Information Visualization, Design, and the Arts
Collision or Collaboration?

Stephen Few, Perceptual Edge
Visual Business Intelligence Newsletter
January 2010

For the last few years I've been watching, with both fascination and trepidation, a rise of interest in information visualization among artists and designers (especially technology-savvy graphic artists). These newcomers to the field, equipped with different perspectives and skills, have much to teach us. They will contribute useful innovations that more traditional practitioners might never imagine. Unfortunately, much that is being produced by our artistic colleagues doesn't qualify as information visualization, even when this is what they insist on calling it.

People who are talented in a particular discipline sometimes march a bit too confidently (perhaps arrogantly) into a different domain, convinced that they know more than traditional practitioners because they have skills that the old-timers lack. It is also true that traditional practitioners often resist help from the outside, which is every bit as arrogant. The result is a collision. But what we have here is an opportunity. We must find a way to make this intersection of expertise and perspectives a rewarding collaboration.

Information visualization has been around long enough to develop into a fairly well-defined discipline. The best-known definition was provided by Stuart Card, Jock Mackinlay, and Ben Shneiderman in their 1999 book *Readings in Information Visualization: Using Vision to Think:*

> *Information visualization is the use of computer-supported interactive visual representations of abstract data to amplify cognition.*

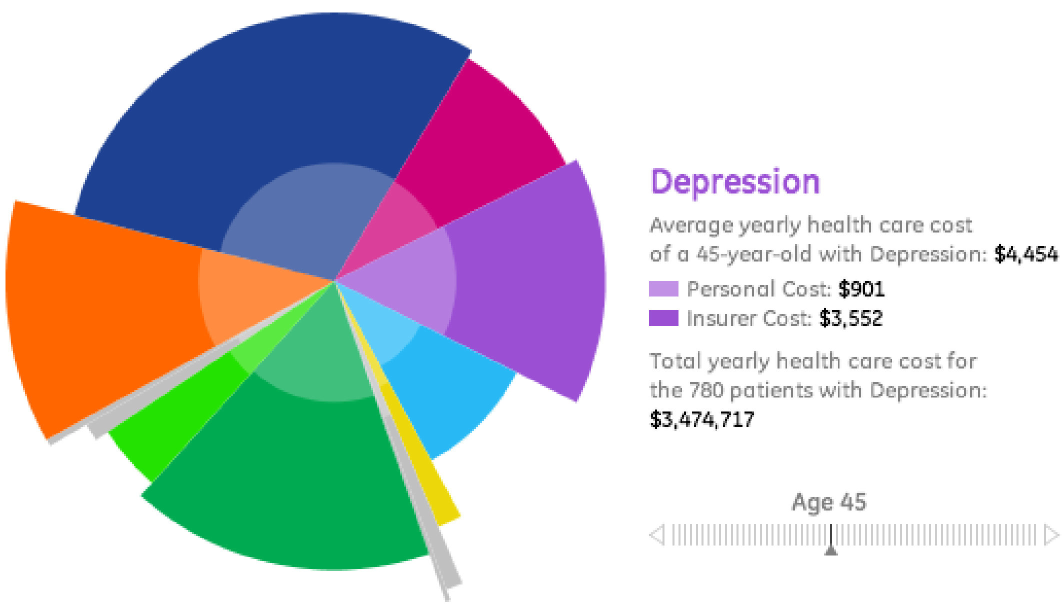
It is about helping people think more effectively about data, with the immediate goal of understanding, and the eventual goal of better decisions. I tend to expand this definition slightly by adding "communication" as a complementary purpose to the amplification of cognition, but I refrain from diluting the definition of "information visualization" by switching to the term "data visualization" when I do so. It isn't appropriate for people to use established terms like these to mean something different in purpose and effect. Something does not qualify as information visualization or data visualization simply because electronic data was used to produce it. When something artistic is produced based on electronic data, it can be worthwhile without leading to a better understanding of the stories contained in the underlying data, but it isn't information visualization. If it's genuinely worthwhile as art but doesn't promote clear and accurate comprehension of the data, let's call it something else–perhaps *information art*. Let's evaluate these works in the same way that we judge other forms of art. Is it beautiful? Does it move us? Does it challenge us? Does it make us think differently? If it does and it also helps us to accurately and usefully understand the data on which it was based, then it qualifies as both information art and information visualization.

Let me illustrate my concern with real examples. I wrote in my blog sometime back about the artistic work of Aaron Koblin of Google's Creative Labs. He uses the term "information visualization" to describe his work, but it rarely qualifies. Seldom does it reveal more than a vague sense of the underlying data. Another person who has attained notoriety for his visual renderings of data is Ben Fry of Seed Media Group, the author of *Visualizing Data*, who was trained in graphic design. Fry is talented, but he's using the term *data visualization* to describe something that's usually quite different in intention and effect–more artistic than informative.

Fry's projects usually visualize massive data sets. The resulting pictures tend to evoke a sense of "Wow, that's a lot of data" and a general message such as "This programming code includes lots of jumps and loops" (Dismap). Rarely do these pictures help us think clearly and thoroughly about the data. If a general sense is all we need, these visualizations succeed as art if viewers find them artful (not everyone does, but that's the nature of art). If we seek greater understanding, however, which is what we expect from information visualization, they leave us gawking and wanting. One exception is Fry's project "All Streets", which displays all the roads in the United States on a map, minus other geo-political features, to reveal a complex web of meaning. This ranges from the disparate distribution of roadways across the nation as a whole, with a clearly delineated and greater concentration in the eastern half, to specific geographical features that emerge from regions where roads are missing, on down to specific clusters of human transit in particular locations. This image is rich. It serves as a meaningful overview that could lead to deeper dives into the data. In this particular example, Fry shows a fine sense of graphical restraint, resulting in a display that allows our eyes to be drawn to meaningful information without distraction.

