Communication through Pictures

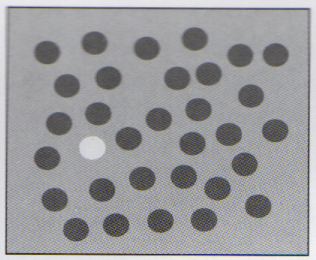
Data Visualization

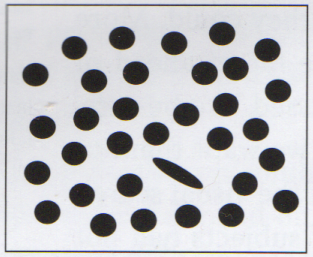
Dr. Andrew Hamilton-Wright

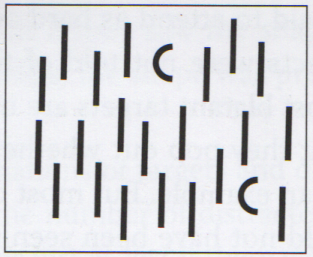
School of Computer Science
University of Guelph
https://github.com/andrewhw/VizThoughts

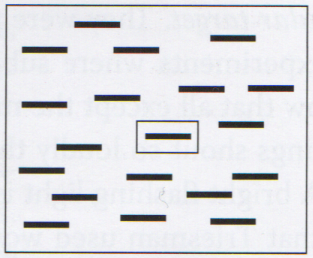
2023-06-26

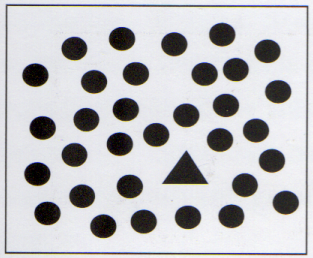
What we can perceive

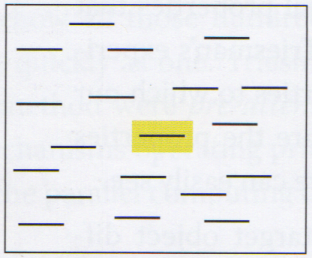

Grey value

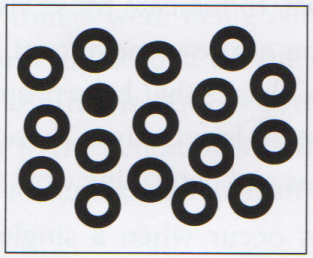

Elongation

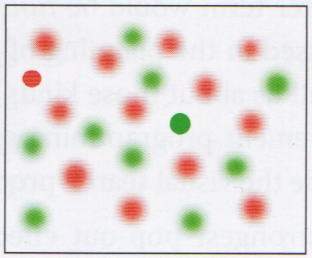

Curvature

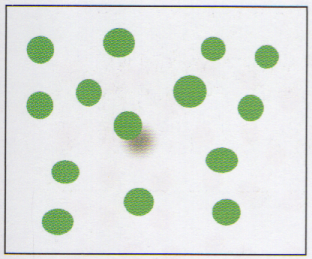

Added surround box

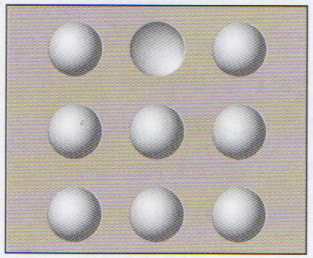

Shape

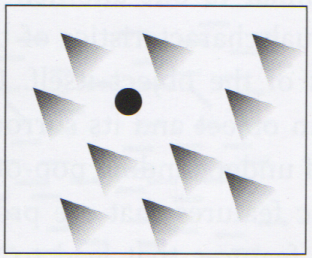

Added surround color

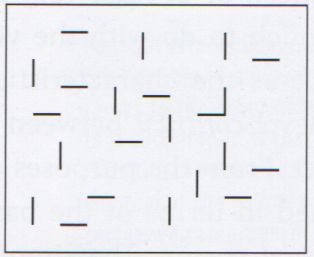

Filled


Sharpness

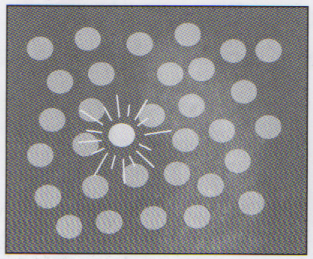

Cast shadow

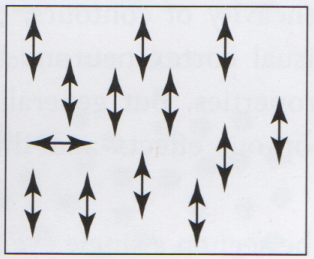

Convex and concave

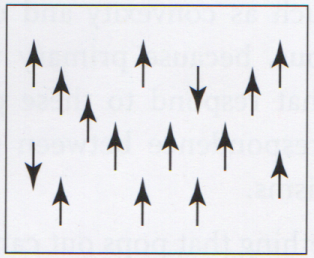

Sharp vertex


Joined lines


Misalignment

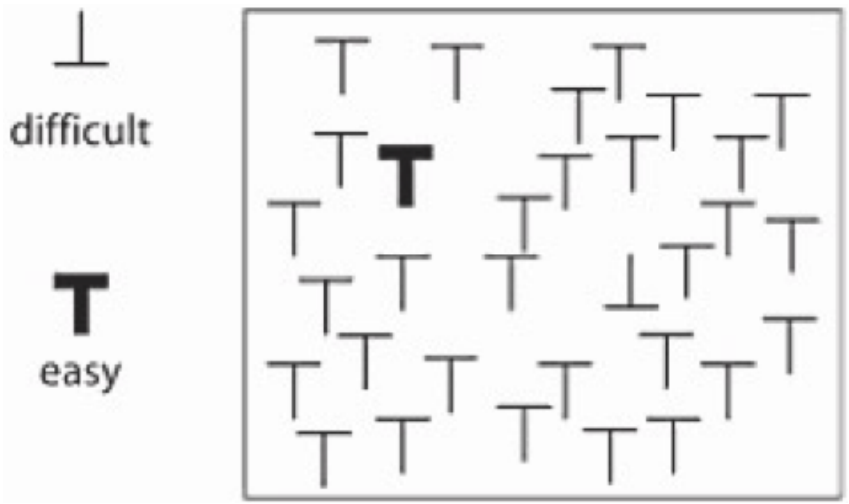

Blinking

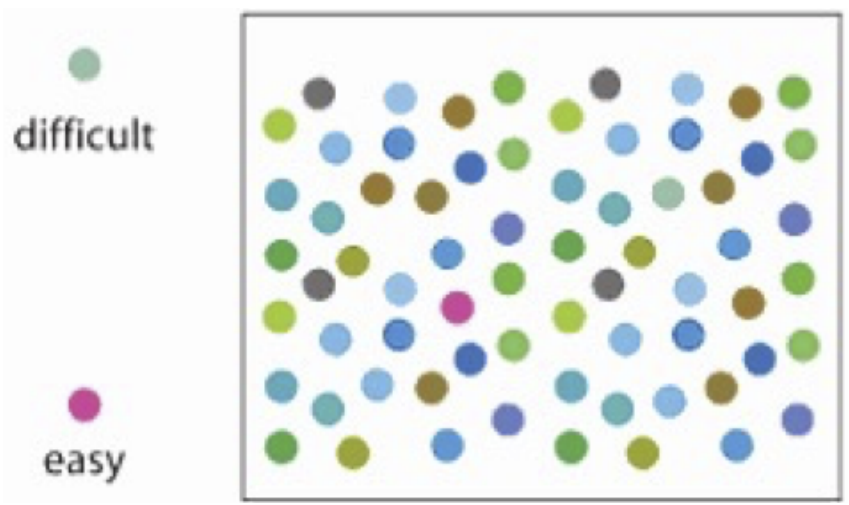

Direction of motion

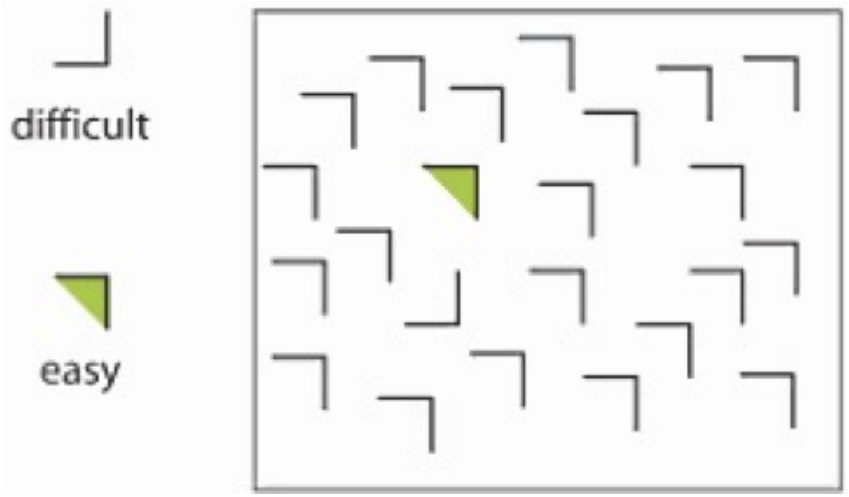

Phase of motion

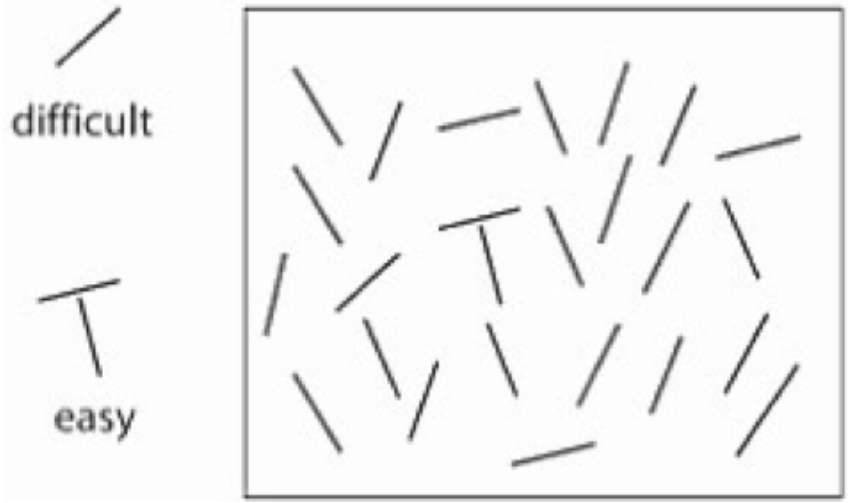
Colin Ware (2008): *Visual Thinking for Design*, Morgan Kaufmann. ISBN: 978-0-12-370896-0

