

Information Visualization Research as Pseudo-Science

Stephen Few, Perceptual Edge
Visual Business Intelligence Newsletter
October/November/December 2015

Research in the field of information visualization is usually mediocre, often severely flawed, and only occasionally well done. In this article, I'll critique a recently published research paper to illustrate some of the flaws that are common and to propose what we can do to improve the work. This is a paper by Michelle A. Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S. Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva, titled "Beyond Memorability: Visualization Recognition and Recall." I've chosen this particular paper, not because it is exceptionally flawed, but because it has received a great deal of attention and the findings that it erroneously claims are potentially harmful to data visualization. This concerns me, and should concern us all.

First, Some Context

Information visualization research tends to fall into a few different categories. From what I've observed, these are as follows:

1. Software Development Studies

These studies focus on developing new forms of display or functionality (visualizations, interfaces, interactions, analytical algorithms, etc.) that can be incorporated into visualization tools to either solve problems in existing tools or to extend their reach. These studies usually include some testing to confirm whether the new features actually work. Most information visualization studies fall into this category. This is perhaps because most information visualization programs reside in Computer Science departments.

2. Experimental Studies

These try to use conventional scientific approaches to determine how (or how well) aspects of information visualization work. This is done by forming and experimentally testing hypotheses.

3. Theoretical Model Studies

These studies create new models for understanding some aspect of information visualization. They often propose new taxonomies for classifying phenomena (visualizations, interactions, purposes for which visualization is done, etc.).

4. Literature Review Studies

These review previous studies about some aspect of visualization to summarize their findings and suggest future work.

It's important to recognize the intentions of a research study before evaluating it, because different purposes involve different methods and must be judged by different standards. Most of the concern that I'll express in this article relates to experimental studies. The merits of any study that makes claims based on empirical observations and measures should be judged according to scientific standards. It is in this particular realm that information visualization research often errs most egregiously.

Before reading further, take some time first to review [Borkin's paper](#) for yourself. Use this as an opportunity to test and hone your own ability to evaluate the reliability and merits of research. If you do this, you will learn a great deal more than you would if you proceeded immediately to my review. You might find problems in the

study that I missed and you also might find that you disagree with aspects of my assessment. Findings of both sorts will be worthwhile.

A Bridge to Nowhere

Research is about discovery. We try to find what we don't already know or to test generally-accepted beliefs. Although we can't always be certain in advance, we seek to discover what might be useful in some way. Not nearly enough of us legitimately pursue this important work, so it's important to focus on what matters. The objective of Borkin's and her co-authors' study—an extension of her prior work titled "What Makes a Visualization Memorable"—was to determine what makes visualizations recognizable after being seen for a period of 10 seconds. The greatest tragedy of this research is that what makes a visualization memorable is actually of no consequence.

In the paper's introduction, Borkin asserts:

Memorability, a basic cognitive concept, has important implications for both the design of visualizations that will be remembered but also lays the groundwork for understanding higher cognitive functions such as comprehension.

Her bold statement is never substantiated. This study does not show that memorability is useful for the design of visualizations. It does not actually address comprehension. Had it done so, it could have been useful. In Borkin's prior study, she wrote:

Clearly, a more memorable visualization is not necessarily a more comprehensible one. However, knowing what makes a visualization memorable is a step towards answering higher level questions like "What makes a visualization engaging?" or "What makes a visualization effective?"