At times when Fry produces something that qualifies as data visualization in its intent, it fails in its execution. One recent example suggests that he skipped some of the curriculum during his data visualization training, especially regarding human perceptual and cognitive systems. I'm referring to an interactive visualization of medical costs that was commissioned by GE Healthcare. If GE wants to give people a sense that they've learned something about healthcare costs without actually informing them, then Fry's visualization delivers. Even though it's not a stretch to imagine some healthcare providers setting out to obfuscate rather than inform us about medical costs, I really doubt that this was GE's intention.

Let's examine, critique, and improve Fry's visualization of healthcare costs to remind ourselves of the principles of data visualization that he has either never learned or has chosen to ignore. Here's his display of this important information:



[Click to link to the original animated display.]

This visualization attempts to display five variables in an unusual pie chart: 1) chronic diseases, 2) patient's ages, 3) the number of patients of each age that were treated for each disease, 4) the average cost per patient for treatment per age and disease, and 5) a breakdown of these costs into the patient's portion and the insurer's portion. Even when pie charts are used as intended to show two variables–names of the parts of some whole and their corresponding percentages–they do so in a way that is hard to decode. When you add three more variables, as Fry has in this example, a pie chart morphs into utter absurdity.

Each slice's color represents a specific chronic disease. To know what a particular color represents, you can either hover over it with your mouse and read the name that appears, or you can click the slice, which causes it to rotate to the right where "Depression" appears in the example above. Once selected, you can read the disease's name along with several values in the section of text that appears to the right of the slice.

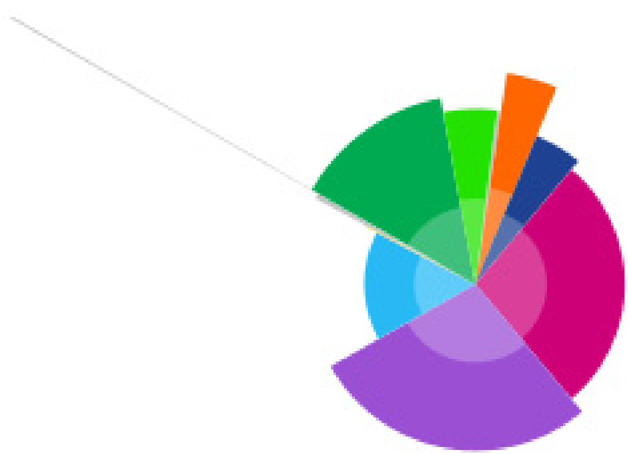
Although it isn't obvious, or even easy discern with time and effort, the pie chart above is divided into 12 slices. Because the purple slice has been selected in this example, as well as age 45 using the slider control on the bottom right, we know that 45 year olds on average pay $901 (the light shade of purple) in annual healthcare costs for the treatment of depression, which, when combined with the $3,552 per person cost paid by insurers (the darker shade of purple), comes to $4,454 per person in total treatment costs.

Notice that the slices extend various distances from the center of the pie. Why do they behave in this abnormal way? What do these distances represent? Is it meaningful to compare the areas of the slices, as we would do with a regular pie chart? How about the angles formed where the slices meet in the center? I played with the chart on GE's website for awhile and couldn't figure out how to read it. Giving up on any obvious interpretation, I looked for instructions. Here's what I read as an introduction to the visualization:

> *To gain a deeper understanding of healthcare costs, we've combined the Medical Expenditure Panel Survey (MEPS) with 500K records from GE's proprietary database. By combining MEPS with GE's data, we gain a more complete picture of the costs associated with chronic conditions.*

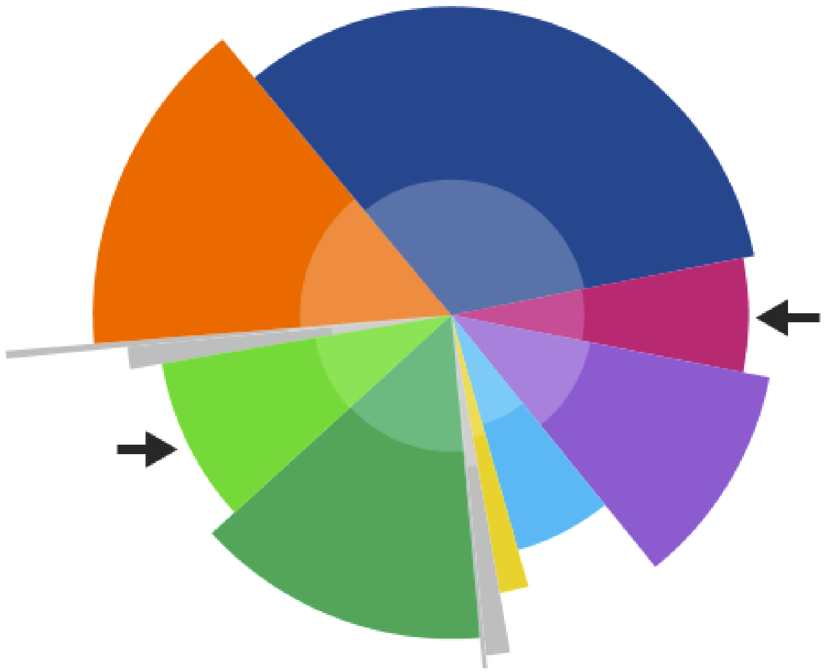
Well, that didn't help. I then found a video on Seed Media Group's website that features Fry talking about his work, including this chart in particular, but it didn't provide any clues either. I decided to see if another set of eyes would do better, so I asked my wife Jayne, a neuropsychologist, to give it a shot. She sat there motionless and apparently dumbfounded for a minute, so I prompted her by saying that she could click on the chart to get more information. For the next few seconds she clicked various slices, watched them spin around, and then made the following declaration: "This is fun and colorful." I then asked, somewhat frustrated, "But can you make sense of it," to which she replied, "Not at all." I suspect that Jayne's experience is typical.

