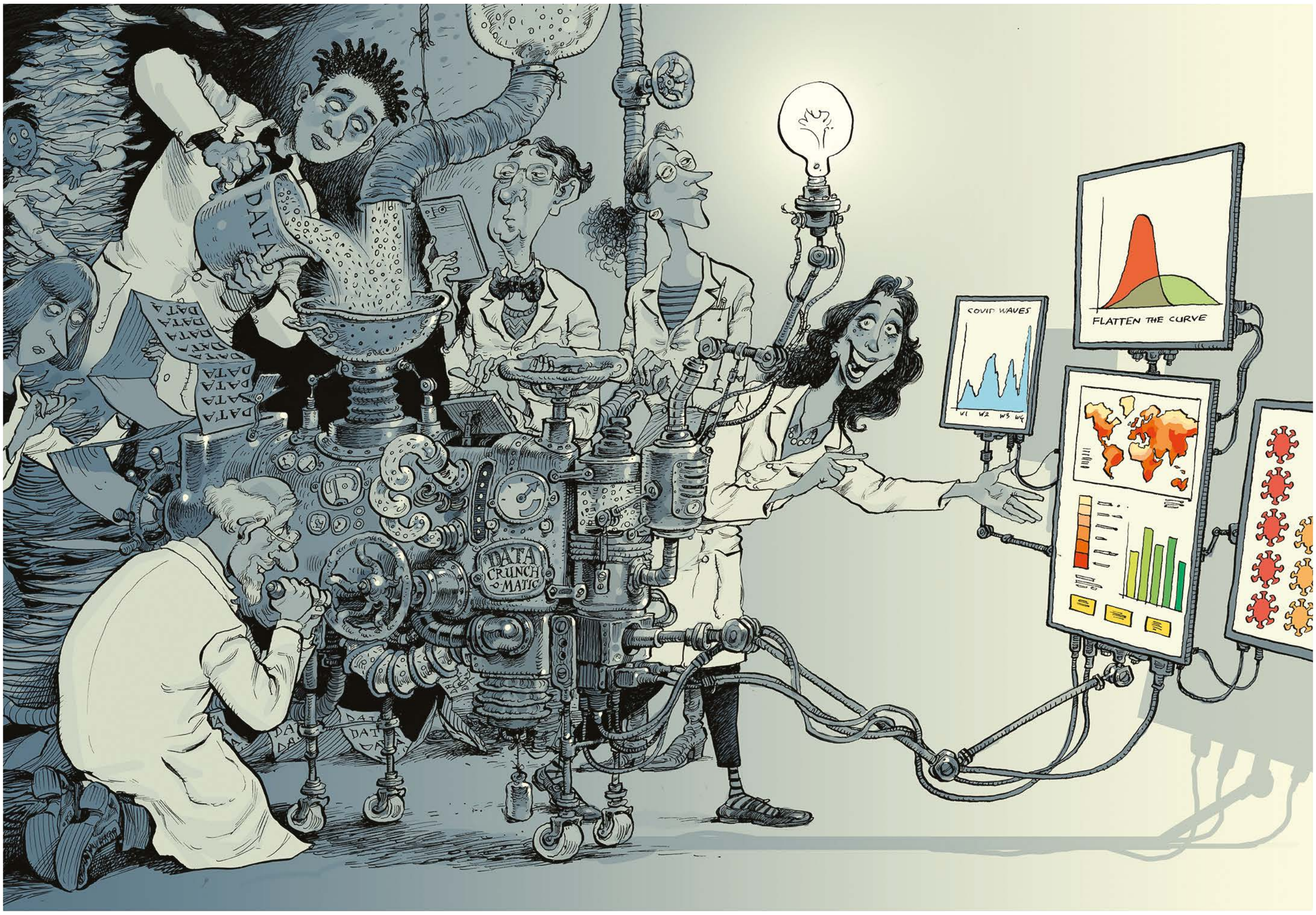


ILLUSTRATION BY DAVID PARKINS

LESSONS FROM THE COVID DATA WIZARDS

Data dashboards have been an important part of pandemic response and planning. What have their developers learnt about communicating science in a crisis? **By Lynne Peeples**

In March 2020, Beth Blauer started hearing anecdotally that COVID-19 was disproportionately affecting Black people in the United States. But the numbers to confirm that disparity were "very limited", says Blauer, a data and public-policy specialist at Johns Hopkins University in Baltimore, Maryland. So, her team, which had developed one of the most popular tools for tracking the spread of COVID-19 around the world, added a new graphic to their website: a colour-coded map tracking which US states were — and were not — sharing infection and death data broken down by race and ethnicity.