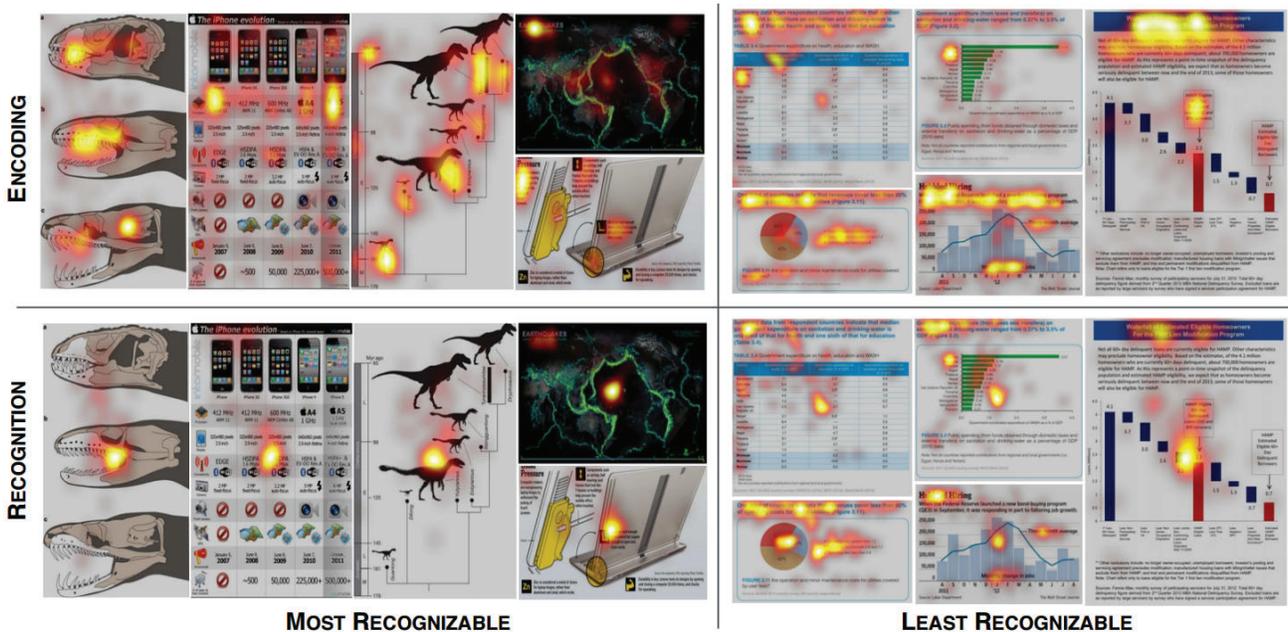
Even though in that study she recognized that "a more memorable visualization is not necessarily a more comprehensible one" and that "What makes a visualization effective?" is a more useful question, she did not address these issues in the follow-up study. Instead, she continued to build on an ephemeral foundation.

Visualizations don't need to be designed for memorability—they need to be designed for *comprehension*. For most visualizations, the comprehension that they provide need only last until the decision that it informs is made. Usually, that is only a matter of seconds. When the comprehension has lasting value, it should be stored in memory, not the visualization. It is true that a visualization can be designed to serve as a mnemonic mechanism to encode a particular message in a memorable way. Borkin's study, however, merely addresses characteristics that make visualizations memorable, not whether or how those characteristics can be used to encode comprehension.

We've known for quite a while that particular characteristics of data visualizations tend to catch our attention and make them sticky. For example, novel images and those that trigger strong emotions tend to stick in memory, but rarely, if ever, in a way that supports comprehension. If I incorporate an image of a kitten into a data visualization, I can guarantee that a test subject would remember seeing that kitten if it is shown to her again a few minutes later. But how is that useful? Unless the visualization's message is that kittens are cute and fun, nothing of consequence has been achieved.

This study involved a series of tasks, divided into three phases: 1) a 20-minute "encoding phase," in which subjects were shown 100 visualizations randomly selected from a pool of 393 for 10 seconds each, 2) a 10-minute "recognition phase," in which subjects were shown these same 100 visualizations randomly intermixed with another 100 previously unseen visualizations for 2 seconds each and asked to indicate whether or not they recognized it from before, and 3) a 20-minute "recall phase," in which the visualizations that were correctly recognized in the previous stage were shown again, one at a time, but were smaller and blurred, and the participants were asked to type descriptions of those visualizations in as much detail as possible. Eye-tracking equipment was used to monitor the subjects' visual fixations during the first and second phases, and the quality of the descriptions that they typed during the third phase was assessed by "visualization experts."

