Visualization

TODAY chart literacy

1. anatomy of a plot
2. scale theory
3. scale perception
4. making comparisons
5. atomic plots

FRIDAY embedding

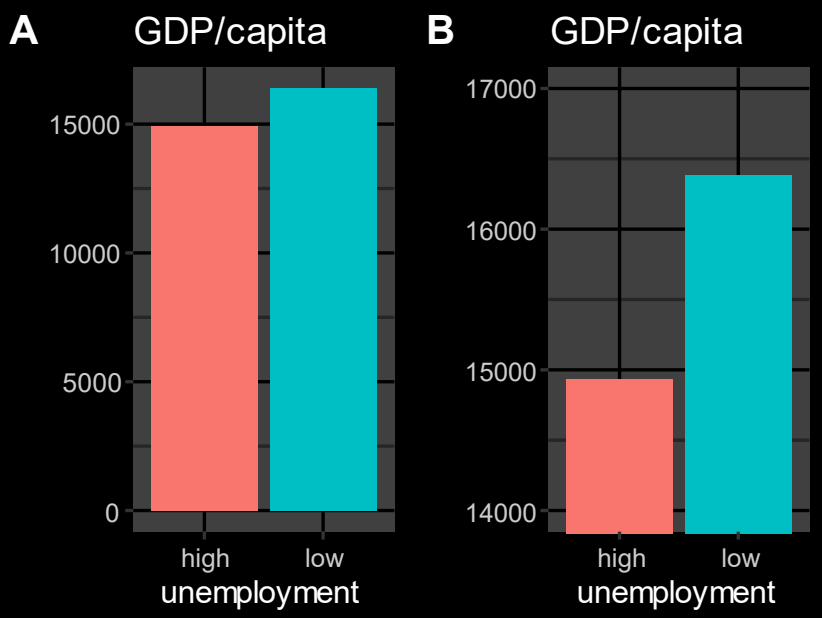
6. content scales
7. PCA, t-SNE, autoencoders

the "grammar" of charts

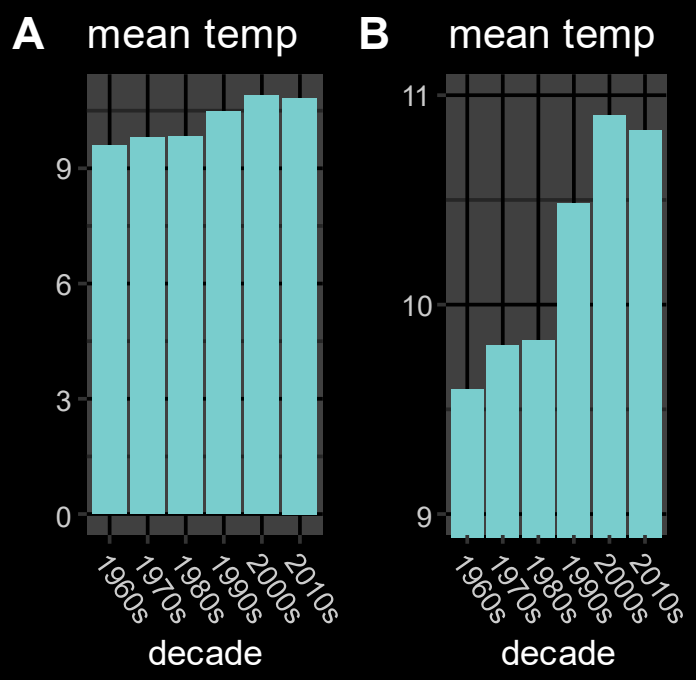
The old y-origin chestnut

Which of these plots is better, A or B? Why?

GDP per capita [PPP USD], split by whether unemployment is <7%



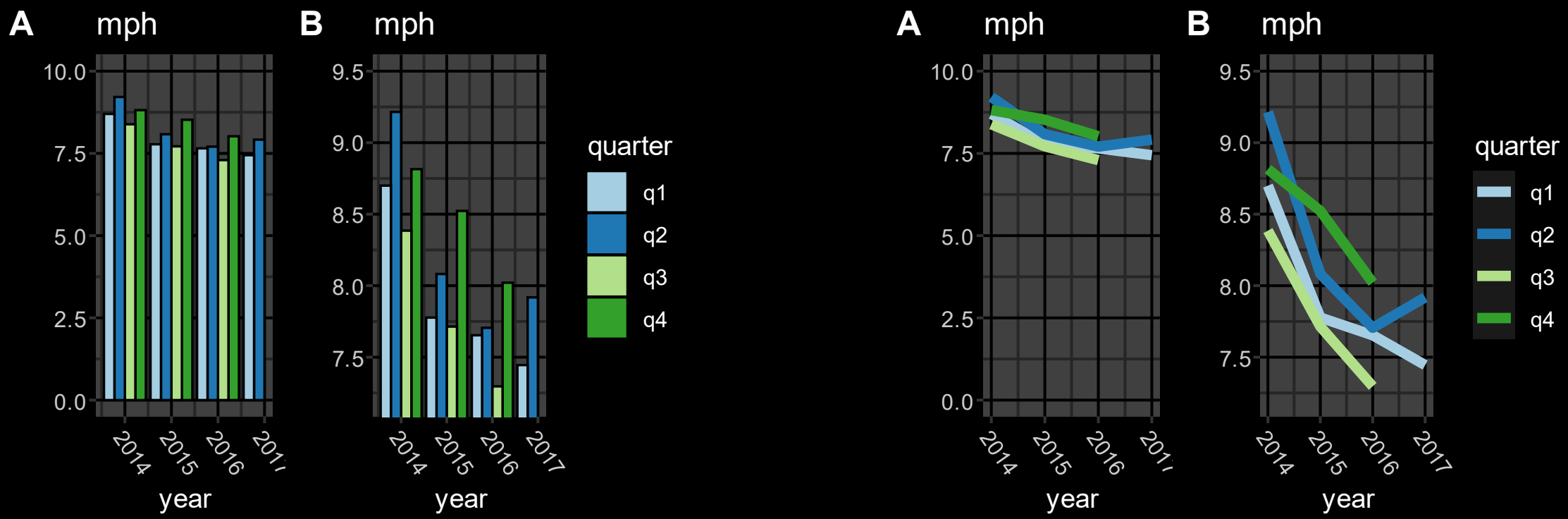
Average annual temperature [°C] in Cambridge



