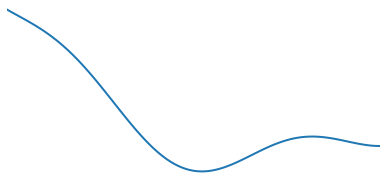
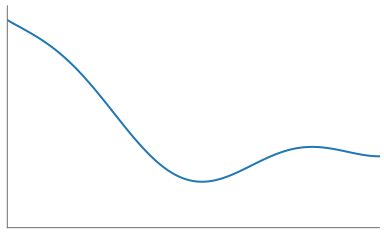
Using axis lines for good or evil

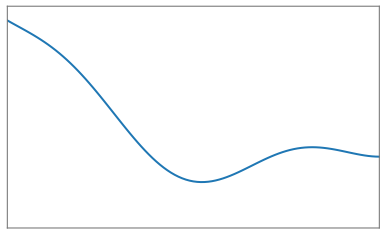
Say you want to plot some data. You could just draw it by itself:



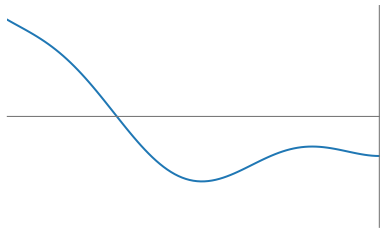
Or you could put lines on the left and bottom:



Or you could put lines everywhere:



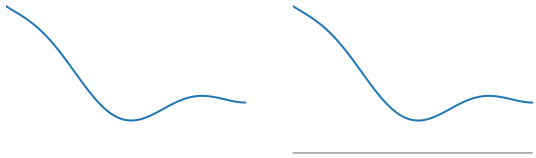
Or, you could be weird:



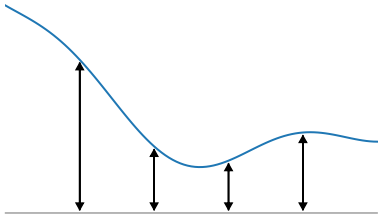
Which is right? Many people treat this as an aesthetic choice. But I'd like to suggest a clear and fairly unambiguous rule.

Principles

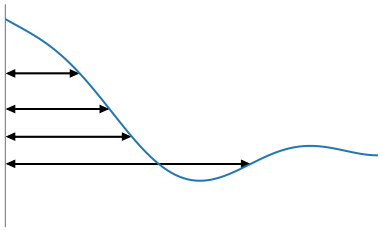
First, you need to accept that all axis lines are optional. I know it's *conventional*, but I promise you your readers will recognize a plot even without those lines. Now, should you draw a line for the x-axis? Compare these plots?



Which is better? That depends on what you're plotting. To answer, mentally picture these arrows:

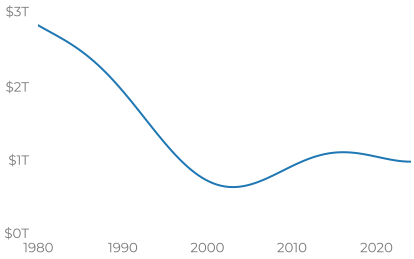


Now, ask yourself, *are the lengths of these arrows meaningful*? When you draw that horizontal line, you invite people to consider those lengths. Should you draw a line along the y-axis? Then it's the same principle. Ask yourself if these lines are meaningful:

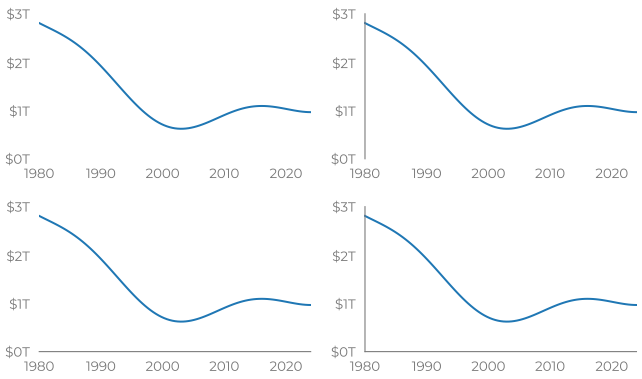


Example: Years v.s GDP

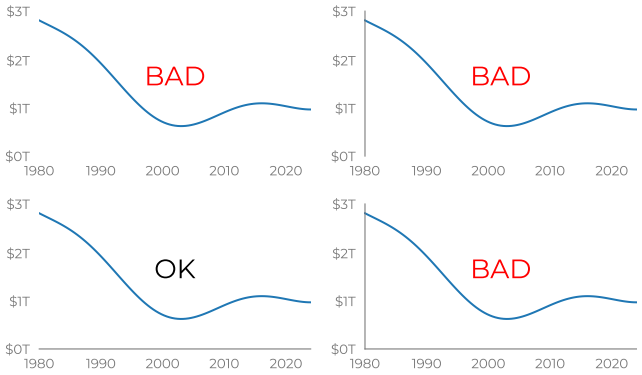
Suppose the x-axis is year, and the y-axis is the GDP of some country.



Which of these four plots is best?



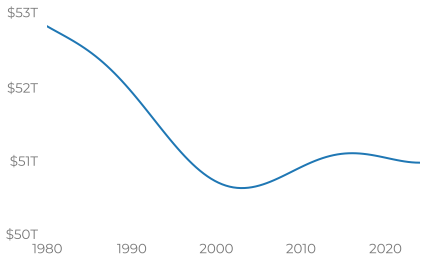
Here's the answer key?



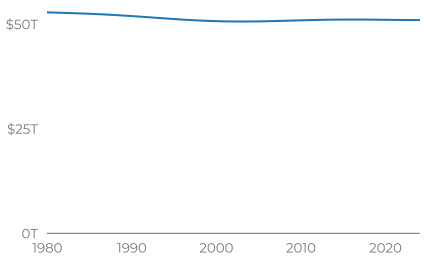
Why? * GDP is an absolute quantity. If GDP doubles, then that means something. Distances between the line and the x-axis mean something. * But 1980 is arbitrary. If you're comparing the year 2020 vs. the year 20, all that matters is that they're 20 years apart. The fact that 2020 is twice as far from 1980 as 2000 is not important, because time did not start in 1980.

Example: Years vs. GDP again

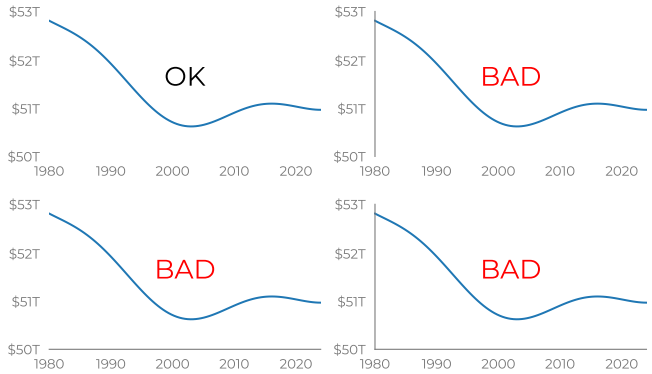
Say we have the same axes as before, but instead of varying between $1T and $3T, GDP varies between $50T and $53T.



What to do? In principle you could force the y-axis to stretch all the way down to zero.



But that doesn't seem like a good idea—you can barely see anything. If you need to start the y-axis at $50T, then fine. (As long as you're not using a bar chart.) But then you need to drop the line:



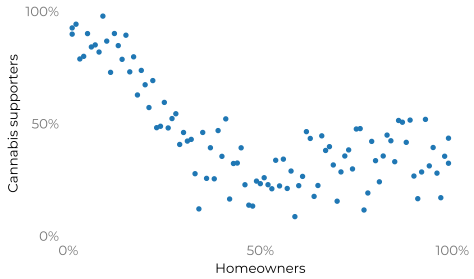
That's because 50k isn't a meaningful value. You don't want people comparing numbers like (GDP in 1980 - 50k) vs. (GDP in 2000 - 50k) because that ratio doesn't mean anything.

