The Visual Display of Quantitative Information

Marvin A. Ruder

Heidelberg University
Winter term 2019/20
Seminar "How do I lie with statistics?"

Abstract

In this report, different types of graphical integrity are presented. We show how newspapers and magazines use different techniques to create graphics that lack graphical integrity and therefore influence the readers' opinion on a matter. We present methods that help creating integer graphics and extract useful information from non-integer graphics.

This report is part of the winter term 2019/20 seminar "How do I lie with statistics?" at Heidelberg University. It is based on the book "The Visual Display of Quantitative Information" by Edward R. Tufte [1].

Chapter 1

Graphical Integrity

When looking up graphical integrity in a dictionary, one will find a definition such as "the quality of being honest and having strong moral principles". It follows that a graphic that lacks integrity is in some way dishonest and immoral, a technique often used by graphic designers to influence the opinion of the audience of a graphic on a certain topic, this opinion being different than the opinion the audience would develop from the data the graphic is based on. In this section, it is discussed how the lack of graphical integrity can be detected, and methods on how to create integer graphics are presented.

1.1 Distortion in Graphics

In general, a graphic can be considered as distorting whenever its visual representation is not consistent with the numerical representation of the data it visualizes. While the numerical representation can be obtained easily and unambiguously, the term "visual representation" can be interpreted in different ways.

Let us assume an example graphic that visualizes data using the area of two-dimensional shapes, as can be found in Figure 1.1.

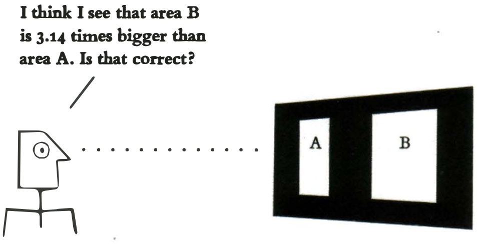

I think I see that area B is 3.14 times bigger than area A. Is that correct?
A
B

Figure 1.1: A graphic using rectangular areas for data visualization.

A simple understanding of visual representation would consider the actual area of the shape that everyone can measure using a ruler as representative for the visualized data.

A different approach would consider the perceived visual effect of the area difference, as shown in Figure 1.1. Here, we assume that people will systematically over- or underestimate a relation in area size.

To decide which concept of visual representation helps at maintaining graphical integrity, we need to analyze the perceived visual effect. For this, we will look at some experimental results. During an experiment, a large number of people were given circles of different sizes and were asked to guess the area of the circles. From the answers, the following power law has been derived:

$$\text{reported perceived area} = (\text{actual area})^x, \quad x = 0.8 \pm 0.3 \qquad (1.1)$$

From the exponent x being between 0 and 1 we can see that the perceived area grows more slowly than the actual area, and the average person underestimates relations in area size. Also, from the comparably large uncertainty factor of 0.3, we can deduct that different people have very different perceptions of area sizes. While some people made no difference between the relation of the circle areas and the relation of the one-dimensional circle diameters (happening at $x = 0.5$), some people even overestimated the relation of the circle areas (at $x > 1$).

The only solution to this dilemma is to not use the perceived visual effect of an area as visual representation in graphics. From this, [1] deducts the following principles that shall enhance graphical integrity:

> "The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.
>
> Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data." [1]

1.1.1 The Lie Factor

When analyzing the integrity of graphics, a measurement of the distortion in graphics would be helpful. For this, the Lie Factor is introduced. It can be computed using the equation

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}. \tag{1.2}$$

It is desirable that both effects are of the same size, following the previous principle. A lie factor of 1 thereby certifies that the designer did a good job and did not use distortion in his graphic. However, if the lie factor is larger or smaller than 1, we can infer that the graphic distorts by overstating or understating the visualized effect in the data respectively. Generally speaking, most distorting graphics involve overstating.

The following example uses the lie factor to describe the distortion involved. In Figure 1.2, the fuel economy standards (in miles per gallon[1]) between 1978 and 1985 set for cars by the U.S. Congress and the Department of Transportation are visualized using the lengths of horizontal lines along a street. While 18 mpg in 1978 are represented using a 0.6 in long line, 27.5 mpg in 1985 are visualized using a line 5.3 in long. From the numbers alone, we can immediately see that this graphic distorts the underlying numbers.

