VISUALISATION

Lecture 6: graph literacy
Lecture 7: machine learning tools

Damon Wischik

This lecture is about arrangement, i.e. how you organize the elements of a chart to convey a message.
It's not about software libraries: you can learn those from StackOverflow and from tutorials, or from a graduate skills class "Visualizing data with R/ggplot2".

What is visualization 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

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

See the distribution = what groups of X values are there? = a bit like unsupervised learning / clustering
Make comparisons = how does Y depend on X? = a bit like supervised learning / regression

There are two audiences

- 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.

Scales

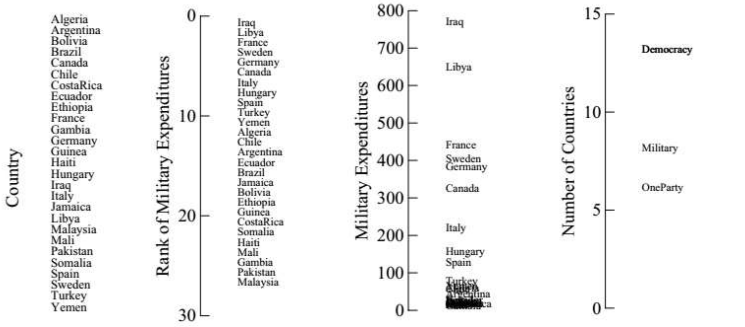
According to *On the theory of scales of measurement* (Stevens 1946) there are four types of scale:

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

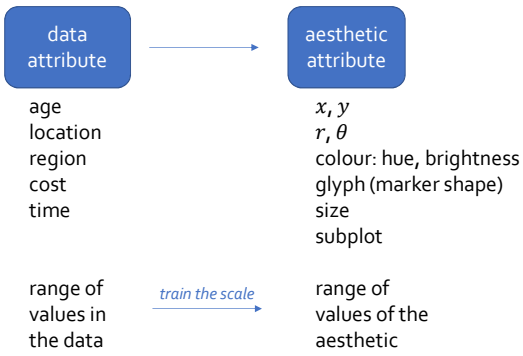
Ratio: we can divide one value by another



This isn't really true, but it's a good place to start.

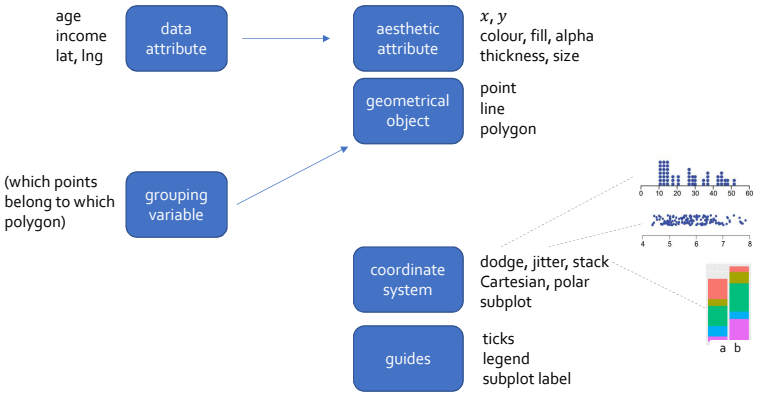
These are the four types of quantity you might find as columns in a CSV.

Scale mapping



- The type of scale constrains the type of chart you should make
  - For x and y interval or ratio a line graph is reasonable;
    for x nominal or ordinal then there are no intermediate values so a bar chart is better
  - Glyph only makes sense for nominal scales
- By choosing scales, you invite the viewer to make comparisons
  - Some aesthetics are easier to compare than others, e.g. y coordinates are easy to compare, hue is harder
  - Choose scales to prioritise the comparisons you want.

Components of a chart

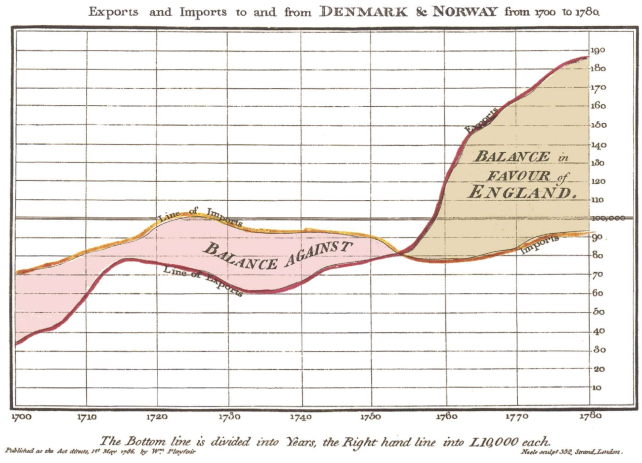


Some tips

- **Colour scales**
  Nominal, ordinal, interval, ratio: these all demand different palettes.
  The psychophysics of colour perception is tricky: don't invent your own scales.

- **Space scales**
  If your data is nominal, think about its order.
  Consider ranking (interval→ordinal) and binning.
  Watch out for over-plotting.

- **Perceptual gotchas**
  Lines banked around 45° are easiest to read.
  Stevens exponent: perceived area = (drawn area)$^{0.8}$.
  Fechner/Weber: you notice %difference in an sensation, not absolute difference.



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.

