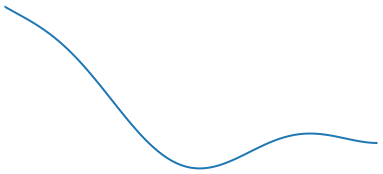
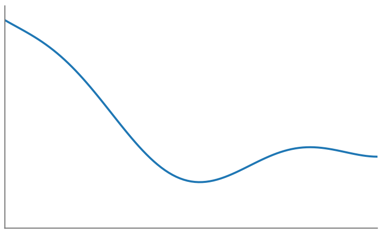
Draw an axis line only when it means something

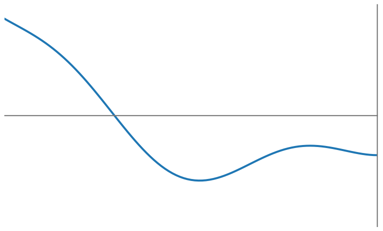
Say you want to plot some data. You could draw it like this:



Or, you could draw it like this:



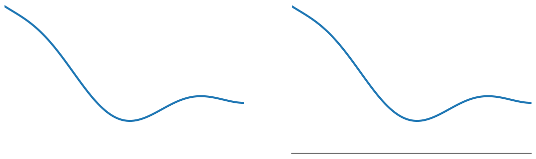
Or, you could do something weird:



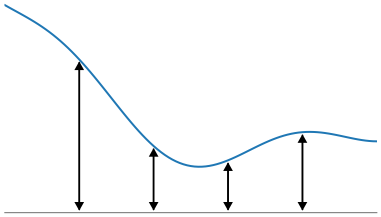
Which is right? Many people seem to treat this as an aesthetic choice. But here, I'd like to give a specific rule for deciding when to include an axis line.

Principles

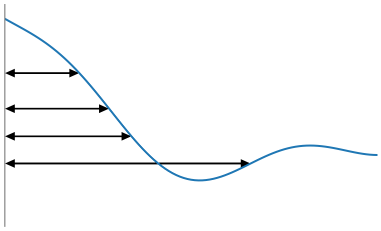
First, you don't have to draw any axis line. It's optional. Your readers aren't so dim that need to draw a box so they don't confuse the plot with the rest of the document. Now, should you add an x-axis? Compare these plots:



Which is better? To answer, mentally picture these arrows:

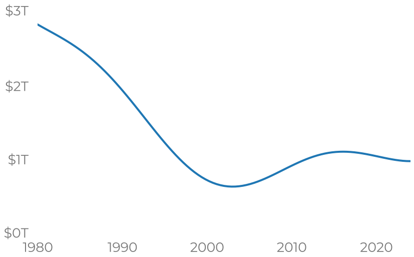


Now, ask yourself, *are the lengths of these arrows important*? When you add that horizontal line, you subtly invite readers to consider those lengths and the ratios between them. Similarly, if you're considering a vertical axis line, you want to ask yourself if the lengths of *these* lines are meaningful:

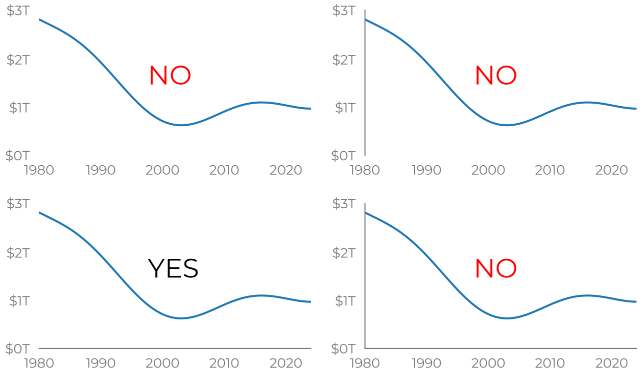


Example: Years v.s GDP

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



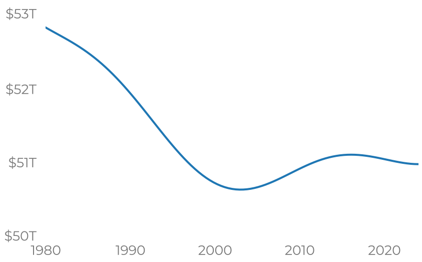
Which axes should you draw?