Using Equation 1.2, we can also quantify this distortion. From 1978 to 1985 the fuel economy standard increases by 53 percent—from 18 mpg to 27.5 mpg. The lengths of the corresponding numbers however increase by 783 percent, from 0.6 in to 5.3 in. The two ratios can be used to calculate the lie factor as

$$\frac{783\%}{53\%} = 14.8. \tag{1.3}$$

When we take the information from Figure 1.2 to create an integer graphic, shown in Figure 1.3, more information is instantly revealed, such as periods of slow or rapid increases of the fuel economy standard. Adding contextual information such as the fuel efficiency of the cars currently in use can help the audience getting a better understanding of the depicted information.

1.2 Data and Design Variation

When humans look at a graphic, they have expectations for its consistency. For example, patterns that extend over some parts of

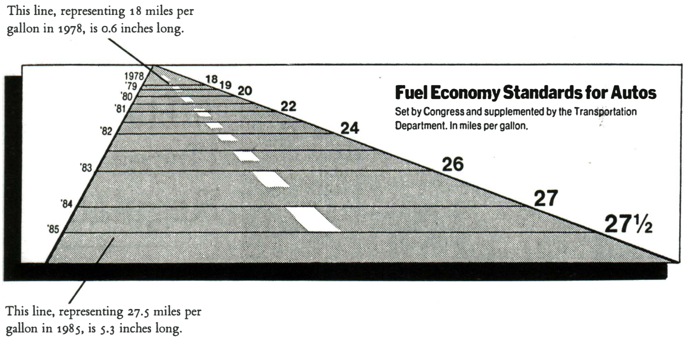


Figure 1.2: *New York Times*, August 9, 1978, p. D-2.

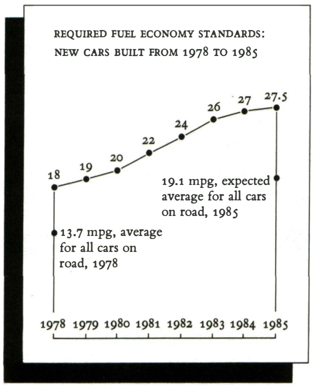


Figure 1.3: An integer graphic showing the information from Figure 1.2.

[1] While the unit *miles per gallon* will confuse most readers in Germany—as the common unit for fuel economy in Germany is *liters per* (typically 100) *kilometers*— the average New York Times reader will not be distracted by this unit, so using this unit does not conflict with graphical integrity.

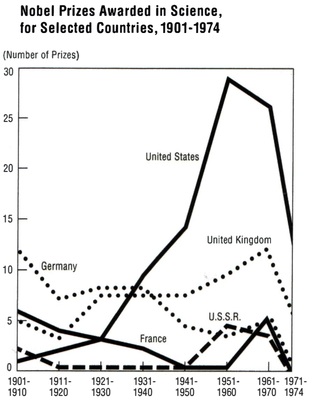


Figure 1.4: National Science Foundation, *Science Indicators, 1974* (Washington, D.C., 1976), p. 15.

a graphic are expected to continue along the whole graphic. Following that, violating this pattern will create wrong expectations and leads to deception.

Let us take a look at the example in Figure 1.4. It visualizes the nobel prizes won by scientists from a number of countries in the time range from 1901 to 1974.

From the rapid decline in the last time interval, one might assume that much less nobel prizes were awarded to the scientists of the evaluated contries during the 70s. As some people may now speculate about the reason for this decline and may conclude that all scientists suddenly became dumb or died during a gigantic scientific accident, the keen and precise observer will discover that the last time interval contains only four years between 1971 and 1974, whereas all other time intervals are ten years wide.

In this example, the consistent width of the time interval is a pattern, which is interrupted with the last time interval. However, the audience expects this last interval to be also ten years long, which creates deception.

