

Visual Perception!

Daniel Anderson

Week 4

Data viz in the wild

- Ian
- Zach
- Ksenia

Hava, Tingyu, and Dillon on deck

Reminder

Your final project proposals are due tonight at midnight.
Please look at the syllabus for the requirements.

Agenda

- Finish up on some text pattern matching/replacement stuff
- Review Lab 1
- Aesthetic mappings and visual encodings of data
- data/ink ratio
- Some do's and don't's (which are all rules of )
- Lab 2

Learning Objectives

- Understand some of the basics of pattern matching for strings
- Understand how decisions you make with your visualizations may help or hinder comprehension

String manipulations

Reminder: base vs tidyverse

The tidyverse package is {stringr}

It is more consistent than base functions and occasionally faster

However, I tend to prefer the base functions, and they are still more commonly seen "in the wild" than **stringr**.

We'll therefore briefly cover both.

Special characters

- `^`: Anchor – matches the start of a string
- `$`: Anchor – matches the end of a string
- `*`: Matches the preceding character **zero or more** times
- `?`: Matches the preceding character **zero or one** times
- `+`: Matches the preceding character **one or more** times
- `{`: Used to specify number of matches, `a{n}`, `a{n,}`, and `a{n, m}`
- `.`: Wildcard – matches any character
- `|`: OR operator
- `[`: Alternates, also used for character matching (e.g., `[:digit:]`)
- `(`: Used for backreferencing, look aheads, or groups
- `\`: Used to escape special characters

More detail

Both of the below are good places to get more comprehensive information

- RStudio cheatsheet
- Regular expression vignette

A few quick examples

Please follow along

```
library(edld652)
d <- get_data("EDFacts_acgr_lea_2011_2019")
d

## # A tibble: 11,326 × 29
##   ALL_COHORT ALL_RATE CWD_COHORT CWD_RATE DATE_CUR
##       <dbl>    <chr>        <dbl>    <chr>      <chr>
## 1       252     80            3     PS        03OCT15
## 2       398     75            47    70-79      03OCT15
## 3      1020     89            51    40-49      03OCT15
## 4       750     91            35    60-69      03OCT15
## 5       128    55-59            15    LT50       03OCT15
## 6       166    90-94            9    GE50       03OCT15
## # ... with 11,320 more rows, and 24 more variables:
## #   ECD_COHORT <dbl>, ECD_RATE <chr>,
## #   FIPST <chr>, FILEURL <chr>, LEAID <chr>,
## #   LEANM <chr>, LEP_COHORT <dbl>,
## #   LEP_RATE <chr>, MAM_COHORT <dbl>,
## #   MAM_RATE <chr>, MAS_COHORT <dbl>,
## #   MAS_RATE <chr>, MBL_COHORT <dbl>, ...
```

Challenge

Do three things:

- Create a new variable that identifies if the LEA is associated with a city or a county
- Drop "City" or "County" from the LEA name (e.g., "Albertville City" would be "Albertville")
- Replace all `.` with **--DOT--** in **FILEURL** (so they are not actual links)

You'll need **base::gsub()** or
stringr::str_replace_all()

05:00

City or county

Ideas?

```
d <- d %>%
  mutate(county = grepl("county$", tolower(LEANM)))
```

Another option

```
d <- d %>%
  mutate(county = grepl("[Cc]ounty$", LEANM))
```

Quick Check

```
d %>%
  select(LEANM, county) %>%
  print(n = 15)
```

```
## # A tibble: 11,326 × 2
##   LEANM           county
##   <chr>          <lgl>
## 1 Albertville City FALSE
## 2 Marshall County TRUE
## 3 Hoover City    FALSE
## 4 Madison City   FALSE
## 5 Leeds City     FALSE
## 6 Boaz City      FALSE
## 7 Trussville City FALSE
## 8 Alexander City FALSE
## 9 Andalusia City FALSE
## 10 Anniston City FALSE
## 11 Arab City     FALSE
## 12 Athens City   FALSE
## 13 Attalla City  FALSE
## 14 Auburn City   FALSE
## 15 Autauga County TRUE
## # ... with 11,311 more rows
```

Remove city/county

- Lots of ways to do this. Use `base::gsub()` or `stringr::str_replace_all()`
- Replace everything after the space with nothing

```
d %>%
  select(LEANM) %>%
  mutate(new_name = gsub(" .+", "", LEANM))
```

```
## # A tibble: 11,326 × 2
##   LEANM           new_name
##   <chr>          <chr>
## 1 Albertville City Albertville
## 2 Marshall County Marshall
## 3 Hoover City    Hoover
## 4 Madison City   Madison
## 5 Leeds City     Leeds
## 6 Boaz City      Boaz
## # ... with 11,320 more rows
```

You can also use "`\s`" to mean "space"

Another way

There are other ways too, of course

```
d %>%
  select(LEANM) %>%
  mutate(new_name = gsub(" City$| County$", "", LEANM))
```

```
## # A tibble: 11,326 × 2
##   LEANM          new_name
##   <chr>          <chr>
## 1 Albertville City Albertville
## 2 Marshall County Marshall
## 3 Hoover City    Hoover
## 4 Madison City   Madison
## 5 Leeds City     Leeds
## 6 Boaz City      Boaz
## # ... with 11,320 more rows
```

Final step

Handling the URLs. This is a bit artificial, but it illustrates escaping, which is important.

- Remember `.` is a special character, so needs to be escaped
- `\` itself is a special character so it needs to be escaped. Functionally, then, you escape special characters with `\\\`, not `\`.

```
d %>%  
  select(FILEURL)
```

```
## # A tibble: 11,326 × 1  
## # ... with 11,320 more rows, and 1 more variable:  
## #   FILEURL <chr>
```

```
d %>%
  select(FILEURL) %>%
  mutate(FILEURL = gsub("\\.", "--DOT--", FILEURL)) %>%
  as.data.frame()
```

```
## 
## 1 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 2 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 3 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 4 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 5 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 6 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 7 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 8 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 9 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 10 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 11 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 12 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 13 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 14 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 15 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 16 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 17 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 18 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 19 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 20 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 21 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 22 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
## 23 https://www2--DOT--ed--DOT--gov/about/inites/ed/edfacts/data-files/a
```

A few more examples

```
some_strings <- c("abc", "a23", "def", "Ad6")
```

Remove the digits

```
gsub("\d", "", some_strings)
```

```
## [1] "abc" "a"    "def"  "Ad"
```

Extract the digits

```
gsub("\D", "", some_strings)
```

```
## [1] ""    "23"  ""    "6"
```

Starts with A

```
grepl("^[aA]", some_strings)
```

```
## [1] TRUE TRUE FALSE TRUE
```

There's lots more

- Please checkout the vignette, in particular
- I end up using back-referencing a lot

```
cities <- c("Eugene, Oregon", "Denver, Colorado")
gsub(
  "^(.+), (.+)\$", # pattern with two groupings
  "The state of \\\2 includes the city of \\\1", # reference the groups
  cities)

## [1] "The state of Oregon includes the city of Eugene"
## [2] "The state of Colorado includes the city of Denver"
```

Review Lab 1

Visual perception

Disclaimer

I'm not a psychologist

I don't really know why we perceive things certain ways.

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

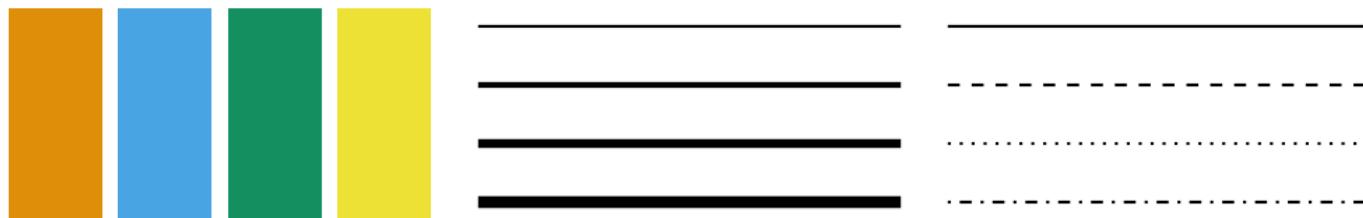
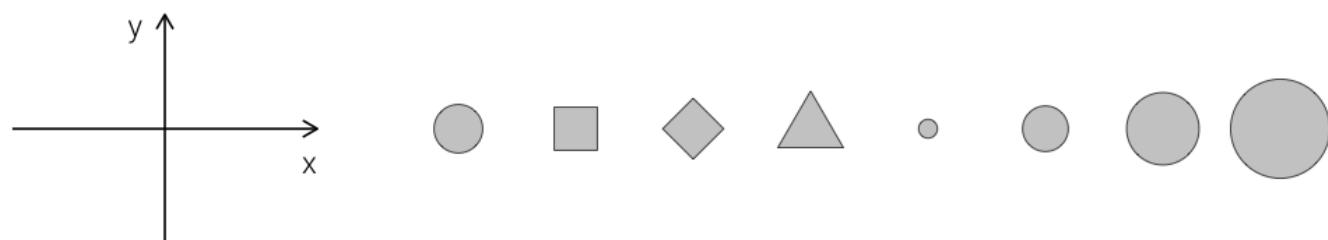
Visual Cues

- **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?

Visual Cues

- **Shape:** *Categorical.* Belonging to which group?
- **Area:** *Numeric.* How big (in two dimensions)?
- **Volume:** *Numeric.* How big (in three dimensions)?
- **Shade:** *Numeric or Categorical.* To what extent? How Severely?
- **Color:** *Numeric or Categorical.* To what extent? How Severely?

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

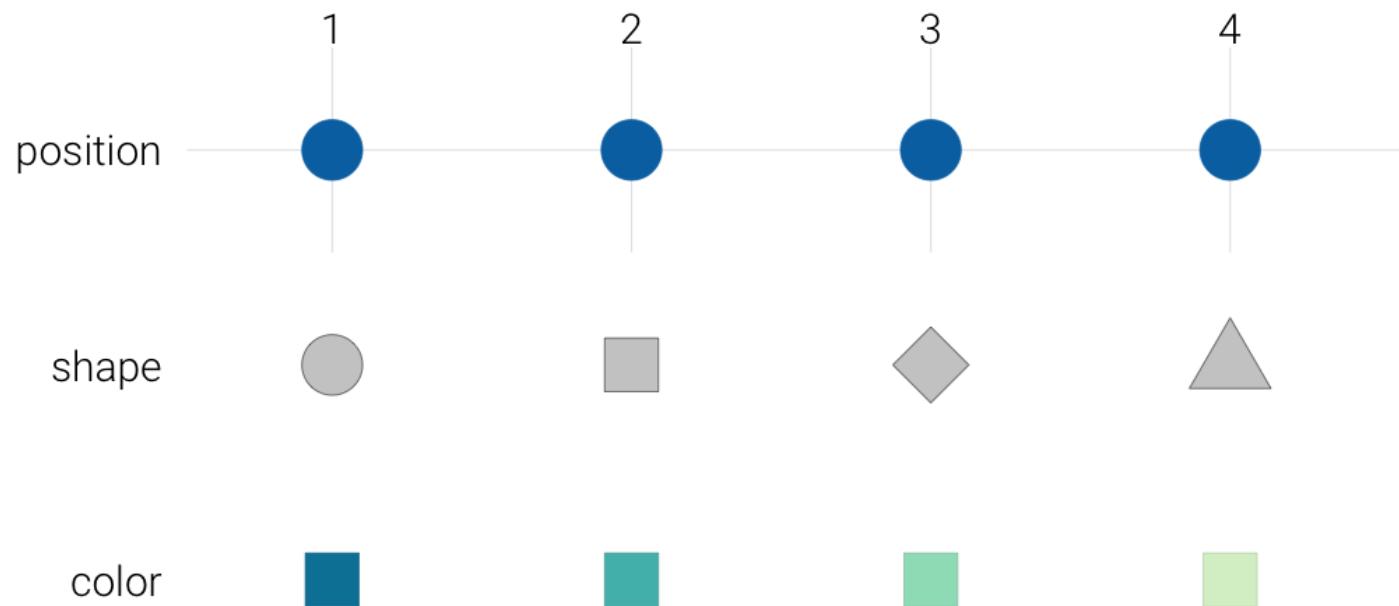
You can assume that month and day can be collapsed to a single *date* column

Scales

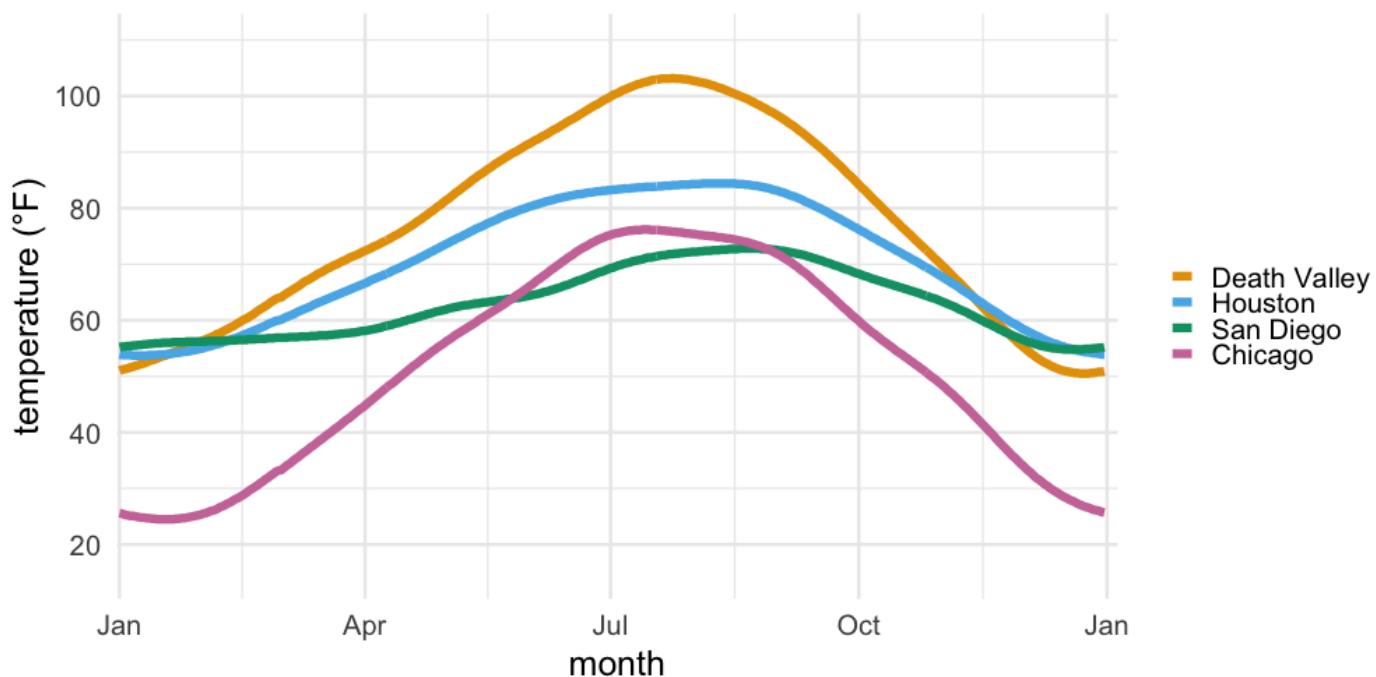
A scale defines a unique mapping between data and aesthetics. Importantly, a scale must be one-to-one, such that for each specific data value there is exactly one aesthetics value and vice versa. If a scale isn't one-to-one, then the data visualization becomes ambiguous.

- Which data values correspond to specific aesthetic values?

