3 Hall-of-Fame / Hall-of-Shame

3.1 Misleading Charts 49
3.2 Home Runs and Strike Outs 52

Unfortunately, the visualization landscape is littered with **misleading charts** and **shameful data displays** that can cause confusion and create dismay in presentation consumers and decision-makers alike. In this chapter, we discuss some of the **most common errors** found in visualizations and provide examples of **effective charts** and graphs that **miss the mark**.[1]

1: The appreciation of data visualizations is definitely a **subjective** activity. To illustrate the potential of consensus, each of the examples will be accompanied by a discussion (trialogue?) among the authors as to the qualities of the chart in question.

3.1 Misleading Charts

Misleading charts are not always maliciously so – sometimes, the errors come by honestly. The chart shown in Figure 3.1, however, is not one of these.

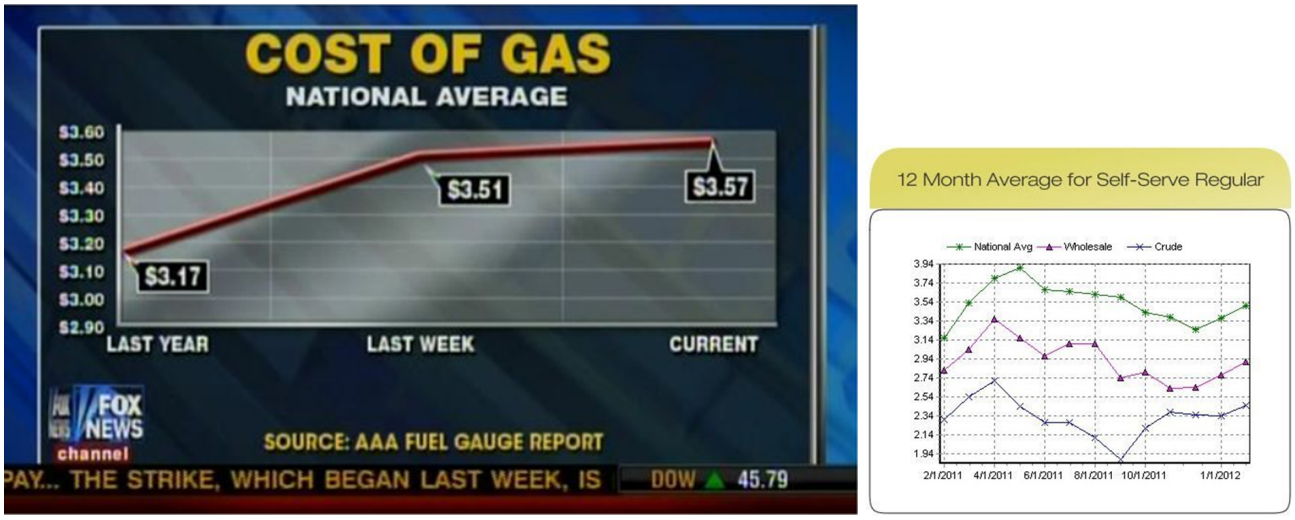


Figure 3.1: A misleading chart from Fox News (Feb 20, 2012): the national average cost of gas in the United States, from Feb 2011 to Feb 2012 (left), and the corresponding monthly series for the same time period, from AAA Gas Prices (right) [mediamatters.org].

What impression is a viewer expected to retain when seeing it flashing briefly on their TV screen? Are they likely to feel that the cost of gas has stayed more or less constant over the last year,[2] or perhaps that it has doubled? What are the intentions of the chart makers? What is the **purpose of the chart**? Upon consulting the "true" chart on the right, does it still seem as though the cost of gas has gone up significantly over the year?

2: A scant 12% increase, in actuality.

How does the Fox News chart fail? Let us count the ways:

- the vertical scale is broken, giving the mistaken impression that the cost in Feb 2011 ($3.17) is about half as large again as the cost in Feb 2012 ($3.57);
- only three time points have been selected [last year (Feb 2011), last week (Feb 13, 2012), and the current date (Feb 20, 2012)], whose cost of gas values erroneously suggest that the national average has been increasing over time;
- the "wonky" horizontal scale misleads the viewer into thinking that the increase has been steady over time;
- the use of only one displayed statistic (the national average) also hints at the increase to have been the same across the US.

To be fair, the chart axes are clearly labeled, as are the gas prices, and any responsibility for misreading the chart ultimately rests with the consumer.

But keep in mind that such a chart would typically only remain on the screen for a short period of time, and even data visualization specialists, which we suspect form at best a small majority of most if not all TV audiences, would barely have the time to parse the chart, let alone produce an interpretation that differs from the intended one.[3]

3: Namely, that gas prices have steadily increased over the year, and are now twice as large as they were last year at the same time.

However, it only takes a glance at the 12-month average cost of self-service regular gas chart (the decidedly less flashy chart provided by the AAA, in Figure 3.1) to destroy this incorrect picture: the price of gas has gone up, then down, then up again over the year, and that pattern is similar for the national minimum, average, and maximum cost of gas, but it showed no global increasing trend.

To an arbitrary observer (and with roughly 10 years of hindsight as of this writing), the monthly variation in 2011 US gas prices hardly seems worthy of mention. In fact, it requires a fair amount of tweaking to transform the inconspicuous chart into something that could become fodder for a national television "newscast". The mind boggles... until one remembers that Barrack Obama was running for re-election in 2012 and that the doubling of gas prices in the last year of his first term could be used to attack the Democratic Party's fiscal and/or foreign policies during their term in power.

