

The Intersection of Data Science and Public Health in the COVID-19 Pandemic



Written by
Jennifer Lane

Published on
January 17, 2022

"All science is data science: one of the foundational principles of science is that it is data driven."

Never has the intersection of data science and public health been so keen as it has been during the COVID-19 Pandemic. Now that we are rounding the corner on a year since the virus reached American shores, we have seen how vital it is that the surrounding data be accurately interpreted, and that the corresponding findings are turned around quickly. To this end, I spoke with the co-founders of the COVID Outlook Project, a volunteer-run website that provides independent data-driven analyses and forecasts about the COVID-19 crisis in the United States for policymakers and individuals. Their goal is to help people make the best decisions for themselves, their families, and their communities during this complicated time of heightened risk.

Q: Do you see data science as a natural continuation emerging out of biostatistics and epidemiology?

Michael LeVasseur: Well, that's a good question. Yes, no, and maybe. In short, I mean, all science is data science, right? One of the foundational principles of science is that it is data driven. Working with big datasets, all these are part of data science. In public health, it's mostly data management, visualization, and simulation work. I think there's a natural evolution, or a natural convergence between computational sciences, and social and Biological Sciences, in that I think it has more to do with the development and advancement of technologies.

Q: In public health, how are data science practices employed?

Michael LeVasseur: I think the easiest starting point is to talk about visualization, which is a big topic in data science. I can tell you that in public health, there isn't much in the curricula that describes visualization. But it's really people like Michael Donnelly who focused on visualization as a way of getting a vast amount of important information into someone's brain in a really quick and easy to understand way. And I really admire that ability because as a scientist, I'm trained to tell you 15 paragraphs of information in order to describe, like, one tiny number. That's my wheelhouse.

Michael Donnelly: So, my background with data science is not in health at all. It's in financial economics, and policy. After the financial crisis of 2008, it was insane the amount of work that went into trying to explain complex phenomena, specifically complex risk, to policymakers in the form of a dashboard. One of the core insights that came out of the financial crisis is that we did a very poor job leading up to the financial crisis, both in government but also in the private sector in understanding the size of risk, and the types of risk. And what we're really talking about were these types of risks that happen very rarely, but have a huge impact. And the financial crisis was one of those events as a result. So one of the things that we really screwed up leading up to the financial crisis was not understanding just how likely those really damaging events were. And not taking the kind of actions we need to, to mitigate risks that happen out on the tails of news distributions. And so when we think about, you know, a normal bell curve distribution, and we think about something that's not in that thick, middle part of the bell curve, but it's kind of often the tails. We look at that, and we go, Oh, that's really tiny. The odds of a banking crisis that takes down the entire U.S. banking system, well, that's really tiny, it happens less than 1% of the time. The question is, how much less? What happened in the financial crisis is that we got that order of magnitude wrong.

Michael LeVasseur: It's really that insight is a big part of what [Donnelly] and I have been trying to spend a lot of time on this year, is helping policymakers understand what happens at those inflection points where small risks have really big impact and kind of how quickly those things can become really significant in day to day life.

Q: That sounds like that translates pretty directly into what we're seeing with the COVID crisis.

Michael LeVasseur: At the federal level, we've been doing pandemic preparedness for decades. We were surprised that it took this long. [The SARS outbreak of 2003] was a bit of a wakeup call. We knew a pandemic was coming but we were convinced it was going to be influenza. And we were convinced it was going to be much worse.

Michael Donnelly: I think this is where we start talking about complexity of risk, right? It's not super straightforward. So as Mike says, any number of aspects of this pandemic could have been worse, like we could have had a higher mortality rate. Mortality rate was relatively low compared to what it could have been, compared to Ebola or something. But then we start getting into second order effects of those changes. So, let's imagine that we had 40 million people infected in the U.S. And then instead of a 1% fatality rate, we had a 5% infection fatality rate. So that's five times more deaths. But what happens in that scenario?

Well, at that fatality rate, I would have been looking out of my window in [New York City], and I would have seen body bags. That sort of visual, I think would have changed the way we would have dealt with this crisis. And so while I think Mike's absolutely right, that this could have been worse, I think the secondary effects are a little bit difficult to totally understand.

Q: This makes me think about the ways in which appropriate data visualization could have had a real impact on how the public treated a mask mandate. When people talk about how comparatively low the COVID fatality rate is, I don't think they adequately understand what a 1% rate actually translates to.

