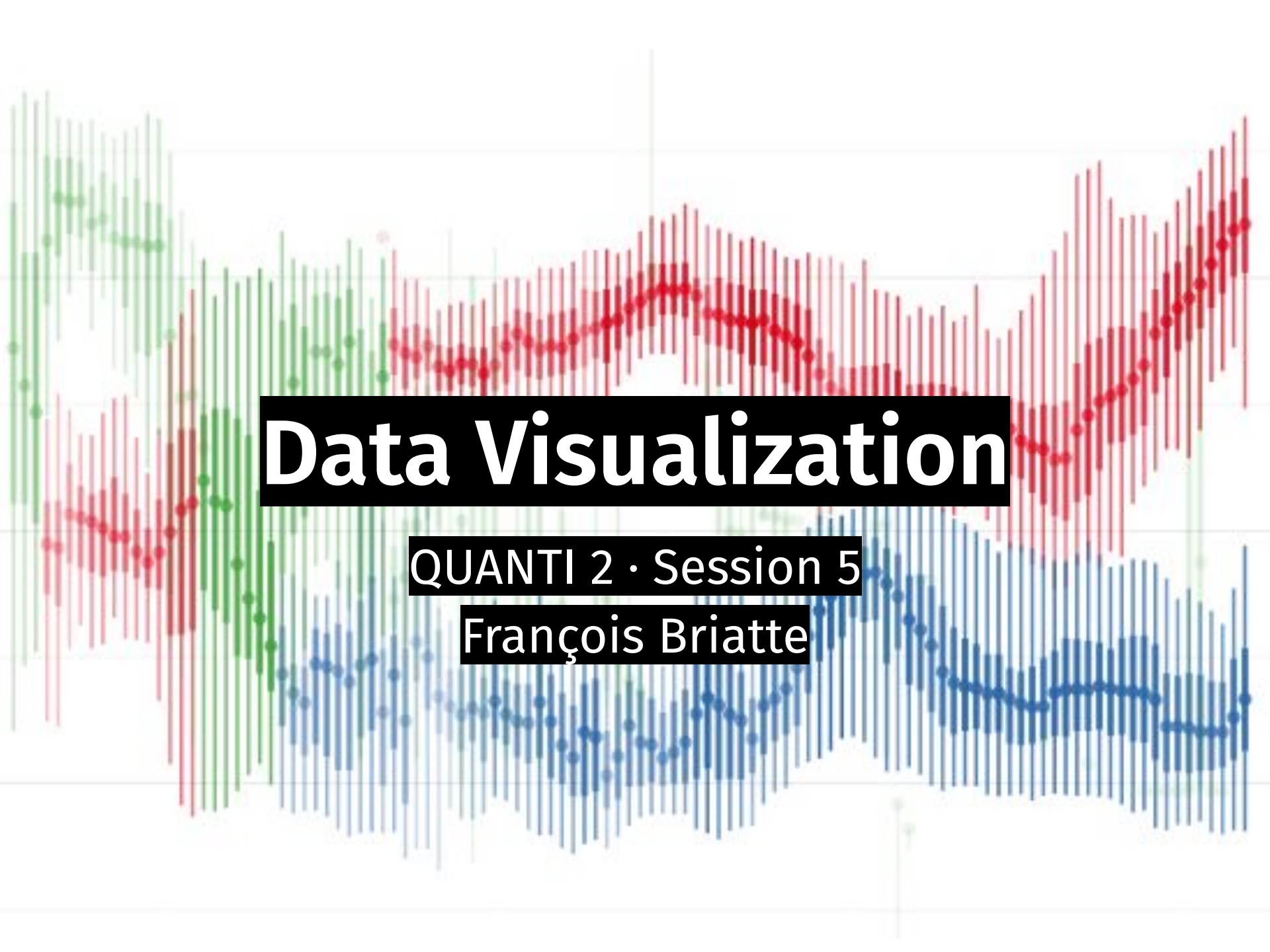


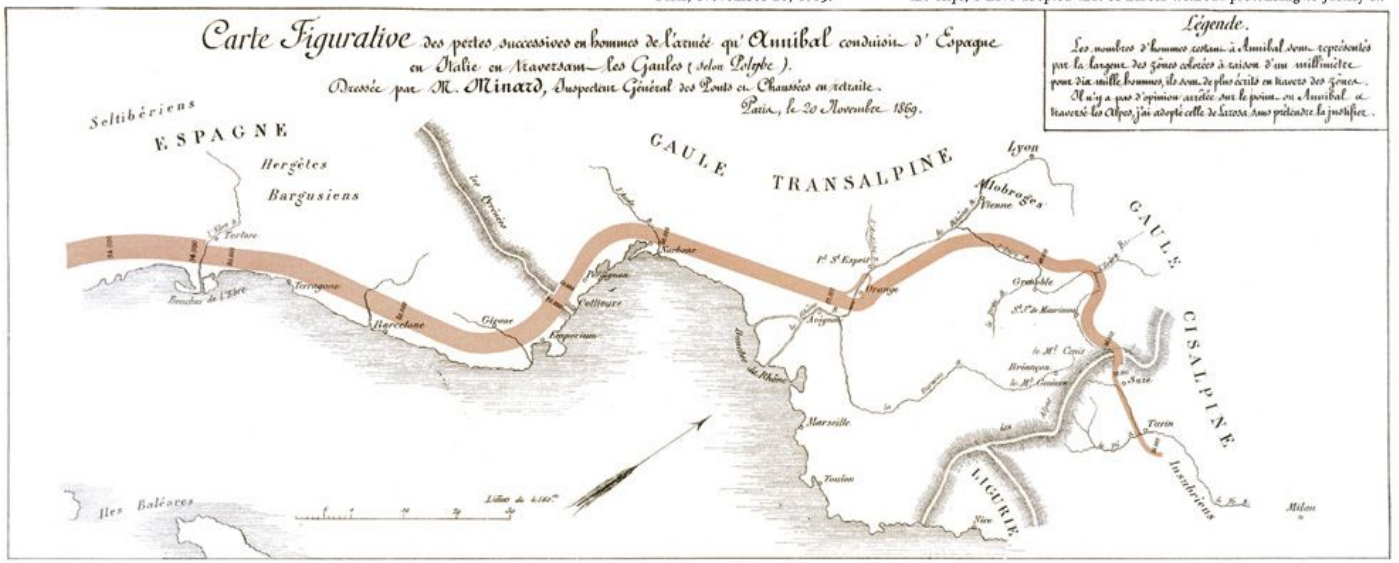
Data Visualization

QUANTI 2 · Session 5

François Briatte

Figurative Map of the successive losses in men of the army that Hannibal led from Spain to Italy while crossing the Gauls (according to Polybius).
Drawn up by M. Minard, Inspector General of Bridges and Roads in retirement.
Paris, November 20, 1869.

Legend.
The numbers of men remaining with Hannibal are represented by the width of the colored zones at a rate of one millimeter for ten thousand men; they are further written across the zones.
There is no final opinion on the point where Hannibal crossed the Alps; I have adopted that of Larosa without pretending to justify it.



