Data Wrangling and Data Analysis
Data Visualization

Daniel L. Oberski & Erik-Jan van Kesteren

Department of Methodology & Statistics

Utrecht University

VERSION: 2023-09-25

Coordinators

Erik-Jan

Ayoub

Lecturer

Daniel

Tutors

Laura

Javier

Mahdi

Anastasia

Jelle

Thom

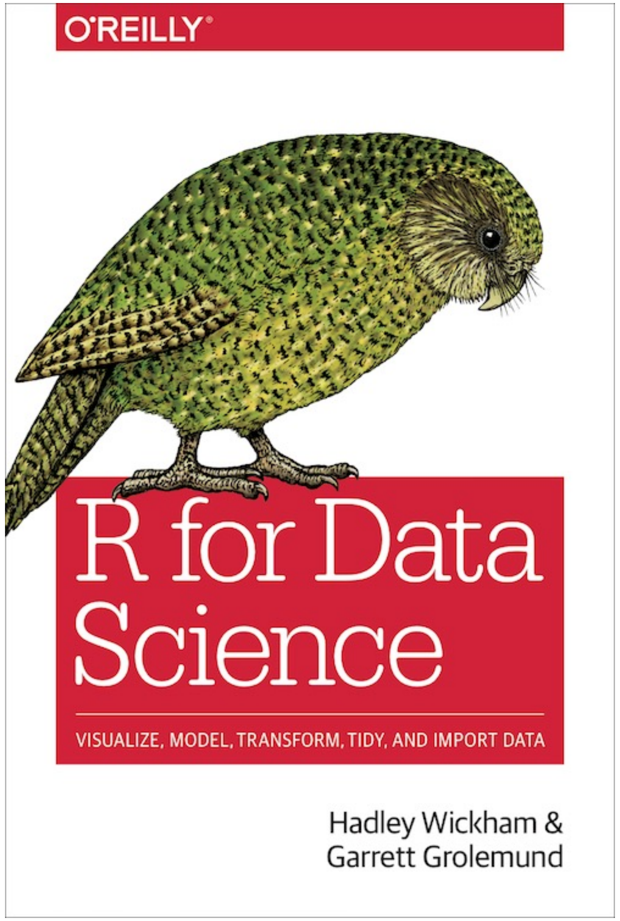
This week

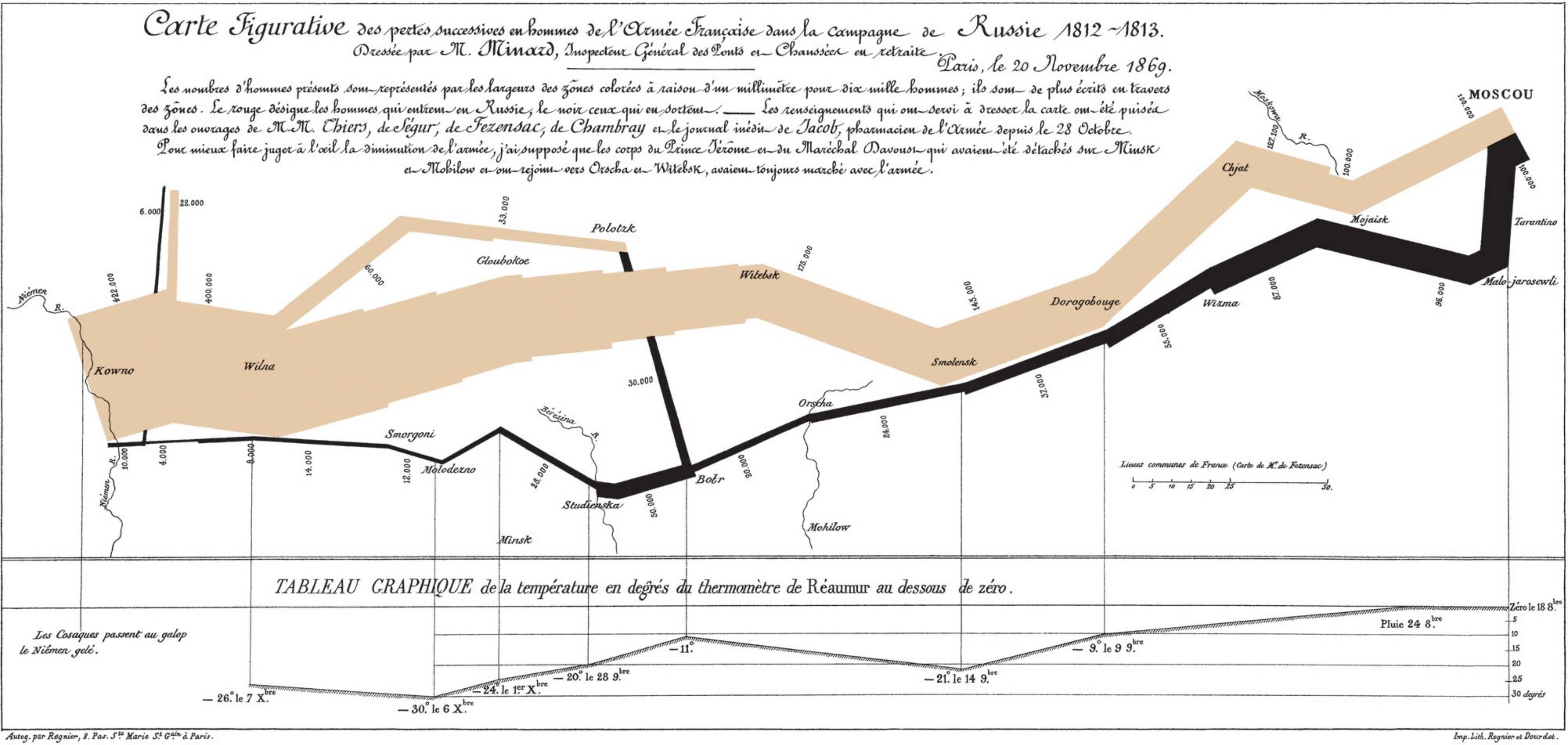
1. (Data preparation 2/2)
2. (Cloud computing guest lecture)
3. **Data visualization principles & Grammar of graphics**

Reading materials for this week

- Chapters from **R for Data Science (R4DS), open access book at:**
- https://r4ds.had.co.nz

- Today: ch 3 visualization
- Tomorrow: chh 3, 5, 7





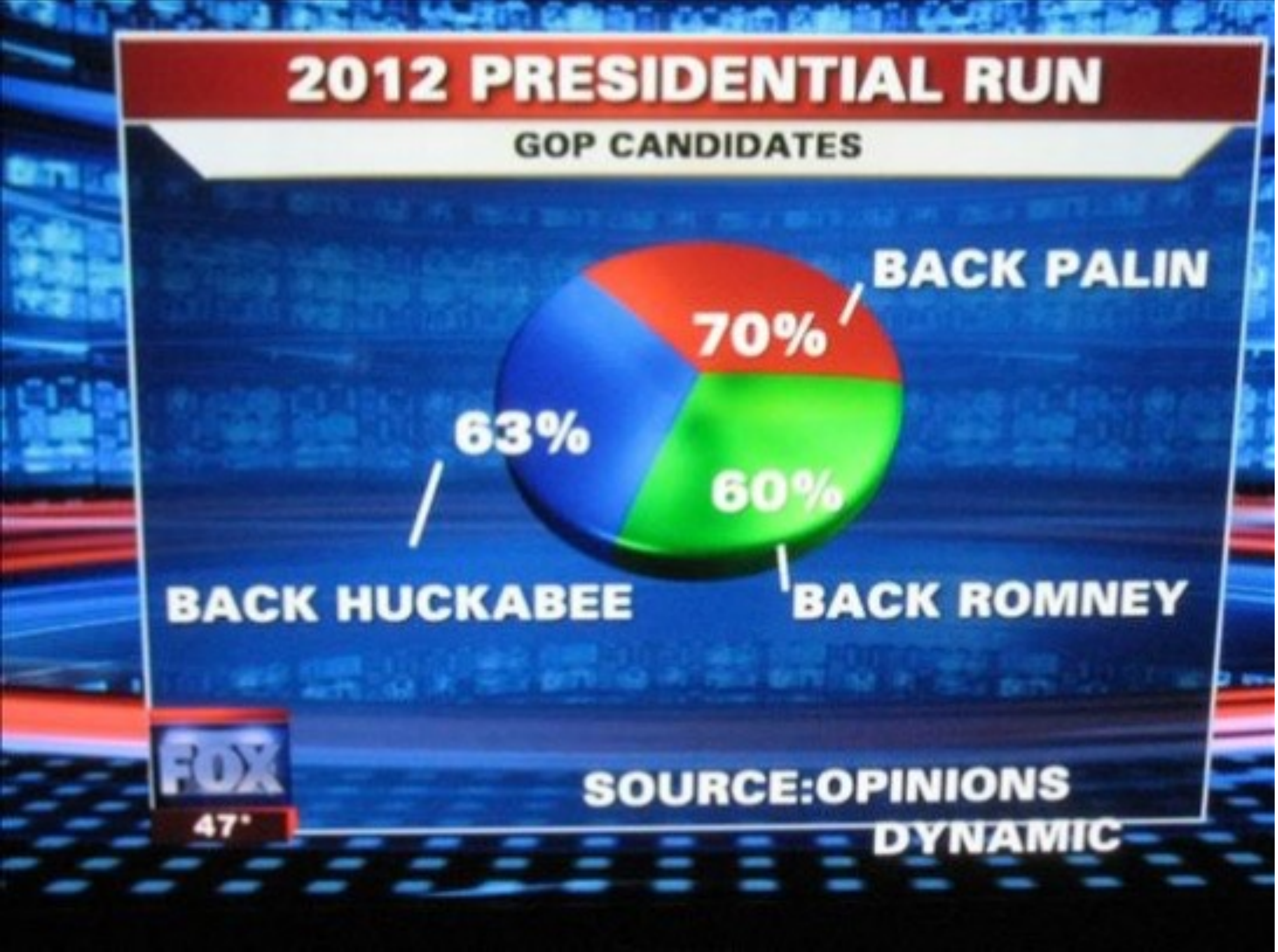
Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.



2012 PRESIDENTIAL RUN

GOP CANDIDATES

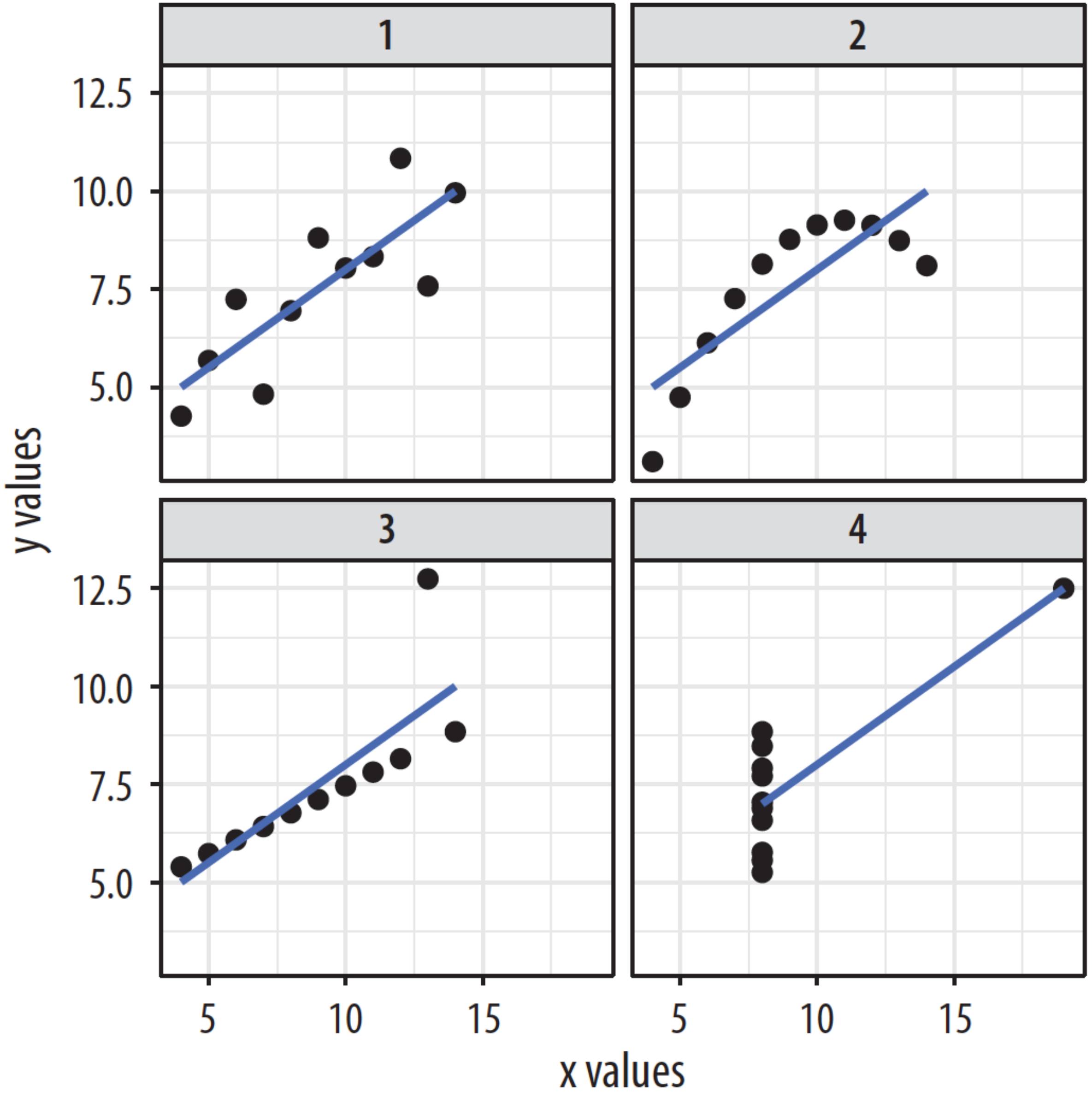
SOURCE:OPINIONS DYNAMIC

Today: visualization principles

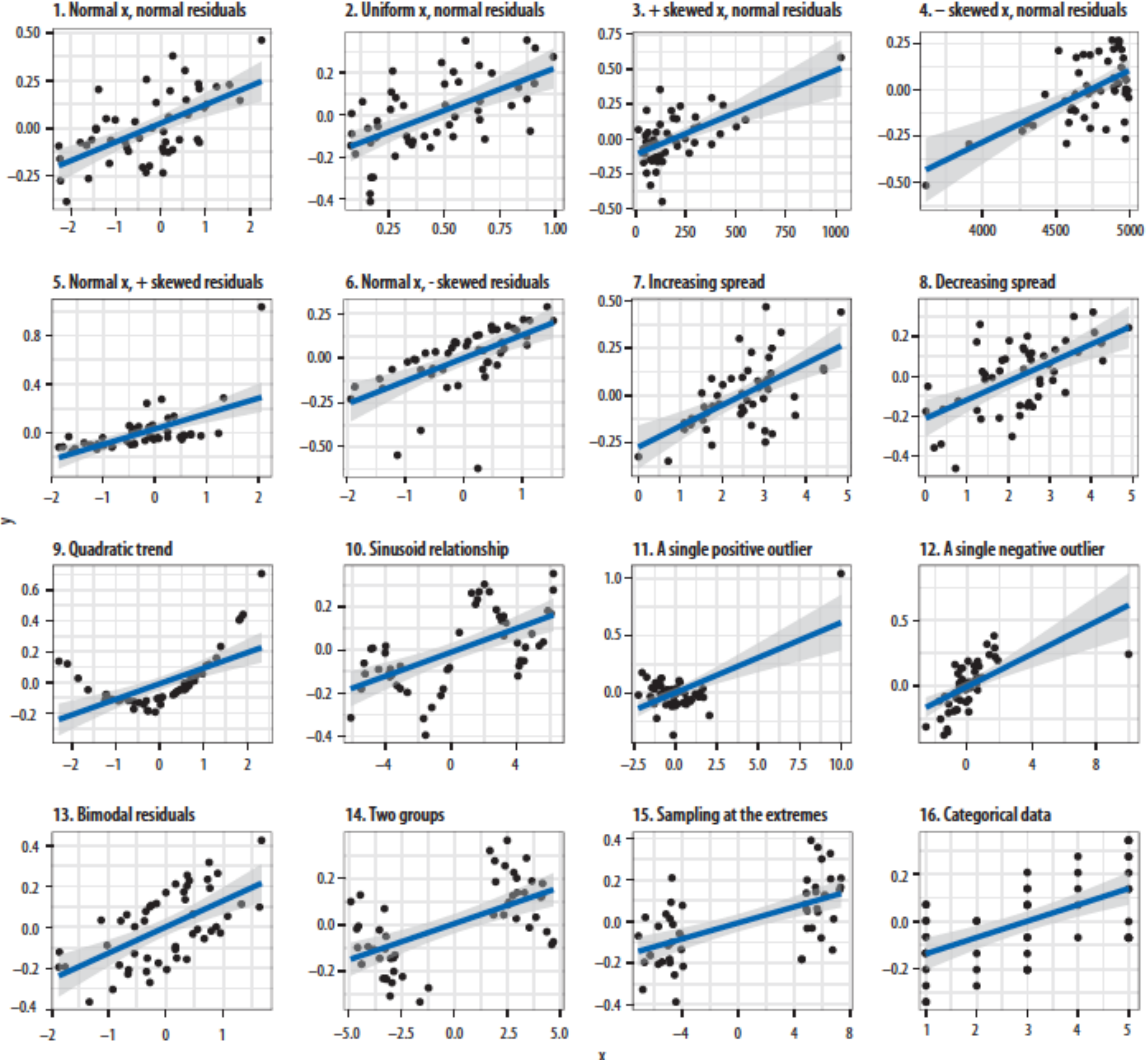
Data visualization

- **For exploration, data analysis ←**
- For communication
- For entertainment

Anscombe’s quartet



Source: Healy (2019)



Source: Healy (2019)

Graphics for data analysis

- The **human retina** can transfer around 10^6 or 10^7 bits per second to the brain;
- **Reading** transfers about 3 words, so $\sim 10^2$ or 10^3 bits/s;
- Potentially (!) visualization is about 4 orders of magnitude more powerful.

