Visual Perception!

Daniel Anderson
Week 4, Class 1

UNIVERSITY OF
OREGON

Agenda

- Aesthetic mappings and visual encodings of data
- data/ink ratio
- Some do's and don't's (which are all rules 👍)

Learning Objectives

- Understand how decisions you make may help or hinder comprehension

Disclaimer

I'm not a psychologist

I don't really know/care why we percieve things certain ways.

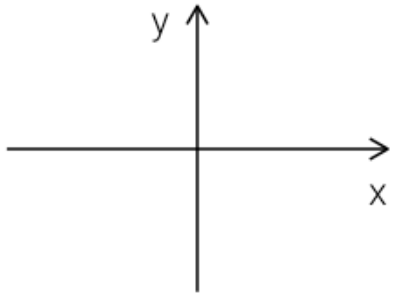
I mainly care that we do, and that your visualizations should account for them.

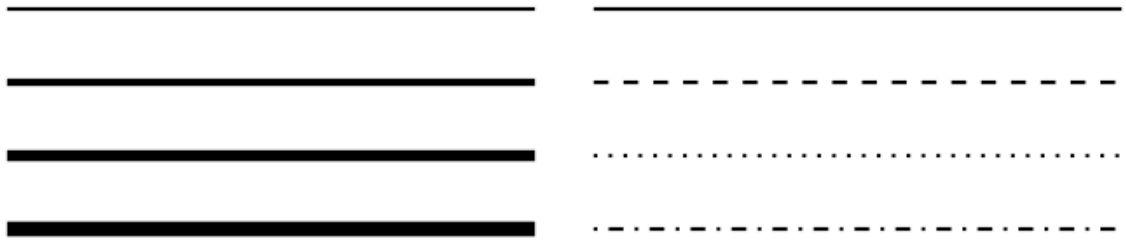
Visual Cues

- **Position:** *Numeric*. Where in relation to other things?
- **Length:** *Numeric*. How big (in one dimension)?
- **Angle:** *Numeric*. How wide? Parallel to something else?
- **Direction:** *Numeric*. At what slope? In a time series, going up or down?
- **Shape:** *Categorical*. Belonging to which group?
- **Area:** *Numeric*. How big (in two dimensions)?
- **Volumne:** *Numeric*. How big (in three dimensions)?
- **Shade:** *Numeric or Categorical*. To what extent? How Severely?
- **Color:** *Numeric or Categorical*. To what extent? How Severely?

Taken from *Modern Data Science with R*, p. 15

Different ways of encoding data





Other elements to consider

- Text
  - How is the text displayed (e.g., font, face, location)?
  - What is the purpose of the text?
- Transparency
  - Are there overlapping pieces?
  - Can transparency help?
- Type of data
  - Continuous/categorical
  - Which can be mapped to each aesthetic?
  - e.g., shape and line type can only be mapped to categorical data, whereas color and size can be mapped to either.

Talk with a neighbor

How would you encode each column of data?

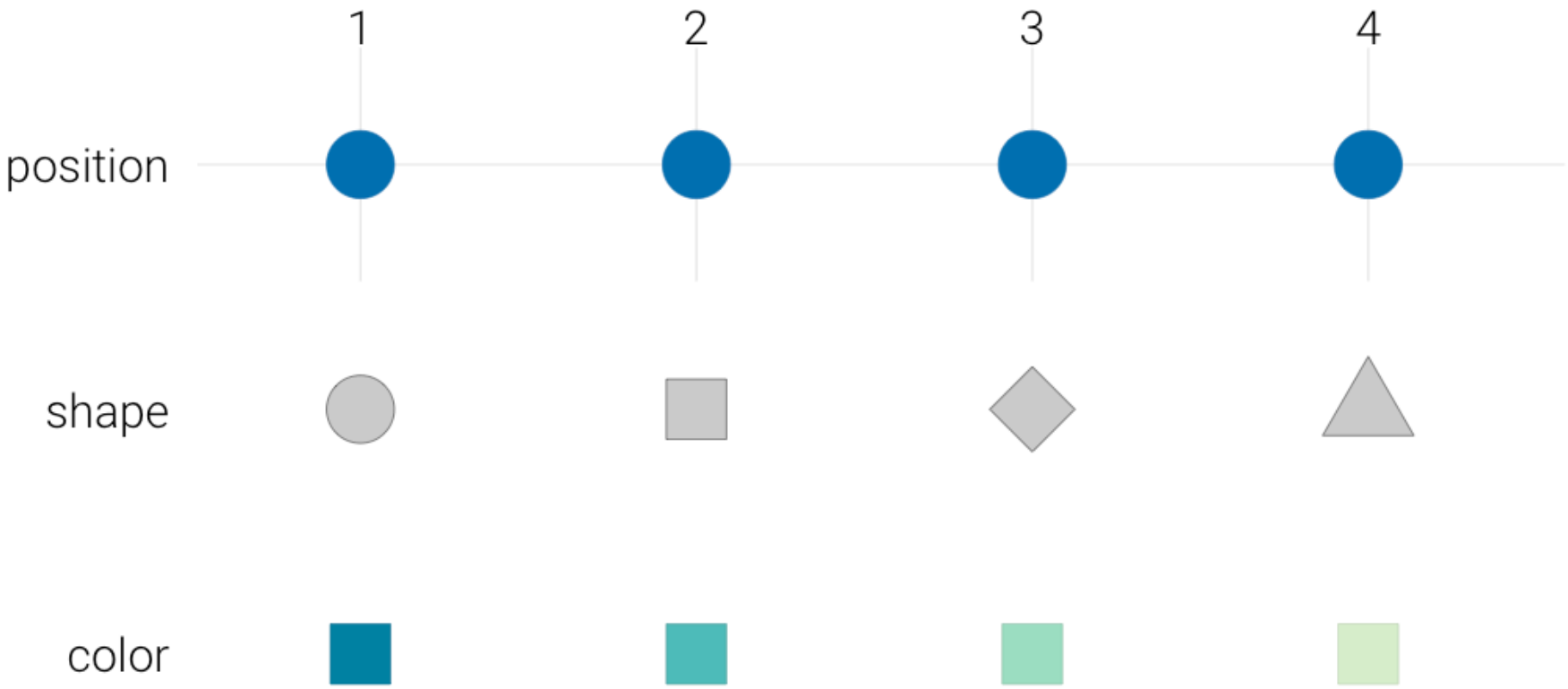
Month	Day	Location	Station ID	Temperature
Jan	1	Chicago	USW00014819	25.6
Jan	1	San Diego	USW00093107	55.2
Jan	1	Houston	USW00012918	53.9
Jan	1	Death Valley	USC00042319	51.0
Jan	2	Chicago	USW00014819	25.5
Jan	2	San Diego	USW00093107	55.3
Jan	2	Houston	USW00012918	53.8
Jan	2	Death Valley	USC00042319	51.2
Jan	3	Chicago	USW00014819	25.3

Scales

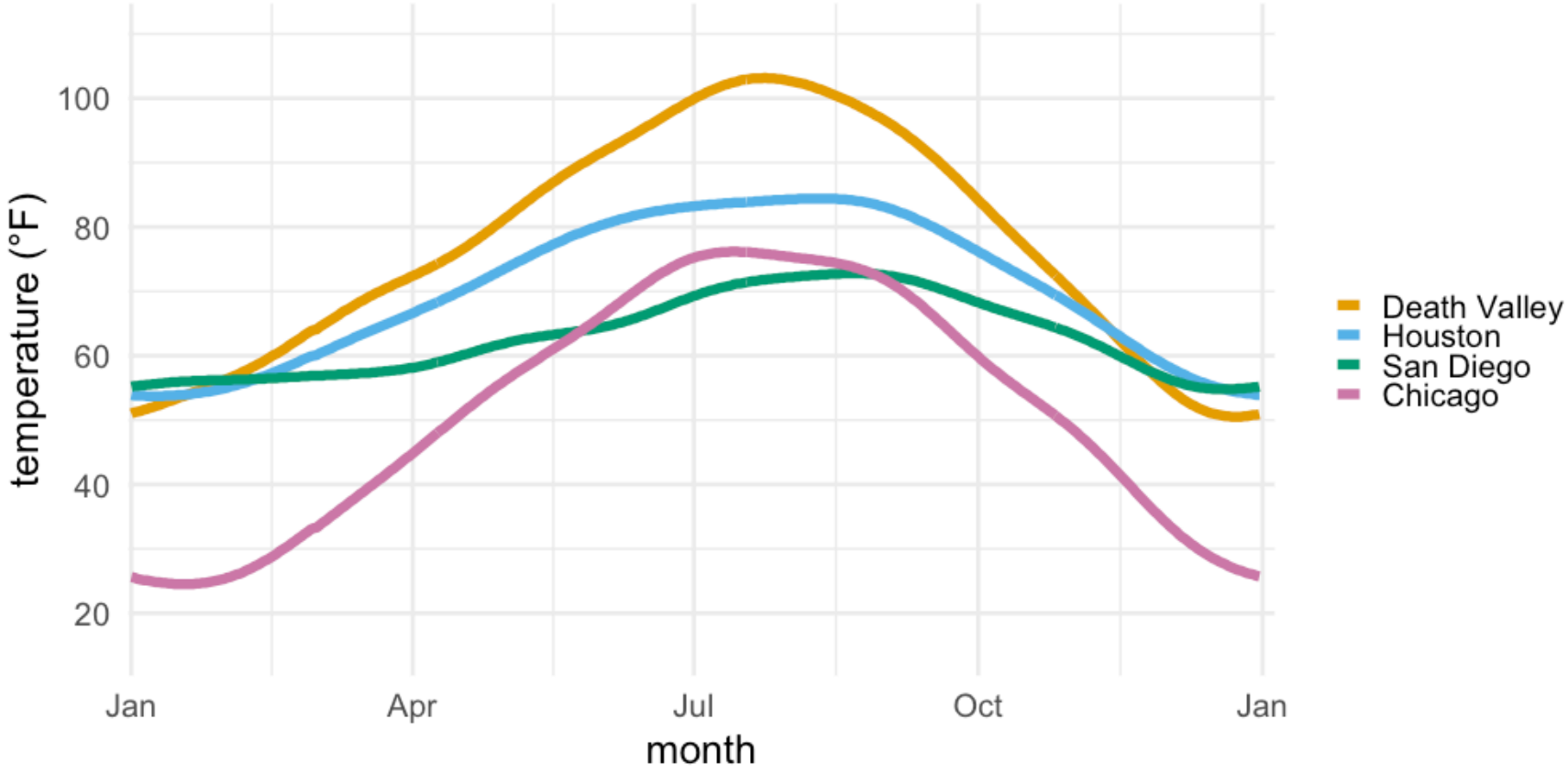
> A scale defines a unique mapping between data and aesthetics. Importantly, a scale must be one-to-one, such that for each specific data value there is exactly one aesthetics value and vice versa. If a scale isn't one-to-one, then the data visualization becomes ambiguous.

- Which data values correspond to specific aesthetic values?