Here, we can split the parts of the visualization into two groups: data variation and design variation. Data variation refers to the different values displayed in the chart varying over time, which is what the audience is interested in. On the other hand, design variation describes the change in the time interval size. As demonstrated here, design variation is an unwanted technique and should be avoided, which leads to the simple principle

> "Show data variation, not design variation." [1]

When design variation is removed from the previous example using time intervals of equal size, the visualization appears far more reasonable, as can be seen in Figure 1.5

Another example shows design variation at an alarming rate. In Figure 1.6, the OPEC oil price is visualized using a bar chart. Immediately visible are the two different parts of the graphic, where bars represent either yearly or quarterly oil prices. In addition to that and without any special notice, the vertical scale of the four quarterly bars is different for every individual bar. The different time and price scales are shown in Table 1.1.

With this variety of scales, it is impossible to compare one bar from 1979 against any other bar in the chart, which makes the graphic quite useless.

In another example, we take a look at design variation in three dimensions. Figure 1.7

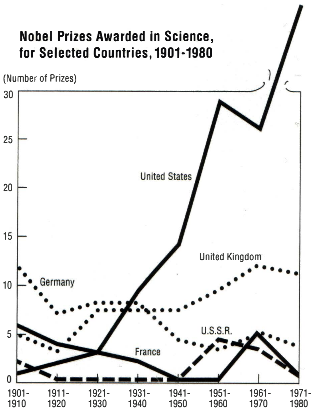


Figure 1.5: An integer graphic showing the information from Figure 1.4.

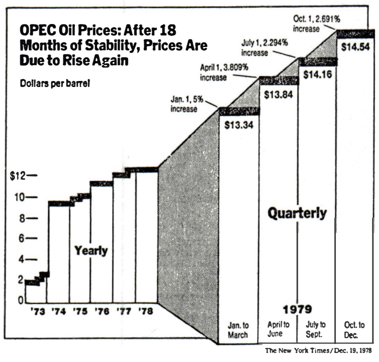


Figure 1.6: *New York Times*, December 19, 1978, p. D-7.

During this time	one vertical inch equals
1973–1978	$8.00
Jan–Mar 1979	$4.73
Apr–Jun 1979	$4.37
Jul–Aug 1979	$4.16
Oct–Dec 1979	$3.92

During this time	one horizontal inch equals
1973–1978	3.8 years
1979	0.57 years

Table 1.1: An overview on the different scales in Figure 1.6.

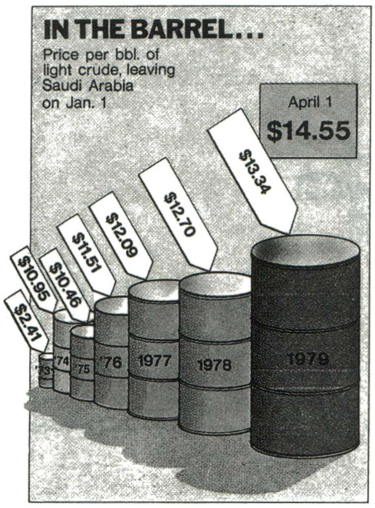
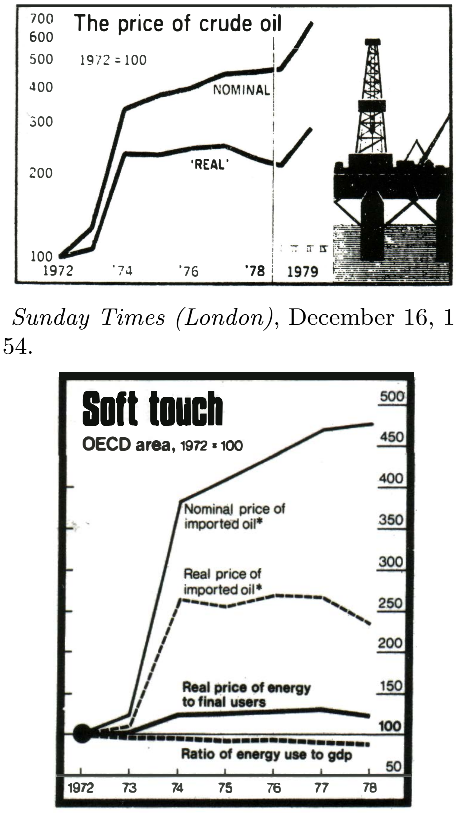


Figure 1.7: *Time*, April 9, 1979, p. 57.

visualizes the same oil prices as discussed before using oil barrels. While the barrel representing the first year appears to be far away from the viewer, the most recent barrel is depicted very close. The usage of perspective is a kind of design variation, resulting in a lie factor of 9.4.