Is this really why Fox News ran the chart? In truth, we do not know. Perhaps it was a comedy of errors: a new intern, unfamiliar with the data visualization used by the company, quickly puts together what they feel is a reasonably accurate chart to accompany a story; the TV host, not expecting this mess to pop-up on the screen heroically putting on a brave face and attempting to rescue the intern's tragic attempt because "the show must go on", the good-natured retraction being unfortunately pre-empted by the commercial break, and so on ...

Somehow, that explanation seems even more far-fetched than the doctoring to which the graph was originally subjected: it is much more likely that this was a deliberate attempt to stretch the truth and manipulate viewer emotions while keeping a veneer of media respectability – something is rotten in the state of Denmark!

The United Kingdom's Norwich North electoral district Liberal Democrats[4] released the brochure of Figure 3.2 in advance of the 2010 UK general election.

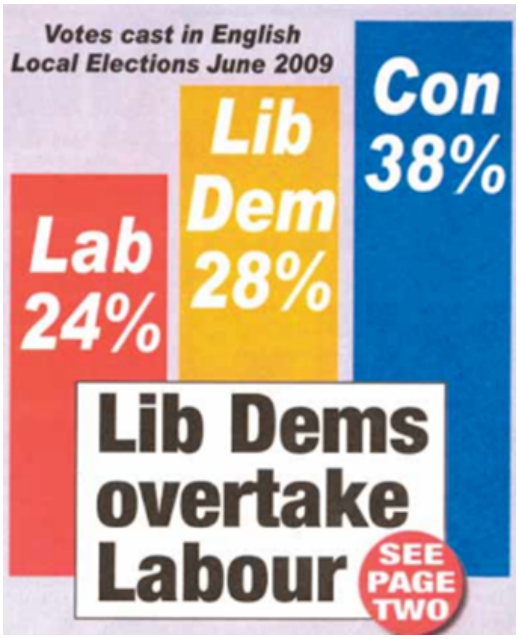


Figure 3.2: A misleading chart displaying the vote distribution in the 2009 English local elections for the combined 27 county councils and 8 unitary authorities contested (simplystatistics.org).

The numbers are valid, as can be ascertained by the 2009 England local election results ; the fundamental issue is **how** these numbers are rendered in the accompanying bar chart.

Although we know that 24% and 28% are much closer together than 28% and 38% are, and yet, the bars do not agree with this – the scale in use certainly seems to suggest that the Liberal Democrats have overtaken Labour, as the boxed text states, and that they are *that* close to pulling even with the Conservatives, so perhaps disaffected Labour voters should consider throwing away their principles and vote for the Liberal Democrats as their best chance to topple the Conservatives. Right?[5]

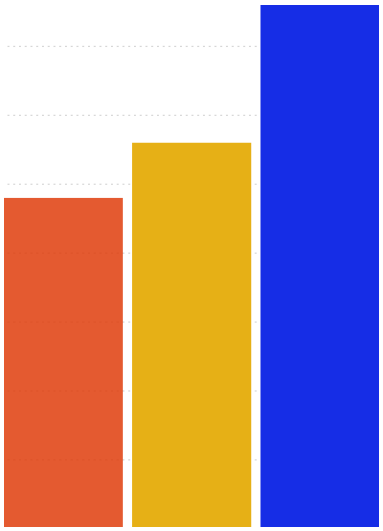
The strategy did not pay off: in 2005, the general election results in Norwich North favoured Labour with 47% of the vote against 31% for the Conservatives and 17% for the Liberal Democrats; in 2010, the Tories came first with 41% of the vote, while Labour and the Lib Dems garnered 31% and 18% of ballots, respectively (UK Parliament).[6]

How do you solve problems like **disingenuous**, **selective** and/or **incompetent reporting** in data visualizations?[7] Analysts should produce charts with:

- consistent scales and units of comparison;
- full time series;
- axes that are not cut-off, and
- numbers that add up.

4: A political party whose policies and stances are more or less aligned with those of the Liberal Party of Canada, but whose influence in the UK political landscape is roughly similar to that of the New Democratic Party in Canada. Roughly.

5: Using a regular linear scale, the bars would instead look as below:



Not quite the same effect, innit?

6: Is this in any way due to the brochure?

7: For low-dimensional datasets, say, a **tabular display** may provide as much information and be less likely to mislead.

What to Watch For As we have seen (and will continue to see), data visualizations can be visually striking, yet **misleading**, charts. Among other things, consumers of such charts should be on the lookout for:

- any sign of tampering with axes and linear scales;
- scaling effects, when representing data points as shapes or volumes;
- cherry-picking and omitting certain data points;
- attempts to distract with graphic design, etc.

Malice is not always (and to be fair, perhaps only rarely) part of the equation, but even the best-intentioned analysts can produce misleading charts if they are not careful.

Adhering to the following guidelines and principles can help mitigate the risks:

- effective data visualizations should **provide insights** and **facilitate understanding**;
- the **basic principles of analytical design** should guide both visualization **design** and **consumption**;
- **creativity** is encouraged, but data and representations must be kept honest;
- attempts to distort **trends** and **conclusions** with flashy visuals should be recognized in displays found "in the wild", and
- whenever possible, **data** and **code** should be made available along with the displays.