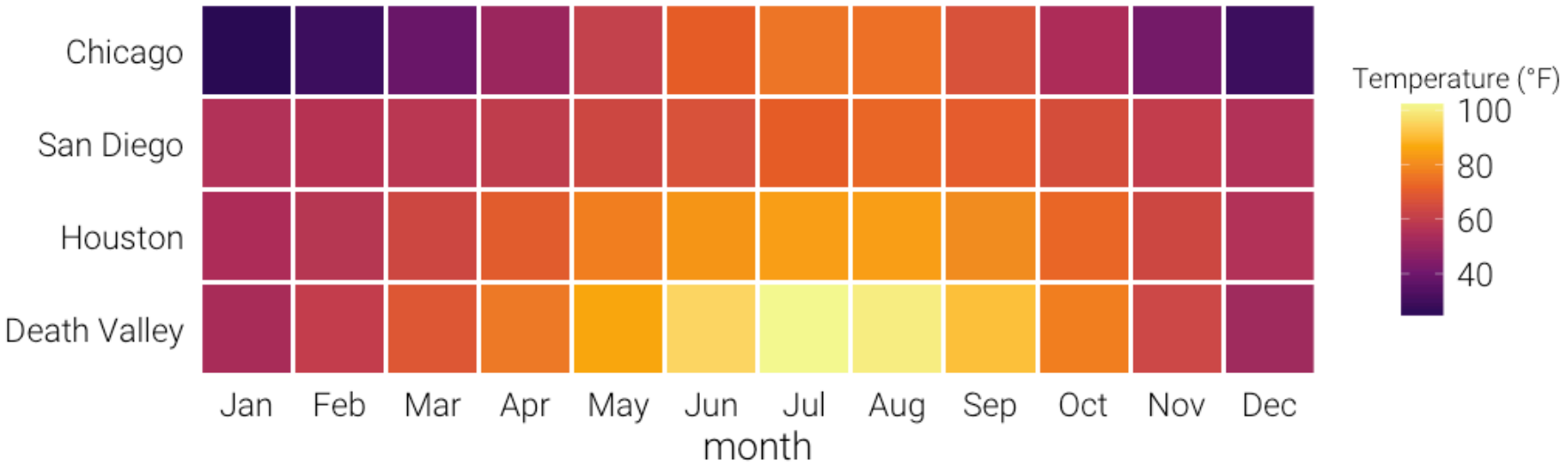
Basic Scales



Putting it to practice

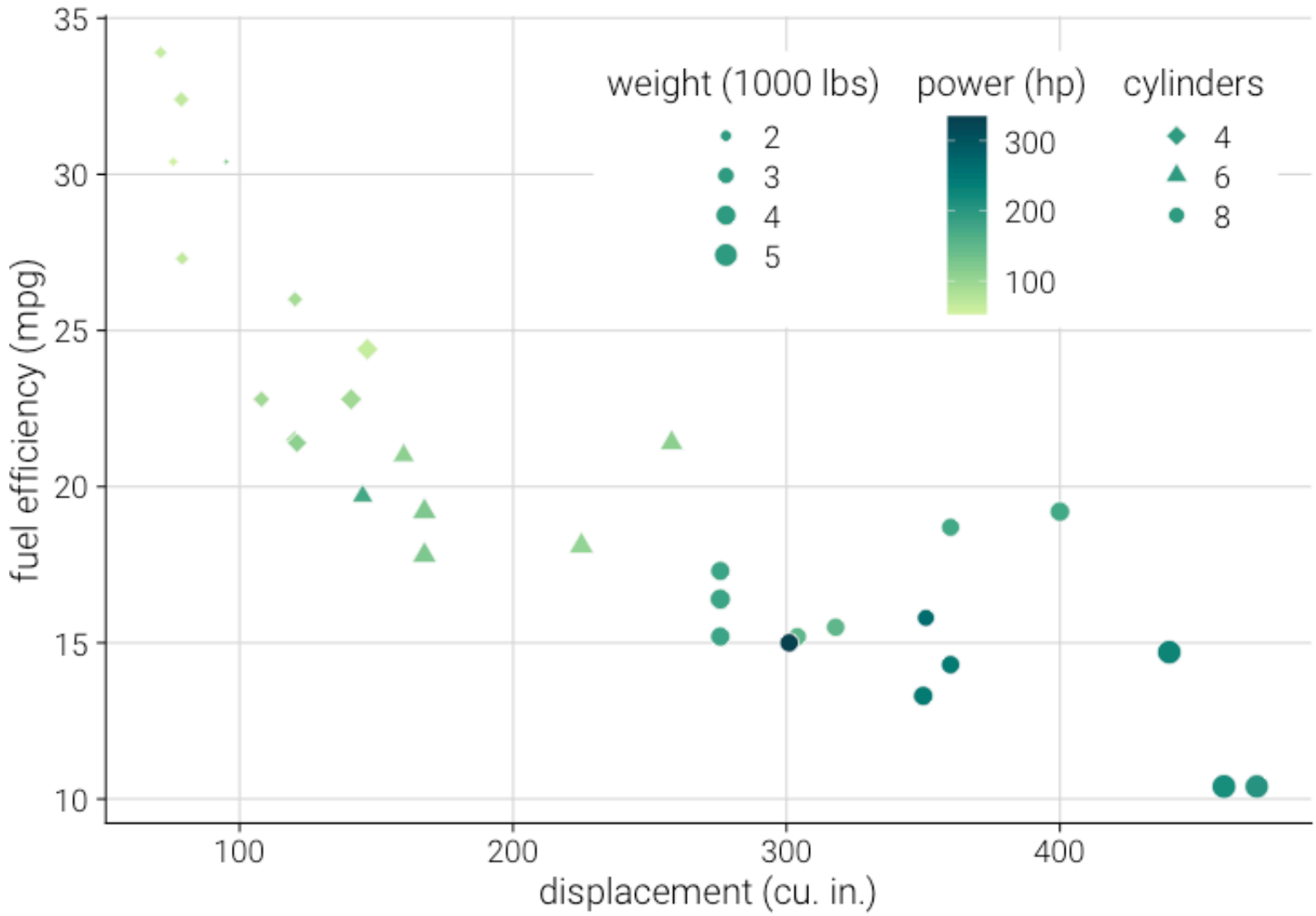


Alternative representation

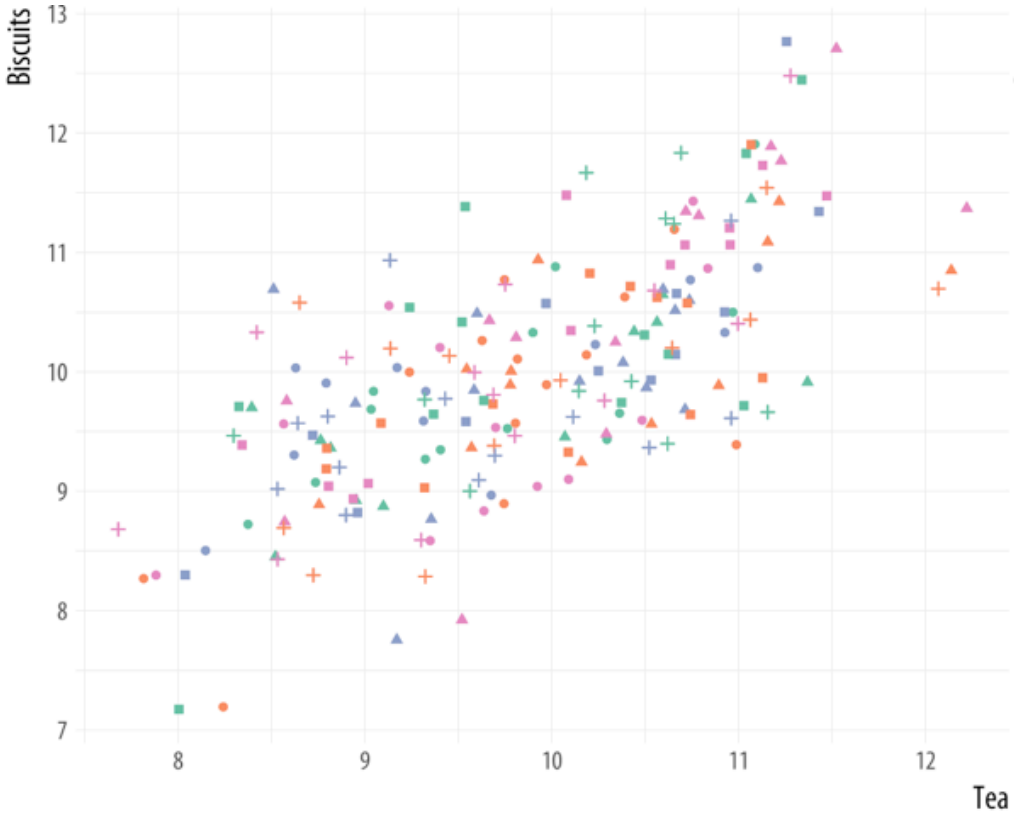


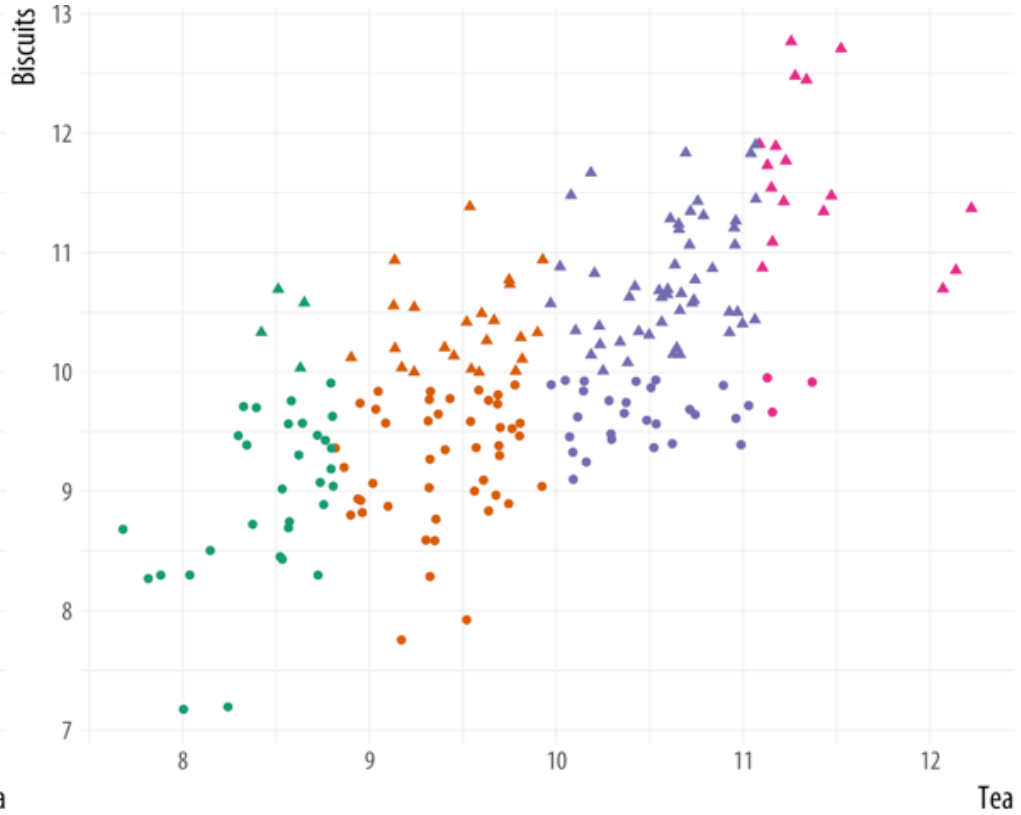
Comparison

- Both represent three scales
  - Two position scales (x/y axis)
  - One color scale (categorical for the first, continuous for the second)
- More scales are possible


fuel efficiency (mpg)
35
30
25
20
15
10
100
200
300
400
displacement (cu. in.)
weight (1000 lbs)
2
3
4
5
power (hp)
300
200
100
cylinders
4
6
8

Additional scales can become lost without high structure in the data





Data ink ratio

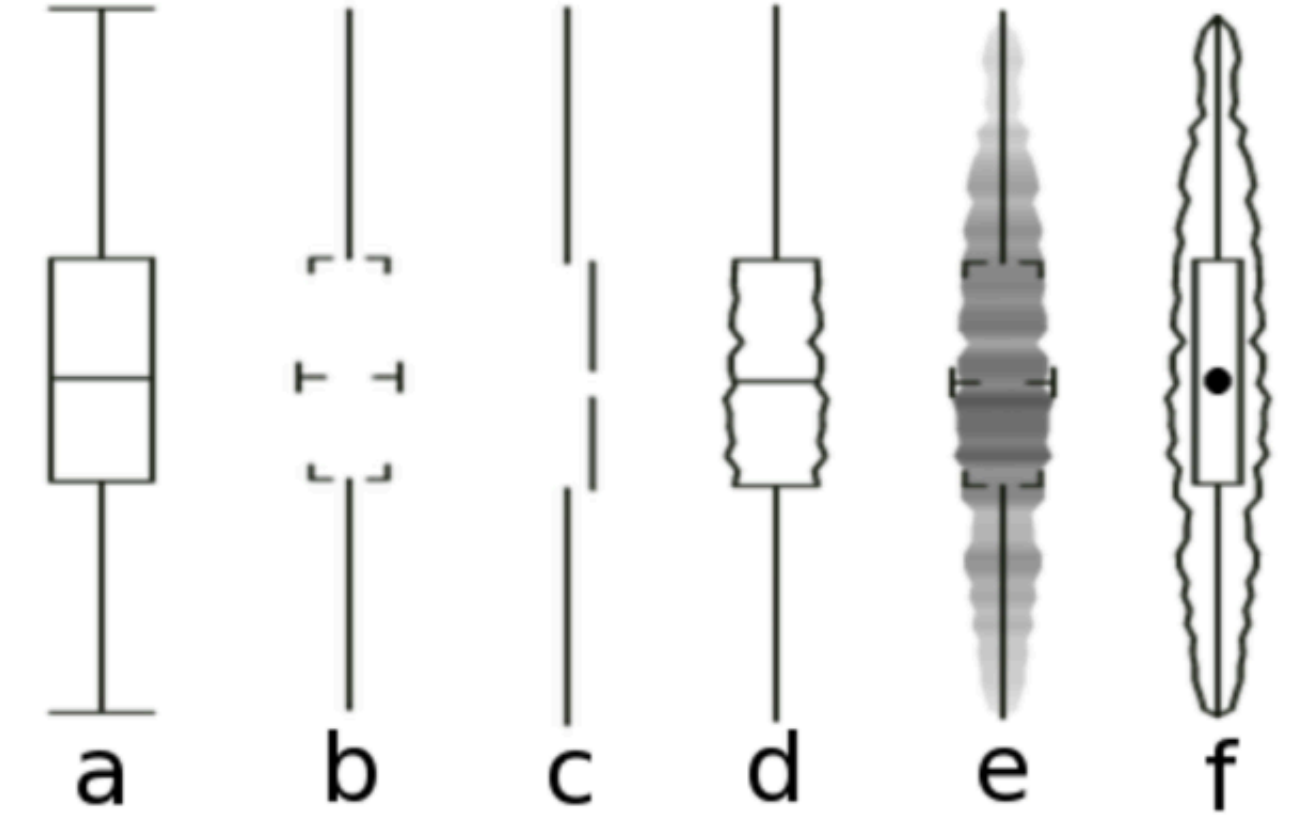
What is it?

> ## Above all else, show the data

-Edward Tufte

- Data-Ink Ratio = Ink devoted to the data / total ink used to produce the figure
- Common goal: Maximize the data-ink ratio

Example



- First thought -might be Cool!

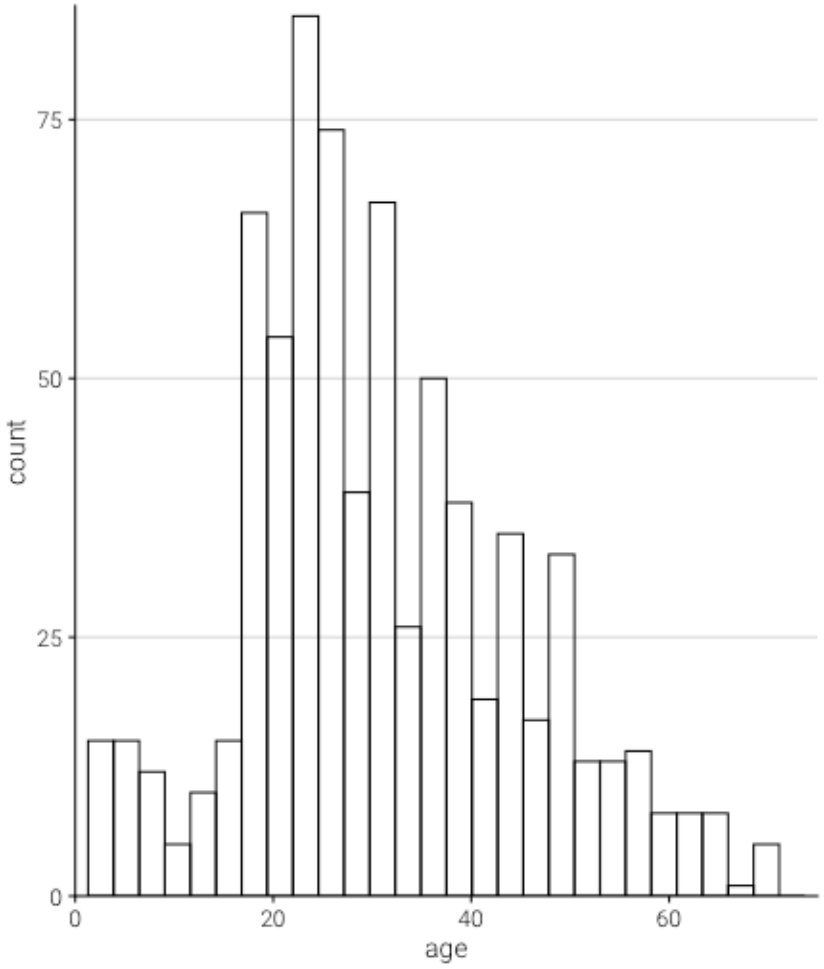

NOT SO FAST, MY FRIEND!