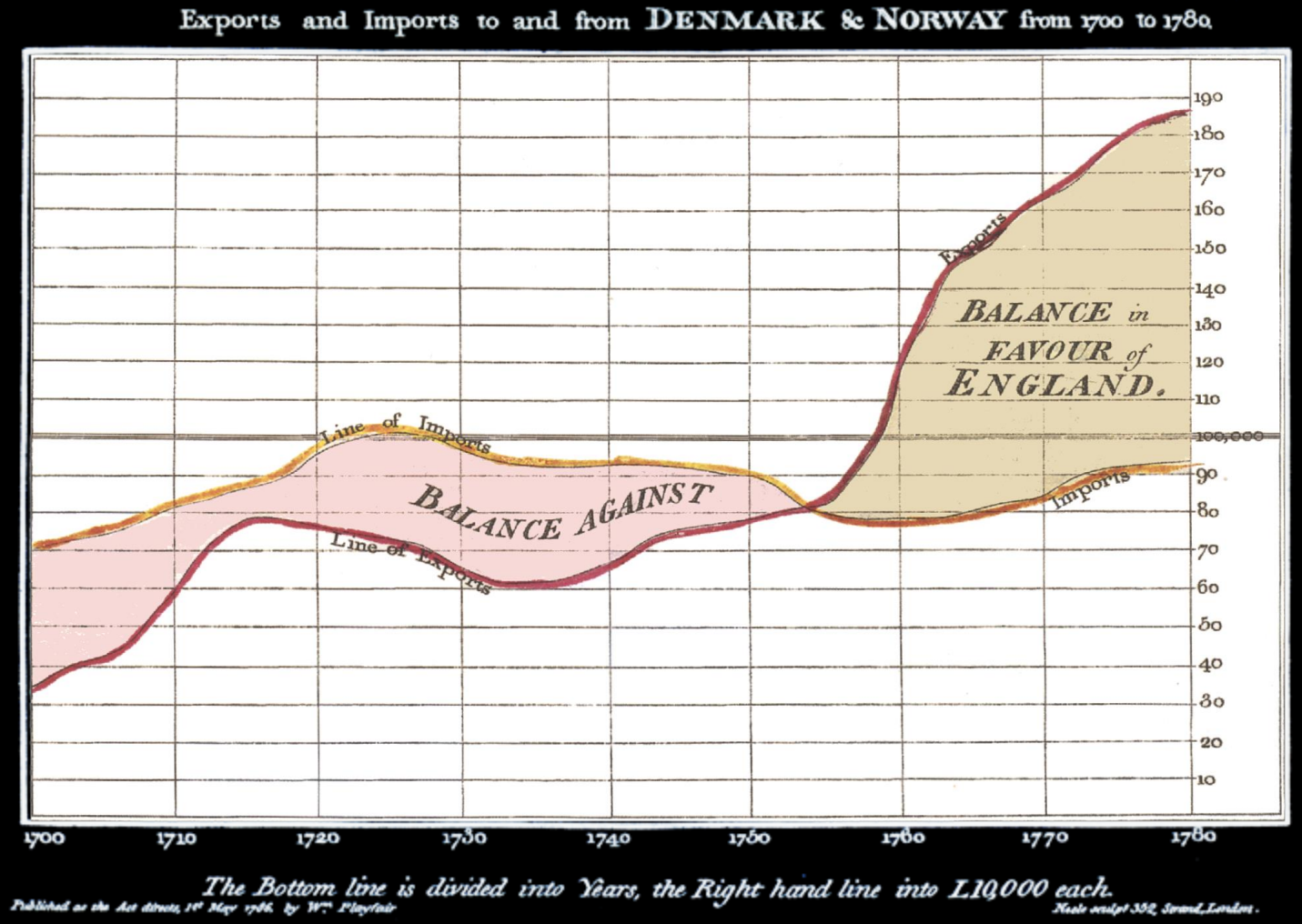
The old y-origin chestnut

Which of these plots is better, A or B? Why?

Average daytime speed in
central London, major roads



0. A short history of visualization



William Playfair (22 September 1759 – 11 February 1823), commonly known as a Scottish engineer and political economist, served as a secret agent on behalf of Great Britain during its war with France. The founder of graphical methods of statistics, Playfair invented several types of diagrams: in 1786 the line, area and bar chart of economic data, and in 1801 the pie chart and circle graph. As secret agent, Playfair reported on the French Revolution and organized a clandestine counterfeiting operation in 1793 to collapse the French currency.

In 1876, William Playfair invented a new language. Between 1876 and 1999, there have been two attempts to work out its grammar. This talk is based on Leland Wilkinson's *Grammar of Graphics*, 1999.

1. Anatomy of a plot

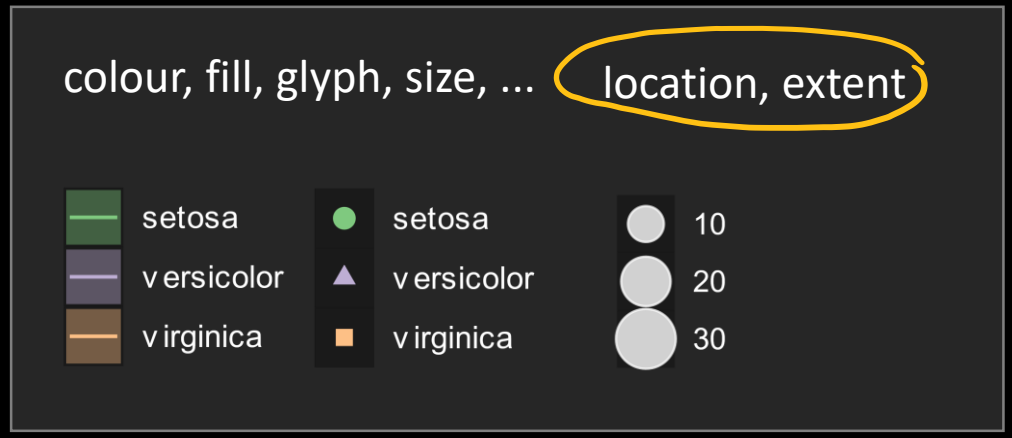
- A plot consists of geoms
- Usually, one row of data ↦ one geom, but some geoms are formed from groups of rows
- Data columns (features) are mapped to geom attributes (aesthetics)

data

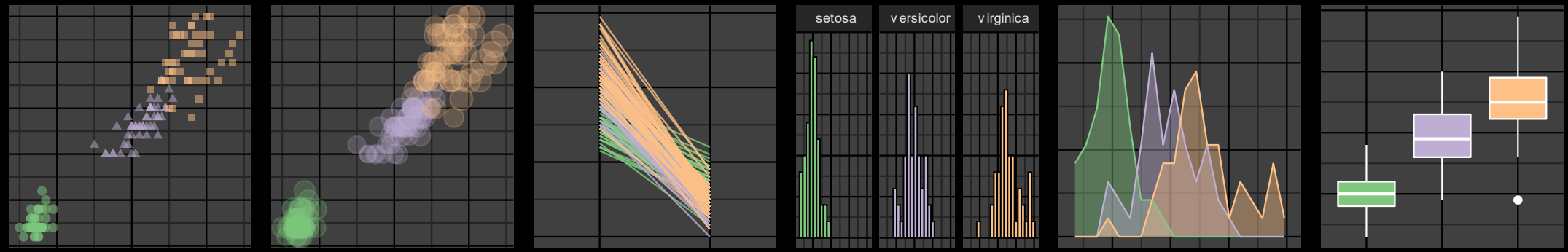
Sepal. Length	Sepal. Width	Petal. Length	Petal. Width	Species
5.0	3.4	1.6	0.4	setosa
6.5	3.0	5.5	1.8	virginica
5.0	3.5	1.3	0.3	setosa
6.7	2.5	5.8	1.8	virginica

