Middlebury

Alex Lyford
Assistant Professor of Statistics
Department of Mathematics
alyford@middlebury.edu

Warner Hall 210
Middlebury, VT 05753

June 2022

Hello, Alumni College Students,

I'm excited to meet you all in September, and thanks for your interest in learning more about big data! Data are becoming more and more ubiquitous, as are the technical and ethical challenges that come along with them. Although it might seem intimidating, many of the common methods of data analysis and visualization are approachable without any formal mathematical, statistical, or computational training. In this course, we'll learn about a range of techniques for visualizing and analyzing data—both responsibly and irresponsibly!

We'll start by thinking about how data are visualized. To preempt our discussion, I would like you to read the article *Make Me Care: Ethical Visualization*, attached to this letter. I also ask you to look at several of the graphs in the *What's Going On in This Graph?* section of the *New York Times*. Are there any that speak to you in particular? Are there any that you don't understand or that you think are fundamentally misleading?

Next, we'll think about how data are collected from a variety of sources and consider how the ways we collect data might affect the things we can responsibly and ethically conclude. To help you think more about the ways we collect data and the potential ramifications, please read the *New York Times* article "They Stormed the Capitol. Their Apps Tracked Them."

Finally, we'll demystify some common data analysis techniques, including a survey of statistical models and machine learning. We'll learn why these approaches are so powerful, but we'll also consider the assumptions they make about the data used to train these algorithms. To see why we need to tread carefully, I encourage you to read the *Time* article "Google Has a Striking History of Bias Against Black Girls," which emphasizes the need for careful consideration of the way we implement these kinds of models. (As a warning, this article contains explicit text content.)

I'm looking forward to a fun Labor Day Weekend and meeting you all at Alumni College! Please feel free to reach out to me if you have any questions or if you'd like to introduce yourself ahead of time.

Can't wait!

Alex Lyford
Assistant Professor of Statistics

Make Me Care: Ethical Visualization for Impact in the Sciences and Data Sciences

Katherine J. Hepworth(✉)

University of Nevada, Reno, NV 89509, USA
khepworth@unr.edu

Abstract. Scientists and data scientists have long aspired to eliminate bias from their visualizations. This paper argues that eliminating bias from visualizations is impossible, efforts to do so have negative real-world consequences for people, and that strategically emphasizing bias in visualizations is not only desirable, but also ethical. The growing public mistrust in science has not been helped by efforts to produce visualizations devoid of bias. This paper further argues that ethical visualization can only be achieved by acknowledging and embracing the treacherous nature of data visualization as a medium, committing to an ethics of care in visualization, investigating the potential for both benefit and harm when visualizing specific data, and then employing strategies to mitigate the harm involved in creating, using, and sharing visualizations. These strategies center around consciously crafting a visual frame (ie bias) for communicating data to a given audience. This paper offers 1) a critical lens on the rhetorical nature of visualizations, 2) the Hippocratic oath as a means of committing to maximizing the benefit, and mitigating the harm, done by visualizations, 3) ethical visualization for impact as a practical strategy for taming treacherous visualizations, and 4) compassionate visualizations as the end goal of following ethical visualization practices.

Keywords: Visualization · Care ethics · Science communication · Compassion

1 Introduction

Visualizations amplify the biases, constraints, and ideological perspectives of the people, organizations, and cultures they originate from. They do this in the form of arguments that are particularly persuasive firstly because they are processed largely unconsciously, and secondly because they present their bias in the guise of impartial, scientific truths. These arguments are usually constructed unconsciously, too, and herein lies the potential for harm. Whether they display big data or small data, whether they are online and interactive or printed and static, whether their intended purposes are scientific, artistic, or somewhere in between, visualizations without consciously crafted arguments always persuade, in the sense of projecting "a set of beliefs about the way the world should be, and present[ing] this construction as truth" [6]. At best, unconsciously constructed arguments in visualizations have a counterproductive effect on what the visualizer is trying to show; at worst they perpetuate harmful, counterfactual public narratives that increase mistrust in science and societal division [14].

A. Marcus and E. Rosenzweig (Eds.): HCII 2020, LNCS 12200, pp. 385–404, 2020.
https://doi.org/10.1007/978-3-030-49713-2_27

Visualizations can be made more ethical—that is less harmful and more effective—by applying reflexivity and critical thinking to the process of constructing them. Ethical Visualization for Impact, the main subject of this paper, is in part a call for scientists and data scientists to pay attention to, and harness, the amplification effect and bias inherent in visualization as a medium.

For example, the NOAA Interactive Sea Level Rise Viewer is a publicly available, online interactive geographic visualization created by the US National Oceanic and Atmospheric Administration (NOAA) to try to educate people about the risks to their communities and homes from climate-driven sea level rise [30]. Several governments and organizations around the world have created similar tools. This one contains lots of data from atmospheric physicists, biologists, oceanographers, and social scientists about the relationships between climate change, sea level rise, and population.

To demonstrate how it works, I'll use the example of the city Norfolk, Virginia in the United States. In terms of population, it's about the same size as Copenhagen, the host city for HCII 2020. Like Copenhagen, it's near the sea, and built around a river. Using the viewer to assess the risk to Norfolk of sea level rise involves navigating a lot of options in the viewer (see Fig. 1). The options are grouped in six key categories that relate to data collected from various research disciplines. The user can move the slider on the left up and down, and the map shows which parts of the city would be under water if the sea level rose. Navigating the variables, it becomes evident in a short amount of time that major areas of downtown Norfolk would be underwater in many scenarios. In fact, Norfolk is one of the most vulnerable cities in the Americas to sea level rise. There's a lot of the city shown as underwater in the view in Fig. 1. Seeing this, I would be tempted to move away, or at least look for a place to live on higher ground. To an impartial viewer, this seems obvious, right?

Technical communication researchers Sonia Stephens and Dan Richards wanted to test this seemingly obvious conclusion, so they conducted usability studies on a sea level rise viewer with people in Norfolk. They found that people consistently made decisions that were less in their interest, and less based on scientifically accurate information, after interacting with a sea level rise viewer [31, 34]. Although their research involves only a small number of participants, the implications of it are huge. A lot of time and money have perhaps been wasted on sea level rise viewers that not only don't do what is intended, but they also put people in greater harm and lead to greater misunderstanding. So, how did this happen?

Richards and Stephens work highlights a present-day, real-world risk of the common assumption, especially when it comes to visualization of data science and scientific findings, that "numbers speak for themselves" [15]. This assumption, rife in computational and scientific research, is based on other assumptions: 1) that what is obvious to researchers will be obvious to their audiences; and 2) that the medium—data visualizations themselves—are objective, something like a blank page on which data can be placed in order to be understood. A growing body of literature demonstrates how untrue this series of assumptions is.

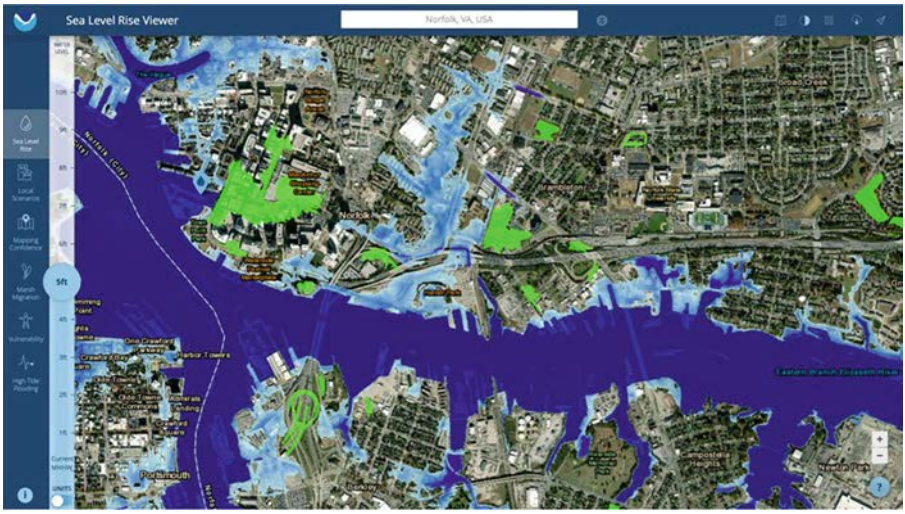


Fig. 1. NOAA Interactive Sea Level Rise Viewer, 2017, 5 feet scenario, zoomed in on Norfolk, Virginia, United States. Visual frame: Sea level rise is too complex and abstract to worry about.

2 Treachery in Visualizations and How to Transmute It

For over a century, scholars working with visualizations have regularly acknowledged that data visualization—here loosely defined as data-informed charts, diagrams, maps, models, and tables with a considered use of visual composition for the goal of impactful communication—is fraught with pitfalls. Visualizations contribute to, reinforce, and thereby amplify, cultural assumptions and stereotypes. For as long as there has been explicit instruction on how to create data visualizations, there have been warnings about the perils of the medium. In his *Graphic Methods for Presenting Facts*, widely recognized as the first instructional manual on data visualization, Engineer Willard Brinton, describes in detail how data visualizations have the potential to mislead [2].