Fig. 1

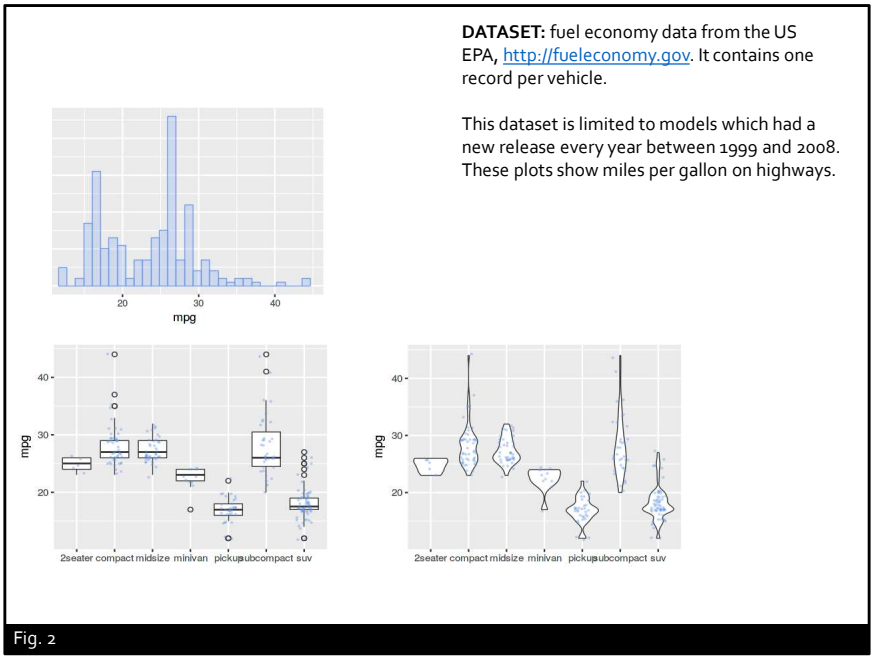


Fig. 2

- mpg is a ratio scale; vehicle type is nominal
- The bottom two charts have two scales, mpg and vehicle type
- The top chart has one scale, mpg

- In the top plot, there are multiple records in the dataset for each mpg. To make them visible, they have effectively been stacked up.
- In the bottom plot, there are multiple records in the dataset for each mpg and vehicle type. To make them visible, they have been randomly jittered.

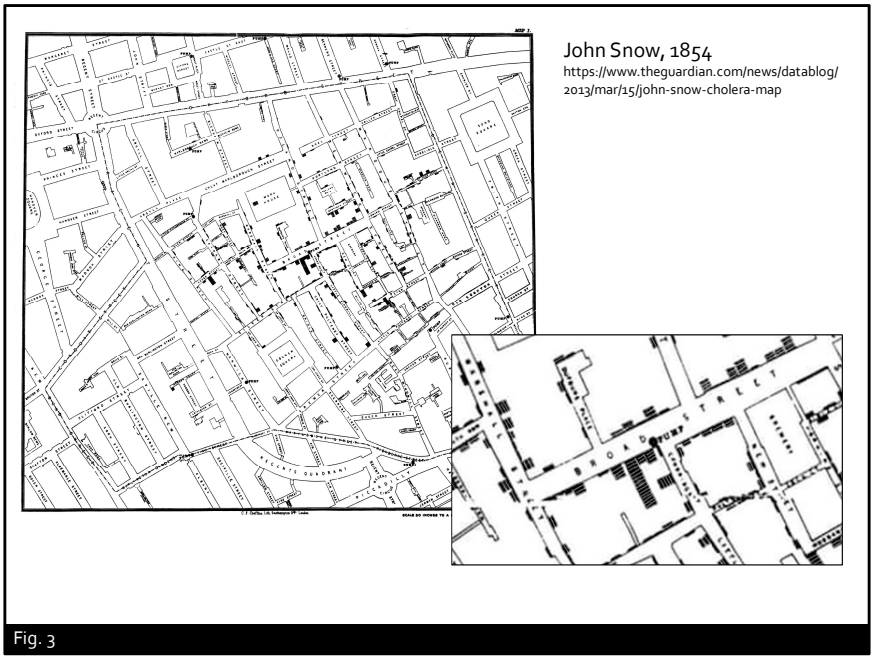
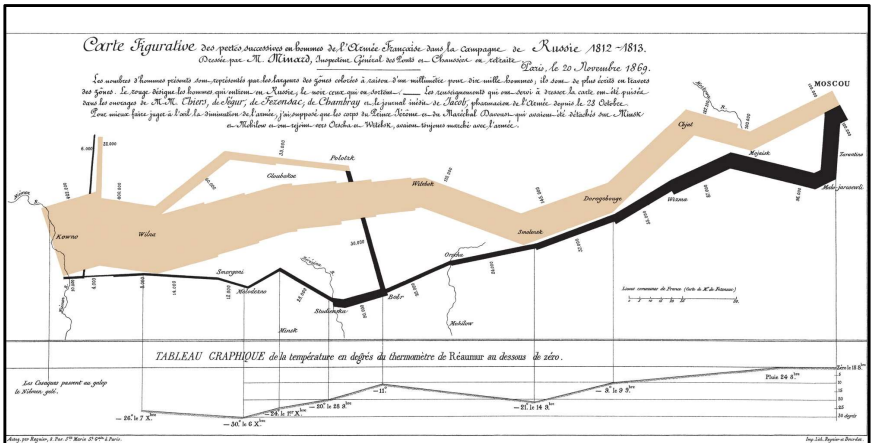


Fig. 3

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



Charles Minard's map of Napoleon's disastrous Russian campaign of 1812. The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates.

Fig. 4

Often the dataset has many attributes for each record, and it's a challenge to plot them all. This plot manages to show seven attributes at the same time.