+ geom

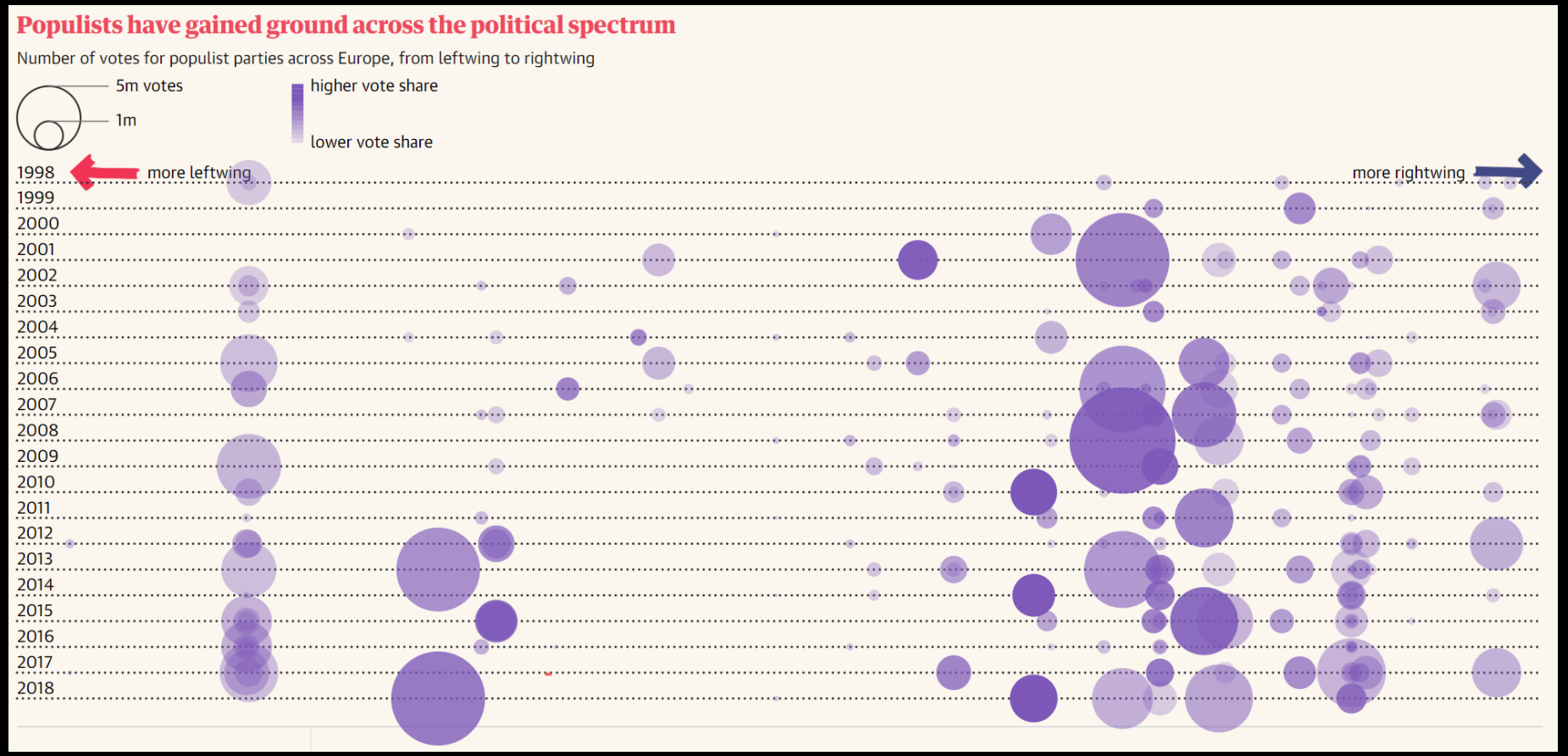
+ aesthetic map



= plot

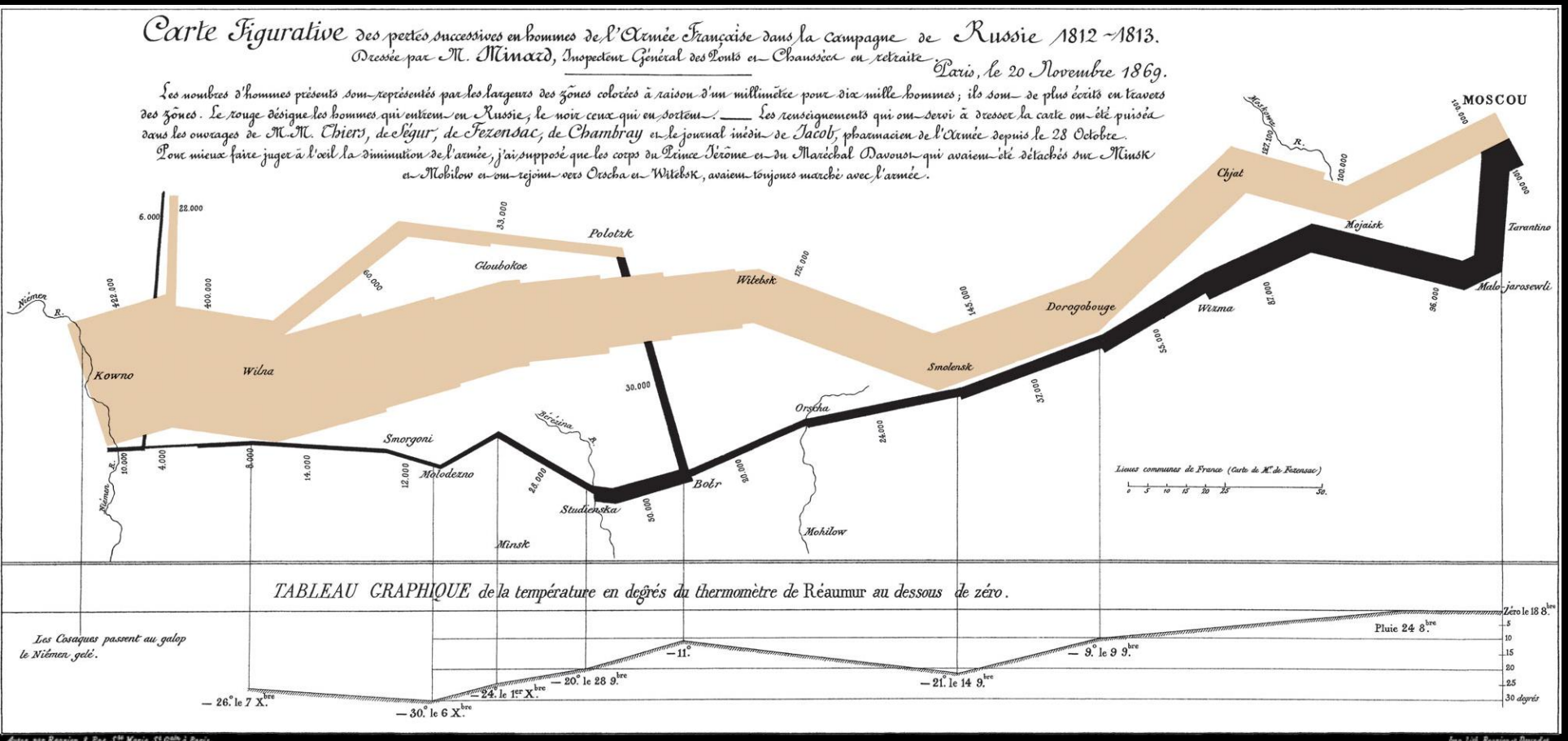


What are the data features and the aesthetic scales?



https://www.theguardian.com/world/ng-interactive/2018/nov/20/revealed-one-in-four-europeans-vote-populist

What are the data features and the aesthetic scales?



Charles Minard's map of Napoleon's disastrous Russian campaign of 1812. The graphic is notable for its representation in two dimensions of five data features:

- the number of Napoleon's troops
- location
- date
- temperature
- direction (advance or retreat)

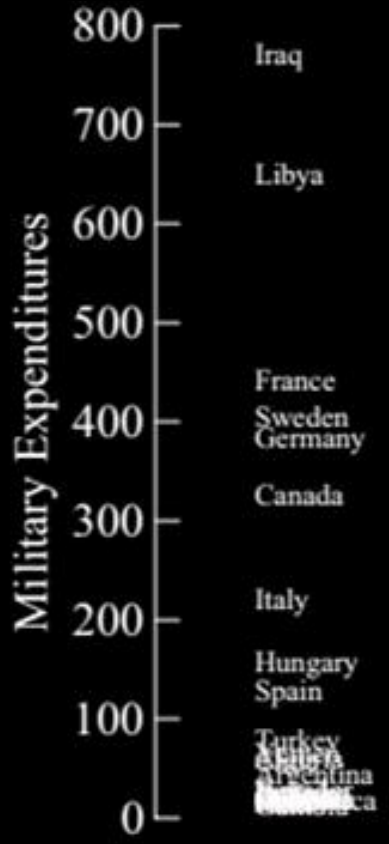
2. Scale theory

According to *On the theory of scales of measurement* (Stevens 1946) there are four types of data scale. (This isn't really true, but it's a good place to start.)

Nominal: no comparison is meaningful

Ordinal: we can ask which is greater, but not measure how much

Interval: we can subtract one value from another



Ratio: we can divide one value by another

Why is military expenditure *interval* rather than *ratio*?

"Money is not a physical or fundamental quantity. It is a measure of utility in the exchange of goods. Research by Kahneman and Tversky (1979) has shown that zero (no loss, no gain) is not an absolute anchor for monetary measurement. Individual and group indifference points can drift depending on the framing of a transaction or expenditure."

Wilkinson, 2005.

2. Scale theory

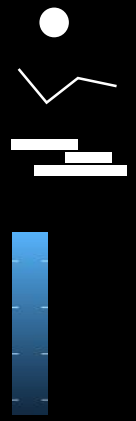
The four data scales work naturally with certain aesthetic scales ...