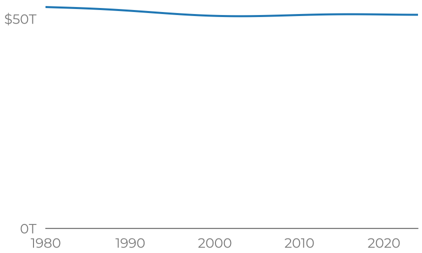
Why? * GDP is an absolute quantity. If GDP doubles, then that means something. * 1980 is an arbitrary reference point. If you're considering the year 2020 vs. the year 2000, all that matters is that they are 20 years apart. The *difference* of 2020 or 2000 from 1980 doesn't really mean anything.

Example: Years vs. GDP again

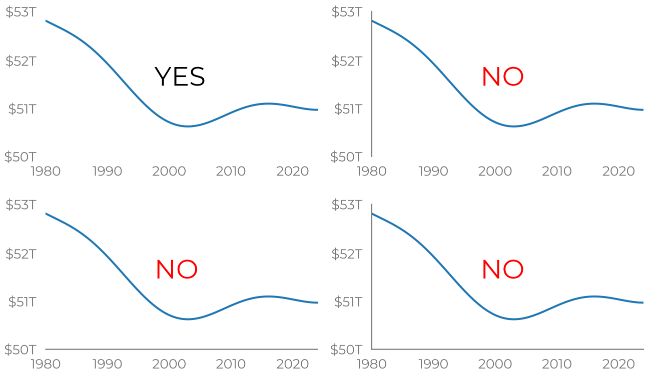
But suppose that the minimum GDP in your plot was more like $50T rather than $0T:



What should you 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. Often you can't fit in zero. In those cases, the right answer changes:



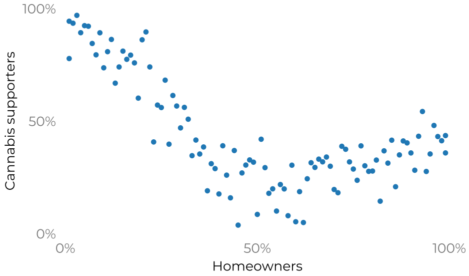
That's because 50k isn't a meaningful value. You don't want people comparing numbers like (GDP in 1980 - 50k) vs. (GDP in 2000 - 50k) because that ratio doesn't mean anything.

Example: Years vs temperature

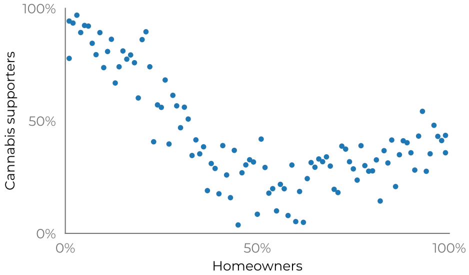
Suppose the x-axis is year and the y-axis is temperature. Should you draw an x-axis at zero? * If the temperature is in Fahrenheit, then no. There are conflicting stories about where the definition of 0 °F even came from. Maybe it was the freezing point of some mixture of salt and water that Daniel Fahrenheit cooked up. 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.

Example: Votes vs. homeowners

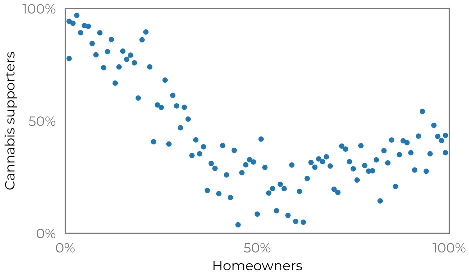
Sometimes you should put lines at the ends of axes, too. Suppose the x-axis is the fraction of homeowners in different counties, and the y-axis is support for legal cannabis. Maybe the data looks like this:



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.



