When carefully looking at the second figure, some peculiarities can be observed. Do you recognize them? (Look at the genus in combination with resistance pattern.)

3. One of the first persons that realized the power of data visualisation was Florence Nightingale. Florence Nightingale was not only a nurse, but also a statistician and a social reformer. During the Crimean War she collected information about the soldier's death causes in the military hospital over a period of two years; the second year after sanitary interventions had been made. She discovered some revealing facts: the majority of them were dying not because of wounds inflicted during the battle, but due to infections like typhus, typhoid and cholera that they had caught inside the hospital. She discovered that hygienic conditions in the hospital were severely neglected. She implemented changes in the military hospitals and was convinced that such changes would have similar effects in the civilian hospitals in London. After a long struggle, she finally managed to receive an audition with the persons that were responsible for the sanity in the London hospitals, but the meeting had to be of very short duration. In this context, she knew her presentation had to be short and concise, but very impactful at the same time. She came up with the Nightingale Diagram or Nightingale Rose[12].

   We have another look at the Nightingale Rose that we showed in the morning. You can find an online version of the figure at `https://www.historyofinformation.com/image.php?id=851`. You also received the data as a csv file `NightingaleRose.csv`[13].

   (a) What are the geometric objects, the aesthetic attributes, scales and the coordinate system. Define the scales; are there any labels and a legend? Has a statistical transformation been applied? Are there any position adjustments? Are there panels (faceting)?

   (b) One of the critiques on the plot is as follows

   > The major flaw with Nightingale Rose Charts is that the outer segments are given more emphasis because of their larger area size. This disproportionately represents increases in value.[14]

   What do you think of this argument? Can you think of other elements that can be improved.

   (c) Try to improve the figure using the data set that was sent to you. You could try a stacked bar chart, a dodged bar chart or a line chart.

[12] An account of her life and her relation with data: `https://theconversation.com/the-healing-power-of-data-florence-nightingales-true-legacy-134649`

[13] It is based on the dataset in the HistData package. It can also be downloaded from `https://vincentarelbundock.github.io/Rdatasets/csv/HistData/Nightingale.csv`. The main change is that the values in *Disease*, *Wounds* and *Other* have been stacked into one single column *perc*, which gives the percentage that died from each of the three types of diseases per month. A column *cause* has been added that provides the cause of death. This long format is almost always needed in `ggplot2` if we want to plot by subgroup

[14] `https://datavizcatalogue.com/methods/nightingale_rose_chart.html`

4. Figure 3 was shown during one of the OUCRU academic meetings. We can see each of the pies as a standalone figure, but it is more interesting trying to discover some patterns in the data. The purpose of this exercise is to find out whether we can create a more informative graph. You received the data as a file `ICUcost.csv`.