Various subdisciplines are now devoted to critiquing the biases inherent in visualizations in multiple fields, including: critical infovis in computer science [10], critical cartography [7, 27–29] and critical GIS [8, 24] in geography, feminist data visualization [11, 12], and ethical data visualization [5, 21, 23]. While some of this literature makes distinctions between visualizations with various levels of haptic engagement, interactivity, or physicality, [7] the definition used here is intentionally expansive enough to include a wide variety of visual representations of data, as *visual means and the goal of having an impact are the key qualities of visualization addressed in this paper*. The following subsections supplement these understandings with a detailed breakdown of the mechanics of argumentation in some key visualization forms, content types, and design elements, demonstrating how these individual elements influence attitudes individually, leading to a greater understanding of how they can reinforce one another to present a cohesive and powerful argument when they are all working in concert.

2.1 Argumentation Through Form

All visualization forms contain the historical and contemporary cultural associations of their invention and prior uses [16, 28, 34]. The shadows of the arguments that were contained within early and common examples of the visualization form linger, to some degree, in present-day uses, even when they contain very different data or are presented in a different context. When used unconsciously, for example, by unquestioningly accepting the "suggested" visualization form after inputting raw data into a partially automated (black box) visualization software, argumentation can be introduced into the new visualization that contradicts the visualizer's goals. However, when prior associations are considered, they can be harnessed to emphasize and amplify goals.

For example, consider Dmitri Mendeleev's periodic table of elements in an early form, the present-day version by the Royal Society of Chemists [31], and Marisa Bate's Periodic Table of Feminism (see Fig. 2). The original periodic table of elements functioned as a convincing argument, or frame, about groups (or periods) of physiochemical properties of atoms being able to be predicted by atomic weight and valence. Since its first introduction, the periodic table of elements been expanded, the visual composition has been formalized, and it has become widely celebrated and used as an instructional tool. It has taken on a broader positivist cultural meaning and argument through these changes [3].

The Royal Society of Chemists' version is an example of a contemporary periodic table, with inclusion of later discoveries (such as inert or noble gases), some reorganization of individual elements (now organized by atomic number), further categorization of elements into groups and blocks, and color coding and visual organization to emphasize the relationships between elements and categories. It supplements these understandings with extensive interactive effects, including the ability to learn about each element through element-specific podcasts, videos, and scientific discovery narratives. This well-known, widely used contemporary periodic table conveys the visual frame that the building blocks of matter are fundamental, orderly, mastered by science, and intelligible to all to seek to know them. Bates' Periodic Table of Feminism leverages this association, applying the same positivist logic to her subject: the history of feminism. Her periodic table, by utilizing the visualization form of the better known periodic table, combined with her compositional and categorization choices, makes a visual argument, or visual frame, that feminists thinkers are fundamental to society—as fundamental as matter is to science, and that they too, are orderly, and intelligible to all who seek to know them.

2.2 Argumentation Through Content

At a more minute level, individual pieces of content and design decisions used within each visualization also contain argumentative qualities and associations that further the overall argument contained within a visualization. Content types and design elements that are explicit about their argumentation are titles, captions, annotations and emphases. Implicitly argumentative content types and design elements include data breaks, categorizations, scale, priming, nudges, priming, sizing, and proportional ink.

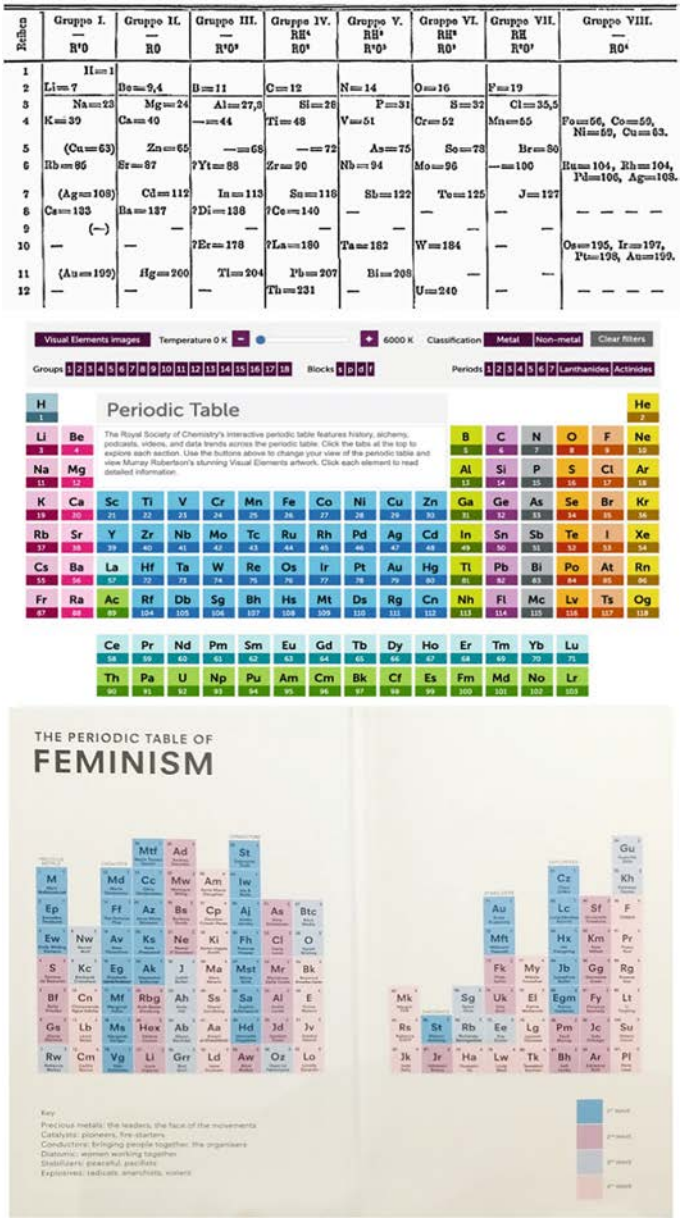


Fig. 2. Periodic tables. Figure 2a (top) Mendeleev's periodic table of elements, 1871. Visual frame: groupings of physiochemical properties (periods) of atoms are able to be predicted by atomic weight and valence. Figure 2b (middle) Periodic table of elements, 2019. Visual frame: the building blocks of matter are fundamental, orderly, mastered by science, and intelligible to all to seek to know them. Figure 2c (bottom) Periodic table of feminists, 2019. Visual frame: feminist thinkers—and by extension, feminism—are as fundamental to society as matter is to science, and they too are intelligible to all who seek to know them.

Titles. Titles and subtitles are the most explicit use of rhetoric in visualizations found in visualizations, both because we are used to reading text as rhetorical, and because titles are usually emphasized to the extent that they are one of the first things people see when looking at a visualization [12]. For example, in Fig. 2a, Mendeleev references his argument in the title of his pioneering visualization. This is despite the term "periodic" not being in common use to describe matter at the time – the naming of his visualization contributed to both the framing of his visualization as well as future understandings of the nature of chemical elements. While a title can make reference to an intended argument, subtitles have the advantage of being able to elaborate upon, and reinforce, a visual frame. They provide additional details and are often able to be longer than captions (in part because they tend to be written in smaller type size, so a greater number of words can fit in the available space), and therefore additional capacity for explicit argumentation.

Captions. Captions are nearly as explicitly argumentative as titles. A caption functions as a linguistic frame for the entire visualization, highlighting what is important to the visualizer, what they want the user/viewer to notice in the visualization [24]. For example, in Fig. 2, I have included each visualization example's author, title, and year —they are the elements I want readers of this paper to contrast and compare, in order to strengthen my own argument. When visualizations are shared on social media, the textual content of the post (including hashtags and emojis) functions the same way as captions. In this context, the number of likes, shares and replies operate to reinforce the argumentative power of the caption content.

Annotations. Annotations can draw attention to key data points and relationships, guide a user through a given sequence of visualizations, and scaffold a user's journey through a visualization; in each of these roles they linguistically and visually direct a user's attention and frame both their interaction activity and the importance they place on annotated elements of the visualization.

Symbols. Symbols used to represent or supplement data in a visualization—such as icons, illustrations, and pictograms—provide simplified visual cues that serve as a shorthand for quickly communicating key concepts. They foster argument by clearly communicating what the visualizer wants the audience to know most, and through the priming effect they have on audiences, conjuring up related visual cultural references and their associated status and power relations. However, they also pose the risk of communicating a universalist sensibility inherited from their modernist origins as pictograms and their cartographic origins as map markers [1, 16].

Emphases. Emphasizing key elements using visual hierarchy, centering, ordering, and negative space. Emphasis is very important for both the intelligibility of a visualization, and for the efficacy of the argument it presents. However, uncritical use of emphasis comes with the risk of fostering "inward-directed worldviews, each with its separate cult centre safely buffered within territories populated only by true believers" [13]. In saying this, art historian Samuel Edgerton was referring to cartography, although his observation applies to visualization generally.

Visual hierarchy draws attention to elements in a specific order, offering a subtler argumentative effect in with a similar effect to annotations. It is the coordinating design

technique that combines the following strategies into a cohesive and intelligible whole. Centering, or determining what is positioned at the center of a visualization upon initial or default view, can be used as a way of declaring what is of most value, and studying what is centered can reveal the biases and world view of the visualization creators and sponsors [16].