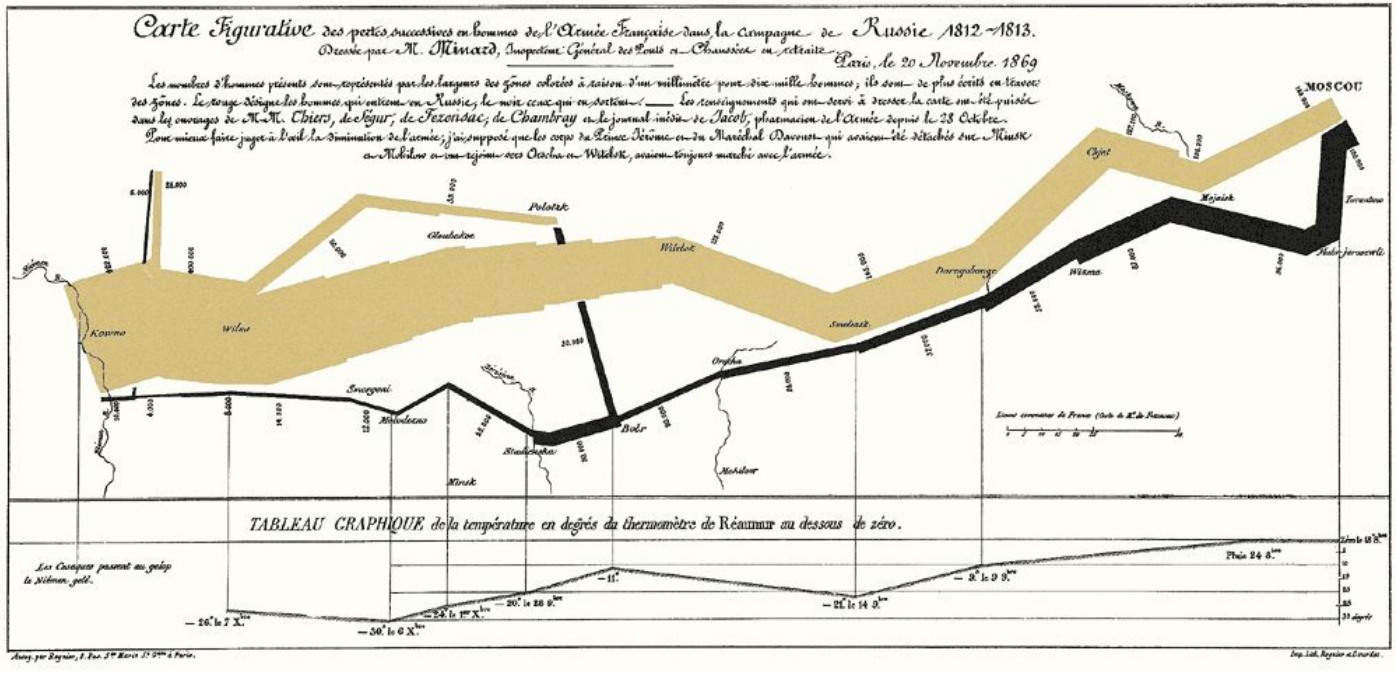
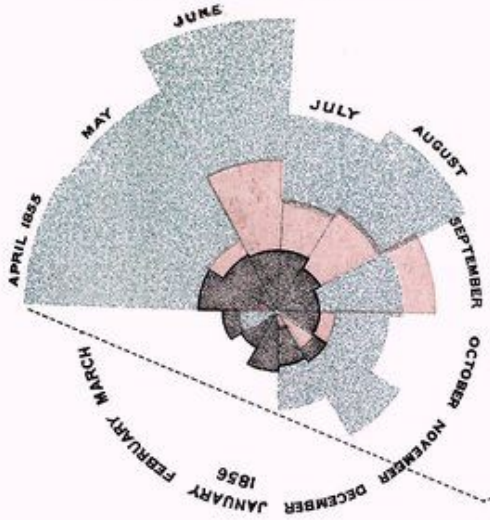
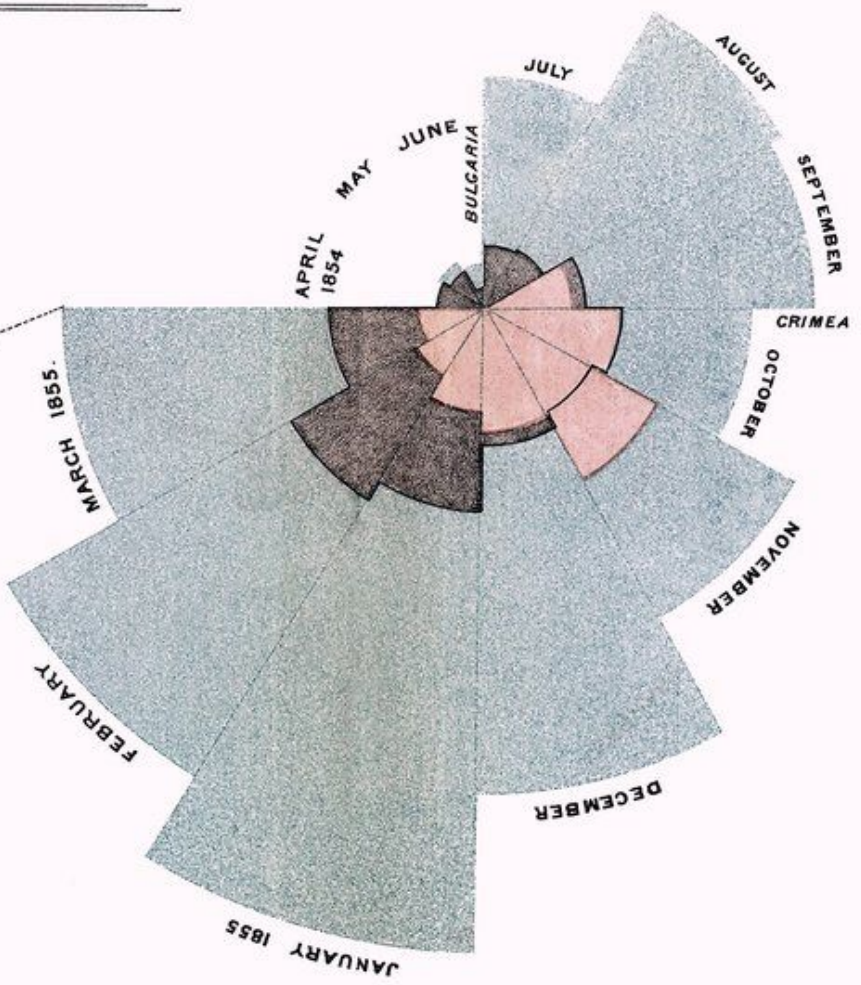


DIAGRAM of the CAUSES of MORTALITY IN THE ARMY IN THE EAST.

2.
APRIL 1855 to MARCH 1856.



1.
APRIL 1854 to MARCH 1855.



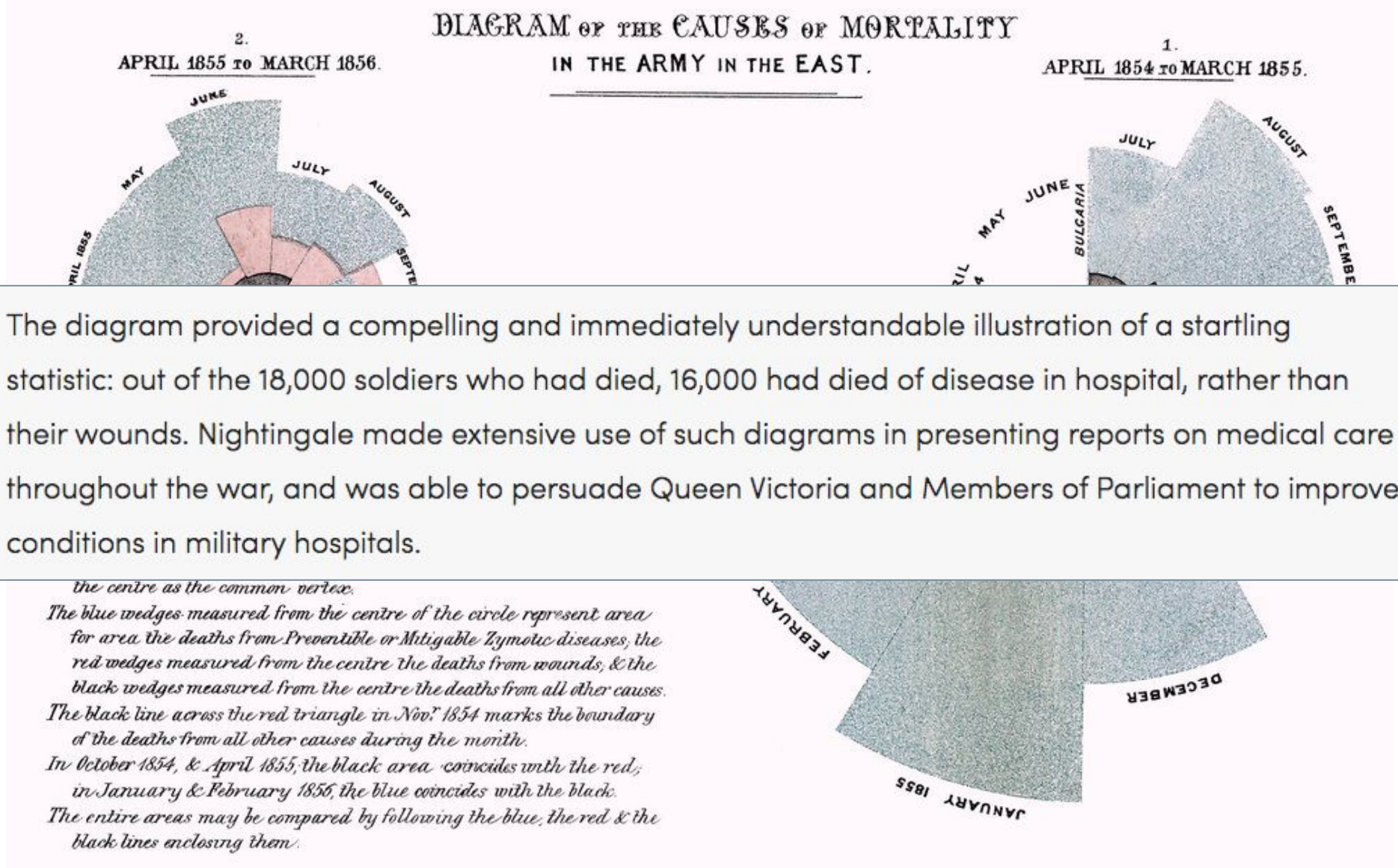
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.

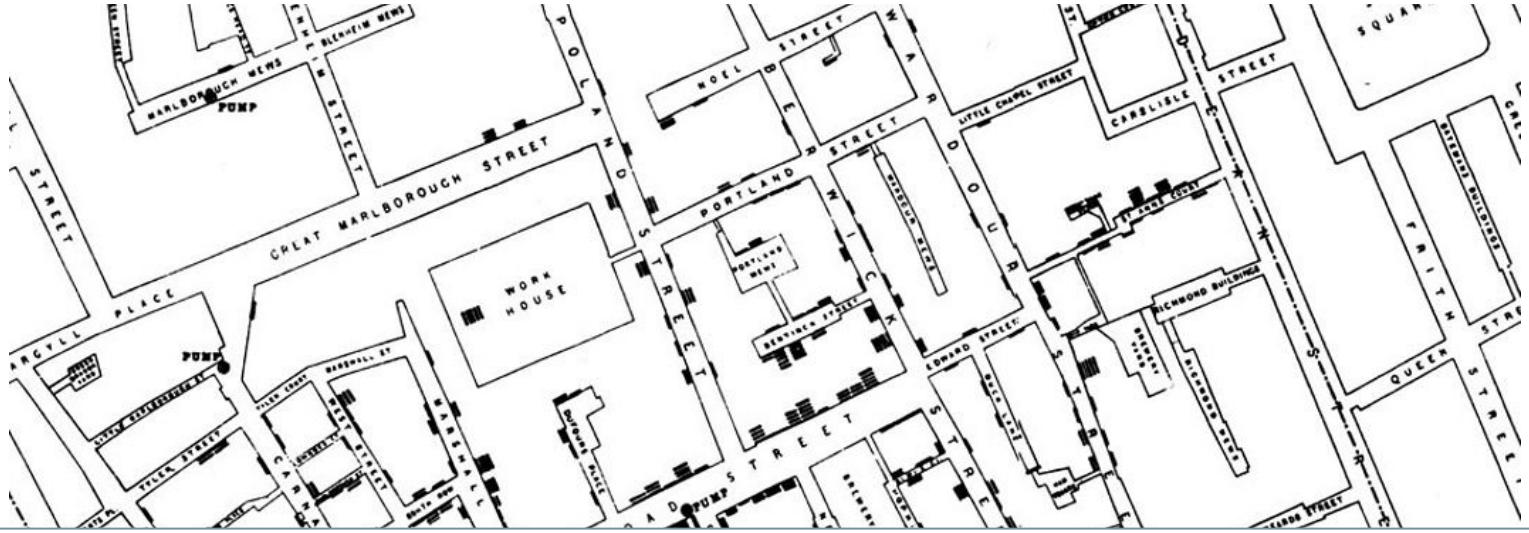
In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

