Visualization

Slides by Damon Wischik

LECTURE ONE — chart literacy

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

LECTURE TWO — embedding

6. unsupervised learning
7. dimension reduction / PCA
8. self-supervised learning / tSNE
9. content scales

the "grammar" of plots

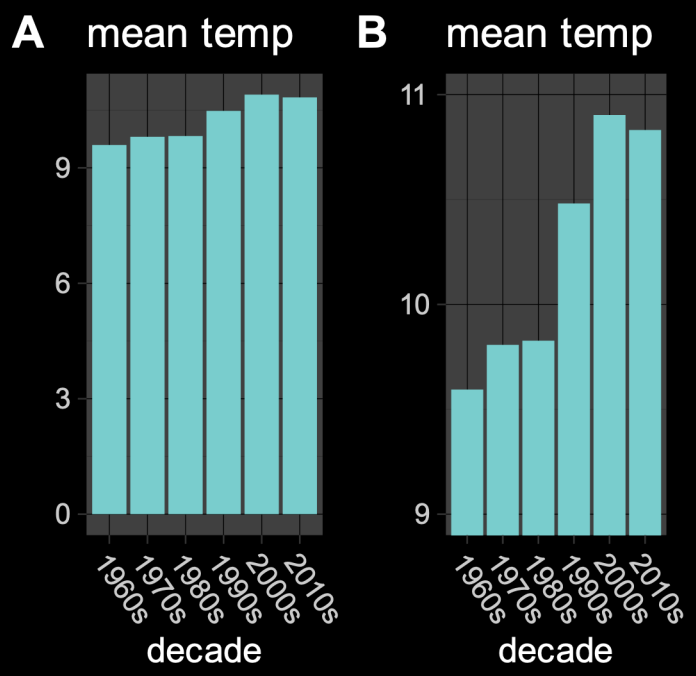
plots for data exploration

Introduction: 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%

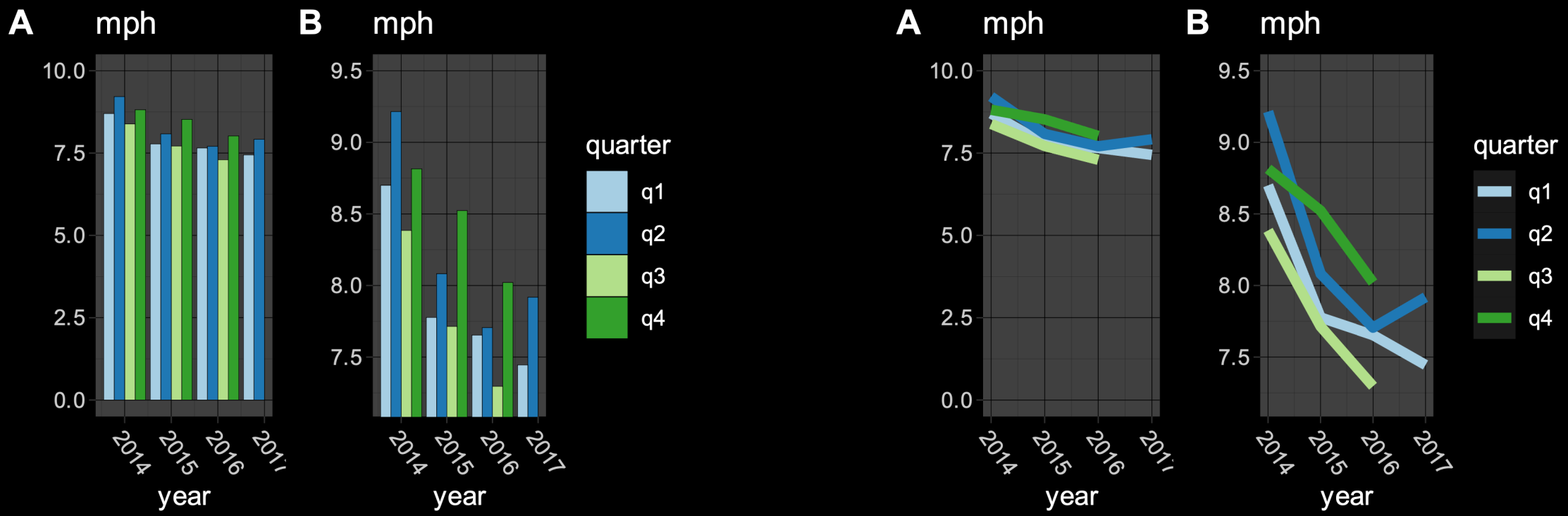
Average annual temperature [°C] in Cambridge



Introduction: the old y-origin chestnut

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

Average daytime speed in
central London, major roads



What is visualisation for?

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

What is visualisation for?

- Summarize the data
- See the distribution / spread / clusters
- Make comparisons / predictions
- Find explanations
- Persuade an audience

What is visualisation for?

- Summarize the data
- See the distribution / spread / clusters
- Make comparisons / predictions
- Find explanations
- Persuade an audience

What groups of X values are there?
(a bit like unsupervised learning / clustering)

What is visualisation for?

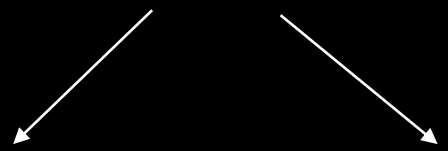
- Summarize the data
- See the distribution / spread / clusters
- Make comparisons / predictions
- Find explanations
- Persuade an audience

What groups of X values are there?
(a bit like unsupervised learning / clustering)

How does Y depend on X? (a bit like supervised learning / regression)

What is visualisation for?

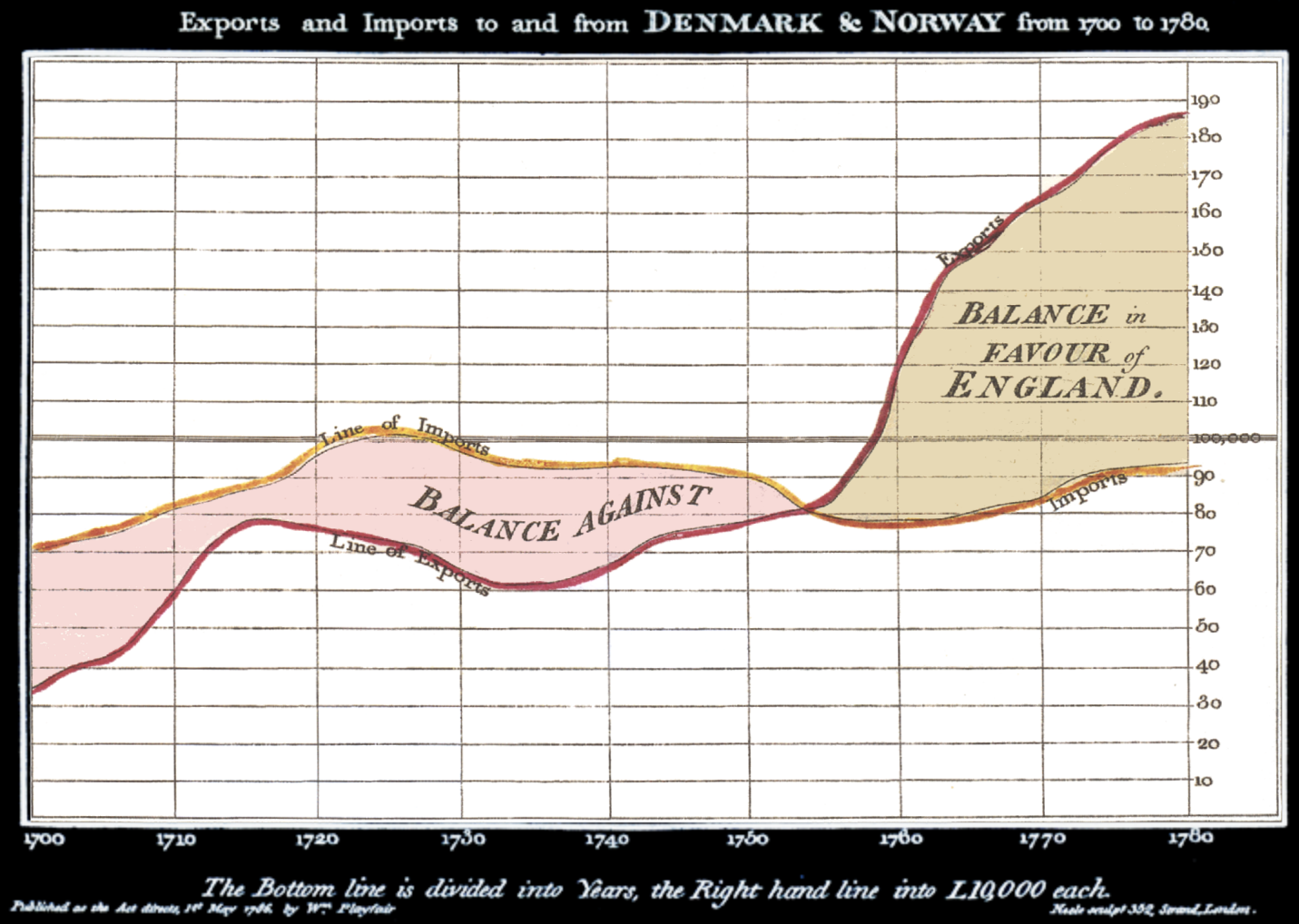
- Summarize the data
- See the distribution / spread / clusters — What groups of X values are there? (a bit like unsupervised learning / clustering)
- Make comparisons / predictions — How does Y depend on X? (a bit like supervised learning / regression)
- Find explanations
- Persuade an audience



You, the data scientist. You should iterate: visualize, see something new, think, repeat.

Your audience, the people you want to persuade. You should think about the comparisons you want your audience to make, and arrange your plots to emphasize them.

Introduction: 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.

William Playfair invented a new language. Between1876 and 1999, there have been two attempts to work out its grammar. This talk is based on Leland Wilkinson's *Grammar of Graphics*, 1999.

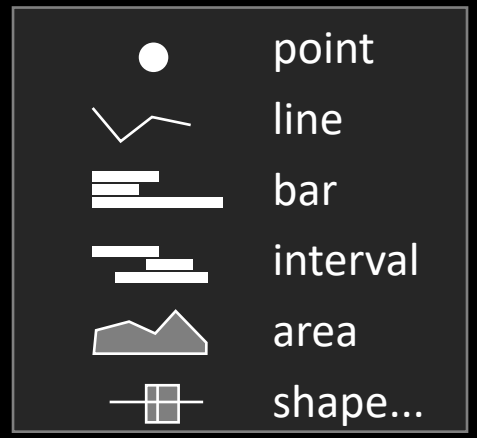
1. Anatomy of a plot

- A plot consists of geoms (geometric objects)
- 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)

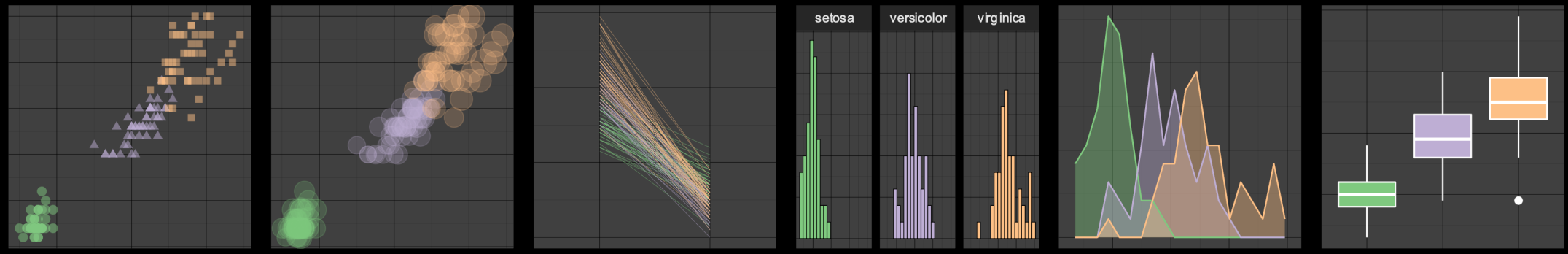
data

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



= plot



1. Anatomy of a plot

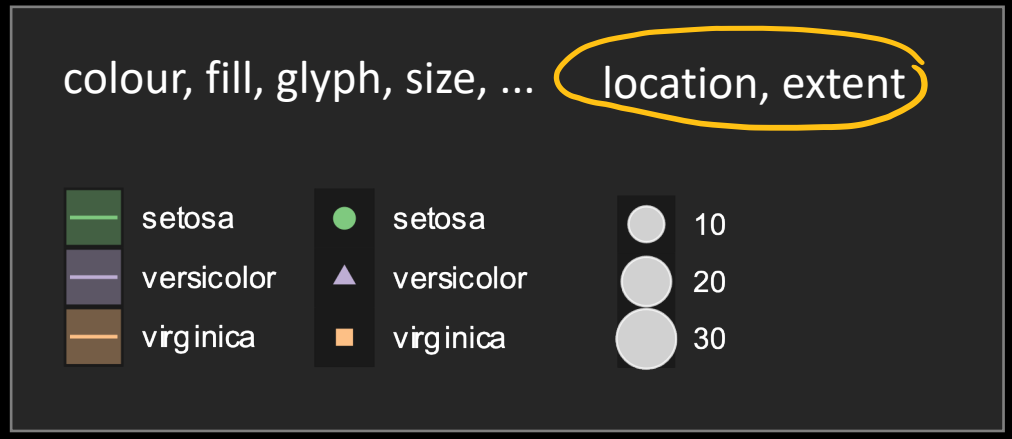
- A plot consists of geoms (geometric objects)
- Usually, one row of data ↦ one geom, but some geoms are formed from groups of rows
- Data columns (features) are mapped to geom attributes (aesthetics)

data

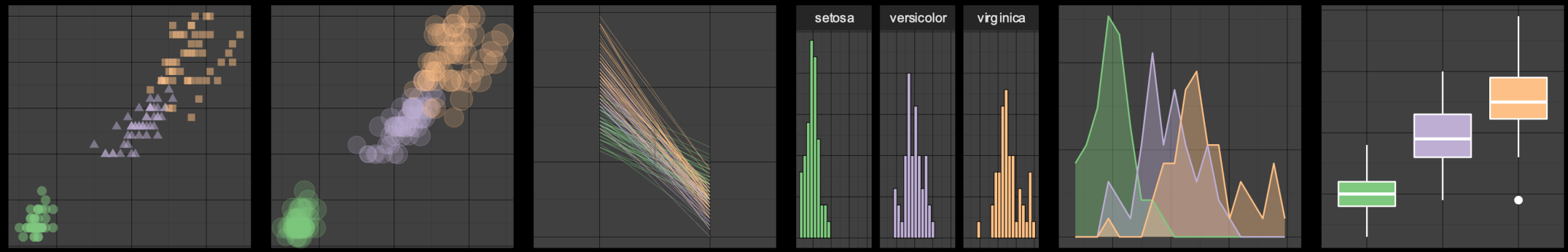
Sepal. Length	Sepal. Width	Petal. Length	Petal. Width	Species
5.0	3.4	1.6	0.4	setosa
6.5	3.0	5.5	1.8	virginica
5.0	3.5	1.3	0.3	setosa
6.7	2.5	5.8	1.8	virginica

