Expert: COVID-19 data visualization 'exciting' and 'scary'

April 14 2020, by Jennifer Chu

Assistant Professor Catherine D'Ignazio uses data and computational techniques to work toward gender and racial equity. MIT News spoke with her about the current boom in Covid-19 data visualizations, and how data visualizers can help us make sense of the pandemic's uncertain numbers. Credit: Diana Levine

The COVID-19 pandemic is generating waves of data points from

around the world, recording the number of tests performed, cases confirmed, patients recovered, and people who have died from the virus. As these data are continuously updated, media outlets, government agencies, academics, and data-packaging firms are racing to make sense of the numbers, using novel design and visualization tools to chart and graph the virus many different contexts.

In general, data visualizations can help people quickly distill an otherwise overwhelming flood of numbers. Catherine D'Ignazio, assistant professor of urban science and planning at MIT, says it is critical that data are visualized responsibly in a pandemic.
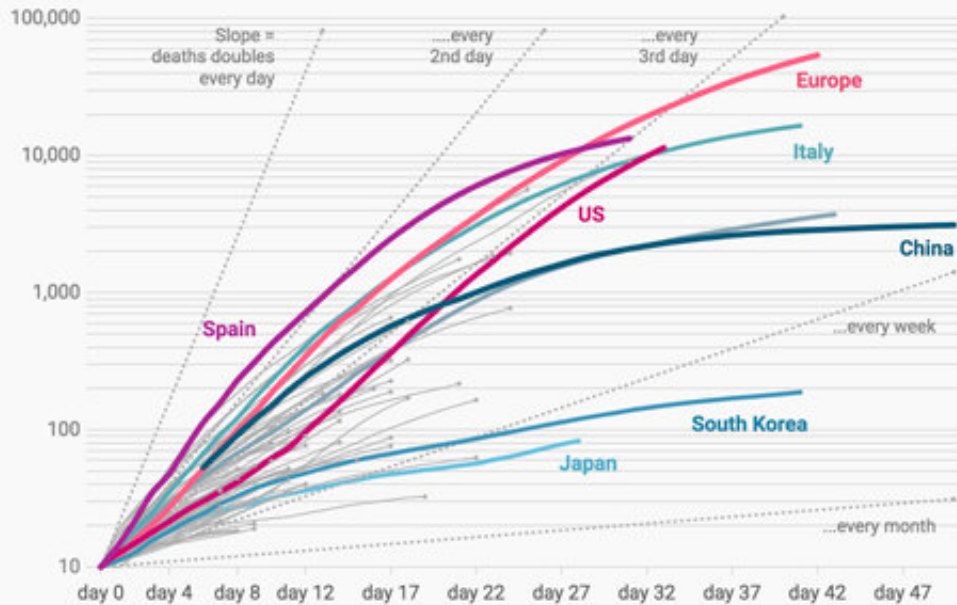
D'Ignazio is the director of the Data and Feminism Lab, where she uses data and computational techniques to work toward gender and racial equity. MIT News spoke with her about the current boom in COVID-19 data visualizations, and how data visualizers can help us make sense of the pandemic's uncertain numbers.

Q: How have you seen data visualization of COVID-19 evolve in the last few months, since the virus began its spread?

A: The first thing I'll note is that there has been an explosion of data visualization. Since the information about the virus comes in numbers—case counts, death counts, testing rates—it lends itself easily to data visualization. Maps, bar charts, and line charts of confirmed cases predominated at first, and I would say they are still the most common forms of visualization that we are seeing in media reporting and on social media. As a person in the field, the proliferation is both exciting, because it shows the relevance of visualization, and scary, because there is definitely some irresponsible use of visualization.

In the US, confirmed coronavirus deaths slow start to flatten. China & South Korea have already a flatter curve.

Cumulative numbers of confirmed deaths due to the COVID-19 disease, in selected countries after the 10th death, last updated with the numbers from yesterday, smoothed with a seven day rolling average. **The helper slopes don't mark corridors.** For example, Europe has a doubling rate of less than three days now – the "...every 3rd day" slope is steeper than the Europe line.

We don't show cases in this chart because they're not as reliably tracked as deaths. You can find a chart showing the doubling rate of confirmed cases here and in our River. It gets updated every day, but we don't recommend using it.

The chart stops at day 45, but more than 60 days have passed since China reported their 10th death. China reports a bit over 3,200 deaths to date. The 10th death in all countries is aligned and therefore approximated. The source didn't report values for China before March 22, so the numbers of days passed before that are estimated.

Chart: Lisa Charlotte Rost, Datawrapper. • Source: Johns Hopkins CSSE • Get the data • Created with Datawrapper

A screenshot of an interactive chart, from Datawrapper, shows cumulative numbers of confirmed deaths due to the COVID-19 disease. Chart: Lisa Charlotte Rost, Datawrapper. Source: Johns Hopkins CSSE. Credit: Datawrapper

Many high-profile organizations are plotting case counts on graduated color maps, which is a big no-no unless you have normalized your numbers. So California, a big and densely populated state, will always appear to be worse off in absolute raw case counts. Conversely, this way of plotting could cause you to miss small states with a high rate of infection since they will be low in relative case numbers and would

always show up in lighter colors on the map.