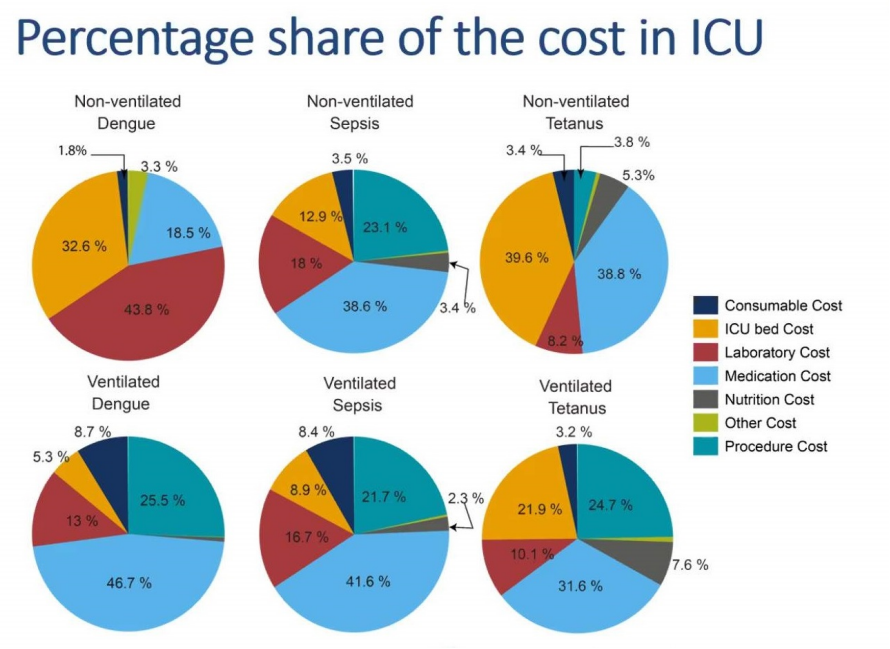



Figure 3: ICU cost

(a) Which coordinate system is used. Mention the geometric object(s) and the aesthetic attributes. Which scale is used; are there any labels and a legend? How many layers does the figure have. Is there a position adjustment? Are there subplots (facets) used?

(b) Change the figure into a stacked barchart. Map `Severity` to the x-axis and `Perc` to the y-axis. Colour the bars by type. In order to make the text more readable, use `coord_flip()`.

(c) Change the figure into a barchart with subplots/facets by ventilation status and disease. Map `Type` to the x-axis. Does the use of colour have added value? Use `facet_grid`, with `Ventilated` along the rows and `Disease` along the columns.

Do you see some patterns in ICU costs? Remove the mapping of `Type` to fill colour. Does this improve the information in the figure?

(d) We think out of the box and use lines instead of bars as geometric objects. Map `Type` to the x-axis, and `Perc` to the y-axis. Map `Disease` to colour. Make separate lines by disease. Make separate panels by ventilation status. You may select `coord_flip()` to increase readability of the labels. Change the size of the lines to 1 or 2.

# 6 Aesthetics: Tufte

Edward Tufte is a statistician and artist who has written influential works on the visualisation of information. He combines the representation of information with the aesthetical component[15].

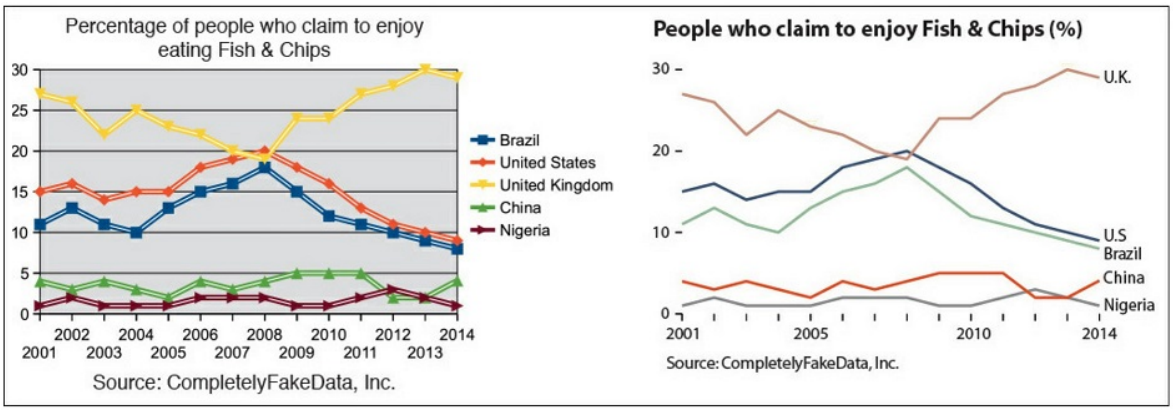
## Edward Tufte: Graphical Excellence

- well-designed presentation of interesting data is a matter of *substance*, of *statistics* and of *design*.
- complex ideas communicated with clarity, precision, and efficiency
- is nearly always multivariate (table is better with $< 30$ data points)
- $$\text{data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graphic}}$$
  - dotplots preferred over barcharts
  - no unnecessary dimensions
- opposite of chartjunk

.44

We give two examples. In slide 45 Fish & Chips preferences are plotted over time for five countries (fake data).

