DS 100: Principles and Techniques of Data Science	Date: February 9, 2018

Discussion #3

Name:

Reading and Evaluating Figures

1. First answer the questions regarding Figure 1 below and discuss your answers.

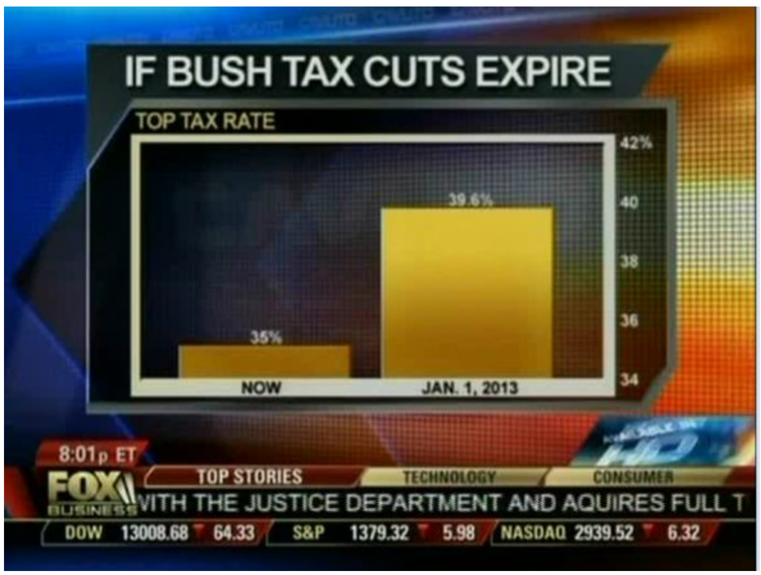


Figure 1: Fox Business chart from 2012, comparing the top marginal tax rate for 2012 and 2013.

(a) What message does this plot convey?

> **Solution:** Fox News wants to portray ending the tax cut as a massive shift in the top tax rate.

(b) How do they physically convey this message?

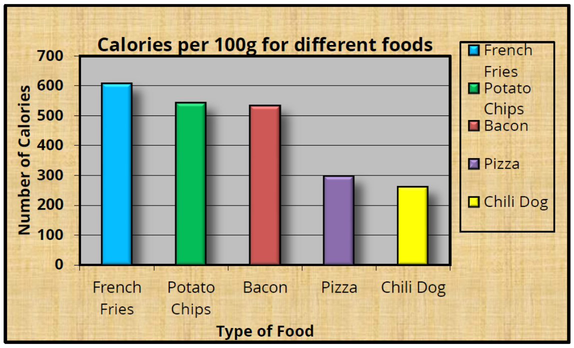
> **Solution:** They make use of the vertical space multiplied by very wide bars. A 13% change in the top tax rate is represented by an area scaling of 400%. The bar on the right takes up roughly 1/5 of the plotting area as opposed to the left, which takes up roughly 1/20.
> It's worth noting that they are reporting the top tax rate (highest bracket).

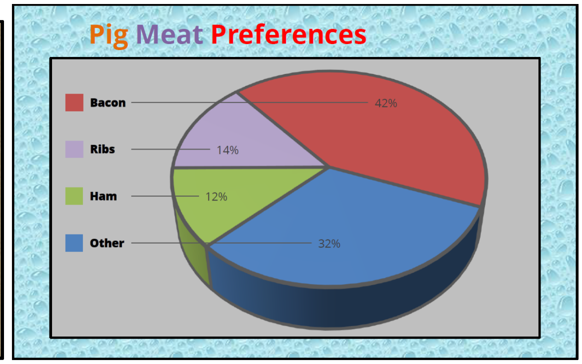
(c) Would you change this chart? If yes, how so? If not, why not?

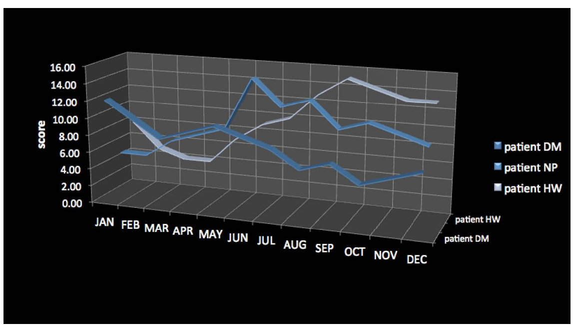
Solution: This chart desperately needs historical context. See Vox video.