What we can perceive - Easy and Hard



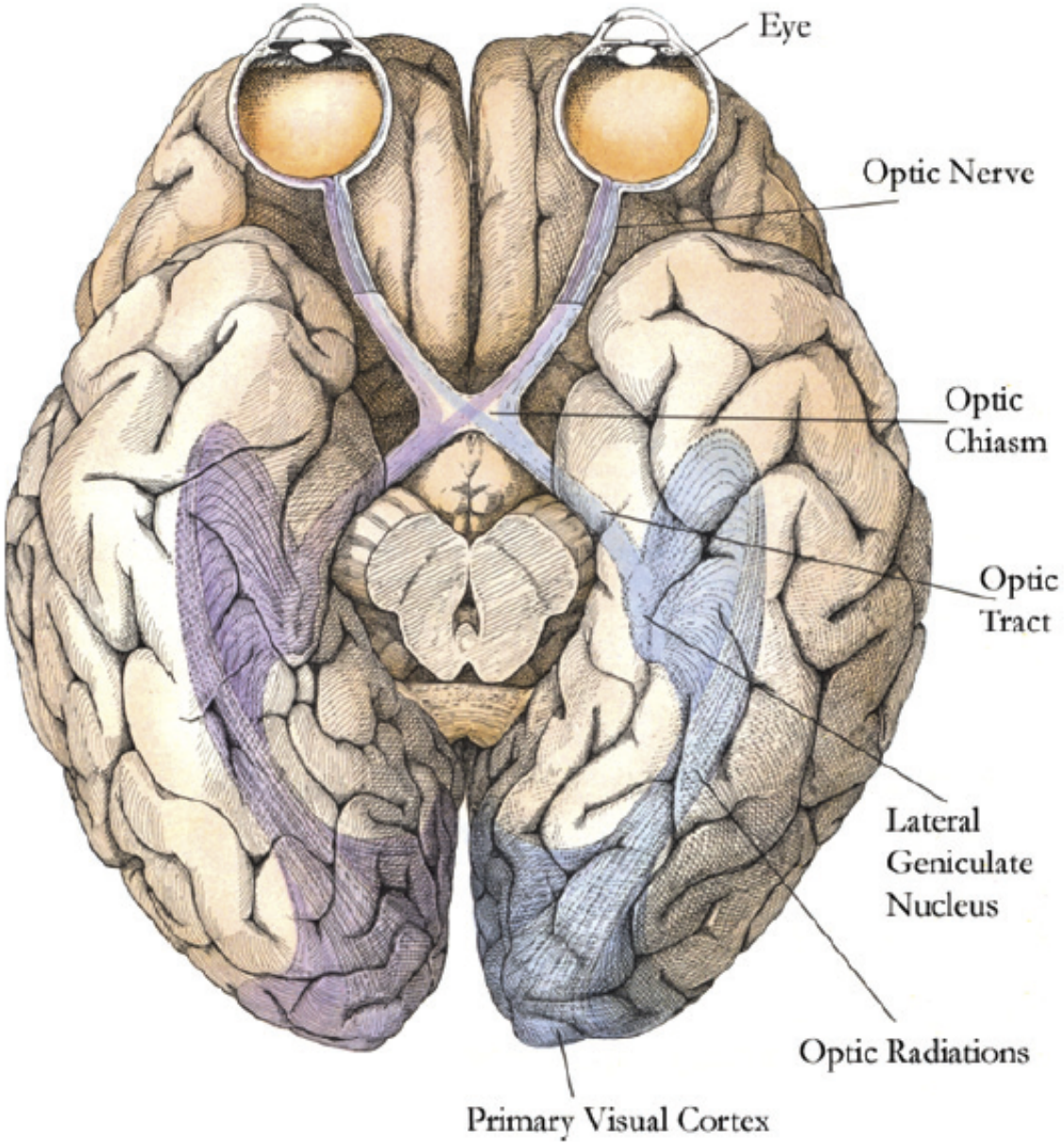






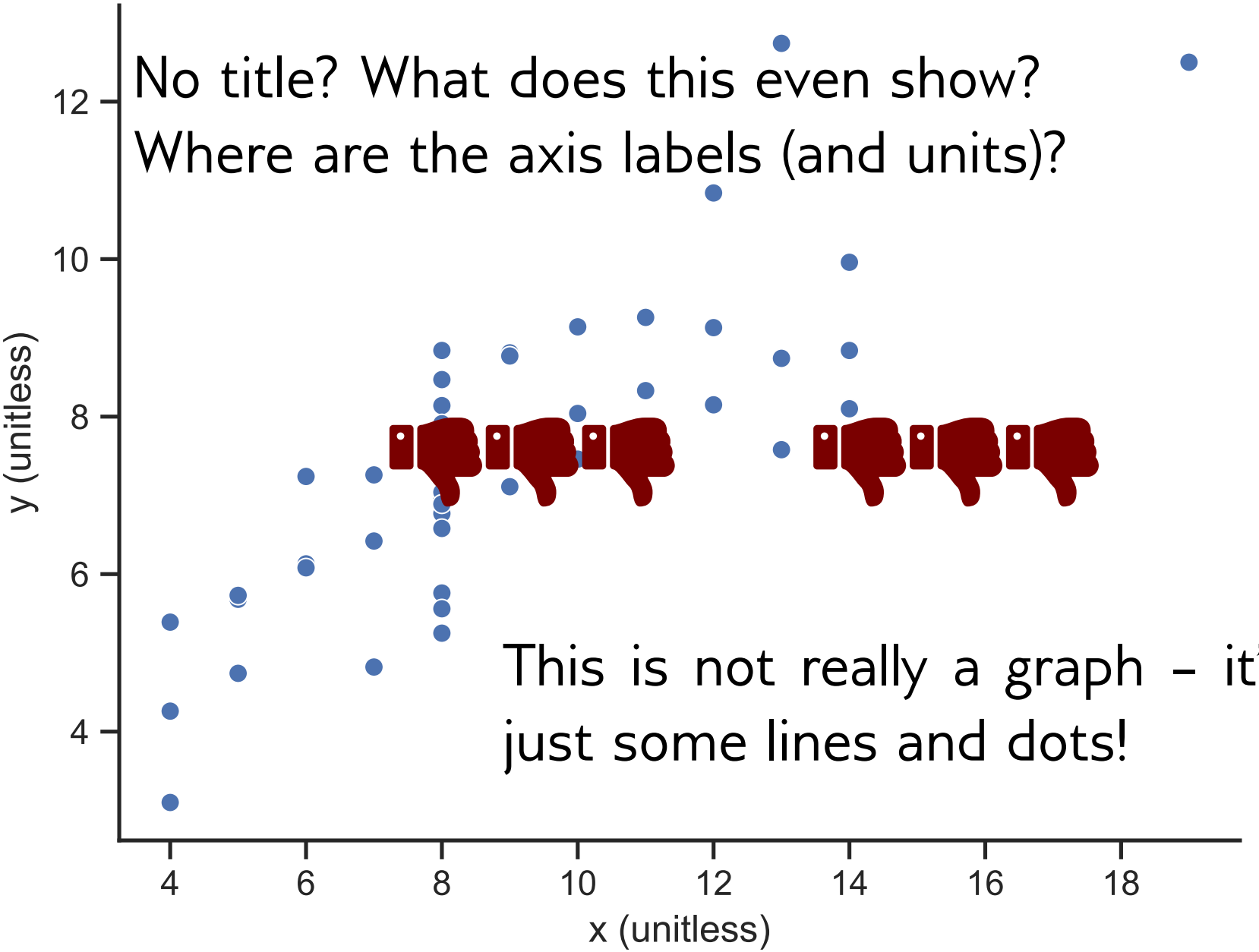
*Ware (2008): *Visual Thinking for Design*

We must design for the perceptual system we have

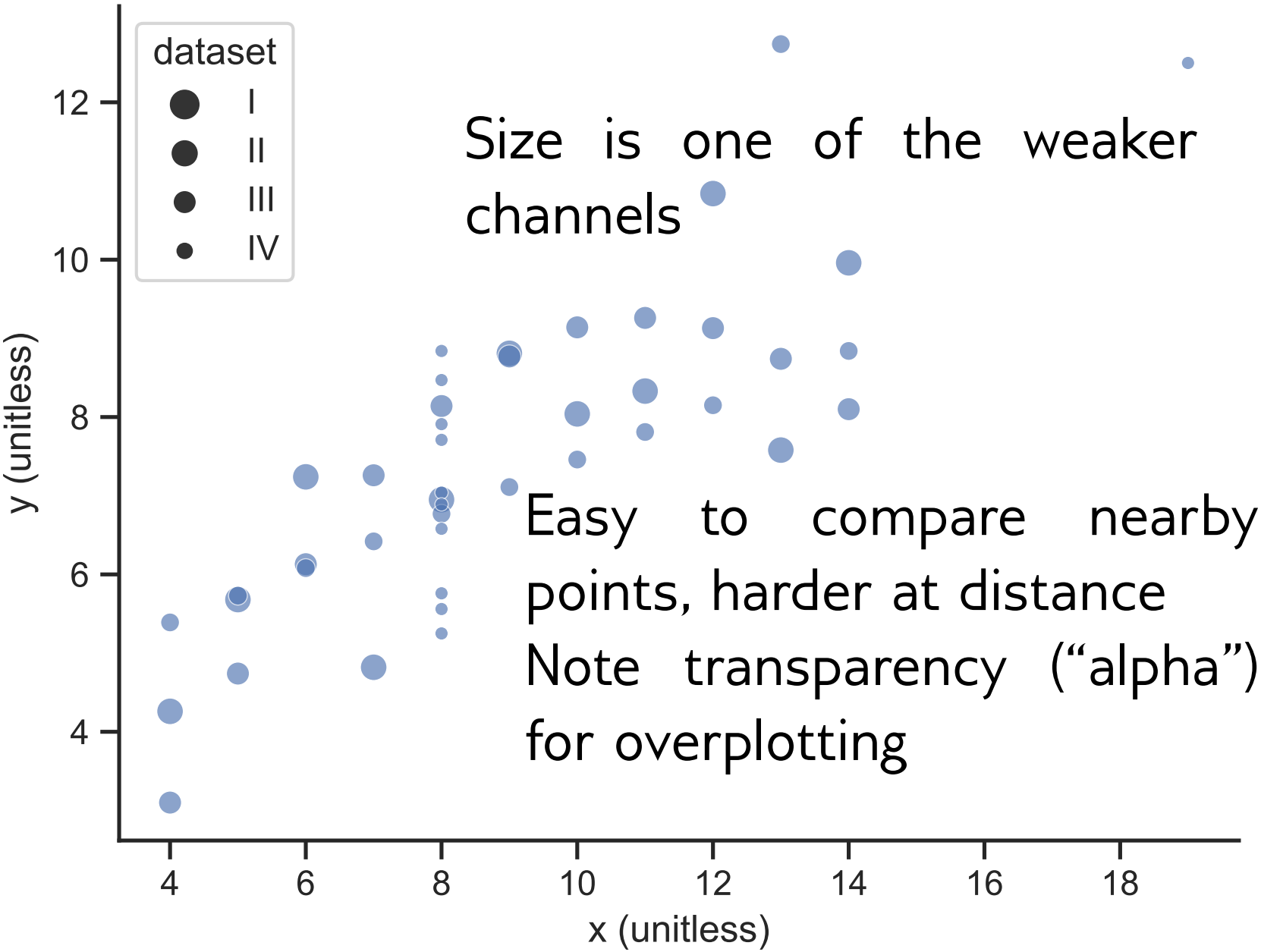


Ware (2008): *Visual Thinking for Design*

A graph?