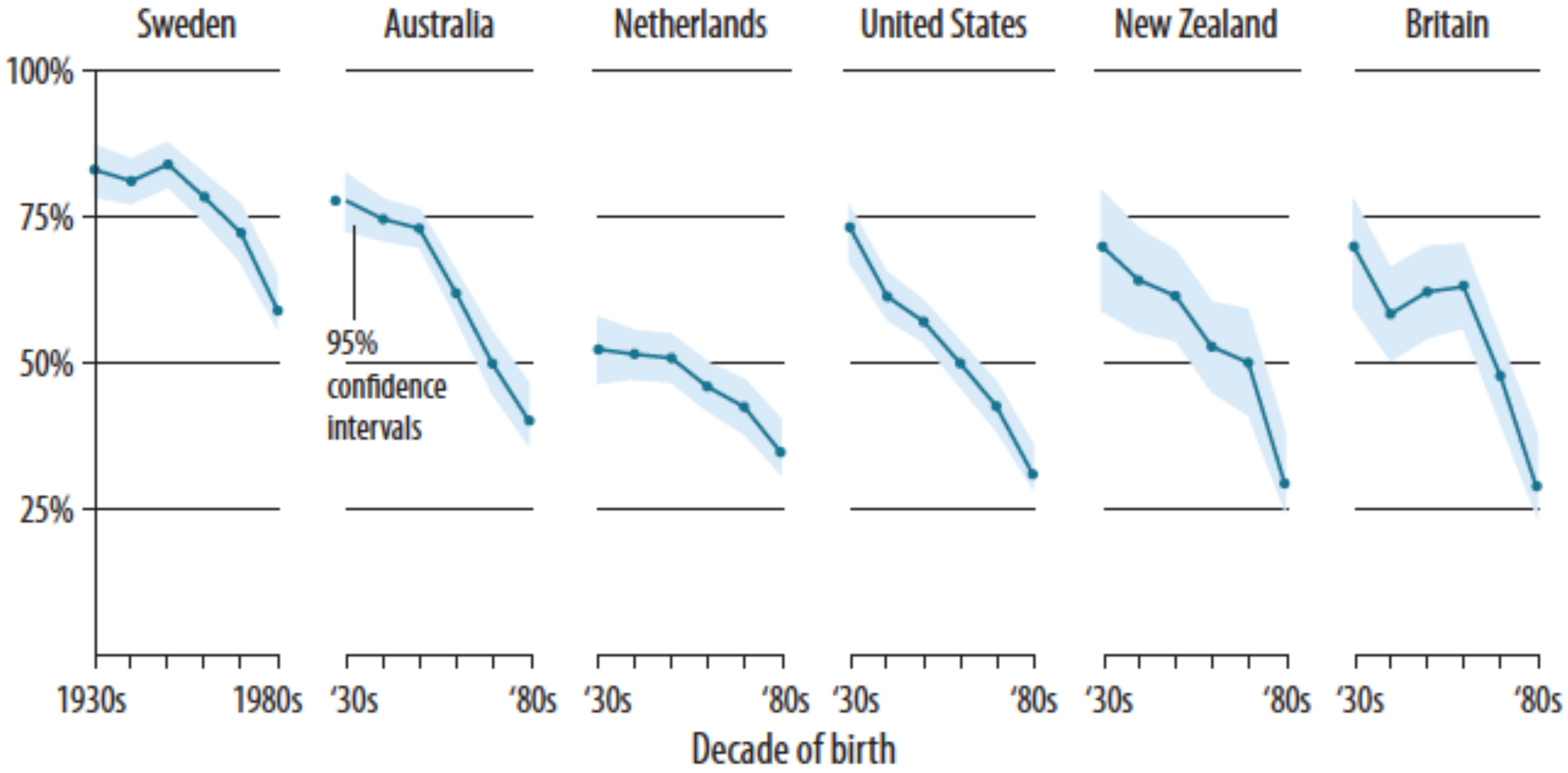
How can we leverage the human visual system to analyze data?

Plotting the right thing

Most common problems:

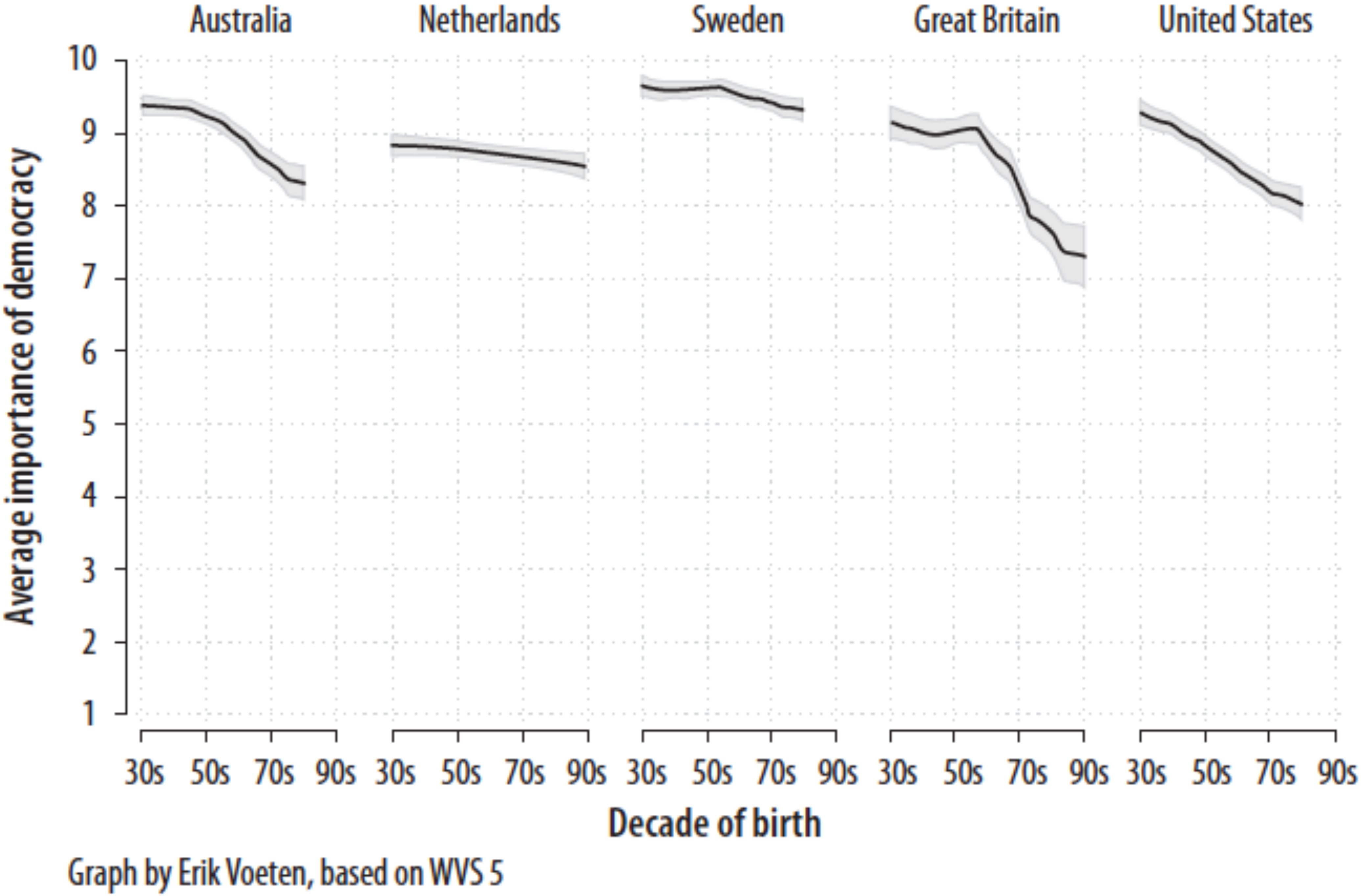
- (Accidentally) misrepresenting what is being plotted
- Omitting baselines

Percentage of people who say it is "essential" to live in a democracy



Utrecht University

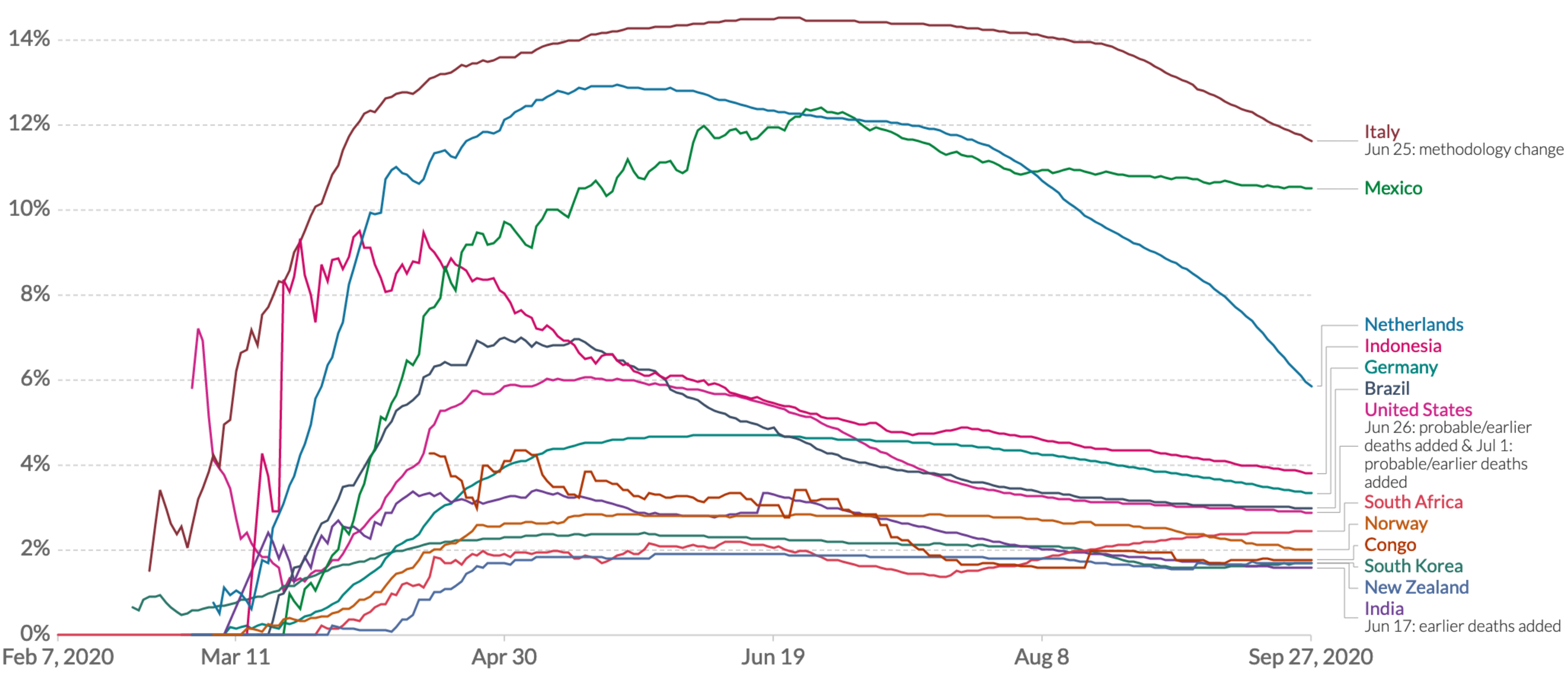
Source: Healy (2019)



Source: Healy (2019)

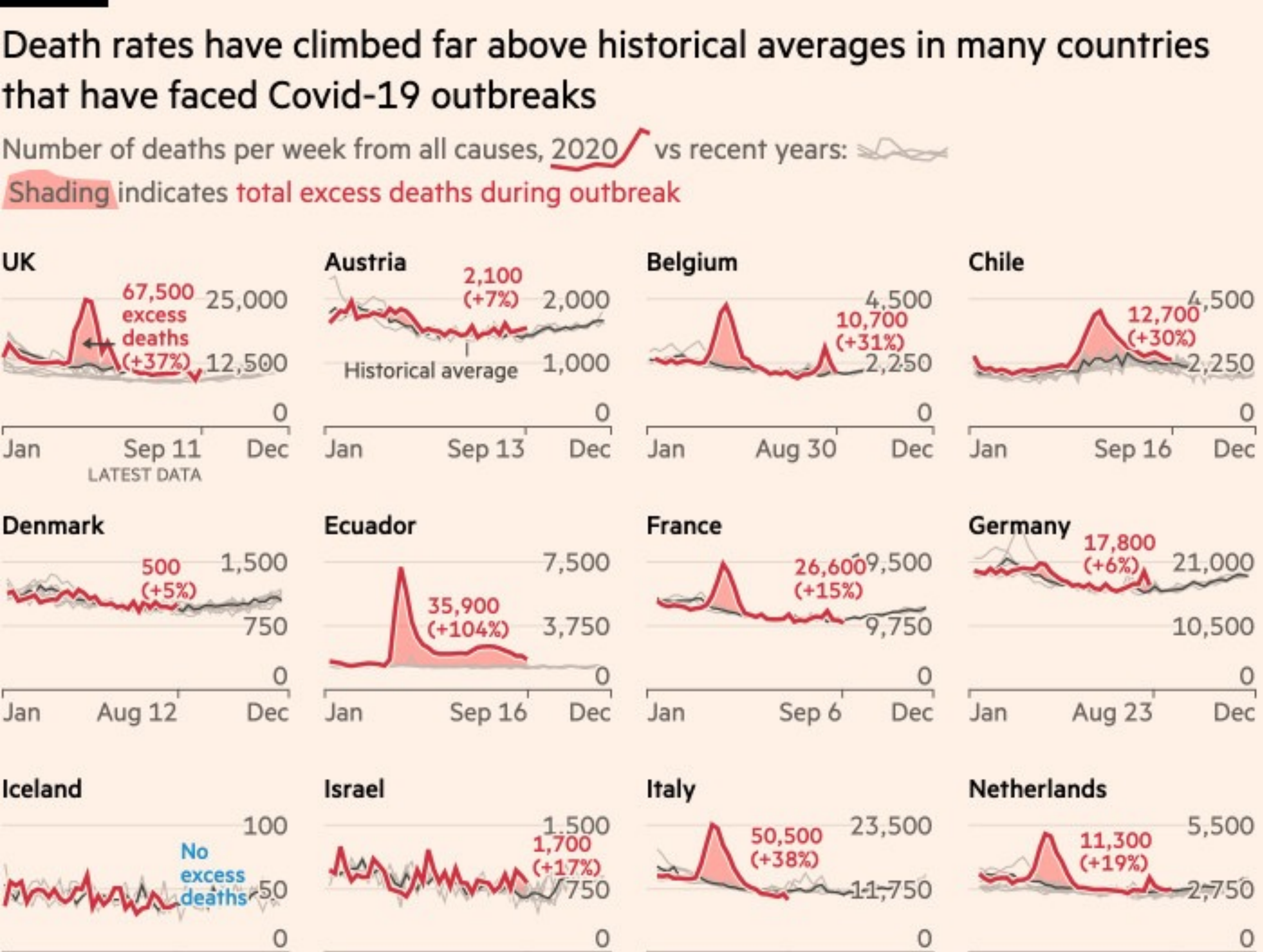
Case fatality rate of the ongoing COVID-19 pandemic