Basic Scales

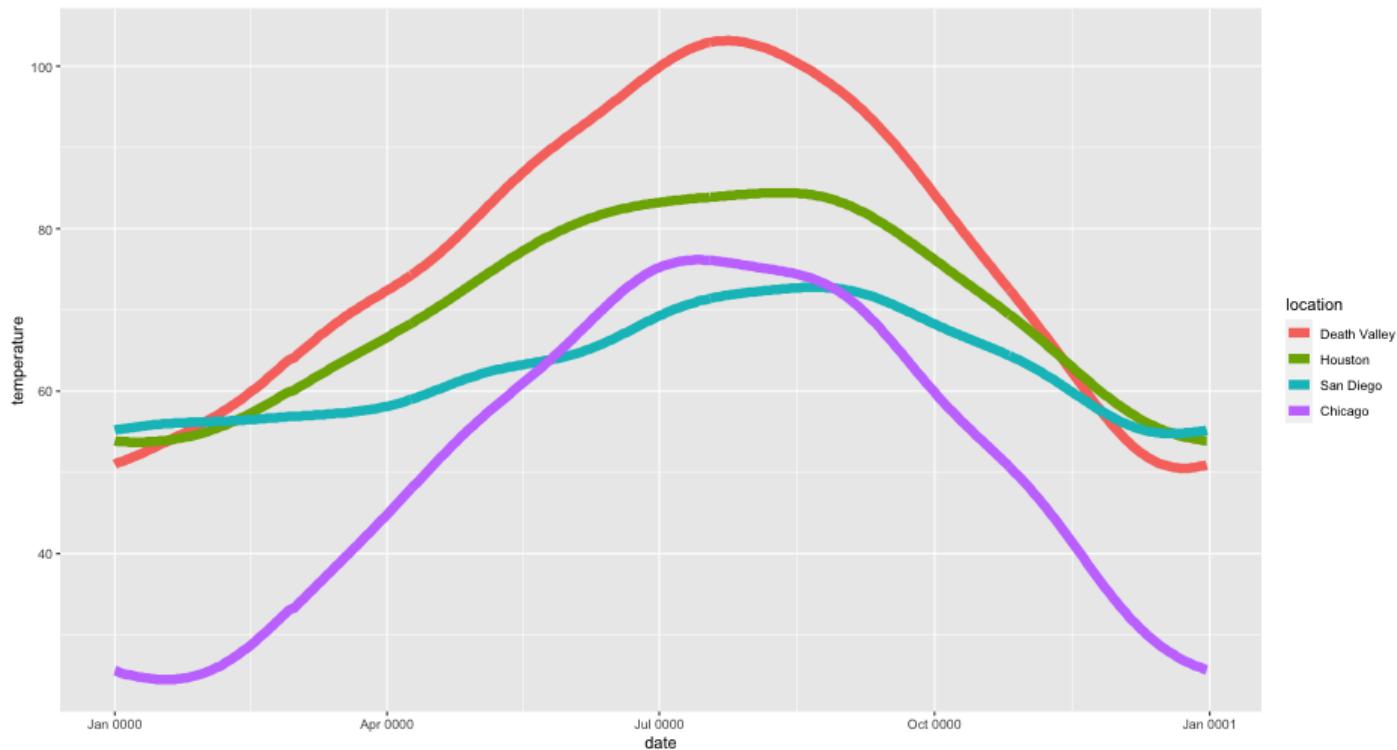


Putting it to practice



Basic code for previous plot

```
ggplot(tempus_long, aes(date, temperature)) +  
  geom_line(aes(color = location))
```

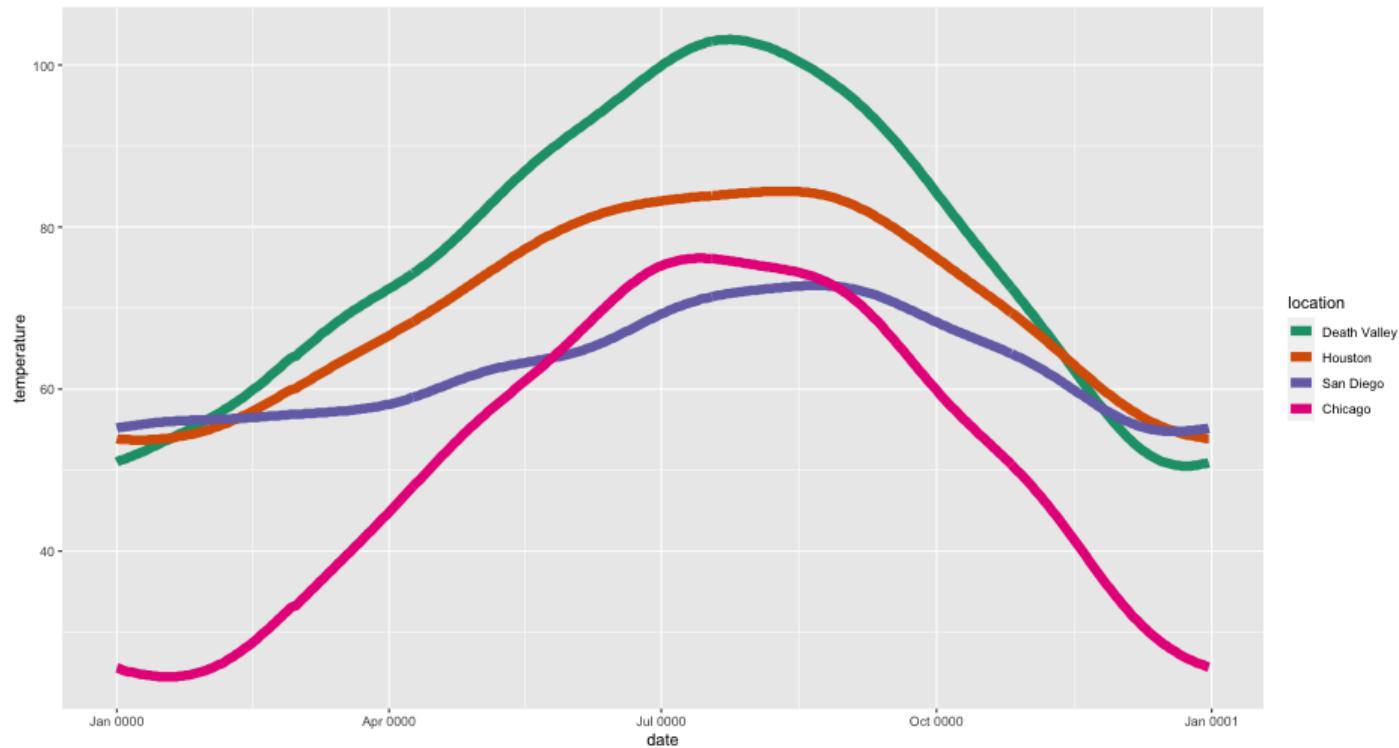


Change colors

If you want to change the colors on the previous plot, you have to change the colors of the scale for the color mapping.

In other words, color is being mapped to data, and you have to change the color scale.

```
ggplot(tempus_long, aes(date, temperature)) +  
  geom_line(aes(color = location)) +  
  scale_color_brewer(palette = "Dark2")
```



One more note on colors

There are lots of different scales and some work better than others. We'll talk about them more next week.

You **do not** use `scale_color_*`() if you are not mapping data to color

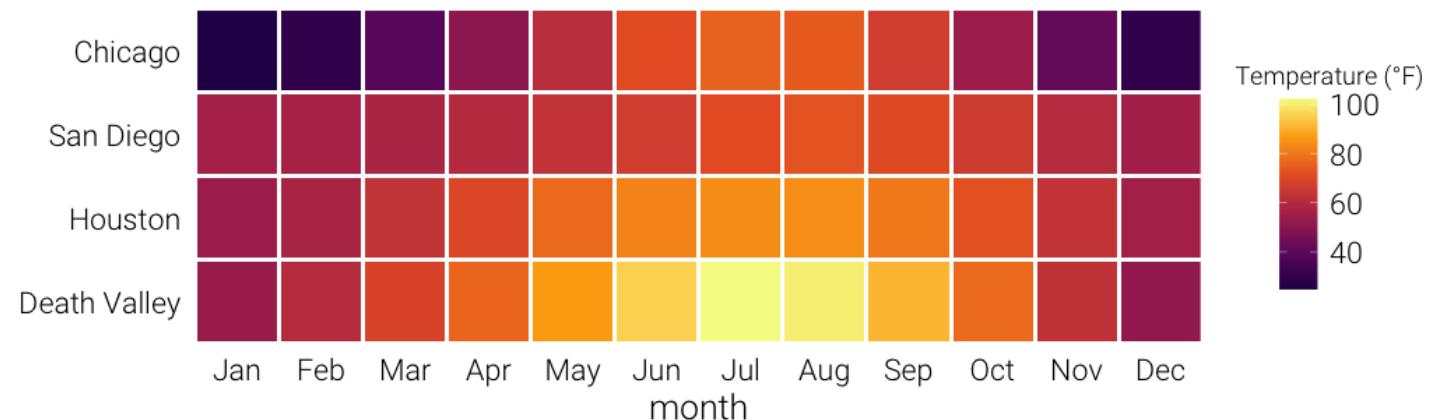
Make sure to keep straight `scale_color_*`() and `scale_fill_*`()

Alternative representation

Can you think of other ways to show this relation?

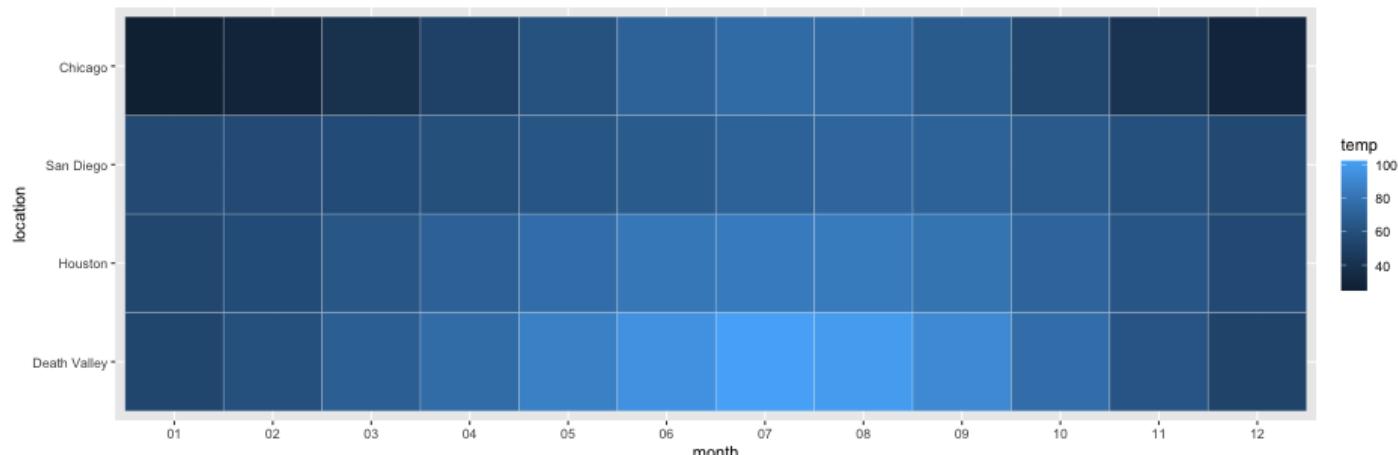
Alternative representation

Same plot as before, but with different scales



Basic code for previous plot

```
temps_long %>%
  group_by(location, month) %>%
  summarize(temp = mean(temperature)) %>%
  ggplot(aes(month, location)) +
  geom_tile(aes(fill = temp),
            color = "white") +
  coord_fixed()
```

