Harvard Data Science Review

Data Science in Times of Pan(dem)ic

Sabina Leonelli

Published on: Jan 29, 2021

Updated on: Jan 07, 2021

DOI: 10.1162/99608f92.fbb1bdd6

ABSTRACT

What are the priorities for data science in tackling COVID-19, and in which ways can big data analysis inform and support responses to the outbreak? It is imperative for data scientists to spend time and resources scoping, scrutinizing, and questioning the possible scenarios of use of their work—*particularly* given the fast-paced knowledge production required by an emergency situation such as the coronavirus pandemic. In this article I provide a scaffold for such considerations by identifying five ways in which the data science contributions to the pandemic response are imagined and projected into the future, and reflecting on how such imaginaries inform current allocations of investment and priorities within and beyond the scientific research landscape. The first two of these imaginaries, which consist of (1) population surveillance and (2) predictive modeling, have dominated the first wave of governmental and scientific responses, with potentially problematic implications for both research and society. Placing more emphasis on the latter three imaginaries, which include (3) causal explanation, (4) evaluation of logistical decisions, and (5) identification of social and environmental need, I argue, would provide a more balanced, sustainable, and responsible avenue toward using data science to support human coexistence with coronavirus.

Keywords: COVID-19, predictive modeling, public health, surveillance, engagement, research planning

1. Introduction: Learning to Live With SARS-CoV-2

Over the coming years the human race will need to learn to live with the SARS-CoV-2 coronavirus—a biological entity that is now irrevocably entangled with our species, an invisible yet decisive part of our ecology and our social life. We are well past hopes of containment, with infections sprouting even in countries that successfully avoided the brunt of the initial contagion, such as Australia, Serbia, and New Zealand. The first wave of lockdowns in East Asia, Europe, and North America also transpired, and a second (worse) wave hit Europe at the time of this article going to press. While progress on developing vaccines is swift, it is doubtful that those will provide long-lasting immunity. Hence, governments, businesses, and communities around the world are grappling with urgent questions concerning how to manage social life, trade, education, communications, travel, and social services beyond the immediate response to a new threat. What are the priorities underpinning alternative construals of 'life with COVID'? How can general guidelines, common infrastructures, and mass-produced technologies help to stop the pandemic in the face of the diversity of geographies, politics, communities, and economic conditions around the world? Whose advice should be followed, whose interests should be most closely protected, which losses are acceptable and which are not?

Data science can play a variety of roles in helping to address these questions, and it is important for researchers working in this area to consider the expectations and assumptions underpinning alternative visions of data use in this emergent domain. The global scale and vast social and economic impact of the pandemic emergency offer a unique opportunity for both technical development and public engagement with data science and related technologies (Meng, 2020). Data have never been more prominent in public discourse, with animated debates in social media and news outlets concerning matters once relegated to technical conversations among policymakers and public health experts: Which clinical data should inform biomedical understandings of the impact of the disease on human health, how to compare death counts and contagion rates across countries, and what implications could extensive population monitoring have on social life and democratic structures?

This crisis has also opened a window for a decisive shift in the use of technology toward a greener, more sustainable, more efficient use of resources to support human life on earth. The pandemic has reminded governments and individuals alike of the fragility of the global economic system and its dependence on planetary health, including humans, nonhumans, and their environment. Moreover, it has offered a chance to take stock of the extent to which digital transformation has affected society and all sectors of the economy, and the challenges and unresolved issues that remain open—especially in locations and sectors where the transformative nature of digitalization has not yet been systematically explored and deployed, such as energy, social services, and health. Relatedly, the crisis has heightened the awareness of opportunities linked to the emergence of artificial intelligence (AI), including the promise of improving our global knowledge base in order to understand ongoing social changes and local vulnerabilities, and developing appropriate interventions. Transnational institutions are responding vigorously to this prospect, with the European Parliament instituting a commission for the development of AI-specific legislation, and UNESCO as well as the United Nations overseeing consultations over the implementation of extensive data mining, machine learning applications, and open science systems.

What, then, are the priorities for data science in tackling COVID-19, and in which ways can big data analysis inform and support these opportunities? A starting point in addressing this question is to note that social values and political commitments typically drive decisions about scientific priorities as much as technical considerations do. Much of COVID-19 research focuses on issues ranked as urgent by funders and policymakers, thus participating in a broader social and economic agenda for the pandemic response. This integration of social and scientific agendas is unavoidable, since scientific findings do not dictate or determine political decisions about what kinds of intervention are warranted within specific local contexts. Hence, politicians cannot rely on proclamations of being 'led by science' to justify their actions. It is of course crucial for politicians to give scientific evidence a central role in informing decision-making and regularly engage with the research community, particularly in a pandemic where reliable knowledge about the characteristics and impact of the disease is essential;

but translating the scientific findings into interventions remains the responsibility of politicians rather than scientists.

At the same time, scientists are responsible for designing, enacting, and producing the kinds of evidence and technology that inform decision makers, which does involve evaluating the political and social implications of their results. Indeed, awareness of the complex links between science and policy does not constitute an invitation for researchers to abdicate accountability toward the broader framing of pandemic response within which they are working. In other words, and in line with data ethics and responsible innovation recommendations (Leslie, 2020), it is imperative for data scientists to spend time and resources scrutinizing and questioning the possible scenarios of use of their work—*particularly* given the fast-paced knowledge production required by an emergency situation such as the coronavirus pandemic.

In this article, I provide a scaffold for such considerations by identifying five different imaginaries for how to use data to support the pandemic response and briefly exploring *the political and socioeconomic priorities associated with them*: (1) population surveillance, (2) predictive modeling, (3) causal explanation, (4) evaluation of logistical decisions, and (5) identification of social and environmental need. I define 'imaginaries of data use' as *the ways in which the data science contributions to the pandemic response are imagined and projected into the future*, and I consider how such imaginaries play out within public discourse, policy evaluations, as well as research practices in data science. As I shall emphasize, imaginaries (1) and (2) are most closely associated with fast, top-down interventions and have therefore dominated the first wave of governmental and scientific responses, in which researchers and policymakers alike were scrambling for short-term solutions. I argue that this strong focus on population surveillance and predictive modeling has problematic implications for both research and society, and that these concerns could be avoided through recourse to slower forms of data-intensive research that draw on multilateral, interdisciplinary, and inclusive exchanges within and beyond the scientific landscape. I conclude that placing more emphasis on the imaginaries (3), (4), and (5) could provide a more effective, sustainable, and responsible avenue toward using data science to support the pandemic response.

2. Imaginaries of Data Use

Imaginaries of data use typically involve a vision for how data and data science can most effectively foster life with coronavirus in the future. Such vision is typically linked to specific expectations about what technical, human, and institutional resources (including methods, skills, and supportive socioeconomic conditions) should ideally be developed and combined in order to effectively use data to address the emergency—without, however, presupposing that such resources are readily available. At the same time, such vision is seldom explicit in scientific or even political reasoning around data science strategies to confront the pandemic, and it is expressed through the practices and priorities of

everyday research work rather than programmatic statements or formalized ideologies. Hence the choice of the term 'imaginaries': rather than just collections of ideas, these are ways in which data science is routinely *imagined* and *performed* by researchers, policymakers, and various publics and stakeholders. They typically do not amount to a coherent plan or a systematic philosophy of data use; they are also not necessarily stable and can rapidly adapt to changing research conditions.

Nevertheless, they play an important role in shaping the future of data science, by informing current allocations of investment and priorities within and beyond the scientific research landscape. They therefore deserve attention and critical reflection as part of ongoing efforts to define and improve data science contributions to global challenges such as those posed by the pandemic. Indeed, as I argue here, there are imaginaries of data use that align particularly well with the urgency to produce fast solutions to a crisis, and yet may not be best suited to identifying sustainable interventions in the long term. Retaining an awareness of the plurality of imaginaries of data use in a pandemic, and the ways in which such imaginaries can complement and support each other, is crucial to advancing data science with and for society.