3.2 Home Runs and Strike Outs

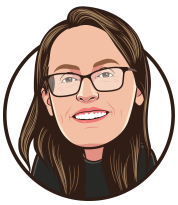
Some of the visualizations presented in this section are perfectly adequate (the **home runs**); some of them fail in any of a number of ways (**strike outs**). Were it possible to classify each of them cleanly in one of the two categories, we would do so as we present them.

Our experience has shown, however, that appreciation of data displays is a **subjective** endeavour;[8] showing a chart to three analysts is likely to yields at least two differing opinions as to its quality!

8: Well, *mostly* subjective.

Case in point, let's see what Jen, Patrick, and Stephen each have to say about the following charts.[9]

9: Feel free to add your opinions as well – the more, the merrier!



Jen: **cognitive scientist** likes bad movies, bad music, and bad charts

Patrick: **mathematician** grumpy; can spot a one pixel gap at ten metres

Stephen: **physicist** will build dashboards out of twine

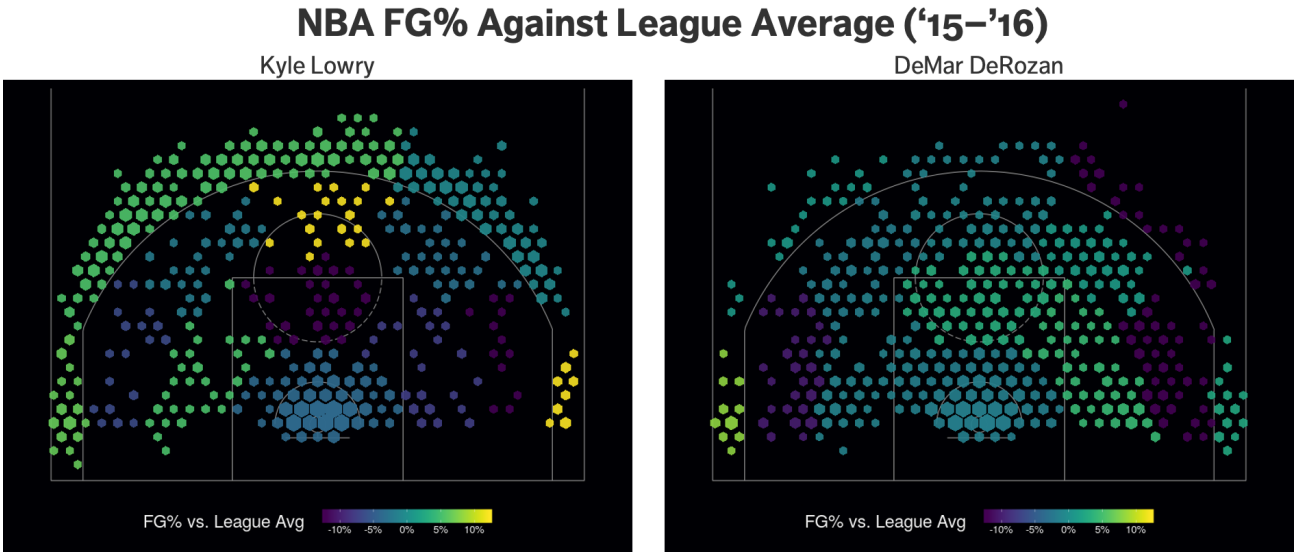


Figure 3.3: NBA FG% against league average ('15-'16) for Raptors Kyle Lowry and DeMar DeRozan [T.W. Schneider]

I have mixed feelings about this chart. At first blush, it conveys a sense of the style of play of a given player, and allows for a comparison between players. I also appreciate the slightly outside-of-the-box colour palette. However, explaining the meaning of FG% against league average and how it relates to what people are seeing in the visualization isn't straightforward, particularly for people who aren't familiar with the game. I view this as a good visualization for people who are already basketball aficionados.

What I appreciate about this chart is that in spite of my relative lack of knowledge of basketball, it is fairly easy to recognize that Lowry and DeRozan likely played different positions (and roles) for the Raptors in 2015-2016, and that we get a sense for where they stand in the league hierarchy... assuming that we know what "FG%" stands for.* A good visualization may require the consumer to conduct additional research in order to bring the insights and context in focus – it can be a two-way street, at times.

Ignoring the fact that I don't know what an FG% is, I find these very intuitive to understand, even with the fairly abstract colour scheme. I follow rugby and similar charts (replace FG% for your favourite variable: kicks, incidents, tackles, whatever) overlaid onto the playing surface is a very efficient way to tell a story. The main issue I have is the comparison aspect (Lowry vs DeRozan): you can tell that there is a difference between the two but I don't know why that is the case (maybe different positions, different coaching, etc. – I know nothing about the NBA). Some additional metadata (position description, dates, coaching staff etc) would really help. I can't help thinking whether it would be possible to combine the two, maybe subtract one players FG% from the other and put that on a chart so it just highights the difference?

* Field goal percentage ; the percentage of successful non-free throw attempts during a game.