Ordering data representations within a visualization includes a vast array of techniques encompassing map projections, data bins and polygons, and arranging data within one or more structured lists. These choices are typically made for pragmatic, logistical purposes in the moment, but nevertheless, like centering, they reinforce social norms of the time and culture in which they are created and commonly used, as in the much-cited case of the Mercator Projection [29]. Ordering information also determines what will be read first and last in a series of data, and therefore what will be given importance by viewers and users subconsciously, by harnessing serial position effects. For example, ordering information in a bar chart alphabetically will compositionally emphasize items with labels starting and ending at the ends of the alphabet.

Areas of a visualization that do not contain data or ornamentation, known in communication design and photography as negative space (or the more problematic, white space) and in critical geography as "silences", convey argument and emphasis through omission [16]. In communication design practice, the use of negative space is commonly understood as a functional technique of emphasis that fosters efficacy; the more space without design elements in a given visualization, the more attention is drawn to those that are present. However, decisions about what is most important to visualize are also decisions about what doesn't matter enough to be communicated. Omissions and negative space communicate a hierarchy of power that perpetuates and amplifies discrimination and cultural biases endemic to the cultures, institutions, and visualizers they are produced by.

Implicit Argumentation. Although the subtler forms of argumentation are seldom obvious, they are nevertheless imbued with meaning, persuading audiences of arguments they are barely aware of in nuances of representation such as data breaks, categorizations, nudges, scale and sizing, and priming. For example, data breaks, or the places where the individual portions of a scale range are cut off, are subtly perceived as significant, whether or not they are carefully considered by the research team. Where data breaks use solid colors to delineate individual data groupings, the data groupings are perceived as more discrete, or separate, than when the visualization uses a gradient color ramp with data breaks. Categorizations also contain subtle argumentation; which categories are chosen to visualize form a visual frame of what is important in the visualization, and the language found in labels used to describe categories reinforce this frame. Nudges, or interaction effects that provide a sort of animated motion-based visual annotation, are implicitly argumentative in that they cajole the user into certain behaviors within the navigation of the visualization, while dissuading others [33]. Scale and sizing of individual elements and the visualization itself both impact audience perceptions about what is significant in the visualization, and how the other elements in the visualization can be, or should be read. Priming, the effect of conjuring up past emotional and visual associations with design elements that reference other cultural and visual associations, is perhaps one of the most pernicious implicit forms of

argumentation in visualizations. For this reason, some scholars rail against use of ornamentation in visualizations [26].

2.3 Committing to Doing No Harm

Each of the elements described above, and their associated argumentative capacities, can be harnessed to present a powerfully coordinated argument. With such power, comes great potential for harm, so much so that some visualization scholars advocate taking *a "do no harm" approach to visualization practice* [4, 18]. If we are to follow a Hippocratic oath approach to doing no harm in visualization, this means both not hurting people, and helping people, when they are in three types of relationships with visualizations. Firstly, *audiences of users of visualizations; doing no harm to these people includes not misleading them*, while helping them includes cultivating their empathy for represented subjects, and using the smallest feasible amount of their cognitive load to communicate effectively (and therefore, not making visualizations harder to navigate than is absolutely necessary). Secondly, *subjects whose data is represented in visualizations; doing no harm to these people includes not harming the flesh and blood humans behind the data*, as well as communicating their personhood, effectively representing their humanity. Thirdly, *people who are significantly impacted by the represented data*; these are people who have a close connection to the represented data, for example, descendants of people whose data is depicted in a given visualization. Doing no harm to these people means *treating the data with a culturally-appropriate sensibility and effectively communicating the humanity of all represented subjects.*

3 Ethical Visualization for Impact

While tactics for avoiding egregious manipulation in visualizations have been suggested by visualization scholars and practitioners, and general principles have been established, few strategies have been offered that can integrate ethical tactics and principles into the day-to-day, practical task of visualizing data. In 2018, with my collaborator Christopher Church, I argued that the ethical visualization workflow could be used by digital humanities scholars as one such strategy [21]. Since the 2018 publication of "Racism in the Machine: Ethical Visualization in the Digital Humanities," this workflow has been iteratively adjusted and developed to be a workable process for fields beyond the digital humanities, and particularly, for scientists and data scientists. The work informing this adjustment has involved review of allied literature, research in historical archives of scientific visualization and information design, and discussions with data scientists, designers, and scientists who rely on visualization in their work. While this research is ongoing and will be further developed to take into account new discoveries in the burgeoning fields of visualization practice and visualization ethics, the Ethical Visualization for Impact process described below represents the current state of this work. It is intended to be of use to people visualizing data in scientific research and the data sciences [19].

The stages of Ethical Visualization for Impact take a complete communication context and whole data pipeline approach to mitigating harm in visualizations, in acknowledgement that decisions made throughout the process of planning, collecting and cleaning data all contribute to shaping the visual frame and resulting argumentation contained within a visualization.

3.1 Discover the Data

The process of sense-making about, or gaining insight into, data in a raw, or minimally processed form, within a research team (defined as an individual or group of people with intimate knowledge of the data, including details such as collection, sources, modeling structure, and processing). This discovery typically occurs immediately after, or in conjunction with, a data collection event (loosely defined as an experiment, data harvest such as a web scrape, or compilation of items such as survey data). Sometimes referred to as 'visualization for analysis,' this stage utilizes visual means to reduce the cognitive load demanded by a given dataset to a degree that is acceptable for the people who are intimately engaged with it, and therefore highly motivated to understand it [20, 22]. In this stage, understanding among the research team is fostered by visual identification of potential correlations, patterns, relationships, and trends. Activity in this step is guided by the question:

What can you find in this data?

The process of discovery in data is typically undertaken using one of two main avenues: 1) a reason-dominant approach, emphasizing investigation for seeking knowledge about the content of the raw data, or *2) an emotion-dominant approach*, emphasizing exploration, to get a felt sense for the meaning that might be made of the data [15]. Both approaches incorporate quick renderings of a range of visualization forms produced iteratively, with very rough visualizations produced quickly initially, typically progressing to more and more detailed versions of a visualization form that make sense of the data for the research team. Sense-making in this stage includes identifying potential correlations, patterns, trends, and outlying data points that could be significant to the research team.

For ethical visualization that has impact, a balance between investigation using reason-driven approaches and exploration using emotion-driven approaches is recommended in this stage. By shifting between analytical and playful modes of sense-making, the research team can gain an understanding of the scope of insights that may be relevant or valuable to their desired audience, as well as to themselves. It is worth noting that although these visualizations make sense to the research team, bringing understanding to a small group of people, this does not mean they will bring understanding to people with less intimate knowledge of the data. To have impact outside of the research team, the final visualizations presented to other people need to pass through many stages of iteration and contextualization in the steps below.

3.2 Scope the Impact

Determine the overall bounds of the visualization activity by identifying what the research team finds most important to the visualize, how important they find it, who they want to visualize data for, and which impact type is going to be most efficacious for reaching that audience.

Assess Stakes. Determining what is at stake in people understanding what the research team is has to share. This will motivate a research team to go through the additional stages required to communicate data effectively to an external (ie outside of the research team) audience.

What is at stake in this visualization effort?

Identify what might be lost or gained if this visualization effort is effective, ineffective, or outright misleading. This can be measured on a matrix of high, medium, or low stakes on various levels of impact: on global, societal, organizational, and personal levels. After determining the stakes, it is possible to make a realistic determination about what an appropriate allocation of resources, time, and effort to this visualization effort might be. In my experience, data scientists and scientists overestimate their available resources for data discovery and dramatically underestimate the time required to visualize findings effectively. A realistic look at the research team's available time, effort, and skills, combined with the stakes of the visualization's efficacy at this stage, will provide the research team with the opportunity to shift resources to the visualization effort. *Ethical visualization demands allocation of an appropriate proportion of the total available resources to visualization for impact.*

Establish Purpose. Noting the reason for visualizing data through this particular visualization. This is a short statement (approximately one sentence) of the research team's motivation for sharing data and intended impact in the research team's own words. It should reference the stakes identified above.

What is your motivation for sharing what you found in this data?

In this step it can also be helpful to identify counter purposes, ie what is the opposite of what the research team intended in sharing this data? Answers to both of these questions provide guides for activity in future steps and benchmarks to use in the last two stages. They are useful in the final stages *to measure the proximity between the impact intended by the research team, and the impact felt by their audience.*

Identify and Learn About Audiences. Determining who the research team wants to have impact on with this visualization is important early on. This can be as simple as identifying one group of people the research team wants to communicate to, or it can include more groups. In professional communication disciplines (such as communication design, marketing, and technical communication) it is common to identify primary, secondary, and even tertiary audiences, and the same can be done in this step.

Who do you want your visualization to reach the most?

Irrespective of how many audiences are identified, it is important to note that most visualizations are seen by many more people, and types of people, than are initially intended. For example, a visualization produced in a thesis may be intended for the eyes of a research student's supervisor only. However, that thesis will likely end up in a publicly available data repository, will be viewed and used by other researchers, and if the findings are significant, people in professions where the findings have value. Therefore, *identifying unintended audiences of a visualization is also helpful, and potentially important for mitigating harmful uses*.