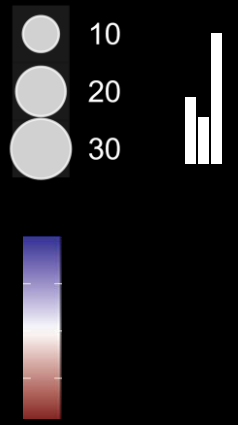
Nominal: no comparison is meaningful

Ordinal: we can ask which is greater, but not measure how much

Interval: we can subtract one value from another

Ratio: we can divide one value by another





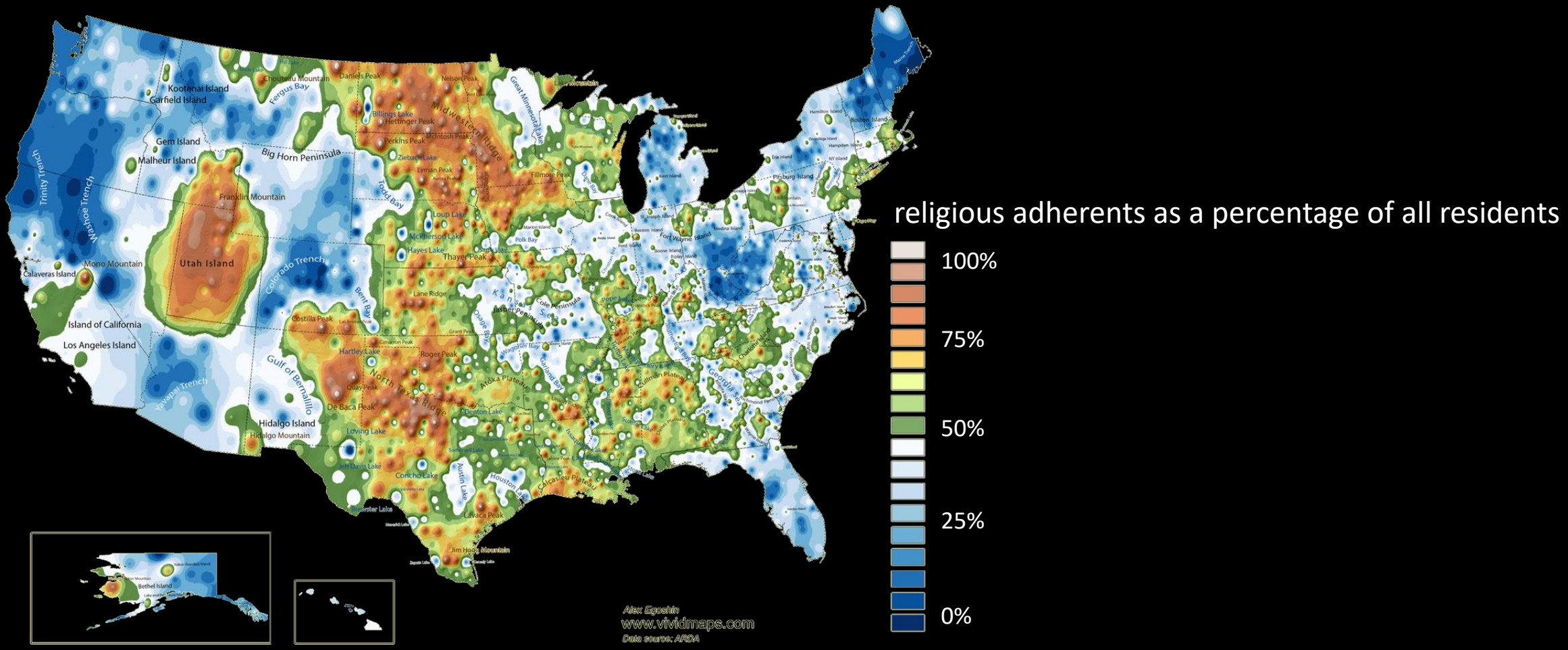
shape
colour choice

location index (panel)
colour sequence

location
extent
colour gradient

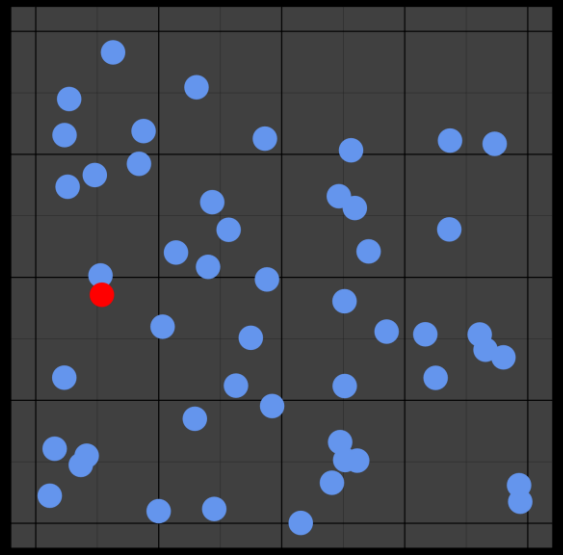
area
size
divergent colours

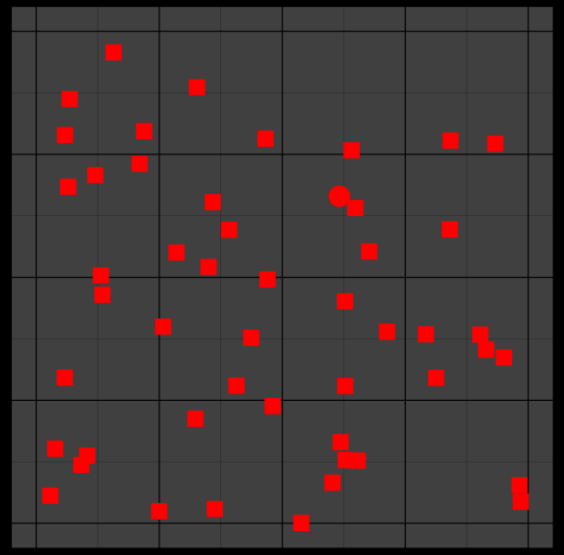
Use aesthetic scales that match your data scale (unless you know what you're doing)



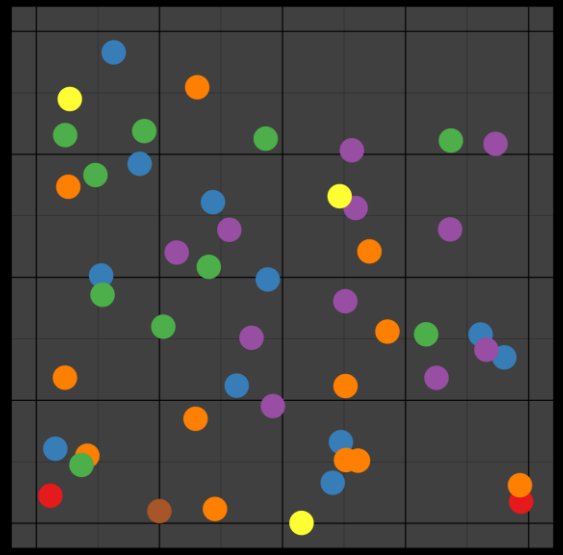
3. Scale perception

Is there a red circle?





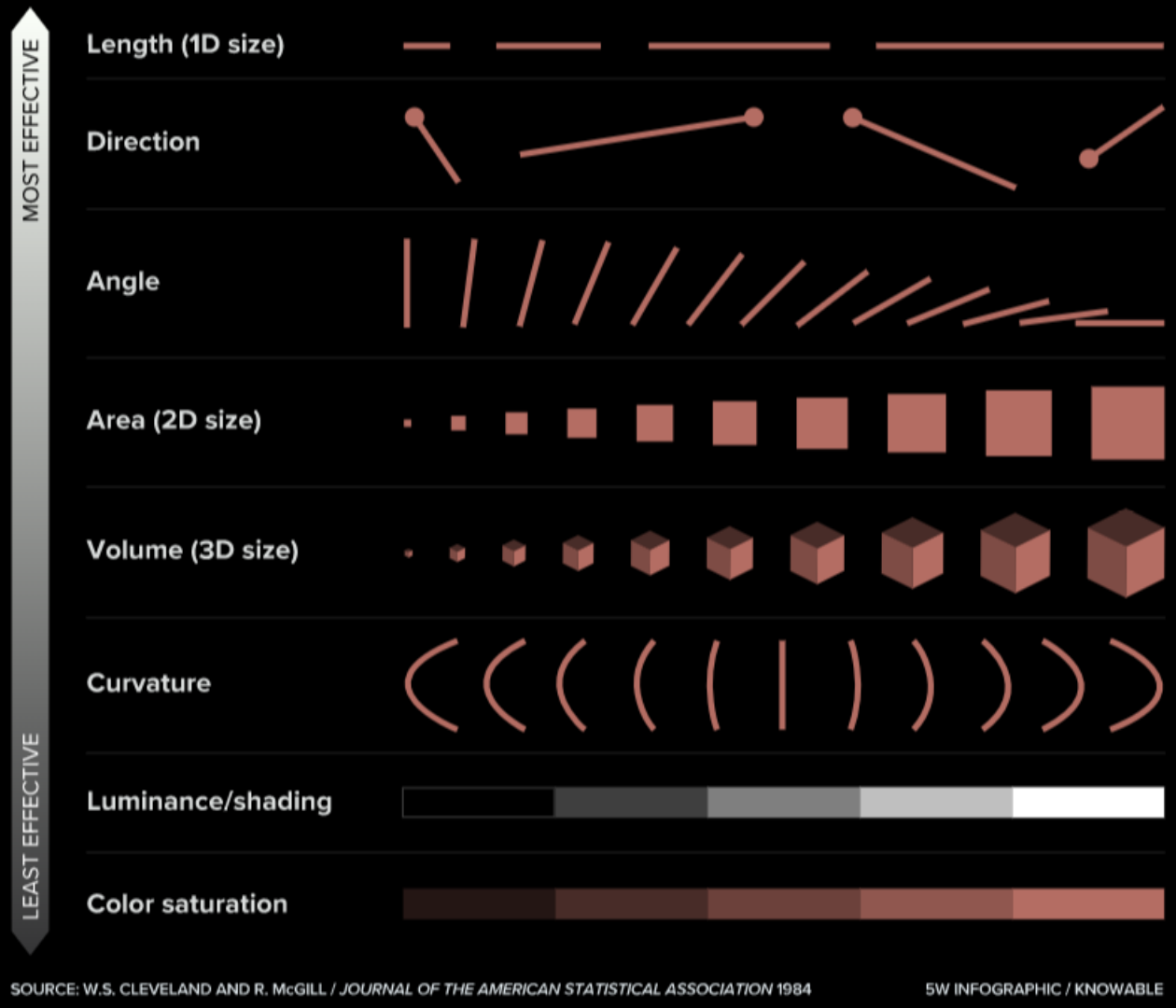
We can see colour differences more easily than glyph.



