

The logo for the Society for Technical Communication (STC), consisting of the letters 'STC' in a bold, yellow, sans-serif font inside a dark blue square.

STC

FEBRUARY 2012

# intercom

THE MAGAZINE OF THE SOCIETY FOR TECHNICAL COMMUNICATION

**UNDERSTANDING  
THE LIMITATIONS OF  
VISUALLY REPRESENTING  
THREE-DIMENSIONAL  
DATA ON SCATTER PLOTS 6**

**TECH COMM 2.0: REINVENTING  
OUR RELEVANCE IN THE 2000S**

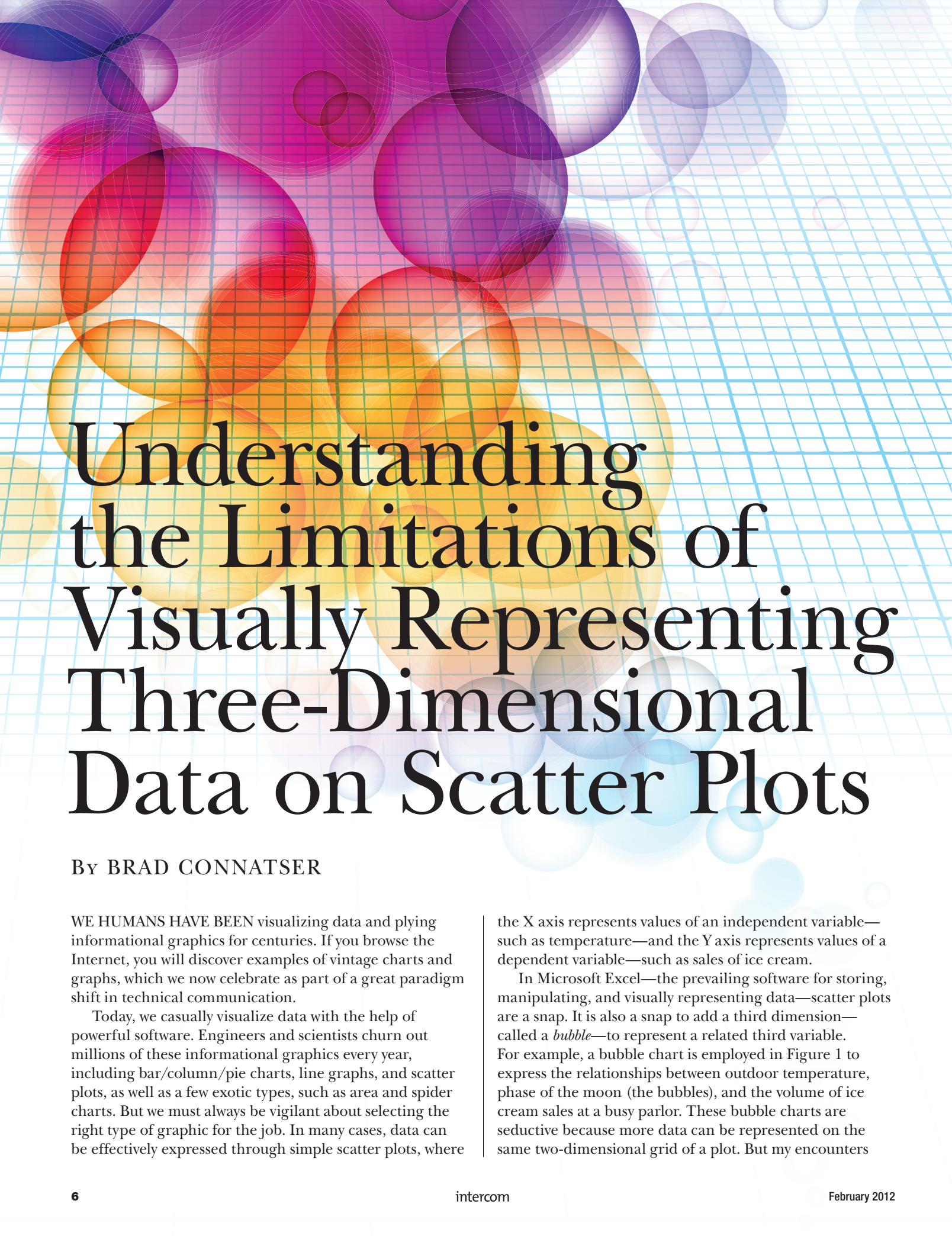
14

**LEADING FROM THE "WRITE"**

19

**REVISING THE REVIEW-  
AND-REVISION PROCESS**

22



# Understanding the Limitations of Visually Representing Three-Dimensional Data on Scatter Plots

BY BRAD CONNATSER

WE HUMANS HAVE BEEN visualizing data and plying informational graphics for centuries. If you browse the Internet, you will discover examples of vintage charts and graphs, which we now celebrate as part of a great paradigm shift in technical communication.