This use of the term 'imaginary' is inspired by social studies of socio-technical imaginaries (Jasanoff & Kim, 2009) and the large body of scholarship recently emerged under the label of critical data studies, which point to the significance of investigating the frictions underpinning data production, management, and use, and their relation to diverging visions for what data science can do for society and how (Borgman, 2012; Edwards et al., 2011; Kitchin, 2014). At the same time, these authors are careful to point to the inchoate, dynamic, and often implicit nature of imaginaries for data science (Poirier et al., 2020), and the difficulties found when attempting to draw rigid boundaries between these imaginaries—difficulties accentuated by the fluidity with which data and related methods, infrastructures, and analytics adapt to changing environments and needs (Leonelli & Tempini, 2020). As I write, the state of emergency associated with the COVID-19 pandemic, and the rapidity with which various forms of data-intensive research and related ideologies are being redeployed toward serving societal responses, are accentuating this fluidity—and, arguably, reshaping the landscape of data science in ways as yet impossible to fully grasp. The analysis that follows, grounded on my own interpretation of the tacit assumptions underpinning developments in this domain, is therefore by no means the only way to parse things out. Indeed, my aim here is not to propose a 'correct' taxonomy of imaginaries of data use, but rather to stimulate discussion around what constitutes such imaginaries, and reflection on the roles they play in researchers' choices and contributions.

2.1. Imaginary 1. Population Surveillance.

As populist politics and social unrest threaten to rise in response to prolonged lockdowns, democratic institutions have come under attack for their perceived inability to tackle the crisis, with their credibility hanging on their ability to marshal medical and social services toward an effective handling

of COVID-19 in the longer term. In line with epidemiological calls for 'tracking and tracing' the spread of infections, this has been interpreted as primarily involving the control and policing of population movements. Contact tracing—recording proximity among individuals, usually through location data extracted from mobile phones—has emerged as a key strategy to lift lockdown measures until an effective vaccine becomes widely available (Ferretti et al., 2020). Hence, population surveillance programs have acquired prominence as a key imaginary for data use informing governmental and scientific responses over the first few months of the pandemic.

According to this imaginary, the combination of new forms of big data and related technologies provide a uniquely effective means to monitor population behavior, including the ways in which individuals may contribute to spreading infection. This helps in locating the foci and manner of contagion spread and effectively limiting it, thus aligning with long-held epidemiological advice to trace, track, and contain the virus. The emphasis on following the virus via enhanced forms of population surveillance has focused the attention of researchers, policymakers, industry, and publics alike on specific types of data collection and analysis, such as: the use of data extracted from mobile phones and social media to locate users, track symptoms as they spread among groups and territories, trace contacts for any infected person to identify possible carriers, monitor whether people testing positive or at risk for COVID-19 do self-isolate as recommended by most governments, and understand how mobility and consumer behaviors—and particularly the flow of people as they travel—have been impacted by lockdown measures.

While these are no doubt significant avenues of investigation, the focus on population surveillance is unavoidably associated with the ethical and social concerns involved in managing vast amounts of sensitive data, which have been exacerbated by the pandemic (Leslie, 2020).[1] In response to the emergency, and in spite of cautionary advice from international organizations such as the European Union (EU) commission (European Commission [EC], 2020a/b), many national contexts have seen a centralization of governmental powers to access, collect, integrate, and systematically analyze data on their subjects, for which tracking apps provide an opportune vehicle. What happened in response to this move varied dramatically across countries. In the United Kingdom and much of Europe, attempts to centralize data collection and reuse caused an immediate backlash, with civil society organizations protesting the resulting potential for infringement of privacy, and the public health sector lamenting the potential loss of trust by prospective users—a serious problem, given that in those countries the effectiveness of the app depends on people's willingness to download it. Big tech industry stepped into the fray in the form of an alliance between Google and Apple, who swiftly repurposed their software to provide a fully decentralized and "privacy-preserving" tracing system (Kitchin, 2020). Some European countries adopted this decentralized solution; others, such as France, refused to privatize this service and developed their own apps in tandem with a sophisticated governance model for the resulting data (Krige & Leonelli, 2021).

Regardless, uptake of the apps by the population remained relatively low. In India, China, and Russia, by contrast, governmental control of data collection was not widely contested, and usage of tracing apps was tied to existing infrastructures for the provision of basic social services. Consider the tracing app Aarogya Setu. Developed by the Indian government as a high-profile public health intervention in response to the pandemic, this app dovetails with longer term Indian efforts to digitize citizenship and related social services via the national ID system Aadhaar. The use of Aadhaar has already been made mandatory by the government for large categories of workers, which paved the way for the wide application of COVID-19 automated monitoring (Sircar & Sardev 2020).

These developments make for a marked increase in the overall scope and depth of surveillance exercised over individuals, as well as creating space for a potential alliance between public and private data sources. Creating a large data pool to study population behavior is undoubtedly highly attractive to data scientists, yet it strengthens existing concern around the conditions under which governmental and industry data on individuals and groups could or should be shared, and who should take responsibility for brokering and monitoring such agreements (Zuboff, 2019). Moreover, in the rush to develop usable technical systems for data collection and reuse, basic sources of bias and inequity across the digital footprint are ever more likely to be ignored or papered over, as already widely documented in relation to racial and gender representation, as well as 'data poverty' in the Global South (D'Ignazio & Klein, 2019; Eubanks, 2018; Milan & Trere, 2020; Noble, 2018). This is particularly problematic considering the disproportionate impact that COVID-19 is having on vulnerable groups, including ethnic minorities and working-class workers who are most exposed to the virus (Kirby, 2020; Tai et al., 2020).

This use of technology is grounded in the expectation that big data and AI can play a key role in solving the epidemiological problem of tracking and tracing virus carriers, and thus act as a 'technological fix' to contain disease transmission and future outbreaks while also dramatically reducing costs. Three key assumptions about data on population movements ground this faith in a technological fix: (1) that the data are reliable and unambiguous in the information they convey, (2) that they are easily transformed into social, public health, and medical intervention (e.g., by testing and isolating contacts found to be at risk), and (3) that they are harmless in their long-term implications for democratic governance. As I shall now argue, all three assumptions are problematic.

First, these data do not speak for themselves. There is no uniformly reliable way to produce, visualize, and evaluate data around COVID-19 contagion and transmission, which places limits on the ease with which the data can be compared and analyzed. Not only are data sources highly uneven, but it is widely recognized that having real-time, reliable information on transmission involves having data about the majority of the population. This is an impossible goal given not only the low levels of trust in such data-collecting apps (evident, for instance, in the European context), but also the low numbers of

people in possession of an up-to-date smartphone with reliable internet connection (over 90% of Italians have smartphones; less than 30% of Indians do). Moreover, many tracing apps based on citizen data ignore the homeless, asylum seekers, and unregistered workers who tend to be excluded from citizenship or otherwise marginalized, and yet can act as a major vector for disease (as already witnessed in the rise of contagion hotspots around German and American factories employing migrants in poor conditions). In the absence of mitigating measures addressing these concerns, we are looking at a white-collar technology for the privileged few—except that, contrary to last generation iPhones, this approach only works if most people can access the technology and be counted. Moreover, achieving reliable and standardizable data interpretation involves comparing the different conditions under which data are produced and collected, including reliable metadata about the different testing strategies adopted by each country (and sometimes each region and municipality), the ways in which deaths and infections are verified and counted, and the resolution at which individuals' movements are tracked and shared. In other words, data are deeply *contextual*. As public health officials have long known, data mining only provides meaningful evidence for social interventions when its results are evaluated in relation to qualitative information, such as interviews with putative contacts to verify the accuracy of the signal and the potential for further transmission in each case.

Second, tracing apps need to be complemented by a health system with the capability to test contacts quickly and effectively and provide adequate local assistance. Given the enormous economic inequality and highly uneven access to medical care in India, it is not surprising that hospitals in major cities like New Delhi and Mumbai were quickly overwhelmed in June and July 2020, the implementation of tracing apps notwithstanding. Complementing monitoring with appropriate care and local expertise is difficult even in high-income countries such as Italy and the United Kingdom, where social and medical services have been decimated by austerity measures. The United Kingdom still has an extensive network of local public health officials, who, however, were not consulted on contact tracing during the first months of lockdown, despite being by far the best equipped workforce to implement it effectively. Instead, in April 2020 the British government hurriedly hired a 'small army' of untrained personnel to support and implement indications emerging from tracking technologies. As long as it remains disconnected from local public health expertise, such implementation risks being patchy and discriminatory, with a great degree of confusion around who will 'monitor the monitors' and how oversight will operate—a situation that was widely discussed as "a masterclass of mismanagement" (Ball, 2020), as demonstrated by the U.K. failure to implement a rigorous testing system well into the fall of 2020.