## Which figure do you prefer?




.45

The other example is from the book *The Visual Display of Quantitative Information* by Edward Tufte. In Figure 4, we show geometric mean magnitude estimates ($\pm$ 1 s.e., 10 observations) of citric acid taste intensity (C). Individuals were pretreated with sucrose (S) or water (H). Values were measured before and after Gymnema sylvestre (GS) or tea was given. The left graph shows the original graph that was used in the paper, whereas the right graph shows how much ink can be removed without losing any information. Since all effects were significant, the stars can be left out if significance is mentioned in the caption of the figure[16].

[15] Edward Tufte has a web site with information on his books and some interesting discussions: http://www.edwardtufte.com/tufte/

[16] Original paper: James T. Kuznicki and N. Bruce McCutcheon. *Cross-Enhancement of the Sour Taste on Single Human Taste Papillae* Journal of Experimental Psychology: General, **108(1)**: 68–89, 1979

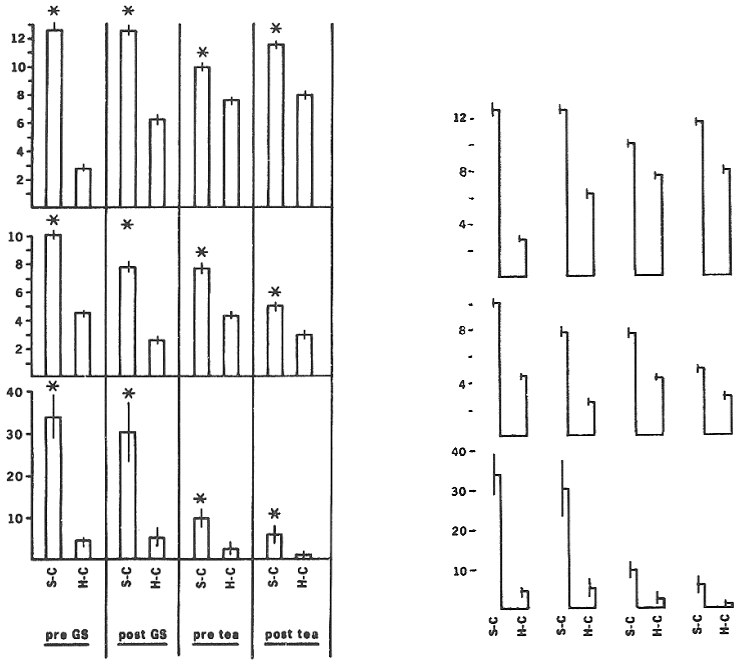
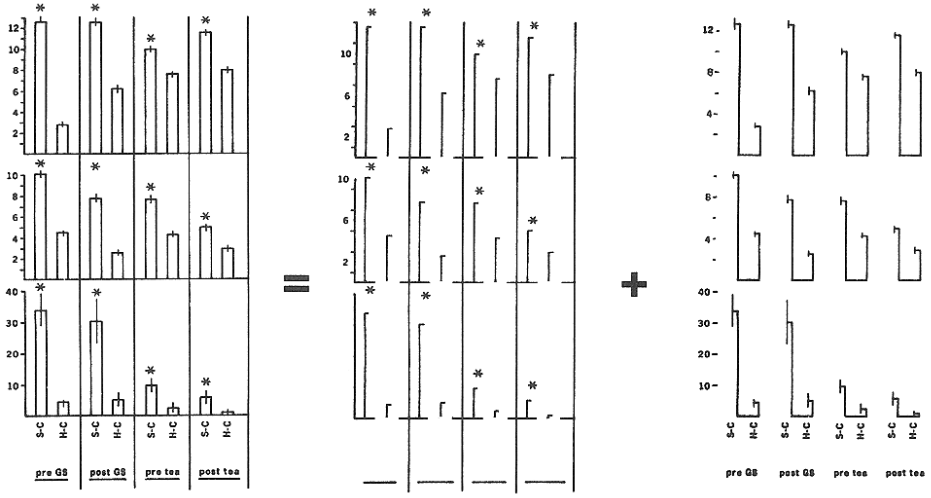



Figure 4: Decreasing data-ink ratio

## Hence




.46

I give one more example of graphical excellence, in which a well designed graph can give important new insights.

