COSC 91/191, Spring 2019
Lecture 7
April 8, 2019

Scribes: Qijia Shao and J. Peter Brady

Today's lecture uses work from Edward Tufte, as well as information from both the Zobel and Dupré books, and from Tom's own notes and experiences.

1 Principles of figure design

The first part of today's lecture came from Edward Tufte's book [1] where Tufte discusses the principles of figure design, but not how to build the graphs or plots in an application.

1.1 Introduction - Graphical excellence

Tufte defines graphical excellence:

1. It is the well-designed presentation of interesting data — a matter of substance, of statistics, and of design. The figure should say something meaningful, something that you are not getting from any other part of the paper. For example, you can have a table and a figure that matches the table. But the figure should tell something that you cannot see clearly from the table.

2. It consists of complex ideas communicated with clarity, precision, and efficiency. Tom suggested that instead of using precision, Tufte may have meant accuracy; or rather than seeing how much information is in the data, make sure the data is correct. Precision is how much information you provide, no matter whether it is useful. Taking a student's GPA as an example, you can present the GPA with 2 decimal places, 4 decimal places, or more decimal places. Accuracy is whether it is correct. If you say one student's GPA is 3.89898, but the GPA is actually 3.5, the GPA is precise but not accurate.

3. It gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space. Time is the reader's time, and *ink* is the information in the figure other than background.

4. It is nearly always multivariate. Multivariate means the relationship two or more variables versus a common axis. The sample graph in Figure 1 shows the relationship between inflation rate and unemployment rate in the United States from 1956 to 1976. The numbers in the figure are years. There had been an idea that the inflation rate and unemployment rate are inversely related. The figure shows that they had no relationship.

   To show how much information could be packed into a graphic, Tom showed the "CARTE FIGURATIVE des perte successives en hommes de l'Armee Français dans la campaigne de Russie 1812–1813," (Figure 2) which demonstrates the strength of Napoleons army during his Russian campaign. The width of the gray band represents how many soldiers were in the army at each location. It begins from the left at the Polish-Russian border, and the size of army was 422,000. In September 1812, they reached Moscow with 100,000 soldiers. The dark part at the bottom is the retreat. At the end, the

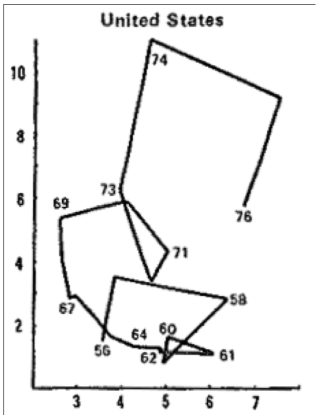


Figure 1: The relationship between inflation rate and unemployment rate in the US from 1956 to 1967.

army returned with only 10,000 soldiers. There are six variables in this graph: the size of the army, two axes for locations during the march, two axes for movement direction, and the temperature over time.

5. It requires telling the truth about the data. In class we discussed how to lie about the data, which the next section discusses.

1.2 Lies about the data

There are several ways that someone can make a graph lie about the data:

1. Shift the axis. Figure 3 starts the Y-axis at $Y = 100$ instead of $Y = 0$; by shifting the axis, the time for size 4 appears to be 4 times larger than it for size 1, though it is actually only 40% higher.
2. Show linear measurements with area. In Figure 4, showing the data in circles with $A = 1$ and $B = 2$ makes B look not twice as big but four times as big. Using area for data that is not area distorts the data. Furthermore, the perceived area of a circle grows more slowly than the actual area. Different people will also perceive the area size differently, so be wary of area to show any important measure. Perceived area = (actual area)^x, where $x = 0.8 \pm 0.3$.
3. Use perspective view when it is not needed. In Figure 5, the fuel economy graph shows an approximate 53% increase in fuel economy by the data, but the graph lines make it look like an increase of about 783% in fuel economy.
4. Omit data and extrapolate. For example, processor speeds have increased dramatically but speeds cannot be extrapolated out to infinity.
5. Show data in terms of the wrong variable. For example, showing economic data over time but not adjusting it for inflation.
6. Stack multivariate data in misleading ways. In Figure 6, I/O time increases in the first three segments but remains constant in the remaining ones. An assumption could be made that since the top of the bar increased, the I/O time was increasing.