I persevered, eventually clicking from slice to slice for awhile, reading the values that appeared to the right of the chart, trying to make sense of the angles, the distances from the center, and the areas of the slices. After testing several hypotheses, I eventually figured it out–or at least I think so. The angle of each slice represents the number of patients in the selected age group that were treated for the disease and the radius of each slice (the distance from the center to the outer edge of the slice) represents the average cost of treating a single patient in the selected age group for the disease. This approach produces some rather oddly shaped pies, such as the following:



The light gray slice that is trying to escape on the left represents a single 20 year old patient who was treated for gastric ulcers. Because the cost of treatment totaled $29,182, the thin gray slice had to grow extraordinarily long to represent that relatively large amount.

So, if you persevere as I did and manage to compare the angles of the slices while ignoring their areas, you can get a vague sense of the percentage of people that were treated for each disease. Then, if you can compare the radius of each slice, you can get a vague sense of the cost per patient for treating each disease. Despite perseverance, your effort would produce unreliable results because human visual perception isn't good at either. If 50 people were asked to compare the same set of angles or radii, their guesses would differ dramatically. I know, because I do this little experiment regularly in my courses. Try it yourself. Using the example below, compare the radii of light green and pink slices (those marked with arrows) to determine which represents the greater per person cost of treatment.



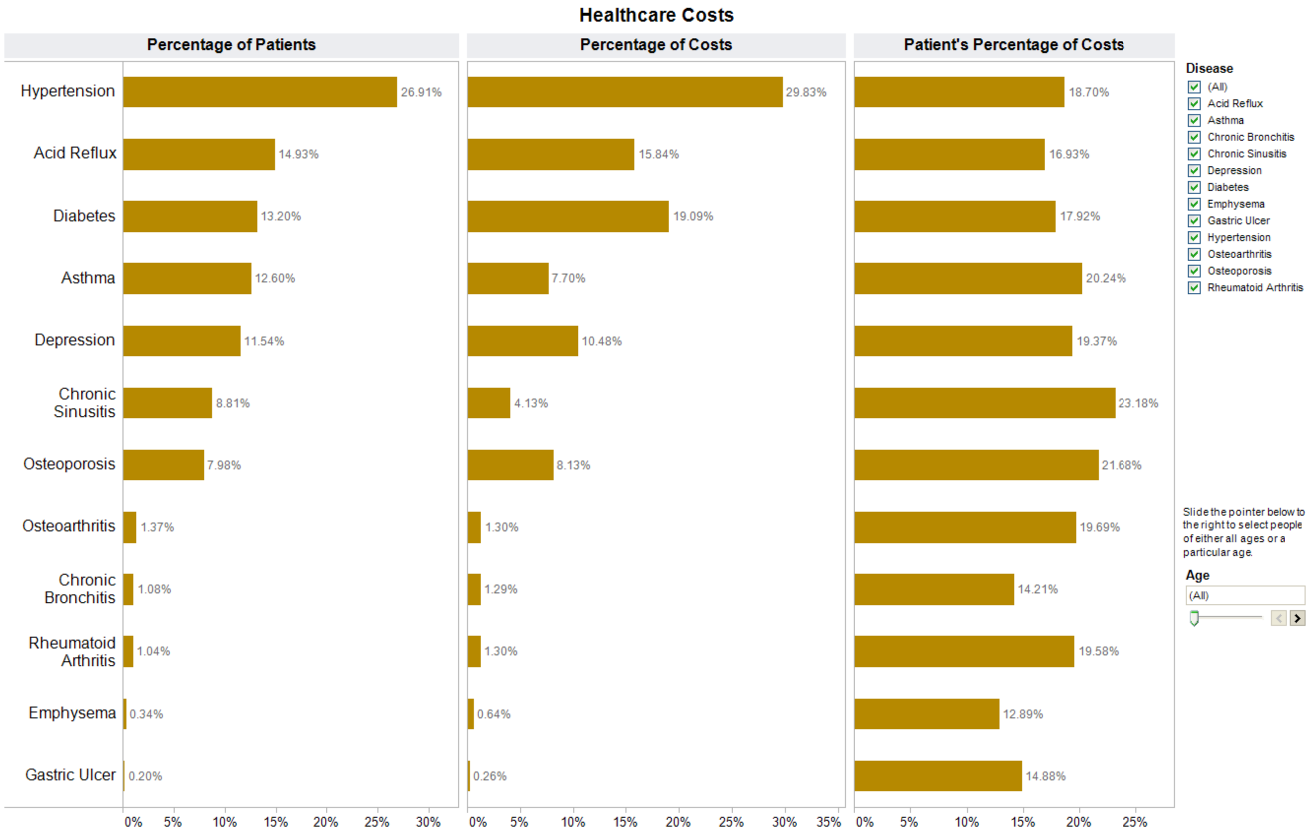
If you answered that the light green slice is greater in value, you guessed wrong. Here are the actual values:

Slice	Treatment Cost
Light green	$4,547
Pink	$4,670

What hope would anyone have of reading this chart? Even with the instructions that I've provided for deciphering it, at best someone who's really interested could click on one slice at a time, read the values that appear as text to the right of the chart, enter them into Excel or some other software, and then construct a new display that actually works.