This brings me to the third assumption, concerning the links between technological fixes and democratic governance. In the absence of the contextual, interpretative, and intervention capabilities that would allow epidemiologically relevant meaning to emerge and help address local outbreaks, all we are left with is surveillance, with data generated in the cause of public health playing a purely

policing function. This is a particularly worrying scenario for India, given the episodes of police brutality used to enforce social distancing rules (Cousins et al., 2020) and ongoing security concerns with patients' personal data held by governmental agencies (Ranjan, 2020) —though it has also surfaced in the United Kingdom, as exemplified by recent debates over local police forces gaining access to tracing data in order to enforce quarantine. To make things worse, the dimming prospects of ever eliminating the SARS-CoV-2 virus make it likely that surveillance measures undertaken this year will outlast the immediate emergency. In early May 2020, use of Aarogya Setu was made mandatory for most public- and private-sector Indian employees, as well as anybody wishing to undertake train or air travel (Greenberg 2020). In China, the government has partnered with social media companies Alipay and WeChat to source data on users' location and body temperature and used them to control entry to public areas, including transport vehicles and offices (Gan & Culver, 2020) —and is considering extending such measures indefinitely (Davidson, 2020; Prado, 2020). Even in Europe, the adoption of measures to counter the abuse of public health data for population surveillance is not without critics, with much debate focusing on how the pandemic emergency may affect the criteria used to differentiate data use from data abuse.

As demonstrated by the history of the census (Thorvaldsen, 2017), the acquisition of high-resolution data documenting individuals' movements, social networks, and interests has long proved valuable to government and industry alike. At the same time, the long-term potential of extensive data collection to exploit masses of personal data is a matter of serious concern for democratic governance; and attempts to acquire data for the purpose of surveillance can backfire, as in the United Kingdom where the problems associated with the app reinforced perceptions of the ineffectiveness of government and technocratic interventions. However one evaluates the results of focusing research efforts on population surveillance, its centrality to data science imaginaries of the pandemic response is a cause for concern. It is not a given that surveillance and monitoring of movements should take priority among the kinds of epidemiological knowledge that help contain the pandemic. As I argue here, other types of data and data analysis can help to identify sources of vulnerability and need in the population in ways that effectively support transmission control, while also fostering the engagement and understanding of marginalized communities.

2.2. Imaginary 2. Predictive Modeling.

Widespread reliance on predictive modeling is often combined with population surveillance as the best way to capitalize on big data analysis. Epidemic modeling aims at prediction, rather than accurate representation of reality, and it is a significant strategy to simulate crucial phenomena such as the possible growth of contagion rates, the impact of specific public health measures, and the characteristics and implications of various post-emergency scenarios. This approach to data use takes a strongly pragmatic attitude, with a wide variety of heterogeneous data used as input for general

models geared to produce actionable predictions (Fuller, 2020). A well-known example is epidemiological models of the contagion curve developed by Neil Ferguson's group at Imperial College London in February and March 2020, which were adopted as an evidence base for the pandemic response in the United Kingdom and United States.

This is not the place to review the heated discussions surrounding these or similar models (Wynants et al., 2020), and particularly the long-standing arguments around the external validity of relevant extrapolations (Cartwright & Hardie, 2012; Fuller, 2019; Reiss, 2019; Rothwell, 2005; Steel, 2008). What interests me is the extent to which epidemic modeling has aligned with the rise of big data, and related expectations that the volume and variety of the data could make up for problems in sourcing, sampling, and calibrating the data.[2] Within this imaginary to data use, data are often understood as mere 'input' for models and, where relevant, machine learning algorithms. The parallel between this mode of envisioning data use and a more general approach to big data epistemology was aptly summarized by Mayer-Schönberger and Cukier (2013) as the triumph of *messiness* and *correlations*. In short, this view goes as follows: since big data are mostly garnered in the absence of the exactitude and accuracy charactering measurement under controlled conditions, as was certainly the case for clinical data on COVID-19 patients collected by different countries, analysts should focus on extracting "a sense of general direction rather than knowing a phenomenon down to the inch, the penny, the atom" (Mayer-Schönberger & Cukier 2013, p. 13). They should favor predictive knowledge deriving from correlation over explanatory knowledge obtained when looking for definite—but ever-elusive—causes.

It is no surprise that the lure of this imaginary of data use resulted in a rush toward producing predictive models based on the COVID-related data pouring in from governmental agencies and medical services. Under pressure to produce predictions that could support government interventions, many researchers reacted to the emergency by rushing to apply existing models to the incoming data, with a strong focus on producing projections for future trends in contagion rates. World-leading epidemiologists such as John Ioannidis and Marc Lipsitch weighed in by offering stark pronouncements on what they saw as the questionable (Ioannidis) or necessary (Lipsitch) nature of social distancing measures. Leading preprint repositories such as bioRxiv and medRxiv were inundated by manuscripts reporting on modeling results and related predictions, many of which had not yet been peer reviewed or validated, and yet were quickly picked up as reliable findings by mainstream media looking for scientific evidence for specific political interventions.

The resulting misuse of results, retractions of findings that turned out to be spurious, and related loss of public trust led to some preprint repositories temporarily banning manuscripts reporting simulated results. As curators wrote to prospective authors: "bioRxiv and medRxiv are not currently posting predictions of drug/therapeutic efficacy/potential for treatment of COVID-19 that are based solely on

in silico work (e.g., molecular dynamics simulations of protein interactions, metabolic network node analysis, etc.), given concerns about drug availability and dangers to the general public. These papers should instead undergo rapid peer review at a journal before dissemination. This has been a difficult decision not arrived at lightly and we understand it may disappoint some authors, but we currently feel this is the most responsible course of action in these exceptional circumstances." (from standard email from bioRxiv and medRxiv to authors, April 2020).

Indeed, as already noted above, data do not speak for themselves without adequate contextualization. COVID-19 relevant data, including seemingly homogeneous data such as localization and mobility data, are highly heterogeneous in their format and resolution, and not easily analyzed without a great deal of preparatory work (Christen, 2019). In fact, the decisions made during data cleaning and wrangling lead to the prioritization of some data types over others, depending on how easily they can be cleaned and fitted into available models. Good examples are testing results, which are widely viewed as essential parameters for epidemic models such as Susceptible, Infected and Recovered (SIR) models and yet are very uneven in the extent and manner in which they are obtained (depending on the scale and targets of testing in each country)—an unevenness that can make a big difference at this scale. Similarly, data on the death toll of the pandemic, despite seemingly straightforward, have proved to be among the hardest to reliably validate due to the diversity of measures used across nations, regions, and provinces, including differences in what and who counts as 'dead' and how an association with COVID was determined, for instance whether or not deaths outside hospitals would be counted and whether a formal test was needed to confirm COVID-19 as a cause.

This kind of messiness can in fact make the data less easily amenable to automated analysis such as performed by machine-learning (ML) algorithms and other forms of predictive modeling. Getting the data right in this case may be secondary to asking the right questions and focusing less on overarching trends and more on local scenarios. As an anonymous reviewer of this article has helpfully pointed out to me, predictive modeling around disease dynamics is arguably best positioned to support qualitative conclusions (e.g., regarding the relative efficacy of proposed interventions within highly well-specified conditions) rather than quantitative predictions (e.g., of the specific numbers of people in various states at time t).[3] It is also crucial to position the results of predictive modeling vis-à-vis other types of outcomes and expertise (Goldstein et al., 2020). Authors of a recent attempt to model the longer term effects of the Indian lockdown, for instance, are careful to point to the limits in their results, and the fact that they constitute only one source of evidence among many needed to take a final decision on the restriction of mobility (Ray et al., 2020). Most obviously, another key source of evidence concerns the broader socioeconomic setting within which predictions are supposed to apply, particularly in national settings where police have taken a heavy-handed approach to enforcing social distancing measures and little assistance has been provided to those affected by the disease and/or its socioeconomic consequences (Cousins et al., 2020). Moreover, within the context of the pandemic response, the

emphasis on correlation over causation is arguably not that helpful. Despite the emphasis on prediction, the search for causes underpinning the observed correlations matters enormously when attempting to understand the interactions between viruses, environments, and human populations, as I discuss in the next section.