+ geom

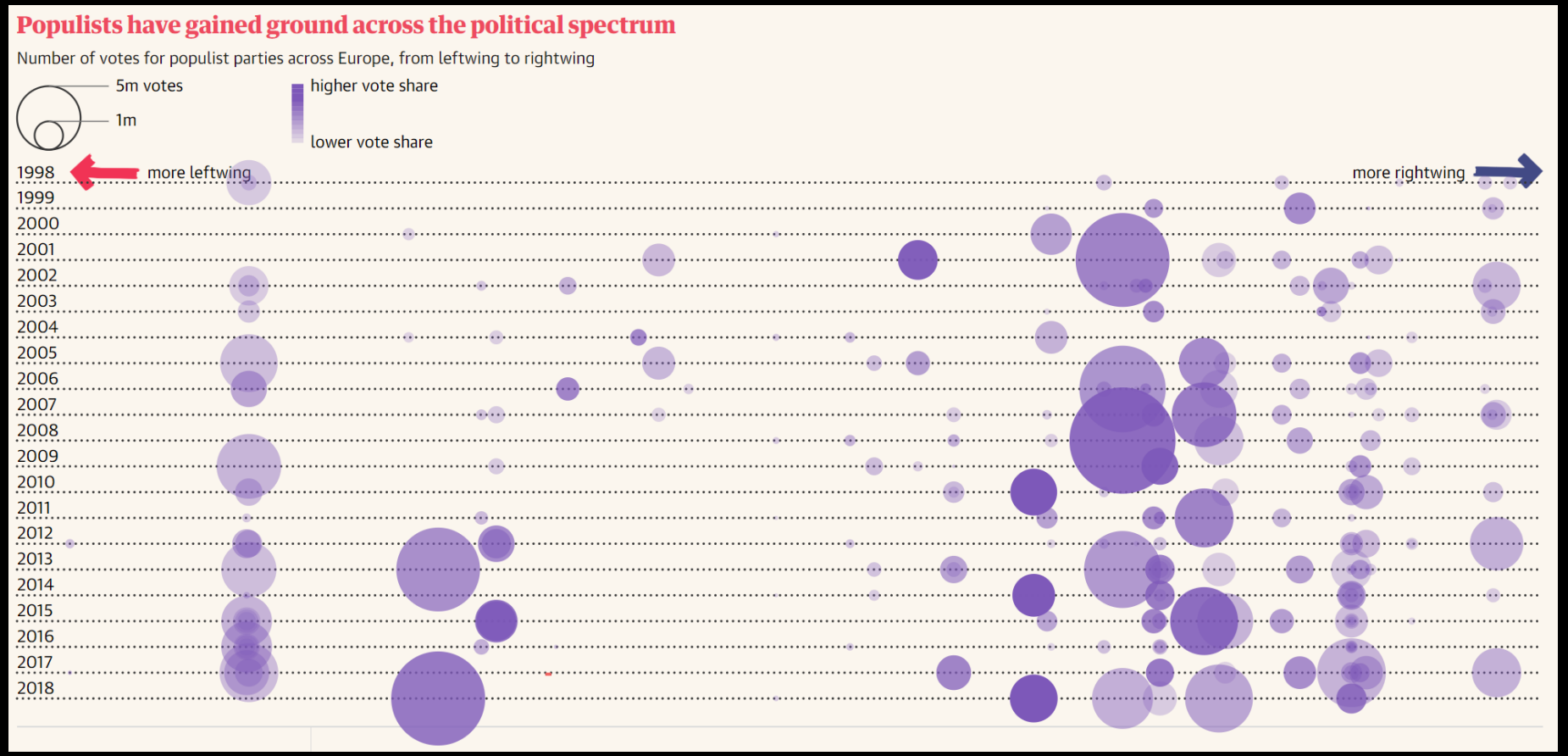
+ aesthetic map



= 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?

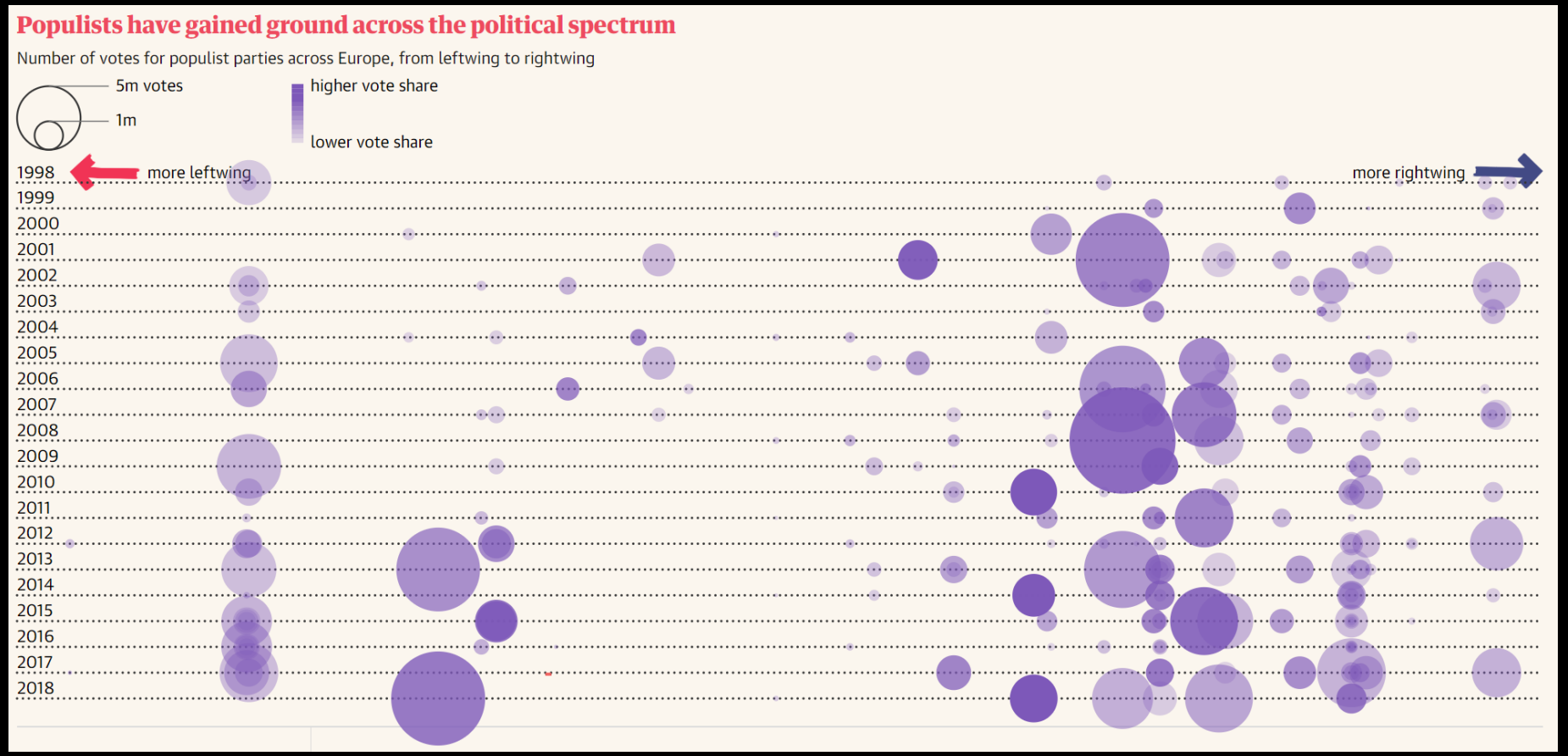
geom: point

aesthetics: size = #votes

α = vote share

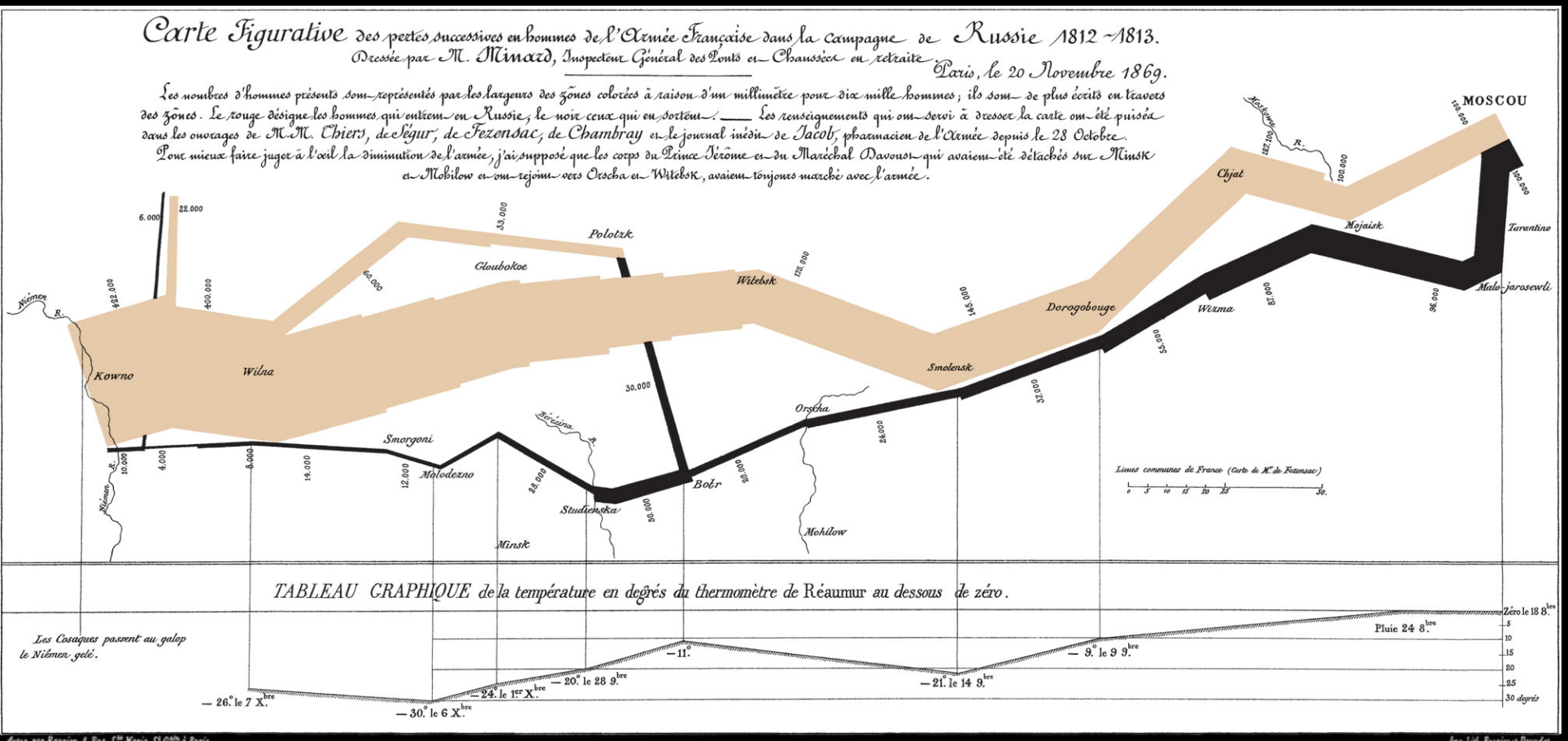
x = political leaning

y = year



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

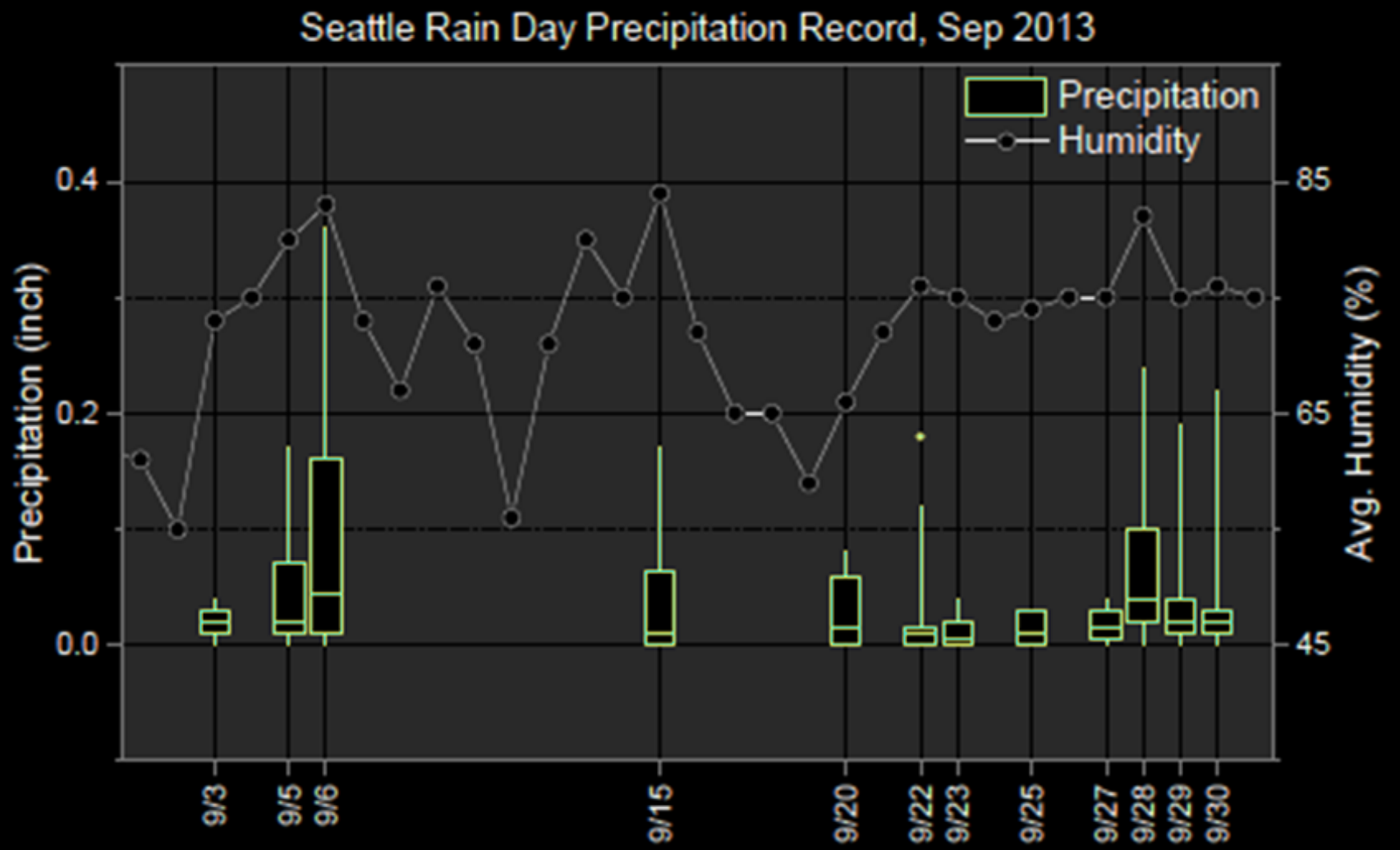
What are the data features and the aesthetic scales?



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)

What are the data features and the aesthetic scales?



Here the y-coordinate aesthetic scale is doing double duty: two different data features are mapped to it, with different mappings. No good data visualization toolkit allows this.