No title? What does this even show?
Where are the axis labels (and units)?
This is not really a graph – it's just some lines and dots!
y (unitless)
x (unitless)
4
6
8
10
12
14
16
18
4
6
8
10
12

Anscombe's data sets - size indicates data set

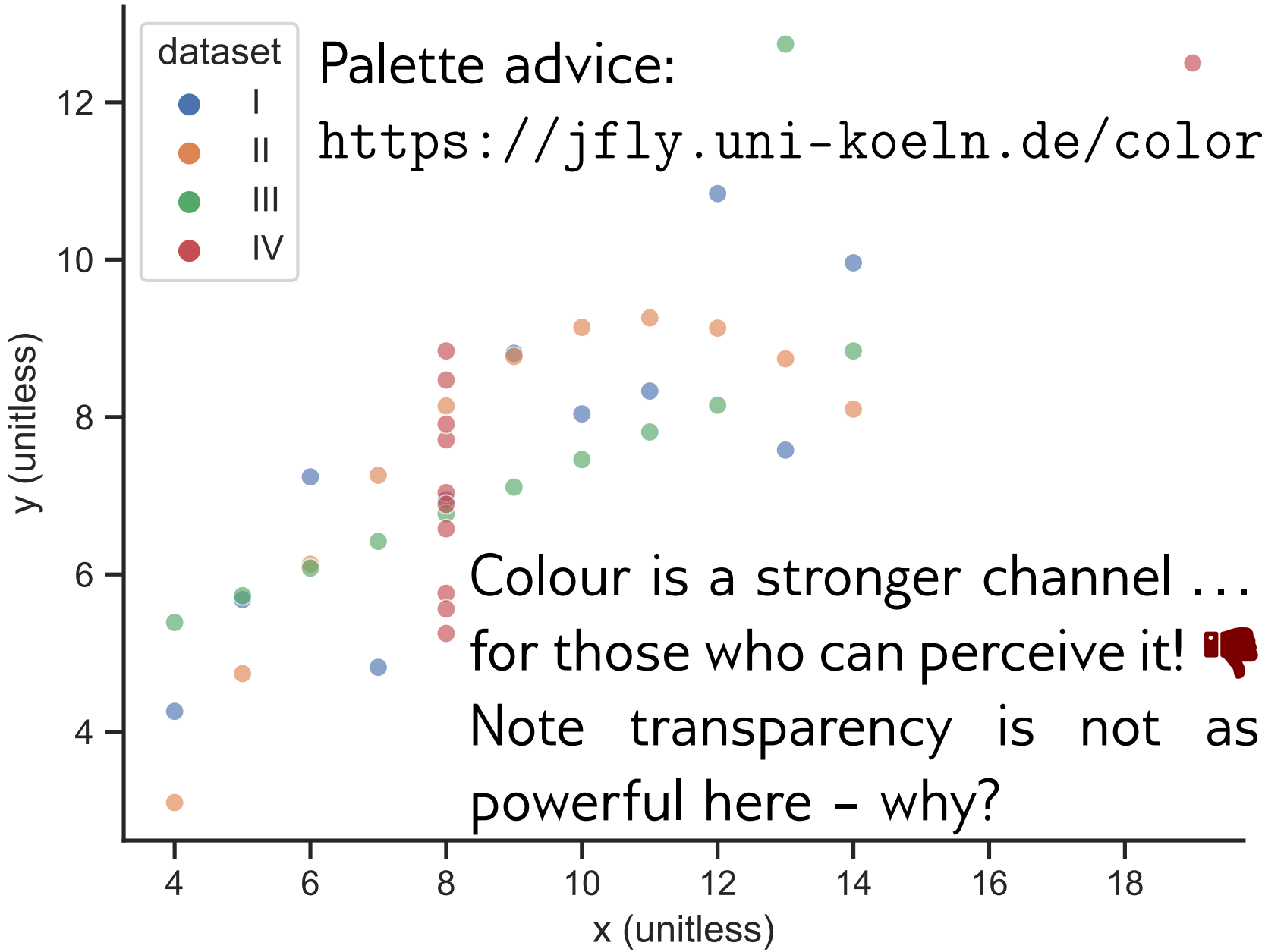


Size is one of the weaker channels

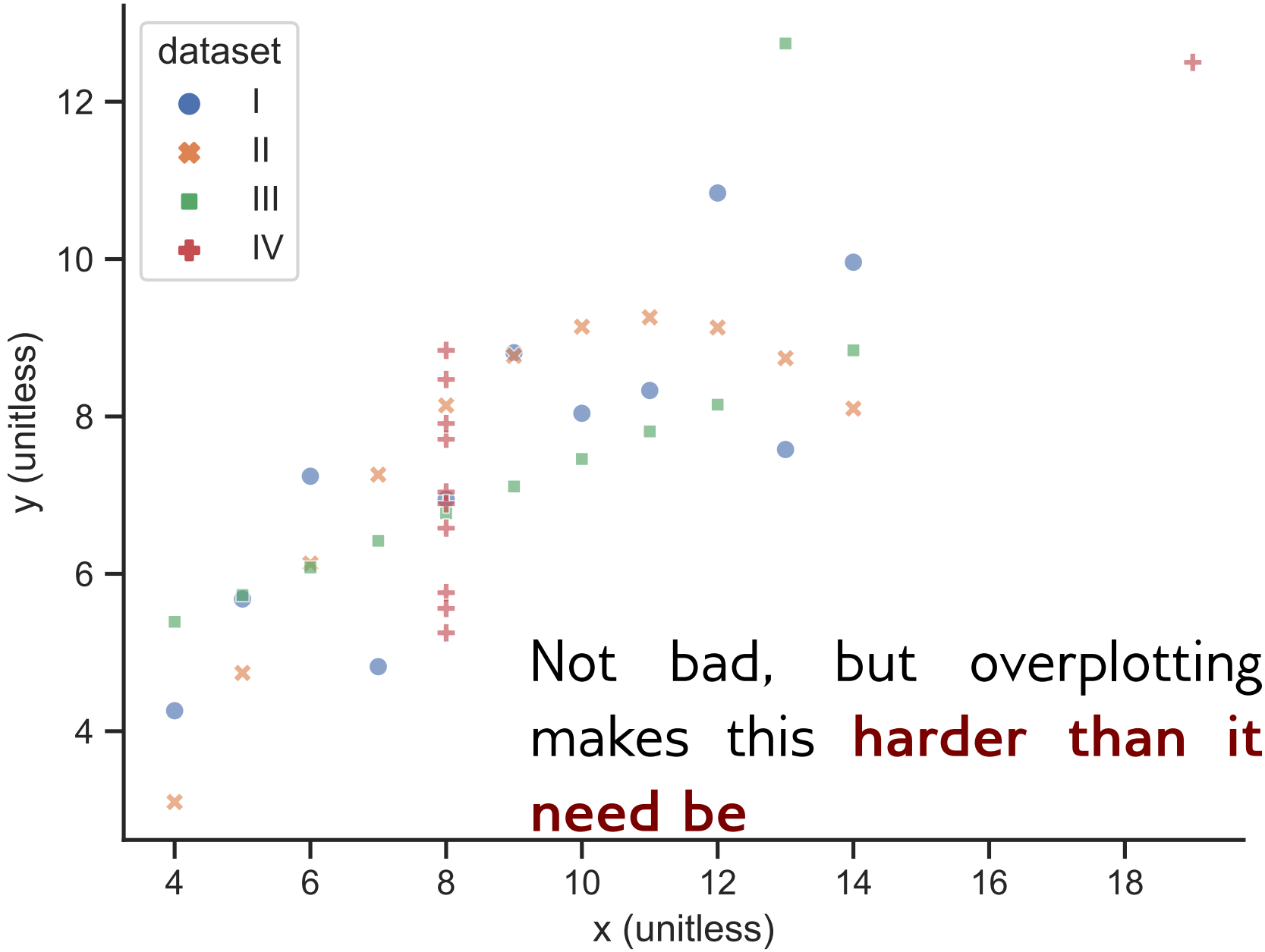
Easy to compare nearby points, harder at distance

Note transparency ("alpha") for overplotting

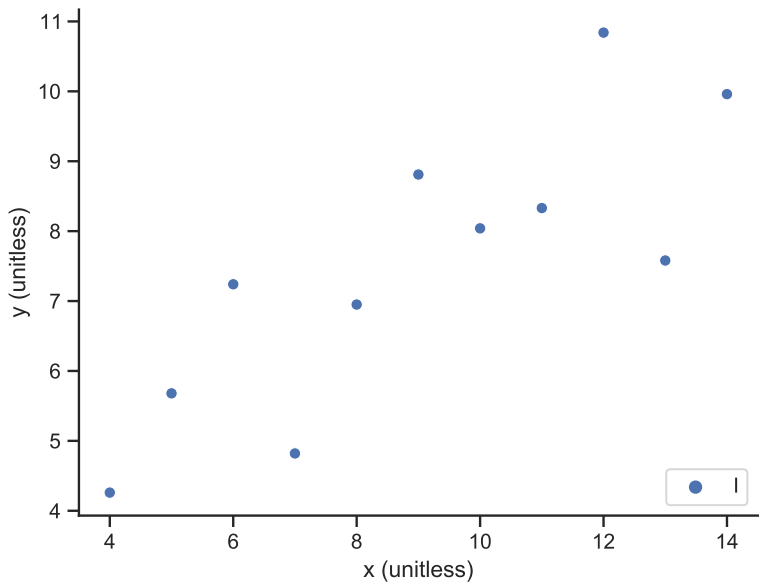
Anscombe's data sets - hue indicates data set

