Data Vis

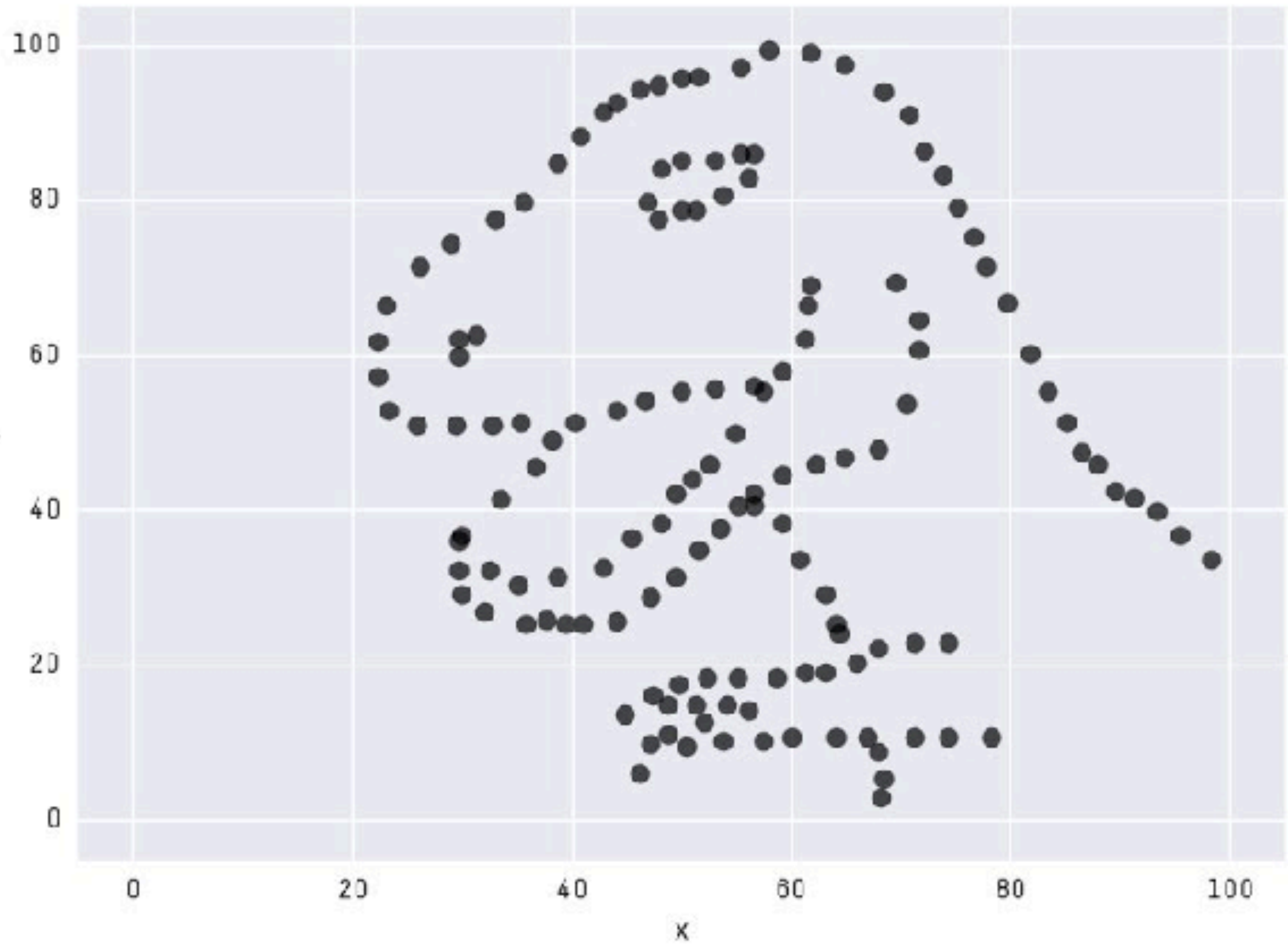
with ggplot2

Why and How

http://bit.ly/2xEUGIu

Przemysław Biecek
http://biecek.pl

Package datasauRus



```
X Mean: 54.26320025
Y Mean: 47.83315781
X SD  : 16.76550109
Y SD  : 26.93553144
Corr. : -0.06445195
```

https://www.autodeskresearch.com/publications/samestats

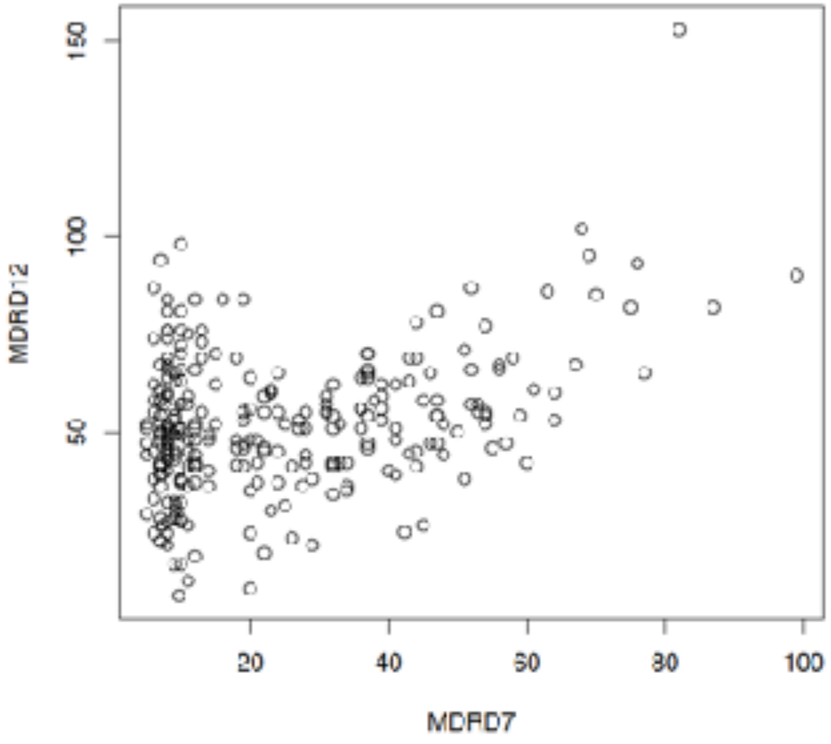
Grammar of Graphics
ggplot2

Three ecosystems for static statistical graphics

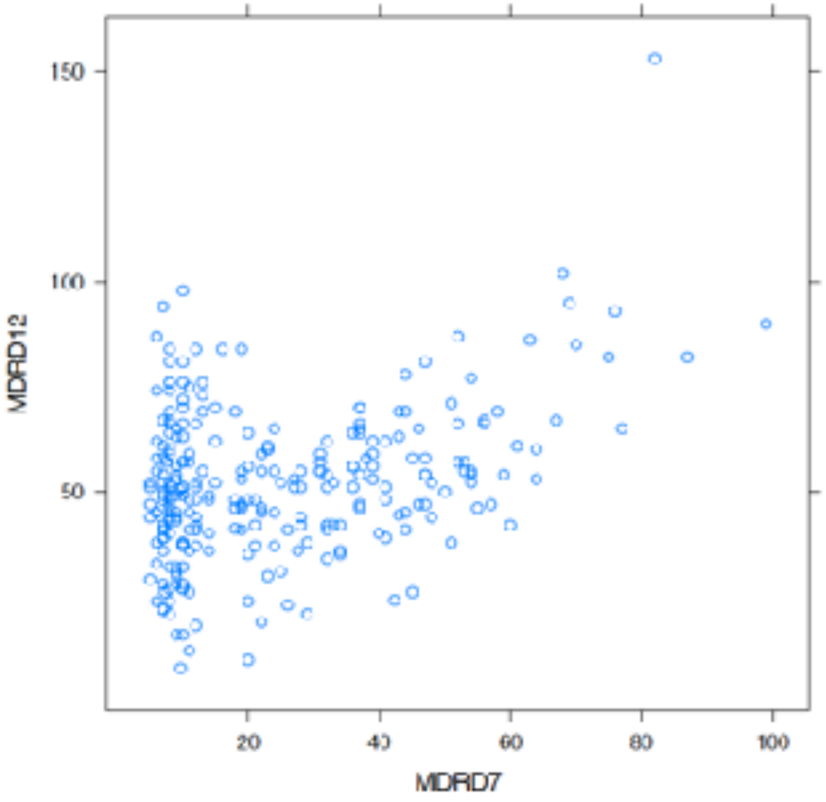
```
library(PBImisc)

# library graphics
plot(MDRD12~MDRD7, kidney)

# library lattice
xyplot(MDRD12~MDRD7, kidney)

# library ggplot2
qplot(MDRD12,MDRD7, data=kidney)
```