The Case Fatality Rate (CFR) is the ratio between confirmed deaths and confirmed cases. During an outbreak of a pandemic the CFR is a poor measure of the mortality risk of the disease. We explain this in detail at OurWorldInData.org/Coronavirus



Source: European CDC – Situation Update Worldwide – Last updated 27 September, 10:05 (London time)

Good example (FT)

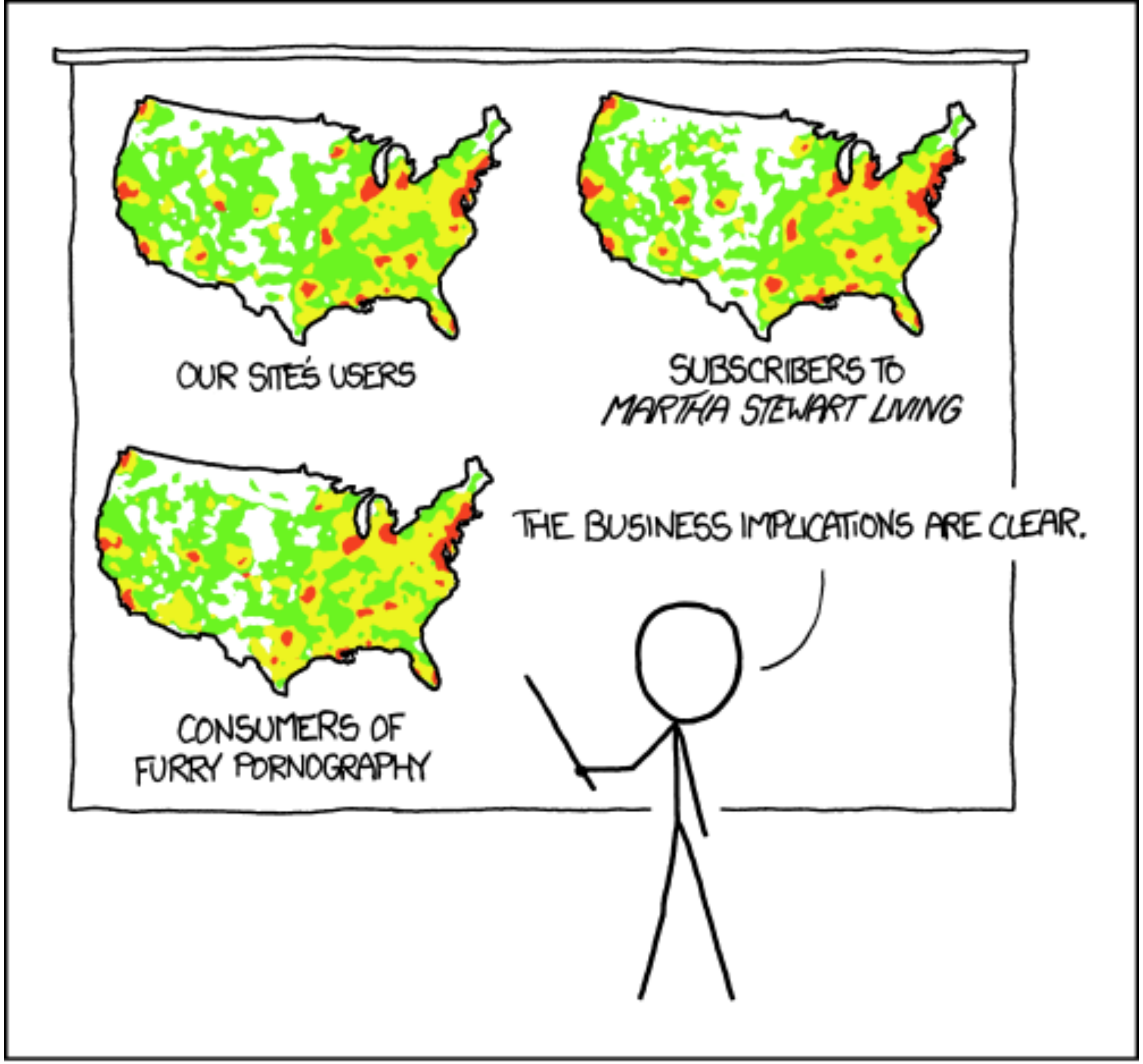


Death rates have climbed far above historical averages in many countries that have faced Covid-19 outbreaks

Number of deaths per week from all causes, 2020 vs recent years:

Shading indicates total excess deaths during outbreak

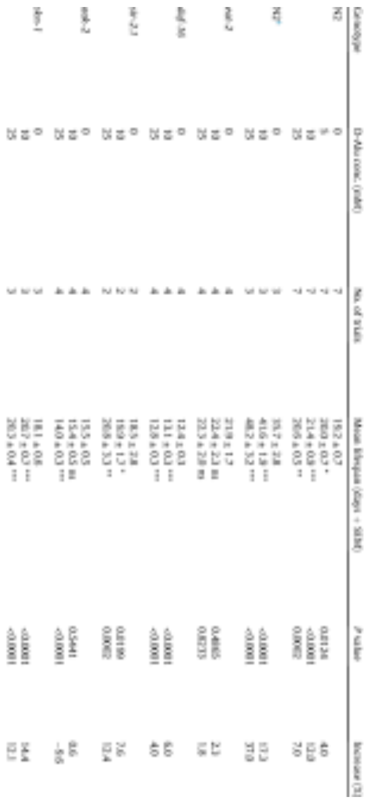
https://www.ft.com/content/a29
8-5eb7-4633-b89c-cbdf5b38693


OUR SITE'S USERS
SUBSCRIBERS TO *MARTHA STEWART LIVING*
CONSUMERS OF FURRY PORNOGRAPHY
THE BUSINESS IMPLICATIONS ARE CLEAR.

PET PEEVE #208:

GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

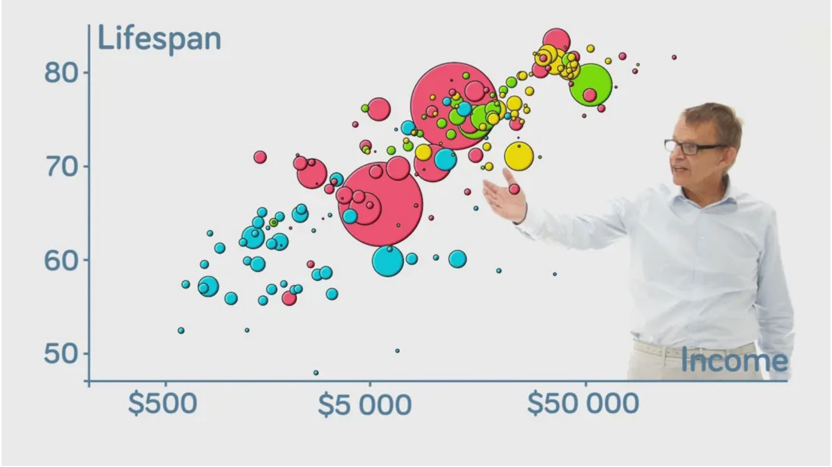
Data



Genotype	D-Alu conc. (mM)	No. of trials	Mean lifespan (days ± SEM)	P value	Increase (%)
N2	0	7	19.2 ± 0.7		
	5	7	20.0 ± 0.7 *	0.0126	4.0
	10	7	21.4 ± 0.8 ***	<0.0001	12.0
	25	7	20.6 ± 0.5 **	0.0002	7.0
N2*	0	3	35.7 ± 2.8		
	10	3	41.6 ± 1.8 ***	<0.0001	17.3
	25	3	48.2 ± 3.2 ***	<0.0001	37.0
eat-2	0	4	21.9 ± 1.7		
	10	4	22.4 ± 2.3 ns	0.4865	2.3
	25	4	22.3 ± 2.8 ns	0.8235	1.8
daf-16	0	4	12.4 ± 0.3		
	10	4	13.1 ± 0.3 ***	<0.0001	6.0
	25	4	12.8 ± 0.3 ***	<0.0001	4.0
sir-2.1	0	2	18.5 ± 2.8		
	10	2	19.9 ± 1.7 *	0.0199	7.6
	25	2	20.8 ± 3.3 **	0.0002	12.4
aak-2	0	4	15.5 ± 0.5		
	10	4	15.4 ± 0.5 ns	0.5441	0.6
	25	4	14.0 ± 0.3 ***	<0.0001	−5.6
skn-1	0	3	18.1 ± 0.8		
	10	3	20.7 ± 0.7 ***	<0.0001	14.4
	25	3	20.3 ± 0.4 ***	<0.0001	12.1

→

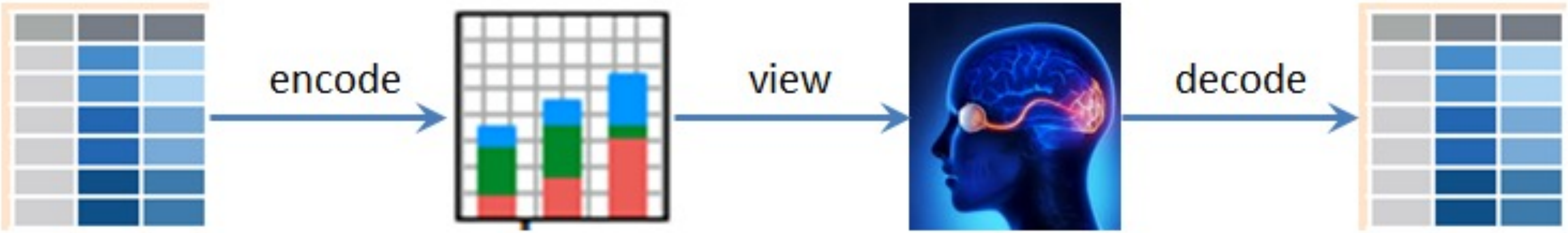
Picture



→

Brain

shutterstock.com · 1666544830



encode

view

decode

Source: **Michael Friendly**, http://euclid.psych.yorku.ca/www/psy6135/#Graphical_Perception

Making pictures that help analyze data

- We’d like to make, not just any kind of picture or graph, but one that transfers some part of the data to our brain
- How do we make sure that the graphs we make transfer:
  - The right part of the data, and;
  - As much of it as possible?

This is where the **“grammar of graphics”** comes in.

Goal is to **specify how data map to picture**, so the correct type and largest amount possible is transferred

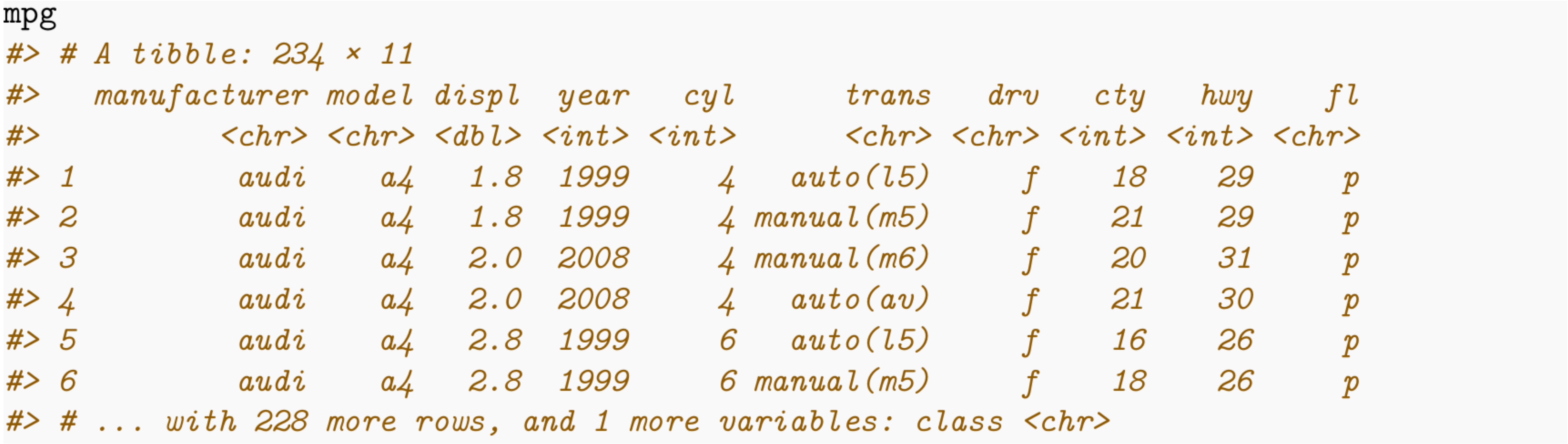
Grammar of graphics (Wickham version)

- http://r4ds.had.co.nz/visualize.html
- Map raw data to following elements:
    - Aesthetics (position, shape, color, ...)
    - Geometric objects (points, lines, bars, ...)
    - Scales (continuous, discrete, ...)
    - Facets (small multiples)
- Additionally, can apply:
    - Statistical transformation (identity, binning, median, ...)
    - Coordinate system (Cartesian, polar, parallel, ...

Grammar of graphics (Wickham version)

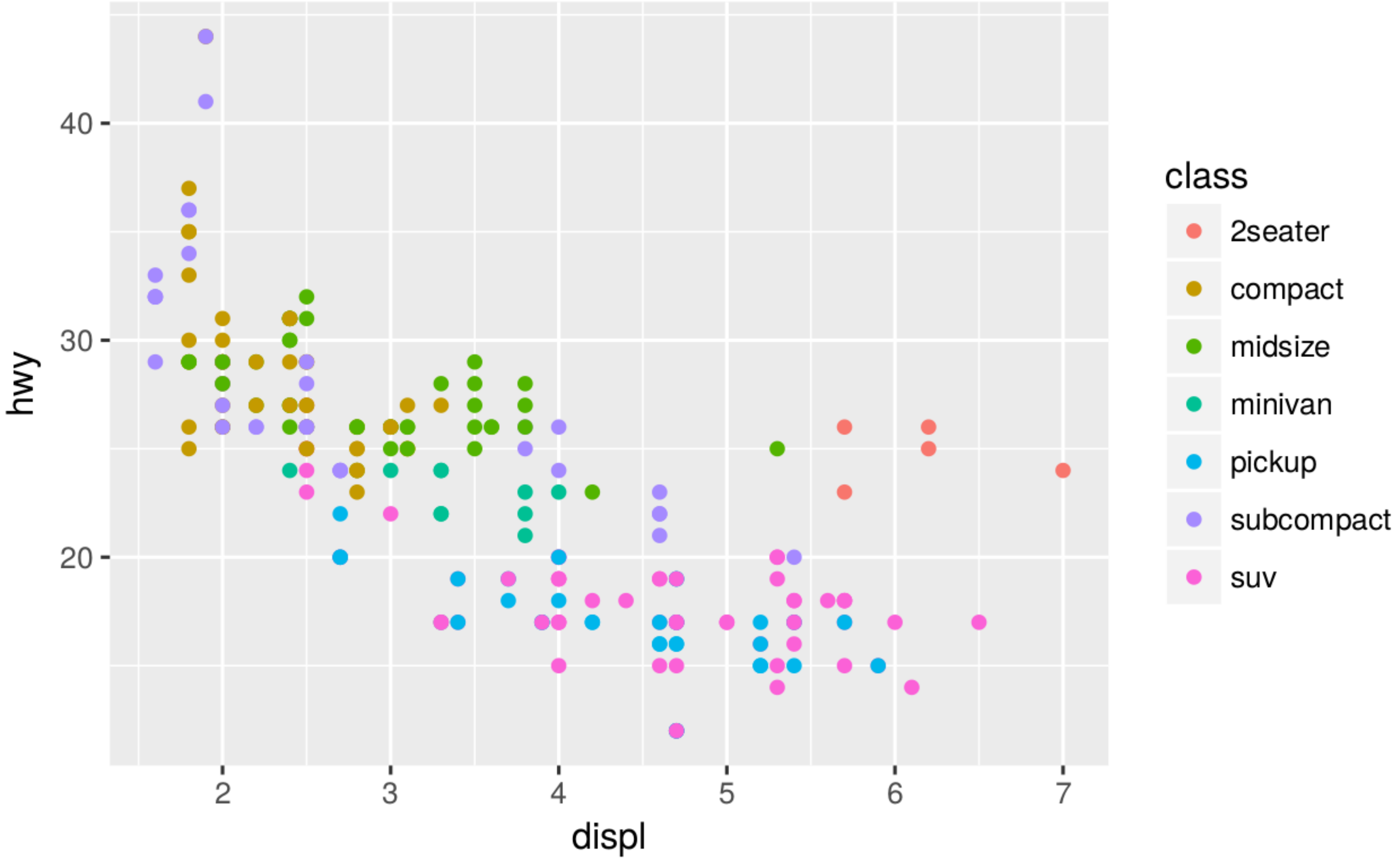
In R, grammar of graphics is implemented in `ggplot`, a function in the `ggplot2` package.