In addition to the examples shown, some newspapers and magazines created good and integer graphics showing the oil price in the 1970s, as in Figure 1.8. We can see that not only the nominal oil price is given, but also the real price adjusted for inflation. From Figure 1.8a, we can see that the real oil price was decreasing from 1974 to 1978, while Figure 1.6 shows increasing bars during that time, making it impossible for the audience to recognize this fact. The example of inflation points out that in some cases, we are interested in removing not only design variation, but also some ways of data variation, from our graphics in order to sustain graphical integrity.



(a) *Sunday Times (London)*, December 16, 1979. p. 54.

(b) *The Economist*, December 29, 1979, p. 41.

Figure 1.8: Good visualizations of oil prices.

1.3 The Case of Skyrocketing Government Spending

Before we will discuss the effect of inflation in the following example graphic from Figure 1.9, we will take a look at some graphical gimmicks that secretly let the values look more dramatically than they actually are.

In the example, we see the increasing total budget expenditures of the State of New York from the fiscals 1966 to 1976. In the bottom right corner, two arrows are pointing upwards at the corresponding bars to indicate that they contain "estimated" and "recommended" values. Although those words could have easily been used without the arrows, the chart designer leads the viewers' attention towards the largest two bars in the graphic and thereby creates a certain impression as if the latest expenditures were larger than they actually are and weightier than those before.

To find the next grapical gimmick, we need to pay attention to the three-dimensional design of the graphic. When looking closely at the top of the fiscal 1974's bar, it appears that this bar, together with the bars to its right, is in front of the bars left to itself. Again, this creates an impression as if the corresponding expenditures, being the highest in the graphic, were somewhat more important.

The last design element distracting the viewers' attention and concentration is the three-dimensional design itself, which creates a very fuzzy and restless graphic. When removing the perspective elements as in Figure 1.10, a much calmer chart emerges.

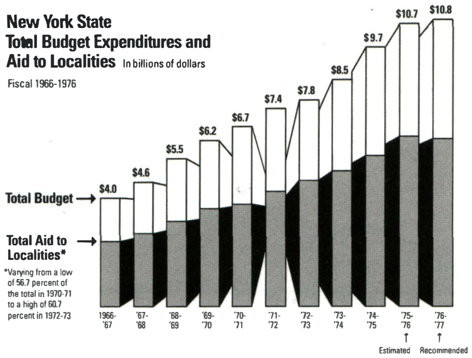


Figure 1.9: *New York Times*, February 1, 1976, p. iv-6.

All of the aforementioned graphical techniques lead the viewers' attention away from the underlying numbers and create a dramatizing effect, overstating the "skyrocketing" increase in government spendings.

Apart from the graphical distractions, two unwanted effects of data variation are part of the chart. The first one, as introduced before, is inflation. 1 US Dollar in 1966 has about the same worth as 2.03 US Dollar in 1977.

Another effect to consider is the growth of the population, as more people using a state's infrastructure or public services justify larger government expenditures. In the State of New York, the population increased by about 10 percent in the corresponding years.

We can take both effects into account by visualizing the per capita budget expenditures in constant dollars, as shown in Figure 1.11. It can be seen that, apart from a 20 percent increase by 1970, the spendings were almost constant, remaining within a 5 percent interval until 1977; a result fundamentally different to the first impression from Figure 1.9.

Therefore, to enhance graphical integrity, the following principle is to be obeyed:

> "In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units." [1]

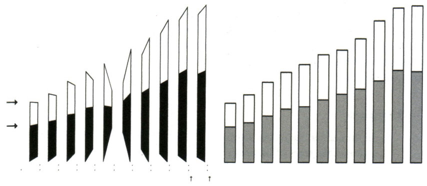


Figure 1.10: *New York Times*, February 1, 1976, p. iv-6, edited by Tufte.

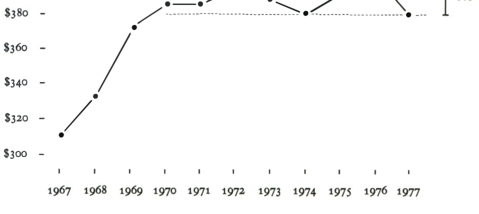


Figure 1.11: An integer graphic showing the information from Figure 1.9.