Find differences between these plots

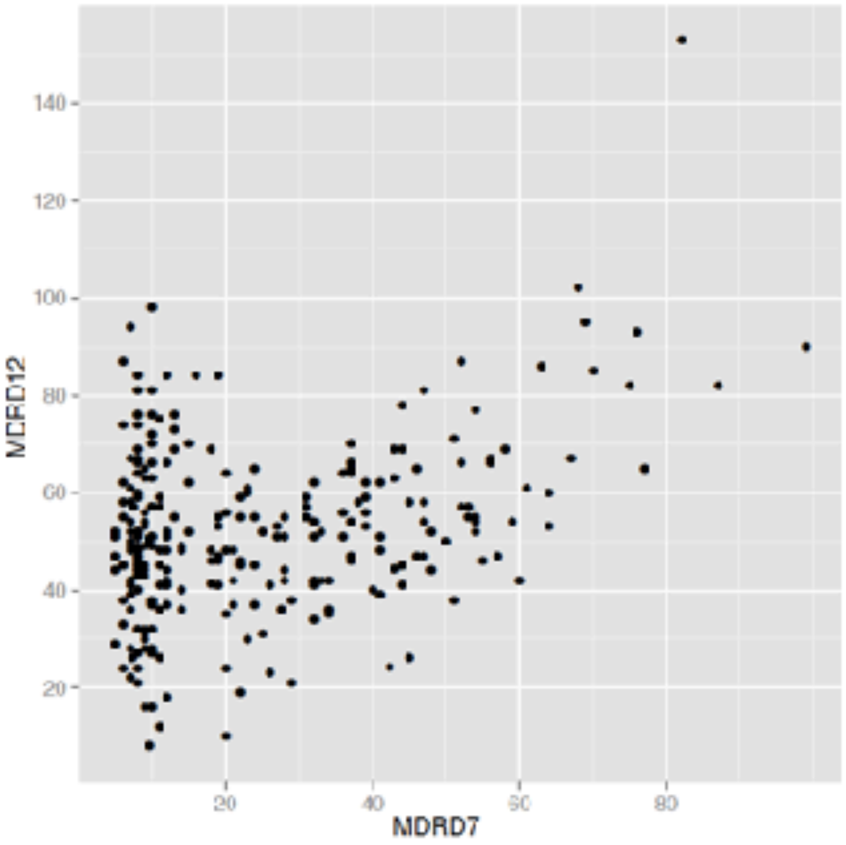
graphics



lattice



ggplot2

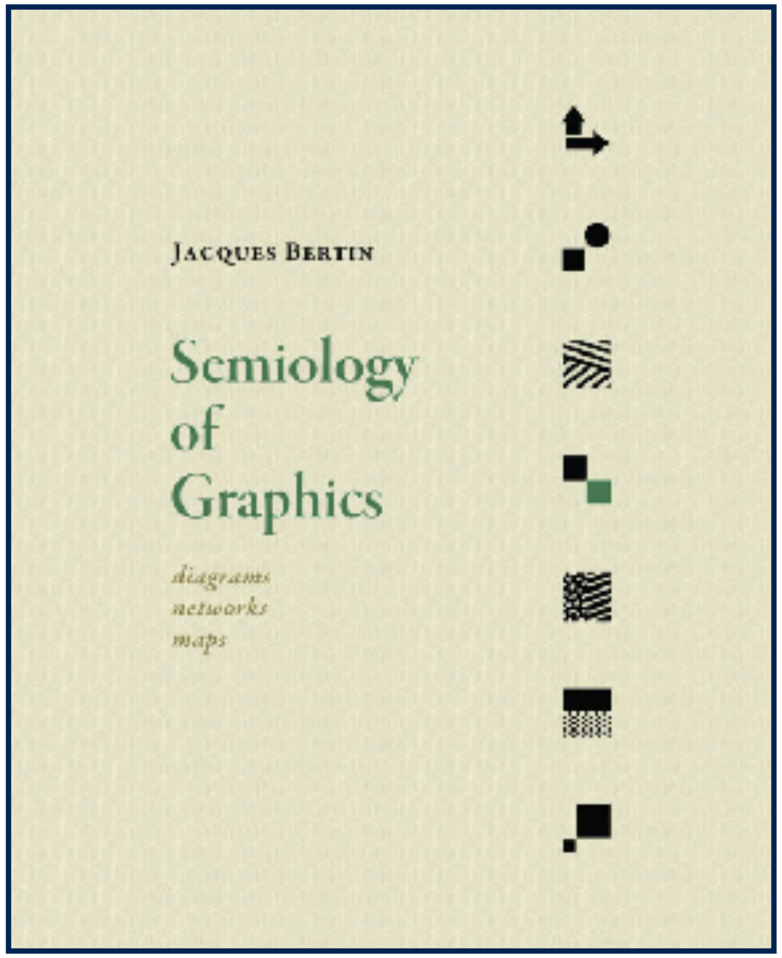


Why ggplot2?

+ Elegant
+ Highly customisable
+ Uniform
+ Natural
+ Expressive
+ Popular

- Steep learning curve
- Slow
- Evolving pretty fast (too fast?)

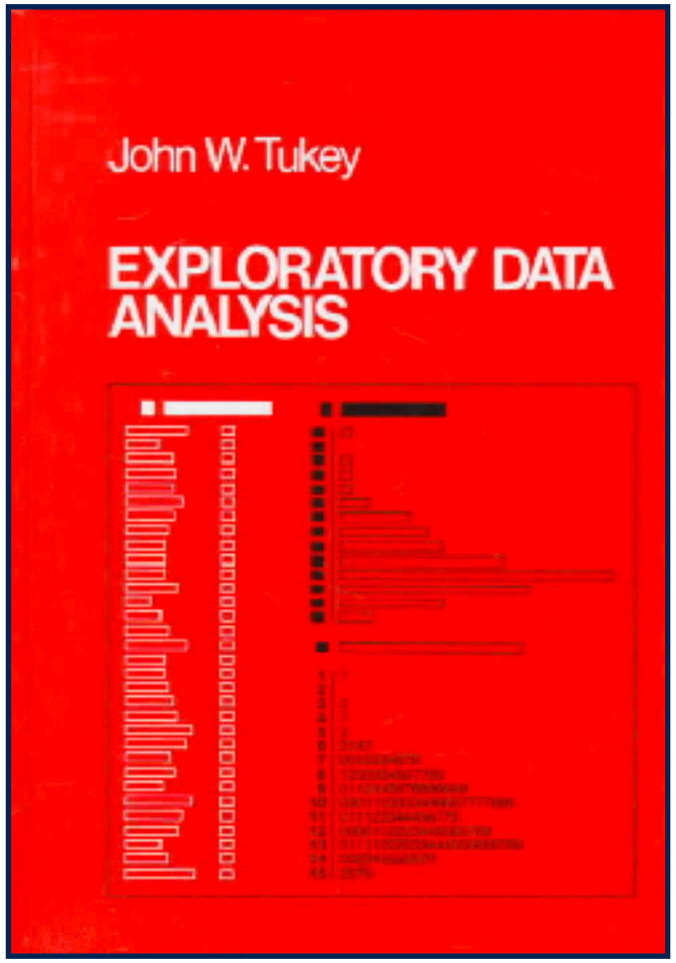
Why ggplot2?



Jacques Bertin

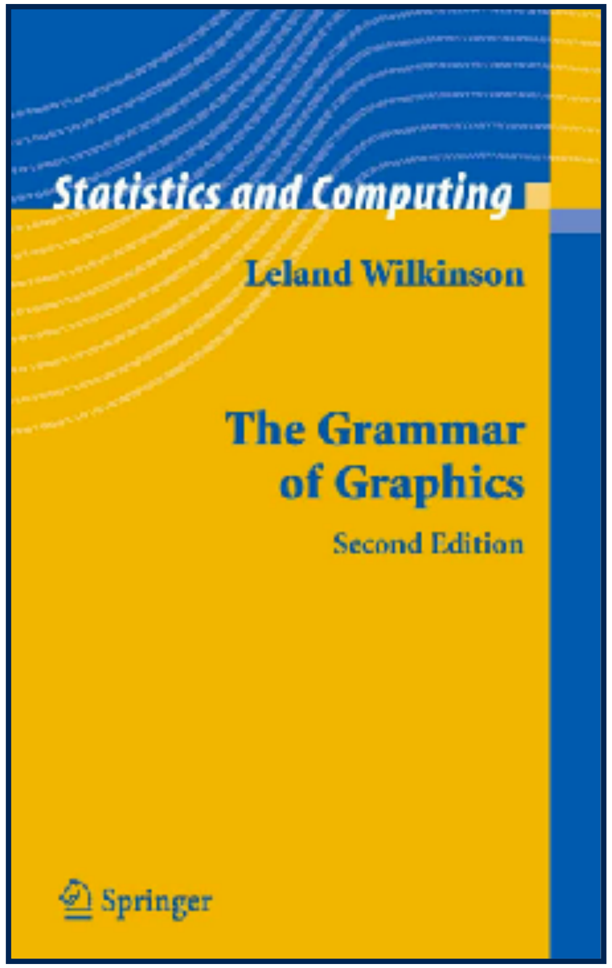
Semiology of Graphics

diagrams
networks
maps



John W. Tukey

EXPLORATORY DATA ANALYSIS



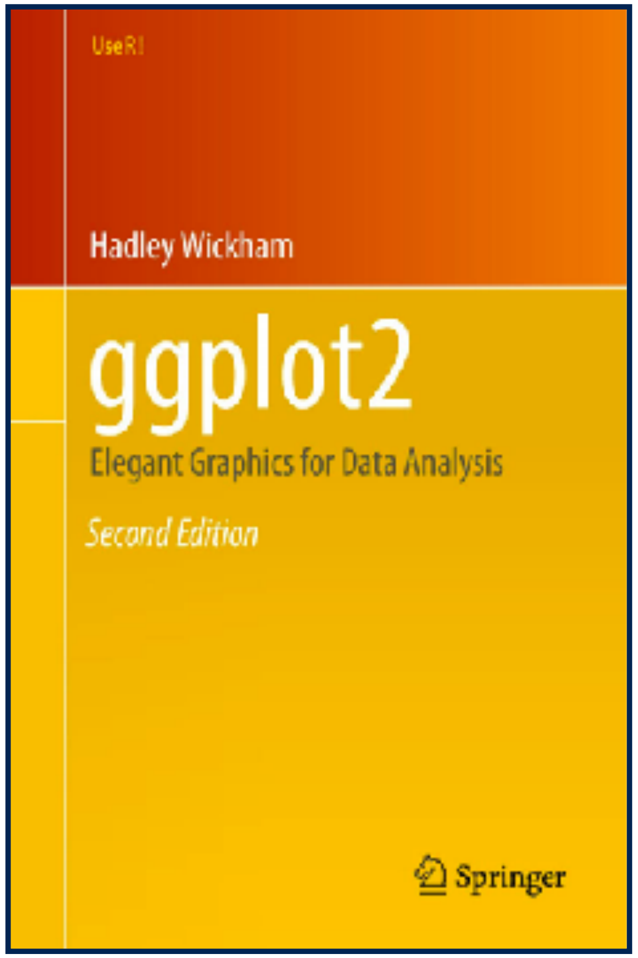
Statistics and Computing

Leland Wilkinson

The Grammar of Graphics

Second Edition

Springer



Use R!