Today, we casually visualize data with the help of powerful software. Engineers and scientists churn out millions of these informational graphics every year, including bar/column/pie charts, line graphs, and scatter plots, as well as a few exotic types, such as area and spider charts. But we must always be vigilant about selecting the right type of graphic for the job. In many cases, data can be effectively expressed through simple scatter plots, where

the X axis represents values of an independent variable—such as temperature—and the Y axis represents values of a dependent variable—such as sales of ice cream.

In Microsoft Excel—the prevailing software for storing, manipulating, and visually representing data—scatter plots are a snap. It is also a snap to add a third dimension—called a *bubble*—to represent a related third variable. For example, a bubble chart is employed in Figure 1 to express the relationships between outdoor temperature, phase of the moon (the bubbles), and the volume of ice cream sales at a busy parlor. These bubble charts are seductive because more data can be represented on the same two-dimensional grid of a plot. But my encounters



with bubble charts have revealed that “data alchemists” are sacrificing clarity for a density that is erroneously assumed to enhance the economy of communication (fewer charts per message). These bubble charts are confounding and may lead authors of technical documents into deep, delusional waters. This article describes the dangers of indiscriminately employing bubble charts to visually represent data in three dimensions.

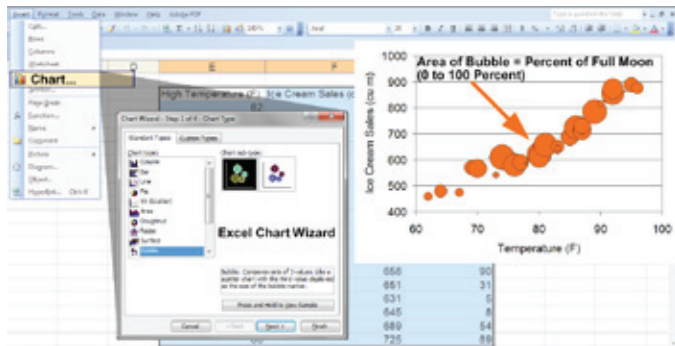


Figure 1. Can you perceive the correlation between phase of the moon and ice cream sales?

### The Efficacy of Scatter Plots

Informational graphics answer questions. For example, does a person’s height seem to affect income? If so, how much? Is the relationship positive or negative? Is it direct (a straight line, for example) or is it complex (a square function, for example)? Is it linear or exponential? In other words, what is the *correlation* between one variable and another?

Like other informational graphics—such as charts and graphs—plots are used to visualize data and express relationships between variables, sometimes vividly narrating a story. A plot can instantly crystallize what the text struggles to describe. In fact, often the spread of data points on a grid cannot be readily described in words, making a scatter plot indispensable to the reader’s understanding of the author’s message. Plots can animate lifeless data points that are stored in a table and reveal the very dependencies that the scientist or researcher is looking for. (Caution is required, of course, to not violate the analytic imperative, “Correlation does not imply causation.”)

Engineers and scientists frequently use informational graphics to convey numerical results of their research. For the most part, they do it masterfully using Microsoft Excel, which has become the chart builder for the everyman. Excel is powerful and makes graphical representation of data easy—perhaps too easy. For example, Excel can endow simple informational graphics with unnecessary bells and whistles, such as pretentious 3D representations of 2D data. Excel lets users craft even exotic charts with ease, such as radar/spider graphs, three-dimensional bar and column charts, and bubble charts.

In its most effective form, the humble scatter plot shows data points on a grid, and each data point represents the

values of two variables. Consider the scatter plot in Figure 2. Students at a welding school must pass their final VAMT test with a score of 600 or more (the range is 0 to 1400). The faculty believe that the number of study hours directly affects the score, and so they issued the exam with a final question that asked the test-taker how many hours he or she spent studying for the exam. Obviously, the data points shown in the figure reveal that as the hours of study increase, the VAMT scores increase.

Adding a trend line to the plot clarifies the X-Y relationship even more. The trend line shows how the two variables correlate (positive, negative, or no correlation at all). The “R-squared” value (correlation coefficient) answers the question: How much of this trend line accounts for the variation in Y? The closer R-squared is to 1.00, the better the trend line fits the data in the plot and the better the correlation between X and Y. (Sometimes, higher R-squared values can be achieved by fitting the plot with a polynomial equation, but that would just complicate the examples in this article.)