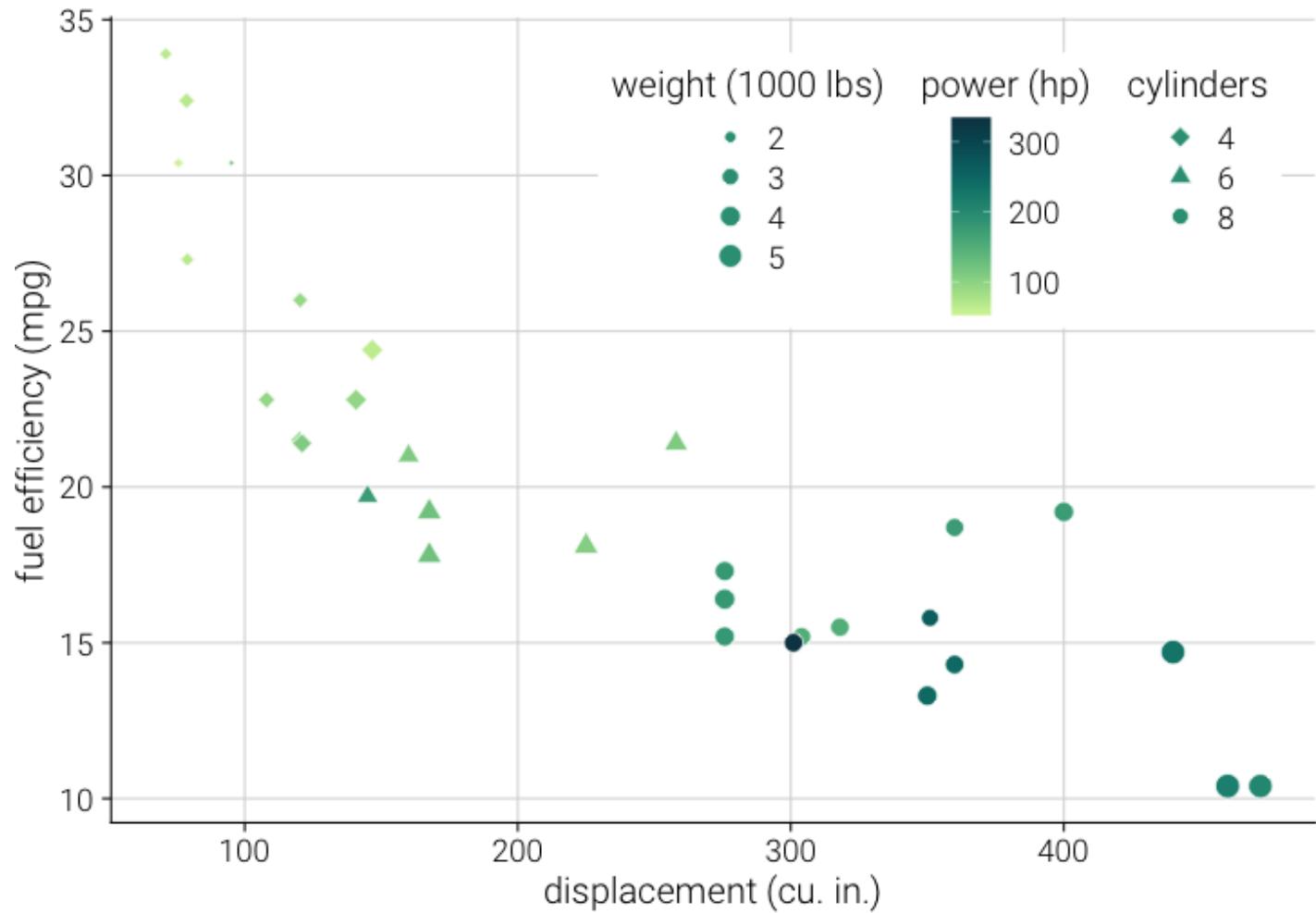


Change the fill

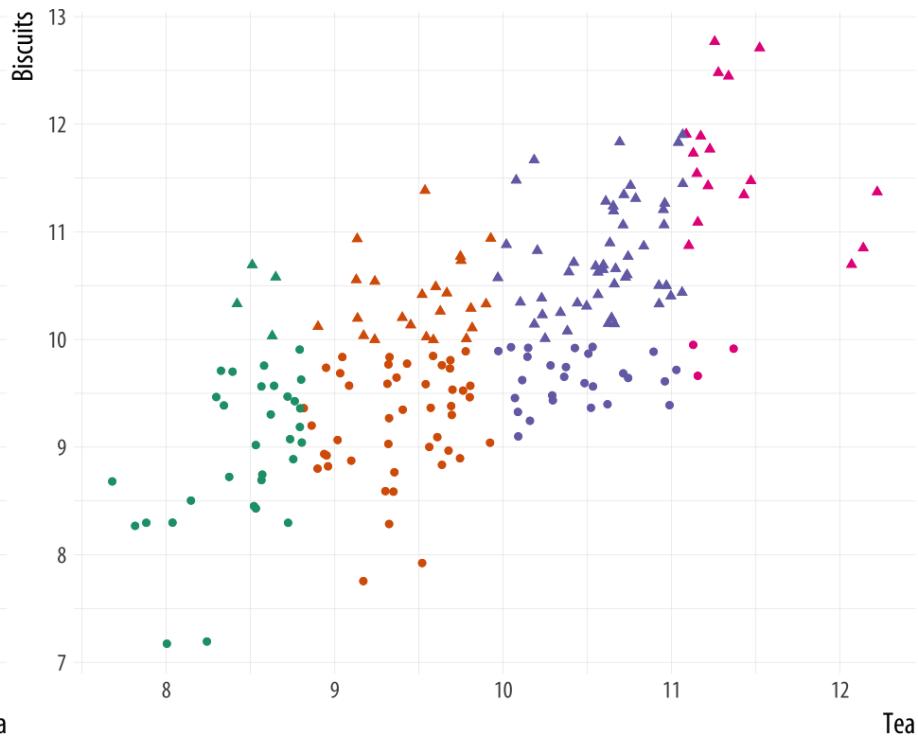
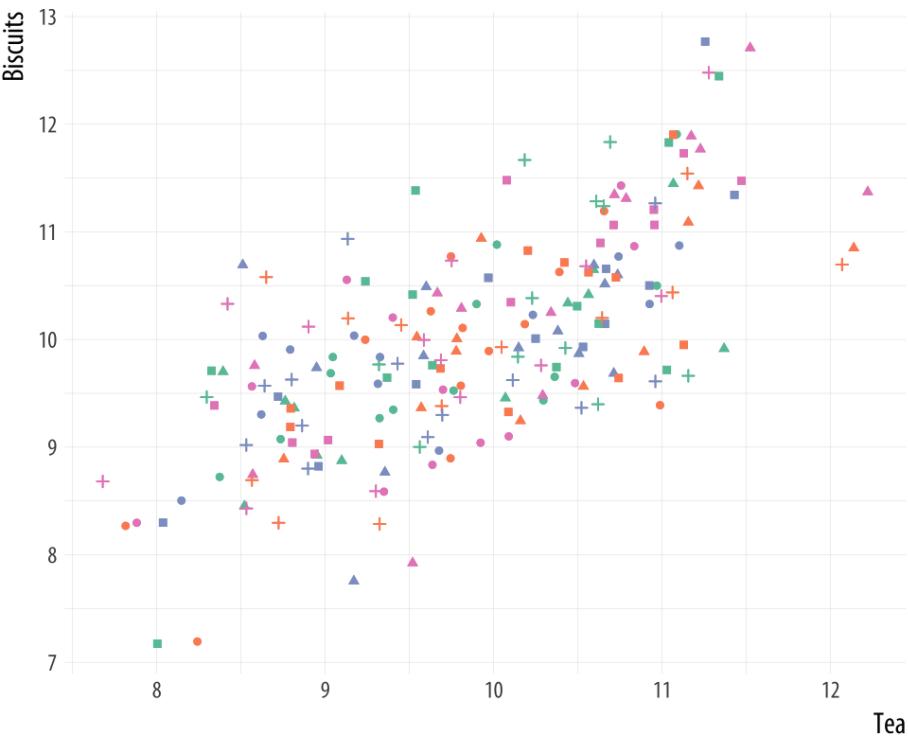
```
temps_long %>%
  group_by(location, month) %>%
  summarize(temp = mean(temperature)) %>%
  ggplot(aes(month, location)) +
  geom_tile(aes(fill = temp),
            color = "white") +
  coord_fixed() +
  scico::scale_fill_scico(palette = "tokyo")
```

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



Additional scales can become lost without high structure in the data



Data ink ratio

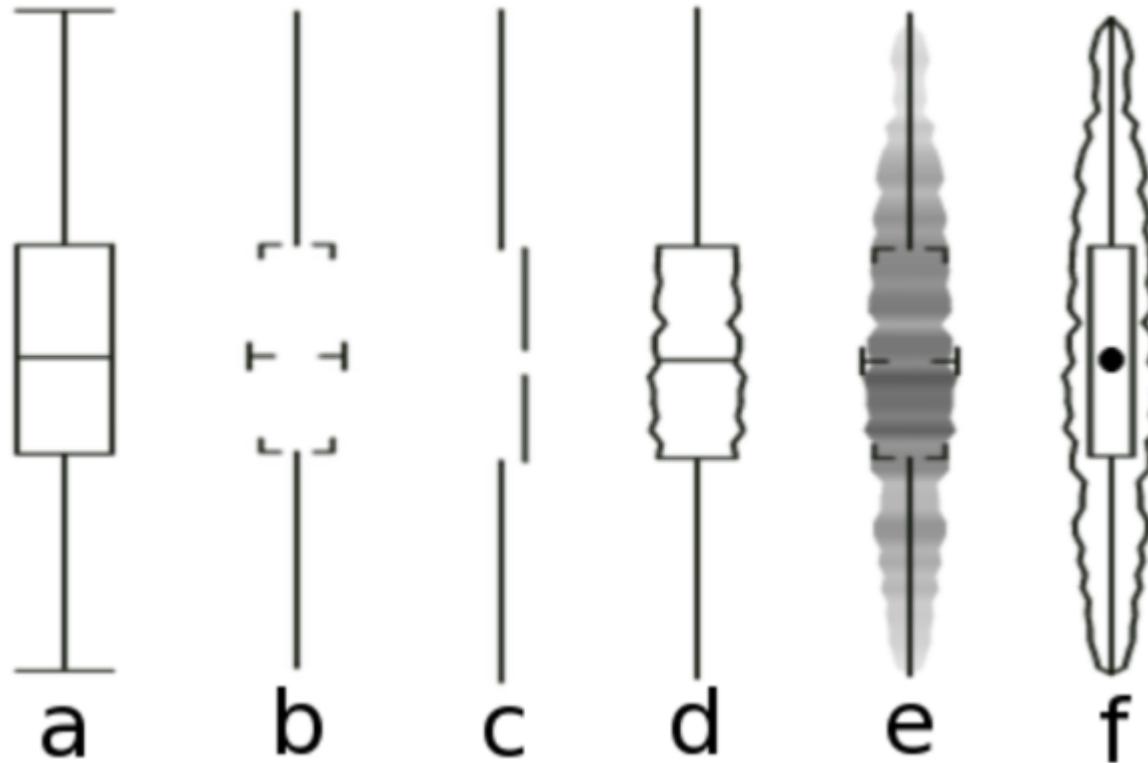
What is it?

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!

A color photograph of an older man with light brown hair, wearing a dark suit jacket, a white shirt, and a red patterned tie. He has a wide-open mouth and appears to be shouting or speaking very loudly. A large, white, rounded rectangular speech bubble is positioned above his head, containing the text "NOT SO FAST, MY FRIEND!" in bold, black, sans-serif capital letters. The background is slightly blurred, showing what looks like a park or garden area with trees and possibly a fountain.

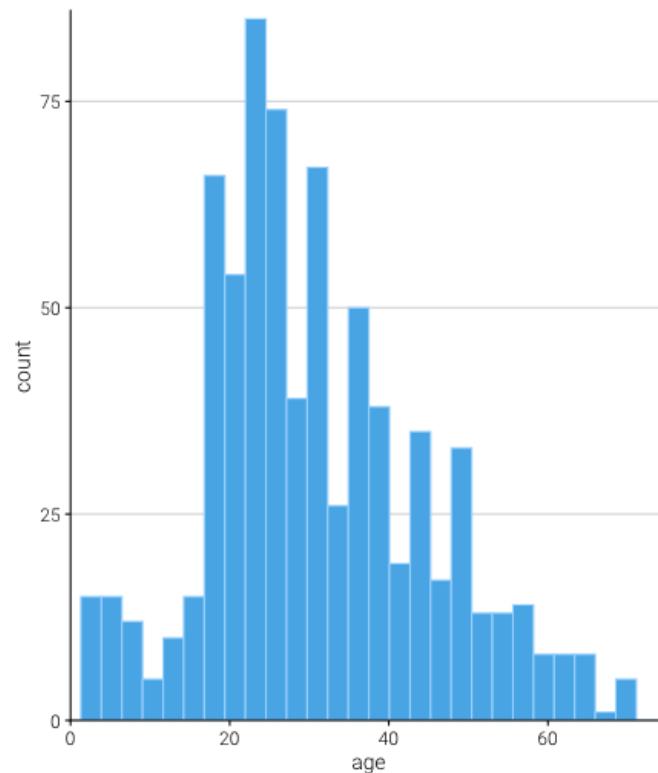
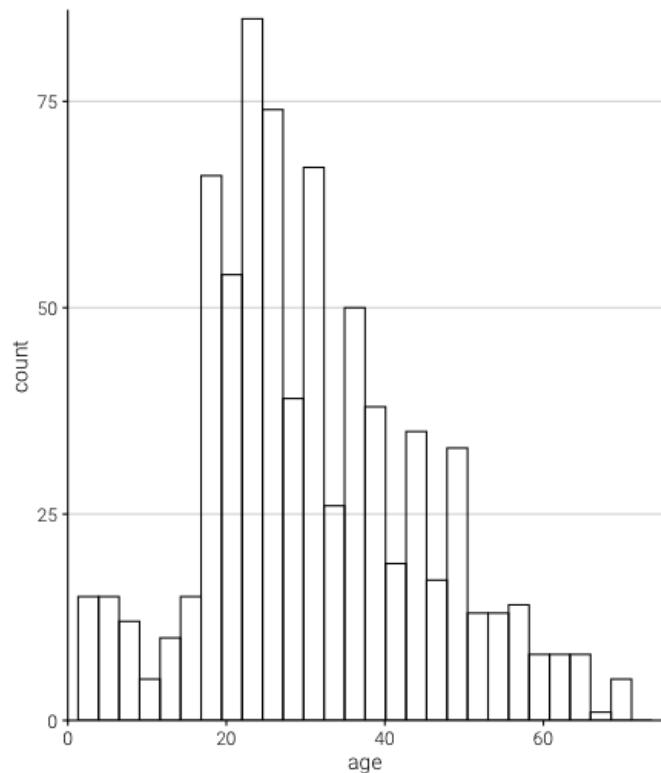
NOT SO
FAST, MY
FRIEND!

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

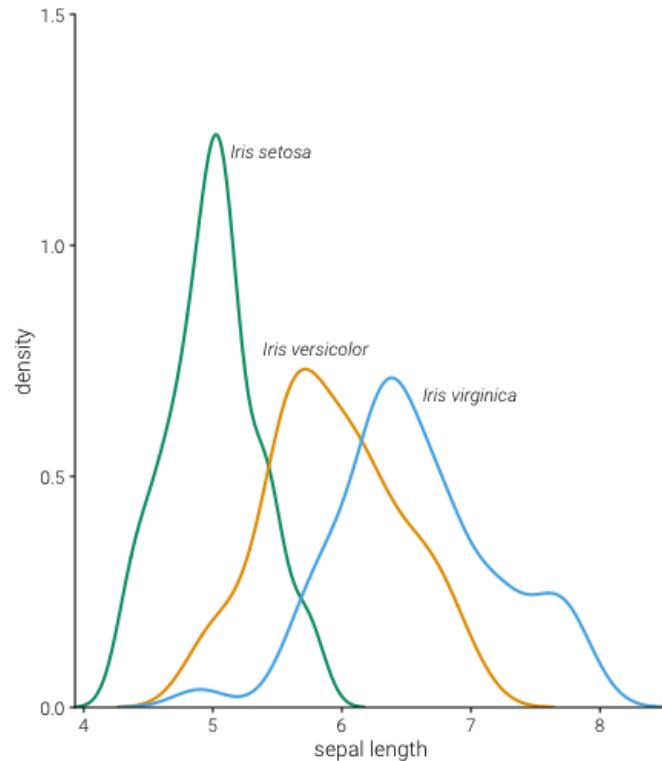
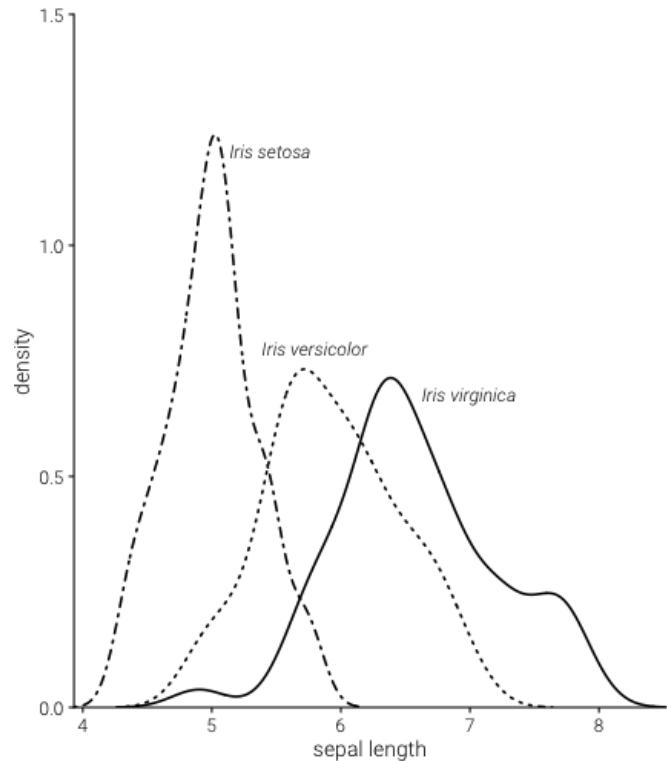
Which do you prefer?

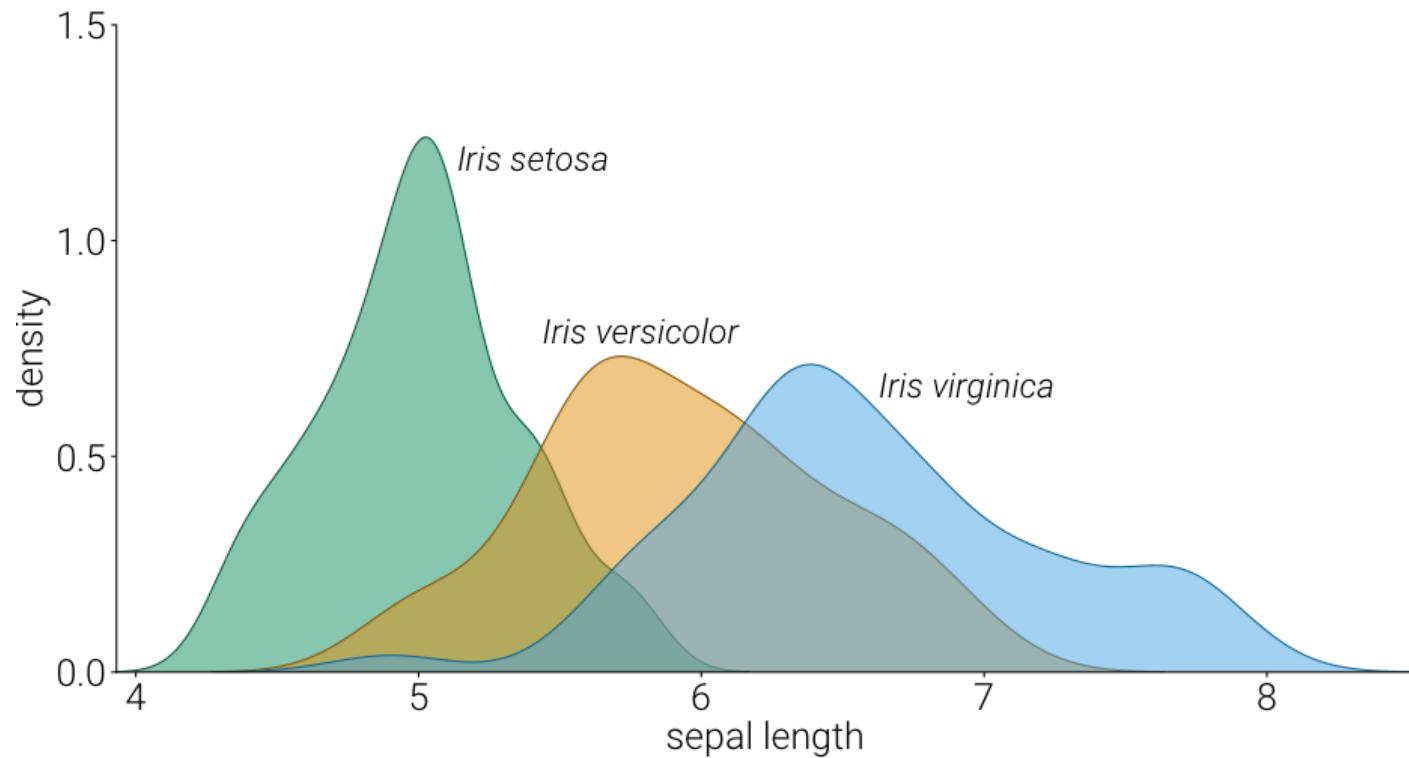


Advice from Wilke

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





MONSTROUS COSTS

Total House and Senate campaign expenditures,
in millions

This?