2.3. Imaginary 3. Causal Explanation.

The use of data to investigate the biological causes and clinical manifestations of COVID-19 infections has been less prominently discussed in policy and the media than in the use of predictive modeling. It has also featured less prominently in data science circles. This may seem surprising since there are many clusters of biostatistical research where causal assumptions and inferences are carefully evaluated, including a plethora of methods for causal inference from observational data (Hernán, 2018). As recently argued in an authoritative review of such work, however, "the scientific literature is plagued by studies in which the causal question is not explicitly stated and the investigators' unverifiable assumptions are not declared. This casual attitude towards causal inference has led to a great deal of confusion" (Hernán & Robins, 2020, p. vii). Furthermore, the governmental attention to predictive modeling as a first port of call at the start of the emergency led to an increase in funding and support for researchers working in that domain. This is particularly troubling in the case of biomedical research on coronavirus, where causal explanation as an imaginary of data use continues to make important contributions to the pandemic response by increasing understanding of how COVID-19 infection is transmitted, the characteristics of the SARS-CoV-2 virus and related vaccines, and the effects of human exposure. Causal understanding is crucial to unravel the variety and interrelations of factors underpinning the disease, including the biological mechanisms of contagion, its social and environmental triggers (e.g., pollution), and the economic conditions for spread and slowdown. In turn, such causal understanding is essential toward informing decisions on how to safely organize society with COVID-19, including measures for social distancing.

Indeed, the analysis of evidence from the medical frontline, including both hospitals and physicians, has proved important in countering speculation around the usefulness of lockdown and social distancing measures, particularly in the face of predictive models that projected negligible public health outcomes from them. The analysis of clinical observations across different medical settings strongly affected by the first wave of the pandemic, such as Italy, China, South Korea, and the United States, demonstrated how for many people, including some not previously thought to be at risk, COVID-19 can be a vicious disease that can affect not only the respiratory system as initially surmised, but also the circulatory, lymphatic, and nervous systems impacted by oxygen deprivation. Numerous observations of silent hypoxia and of the surprisingly long-term effects of COVID-19 on patients have revealed the severe toll that the disease exacts on some of its victims, as well as the fact that it tends to remain hidden until requiring very lengthy—and in many cases ineffective—hospitalization. These

surprising characteristics of the disease are crucial parameters to take into account within predictive models, since they cause unprecedented pressure on wards overwhelmed with infectious, severely ill patients—potentially leading to the collapse of the medical system and a long tail of damaged patients, with effects that go well beyond the death toll exacted by COVID-19.

The acquisition of causal understanding requires two types of data work in addition to the predictive modeling discussed above. One is the creation of data under controlled and/or well-monitored conditions, with the specific aim of test correlations spotted in the data. Under this heading we find the hundreds of clinical trials jumpstarted in the first months of the pandemic to verify symptoms, potential treatments, and vaccination programs; exploratory experimentation on the virus and nonhuman hosts conducted in the lab; behavioral studies exploring the effectiveness of public health messaging (Sanders et al., 2020); as well as natural experiments conducted through access to population data in combination with sophisticated analytic tools, and without researchers retaining control over the condition of the experiment (Craig et al., 2017) —a particularly salient methodology given the urgency and the scale of the pandemic. Taken together, these data sources are most likely to satisfy the seminal Bradford Hill Criteria for inferring causation from association, which are respected and used by all branches of biomedicine and are strongly committed—particularly considering the latest advances in data science—to the use of a plurality of kinds of research as evidence base (Fedak et al., 2015).

The other type of data work is the integration of quantitative measurements and qualitative observations, including case reports as well as clinical and social observations not typically encompassed by predictive modeling. This is made particularly laborious by the lack of relevant data infrastructures, and thus of access to relevant data. Public health emergencies such as pandemics have virtually no dedicated global infrastructures bringing together observations, measurements, and case reports emerging from various fields and locations. This is due to a variety of factors, including the significant national and regional differences in approaches to the sharing of sensitive data; the scarcity of transnational venues dedicated to the open discussion of how to weigh evidence of multiple kinds; and the heterogeneity in formats for such data, with many medical systems still relying on analogue archival systems which, while guaranteeing physical control over data movements and thus increasing data security, make it virtually impossible to share the data with a wider audience for monitoring or comparative purposes.

The situation of public health data is in sharp contrast with the long history of climate data collection and related infrastructures, which goes back several centuries and includes a complex set of governance structures overseeing data analysis and reporting, and particularly the interpretation of predictions acquired through modeling (Edwards, 2010). Several efforts have been initiated during the pandemic to address this issue at both the international and national levels, including, for instance, the

COVID-19 Working Group of the Research Data Alliance, which, thanks to savvy use of support by many world-leading experts as well as national and international organizations (such as the EC, the International Council for Science, and the World Health Organization[WHO]), was able to produce a set of guidelines and resources to support coronavirus-related data sharing by the end of June 2020 (see Krige & Leonelli, 2021). More investment in such transnational collaboration, as well as material infrastructures and social institutions to support it, is arguably central to upholding the causal explanation imaginary of the usefulness of data science—as well as the three remaining imaginaries that I will briefly present in the coming sections.

2.4. Imaginary 4. Evaluation of Logistical Decisions.

Beyond attempting to gauge the implications and causes of the pandemic, data science can be used to inform the logistical and organizational demands of the 'new normal' associated with life with coronavirus. Evaluating the consequences of adopting specific technologies, platforms, architectures, and management models is crucial to the reorganization of medical and social services, as well as to post-lockdown arrangements in all working spaces, leisure facilities, and education establishments, not to speak of urban spaces more generally.

Just as surveillance strategies implemented this year may long outlast the emergency, so do decisions made about which workflow, organizational, and technological infrastructures to adopt in response to the pandemic—a situation that science and technology scholars have long labeled a 'technological lock-in.' And indeed, some data scientists are using agent-based modeling and other systems-theoretical methods to address organization-level resource allocation and assess the implications of logistical interventions—most notably at the start of the pandemic, by triangulating the effects of measures such as mask wearing with data on human movements within public spaces, to assess the impact of such measure on infection rates (e.g., Petrônio et al., 2020). It is therefore strange that the imaginary of data science as means to evaluate logistical decision has not featured more prominently in policy-making and public discourse around the outbreak response.

Consider the choice of which online communication tools to adopt in schools, hospitals, industries, and social services. This choice does not need to be blind or informed solely by the current popularity levels of a given service provider; after all, Facebook is used by 2.7 billion people, and yet there are well-documented reasons to mistrust the ways in which this platform handles user data and manages misinformation campaigns—which may well extend to its Messenger and WhatsApp communication services. Rather, the analysis of data pertaining to specific locations and types of activity (including broadband availability, user preferences and needs, and the past performance of the provider) can usefully inform the choice of communication tools to suit the situation at hand. Similarly, data science can be used to inform workers' movements across offices, ensuring that public spaces within any one organization can be utilized in full compliance with public health guidelines and without curtailing

civil rights; to model traffic and pedestrian movements around schools, stations, and airports, thus helping to avoid congestion and dangerously dense crowds; and to help coordination between volunteers and private as well as public organizations toward various forms of pandemic response, including the very collection and validation of outbreak data.

Exemplifying the latter strategy are the many citizen science initiatives hastily assembled to crowd-source data from the population, such as assembled under the website of Citizen Science (www.citizenscience.org/covid-19). These initiatives constitute community-based processes that play a crucial role in attracting new evidence on data-poor subjects and in validating (or countering) results obtained through other forms of research (Bowser et al., 2020). For instance, a partnership between Harvard University, Boston Children's Hospital, and the Skoll Global Threats Fund was able to hastily refashion its ongoing citizen science project on influenza ("Flu Near You," https://flunearyou.org/#!/) toward garnering coronavirus reports directly from U.S. residents ("COVID Near You," https://www.covidnearyou.org). Thanks to the long-standing relation between the Flu Near You project and international public health officials running similar projects, as well as the Centers for Disease Control and Prevention (CDC), these data can now be promptly and securely shared in ways that strongly support research efforts to collect and analyze data.

It is concerning that this type of data use imaginary has not emerged more prominently—and been more heavily supported—as part of national and international approaches to the pandemic response. There is a dire need for more, and more inventive, 'reflexive' uses of data science to investigate the pros and cons of specific forms of data management and smart working for the long term. A data-informed reshaping of data management practices could be framed as a central element of digital transformation programs for virtually all industries and services, which would help to develop solutions in tune with the specific objectives of each sector (many of which need anyhow to be reimagined at this time, such as decarbonization targets for the energy sector or alternative forms of smart working). Industry organizations are already moving in this direction, with ongoing discussion around data strategies to optimize working conditions under the pandemic. For instance, this was the theme of the IDC Digital Forum that took place in Italy in September 2020, in which I participated as an external speaker and which included companies ranging from transport services to city planning (https://www.idc.com/we/events/67327-idc-data-strategy). The opportunity for fruitful interactions between data scientists and more traditional management structures is wide open. Similarly, it is clear that regular exchanges between government agencies, social services, and data scientists are highly beneficial to all parties, whether or not under a state of emergency. This in turn requires the development and nurturing of effective channels of communication—a point I shall come back to in the following.