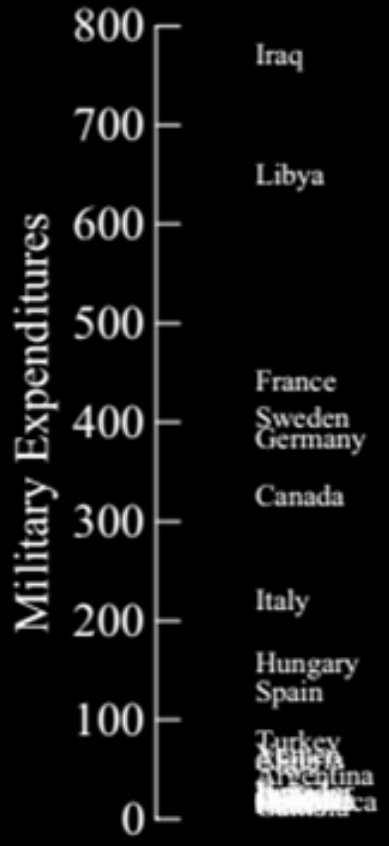
2. Scale theory

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

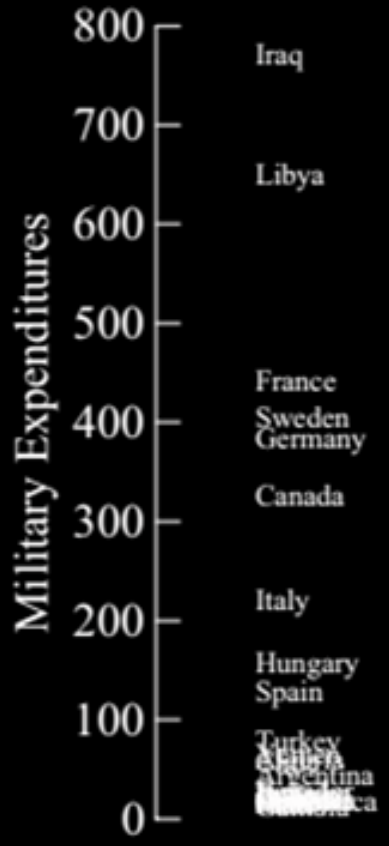
2. Scale theory

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

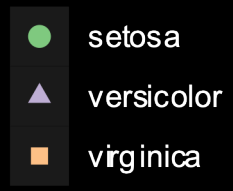
The four data scales work naturally with certain aesthetic scales ...

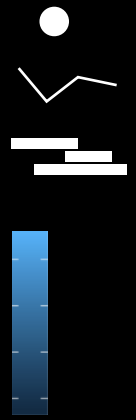
Nominal: no comparison is meaningful

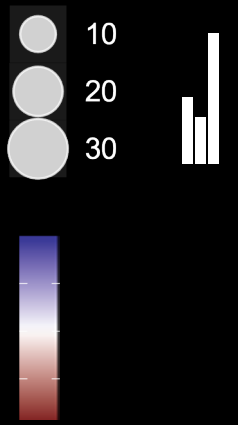
Ordinal: we can ask which is greater, but not measure how much

Interval: we can subtract one value from another

Ratio: we can divide one value by another







shape
colour choice

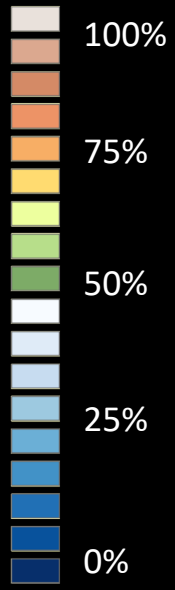
discrete location (subplot)
colour sequence

location
extent
colour gradient

area
size
divergent colours

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

religious adherents as a percentage of all residents



Nominal: no comparison is meaningful

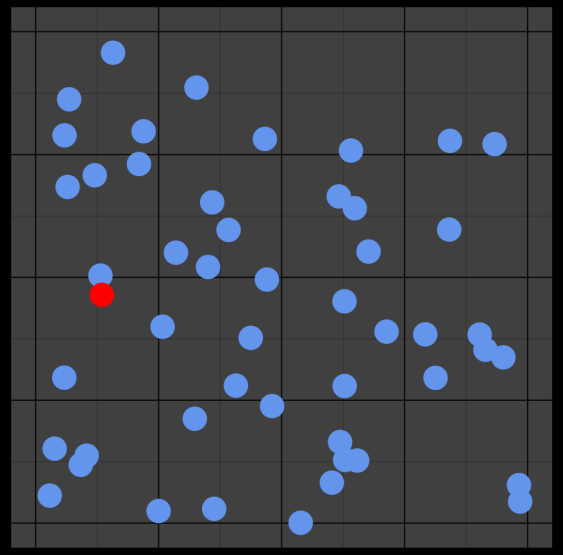
This is dumb!

How can I say “no comparison is meaningful” — and at the same time render onto a y scale?

See next lecture, on *embeddings*.

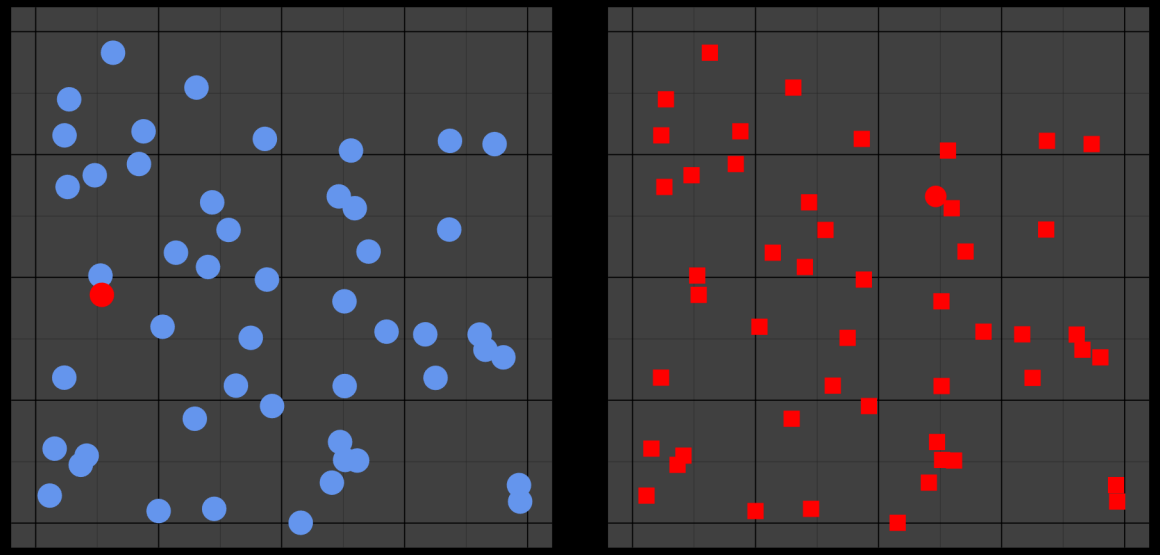
3. Scale perception

Is there a red circle?



3. Scale perception

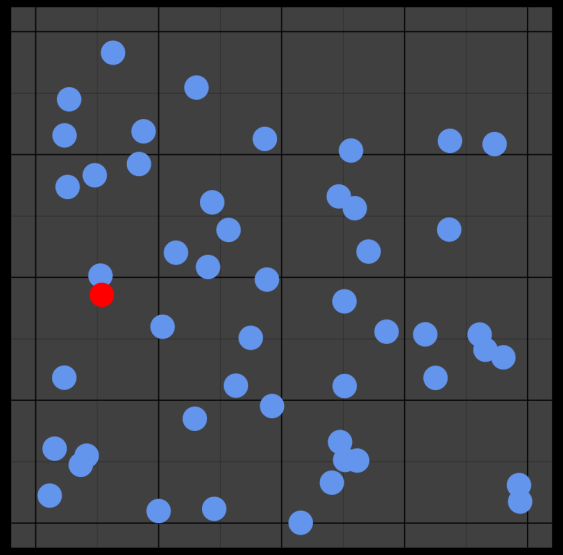
Is there a red circle?

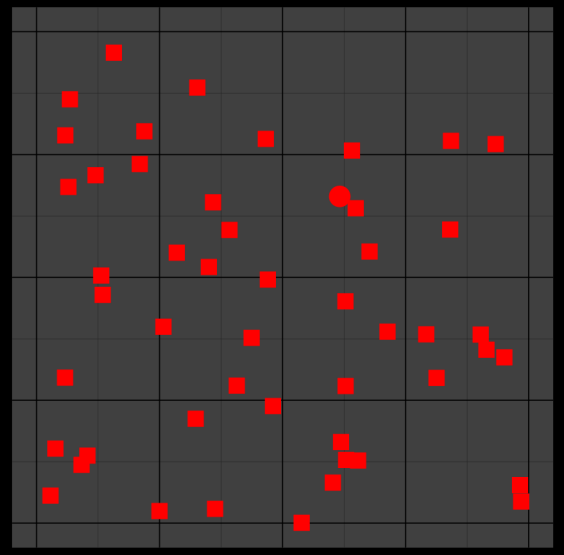


We can see colour differences more easily than glyph differences.

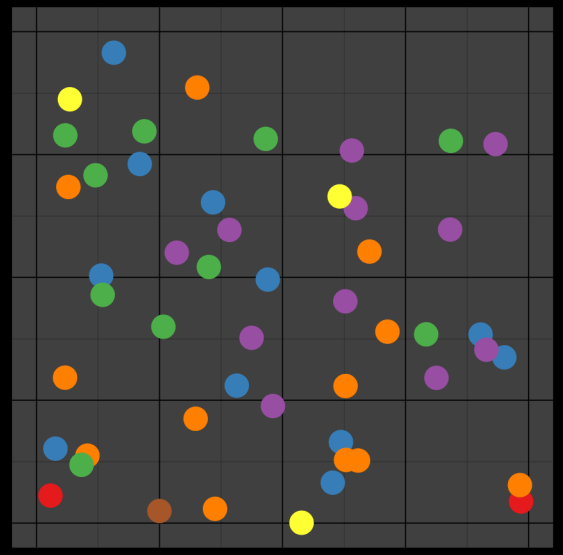
3. Scale perception

Is there a red circle?





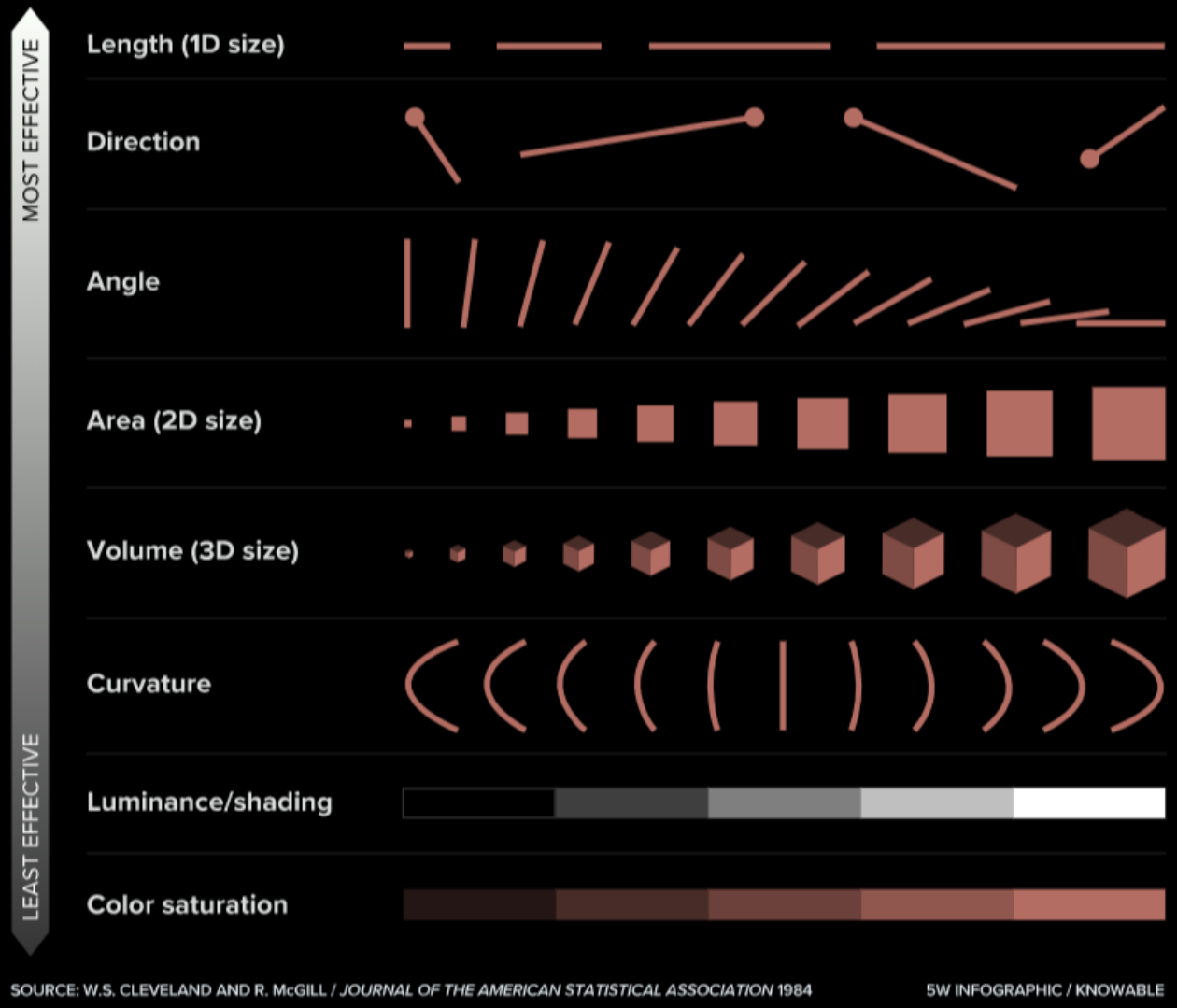
We can see colour differences more easily than glyph differences.



Less is more.

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

Some scales are more effective than others at communicating differences.