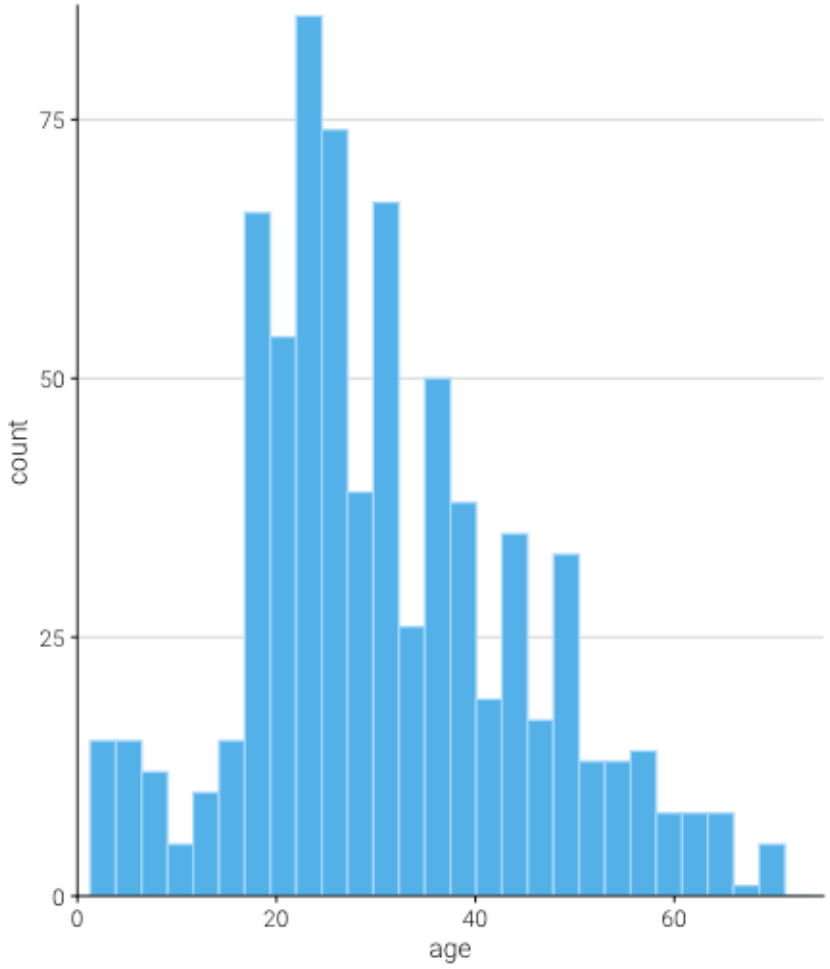
Minimize cognitive load

- Empirically, Tufte's plot was **the most difficult** for viewers to interpret.
- Visual cues (labels, gridlines) reduce the data-ink ratio, but can also reduce cognitive load.

Another example

Which do you prefer?

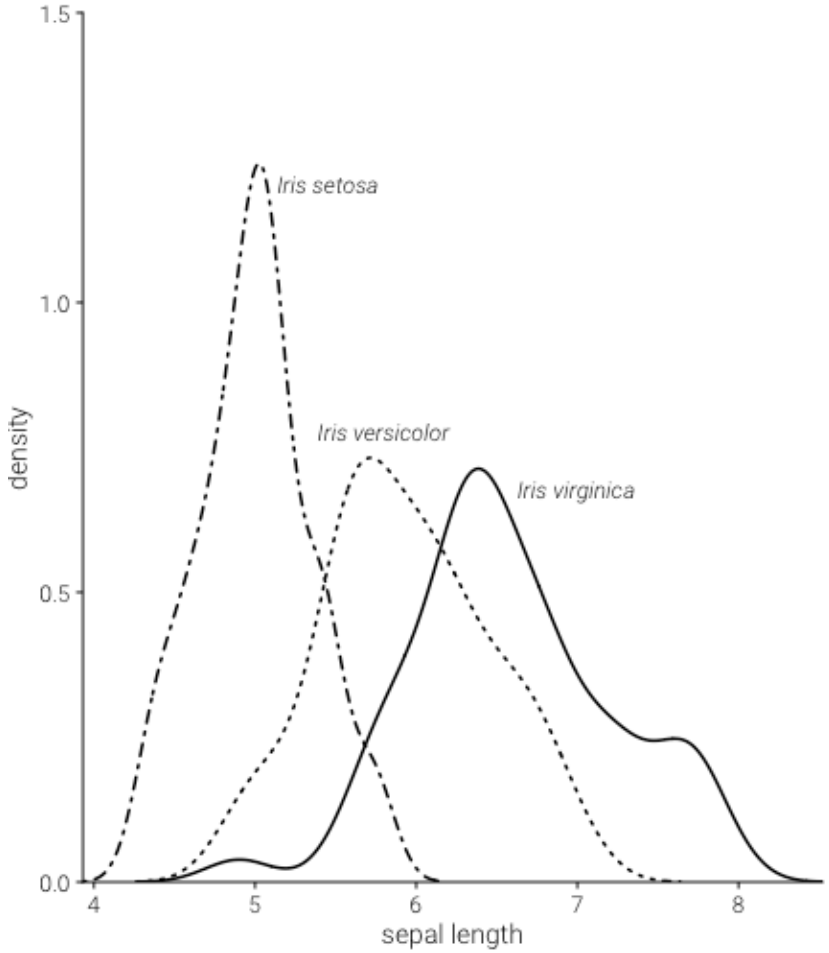


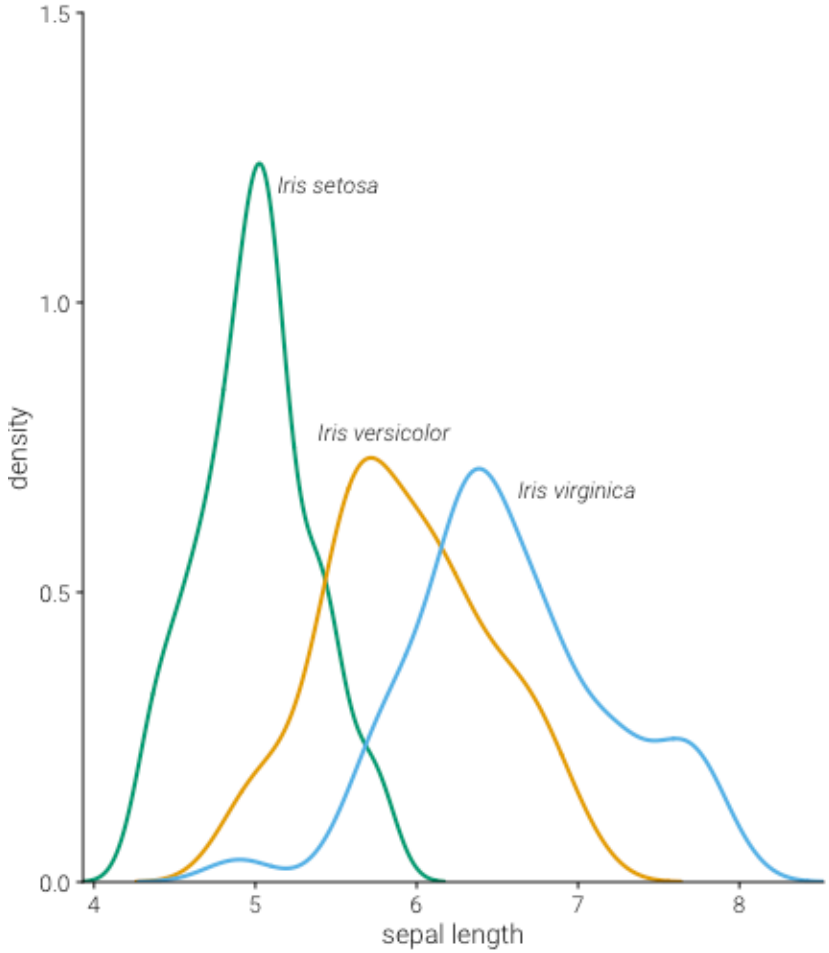


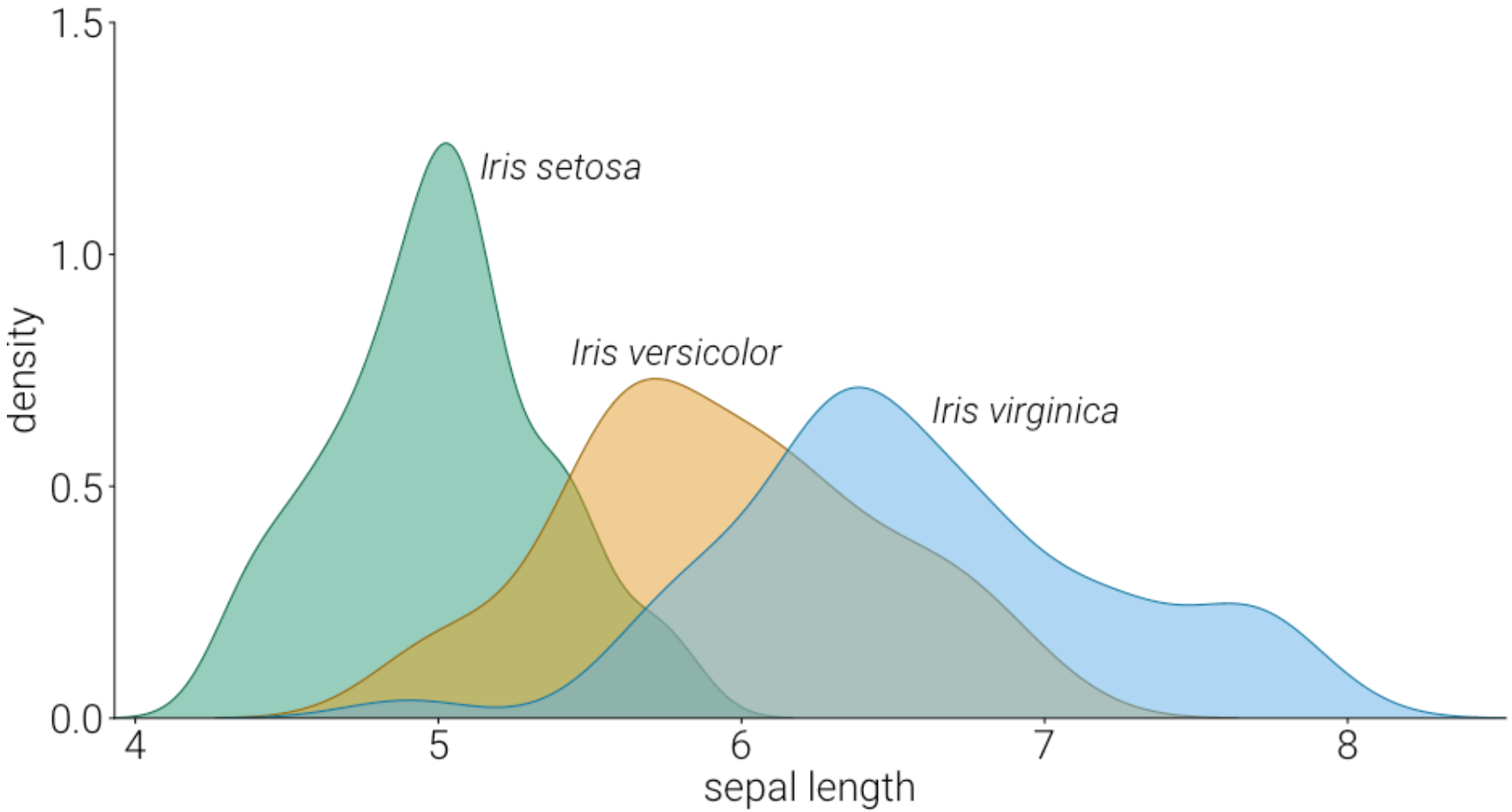
Advice from Wilke

> Whenever possible, visualize your data with solid, colored shapes rather than with lines that outline those shapes. Solid shapes are more easily perceived, are less likely to create visual artifacts or optical illusions, and do more immediately convey amounts than do outlines.

Another example






1.5
1.0
0.5
0.0
density
Iris setosa
Iris versicolor
Iris virginica
4
5
6
7
8
sepal length

What about plots like these?

The takeaway?

- It can often be helpful to remove "chart junk"
  - Remove background
  - Unnecessary frills
  - Certainly don't use 3D when it's not clearly warranted

But...

- Infographics can often be more memorable

Quick/easy compromise

In some cases, it may be easy and more memorable to use glyphs instead of points or squares

- Install packages

```
install.packages("extrafont")
devtools::install_github("wch/extrafontdb")
devtools::install_github("wch/Rttf2pt1")
devtools::install_github("hrbrmstr/waffle")
```

- Create data

```
parts <- c(`Un-breached\nUS Population` = (318 - 11 - 79),
           `Premera` = 11,
           `Anthem` = 79)
```

Basic plot

```
library(waffle)
waffle(parts,
       rows = 8,
       colors = c("#969696", "#1879bf", "#009bda"))
```

Glyph plot

- Download and install `fontawesome-webfont.ttf` on your machine locally (see here)
- Import new fonts (including glyphs, via font awesome)

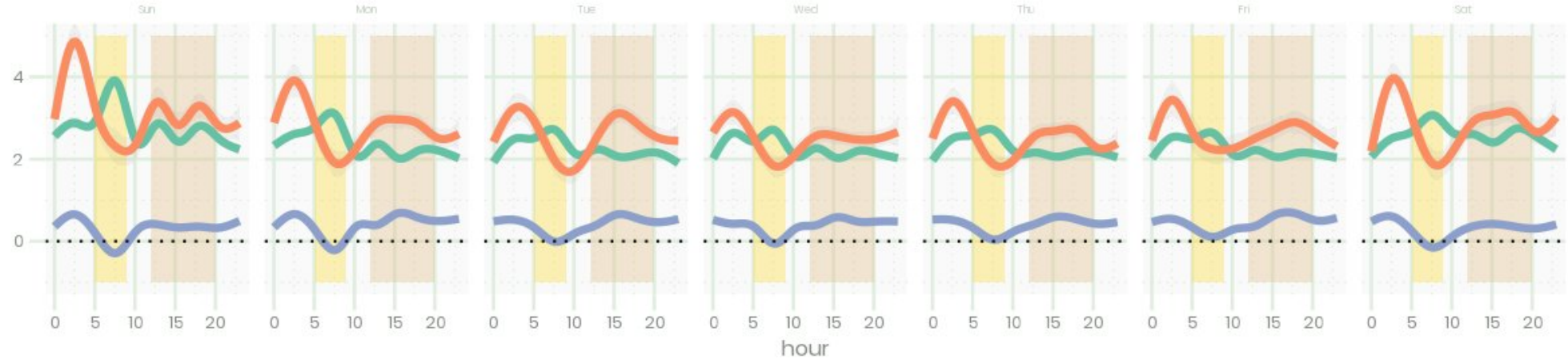
```
library(extrafont)
font_import()
loadfonts()
```

```
waffle(parts/10,
       rows = 3,
       colors = c("#969696", "#1879bf", "#009bda"),
       use_glyph = "medkit")
```


Un-breached
US Population
Premera
Anthem

#rstats aren't early morning grinders

the average number of retweets and hashtags per #rstats tweet, by day and time



In general, more hashtags = more retweets. Early morning tweets usually contain more hashtags than tweets later in the day. Fewer people are on twitter at this time, so the apparent effectiveness of hashtags at provoking retweets is reduced. This begs the question: Why do people use more hashtags in the morning?

Hashtags used in the mid-afternoon, when most people are on twitter, are almost twice as effective at provoking retweets! Also, while most days see a single pronounced increase in tweets, Sunday afternoon sees two distinct bumps in #rstats tweets.