Michael Donnelly: I think one of the interesting challenges for data scientists involved in prediction and risk communication around human events is that it's meaningfully different from other prediction exercises, like meteorology, for instance. We have invested so much in the National Oceanic Oceanographic and Atmospheric Administration, who are the ones who do the hurricane predictions. And by investing so much in them, their predictions have become so much better over the last few decades. And they're incredibly accurate. The thing is, hurricanes do not read weather.com. They don't check the weather and then say, oh, they're going to protect southern Florida and North Carolina, I better go hit Mississippi. Hurricanes don't do that. When we do that, about human events, be it financial crises or pandemics, we run this problem called moral hazard or risk compensation. And people say, oh, the CDC has got me covered. I don't need this stinking mask.

Michael LeVasseur: They look at the dashboard, they say, oh, positivity is down from 10% to 9%. I can go out and hang out with my friends today. Treat it like a weather report. Right? One of the questions is: what does the public do with this information? Are we doing a disservice to the public by giving them all this information? I don't know the answer to that. But I think that it's a question that we need to have a conversation about if the public doesn't know how to interpret data, which is part of the responsibility for a data scientist.

Q: Bias has been sort of a hot button topic in the world of Artificial Intelligence and Data Science. Can you talk a bit about that and how it applies to public health?

Michael LeVasseur: So, when we're talking about bias, from a public health standpoint, what we're really talking about is: how different is it from truth? Most of the time, we don't really know what truth is, we're trying to estimate it. So the biases in public health are not necessarily the same as the word bias when we're talking about, for example, facial recognition software, being biased against people with darker skin. It's more of a messy data issue. In terms of the COVID crisis, depending on your state, your jurisdiction, most testing is available to you free of charge. That said, I do know people in Chicago when they were going through their peak in the summer, who would have to wait seven to 10 days in order to get their test result back from the publicly available labs. Your other option was to pay a private lab \$150, and get your results the next day. So, what is more important from a public health standpoint? That someone knows their status the next day or that it's affordable? I can't choose between those things, it needs to be affordable, it needs to be accessible. But a 4–14 day lag in reporting — I mean, that's the entire window for quarantine for most people. And it has an impact on their behavior, which has an impact on the people around them.

Michael Donnelly, MSc

Michael Donnelly is a data scientist with nearly a decade of experience in data analysis, data science management, and time series forecasting. Michael is a graduate of Vassar College and holds a Master's Degree from the London School of Economics. He has also managed the development and operations of research units in the public and private sectors. Michael's analysis in early March 2020 was pivotal in convincing public authorities in New York City and New York State to prepare for a massive public health crisis. His work on the novel coronavirus and COVID-19 have been reported on in the Financial Times, WNYC's The Gothamist, The New York Post, and Politico, among many others.

Michael LeVasseur, PhD, MPH

Prof. Michael LeVasseur is an infectious disease epidemiologist and assistant professor in the Department of Epidemiology and Biostatistics at Drexel University's Dornsife School of Public Health. Michael is also the lead epidemiologist for Drexel University's COVID-19 testing program and runs a weekly interdisciplinary COVID-19 journal club exploring various aspects of the pandemic from the molecular through the societal. He received his Bachelor's Degree in liberal arts with concentrations in molecular biology and medical anthropology from Sarah Lawrence College, his Master's in Public Health with a concentration in epidemiology from the CUNY Hunter School of Public Health, and his PhD in Epidemiology from Drexel University's Dornsife School of Public Health. He has worked at the Bureau of Environmental Safety and Policy at the New York City Department of Health and Mental Hygiene the HIV Behavioral Research Center at Columbia University, Weill Cornell Medical College, the Center for Health Incentives and Behavioral Economics at UPenn, and the Center for Firefighter Injury Research & Safety Trends at Drexel University. Michael's research interests include HIV epidemiology, data science, and sexual and gender minority health disparities.

Jennifer Lane Bustance — Writer

Jennifer Lane is a California-based novelist, playwright, and teaching artist. She is currently filling in for Bahija Humphrey, CEO of the Data Science Alliance, while she is out on maternity leave. MFA: Columbia University; BA: Sarah Lawrence College. For more information, please visit jennifer-lane.net. Originally published at <https://bahija.medium.com> on January 25, 2021.