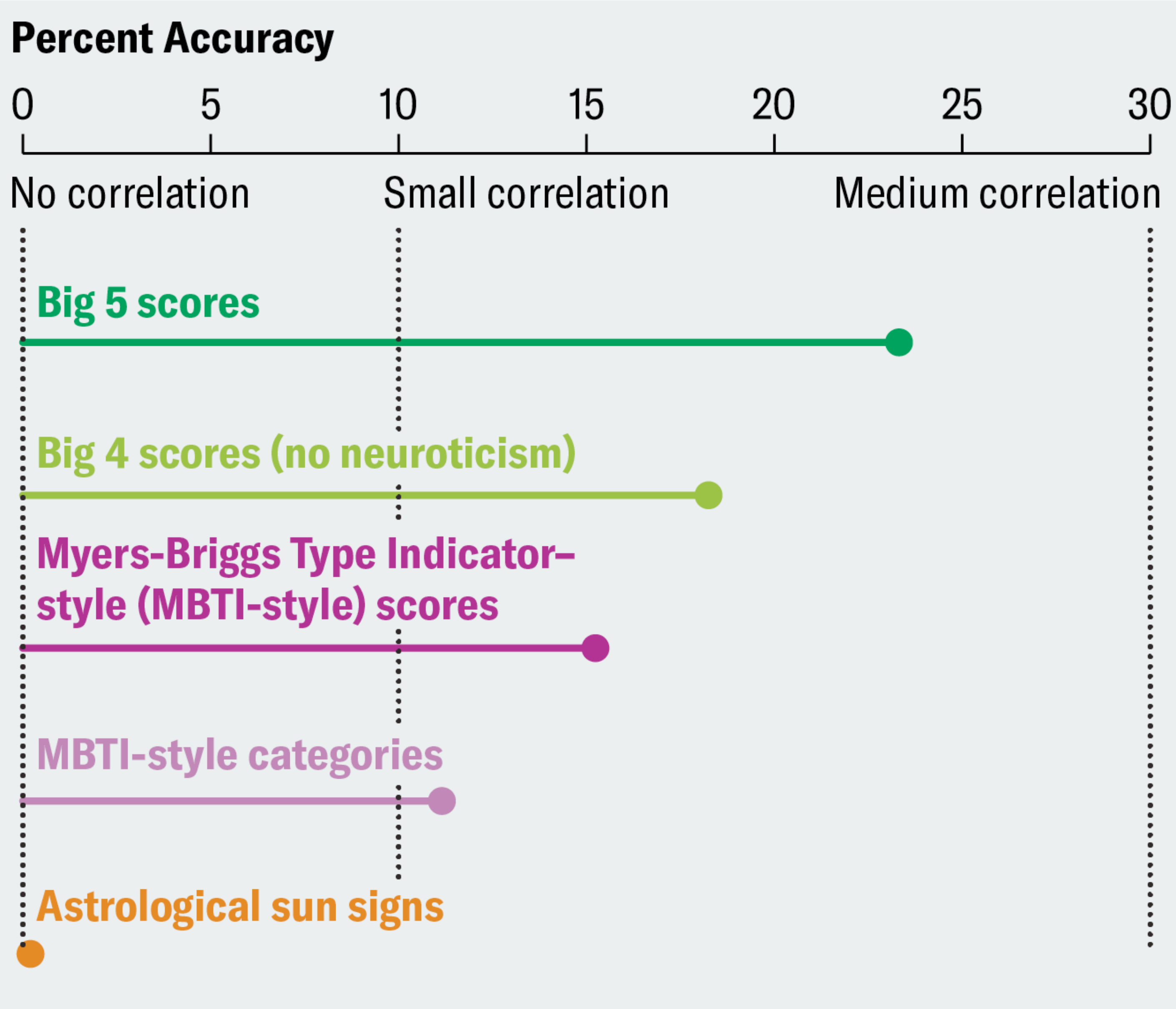
The answer is that both ends of both axes are meaningful. So you probably want a box around the entire plot.