After identifying audiences in order to effectively communicate with them, it is important to develop an empathetic understanding of those audiences. Identifying details about the audience can be done in three ways: interpersonally (by meeting and asking them questions), through profiling (by using a combination of available data on the audience and thinking tools), and through identifying distance from researchers. The first two approaches are professional skills that some professions take years to develop. Interpersonal approaches originate in anthropology and design research, and commonly rely on ethnographic interviewing. Profiling approaches originate in advertising and strategic communication, and commonly rely on building an audience profile that is synthesized from demographic and psychographic insights.

While working with people from these professions would be ideal in the case of large visualization projects with high stakes of global significance, and large budgets, a simpler third option is presented here for lower stakes visualization for impact. *If the research team can identify the distance between themselves and their audiences,* in terms of geography, language, discipline, age, expertise, cultural background, and prior familiarity with the subject, *they will gain good enough insight into the difference between the context in which they operate and the context in which their visualization is likely to be received.*

Determine Impact Type. *Five ways visualization can have an impact on audiences are: management, dissemination, entertainment, decoration, and expression.* Visualization for management has impact by aiding decision making. It is generally performed by people motivated to surveil and understand something. One micro level example is people wishing to surveil themselves (in the case of the quantified self movement), a medium level example is policy makers (using data dashboards to inform policy decisions), and a macro level example is found in military personnel studying security threats within and to a national population (in the case of human dynamics) [9, 18].

Visualization for dissemination has impact by fostering understanding of something the research team has found with people outside of the research team. For example, visualizations contained in research reports, research presentations, flyers, brochures, and promotional materials for research initiatives are all types of visualization for dissemination. Visualization for entertainment has impact by telling stories about the research team's data. Examples of visualization for entertainment include science-informed fictional movies, data journalism, and science education for children. Visualization for decoration has impact by crafting pleasurable experiences for people. Examples include when data-informed reactive artworks are put in public spaces, and when items such as fashion or soft furnishings (pillows, blankets, curtains etc.)

incorporate data visualization into the design of their surfaces or functionality. Visualization for expression has impact by demonstrating virtuosity in a given medium, or making fine artworks, and sometimes both. An example of this kind of impact is found in data visualization being increasingly used in fine art and contemporary art practice using a broad range of media.

Which impact type will be most effective for your audiences?

The first two impact types, management and dissemination, deliver impact primarily through information; relying primarily on reason, and using felt sense and aesthetic experience as secondary sources of information retention. The remaining types of impact – entertainment, decoration, and expression – deliver impact primarily through impression (felt sense and experience of aesthetics); they rely on experiential factors as the primary mode of communication and use reason as a secondary support. While ethical visualization will demand utilization of both information and impression, identifying the most effective mode for your audience, purpose, and stakes will guide the development of future stages. *Each of these impact types can be used either ethically or unethically. Following these stages of Ethical Visualization for Impact is recommended to prevent the unethical use of any impact type.*

3.3 Develop the Frame

Critically assessing the reason for data exploration and visualization, and forming both a verbal and visual sense of the appropriate expression of the frame. The frame is a persuasive statement (argument or explicitly stated bias) communicated through coordinated use of visual, textual, and experiential design elements that is believable and relevant to the audience. It integrates the discovered data and the intended impact of the research team with the needs and goals of the audience. The result of this stage is a persuasive summary statement or sketches that are research informed and intentionally persuasive, in order to harness the inherent biases in visualization that can amplify an ethically derived message.

Empathize with Audiences. Understanding of the intended and unintended audiences gained in the previous stage is expanded here into empathizing with your audiences, and summarizing their motivations and needs.

What are your audiences' greatest needs?

For each identified audience, immediate and existential needs can be identified by interviewing, observing, profiling, and surveying. The extensiveness of such audience research will be determined by stakes, and the resultant resources allocated to the visualization effort. Immediate needs are situational and usually reason-based, as in the case of someone who is running late for a meeting but needs to use a building map to find the meeting room. Existential needs are more abstract, emotionally mediated, and demand reassurance to enforce primal feelings of security and avoid primal feelings of abandonment. For example, members of a given audience might need to feel justified in a particular course of action, which can be thought of as a need for security. A person making a high-level policy decision may need a data dashboard to reassure them a

course of action is appropriate. On a more personal level, a person who is trying to give up smoking may need a visualization within a behavior change app to reassure them they are making positive progress.

Utilization of existential needs in crafting visual frames is very effective, and should be used with great caution. Much harm can be done by using this approach (in terms of fueling extreme emotional reactions about visualized subjects), and best practices in this particular area are yet to be established.

Formulate Goal. Stating the goal for a visualization incorporates insights gained in previous steps, combining the data, the research team's intended impact and the audiences' needs. It expands upon the purpose identified earlier in the process in a somewhat formulaic manner.

What is the goal for the visualization?

This answer can be formulated as follows: *[impact type] of [data content] for [impact purpose] to/for [intended audience] about [audience need] so that [purpose].* Contra-goals can also be identified in this stage using the same process, to identify worst case scenarios of what might happen if the visualization was used by the unintended audiences developed earlier in the process. *Ethical visualization is goal-driven, as well as contra-goal-driven, meaning that it considers both the potential for benefit and for harm engendered in the design of a visualization.*

Create Frame. Using the goal as the basis for creating a frame (a persuasive argument that is believable and relevant to the audience) around data. Sub-frames can also be developed in this process, to give further nuance to achieving the goal outlined in the previous step.

Which frame best achieves this goal?

Wherever possible, for example in cases where there is a short distance between the research team and the audience, and in cases where there is a large amount of resources available, the frame should be tested with the intended audiences to determine believability and relevance. Where testing is possible, the frame can be refined to increase its relevance to the desired audience. The end result of this process will be a frame in either the form of a persuasive statement or a sketched visualization form detailing key persuasive elements (such as title, caption, or data emphasis). The sketches produced in this step differ from those produced in 3.1 in that the focus of the visualization is now on communicating data outside of the research team.

Review Literature about Frame. Searching for, identifying, and reviewing available literature about the created frame. Various fields, including communication studies, health communication, and science communication offer detailed testing and analysis of common frames for specific audiences and contexts, as well as hypotheses about lesser known frames. Familiarizing themselves with this literature can provide the research team with valuable insight into what kind of framing that is similar has worked in the past.

What does the literature say on this frame?

The results of this interdisciplinary literature review inform the following steps, thereby allowing the research team "to compensate for the data set's shortcomings by seeking out and including new information, or to limit the scope of the visual argument to be produced with said data" [21].

3.4 Prepare the Dataset

Creating a custom dataset specifically for the visualization for impact from available, reputable sources. This a custom dataset, created for each visualization for impact both ensures the data is relevant to the created frame and minimizes the risk of perpetuating unwanted biases from other researchers, institutions and organizations by using their unaltered datasets. This custom dataset includes, but is not limited to, the data in the initial data collection event. Depending on the frame decided upon, the data from the original data collection event may be deemed far too extensive and in need of pruning. It will also likely need supplementing from other sources in order to be impactful.

Combine Sources. Creating a preliminary, visualization-specific aggregated dataset, by 1) identifying the data from the original data collection event that is most valuable for communicating the frame, 2) transferring that data into a new set, and 3) identifying reputable datasets and data sources to supplement the original data collection, and 4) adding as much potentially valuable data to the new, aggregated set as is feasible.

What sources will you draw from?

The search for reputable datasets and addition of data to the new, aggregated dataset is guided by the frame and goal identified in the previous stage.

Improve Veracity. Using computational methods to increase the accuracy of the aggregated dataset. This action is guided in this stage by the question:

Will the data hold up under scrutiny?

Computational methods are used to clean, normalize, and refine the aggregated dataset, and manual oversight of the results of these processes can identify accuracy and anomalies. The key to impactful normalization in the context of the visualization-specific aggregated dataset is supplementing existing data with *normalizing data that is meaningful to the intended audiences* [31]. The appropriateness and authenticity of data is also checked in this stage, and questionable data points are pruned.

Structure Data. Creating a working dataset by adding necessary descriptions and context in the form of metadata, as well as appropriate structuring for the required visual frame. For example, if the chosen visual frame includes a network chart, the data will likely need significant structural adjustment to make this possible.

Is the dataset intelligible and navigable?

Additional organization, supplementation, and categorization can be added in this step through processes such as topic modeling.

Refine Frame. Adjusting the frame based on the content of the visualization-specific dataset. The process of improving veracity and structuring data presents opportunities for different framing, and sometimes makes the original frame not possible to claim in an unaltered form. The research team explores the refined, custom dataset for the objective of answering the following question:

How does the frame need to be adjusted?

The chosen visual frame may need to be adjusted, given the available data. In some cases, the activities of this stage may have provided additional insights that strengthen the original conception of the visual frame. In others, the visual frame may need to be made more modest in its claims, when less supporting data has made it through the rigorous preparation process than would be needed to support the original frame.

3.5 Visualize the Frame

Identifying the most appropriate context, media, and medium for presenting the frame within the final visualization.

Review Literature About Ethics. Finding and reviewing the latest literature relevant to ethical visualization (in the range of allied fields discussed earlier). This enables the research team to learn about current best practices.

What are the latest ethical recommendations?

By completing this review as the first step of visualization, the research team can ensure they have up-to-date methods and are applying recommended best practices.