The answer to the faculty’s question is that hours of study account for about 75% of the variation in VAMT scores, which is a pretty good correlation. What accounts for the other 25%? There are many possibilities, such as students’ fibbing about the number of hours that they spent studying (self-reporting error), general test-taking ability, how much sleep that the student had the night before the test, and the student’s GPA.

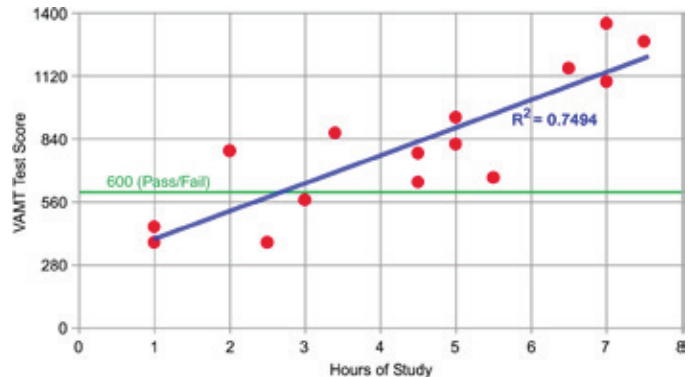


Figure 2. A simple scatter plot with a trend line and R-squared value

### The Seduction of Bubbles

In informational graphics, bubbles are not really bubble-like at all. They are simple circles, and each circle takes the place of a data point on a scatter plot. In a typical bubble chart, there is data underlying the bubbles: a grid with an X axis and Y axis. The complex meaning of a bubble is determined not only by its absolute position on this two-dimensional grid but also by its relative area (remember  $\pi r^2$ ?), which adds a third dimension to the scatter plot (Z).

Sometimes, bubble charts do work. Figure 3 shows what I consider an effective bubble chart, called a *bubble map* or *cartogram*. Each bubble represents the number of hours that an average person naps during a year. The reason

that this bubble map works is that there is no underlying data on the X-Y grid. The data points under the bubbles (representing the locations of U.S. cities) do not compete for your attention and processing resources. But even with a serviceable bubble chart—such as this one—you do not have the precision of absolute values, so you have to point out some critical values—such as the maximum and minimum values—or perhaps install a label on every bubble (this would be clumsy and produce clutter).

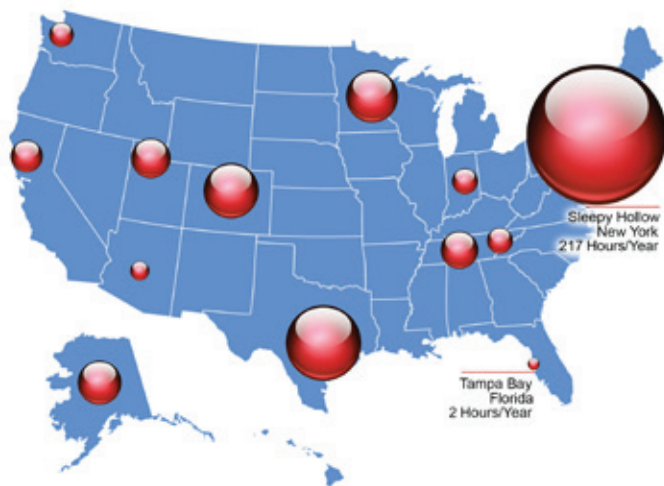


Figure 3. A bubble map indicating relative time spent napping per year per person

Bubble charts simply do not readily impart precise data and therefore do not lend themselves to science. They are better suited for show (in the manner of *USA Today*). In the napping bubble map (Figure 3), the bubbles are used to loosely represent data—that is, they impart a *notion*. If precision in the bubble map is required, a table of data points tying cities to hours can be added.

A more questionable case for bubbles follows: Consider an experimental test whose results seem to lend themselves to a bubble chart. Scientists are experimenting with a new rocket design and want to know the interactions between nozzle temperature (X), rocket acceleration (Y), and fuel temperature (Z). Instead of making multiple plots to visualize the resulting data, they decide to compress the results into a single bubble chart, with the areas of the data bubbles representing the values of fuel temperature. For reasons discussed here, this is not a sound decision.

If bubbles have such limited use in expressing quantitative data, then why use them? Engineers tend to compress information in both graphics and words. In their prose, they favor what I call a *portmanteau*, a heap of adjectives and nouns. Literally, a portmanteau is a large leather suitcase. When writers pack things into a suitcase, they squeeze out the little words that help the reader understand the relationships between the big words—the connective tissue of the prose, words such as “of,” “in,” and “for,” which are vital for fluid reading. The reader has to unpack this dense suitcase

to recover the lost connectives and the proper arrangement of phrases. For example, “employed custom power device energy shortfall makeup technology” can be more effectively expressed as “the technology that the custom power device employs to make up for energy shortfalls.” Oops. We increased the word count from 8 to 14 (kind of like breaking up a single bubble chart into several scatter plots), but now you can actually understand what the author is saying.