Hadley Wickham

ggplot2

Elegant Graphics for Data Analysis

Second Edition

Springer

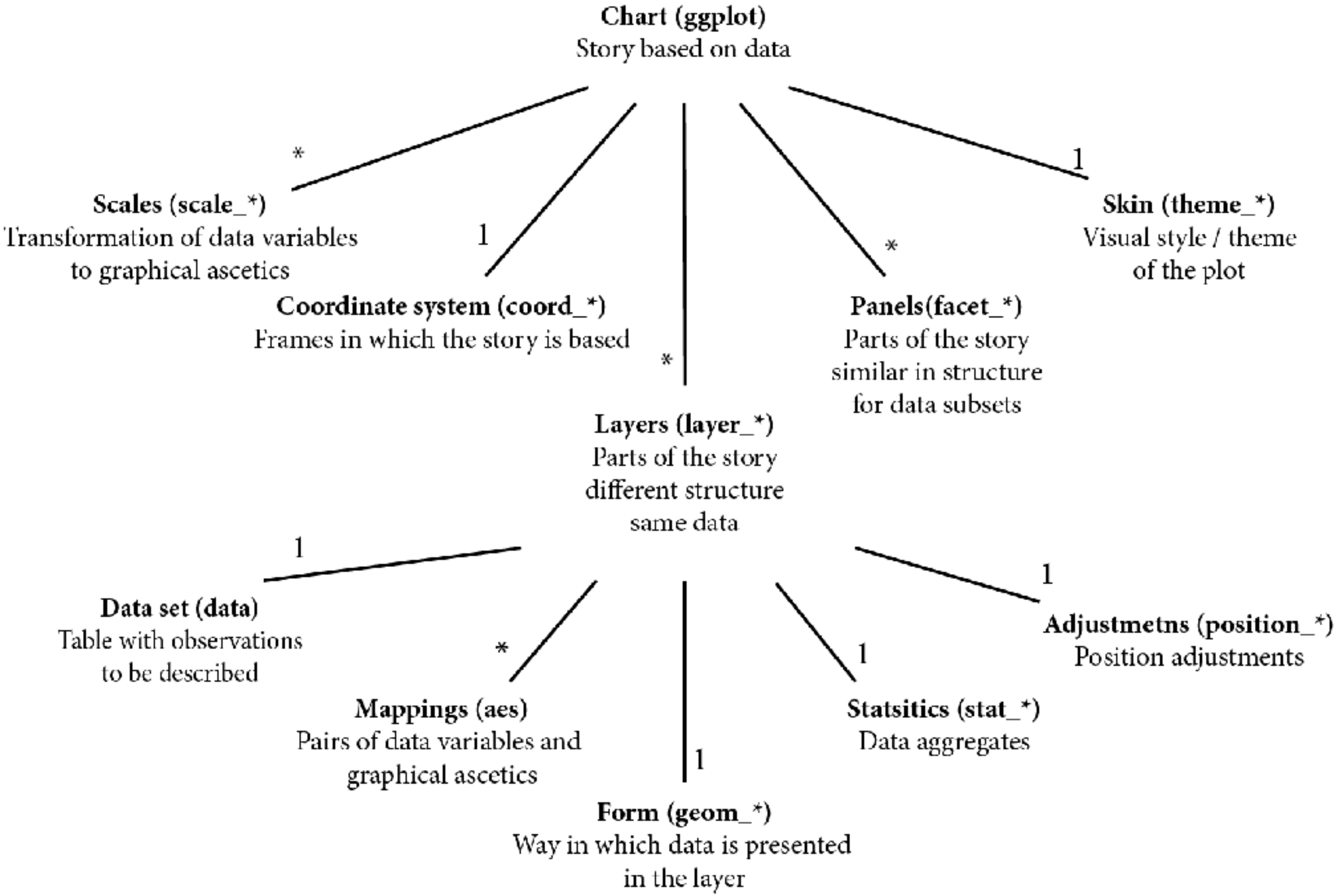


Chart (ggplot)
Story based on data

*
Scales (scale_*)
Transformation of data variables
to graphical ascetics

1
Coordinate system (coord_*)
Frames in which the story is based

*
Layers (layer_*)
Parts of the story
different structure
same data

*
Panels(facet_*)
Parts of the story
similar in structure
for data subsets

1
Skin (theme_*)
Visual style / theme
of the plot

1
Data set (data)
Table with observations
to be described

*
Mappings (aes)
Pairs of data variables and
graphical ascetics

1
Form (geom_*)
Way in which data is presented
in the layer

1
Statsitics (stat_*)
Data aggregates

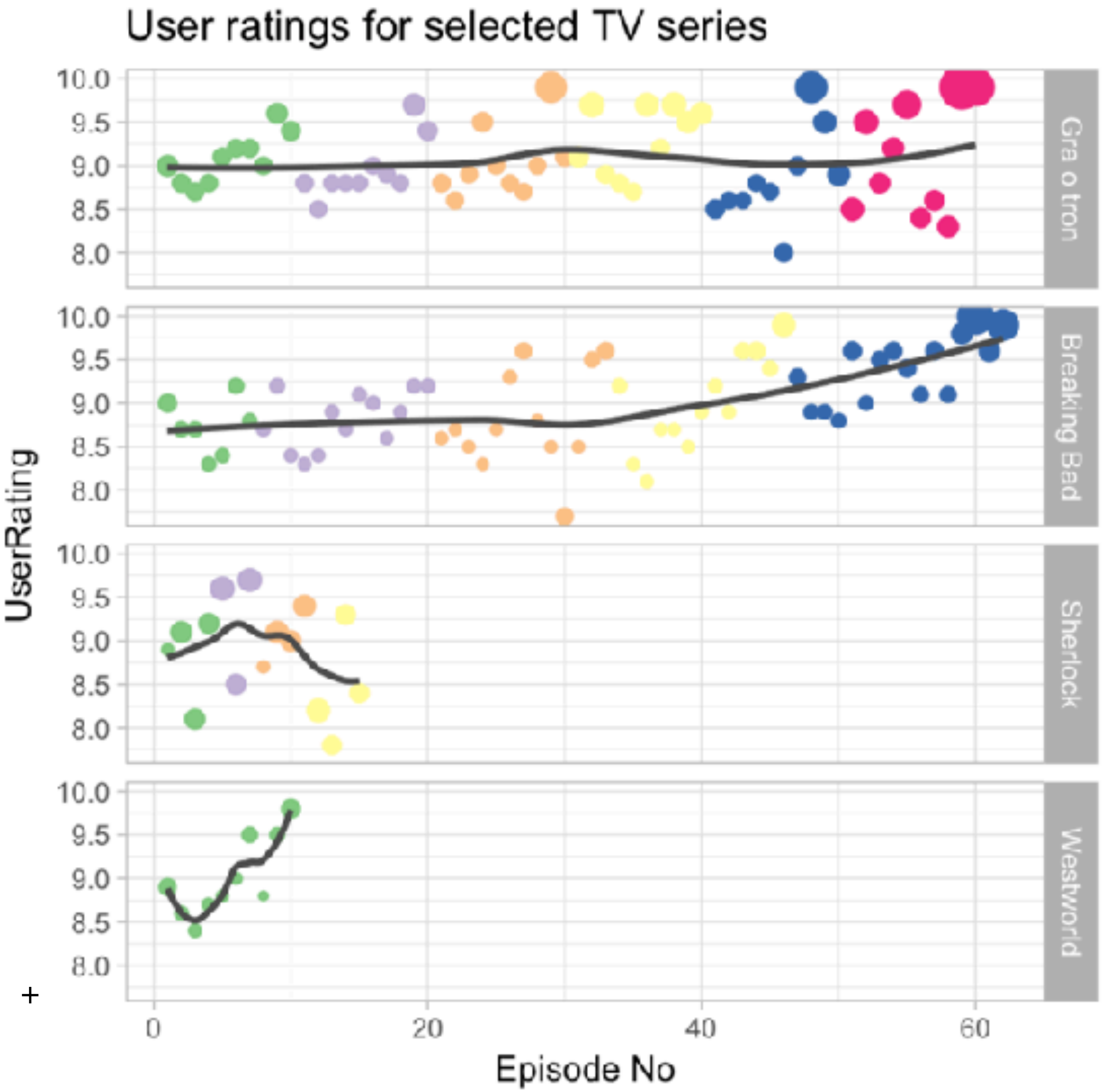
1
Adjustmetns (position_*)
Position adjustments

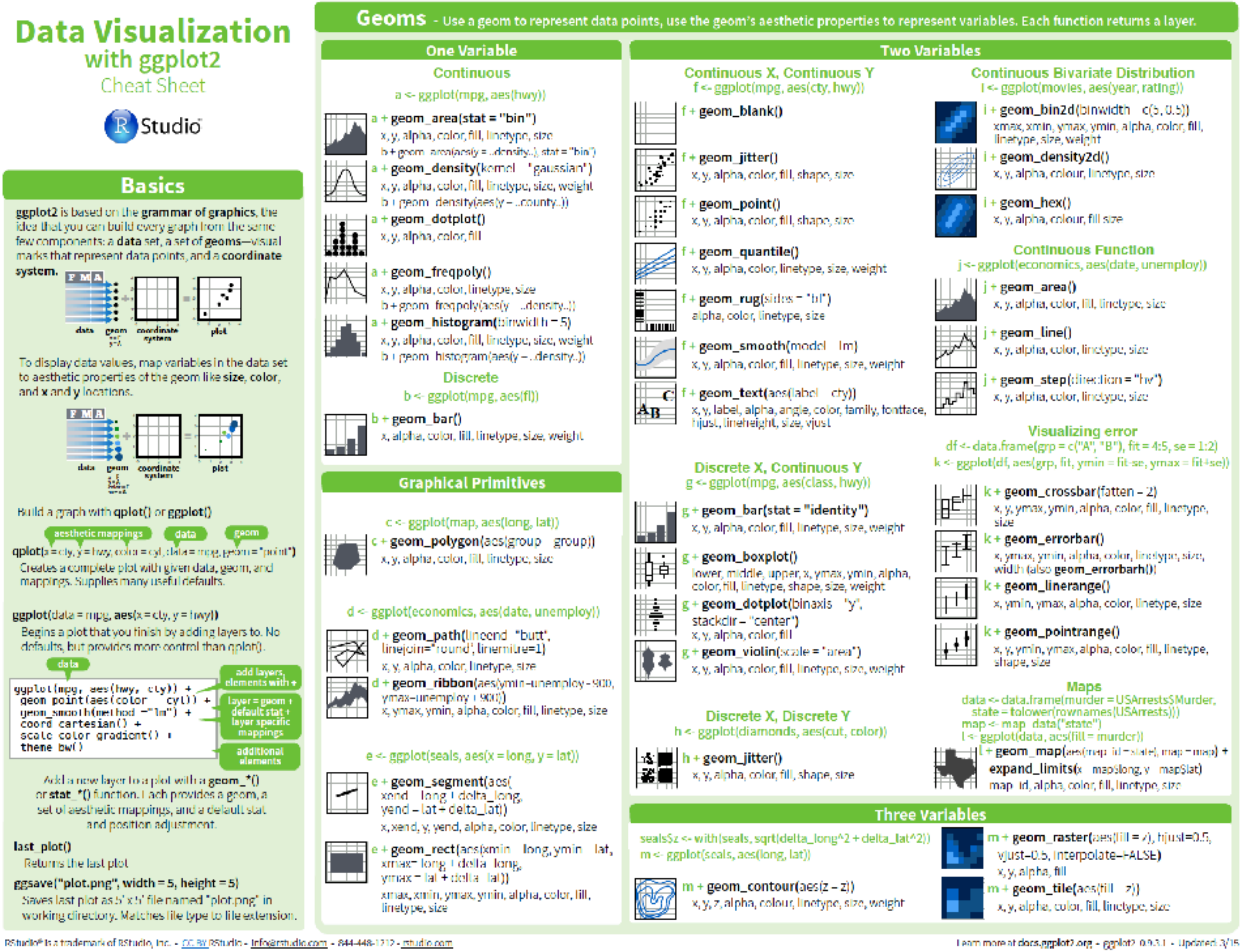
http://biecek.pl/Eseje/indexGramatyka.html

[hands on live R]

```
# Best 250 series http://www.imdb.com/chart/toptv/
## (1) read the new data with archivist
library(archivist)
series2017 <- aread("mi2-warsaw/RLadies/arepo/45aa16dc4dbf0d87e3e40eb9dc9d18ae")

## (2) or read the old data with Pogromcy Danych
library(PogromcyDanych)
serialeIMDB

## (3) or scrap the data from IMDB database
library(rvest)
library(dplyr)
# read links and titles
page <- read_html("http://www.imdb.com/chart/toptv/")
series <- html_nodes(page, ".titleColumn a")
titles <- html_text(series)
links <- html_attr(series, "href")
codes <- sapply(strsplit(links, split = "/"), `[`, 3)

allSeries <- lapply(seq_along(codes), function(i) {
  tab <- read_html(paste0("http://www.imdb.com/title/",codes[i],"/epdate?ref_=ttep_ql_4")) %>%
    html_node("table") %>%
    html_table()
  data.frame(Serie = titles[i], tab[,1:4], Season = gsub(tab[,1], pattern="\\..*", replacement=""))
```

```
selected <- c("Gra o tron", "Breaking Bad",
              "Sherlock", "Westworld")

dat <- series2017 %>%
  filter(Serie %in% selected)

ggplot(dat, aes(id, UserRating)) +
  geom_point(aes(color=Season, size=UserVotes)) +
  geom_smooth(se=FALSE, color="grey30") +
  facet_grid(Serie~.) +
  theme_light() + theme(legend.position="none") +
  scale_color_brewer(palette = 1, type = "qual") +
  ggtitle("User ratings for selected TV series") +
  xlab("Episode No")
```





Data Visualization with ggplot2 Cheat Sheet

RStudio

Basics

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

F M A
data
geom
coordinate system
plot

To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.