Determine Context. Considering contextual factors including the type of media most appropriate for presenting the visualization in, and the level of interactivity that will best support the created visual frame. Printed documents such as posters and reports can be much more directed in terms of who sees the visualization and may therefore be more appropriate for very sensitive or highly controversial visual frames. Online publication has both greater potential for impact and greater potential for harm, although neither is a given. Online publication is the hardest media in which to control audience reception. Higher levels of interactivity can come foster engagement and a greater sense of agency in audiences. However, as seen in the case of sea level rise viewers, this is not a given, and more interactivity can also lead to more erroneous conclusions.

What is the most appropriate media context for your frame?

Another contextual factor to consider is which data to visualize for the purpose of normalizing representations of the data. For example, including census data on the total population in a given place, alongside data about number of murders by firearms in the same place, is a normalized data representation. It is as important that visualizations be normalized as it is that data be normalized; *people make sense of the key findings presented in data visualizations by comparing them with associated data that is meaningful in the context of their day-to-day lives*.

An audience-related contextual factor is the design elements and media formats that are relevant, familiar, and engaging for your audiences. “The persuasive and culturally bound associations those audiences necessarily have with design elements, explanatory text, headers, legends and interaction experiences need to be considered. The choice of colors and color ramps, as well as graphic or cartographic elements like political boundaries, invariably influence the argument produced by the visualization, as do map default views at certain screen widths, and zoom options” [21].

Design Visualization. Designing and testing rapidly iterated visualizations and restructuring the dataset as necessary for creating more refined visualizations. This process starts by “creating test visualizations (these are more rudimentary than, and distinct from, alpha prototypes)” [21]. These visualizations are intentionally quick and intended to test how effectively the composition, data selection, normalization, and legibility communicate the frame.

What design decisions will visualize the frame effectively?

The extensiveness of testing undertaken in the design process should be determined by the stakes and available resources. After test visualizations, a range of prototypes are designed and tested, moving from alpha prototype level through various stages of refinement to a high fidelity visualization or set of visualizations. Selection of a set of final visualization forms is important in this step, as is selecting design elements (for example, colors, typefaces, grids, interaction affects, and transitions), functionality (for example, annotation, customization, filtering, transition between visualizations, and exporting), and guidance (for example annotations, captions, navigation pathways, and nudges) that communicate the frame developed in stage 3.3.

Test Visualization. Administering a final round of user-testing on the close-to-finalized visualization to determine likely reception and efficacy of the visual frame. The success of this step will be aided by testing in as close an approximation to the media, format, functionality, and context that the finished visualization will have.

Does your visualization communicate the frame?

Completion of a round of pre-release testing provides an opportunity for pre-release correction to improve efficacy, and also offers the opportunity to mitigate any potential unforeseen harms that may be discovered. For data, audiences, and contexts where stakes are particularly high, in this step the visualization can also undergo pre-release testing with unintended audiences who may encounter it in the published format. For example, a visualization published online is available to all internet users with sufficient computational power and internet speeds to view it, and therefore could be tested for unintended effects with a random sampling of people with access to such technology. Feedback on the visualization will allow tweaking of design elements and functionality to ensure the visualization communicates the frame as closely as possible.

3.6 Publish the Visualization

Publishing the visualization is the end goal of a standard visualization process. However, for ethical visualization for impact, the issue of measuring impact remains at this point, along with the issue of highlighting the visualization as particularly ethical.

Release Visualization. Publishing the visualization, and, wherever possible without doing harm, the dataset on which the visualization is based.

Will publishing base data do harm to any intended or unintended audience?

Where there are potentially harmful consequences of publishing the dataset, or where rights to the data are not owned by the research team, extensive citation of where audiences can find the datasets is important for demonstrating veracity and thereby increasing trustworthiness of the visualization.

Report Process. Publicizing the process undertaken to complete this particularly ethical visualization in "show your work" documentation.

How can your ethical process be best demonstrated?

This is important for increasing both the reputation of the research team and disambiguating the visualization from poor science and disreputable creators of similar visualizations [12].

Measure Efficacy. Administering further user-testing after the visualization is published, to measure its actual impact.

What is the felt impact of your published visualization?

Simple measurement at this stage can be in the form of surveys, while detailed testing could take the form of ethnographic observation, eye tracking, think aloud protocols, or qualitative interviews.

Feedback Results. Reflecting on the efficacy of the findings, including measuring the extent to which the published visualization's intended and actual impact are in line with one another. This is important for multiple reasons. It allows highlighting, and therefore correction of any major and potentially harmful experienced impacts from the visualization (such as a data dashboard that encourages users to make decisions against their own interests). It also provides valuable information to the research team about the differences required in their communication amongst themselves and with their audiences. This can be surprising, especially when the audiences seem close to the research team (for example, a colleague in the department) but have very different experiences and understandings of the visualization than the research team intended.

How do the visualization's intended impact and felt impact compare?

Research teams can also supplement their "show your work" documentation with this added information as it becomes available, providing periodic reporting on both the ethical intentions and rigor of the research team's communication efforts and their commitment to having significant broader impacts. When there is an avenue for the updates to visualization documentation, additional information can also be updated,

such as details of when new data is added, as in the "data updates" section of the NOAA Sea Rise Level Viewer.

4 Conclusion: Visualizations that Foster Compassion

Visualization is a technology that has impact because of its amplification effect that broadcasts the perceptions of its designers, data scientists, data repositories, institutional affiliations, and funders through visual and haptic means. Similar to the widespread critical commentary regarding artificial intelligence, data visualization can perpetuate negative personal and institutional biases, and, as in the case of sea level rise viewers, can at times appear to add support to inaccurate public assumptions. This is particularly true in cases when the arguments inherent in visualizations are not consciously crafted—when visualizers attempt to "let the data speak for itself." Visualizations have such capacity to be so problematic because they are human-made tools, and we ourselves contain a multitude of biases. However, as well as being riddled with biases, people are also riddled with compassion and empathy, positive qualities that can also permeate our human-made tools – including the visualizations we produce – if we let them.

The time for attempting to remove bias from visualizations is past. As our realities are increasingly shaped by computational processes mediated by visualizations, we need new, radically different approaches to visualization that prioritize ethical visualization practices in the service of producing visualizations that foster compassion. *Visualizations produced using the Ethical Visualization for Impact process exhibit four compassion-fostering qualities: they are humane, effective, trustworthy, and empowering*. In terms of humaneness, ethical visualizations build empathy in audiences and honor human dignity, including the dignity of users, represented subjects, and people particularly effected by the visualization. Ethical visualizations are also humane in the sense of being relatable. They foster understanding of how the data relates to them and their world, by creating a visual frame that gives users a compelling, humane bias as well as contextual information through devices such as normalizing representations.

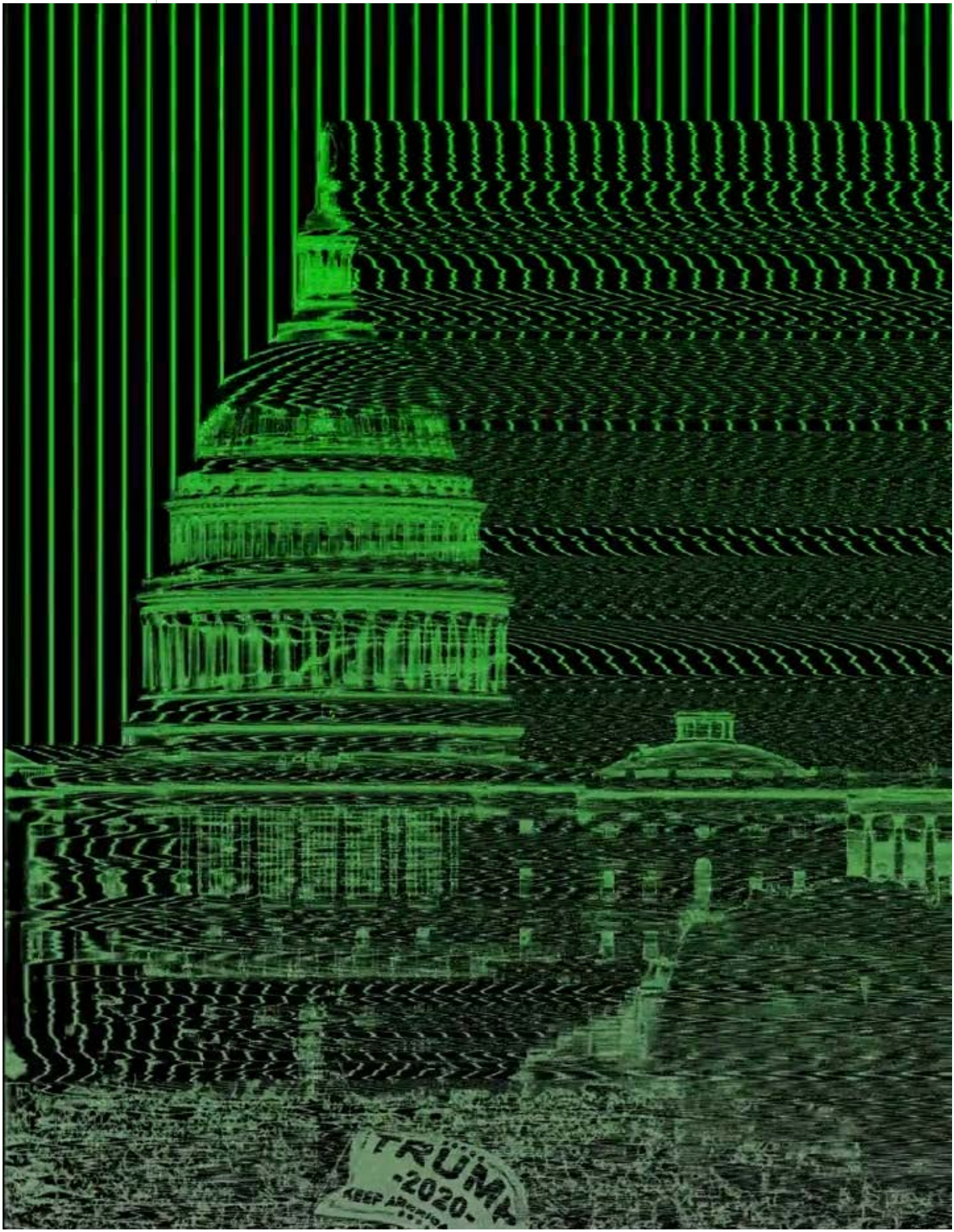
Ethical visualization is effective in the sense that it meets the goals of the research team as well as the needs of individual users through an appropriate visual frame. Visualizations signal that they are trustworthy when they have a consciously-crafted, coordinated visual frame, when they clearly state both what is known and what is not (through citations, margins of error, and caveats), show their process of visualization development, and clearly state their data sources and affiliations (authorship, funding, etc). Ethical visualizations are empowering when they give users agency to navigate data using appropriate scaffolding, within the confines of the visual frame, and give users a felt sense of agency through a welcoming user experience, that is, one that utilizes minimal cognitive effort and attention to achieve the above. As experts with intimate knowledge of the power of our data, we have a responsibility to communicate its importance to our key audiences ethically, that is, in a way that makes them care.