Example: Years vs temperature

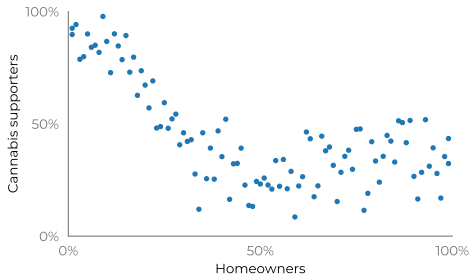
Suppose the x-axis is year and the y-axis is temperature. Should you draw an x-axis at zero? * If the temperature is in Fahrenheit, then no. There are conflicting stories about where the definition of 0 °F even came from. (Maybe it was the freezing point of some mixture of salt and water that Daniel Fahrenheit made?) That's almost certainly irrelevant to whatever you're plotting. * If the temperature is in Celsius, then maybe. If differences from the freezing point of water are relevant to whatever you're plotting, then show it. Otherwise don't. * If the temperature is in Kelvin, and the plot includes 0K, then yes. Or say you're plotting degrees in Fahrenheit and differences from freezing are very important. Then put a line at *32°F*. There's special about zero. The goal is to have something against which comparisons are meaningful. That's most often true for 0 or for 1. And in a philosophical sense, when you claim some value is good to compare against, you're claiming its a "kind" of 0 or 1. ("Identity elements for the two operations in a mathematical ring" if you're into that kind of thing.) But use your judgement

Example: Votes vs. homeowners

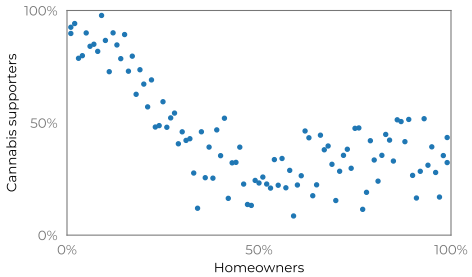
Sometimes you should put lines at the ends of axes, too. Suppose the x-axis is the fraction of homeowners in different counties, and the y-axis is support for legal cannabis. Maybe the data looks like this:



Should you draw axis lines? Well, comparisons to 0% are highly meaningful in both axes, so you might add them both.



That's fine. But comparisons to 100% in either direction are *also* meaningful. So in this case, you really do want



a full box around the plot.