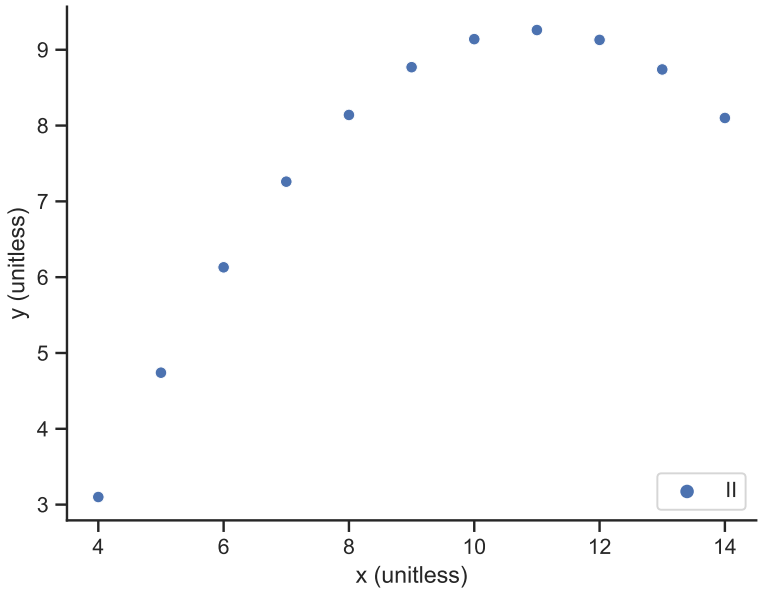


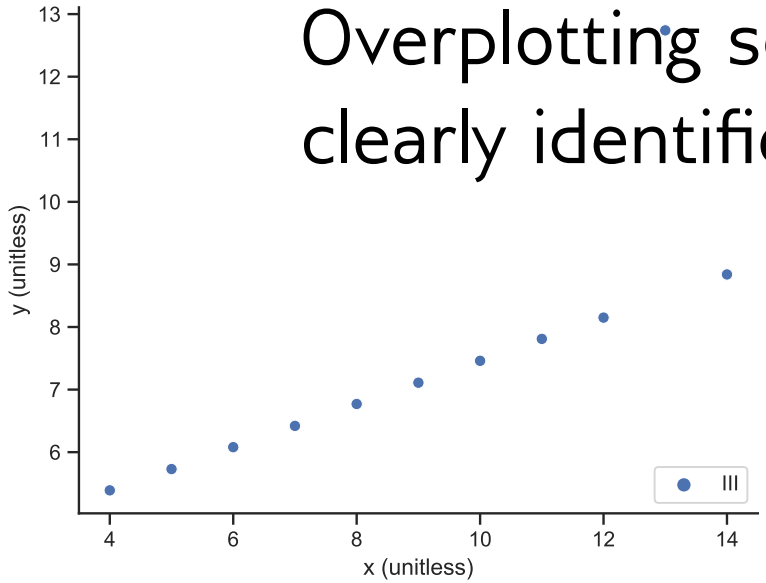
Anscombe's data sets - hue and style indicate data set

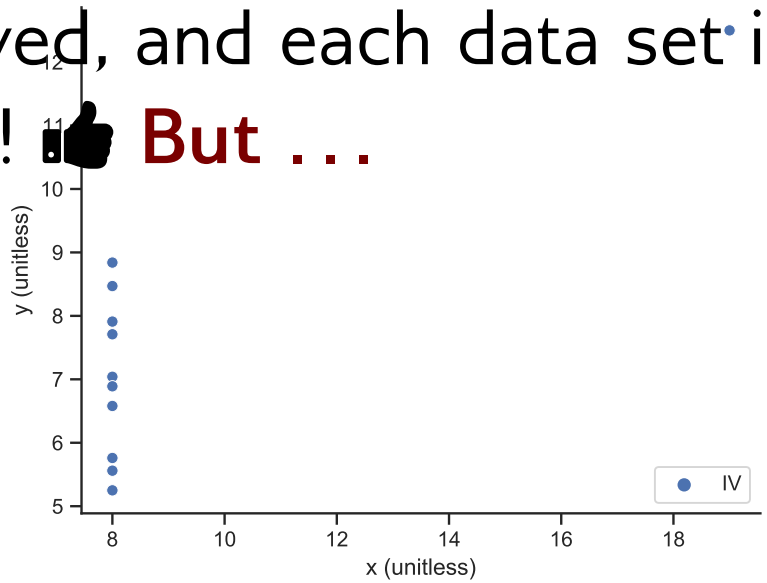


Anscombe's data sets – separate plots, one figure



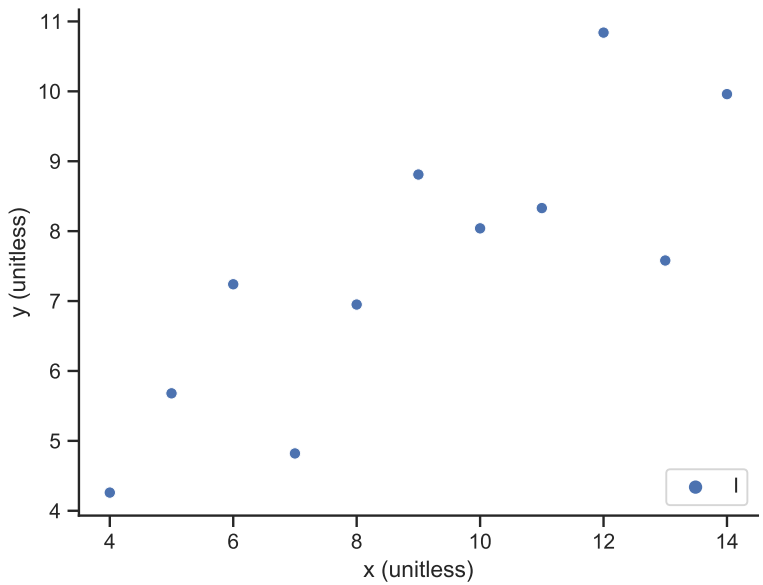


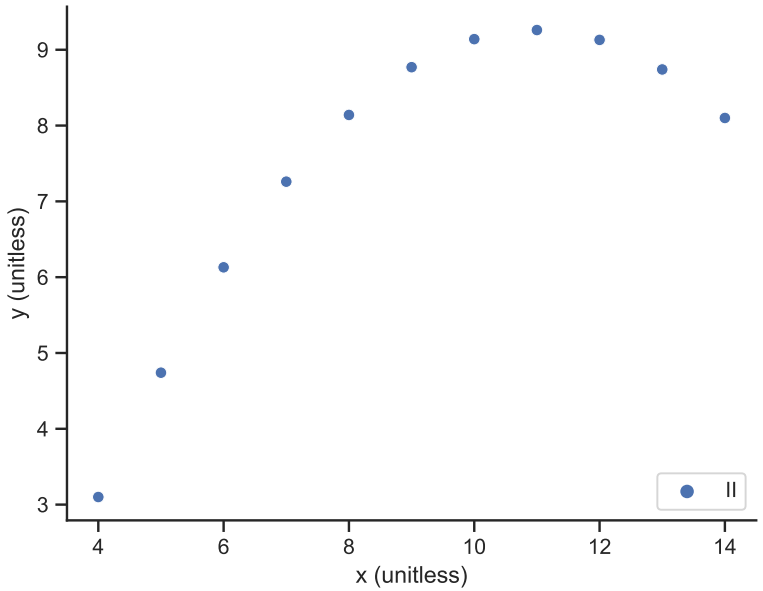


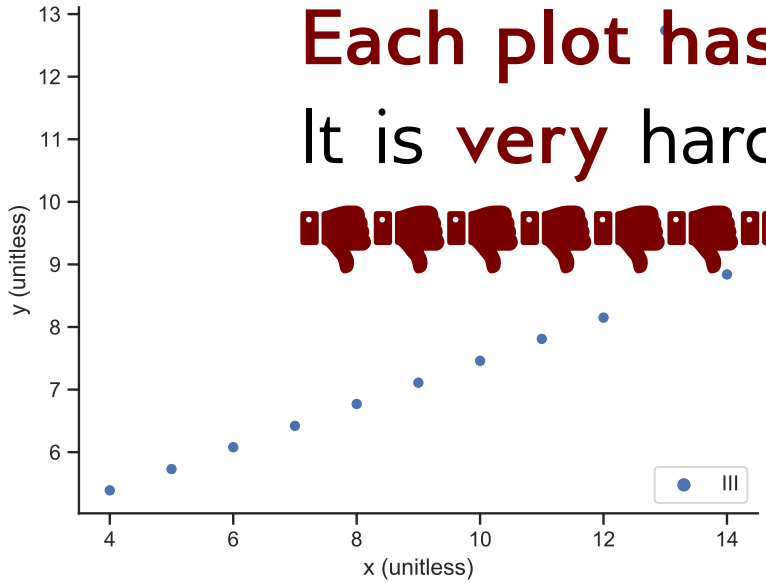


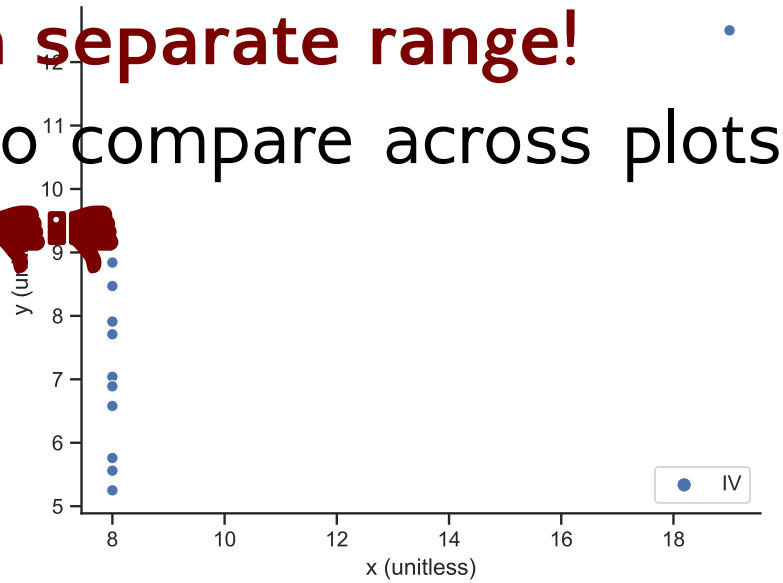
Overplotting solved, and each data set is clearly identified! 👍 **But . . .**