They posted the map to their data dashboard — the Coronavirus Resource Center — in mid-April 2020 and promoted it through social media and blogs. At the time, just 26 states included racial information with their death data. "Then we started to see the map rapidly filling in," says Blauer. By the middle of May 2020, 40 states were reporting that information. For Blauer, the change showed that people were paying attention. "And it confirmed that we have the ability to influence what's happening here," she says.

COVID-19 dashboards mushroomed around the world in 2020 as data scientists and journalists shifted their work to tracking and presenting information on the pandemic – from infection and death rates, to vaccination data and other variables. "You didn't have any data set before that was so essential to how you plan your life," says Lisa Charlotte Muth, a data designer and blogger at Datawrapper, a Berlin-based company that helps newsrooms and journalists to enrich their reporting with embeddable charts. "The weather, maybe, was the closest thing you could compare it to." The growth in the service's popularity was impressive. In January 2020 – before the pandemic – Datawrapper had 260 million chart views on its clients' websites. By April that year, that monthly figure had shot up to more than 4.7 billion.

Policymakers, too, have leaned on COVID-19 data dashboards and charts to guide important decisions. And they had hundreds of local and global examples to reference, including academic enterprises such as the Coronavirus Resource Center, as well as government websites and news-media projects.

The New York City Department of Health was among Datawrapper's clients. And Blauer notes that she has hosted regular webinars with several US mayors, walking them through her team's metrics. She is confident, she says, that the "data informed policy".

The architects of these dashboards put in long hours and faced considerable challenges, including incomplete and inconsistent data, misconceptions and misunderstandings about how the information was collected, and efforts to twist the messages that the dashboards present. As these data wranglers continue to try to inform individuals and public-health officials, they are learning lessons that will help to navigate the next stage of the pandemic, as well as other social and public-health issues – from crime to climate change.

Hard data

The Johns Hopkins dashboard originated at a meeting between Lauren Gardner, who studies civil and systems engineering at Johns Hopkins, and her PhD student, Ensheng Dong. In early January 2020, Dong began closely following cases of a type of pneumonia emerging in his home country, China. "He was hearing things directly from his friends and family," says Gardner.

Dong was concerned for their well-being, and began pulling data from a Chinese website, DXY.cn. He and Gardner spent days and nights tracking the information on a Google sheet. Then they built a map to go alongside that dynamic spreadsheet and made both available to the public. "We literally decided this one afternoon and built the initial version of the dashboard that night," says Gardner. "It seemed like a manageable, simple task, given the scale of the problem at the time. Of course, we didn't know the scale that this would grow to." Just weeks later, the website had upwards of 4 billion queries a day.

Gardner and Dong eventually moved the data to a GitHub repository, a cloud-based data-storage and -management space that maintains a file history. Their initial global map, with its recognizable red dots proportional to case counts, is still updated every hour. Blauer and others joined the effort early and expanded it with a multi-layered, interactive dashboard to help people digest the data.

Ideally, data that are this important for public health should be freely available, machine-readable and standardized. From the start, the team realized that they were not. Compiling complete and consistent COVID-19 data was "very manual and very messy", says Gardner. "We were scrambling, collecting and validating reported data as fast as we could." Because COVID-19 data were not yet provided on any public-health agency's website, they looked elsewhere, including on Facebook and Twitter posts and in one-off news and media announcements. Even after agencies launched official data pages, both sourcing and formatting remained an issue. Gardner says that some of the data the team collects are still not machine-readable. "There should be a standardized way in which the data is provided and shared publicly, as well as what is shared," says Gardner. "That would've made our job a lot easier."