Example data set: cars



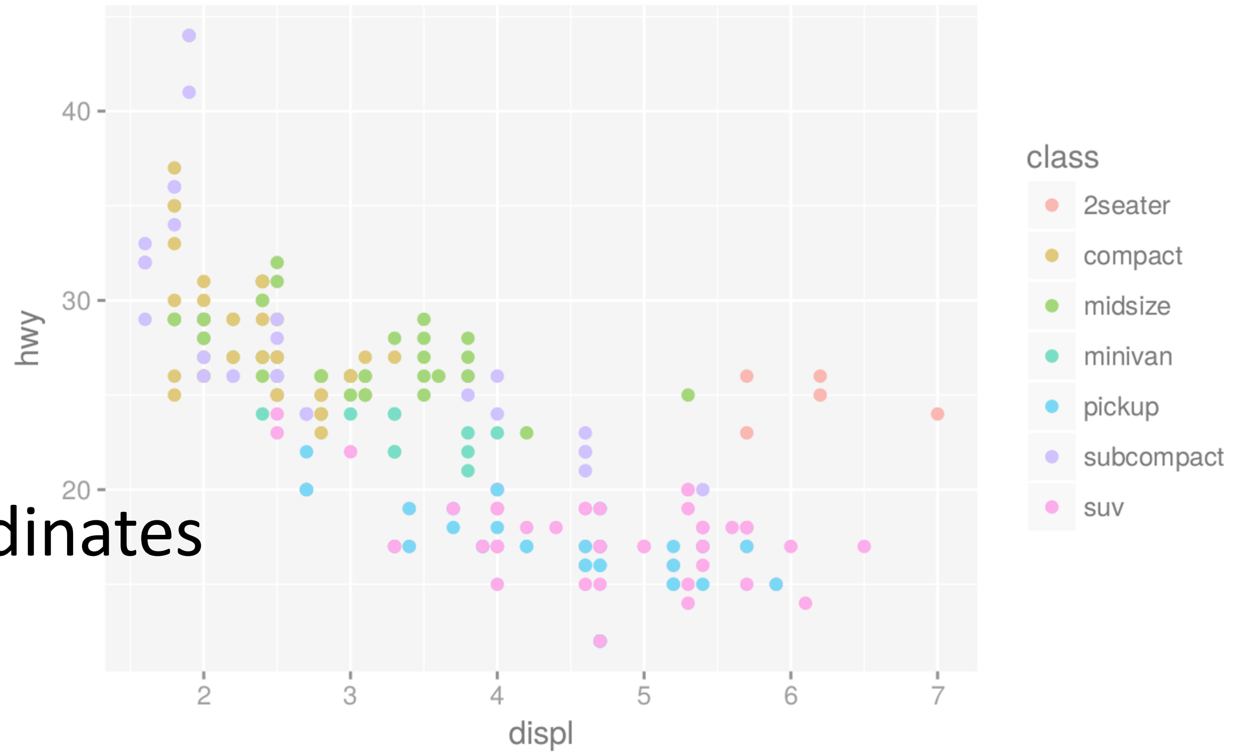
```
mpg
#> # A tibble: 234 × 11
#>   manufacturer model displ  year   cyl      trans   drv   cty   hwy    fl
#>          <chr> <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>
#> 1         audi    a4   1.8  1999     4   auto(l5)     f    18    29     p
#> 2         audi    a4   1.8  1999     4 manual(m5)     f    21    29     p
#> 3         audi    a4   2.0  2008     4 manual(m6)     f    20    31     p
#> 4         audi    a4   2.0  2008     4   auto(av)     f    21    30     p
#> 5         audi    a4   2.8  1999     6   auto(l5)     f    16    26     p
#> 6         audi    a4   2.8  1999     6 manual(m5)     f    18    26     p
#> # ... with 228 more rows, and 1 more variables: class <chr>
```

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy,
                                           color = class))
```

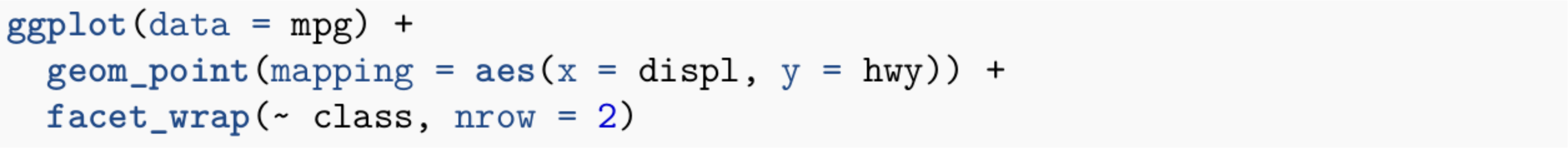


```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy,
                           color = class))
```

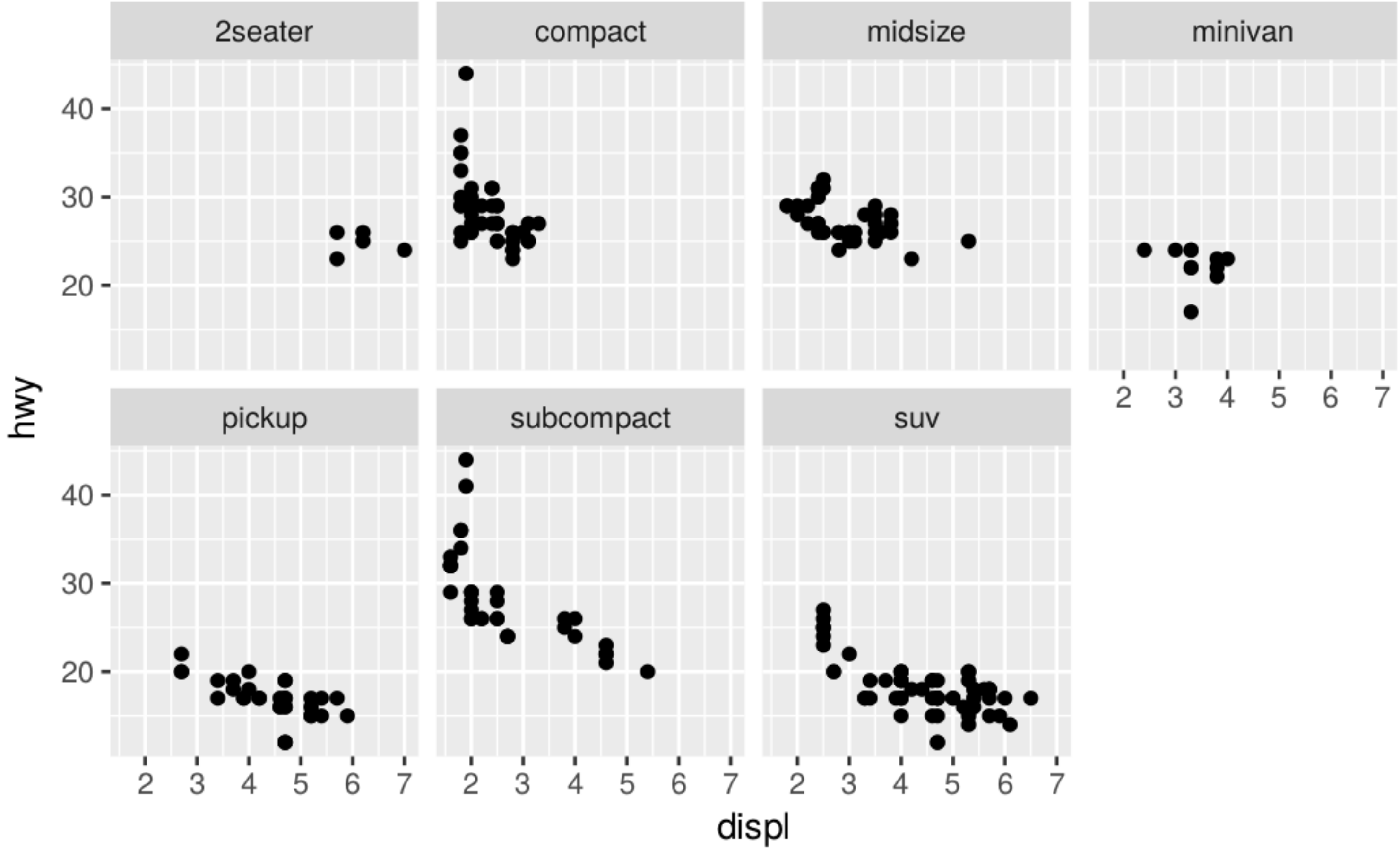
- Aesthetics:
  - x-position mapped to engine size
  - y-position mapped to fuel efficiency
  - color mapped to car type
- Geometric objects: points
- Transformation: identity
- Scales: continuous, cartesian coordinates
- No facets



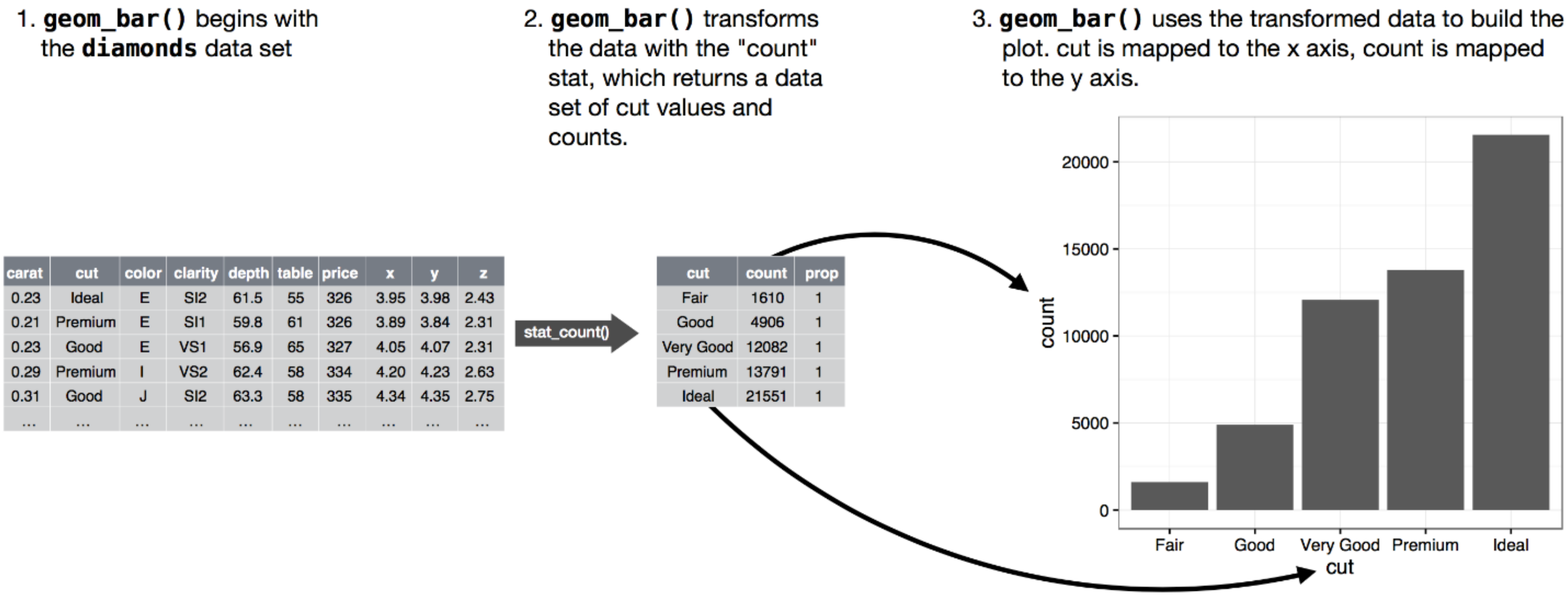
Facets



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)
```



Transformation (stats)



1. **geom_bar()** begins with the **diamonds** data set

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...

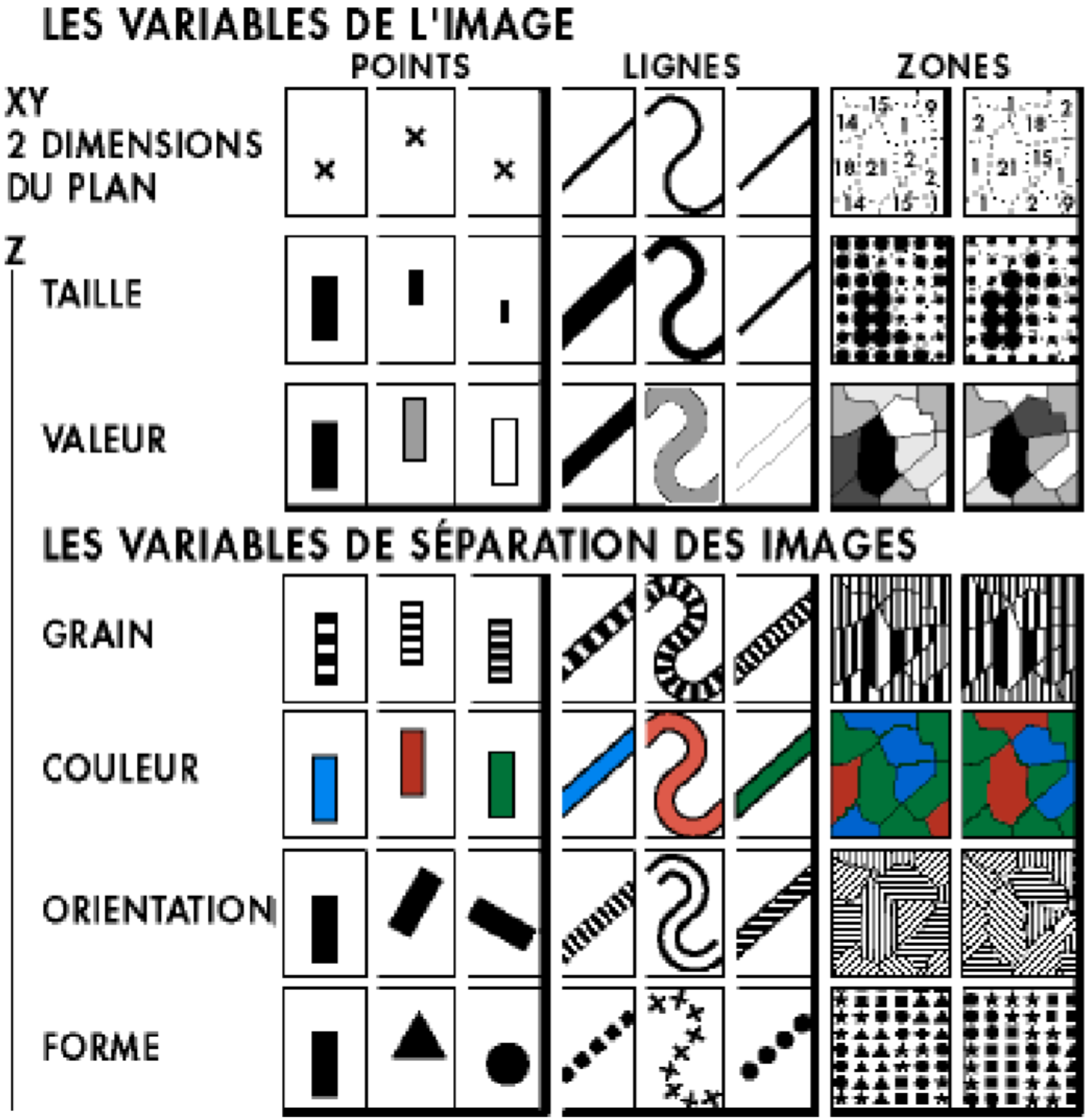
2. **geom_bar()** transforms the data with the "count" stat, which returns a data set of cut values and counts.

cut	count	prop
Fair	1610	1
Good	4906	1
Very Good	12082	1
Premium	13791	1
Ideal	21551	1