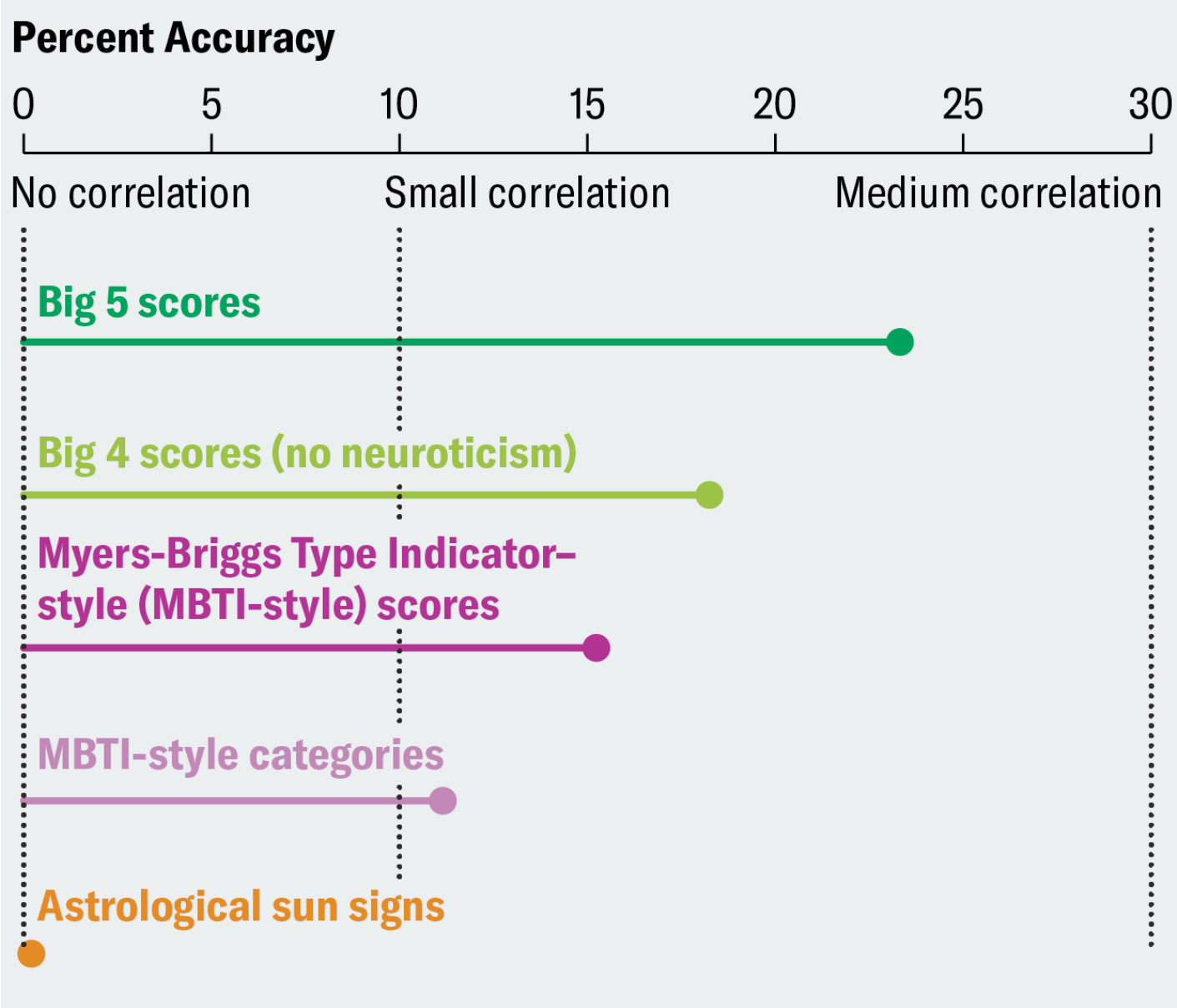
The answer is that both ends of both axes are meaningful. So you probably want a box around the entire plot.