Anscombe's data sets – separate plots, one figure





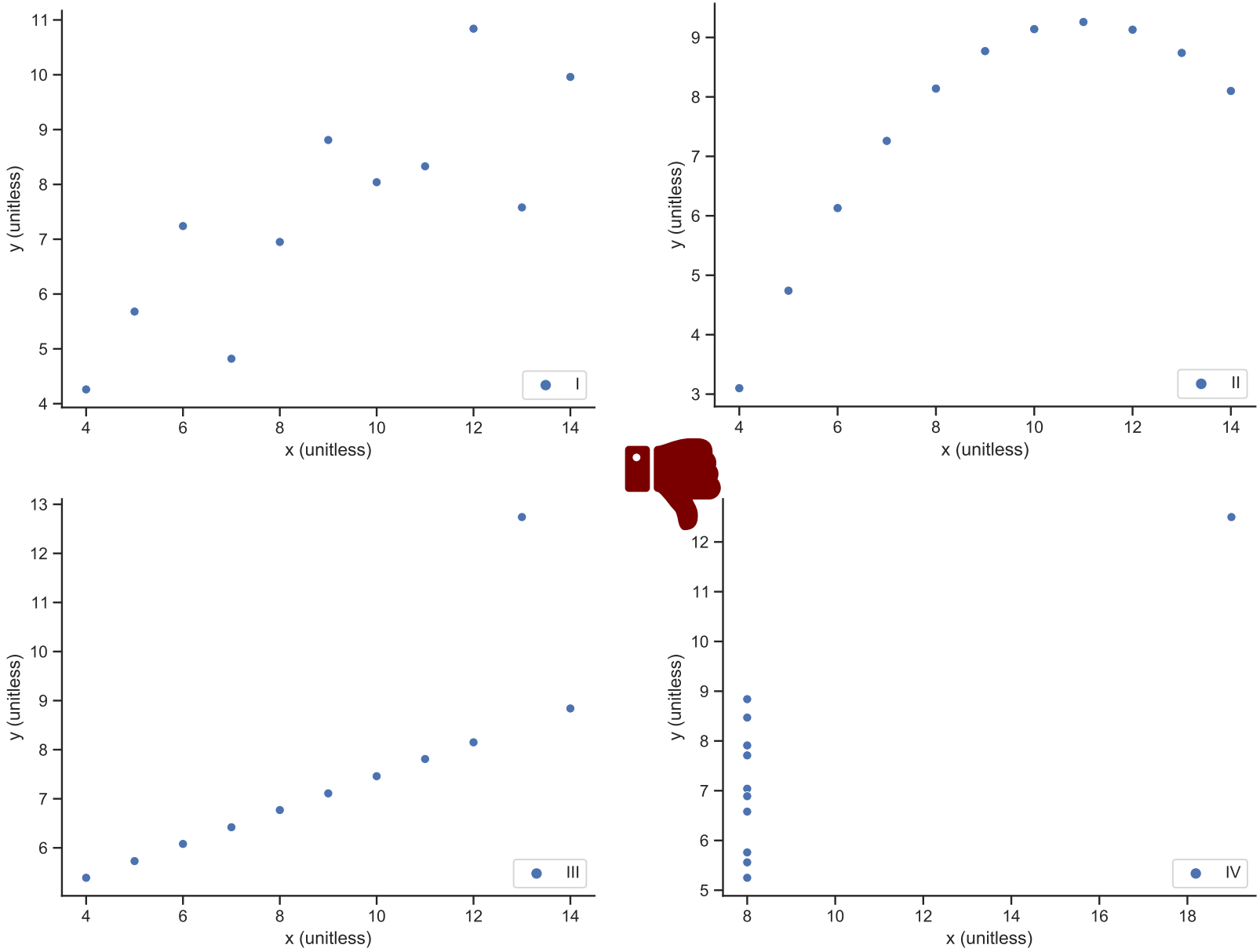




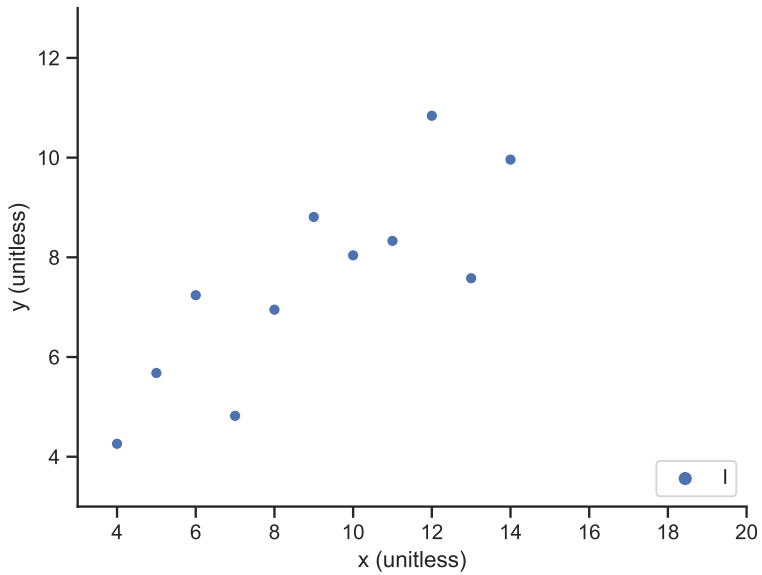
Each plot has a separate range!

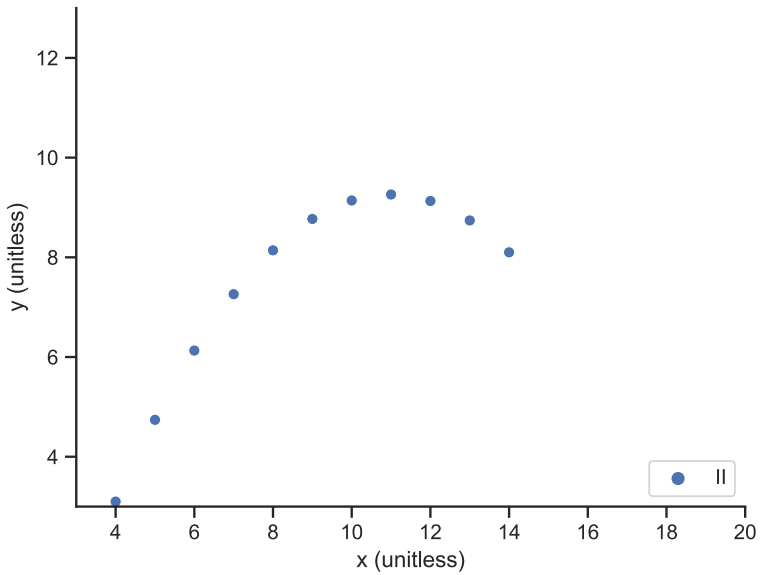
It is **very** hard to compare across plots.

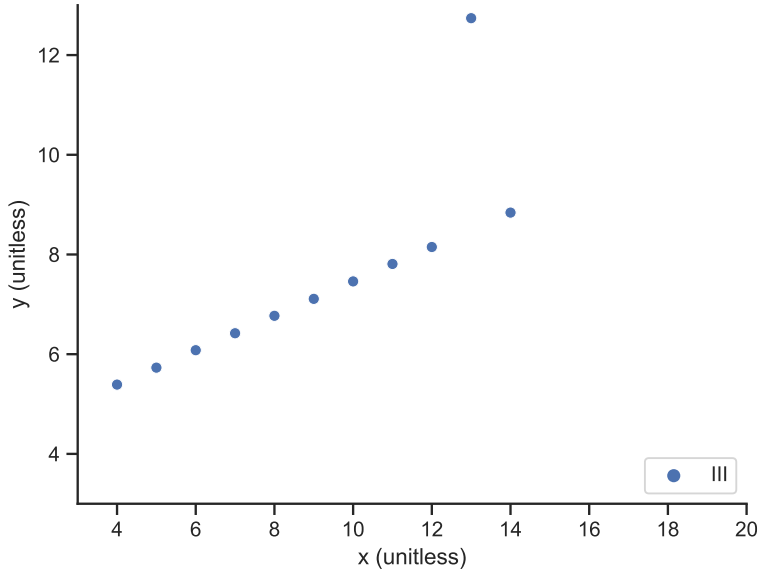
Anscombe's data sets – from this...

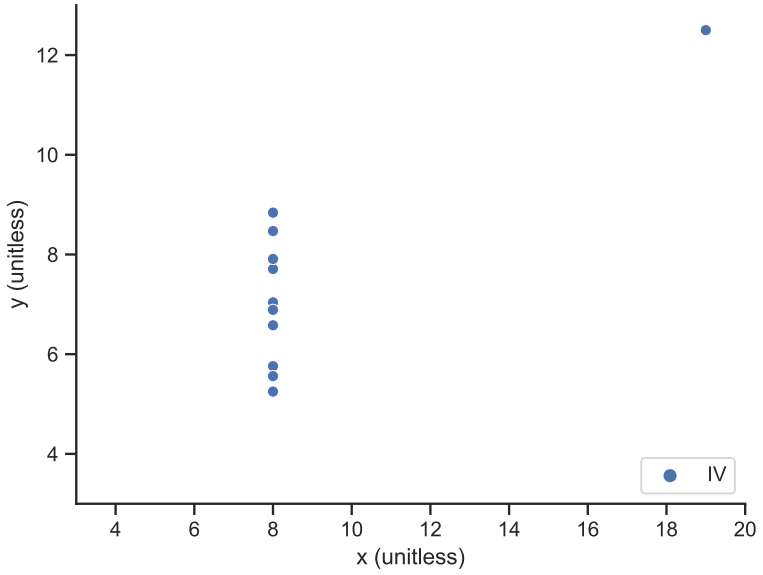


Anscombe's data sets – to this! Axes fixed!