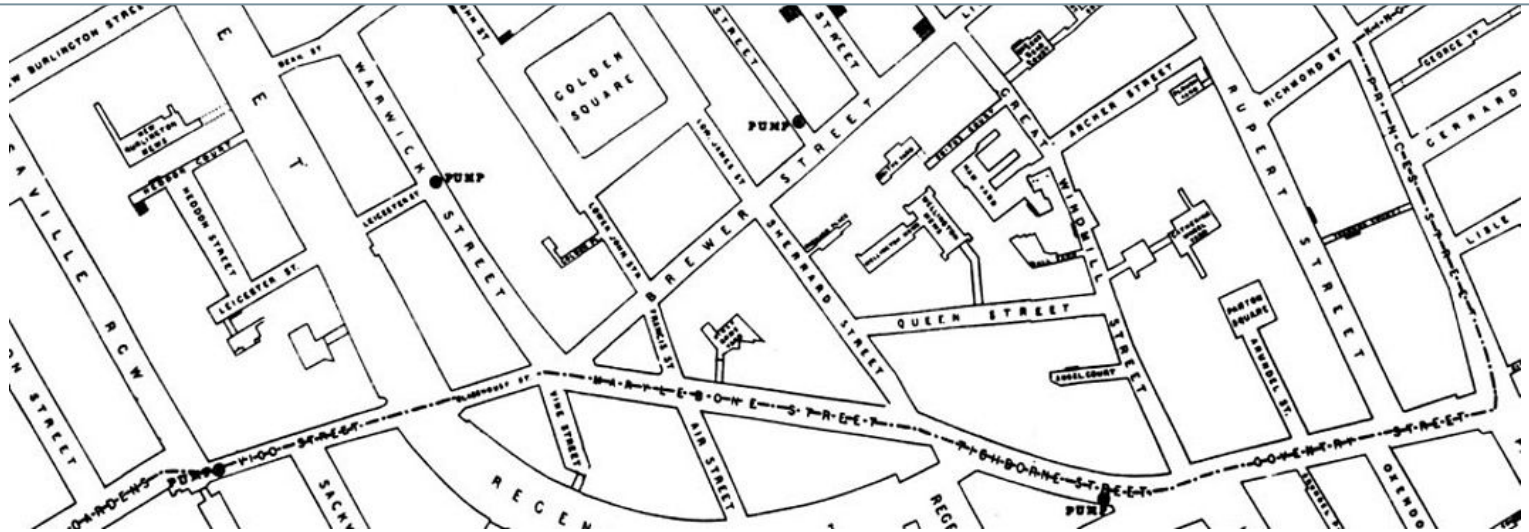


The diagram provided a compelling and immediately understandable illustration of a startling statistic: out of the 18,000 soldiers who had died, 16,000 had died of disease in hospital, rather than their wounds. Nightingale made extensive use of such diagrams in presenting reports on medical care throughout the war, and was able to persuade Queen Victoria and Members of Parliament to improve conditions in military hospitals.


GREAT MARLBOROUGH STREET
MARLBOROUGH MEWS
PUMP
ARGYLL PLACE
WORK HOUSE
POLAND STREET
PORTLAND STREET
NOEL STREET
WARDOUR STREET
LITTLE CHAPEL STREET
CARLISLE STREET
BROAD STREET
EDWARD STREET
RICHMOND BUILDINGS
QUEEN STREET
KING STREET
MARSHALL STREET
GOLDEN SQUARE
GREAT PULTENEY STREET
LITTLE PULTENEY STREET
PETER STREET
OLD COMPTON
REGENT STREET
BURLINGTON STREET
WARWICK STREET
BRIDLE STREET
BREWER STREET
SHERRARD STREET
ARCHER STREET
RUPERT STREET
RICHMOND ST
GEORGE ST
AVILLE ROW
QUEEN STREET
AIR STREET
VINE STREET
PICCADILLY



Snow used his map to convince local authorities to remove the handle of the Broad Street pump. Though the cholera epidemic was already on the wane when he did so, it is possible that the disabling of the pump prevented many deaths from future waves of the disease.

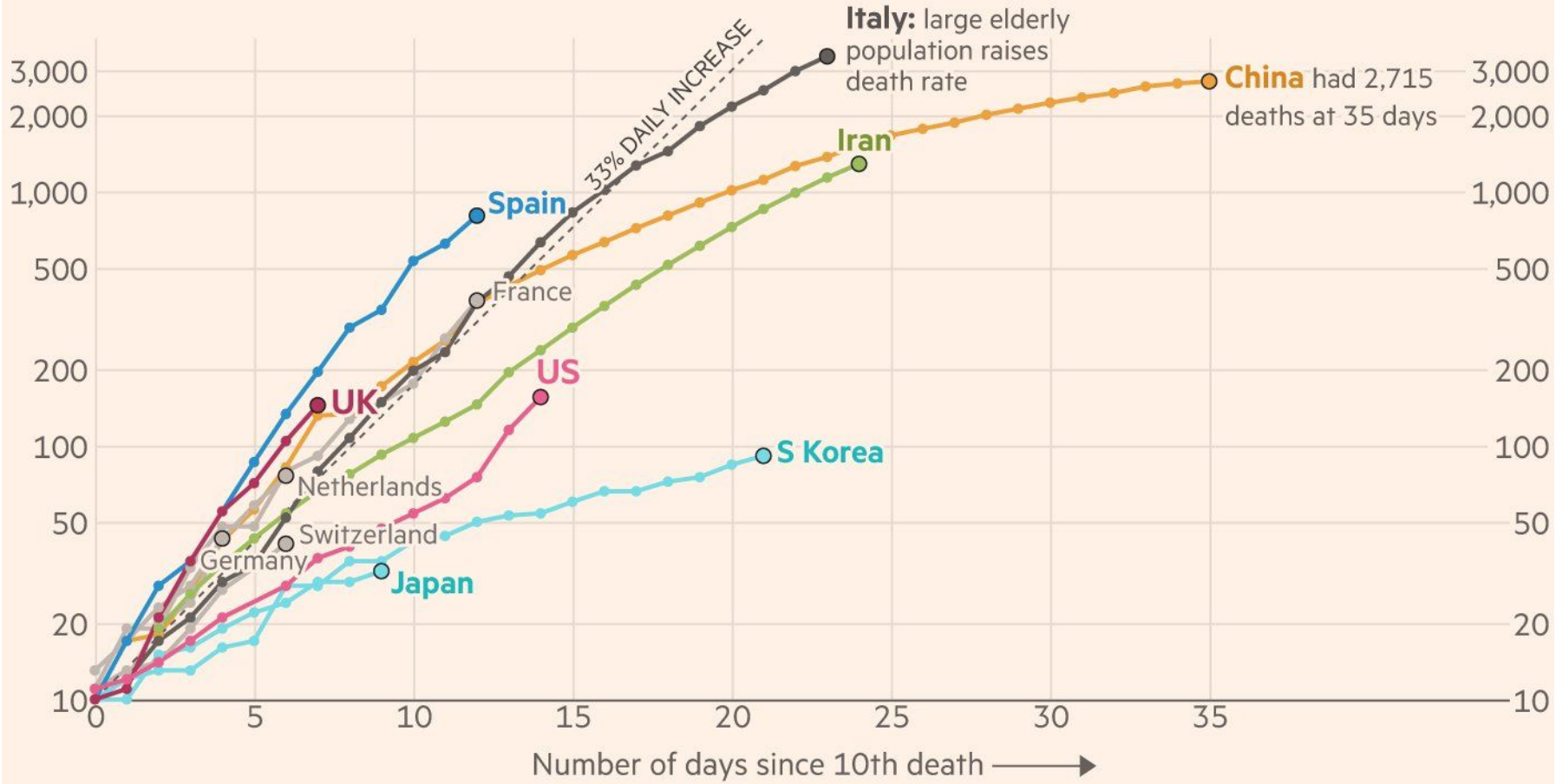


What this all means

- Your future jobs will revolve around information, knowledge, expertise. You will need to show verbal, written and **visual communication skills**
- Visualization is useful for everything that you will want to do when manipulating quantitative information: **description, inference, prescription**
- Think of visualization as a natural branch of what you already know how to do, i.e. **descriptive statistics** and **statistical models**, plus things like **maps** and **networks**

Coronavirus deaths in Italy and Spain are increasing much more rapidly than they did in China

Cumulative number of deaths, by number of days since 10th death



FT graphic: John Burn-Murdoch / @jburnmurdoch

Source: FT analysis of Johns Hopkins University, CSSE; Worldometers. Data updated March 19, 19:00 GMT

How to get there

How to draw an owl

1.

2.