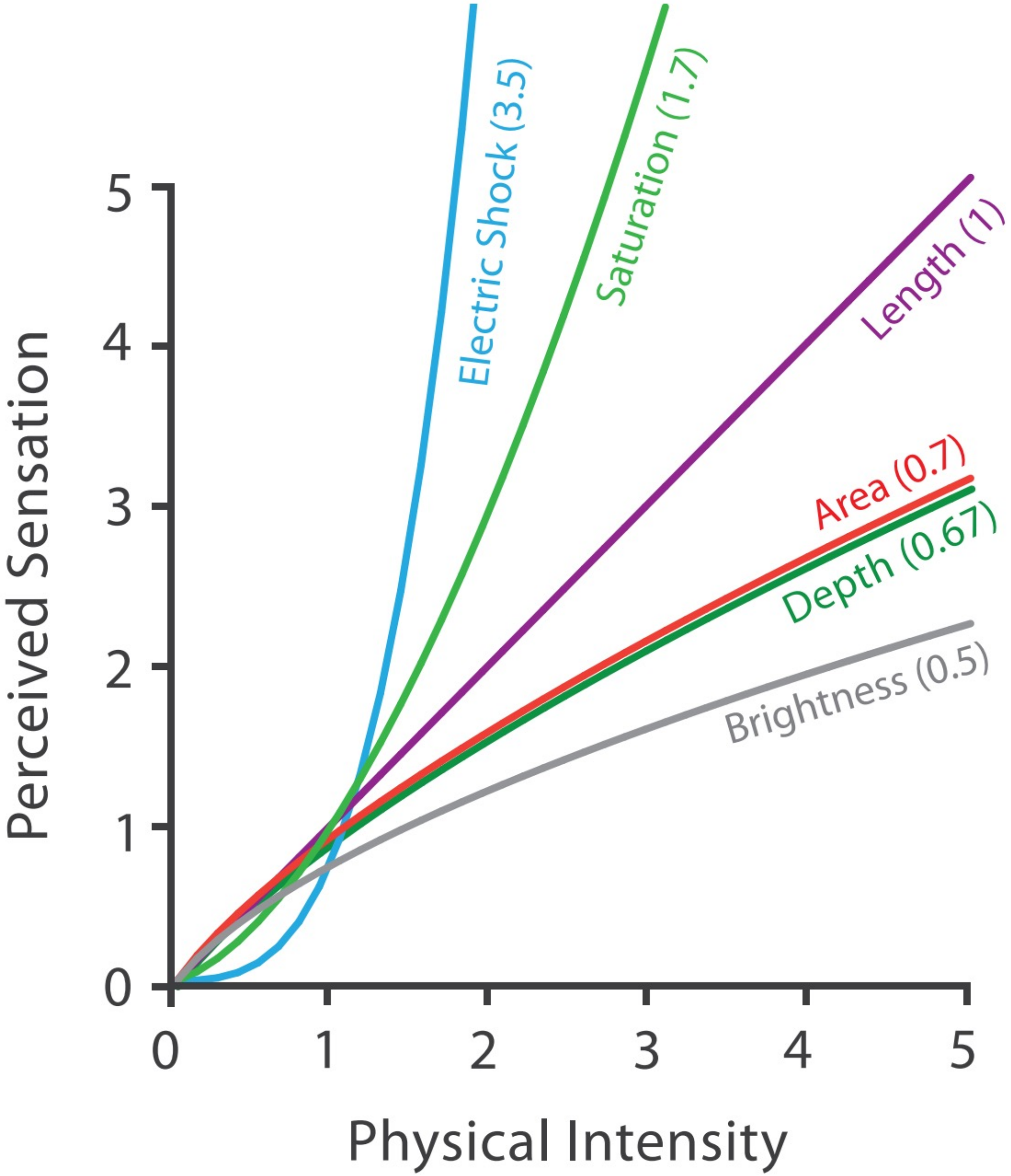
3. **geom_bar()** uses the transformed data to build the plot. cut is mapped to the x axis, count is mapped to the y axis.

Utrecht University

What should I choose?

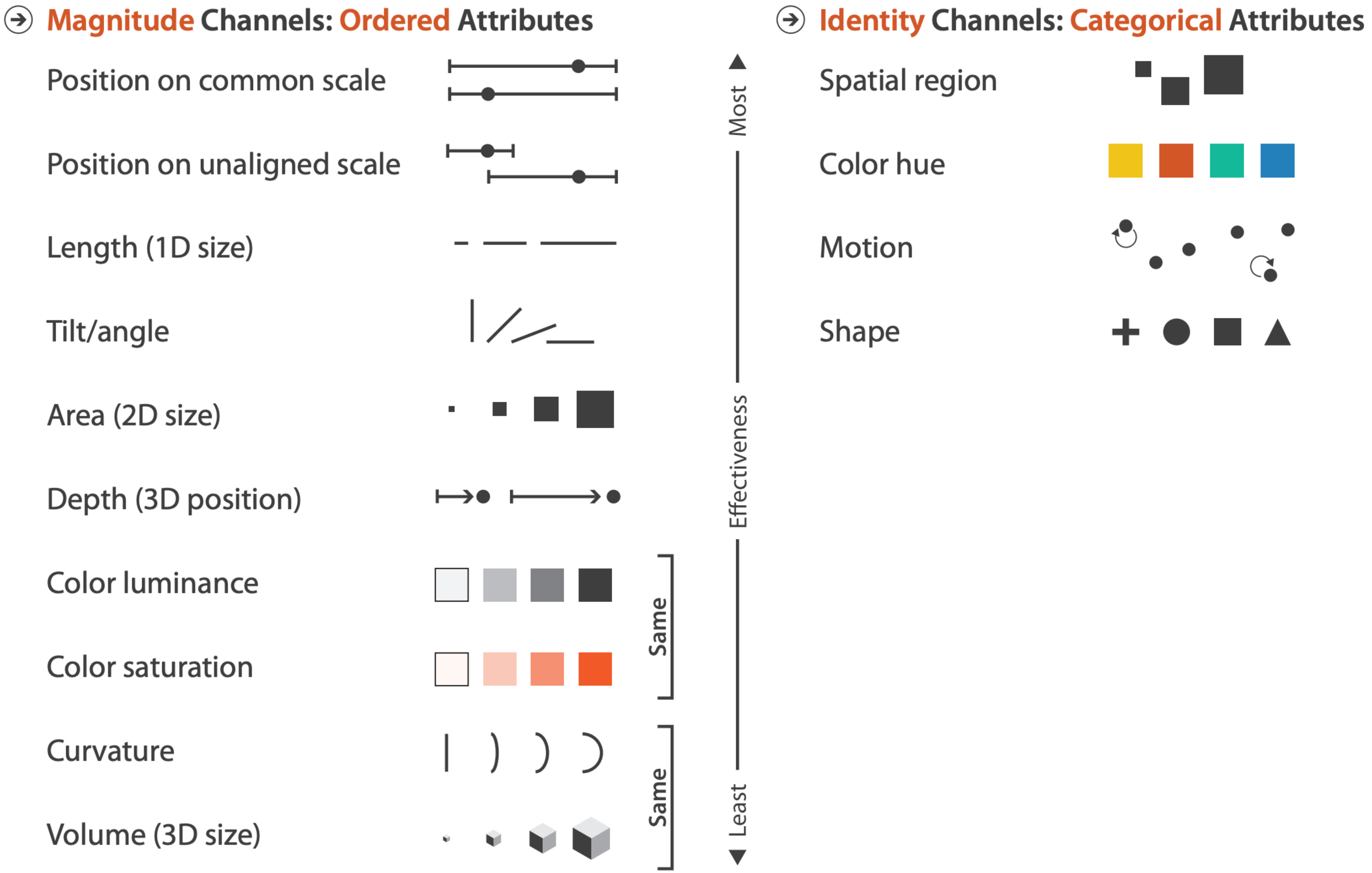

LES VARIABLES DE L'IMAGE
POINTS
LIGNES
ZONES
XY
2 DIMENSIONS
DU PLAN
Z
TAILLE
VALEUR
LES VARIABLES DE SÉPARATION DES IMAGES
GRAIN
COULEUR
ORIENTATION
FORME

Steven's Psychophysical Power Law: $S= I^N$



Source: **Tamara Munzer** (2014). Visualization Analysis and Design.

Channels: Expressiveness Types and Effectiveness Ranks



Magnitude Channels: Ordered Attributes

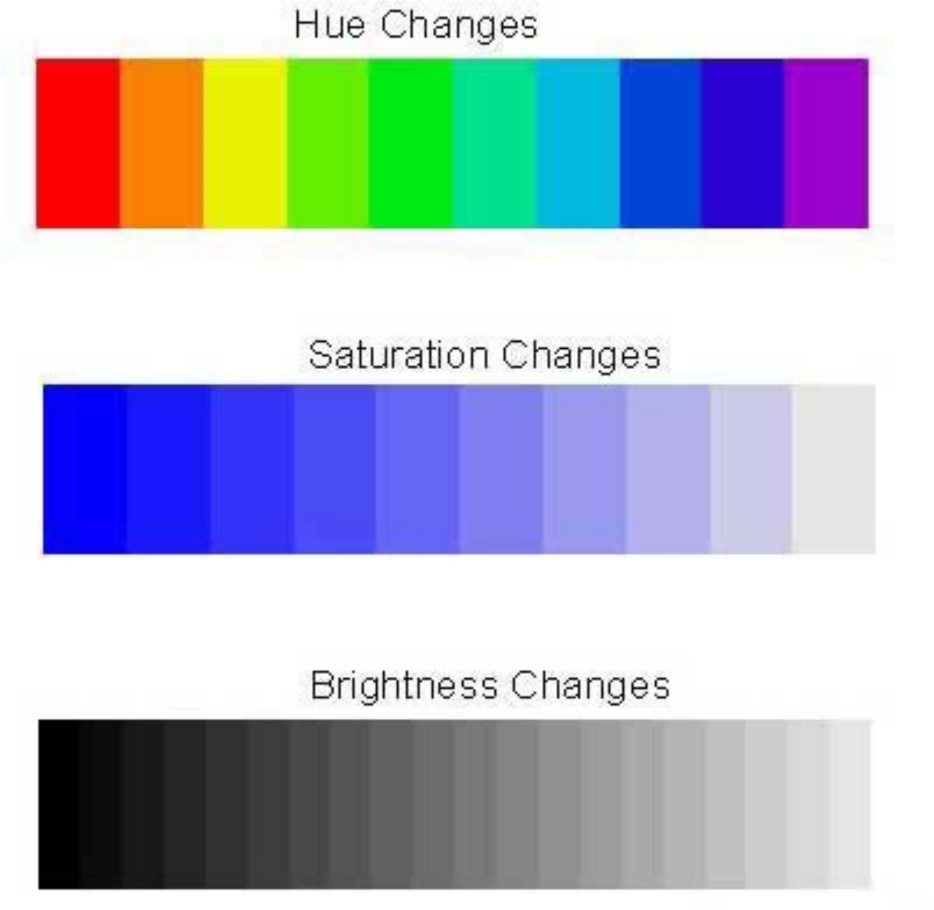
- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

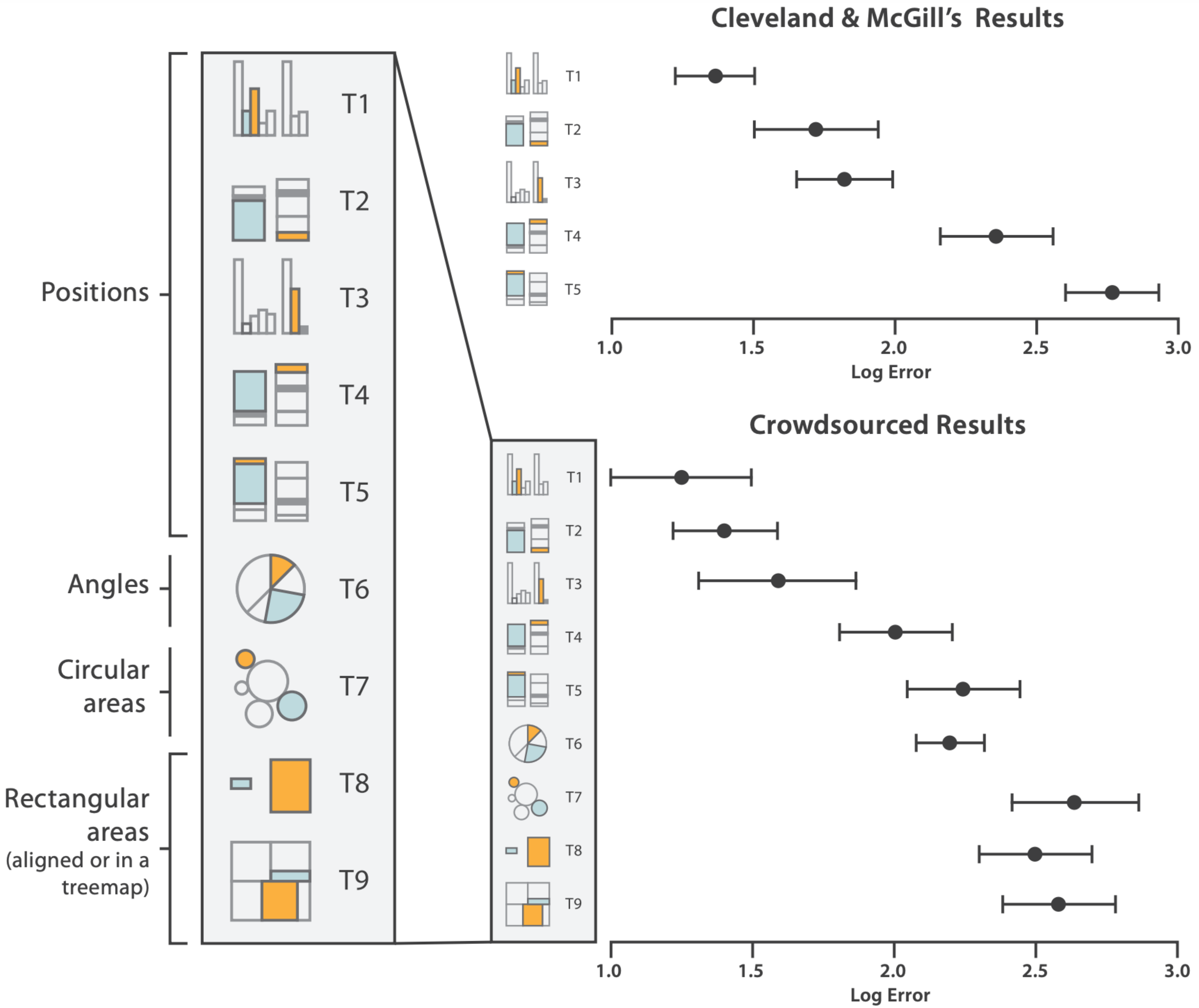
Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape

Source: **Tamara Munzer** (2014). Visualization Analysis and Design.

Color: hue-saturation-brightness (HSB)





Source: **Tamara Munzer** (2014). Visualization Analysis and Design.

How many 5s in this display?

1561321203658413076510374627
4173127527327592732990709742
1703707774179527931749270973
4019743217909370945179279417

How many 5s in this display?

1**5**613212036**5**841307 6**5**10374627
4173127**5**27327**5**92732990709742
1703707774179**5**27931749270973
401974321790937094**5**179279417

Numerals differ only in shape, and are high-level symbols
You have to literally scan them **all** & count the 5s.
The distinction of color is immediate & **pre-attentive**
You only have to scan & count the 5s.