2.5. Imaginary 5. Identification of Social and Environmental Need.

Enhancing opportunities to identify and address social and environmental need is seemingly an obvious imaginary of data use that, however, has been largely overlooked in the public sphere during the first months of the pandemic. This may be partly due to the expectation that governments would already have a sense of what may be needed to respond to the coronavirus outbreak, when in fact the novel features of this virus and related social disruption were unlike the pandemic scenarios that most governments and international organizations had been preparing for. It may also be partly due to the widespread expectation that big data and related analytics are good 'fuel' for novel high-tech solutions (such as the tracing apps and the related imaginary of surveillance) but have less to offer when it comes to less gadget-focused demographic, epidemiological, and social understanding. It is certainly true that assessing what social and environmental concerns have emerged from the pandemic, for whom, and in which forms, is not typically conducive to developing easy fixes in the form of marketable products with a clear and measurable impact.

The lack of research incentives toward longer term, complex solutions has been aggravated by the marked disregard that some governments displayed for research attempting to understand the social circumstances and implications of public health interventions. Contrary to Germany, where the national response committee included philosophers and social scientists from the get-go, countries like the United Kingdom and the United States favored the expertise of modelers and epidemiologists over the skills of local public health officials, anthropologists, and sociologists. It could be argued that it makes sense to prioritize sources prepared to recommend urgent interventions over fields focused on longer term analysis—and yet, urgent interventions implemented without a sense of their broader social implications can be as dangerous as lack of action, as well as radically reducing the opportunities for improvement and advancement that may accompany the current social upheaval. In this sense, the pandemic response not only needs to learn from ongoing efforts to address the existential threats posed by climate change, but needs to be intertwined with the apparatus of scholarship, data infrastructures, and methodological approaches set up to investigate planetary health (defined as "the health of human civilisation and the state of the natural systems on which it depends" by Whitmee et al., 2015, p.1973; see also Pàllson, 2020, and https://www.planetaryhealthalliance.org/planetary-health).

An early example of data science used to identify social need concerned the demographics of the impact of the pandemic, which revealed the dramatic imbalance between the high death toll suffered by ethnic minorities and disadvantaged groups and the much lower toll suffered by wealthier and/or white individuals (Kirby, 2020). This was particularly pernicious in the case of frontline workers: in the United Kingdom, for instance, six out of 10 medical staff who died from exposure to the virus were black, Asian, and minority ethnic (BAME). This finding worked as an alarm bell to sensitize politicians

and the wider community toward the injustices amplified and expanded within the pandemic context. This attention to social circumstances as crucial to understanding the pandemic—and shaping any response—chimes with calls to think about the coronavirus outbreak as a 'syndemic': "a set of closely intertwined and mutual enhancing health problems that significantly affect the overall health status of a population within the context of a perpetuating configuration of noxious social conditions" (Bambra et al., 2020, p. 13). According to this view, the risk factors associated with a pandemic are intertwined with and exacerbated by specific social factors, in ways that exacerbate existing situations of disadvantage. In turn, the analysis of inequities in human societies becomes constitutive of research on the spread and dynamics of the outbreak. Preliminary research in countries like Austria has confirmed the extent to which research on the economic and social consequences of the pandemic—including people's own understandings and experiences—complements and strengthens research on outbreak responses, including both epidemiological studies and public attitudes to governmental policies (Prainsack, 2020).

Another promising strand of research concerns the levels of exposure to pollutants by different parts of the population, with several recent studies demonstrating that air pollution contributes significantly to the spread of the virus and that ethnic minorities are disproportionately exposed to harmful chemicals, regardless of income and background (a phenomenon dubbed "environmental racism," Washington, 2020—and examined as part of extensive scholarship on the social determinants of disease, e.g., Abrams & Szefler, 2020, and Van Bavel et al., 2020). Data science can make enormous strides in supporting this kind of research, due to the novel opportunities to cross-reference, integrate, and mine data sourced from very different fields, phenomena, and locations. Even studies using data to identify problems seemingly unrelated to the pandemic, such as energy poverty (a high proportion of income being needed for a family to be comfortable and warm in their accommodation), turn out to provide important clues for pandemic-related policies and long-term social shifts—for instance, by formulating energy-saving tips for people whose finances and health have been compromised by the outbreak.

As pointed out earlier in relation to the need to contextualize data, this imaginary of data use requires a mix of quantitative and qualitative data sources, encompassing the expertise of data subjects as well as data analysts. In other words, it is not only data on social determinants that matter, but also data about the practices and experiences of people.[4] This can involve comparisons between data extracted from social media and data collected from local volunteering groups that provide mental health support; or complementing mortality data across regions with testimonies from local medical services and transparent information about which key workers have had access to protective equipment (a seemingly obvious approach, except in the United Kingdom where medical staff was explicitly barred from complaining about lack of equipment on public platforms). These forms of data and data analysis help to document the differential impact of lockdown restrictions on women and ethnic minorities,

and inform policies explicitly geared toward supporting these groups. Incorporating such expertise is key to obtaining robust data and insights about the social impact of COVID-19.

This imaginary of data use also requires openness to comparing and cross-referencing a variety of different situations, within and beyond national borders. In February and March 2020, the data that emerged from Italian medical institutions provided tragic factual insight into the material consequences of containment failure in the early stages of disease outbreak. Balkan countries like Greece and Slovenia, strongly attuned to the experiences of other Mediterranean countries and aware of the relative weaknesses of their own medical systems, were quick to act on such knowledge, resulting in early lockdowns and very low numbers of fatalities—especially when compared to the tens of thousands who died in the United Sates, Brazil, and United Kingdom between March and June 2020. Despite the efforts to compile and update comparative data analysis by the WHO and many other international agencies, the initial impetus within too many countries in the Global North was to focus on national-level data rather than transnational comparisons, and quickly think through appropriate solutions at the national level. By contrast, what underpins this imaginary is a vision of data use as a window into understanding multiple social situations and fostering solidarity across them. Big and open data can be enormously helpful in understanding different realities and evaluating alternative futures, by improving knowledge of other ways of life, empowering diverse voices, and, perhaps most importantly, enabling comparisons across regional contexts. In other words, this imaginary of data use can serve as an effective antidote to inward-looking, politically controlled social perspectives.

3. Reframing Data Science to Facilitate Effective Interventions

The first months of the pandemic saw many calls for emergency data science as requiring a dramatic acceleration in the pace of research, with solutions needed urgently and researchers under enormous pressure to deliver socially transformative results within hours. As a counter to such pressure, I am arguing that less rushed and more engaged forms of data-driven research not only remain crucial, but have in fact become ever more significant given the potential longer term impacts of the paths taken in the coming months. Pandemic data science requires pointed reflection on long-term strategy, not blind panic and knee-jerk reactions.

I do not mean to advocate that data science undertaken under emergency conditions cannot be fast, nor to endorse a tired and simplistic juxtaposition between fast and 'safe' science (whatever the latter may mean). On the contrary: It is perfectly possible to develop transnational, socially attuned, effective solutions—including extensive collaborative networks—in the space of a few weeks. Many of the examples given above, including the COVID Near You initiative and the COVID-19 Working Group of the Research Data Alliance, demonstrate how this can be done. What those examples also demonstrate, however, is the decisive role played by existing, longer term infrastructures, networks,

and venues for collaborations among diverse experts and relevant communities. Developing such resources often takes significant investment over decades and sometimes yields no immediate returns, yet it is key to the fast deployment of data collection and socially engaged analytic services in times of crisis. Perhaps most significantly, developing data science on the basis of such socially robust institutions and infrastructures arguably helps provide solutions that, while not the fastest to emerge, are actually fast to implement—because they are already aligned with social needs and expectations, and thus have more resilience and built-in flexibility than top-down interventions evaluated in the abstract by a small group of experts.