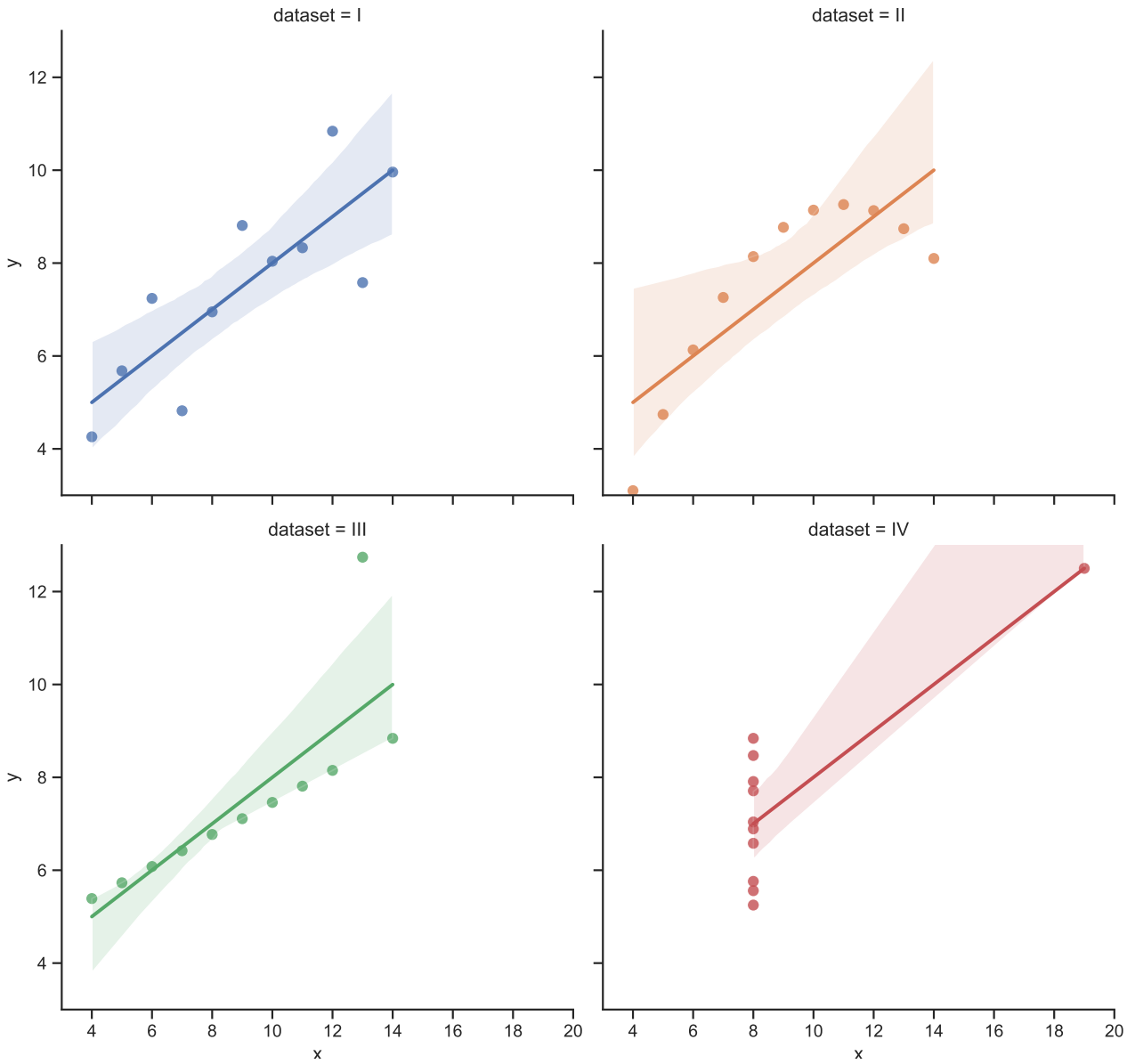








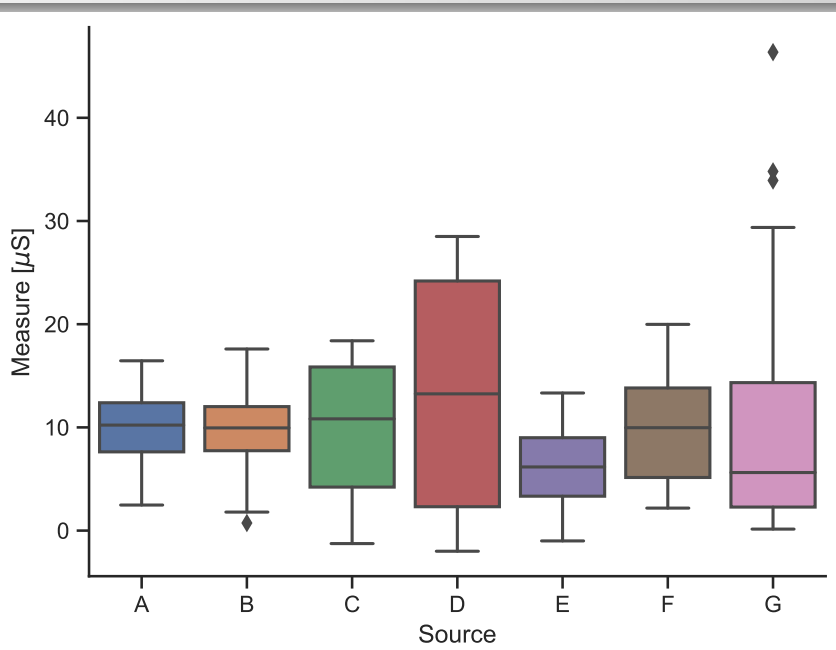
Anscombe's data sets - linear model plot

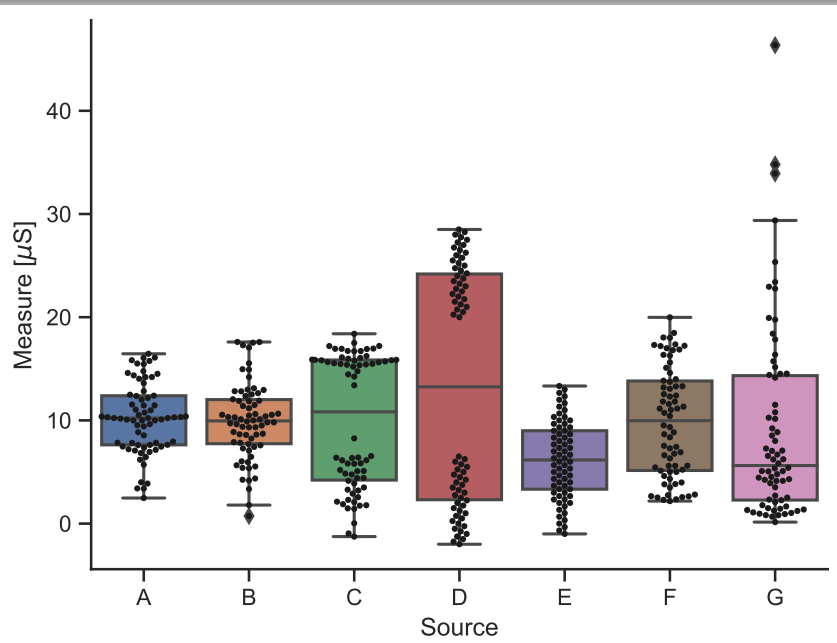



```
sns.lmplot(x="x",
y="y",
col="dataset",
hue="dataset",
data=df,
col_wrap=2)
```

- common axes? 👍
- id data sets? 👍
- colour as highlight only 👍

Boxplots



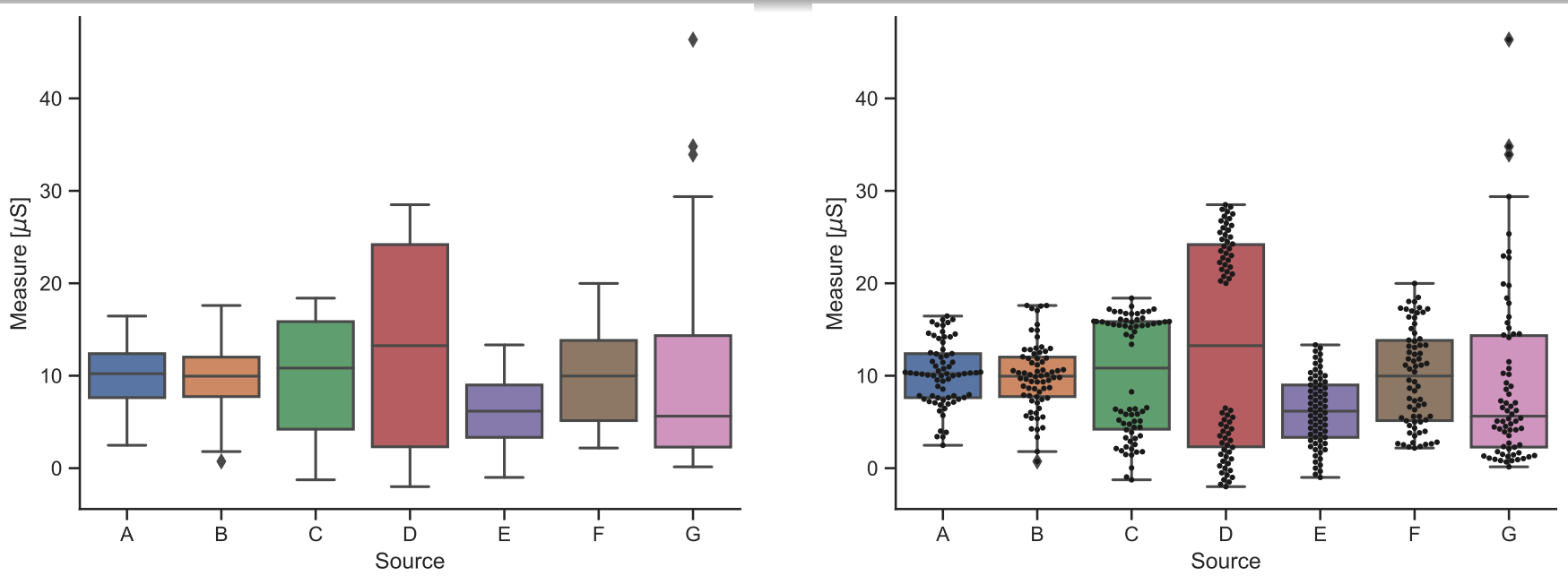


- A boxplot provides a "5 number summary": median, lower/upper quartiles, min/max. Are these useful for **your** data?
- A boxplot will encourage you to think of your data as centrally tended, even when it is not.

Always plot the points to see what is going on.

```
sns.boxplot(x="Source", y="Measure", data=df)
sns.swarmplot(x="Source", y="Measure", data=df, color="0.25")
```