Because I'm personally interested in healthcare costs, I was determined to examine this information in an effective way. Rather than reconstructing the data one slice at a time, I downloaded the application from the Web, extracted the data from the programming code, loaded it in Excel, restructured it in a way that could be used for analysis, and then read the file into Tableau. Using displays that I constructed in Tableau, I'll illustrate how GE could be informing the public about the costs of healthcare.

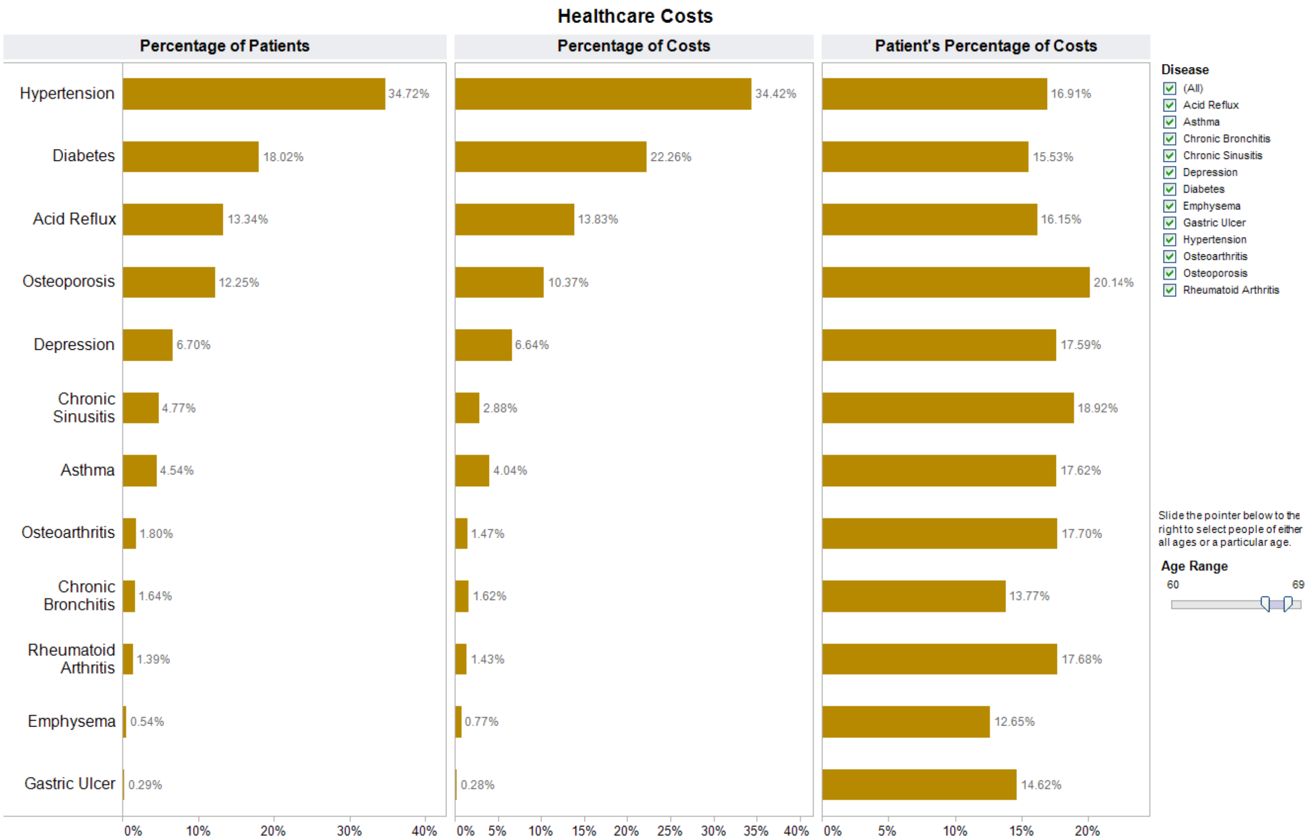
I'll begin with a simple example and then build to one that's more useful for exploring and making sense of these healthcare costs. Pie charts–even those that are designed normally–are hard to use because they force us to compare angles and areas, which we can't do very well. The two most effective ways to represent quantities graphically involve encoding them as positions on a 2-D surface, such as data points along a line, or encoding them as lengths of simple objects such as bars. The display on the next page provides a simple and effective way to compare the number of patients, percentage of costs, and the patients' portion of the costs for each of the 12 chronic diseases.



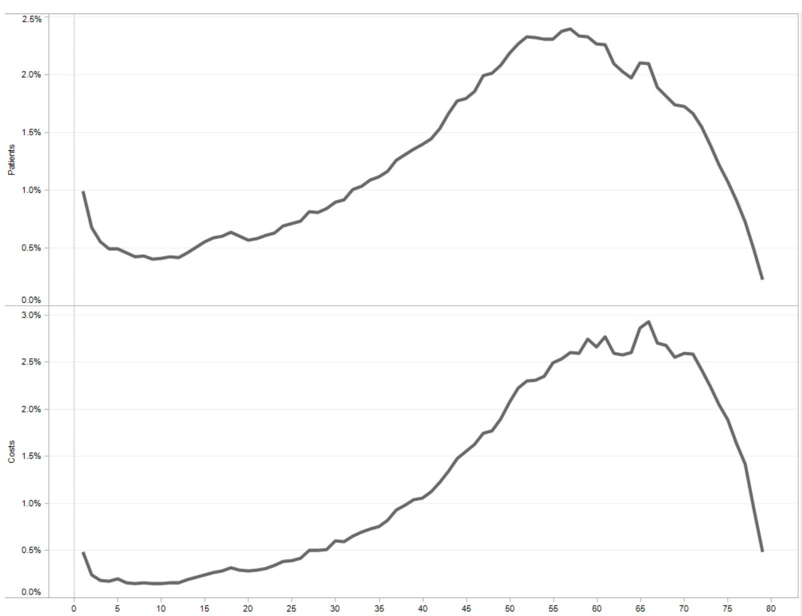
The “Age” selection control near the bottom right allows us to see either the total of all ages (the current selection) or one age at a time. Here are a few aspects of this design that make it meaningful and easy to use:

- Bars are used to graphically represent the values, which makes them quick and easy to accurately compare.
- Each set of values is expressed in terms that are relevant and easy to understand.
  - Patients are expressed as percentages rather than counts, because the counts are based on a sample of around half a million patients, which isn’t particularly relevant to the public.
  - Costs are expressed as percentages rather than dollar amounts, once again because the total dollar amount is for a sample of patients only.
- In addition to seeing one age at a time, values for all ages combined can now be seen.
- Diseases that are not of interest can be easily removed from the display by unchecking their boxes on the upper right.
- The fact that costs for treatment differ among these diseases is easy to see because the diseases are sorted in order of the number of patients from high to low, but costs associated with those diseases clearly don’t exhibit the same high to low sequence. Notice, for instance, that the cost of treating diabetes is significantly greater than that of acid reflux, even though fewer people suffer from the disease.
- The fact that the average portion of total treatment costs that patients pay differs from disease to disease is easy to see, ranging from 12% to 23%.

One of the major limitations of this design is that you can only look at either all ages combined or one age at a time, but never at a specified range of ages, such as patients in their 60s. By changing the age selection control to one that allows ranges to be specified, as I’ve done in the next example, we can now look at all ages combined, an individual age, or a specified range of ages. In the example below, I've selected patients from 60 to 69 years old.

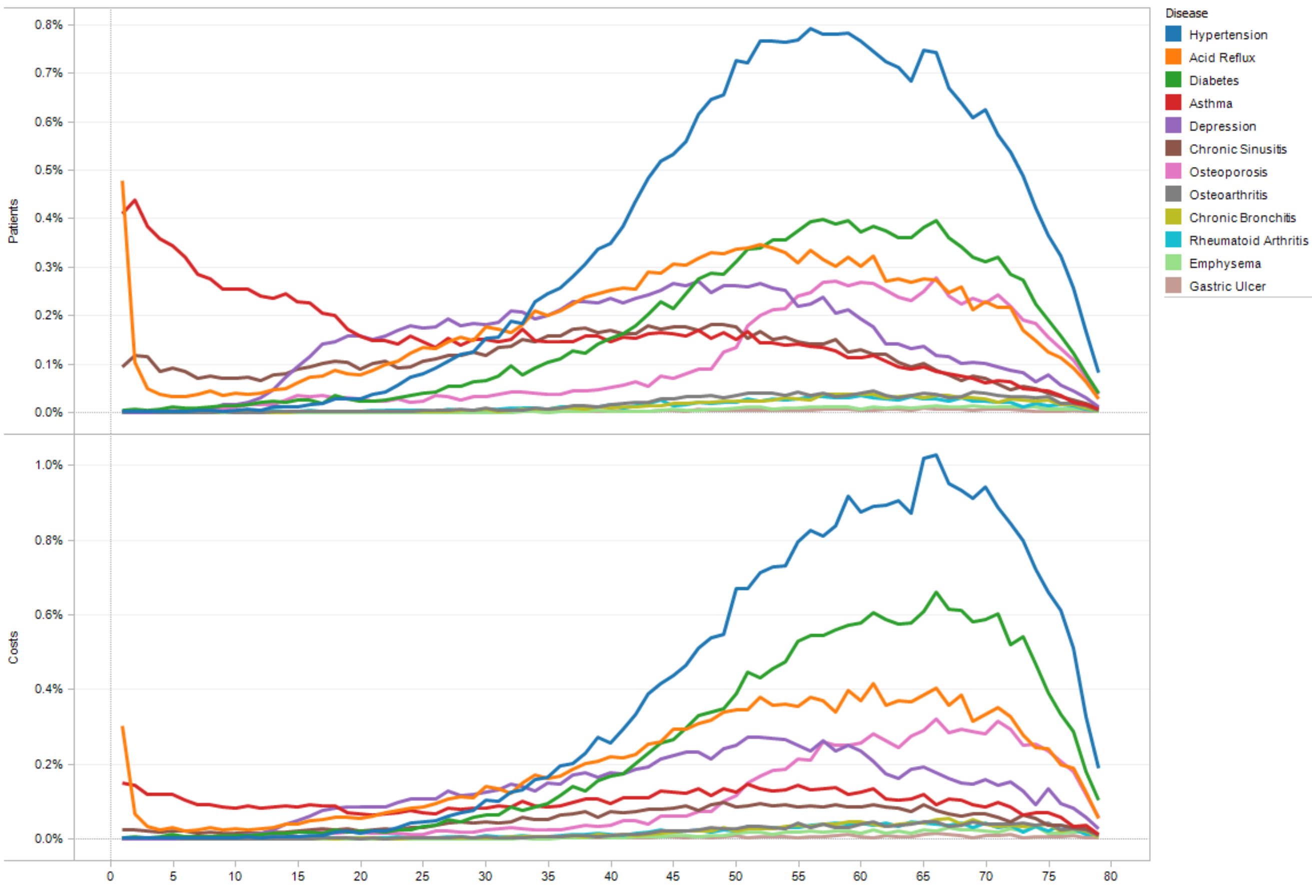


Using the two displays that I've shown so far, we would struggle to see how these diseases affect various age groups differently, because we can only see one set of values for the selected age group. For instance, you probably haven't noticed that Acid Reflux, which is second only to hypertension in the number of people that it affects overall, fell to third place when we narrowed the age range to people in their 60s. It would be useful to see the number of patients and healthcare costs associated with each age at once. One way this can be done is by using line graphs, as illustrated below.