> **"IT SEEMED LIKE A MANAGEABLE, SIMPLE TASK ... WE DIDN'T KNOW THE SCALE THAT THIS WOULD GROW TO."**

Blauer has been vocal, blogging about the need for greater accessibility and standardization of data, including the use of consistent categories and naming conventions for age, gender, race and ethnicity. "Demographic data is a complete mess," she says. Racial and ethnic categories are tricky because they are regarded differently in different countries. But even in a single US state, Blauer found category definitions changed depending on whether they were linked to vaccination rates, cases or deaths. She has made creative moves to fill in the blanks,such as when her team revealed which states were and weren't collecting race and ethnicity data. Blauer and her team have confirmed that the pandemic has had inequitable impacts. As of September 2021, for example, Black residents of Washington DC made up 45% of the population, but 76% of COVID-19-associated deaths.

The Johns Hopkins team was not alone in its struggles. Hannah Ritchie, head of research at the non-profit organization Our World in Data in Oxford, UK, recalls her efforts to copy data from PDFs. She also points to some of the consequences of incomplete and inconsistent data. For example, differences in access to COVID-19 testing data can result in inaccurate comparisons. "It can often lead you to conclude that some countries have not been touched by the pandemic," says Ritchie. "That is just not true."

Ritchie also fears that the gains that have been made in data collection and visualization could easily be lost before the global pandemic is over. "A lot of these data projects are seen as one-off things," she says. "As rich countries start to get more back to normal because of high vaccination rates, for example, will they turn around and just let these projects die?" Some dashboards have already stopped their efforts. And government efforts to collect and display data in real time are slowing in many parts of the world.

Raising 'graphicacy'

Whether they have realized it or not, people around the world have essentially been enrolled in a two-year mathematics course. Core to this curriculum is 'graphicacy', or the ability to understand data presented in graphs and figures. "I think the pandemic helped to bring the graphicacy of the general public to a higher level," says Maarten Lambrechts, a data-visualization consultant based in Diest, Belgium. "A lot more people now have a better understanding of how a chart works and how they should interpret it."

COVID-19 dashboards were vehicles for this public education. Data scientists took to mainstream news outlets and social media to explain and contextualize the visualizations they posted on dashboards. Soon, terms such as 'flattening the curve' and 'log scale' became dinner-table topics for many people.

Among the shared themes for the dashboards were simplicity and clarity. Whether you are producing visuals and analytical tools for policymakers or for the public, Blauer says, the same rules of thumb apply. "Don't overcomplicate your visualization, make the conclusions as clear as possible, and speak in the most basic of plain-language terms," she says.

Yet, as other data scientists point out, presenting data simply might not be enough to ensure viewers get the message. For one thing, attention to detail matters. Ritchie recalls how she and her team spent hours focused on the titles and subtitles of charts, "because that is ultimately what most people will look at". And in those titles and subtitles, the analysts made sure to specify 'confirmed' deaths or 'confirmed' cases. "An emphasis on 'confirmed' is really important because we know that it's an underestimate of the total," says Ritchie. "It might seem very basic, but it's really crucial to how you understand the data and the scale of the pandemic."

The best data visualization might also not be the one that is most pleasing to the eye. Data scientist Pouria Hadjibagheri led the team working on the UK COVID-19 dashboard run by the UK Health Security Agency, formerly known as Public Health England, in London. His team gathered and processed nearly one billion records a day from 26 different data sources. His audience includes individuals with visual impairments, epilepsy, attention-deficit hyperactivity disorder or other conditions. "All of these people have equal rights to see the data," says Hadjibagheri, who was the lead software developer for the agency until June 2021, but continues to work on the dashboard.

His team regularly surveys dashboard users to see how well the visualizations are understood. In February 2021, for example, he asked users which of the three vaccination visuals they liked the best. Importantly, the team also quizzed the participants on their interpretation of those graphics. People tended to prefer a bar-chart option that looked "simple and very nice", says Hadjibagheri, but they couldn't accurately identify the proportions it represented. Ultimately, the dashboard team opted for a 'waffle' chart design that offered a more granular view of the data, because the polling showed that users understood it better (see 'User testing').

> "THERE ARE SO MANY NUMBERS, AND SO MANY SOURCES AND TABLES TO PICK FROM … IT'S A REAL THREAT."