Build a graph with **qplot()** or **ggplot()**

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

ggplot(data = mpg, **aes**(x = cty, y = hwy))
Begins a plot that you finish by adding layers to. No defaults, but provides more control than qplot().

data
add layers, elements with +
layer = geom + default stat + layer specific mappings
additional elements

```
ggplot(mpg, aes(hwy, cty)) +
 geom_point(aes(color = cyl)) +
 geom_smooth(method ="lm") +
 coord_cartesian() +
 scale_color_gradient() +
 theme_bw()
```

Add a new layer to a plot with a **geom_*()** or **stat_*()** function. Each provides a geom, a set of aesthetic mappings, and a default stat and position adjustment.

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

One Variable

Continuous

a <- ggplot(mpg, aes(hwy))

a + **geom_area(stat = "bin")**
x, y, alpha, color, fill, linetype, size
b + geom_area(aes(y = ..density..), stat = "bin")

a + **geom_density**(kernel = "gaussian")
x, y, alpha, color, fill, linetype, size, weight
b + geom_density(aes(y = ..county..))

a + **geom_dotplot()**
x, y, alpha, color, fill

a + **geom_freqpoly()**
x, y, alpha, color, linetype, size
b + geom_freqpoly(aes(y = ..density..))

a + **geom_histogram(**binwidth = 5)
x, y, alpha, color, fill, linetype, size, weight
b + geom_histogram(aes(y = ..density..))

Discrete

b <- ggplot(mpg, aes(fl))

b + **geom_bar()**
x, alpha, color, fill, linetype, size, weight

Graphical Primitives

c <- ggplot(map, aes(long, lat))

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

d <- ggplot(economics, aes(date, unemploy))

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

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

e <- ggplot(seals, aes(x = long, y = lat))

e + **geom_segment**(aes(xend = long + delta_long, yend = lat + delta_lat))
x, xend, y, yend, alpha, color, linetype, size

e + **geom_rect**(aes(xmin = long, ymin = lat, xmax= long + delta_long, ymax = lat + delta_lat))
xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

Two Variables

Continuous X, Continuous Y

f <- ggplot(mpg, aes(cty, hwy))

f + **geom_blank()**

f + **geom_jitter()**
x, y, alpha, color, fill, shape, size

f + **geom_point()**
x, y, alpha, color, fill, shape, size

f + **geom_quantile()**
x, y, alpha, color, linetype, size, weight

f + **geom_rug**(sides = "bl")
alpha, color, linetype, size

f + **geom_smooth**(model = lm)
x, y, alpha, color, fill, linetype, size, weight

f + **geom_text**(aes(label = cty))
x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

Discrete X, Continuous Y

g <- ggplot(mpg, aes(class, hwy))

g + **geom_bar(stat = "identity")**
x, y, alpha, color, fill, linetype, size, weight

g + **geom_boxplot()**
lower, middle, upper, x, ymax, ymin, alpha, color, fill, linetype, shape, size, weight

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

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

Discrete X, Discrete Y

h <- ggplot(diamonds, aes(cut, color))

h + **geom_jitter()**
x, y, alpha, color, fill, shape, size

Continuous Bivariate Distribution

i <- ggplot(movies, aes(year, rating))

i + **geom_bin2d**(binwidth = c(5, 0.5))
xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size, weight

i + **geom_density2d()**
x, y, alpha, colour, linetype, size

i + **geom_hex()**
x, y, alpha, colour, fill size

Continuous Function

j <- ggplot(economics, aes(date, unemploy))

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

j + **geom_line()**
x, y, alpha, color, linetype, size

j + **geom_step**(direction = "hv")
x, y, alpha, color, linetype, size

Visualizing error

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

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

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

k + **geom_linerange()**
x, ymin, ymax, alpha, color, linetype, size

k + **geom_pointrange()**
x, y, ymin, ymax, alpha, color, fill, linetype, shape, size

Maps

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

l + **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

seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2))
m <- ggplot(seals, aes(long, lat))

m + **geom_contour**(aes(z = z))
x, y, z, alpha, colour, linetype, size, weight

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

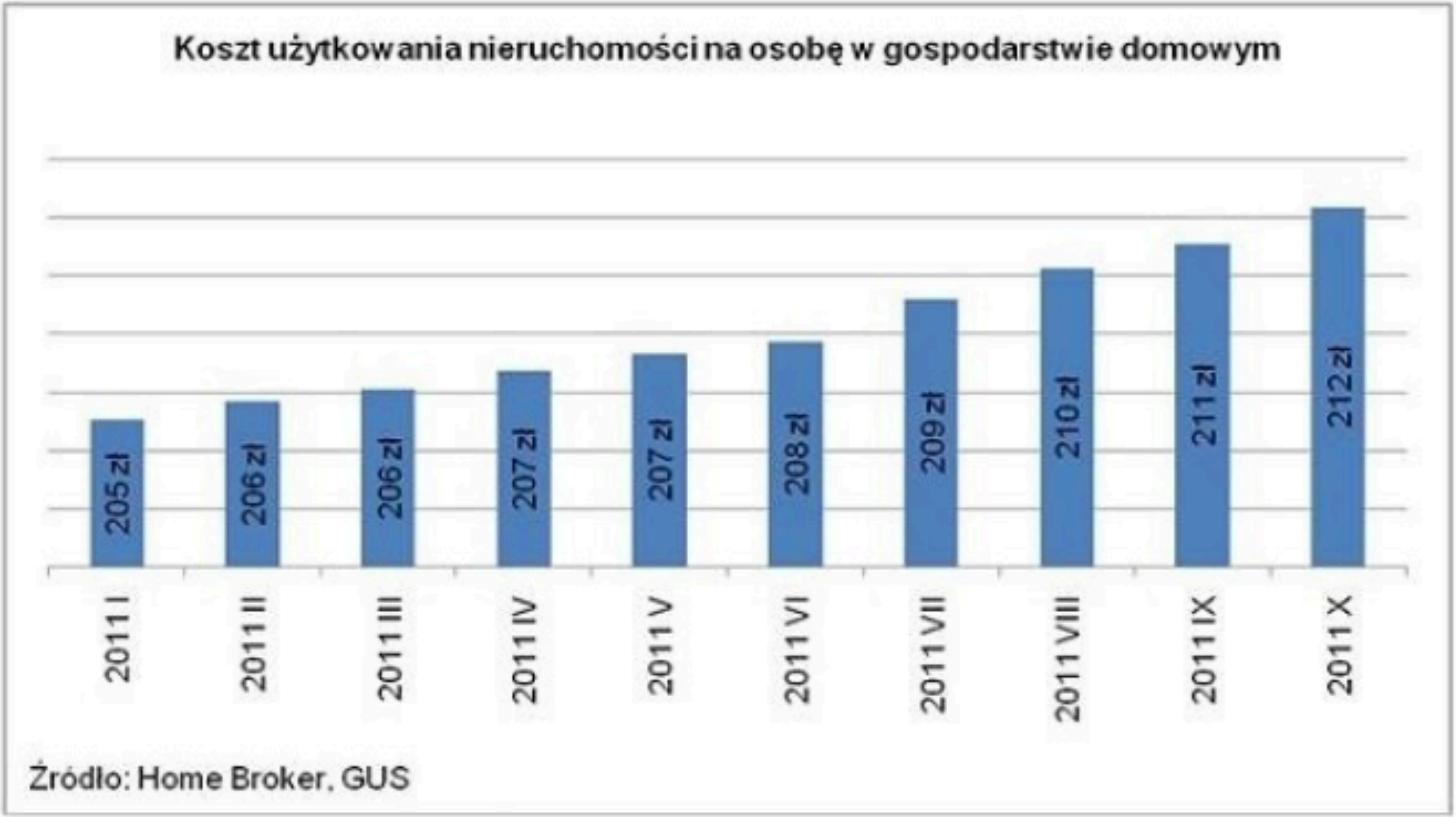
m + **geom_tile**(aes(fill = z))
x, y, alpha, color, fill, linetype, size

RStudio® is a trademark of RStudio, Inc. • CC BY RStudio • info@rstudio.com • 844-448-1212 • rstudio.com

Learn more at docs.ggplot2.org • ggplot2 0.9.3.1 • Updated: 3/15

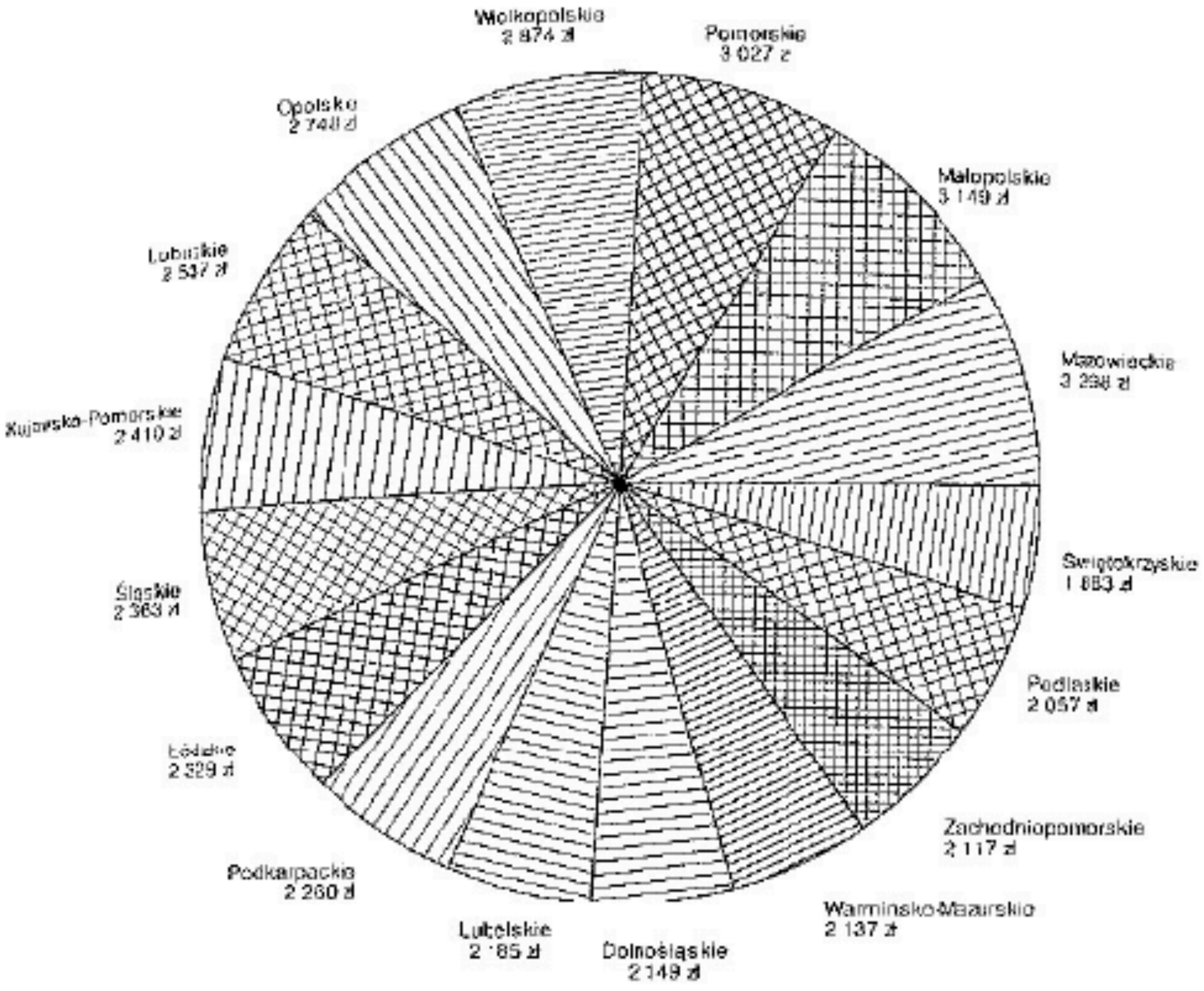
https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