Less is more.

Fechner/Weber: we notice %difference in a sensation, not absolute difference.

Some scales are more effective than others at communicating differences.



Location is the easiest scale from which to read off differences

Note that the subplot index (in a multipanel plot) is also a type of location scale

Area is dangerous

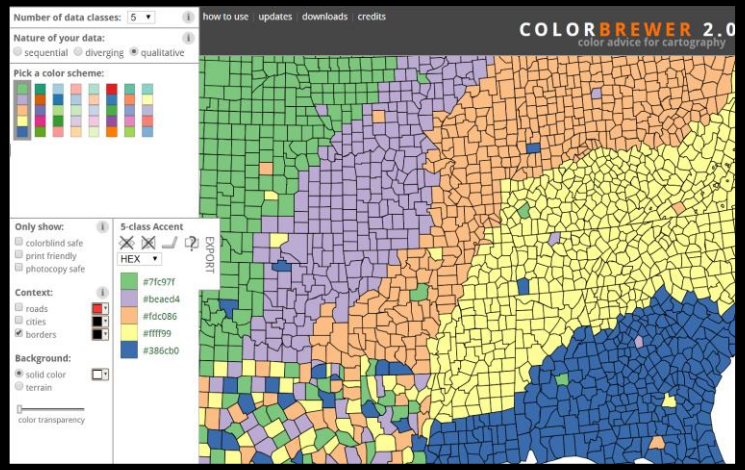
Stevens exponent: perceived area = (drawn area)$^{0.8}$.

Memory is also an aesthetic scale, used in user-hostile slideshows

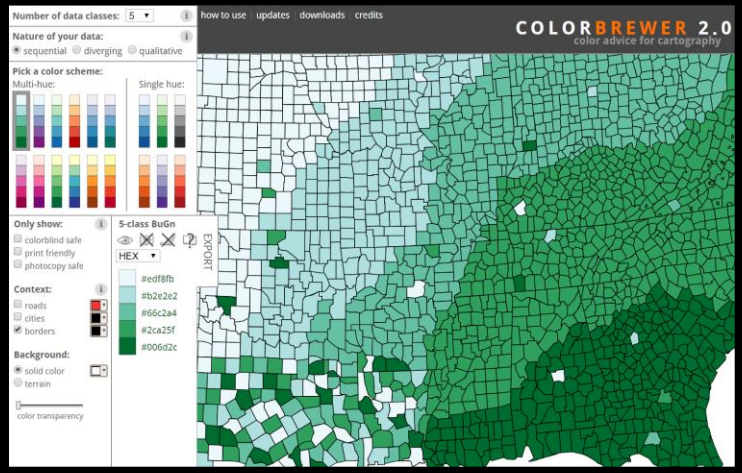
Why scientists need to be better at data visualization
https://www.knowablemagazine.org/article/mind/2019/science-data-visualization

Human perception of colour is tricky. Best not invent your own colour scales.

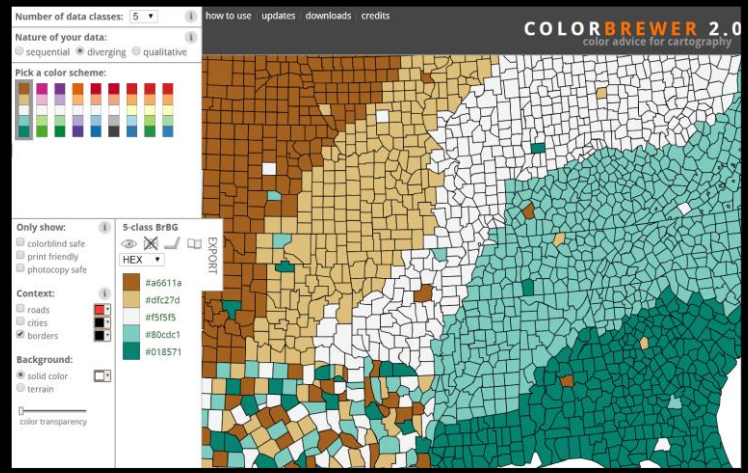
Nominal: no comparison is meaningful



Ordinal: we can ask which is greater, but not measure how much



Ratio: we can divide one value by another



4. Making comparisons

Since no model is to be believed in, no optimization for a single model can offer more than distant guidance. What is needed, and is never more than approximately at hand, is guidance about what to do in a sequence of ever more realistic situations. The analyst of data is lucky if he has some insight into a few terms of this sequence, particularly those not yet mathematized. [...] The main tasks of pictures are then: to reveal the unexpected, to make the complex easier to perceive. Either may be effective for that which is important above all: *suggesting the next step in analysis, or offering the next insight*.

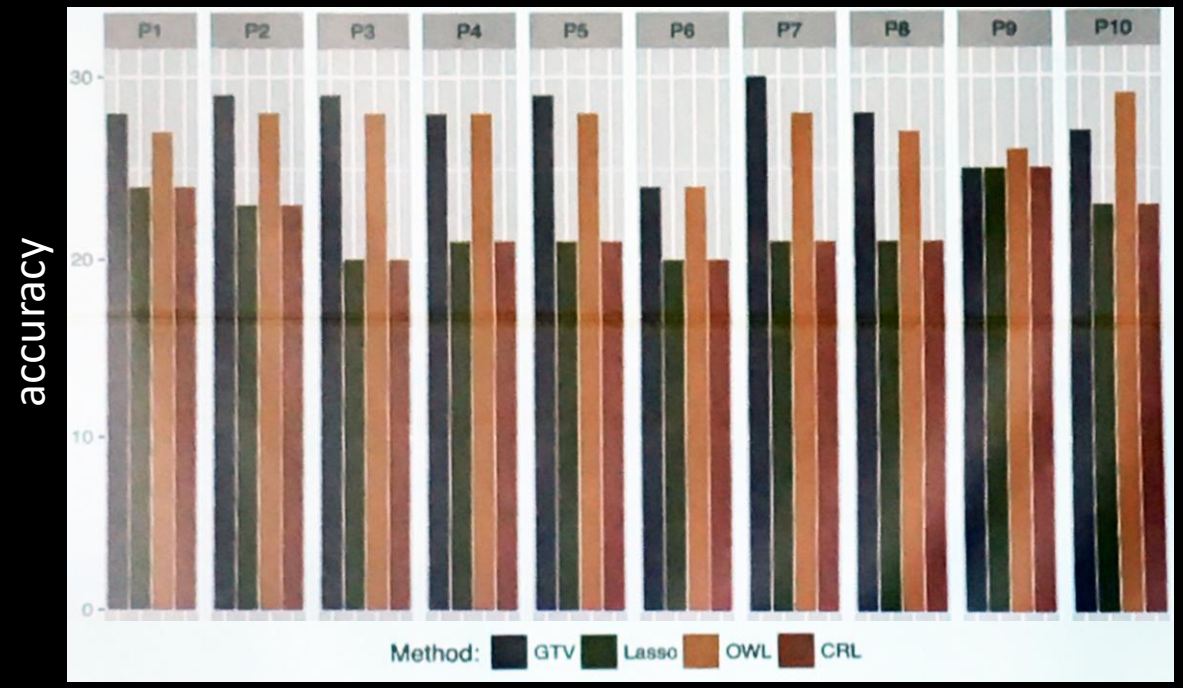
Mathematics and the picturing of data, John Tukey, 1975

- Plots invite the viewer to make comparisons (how does feature A depend on B, C, or D?)
- Have you put your primary comparators on the best-perceived scales?