1.4 Visual Area and Numerical Measure

As already discussed in section 1.1, several problems can occur when numbers are visualized using areas of figures. This holds particularly true when the visualization is done wrong. The core problem is that in most cases, one-dimensional data is presented using two or more dimensions. When doing so, the area needs to be proportional to the data it represents, as concluded in section 1.1. Very often though, the width and height of an area are scaled linearly with the data, so the area increases quadratically, leading to misrepresented data.

When looking at Figure 1.12, we see that the size of the doctor is supposed to be consistent with the "percentage of doctors devoted solely to family practice". When we compare the numbers from 1990 and 1964, we can say that the doctor on the left should be 2.25 times larger than the right hand side doctor. However, it can be seen with the naked eye that the difference in height of the doctors already roughly corresponds to the factor of 2.25, so the difference in area is close to a factor of 5. As the human eye pays attention to the area of a depicted object rather than to its height, the 1964 doctor also appears to be far larger in comparison to his 1990 colleague than he should be. In our example, a lie factor of 2.8 results from the falsely scaled doctors.

The same mistake that has been done in the previous example is even more serious when it is done in three dimensions. Let us take a look at the example from Figure 1.7 again. Its lie factor of 9.4 that has been calculated in section 1.2 is based on the assumption that the viewer is supposed to interpret the printed area of an oil barrel as a measure for the underlying data. But if we take the three-dimensional depiction seriously and assume that the volume of a barrel shall correspond to the data, an astounding increase of 27000 percent can be calculated between the smallest and the largest barrel, whereas the underlying numbers increase by only 454 percent. Following that, a lie factor of 59.4 can be calculated, which must be a record.

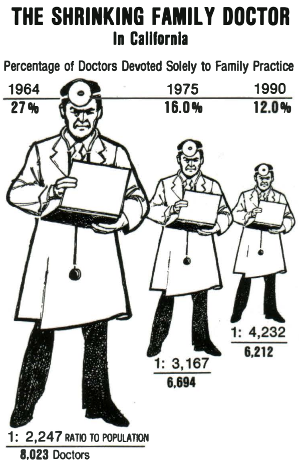

THE SHRINKING FAMILY DOCTOR
In California
Percentage of Doctors Devoted Solely to Family Practice
1964
27%
1975
16.0%
1990
12.0%
1: 2,247 RATIO TO POPULATION
8,023 Doctors
1: 3,167
6,694
1: 4,232
6,212

Figure 1.12: *Los Angeles Times*, August 5, 1979, p. 3.

As we have already seen in section 1.1, large differences occur in how people perceive the size of areas and different impressions may arise from the same graphic. Following that, showing one-dimensional information using two or even more dimensions is an inefficient technique with a large perception bias which is also often done wrong. To enhance graphical integrity here, we need to follow the principle

> "The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data." [1]

However, not all two-dimensional representations of data are misleading. With Figure 1.13,

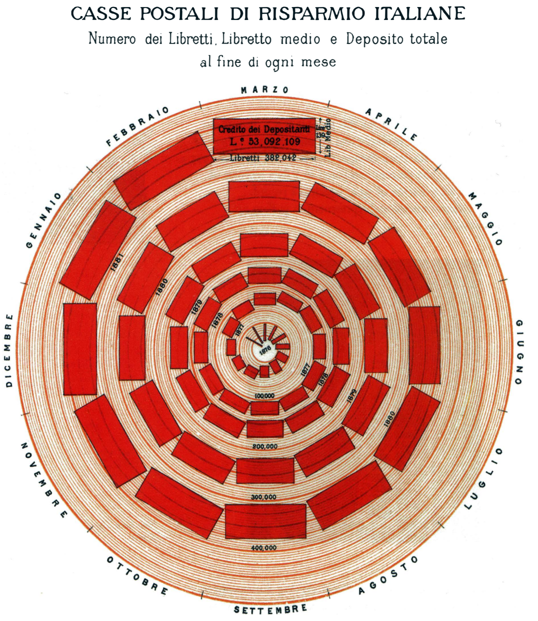


Figure 1.13: Antonio Gabaglio, *Teoria Generale della Statistica* (Milan, second edition, 1888).

we have a good example of using two dimensions for data visualization. The rectangles show both the number and average deposit value of postal savings books in Italy from different months in the late 19th century. The area of the rectangles thereby show the total deposit of all postal savings books. In this example, two dimensions are used to visualize two-dimensional data, in accordance with the principle above.