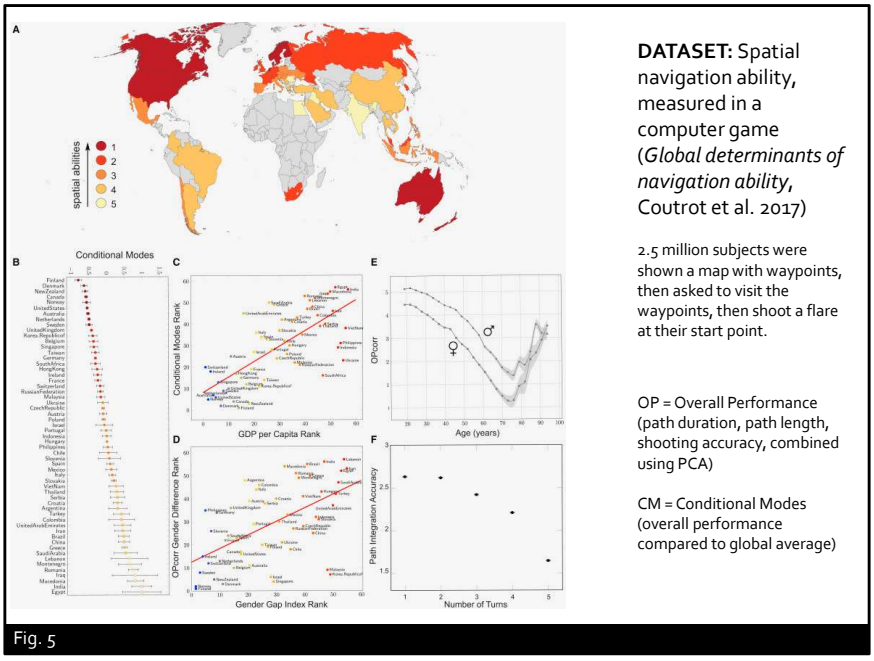
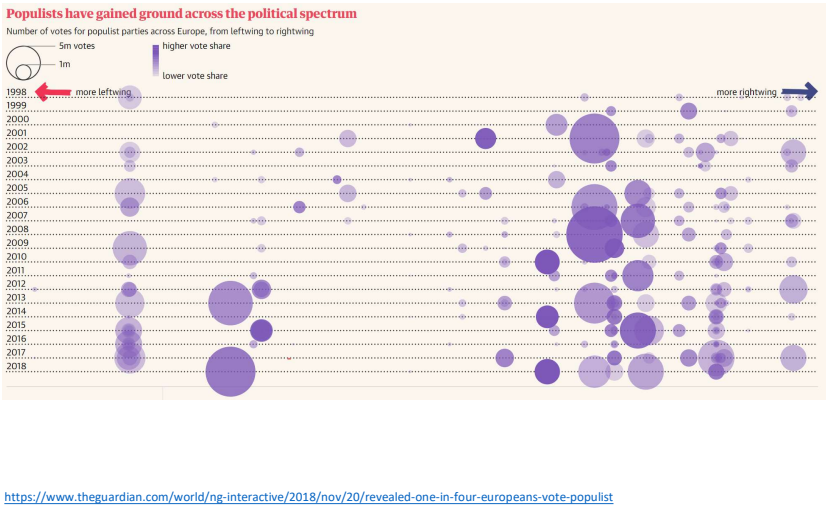


Fig. 5

More commonly, to show multiple attributes, you need to use multiple subplots.

The OP (overall performance i.e. spatial ability) score is interesting. It's a composite score made up of three separate attributes. Next lecture, we'll discuss ways to create composite attributes like this.

The y-axis in blot B shows country. This is a nominal scale, so we're free to order it how we see fit. Here it has been ranked (i.e. turned into an ordinal scale) by spatial ability.



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

Fig. 7

This plot shows four attributes (year=ordinal, vote share=interval, number of votes=ratio, political leaning=interval?)

What are we meant to learn from it? I can't make any sense of it. Are parties shifting their political leaning? Are all the dots different parties? Is populism getting worse?

Why would Labour clarify its stance on Brexit? Ambiguity is working

Anand Menon, The Guardian, 2 May 2018

https://www.theguardian.com/commentisfree/2018/may/02/labour-clarify-position-brexit-vote-share-leave-remain

What is apparent is the relative lack of Kensingtons or Canterburys waiting to fall to a Labour party expressing clearer opposition to Brexit. Moreover, it is in the bottom right quadrant – places where Labour gained more than average on the Conservatives last time, and that voted leave in the referendum – where the highest number of potential Labour gains are clustered.

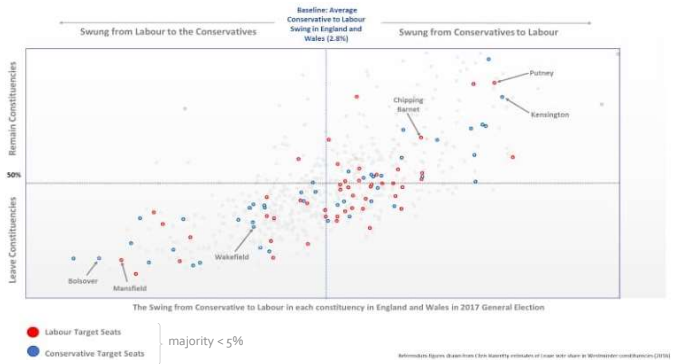


Fig. 8

This is another plot that is almost impossible to interpret without text. Even with text it's confusing. The text suggests that the x-coordinate has some meaning for how likely the seat is to change parties, but I don't see why that should be true; and without it the plot is useless. I would have chosen a plot where the x-coordinate is something unambiguous, like vote share.

Fig. 9

Colour scales are tricky because of perceptual issues. I instinctively see blue as water, and grey as a neutral background. This leaves Spain and Scandinavia and the oceans as the “land mass”!