1. Draw some circles

2. Draw the rest of the fucking owl

Learning blocks

- **Fundamentals**

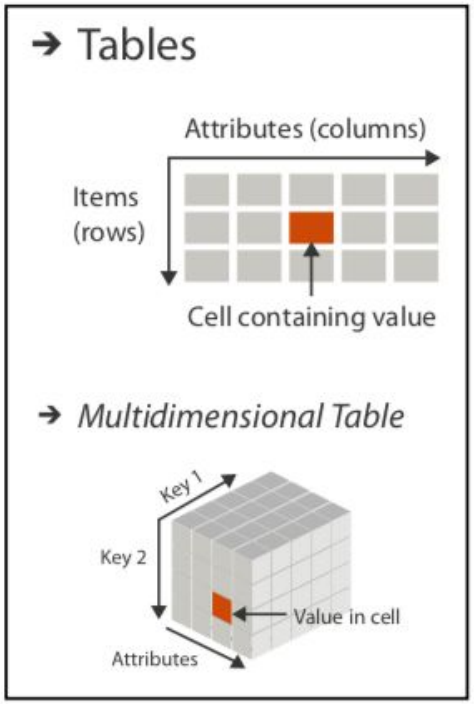
  e.g. Jacques Bertin, Otto Neurath, John Tukey

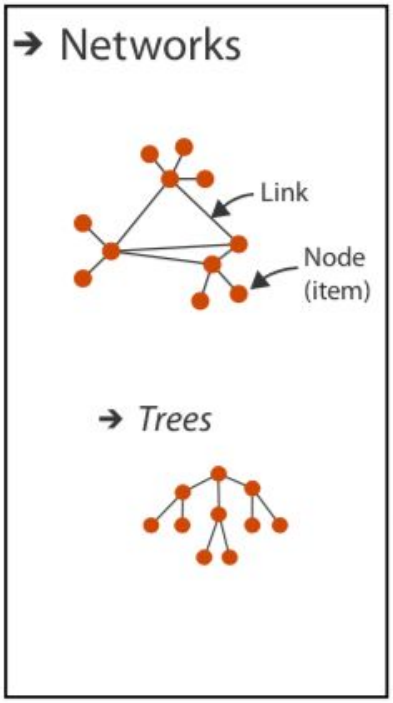
  - ‘Dataviz’ (data stories) · see *e.g.* Alberto Cairo
  - Computational graphics · see *e.g.* SIGGRAPH
- **Graphics in R**

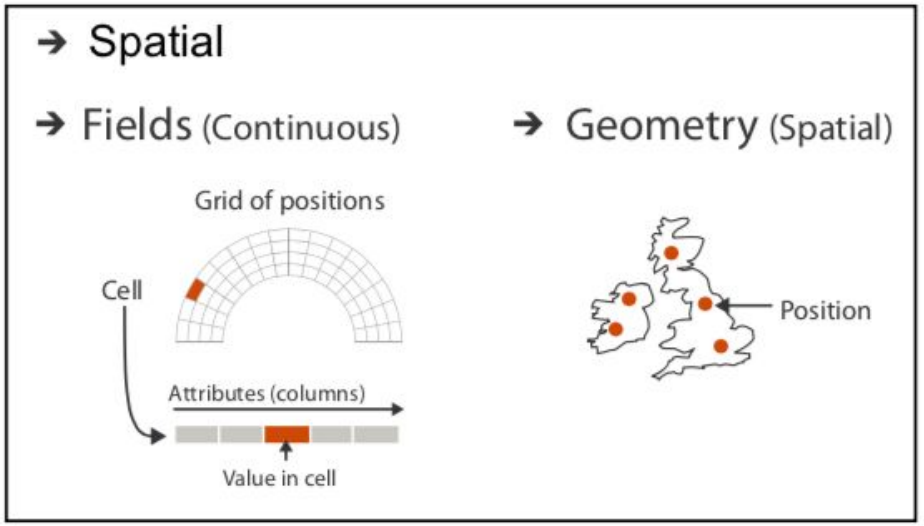
  via plotting systems and graphics devices

  - **base R**, **default colors** and **lattice**
  - **ggplot2** (part of the tidyverse)