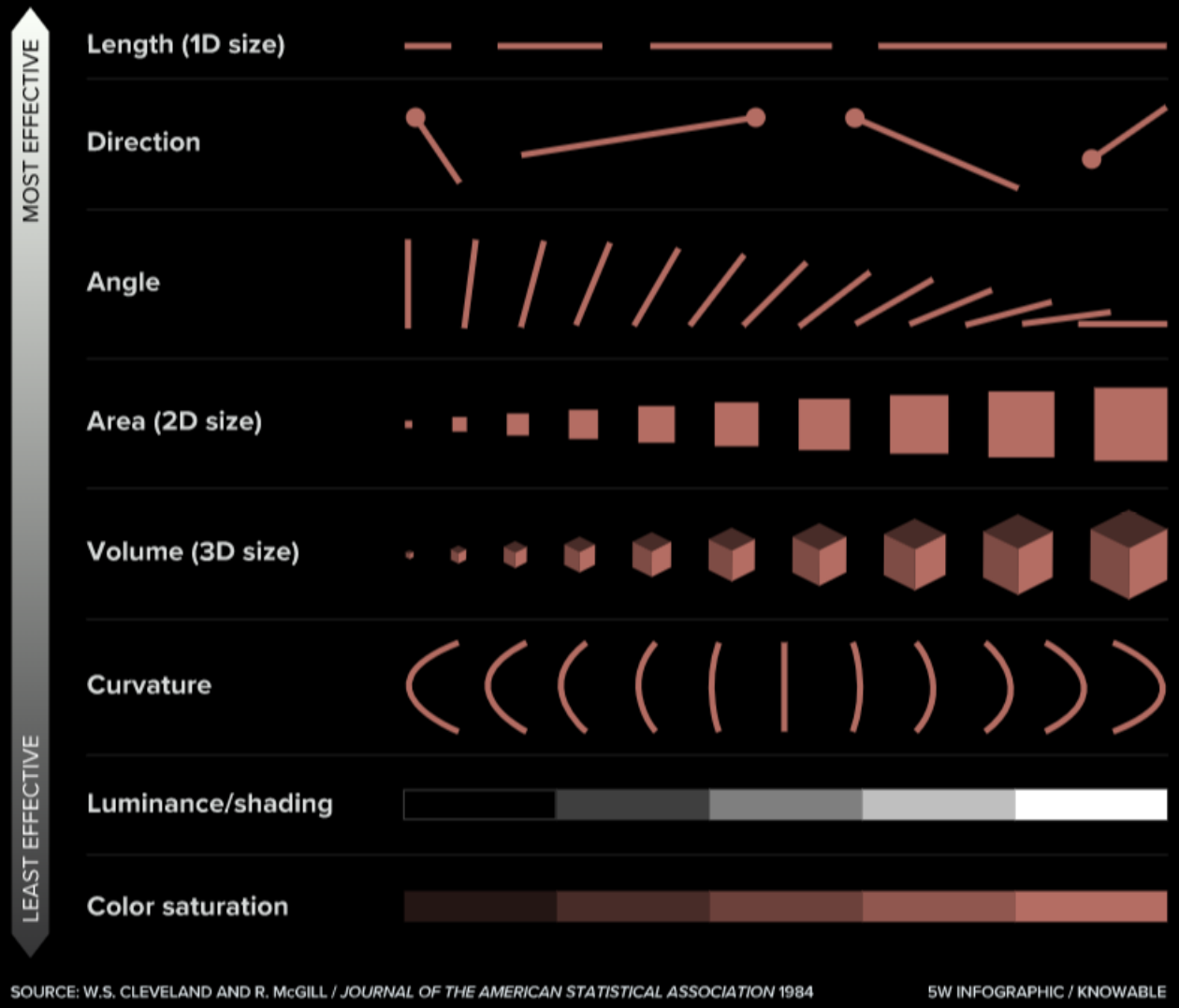


Location is the easiest scale from which to read off differences

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

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

Some scales are more effective than others at communicating differences.



Location is the easiest scale from which to read off differences

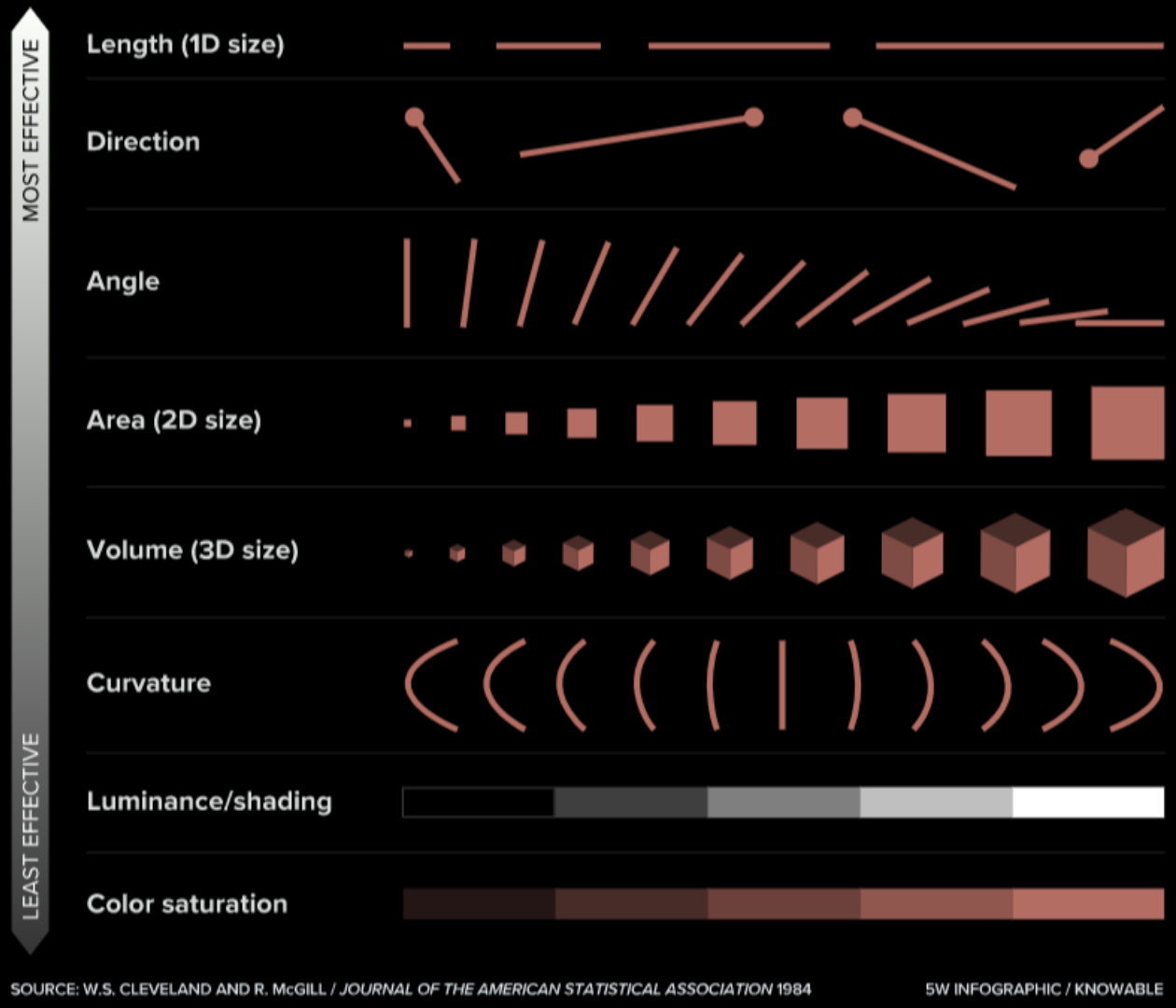
Note that the subplot index (in a multipanel plot) is also a type of location scale

Area is dangerous

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

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

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

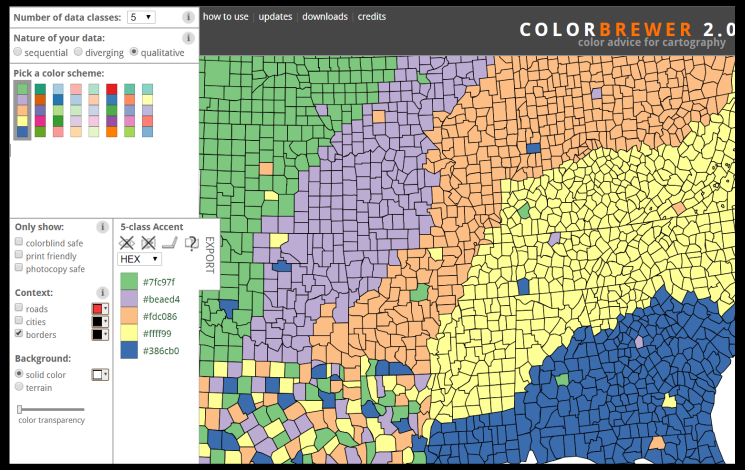
Human perception of colour is tricky because of perceptual issues

It’s not surprising if you see blue as water, and grey as a neutral background.

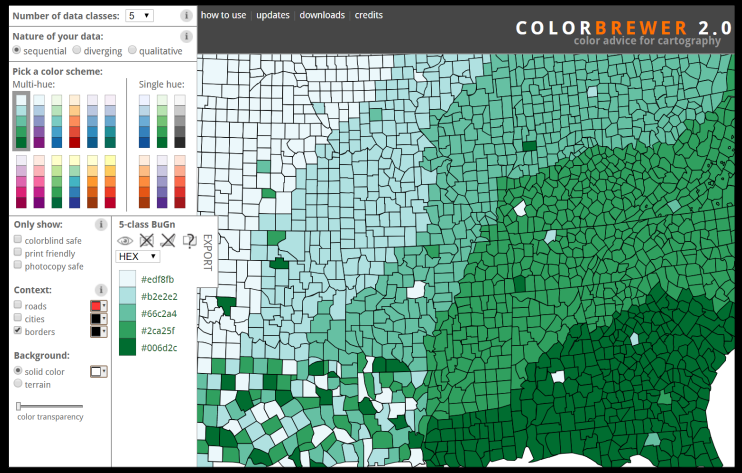
This leaves Spain and Scandinavia and the oceans as the “land mass”!

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

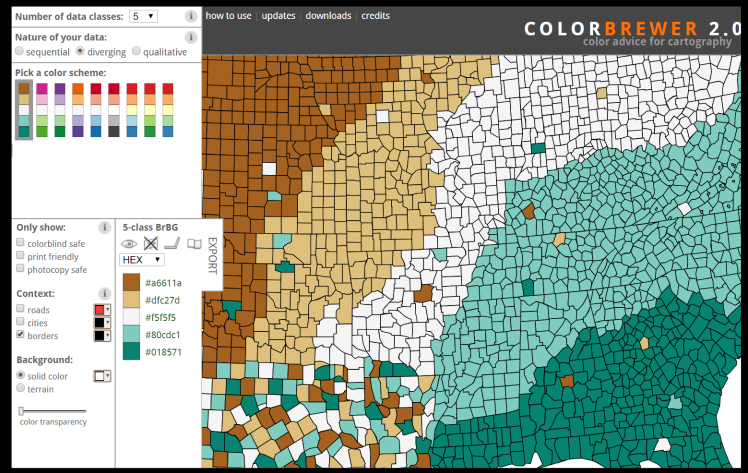
Nominal: no comparison is meaningful



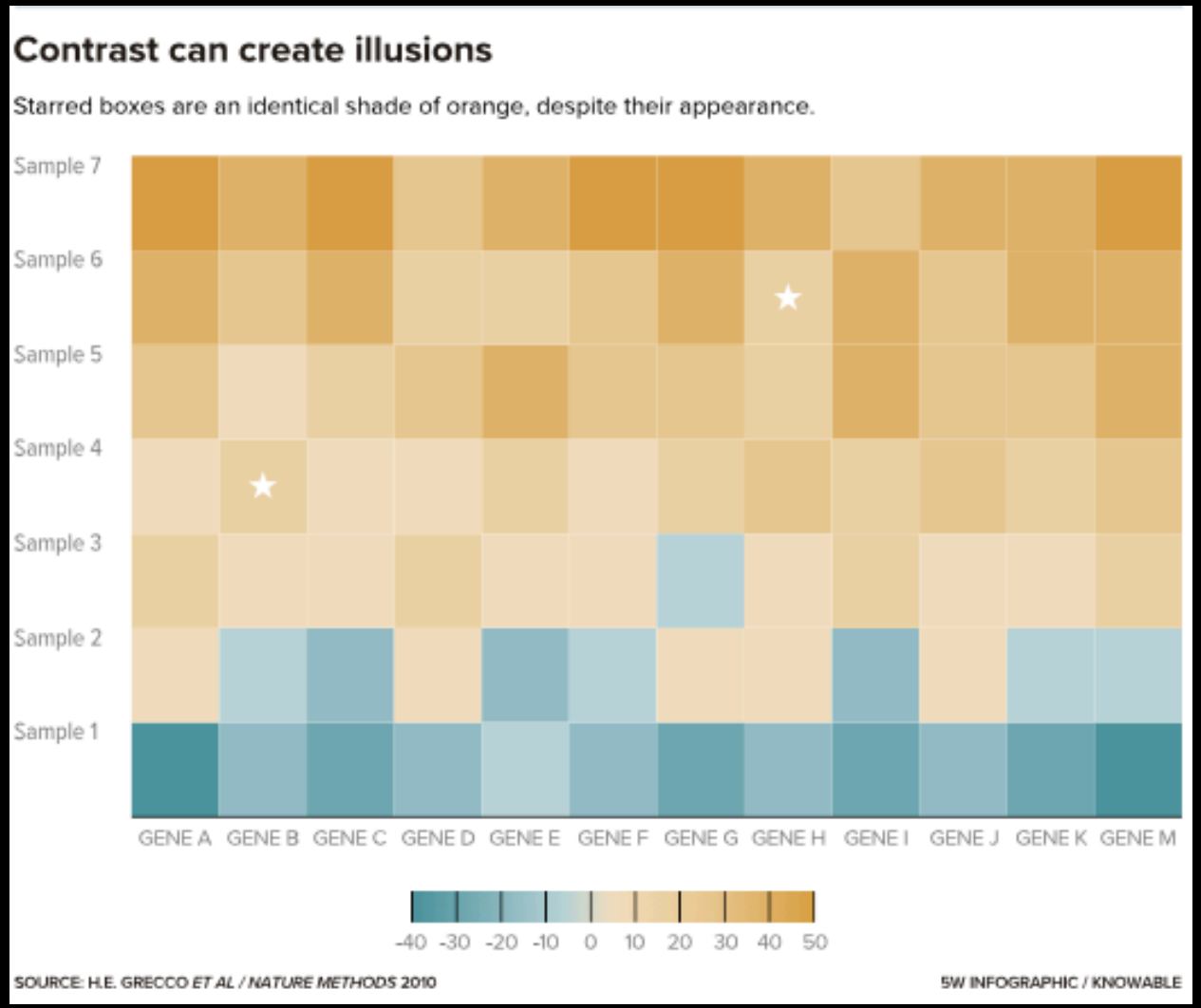
Ordinal: we can ask which is greater, but not measure how much



Ratio: we can divide one value by another



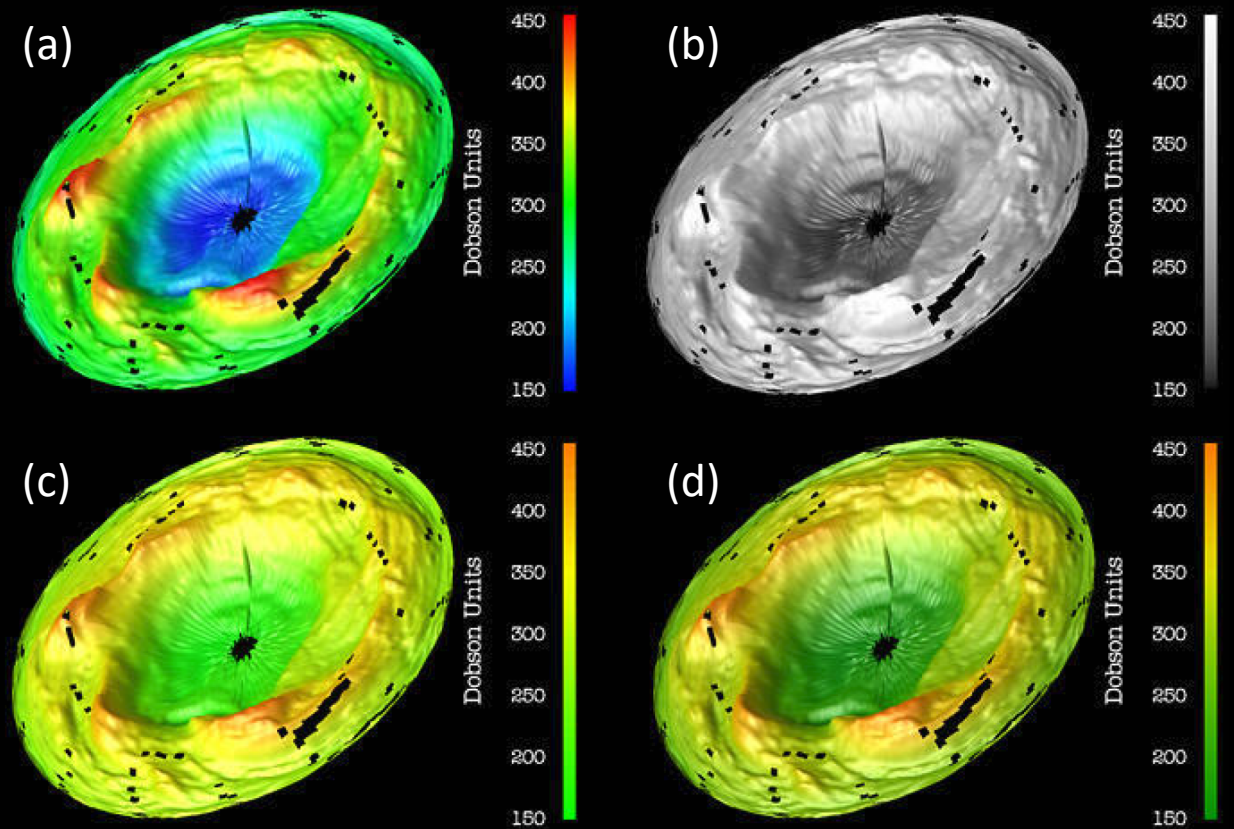
Human perception of colour is tricky. Contrast can create illusions.