The following figure from the paper shows what subjects predominantly focused on during the encoding and recognition phases for the most recognizable and least recognizable visualizations.



Notice that the most recognizable visualizations all had unique visual images that made them easy to remember, while the least recognizable consisted of fairly standard graphs with a considerable amount of text, which required time to read. A standard chart, such as a bar graph, typically includes no novelty that would make it sticky. Furthermore, text would not be memorable unless something that the subject had time to read during the 10-second encoding phase was surprising or seemingly important. The findings reported in this figure were unsurprising.

Borkin states, “The recognizable visualizations tend to contain more human recognizable objects (objects are present in 74% of the top third most recognizable visualizations) compared to the least recognizable visualizations.” No “human recognizable objects” could be incorporated into the least recognizable visualizations on the right to make them more rapidly digestible and useful. In other words, this observation is of no use to us.

Take a moment to try a little experiment. Try to retrieve from memory *useful* details about any specific data visualizations that you’ve seen in the past. Can you do it? If so, how many can you retrieve? If you’re like me, the answer is, “Not many.” In fact, even though I’ve viewed tens of thousands of data visualizations in my life and have created thousands of them, I can only recall a few useful visual details from them, and those are simple visual patterns that exist in graphs that I’ve used countless times in my courses and lectures. For example, in a line graph that I use to illustrate particular benefits of data visualizations over tabular data displays, I remember the cyclical quarterly pattern of domestic sales in the form of a hockey stick and the relatively flat pattern of international sales, except for the dip in August. A memory of this type is potentially useful, but limited, for I can’t remember any of the actual values. What I’m pointing out with this little experiment is that the usefulness of data visualizations is not based on our ability to store their visual components in memory.

Borkin’s study illustrates a fundamental problem in many visualization research studies: the researchers do not understand what people actually do with data visualizations or how visualizations work perceptually and cognitively. Consequently, they don’t know what’s worth studying. Everyone who does research in the field of data visualization must spend some time actually working as a practitioner in the field. Relatively few do.

In data visualization research, we should identify objects of study that are potentially useful. Once a worthwhile hypothesis has been generated, a research study must then be properly designed to yield trustworthy results. As in Borkin’s prior study, the design of this new study was fatally flawed. The flaws are statistical, methodological, and logical in nature. Let’s examine them one at a time.

Statistical Unreliability

I'll begin with a fundamental flaw that would cause many scientists to read no further than the abstract of Borkin's paper, where it reveals that only 33 test subjects were involved in the study. The minimum number of subjects in an empirical study that are needed to produce reliable results depends on the nature of the study, the size of the population, and the degree to which you want your results to be reliable, which is called *statistical power*. (If you aren't familiar with statistical power, read *Statistics Done Wrong: The Woefully Complete Guide*, by Alex Reinhart.) There is no evidence that Borkin attempted to determine a reliable sample size. Nowhere was statistical power even considered. I suspect that the size of the sample was determined by convenience alone.

Some empirical studies can be done with samples of this size or even smaller, but only when it has been established that variation in the domain under investigation is slight. For example, psychological studies of low-level perception can be done with relatively small samples because this level of brain activity varies little from person to person. For a study of this type, however, a sample of only 33 subjects will not likely produce statistically reliable results. Because Borkin has not provided a convincing rationale for a sample of this small size, this study could at best serve as a prototype to suggest the need for a legitimate, properly designed, statistically valid scientific study.

Why are visualization research studies still plagued with these small samples when it is well known that they are inadequate? It is for one reason only: small samples are convenient. In science, however, trading validity and reliability for convenience is not an option.

To add insult to injury, this 33-participant study made no attempt to make the sample representative of a particular population. It consisted of 17 females and 16 males from the communities of Cambridge and Boston, as if balancing the gender mix was all that was needed to make the sample valid. This is a common problem in particular areas of research, such as psychology.

