ETHICAL INFOGRAPHICS

In data visualization, journalism meets engineering

BY ALBERTO CAIRO — UNIVERSITY OF MIAMI

LET ME START by boring you to death with some relevant theoretical considerations. Bear with me for just a few paragraphs. You'll have much more fun if you grasp where I'm coming from. At the core of the discussion on what constitutes morally good behavior when designing news graphics and data visualizations lies a much more fundamental question: What is a visualization, anyway?[1] It's information encoded as charts, maps, diagrams and illustrations that facilitate communication, analysis and understanding. A visualization is much more than an unidirectional channel between a "teller" and a "viewer." Rather, visualizations are artifacts that designers contrive for people to explore and comprehend datasets, geographic realities or complex entanglements of ideas and events.

Therefore, creating a visualization isn't just an act of journalism, but also of engineering. To think about the ethics of news visualization — a field waiting to be developed — is to go beyond the grand themes traditionally covered in the literature on journalism ethics and to explore matters of effectiveness and efficiency.

Journalists who design visualizations need to address questions related to what they should display and why, but also pay increased attention to how they should display it. In other words, visualization designers must think about the structures, styles and graphic forms that let audiences access information successfully in every situation.
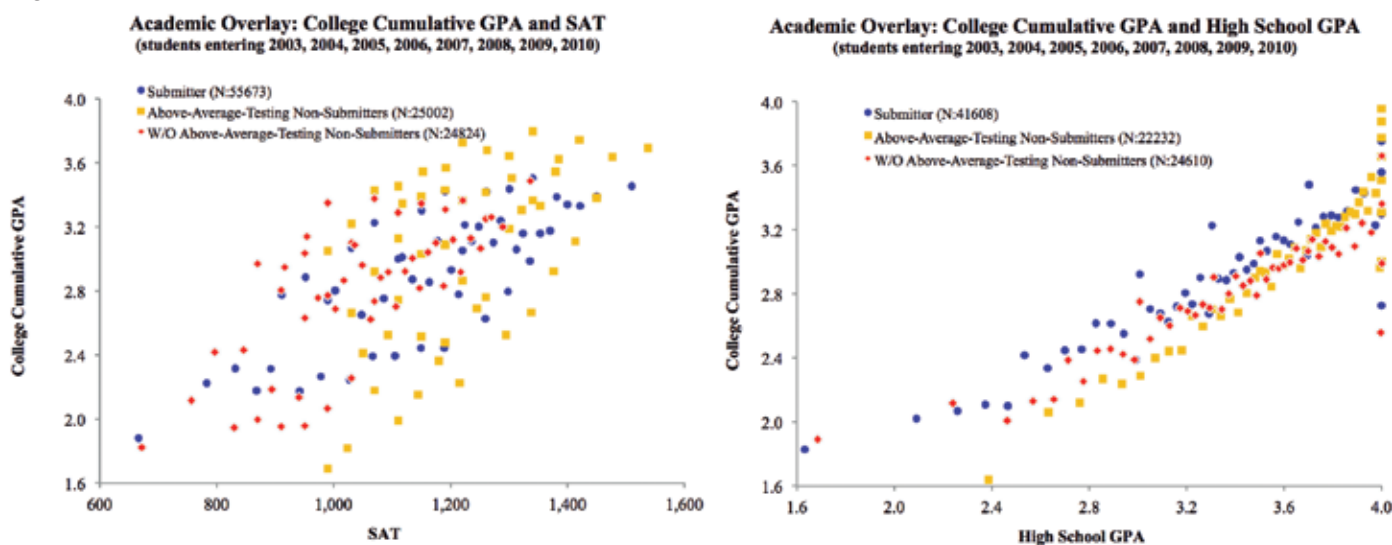
A disclaimer: I'm an unashamed rule utilitarian.[2] I won't take the time to defend my stance here. I've been told that I can only write up to 1,800 words. It'll be enough for you to remember that the cornerstone of this article is that the purpose of journalism is to increase knowledge among the public while minimizing the side effects that making that knowledge available might have. Take the time to digest the previous sentence: Increasing understanding while minimizing harm.