A #TidyTuesday adventure
Data from rtweet.info
Analysis @Tanner Koomar
Design @Tanner Koomar
https://github.com/tkoomar/tidytuesday/blob/master/work/2019-01-01.md

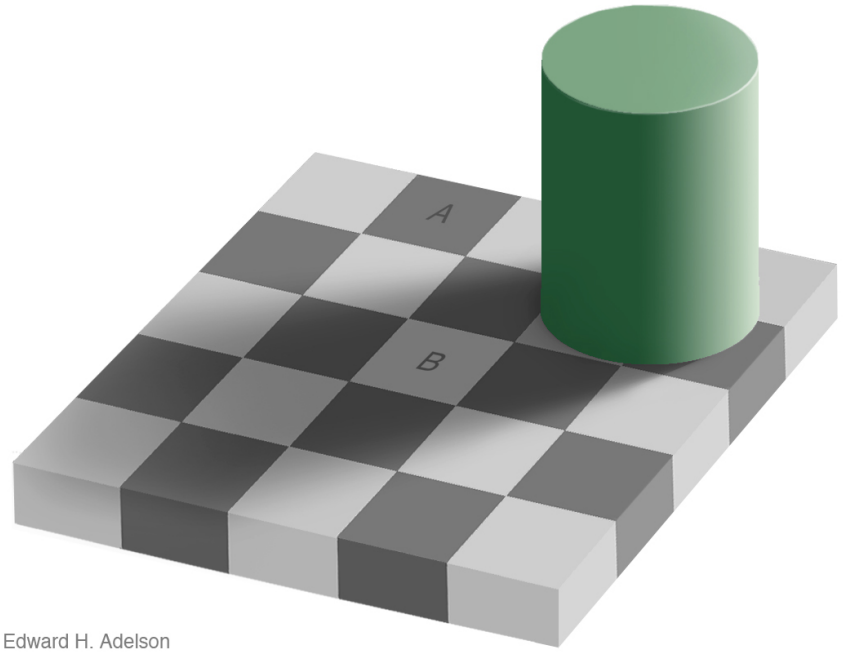
You can create them!

- Create plots
- Use illustrator or similar to put them together
- Add some annotations
- Consider using glyphs for greater memory

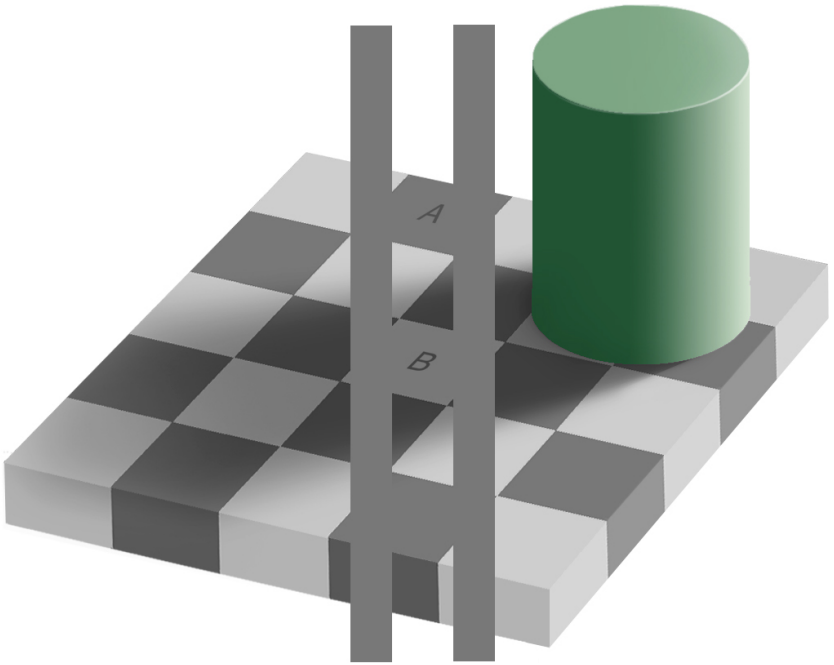
Alternatively

Consider pagedown!

More visual properties

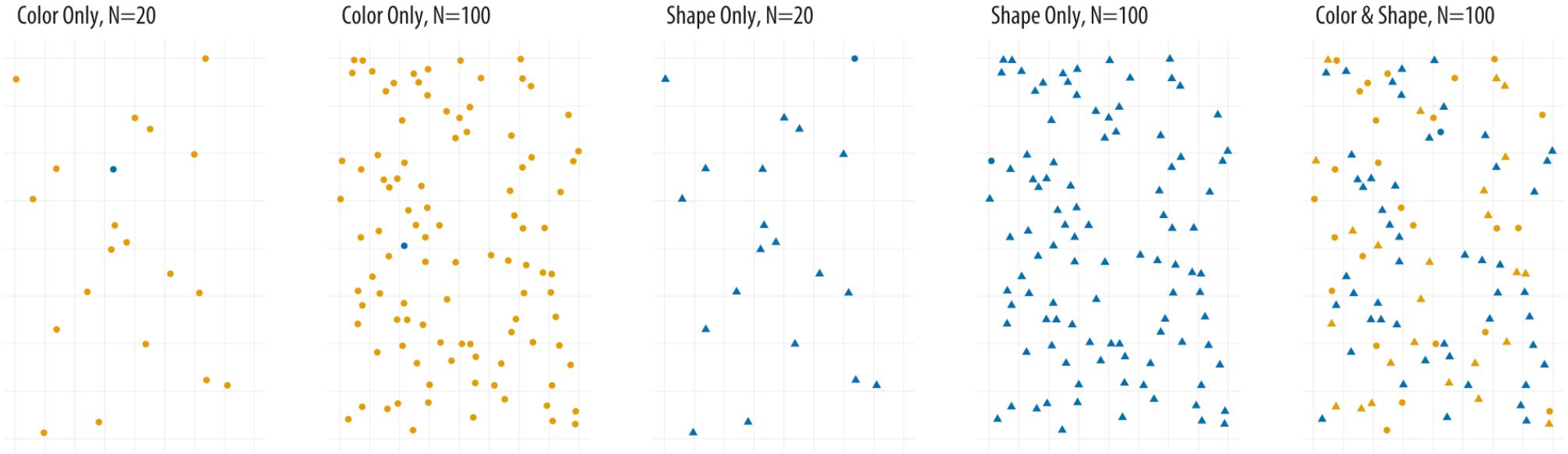


Edward H. Adelson

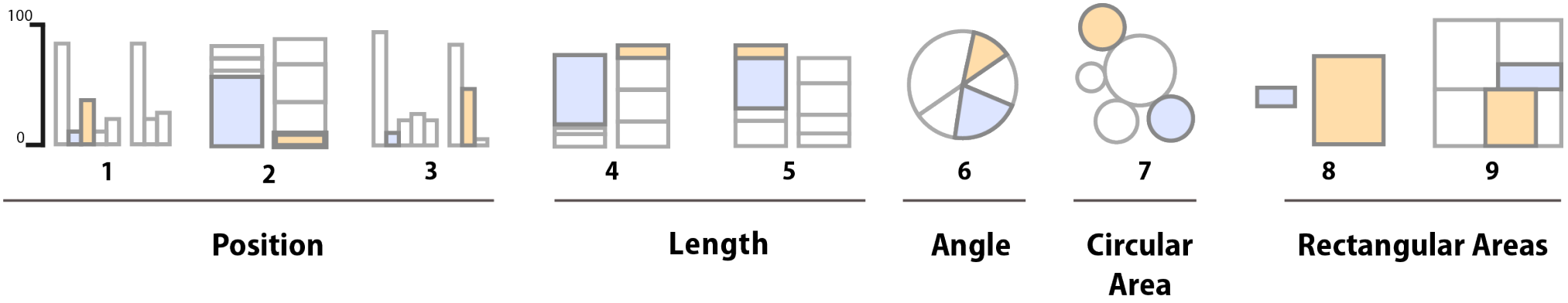


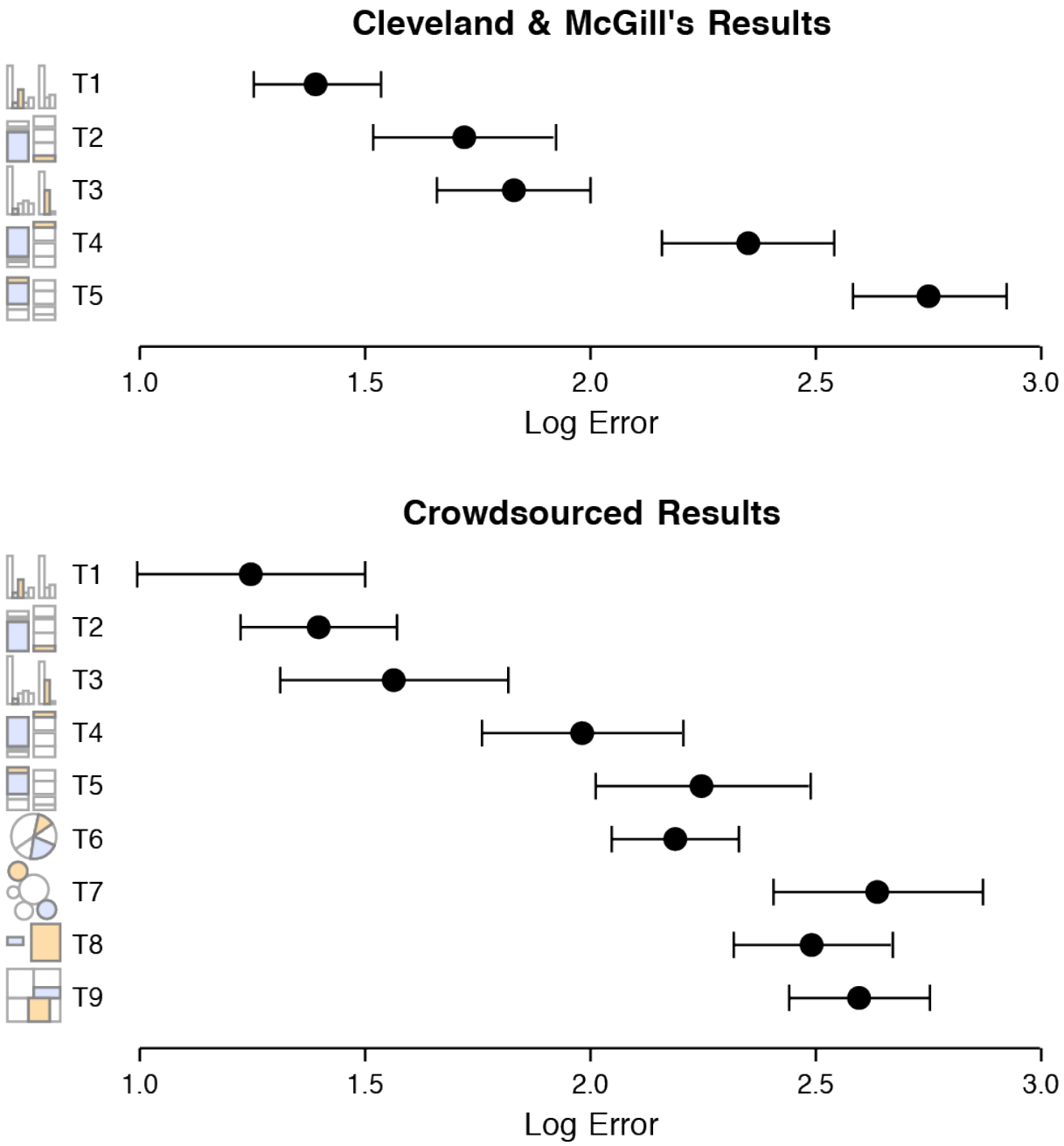
What "pops"?

Where's the blue circle in each plot?

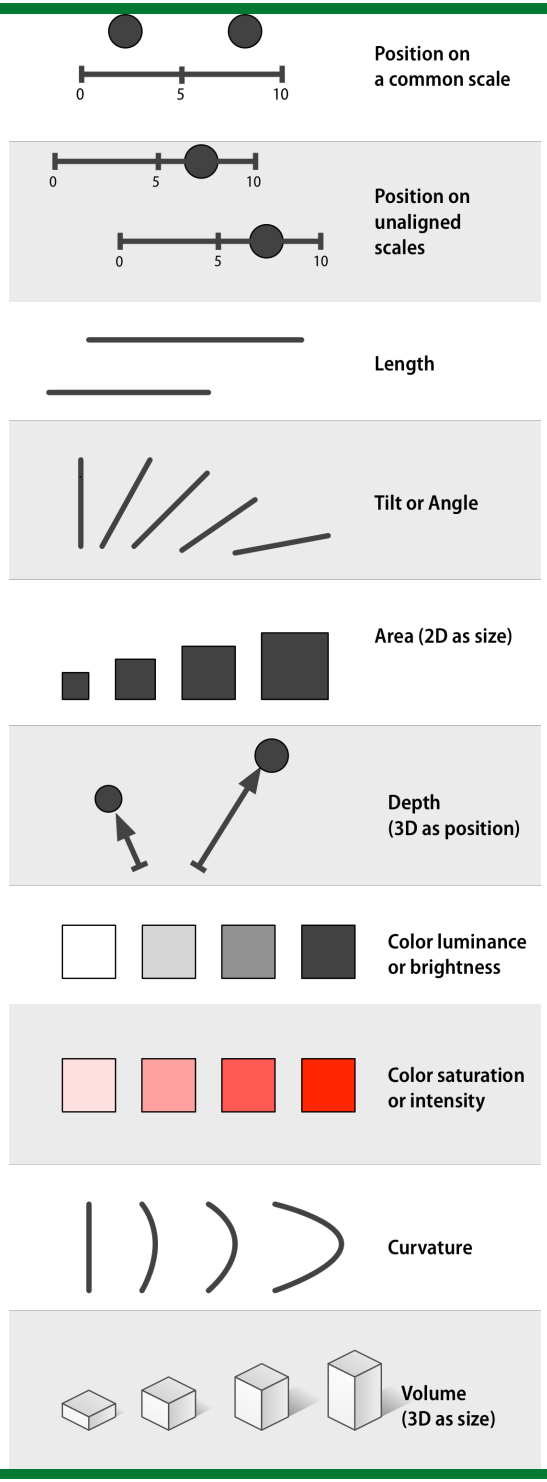


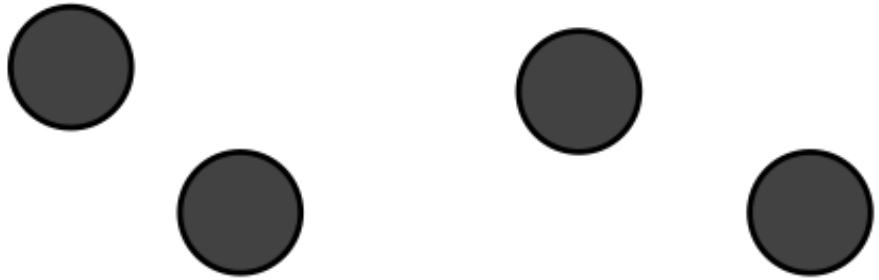
What are we good at perceiving?