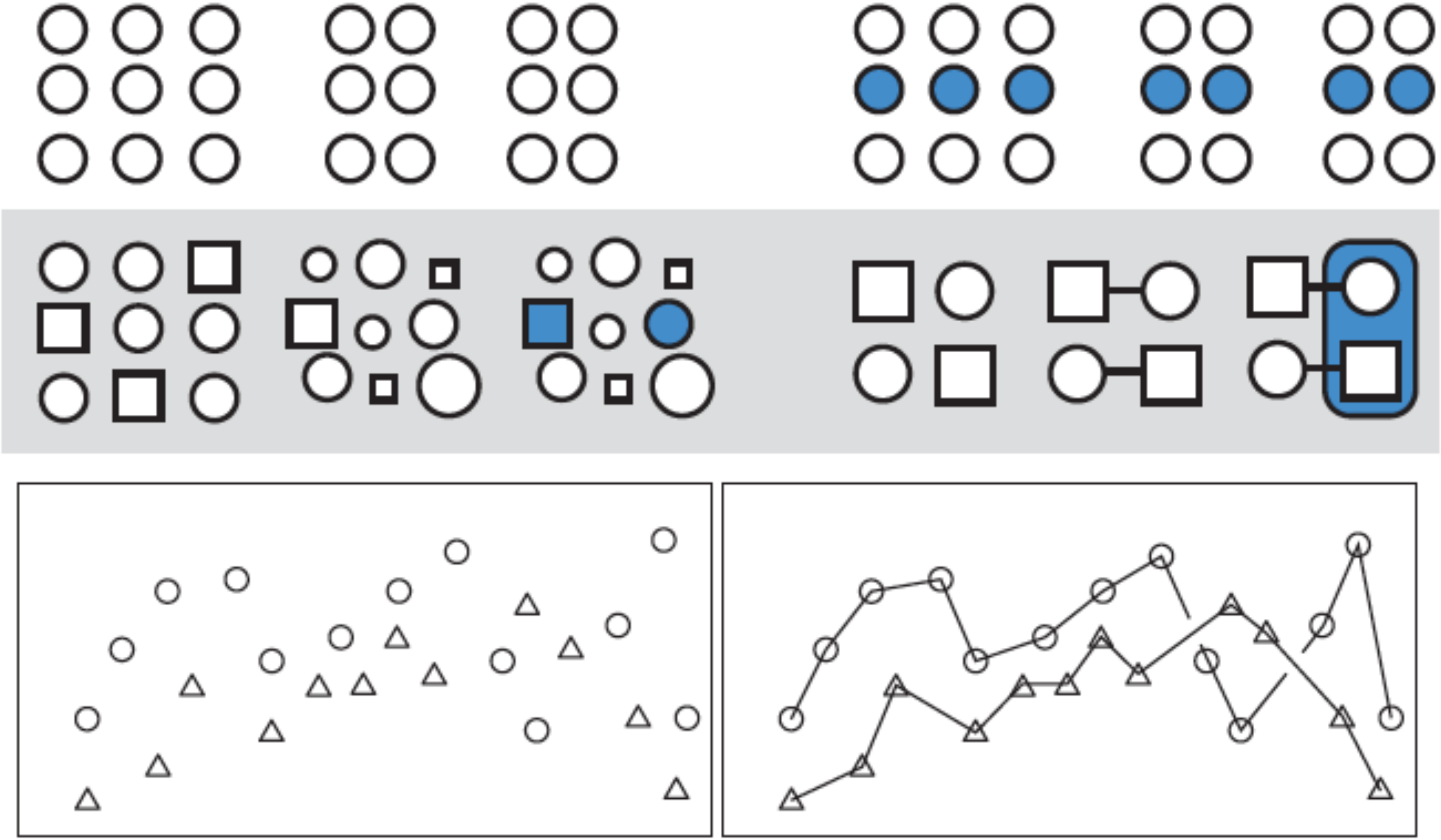
This is why color is an important visual attribute for a categorical variable in graphs

Utrecht University

Source: **Michael Friendly**, http://euclid.psych.yorku.ca/www/psy6135/#Graphical_Perception

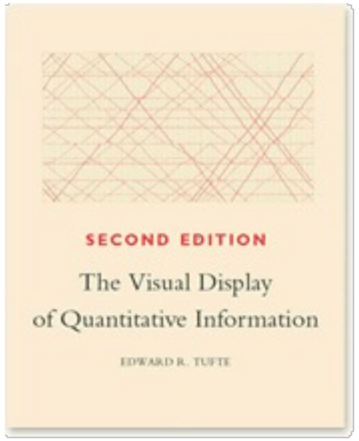
Gestalt principles of relatedness

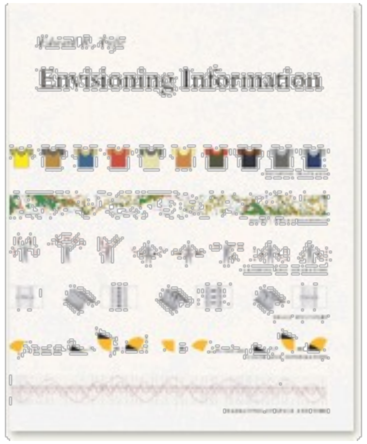
- **Proximity**: Things that are spatially near to one another seem to be related.
- **Similarity:** Things that look alike seem to be related.
- **Connection**: Things that are visually tied to one another seem to be related.
- **Continuity**: Partially hidden objects are completed into familiar shapes.
- **Closure**: Incomplete shapes are perceived as complete.
- **Figure and ground**: Visual elements are taken to be either in the foreground or in the background.
- **Common fate**: Elements sharing a direction of movement are perceived as a unit.

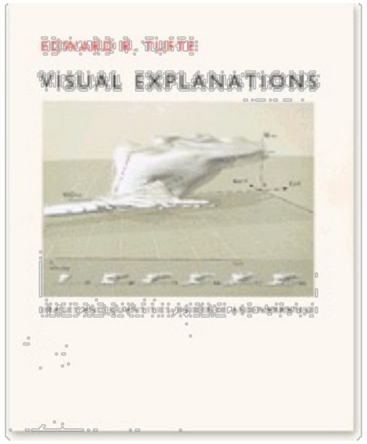


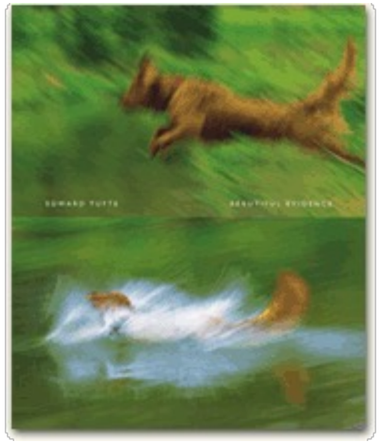
Source: Healy (2019)

Some (distilled) principles from Tufte

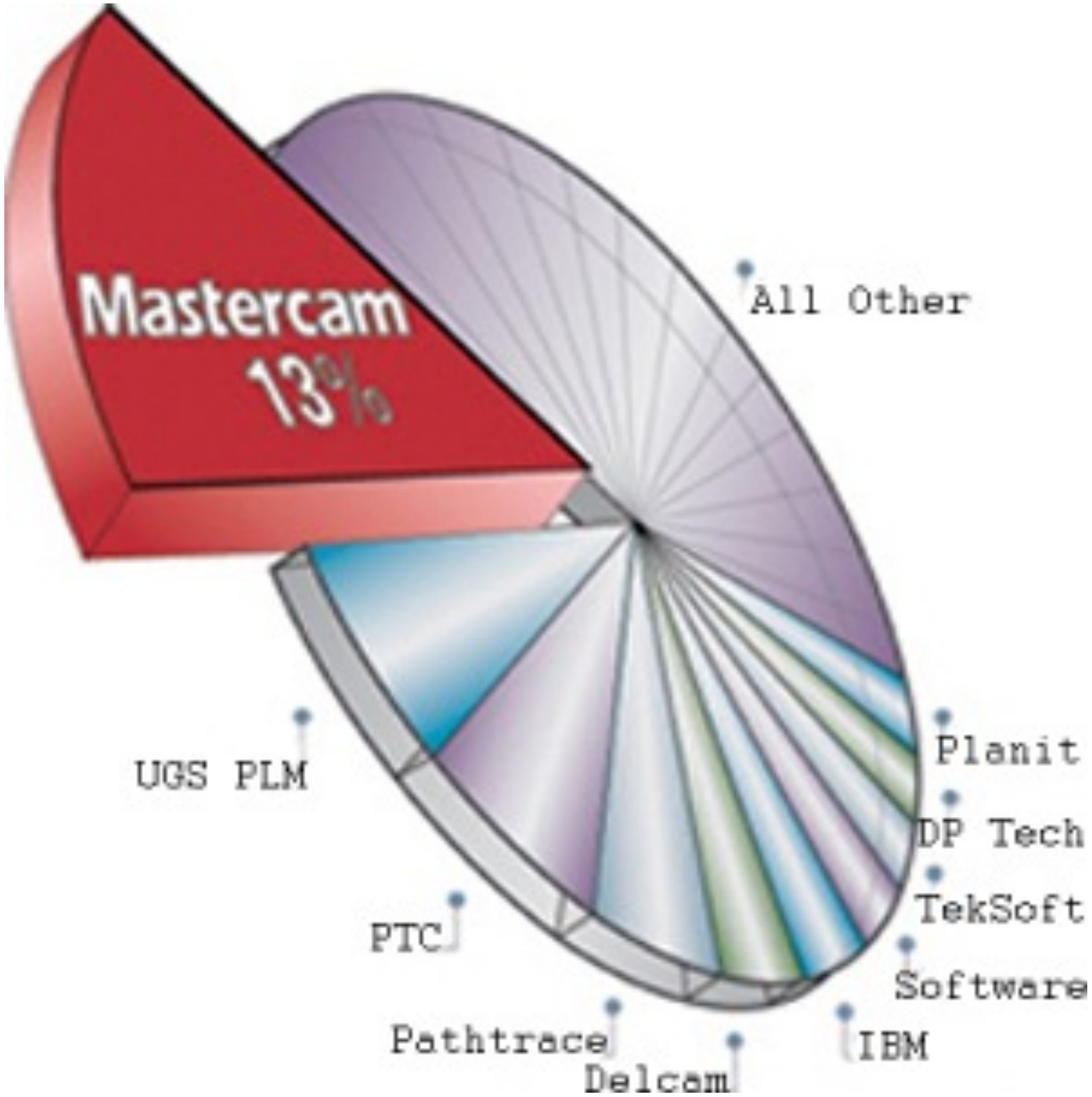








- Ask how data maps to perception
- Ask which comparisons you want, guide eye to those
- Maximize data-to-ink ratio
- Present more data (without losing interpretability)
- Use levels of detail
- (Remember narrative)


Mastercam
13%
All Other
Planit
DP Tech
TekSoft
Software
IBM
Delcam
Pathtrace
PTC
UGS PLM

Nightingale Rose / Coxcomb chart

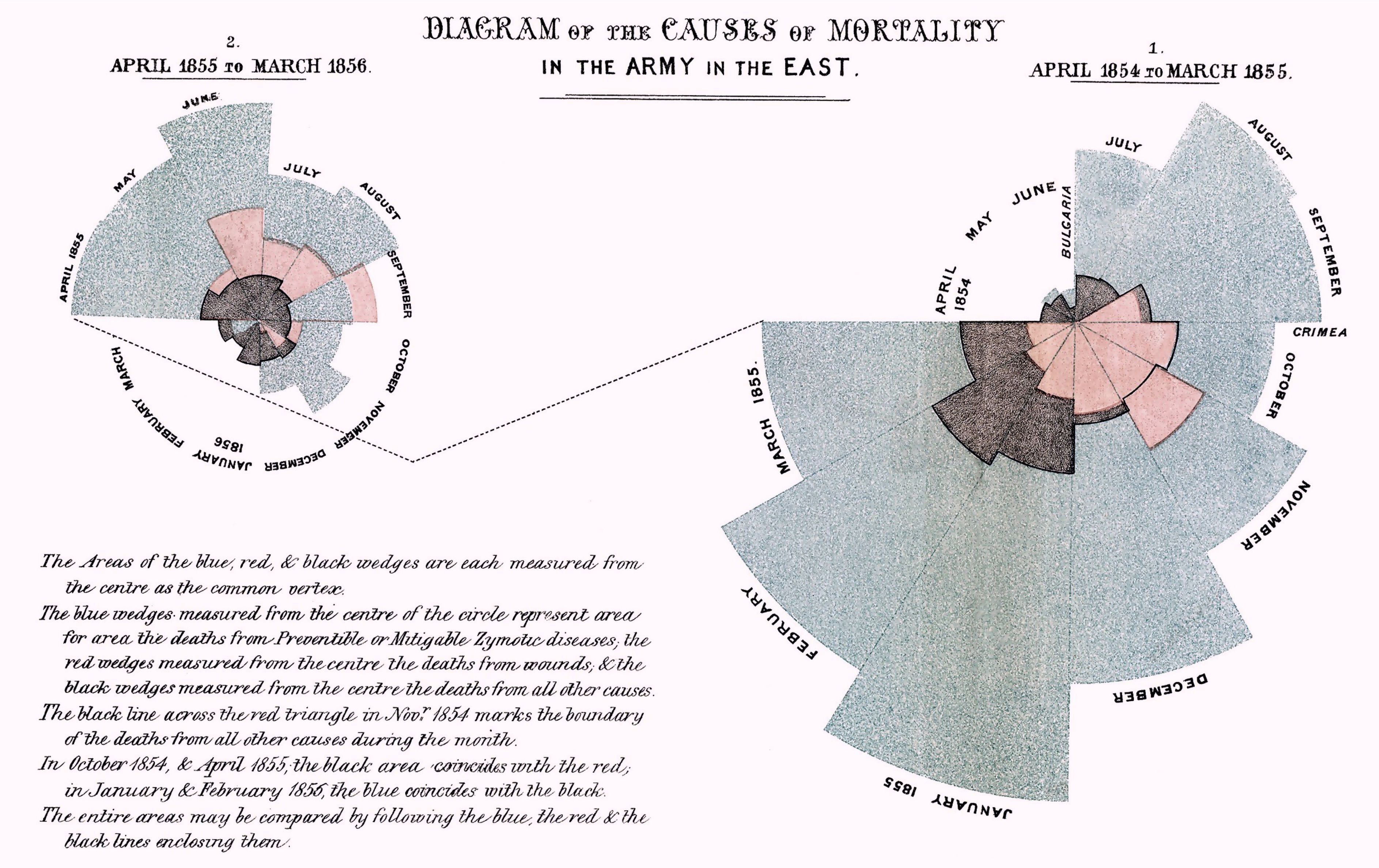


DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST.

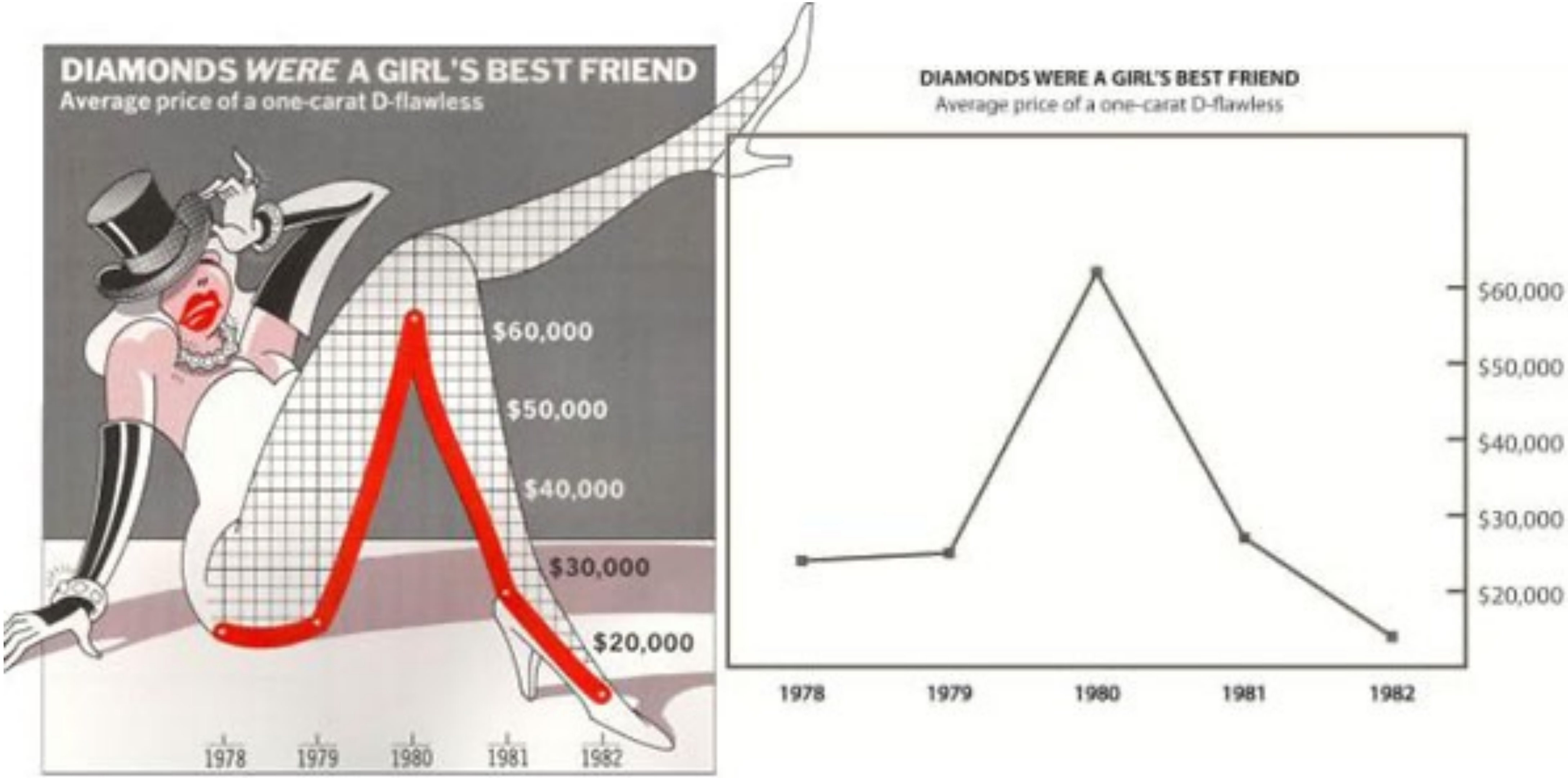
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