Lines can also be used for evil

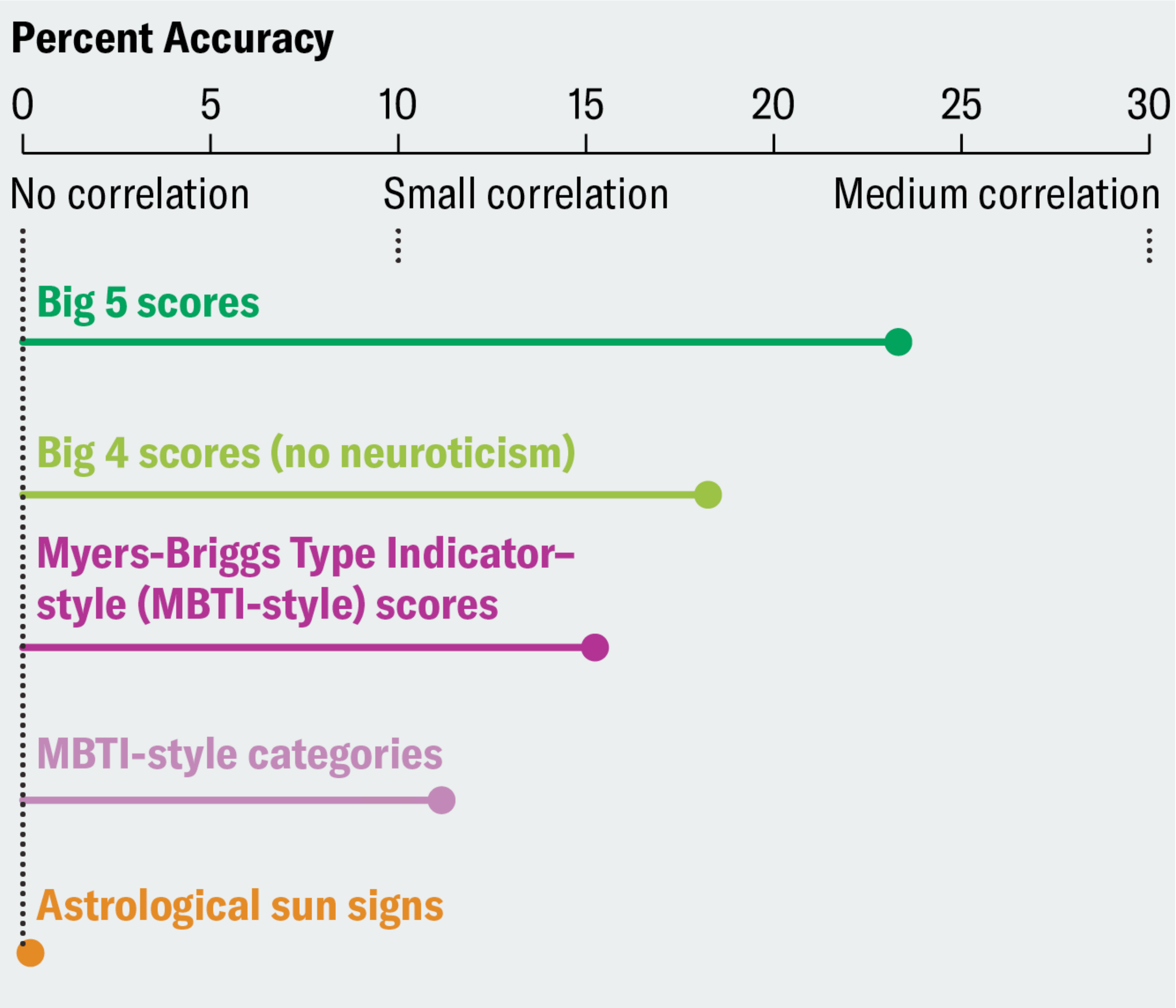
Lots of people hate the Myers Briggs personality test—usually suggesting that it's nonsense and people should use the Big Five test instead. I've long argued this was misguided. My basic claim was that if you take the Myers Briggs *scores* (without discretizing them) then this is similarly informative to the Big Five but dropping neuroticism. Basically, I claimed that Myers Briggs is just as good as the Big Four. So I was excited to see this very recent research that directly tests this claim. These researchers had a bunch of people take various personality tests and then rate themselves on 40 different life outcomes, e.g. how many friends they have or how happy they are. They then computed how well you could predict different life outcomes from the different tests:

| Test | Accuracy | |-|-| | Big 5 | 0.23 | | Big 4 | 0.18 | | MBTI scores | 0.15 | | MBTI categories | 0.11 | | Astrology | 0.002|

Here, accuracy is an "R² value"—0 means the test is worthless, 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:

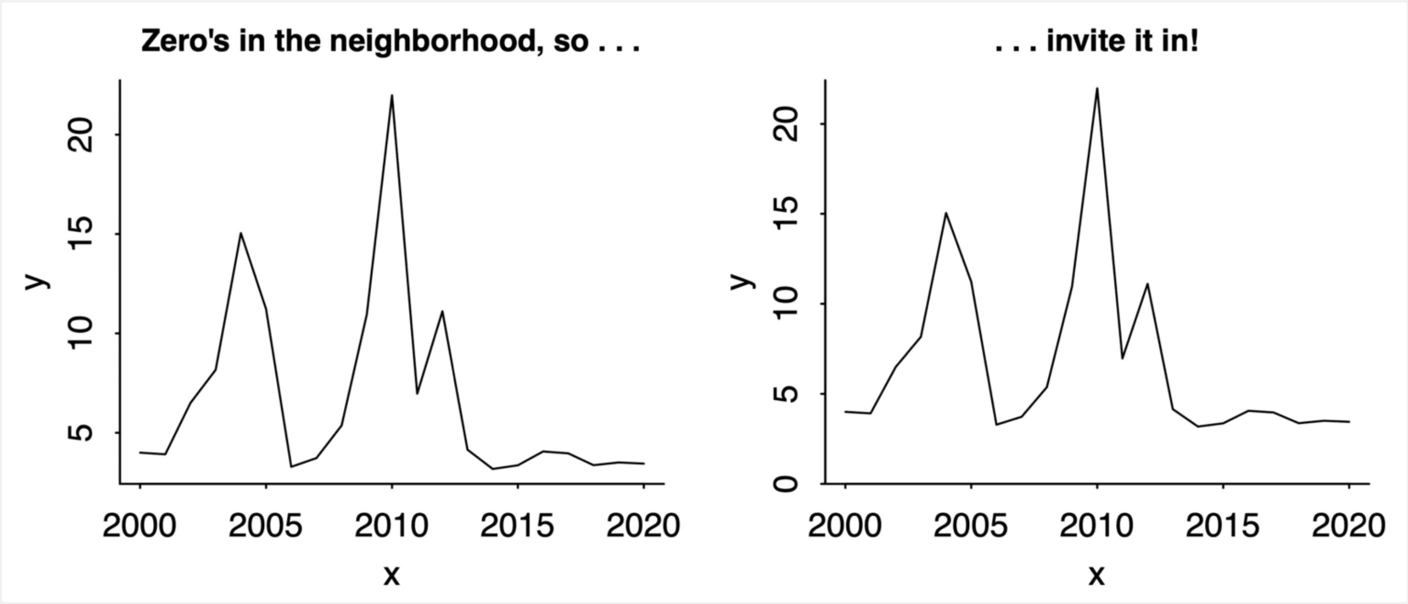


Look at that "small correlation" line. It's genius—your eye naturally compares the dots to the dotted line in the middle, giving the impression that the Big Four scores as almost twice as good as the MBTI. But, of course, the difference between a correlation and an arbitrary threshold for a "small correlation" is not a meaningful quantity. A plot that follows the rules I laid out here is much less likely to mislead.