The problematic juxtaposition of fast versus safe data science parallels another, more widely discussed false dichotomy: that between civil liberties and public health, which have often been pitted against each other when evaluating whether infringements of key rights such as data privacy were warranted by the need to effectively track the spread of disease. As highlighted by extensive inquiries run by the Ada Lovelace Institute, the European Commission, and the American Civil Liberties Union, among others, there is no principled reason to pit these two key concerns against each other, particularly given that an understanding and fundamental respect for civil liberties and social concerns strengthens both data production and data analysis (Ada Lovelace Institute, 2020; EC, 2020a; Guarglia & Schwartz, 2020; Kitchin, 2020; Stanley & Granick, 2020; Allen et al 2020; National Academies 2020). Rather, these tensions derive from a practical obstacle: that is, the lack of venues, incentives, and time for data scientists to engage with research on data governance and ethics as well as nonacademic stakeholders that can voice social concerns—and to explore how such engagement informs the development of algorithms and data models.

This in turn involves abandoning the temptation of the low-hanging fruit by exploiting existing research strengths in high-powered locations, and instead devoting more resources toward involving multiple stakeholders in the collection, validation, and reuse of data sets and models. The blatantly transformative role of digitalization for all parts of society, and increasing public awareness of its shortcomings, provides a fertile terrain for dialogue. Precisely because of the urgency of an emergency response, investing in community engagement in data collection, validation, and processing is not 'a waste of time': such engagement makes data more robust and the resulting knowledge more reliable (Milan & Trere, 2020). As reported by Milan and Trere (2020, p.3), "Chenoweth and colleagues (2020) have documented over 154 strategies of collective action specifically related to COVID-19. Their preliminary mapping displays the incredible richness of these novel online, offline and hybrid repertoires of contention, that include grassroots tactics of 'data making' (Pybus et al., 2015) at the margins, where vulnerable groups and their allies become active producers and consumers of alternative narratives to reclaim their visibility amid the pandemic." At the same time, researchers with long-standing pedigrees in community engagement are now in a position to conduct highly innovative and impactful studies that simply would not be possible without that existing network. For

instance, current understandings of migration patterns have been boosted by studies of how refugee communities cope with the new challenges presented by the pandemic (Milan, 2020).

It is thus possible and desirable for researchers to move away from data collection as a top-down exercise in surveillance, and toward collaborative, engaged forms of data work that seek to understand social and environmental needs, evaluate research directions, and construct appropriate tools in dialogue with relevant communities. Community engagement is crucial to obtaining robust data as well as robust data use and outputs; this can be enormously strengthened by collaboration with qualitative social scientists and humanists who specialize in contextualizing data and evaluating the implications of proposed technical solutions.

The argument for a more socially robust and environmentally sustainable data science of service to the pandemic response brings us back to the fundamental question of who counts as a data scientist in this context. It is clear that there are many crucial roles for data science to play at this time, which demand an ever-expanding range of skills and expertise. This is in line with the 'ecosystem view' of data science espoused in the editorial to the first issue of *HDSR* (Meng, 2019), which sees this domain as a catalyst for contributions from several different disciplines, including both STEM and SHAPE subjects (respectively Science, Technology, Engineering and Mathematics, and Social Science, Humanities and the Arts for People and the Economy). To devise data solutions and adequate visualizations for the pandemic response, data scientists need expertise well beyond the technical realm of computer science and data analytics, including epidemiology and public health, biology and genomic analysis, public policy and governance, social science, cultural studies, behavioral science, and mental health.

This expansion of the range and scope of data-intensive analysis comes with a wider range of accountabilities. Data scientists need to abandon the myth of neutrality that is attached to a purely technocratic understanding of what data science is as a field—a view that depicts data science as the blind churning of numbers and code, devoid of commitments or values except for the aspiration toward increasingly automated reasoning. Data science is sometimes regarded as a methodological field, a sort of generalist toolkit that can be credibly and reliably put to the service of a vast array of goals (as Xiao-Li Meng put it to me when commenting on this article, a "tool discipline" ready to serve any master). And it is certainly possible for data scientists to behave in this way, by taking no interest in the broader context and political interest underpinning their work, and churning numbers for the highest bidder. The high level of confusion and contradictory advice emerging from the same data set speak not only to technical disagreements on how to visualize, analyze, and interpret data, but also to different stances on what masters are worth service, and whose interests are served. *Data science needs to stop feigning neutrality, and instead work collaboratively with domain experts and relevant communities toward forging socially beneficial solutions.* As convincingly argued by prominent scholars across virtually all fields, including the emerging field of critical data studies to which the journal *Big*

Data & Society is dedicated, it is imperative that data scientists take responsibility for their role in knowledge production.

Those researchers who work on surveillance and predictive models should ask themselves what actions can be prompted by their work, and whether their recommendations can be realistically implemented. Whether or not a given government is capable of providing local public health support matters when deciding the technical specifications of a tracing tool. Democratic and accountable ways to imagine and implement data use require eschewing the technocratic mindset that underlies testing and tracking regimes in too many places today and investing in different forms of data analysis and infrastructures—including transparent and accountable forms of governance for the sharing of sensitive data across public and private organizations. Such investments strengthen the reliability and comprehensiveness of data samples, thus also facilitating the responsible use of data for surveillance, prediction, and explanation.

4. Conclusion: Fast Data Science Need Not Be Rushed

At the time of writing, many new data science projects are being developed in the wake of stimulus packages set up to respond to the economic distress caused by the COVID-19 pandemic. It is desirable that this emerging data science work combine the different imaginaries discussed above, so that the focus on surveillance and prediction is accompanied by a serious attempt to understand the cultural, social, and environmental contexts in which research is performed. It is perfectly feasible to stretch one's imagination to consider several of the imaginaries discussed here at once, especially when working with an interdisciplinary and engaged group—as exemplified by some of the initiatives reported in this article, these imaginaries need not be mutually exclusive. Working in one of these modes, without taking time to consider others, creates significant risks for data science research: It fosters conservatism and a tendency to apply existing methods without evaluating their adequacy to the research context at hand; it discourages consultations across different stakeholders around which data to collect, how to share them and with whom, and for which purposes; and it reduces researchers' ability to combine sophisticated analytics (such as involved in predictive modeling) with cutting-edge insights on what such methods could achieve for society (as obtained through logistical analysis and engagement with relevant communities). By contrast, the ability to combine different imaginaries of data use can help data scientists to look beyond short-term solutions and develop robust, novel approaches to the concerns at hand.

The failure of most governments to adequately prepare for this pandemic can be interpreted as a failure of imagination—an interpretation underwritten by the WHO's admission that diseases other than influenza were not given appropriate attention during the last decade of preparations against global outbreaks. The same danger holds for the pandemic response: Tired appeals to support the achievement of a supposedly uniform 'new normal' across the world could constitute a severe

drawback to scientific advancement as well as social well-being, especially given the unfolding environmental crisis and the possible emergence of new pandemics in the near future. There is a real risk of recycling technical solutions with no long-term sustainability in the hope of eventually stumbling into an easy and effective 'technological fix' for COVID-19, such as an effective vaccine or a well-functioning tracing app. The much-touted political tendency toward homogeneous solutions and a 'return to normality' goes together with the technical emphasis on accelerating research to produce easy fixes. Across all five imaginaries of data use discussed above, what emerges instead is the significance of data science in fostering a localized, situated, procedural understanding of the conditions and behaviors most likely to stem transmission and improve (not just human, but planetary) health. This involves spending time and resources to consider which priorities data projects need to heed and how research needs to be organized to serve those priorities, with a clear focus on creating and maintaining avenues for data scientists to engage with other experts and relevant communities. It also involves a reimagining of social life as well as data work: Both are multiple, situated, and contextual in their most robust forms; both call for dialogue between many perspectives and forms of expertise in order to achieve sustainable solutions.

Why is surveillance and monitoring of movements taking priority in contemporary public discourse, particularly in the United Kingdom and the United States? Credible and useful knowledge can be obtained via the analysis of many different types of data for explanatory, exploratory, or comparative purposes. Yet, a conversation about alternative applications of data science, and the ways through which data should be sourced in the first place, occupied a vanishing space in relation to the technocratic regime that has taken hold of much of the scientific and political response strategy in the first months of the pandemic. This technocratic regime strongly aligns with the exceptionalist, nationalistic, top-down, and paternalistic narratives favored by some prominent politicians; the political unwillingness to devote resources toward supporting crucial institutions such as schools, social and environmental services, and local councils, working with communities on the ground, and thinking about locally adaptive solutions rather than 'one size fits all' (see also Jennings & Ellis, 2016); and a view of 'public trust' as fickle, unreliable, yet pliable—something to be monitored, controlled, and directed in the right ways, very much like individuals during a pandemic.