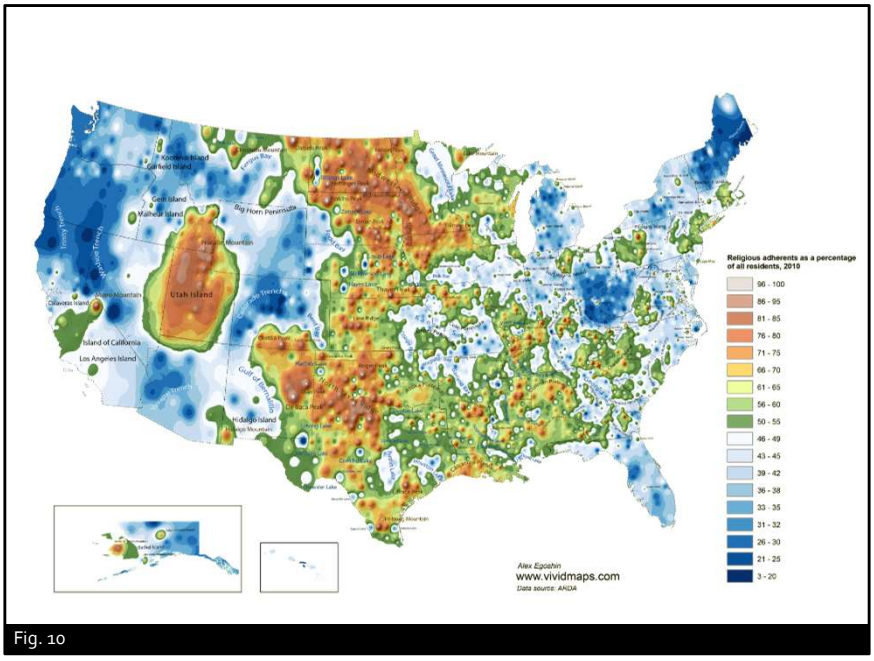


Fig. 10

Percentages are an interval scale (i.e. you can subtract them, but not multiply or divide).
But the scale used here has a clear 0 level (at 49% religious), which is only suitable for a ratio scale.
The author is making a point, saying that the sinners underwater might as well be drowned.

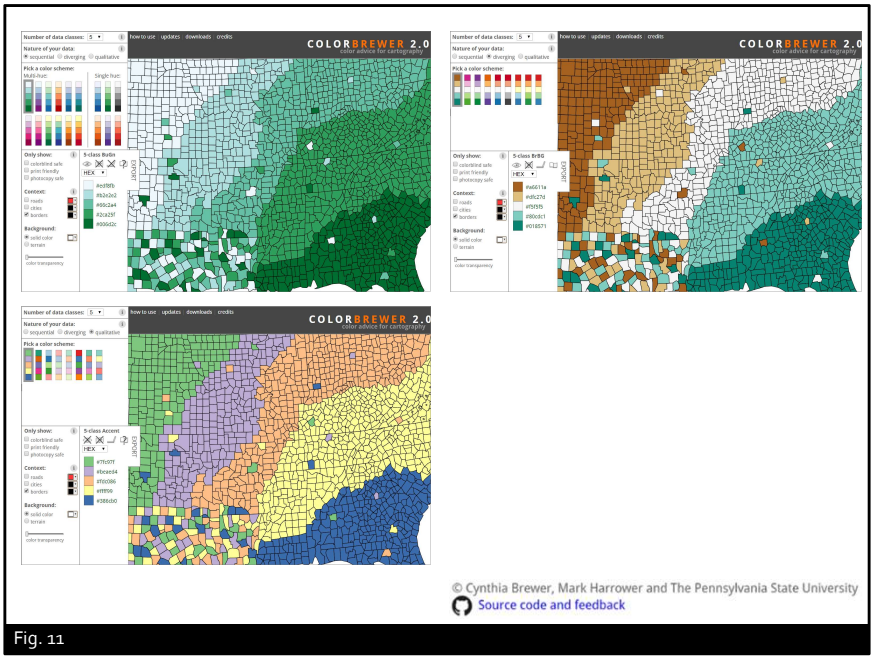
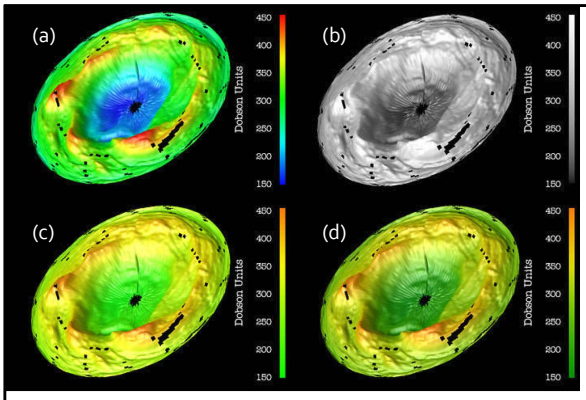


Fig. 11

Here are the types of colour scale you should be using:

- for ordinal or interval data, use a sequential colour scale e.g. shades of green
- for ratio data, use a divergent colour scale e.g. neutral hue for 0, one hue for +ve another for –ve, intensity for the absolute value
- for nominal data, use a qualitative colour scale e.g. a bunch of different colours each of equal visual weight

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



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)

Fig. 12

a. A rainbow palette is almost always inappropriate
b. The human eye is very good at detecting high frequency differences (differences over a small spacescale) from brightness, but it's hard to tell the difference between the crater rim and the rest of the boundary
c. The human eye is good at detecting low frequency differences (differences over a large spacescale) from hue, but the striations on the side of the "crater" are less visible than in (b)
d. By combining hue and brightness cleverly, the human eye can perceive much more

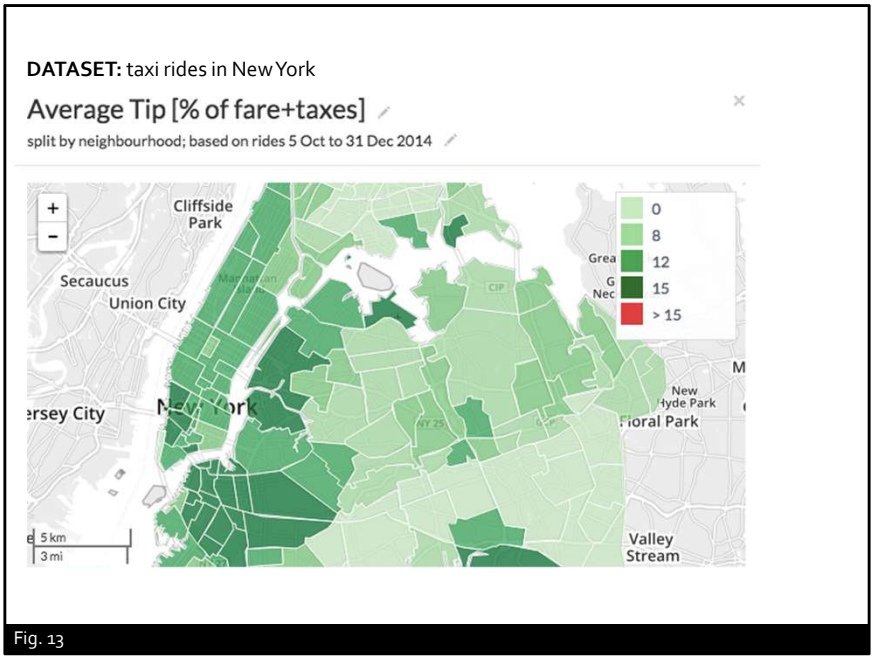
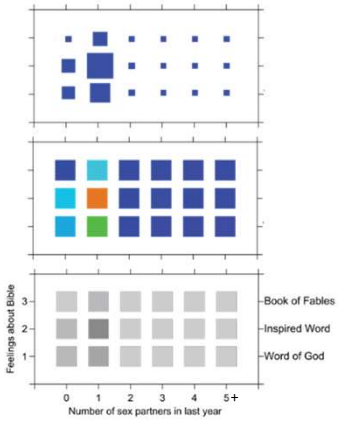