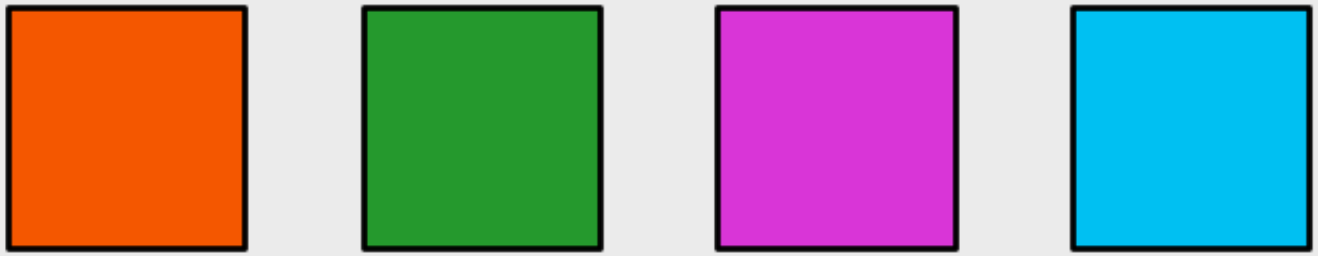

Cleveland & McGill's Results
T1
T2
T3
T4
T5
1.0
1.5
2.0
2.5
3.0
Log Error
Crowdsourced Results
T1
T2
T3
T4
T5
T6
T7
T8
T9
1.0
1.5
2.0
2.5
3.0
Log Error