Data abstraction

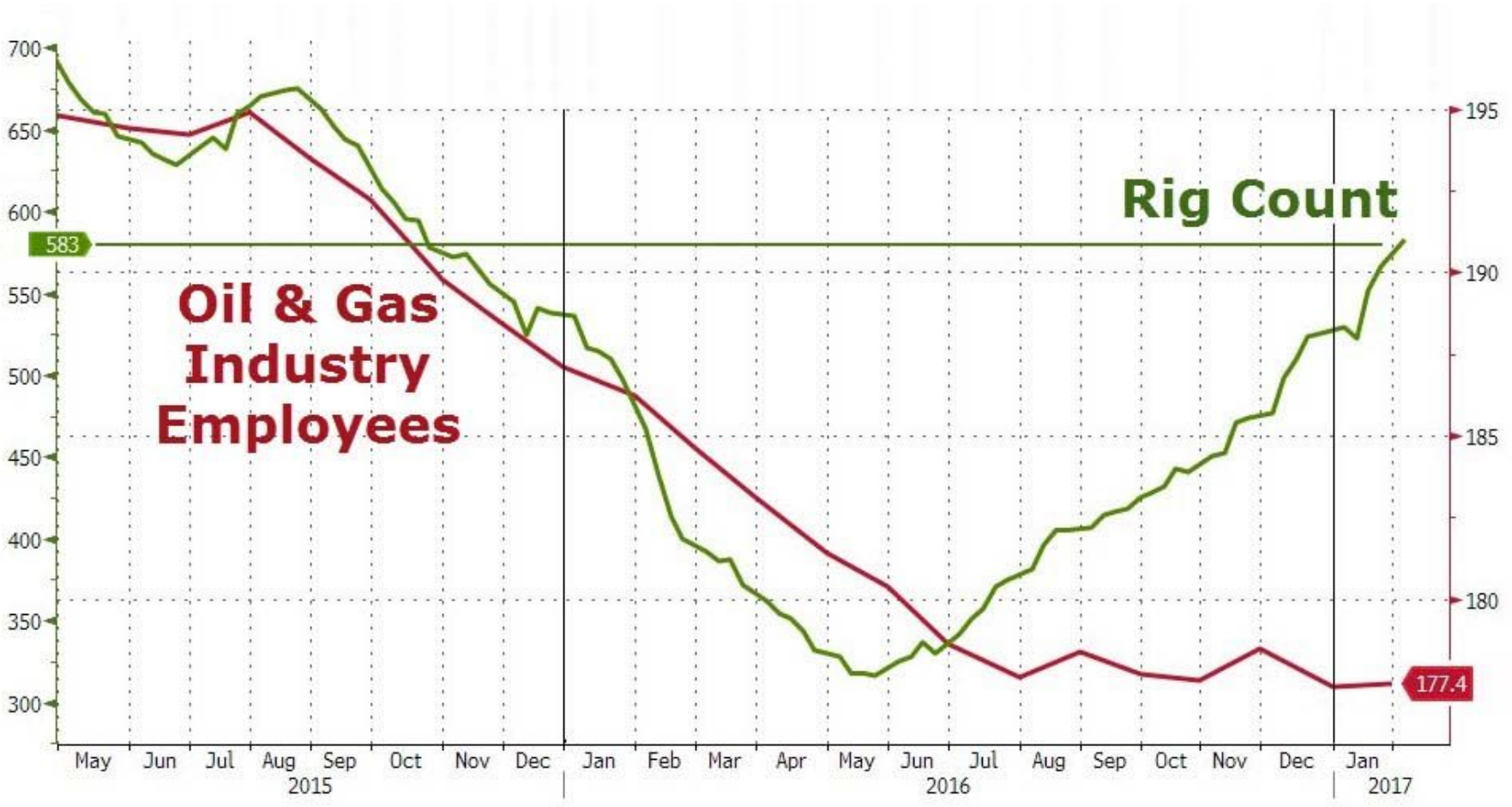






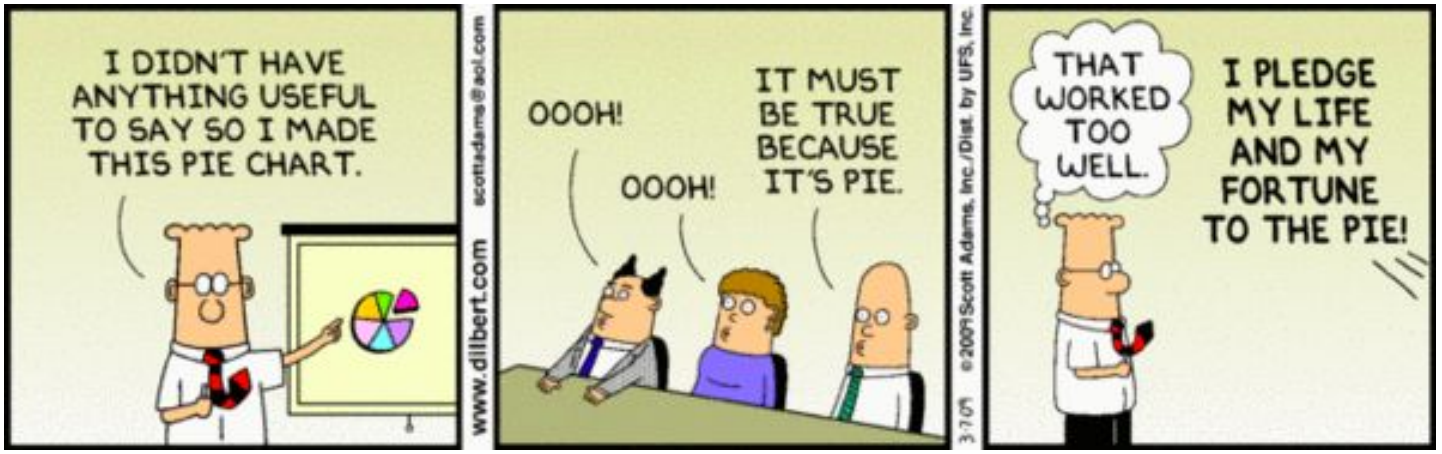
by Tamara Munzner

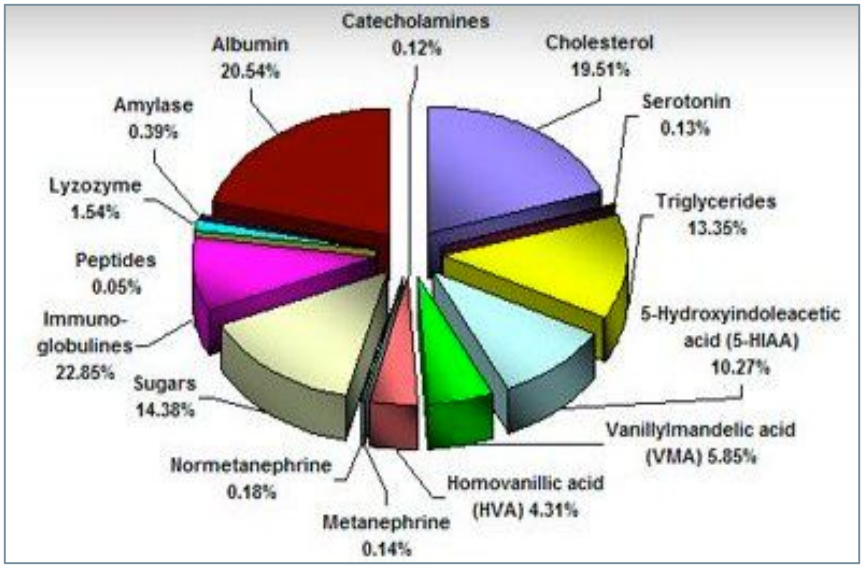
Do *not* use double axes

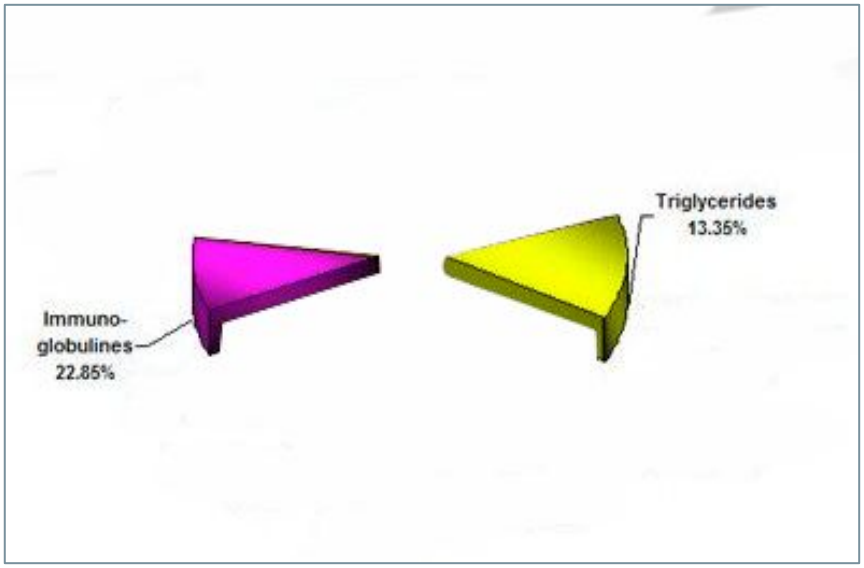


Do *not* use pie charts

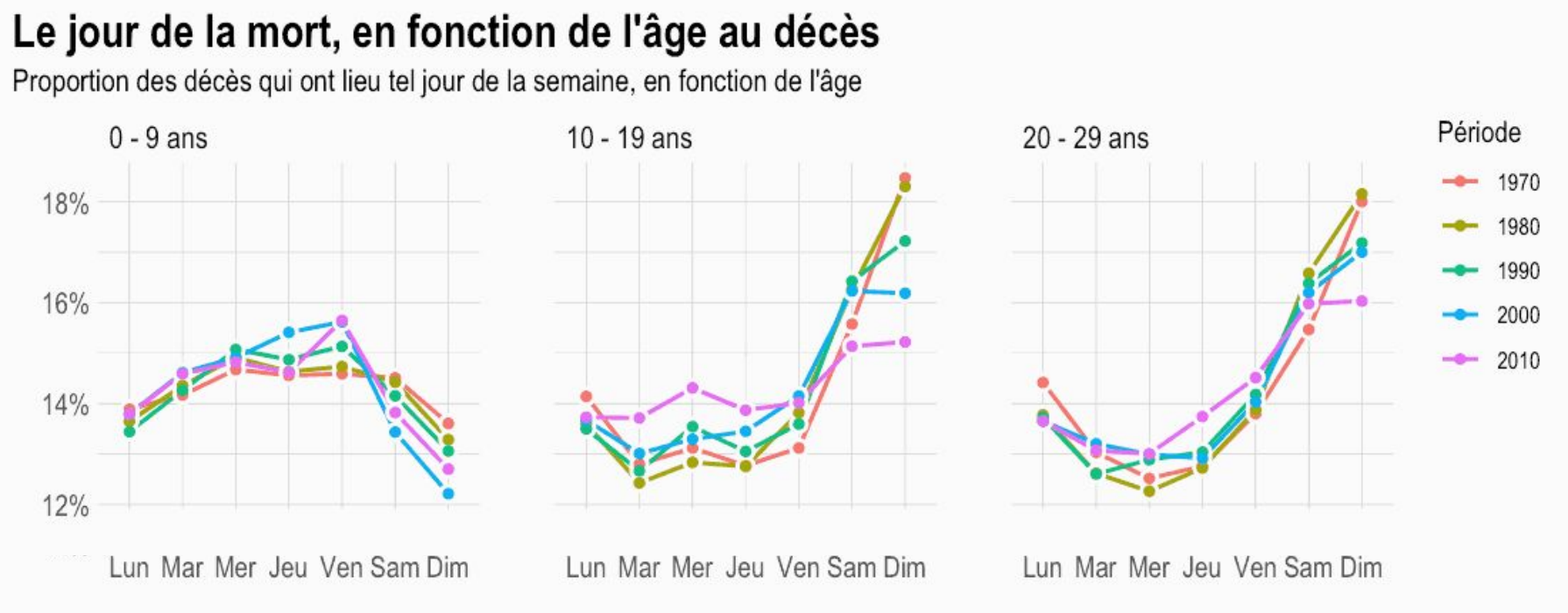
- Polar coordinates are (impossibly) hard to read
- 3-dimensional pie charts have their own place in Hell
- Pie charts generally have low data-ink ratios



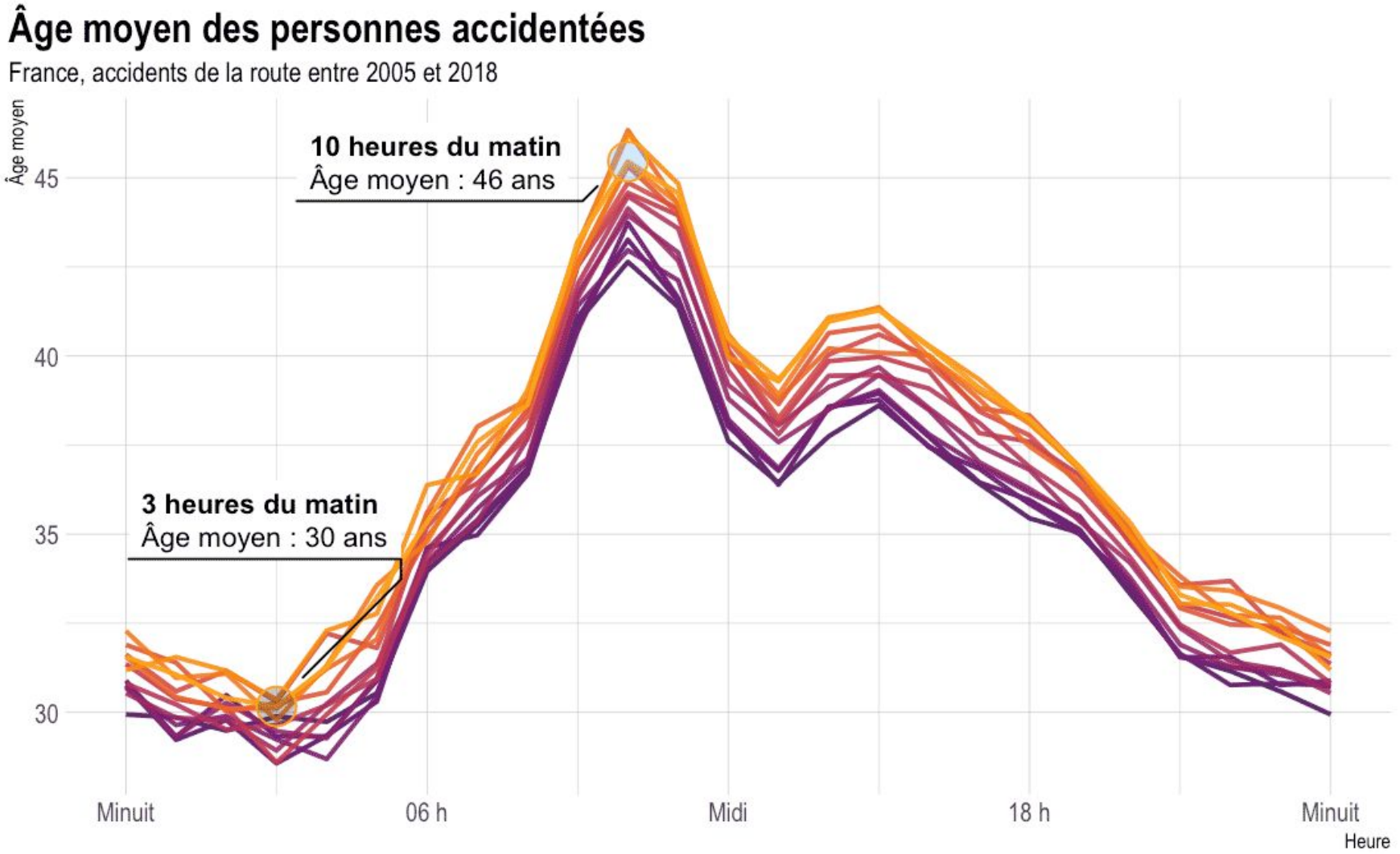

Albumin
20.54%
Catecholamines
0.12%
Cholesterol
19.51%
Amylase
0.39%
Serotonin
0.13%
Lyzozyme
1.54%
Triglycerides
13.35%
Peptides
0.05%
5-Hydroxyindoleacetic
acid (5-HIAA)
10.27%
Immuno-
globulines
22.85%
Sugars
14.38%
Vanillylmandelic acid
(VMA) 5.85%
Normetanephrine
0.18%
Homovanillic acid
(HVA) 4.31%
Metanephrine
0.14%