Hadjibagheri and other dashboard architects have also been engaged in a two-year crash course of sorts, on how best to present information to the public. "Science communicators have had about two years now to sort of stress-test all of their methods of communicating to a general audience," says John Burn-Murdoch, chief data reporter at the *Financial Times* in London. "And I think there have definitely been improvements."

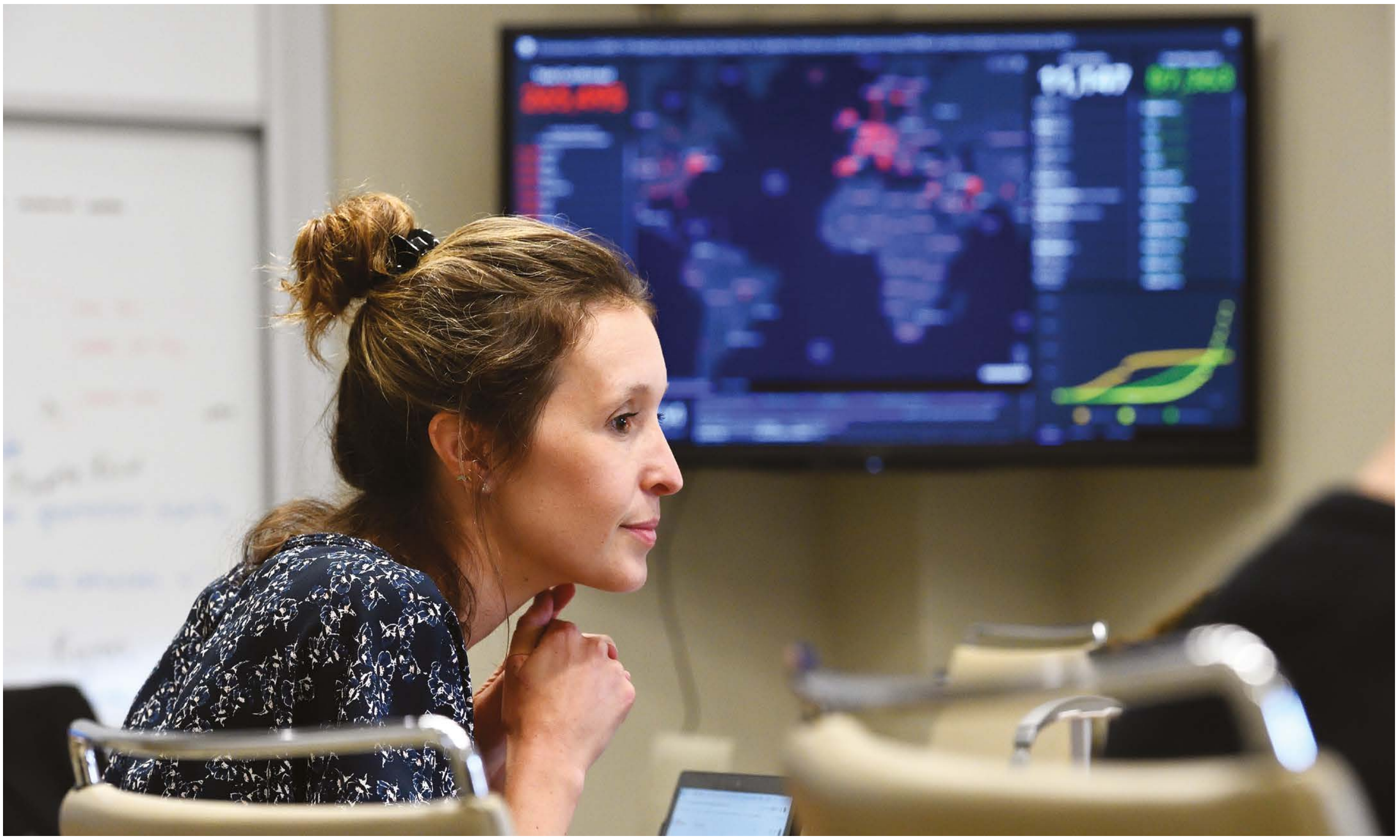
Trust and transparency

Other common lessons learnt: let people see behind the scenes, share the data sources, and teach people what data methodologies were used and why. And always be transparent about any gaps or errors that could lead viewers astray.