Ordered data mappings: Ranked

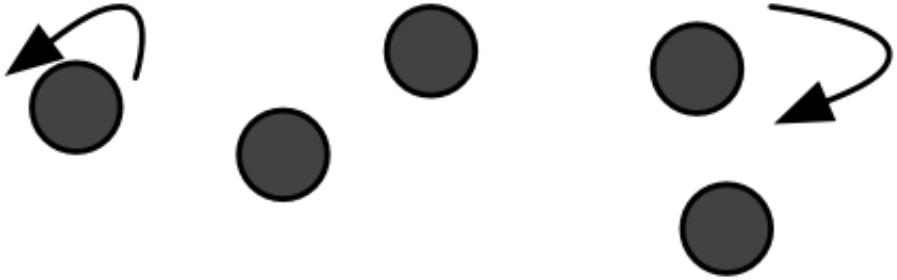




Position in space



Color hue



Motion

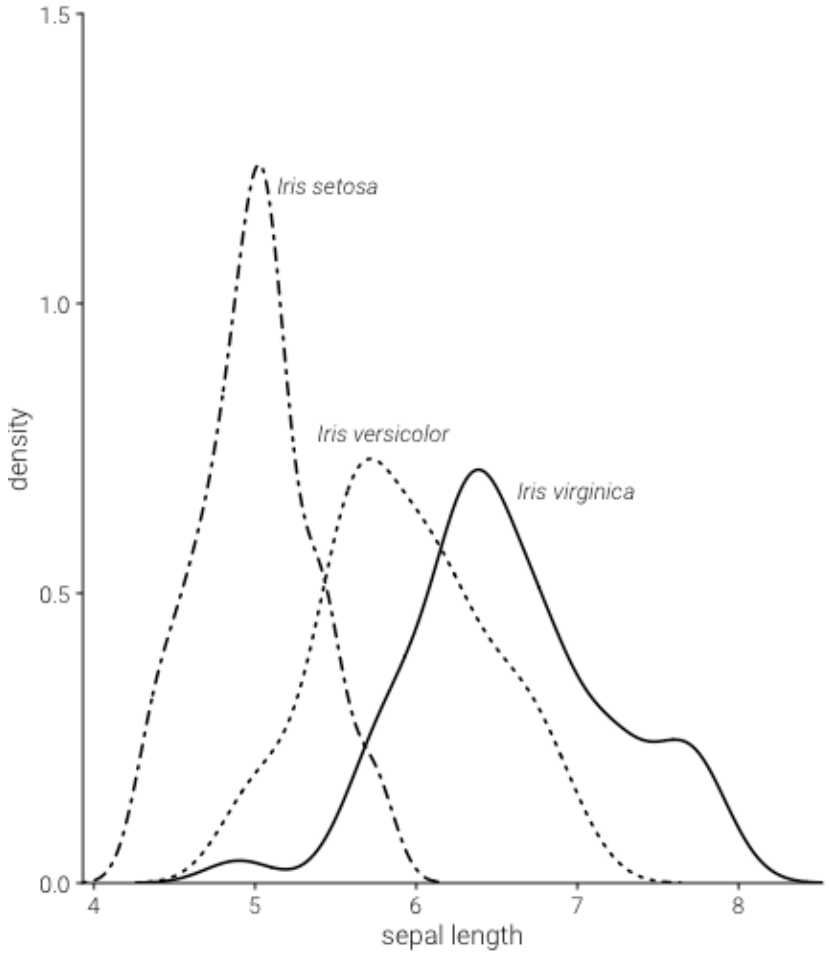
Shape

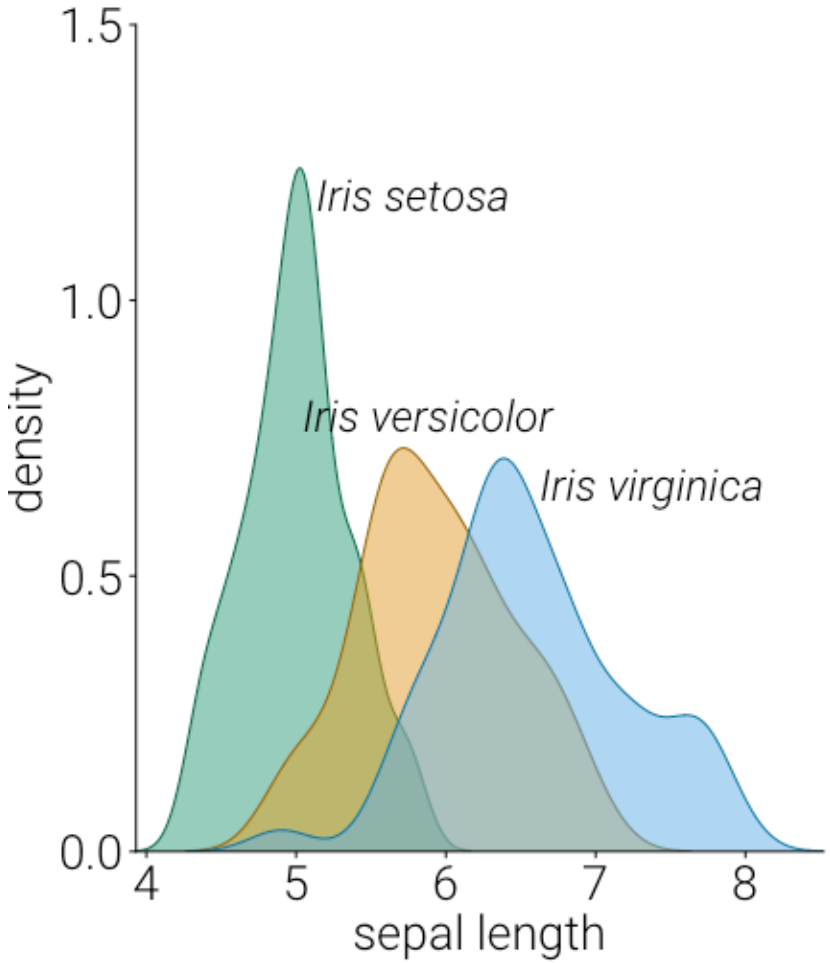
Unordered data mappings: Ranked

Some things to avoid

Line drawings

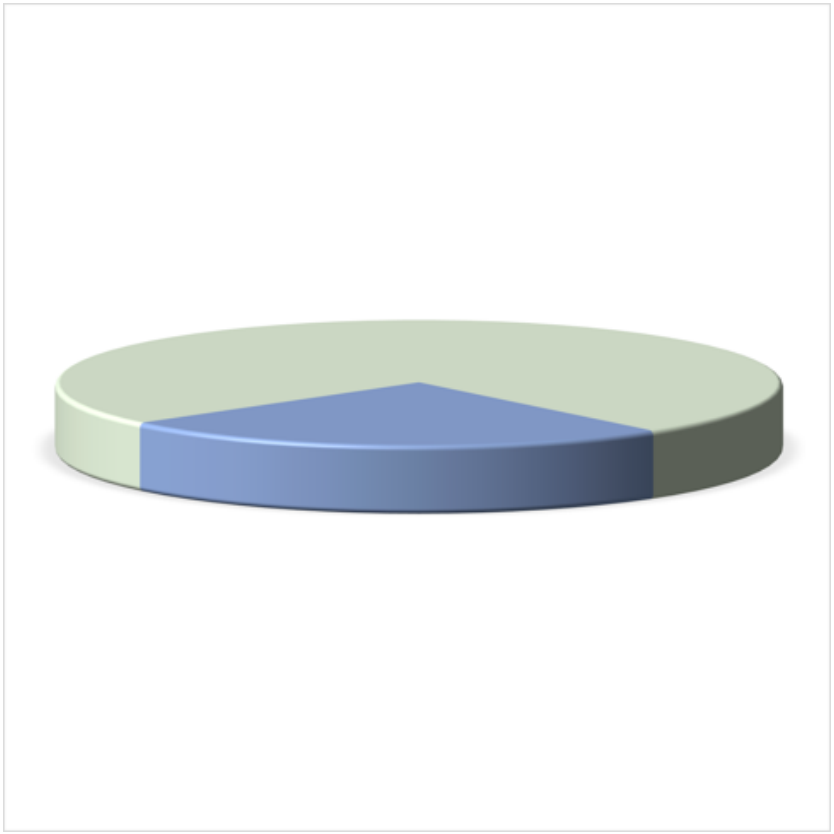
As discussed earlier

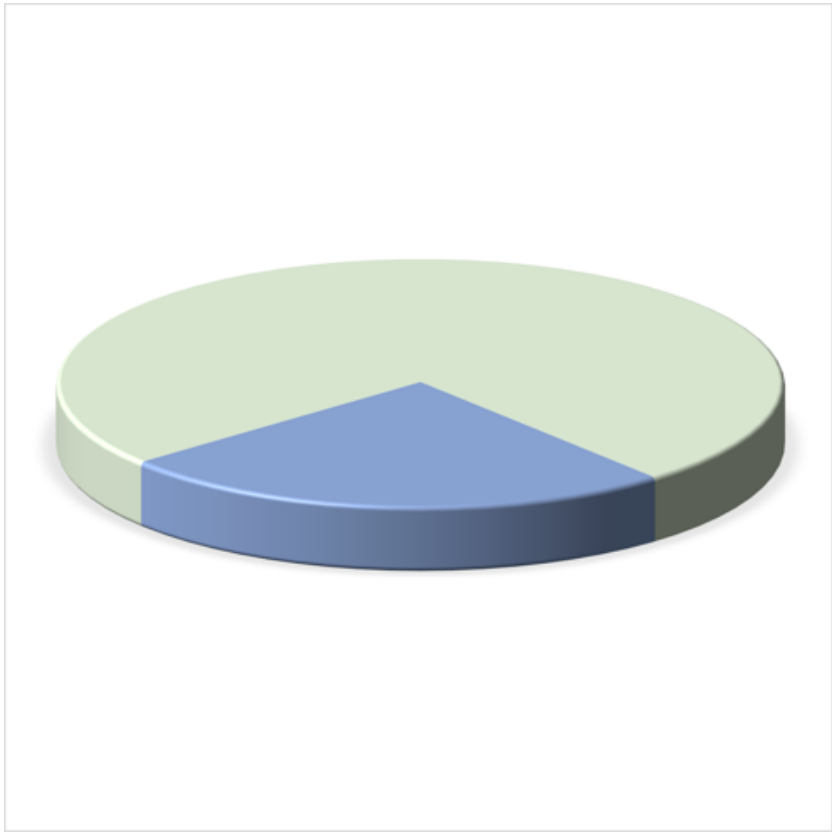




Much worse

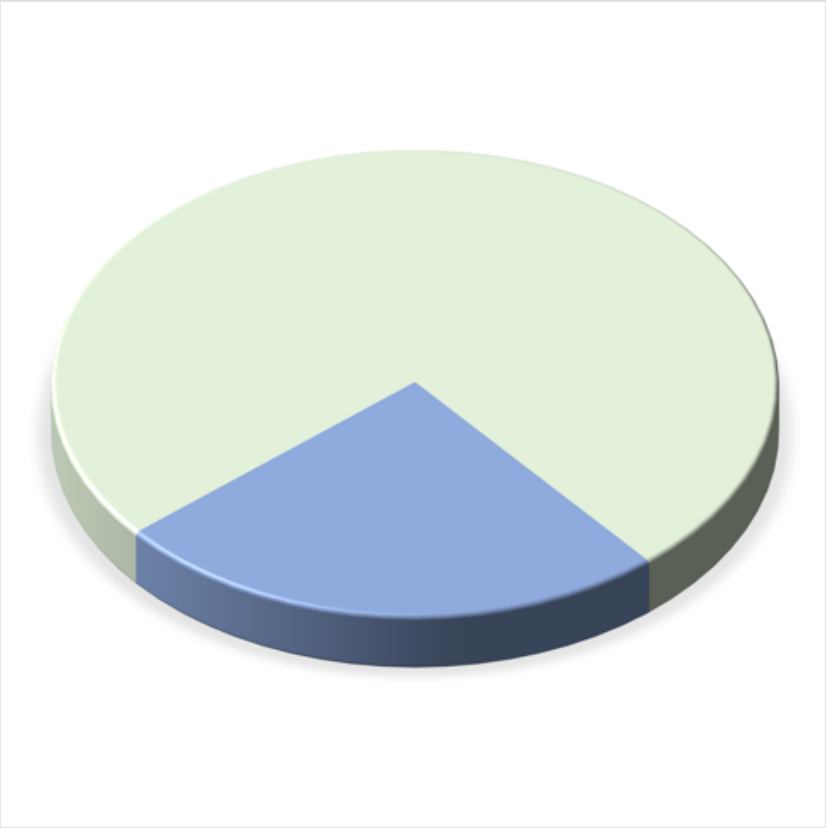
Unnecessary 3D

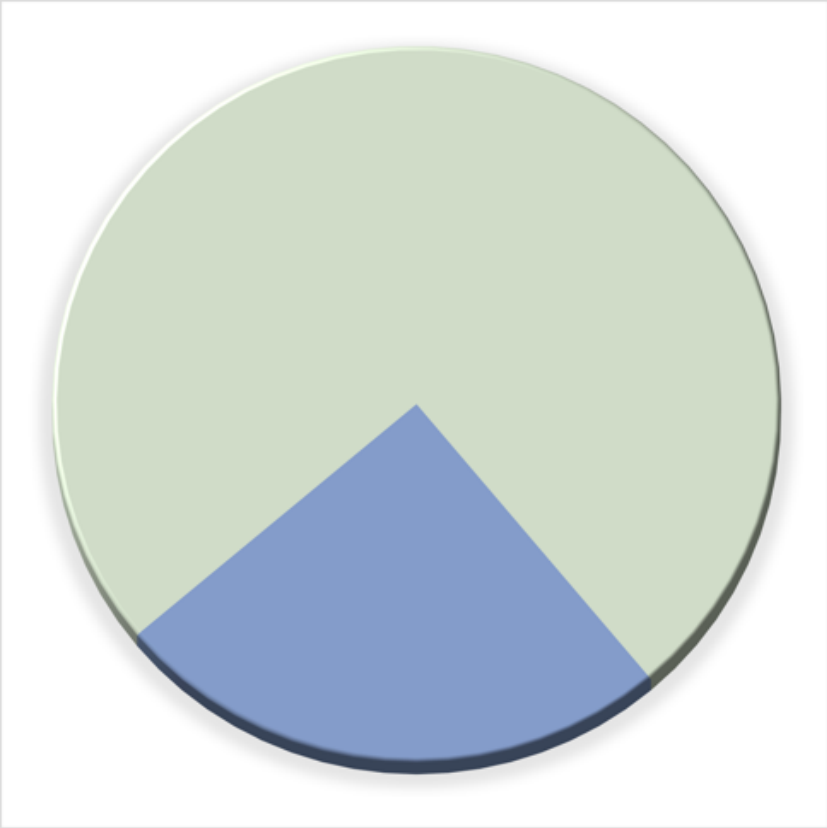




Much worse

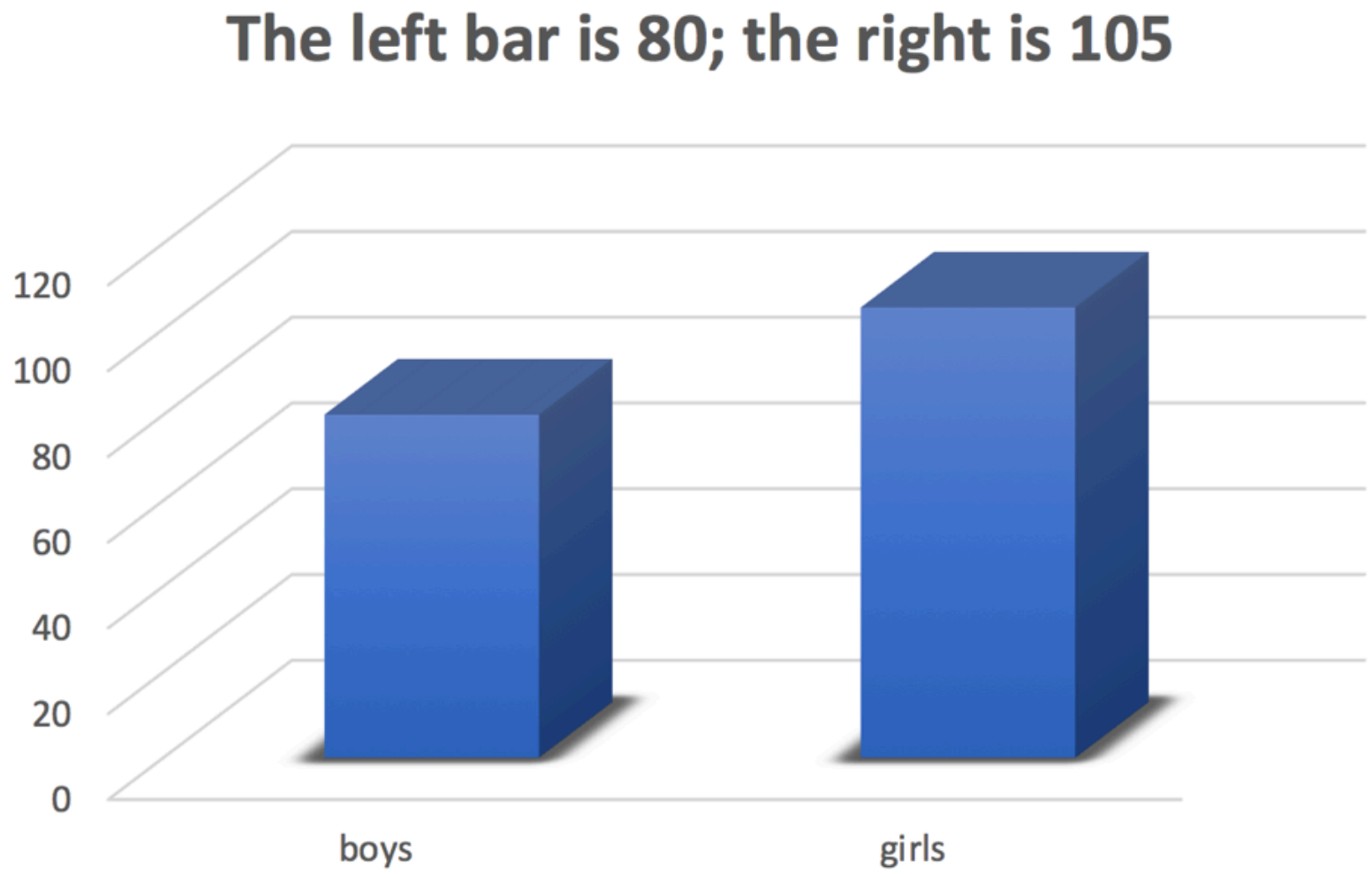
Unnecessary 3D





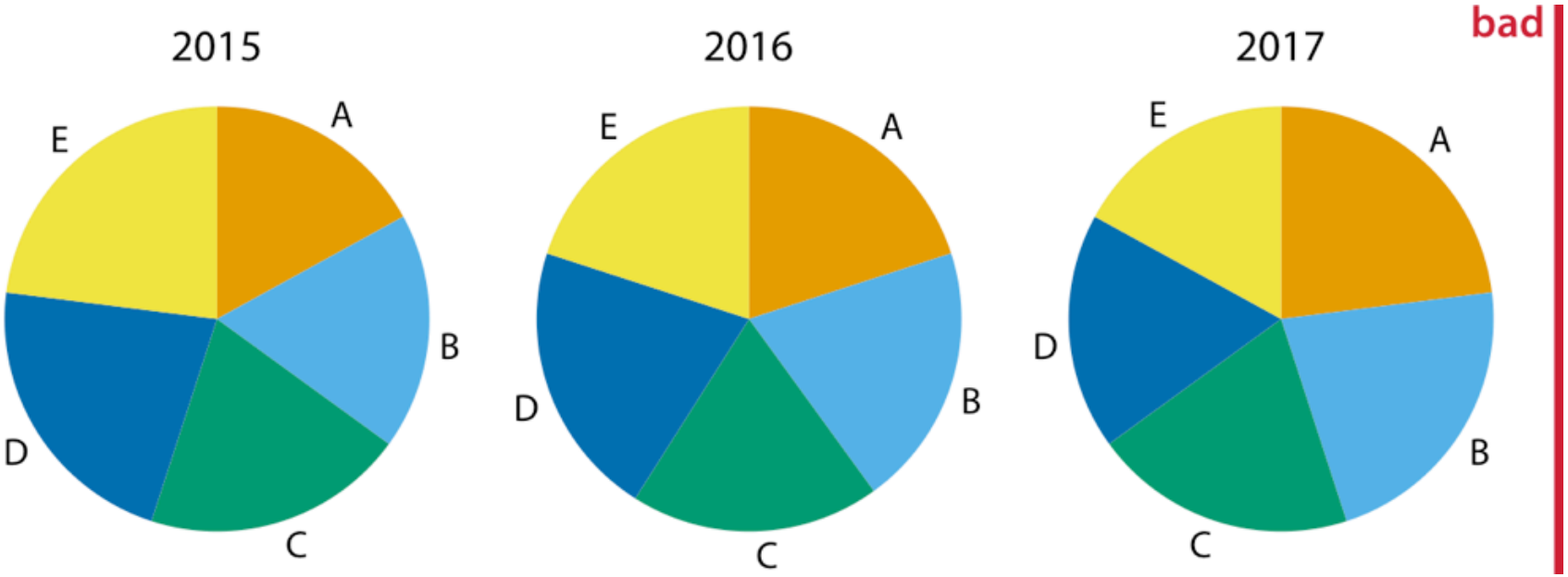
Horrid example

Used relatively regularly



Pie charts

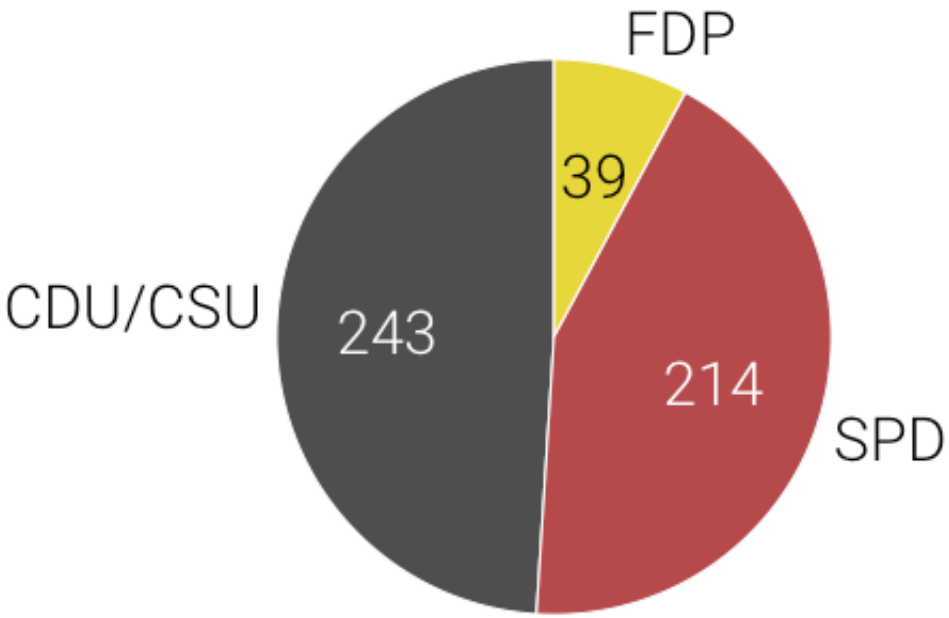
Especially w/lots of categories



Alternative representation

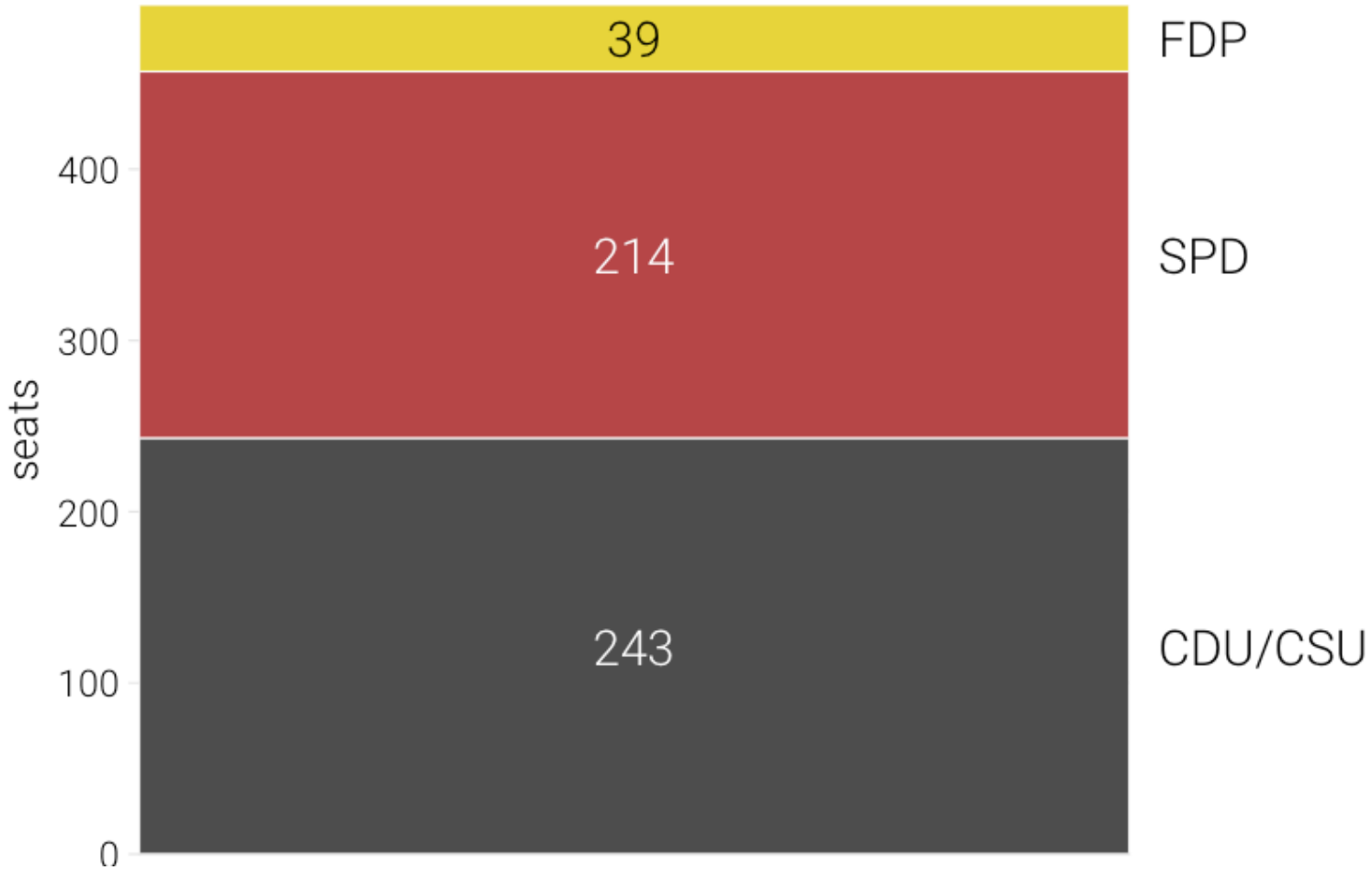
A case for pie charts

- n categories low,
- differences are relatively large
- familiar for some audiences