This understanding of public trust could be understood as the opposite of trust in scientific claims, which is supposed to spring from context-independent qualities such as reliance on well-established methods, empirical data, and logically sound reasoning—qualities that confer trustworthiness on the outputs of research. And yet, these two visions of trust align in one crucial respect: the exclusion of social, contextual factors from evidential reasoning, and thus a disregard for the conditions under which data are generated and interpreted and the wide varieties of expertise and consultations required to understand and improve such conditions (Leonelli, 2019). This is a misguided and autocratic view of science, to match a misguided and autocratic view of public trust. It is crucial for

data scientists to be alert to manifestations of these pernicious views and structure their research and goals as a counter to this. The ideology of surveillance that so far dominated public discourse around using big data to tackle the pandemic is not the most useful, imaginative, and sustainable approach for data scientists to embrace in the longer term. A multidisciplinary, reflexive, socially attuned, and engaged approach to research can go a long way toward fostering robust, reliable, and responsible outcomes, despite requiring more time and resources to set up. Emergency data science can be fast, but should never be rushed.

Disclosure Statement

This research benefited from funding from the Alan Turing Institute under EPSRC grant EP/N510129/1, the European Research Council under ERC award 335925.

Acknowledgments

This paper benefited from relevant discussions with: Xiao-Li Meng, Kaushik Sunder Rajan, Brian Rappert, Michel Durinx, Rachel Ankeny; participants to presentations of these ideas hosted by the University of Cambridge, IV ESCO Tech Forum, FORCE11, R2B OnAir 2020, IDC Italia and Ammagamma; and members of the "Exeter Data Crunch" hosted by the Exeter Centre for the Study of the Life Sciences (Egenis) and the Institute for Data Science and Artificial Intelligence. I am also indebted to four anonymous referees for extremely helpful comments.

References

Abrams, E. M., & Szefler, S. J. (2020). COVID-19 and the impact of social determinants of health. *The Lancet Respiratory Medicine 8(7): p.659-661.* https://doi.org/10.1016/S2213-2600(20)30234-4

Ada Lovelace Institute. (2020). *Exit through the app store? Rapid evidence review.* https://www.adalovelaceinstitute.org/our-work/covid-19/covid-19-exit-through-the-app-store/

Allen, D., Block, S., Cohen, J., Eckersley, P., Eifler, M., Gostin, L., Goux, D., Gruener, D., Hart, V., Hitzig, Z., Krein, J., Langford, J., Nordhaus, T., Rosenthal, M., Sethi, R., Siddarth, D., Simons, J., Sitaraman, G., Slaughter, A. M., Stanger, A., … Glenweyl, E. (2020). *Roadmap to pandemic resilience.* Edmond J. Safra Centre for Ethics at Harvard University. https://ethics.harvard.edu/covid-roadmap

Ball, J. (2020). The UK's contact tracing app fiasco is a master class in mismanagement. *MIT Technology Review.* June 19 2020. https://www.technologyreview.com/2020/06/19/1004190/uk-covid-contact-tracing-app-fiasco/

Bambra, C., Riordan, R., Ford, J., & Mathews, F. (2020). The COVID-19 pandemic and health inequalities. *Journal of Epidemiology & Community Health, 74*(11), 964–968. https://doi.org/10.1136/jech-2020-214401

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Bowser, A., Parker, A., & Long, A. (2020, June 22). Citizen science and COVID-19: The power of the (distanced) crowd [Blog post]. *CTRL Forward.* https://www.wilsoncenter.org/blog-post/citizen-science-and-covid-19-power-distanced-crowd

Cartwright, N. (2012). Will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science, 79*(5), 973–989. https://www.journals.uchicago.edu/doi/abs/10.1086/668041

Cartwright, N., & Hardie, J. (2012). *Evidence based policy: A practical guide to doing it better*. Oxford University Press.

Chenoweth, E., Choi-Fitzpatrick, A., Pressman, J., Santos, F. G., & Ulfelder, J. (2020). *Methods of dissent & collective action under COVID: A crowdsourced list*. Crowd Counting Consortium. https://docs.google.com/spreadsheets/d/179hz-OKrfcAr3O0xi_Bfz9yQcK917fbLz-USxPZ3o_4/edit - gid=0

Christen, P. (2019). Data linkage: The big picture. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.84deb5c4

Cook, S. N., & Wagenaar, H. (2012). Navigating the eternally unfolding present: Toward an epistemology of practice. *The American Review of Public Administration*, *42*(1), 3–38. https://doi.org/10.1177/0275074011407404

Cousins, T., Leonelli, S., Pentacost, M., & Rajan, K. S. (2020). Situating the biology of COVID-19: A conversation on disease and democracy. *The India Forum*. 19 June 2020. https://www.theindiaforum.in/article/situating-biology-covid-19

Craig, P., Katikireddi, S. V., Leyland., A., & Popham, F. (2017). Natural experiments: An overview of methods, approaches, and contributions to public health intervention research. *Annual Review of Public Health, 38*(1), 39–56. https://doi.org/10.1146/annurev-publhealth-031816-044327

Davidson, H. (2020, May 26). Chinese city plans to turn coronavirus app into permanent health tracker. *The Guardian*. https://www.theguardian.com/world/2020/may/26/chinese-city-plans-to-turn-coronavirus-app-into-permanent-health-tracker

D'Ignazio, C., & Klein, L. F. (2019). *Data feminism*. MIT Press.

Edwards, P. (2010) *A vast machine: Computer models, climate data and the politics of global warming*. MIT Press.

Edwards, P., Mayernik, M. S., Batcheller, A., Bowker, G., & Borgman, C. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science, 41*(5), 667–690. https://doi.org/10.1177/0306312711413314

European Commission. (2020a). *Coronavirus: A common approach for safe and efficient mobile tracing apps across the EU*. https://ec.europa.eu/digital-single-market/en/news/coronavirus-common-approach-safe-and-efficient-mobile-tracing-apps-across-eu

European Commission. (2020b). *Guidelines on the use of location data and contract tracing tools in the context of the COVID-19 outbreak*. https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_20200420_contact_tracing_covid_with_annex_en.pdf

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fedak, K. M., Bernal, A., Capshaw, Z. A., & Gross, S. (2015). Applying the Bradford Hill criteria in the 21st century: How data integration has changed causal inference in molecular epidemiology. *Emerging Themes in Epidemiology*, *12*, Article 14. https://doi.org/10.1186/s12982-015-0037-4

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dorner, L., Parker, M., Bonsall, D., & Fraser, C. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science, 368*(6491), Article eabb6936. https://doi.org/10.1126/science.abb6936

Fuller, J (2020, May 5). Models v. evidence. *Boston Review.* http://bostonreview.net/science-nature/jonathan-fuller-models-v-evidence

Fuller, J. (2019). The myth and fallacy of simple extrapolation in medicine. *Synthese.* https://doi.org/10.1007/s11229-019-02255-0

Fuller, J., & Flores, L. J. (2015). The risk GP model: The standard model of prediction in medicine. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 54*, 49–61. https://doi.org/10.1016/j.shpsc.2015.06.006

Gan, N., & Culver, D. (2020, April 16). *China is fighting the coronavirus with a digital QR code. Here's how it works*. CNN Business. https://edition.cnn.com/2020/04/15/asia/china-coronavirus-qr-code-intl-

hnk/index.html

Goldstein, N. D., LeVasseur, M., & McClure, L. A. (2020). On the convergence of epidemiology, biostatistics, and data science. *Harvard Data Science Review, 2*(2). https://doi.org/10.1162/99608f92.9f0215e6

Greenberg, A. (2020, June 5). India's Covid-19 contact tracing app could leak patient locations. *Wired.* https://www.wired.com/story/india-covid-19-contract-tracing-app-patient-location-privacy/

Guarglia, M., & Schwartz, A. (2020, March 10). *Protecting civil liberties during a public health crisis*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2020/03/protecting-civilliberties-during-public-health-crisis

Hernán, M. A. (2018). The C-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health, 108*(5), 616–619. https://doi.org/10.2105/AJPH.2018.304337

Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *Chance*, *32*(1), 42–49. DOI: 10.1080/09332480.2019.1579578

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.