Figure 1

Kidnappings in Nigeria, 1982-present

Kidnapping rate per 100,000 people, 1982-present
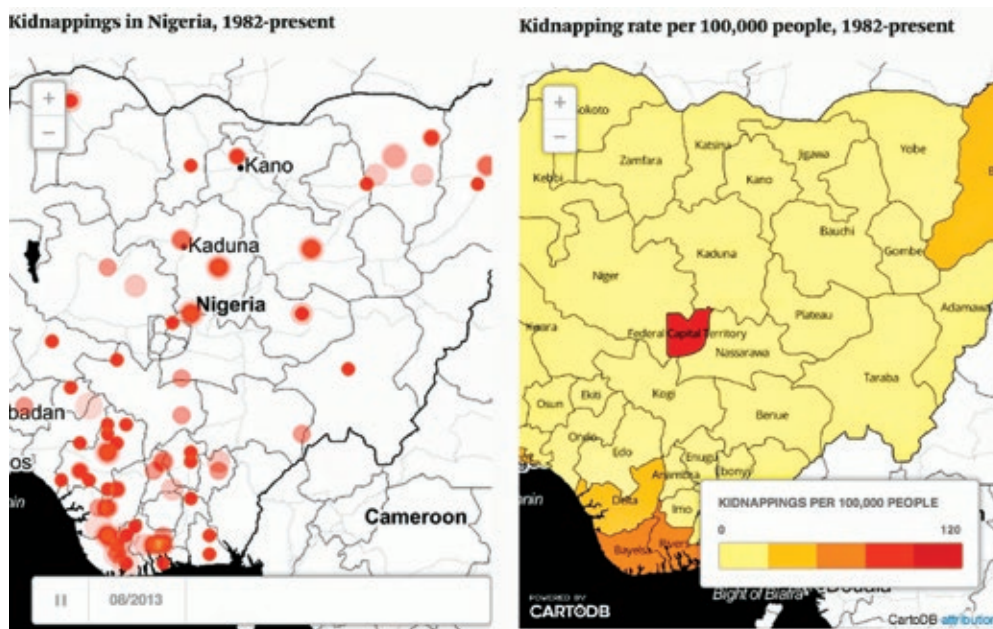
FiveThirtyEight used data from the Global Database of Events, Language and Tone, which tracks news stories about kidnappings, to create both graphics.

FiveThirtyEight

Figure 2

Academic Overlay: College Cumulative GPA and SAT
(students entering 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010)

- Submitter (N:55673)
- Above-Average-Testing Non-Submitters (N:25002)
- W/O Above-Average-Testing Non-Submitters (N:24824)

Academic Overlay: College Cumulative GPA and High School GPA
(students entering 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010)

- Submitter (N:41608)
- Above-Average-Testing Non-Submitters (N:22232)
- W/O Above-Average-Testing Non-Submitters (N:24610)

Correlation between SAT scores and GPA in college and between high school GPA and college GPA.

The journalistic virtues and deontological rules we often hear and read about (academics in journalism ethics love them, but they rarely reason why we should embrace them in the first place) can be reworded as utilitarian guidelines for conduct. We don't "tell the truth" just for the sake of telling the truth, and we are not "honest" because "honesty" is a self-evident "virtue." Arguably, such notions are risible.

Rather, we are honest and we tell the truth because we have evidence to demonstrate that doing so increases the public's understanding about relevant matters. Better information is correlated to well-being. Generalized knowledge can lead to generalized happiness.

Now, as promised in the first paragraph, let's proceed to the fun part, in which I would like to propose several topics for further discussion.

What to show

Although the main focus of this article isn't how we should choose topics to cover or how to collect and process data, I would like to make a couple of points. First, I feel that news publications that conduct data journalism rely too much on existing and easily accessible databases, a fact that may introduce biases in their story choices. Second, those stories are sometimes built without assessing the quality of their sources or applying proper reporting and analysis methods. This can lead to disastrous results.[3]

In May, Mona Chalabi, lead writer for the FiveThirtyEight's Data-Lab, published two blog posts which claimed that kidnappings in Nigeria have accelerated in the past 30 years. The evidence provided to support the article's thesis were a couple of bar graphs and two animated maps. See Figure 1.