1.5 Context is Essential for Graphical Integrity

Quite often, graphics which seek to influence the opinion of the audience do not show false information, but omit certain information such that only information that the chart designer likes is presented.

Figure 1.14 shows the traffic deaths in the state of Conneticut at two points in time, 1955 and 1956, and provides the information that in 1956 the police started to prosecute drivers exceeding speed limits more strictly. The data point from 1955 is larger than the point from 1956, showing a decrease in traffic deaths. At first sight, one might conclude that the stricter police enforcement is responsible for the traffic death decline and therefore was quite successful. But without any knowledge about traffic deaths in other years, one cannot be certain that the decrease is significant and not similar to other changes in the years before or after.

To help us interpret the context of certain information, Figure 1.15 provides us with three possible context data sets for our two data points. Very different interpretations arise from the different scenarios.

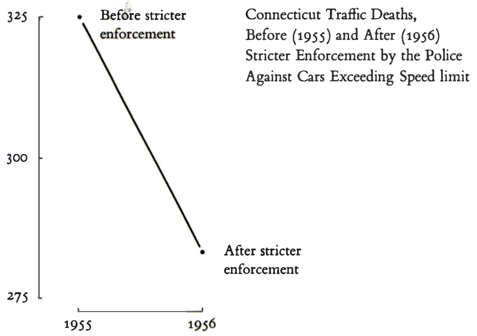


Figure 1.14: Traffic deaths in Conneticut from 1955 and 1956.

In the first scenario, a increase or decrease by the same amount as in Figure 1.14 occurs every year and the change would not be significant, so the police enforcement would not have caused the decline in traffic deaths.

The next scenario shows a peak—the traffic death toll increases in one year and decreases by the same amount in the following year. A possible reason for this could be a one-time event such as extreme weather conditions or a mass collision. Following that assumption, the numbers would go back to normal in the following year by default, with or without the stronger enforcement against speeding. Again, the police enforcement would not have caused the decline in traffic deaths.

However, the last scenario shows a constant high number of traffic deaths in the years up to 1955 and a constant low number of traffic deaths in the years after 1956. Here, the police enforcement might have caused the decline in traffic deaths, however, we have insufficient information on the matter to say that for sure.

Figure 1.16 shows the information from Figure 1.14 together with the traffic deaths from 1951 to 1959. From there, we can see that the decrease is of a similar magnitude to other years'. Also, the 1955 traffic deaths are the highest in this decade, so the scenario is similar to a peak as described before.

When we add even more data from the same time, but from different states, we see in Figure 1.17 that not only did the 1956 numbers decline in Conneticut, but also in three other states. It is therefore very likely that the declining traffic deaths have nothing to do with the stronger police enforcement in Conneticut,

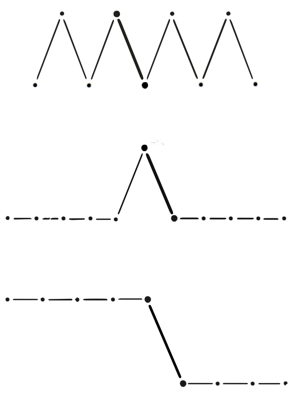


Figure 1.15: Contextual information for Figure 1.14.

but are provoked by some cause that affects more than one state.

Here, using the contextual data points in Figure 1.17, we were able to conclude what would not be possible from the two data points we started with. Therefore, when graphical integrity shall be preserved, we need to follow the principle that

> “Graphics must not quote data out of context.” [1]

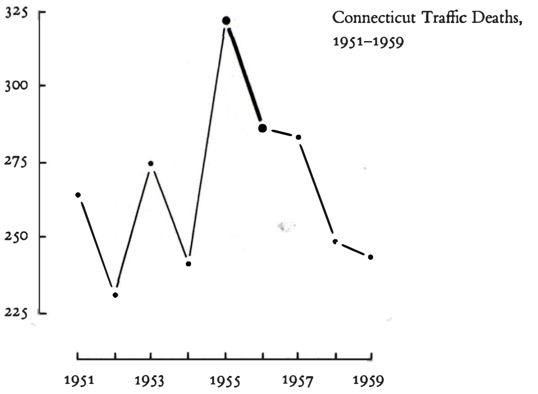


Figure 1.16: Traffic deaths in Conneticut from 1951 to 1959.