The UK dashboard got some fine-tuning in the aftermath of a mishap on 7 March 2021. That day, death data from England had not been processed by the agency's usual deadline. The team decided to go ahead with an update and to post a prominent explanation for why there were no deaths listed. "We thought it was so bizarre a number at that point in time that people would be forced to look at the website," says Hadjibagheri.

They were wrong. Within minutes, several news agencies, including the BBC and Reuters, posted stories with the new data – and without the added caveat. Hadjibagheri quickly logged on to Twitter and corrected their mistakes. "We knew people were paying close attention. But we weren't expecting broadcasting agencies to not apply a sensibility check," he says.

The team subsequently set up automatic feeds so that organizations could subscribe to their announcements directly. It also decided to delay posting data with big gaps in the future. Hadjibagheri had been increasing his



WILL KIRK

Lauren Gardner works on the Coronavirus Resource Center dashboard at Johns Hopkins University in Baltimore, Maryland.

social-media activity since about November 2020. He says that trust in his service, based on user surveys, subsequently rose by nearly 35%.

Engagement is important. In parallel with his widely shared data visualizations, Burn-Murdoch took to social media as well. In March 2020, he posted a six-part Twitter thread explaining the log scale and why he chose to use it to show the rise in case numbers in different countries. "I've always seen my job, fundamentally, as communication," says Burn-Murdoch, who regularly responds to people's questions about charts and concepts.

Social media has also given data scientists a chance to draw back the curtain around what they do, answer questions, correct misunderstandings and provide context for their data – promoting transparency and trust. But social media can rapidly spread inaccurate messages and potentially degrade trust in dashboards.

Take, for example, the wave of misinformation that rippled worldwide after a guest of US podcaster Joe Rogan said that COVID-19 case rates were higher among vaccinated people. Brazilian President Jair Bolsonaro reported the same thing on social media and included baseless claims that vaccinated people also had a higher risk of contracting AIDS. They each referenced data published by the UK Health Security Agency, which did show that SARS-CoV-2 infection rates among fully vaccinated people aged 40 years and older were higher compared with unvaccinated people in the same age group.

The numbers, however, were based on an inaccurate estimate of the pool of unvaccinated people. When a more appropriate estimate was used, case rates among unvaccinated individuals were shown to be higher than for vaccinated people in nearly all age groups. The agency was sharply criticized for releasing the misleading figures. Recognizing early on that the data interpretation is not straightforward, Hadjibagheri says that his team never published visualizations of the breakdown of cases and deaths based on vaccination status.

Indeed, the COVID-19 pandemic has shown how easy it can be to twist data to fit a narrative. The dashboards have provided a lot of information. "There are so many numbers, and so many sources and tables and papers to pick from," says Lambrechts. "It's a real threat."

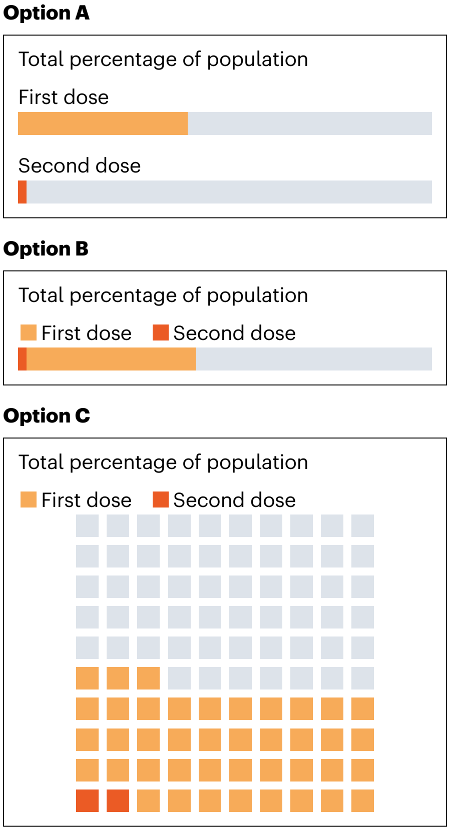
Bill Hanage, an epidemiologist at the Harvard School of Public Health in Boston, Massachusetts, warns that people can also "fall prey to motivated reasoning".