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

Compromise?

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

- Install packages

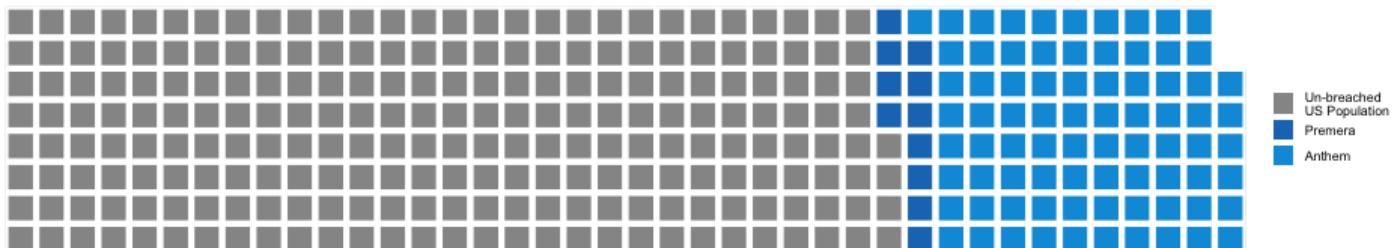
```
install.packages("extrafont")
remotes::install_github("wch/extrafontdb")
remotes::install_github("wch/Rttf2pt1")
remotes::install_github("hrbrmstr/waffle")
```

- 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

Doesn't seem to work anymore... 