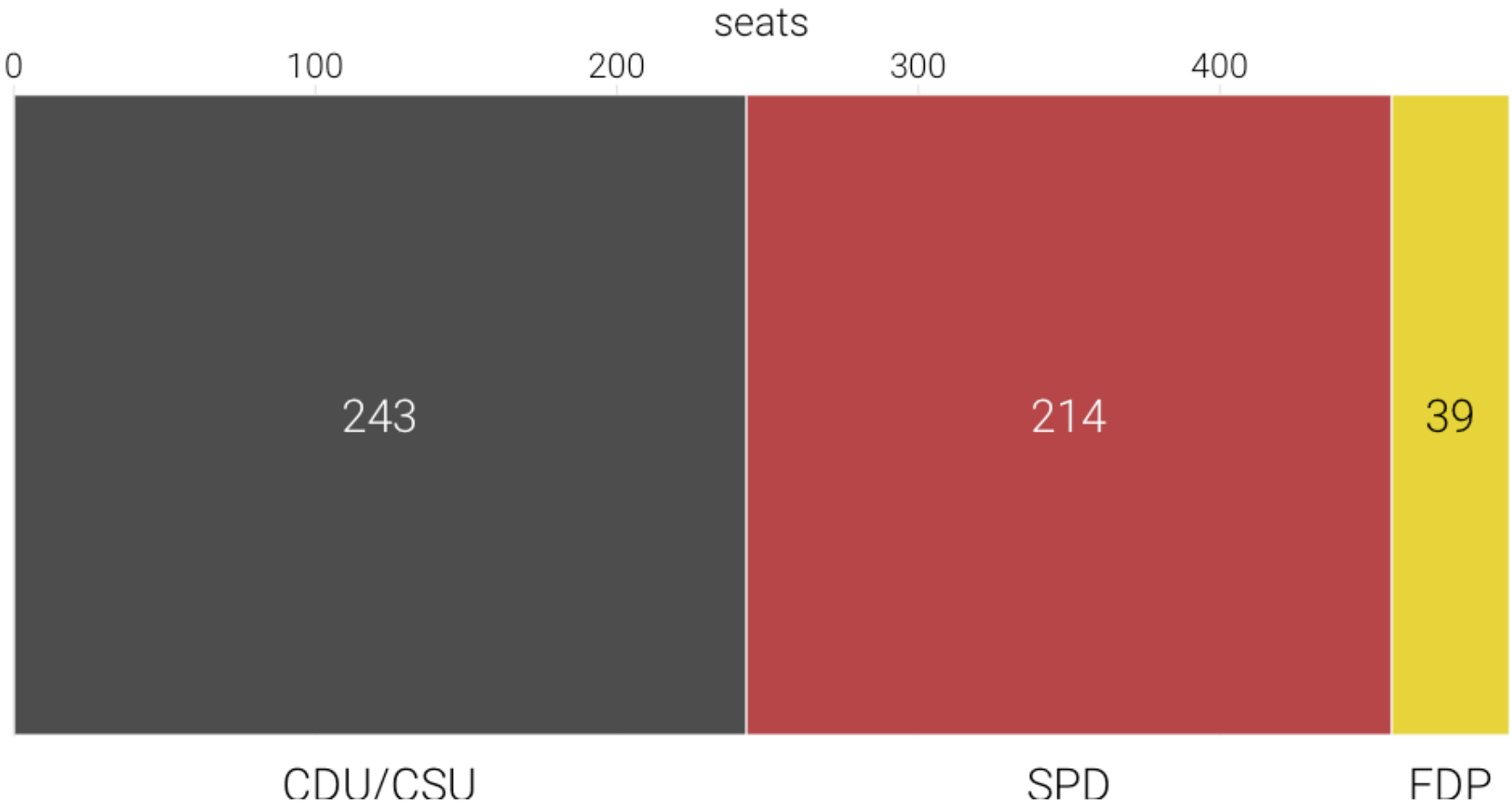


The anatomy of a pie chart

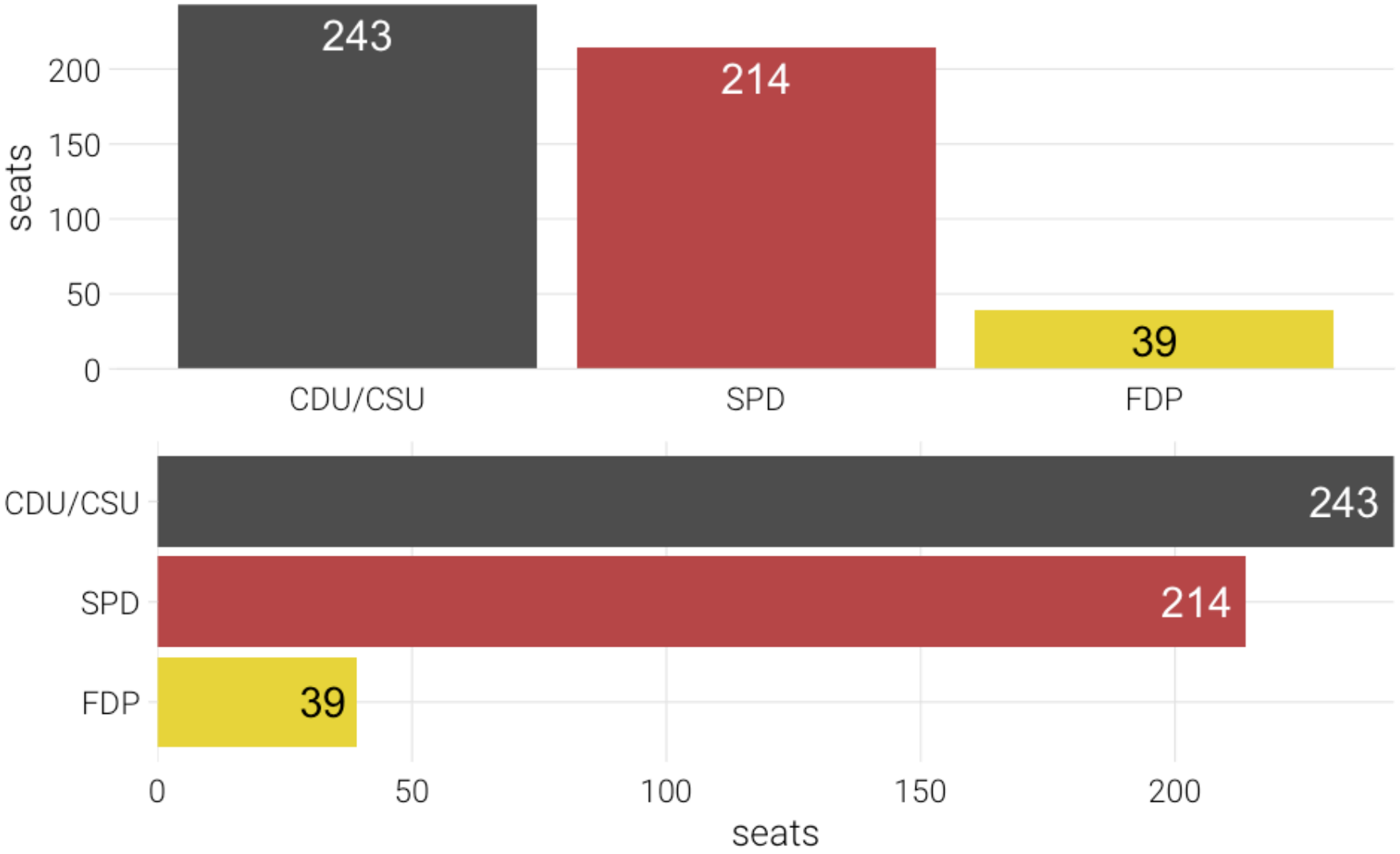
Pie charts are just stacked bar charts with a radial coordinate system



My preference

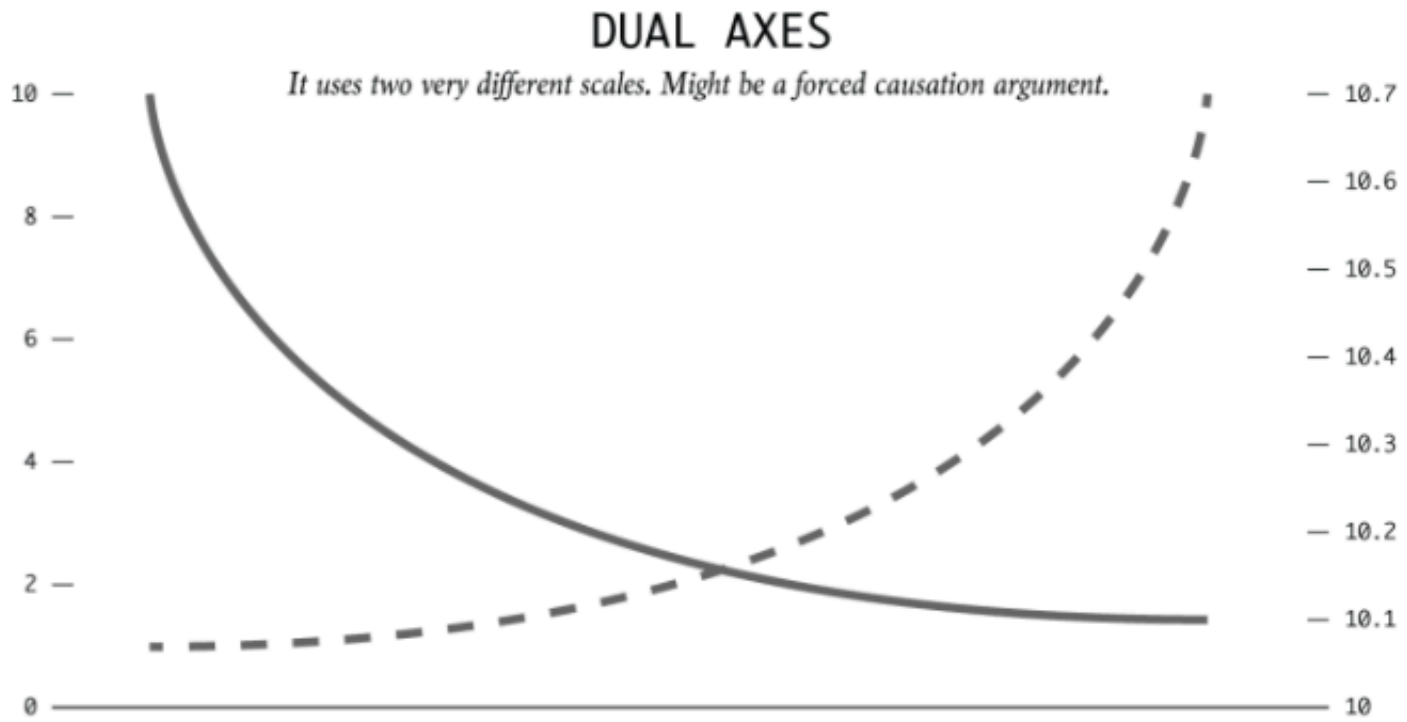


Or one of these



Dual axes

- One exception - if second axis is a direct transformation of the first
  - e.g., Miles/Kilometers, Fahrenheit/Celsius

