Editorial
Big Data Visualization: Promises & Pitfalls

Katherine Hepworth

University of Nevada, Reno
khepworth@unr.edu

A few weeks ago, I was having dinner with a friend when a controversial subject came up. My friend had an extremely strong opinion about the harm caused by vaccination, and his argument went something like this: "I've seen the data. There was an infographic laying it all out." He couldn't remember specific numbers from the visualization he'd seen or the author of the article. He couldn't even remember the name of the publication, but the data visualization's overall argument was firmly lodged in his mind. His situation is not unique, and it provides telling insights on how we, as humans, perceive and respond to big data visualization.

SEEING IS BELIEVING BECAUSE SEEING IS SEDUCTION

People want to believe. From faith practices and fortune-tellers to science and data, the desire to believe in grand visions, and small facts that support those visions, is undeniably human. And we have a tendency to believe what we see more than what we hear, read, or feel. No matter how much evidence we have that seeing is not necessarily believing, the experience of seeing is strongly correlated with truth.

This human quality is as true of the most rigorous researchers, as it is of small children. Seeing things concretizes them in our mind's eye. It makes them seem more permanent and more real. We're all less skeptical of information presented in a visualization than information presented entirely as numbers and text. No matter how abstract, complex, or multifaceted we know a given research problem is, information distilled into a simplified visualization is

seductive. Among researchers there is little acknowledgement of or reflection upon the seductive quality of data visualization. This oversight has dangerous implications for research quality, and the human subjects represented through research data visualizations.

DATA VISUALIZATION DEFINED

Data visualization, in sum, is the reduction and spatial representation of datasets in such a way as to make them more intelligible than in their pre-visualization, tabular format. Data visualization includes broad categories of spatial representation such as maps, charts, tables, and infographics. In the realm of research tools, data visualization is uniquely persuasive. It's a fundamentally political practice, one that constantly molds beliefs, behaviors, and emotions, predominantly at a subconscious level. Sociologist Nikolas Rose has called them ''little machine[s] for producing conviction in others" (Rose, 2008, p. 36). This persuasive quality of data visualization, is sometimes portrayed as inherently bad. A recent National Geographic article described it as "weaponized data visualization" as if your next chart could be hiding a secret cache of weapons (McGhee, 2015).

This militaristic language is representative of our certainty that "show me, don't tell me" must be true and the extreme suspicion with which subjectivity in data visualization is held. We rely on data visualization for so much: reporting on sensor networks, displaying critical data, facilitating communication, and helping us make good decisions, both in research and in our everyday lives. Consider, for a moment, life without Google Maps or GPS navigators. While geographic maps are a convenience that we would miss, data visualization also affects our democratic rights in fundamental ways. In the 2000 presidential race, for example, the poor visual organization of the butterfly ballot in Palm Beach County in Florida is said to have caused thousands of voters to vote for someone they did not intend to, costing Al Gore the Presidency (Wand et al., 2001). The thought that so many these valuable functions depend on a subjective, fallible medium is uncomfortable.

I'll explain the persuasive power of visualization in order to demystify it. This power relates to the potency of immersive

communication as compared with literate communication. Literate communication is of course related to language, reason, and critical thinking. It encompasses all forms of communication that require fluency in written, spoken, or read language. While most communication requires some degree of literacy, certain forms of communication depend on it much more than others. Academic articles, long-form journalistic writing, and detailed instructional manuals are all forms of communication that depend primarily on literate communication. This concept of literacy is the one communication researchers thrive on, and feel most comfortable with.

IMMERSIVE, EXPERIENTIAL COMMUNICATION

Then there is immersive communication, the experiential quality. It's related to aesthetics and emotion. All communication leaves us feeling a certain way, due to many small details that affect our emotional reception; this emotional response is the immersive component. When listening to someone, the cadence, tone, and volume of the words spoken all affect our emotional response to those words. When reading highly literacy-dependent communications such as academic articles, immersive aspects of the communication include the quality of the paper the text is written on, or the brightness of the screen, the font style, size, and weight, and the compositional layout of the article. Text that is published in a hard-to-read font, too-small font, or laid out with lines too close together tend to illicit negative responses irrespective of the content. Generally, the more visually or aurally dependent a communication is, the more it relies on immersive communication. Data visualizations, in-person debates, and video footage are all communication formats that rely heavily on immersive communication.