Because of these fundamental statistical flaws, no reliable findings can be claimed. Nevertheless, we'll press on to examine other ways in which the study was flawed.

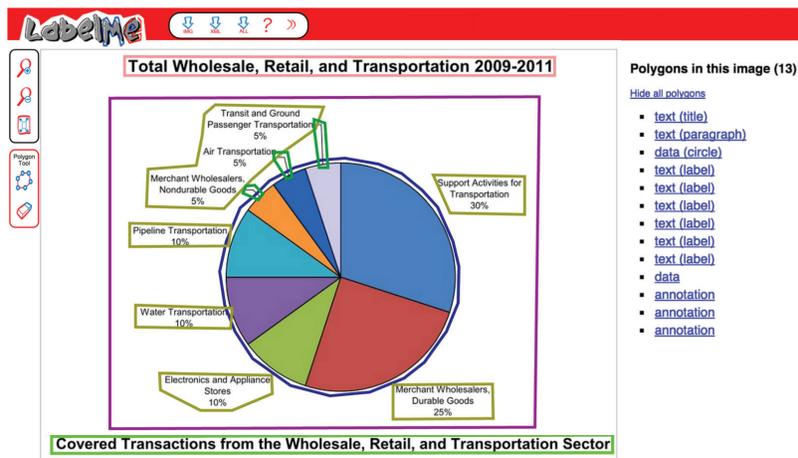
Methodological Misdirection

The scientific method directs us to design empirical studies in particular ways. Several of the fundamental rules for designing a study of this type, however, were broken.

1. When subjects are selected to form a representative sample of a particular population, randomization is often used to achieve this result. Rather than randomly selecting the test subjects, however, this study randomly selected the visualizations that subjects were shown. During the initial encoding phase, each subject was shown 100 visualizations that were randomly selected from a set of 393. In other words, no two subjects saw the same set of visualizations, making it difficult to compare their responses.
2. During the encoding phase, subjects were shown each of the randomly selected visualizations for 10 seconds each, separated by a half-second "fixation cross to clear their field of vision." A period of 10 seconds is not sufficient to read most useful data visualizations for comprehension. The choice of a 10-second period of time, which Borkin referred to as "prolonged exposure," likely stems from her lack of familiarity with the way people actually use data visualizations. It would be unusual for a visualization to be examined for only 10 seconds if the viewer were actually trying to glean information from it. This brief period of time may only be described as prolonged in comparison to the one-second "at-a-glance" period of exposure that was used in the prior study. Neither study simulates the way that people routinely view visualizations. This new study appears to have been done in response to criticism of the prior study that a one-second exposure could not yield meaningful findings. Borkin wrote, "In this paper, we extended encoding to give participants enough time to process the content (textual components, messages, and trends) of a visualization." In fact, a data visualization that can be sufficiently processed in 10 seconds is rarely useful.
3. The paper does not reveal the specific instructions that subjects were given for the encoding phase. This is a critical omission, because the nature of the task that subjects were assigned certainly led

them to encode information from the visualizations in particular task-driven ways.