References

1. Bakker, W.: Pictopolitics: Icograda and the international development of pictogram standards: 1963-1986. In: Frascara, J. (ed.) Information Design as Principled Action, pp. 114–145. Common Ground Publishing, Champaign (2015)
2. Brinton, W.C.: Graphic Methods for Presenting Facts. The Engineering Magazine Company, New York (1914)
3. Brito, A., Rodríguez, M.A., Niaz, M.: A reconstruction of development of the periodic table based on history and philosophy of science and its implications for general chemistry textbooks. J. Res. Sci. Teach. **42**, 84–111 (2005). https://doi.org/10.1002/tea.20044
4. Cairo, A.: Ethical infographics. IRE J. **37**, 25–27 (2014)
5. Cairo, A.: How Charts Lie: Getting Smarter about Visual Information, 1st edn. Norton & Company, New York (2019)
6. Cosgrove, D.E.: Social Formation and Symbolic Landscape. University of Wisconsin Press, Madison (1998)
7. Crampton, J.W.: Maps as social constructions: power, communication and visualization. Prog. Hum. Geogr. **25**, 235–252 (2001). https://doi.org/10.1191/030913201678580494
8. Crampton, J.W.: Mapping: A Critical Introduction to Cartography and GIS. Wiley, Hoboken (2011)
9. Crampton, J.W.: Collect it all: national security, Big Data and governance. GeoJournal **80**, 519–531 (2015). https://doi.org/10.1007/s10708-014-9598-y
10. Dork, M., Feng, P., Collins, C., Carpendale, S.: Critical InfoVis: exploring the politics of visualization. Presented at the CHI 2013. Extended Abstracts. ACM, Paris (2013). https://doi.org/10.1145/2468356.2468739
11. D'Ignazio, C., Klein, L.: Feminist data visualization. Presented at the IEEE VIS Workshop on Visualization for the Digital Humanities. Springer, Heidelberg (2016)
12. D'Ignazio, C., Klein, L.F.: Data Feminism. The MIT Press, Cambridge (2020)
13. Edgerton, D.: From Mental Matrix to Mappamundi to Christian Empire: The Heritage of Ptolemaic Cartography in the Renaissance. Art and Cartography: Six Historical Essays (1987)
14. Hall, P., Heath, C., Coles-Kemp, L.: Critical visualization: a case for rethinking how we visualize risk and security. J. Cyber. Secur. **1**, 93–108 (2015). https://doi.org/10.1093/cybsec/tyv004
15. Hall, P.A.: Bubbles, lines and string: how visualisation shapes society. In: Atzmon, L., Triggs, T. (eds.) The Graphic Design Reader. Bloomsbury Academic, London (2017)
16. Harley, J.B.: Maps, knowledge, and power. In: Cosgrove, D., Daniels, S. (eds.) The Iconography of Landscape: Essays on the Symbolic Representation, Design and Use of Past Environments, pp. 277–312. Cambridge University Press, Cambridge (1988)
17. Hepworth, K.: Governmentality, technologies, & truth effects in communication design. In: Vermaas, P.E., Vial, S. (eds.) Advancements in the Philosophy of Design. DRF, pp. 497–521. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73302-9_23
18. Hepworth, K.: A panopticon on my wrist: the biopower of big data visualization for wearables. Des. Cult. 1–22 (2019). https://doi.org/10.1080/17547075.2019.1661723
19. Hepworth, K.: Ethical Visualization for Impact. Ethical Visualization for Impact (2020). https://kathep.github.io/ethics/. Accessed 18 Mar 2020
20. TEAM-Based Approach. Dialectic **2**. http://dx.doi.org/10.3998/dialectic.14932326.0002.104
21. Hepworth, K., Church, C.: Racism in the machine: visualization ethics in digital humanities projects. Digital Hum. Q. **012** (2018)

22. Hepworth, K., Ivey, C.E., Canon, C., Holmes, H.A.: Embedding online, design-focused data visualization instruction in an upper-division undergraduate atmospheric science course. J. Geosci. Educ. 1–16 (2019). https://doi.org/10.1080/10899995.2019.1656022
23. Kostelnick, C.: Humanizing Visual Design: The Rhetoric of Human Forms in Practical Communication, 1st edn. Routledge (2019)
24. Leszczynski, A.: Spatial big data and anxieties of control. Environ. Plan. D **33**, 965–984 (2015). https://doi.org/10.1177/0263775815595814
25. Lidwell, W., Holden, K., Butler, J.: Universal Principles of Design: 100 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions, and Teach Through Design. Rockport, London (2003)
26. Manning, A., Amare, N.: Visual-rhetoric ethics: beyond accuracy and injury. Tech. Commun. **53**, 195–211 (2006)
27. Monmonier, M.S.: Maps, distortion, and meaning, Resource paper - Association of American Geographers, Commission on College Geography , vol. 75, no. 4. Association of American Geographers, Washington (1977)
28. Monmonier, M.: How to Lie with Maps. University of Chicago Press, London (1991)
29. Monmonier, M.: Rhumb Lines and Map Wars: A Social History of the Mercator Projection. University of Chicago Press (2010)
30. NOAA: NOAA Interactive Sea Rise Level Viewer (2017)
31. Richards, D.P.: Not a cape, but a life preserver: the importance of designer localization in interactive sea level rise viewers. Commun. Des. Q. Rev. **6**, 57–69 (2018). https://doi.org/10.1145/3282665.3282671
32. Royal Society of Chemistry: Periodic Table (WWW Document) (2019). https://www.rsc.org/periodic-table. Accessed 18 Mar 2020
33. Schüll, N.D.: Data for life: wearable technology and the design of self-care. BioSocieties **11**, 317–333 (2016). https://doi.org/10.1057/biosoc.2015.47
34. Stephens, S.H., DeLorme, D.E.: A framework for user agency during development of interactive risk visualization tools. Tech. Commun. Q. **28**, 391–406 (2019). https://doi.org/10.1080/10572252.2019.1618498
35. Tufte, E.R.: Visual Explanations: Images and Quantities, Evidence and Narrative. Graphics Press, Cheshire (1997)

The New York Times https://www.nytimes.com/2021/02/05/opinion/capitol-attack-cellphone-data.html



Opinion

They Stormed the Capitol. Their Apps Tracked Them.

Times Opinion was able to identify individuals from a trove of leaked smartphone location data.

By Charlie Warzel and Stuart A. Thompson
Mr. Warzel and Mr. Thompson are writers in Opinion. They previously reported on smartphone tracking for the series “One Nation, Tracked.”

Feb. 5, 2021

In 2019, a source came to us with a digital file containing the precise locations of more than 12 million individual smartphones for several months in 2016 and 2017. The data is supposed to be anonymous, but it isn’t. We found celebrities, Pentagon officials and average Americans.

It became clear that this data — collected by smartphone apps and then fed into a dizzyingly complex digital advertising ecosystem — was a liability to national security, to free assembly and to citizens living mundane lives. It provided an intimate record of people whether they were visiting drug treatment centers, strip clubs, casinos, abortion clinics or places of worship.

Surrendering our privacy to the government would be foolish enough. But what is more insidious is the Faustian bargain made with the marketing industry, which turns every location ping into currency as it is bought and sold in the marketplace of surveillance advertising.

Now, one year later, we’re in a very similar position. But it’s far worse.

A source has provided another data set, this time following the smartphones of thousands of Trump supporters, rioters and passers-by in Washington, D.C., on January 6, as Donald Trump’s political rally turned into a violent insurrection. At least five people died because of the riot at the Capitol. Key to bringing the mob to justice has been the event’s digital detritus: location data, geotagged photos, facial recognition, surveillance cameras and crowdsourcing.