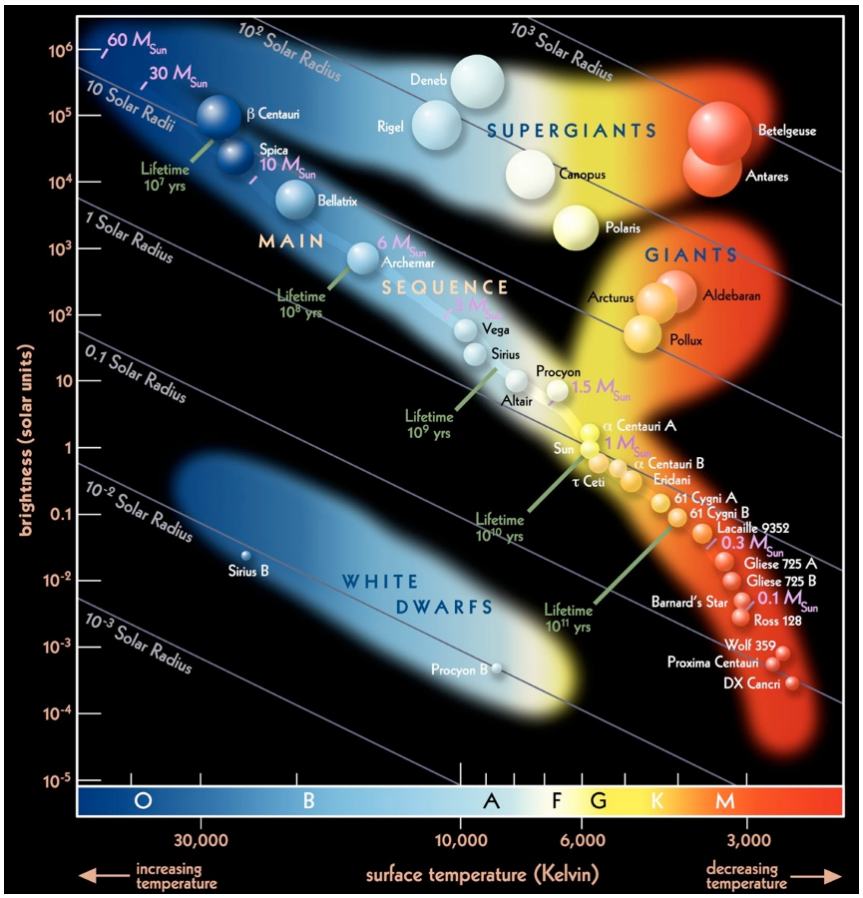


Figure 3.4: The celebrated Hertzprung-Russell diagram of astrophysics [European Southern Observatory].

I appreciate this visualization conceptually, and I think it's an exemplary use of a bubble diagram, particularly in the way it integrates the different variables involved. That said, I'm not a huge fan of the aesthetics of this particular version of the diagram. I like the way the various elements and labelling are incorporated, and realize the colours are intended to be relatively realistic, but there's something about the style of the visualization that I find a bit jarring.

An all-time favourite. The underlying structure that emerges would be impossible to predict just by looking in a telescope: there seems to be 4 main clusters/groups of stars (note that not all stars appear in the diagram), and there is a clear association between lifetime, mass, radius, brightness, and surface temperature for stars on the Main Sequence. Look at it! LOOK AT IT!

As somebody with a physics background I am (of course) positively biased with respect to this visualization. Minard manages to layer in many dimensions into an easy to understand diagram in *The March to Moscow*; the same happens here – 5 variables (radius, temperature, brightness, mass, lifetime), 1 classification type (White Dwarfs, Giants, SuperGiants and Main Sequence) and finally star name! Having some background in the subject makes a difference in how quickly you understand the chart, but when I use this an example in a class it never takes long to explain.