- 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")
```

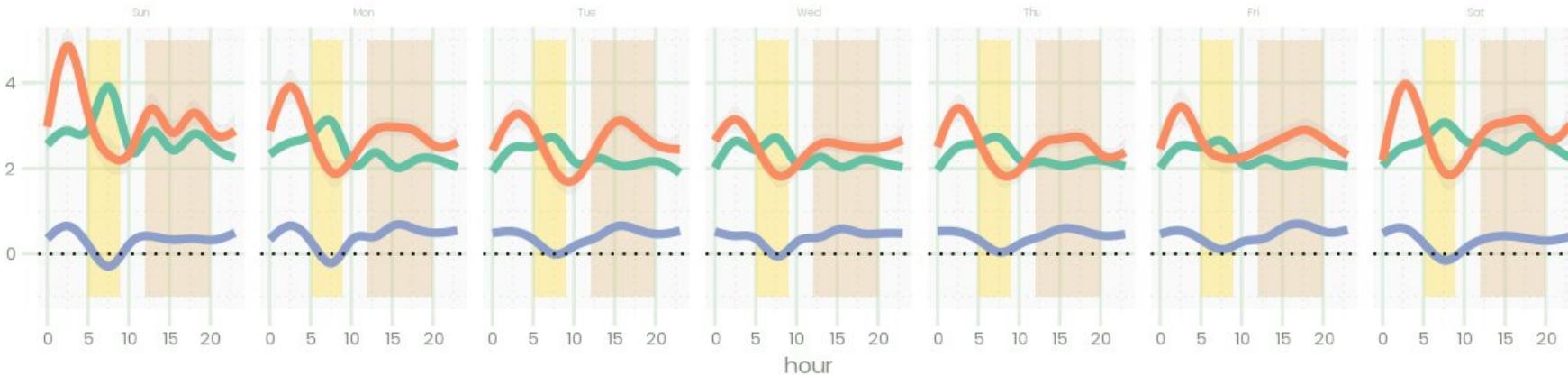
Should look like this



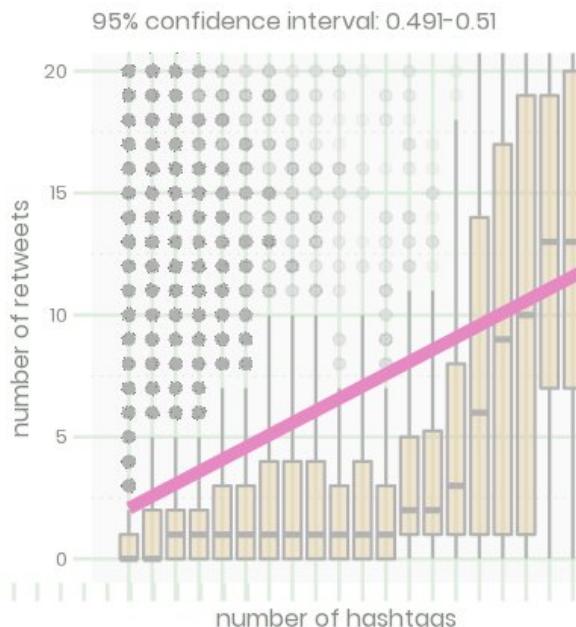
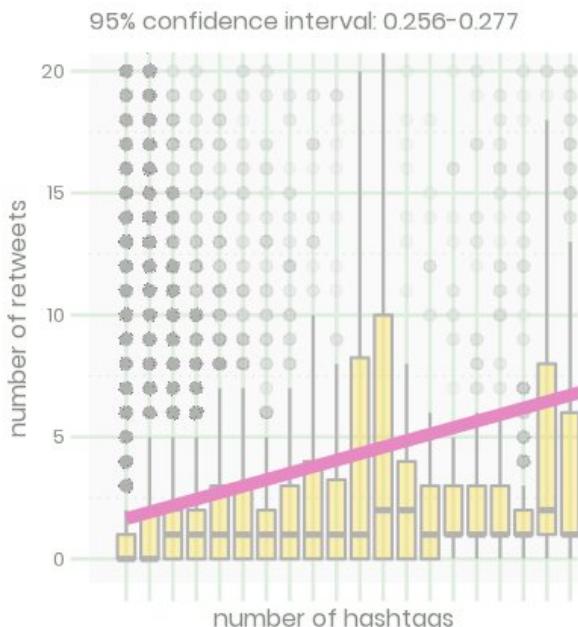
Despite glyphs not (easily) working anymore, I still recommend you [check it out](#). It's a neat package and does have some integration with ggplot2 now.

#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?

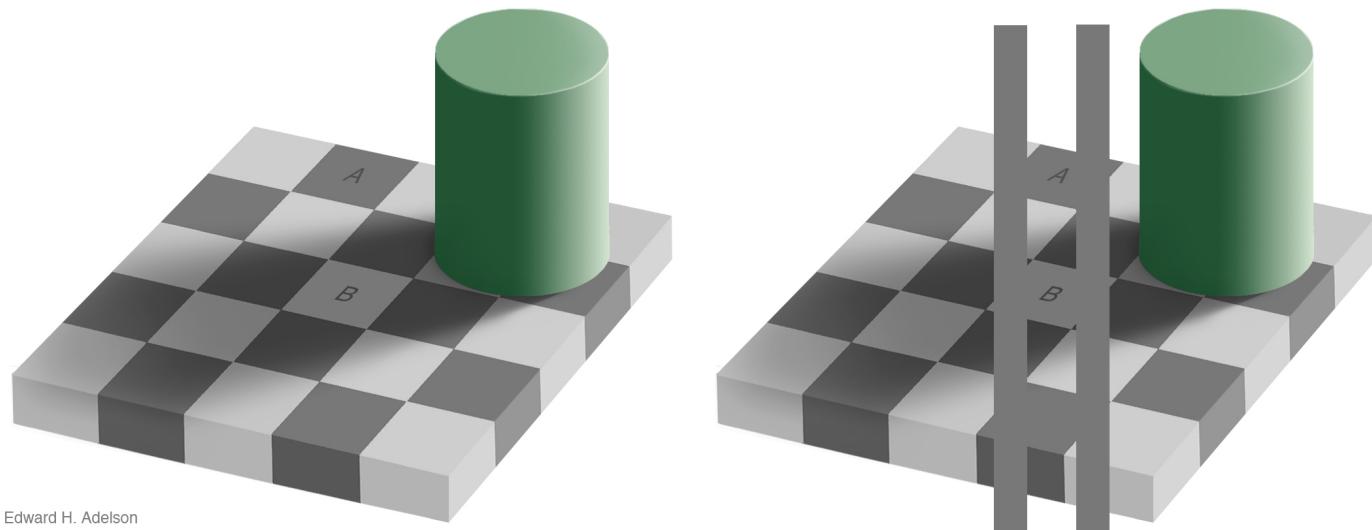


Hashtags used in the mid-afternoon, when most people are on twitter, are almost twice as effective at provoking retweets! Also, while most days see a single pronounced increase in tweets, Sunday afternoon sees two distinct bumps in #rstats tweets.

You can create them!

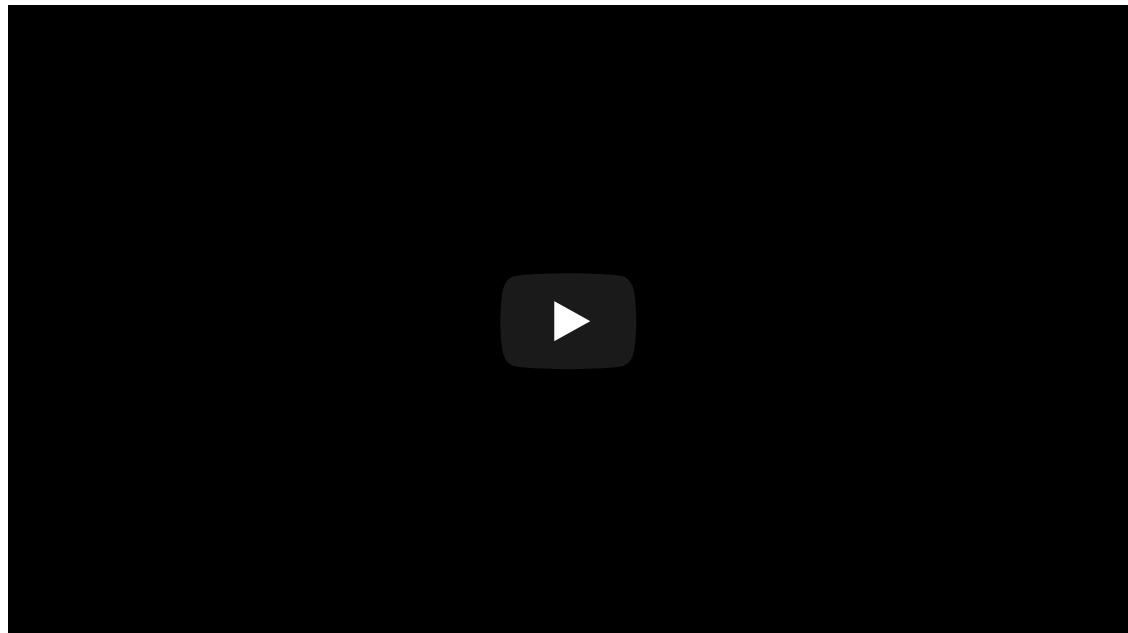
- Create plots
- Use illustrator or similar to put them together
- Add some annotations
- Consider using glyphs for greater memory
- You can do a lot in R without going to illustrator etc. by just using **{patchwork}** or **{cowplot}**

More visual properties



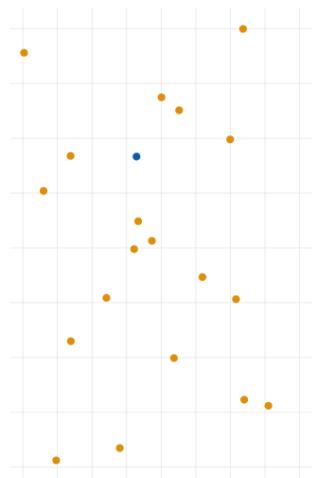
Edward H. Adelson

Or in real life

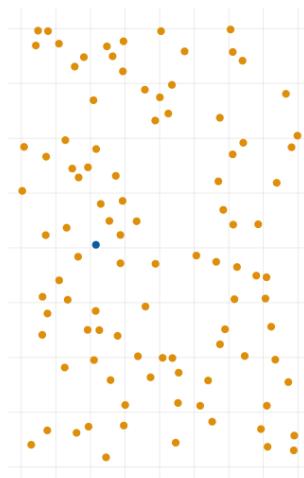