Triglycerides
13.35%
Immuno-
globulines
22.85%

Use small multiples (facets)



Use annotations

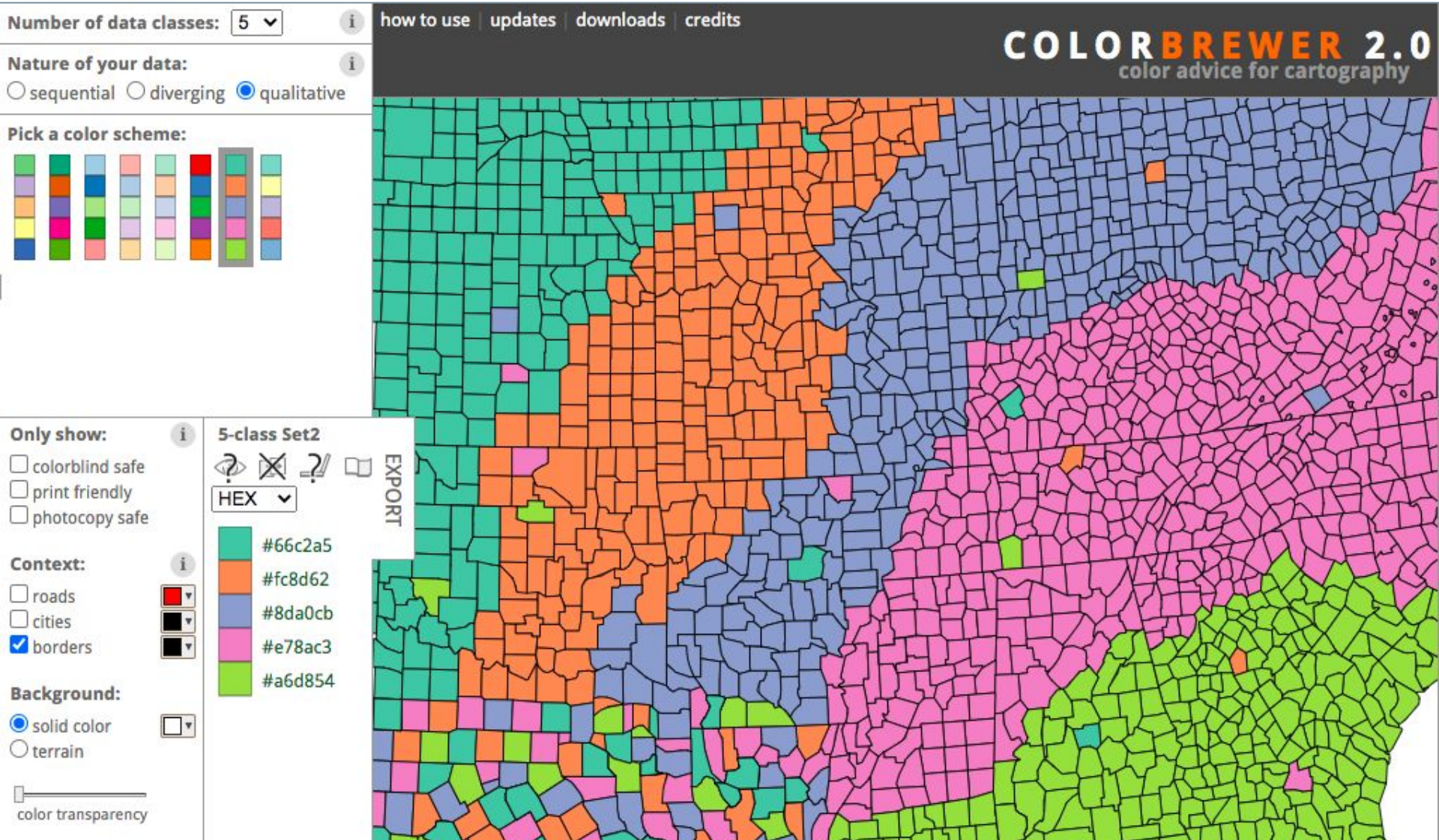


Plots with `ggplot2`

Your plots are layers

- **Data** – a data frame
- **Aesthetics** – mappings
- **Geometries** – what to draw
- **Facets** – small multiples
- **Statistics** – transformations
- **Coordinates** – planes
- **Theme** – cosmetics

You'll also need colors



Find which plot you need

The R Graph Gallery

Distribution


Violin


Density


Histogram


Boxplot


Ridgeline

Correlation


Scatter


Heatmap


Correlogram


Bubble


Connected scatter


Density 2d

Ranking


Barplot


Spider / Radar

Wordcloud


Parallel


Lollipop


Circular Barplot

Find example code

Data Visualization with ggplot2 : : CHEAT SHEET

Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same components: a **data** set, a **coordinate system**, and geoms—visual marks that represent data points.

To display values, map variables in the data to visual properties of the geom (**aesthetics**) like **size**, **color**, and **x** and **y** locations.

Complete the template below to build a graph.

required

```
ggplot(data = <DATA>) +
<GEOM_FUNCTION>(mapping = aes(<MAPPINGS>),
  stat = <STAT>, position = <POSITION>) +
<COORDINATE_FUNCTION> +
<FACET_FUNCTION> +
<SCALE_FUNCTION> +
<THEME_FUNCTION>
```

Not required, sensible defaults supplied

ggplot(data = mpg, **aes(**x = cty, y = hwy**))** Begins a plot that you finish by adding layers to. Add one geom function per layer.

aesthetic mappings | data | geom

qplot(x = cty, y = hwy, data = mpg, geom = "point**)** Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

last_plot() Returns the last plot

ggsave("plot.png", width = 5, height = 5) Saves last plot as 5' x 5' file named "plot.png" in working directory. Matches file type to file extension.

Geoms

Use a geom function to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

GRAPHICAL PRIMITIVES

```
a <- ggplot(economics, aes(date, unemploy))
b <- ggplot(seals, aes(x = long, y = lat))
```

a + geom_blank()
(Useful for expanding limits)

b + geom_curve(aes(yend = lat + 1, xend=long+1),curvature=1**)** - x, xend, y, yend, alpha, angle, color, curvature, linetype, size

a + geom_path(lineend="butt", linejoin="round", linemitre=1**)**
x, y, alpha, color, group, linetype, size

a + geom_polygon(aes(group = group**))**
x, y, alpha, color, fill, group, linetype, size

b + geom_rect(aes(xmin = long, ymin=lat, xmax= long + 1, ymax = lat + 1**))** - xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

a + geom_ribbon(aes(ymin=unemploy - 900, ymax=unemploy + 900**))** - x, ymax, ymin, alpha, color, fill, group, linetype, size

LINE SEGMENTS

common aesthetics: x, y, alpha, color, linetype, size

b + geom_abline(aes(intercept=0, slope=1**))**
b + geom_hline(aes(yintercept = lat**))**
b + geom_vline(aes(xintercept = long**))**

b + geom_segment(aes(yend=lat+1, xend=long+1**))**
b + geom_spoke(aes(angle = 1:1155, radius = 1**))**

ONE VARIABLE continuous

```
c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)
```

c + geom_area(stat = "bin")
x, y, alpha, color, fill, linetype, size

c + geom_density(kernel = "gaussian"**)**
x, y, alpha, color, fill, group, linetype, size, weight

c + geom_dotplot()
x, y, alpha, color, fill

c + geom_freqpoly() x, y, alpha, color, group, linetype, size

c + geom_histogram(binwidth = 5**)** x, y, alpha, color, fill, linetype, size, weight

c2 + geom_qq(aes(sample = hwy**))** x, y, alpha, color, fill, linetype, size, weight

discrete

```
d <- ggplot(mpg, aes
```

d + geom_
x, alpha, color, fill, linetype, size, weight

TWO VARIABLES

continuous x , continuous y

```
e <- ggplot(mpg, aes(cty, hwy))
```

e + geom_label(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE**)** x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

e + geom_jitter(height = 2, width = 2**)**
x, y, alpha, color, fill, shape, size

e + geom_point(), x, y, alpha, color, fill, shape, size, stroke

e + geom_quantile(), x, y, alpha, color, group, linetype, size, weight

e + geom_rug(sides = "bl"**)**, x, y, alpha, color, linetype, size

e + geom_smooth(method = lm**)**, x, y, alpha, color, fill, group, linetype, size, weight

e + geom_text(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE**)**, x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

discrete x , continuous y

```
f <- ggplot(mpg, aes(class, hwy))
```

f + geom_col(), x, y, alpha, color, fill, group, linetype, size

f + geom_boxplot(), x, y, lower, middle, upper, ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight

f + geom_dotplot(binaxis = "y", stackdir = "center"**)**, x, y, alpha, color, fill, group