Do not lie

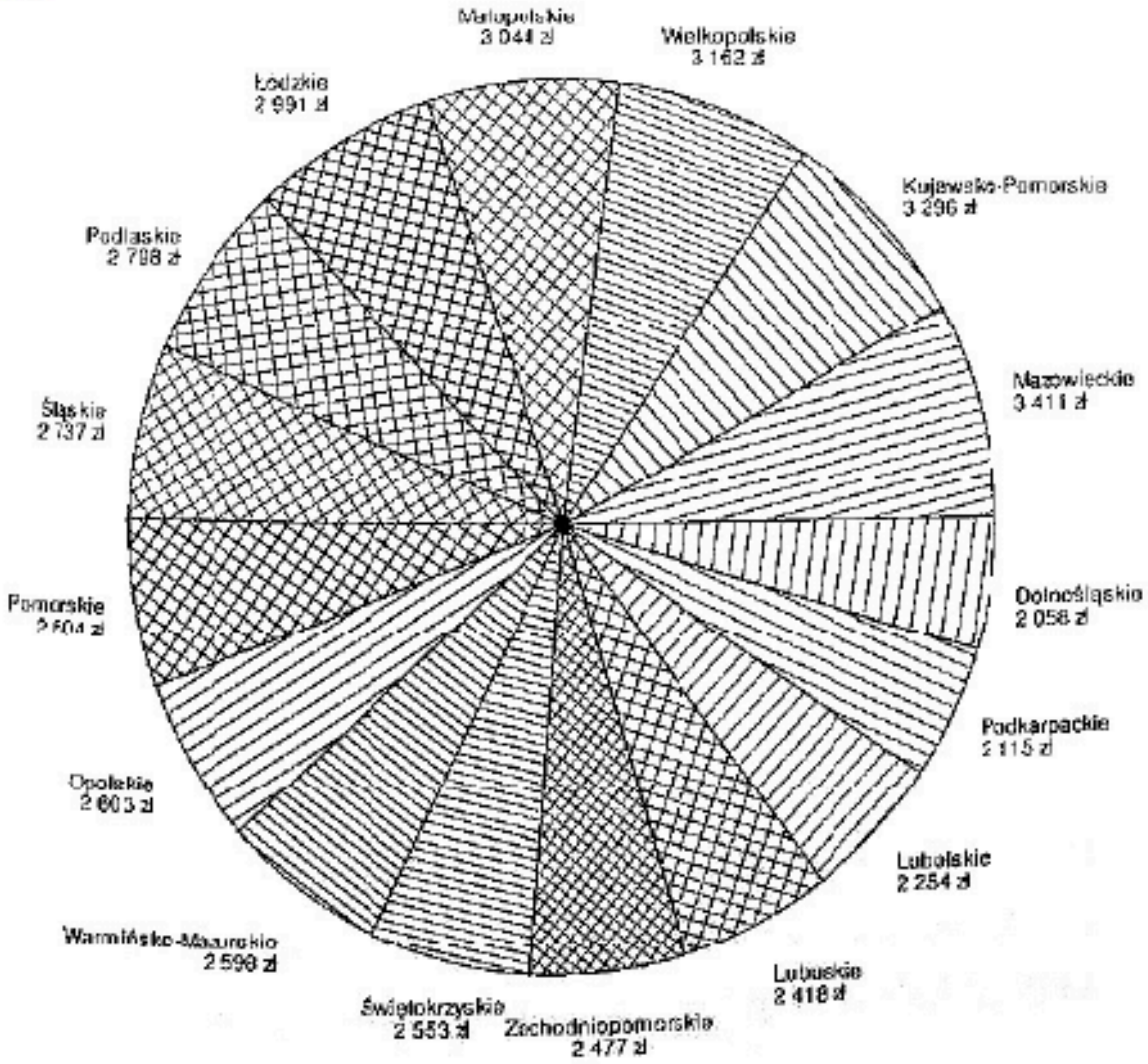


Źródło: Home Broker, GUS

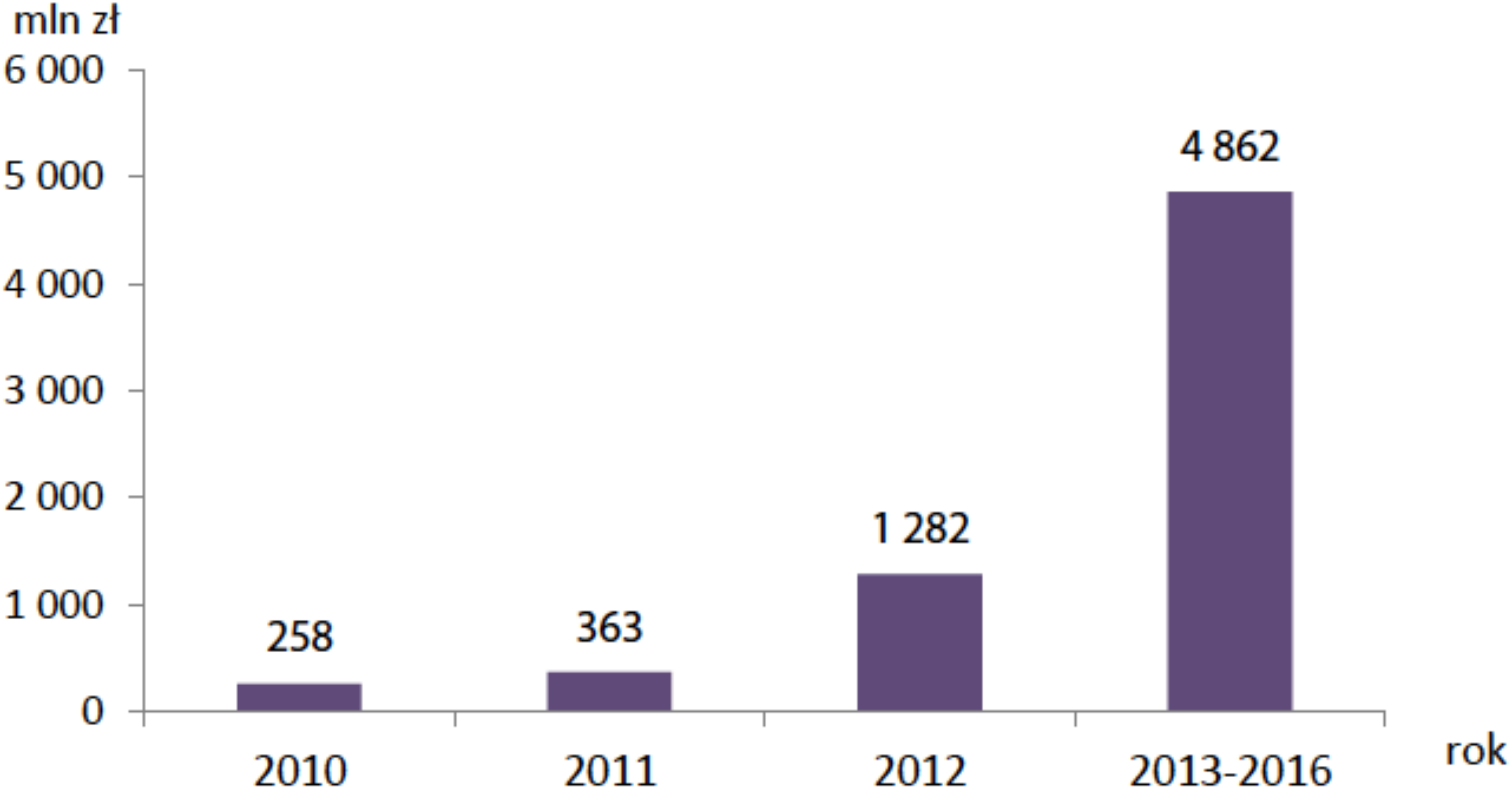
Kobiety



Mężczyźni



I.3 Wydatki deklarowane przez przedsiębiorców na B+R w programach NCBiR w latach 2010-2016.