The two starred squares on this heat map are identical shades of orange, indicating identical values in terms of gene activity. But differences in the colour of neighbouring squares means that the starred ones don't look identical, which can be misleading.

https://knowablemagazine.org/article/mind/2019/science-data-visualization

Hue is good for showing low-frequency effects, brightness is good for high-frequency effects

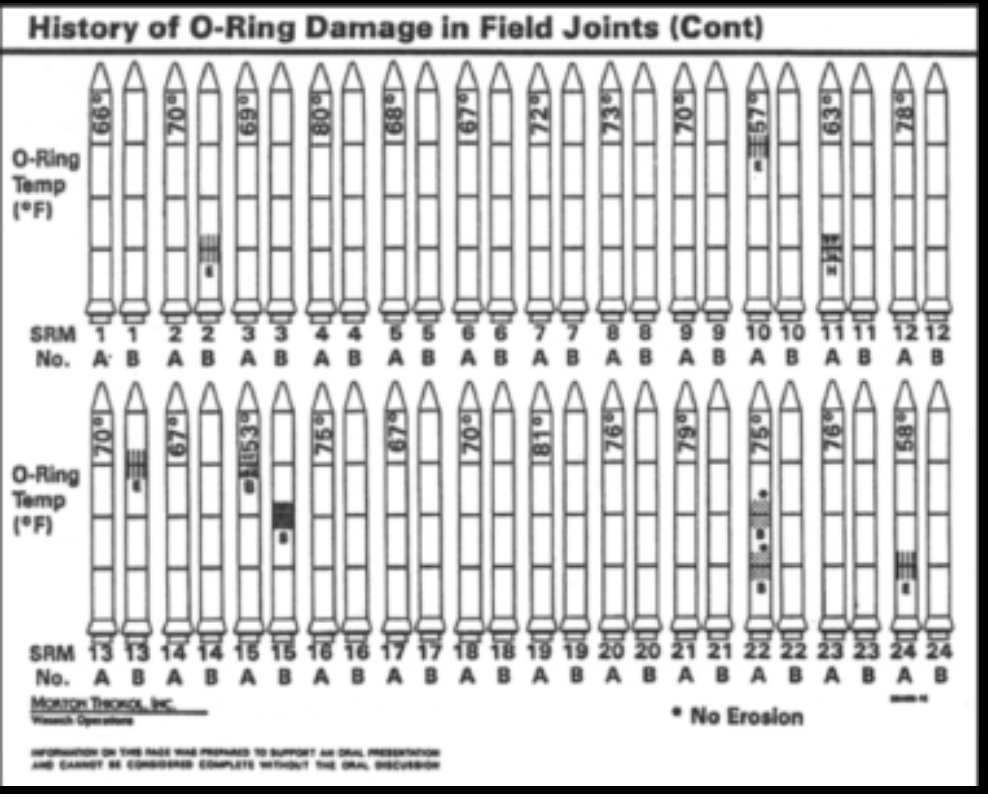


DATASET: total column density of ozone above the southern hemisphere (*Why Should Engineers and Scientists Be Worried About Color?* Rogowitz and Trienish, 1998)

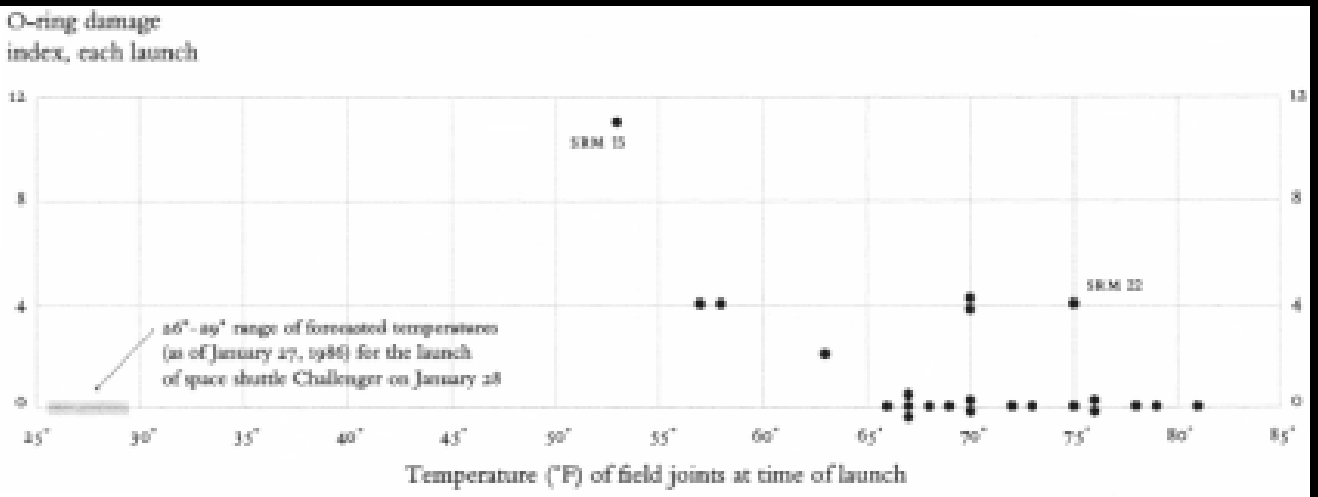
(a) rainbow palette
(b) brightness palette
(c) divergent hue palette
(d) combines (b) and (c)

Do not use the rainbow palette, except to show rainbows.

Chartjunk obscures the scales. (From Tufte, *Visual Explanations.)*



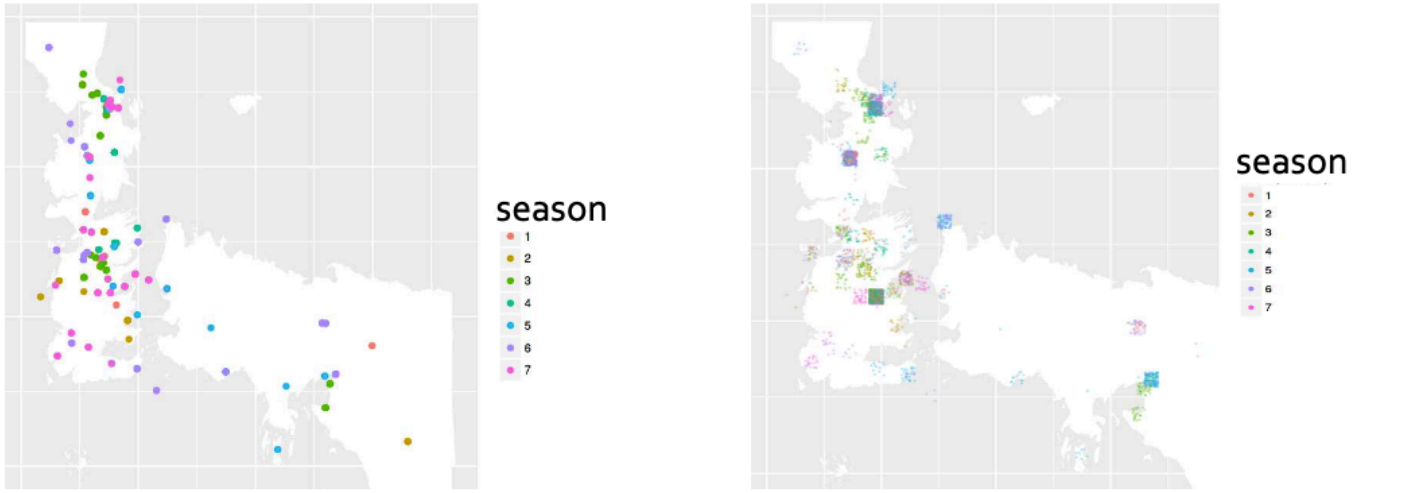
This is the famous slide that persuaded NASA management that it was safe to launch the Challenger shuttle.

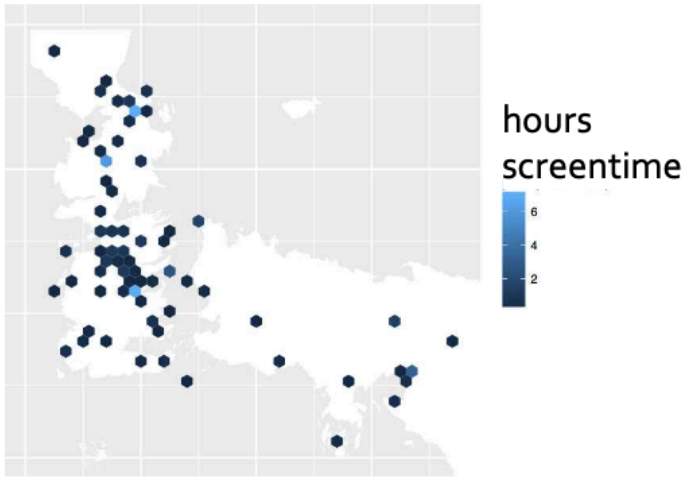


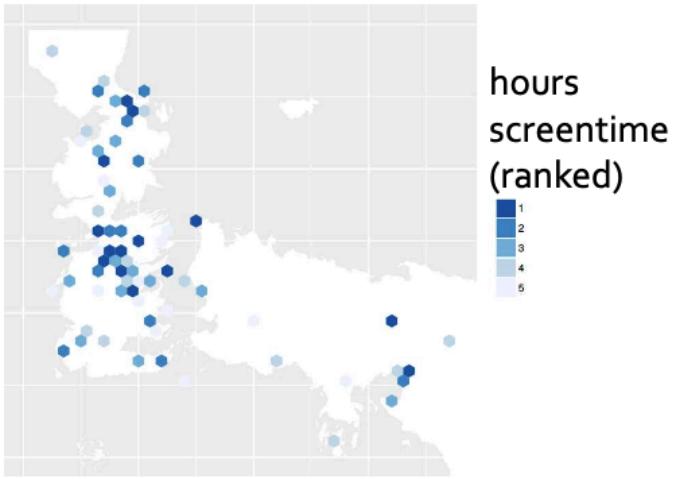
Tufte's redrawn chart makes it clear just how dangerous the launch was likely to be.

Overplotting

Dataset: scenes in *Game of Thrones*. Attributes are season, screentime, location.







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.*

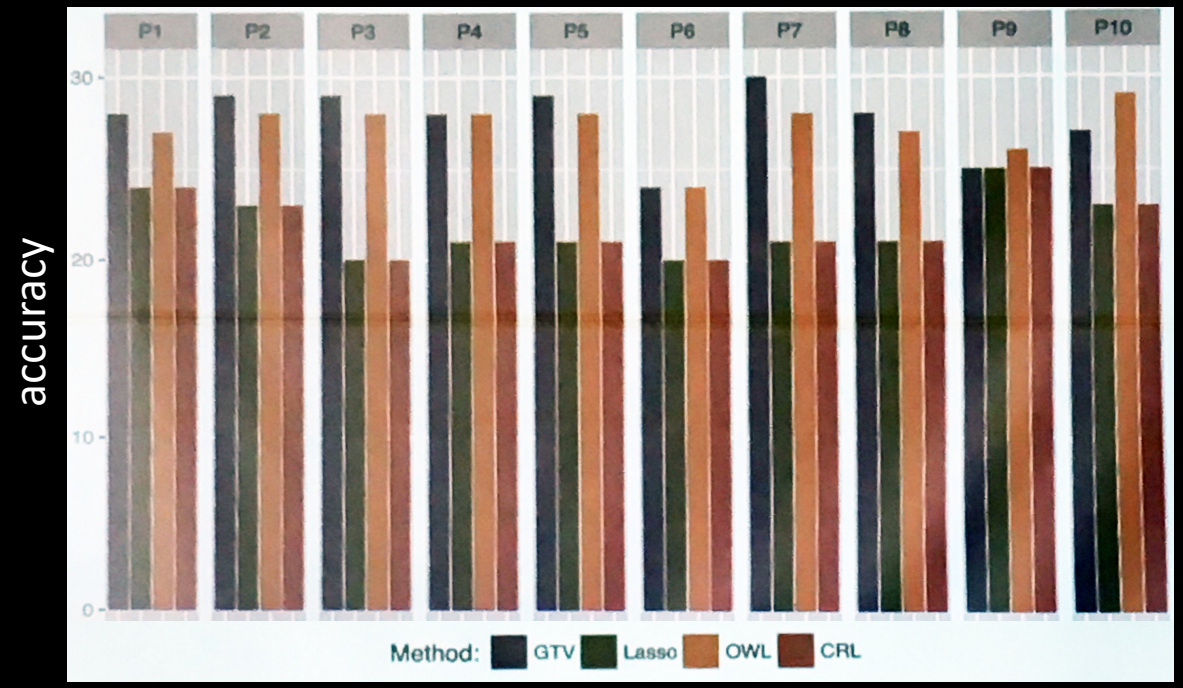
Mathematics and the picturing of data, John Tukey, 1975

- Plots are inviting the viewer to make comparisons (how does feature A depend on B, C, or D?)
- So you should 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.