The compulsive linguistic thrift evident in portmanteaus extends to informational graphics in many ways, such as loading a graph with too many series of data points, endowing empty graph space with copious notes, and, of course, spawning dense bubble charts. But fewer words and fewer charts do not translate into “easier to read.”

To illustrate how this data compression works, let’s revisit our example scatter plot shown in Figure 2. What if the faculty of this welding school also tabulated each student’s GPA, with the notion that GPA probably positively influences VAMT test scores? Should the faculty employ the bubble chart to express that relationship? If they did, it would look something like the bubble chart shown in Figure 4. The areas of the bubbles vary from a GPA of 2.1 to 3.9. The full meaning of a bubble is determined not only by its position on the X-Y grid of the scatter plot (dimensions 1 and 2) but also by its size (dimension 3, area). Can you readily ascertain the relationship between GPA and the VAMT scores? Even when you are primed to expect a linear positive relationship—that is, GPA positively influences the test scores—can you see the relationship? Not only do these bubbles fail to translate into meaning, but they also obscure the X-Y plot pattern to some degree.

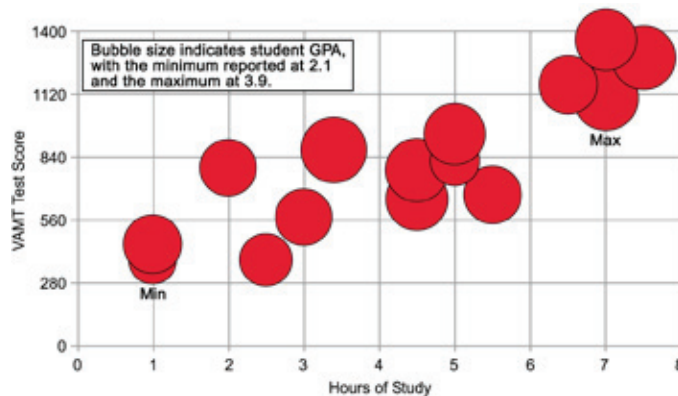


Figure 4. A scatter plot turned bubble chart with a few mouse clicks in Excel

Does adding a bubble dimension to an existing scatter plot make the plot richer? Or does it overcomplicate and obscure the message of the data? In the next section, I will describe in detail how bubbles can fail a well-intending author.

## The Troubles with Bubbles

Consider this authentic exchange between a technical writer and an engineer: The project engineer presented her research results to a roomful of colleagues, emphasizing

a few bubble charts. She expected feedback at one point in the presentation but was met with silence. After an awkward pause, the attendees nodded to indicate some level of understanding. The writer asked the engineer what she thought was going on during that silence. “Confusion,” suggested the writer, and the engineer agreed. Is it reasonable to drag the reader through a briar patch of confusion to conserve a few square inches of a technical report or reduce the slide count in a PowerPoint deck?

An author can make all sorts of facile claims with bubble charts—like reading tea leaves—because the audience is unlikely to challenge him or her. To read a bubble chart is to feel frustration, and to study one is to feel inadequate. The adorable term *bubble* may be appealing, but it connotes a playfulness that is belied by the complexity of bubbles. In a while, you will see how the suds floating on a scatter grid are dishwasher murky and interfere with the X-Y message that is so beautifully conveyed by unadorned scatter plots.

So far, my general censure of the bubble chart has not included any failure mechanisms that one can point to as solid reasons to refrain from using bubble charts. Bubble charts do indeed fail in several ways. Below I discuss four failure modes in particular.

### 1. Absolute Values

Imagine—or better yet, garner from Figure 4—that each bubble in a bubble chart is a data point on an underlying X-Y grid. A bubble’s location on this grid is determined easily enough by the cross-section between the X-axis value and the Y-axis value. Yet how does one determine the size of a bubble? The minimum and maximum circle areas are marked, but what about the other bubbles? Consider the bubble at VAMT 570 and 3 hours of study. It is bigger than the minimum (2.1) and smaller than the maximum (3.9), but is it helpful to know that it is merely somewhere between 2.1 and 3.9? (By the way, the value of that bubble is 2.9.) This lack of absolute values makes bubble charts a poor candidate for scientific communication.

One can remedy this vagueness by labeling each bubble with its absolute value, but labels clutter the plot, and if you are going to label each bubble, you might as well put the data in a table. Lots of overlaid text will certainly diminish the visual impact of an informational graphic.