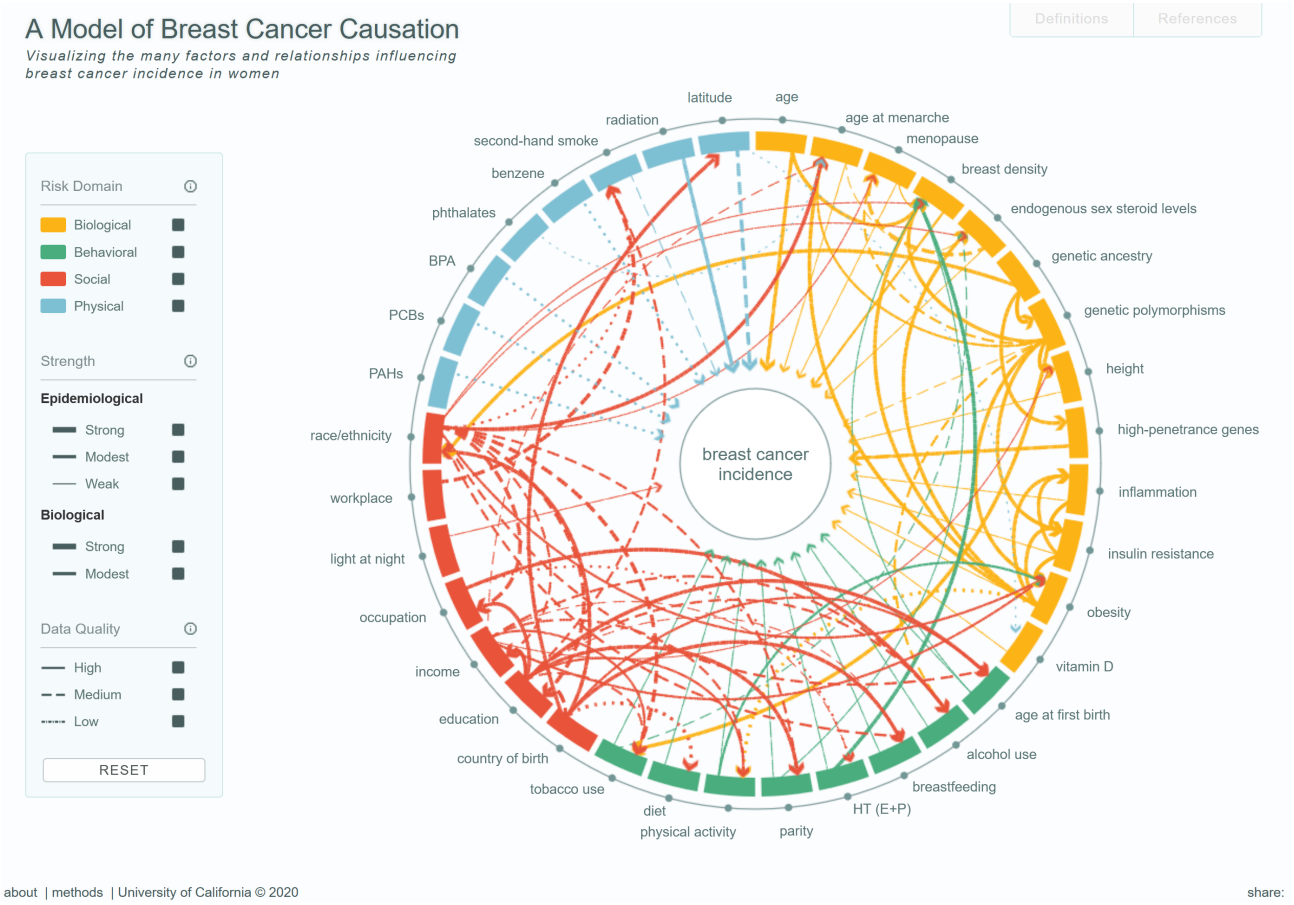


Figure 3.5: A model of breast cancer causation, University of California .

This is an interesting interactive chart; it contains only 4 colours and its documentation is easily accessible (methods, definitions, references), but can we really infer causality from the diagram? One aspect I quite appreciate is the explicit encoding of data quality in the chart: not all the links are created equal – *caveat emptor*. Another interesting item is that not all displayed associations are linked to breast cancer incidence, leaving some room for exploration.

This one doesn't work in print and has to be interacted with online. Even then I have a hard time making it work for me. Showing links between categories is interesting but it doesn't tell me much about the subject, at least not without some additional work (what IS a "Strong Epidemiological?"). Circular chord charts help with regard to real estate, and the use of circles forces whitespace onto the page which is a good thing: "pretty but not that helpful".

In my opinion this is a rare instance where the animated version of the visualization works much more effectively than the static version of the visualization, which I find very busy. That said, I think displaying the causal network in circular fashion works very well – the network is effectively contained in the space. The colours indicating the different categories of risk do help to make the visualization more navigable.

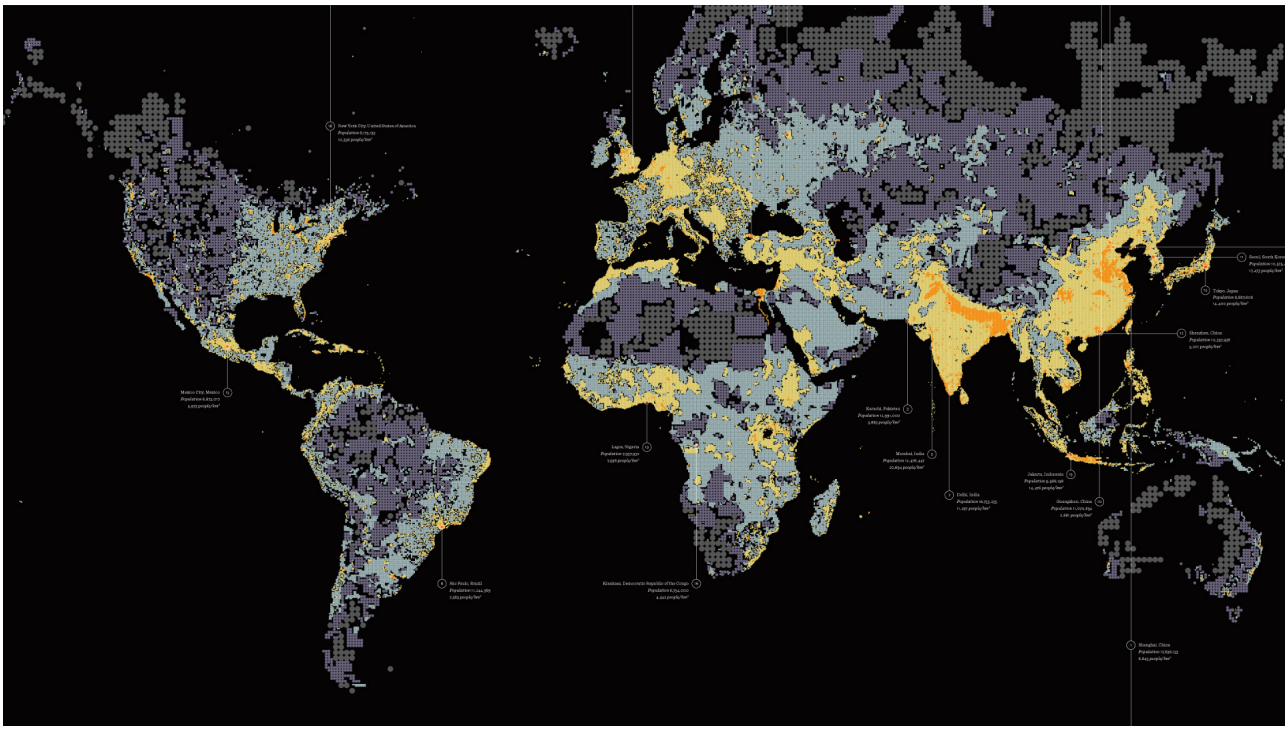


Figure 3.6: World's population "dencity" map [Fathom].