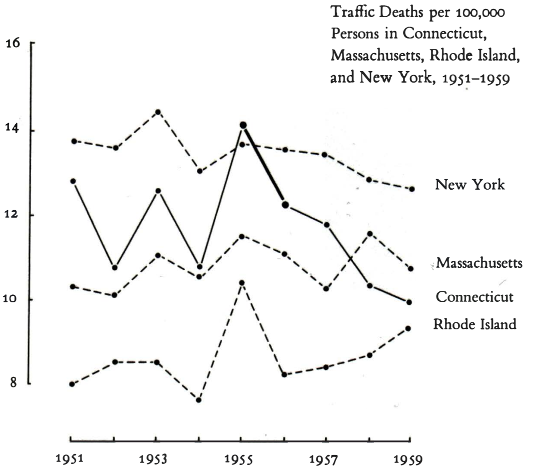


Figure 1.17: Donald T. Campbell and H. Laurence Ross, “The Connecticut Crackdown on Speeding: Time Series Data in Quasi-Experimental Analysis,” in Edward R. Tufte, ed., *The Quantitative Analysis of Social Problems* (Reading, Mass., 1970), 110-125.

Chapter 2

Sources of Graphical Integrity

In the previous chapter we have seen that even respected and well-known newspapers or magazines print graphics that lack graphical integrity and show a distorted picture of reality. So why do artists draw graphics that lie, and why do the world's major newspapers and magazines publish them? Do they really want to manipulate their readers picture of the world, do they use misleading graphics unintentionally, or are they just not interested in integer graphics?

In [1], three explanatory approaches are presented to answer the questions above.

The Lack of Quantitative Skills of Professional Artists Most designers that create charts and graphics for newspapers studied fine arts, so they are skilled to let graphics look appealing and interesting. However, artists usually have no experience with analyzing data and simply do not know how to achieve graphical integrity. Their goal is to create fancy graphics, not accurate graphics.

The Doctrine That Statistical Data Are Boring From this doctrine, it follows that graphics are used for the purpose of catching the reader's attention and being entertaining, rather than informing. A chart specialist of the Time magazine, who also was an art-school graduate, was quoted: "The challenge is to present statistics as a visual idea rather than a parade of numbers." From here, we can see that the designer had no intentions in graphical integrity, a consequence of letting designers create graphics and not statisticians, who know how to work with numbers and present them in an accurate way, or authors, who know the matter of the article the graphic belongs to and can put the numbers in a context.

The Doctrine That Graphics Are Only for the Unsophisticated Reader Some editors use graphics to entertain those people from an audience from whom it is assumed that they will not understand the information or the words in the articles. While the intelligent readers shall read the text, the less intelligent readers are supposed to get a brief idea of the matter from the graphic that is created in a way that makes it simple to understand the information presented. For this reason, overstating effects is a commonly used technique in order to emphasize certain circumstances. A publisher of a magazine intended for children, who are a good example for less sophisticated readers, once said that they "produced an article that was longer on graphics than on information. We had feared children might be overwhelmed by too many facts." Because of this fear, the publisher did not make any effort to create integer graphics, since the children will see no difference between those and non-integer graphics.

From the approaches above, [1] comes to the conclusion that

> "Graphical competence demands three quite different skills: the substantive, statistical, and artistic. Yet now most graphical work, particularly at news publications, is under the direction of but a single expertise—the artistic. Substantive and quantitative expertise must also participate in the design of data graphics, at least if statistical integrity and graphical sophistication are to be achieved." [1]

Bibliography

[1] Edward Rolf Tufte. *The Visual Display of Quantitative Information.* 2nd ed. Cheshire, Conneticut, U.S.A.: Graphics Press, 2001. Chap. 2–3, pp. 53–87. 197 pp. ISBN: 978-0-9613921-4-7.