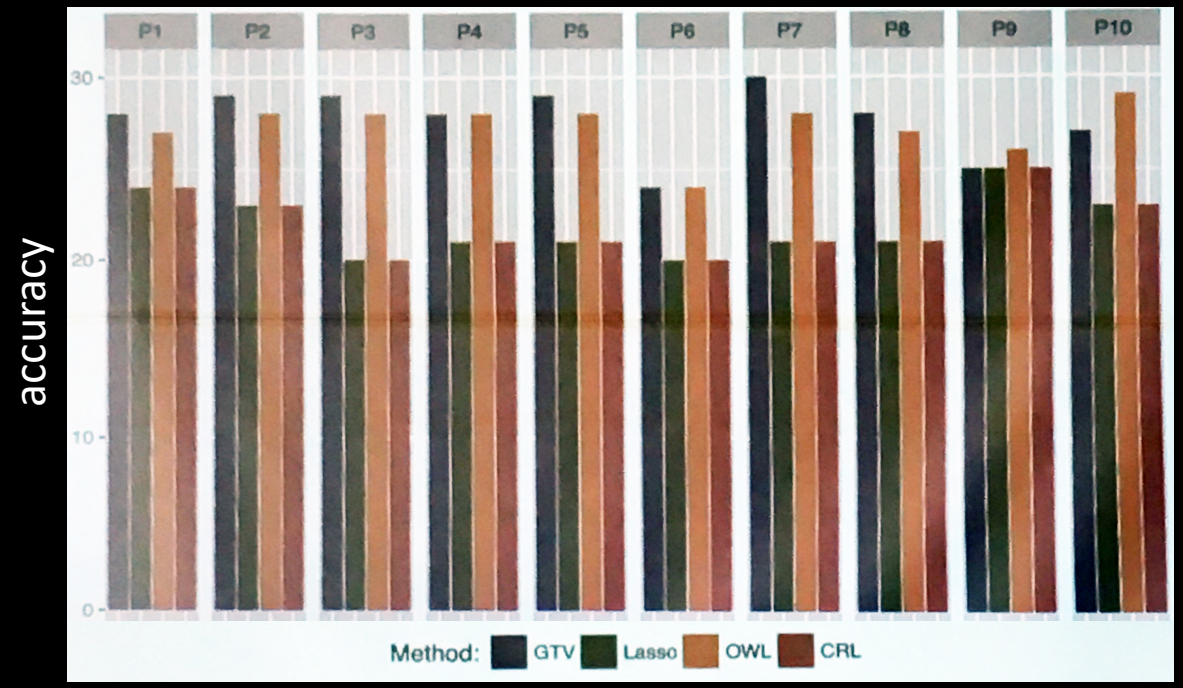
patient ID	classification algorithm	accuracy score
p2	lasso	0.228
p3	owl	0.279
p3	crl	0.197
⋮	⋮	⋮



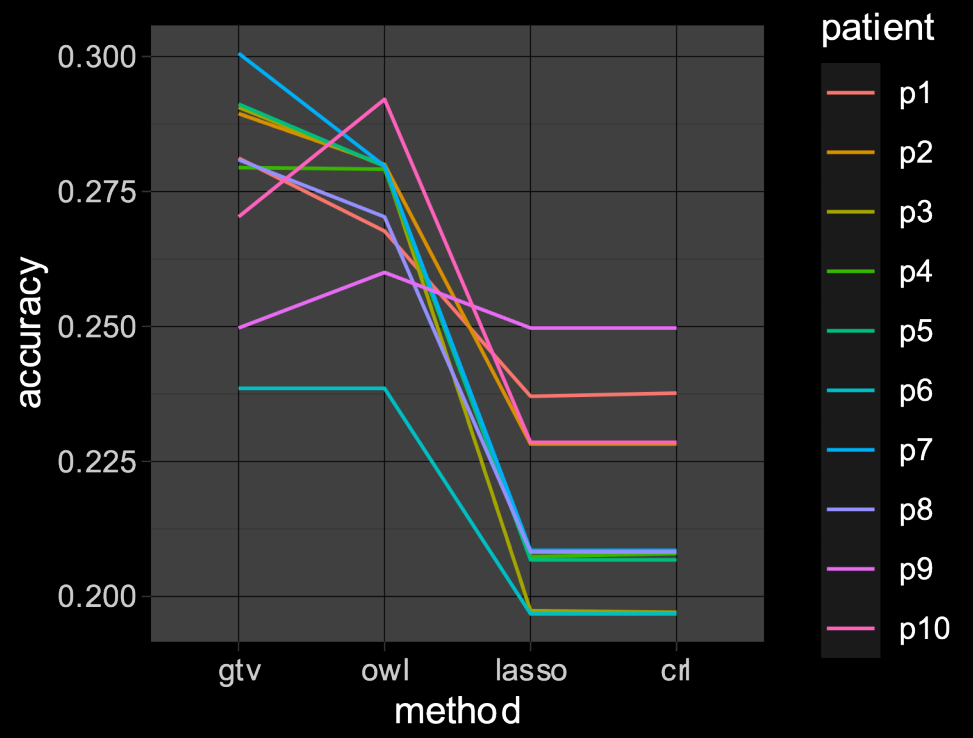
What comparisons does the plot invite?

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
⋮	⋮	⋮

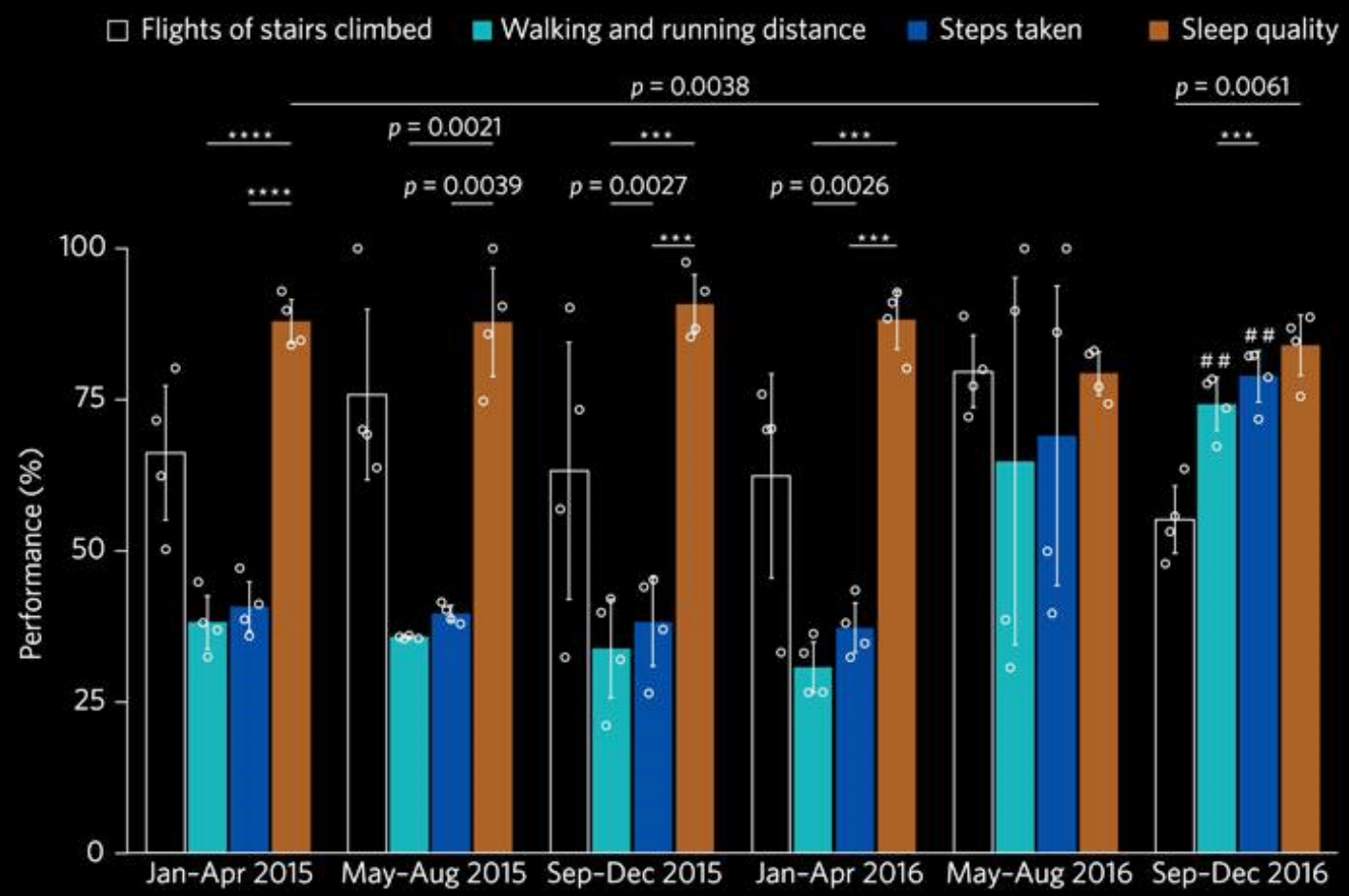


- The main comparison is "how does accuracy depend on algorithm?"
  So put this on x and y scales.
- The “line” geom used here helps make between-patient comparisons.
  Between-patients is less important than between-algorithms, so a hue scale is fine for patient ID.



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

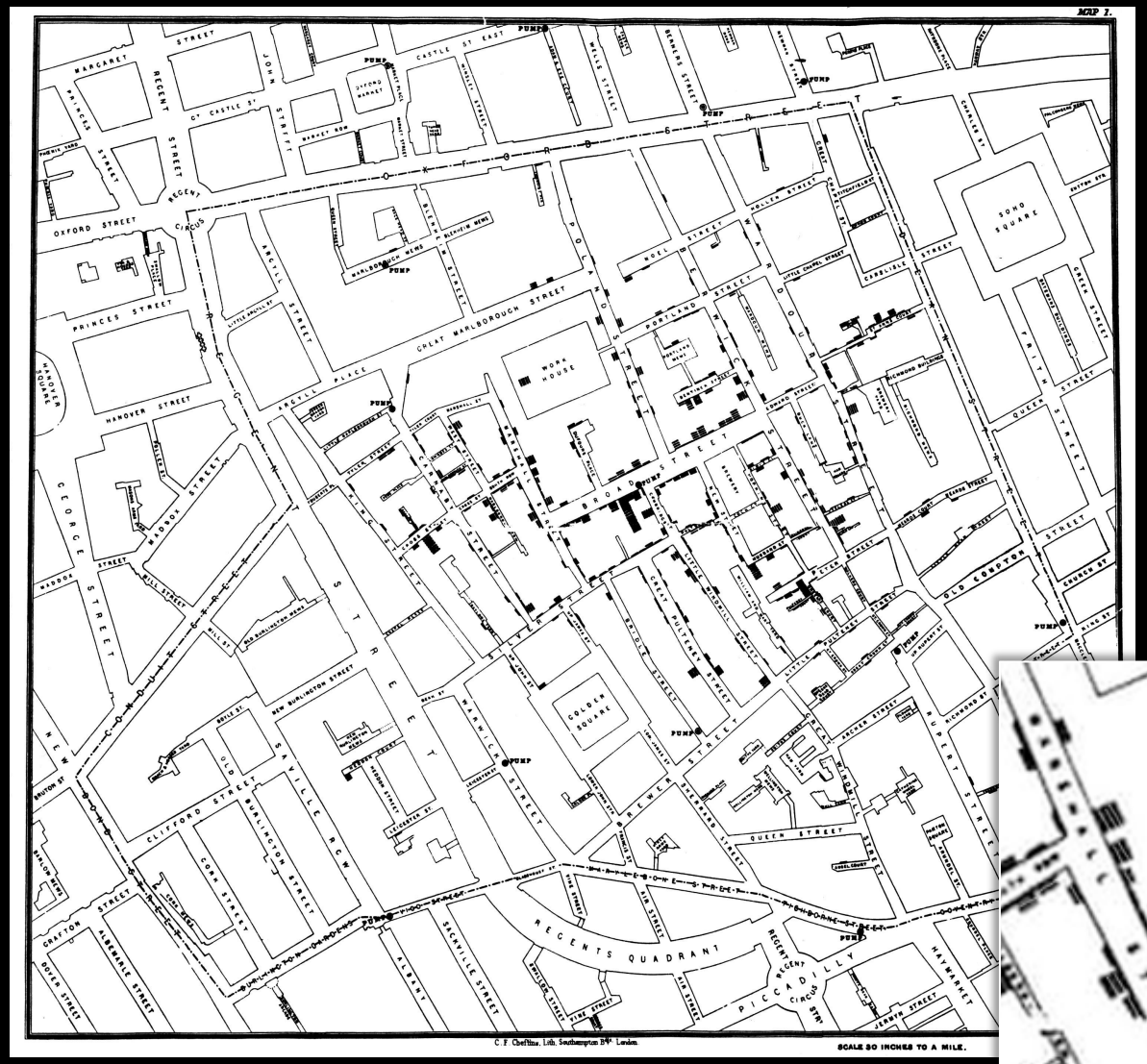
What comparisons does the plot invite?

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.

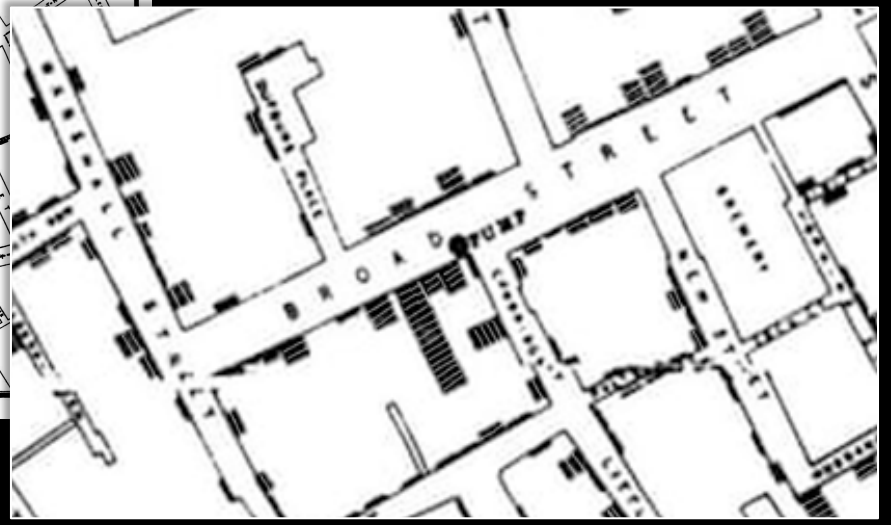
This is a very effective way to compare waveforms from one period to the next.