From Trump’s Rally to Congress
This time-lapse animation shows smartphones as they moved from Donald Trump’s rally to the Capitol.

White House

WASHINGTON, D.C.

Rally Stage

U.S. Capitol

Satellite Imagery: Microsoft Corporation, Maxar.

The sacking of the Capitol was a shocking assault on the republic and an unwelcome reminder of the fragility of American democracy. But history reminds us that sudden events — Pearl Harbor, the Soviet Union testing an atomic bomb, the Sept. 11 attacks — have led to an overreach in favor of collective security over individual liberty that we'd later regret. And more generally, the data collected on Jan. 6 is a demonstration of the looming threat to our liberties posed by a surveillance economy that monetizes the movements of the righteous and the wicked alike.

The data we were given showed what some in the tech industry might call a God-view vantage of that dark day. It included about 100,000 location pings for thousands of smartphones, revealing around 130 devices inside the Capitol exactly when Trump supporters were storming the building. Times Opinion is only publishing the names of people who gave their permission to be quoted in this article.

About 40 percent of the phones tracked near the rally stage on the National Mall during the speeches were also found in and around the Capitol during the siege — a clear link between those who'd listened to the president and his allies and then marched on the building.

While there were no names or phone numbers in the data, we were once again able to connect dozens of devices to their owners, tying anonymous locations back to names, home addresses, social networks and phone numbers of people in attendance. In one instance, three members of a single family were tracked in the data.

The source shared this information, in part, because the individual was outraged by the events of Jan. 6. The source wanted answers, accountability, justice. The person was also deeply concerned about the privacy implications of this surreptitious data collection. Not just that it happens, but also that most consumers don't know it is being collected and it is insecure and vulnerable to law enforcement as well as bad actors — or an online mob — who might use it to inflict harm on innocent people. (The source asked to remain anonymous because the person was not authorized to share the data and could face severe penalties for doing so.)

"What if instead of going to you, I wanted to publish it myself?" the source told us. "What if I were vengeful? There's nothing preventing me from doing that. It's totally available. If I had different motives, all it would take is a few clicks, and everyone could see it."

There is an argument to be made that this data could be properly used by law enforcement through courts, warrants and subpoenas. We used it ourselves as a journalistic tool to bring you this article. But to think that the information will be used against individuals only if they've broken the law is naïve; such data is collected and remains vulnerable to use and abuse whether people gather in support of an insurrection or they justly protest police violence, as happened in cities across America last summer.

The data presented here is a bird's-eye view of an event that posed a clear and grave threat to our democracy. But it tells a second story as well: One of a broken, surreptitious industry in desperate need of regulation, and of a tacit agreement we've entered into that threatens our individual privacy. None of this data should ever have been collected.

This is Ronnie Vincent.

We traced a phone inside the Capitol to Mr. Vincent's home in Kentucky. Confirming his identity led us to his Facebook page, where we found a few photos of him standing on the steps of the building during the siege. Another photo shows a crowd standing in front of the Capitol, its doors wide open.

At the Capitol

Smartphones tracked between 2 p.m. and 5 p.m. record the seige on the Capitol.

Satellite imagery: Microsoft Corporation and DigitalGlobe.

"Yes we got inside. One girl was shot by the DC cops as she was knocking on the glass. She probably will die. We stopped the voting in the house," he wrote.

Shortly after he posted the photos, Mr. Vincent, a pest control business owner in Kentucky who goes by the nickname Ole Woodsman, took them down. When we reached him by phone, he insisted he never entered the Capitol.

"There is no way that my phone shows me in there," he said. Yet it did.

A now-deleted Facebook post by Ronnie Vincent shows pro-Trump demonstrators advancing toward the open doors of the Capitol.
Screenshot from Facebook

For all its appearance of omniscience, the data can be imprecise. In a situation such as the Capitol riot, exact locations matter. A few feet can be the difference between a participant who committed a serious crime and an onlooker.

While some location data is accurate to within a few feet, other data is not. Location companies can work with data derived from GPS sensors, Bluetooth signals and other sources. The quality depends on the settings of the phone and whether it is connected to Wi-Fi or a cell tower. Issues like population and building density can sometimes play a role in the quality of the data.

Mr. Vincent told us that when he wrote "we got inside," he meant "we the people got in."

He added, "I did not go in."

Can we say definitively Mr. Vincent was inside the Capitol on Jan. 6? No, and that is one of the problems with this type of data.

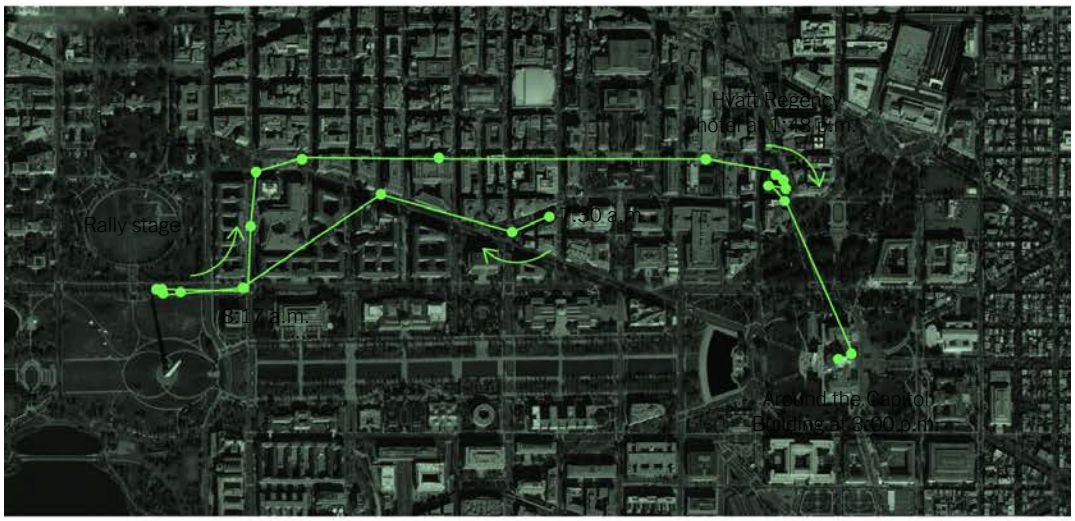
Ronnie Vincent's Journey

It was easier to identify Mr. Vincent — and discover the path he took to get to Washington, D.C. — because an email was matched to the phone's anonymous advertising ID.

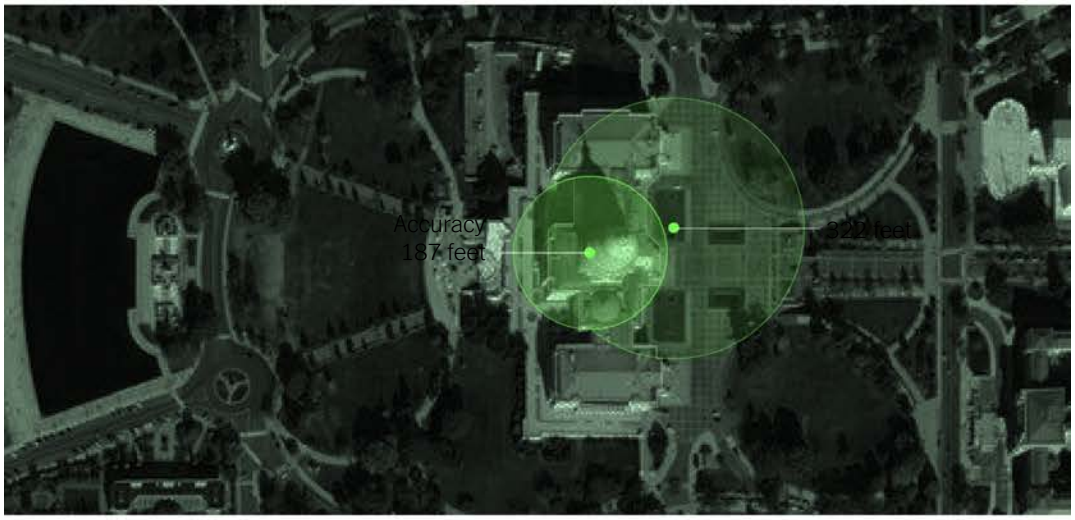
The trip to Washington, D.C.

Home location has been obscured.