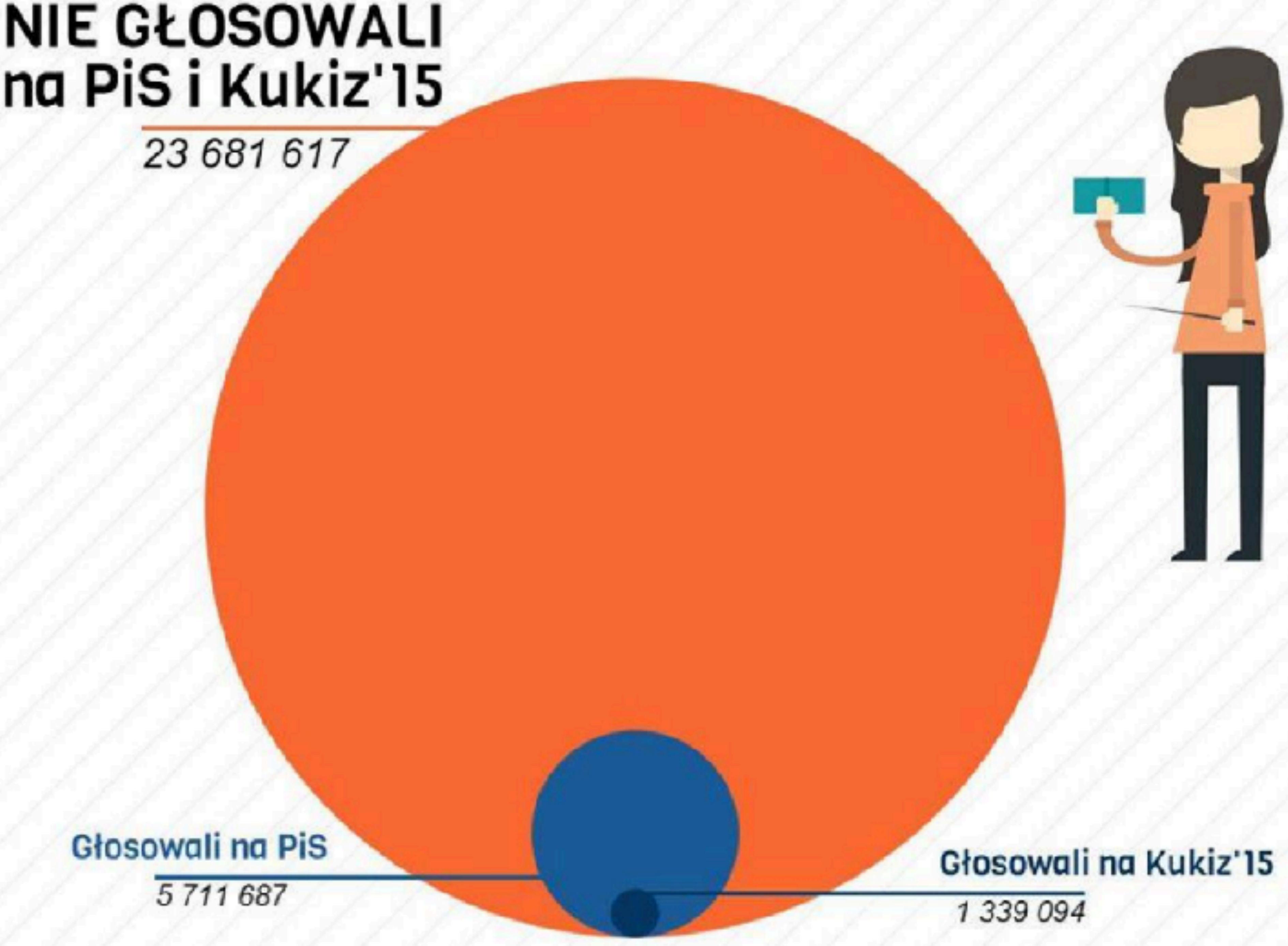


źródło: Narodowe Centrum Badań i Rozwoju

30 732 398 Polaków

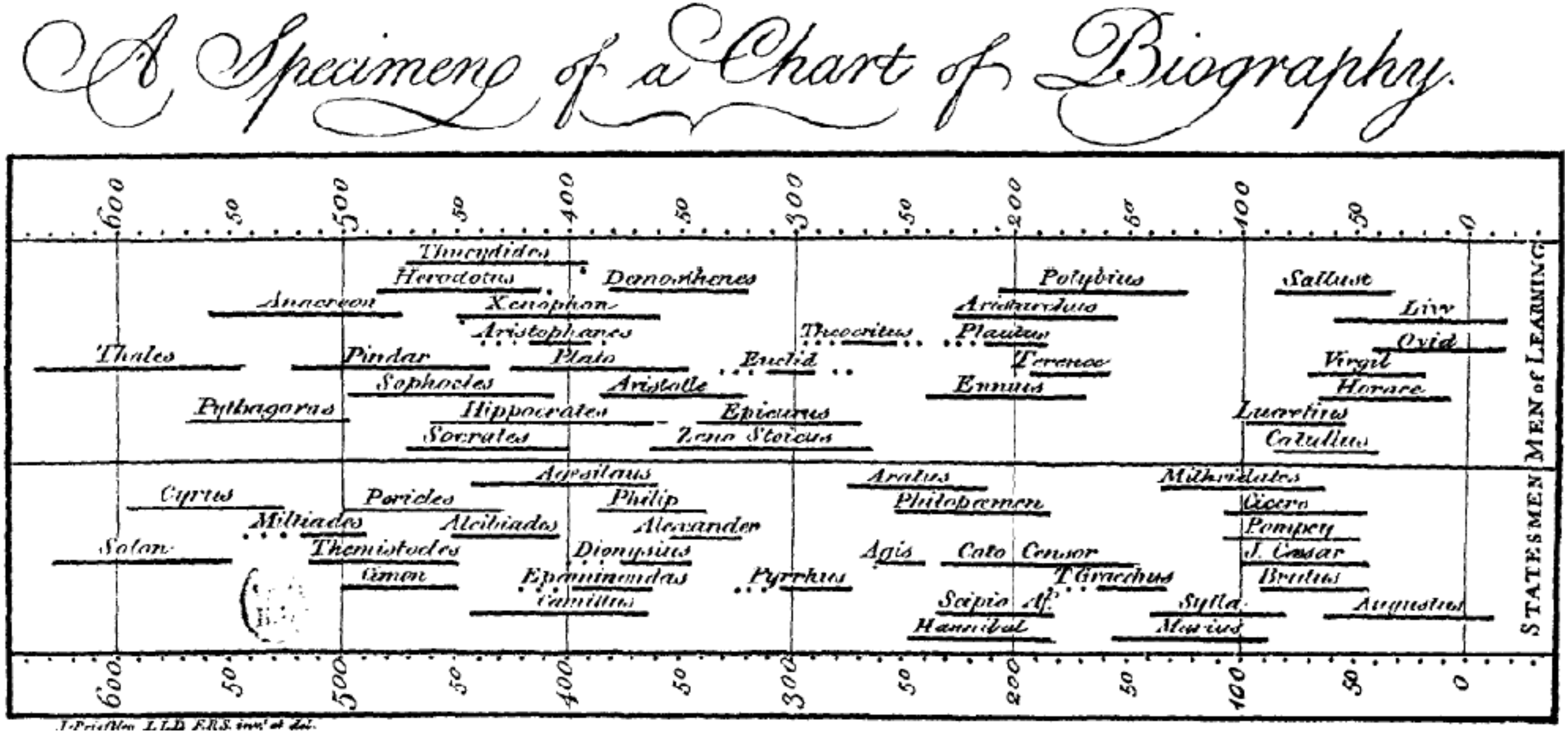
UPRAWNIONYCH DO GŁOSOWANIA

* Dane: Państwowa Komisja Wyborcza

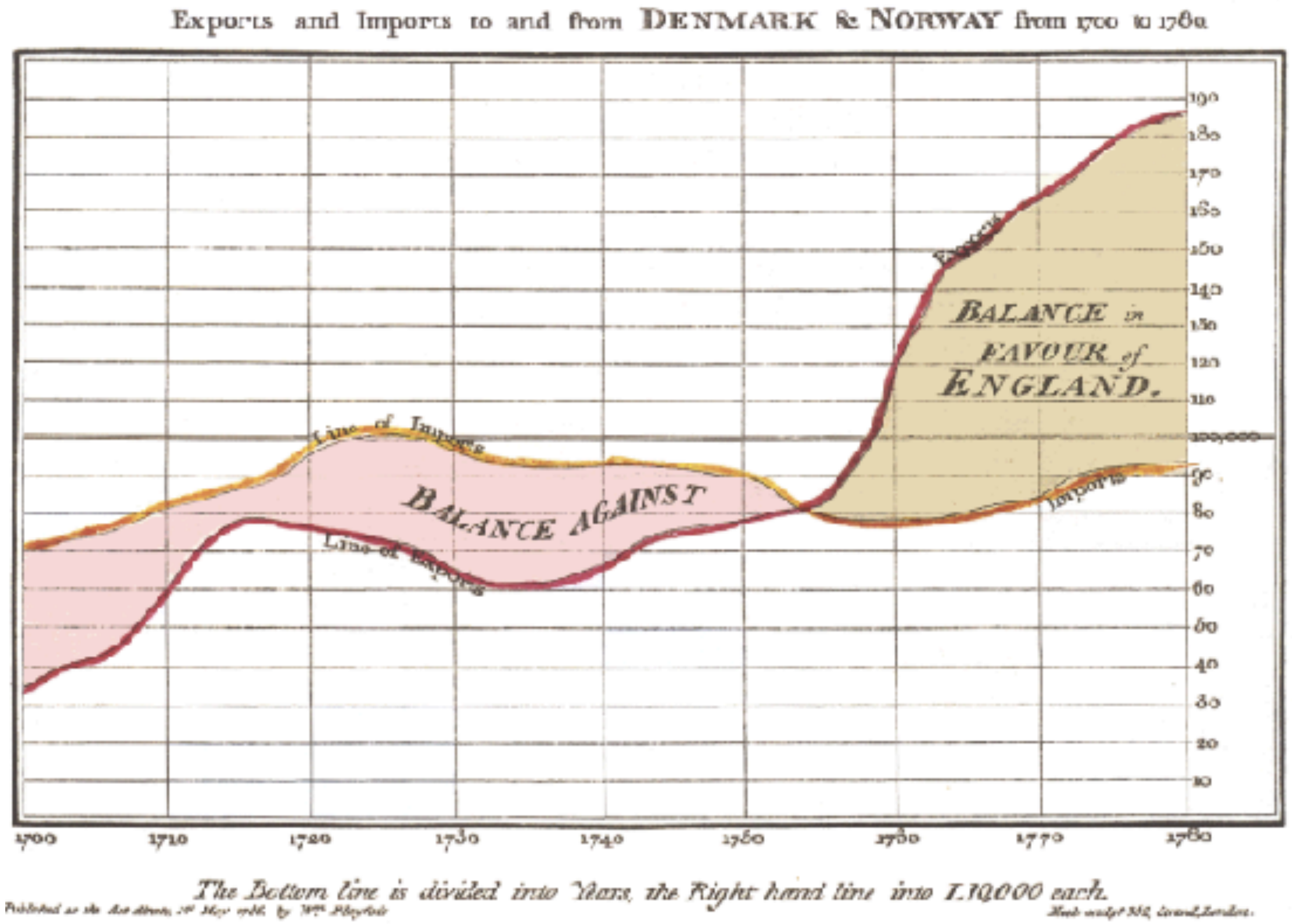


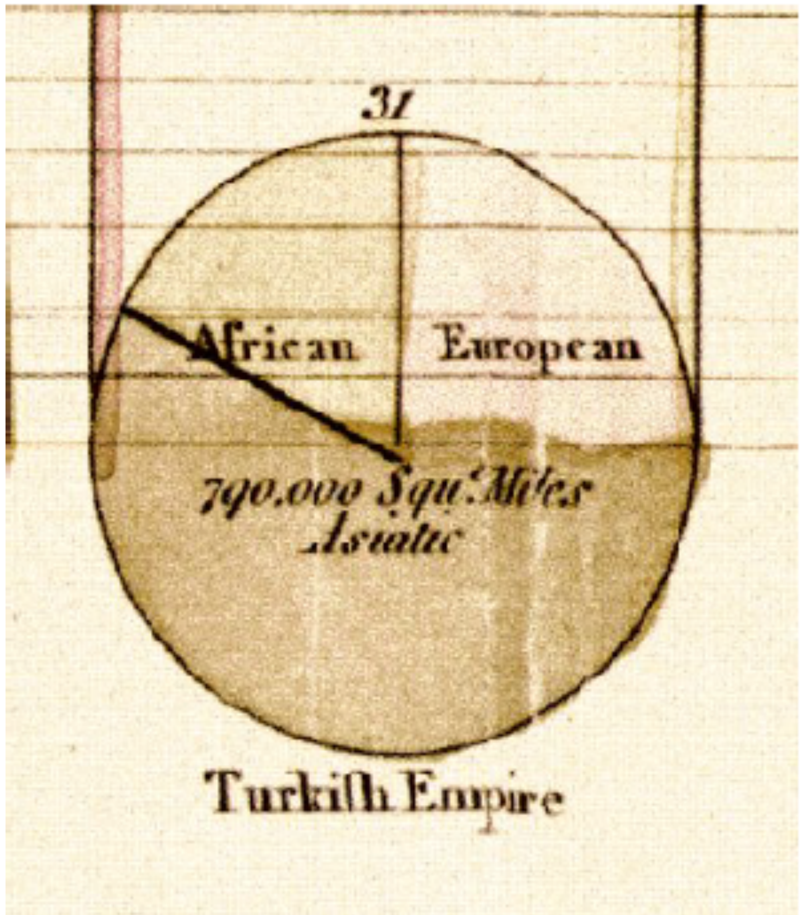
History

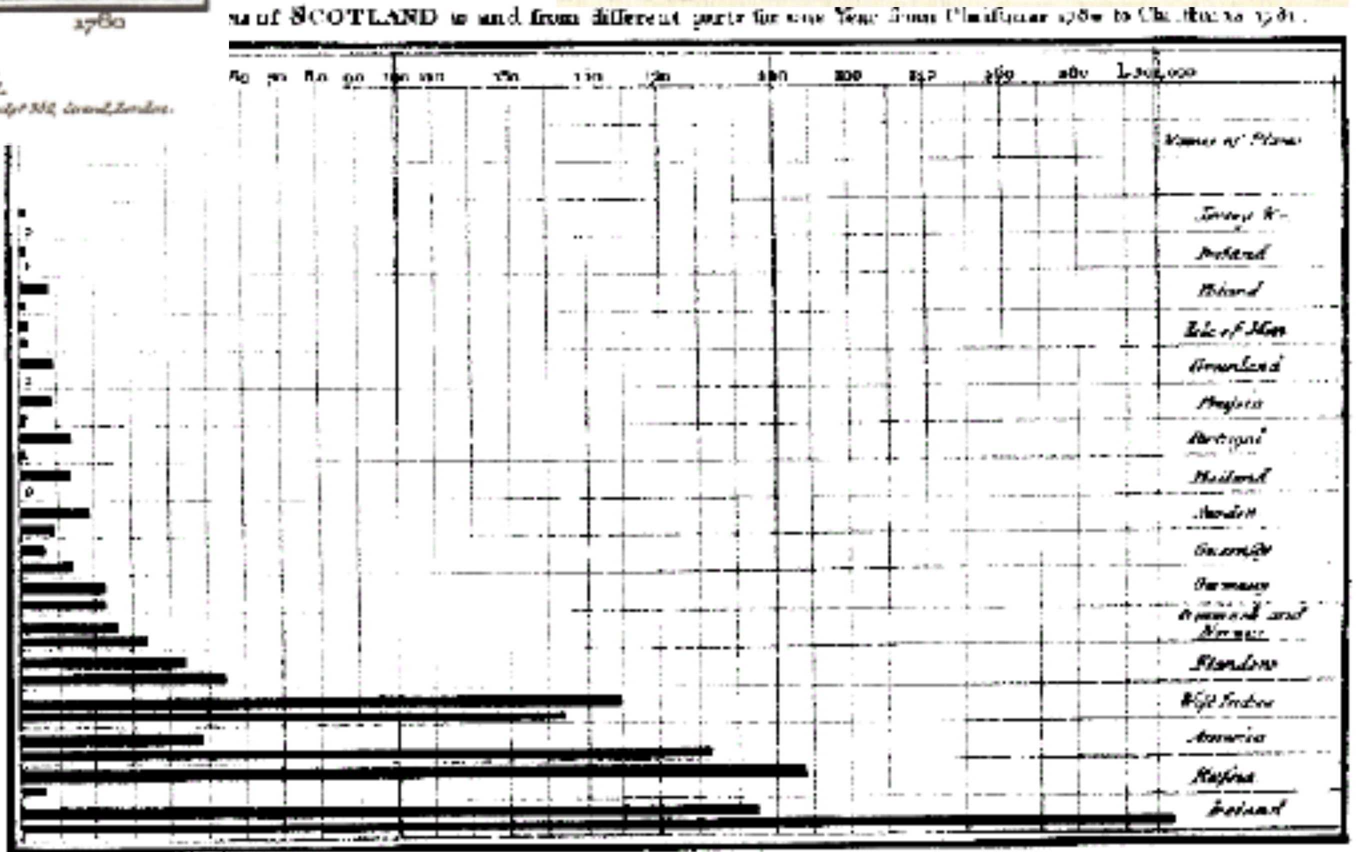