The potency of such immersive communication is connected to the seductive power of emotions and the ancient fight or flight responses in our brains. Strong emotions trigger what psychologist Daniel Goleman refers to as "emotional hijack" (2012) in which the reasoning parts of our brains are temporarily overridden by emotional responses, which come with overwhelming feelings of righteousness and certainty. The triggering of this physiological

response by immersive communication is the reason advertising is effective, the reason we have art, and the reason we love music. Immersive communication is uncomfortable to contemplate for most researchers. Some may even consider the phrase "immersive communication" a contradiction. But its existence is undeniable.

In the anecdote I relayed earlier, my friend had no recollection of the literate aspects of the data visualization he saw on the subject we were discussing. But he clung firmly to the visceral experience of the immersive communication it conveyed. Immersive communication in research is unavoidable simply because the immersive aspect of communication is unavoidable. No matter how clinical or remote a communicative exchange is, we're left with an experience of it.

Most research processes and outputs focus on the literate aspects of communication, with little to no attention given to immersive aspects. For example, visual presentation of research findings in academic articles is most commonly kept to a minimum and used exclusively for illustrative purposes. It is possible to communicate complex research findings and arguments in visual formats such as maps, infographics, and charts. But primarily visual formats are not considered acceptable research outputs by most academic publications unless they are presented as supplements to an argument written out in standard academic language. This situation results in research findings being experienced as flat, and boring.

WHY DATA VISUALIZATIONS ARE ENJOYABLE

Data visualizations are so compelling because the medium demands equal footing between literate and immersive communication. Without significant attention given to immersive visual elements, such as color, symbols, and spatial relationships, data visualizations are unintelligible. Individual colors, color combinations, line thicknesses, patterns, icons, font styles, weights, and sizes all affect our emotional response to communication. Consider, for example, the differences between presenting statistical information about US states in a black and white table in an academic article, and that same information presented online in

a geographical map with statistical significance represented by color tints. The considered selection and spatial arrangement of these immersive elements contributes to data visualizations being inherently more pleasant than most research outputs, both to look at, and to work with. And so, we like them. It's this liking that makes data visualizations inherently more persuasive. More productive of experiential, visceral responses that sway our beliefs in ways we're either completely unaware of, or only vaguely realize.

Subjectivity and persuasion are inevitable in big data visualization. But persuasive doesn't necessarily mean inaccurate or immoral. Immersive communication is inherently persuasive simply because it is emotionally, rather than rationally, parsed. Despite what common terms such as "visual language," "visual grammar," and "visual rhetoric" suggest, visual elements cannot be parsed in the same way as language. Visual presentation of information is far more complex, and far less rule-bound, than language.

In the mid-twentieth century, communication research was dominated by the idea that all communication, including visual communication, was as rule-bound as language. This idea was epitomized in semiotics, a theory that envisaged all communication as one-way transactions involving three elements: sender, message, and receiver. While a few researchers still hold this view, many others have repeatedly shown it to be false. Across many fields—anthropology, communication studies, graphic design, and sociology—the prevailing understanding is now that communication depends on interpretation and mediation. A vast array of individual, cultural, and environmental factors contribute to visual meaning in any given context. These include associations with particular visual elements stemming from cultural and subcultural, community-wide, societal, and national traits. There are also important institutional contexts.

COLLABORATION IS KEY

No one, no matter how expert in a particular visual discipline, can understand all uses of particular visual conventions. No amount of visual comprehension at one particular moment in time guarantees

continued comprehension. Visual meaning changes at the same pace, and as intricately, as human culture and physical environment do. For example, two visualization strategies in common usage today, fever charts and donut charts, were unknown only decades ago, while many other charts have fallen out of common usage.

The changing nature of visual meaning affects individual elements within visualizations too. The symbolism of color, for example, is different and at times opposite, from culture to culture and also changes over time within cultures. For example, red is regularly used in so-called "universal" signage worldwide — in airports, train stations, and many public buildings — to indicate danger. While this makes sense in a western context, where red has traditionally been identified with anger and danger it is not so intuitive from a Chinese perspective since red is more commonly associated in Chinese culture with good fortune and luck.

These varied and shifting meanings are part of the reason why data visualization is such a collaborative and interdisciplinary field. Collaboration allows multiple perspectives as well as multiple skill sets and disciplinary backgrounds. Data visualizations draw from the disciplines of graphic design (my area), interaction design, technical communication, data analytics, statistics, math, psychology, and computer science. Collaborations between artists, designers, historians, journalists, scientists, and technical communicators who draw from these disciplines make best practice big data visualizations.