**Trends in cancer survival.** In a study that appeared in the Lancet[17], the authors estimated trends in long-term survival rates for different types of cancer. The table in slide 47 shows the estimates of long-term relative survival, i.e. the

[17] Hermann Brenner, "Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis," The Lancet, 360 (October 12, 2002), 1131-1135

survival in the hypothetical situation in which cancer is the only cause of death. Can you observe any difference by type of cancer?

| | Relative survival, % (SE) | | | |
|---|---|---|---|---|
| | 5 years | 10 years | 15 years | 20 years |
| **Cancer site** | | | | |
| Oral cavity and pharynx | 56.7 (1.3) | 44.2 (1.4) | 37.5 (1.6) | 33.0 (1.8) |
| Oesophagus | 14.2 (1.4) | 7.9 (1.3) | 7.7 (1.6) | 5.4 (2.0) |
| Stomach | 23.8 (1.3) | 19.4 (1.4) | 19.0 (1.7) | 14.9 (1.9) |
| Colon | 61.7 (0.8) | 55.4 (1.0) | 53.9 (1.2) | 52.3 (1.6) |
| Rectum | 62.6 (1.2) | 55.2 (1.4) | 51.8 (1.8) | 49.2 (2.3) |
| Liver and intrahepatic bile duct | 7.5 (1.1) | 5.8 (1.2) | 6.3 (1.5) | 7.6 (2.0) |
| Pancreas | 4.0 (0.5) | 3.0 (0.5) | 2.7 (0.6) | 2.7 (0.8) |
| Larynx | 68.8 (2.1) | 56.7 (2.5) | 45.8 (2.8) | 37.8 (3.1) |
| Lung and bronchus | 15.0 (0.4) | 10.6 (0.4) | 8.1 (0.4) | 6.5 (0.4) |
| Melanomas | 89.0 (0.8) | 86.7 (1.1) | 83.5 (1.5) | 82.8 (1.9) |
| Breast | 86.4 (0.4) | 78.3 (0.6) | 71.3 (0.7) | 65.0 (1.0) |
| Cervix uteri | 70.5 (1.6) | 64.1 (1.8) | 62.8 (2.1) | 60.0 (2.4) |
| Corpus uteri and uterus, NOS | 84.3 (1.0) | 83.2 (1.3) | 80.8 (1.7) | 79.2 (2.0) |
| Ovary | 55.0 (1.3) | 49.3 (1.6) | 49.9 (1.9) | 49.6 (2.4) |
| Prostate | 98.8 (0.4) | 95.2 (0.9) | 87.1 (1.7) | 81.1 (3.0) |
| Testis | 94.7 (1.1) | 94.0 (1.3) | 91.1 (1.8) | 88.2 (2.3) |
| Urinary bladder | 82.1 (1.0) | 76.2 (1.4) | 70.3 (1.9) | 67.9 (2.4) |
| Kidney and renal pelvis | 61.8 (1.3) | 54.4 (1.6) | 49.8 (2.0) | 47.3 (2.6) |
| Brain and other nervous system | 32.0 (1.4) | 29.2 (1.5) | 27.6 (1.6) | 26.1 (1.9) |
| Thyroid | 96.0 (0.8) | 95.8 (1.2) | 94.0 (1.6) | 95.4 (2.1) |
| Hodgkin's disease | 85.1 (1.7) | 79.8 (2.0) | 73.8 (2.4) | 67.1 (2.8) |
| Non-Hodgkin lymphomas | 57.8 (1.0) | 46.3 (1.2) | 38.3 (1.4) | 34.3 (1.7) |
| Multiple myeloma | 29.5 (1.6) | 12.7 (1.5) | 7.0 (1.3) | 4.8 (1.5) |
| Leukaemias | 42.5 (1.2) | 32.4 (1.3) | 29.7 (1.5) | 26.2 (1.7) |

The Lancet, 360 (October 12, 2002), 1131-1135

.47

If you look at the table for some time, for sure you will observe differences. You will notice that survival is highest for prostate cancer, and lowest for cancer of the pancreas. What's more difficult to observe are the trends. Are there types of cancers for which survival beyond five years still goes down? For this, a visualisation of the numbers is much more insightful, as in Figure 5.
We can even make a finer distinction as in the figure in slide 48. Have a look at the choices that were made in the creation of this figure. Do you think they were good choices? Do you have suggestions for improvement?