Where's the blue circle in each plot?

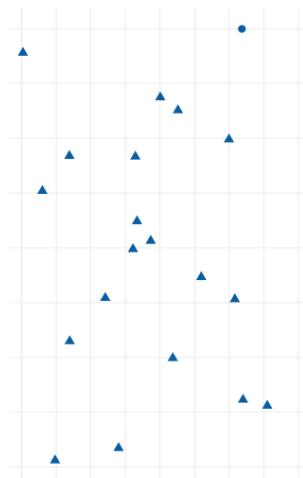
Color Only, N=20



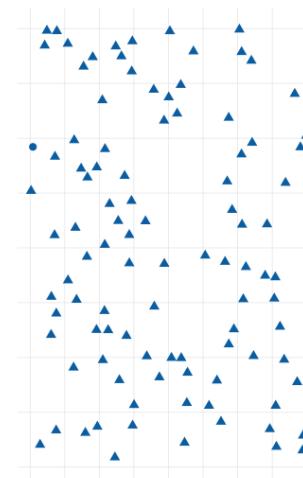
Color Only, N=100



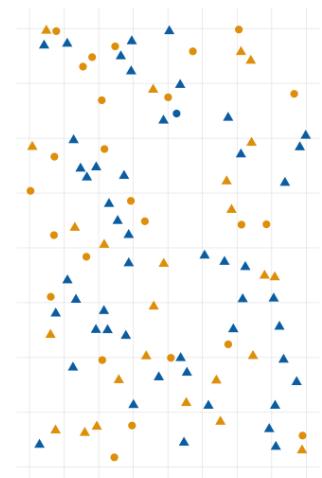
Shape Only, N=20



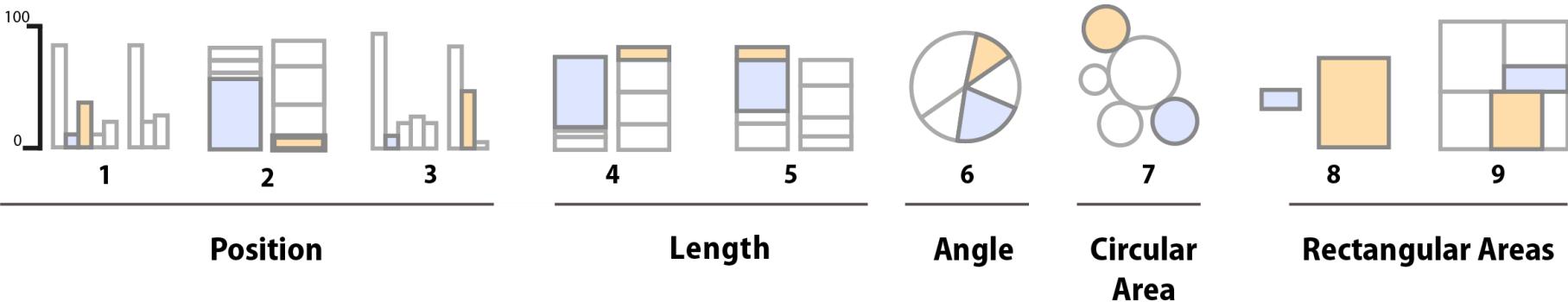
Shape Only, N=100



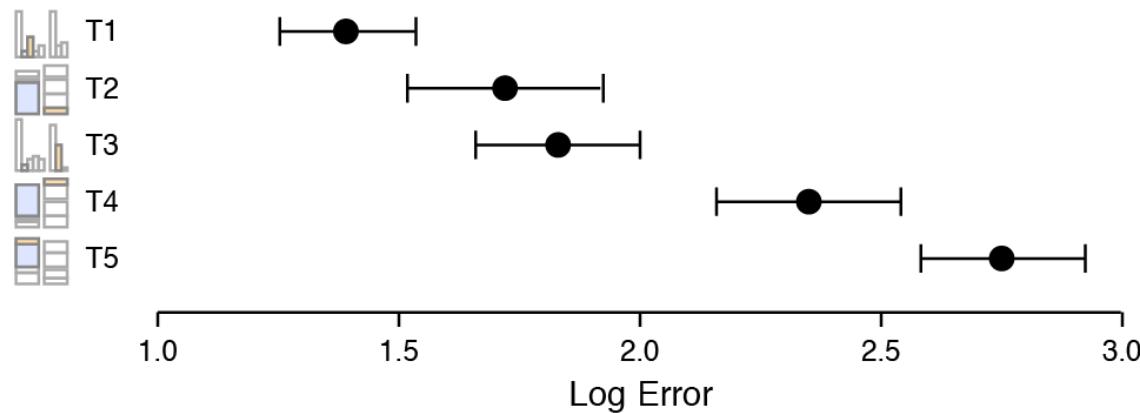
Color & Shape, N=100



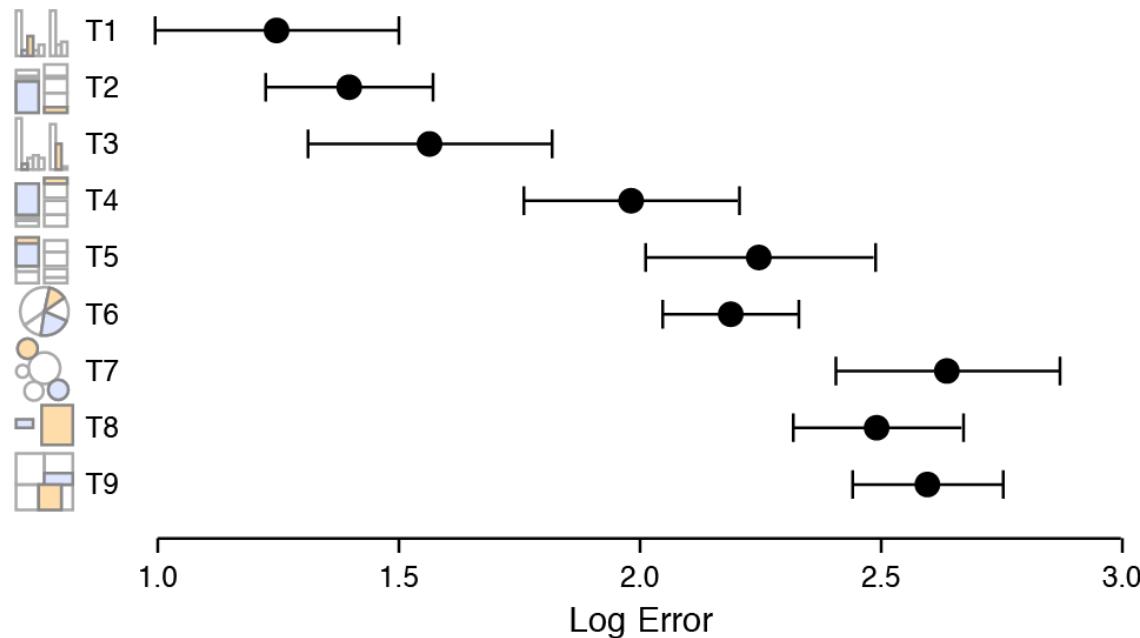
What are we good at perceiving?



Cleveland & McGill's Results



Crowdsourced Results

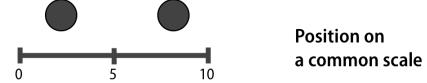


Ordered

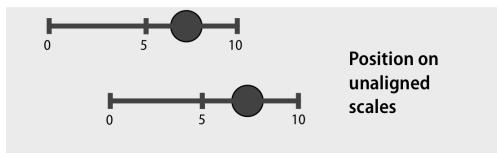
data

mappings:

Ranked



Position on
a common scale



Position on
unaligned
scales

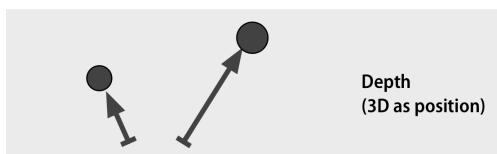
Length



Tilt or Angle



Area (2D as size)



Depth
(3D as position)



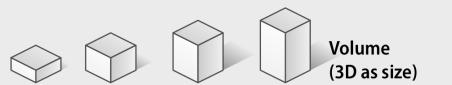
Color luminance
or brightness



Color saturation
or intensity

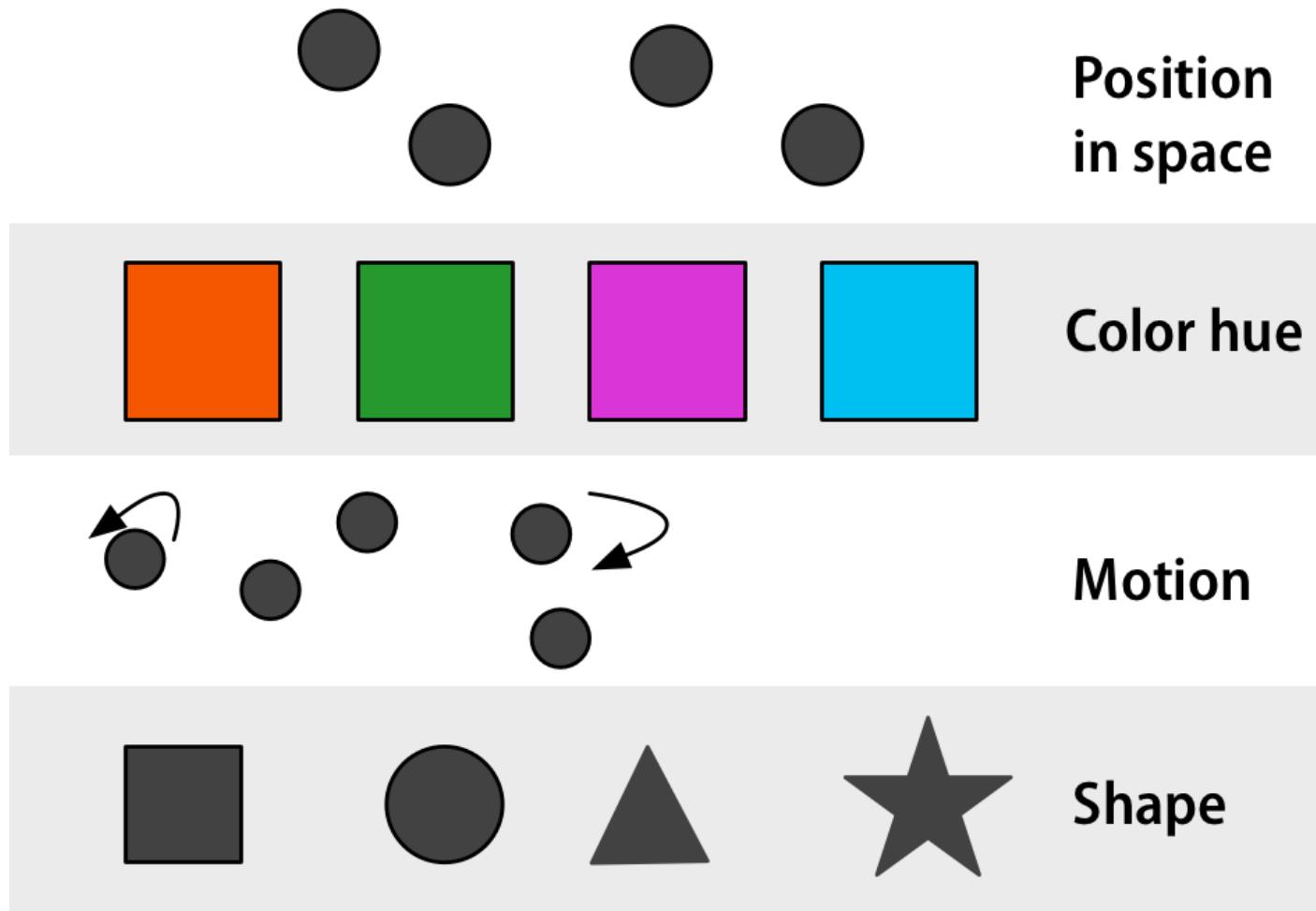
|))>

Curvature



Volume
(3D as size)

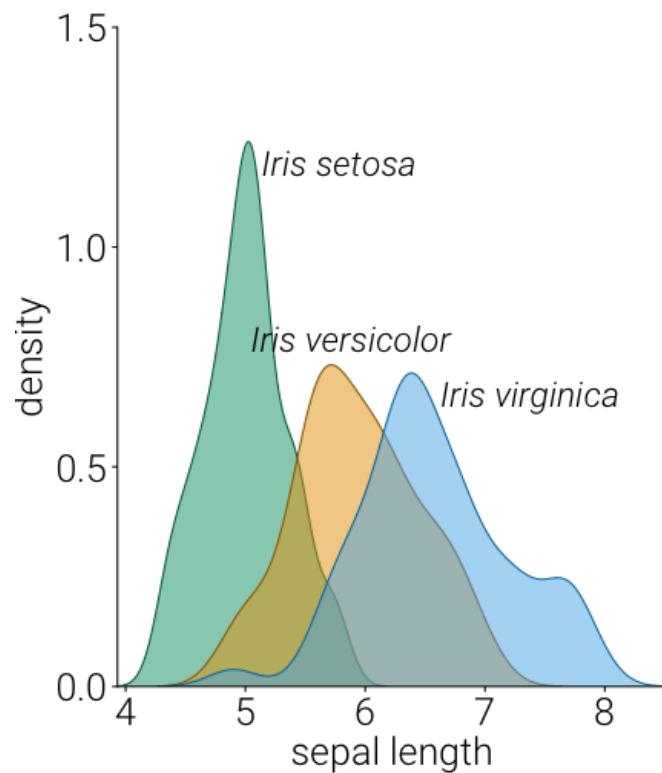
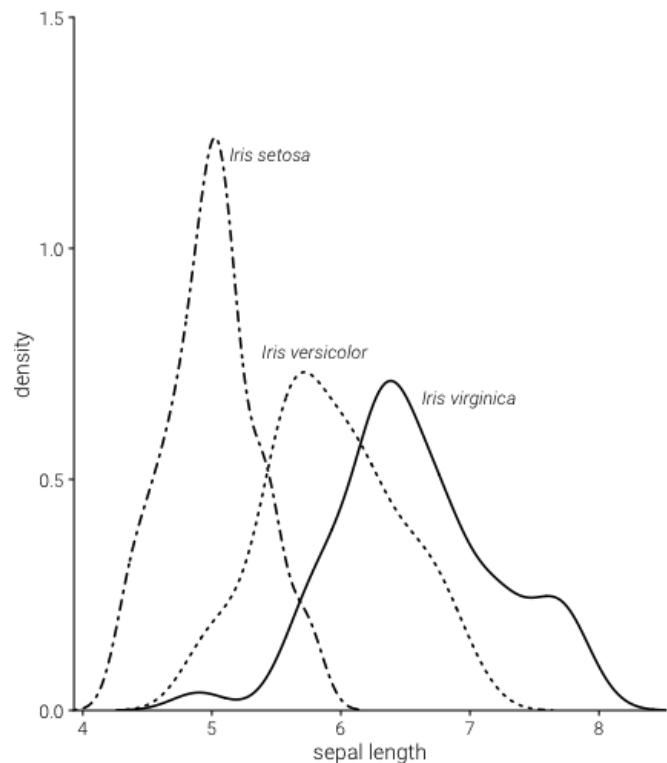
Unordered data mappings



Some things
to avoid

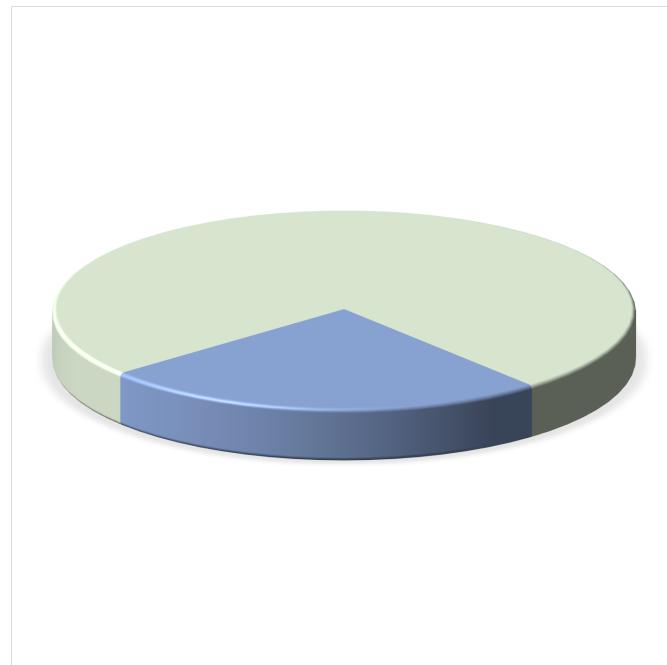
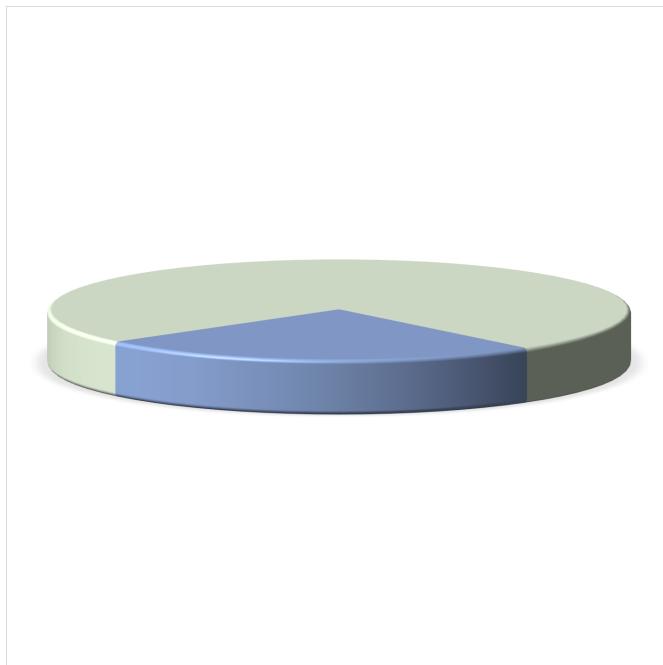
Line drawings