Working with big data shifts the role of data visualization in research. It moves from an optional research output to a necessity for data exploration. With small datasets of say, hundreds or even thousands of data points, it is possible for many researchers to navigate the data in a numerical format to identify patterns or answers to research questions. In a big data context, where data points are at least in the millions, this kind of scanning of numerical data becomes a physiological impossibility. Instead, researchers must rely on visualizations to query and summarize data in order to answer research questions. The subjective, persuasive practice of data visualization becomes an essential part of the research process.

TWO DIRECTIONS FOR BIG DATA VISUALIZATION

Working with big data also changes the nature of visualization. Because big data sets are, by definition, harder to comprehend than small data, big data visualizations need to be better at making data comprehensible. Two trends in best practice in big data visualization, bear this out: interactivity, and real-time updating.

Trend 1: Interactivity

Interactive elements help researchers explore their data sets at various levels of complexity and in various spatial configurations. An increasing number of data analysis tools — NVivo and R, for example — contain interactive visualization modules that allow researchers to visualize a particular set of data points in many different spatial organizations. By hovering or clicking on data groupings and individual data points, the researcher is lead to other visualizations of his or her data, providing new perspectives that may not have been arrived at without this interactivity. Being able to switch between so many views quickly and easily allows researchers to explore their data in previously unimaginable ways.

Still other tools have been created — Graphiq, Immersion, and Tableau, for example — that use interactive visualizations as the primary form of data analysis. These tools hide the numerical complexity of data, instead presenting interactive visualizations as the sum of research findings. Each of the sometimes thousands of individual perspectives enabled by these interactive data visualizations temporarily reduces complexity, cutting through otherwise impenetrable fogs of data, and each providing a unique perspective (Salvo, 2012). This increase in data visualization interactivity has been made possible through a combination of advancements in interaction design, software usability, and Web 2.0 technologies.

Trend 2: Real-Time Updating

The second trend in big data visualization is real-time updating. These interactive visualizations incorporate new data added to datasets instantaneously, through API access to publicly available online databases. This technology allows visualizations to become

integrated with data generating systems, extending the value and therefore shelf-life of visualizations well past their development stage. These are the visualizations that become valuable exploratory research tools.

Interactivity and real-time updating transform visualization, which has traditionally been a two dimensional exercise using static graphics. These traditional, static graphs and maps are snapshots of data frozen in time. They can be very effective research findings; they can even be beautiful. But best practice big data visualization has moved beyond the static page to applications and websites that display live feeds of data continuously in four dimensions, the fourth dimension being time.

At their best, big data visualizations put data into a human context by relating scientific and statistical insights to environmental and social contexts. They highlight perspectives about our world, and our societies, that we can collectively benefit from. Most importantly, they build understanding between users and the people who interact with the data generating systems we study by fostering respect and empathy for people and situations of which we would otherwise be unaware.

Example 1: Histography

The "Histography" project by interactive designer Matan Stauber (see Figure 1), for example, visualizes the entire contents of Wikipedia as an interactive timeline. Shaped like a reflected histogram, it records every page on Wikipedia as a single dot.

Figure 1: Histography big data visualization site, organizing. Top image shows one event on rollover, bottom image shows extra information on click.

Videos and photos are pulled from Wikipedia, and converted into enticing visual rollover effects for each documented event. Histography encourages us to choose pages randomly, and we're rewarded with surprising footage of people, experiments, and events from the past. Clicking on a rollover image brings up both more details from Wikipedia and the option to read the entire article within Histography. As new articles are added to Wikipedia, Histography automatically updates. Upon selecting a date range,

Histography highlights the events within that range that are mentioned in the greatest number of Wikipedia pages.

By presenting such a comprehensive timeline, Histography contextualizes important historical events and scientific discoveries in a way that easily counteracts revisionist histories based on religious beliefs and pseudo-science. It debunks myths at the same time as being entertaining, thoughtful, and fun. Histography also presents a completely original way to investigate historical events, one that gets around a significant challenge in historical investigation: avoiding confirmation bias, the unconscious tendency to look for information that confirms one's preconceived ideas. Histography is by no means perfect. Given that Wikipedia, on which it's based, is renowned for unreliable content, Histography can't be used as an authoritative source. But it does show the potential of big data visualization to be accessible, enjoyable, beautiful, and in the public interest.