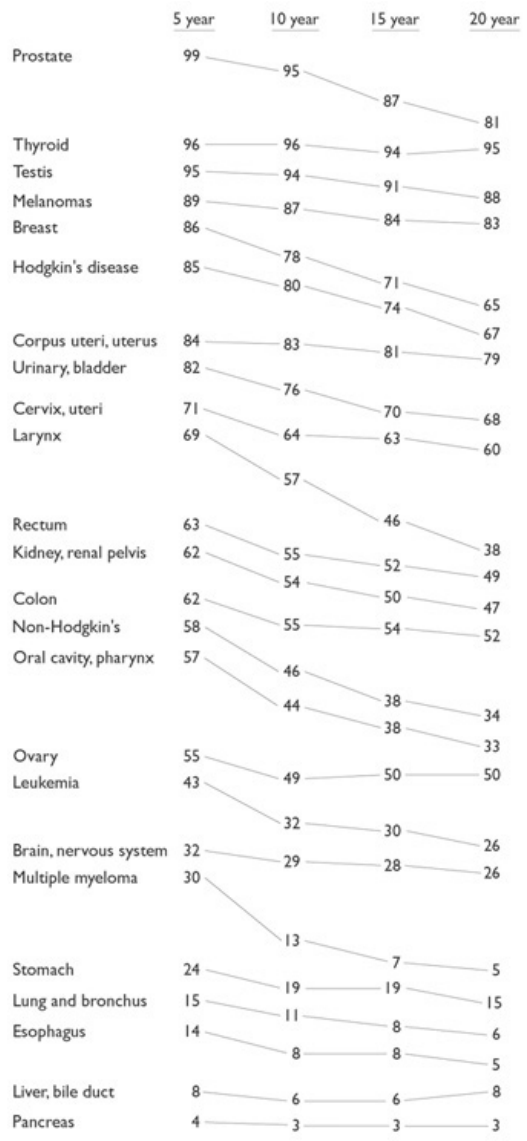
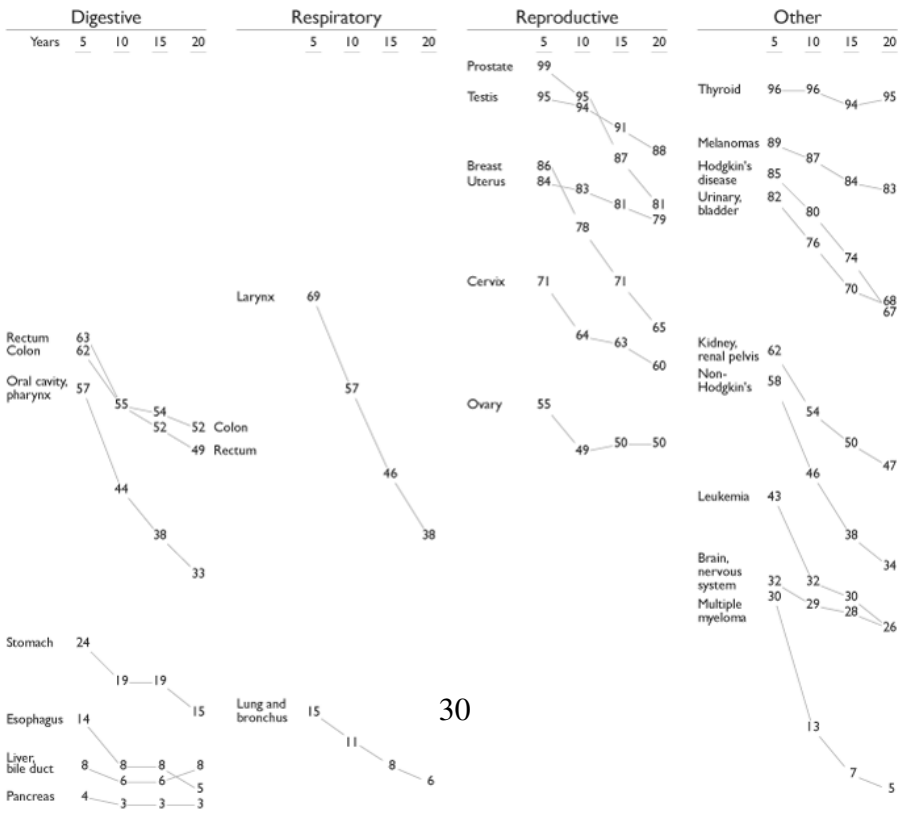