4. During the recognition phase, subjects were shown the same 100 visualizations “as well as 100 filler visualizations” in random order for two seconds each with a half-second “fixation cross between visualizations.” For each visualization, they were instructed to press the spacebar if they recognized it from the previous encoding phase. Recognition in this setting, however, can be highly influenced by the nature of the filler visualizations, but no effort was made to control this influence. For example, visualizations that are similar can be easily confused with one another and therefore a subject might be inclined to designate a filler visualization as recognized. On the other hand, if filler visualizations are quite different from target visualizations, they could be easily differentiated from them, encouraging subjects to select all visualizations as recognized that are not obviously unfamiliar. Given the fact that both the target visualizations that were initially shown and the filler visualizations that were added during the recognition phase were randomly selected, no attempt was made to control this influence. During the recognition phase, after each visualization was shown for two seconds and the subject labeled it as recognized or not, “A feedback message was presented (i.e., correct, incorrect).” Why? The purpose of the study was not to help subjects improve their recognition skills. This feedback could have exercised inappropriate influence over recognition performance.
5. During the recall phase, “each visualization was presented at 60% of its original size and blurred by a 40-pixel wide Gaussian filter to make the text unreadable. The purpose of blurring visualizations was to allow the visualization to be recognizable, but not contain enough detail to enable the extraction of any new information.” Subjects were then asked to type a description of the visualization “in as much detail as possible.” I would have liked to see an example of a blurred visualization, but none were shown in the paper. This set up did not strictly limit the subject to recalling information from memory, for some characteristics of the visualization could be detected when blurred, but others could not. This would lead us to expect subjects to mostly recall information related to the visual elements that could still be detected even when blurred. Borkin admits that this setup “has its limitations, including the potential to bias participants to more easily recall visual elements.”
6. During the recall phase, subjects were only shown those visualizations that they correctly identified during the recognition phase. Because subjects were only shown the visualizations that they correctly identified, which varied from subject to subject, each was shown differing numbers of visualizations, which gave different subjects varying amounts of time to describe the individual visualizations during the 20-minute recall phase. This difference would have had an impact on the quality and quantity of subjects’ descriptions.
7. Prior to the study, visualizations were stratified by three “visualization experts” into “polygons” that supposedly matched specific “visual elements.” These polygons were then used to determine the content that was viewed using eye-tracking technology. These polygons, however, did not necessarily match the way that we stratify a visual scene into perceptual units. The example below, which appeared in the paper, groups some of the contents in ways that seem arbitrary. Notice that the three labels in the upper left of the pie chart were treated as a single visual element and that all the slices of the pie were grouped together, even though they represent separate values.



8. The study reports findings that have nothing to do with its purpose. For example, it reports differing amounts of “data redundancy” and “message redundancy” among the following sources of visualization: government, infographic, news, and scientific. The highest rates of message redundancy supposedly exist in infographic and news media visualizations, which is “probably due to both venues prioritizing clear communication of the visualization message.” Also, according to Borkin, the “scientific visualizations in our sample had the least message redundancy and no data redundancy.” Another example is the calculation of the percentage of space in visualizations that contains data per these same sources. Apparently, science visualizations include a greater percentage of data than the other sources. These source categories and observations regarding them are irrelevant to the study—distractions. They suggest an unfocused approach to observation and data gathering in this study.

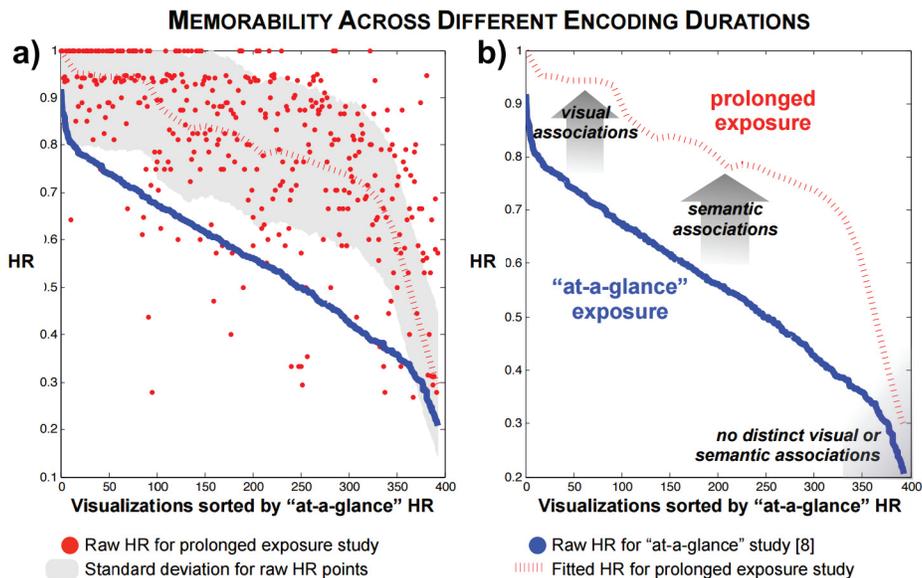
Logical Fallacies

I will frame this section by quoting a well-known syllogism:

*All philosophers are morsels
Socrates is a philanthropist
Therefore, Socrates is a Mormon.*

Obviously, this is a perversion of the well-known syllogism, but that’s intentional, for I will show that this study makes several claims that don’t flow logically from the work that was done. Each of the following begins with a claim made by Borkin.