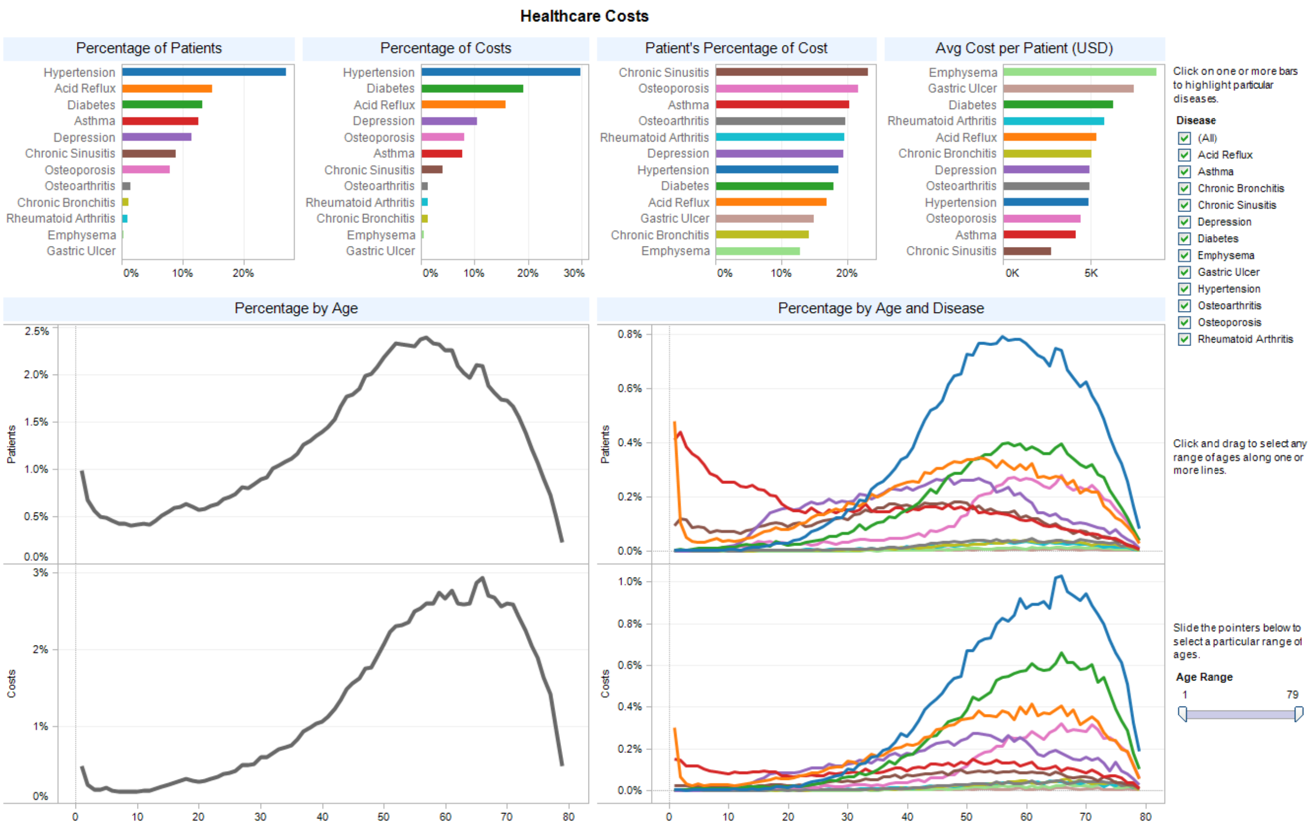
Looking at these two graphs, we can now see that, excluding one year olds, there is a fairly consistent rise in the number of patients and costs associated with these diseases as age increases that peaks around age 57 for the number of patients and age 66 for costs, and then falls off sharply until age 79. We can also see that there is a noticeable rise in both the number of patients and costs for ages 65 and 66 following a decrease from ages 62 through 64. These patterns weren't visible using the previous displays.

The story gets more interesting when we break these age-related patterns down by disease, which I've done in the following example.