Chart of Biography Joseph Priestley (1765)



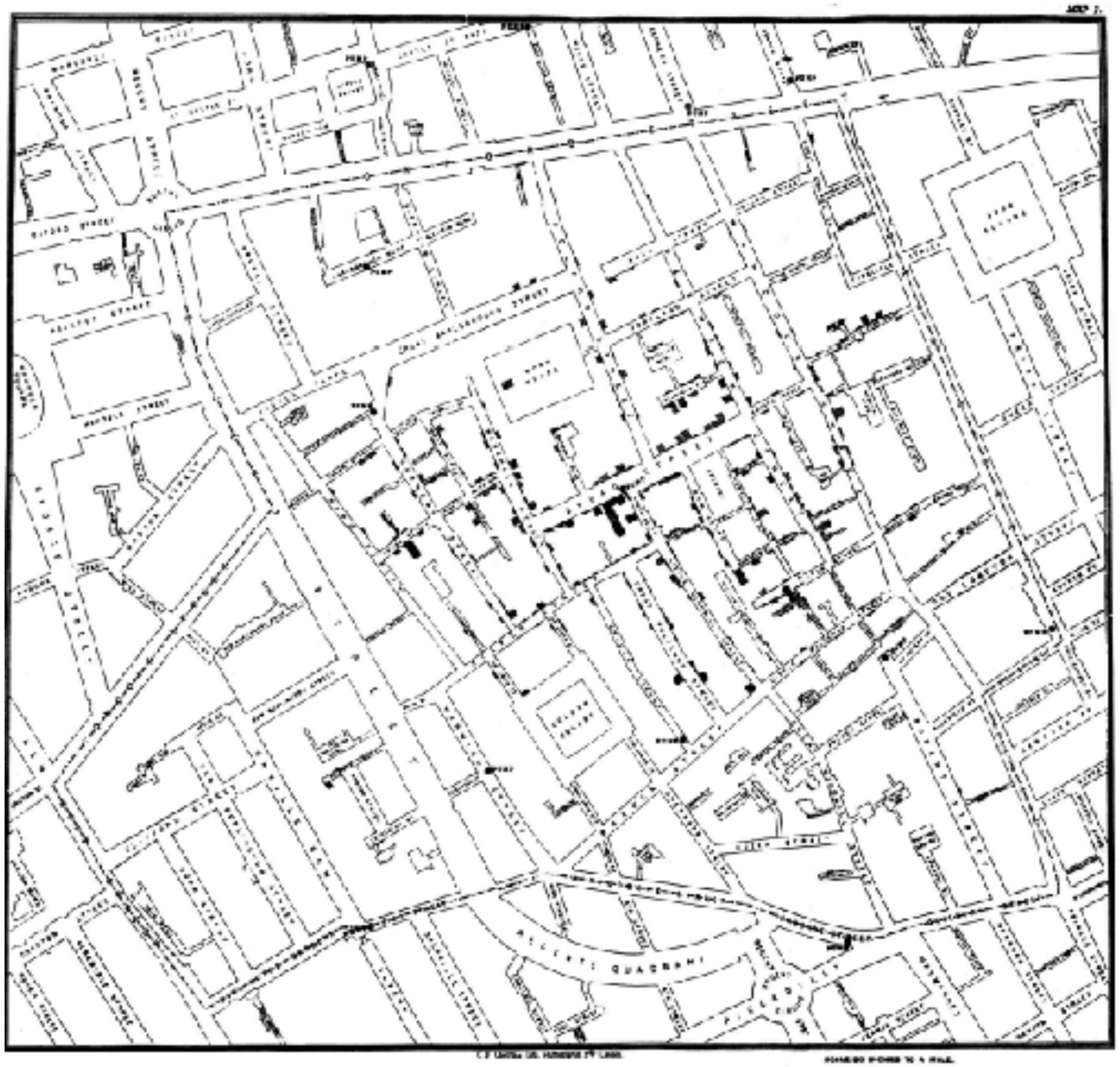
Commercial and Political Atlas William Playfair (1786)







On the Mode of Communication of Cholera. John Snow. 1855



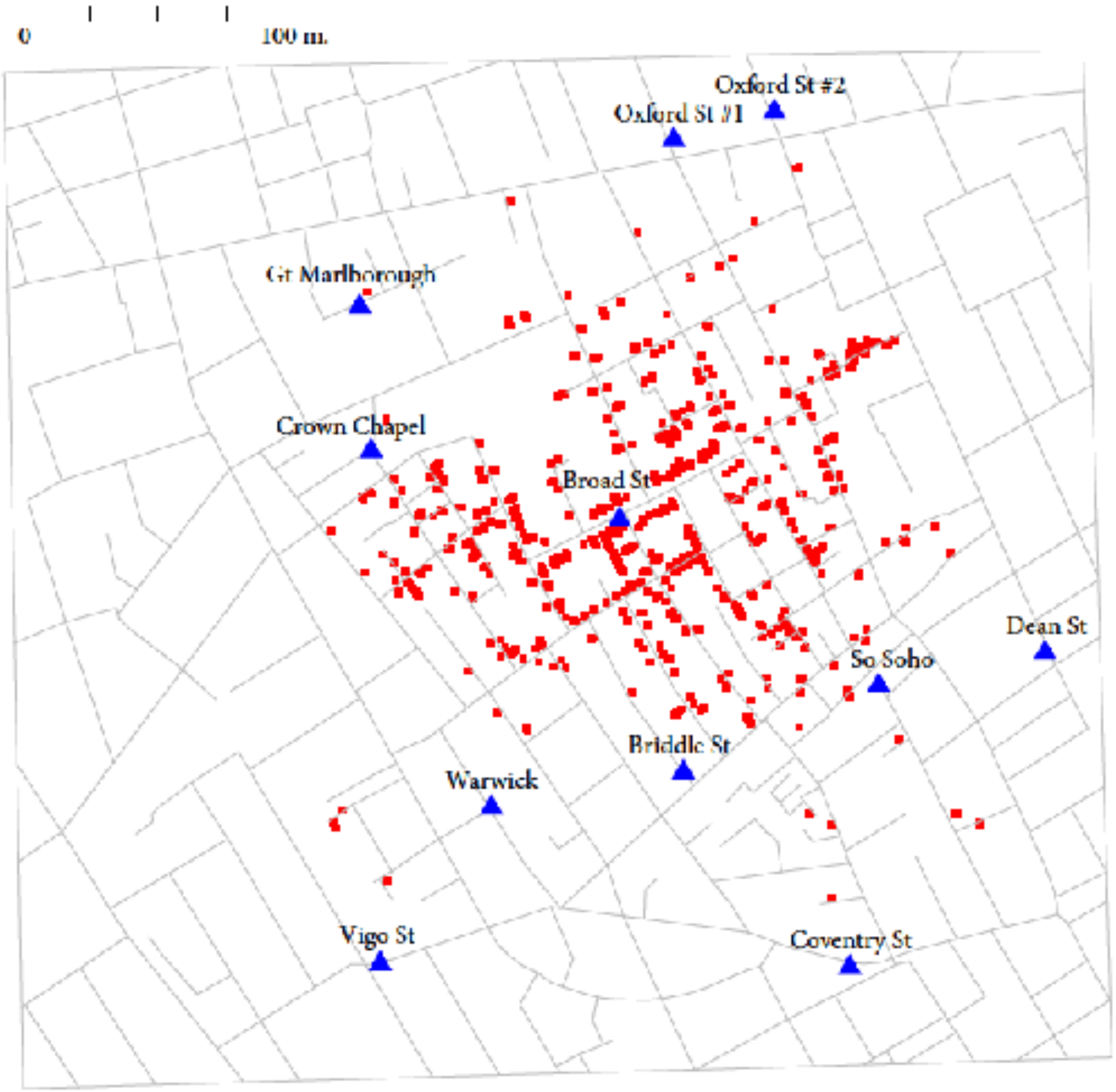


Diagram of the causes of mortality in the army in the East. Florence Nightingale. 1858