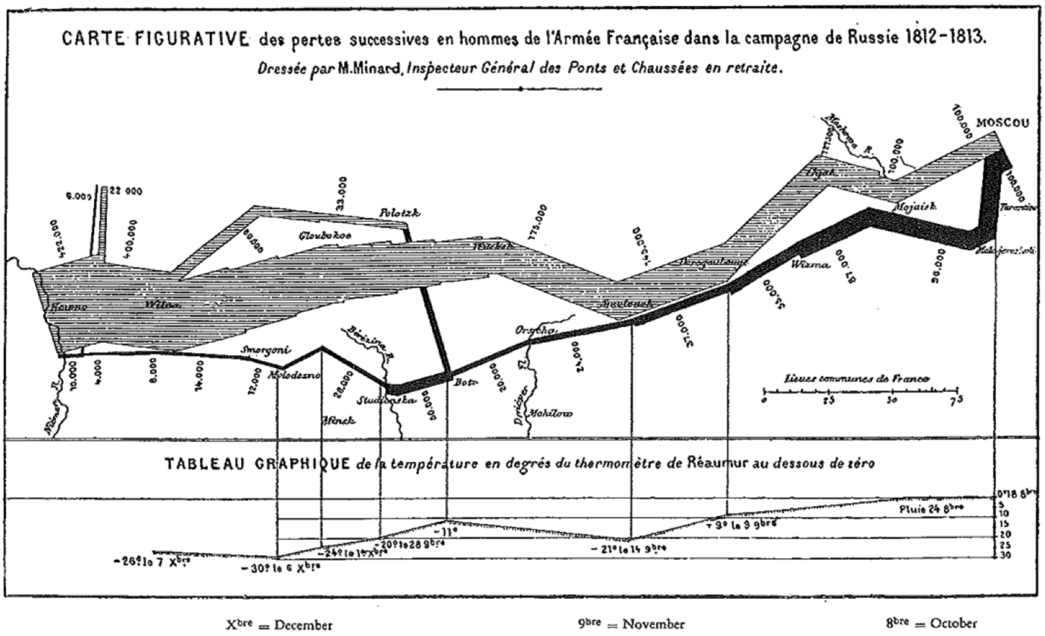


Figure 2: The French Army's running loss of troop strength during Napoleon's Russian campaign.

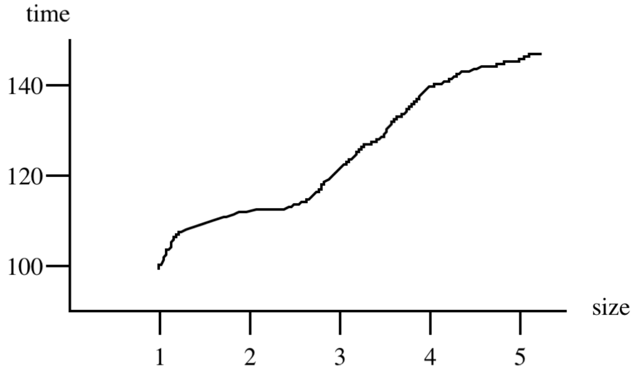


Figure 3: Shifting the axis. It makes the change in time look much greater than it actually is.

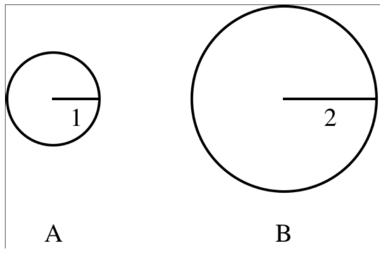


Figure 4: Misuse of area to show the linear relationship. Circle B's radius is twice that of circle A, but B's area is 4 times larger.

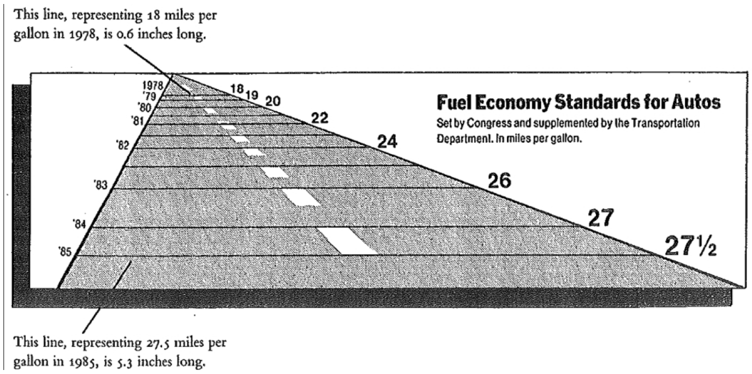


Figure 5: The perspective view makes for misleading results. The change in fuel economy standards for autos is significantly lower than shown by the graph lines.