I'm a fan of this visualization. The graphical elements draw me in, and I like the choice to stay away from a more semantically loaded colour palette. Initially, the choice to have low resolution – larger dots – indicate low population seemed like a strange one, but with a continued review, I like that this choice adds an interesting emphasis to the parts that have low population density, and highlights the ways in which these parts are connected to other higher density parts. This visualization is one where you are rewarded by both a cursory and more in-depth examination of the details.

I like charts that help me re-cast what I know about my world in a different light. As a resident of a small Canadian village, I'm aware that I live in a sparsely populated area of the world – I think of Ottawa as a Big City. I've visited bigger cities, so I know that the 2 people per square kilometre that is my everyday life is not the norm... but I do not "know" it. Fathom's "dencity" chart really drives home that point. From a technical perspective, I find the counter-intuitive use of smaller (and more numerous) dots to indicate higher population density (and *vice-versa*) well done. Furthermore, while I can easily spot the well-known outlines of continents, the distortions created by colouring both "uninhabited" land and oceans in black draws my attention to the fact that this is *not* the usual fare.

I like "heat type" maps, I find them very intuitive but the colour scheme in this one really puts me off. In terms of data classification I get that the oceans have no population and are therefore black, but in this case removing them from the classification altogether and making them white (assuming white has no meaning in the density scale) would significantly reduce the cognitive load and make everything else clearer and easier to understand. The size of the dots is an interesting one, I get it that it is used to add emphasis on lower density areas but my brain just thinks of low resolution / text interface games (ok so I play too many legacy games!) which for me is distracting but from a technical perspective I still think it is an innovative, interesting decision that works for most folks.

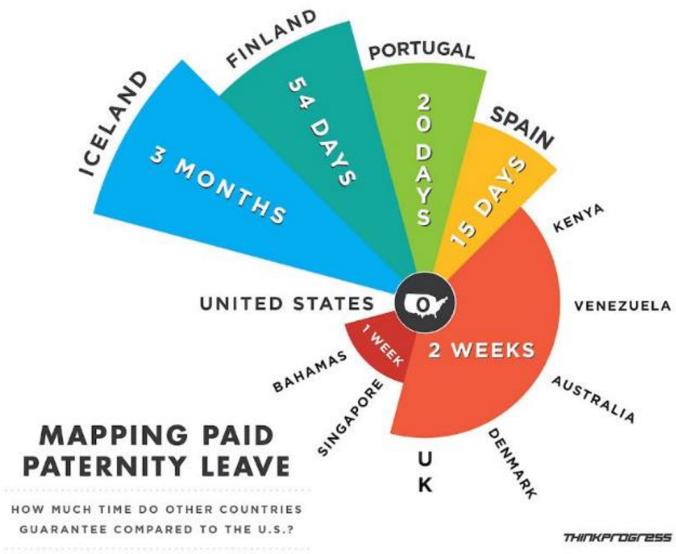


Figure 3.7: Mapping paid paternity leave in countries around the world, in relation to the United States, c. 2010 [ThinkProgress].

If I don't think too hard about this visualization, it kind of works. There are countries, there are amounts of time, and the amount of time decreases as you spiral inwards and arrive at the punchline: United States – 0 Paternity leave. Considered in this fashion, I would say it works more as an infographic than a data visualization. From a data visualization point of view, there are some issues. Does the volume or arc length mean anything consistent? Why have these countries been chosen, and not others? These are just a few of the questions that spring to mind with respect to the visualization choices made here.

Yuk, this one hurts my brain. I get what it is trying to show, and in fact I get the message quite quickly so from that perspective it works but, well, yuk. Why those countries? Why those colours? Why does it spiral? Why the change in unit resolution? Why ALL CAPS? Why the change in font size? Finally why no Canada (up to 8 months / 35 weeks of paid paternity leave depending on circumstances)? There are SO many other, better ways of displaying this data. Yuk.

I'm a father of three children; I was a graduate student when Elowyn, my eldest child, was born and so did not get paid parental leave; I took 4 weeks of parental leave (2 weeks paid by a combination of my employer and the Government of Canada) when Llewellyn, my middle child was born, and 4 weeks of parental leave (under the same set-up) and 4 weeks of (paid) accumulated personal leave when Gwynneth, my youngest child, was born.[†] In retrospect, I wish I had had access to more parental leave. So I get what this chart is saying, and I agree with it fully: it is ridiculous that no paid paternity leave is guaranteed in the United States, nominally a First World country (at least, whenever the display was prepared); the US does not compare advantageously to the handful of countries against which it is pitted in the chart.

But what is going on with the chart? The low data density and cherry picked countries (only 12 countries are displayed, all of which but 1 offering some paternity leave, when about half of the world's population is in the same boat as the US in 2018), the high "chart junk ratio" [2] of the radiating pie slices (too much ink for the data being displayed), the odd scaling effect (if Singapore's radius represents 1 week and Kenya's radius 2 weeks, then Iceland's radius definitely does not correspond to 3 months). Why not use a bar chart or a tabular display instead?

I'll say this for the chart, however: absolutely love the earthy tones of the colour palette.

[†] The 2023 Canadian parental leave rules are too complex to write up in this space (maternity leave also enters the equation), but they are much more generous than what is seen in the chart.