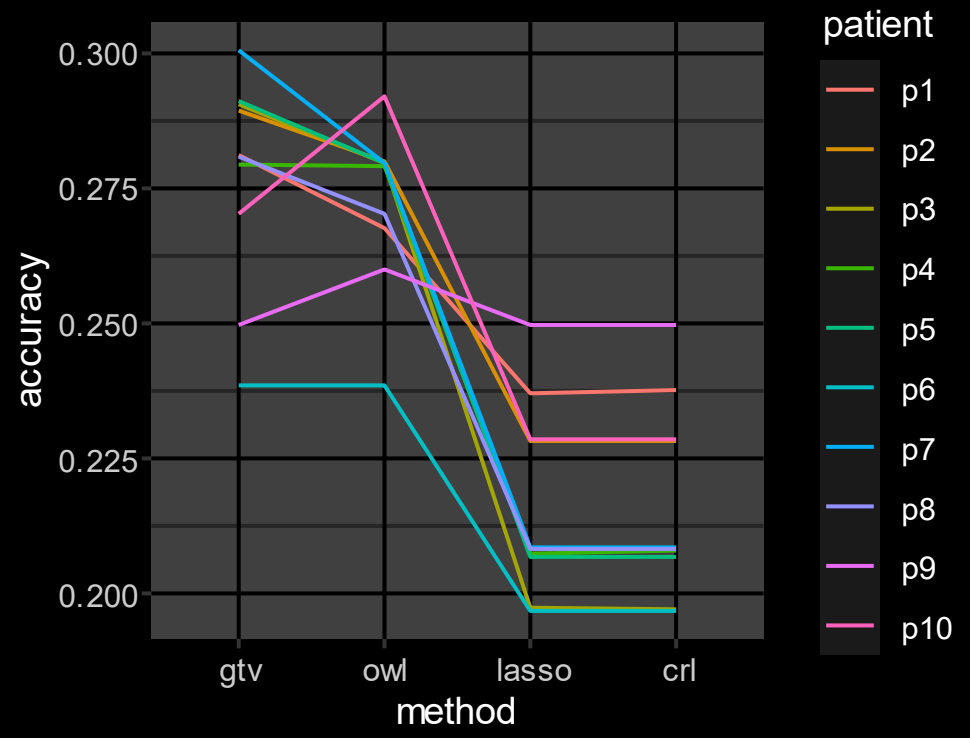
What comparisons does the plot invite?

DATASET: medical data for 10 patients was processed by 4 classification algorithms, and each algorithm was scored on a holdout dataset of size 30, to measure its prediction accuracy.

patient ID	classification algorithm	accuracy score
p2	lasso	0.228
p3	owl	0.279
p3	crl	0.197
⋮	⋮	⋮



- The main comparison is "how does accuracy depend on algorithm?"
  So put this on x and y scales.
- This geom (line) is based on row groups, to help make within-patient comparisons.
- The comparison between patients is less important. So hue is fine.



What comparisons does the plot invite?

Scientists love hypothesis tests, and they love bar charts.
But it is bad form to combine them! If your bars do not
show the comparisons you want to make, find a better plot.

https://www.nature.com/articles/s41551-017-0079

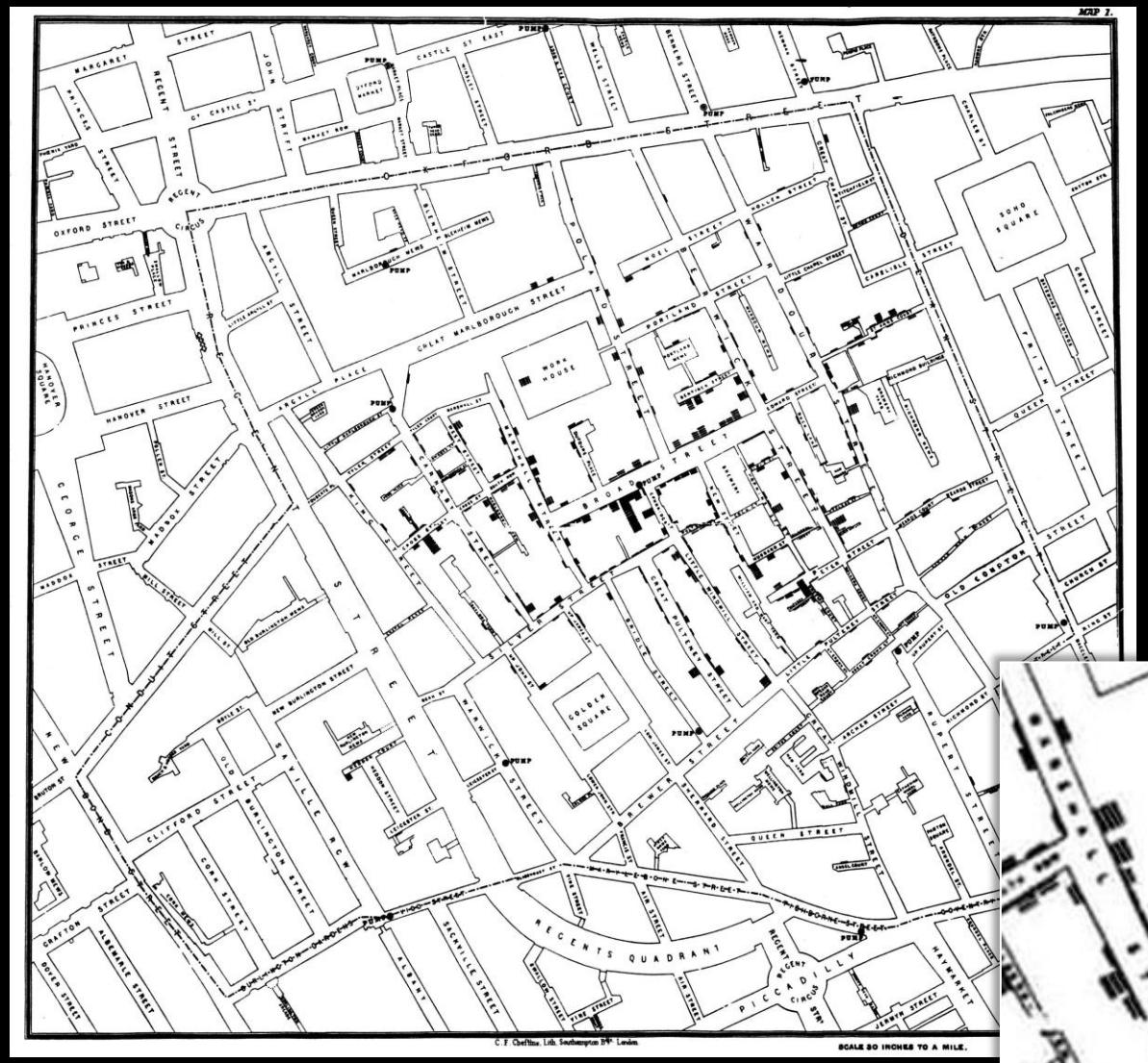
What comparisons does the plot invite?

This plot shows signal strength from a pulsar. Each line spans a period in time, and the periods are arranged in order of time, with occlusion. This is a very effective way to show “precise period, fairly arbitrary difference from one period to the next”.

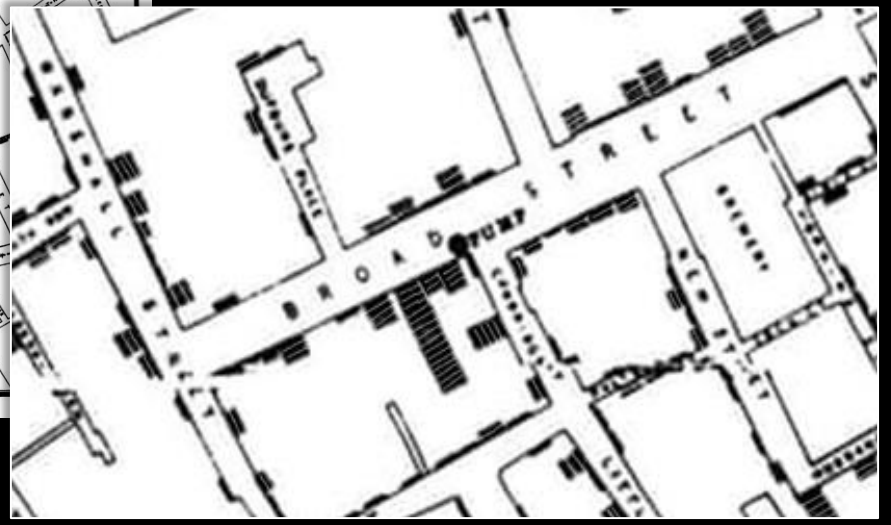
Joy Division’s album *Unknown Pleasures*, 1979

https://blogs.scientificamerican.com/sa-visual/pop-culture-pulsar-origin-story-of-joy-division-s-unknown-pleasures-album-cover-video/