Lines can also be used for evil

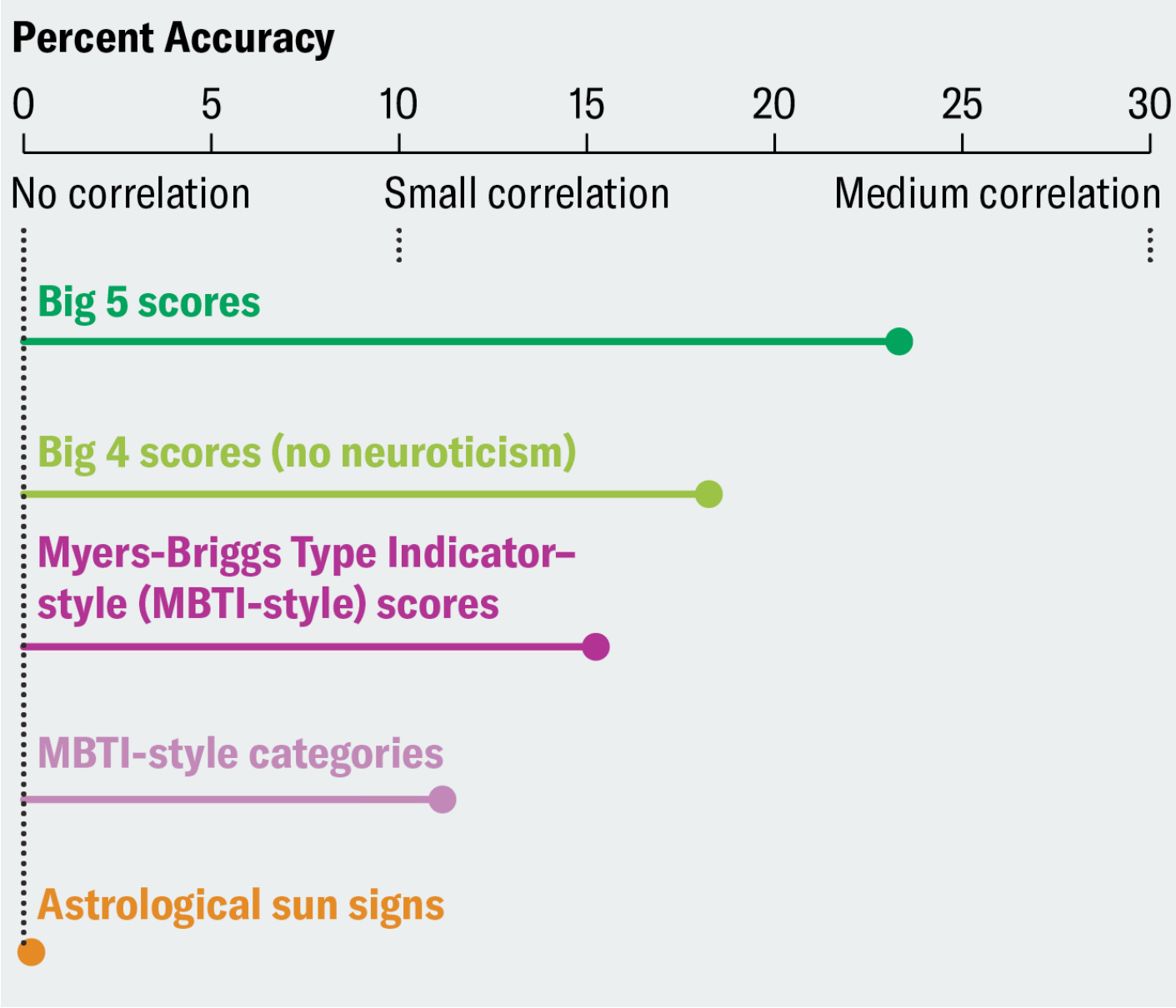
Lots of people hate the Myers Briggs personality test—usually suggesting that it's nonsense and people should use the Big Five test instead. I've long held this was misguided. My argument was that if you take the Myers Briggs *scores* (without discretizing them into categories) then this is similarly informative to the Big Five test without neuroticism, or "Big four". So I was excited to see some recent research that tests this claim. A bunch of people had to take various personality tests and then rate themselves on 40 life outcomes, e.g. how many friends they had. They then computed the correlations between the personality tests and the life outcomes:

Test	Correlation
Big 5	0.23
Big 4	0.18
MBTI scores	0.15
MBTI categories	0.11
Astrology	0.002

Here, the correlation is an "R^2 value"—0 means a test is worthless, and 1 would indicate and a 1 would mean perfect prediction. So, to my surprise, the big 4 did a little better than MBTI scores. But we are here to talk about *figures*, not psychology. So look at how the above numbers were pictured in Scientific American:

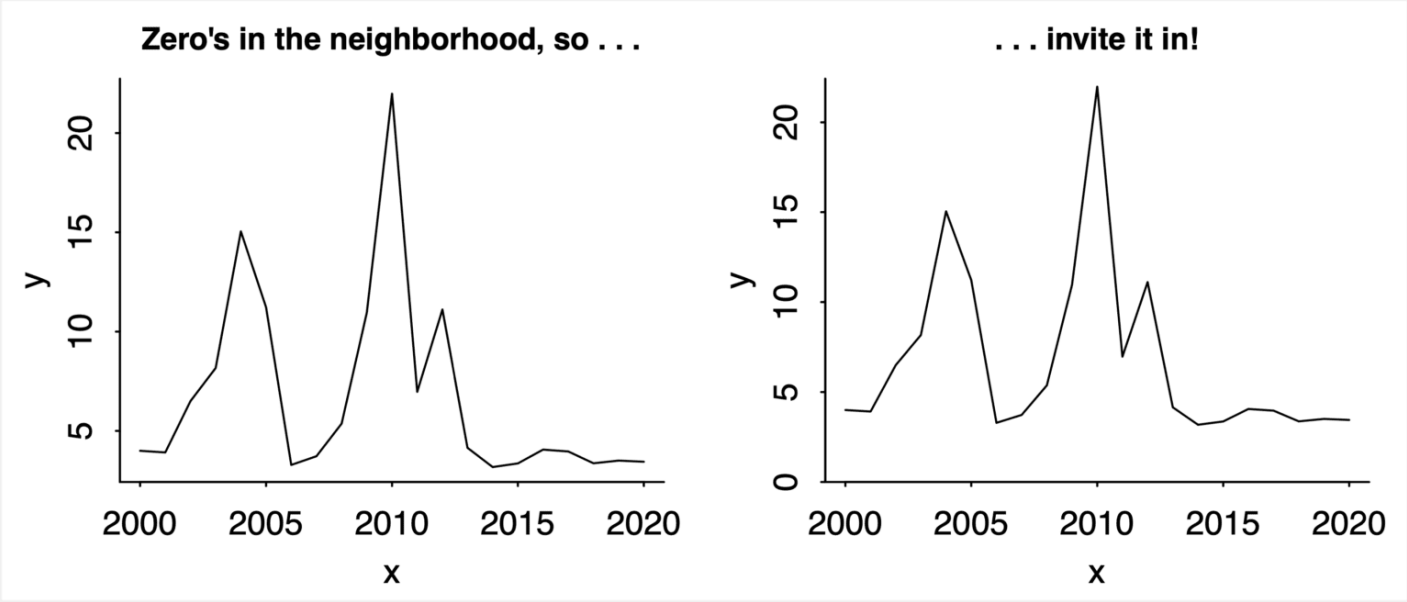


That "small correlation" line is *genius*—your eye naturally compares the dots to it, rather than the "no correlation" line, giving the impression that the Big Four is twice as good as the MBTI. Of course, the difference between a correlation and a "small correlation" threshold is not anything that anyone could ever conceivably care about. A plot that follows the rules I laid out here is much less misleading:

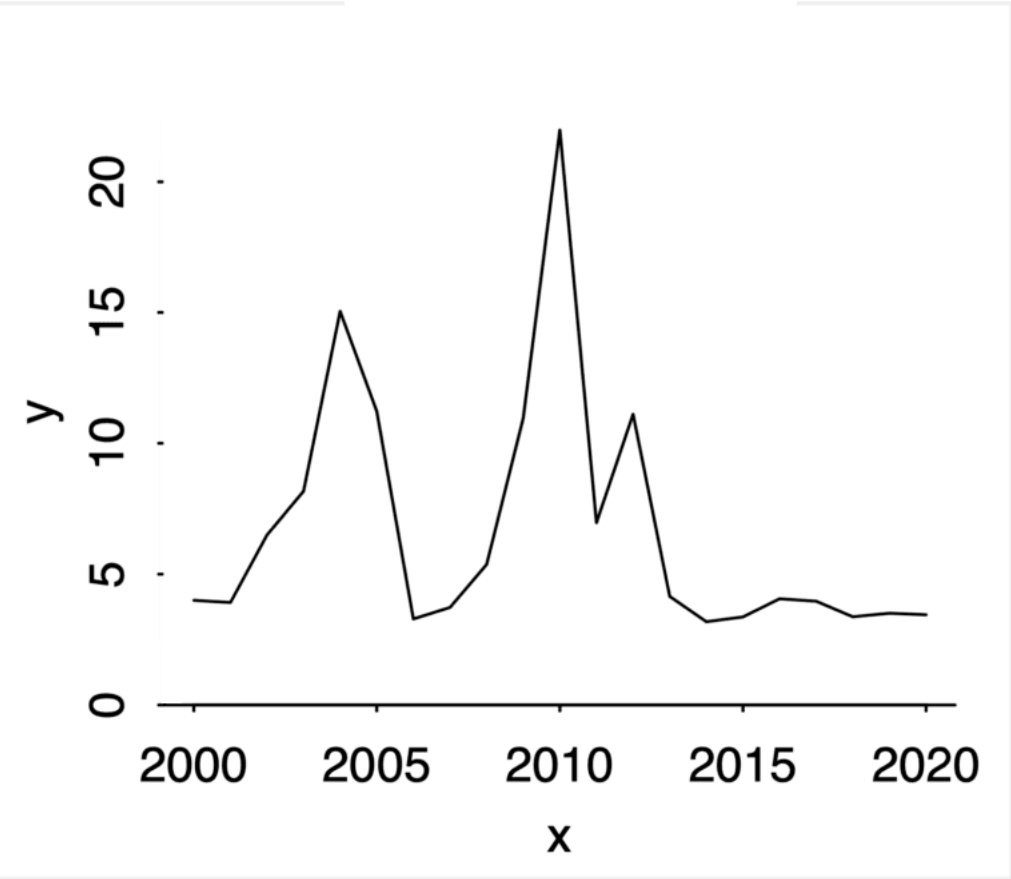


Related advice

Andrew Gelman gives related advice of, "If zero is in the neighborhood, invite it in". The idea is that if your plot *almost* includes zero, and zero is meaningful, then make the plot go all the way to zero:



I agree with this advice, though I'm not sure about that vertical line. If "time since the Slim Shady LP came out" is relevant, then fine. Otherwise, I think this is (slightly) better:



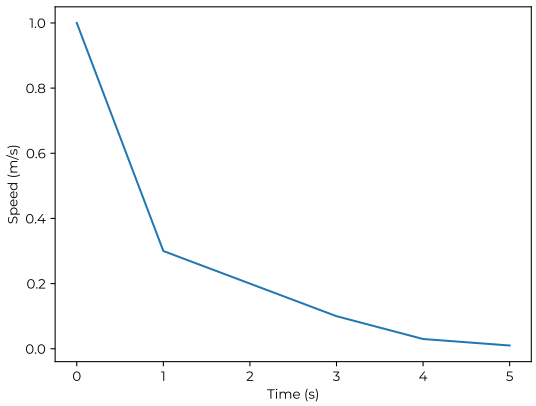
Axis lines are not tickmarks

On a related note, don't conflate drawing an axis line with drawing *tick marks*. See how the previous plot has tick marks for the y-axis, even though there's no line along the y-axis. You can do that.

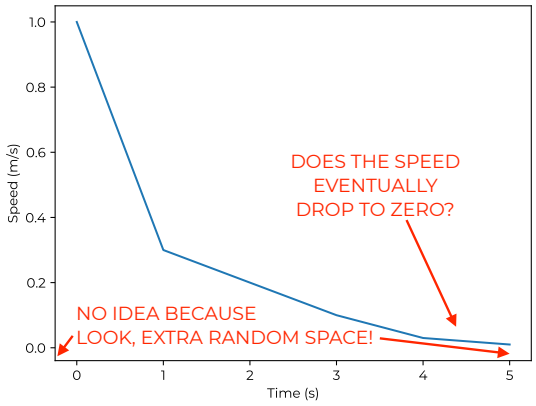
Matplotlib's tragic defaults

Matplotlib is a popular plotting library for making visualizations. Here's an example of calling it:

```
python from matplotlib import pyplot as plt plt.rcParams["font.family"] = "Montserrat" # ALWAYS time = [0,1,2,3,4,5] speed = [100,50,30
```



I don't want to quibble with the default of adding axis lines everywhere. After all, there has to be *some* default, and if the right choice depends on the semantics of the data, you can't expect the plotting library to guess that. Except, I *can't* not quibble because there's a more serious problem. Do you see it? Here's a little hint:



If you want to make sure the data is visible, then don't draw axis lines at all. To draw them, but place them at random locations basically *guarantees* that the default isn't right—either the axis is drawn somewhere meaningless or it shouldn't exist at all.