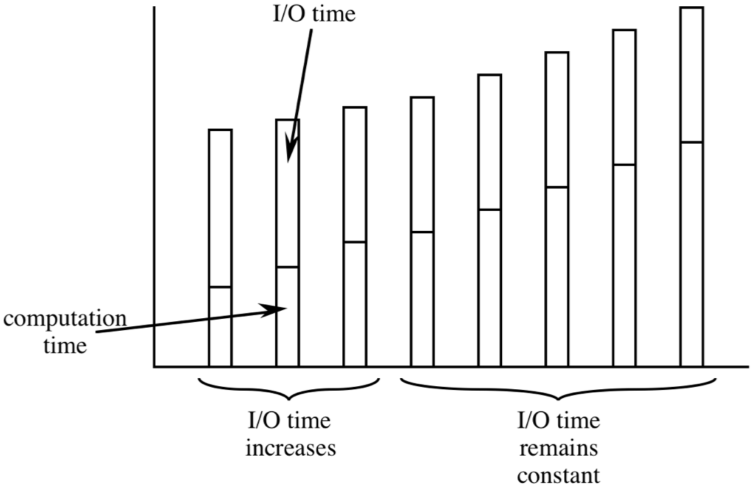


Figure 6: I/O time and computation time. Stacking the I/O time on the computation time makes the I/O time appear to increase.

7. Use area in a way not related to area or in a misleading way. Figure 7 shows incidence rates of cancer by county. Comparing San Bernadino and Los Angeles Counties in California, you might think they have the same large incidence of cancer, except that the map doesn't take population density into account; San Bernadino County is sparsely populated compared with Los Angeles County. With this map, you see the area but not the population.

1.3 Maximizing effectiveness

Tufte discusses *data-ink*, which he defines as the non-erasable core of the graphic. He then looks at the data-to-ink ratio, which is the data-ink versus the total ink used to print; optimally this ratio should be close to 1. For example, adding borders, sea serpents, and smiling suns to a map reduces the ratio.

Tom then challenged the class to find six ways that Figure 8 is telling us its value. The class found seven:

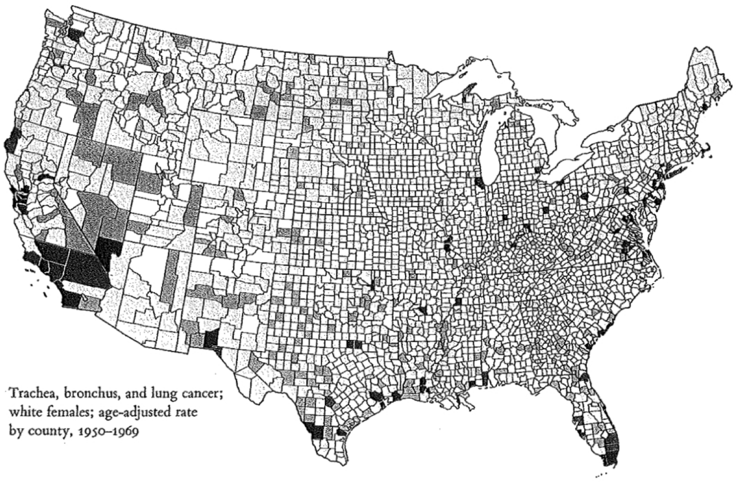

Trachea, bronchus, and lung cancer;
white females; age-adjusted rate
by county, 1950–1969

Figure 7: Cancer rates in the United States by county. The map displays cancer incidence without taking population density into account.