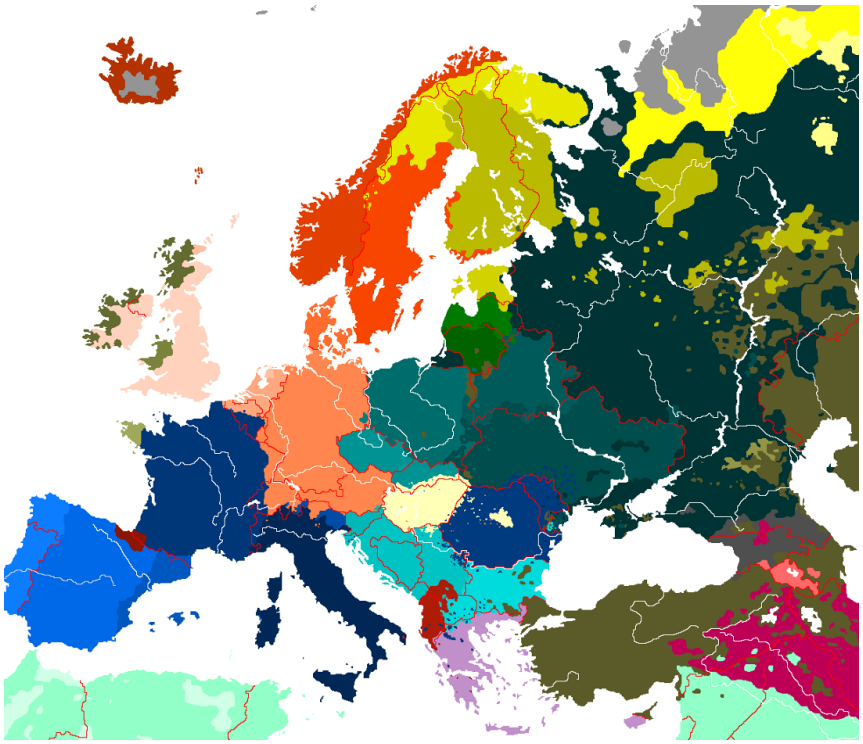


Figure 3.8: Language map of Europe [author unknown].

If I had the ability to interact with this map I think it would work for me much better. It's really designed for data exploration and the ability to select different language families, or look how languages migrate over time would be really interesting. As a static image with no legend it's essentially impossible to understand without significant input from other metadata or somebody explaining the concepts behind it.

There could be some interesting data behind this map, but I have a hard time getting past some of the design choices of the visualization itself. I really want to put on my sunglasses when I look at it. The lack of legend makes drawing conclusions difficult if not impossible.

If you know anything about languages in that part of the world, it is not too difficult to see that colour groups correspond to language families: blue for Romance languages (Portuguese, Spanish, Catalan, French, Italian, and Romanian) a light beige for Magyar (an isolated island within Easter Europe, but in a shade akin to that of Finnish and Estonian), red-orange hues for the Germanic languages (German, Dutch, Flemish, English, and the Scandinavian dialects), teal colours for the Slavic languages, earth green for the Celtic languages, etc. But a lot of that information has to be inferred: there is no encoding in this version of the chart (compare with Figure 2.12, say). Further investigation brings out a number of oddities: there are areas in Iceland and Russia where the map is not coloured, but the map is coloured everywhere else. It's not much of a leap to conclude that these are regions with no population, but surely there are similar regions in the Alps or the Pyrenees... at the very least, population density could be incorporated into the chart.

Another interesting aspect is that of secondary languages: Bretagne (Northwest of France) has Breton (one assumes) as its language, but at most 20% of its population speaks the language (see Welsh for a similar example). Is the colouring meant to represent historical languages instead? If so, why aren't the other historical languages of France represented as well (such as Occitan)?

At first glance, this is a good chart. But a second pass produces enough anomalies to make me doubt that this is a good *data* map. I would need more evidence (or a data source) to really buy into it.