- “Within the context of our experimental design, we are able to more deeply understand the specific cognitive processes involved in recognition and recall.” In fact, no new understanding of cognitive processes emerged from this study.
- “A 10 second duration [during the encoding phase of the study] proved to be of sufficient length for a participant to read the visualization’s title, axes, annotations, etc., as well as explore the data encoding, and was short enough to avoid too much redundancy in refixations as well as explorative strategies.” The statement was never substantiated. On what evidence did she base this claim?
- “The increase in memorability across experiments [the prior vs. the current study] for some visualizations can be explained by additional semantic associations (e.g., title, text, etc.) being invoked.” This comment, which appears in the caption of the figure below, seems to pertain to the portion of the red fit model to the right of the “semantic associations” arrow. This claim appears to be pure conjecture. No evidence was provided to support it.



- “By considering the eye movements made on these visualizations during 10 seconds of encoding and at recognition time, we can see what parts of a visualization are encoded and what parts are required for recognition.” This claim cannot be made. Eye tracking merely revealed where the subjects looked during each of these two phases, not what parts were required for recognition.
- “These two metrics (eye movements and textual descriptions) together are evidence for which elements of a visualization contribute to a participant’s memory and understanding of a visualization.” In fact, none of the metrics that were used in this study reveal anything about understanding. The descriptions that subjects typed during the final recall phase were later subjectively scored by researchers, and not in a way that could reliably serve as a measure of understanding. Measuring understanding is difficult, which any good teacher can tell you.
- In the final paragraph of the paper, the following quote appears from a book written by a friend of mine, Jean-luc Doumont:

Effective communication is getting messages across. Thus it implies someone else: it is about an audience, and it suggests that we get this audience to understand something. To ensure that they understand it, we must first get them to pay attention. In turn, getting them to understand is usually nothing but a means to an end: we may want them to remember the material communicated, be convinced of it, or ultimately, act or at least be able to act on the basis of it. (Trees, Maps, and Theorems: Effective Communication for Rational Minds, Jean-luc Doumont, Principia, 2009)

Doumont is a true expert in communication. It feels wrong to include his words because this study is about the recall of particular elements of visualizations following a 10-second viewing period, not about communication. It did not measure understanding, which, as Doumont wisely points out, is the goal of communication.

What Should We Do?

In the realm of information visualization research, Michelle Borkin’s paper is not uniquely flawed. Rather, it is typically flawed. The flaws stand out sharply when compared to a well-designed study. For example, this year a study was published titled “Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations” by Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer, which is an exemplar of information visualization research. It serves as a counterpoint to the flaws that I’ve identified in Borkin’s paper. It addresses a real need and does so in a way that demonstrates an intimate, experience-based understanding of that need. The authors recognized that most data analysts fail to explore and examine their data as broadly as they should. This is in part because most analytical tools are designed to encourage depth rather than breadth of analysis. In response to this critical need, the authors proposed and tested ways that the interfaces of tools can be designed to encourage and support greater breadth of analysis. Because this study, like Borkin’s, involved a small sample of test subjects, it thoughtfully refrained from claiming conclusive findings. Although tentative in a scientific sense, its findings are useful as the basis for further studies, which is all that the authors claimed.

The causes of the flaws in Borkin’s study are systemic. As such, to fix them, we must address their systemic roots. Here are the main problems, as I understand them.