### 2. Relative Values

Now imagine that you really do not care that much about the absolute values of the bubbles. You want your audience to compare the circle areas to get an overall feel about them. Back to Figure 4: As X increases, what happens to the bubble sizes? What about the correlation with Y? As with anything that involves psychology, there is complexity at play here that constrains the straightforward interpretation of the bubbles.

First, there is a stress between the comparison of the bubbles’ diameters and the comparison of their areas. In

fact, you can tell Excel whether you want a simple diameter to represent the bubble value or a square function to represent the value. The “circular logic” that you choose greatly affects the relative sizes of the bubbles. As shown in Figure 5, when “Size Represents Area” is selected, the bubbles are significantly larger than when “Size Represents Diameter” is selected, but not in a linear way. The size of the smallest bubble changes significantly, whereas the size of the biggest bubble does not change at all. Which ratio of large-to-small bubble is right? A 1.36-to-1 ratio of maximum to minimum or a 1.81-to-1 ratio? (Spoiler: The area of a circle is proportional to the square of its radius, and this square function requires you to select “Size Represents Area” for fidelity.)

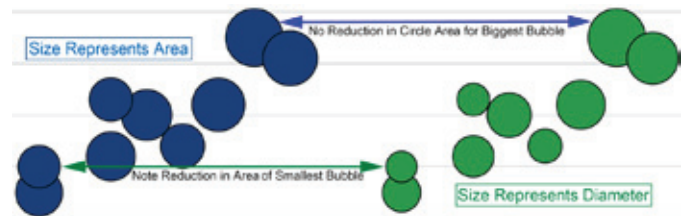


Figure 5. Microsoft Excel enables you to equate the size of a bubble to its tabular value directly (as the width of the bubble) or through a square function (as the area of the bubble)

Even if you have good contrast between the sizes of the bubbles (the smallest ones are much smaller than the biggest ones), the visual information must still pass through psychological filters that can trip up the comparator in your brain and distort relative size. One bubble might seem to have a value of 400 on one side of the chart but seem to have a value of 600 on the other side.

The Ebbinghaus illusion, shown in Figure 6, vividly demonstrates this complication. Because of its environment, the center circle on the left looks much bigger than the center circle on the right, even though they are identical in size. Human perception is saddled with all kinds of eye-brain responses for size, location, color, pattern, texture, and so on. Without gridlines to maintain scale and parallelism, we are subject to the uncertainty of squishy relativism.

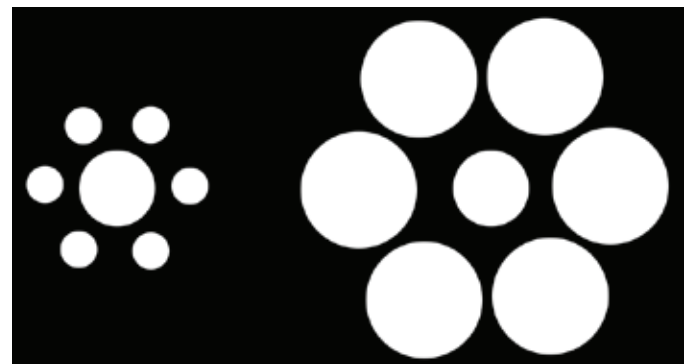


Figure 6. The Ebbinghaus illusion, illustrating how the environment of a circle can influence the perception of relative size



### 3. Eclipsing

You do not need celestial objects to experience an eclipse. Just go to Google Images, type in “bubble chart,” and you will find examples. Compared to the other three failure modes, this one is relatively minor because it can be manually ameliorated. Without such intervention, eclipsing can hide data or distort the value of data. For example, one bubble might overlay another bubble and distort the total areas of both, making them appear smaller or larger than they really are.

Eclipsing can be minimized by manually laying out the bubbles in an image processor such as Photoshop. Consider the bubble map in Figure 7, which contains significant eclipsing. Each bubble represents the population in a U.S. state capital. In the Northeast, some large bubbles would cover smaller ones if the larger ones were not purposively layered under smaller ones. But remember that this example was built by hand, not by an automated process in Excel. In Excel, eclipsing is not automatically controlled.