f + geom_violin(scale = "area"**)**, x, y, alpha, color, fill, group, linetype, size, weight

discrete x , discrete y

```
g <- ggplot(diamonds, aes(cut, color))
```

g + geom_count(), x, y, alpha, color, fill, shape, size, stroke

continuous bivariate distribution

```
h <- ggplot(diamonds, aes(carat, price))
```

h + geom_bin2d(binwidth = c(0.25, 500)**)**
x, y, alpha, color, fill, linetype, size, weight

h + geom_density2d()
x, y, alpha, colour, group, linetype, size

h + geom_hex()
x, y, alpha, colour, fill, size

continuous function

```
i <- ggplot(economics, aes(date, unemploy))
```

i + geom_area()
x, y, alpha, color, fill, linetype, size

i + geom_line()
x, y, alpha, color, group, linetype, size

i + geom_step(direction = "hv")
x, y, alpha, color, group, linetype, size

visualizing error

```
df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)
j <- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+se))
```

j + geom_crossbar(fatten = 2**)**
x, y, ymax, ymin, alpha, color, fill, group, linetype, size

j + geom_errorbar(), x, ymax, ymin, alpha, color, group, linetype, size, width (also **geom_errorbarh()**)

j + geom_linerange()
x, ymin, ymax, alpha, color, group, linetype, size

j + geom_pointrange()
x, y, ymin, ymax, alpha, color, fill, group, linetype, shape, size

maps

```
data <- data.frame(murder = USArrests$Murder,
state = tolower(rownames(USArrests)))
map <- map_data("state")
k <- ggplot(data, aes(fill = murder))
```

k + geom_map(aes(map_id = state), map = map**)**
+ expand_limits(x = map$long, y = map$lat**)**,
map_id, alpha, color, fill, linetype, size

THREE VARIABLES

```
2 + delta_lat^2)); l <- ggplot(seals, aes(long, lat))
```

l + geom_raster(aes(fill = z), hjust=0.5, vjust=0.5, interpolate=FALSE**)**
x, y, alpha, fill

l + geom_tile(aes(fill = z)**)**, x, y, alpha, color, fill, linetype, size, width

Learn the syntax

ggplot2 part of the tidyverse 3.3.2

Search...

Reference

Reference

Plot basics

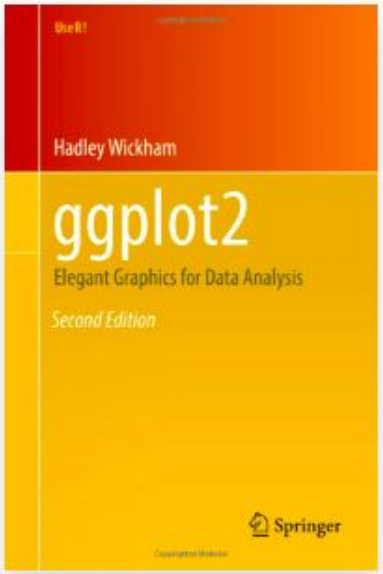
All ggplot2 plots begin with a call to `ggplot()`, supplying default data and aesthetic mappings, specified by `aes()`. You then add layers, scales, coords and facets with `+`. To save a plot to disk, use `ggsave()`.

`ggplot()`	Create a new ggplot
`aes()`	Construct aesthetic mappings
`` `+`(<gg>) `%+%` ``	Add components to a plot
`ggsave()`	Save a ggplot (or other grid object) with sensible defaults
`qplot()` `quickplot()`	Quick plot

Use the (excellent) documentation

Welcome

This is the on-line version of work-in-progress **3rd edition** of "ggplot2: elegant graphics for data analysis" published by Springer. You can learn what's changed from the 2nd edition in the Preface.



While this book gives some details on the basics of ggplot2, it's primary focus is explaining the Grammar of Graphics that ggplot2 uses, and describing the full details. It is not a cookbook, and won't necessarily help you create any specific graphic that you need. But it will help you understand the details of the underlying theory, giving you the power to tailor any plot specifically to your needs.

The book is written by Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen.

Preface to the third edition »

The book covers a few more things

Good ways to get started

- *R for Data Science*, ch. 3 (Data Visualization)
- Video tutorials by Thomas Pedersen: part 1, part 2 (2020)
- Detailed guide to the bar chart in R with ggplot2 (2019)
- A ggplot2 tutorial for beautiful plotting in R (2018)
- Data visualization using ggplot2 (2016)

There are also lots and lots of blogs (e.g. Jason Timm's) showing beautiful examples of `ggplot2` in action, and lots answers to lots of questions on StackOverflow

Example extensions to ggplot2

For various uses

- GGally (lots of different plots and tables)
- `ggfortify` (excels with e.g. PCA results)

For regression models

`dotwhisker` Dot-whisker plots for regression results

`interplot` Interaction terms in regression models

`ggeffects` Marginal effects for regression models

Activity / Homework

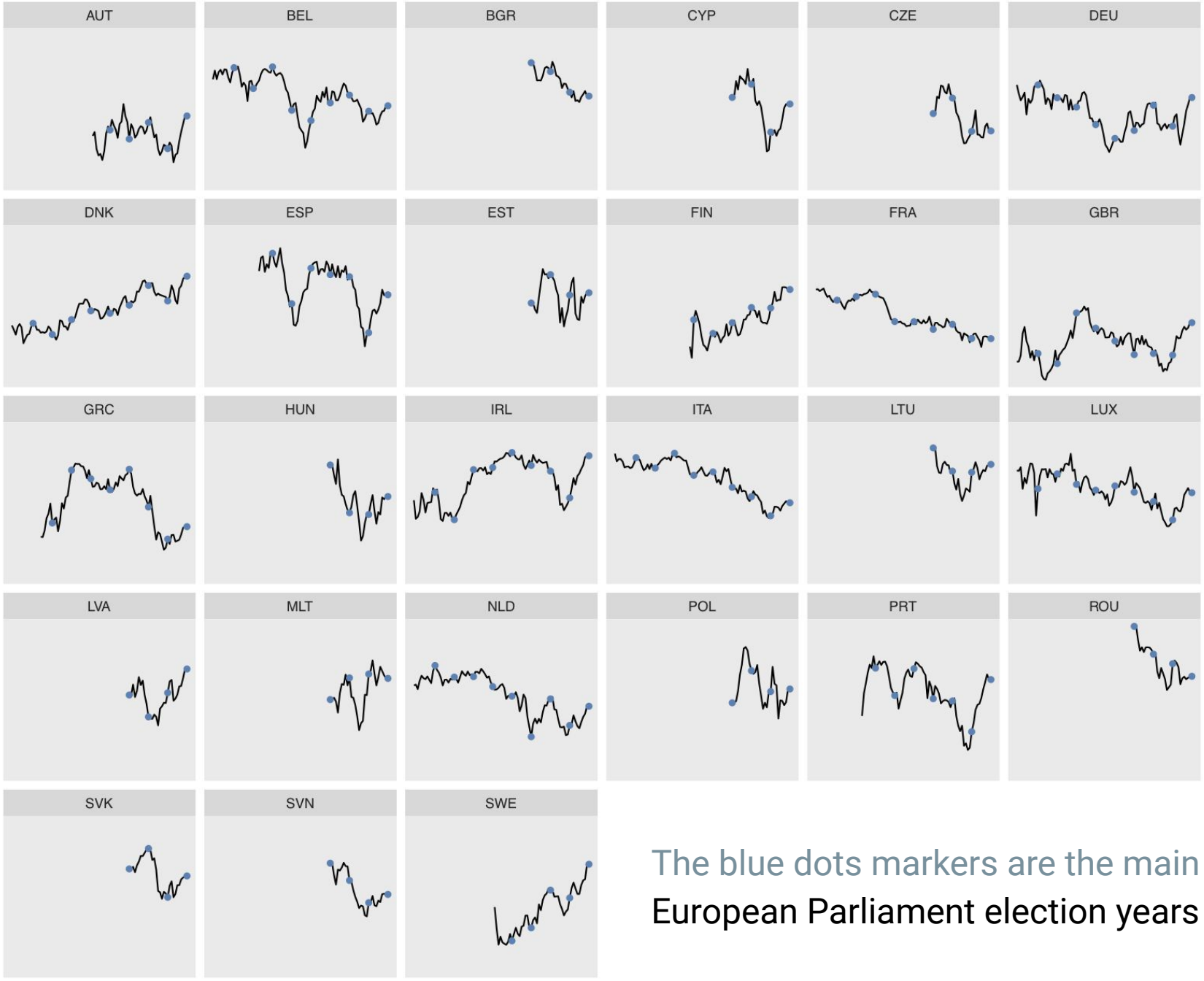
Practice material

Unzip the `quanti2-s5.zip` archive

Open the `s5-code.Rproj` R project

Follow along as we go through

1. **Plotting descriptive statistics** (Session 4 Activity)
2. **ggplot2 fundamentals** with Anscombe's quartet
3. **Plotting model results** with **broom** **and ggplot2**