As discussed earlier



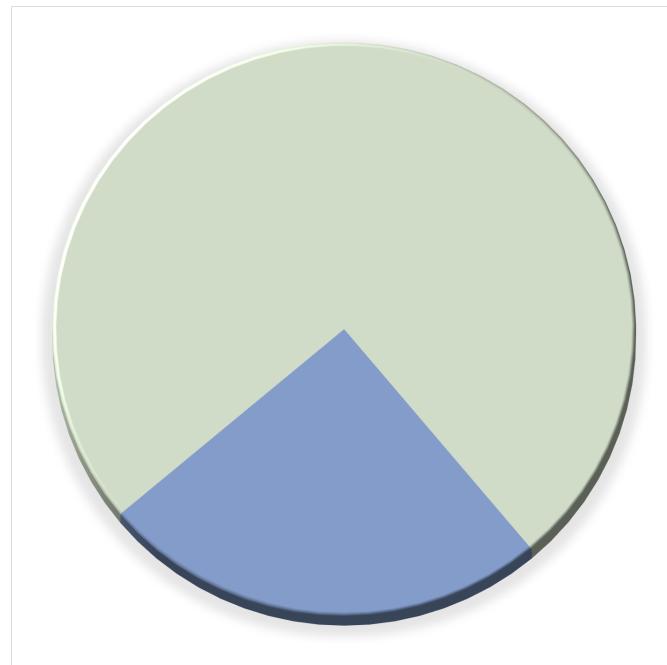
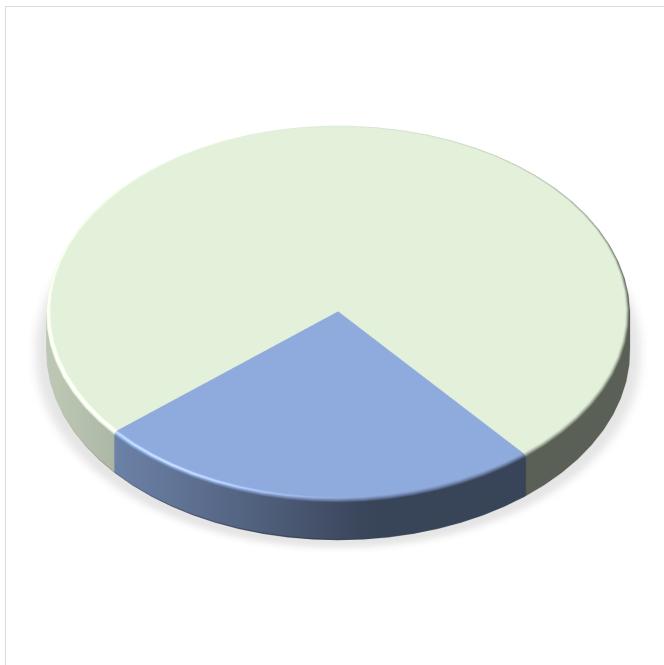
Much worse

Unnecessary 3D



Much worse

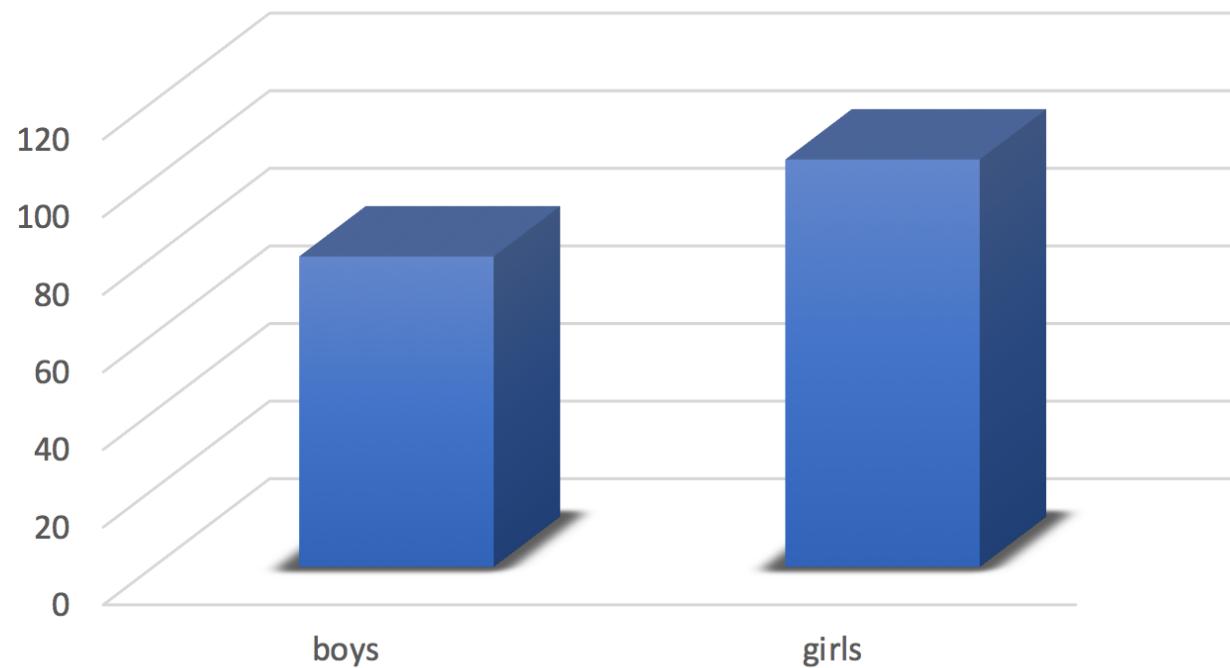
Unnecessary 3D



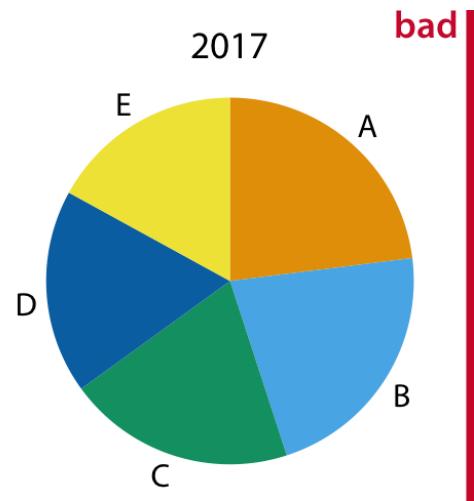
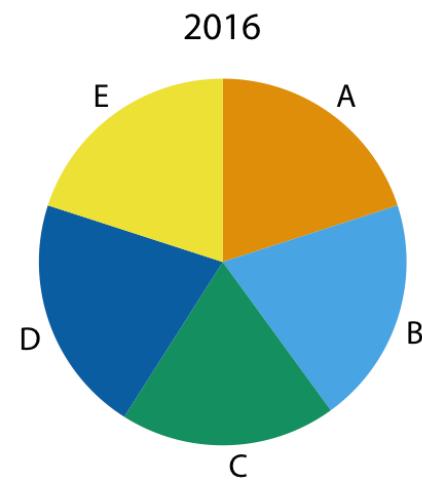
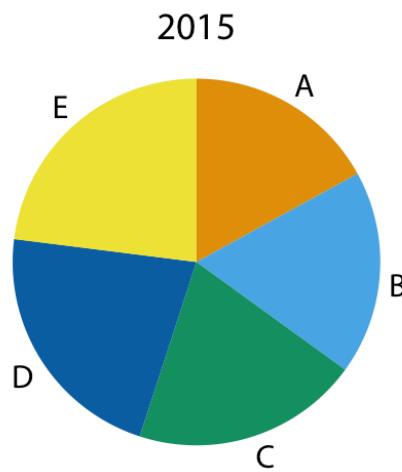
Horrid example

Used relatively regularly

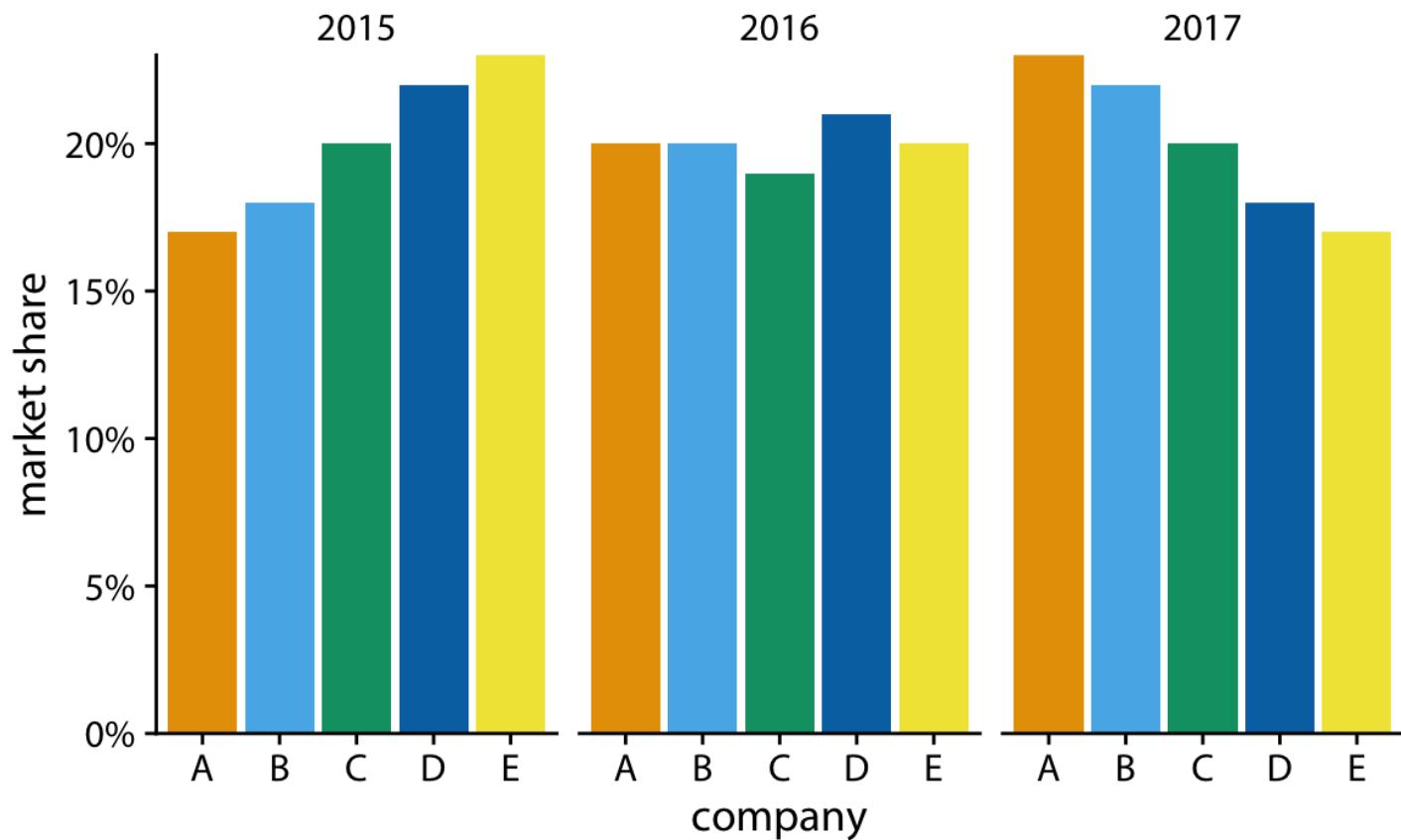
The left bar is 80; the right is 105



Pie charts w/lots of categories

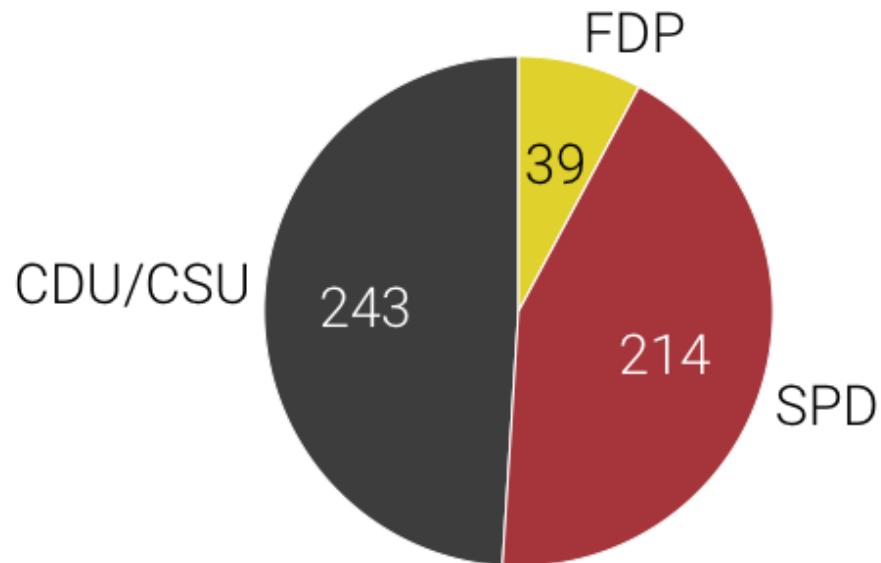


Alternative representation



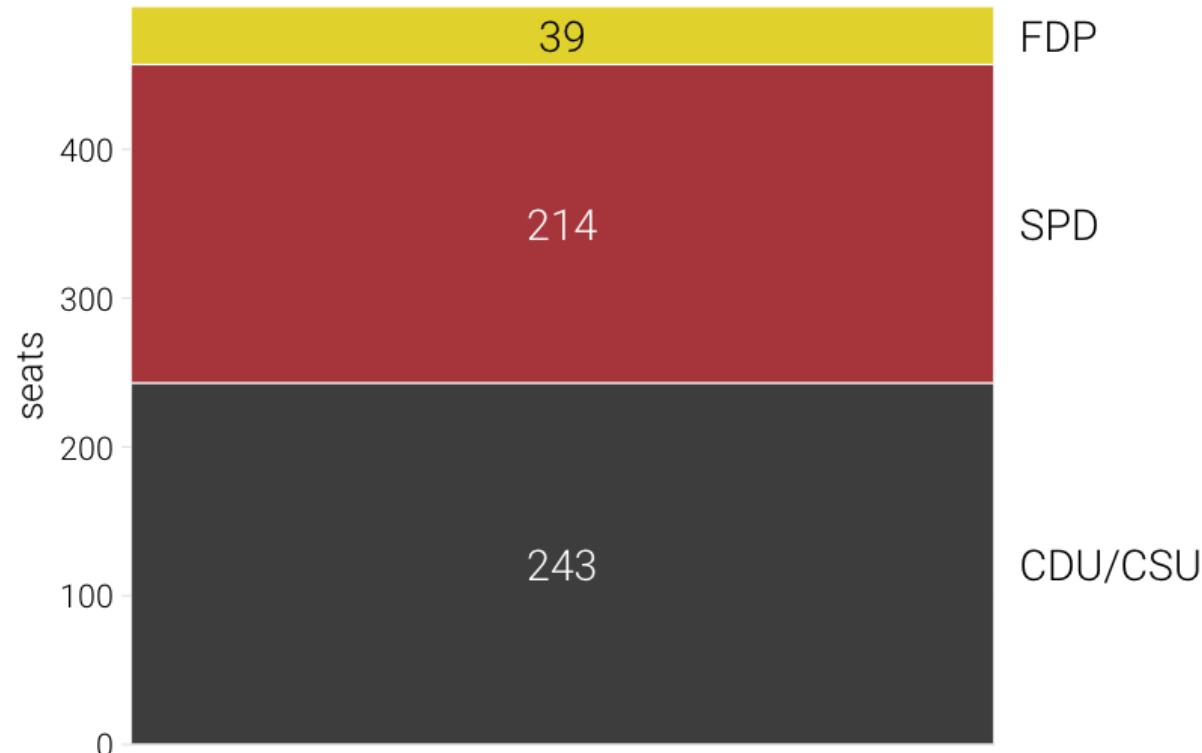
A case for pie charts

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

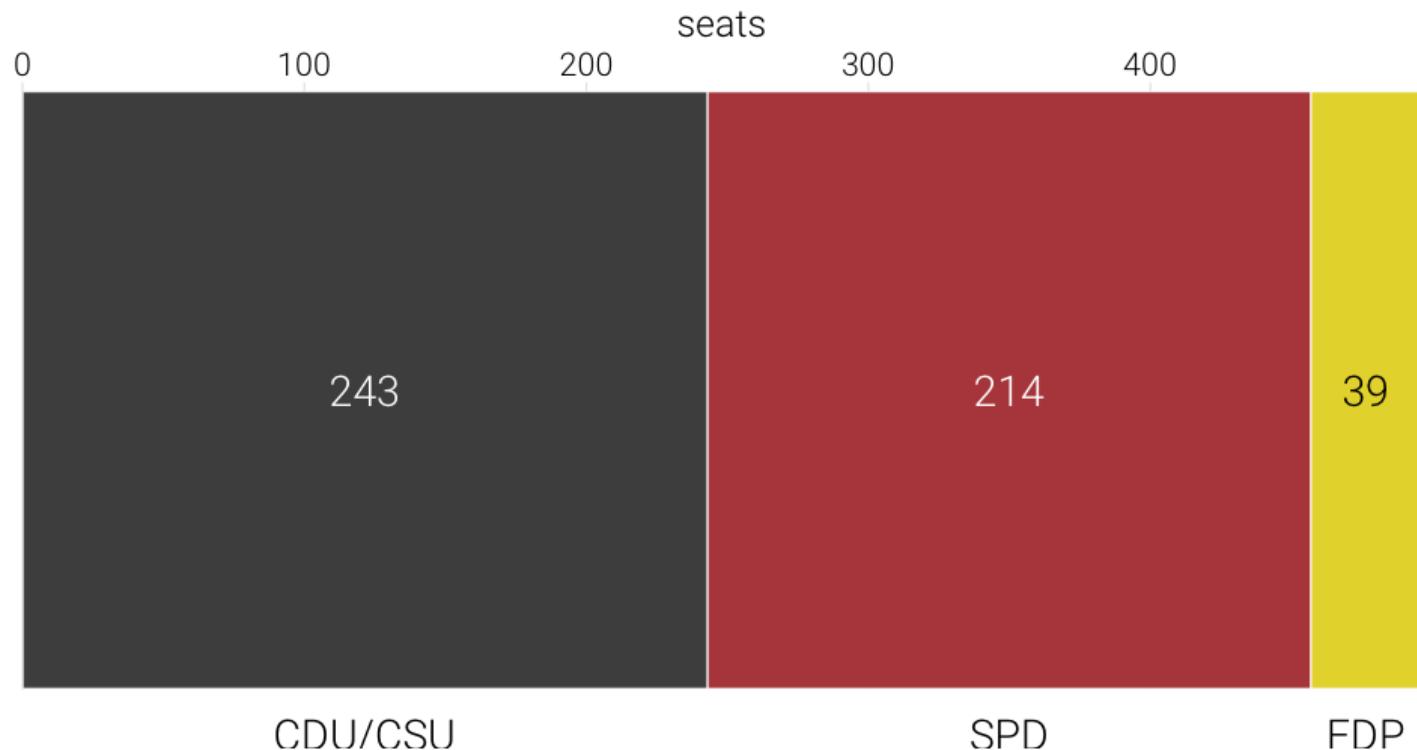


The anatomy of a pie chart

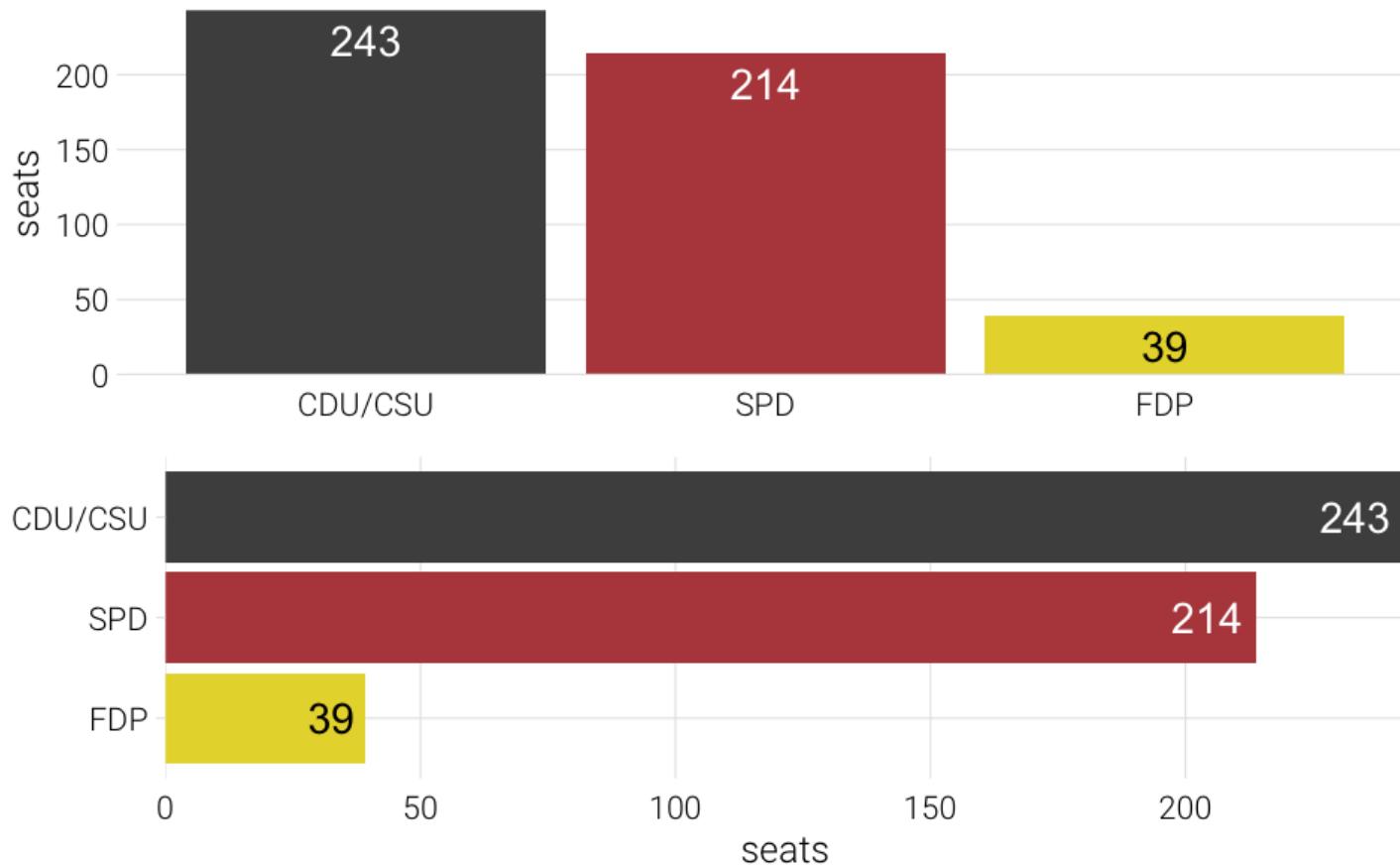
Pie charts are just stacked bar charts with a radial coordinate system



Alternative representation

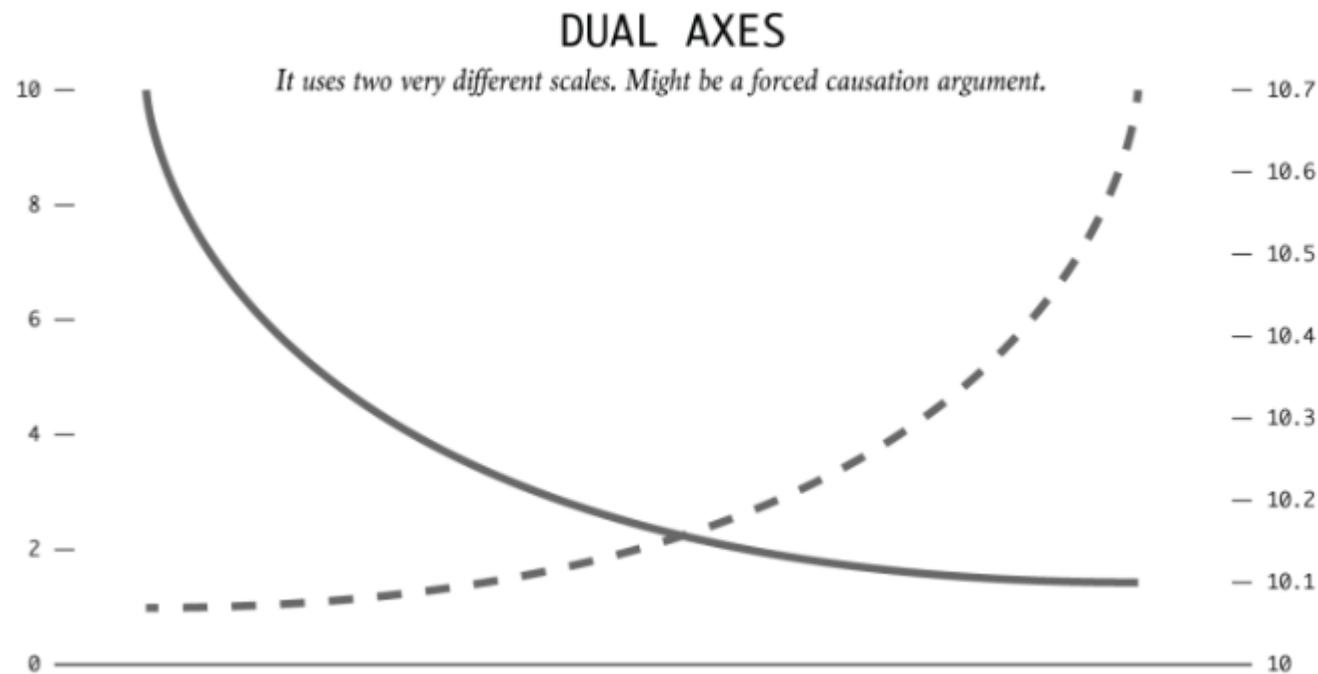


Or one of these

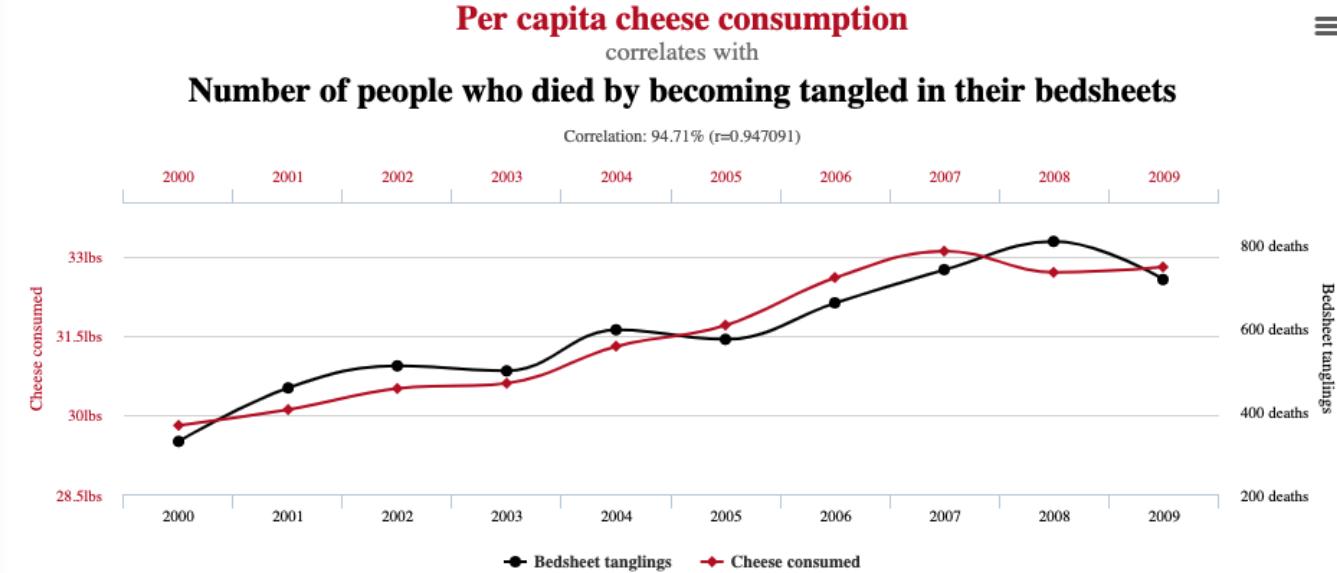


Dual axes

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



Another example

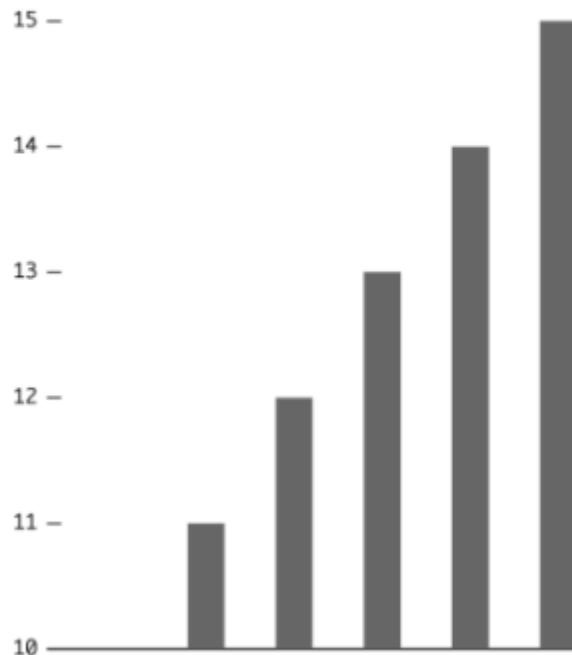


See more examples [here](#)