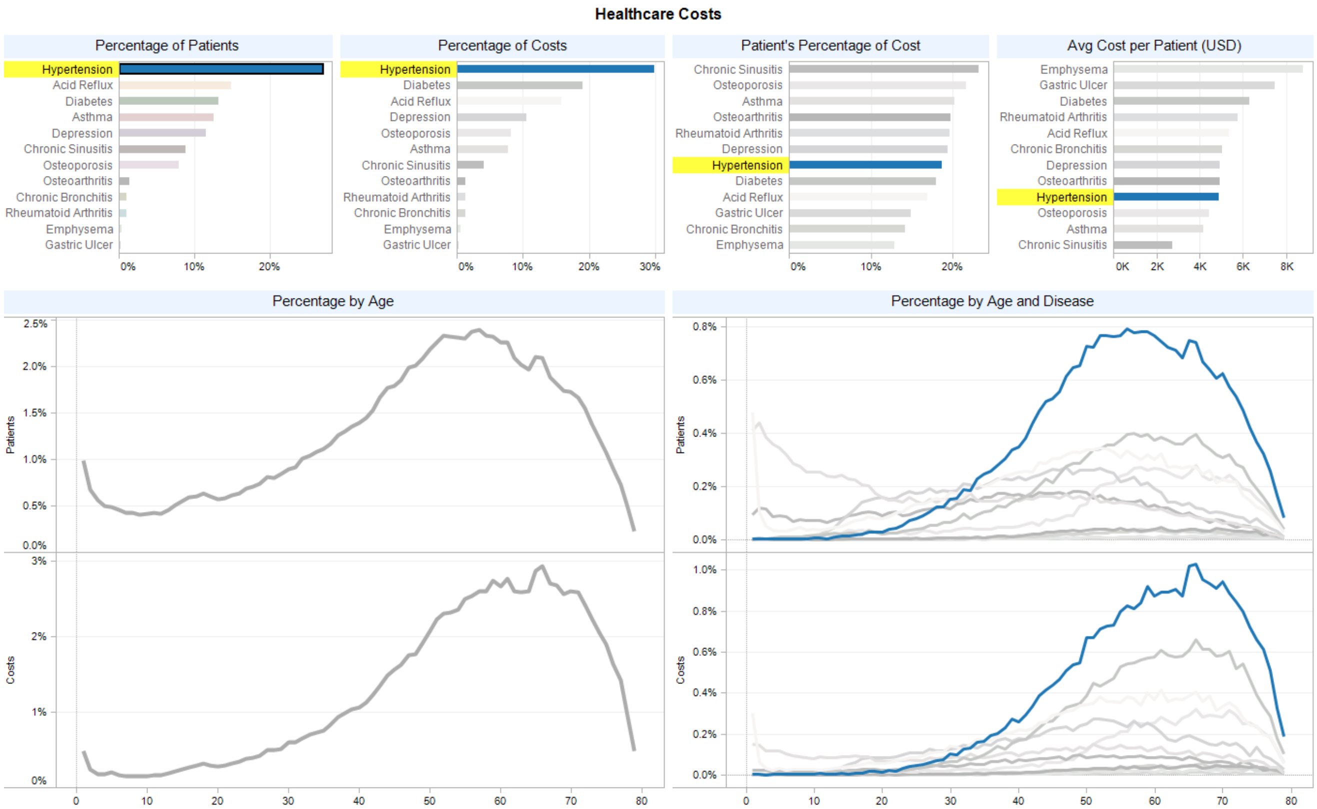


Now we can see differences in the effects of these diseases by age. For example, notice the large number of one year olds that are treated for acid reflex (the orange line), which drops off to relatively few for two year olds and then begins to rise slowly to a peak in the 50 to early 60 year range. Notice also how asthma (the red line) is primarily a disease that affects the young, yet treatment costs remain fairly consistent until old age.

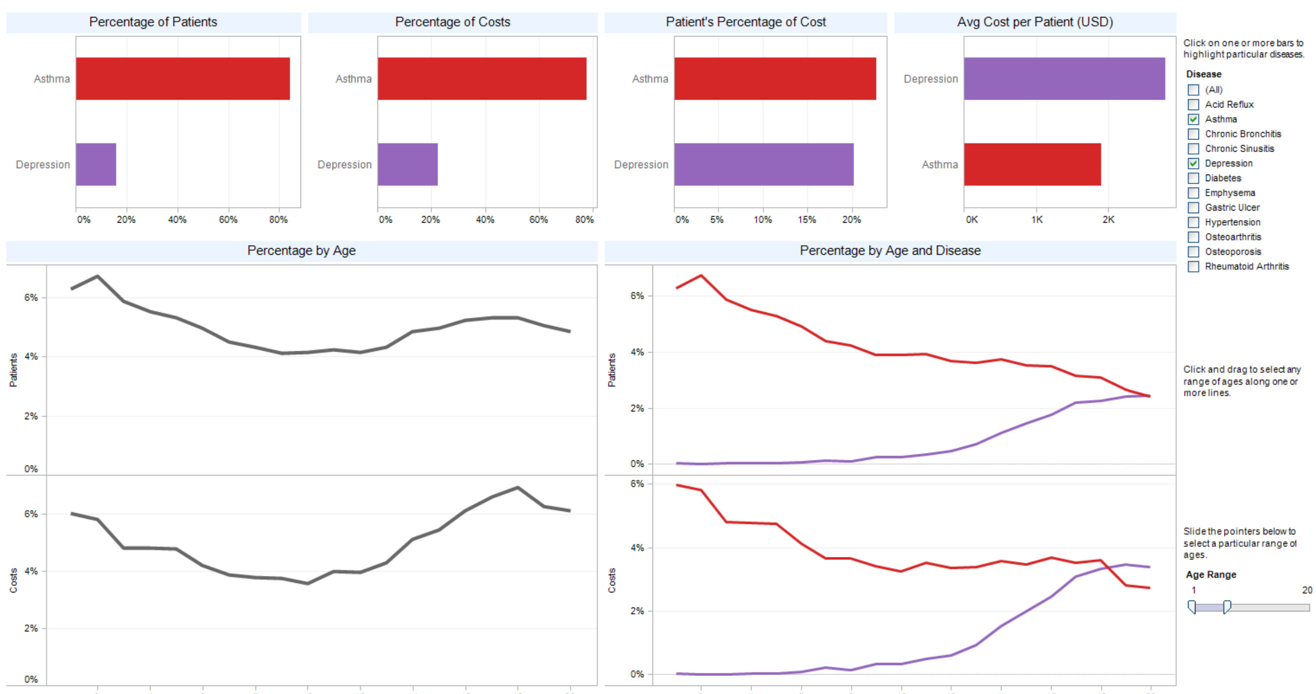
For even greater analytical insight, we need a display that combines everything that we've seen so far, and perhaps more, together. Here's one way this can be done:



In addition to the views that we've already seen, a graph that shows the average cost per patient per disease appears in the top right. Even though this display consists of eight graphs, it isn't difficult to understand, because each individual graph is simple. By seeing them together, we can discover interesting facts and patterns that would remain hidden if we could view only one graph at a time. For example, notice how many patients are affected by hypertension overall (the blue bar in the upper left graph), resulting in huge treatment costs (second bar graph), even though the per patient cost is relatively low (rightmost bar graph), and the patients' portion of costs is near the middle (third bar graph), yet hypertension has little effect on the young (the two line graphs on the right). In the example on the next page, I've featured this story about hypertension by selecting it in the leftmost bar graph, which caused it to be highlighted in every graph that displays values per disease.