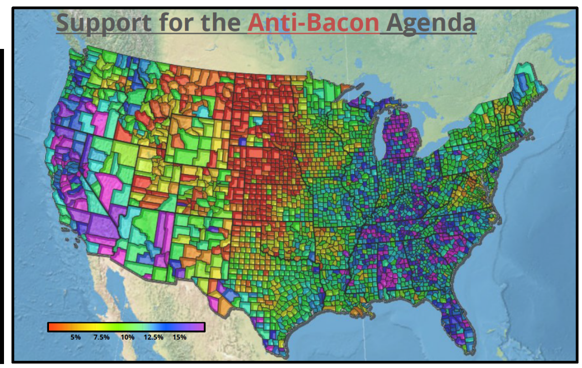
2. Watch Vox's video given at this link watch `http://bit.ly/data100-yaxis`. Discuss your opinions on this video.

3. Discuss the problems with keeping the visualizations below as they are. Color versions are given in the document found on the course website.







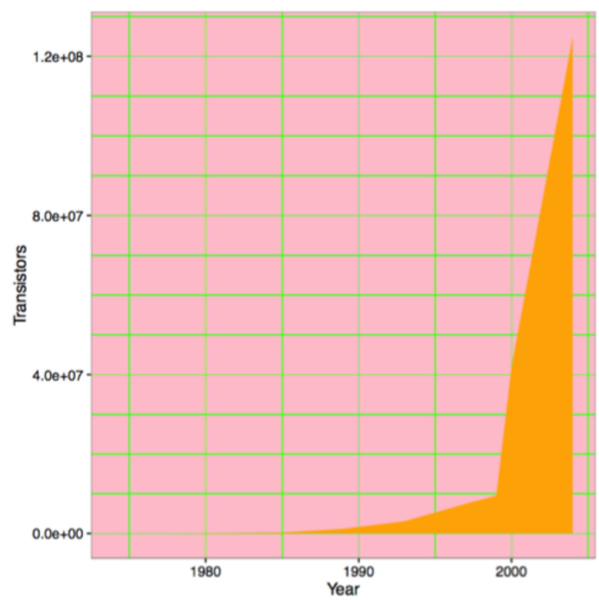


Solution: Refer to the darkhorse pdfs for improvements to the plot.

4. The chips that are present in your computer contain electrical components called transistors. Intel is one of the leading manufacturers of these chips; they released the first chip for home computers in 1979.

   We'd like to visualize the improvements in chips that Intel has made since 1979.

   Download a copy of the discussion from the course website to see the image in color.



   (a) What is this plot trying to communicate?

   Solution: Presumably Moore's Law and exponential growth of computing power in terms of number of transistors.

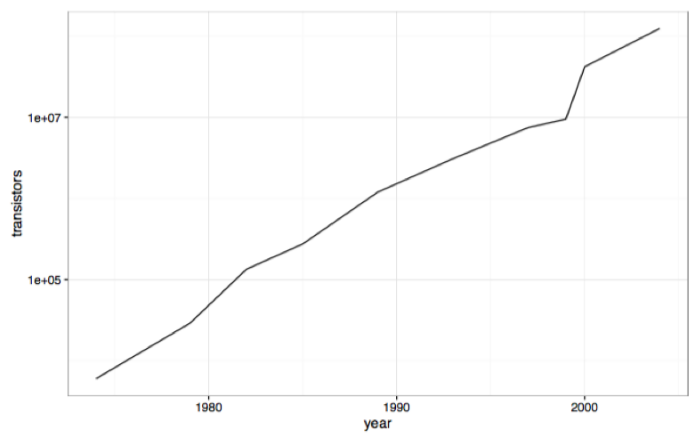
   (b) How does it communicate it? Discuss what data are being mapped to what physical properties of the plot.

   Solution: Number of transistors is mapped to a vertical position. Time is mapped to a horizontal position.

   (c) Are there visual elements that seem unnecessary? List them.

   Solution: The background colors... actually all the color. Possibly the grid (we could make it lighter gray to have it serve as a guide). The fill under the curve doesn't serve any purpose either.

5. Now we look at an attempt to improve the plot.



(a) What is clear in this plot that wasn't clear in the plot before?