The day of the protest



At the Capitol



Note: Location pings may not be precise. Satellite imagery: Microsoft Corporation, Earthstar Geographics, SIO, Maxar • By The New York Times

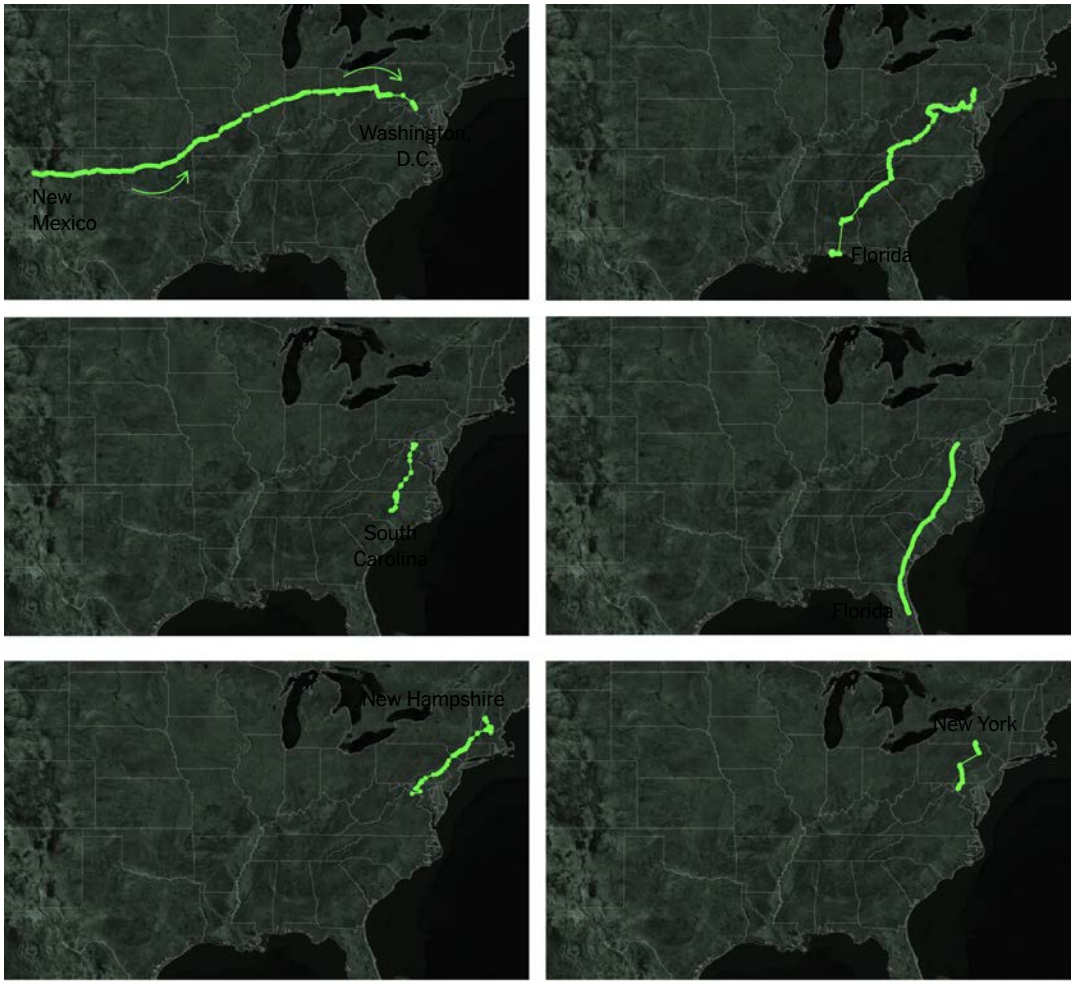
While the power and scope of this commercial surveillance come into sharp focus when we look at the specific time of the attack on the Capitol, it's important to remember that it is recording the movements of millions of Americans all day, all night, all year, wherever they are.

The data set Times Opinion examined shows how Trump supporters traveled from South Carolina, Florida, Ohio and Kentucky to the nation's capital, with pings tracing neatly along major highways, in the days before the attack. Stops at gas stations, restaurants and motels dot the route like bread crumbs, each offering corroborating details.

In many cases, these trails lead from the Capitol right back to their homes.

Trump Supporters Go to Washington

While protesters may have felt anonymous, their journeys to Washington and back were recorded in meticulous detail by apps on their phones.



Satellite imagery: Microsoft Corporation and Earthstar Geographics.

In the hands of law enforcement, this data could be evidence. But at every other moment, the location data is reviewed by hedge funds, financial institutions and marketers, in an attempt to learn more about where we shop and how we live.

Unlike the data we reviewed in 2019, this new data included a remarkable piece of information: a unique ID for each user that is tied to a smartphone. This made it even easier to find people, since the supposedly anonymous ID could be matched with other databases containing the same ID, allowing us to add real names, addresses, phone numbers, email addresses and other information about smartphone owners in seconds.

The IDs, called mobile advertising identifiers, allow companies to track people across the internet and on apps. They are supposed to be anonymous, and smartphone owners can reset them or disable them entirely. Our findings show the promise of anonymity is a farce. Several companies offer tools to allow anyone with data to match the IDs with other databases.

How "Anonymous" Pings Could Be Identifiable

The "anonymous" mobile advertising ID can be matched across databases ...

MARKETING DATABASE	APP DATABASE	ANONYMOUS LOCATION DATABASE
First name	Phone number	**Mobile Ad ID**
Last name	**Mobile Ad ID**	Precise location
Home address	Email address	Date
Mobile Ad ID	App name	Time

... creating a new deanonymized database

NEW DATABASE
Mobile Ad ID
First name
Last name
Home address
Phone number
Email address
App name
Precise location

Date

By The New York Times

We were quickly able to match more than 2,000 supposedly anonymous devices in the data set with email addresses, birthdays, ethnicities, ages and more.

One location data company, Cuebiq, publishes a list of customers that may receive the ID with precise smartphone locations. Companies listed there include household names like Adobe and Google, alongside a litany of lesser-known upstarts, like Hivestack, Mogean, Pelmorex and Ubimo.

In an emailed statement, Cuebiq said it prohibits attempts to merge location data with personally identifiable information and requires customers to undergo yearly third-party audits.

Smartphone users will never know if they are included in the data or whether their precise movements were sold. There are no laws forcing companies to disclose what the data is used for or for how long. There are no legal requirements to ever delete the data. Even if anyone could figure out where records of their locations were sold, in most states, you can't request that the data be deleted.

Their movements could be bought and sold to innumerable parties for years. And the threat that those movements could be tied back to their identity will never go away.

Mark Peterson for The New York Times

If the Jan. 6 rioters didn't know before, they surely know now the cost of leaving a digital footprint. Tip lines at the Federal Bureau of Investigation have been flooded for weeks in an effort to identify participants, and detectives in Miami and other police departments are using facial recognition software. Amateur investigators on TikTok, Instagram and other platforms have launched their own identification efforts.

Law enforcement has used cellphone footage from the siege to identify participants. As of February 4, there were 181 federal cases pending against individuals involved in the Capitol Hill siege, according to an analysis by George Washington University's program on extremism. Affidavits show that federal investigators were easily able to cross-reference footage with public social media posts.

A leak of data from the social media platform Parler also helped investigators and journalists place rioters in the building, using posts that were geotagged with GPS location data. For some, like 38-year-old Oath Keepers member Jessica Watkins, there was no need for precise location data. Her words tell the story: "Yeah. We stormed the Capitol today. Teargassed, the whole, 9. Pushed our way into the Rotunda. Made it into the Senate even," she wrote on Parler.

Which is to say that law enforcement may not need this data. But as a recent New York Times report shows, military agencies use these data sets — without a warrant, no less. How? They purchase it. Because we have seen what's in the data, that revelation is deeply troubling.

While some Americans might cheer the use of location databases to identify Trump supporters who converged on the Capitol, the use of commercial databases has worrying implications for civil liberties. The American criminal justice system is set up for a judge or jury to determine whether, in fact, Ronnie Vincent broke any laws on Jan. 6. But the data leads us directly to him, and in the hands of law

enforcement officials — or rogue employees of the company that collected the data — it could narrow their search for participants and offer clues about their activity.

To focus attention only on those people present at the deadly sacking of the Capitol is to lose sight of the larger context of the campaign of incitement and lies from Mr. Trump, right-wing media and members of Congress that set the stage for it. Just as focusing on the movements of Mr. Vincent's cellphone is to lose sight of the larger surveillance ecosystem that he — and all of us — are trapped in.

The location-tracking industry exists because those in power allow it to exist. Plenty of Americans remain oblivious to this collection through no fault of their own. But many others understand what's happening and allow it anyway. They feel powerless to stop it or were simply seduced by the conveniences afforded in the trade-off. The dark truth is that, despite genuine concern from those paying attention, there's little appetite to meaningfully dismantle this advertising infrastructure that undergirds unchecked corporate data collection.

This collection will only grow more sophisticated. This new data set offers proof that not only is there more interest in location data than before, but it is also easier to deanonymize. It gets easier by the day. As the data from Jan. 6 eerily demonstrates, it does not discriminate. It harvests from the phones of MAGA rioters, police officers, lawmakers and passers-by. There is no evidence, from the past or current day, that the power this data collection offers will be used only to good ends. There is no evidence that if we allow it to continue to happen, the country will be safer or fairer.

In our previous investigation, we wrote that Americans deserve the freedom to choose a life without surveillance and the government regulation that would make that possible. While we continue to believe the sentiment, we fear it may soon be obsolete or irrelevant. We deserve that freedom, but the window to achieve it narrows a little more each day. If we don't act now, with great urgency, it may very well close for good.

The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com.

Follow The New York Times Opinion section on Facebook, Twitter (@NYTopinion) and Instagram.

Charlie Warzel, a New York Times Opinion writer at large, covers technology, media, politics and online extremism. He welcomes your tips and feedback: charlie.warzel@nytimes.com | @cwarzel

Stuart A. Thompson is a writer and editor in the Opinion section. @stuartathompson

A version of this article appears in print on , Section A, Page 23 of the New York edition with the headline: Capitol Mob's Phone Apps Betrayed Them