These graphics don't show kidnappings, though, but the number of news stories about kidnappings, a variable that was obtained from the Global Database of Events, Language and Tone. But "News stories about kidnappings" is a poor proxy variable for actual kidnappings if the data is not normalized somehow, perhaps by adjusting for the amount of news stories related to the beat. This is similar to inferring armed robbery rates in Miami based on the occurrence of the word "robbery" in the Miami Herald. It would be absurd, as the Herald may be covering middle-class neighborhoods better than it covers poor ones (I'm making this up.) Neglecting confounding variables is risky.

The two posts that Chalabi wrote didn't even put the data in context, for instance by mentioning that Boko Haram, the group that kidnapped hundreds of Nigerian girls recently, was founded in 2002. There's a surge in kidnapping reports after that year.

Several data-savvy readers reprimanded FiveThirtyEight. One of them, Erin Simpson, wrote on Twitter: "Validate your own data. It's not true just because it's on a goddamn map."[4] Simpson's observation is poignant, as we have some tentative evidence[5] to think that

Stories are sometimes built without assessing the quality of their sources or applying proper reporting and analysis methods. This can lead to disastrous results.

people are less skeptical of new information when it is presented as a chart or a data map than when it comes just as text. Visuals are persuasive and may provide a sense of certainty, for better or worse. FiveThirtyEight has since added an Editor's Note admitting the errors.

How to show it

When I say that news graphic designers are not just journalists, but also information engineers, I imply that we should think a bit more as software developers or industrial designers. In other words, for us the information is as important as the effectiveness and efficiency of the displays we devise to convey it. When we create a visualization, we're giving information a visual shape, we model it, we sculpt it. There should be a connection between the forms we choose and the tasks that our visualization is intended to facilitate.

Let's suppose that you're working on a story about the weak correlation between SAT scores and GPA in college, and the strong one between high school GPA and college GPA.[6] You have all the data and you know where all students in the sample came from.

How should you encode the data? The right procedure is to think about the purpose of the graphic ("to display relationships") and then choose the graphic form accordingly. In this case, a couple of scatter plots similar to Figure 2 will do the job quite nicely — after you have styled them for publication, of course.

Not all visualization designers proceed this way. Instead, they sometimes rely on their personal aesthetic preferences and choose graphic forms based just on their visual appeal. Beauty and creativity have a role in visualization, for sure. We need to experiment with new graphic forms — our current visual grammar and vocabulary are quite limited — but not at the expense of clarity and efficiency. Would it make sense to present the same data mentioned above on a map? No, because the purpose of this particular graphic is to display correlations, not to reveal geographic patterns.[7]

You may be asking yourself at this point: How do we know if a graphic form is better at fulfilling a particular purpose than another one? The answer is that we can partially base our decisions on existing research on statistics, computer science, and cognitive psychology. I can't go into many details here[8], but we already know that some shapes are indeed better for certain functions. To test this notion, take a look at Figure 3.

I'm pretty sure that you're able to estimate the height of most lines in the dot plot quite accurately. You cannot do the same in the bubble chart or in the pie chart. Why not? It has to do with the fact that the human brain is much better at comparing a single dimension (length, width, or height) than two (width and height or, in other words, area.)[9]

How is all this connected to ethics? Is it because clarity, efficiency, depth and caring for the functionality of graphics are self-evident goals? Not at all. The reason is that there should be a relationship between evidence (when evidence is available) and moral choices.

When a designer chooses a graphic form to represent data just because she likes it, while ignoring evidence that may lead her to choose a more appropriate one, her act is morally wrong. It's not wrong just because she's not been virtuous or because there is a deontological rule against inappropriate charts, but because her act will likely have negative consequences, such as confusion, obfuscation and misunderstanding.

Visualizations that are beautiful, fun, clear, functional and concise — but not devoid of depth and nuance — help us see things that we wouldn't be able to otherwise. Good visualizations provide insight. They enlighten us. And knowledge has the potential to lead people to live better, happier and more fulfilling lives. There can be no higher law in journalism, not even telling the truth or shaming the devil.