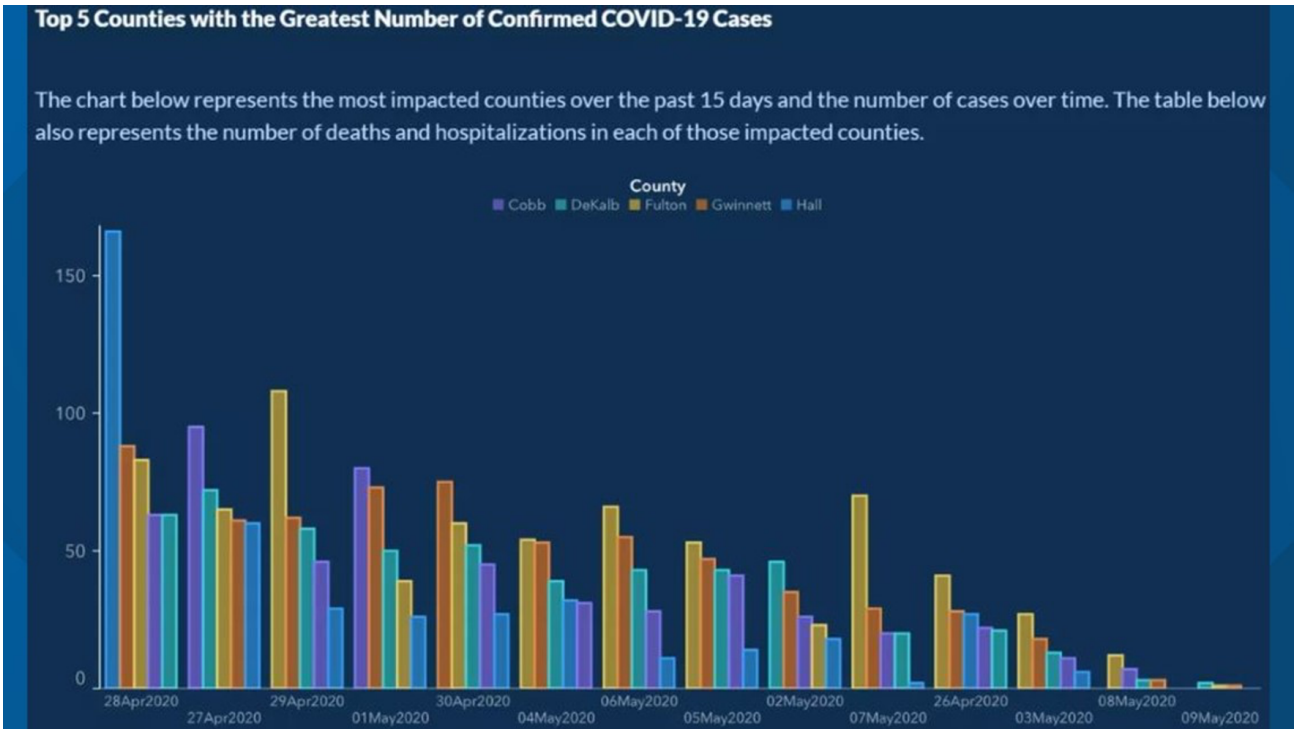


Figure 3.9: The 5 Georgia counties with the greatest number of confirmed COVID-19 cases; April-May 2020 [Georgia Department of Public Health].

Baaaaaaaad graph! Bad graph!! Assuming this was done deliberately, and isn't just an artefact of some poor default chart generating algorithms, this visualization banks on people's familiarity with bar charts and uses this against them, relying on people not to look too closely at the details and just pay attention to the overall pattern of the chart. The name of the game when it comes to creating the pattern, however, is shuffling: shuffle the dates, shuffle the ordering of the bars.

A professional disgrace.

Bad things: Choice of chart, value vs axis sort, colour scheme (framing elements and series), sub heading explanation, overall concept.

Good things: *tumbleweed* .

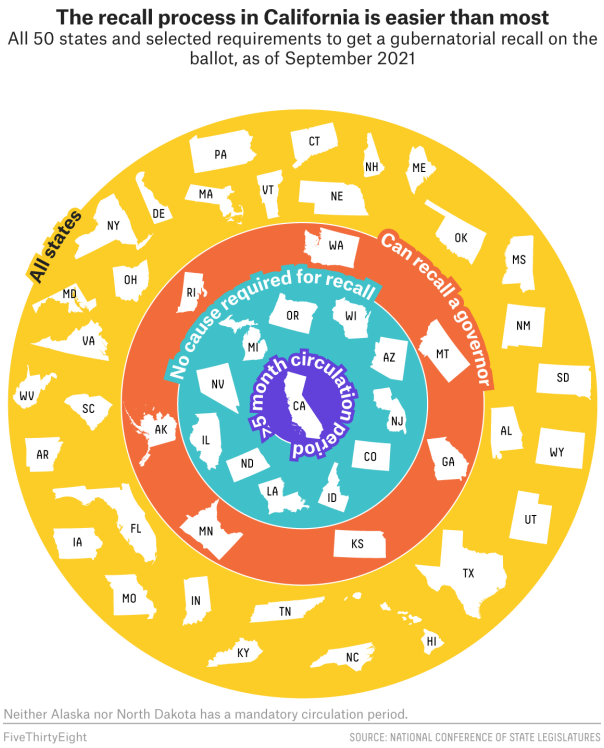


Figure 3.10: Comparison of the gubernatorial recall process in the 50 US states [FiveThirtyEight.com].

One glance at this chart and I can't get the "wheels on the bus" song out of my head. Similar to the paternity leave chart, this feels more like an infographic than a data visualization. The inner circles aren't too bad, because they have a relatively clear and meaningful label, but the outer ring label (All States) is confusing rather than helpful.

I get the point: California's recall process is unique.

But why the concentric circles? The 3 outer layers make sense: among all states, the majority cannot recall their Governor (outer shell, in yellow); among those states where recalls are possible, some can do it without cause (teal shell and ... California?). Is California (">5 month circulation period") a special case of "No cause required for recall" that does not apply to Nevada, Illinois, and the likes? There is a footnote in the chart saying that "neither Alaska nor North Dakota has a mandatory circulation period", which suggests that all the other states in the teal shell (and California) have a circulation period, whatever that might be. Perhaps California is the state with the smallest circulation period (at least 5 months), which is why it is "unique"... but then what about Alaska and North Dakota?

I think the information could be presented in a more reasonable way. But it could just be because I don't know enough about the context and I'm confused; presumably, this makes more sense to informed Americans.

It's a data visualization pretending to be an infographic, ... there are so many other (and better) ways of telling this story in a much simpler manner. The one thing that concentric circles do give is a focus towards the middle, so *that* works; but not much else in this display works.

The shapes of the states only add clutter, the circles force the text to be circular which reduces readability, the category labeling is therefore almost impossible to read... and doesn't make sense regardless (unless you know the different mechanisms involved).