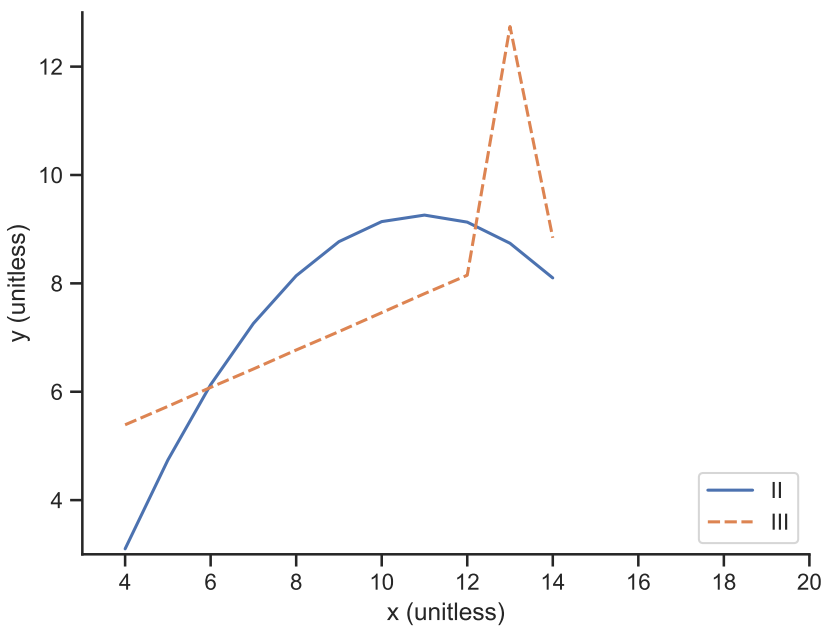
Boxplots and whiskers

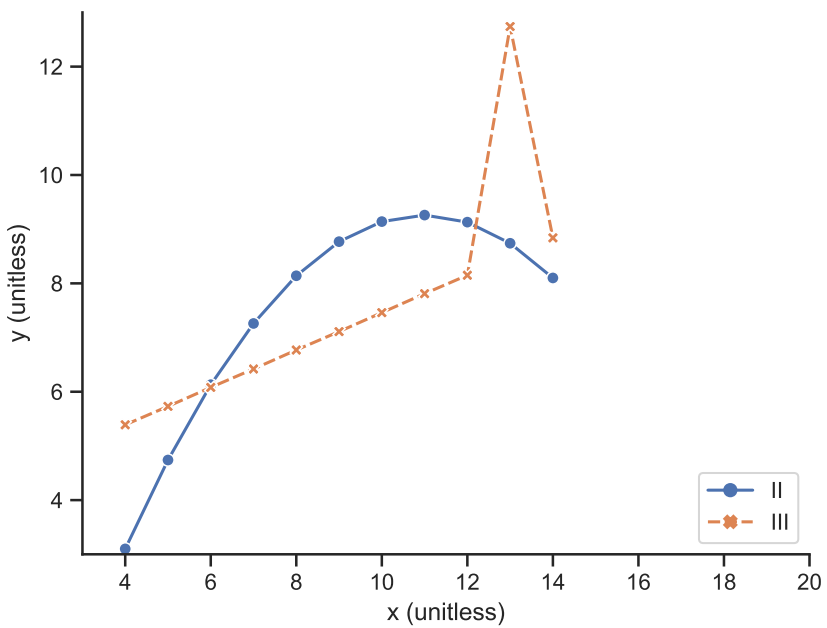


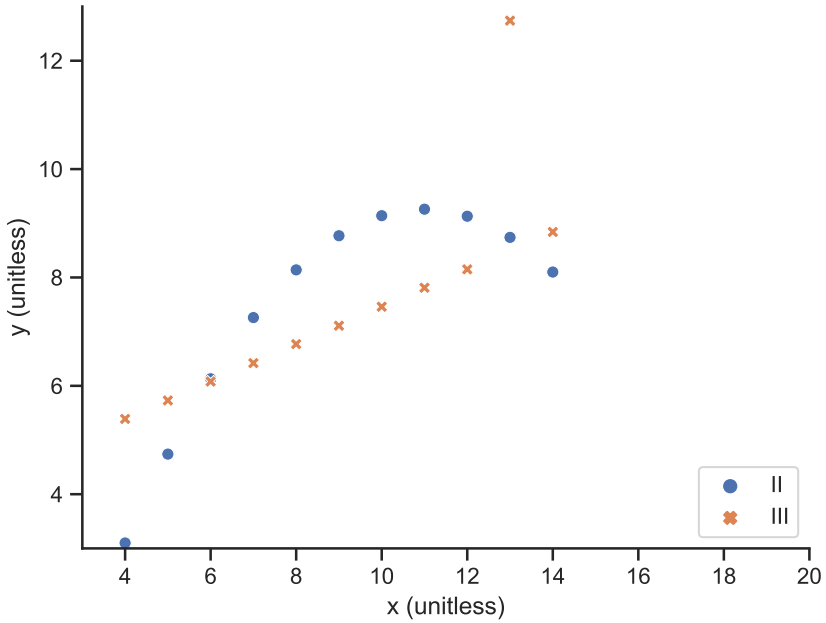
- Make sure you know where your whiskers go to!
- **R/matplotlib(seaborn)**: box=quartiles, whisker = furthest point within 1.5 of quartile length (on that side)
- **Excel**: default = whisker to max/min
- **Others**: various, and commonly 2 standard deviations

Be sure you know what your package does!

Scatters and Lines

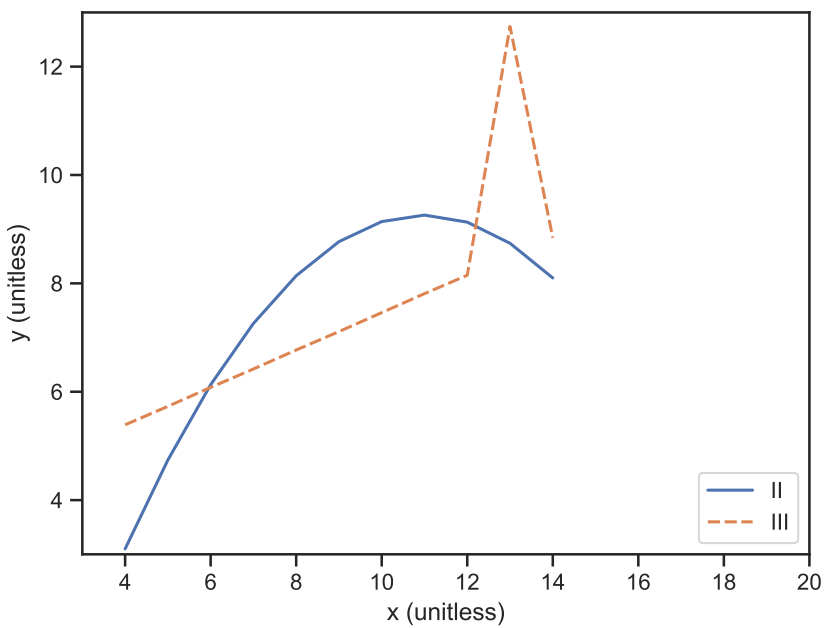


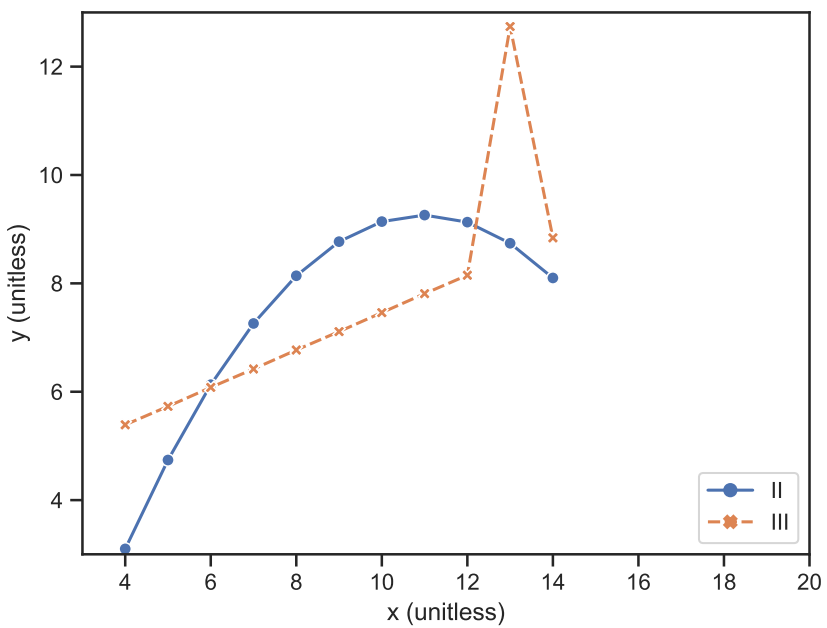


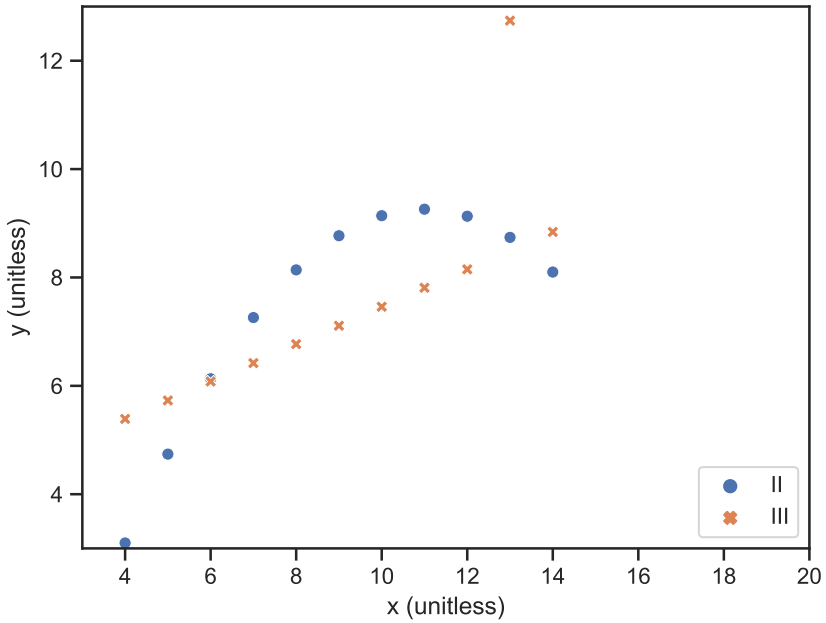


- lines make association between points much more visible, but …
- lines imply a series – don't use them if there isn't one
- don't hide your sample points

Scatters and Lines

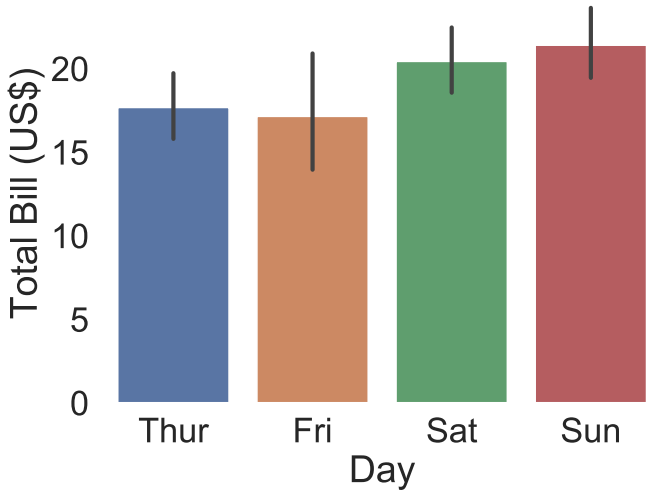






- note that the "data box" makes the outlying $\times$ point easier to see as it is distinguished from the plot above ...