5. The atomic theory of plotting



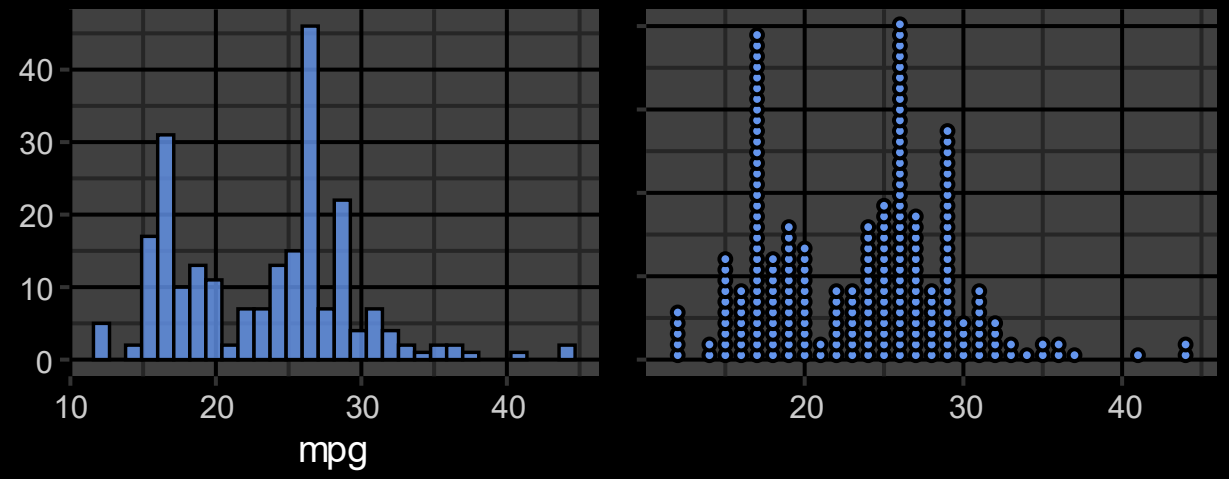
Each bar is a person who died from cholera. The bars have been stacked.

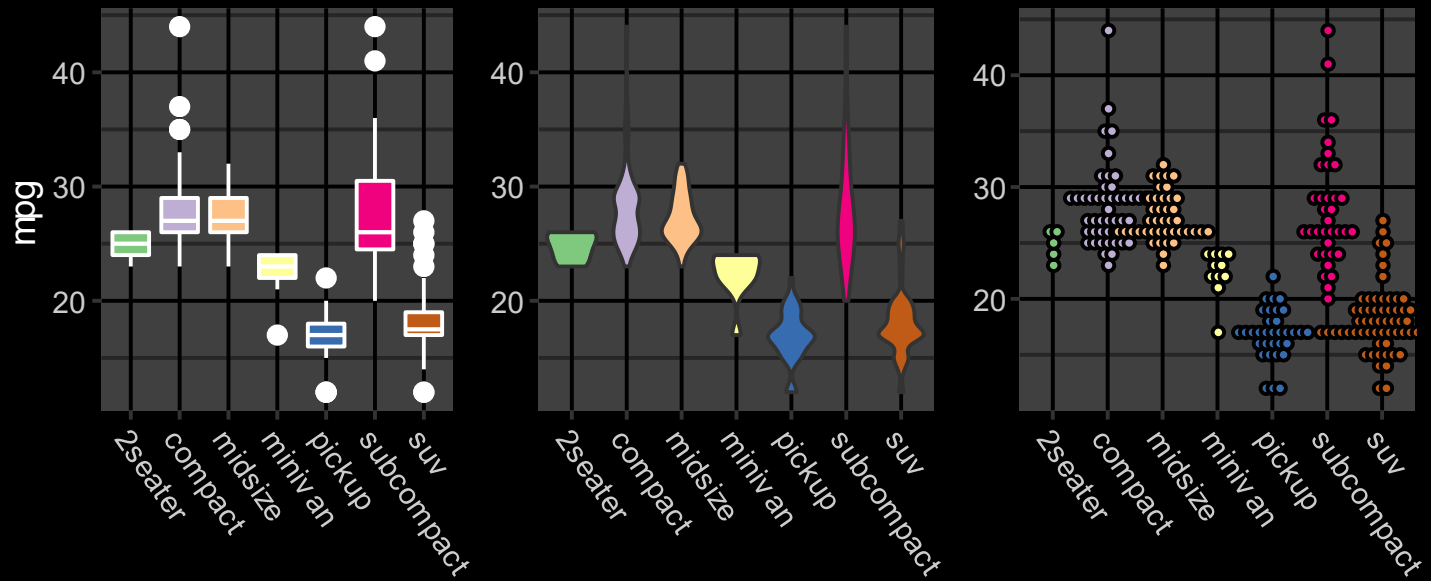


John Snow, 1854 https://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map

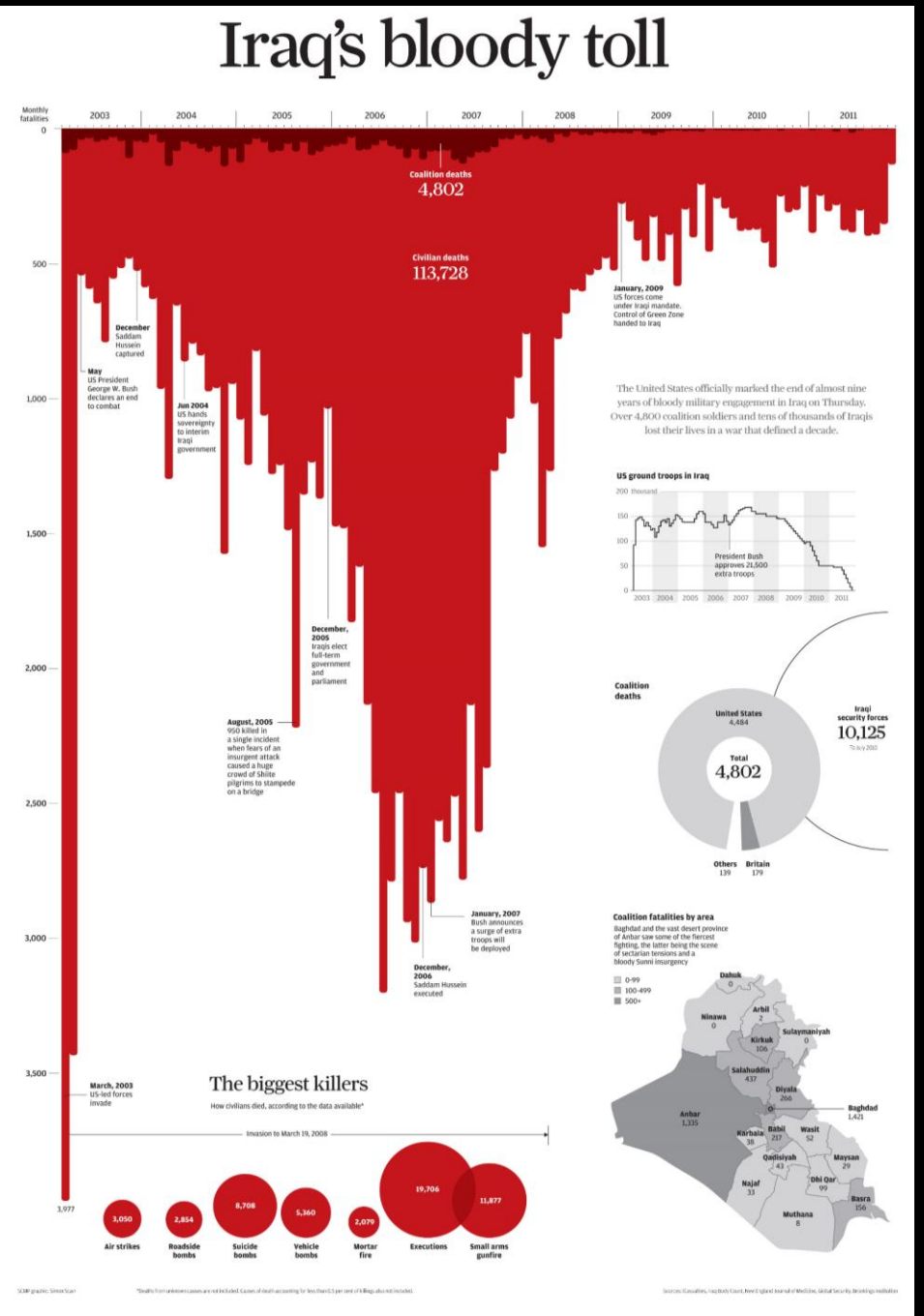
In the best plots, every dot of ink is a datapoint. (This is why histograms are easy to read.)

Dataset: miles per gallon for a variety of cars





In the best plots, every dot of ink is a datapoint.