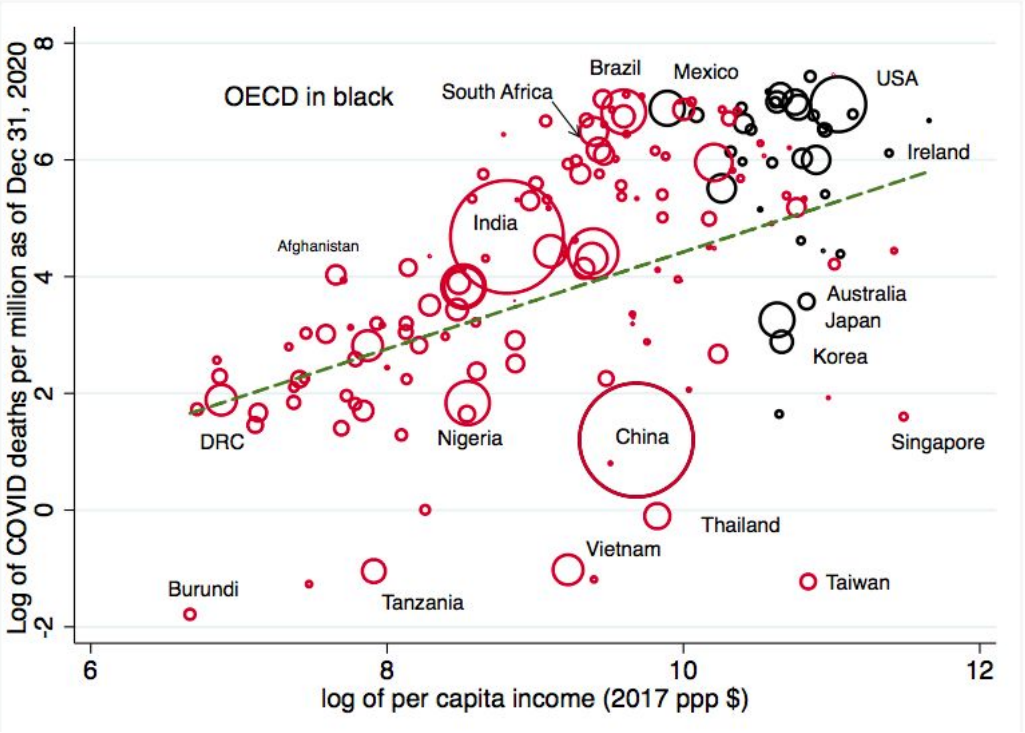


The blue dots markers are the main European Parliament election years

Easy activity for Session 5 (due for Session 11)

- Read this recent working paper by Angus Deaton
- Reproduce Figure 1 as closely as possible, without worrying too much about labelling all countries

Hint: the data come from the World Bank Development Indicators



Hard activity for Session 5 (due for Session 11)

- Get the Quality of Government Time Series dataset
- Using the full (pooled) country-year dataset, regress life expectancy on female education at age 15-24, current GDP/capita, having never been colonized, and being a democracy
- Plot the coefficients using point estimates with 95% confidence intervals
- Produce similar coefficients for each year, and plot them to see how they change through time

Useful resources

Data Visualization in Sociology

Kieran Healy and James Moody

Annu. Rev. Sociol. 2014. 40:105–28

First published online as a Review in Advance on June 6, 2014

The *Annual Review of Sociology* is online at soc.annualreviews.org

This article's doi: 10.1146/annurev-soc-071312-145551

Keywords

visualization, statistics, methods, exploratory data analysis

Abstract

Visualizing data is central to social scientific work. Despite a promising early beginning, sociology has lagged in the use of visual tools. We review the history and current state of visualization in sociology. Using examples throughout, we discuss recent developments in ways of seeing raw data and presenting the results of statistical modeling. We make a general distinction between those methods and tools designed to help explore data sets and those designed to help present results to others. We argue that recent advances should be seen as part of a broader shift toward easier sharing of code and data both between researchers and with wider publics, and we encourage practitioners and publishers to work toward a higher and more consistent standard for the graphical display of sociological insights.

Data Visualization

Use R, ggplot2, and the principles of graphic design to create beautiful and truthful visualizations of data

PMAP 8921 • May 2020
Andrew Young School of Policy Studies
Georgia State University

Instructor

Dr. Andrew Heiss
357 Andrew Young School
aheiss@gsu.edu
@andrewheiss

Course details

Every day
May 11–June 1, 2020
Whenever
Slack

Contacting me

E-mail and Slack are the best ways
contact with me. I will try to respon
course-related e-mails and Slack m
within 24 hours (*really*), but also re
life can be busy and chaotic for eve

Gaston Sanchez

Introduction to Data Visualization

- Introduction
- Classic Examples
- Visualization Basics
- Visual System
- Visual Perception
- What is Color
- Color Vision
- Effective Charts
- Various Examples
- Graphing Process
- Entertainment

get it on GitHub

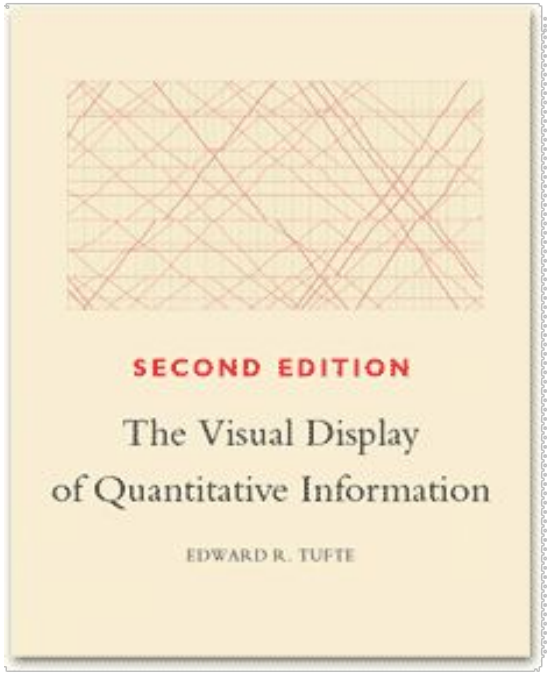
gastonstat / intro2datavis

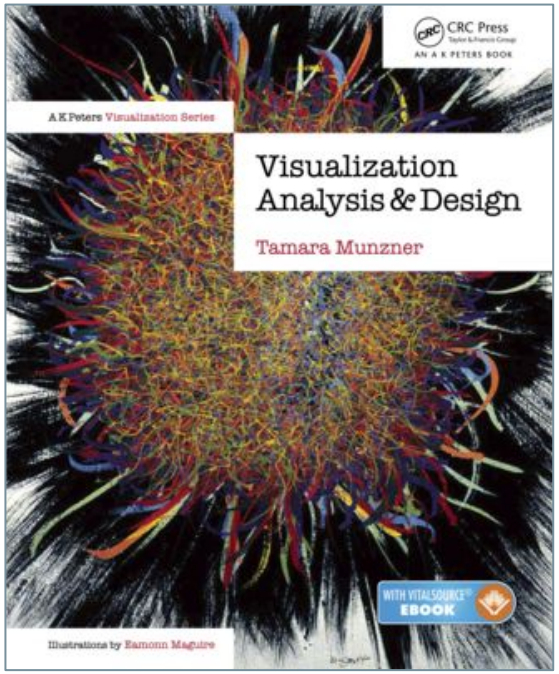
click this 

and select ‘Download ZIP’

Useful books

Tufte *The Visual Display of Quantitative Information*





Munzner *Visualization Analysis and Design*

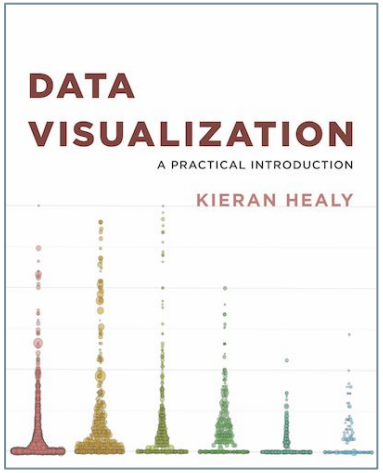
Useful books that use R

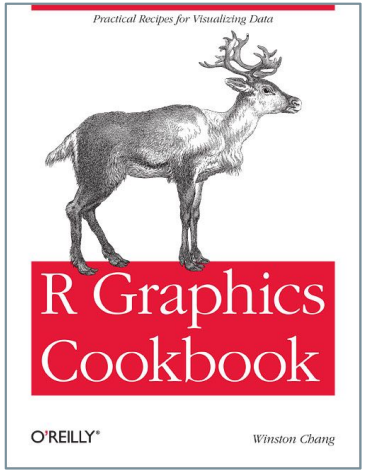
Healy *Data Visualization*

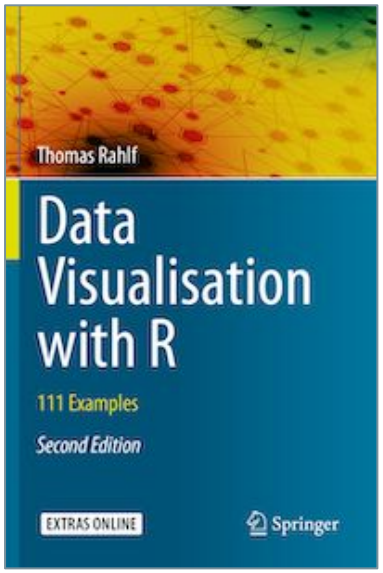
Chang *R Graphics Cookbook*

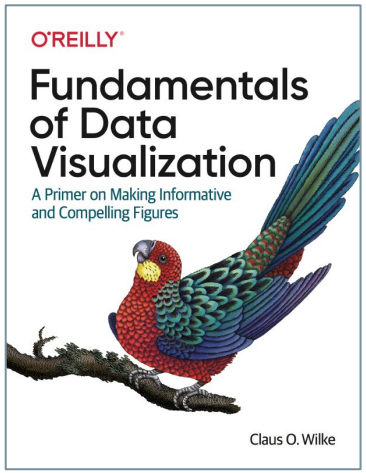
Wilke *Fundamentals of Data Visualization*

Rahlf *Data Visualisation with R — 111 Examples* (using base R)









A few more inspiring websites

Gapminder (RIP Hans Rosling)

Our World in Data (Max Roser)

Flowing Data (Nathan Yau)

Observable (uses d3.js — Mike Bostock *et al.*)

The R Graph Gallery (Yan Holtz)

Visions Carto (Philippe Rivière *et al.*)

Thanks for your attention

See you again for

Session 11 (topic TBD)