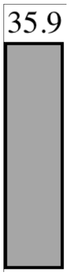
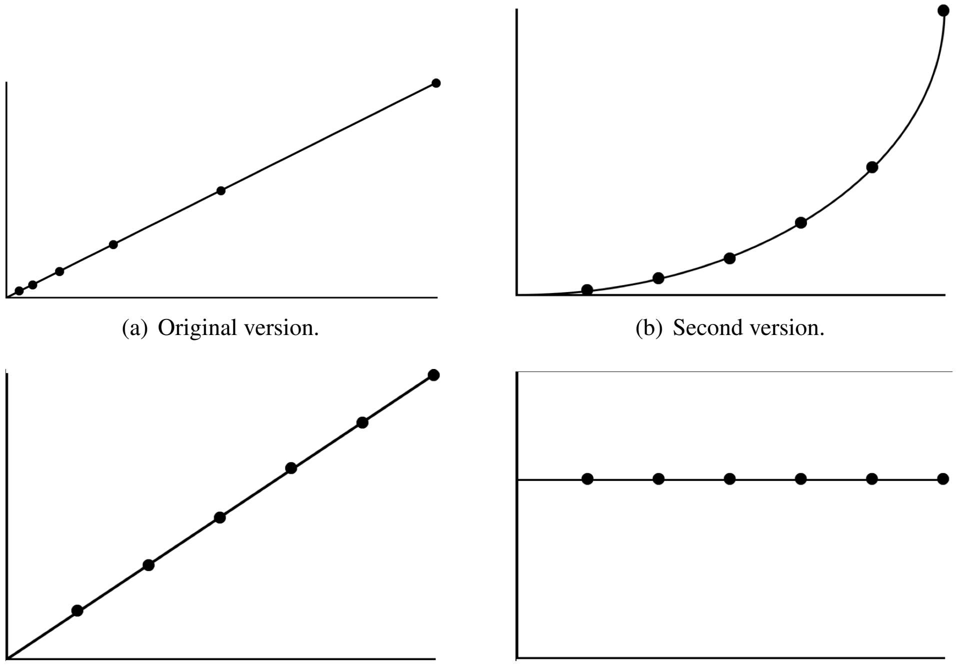


Figure 8: Seven ways to demonstrate the value 35.9.

1. The value 35.9.
2. The gray scale shading. (That was a new one for Tom.)
3. The height of the left edge.
4. The height of the right edge.
5. The position of the top line.
6. The height of the shading.
7. The height of the number itself.

According to Tufte, we could erase six of the seven items and we would still have the information we needed. The final definition from Tufte is *chartjunk*, the extraneous and unnecessary decoration that is unneeded (like sea serpents around the edge of a map.)



(a) Original version.

(b) Second version.

(c) Third version.

(d) Final version.

Figure 9: The evolution of one figure. **(a)** The original graph showing a linear relationship. **(b)** The graph using a log scale for the x-axis. **(c)** The graph using a log scale for both the x-axis and y-axis. **(d)** The final version of the graph normalized with $y/f(s)$ vs. x.

1.4 Tom's pet peeves about graphs and charts

Here are some of the things that Tom said you should try not to do when making graphs or charts:

1. Too small to read. The graph is tiny or the text has been reduced to the point it's unreadable. Make sure that graphs are readable and the text is legible.

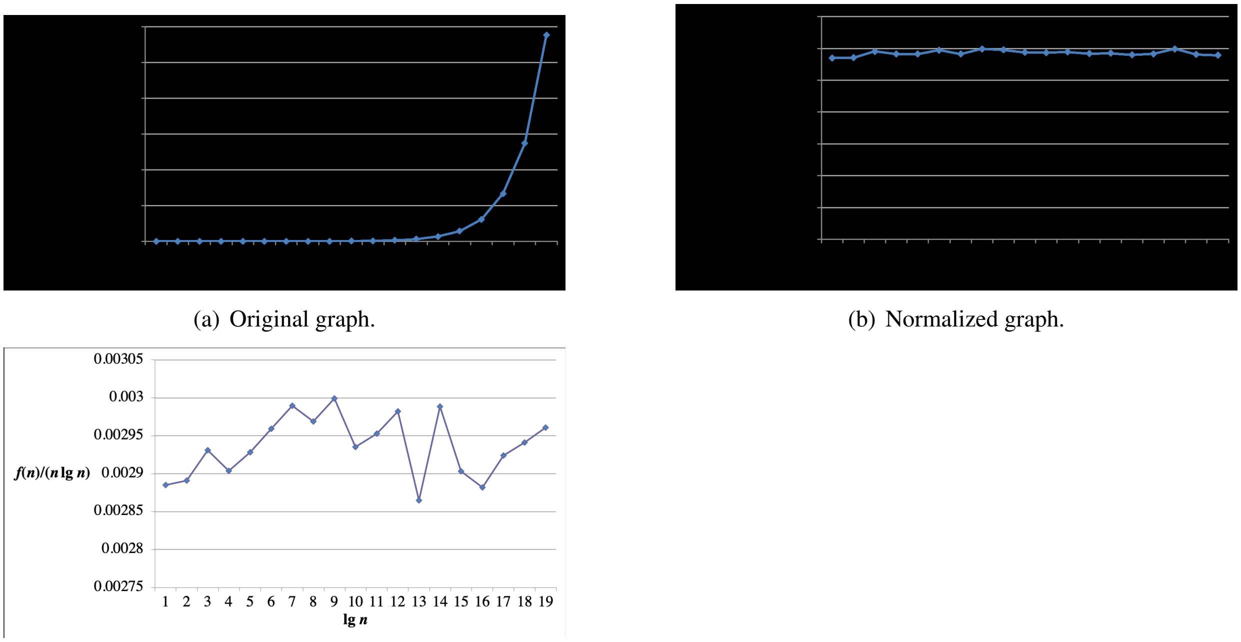
2. Multiple lines that are too similar and cannot be differentiated. The main reason this happens is an originally colored graph was converted to grayscale or black and white without regard to how it might look.