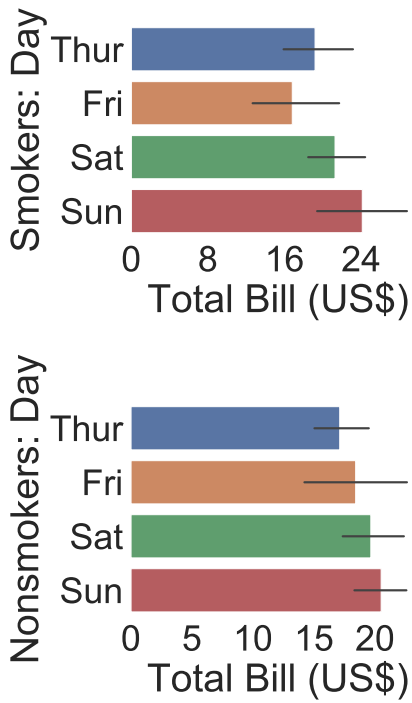
Categorical Data



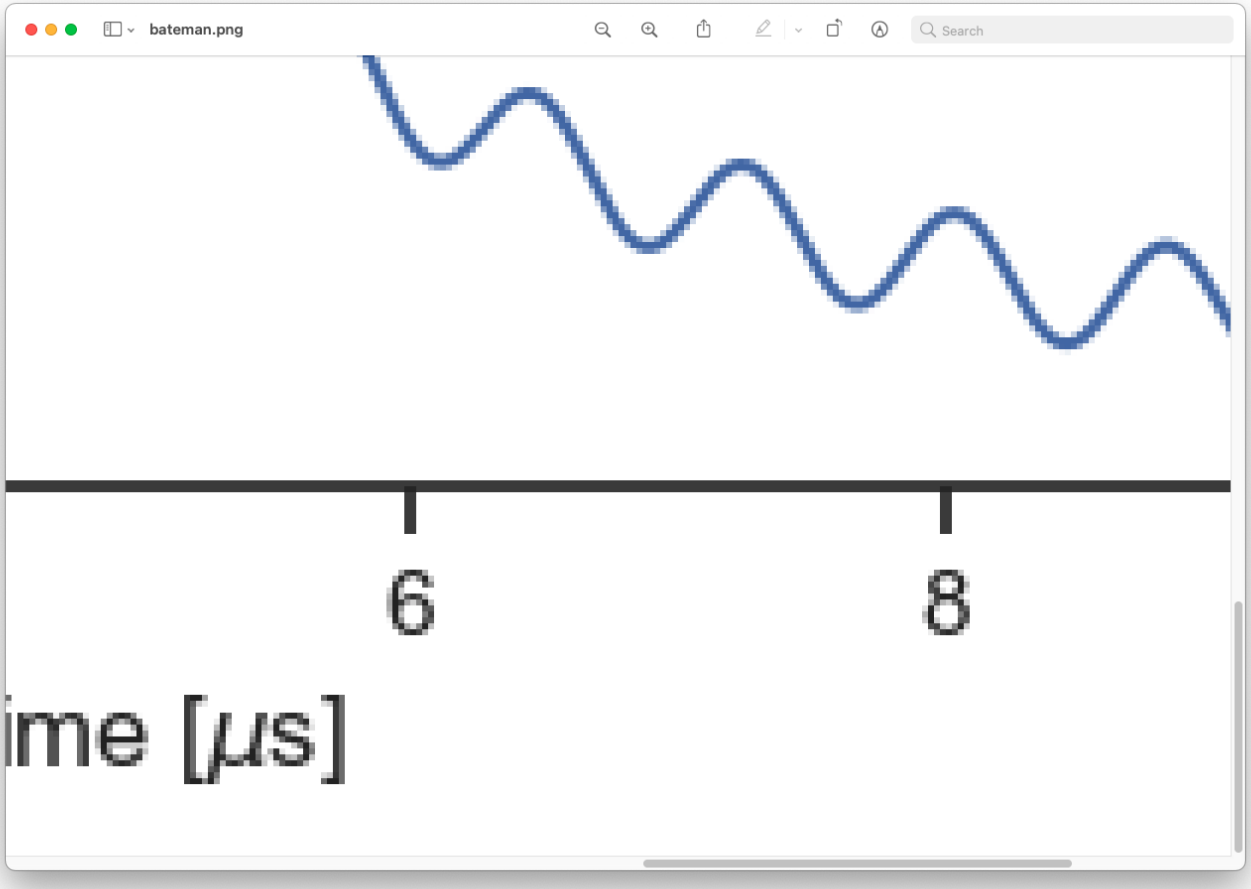
- use barplots for categorical data
- works both vertically and horizontally:
  - only time you should put a **dependent variable** on the *x* axis!

Far, far better than pies!

- estimating relative size of pie sections is hard – visual cortex does not extract fine distinctions between angles under rotation
- but bars are easy to compare!
- almost impossible to compare **two** pies
- they take up more space for the same data, and need to be much larger to be understood
- "3D" effects on pies make this all worse



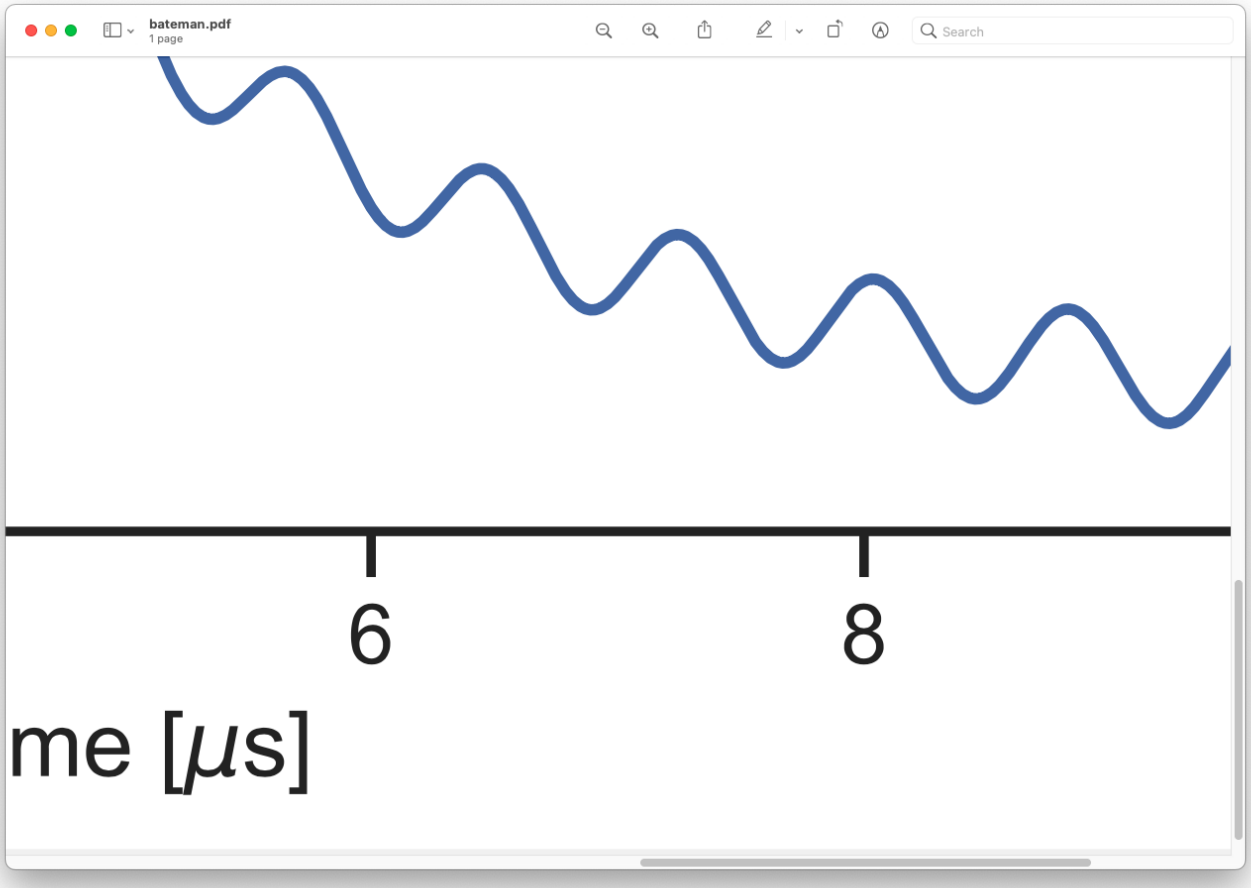
File format – raster graphics



PNG, JPEG, GIF, TIFF *etc.* 👎

These are called "raster graphic" formats. Their data is in **pixels**. They all look like crap when zoomed in. **If you are drawing your own figure, it doesn't have to be like this!**

File format – vector graphics



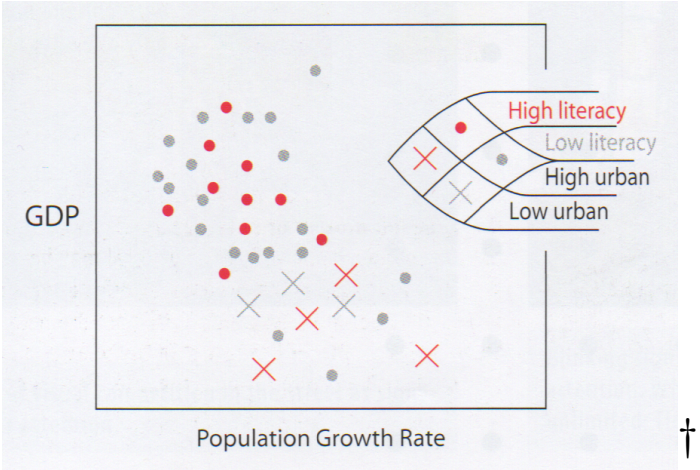
PDF, **EPS** 👍

These **vector graphic** data formats can store "**pen strokes**" instead of ~~pixels~~, so the image can withstand arbitrary zoom.

Use everywhere, but **especially on POSTERS!**

Perceptual Design Constraints and Strategies

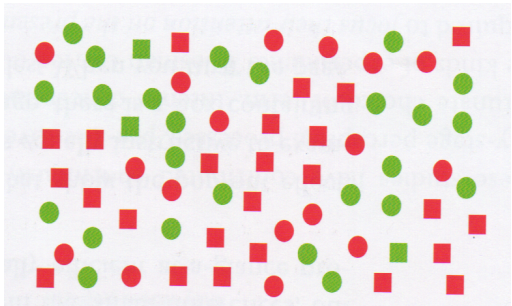
- to make something easy to find, make it different from surroundings according to a primary visual channel
- to make several things easy to search, use multiple channels



†

N.B.

Avoid overloading channels with conflicting information



"There are three green squares in this pattern. The green squares do not show a pop-out effect, even though you know what to look for. The problem is that your visual cortex can either be tuned for the square shapes or the green things, but not both"‡

†Ware (2008): *Visual Thinking for Design*

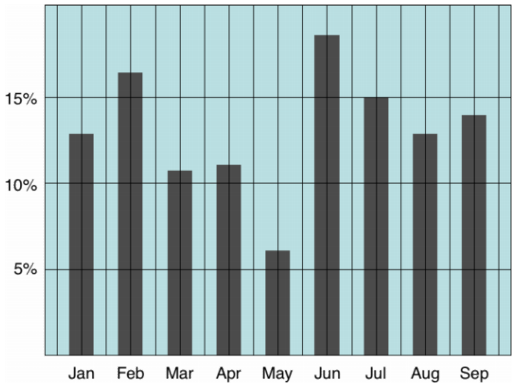
‡Ware (2008): *Visual Thinking for Design, pp. 30.*

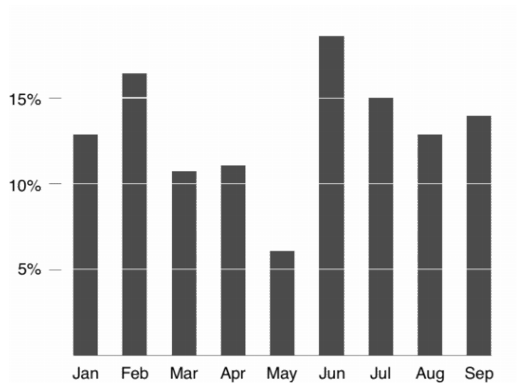
Perceptual Design Constraints and Strategies

- using two channels for same distinction makes it easier
- we only have 8-10 channels, and only ~ 3 steps per channel
- colour is powerful but hard to use well:
  - common perceptual difference: red/green + blue/green
  - "angry fruit salad" is awful – and contrast is critical
  - projector has less dynamic range than a monitor
- intensity is accessible, powerful and easy to use
  - combine intensity with colour in gradients
  - ensure **monotonic** change in intensity

Improving the visualization

- reduce the clutter – improve the "data-ink ratio"*





- choose distinct graphical elements; separable by channel; avoid overlap
- ensure data series are referenced; labels + legend must be clear of the data
- use all your space; log axes may be appropriate to spread information
- **provide explanations in the text and draw conclusions**
  - Don't be afraid to tell the reader what you want them to see and learn from your diagram

*E. Tufte (1983): *The Visual Display of Quantitative Data*, Graphics Press. ISBN: 0-961-3921-2-6

Thank-you for your attention!

`https://github.com/andrewhw/VizThoughts`