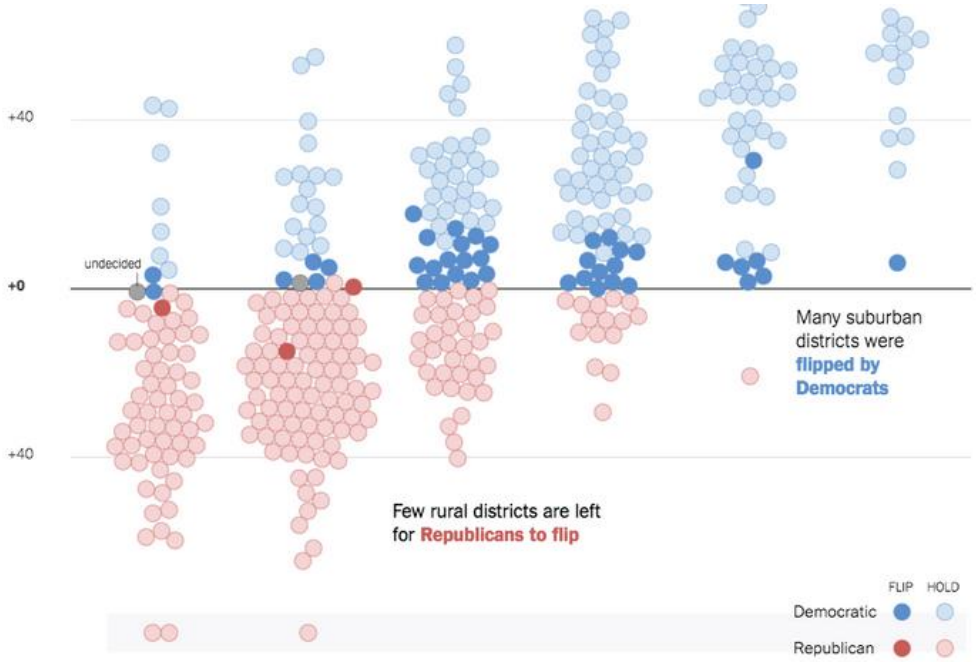
South China Morning Post
https://www.scmp.com/infographics/article/1284683/iraqs-bloody-toll

In the best plots, every dot of ink is a datapoint.

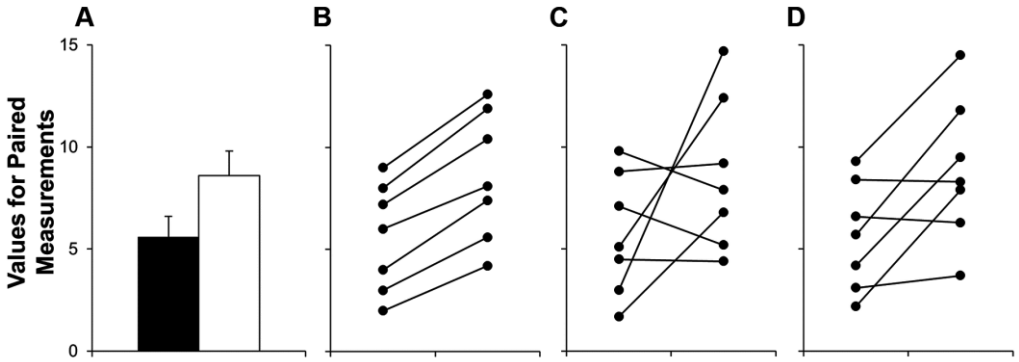
NASA earth observatory
https://earthobservatory.nasa.gov/images/87551/london-at-night

In the best plots, every dot of ink is a datapoint.

The New York Times makes great use of interactive dotplots, to tell stories.



Show us the dots! cry editorials in scientific journals. It's too easy to lie with aggregated data.

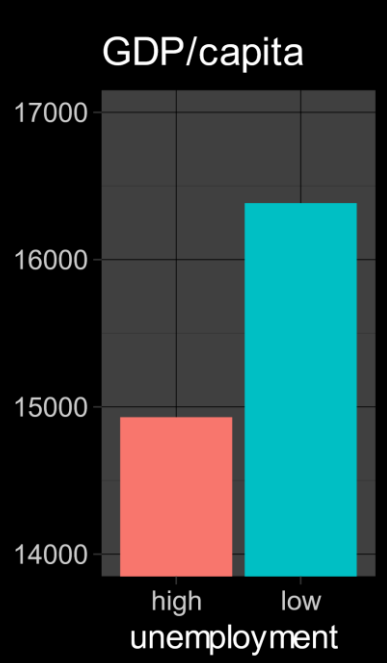


Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

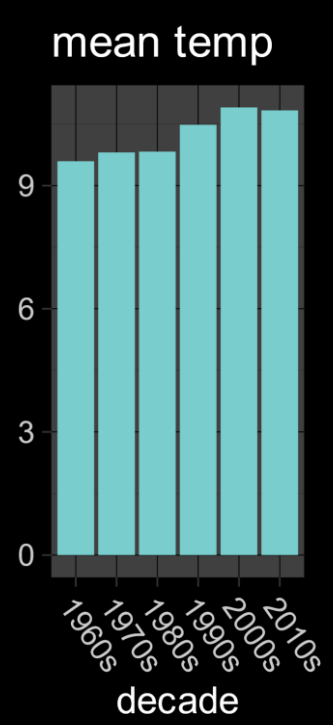
Show the dots in plots
https://www.nature.com/articles/s41551-017-0079

Rules for atomic plots

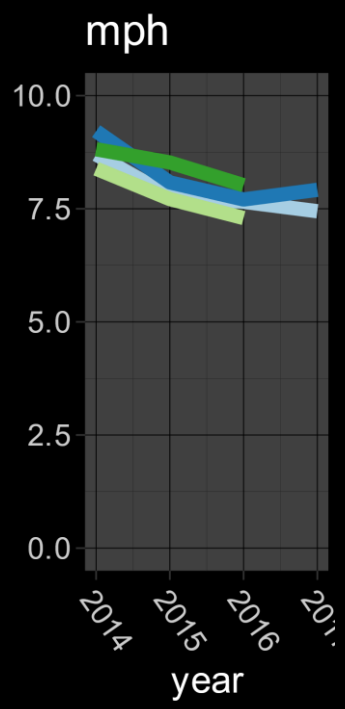
- If your data scale is accumulative:
  show it with a histogram, and let the size be accumulated mass.
- If your data is not accumulative:
  don't use bars, because they convey the impression of mass.



Average annual temperature [°C] in Cambridge



Average daytime speed in central London, major roads

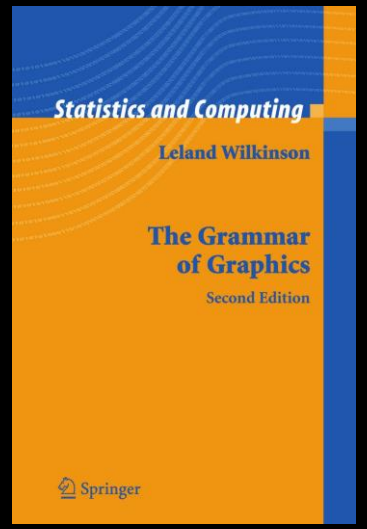


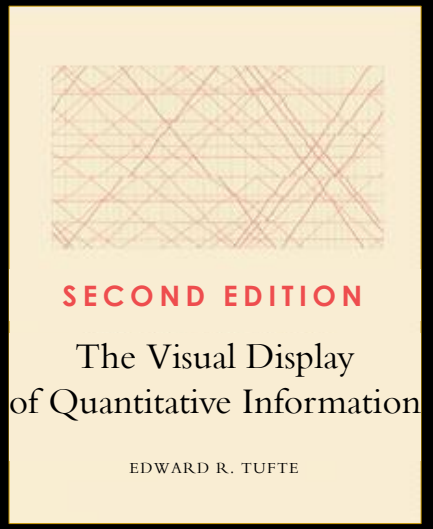
We have studied the grammar of graphics.

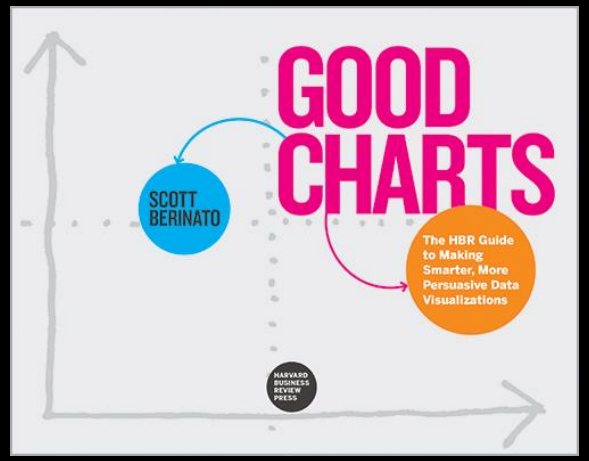
Grammar doesn't tell you how to create great charts. But it does give you tools to think systematically about your charts.

You also need • style • the skill to tell a story • good software libraries.

rhetoric = grammar + style + reason / arrangement







R + ggplot2

Javascript + D3

Vega Lite

and many many badly conceived libraries ...

matplotlib

When to use pie charts