3. Multiple lines that are drawn too close together so that they cannot be easily distinguished.

4. The legends mismatch the dominant order of the lines.

Tom then told a story about working with one of his first senior thesis students in the mid-1990s. The permutations they were working with were defined only for input sizes that were powers of 2. The running times were linear in the input size. They wanted to plot running time versus input size, but they had a problem with a lot of points near the origin; resolution was lost at small input sizes (Figure 9(a)).

The first change they made to the plot was to make a log scale for the input size, but that makes the plot look exponential and not a linear relationship (Figure 9(b)). Next they used a log-log plot, which was submitted with the paper for review (Figure 9(c)). The reviewer had issues with this plot because they were not showing a linear relationship, but a polynomial one. The solution they ended up using in the paper was



(a) Original graph.

(b) Normalized graph.

(c) Overcorrected graph.

Figure 10: A sequence of the steps to normalize a sample graph. **(a)** A sample log graph created in Excel. **(b)** The sample log graph normalized by $n \lg n$. **(c)** The normalized graph where Excel's default scaling capped the Y-axis, removing the zero point.

to have two linear plots: one displaying the whole range, and one zoomed in to show the lower left area of the plot.

Many years later, Tom came up with a better solution to this problem by using normalization. In Figure 9(d)), to show $y = f(x)$, they plot $y/f(x)$ versus x and get a flat line. Normalization removes issues with curve, slope, and scale.

One warning is to make sure you choose the correct scale for Y-axis. Figure 10(a) shows a curve Tom created in Excel and Figure 10(b) shows the same data that was normalized with $n \lg n$. Excel's default scaling is known to cap ranges in Y which can distort the curve being presented; the overcorrected graph is seen in Figure 10(c).

2 Captions

Tom prefers long captions (a.k.a. "*Scientific American* style") that completely describe the figure. Some paper referees complain that long captions are repeating what is in the text; Tom agrees with them but this is not the reason that we should not use long captions. Dupré says to use tags in either all captions or no captions. A tag is a noun phrase. Tom doesn't think we need the rule that Dupré says. It is OK to vary by the figure but everything after the first thing (a tag or sentence) should be sentences, which can be described by the regular expression (tag | sentence) sentence*(part (tag | sentence) sentence*)*.

Figure 11 demonstrates the tag-followed-by-sentences method as well as integrating the parts of the figure into the caption. A more traditional approach is shown in Figure 12 where you have a tag, followed by tags for (a) and (b), then two sentences describing the operation. Figure 13 is an example where a tag