Jasanoff, S., & Kim, S. H. (2009). Containing the atom: Sociotechnical imaginaries and nuclear power in the United States and South Korea. *Minerva, 47*(2), 119–146. https://doi.org/10.1007/s11024-009-9124-4

Jennings, B., & Ellis, B. A. (Eds.) (2016). *Emergency ethics: Public health preparedness and response*. Oxford University Press.

Kirby, T. (2020). Evidence mounts on the disproportionate effect of COVID-19 on ethnic minorities. *The Lancet, 8*(6), 547–548. https://doi.org/10.1016/S2213-2600(20)30228

Kitchin, R. (2020). Civil liberties or public health, or civil liberties and public health? Using surveillance technologies to tackle the spread of COVID-19. *Space and Polity.* Advance online publication. https://doi.org/10.1080/13562576.2020.1770587

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. SAGE.

Krige, J., & Leonelli, S. (2021, in press). Mobilizing the translational history of knowledge flows: COVID-19 and the politics of knowledge at the borders. *History & Technology*.

Leonelli, S. (2019). *La Recherche Scientifique à l'Ère des Big Data: Cinq Façons Donc les Données Massive Nuisent à la Science, et Comment la Sauver.* Éditions Mimésis.

Leonelli, S., & Tempini, N. (2020). *Data journeys in the sciences*. Springer.

Leslie, D. (2020). Tackling COVID-19 through responsible AI innovation: Five steps in the right direction. *Harvard Data Science Review* (Special Issue 1-COVID-19). https://hdsr.mitpress.mit.edu/pub/as1p81um

Mayer-Schoenberger, V., & Cuckier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. John Murray.

Meng, X.-L. (2019). Data science: An artificial ecosystem. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.ba20f892

Meng, X.-L. (2020). COVID-19: A massive stress test with many unexpected opportunities (for data science). *Harvard Data Science Review* (Special Issue 1-COVID-19). https://doi.org/10.1162/99608f92.1b77b932

Milan, C. (2020). Refugee solidarity along the Western Balkans route: New challenges and a change of strategy in times of COVID-19. *Interface: A Journal for and About Social Movements* 12(1): p.208-2012. https://www.interfacejournal.net/wp-content/uploads/2020/05/Chiara-Milan.pdf

Milan, S., & Trere, E. (2020). The rise of the data poor: The COVID-19 pandemic seen from the margin. *Social Media + Society, 6*(3). https://doi.org/10.1177/2056305120948233

National Academies of Sciences, Engineering, and Medicine. (2020). *Evaluating data types: A guide for decision makers using data to understand the extent and spread of COVID-19*. National Academies Press. https://doi.org/10.17226/25826

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Pàllson, G. (2020). *The human age: How we created the Anthropocene epoch and caused the climate crisis*. Welbeck Publishing Group.

Petrônio, C. L. S., Batista, P. V. C., Hélder, S. L., Alves, M. A., Guimarães, F. G., & Silva, R. C. P. (2020). COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals, 139*, Article 110088. https://doi.org/10.1016/j.chaos.2020.110088

Poirier, L., Fortun, K., Costelloe-Kuehn, B., & Fortun, M. (2020). Metadata, digital infrastructure, and the data ideologies of cultural anthropology. In J. Crowder, M. Fortun, R. Besara, & L. Poirier (Eds). *Anthropological data in the digital age* (pp 209–237). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-24925-0_10

Pybus, J., M. Coté, & T. Blanke (2015). Hacking the social life of big data. *Big Data & Society*, 2(2). https://doi.org/10.1177/2053951715616649

Prado, S. (2020, April 6). *Coronavirus: Surveillance helps but the programs are hard to stop*. Bloomberg. https://www.bloomberg.com/news/articles/2020-04-06/coronavirus-surveillance-helps-but-the-programs-are-hard-to-stop

Prainsack, B., Kittel, B., Kritzinger, S., Boomgaarden, H. (2020, June 16). *The coronation of Austria, Part 5: Time for a syndemic perspective?* Medium. https://medium.com/@bprainsack/the-coronation-of-austria-part-5-time-for-a-syndemic-perspective-57b307273af1

Ray, D., Salvatore, M., Bhattacharyya, R., Wang, L., Du, J., Mohammed, S., Purkayastha, S., Halder, A., Rix, A., Barker, D., Kleinsasser, M., Zhou, Y., Bose, D., Song, P., Banerjee, M., Baladandayuthapani, V., Ghosh, P., & Mukherjee, B. (2020). Predictions, role of interventions and effects of a historic national lockdown in India's response to the COVID-19 pandemic: Data science call to arms. *Harvard Data Science Review* (Special Issue 1-COVID-19). https://doi.org/10.1162/99608f92.60e08ed5

Ranjan, R. (2020, May 11). Madhya Pradesh app to track patients leaks personal data, taken offline. *Hindustan Times*. https://www.hindustantimes.com/india-news/mp-app-to-track-patients-leaks-personal-data-taken-offline/story-WO7ATpaxOMDTsmUxSKduUO.html

Reiss, J. (2019). Against external validity. *Synthese, 196*, 3103–3121. https://doi.org/10.1007/s11229-018-1796-6

Rothwell, P. M. (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet, 365*(9453), 82–93. https://doi.org/10.1016/S0140-6736(04)17670-8

Sanders, M., Stockdale, E., Hume, S., & John, P. (2020). Loss aversion fails to replicate in the coronavirus pandemic: Evidence from an online experiment. *Economic Letters* 109433. https://doi.org/10.1016/j.econlet.2020.109433

Sircar, S., & Sachdev, V. (2020, May 2). *Not just red zones, new rules make Aarogya Setu mandatory for all*. The Quint. https://www.thequint.com/tech-and-auto/aarogya-setu-app-mandatoryfor-containment-zone-red-zone-orange-zone-all-employees

Stanley, J., & Granick, J. S. (2020, April 8). *The limits of location tracking in an epidemic*. ACLU. https://www.aclu.org/sites/default/files/field_document/limits_of_location_tracking_in_an_epidemic.pdf

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.

Tai, D. B. G., Shah, A., Doubeni, C. A., Sia, I. G., & Wieland, M. L. (2020). The disproportionate impact of COVID-19 on racial and ethnic minorities in the United States. *Clinical Infectious Diseases*, Article ciaa815. https://doi.org/10.1093/cid/ciaa815

Thorvaldsen, G. (2017). *Censuses and census takers: A global history*. Routledge.

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., & Drury, J. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour* 4(5), P460-471. https://doi.org/10.1038/s41562-020-0884-z

Washington, H. A. (2020). The fight against environmental racism needs data. *Nature, 581*(708), Article 241. https://doi.org10.1038/d41586-020-01453-y

Whitmee, S., Haines, A., Beyrer, C., Bolz, F., Capon, A. G., de Souza Dias, B. F., Ezeh, A;, Frumkin, H., Gong, P., Head, P., Horton, R., Mace, G. M., Marten, R., Meyers, S. S., Nishtar, S;, Osofsky, S. A., Pattanayak, S. K., Pongsiri, M. J., Romanelli, C., Soucat, A., Vega, J., Yach, D. (2015). Safeguarding human health in the Anthropocene epoch: Report of The Rockefeller Foundation–Lancet Commission on planetary health. *The Lancet, 386*(1007), P1973–2028. http://dx.doi.org/10.1016/S0140-6736(15)60901-1

Wynants, L., Van Calster, B., Bonten, M. M. J., Collins, G. S., Debray, T. P. A., De Vos, M., Haller, M. C., Heinze, G., Moons, K. G. M., Riley, R. D., Schuit, E., Smits, L., Snell, K. I. E., Steyerberg, E. W., Wallisch, C., & van Smeden, M. (2020). *Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal*. BMJ 2020; 369 :m1328 doi: https://doi.org/10.1136/bmj.m1328

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for the future at the new frontier of power*. Profile Books.

---

otherwise indicated with respect to particular material included in the article. The article should be attributed to the authors identified above.

Footnotes

1. While there are many ways to define 'sensitive' data, including legal definitions such as that offered by the General Data Protection Regulation in the EU, I understand this term broadly as data meant to capture information about individuals or groups that could be used to harm these individuals or groups. ↩

2. Note that this section focuses especially on epidemiological prediction rather than other types of predictive modeling (like clinical diagnosis and prognosis). ↩

3. Philosophers of science have made a similar point in relation to predictive modeling in evidence-based medicine (e.g. Cartwright, 2012; Fuller & Flores, 2015). ↩

4. For a review of understandings of 'practice' of relevance to data collection and use, see Cook and Wagenaar (2012). ↩