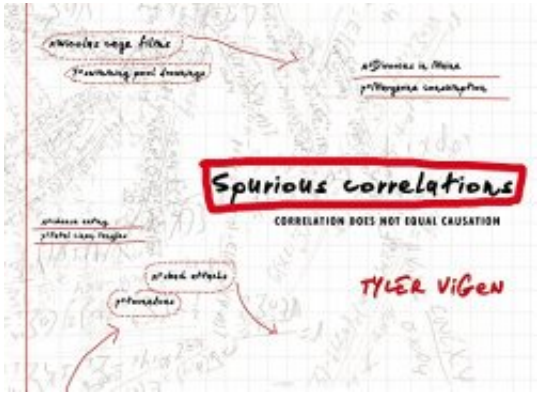


More examples

tylervigen.com

about | twitter | email | subscribe

Spurious correlations

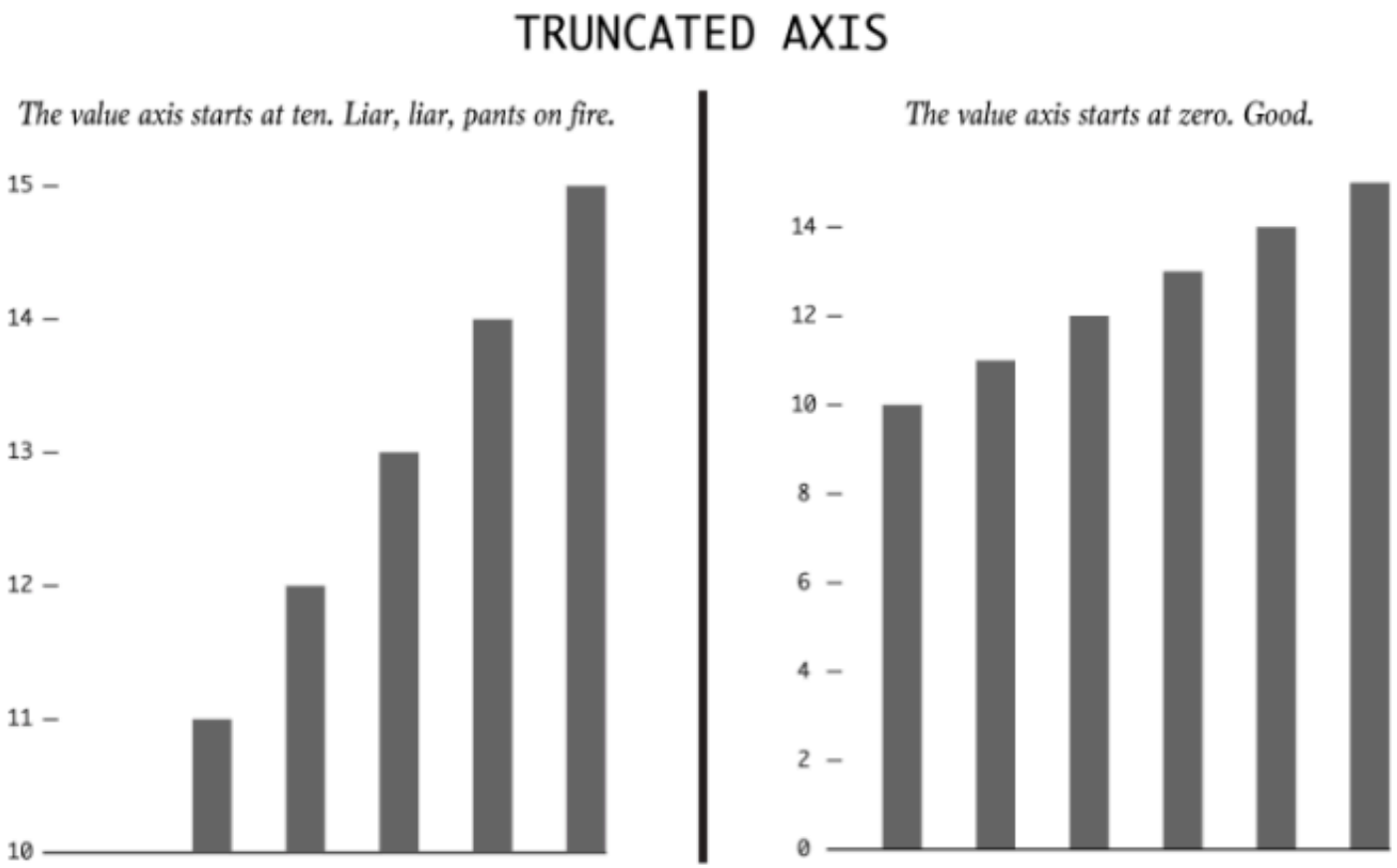


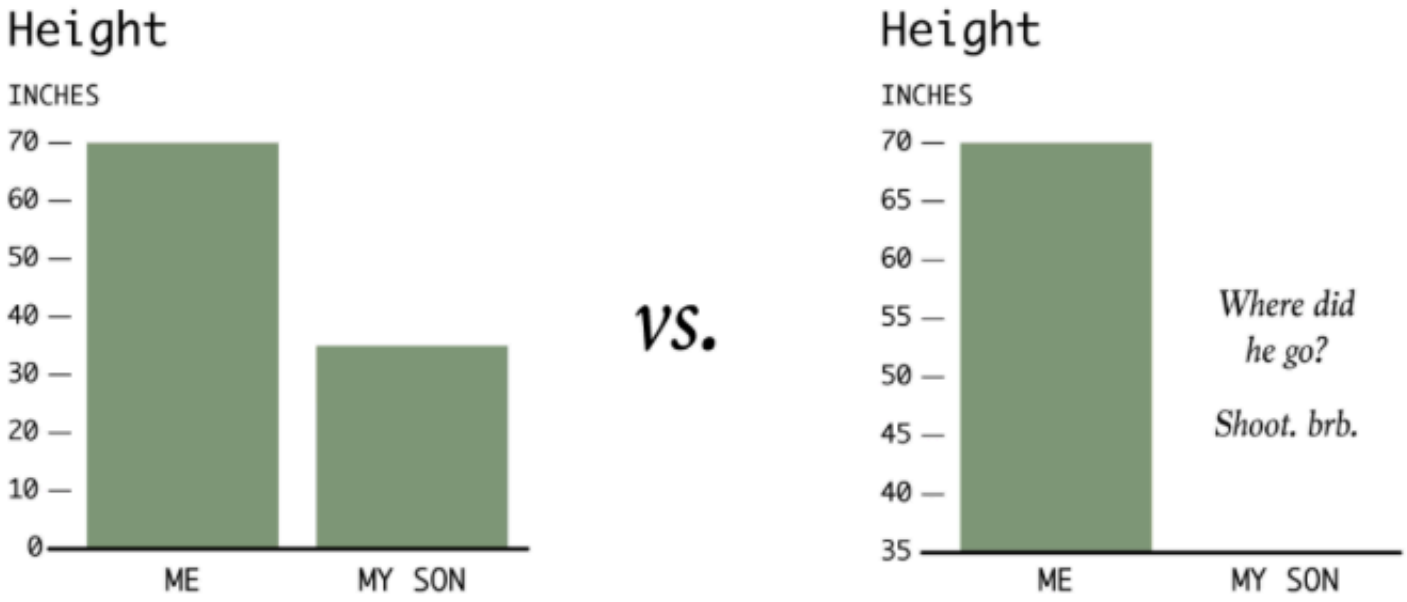
Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

Amazon | Barnes & Noble | Indie Bound

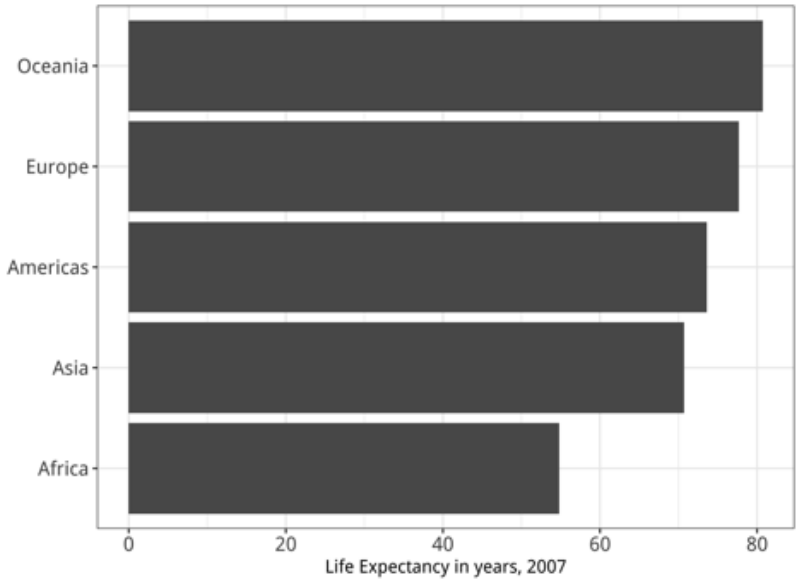
Truncated axes

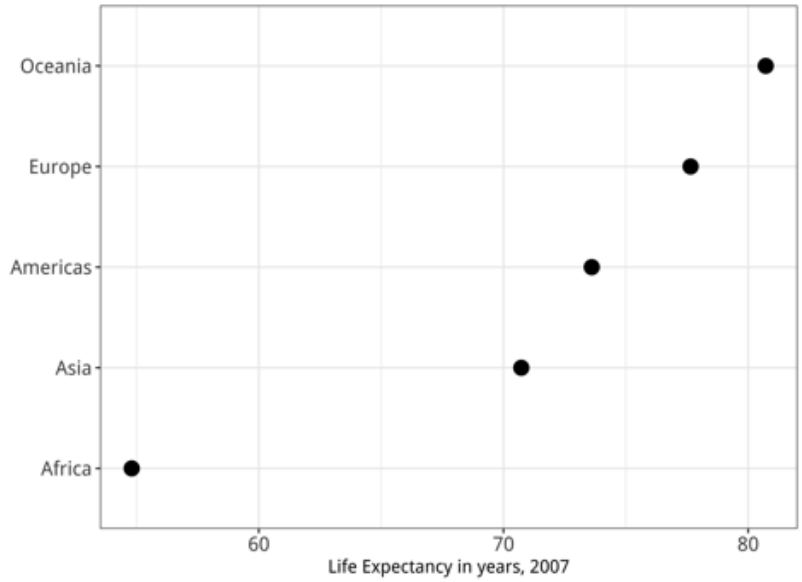


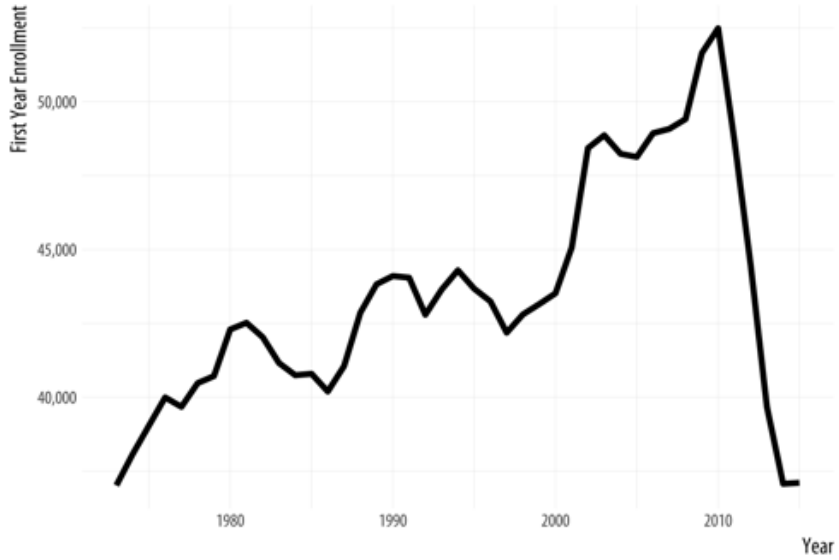

Height
INCHES
70
60
50
40
30
20
10
0
ME
MY SON
vs.
Height
INCHES
70
65
60
55
50
45
40
35
ME
MY SON
Where did he go?
Shoot. brb.

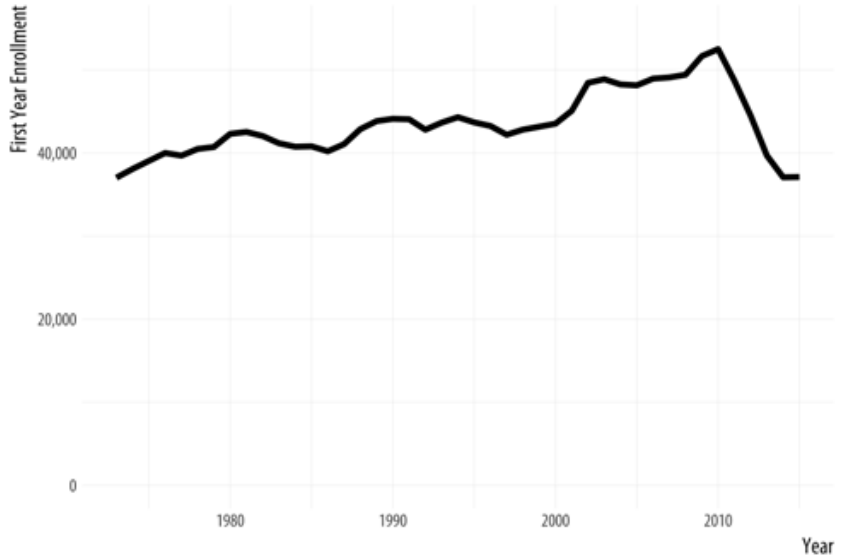
Not always a bad thing

> It is tempting to lay down inflexible rules about what to do in terms of producing your graphs, and to dismiss people who don't follow them as producing junk charts or lying with statistics. But being honest with your data is a bigger problem than can be solved by rules of thumb about making graphs. In this case there is a moderate level of agreement that bar charts should generally include a zero baseline (or equivalent) given that bars encode their variables as lengths. But it would be a mistake to think that a dot plot was by the same token deliberately misleading, just because it kept itself to the range of the data instead.

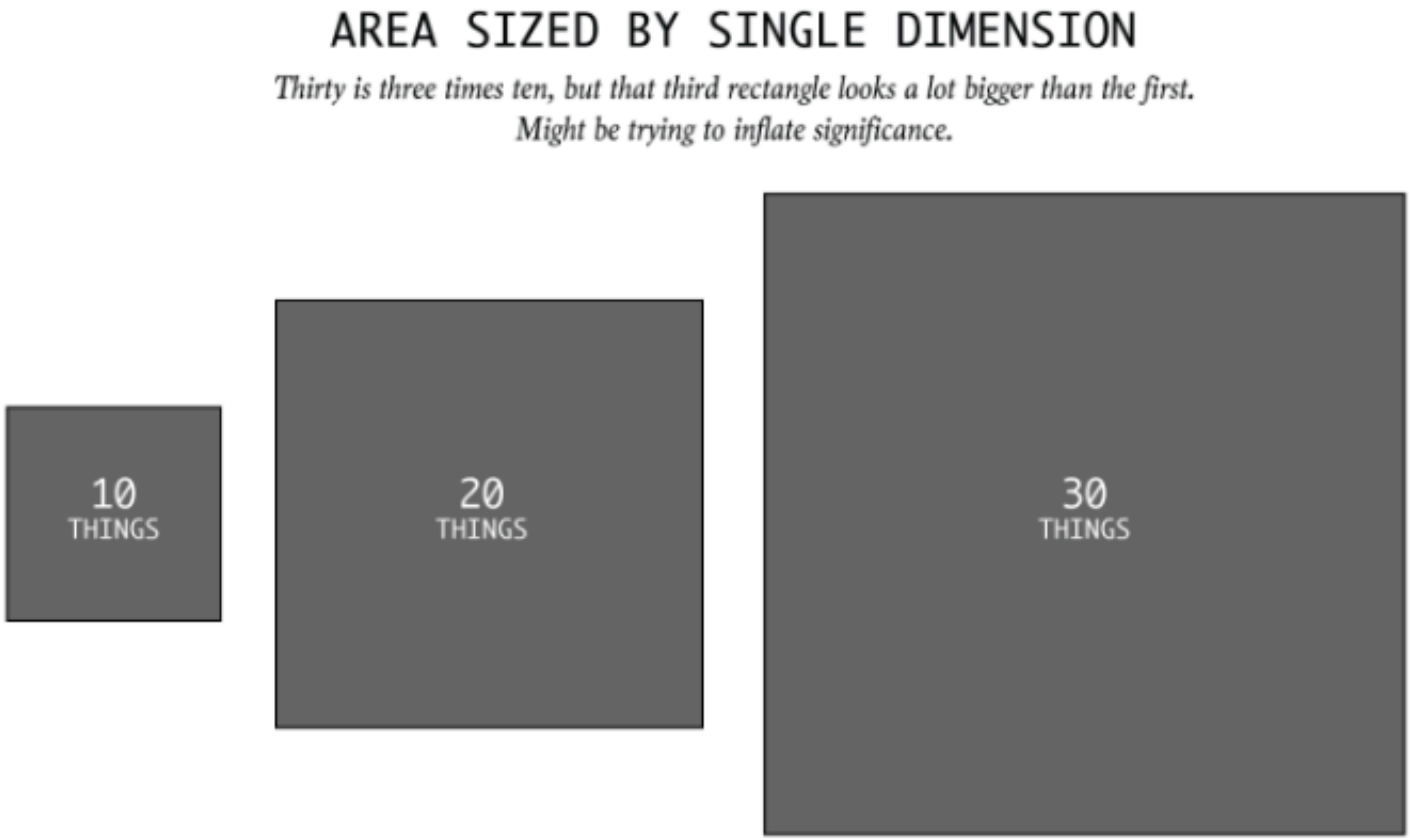

Oceania
Europe
Americas
Asia
Africa
0
20
40
60
80
Life Expectancy in years, 2007


Oceania
Europe
Americas
Asia
Africa
60
70
80
Life Expectancy in years, 2007

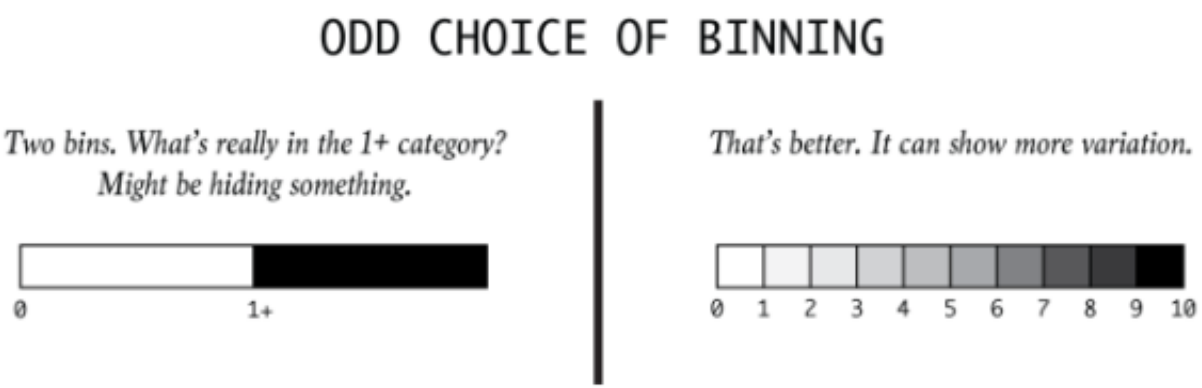

First Year Enrollment
50,000
45,000
40,000
1980
1990
2000
2010
Year


First Year Enrollment
40,000
20,000
0
1980
1990
2000
2010
Year

Scaling issues



Poor binning choices



Conclusions

- Essentially never
  - Use dual axes (produce separate plots instead)
  - Use 3D unnecessarily
- Be wary of
  - Truncated axes
  - Pie charts (usually (always?) use bars instead)
- Do
  - Minimize cognitive load
  - Be as clear as possible