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

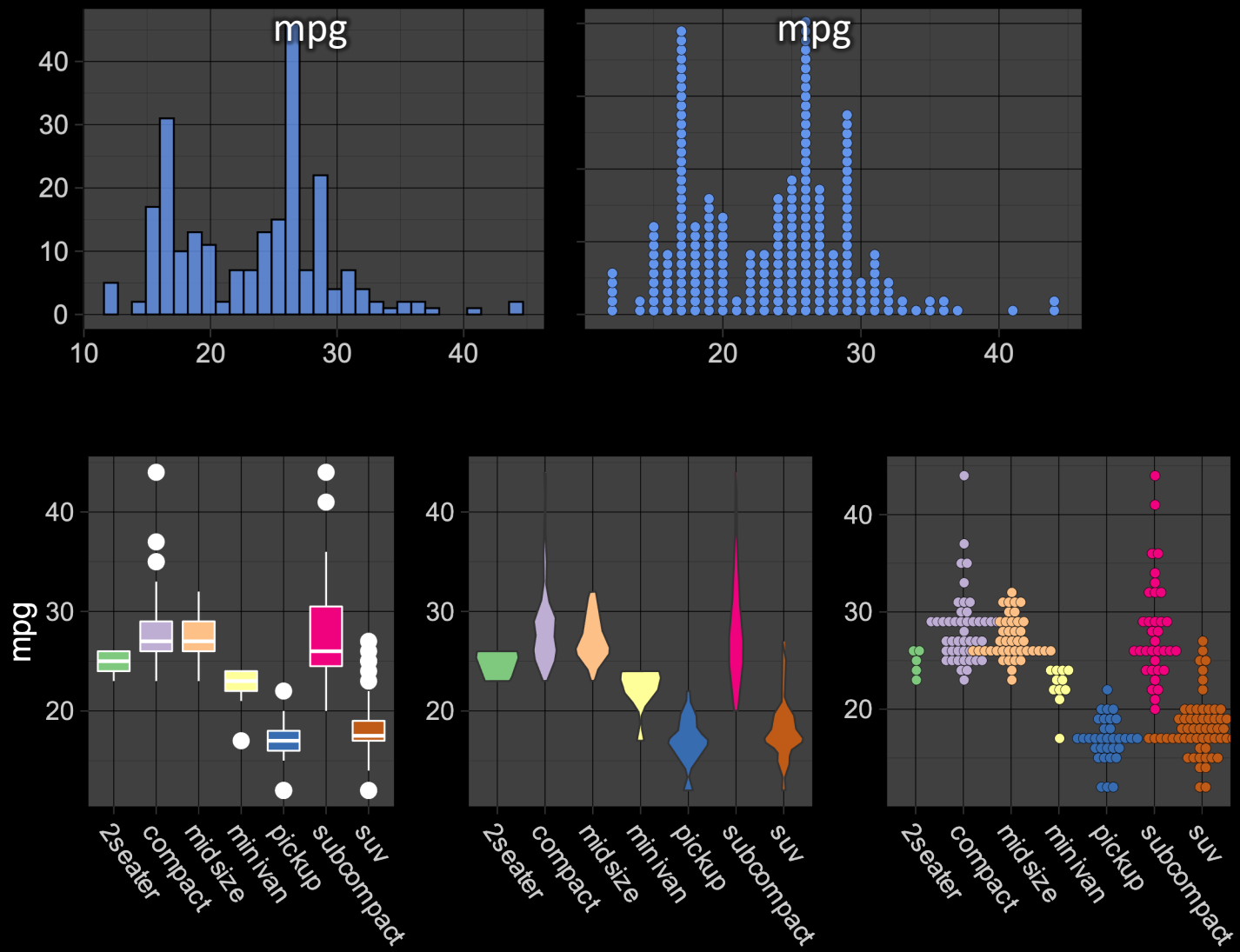


Each bar is a person who died from cholera. The bars have been stacked.

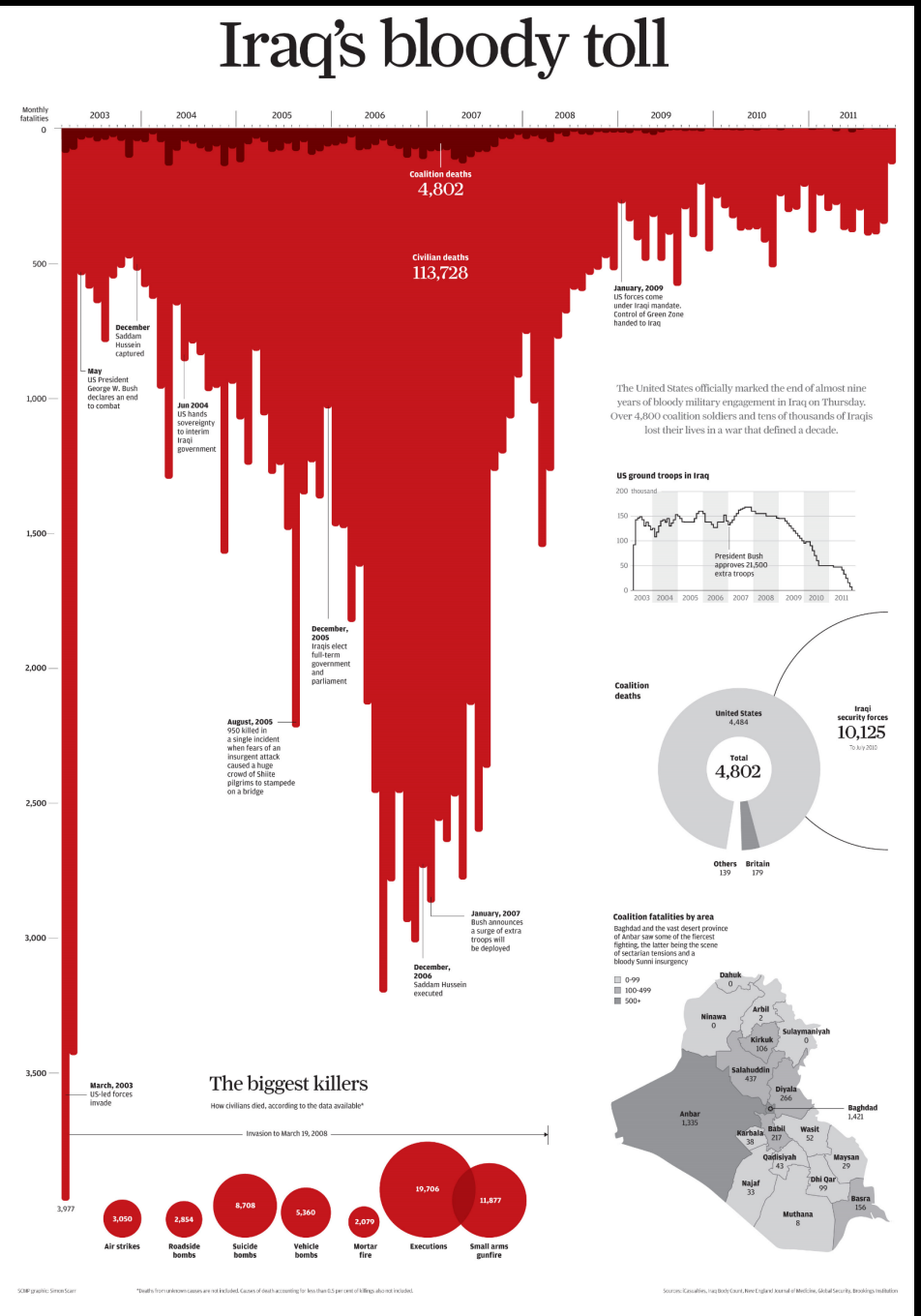


John Snow, 1854 https://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map

In the best plots, every dot of ink is a datapoint.
This is why histograms are easy to read.
Here are several different styles of histogram, from a dataset of miles per gallon (mpg) for a variety of cars.

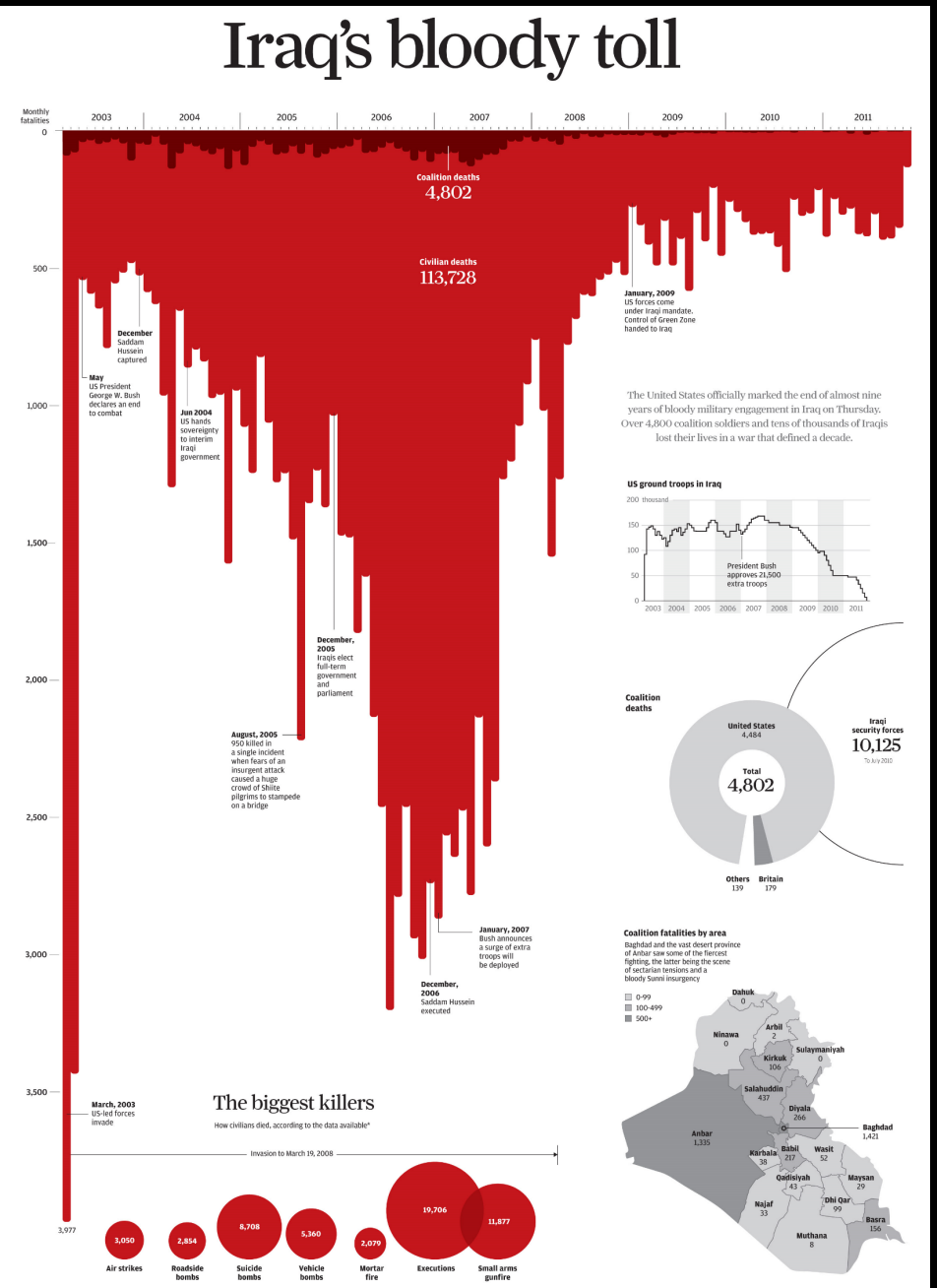


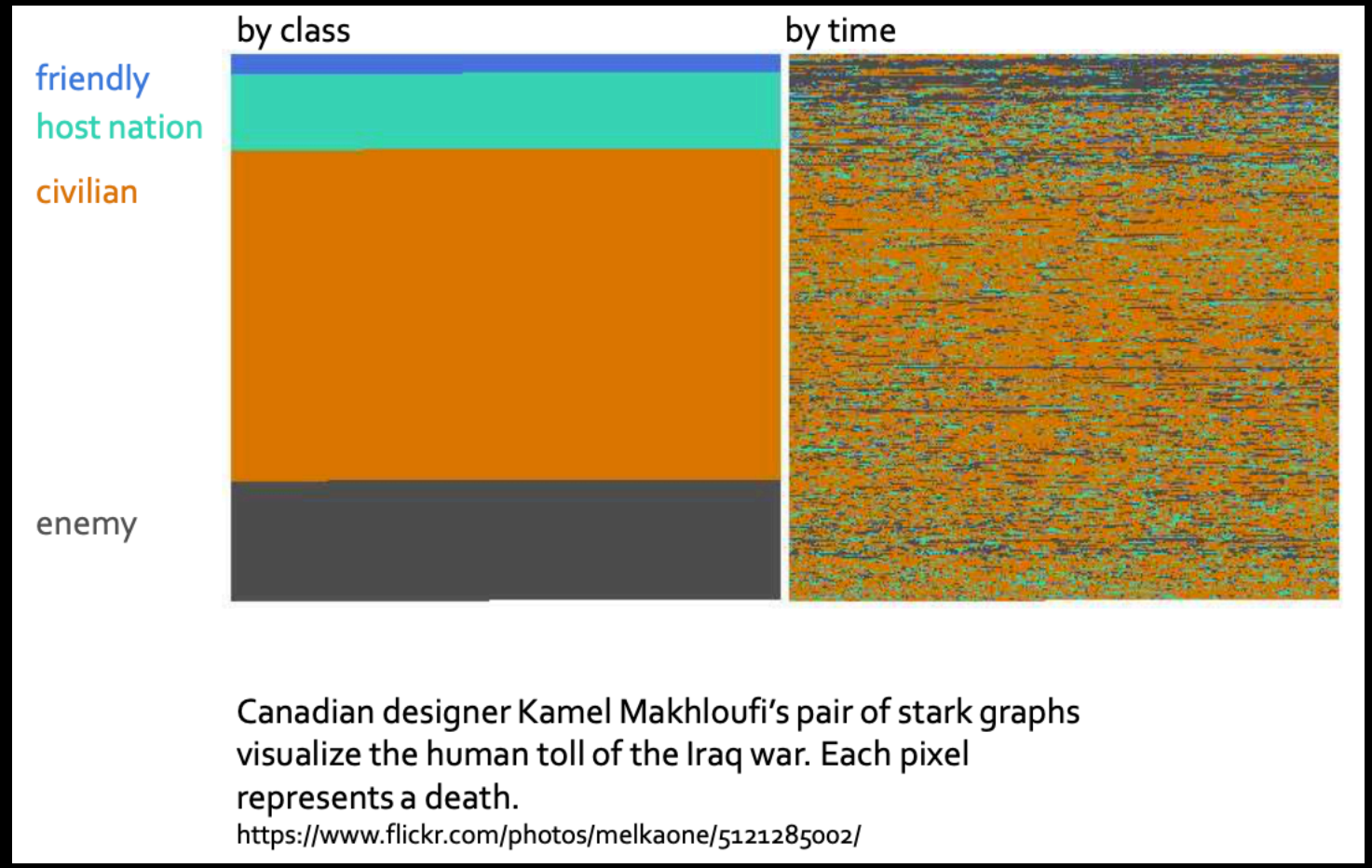
In the best plots, every dot of ink is a datapoint.



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.