> **Solution:** The growth is indeed exponential in time

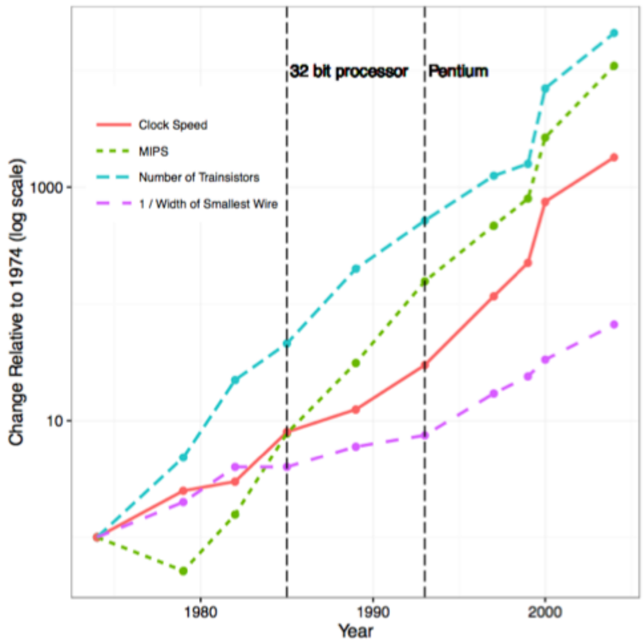
(b) What design decisions made understanding the data easier?

> **Solution:** A log transformation on the y-axis. Removal of chartjunk elements.

(c) Are there still properties of the plot that seem like they need more context?

> **Solution:** Is the number of transistors really a good measure of processing power? What is that jump we see in 2000?

6. Here's a third attempt to improve the plot. MIPS stands for Millions of Instructions Per Second. Download a copy of the discussion from the course website to see the image in color.



(a) What is clear in this plot that wasn't clear in the plot before?

> **Solution:** More metrics other than transistor count are given. Historical context is given. A reference is given for the log scale.

(b) What design decisions made understanding the data easier?

> **Solution:** There's a double-differentiation for the 4 curves: line type and color. This makes it easier for print and color-blind people. Human-readable numbers are given on the y-axis. Reference lines are given for historical events.

(c) Are there still elements that seem like they deserve context?

> **Solution:** What is that jump in 2000 (release of Pentium 4 architecture)? Why did MIPS drop in the late 70s?

(d) Why do we get to draw the four curves on the same plot? The four curves seem to represent data with different units.

Solution: Refer to the discussion that log should be unitless. Hence a reference should always be given. After all, when we plot a 1 on a log_{10} scale, we mean ”10 times bigger”... but ”bigger than what?”. Here they make use of that to use 1974 as a reference year.

Logarithmic Transformations

7. In our discussion, $\log$ will refer to the natural logarithm (also written as ln in other texts). A logarithm in base b will be written as $\log_b$. Sometimes, it will be convenient to write exponential as $\exp$ i.e. $\exp(x) = e^x$. Let $a, b \in \mathbb{R}$ be real numbers and $x, y \in \mathbb{R}_+$ be positive real numbers. Which of these statements about logarithms are true?

   A. $\log xy = \log x \log y$

   B. $\log(x + y) = \log x + \log y$

   C. $\log 1^a = 0$

   D. $\log \frac{x^a}{y^b} = a \log x - b \log y$

   E. $\log_x y^a = a \log y / \log x$

   F. $a^x \exp(b^y) = \exp(x \log a + by)$

8. When visualizing the data, what are some reasons for performing log transformations on the data?

Solution: Possible discussion:

The logarithm linearizes multiplication and scaling. Ink and paper exist in the physical realm, and unless you're traveling near the speed of light, 1D distances act pretty linearly.

Let's say $x = \$1$ and $y = \$1000$. Now imagine that you're going to plot this as-is with a physical scaling of 1cm per dollar. You would need a piece of paper 10 meters wide to just barely fit the two data points on the same plot! On the other hand, if you were to plot $\log_{10} x = 0$ and $\log_{10} y = 3$ instead, we'd only need 3cm of space!

Another issue is the interpretation of scaling. Income is a fairly canonical example. For a person making $20,000 per year, a $5,000 bonus would probably be a fairly big shock to their lifestyle. For a person making $30,000,000$ (how quickly can you read those 0's?), a $5,000 bonus would probably be insulting. Somehow the logarithm captures a better scale than just naively thinking about the raw values.

An aside: What happened to the units? Technically, when we use transcendental functions for our transformations, we actually want unitless arguments. In the logarithm case, we simply divide the data by a reference $\$1$ before we take the log. In practice, most everyone ignores the subtle point...