Example 2: The Drone Papers

At their worst, big data visualizations help a few corporations make vast profit off of human suffering. The big data of modern surveillance and warfare is also big business. This is the side of big data revealed in the Drone Papers, obtained by investigative news outlet *The Intercept*. The "Finish ops" represented on this map (see Figure 2) are operations to assassinate individuals using missiles. This is the big data of dehumanization.

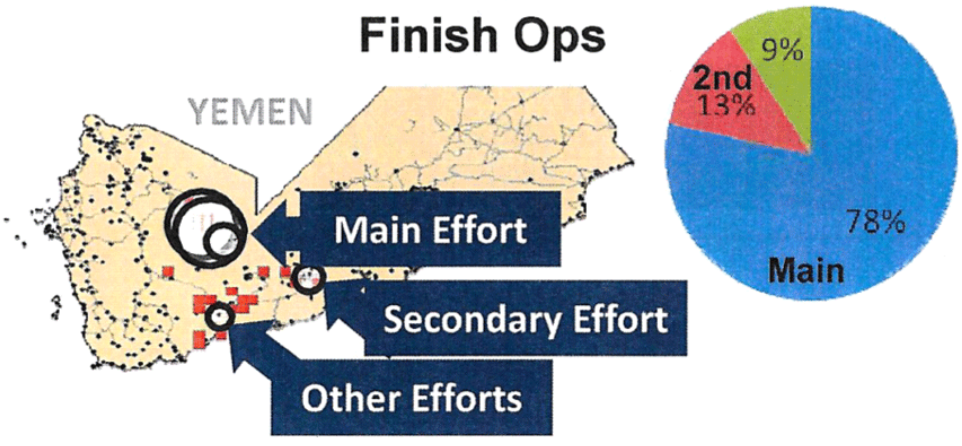

Figure 2: Images from the Drone Papers, US drone program documents obtained by *The Intercept*. This image shows drone strikes in Yemen.

In another image from the Drone Papers (see Figure 3), a similar kind of dehumanization is created visually.

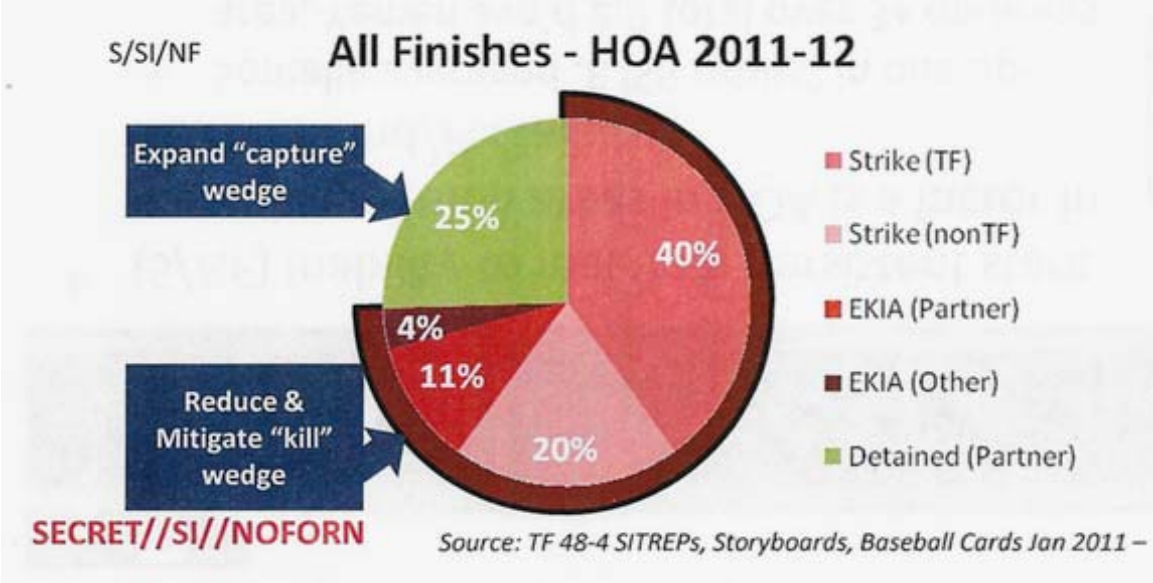

Figure 3: Images from the Drone Papers, US drone program documents obtained by *The Intercept*. This image shows deaths and injuries from drone-delivered missiles.