Fig. 13

It's easy to compare the brightness of nearby regions, much harder to compare the brightness of regions far apart. To make the overall comparison easier, I've split the data attribute (average tip) into quintile bands, and limited myself to 5 shades of brightness. This is a bit like drawing contour lines on a map.



DATASET: The General Social Survey asked

- How many sex partners have you had in the last 12 months?
- Which of these statements comes closest to describing your feelings about the Bible?

Respondents were 1606 US residents. (Davis et al. 1993)

Fig. 14

A data attribute can be mapped to many different aesthetic attribute. Here, the data attribute is "count" for each combination of (# sex partners) and (view of bible).

- A size scale (top) is natural for ratio data like "count". It's fairly easy to read off "This count is 3 times bigger than that".
- A brightness scale (bottom) is more appropriate for interval data. We can read off "This count is between those other two counts", but hard to read off "This count is 3 times bigger than that".
- A hue scale (middle) doesn't make sense here. Is green better or worse than blue? How about red?

DATASET: A survey of U.S. scholars (Morton and Price, 1989). Surveyed were 5,385. Respondents numbered 3,835. Respondents answered the question "How often, if at all, do you think the peer review refereeing system for scholarly journals in your field is biased in favour of males?"

	rarely	infrequently	occasionally	frequently	not sure
male respondents	851	426	284	199	1078
female respondents	80	110	170	319	319

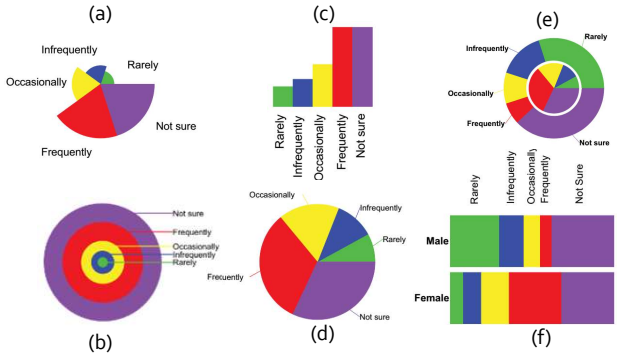
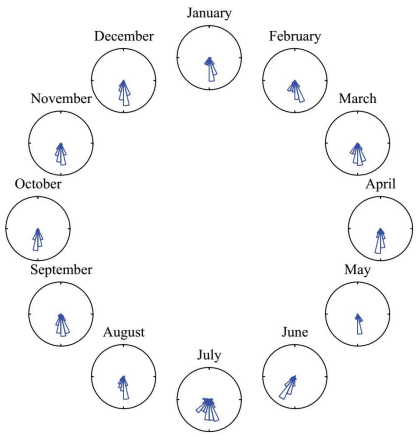


Fig. 15

Once you decide on the scale mapping, the rest of the visualisation just sorts itself out.

a. count→ r, response→ θ. Since response is nominal, each is given its own θ coordinate, equally spaced.
b. count→ r, equal θ, and stack different responses (so that r accumulates).
c. count→ y, response→ x.
d. count→ θ, equal r, and stack different responses (so that θ accumulates).
e. count→ θ, gender→ r, and stack different responses (so that θ accumulates).
f. count→ x, gender→ y, and stack different responses (so that x accumulates).



Predominant wind direction

Fig. 16

The scale mappings are

- wind direction → θ
- count (days with that wind direction) → r
- month → θ for the subplot

In other words, plots with subplots are just another example of a scale mapping.

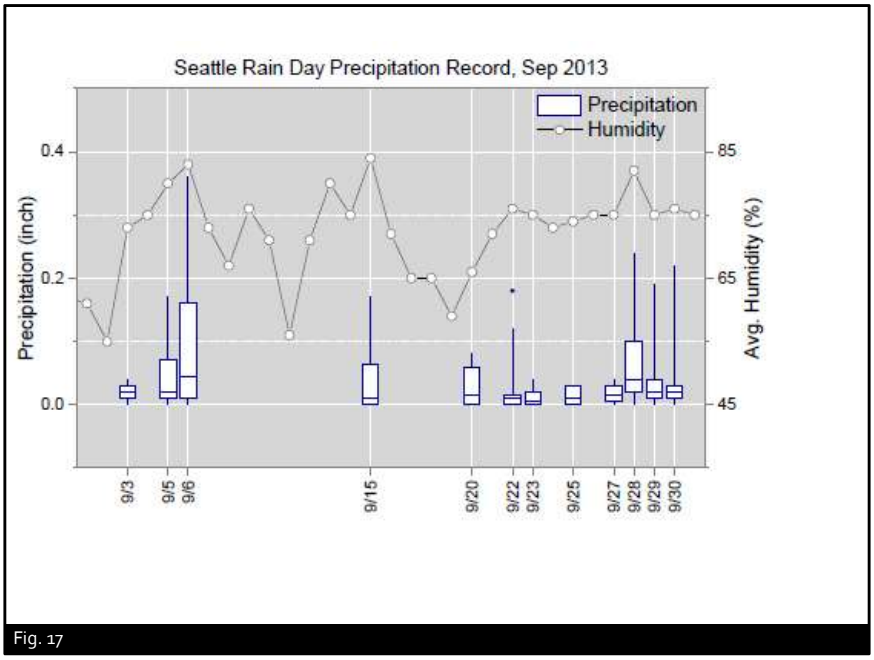


Fig. 17

What's the scale mapping here? There are in fact two different data attributes both mapped to the same aesthetic attribute (y axis). No good data visualization toolkit lets you do this, because it breaks the fundamental idea that a scale mapping is one-to-one, i.e. it doesn't permit you to read off data values from aesthetic values. Instead, you should show two separate subplot, one beneath the other, and make them share their x axis range.