Canadian designer Kamel Makhloufi's pair of stark graphs visualize the human toll of the Iraq war. Each pixel represents a death.
https://www.flickr.com/photos/melkaone/5121285002/

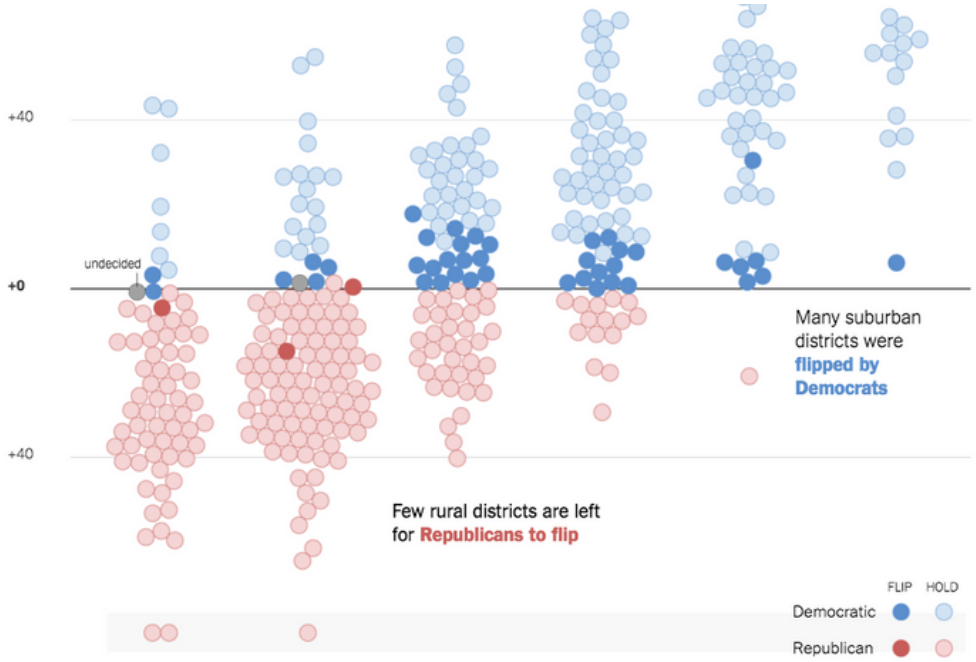
South China Morning Post
https://www.scmp.com/infographics/article/1284683/iraqs-bloody-toll

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

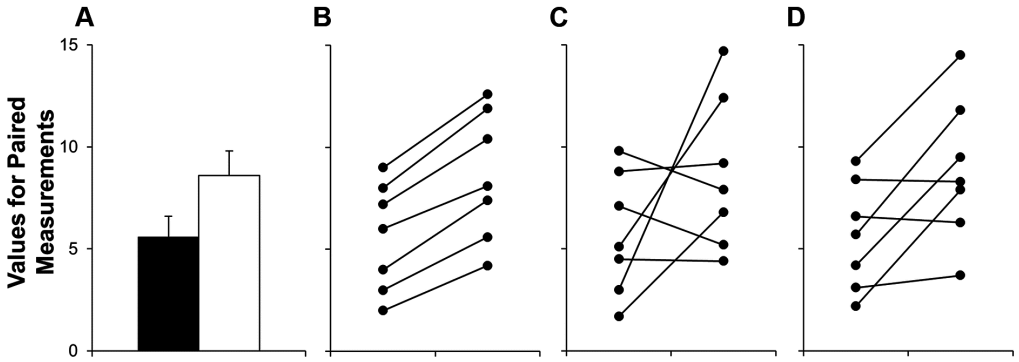
NASA earth observatory
https://earthobservatory.nasa.gov/images/87551/london-at-night

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.

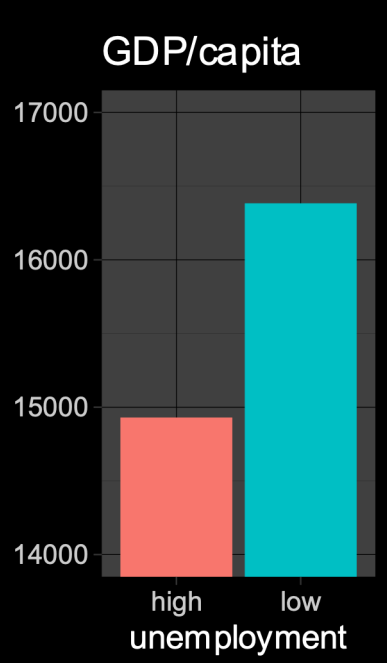


Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

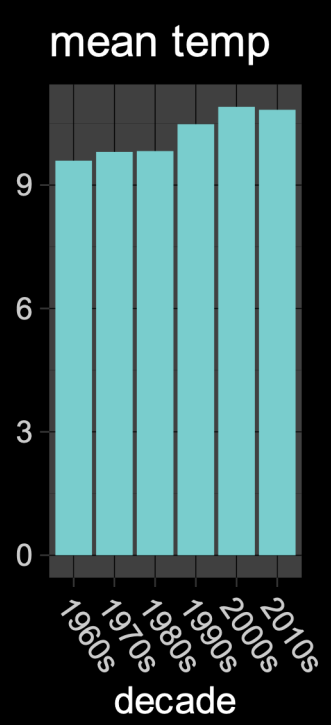
Show the dots in plots
https://www.nature.com/articles/s41551-017-0079

Rules for atomic plots

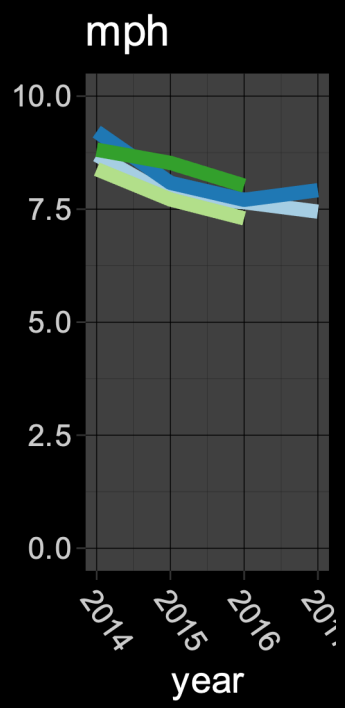
- If your data scale is accumulative:
  show it with a histogram, and let the size be accumulated mass.
- If your data is not accumulative:
  don't use bars, because they convey the impression of mass.



Average annual temperature [°C] in Cambridge



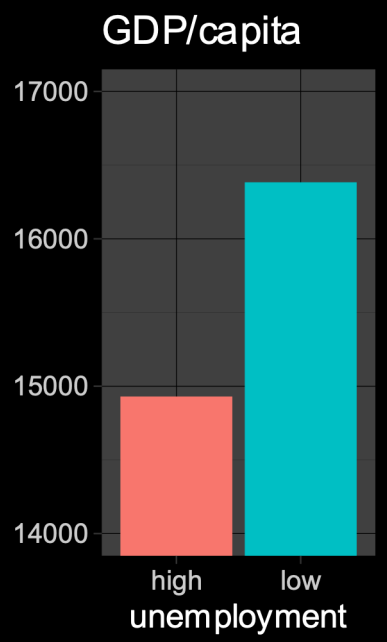
Average daytime speed in central London, major roads



Rules for atomic plots

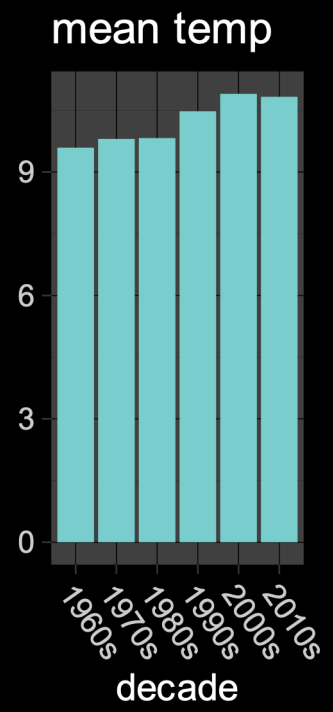
- If your data scale is accumulative:
  show it with a histogram, and let the size be accumulated mass.
- If your data is not accumulative:
  don't use bars, because they convey the impression of mass.

GDP is accumulative,
so we should let size = GDP,
so y-axis should start at 0.



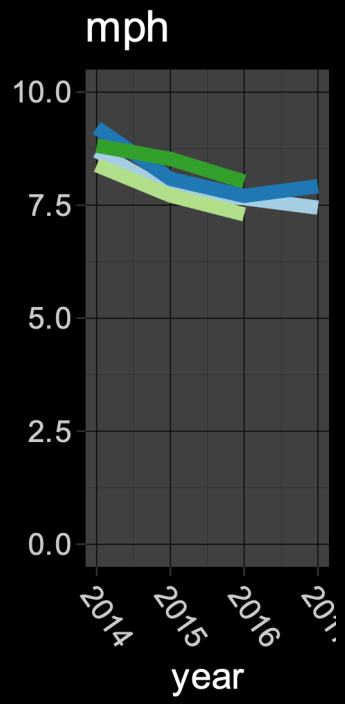
Temperature is not accumulative, so don't use bars.

Average annual temperature [°C] in Cambridge

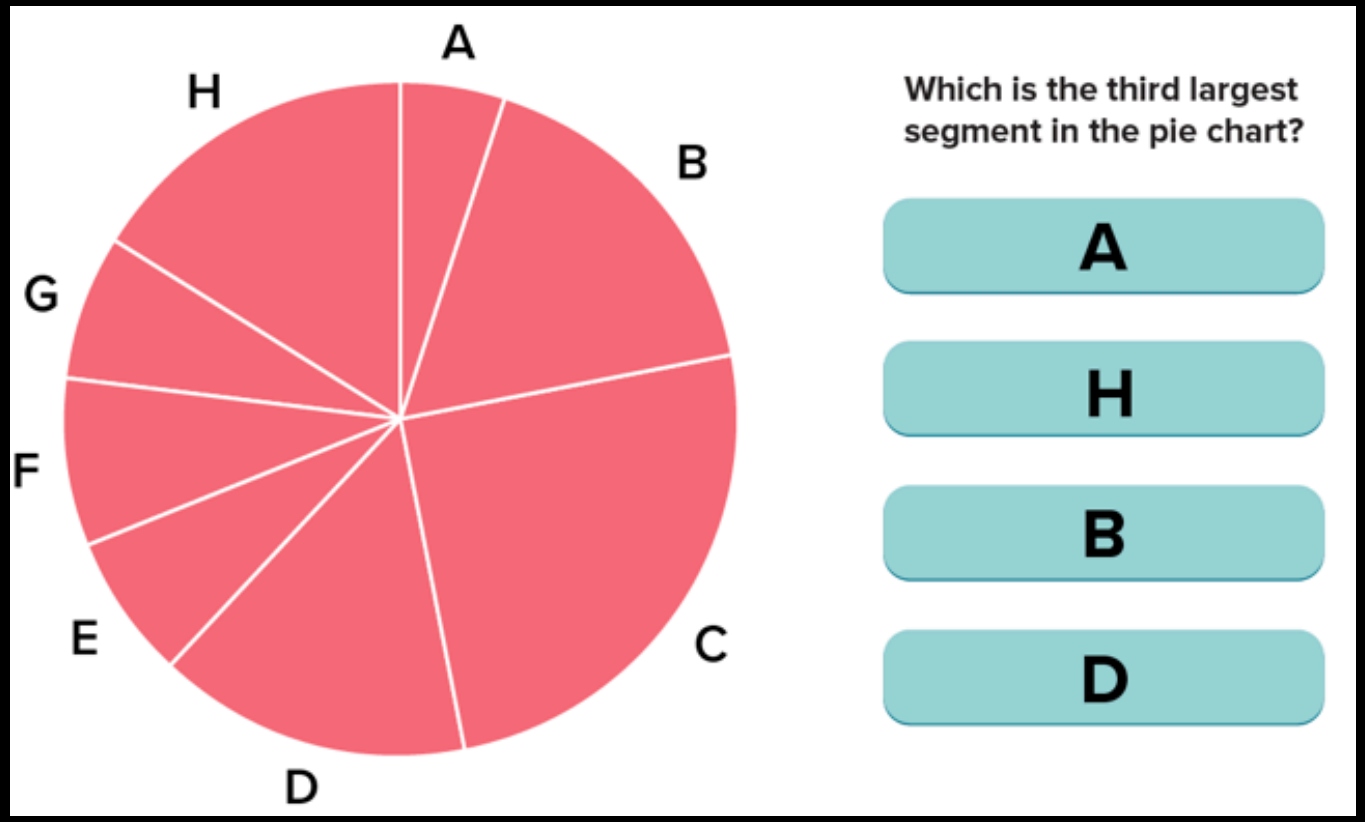


Correct!

Average daytime speed in central London, major roads



EPILOGUE When to use pie charts

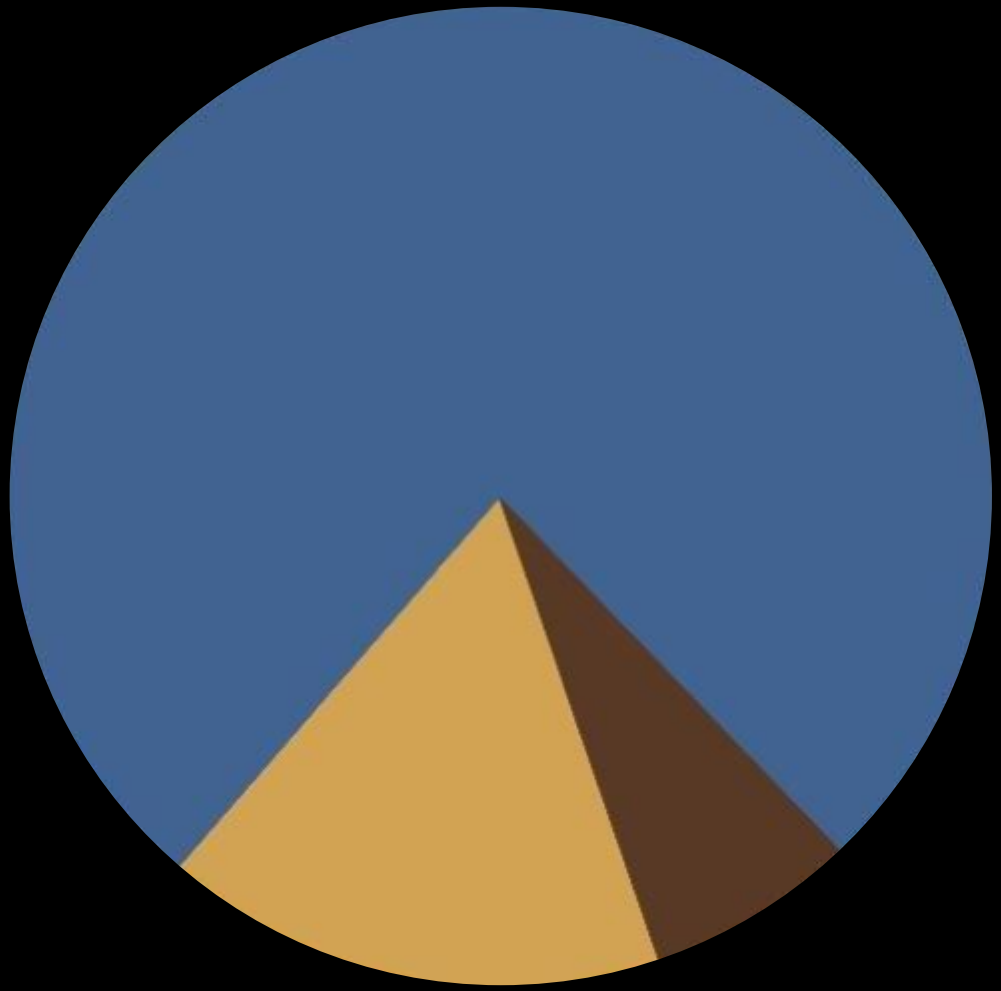


"The only design worse than a pie chart is several of them"

E. Tufte, *The Visual Display of Quantitative Information*, 1983

https://knowablemagazine.org/article/mind/2019/science-data-visualization

EPILOGUE When to use pie charts



EPILOGUE When to use pie charts