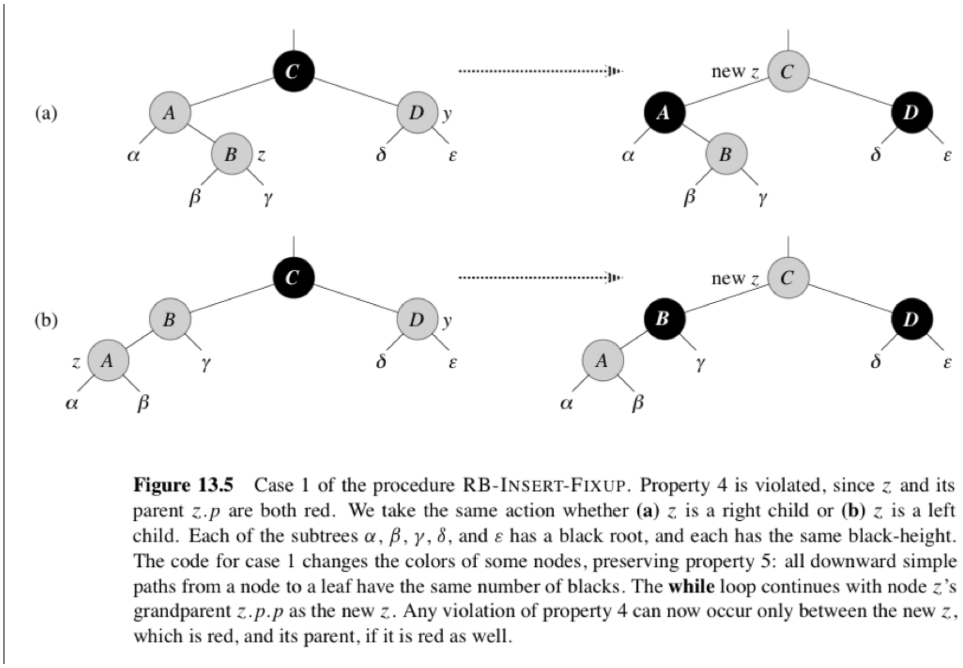


Figure 13.5 Case 1 of the procedure RB-INSERT-FIXUP. Property 4 is violated, since z and its parent $z.p$ are both red. We take the same action whether **(a)** z is a right child or **(b)** z is a left child. Each of the subtrees α, β, γ, δ, and ε has a black root, and each has the same black-height. The code for case 1 changes the colors of some nodes, preserving property 5: all downward simple paths from a node to a leaf have the same number of blacks. The **while** loop continues with node z's grandparent $z.p.p$ as the new z. Any violation of property 4 can now occur only between the new z, which is red, and its parent, if it is red as well.

Figure 11: Sample of *Scientific American* style captions. The caption contains a tag followed by sentences that integrate the parts of the figure into the caption.

is used for the main captions but sentences only for the parts. Dupré would object to this style but Tom thinks it's OK. Dupré goes on to say, "do not give long-winded explanations of a figure in the text, and give inadequate information in the caption." Tom believes the figure should be described in the text and in the caption.

3 Location

Tom discussed how to place figures in a paper.

- Do not place a figure before its first reference.
- Try to put the figure and its first reference the same page. If you cannot do that, put the figure on the next available page.
- Make the figures float. Tom has found that letting the figures float in his book has worked out well.

4 Miscellanea

We've tried to compile the fun, arcane, and possibly less-than-useful information here.

Tom used the word *hark* only once in class today.

In the graphic of Napoleon's losses in Russia, the temperature scales uses degrees Réaumur (Re), which was in use in France at the time. To convert to Celsius, $C = Re \times 1.25$. The Réaumur scale was invented