We can use the data selection controls for disease and age to focus on particular subsets of data without distraction from anything that doesn't interest us in the moment. For instance, let's take a closer look at asthma and depression among young people, ages 1 through 20, by filtering out all other data.



Notice the steady decline in asthma from age two on, despite the fact that overall it affects the most young people by far. Notice also the steady rise in depression that begins at age nine and continues without relief, despite the fact that costs decrease slightly from ages 19 to 20.

Rather than relying on my examples, test the effectiveness of this display for yourself. You can view a live, web-based version by clicking this link. (Note: The interactive, Web-based version of this display is powered by Tableau Public, a free offering from Tableau coming soon.)

Here are a few of the principles that I've illustrated, which would have led to a better visualization of healthcare costs, had Fry followed them.

1. 2-D position and the lengths of simple objects such as bars encode quantitative values in ways that are easy to perceive; angles and areas do not, and therefore should be used only when you can't use better means.
2. We cannot build a picture in our heads of a pattern that is formed by multiple values (such as the average cost of healthcare for patients of each age from 1 to 79 years old) by looking at one value at a time.
3. Lines do a good job of showing the pattern formed by a set of values across a continuous range such as patients' ages, and do so in a way that allows us to compare patterns when multiple data sets are represented at once (such as one line per disease).
4. Multiple graphs shown together are often a better solution than a single graph, especially when several variables are involved.
5. Several simultaneous views of the same data set, each showing the data from a different perspective, make it possible to see relationships that can't be seen from one perspective only or from viewing different perspectives independently.
6. The ability to easily filter out data that doesn't concern us at the moment makes it easier to focus without distraction on what does concern us.

Fry's design was probably influenced by graphical design principles that don't apply to data visualization–at least not to this particular set of data, for this purpose, in the way that he applied them.

Let's return to where we began before veering off into a critique and recreation of Fry's healthcare visualization. I wrote this article to welcome an opportunity to take information visualization into new territory through the help of designers and artists who are experimenting with it as a mode of expression. I don't want to end on a negative note. Despite my critique of Fry's visualization, I don't want to discourage him or any other designer or artist who wants to work in the field of data visualization. What I do want is to promote a collaboration that mutually respects the expertise of newcomers and old-timers alike. Newcomers are remiss when they ignore the hard-learned lessons that have emerged through many years of research and real-world experience. Those who have worked in the field for awhile should welcome new perspectives and help newcomers to avoid mistakes of the past.

I am not alone in my hopes and concerns. I want to mention one person in particular who seems to share them–Manuel Lima–because like many of the newcomers I've been describing, he came to information visualization a few years ago from a background in design. Lima brings a fresh and useful perspective to the field. In his case, the opportunity for advancing the field is not being frittered away. He began by taking time to understand the field–its objectives and effective design practices–without sacrificing the complementary perspective that he brings as a designer. His website Visual Complexity features some of the most interesting work that is being done to by designers of various backgrounds, including artists. A recent video interview produced by *digup.tv* provides an evocative glimpse into Lima's world.

Last August, Lima expressed concerns similar to mine in the form of a manifesto. He argued that designers and artists who wish to contribute to information visualization should not lose sight of its goals or ignore its useful principles. Although Lima and I have come to information visualization from very different backgrounds and perspectives–he from industrial design and I from years of working in the field of business intelligence–our perspectives are united in the belief that information visualization should solve real problems and do so effectively. Here's an excerpt from the beginning of Lima's manifesto:

> *Over the past few months I've been talking with many people passionate about Information Visualization who share a sense of saturation over a growing number of frivolous projects. The criticism...usually goes along these lines: "It's just visualization for the sake of visualization", "It's just eye-candy", "They all look the same"...The recent outburst of interest for Information Visualization caused a huge number of people to join in, particularly from the design and art community, which in turn lead to many new projects and a sprout of fresh innovation. But with more agents in a system you also have a stronger propensity for things to go wrong.*

While not a harsh critique of these efforts, Lima feels that "it's important to reemphasize the goals of Information Visualization, and at this stage make a clear departure from other parallel, yet distinct practices."

The intersection of the arts, design, and information visualization need not be a collision strewn with the wreckage of failed efforts. There's too much at stake. All of us who wish to help people tame the beast of data overload can work together, each contributing our unique talents and perspectives. This requires mutual respect. Newcomers can't ignore the hard-earned wisdom of those who have labored long, nor can those who are thoroughly versed in the discipline look on newcomers as interlopers. Let's collaborate to give the world visualizations that enlighten.

Discuss this Article

Share your thoughts about this article by visiting the Information Visualization, Design, and the Arts: Collision or Collaboration? thread in our discussion forum.

About the Author

Stephen Few has worked for over 25 years as an IT innovator, consultant, and teacher. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides training and consulting services, writes the bi-monthly *Visual Business Intelligence Newsletter*, speaks frequently at conferences, and teaches in the MBA program at the University of California, Berkeley. He is the author of three books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, *Information Dashboard Design: The Effective Visual Communication of Data*, and *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. You can learn more about Stephen's work and access an entire library of articles at www.perceptualedge.com. Between articles, you can read Stephen's thoughts on the industry in his blog.