Related advice

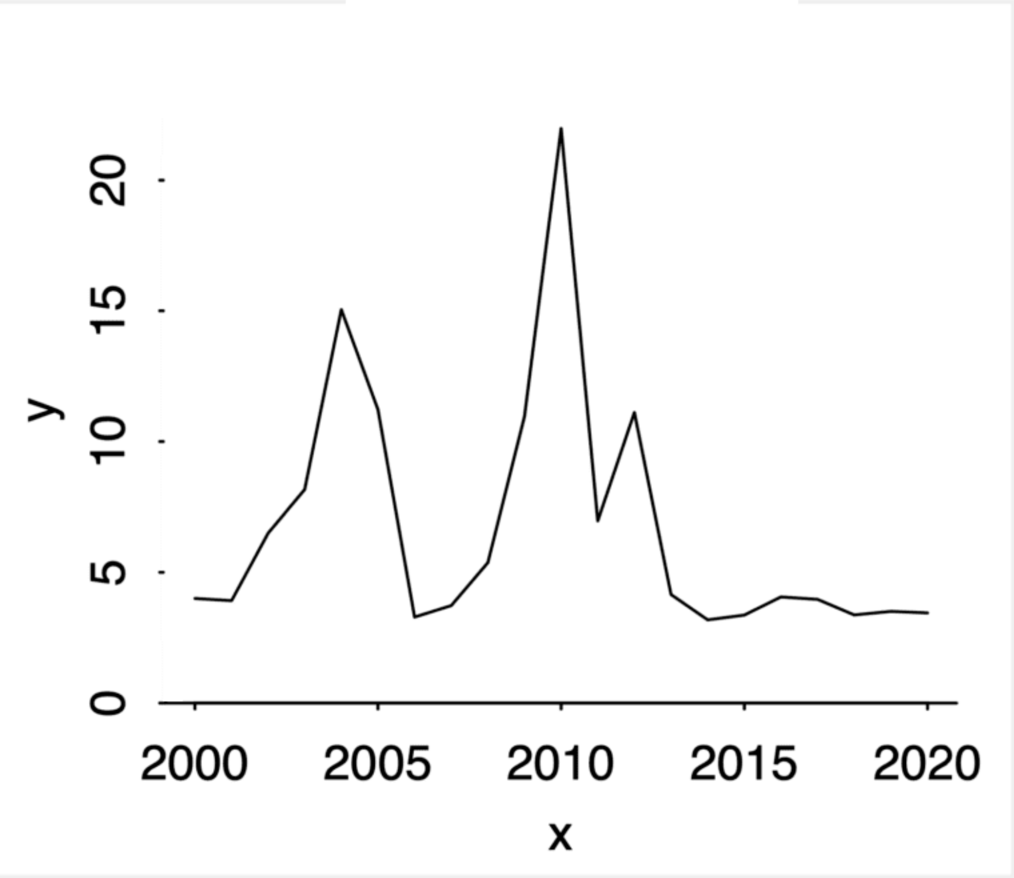
Andrew Gelman gives related advice of, "If zero is in the neighborhood, invite it in". The idea is that if your plot *almost* includes zero, and zero is meaningful, then make the plot go all the way to zero:



Zero's in the neighborhood, so . . .

. . . invite it in!

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

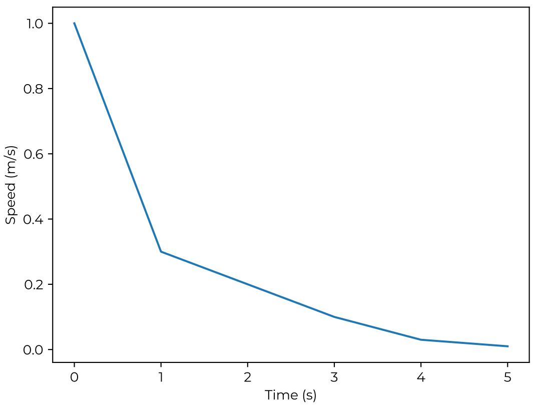
On a related note, don't conflate drawing an axis line with drawing *tick marks*. See how the previous plot has tick marks for the y-axis, even though there's no line along the y-axis. You can do that.

Matplotlib's crime againt plotting

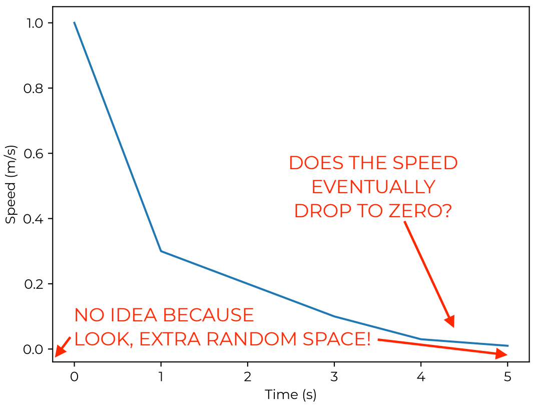
Matplotlib is a popular plotting library for making plots in the Python programming language. Here's an example of calling it:

```
python from matplotlib import pyplot as plt time = [0,1,2,3,4,5] speed = [100,50,30,20,10,1] plt.plot(time,speed) plt.xlabel('Time (s)'
```

And here's the result:



I don't want to quibble with the default of adding axis lines to all four sides. After all, there has to be *some* default, and according to my own advice here, that decision depends on the semantics of the data. But I *can't* not quibble because there's a more serious problem. Do you see it? Here's a little hint:



DOES THE SPEED EVENTUALLY DROP TO ZERO?

NO IDEA BECAUSE LOOK, EXTRA RANDOM SPACE!