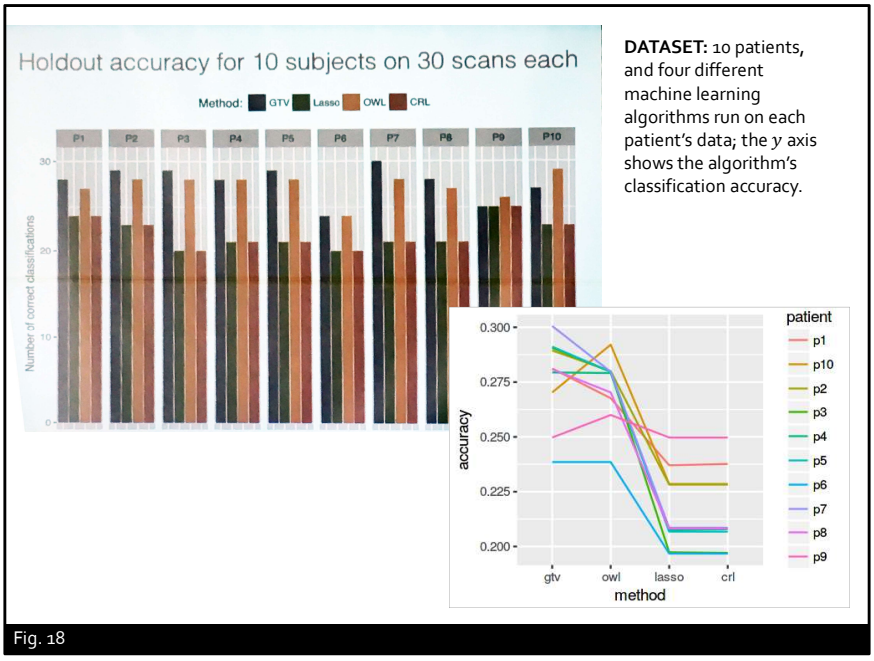
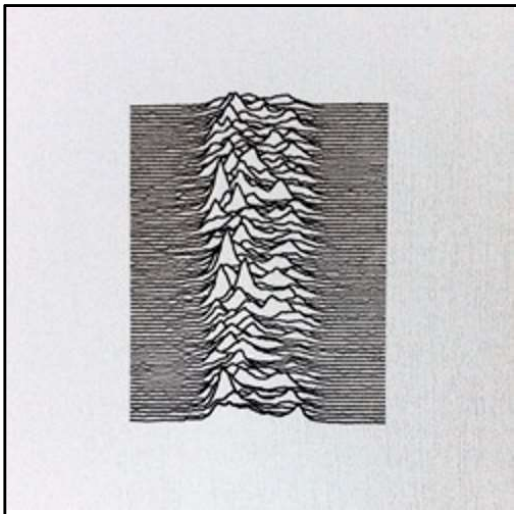


Fig. 18

The original plot (top left) makes it hard work to compare methods. It's easy to compare methods within a patient; but to get an overall sense our eyes have to keep flicking back and forth. The original plot emphasizes the patient, which is a useless thing to emphasize, since we (presumably) want to decide which method is best to use on new patients.

The redrawn plot (bottom right) chooses aesthetic scales to highlight the comparisons we want to convey. It's very easy to compare y values, so (as in the original) we'll map accuracy to the y axis. We want to compare accuracy across methods, so we'll set the x axis to be method. This gives 10 points at each method, and we might as well group them (as in the original), but all we have left is less easy-to-read scales such as colour. Patient is a nominal scale, so the colour scale is qualitative. There's no inherent ordering to the methods, so we might as well order them to convey some extra meaning. I've ordered them by overall accuracy. This shows very clearly that lasso and crl have identical performance, and also that one of the patients is an outlier (p9) compared to the others.

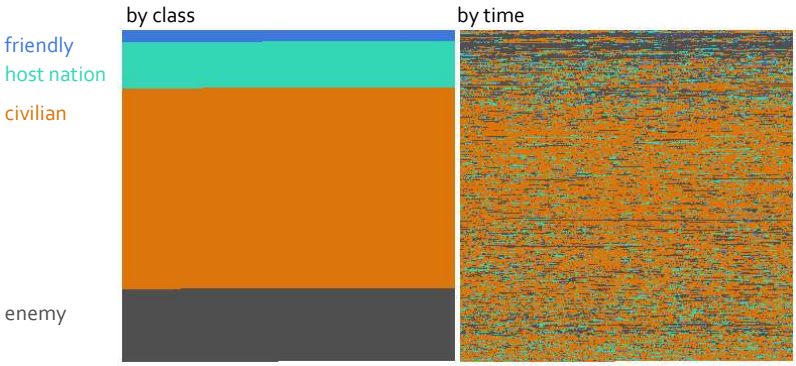
The plot we've produced is also known as a "parallel coordinates plot".