In this image, E.K.I.A. stands for "enemy killed in action." This is the default government classification for everyone killed by United States-operated, drone-delivered missiles. The word "enemy" is used despite the US military's own finding that only 10% of the people killed in this way are their intended targets. Civilians, bystanders, ordinary people going about their everyday activities, just like you and I, are listed as "enemies killed in action." It bears repeating because it's so unbelievable. These big data visualizations are shocking only in their banality. The language is ambiguous at best, and intentionally misleading at worst. The visualizations are crude and simplistic, obfuscating the big data underpinnings of the drone program. And I argue that they're intentionally so, reducing such large scale loss of life, to administrative details.

Most big data visualization falls between these two extremes. Frequently for profit, but without causing harm. One growing trend, in big data visualizations falling between the extremes, is combining interactive technology with design and gaming principles to incorporate big data into interactive, entertaining narratives.

Example 3: Interactive Wall, US Tennis Open 2013

For example, design agency Hush made a fifteen-foot interactive wall for attendees of the US Tennis Open to enjoy. The interactive wall allowed tennis fans to test their knowledge of player statistics in gameplay that is something like Wii Fit meets Tinder. This beautiful, fun interface wove vast sums of tennis data, that would otherwise be overwhelming, into an inviting and enjoyable experience.

The interactive wall epitomizes the pleasurable potential, as well as the economic realities, of big data visualization. Created as an elaborate promotional tool for IBM and the US Open, the interactive wall was only enjoyed by those who could afford expensive tickets to the exclusive sporting event. This one example is representative of the elitism and profit motive currently driving corporate research into big data visualization (Schwartz, 2015).

A CALL FOR EMPATHETIC BIG DATA VISUALIZATIONS

While big data visualization has the potential to benefit large amounts of people and tackle global issues, current development and spending indicates this won't be its main use. Instead, big data visualization is likely to continue to be developed for the use, and benefit, of relatively few, privileged communities. For better or worse, we are one of those communities.

In this piece, I've provided examples of a few different uses of big data visualizations. To keep the inevitable persuasion in big data visualizations within moral bounds, we must apply the same ethical rigor to our visualizations as we do to other aspects of our research. Renowned information designer Alberto Cairo has advocated for development of a field of "ethics of data visualization" (Cairo, 2014). He argues that the fundamental goal of data visualization should be to make people better informed. I would add to this that big data visualization should help people gain empathy for each other's situations. Big data visualizations can only fulfill this aim when the teams working on them have large enough, and critical enough, understandings of the

persuasive, visual elements from which they are built. When we're well informed, we can use the persuasive qualities of big data visualization to better inform others and to foster empathy.

We have a responsibility to educate ourselves about the literate and immersive aspects of various design elements within the communities where we want to share data. We owe this to ourselves, the agencies who fund our research, and the public. We also have a responsibility to gain enough understanding of the practical and ethical considerations of visualization in those communities, to wield the persuasive power of big data visualization for the common good. An important step on this path is acknowledging the persuasive nature of big data visualization and the inherent risks in its use. To understand, and take seriously, how visualizations affect both our conscious and unconscious judgments is to reduce the risk of misuse.

REFERENCES

Cairo, A. (2014). Ethical Infographics. *IRE Journal*, *37*(2), 25–27.

Goleman, D. (2012). *Emotional Intelligence: Why It Can Matter More Than IQ*. Random House Publishing Group.

Rose, N. (2008). *Powers of Freedom: Reframing Political Thought*. Cambridge, MA: Cambridge University Press.

McGhee, G. (2015, October 16). The "Rules" of Data Visualization Get an Update. Retrieved from http://news.nationalgeographic.com/2015/10/151016 data-points-alberto-cairo-interview/

Salvo, M. J. (2012). Visual Rhetoric and Big Data: Design of Future Communication. *Communication Design Quarterly Review*, *1*(1), 37–40. https://doi.org/10.1145/2448917.2448925

Schwarz, J. (2015, October 23). Drones, IBM, and the Big Data of Death. Retrieved from https://theintercept.com/2015/10/23/drones-ibm-and-the-big data-of-death/

Wand, J. N., Shotts, K. W., Sekhon, J. S., Meban Jr, W. R., Herron, M. C., & Brady, H. E. (2001). The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida. *American Political Science Review*, *95*(4), 793–810.