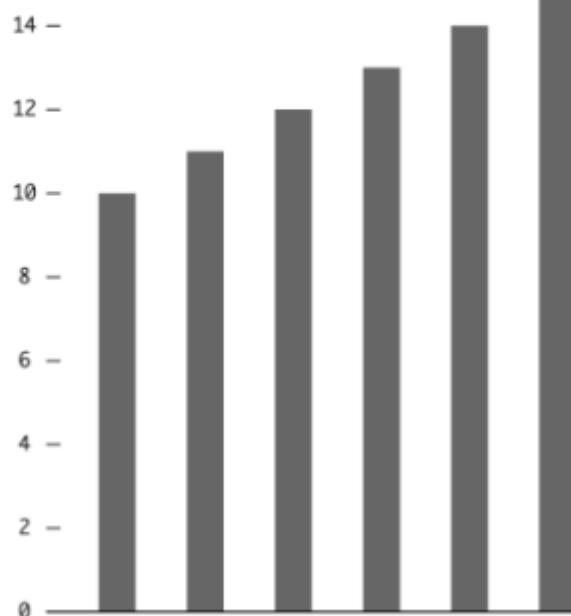
Truncated axes

TRUNCATED AXIS

The value axis starts at ten. Liar, liar, pants on fire.

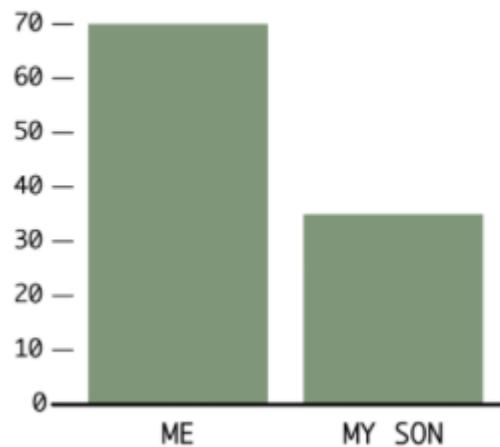


The value axis starts at zero. Good.



Height

INCHES



VS.

Height

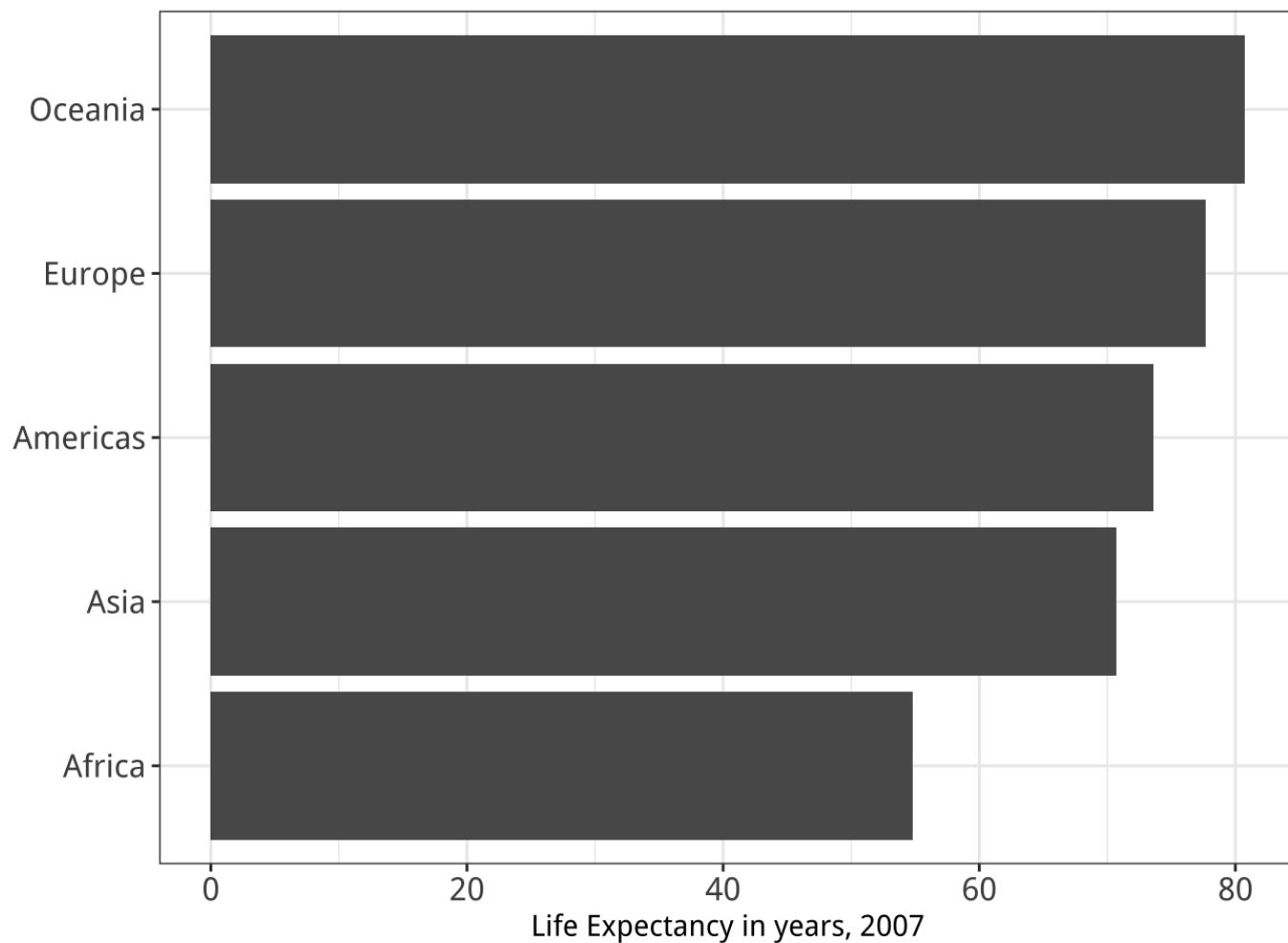
INCHES



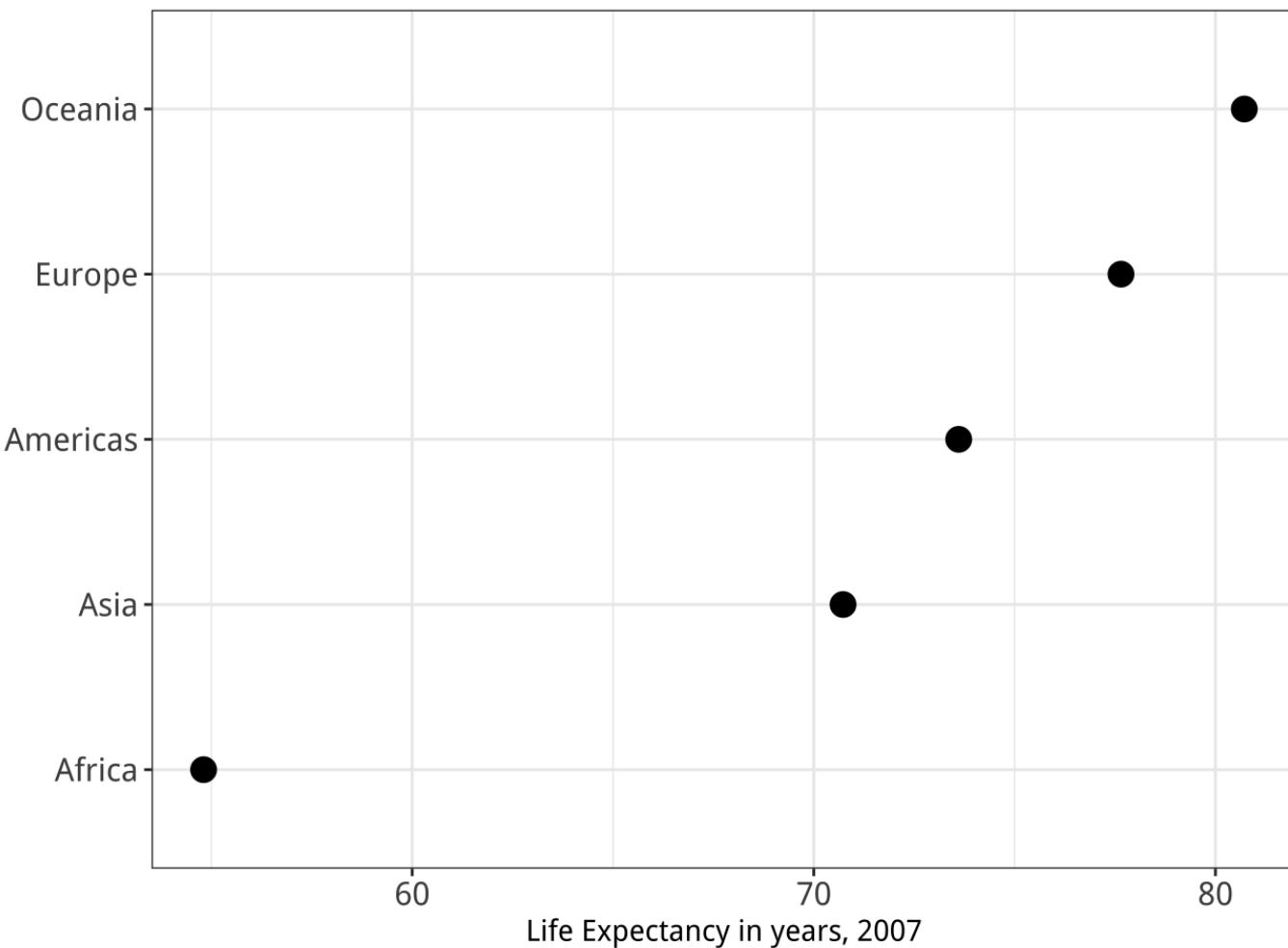
Not always a bad thing

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.

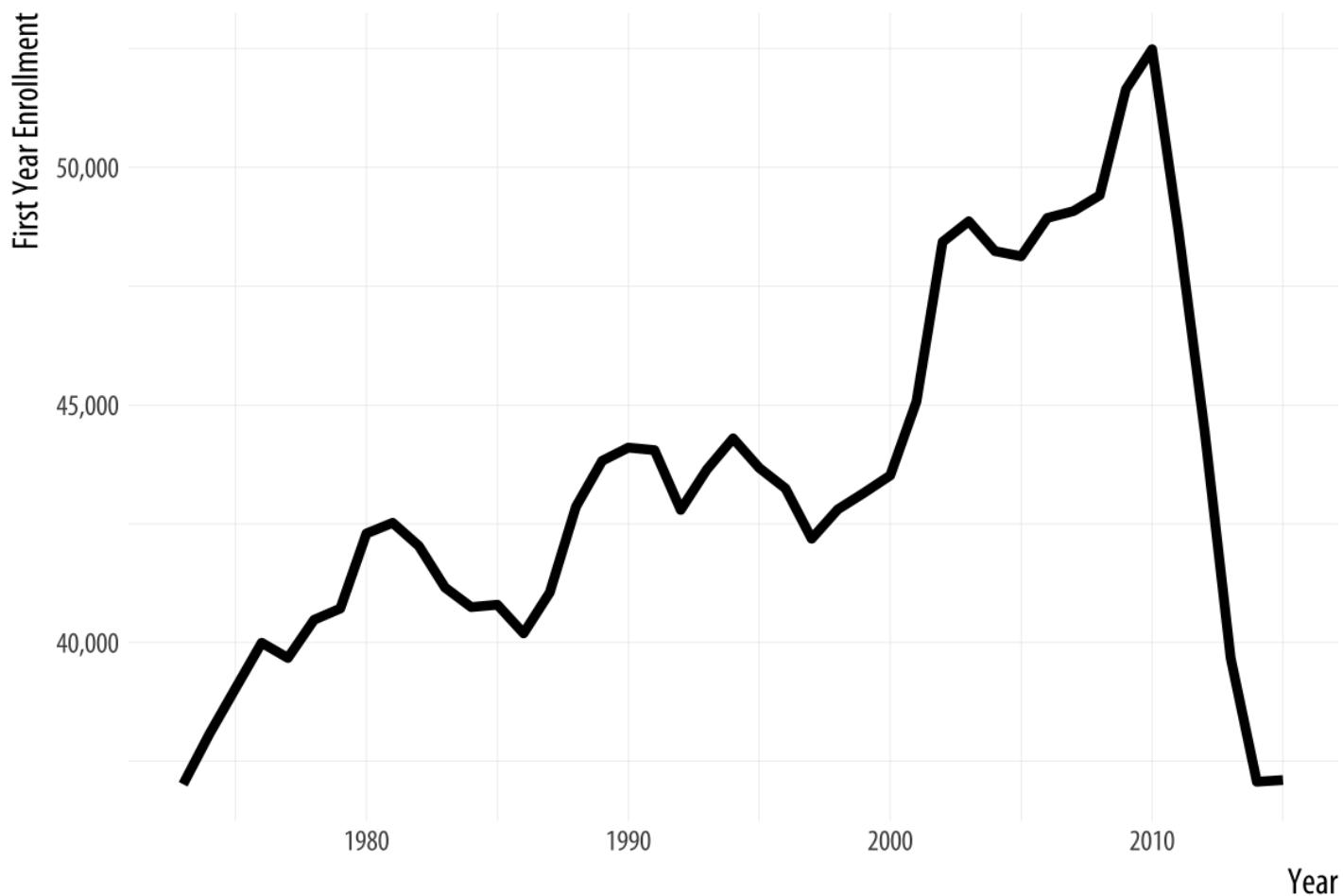
Bars



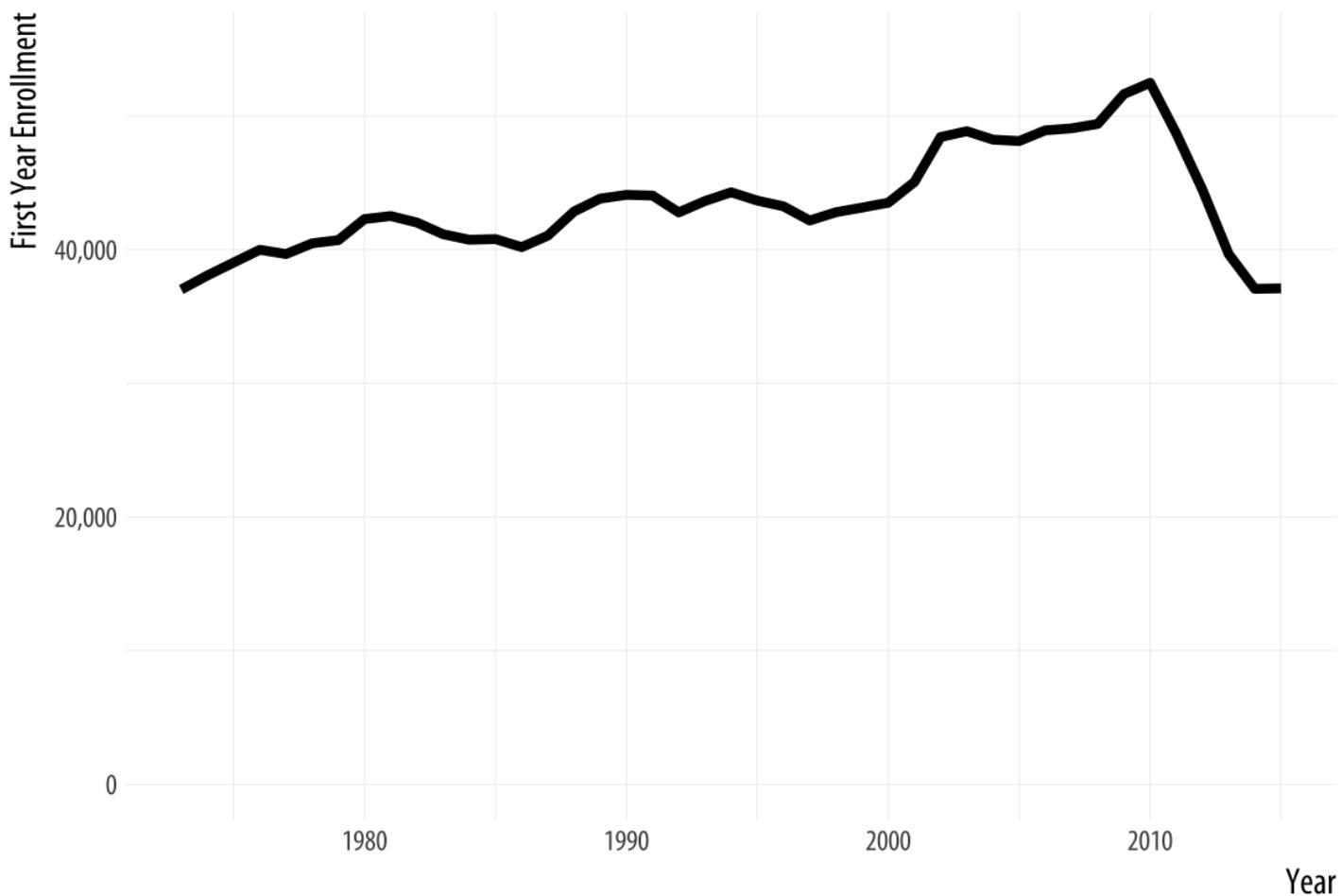
Points



Law school enrollments



Start at zero

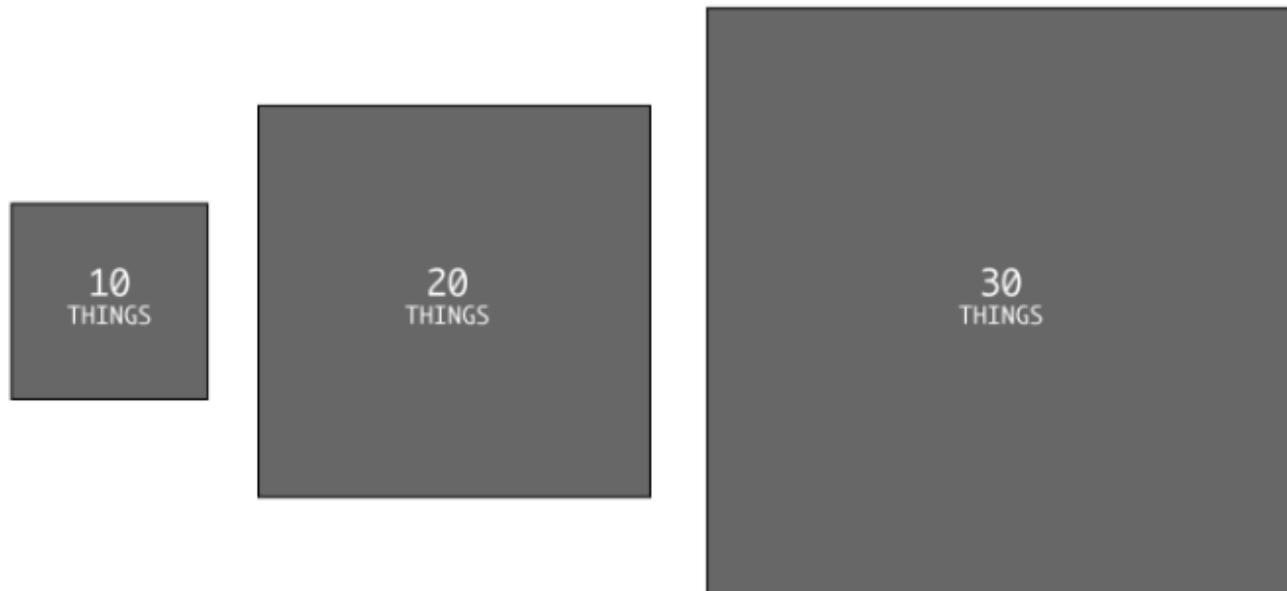


Scaling issues

AREA SIZED BY SINGLE DIMENSION

Thirty is three times ten, but that third rectangle looks a lot bigger than the first.

Might be trying to inflate significance.



Poor binning choices

ODD CHOICE OF BINNING

*Two bins. What's really in the 1+ category?
Might be hiding something.*



That's better. It can show more variation.



Conclusions

Essentially never

- Use dual axes (unless they are direct transformations, just produce separate plots instead)
- Use 3D unnecessarily

Be wary of

- Truncated axes

Do

- Minimize cognitive load
- Be as clear as possible

Lab 2