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/

Fig. 19

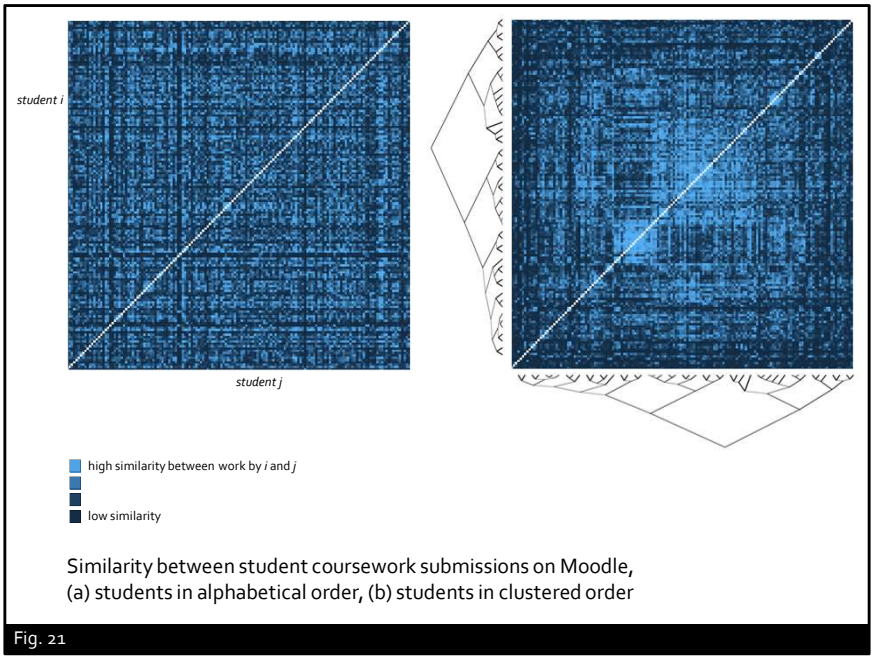
This plot shows signal strength from a pulsar. Each line spans a period in time, and the periods are arranged in order of time, with occlusion. This is a very effective way to show "precise period, fairly arbitrary difference from one period to the next".



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/

Fig. 20

This chart counts as Art. Each pixel is a death, and the artist is making a point about the inherent individuality and meaningless of these deaths. They can be positioned arbitrarily, to make an artistic point.



Similarity between student coursework submissions on Moodle,
(a) students in alphabetical order, (b) students in clustered order

Fig. 21

"Student id" is a nominal value. Here, I clustered students based on similarity (see IA Algorithms). When we replot the heat map, with students arranged according to the clustering, it's easier to see patterns i.e. groups of students who probably work together.

Moral: Whenever you map a nominal value to an ordered aesthetic, you might as well reorder it to make a point.

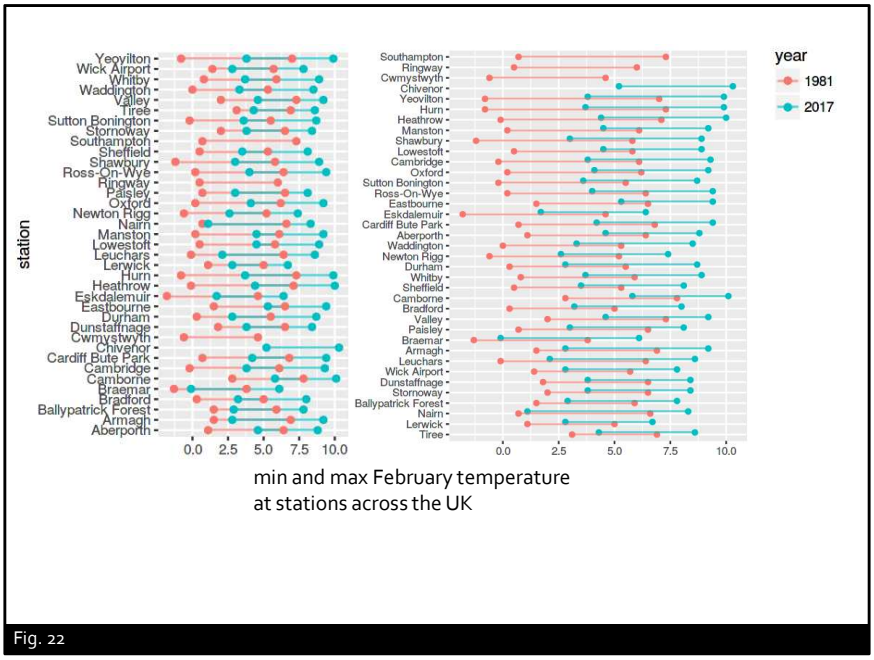
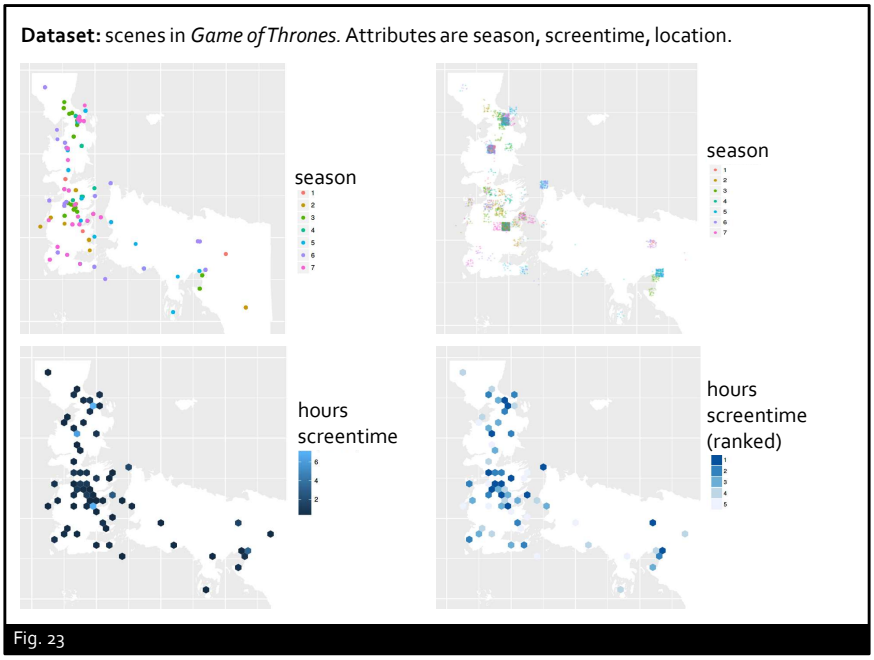


Fig. 22

Another example of ordering a nominal value.