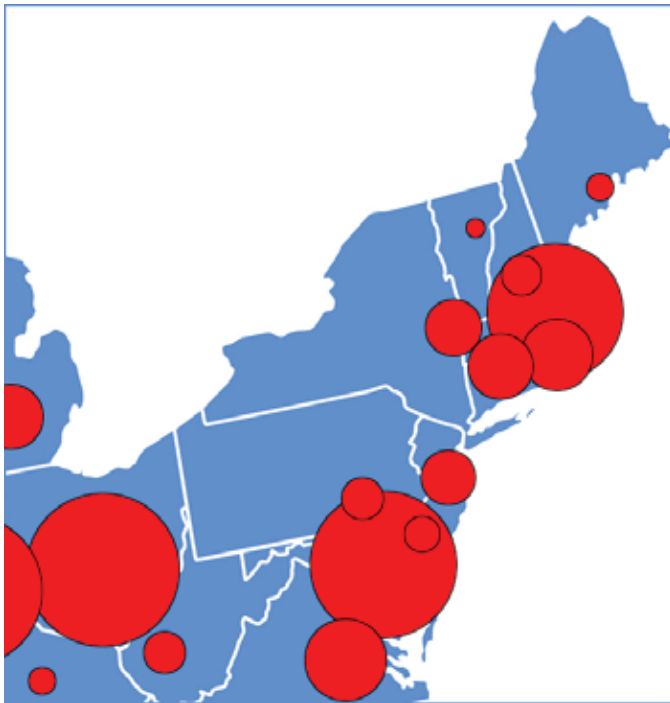


Figure 7. Eclipsing in a bubble map

### 4. Cognitive Loading

Bubble charts simply do not enable readers to understand multiple data concepts simultaneously. However, if devotees of bubble charts claim them for the sake of their readers, should people who know better take arms against a sea of bubbles? Should technical communicators tilt at the counterproductive density of bubble charts? To do so, one needs to bear not a sword but an unassailable logic. This last of the four failure modes of bubble charts, I believe, is the most powerful and cogent because there is

a large quantum of literature on the subjects of cognitive linguistics, perception, and cognitive psychology that apply to the interpretation of bubble charts.

The people who use bubble charts want—and sometimes really, really want—bubble charts to convey information in three dimensions and in multiple domains in the way that the many master strokes of Renoir convey the meaning of his painting. But bubbles and X-Y data do not coalesce in that way. You must parse the X-Y data points *or* the bubble areas, but not both. Sadly, the author of a bubble chart does not achieve the assumed effect.

Bubbles are process hogs in that they demand almost all of the reader’s cognitive load to parse them. You must leave the X-Y to focus on the bubbles. The reverse is just as true. Why is this so? I believe that the answer lies within the way that people categorize objects and concepts in their world. Our drive to categorize things is strong and compelling. Think of a category as a domain of thought or perception (George Lakoff’s *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind* is a wonderful study of this concept). We can generally recognize and process only one domain at a time. That is, the brain can have only one focal point at one time. Nothing, in my opinion, illustrates this concept better than the gestalt switch.

A gestalt switch presents an ambiguous image that contains two domains, such as the two shown in Figure 8. In one domain, you can see a chalice (white on black). In the second domain, you can see the profiles of two people facing each other (black on white). However, you cannot see both gestalts at once. Presented with two such perceptual tasks, the brain must switch between them.

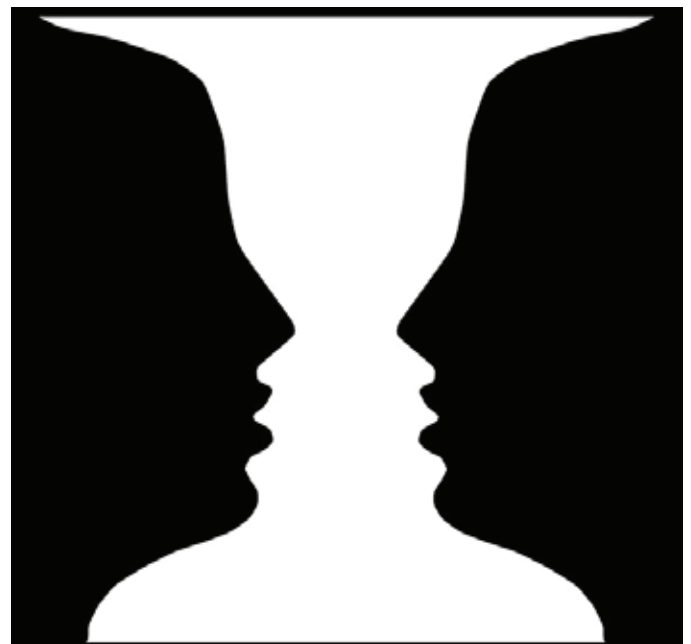


Figure 8. Classic example of a gestalt switch

The bubble chart is a gestalt switch. One domain is the X-Y scatter plot, and the other is the constellation of bubbles. Able to focus on only one domain at a time, the reader cannot perceive a bubble chart as one coherent message. What the author of the bubble chart wants to happen—some sort of intelligent, cross-domain coalescence—simply cannot happen. The bubbles and the scatter plot can coalesce no more than the facial profiles and the chalice in Figure 8 can coalesce into one super gestalt.