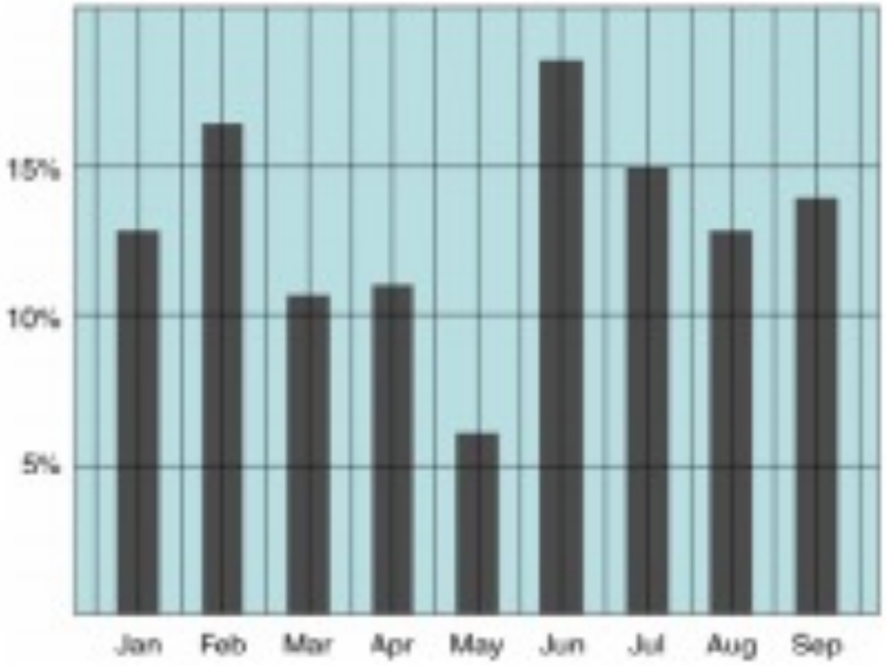
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

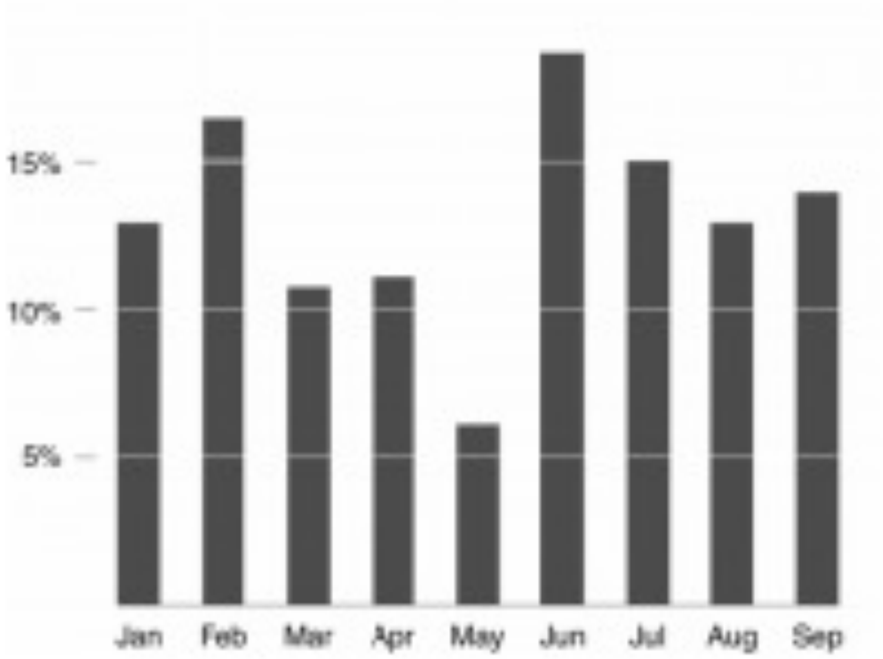
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.

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

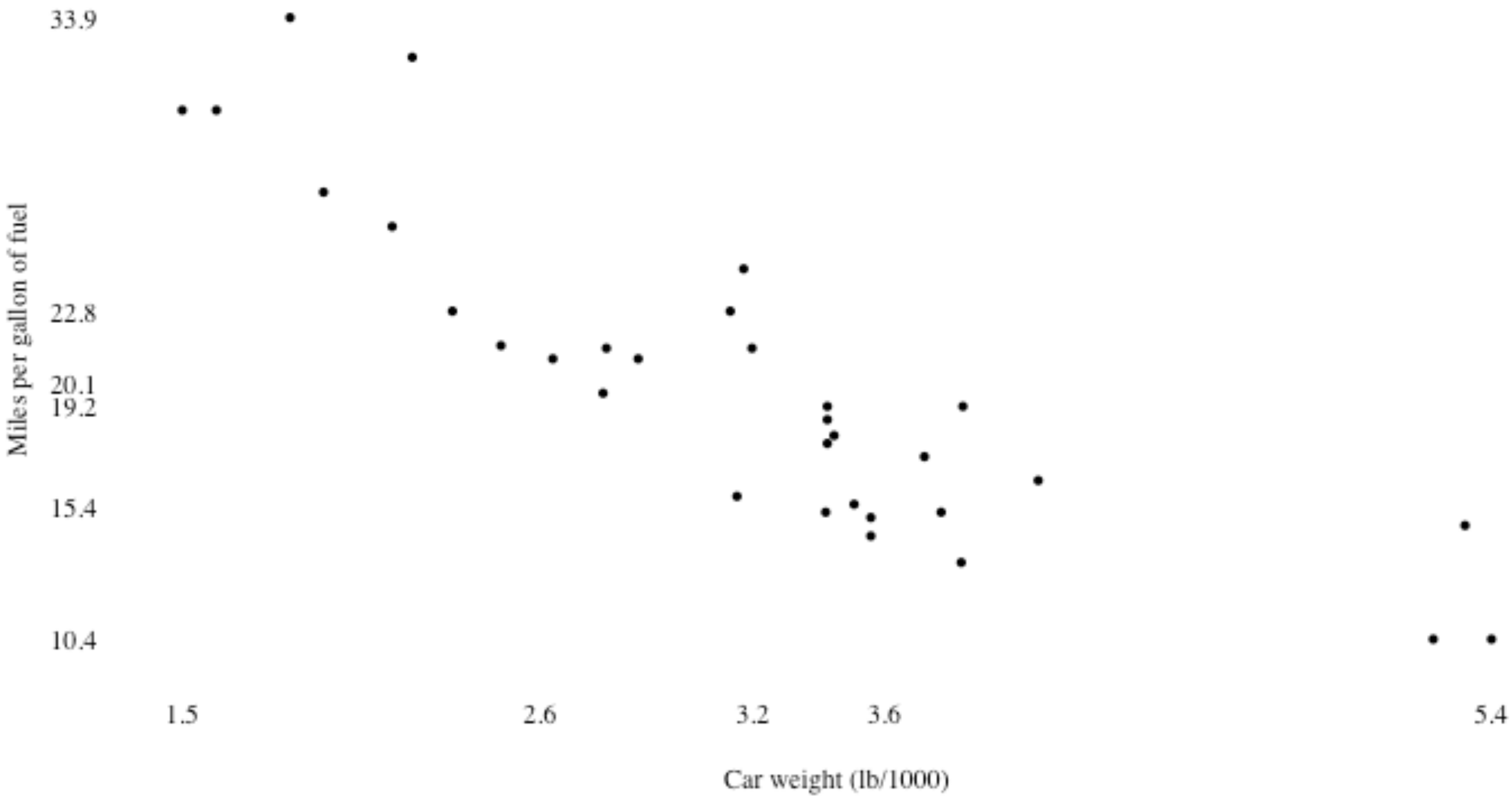
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

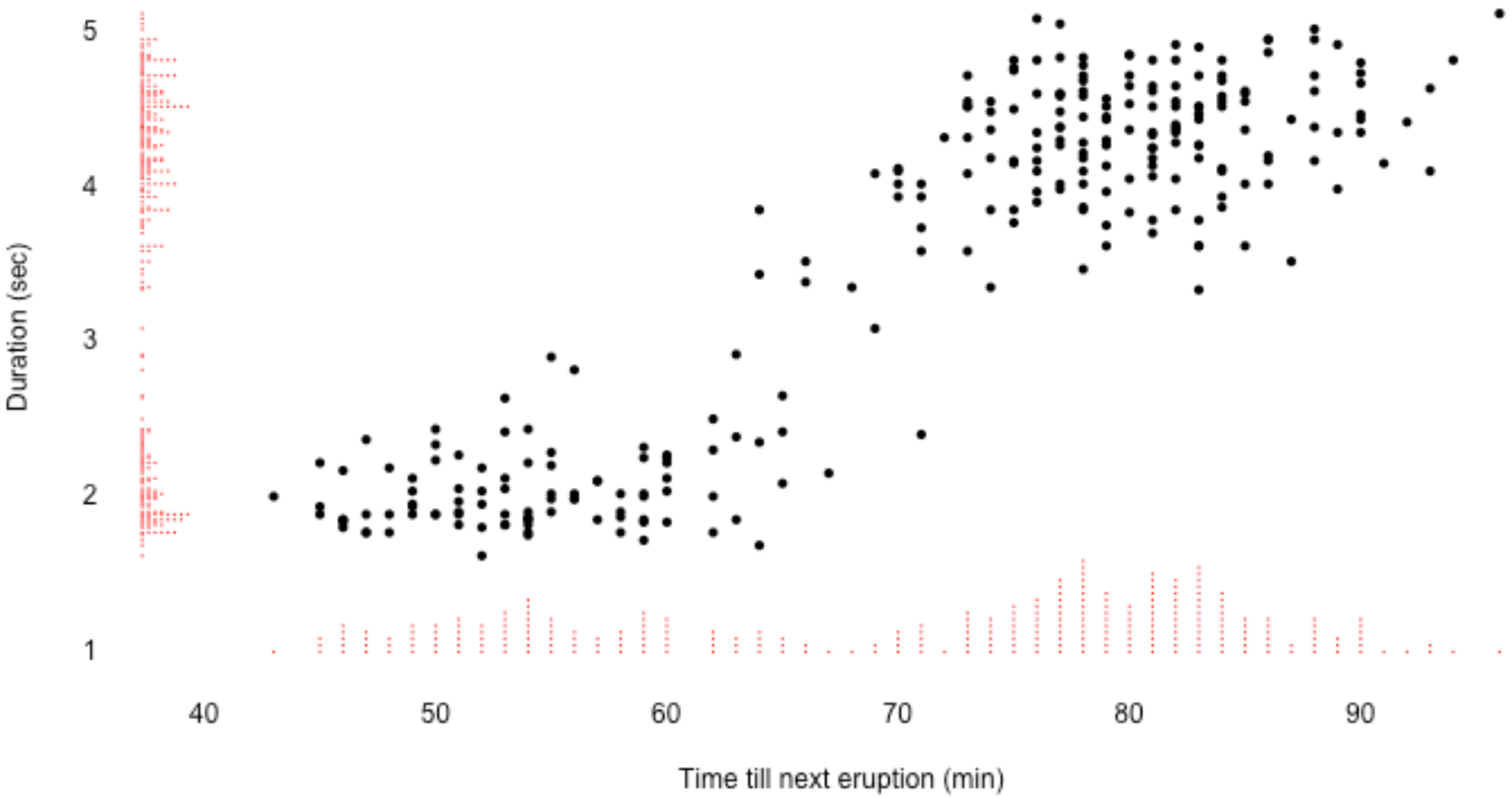

DIAMONDS *WERE* A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless
$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982
DIAMONDS WERE A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless
$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982





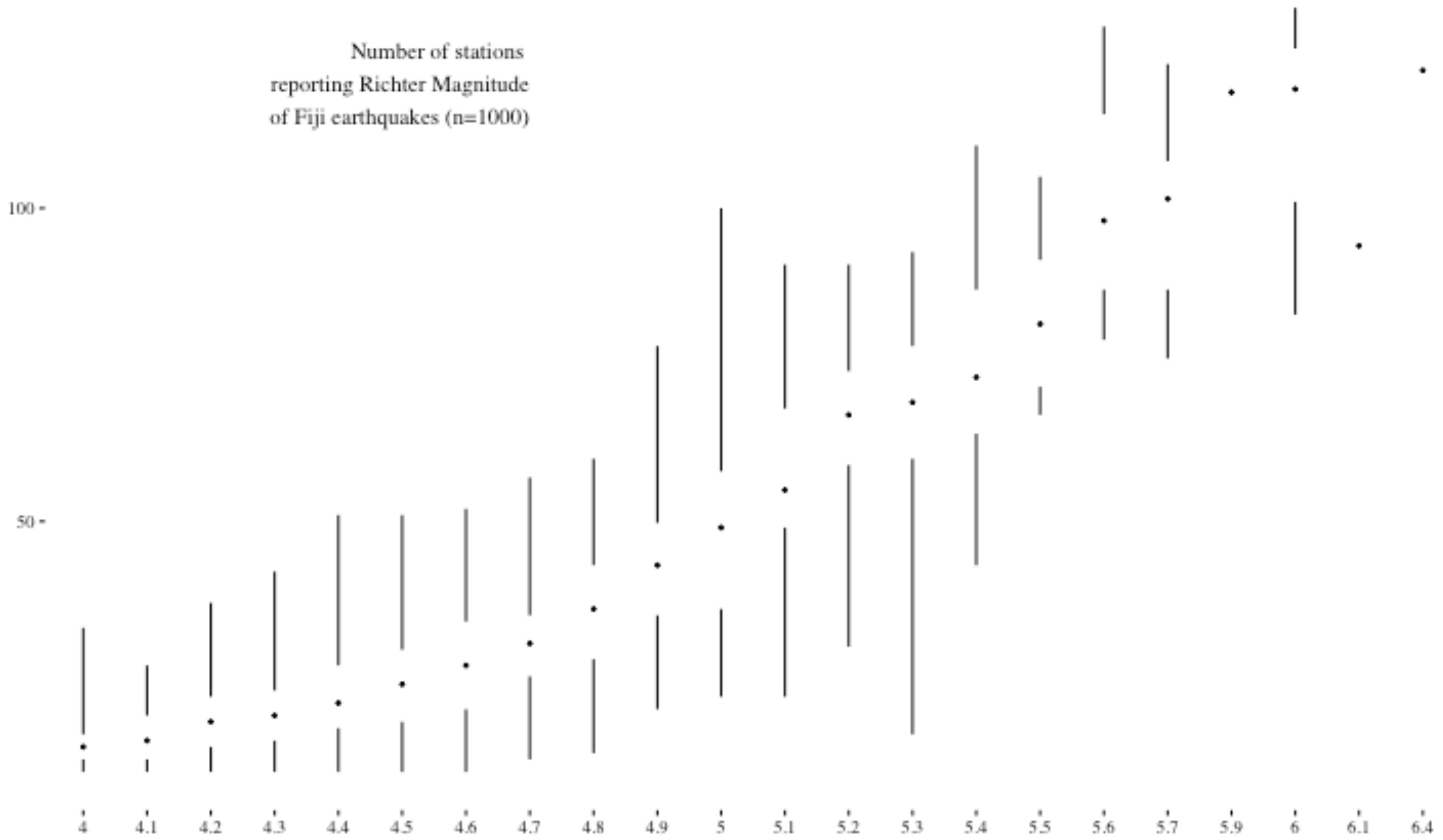
Low Data/Ink High Data/Ink


33.9
22.8
20.1
19.2
15.4
10.4
Miles per gallon of fuel
1.5
2.6
3.2
3.6
5.4
Car weight (lb/1000)



Duration (sec)
5
4
3
2
1
40
50
60
70
80
90
Time till next eruption (min)

```
ggplot(quakes, aes(factor(mag), stations)) +
  theme_tufte() +
  geom_tufteboxplot(outlier.colour = "transparent") +
  theme(axis.title = element_blank())
```