- If I wanted the viewer to use the y axis for alphabetic lookup, I'd use the left plot (and also left-align the labels)
- But there are so few stations that alphabetic lookup is easy even without alphabetic sorting. Instead, I might as well use the order to convey some information. In this case, I chose the order to show ranking of difference between 1981 and 2017.



A common problem is how to deal with overplotting.

- In the top left plot, there are so many points at each location that they obscure each other and we can't compare locations.
- In the top right plot, I used jittering, and mapped screentime to opacity. The total ink tells us how much screentime there was; the density of points tells us how many scenese; and the colouring tells us the breakdown by season. This plot is very rich but rather confusing.
- In the bottom left plot, I cut out most of these dimensions, and binned the data instead. This is less rich, and less overwhelming.
- In the bottom right plot, I changed the colour scale: instead of mapping screentime to brightness, I ranked the hexes by total screentime and cut it into 5 bins (as for the New York taxi data). This shows a bit more detail, but it loses information about magnitude.

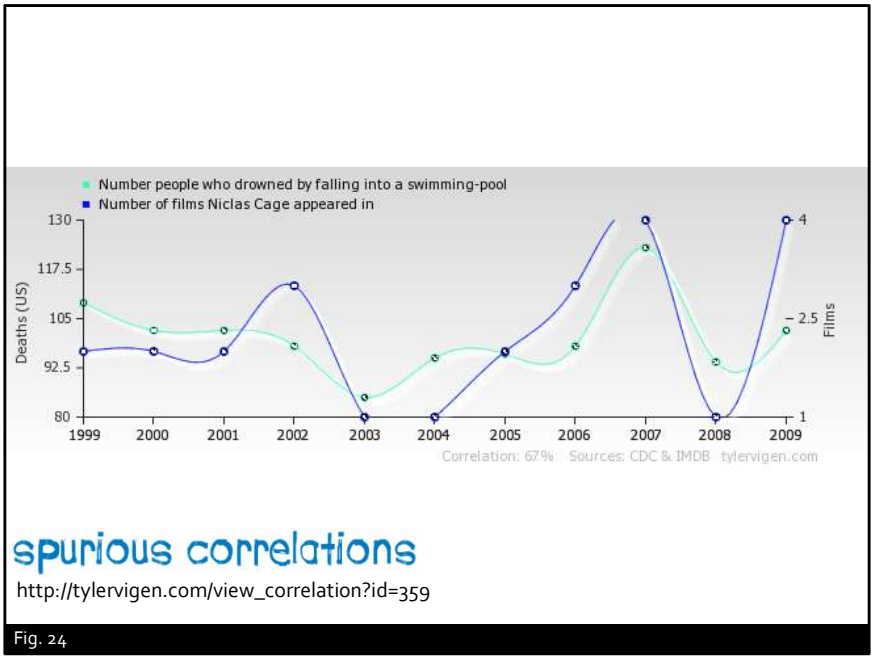


Fig. 24

Very often in data visualization, our goal is to discover something new and meaningful in the data.

The correlation in this plot is not meaningful! Partly, it's our job always to be on the lookout for meaningless plots. Partly, we shouldn't abuse the y axis by mapping it to two different data attributes, since this tends to worsen the illusion of meaning.

Singapore's MRT Circle Line was hit by a spate of mysterious disruptions (emergency breaking due to signal loss) in recent months, causing much confusion and distress to thousands of commuters. But the incidents—which first happened in August—seemed to occur at random, making it difficult for the investigation team to pinpoint the exact cause.
(*How the Circle Line rogue train was caught with data*. Lee Shangqian, Daniel Sim, Clarence Ng, data.gov.sg, 2016)
https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

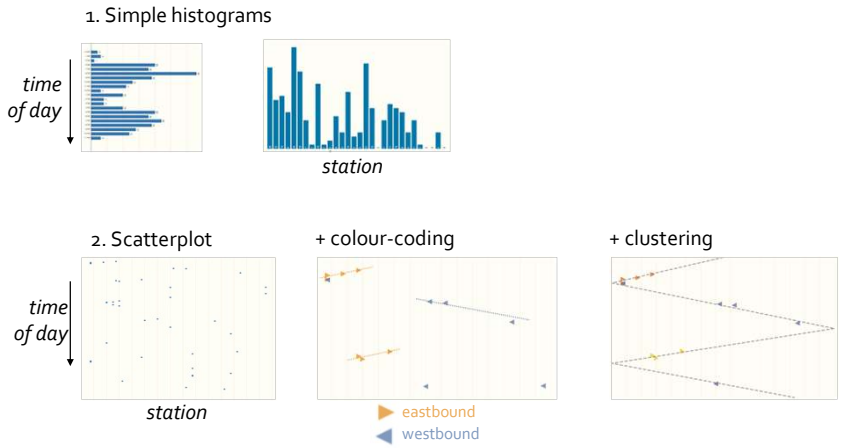
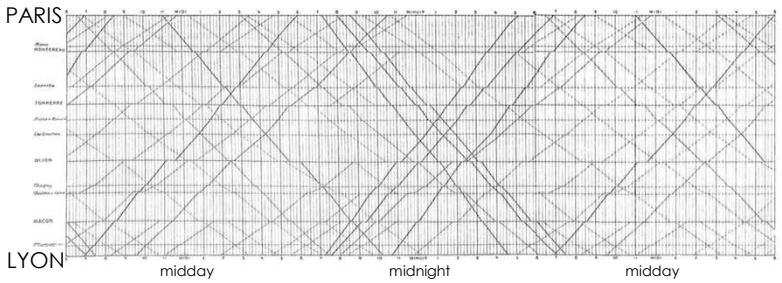


Fig. 25

Here are some plots summarizing a much more serious and interesting hunt for meaning. Read the blog post for full details.



Railway timetable plot
E. Marey, 1880

Fig. 26

The space-time plot in Figure 25 was inspired by this railway timetable plot. The scale mapping is time→ x, location along the line→ y, group points by train id. The diagonal lines are trains in each direction, and you can clearly see express versus slow trains, and the long stops at Dijon.

The x axis rolls over, showing 36 hours. This is helpful redundancy, because we get to see the full journey of every train (assuming the timetables are the same each day).