Perhaps someone will argue that the meaning of a bubble chart will come about through multitasking. After all, even children can multitask, wielding their various communication technologies concurrently—or so it seems. New science on multitasking—including research studies that involve brain scans of active subjects—may be dashing the unwarranted optimism of self-professed multitaskers. Multitasking is really serial-tasking at a fast rate of change. The brain’s ability to focus on two or more cognitive tasks at once is a youthful misinterpretation of what is happening in the brain.

Jon Hamilton, in his NPR story “Think You’re Multitasking? Think Again,” gently builds the background for why we “simply can’t focus on more than one thing at a time” ([www.npr.org/templates/story/story.php?storyId=95256794](http://www.npr.org/templates/story/story.php?storyId=95256794)). Similar tasks—such as interpreting a scatter plot and interpreting bubbles—compete to use the same part of the brain. It is similar to writing an email and talking on the phone—each language performance interferes with the other.

**Real-world Example**

This is the part of the article where I sigh and present to you the demoralizing evidence of my case. The bubble chart shown in Figure 9 is festive, like grandma’s Christmas tree, loaded with shiny colored lights but with no discernable patterns. The example bubble chart is analogous to a real, published chart, but the variables are fictionalized here.

Here is the source of the data. Seventy-nine subjects participated in a small pharmaceutical trial. All subjects were diagnosed with chronic, treatment-resistant depression and an accompanying weakness in the hands called DCO. Subjects were given either a new drug, the drug plus an herbal supplement, or a placebo. Previous trials indicated that the level of oxidation in the bloodstream negatively affected the size of relevant brain receptors (an indicator of clinical depression). Therefore, the level of oxidation in the bloodstream—along with receptor size and age of the subject—was also measured. Researchers wanted to know three things: 1) whether the experimental drug or drug/supplement combination affects receptor size, 2) whether age affects the efficacy of the drug or drug/supplement combination, and 3) as a second-order effect, whether oxidation in the bloodstream affects the receptor size.

The results of the trial are visualized in Figure 9. Can you see the message? What is this bubble chart trying to say? By convention, the X axis is the independent variable, and the Y axis is the dependent variable. So what is the Z axis (bubble area)? The researchers envisioned the Z variable (oxidation) as an independent variable that affects the Y variable (receptor size). So Z, in effect, is like X, an independent variable. However, I think that our instinct is to consider the X axis as the independent variable and consider the bubble size as a function of X. But that is not the case here. That is not the only confusing element of this real-world example. Ideally, a graphic answers one question, but this chart tries to answer six questions involving three series (treatment 1, treatment 2, and placebo) over three dimensions (X, Y, and Z) and two domains (X-Y scatter plot and bubble series). It’s a graphical portmanteau.

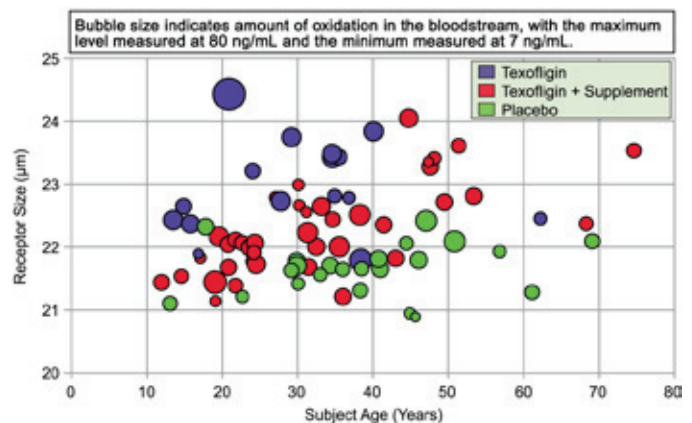


Figure 9. A real-world example of a multi-series bubble chart (also known as a mess)

Now, take a few moments to make sense of this chart. First, eliminate the blue series (texoflign) and the green series (placebo) to make things a little more digestible. Now, two questions remain: 1) Does age affect receptor size, and 2) does the size of the red circles (texoflign + supplement) affect receptor size? Let us focus on the bubbles (question 2). As the note atop the chart indicates, the ratio of big circle to little circle is 80 to 7 (about 11.5), but it appears to be greater than that. It is hard to estimate the relative areas of these bubbles, and even harder to correlate them to the Y axis.