Dashboards are a great place to start thinking about data, he says, but people shouldn't stop testing their assumptions. "Look at all the people who trumpet the low reported mortality in sub-Saharan Africa, without asking about testing capacity."

He points to a lesser-known category of data dashboards that has proved powerful – and perhaps less prone to manipulation. Wastewater samples provided one of the earliest signals that Omicron was emerging in South Africa, he notes. Many US municipalities, including several counties in Massachusetts and North Carolina, have created public dashboards for wastewater data, helping to track infections and predict new outbreaks. "Wastewater can't lie," he says.

USER TESTING

When the developers of the UK COVID-19 dashboard surveyed users, asking which graphical presentation they liked best, people gravitated towards option A — a double bar graph. But after quizzing users on the numbers these graphics aimed to display, almost 90% found option C — the waffle graph — more useful.



Option A

Option B

Option C

SOURCE: P. HADJIBAGHERI/UK HEALTH SECURITY AGENCY

Beyond COVID-19

Data scientists are hopeful that the time and money invested in COVID-19 data dashboards, and the global education they provided, will pave the way for similar efforts to address other important issues. "Creating standards that are easily adoptable, like measuring cases and deaths for COVID, will be really important when we're trying to do the hard work of eradicating poverty and improving climate conditions," says Blauer. "The big question is, are policymakers willing to do it?"

For example, the data hosted on US dashboards such as Mapping Police Violence, run by the non-profit US police-reform advocacy group Campaign Zero in New York City, suggest that racial bias plays a part in police violence. "The data is telling us that there are kinds of officers who are going to engage in an officer-involved shooting," says Blauer. Yet incident reporting is piecemeal, voluntary and not standardized across the country.

And the data are complex. Researchers have yet to reach a consensus on how to account for differing rates and types of police encounter – from traffic incidents to active-shooter situations – for different races and ethnicities. "How do you attract better, more empathetic officers? Getting that data is almost impossible," says Blauer.

Much like COVID-19, climate change is a global problem that no country can tackle on its own. Also, as for COVID-19, crucial information on the subject can be difficult to access. The International Energy Agency (IEA) in Paris, an intergovernmental organization established as part of the Organisation for Economic Co-operation and Development, is the world's most authoritative and comprehensive source of global-energy data. Yet much of the IEA's data remain locked behind a paywall. Currently, many people rely on information provided by the energy firm BP, which is missing data from many countries and lacks key metrics. "So we have people trying to understand climate change based on data published by a petrochemical company," says Ritchie, who is among many scientists pushing the IEA to become more forthcoming with its data. The agency has shown signs of relenting. It has started offering free reports and providing more free data, according to Jad Mouawad, an IEA spokesperson. "We are now exploring options to further increase the amount of data that is available for free to more users while at the same time maintaining the financial stability of the agency," says Mouawad.

Meanwhile, improvements in technologies, such as remote sensing using satellites, will show in near-real time and in greater detail "how we are changing our planet", says Lambrechts. An increasing number of climate data dashboards are popping up, including one from the UK government that shows changes in metrics ranging from temperature and rainfall to the area of woodland in the United Kingdom. "The thing about climate change is that we think it's not happening in the present, when it is actually happening in the present," says Angela Morelli, co-founder of the InfoDesignLab in Oslo, which produces data visualizations for the Intergovernmental Panel on Climate Change.

Hadjibagheri, who was recently approached by the Ministry of Justice in the United Kingdom regarding a potential dashboard for sex crimes, suggests that the world can build on the data lessons and achievements from COVID-19. "The infrastructure built, the services built, the culture built – those can all be adopted," he says. Specifically, he argues that every country needs an independent data-science office, with the expertise and ability to handle and publish large volumes of data. "This is data that we can use to our advantage."

Lynne Peeples is a science journalist in Seattle, Washington.

Correction
This Feature misspelt the name of Maarten Lambrechts.

Corrected 28 March 2022