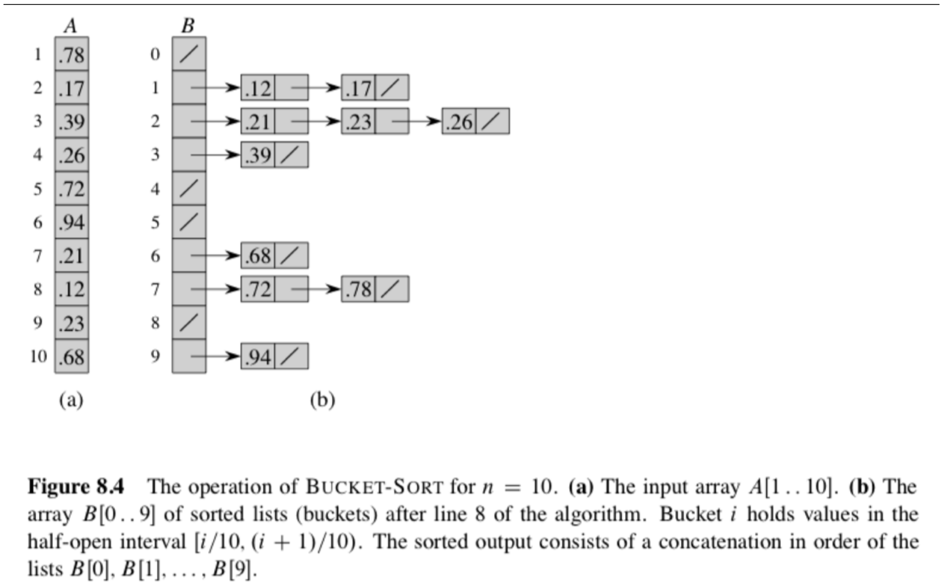


Figure 8.4 The operation of BUCKET-SORT for $n = 10$. **(a)** The input array $A[1..10]$. **(b)** The array $B[0..9]$ of sorted lists (buckets) after line 8 of the algorithm. Bucket i holds values in the half-open interval $[i/10, (i+1)/10)$. The sorted output consists of a concatenation in order of the lists $B[0], B[1], \ldots, B[9]$.

Figure 12: A traditional approach to *Scientific American* style captions. The figure's caption contains a tag followed by tags for (a) and (b). Two sentences following the tags completely describe the operation.

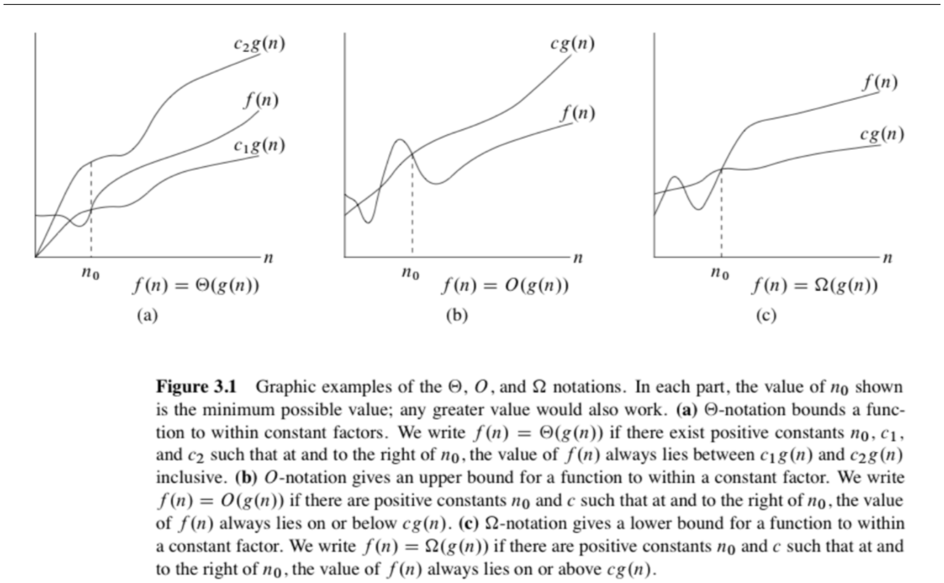


Figure 3.1 Graphic examples of the Θ, O, and Ω notations. In each part, the value of n_0 shown is the minimum possible value; any greater value would also work. **(a)** Θ-notation bounds a function to within constant factors. We write $f(n) = \Theta(g(n))$ if there exist positive constants n_0, c_1, and c_2 such that at and to the right of n_0, the value of $f(n)$ always lies between $c_1g(n)$ and $c_2g(n)$ inclusive. **(b)** O-notation gives an upper bound for a function to within a constant factor. We write $f(n) = O(g(n))$ if there are positive constants n_0 and c such that at and to the right of n_0, the value of $f(n)$ always lies on or below $cg(n)$. **(c)** Ω-notation gives a lower bound for a function to within a constant factor. We write $f(n) = \Omega(g(n))$ if there are positive constants n_0 and c such that at and to the right of n_0, the value of $f(n)$ always lies on or above $cg(n)$.

Figure 13: One main tag and sentences for a caption. The figure has a tag and sentence for the main caption, but only sentences for the parts.

in 1731, but fell out of use by the late 19th century. [1]

Finally, Edward Tufte gives one-day seminars around the country[2] for $380 per person ($220 for students and faculty) and his four books are included. The four books retail on Amazon for about $150.

References

[1] Tufte, Edward R., The Visual Display of Quantitative Information, Second Edition. Graphics Press, Cheshire, Conn. 2001.

[1] Encyclopædia Britannica online, https://www.britannica.com/science/Reaumur-temperature-scale

[2] https://www.edwardtufte.com/tufte/courses