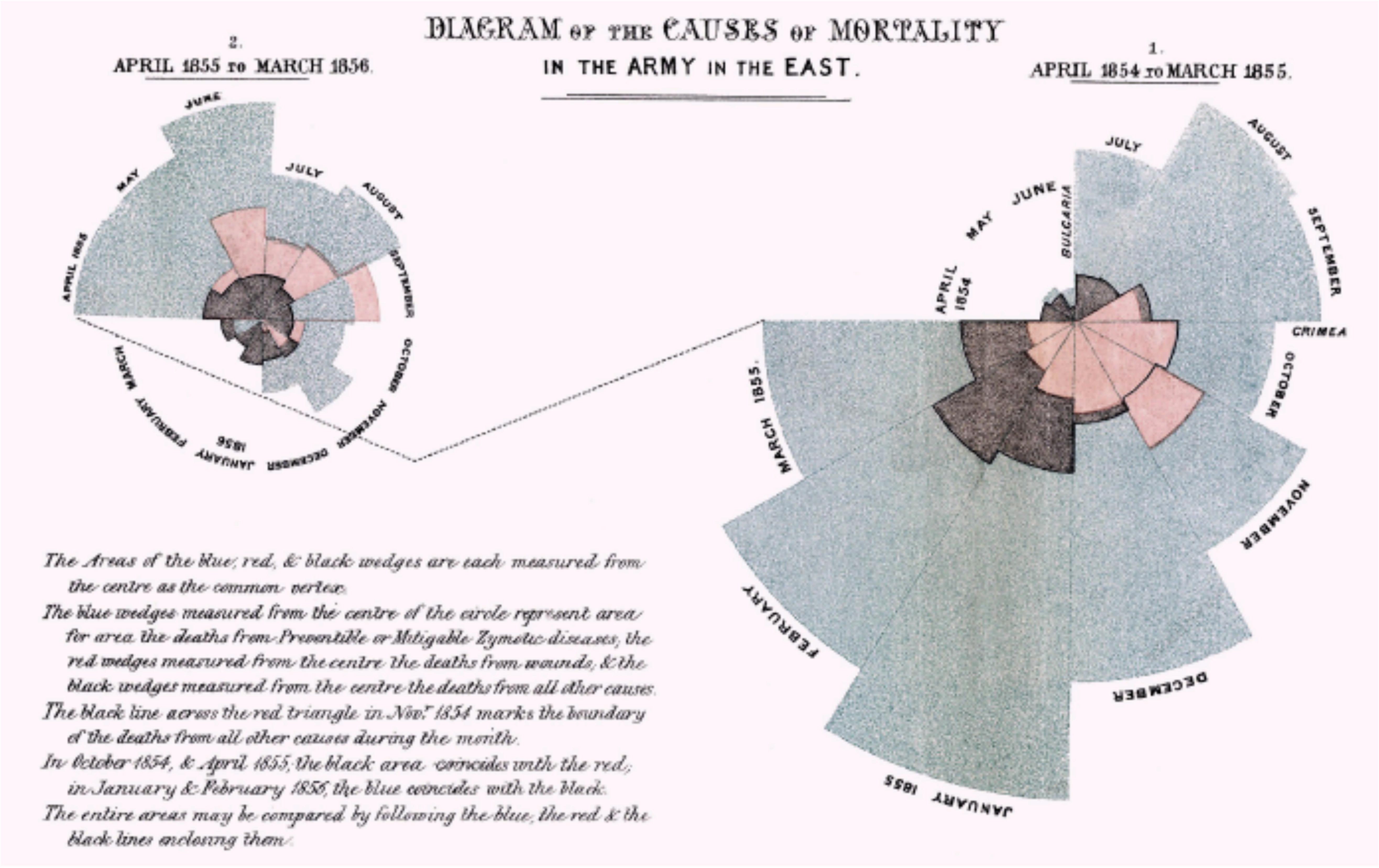


DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.

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.

More resources

Przemysław Biecek

Odkrywać! Ujawniać! Objaśniać!

Zbiór esejów o sztuce prezentowania danych

Discover! Reveal! Describe!

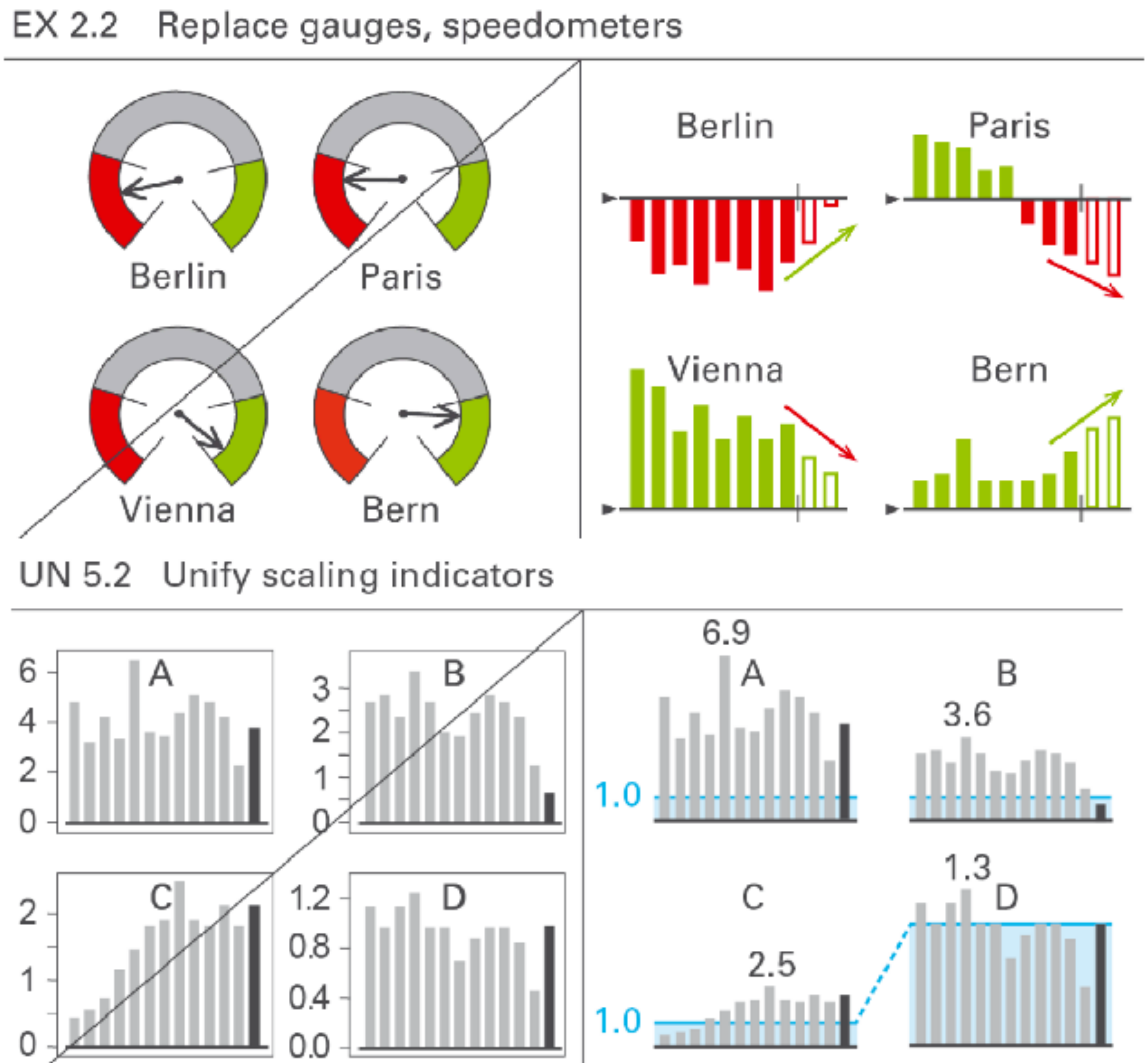
Essays about the art of data presentation

http://biecek.pl/Eseje

(english version will be there soon)

International Business Communication Standards

http://www.ibcs-a.org/



EX 2.2 Replace gauges, speedometers

UN 5.2 Unify scaling indicators

Docs ggplot2 http://docs.ggplot2.org/current/
Cookbook for R http://www.cookbook-r.com/Graphs/
Docs ggvis http://ggvis.rstudio.com/
Great blog http://flowingdata.com/
Graphs in NYT http://kpq.github.io/chartsnthings/
Nature Methods, Points of View http://clearscience.info/wp/?p=546

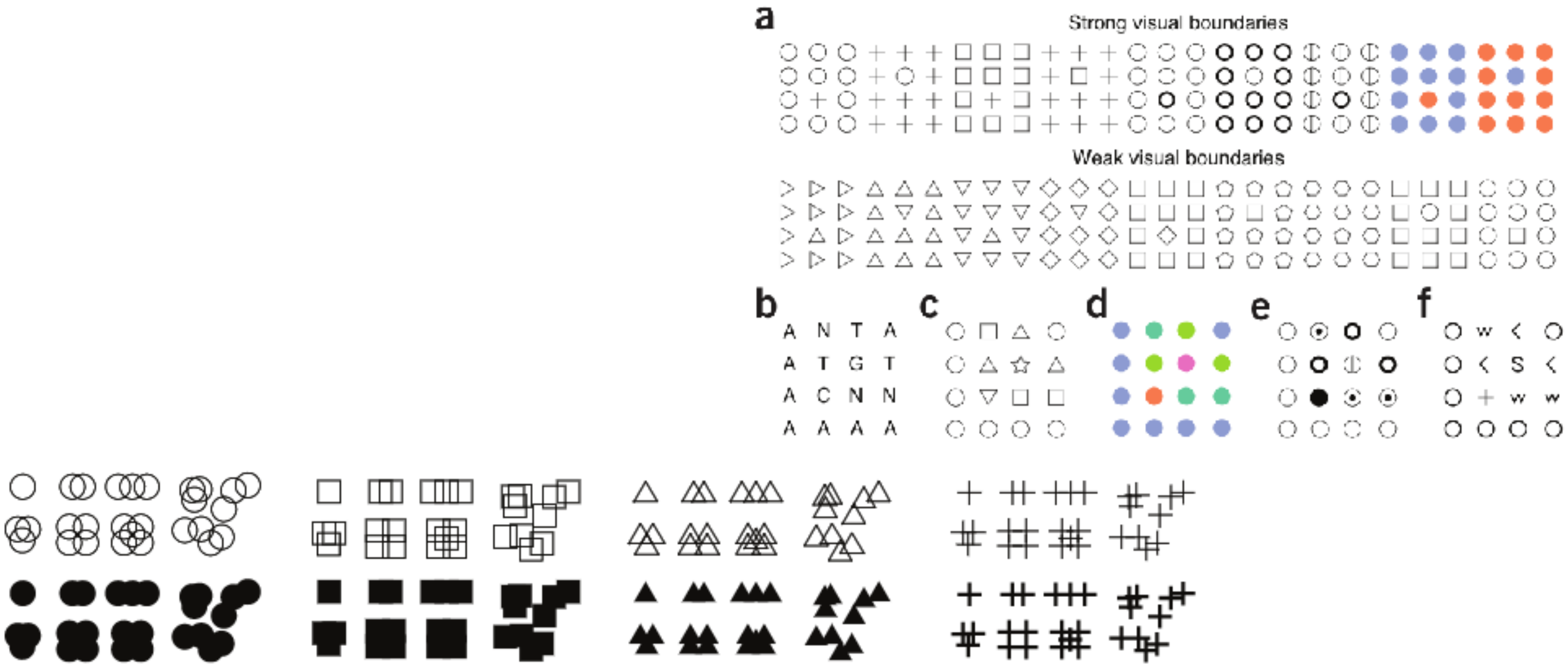


Figure 1 | The hollow circle is a flexible and robust plotting symbol.

[bio project]