1. Most information visualization research is done by people who have not been trained in the scientific method.

Even though most information visualization research is done in university departments with “science” in their names—predominantly “Computer Science”—most of the work is not done scientifically. What goes by the name “information visualization research” today falls into two general categories: 1) empirical research that should be conducted according to the scientific method, and 2) software development that seeks to create new or improved technologies for information visualization. Because the latter is not research in the strict sense, the standards that govern it need not involve the scientific method. The former, however, must conform to the scientific method. As such, those who do this work and those who supervise it must be scientifically trained. Until they are, most of this empirical research will remain pseudo-science. Even though information visualization does not and probably should

not qualify as an independent branch of science, it relies on other scientific disciplines (cognitive psychology, statistics, etc.) to inform its work. Consequently, work that draws on these other scientific disciplines must adhere to the scientific method. Until it does, it will remain an object of derision in the scientific community.

2. Academics are forced to publish research papers.

This particular problem is not unique to information visualization. Many disciplines are plagued by the fact that they function as publication mills with insufficient controls for research quality. When you cannot advance in your field without publishing, publication becomes your goal rather than the pursuit of meaningful, valid, and useful research.

3. Information visualization publications do not adequately vet research studies.

Much of the blame for flawed research belongs to those who review the work, giving the green light for publication to papers that don't deserve it. Most of the information visualization research that is published should have never passed inspection. When the authors of flawed research are rewarded with publication by a dysfunctional peer review process, and later given advanced degrees and faculty positions, it is no wonder that their work never improves. Why should it? Those who review flawed research and give it a passing grade are just as responsible as those who do the research. The role of reviewer is critical, but the process is lax. If the process were not conducted anonymously, perhaps reviewers would do their job more responsibly. Perhaps not. The truth is, most reviewers are no more acquainted with the scientific method than those who do the research.

4. The information visualization community is complacent.

Why is it that few people besides me are critiquing flawed information visualization research? I'm certainly not unique in my understanding of science. Several of my friends and colleagues do excellent research in the field. They're aware of the same flaws I am, so why aren't they speaking up? I suspect that for many, it is fear of recrimination or distaste for conflict that keeps them silent. I loathe recrimination and conflict as much as anyone. But I want information visualization to contribute as much as it can to the world, so I work to improve it. My lone voice, however, cannot turn the tide. Others who are respected in the field must speak up as well. Imagine what a difference we could make if we raised our voices in a chorus of guidance to elevate the work of our field. Oh, how I long for compatriots.

I've made these suggestions before, and I'll probably make them again, but I won't be around to do this forever. We'll never progress by setting the bar this low. Borkin didn't produce a flawed study because she lacks talent. As a doctoral student she did a study titled "[Evaluations of Artery Visualizations for Heart Disease Diagnosis](#)" that was exceptionally worthwhile and well done. In that study, she showcased her strengths. I suspect that her studies of memorability were dysfunctional because she lacked the experience and training required to do this type of research. She is now an Assistant Professor at Northeastern University, teaching the next generation of students. I'm concerned that she will teach them to produce pseudo-science. This is a depressing cycle. Too many academics are supervising research studies that fall outside of their areas of expertise. Isn't it time to break this cycle? To do so, we must speak up. You can begin now by responding to this article. Whether you agree with my perspective or not, a public dialog needs to begin. Critique my critique, if you feel that I've missed the mark in any way. If you share my perspective, do so publicly. This isn't about us; it's about a fledgling field of study that is struggling to emerge from the nest. Let's give it wings.

Discuss this Article

Share your thoughts about this article by visiting the [Information Visualization Research as Pseudo-Science](#) thread in our discussion forum.

About the Author

Stephen Few has worked for over 30 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 quarterly [Visual Business Intelligence Newsletter](#), and speaks frequently at conferences. He is the author of four books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, Second Edition, *Information Dashboard Design: Displaying Data for at-a-Glance Monitoring*, Second Edition, *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, and *Signal: Understanding What Matters in a World of Noise*. 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](#).