Figure 5: Relative survival by type of cancer



# 7 Final words

## Main Message

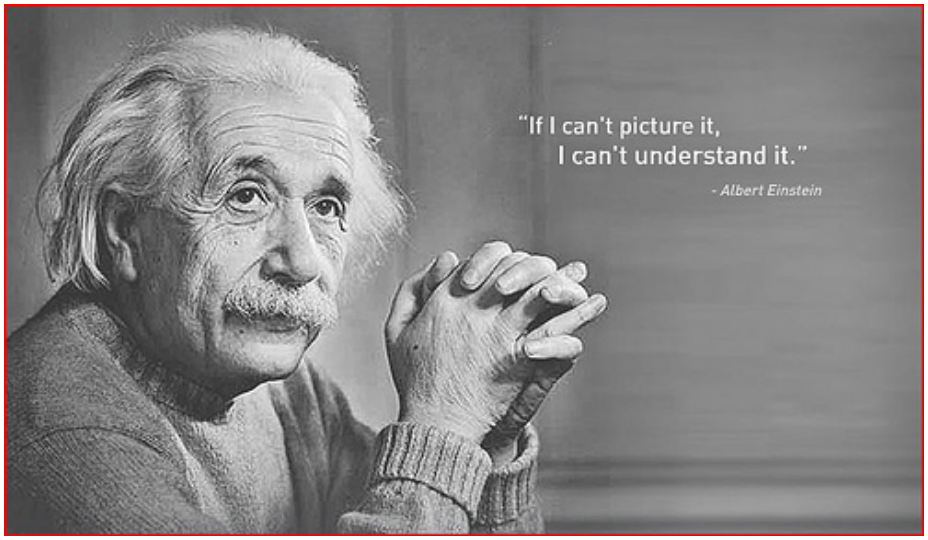
- Last two decades: new ideas on what is the most informative way to display quantitative information
  - Based on scientific knowledge on perception
  - Guidelines that help improving graph quality. "Mistakes" due to ignorance
  - No "right or wrong", only "better or worse"
- Graph creation iterative and laborious process, with trial and error. Creating a half-page figure takes same time as writing half-page text
- Several high quality newspapers and magazines (New York Times, Financial Times, Washington Post, The Economist) are ahead of many scientists and institutions (including WHO)

.49

## Graphics principles, the STRATOS initiative


From
https://graphicsprinciples.github.io/threelaws.html
STRATOS (STRengthening Analytical Thinking for Observational Studies) visualisation panel

.50




.51

# References


[1] Alberto Cairo. *The Truthful Art*. Pearson Education, 2016. http://www.thefunctionalart.com/

[2] William S. Cleveland. *The Elements of Graphing Data*. Hobart Press, Summit, New Jersey, 1994.

[3] Kieran Healy. *Data Visualization*. Princeton University Press, New Jersey, 2019. https://socviz.co/

[4] Kate Strachnyi. *ColorWise*. O'Reilly, 2023.

[5] Edward D. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, 2001.

[6] Klaus Wilke. *Fundamentals of Data Visualisation*. O'Reilly, 2019 https://serialmentor.com/dataviz/.

[7] Leland Wilkinson. *The Grammar of Graphics*. Springer, New York, 2005.