Tufte wisdom

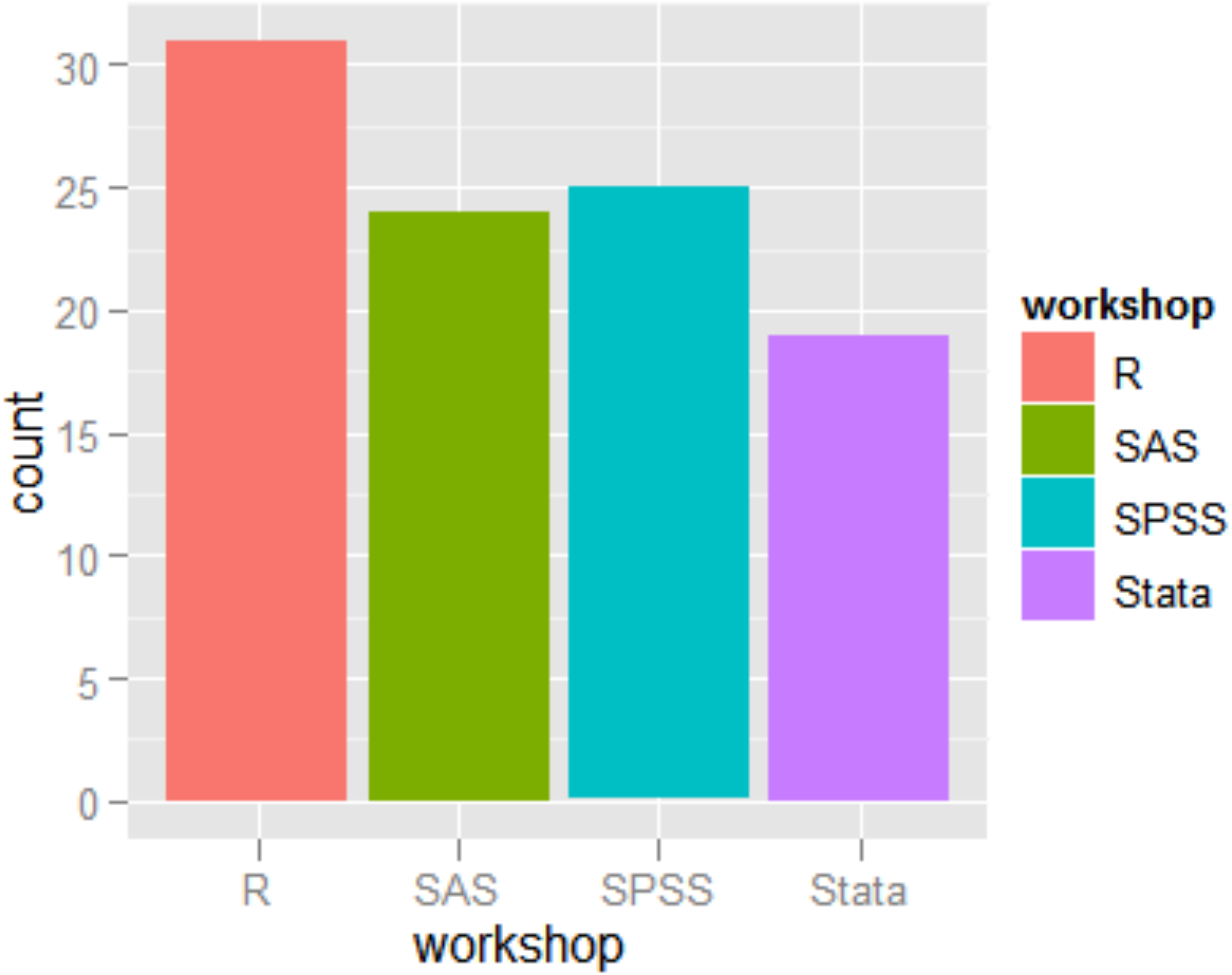
- Tufte’s principles are more oriented to communication and can be taken too far

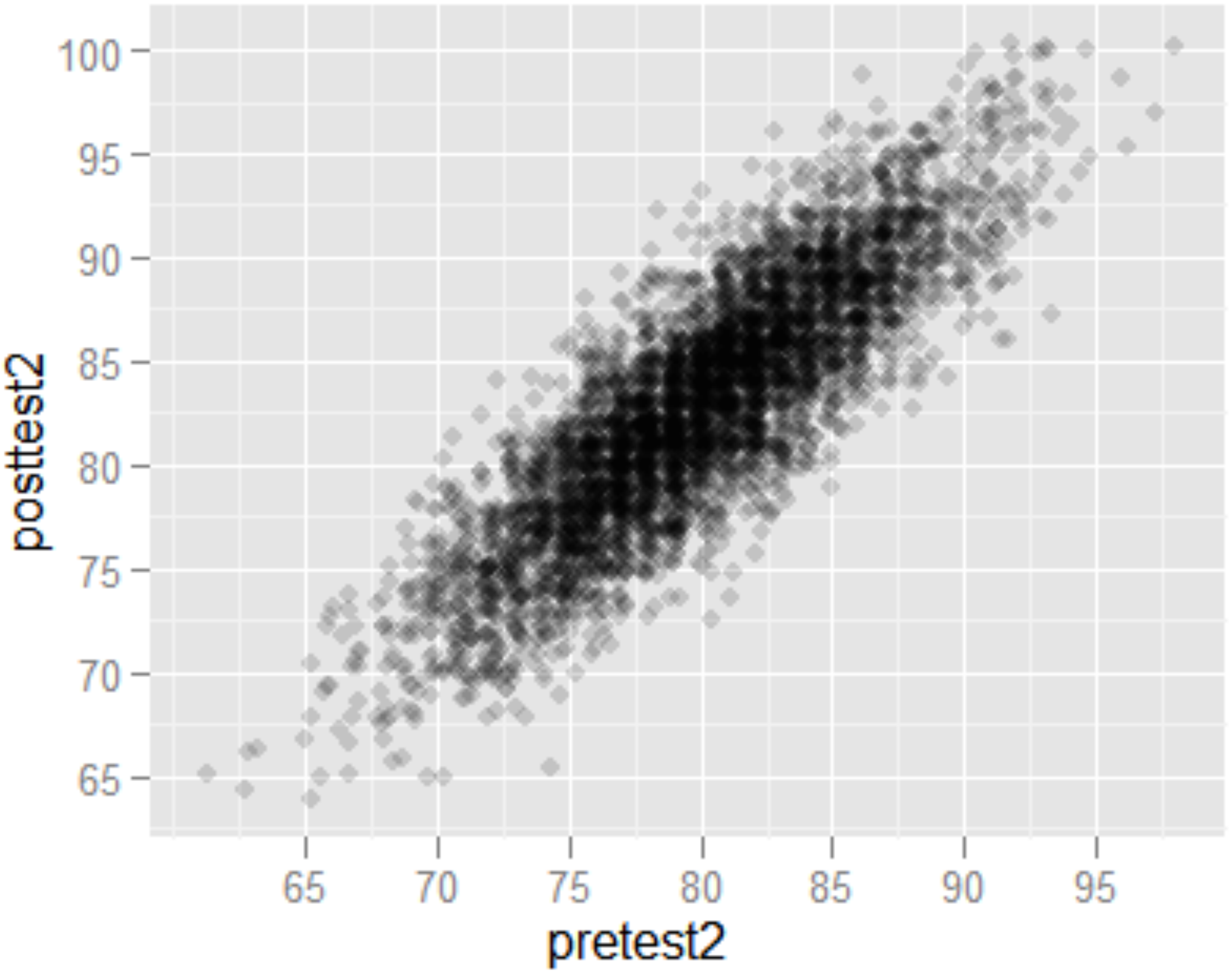
- Better data/ink → display more information without overload;
- Thinking about perception can help you choose better geoms, aesthetics.

Some practice

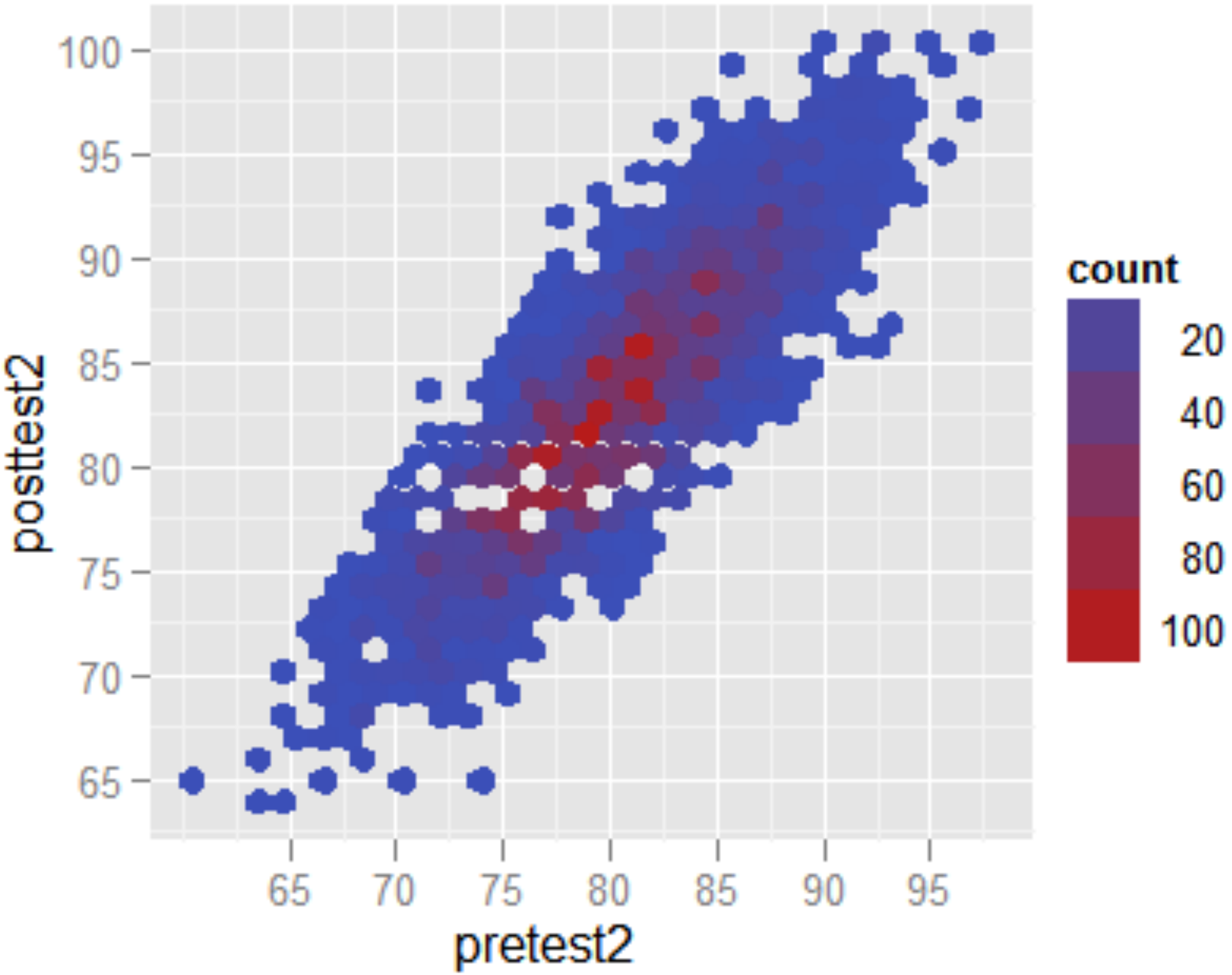
Answer these questions:

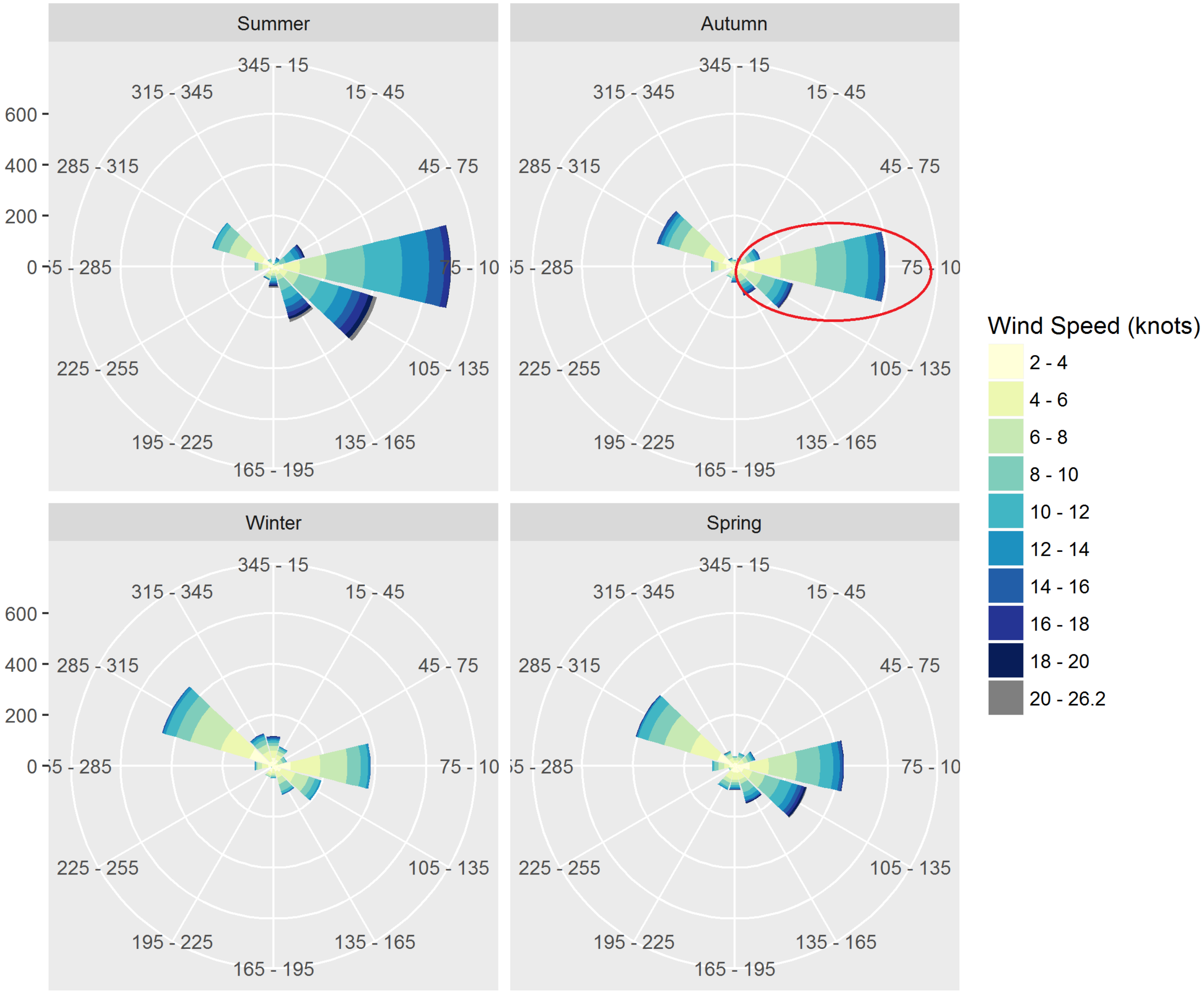
- Are we plotting the right thing?
- What are: aesthetics, geom, scale, facets, transformation, coordinate system
- How is data/ink?
- Is perception considered optimally?
- Can you think of questions you can’t answer from this plot which are in the data?


count
30
25
20
15
10
5
0
R
SAS
SPSS
Stata
workshop
workshop
R
SAS
SPSS
Stata



posttest2
100
95
90
85
80
75
70
65
pretest2
65
70
75
80
85
90
95


posttest2
100
95
90
85
80
75
70
65
pretest2
65
70
75
80
85
90
95
count
20
40
60
80
100


Summer
Autumn
Winter
Spring
345 - 15
15 - 45
45 - 75
75 - 105
105 - 135
135 - 165
165 - 195
195 - 225
225 - 255
255 - 285
285 - 315
315 - 345
600
400
200
0
Wind Speed (knots)
2 - 4
4 - 6
6 - 8
8 - 10
10 - 12
12 - 14
14 - 16
16 - 18
18 - 20
20 - 26.2

Conclusion

Conclusion

- Data visualization is data analysis + psychology;
- Sticking to **basic principles** helps:
  - **Map data** to aesthetics, geoms, scales, facets;
  - Perception research guides choices;
  - **Which comparisons** do I want?
  - Maximize **data-ink** (within reason).
- Some standard recipes (e.g. “barplot”, “histogram”, “line graph”), but pros do not need the cookbook…
- Don’t believe everything you hear (“do’s and don’ts”)