The meaning of this chart certainly does not crystallize. If the reader has to study the graph and exquisitely parse it to create meaning, then perhaps there is another reader-friendly method of demonstrating relationships or lack of relationships. So, given the questions that these researchers want to answer, what is the best way to impart the correlation between the oxidation and receptor size?

## Revisualizing the Data

Like unpacking a portmanteau in words, we can unpack the graphical portmanteau in Figure 9 and create a single scatter plot that answers one question: Is there a relationship between oxidation in the bloodstream and the size of brain receptors?

Bursting the bubbles is relatively easy. First, locate the oxidation and receptor data (in a spreadsheet or table). Create a new scatter plot based on those two variables (sans bubbles). In Excel, this takes a few mouse clicks. The new plot has oxidation as its X axis (the independent variable) and receptor size as its Y axis (the dependent variable), shown in Figure 10. For even greater clarity, we can add a trend line plus an R-squared value. Now, instead of suggestive or notional data, we have data that is accountable to the grid—real numbers.

It turns out that oxidation accounts for about 17% of the variation in the receptor size, and the correlation is positive—a result that is opposite from the expected effect. The reader could never glean such an overt answer from bubbles.

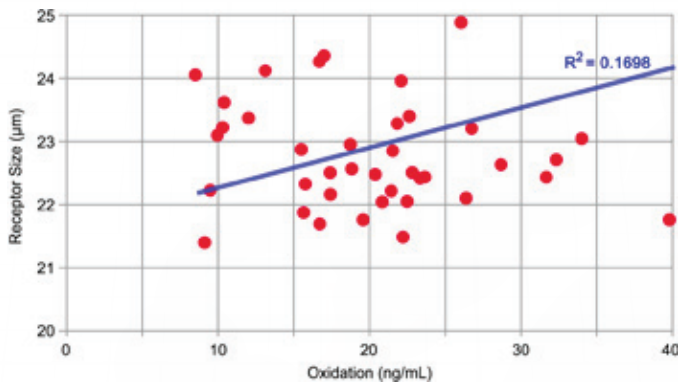


Figure 10. A bubble dimension separated from a bubble chart and visualized as a lucid scatter plot

In the case where the researchers want to visualize the relationship between oxidation (Z) and age of the subject (X), then the Y axis can be easily split—one on the left for receptor size and one on the right for oxidation. (However, this might overcomplicate the graphic.) The result, shown in Figure 11, indicates a modest correlation between age and receptor size but practically no correlation between age and oxidation.

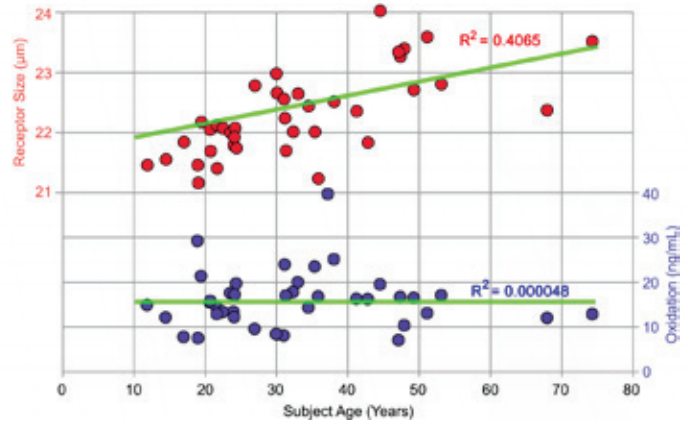


Figure 11. A bubble chart visualized as a scatter plot with two Y axes

## Conclusion

Perhaps in the future—when cars fly and the human brain has developed the ability to truly multitask in multiple domains—bubble charts will be the norm. But today is not that time. Bubbles may seem clever, but stamped out thoughtlessly in programs like Excel, they can be counterproductive to efforts to clearly visualize data.

Bubble charts have no place in the scientific method or in research that endeavors to quantify results. In fact, the cheerful employment of bubbles borders on superstition, in that the belief in their efficacy is not based on reason. Such misconstrued efficacy goes unchallenged.

It must be obvious by now that I am on a mission to prevent the abuse of the bubble. Bubbles in the wrong hands are nothing more than *chartjunk*, a term coined by Edward Tufte that refers to superfluous elements in informational graphics. I hope I have given the reader a vocabulary and sweet reason to defend a position against the use of bubble charts when quantitative data is at stake. Tell a bubble-lover today that the emperor wears no clothes. **f**

BRAD CONNATSER has been writing and editing technical documents for two decades. He has authored dozens of articles in trade magazines and peer-review journals. Brad taught technical communication and English composition at Temple University, Maryville College, and Pellissippi State. He served in various capacities in the STC East Tennessee Chapter.