Second, as the crisis has developed, media outlets are mapping things other than simply case counts or death rates. There have been many versions of the "flatten the curve" chart. This one is interesting because it's not about plotting specific numbers, but about explaining a public health concept to a broad audience with a hypothetical chart. The best visual explanation I've seen of the flatten the curve concept is from The Washington Post and comes with simulations and animations that explain virus transmission. There have also been a number of visualizations of how social distancing has changed people's mobility behavior, shifting traffic patterns, and even a global satellite map where you can see how COVID-19 has reduced urban pollution over the past three months.

Finally, this crisis is posing some difficult visual communication problems: How do you depict exponential growth in an accessible way? How do you visually explain the uncertainty in numbers like case counts, where we (at least in the U.S. context) have not done nearly enough testing to make them a reliable indicator of actual cases?

Journalists and health communicators have responded to these challenges by developing new visual conventions, as well as making heavy use of explanations and disclaimers in the narratives themselves. For example, the chart below, by Lisa Charlotte Rost for DataWrapper, uses a log scale on the y-axis for showing exponential rates of change. But note the dotted reference lines, labeled "deaths double every day" or "...every 2nd day." These annotations help to highlight the use of the log scale (which otherwise might go unnoticed by readers) as well as to explain how to interpret the different slopes of the lines. Likewise, Rost is explicit about only making charts of death rates, not case counts, because of the variation in availability of tests and vast underreporting in many countries. Whereas actual cases may or may not be detected and

counted, deaths are more likely to be counted.

Q: What are some things people should keep in mind when digging into available datasets to make their own visualizations?

A: This is such a great question, because there has been a proliferation of visualizations and models that are not only erroneous but also irresponsible in a public health crisis. Usually these are made by folks who do not have expertise in epidemiology but assume that their skills in data science can just be magically ported into a new realm. I'd like to shout out here to Amanda Makulec's excellent guidance on undertaking responsible data visualizations in a public health crisis. One of her main points is to consider simply not making another COVID-19 chart. What this points to is the idea that data scientists and visualization designers need to take their civic role very seriously in a pandemic. Following Makulec's line of reasoning, designers can think of the visualization they are making in the context of decision support: Their visualization has the power to help people decide whether to reject public health guidance and go out, to stay home, to feel the gravity of the problem without being overwhelmed, or to panic and buy up all the toilet paper.

Data visualization carries the aura of objectivity and authority. If designers wield that authority irresponsibly—for example, by depicting case counts with clean, certain-looking lines when we know that there is deep uncertainty in how case counts in different places were collected—it may deplete public trust, lead to rejecting public health guidance like social distancing, or even incite panic.

This carries over into all manner of visual choices that designers make. For example, color. Visualizations of COVID-19 cases and deaths have tended to use red bubbles or red-colored states and provinces. But color

has cultural meaning—in Western cultures, it is used to indicate danger and harm. When a whole country is bathed in shades of red, or laden with red bubbles that obscure its boundaries, we need to be very careful about sensationalism. I'm not saying "never use red"; it is warranted in some cases to communicate the gravity of a situation. But our use of charged colors, particularly during a pandemic like this, involves making very careful ethical decisions. How serious is the risk to the individual reader? What do we want them to feel from viewing the visualization? What do we want them to do with the information in the visualization? Are those goals aligned with public health goals?

Rather than reducing complexity (to generate sensationalist and attention-grabbing clicks), some of the most responsible visualization is working to explain the complexity behind our current crisis. This is the case in the above graphic. The journalists walk us through why even calculating a simple input like the fatality rate depends on many other variables, both known and unknown.

All that said, public health communication really does need good visualization and data science right now. One of the exciting developments on the responsible-vis horizon is a new program from the Data Visualization Society that matches people with visualization skills to organizations working on COVID-19 that need their help. This is a great way to lend a hand, concretely, to the organizations who need help communicating their data during this crisis.

Q: How can we as individuals best make sense of and put into context all the data being reported, almost by the minute, in some cases?

A: One of my students said something wise to me this week. As she was describing her obsession with checking the news every couple minutes,

she reflected, "I realized that I'm looking for answers that I cannot find, because nobody knows them." She's right. At this point, nobody can truly answer our most basic questions: "When will this end? Will I lose my job? When will my kids return to school? Are my loved ones safe? How will my community be changed by this?"

No amount of data science or data visualization can solve these questions for us and give us the peace of mind we are craving. It is an inherently uncertain time, and we are in the middle of it. Rather than obsessively seeking information like case counts and scenario models to give us peace, I have been telling students to practice self-care and community-care as a way to direct their attention to things they have more control over. For example, in our local communities, COVID-19 is already having a disproportionate impact on the elderly, on health care workers, on first responders, on domestic workers, on single parents, on incarcerated people, and more. Below is one effective graphic that highlights these disproportionate impacts.

As the graphic shows, there is a great dimension of privilege in the people who are able to work from home: The vast majority of folks who can earn money from home are in the richest 25 percent of workers. This attention to how power and privilege play out unequally in data is also a throughline in Lauren F. Klein and my recently published book, "Data Feminism." A feminist approach demands that we use data science and visualization to expose inequality and work toward equity.

So while it is important to (responsibly) track and visualize death rates from COVID-19, how do we also focus our attention on efforts to support the groups who are most directly and unfairly impacted by this crisis, to get them the care, equipment and the economic security that they need? The reality—even amidst this great uncertainty—is that we can all take action now, in our local communities, to support each other.

Provided by Massachusetts Institute of Technology