Alberto Cairo is a professor of the professional practice at the School of Communication of the University of Miami and director of the visualization program of UM's Center for Computational Science. He's the author of the book 'The Functional Art: An Introduction to Information Graphics and Visualization.'

Notes:

(1) For the sake of brevity, when I write "visualization" I refer to any kind of display that consists of graphic representations of information, from traditional static print infographics to interactive visual displays intended to explore a data set.

(2) To learn more about "rule utilitarianism" read http://plato.stanford.edu/entries/consequentialism-rule/ In case you've forgotten your Journalism Ethics 101 class, here's a reminder of the differences between consequentialism, deontology, and virtue theory (roughly speaking, the three large branches of normative ethics): http://www.trinity.edu/cbrown/intro/ethical_theories.html

(3) Two examples: First, the one discussed in this article, from FiveThirtyEight: http://www.thefunctionalart.com/2014/05/when-plotting-data-ask-yourself.html The other comes from vox.com: http://www.thefunctionalart.com/2014/04/my-new-infographics-motto-its-more.html

(4) "If A Data Point Has No Context, Does It Have Any Meaning?" https://storify.com/AthertonKD/if-a-data-point-does-not-have-any-me

(5) Blank slates or closed minds? http://www.dartmouth.edu/~nyhan/opening-political-mind.pdf and 'Lay understanding of probability distributions' http://journal.sjdm.org/13/131029/jdm131029.pdf

(6) See 'Defining Promise: Optional Standardized Testing Policies in American College and University Admissions' http://www.nacacnet.org/research/research-data/nacac-research/Documents/DefiningPromise.pdf

(7) There's a reason why the title of my first book in the American market is "The Functional Art."

(8) If you're interested in those details, I'd recommend to look into the work of Colin Ware, Stephen Few, Stephen Kosslyn, Naomi Robbins, and William Cleveland, among others.

(9) You should not infer from this that bubbles are always wrong. Think of a map showing electoral results all over the U.S. to identify regional voting patterns. In this case, bubbles may come in handy.

Figure 3
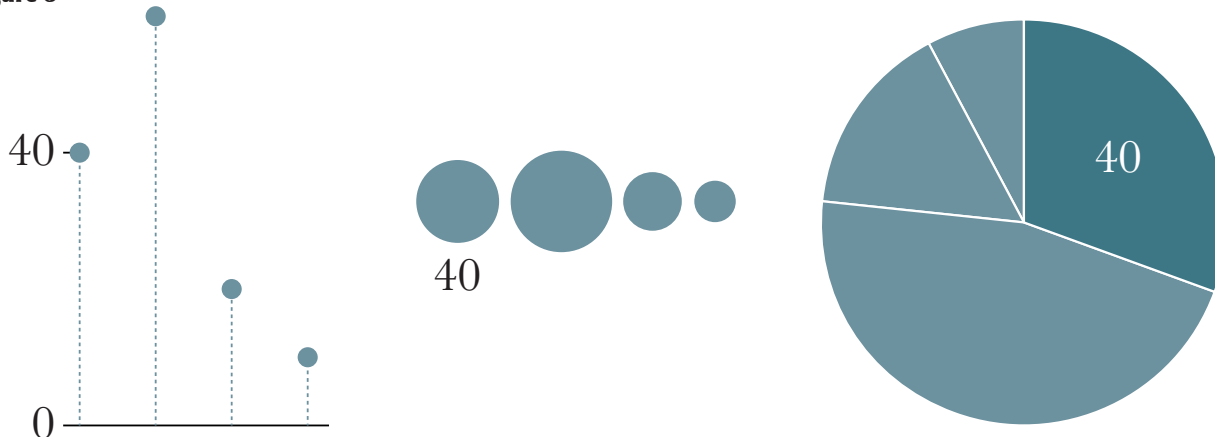
Try to estimate the values represented on each chart based on the only value shown. You can probably do this quite well on the dot plot, but not on the bubble chart or the pie chart. The values are 40, 60, 20 and 10.