**Harvard Data Science Review • Special Issue 1: COVID-19:
Unprecedented Challenges and Chances**

COVID-19: A Massive Stress Test With Many Unexpected Opportunities (for Data Science)

Xiao-Li Meng^{1,2}

¹Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge,
Massachusetts, United States of America,

²Harvard Data Science Review, Harvard Data Science Initiative, Harvard University, Cambridge,
Massachusetts, United States of America

Published on: May 14, 2020

DOI: <https://doi.org/10.1162/99608f92.1b77b932>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

When I conceived my first editorial title this year, [“2020: A Very Busy Year for Data Science \(and Scientists\),”](#) COVID-19 was an unknown term to me. Had I been given any hint of its potential for tormenting the human society, I might have at least contemplated the possibility of ‘an unexpected year for data science.’ Indeed, not just the challenges but also the opportunities fall into the category of the unexpected.

All COVID-19-related challenges are virtually the same in nature: a massive stress test on a global scale. The pandemic pushes essentially every system to its extreme, exposing the good, the bad, and the ugly of its inner workings. It tests each system’s resilience (or lack thereof) as a well-functioning integrant of the human ecosystem. We have seen this stress test being conducted in public health and medical systems; in social and economic systems; in political and administrative systems; in law enforcement and regulatory systems; in educational systems; in production and supply chains; and in tourism and service industries, just to name a few of the most obvious ones. I am sure that each of you have your own extensive list, which may vary a great deal depending on your locality, in the physical, cultural, and social senses of the word.

Thinking more positively, the mantra ‘never waste a crisis’ exhorts us to make the best out of the unprecedented global experiments foisted upon us by COVID-19. Consider the study of environmental impacts of human activities. Imagine a global A/B test that imposes home confinement on the vast majority of human beings for a month and then measures the reduction of the carbon footprint compared to that from the preceding month. Prior to COVID, would anyone in their right minds have actually proposed such a research plan to an Institution Review Board (IRB), much less a government or funding agency? And yet, we have now been living out exactly this scenario for more than a month. How could we possibly waste such an otherwise unthinkable opportunity, especially when it comes at the incalculable cost of the many lives and livelihoods destroyed? Is there not a better way to try to make all this suffering count?

This opportunity to study environmental impact is merely one of the many extreme social experiments in progress that would have otherwise been deemed inconceivable, unfeasible or even unethical. Again, just to list a few obvious global scale studies that have presently become possible: the effectiveness of working from home (and the impacts of this on work-life balance); the utility of virtual education or home-schooling versus that of classroom-based learning; the impact of social isolation on mental health and personal relationships; the consequences of establishing societal control regimes and human monitoring versus protecting individual autonomy (and privacy); a global comparison of public health and medical (insurance) systems, as well as political, regulatory, legal, and economic systems. Even for experiments that would otherwise be possible, the pandemic may still provide the best window of opportunities to implement them. The article by Alexander Podkul, Liberty Vittert, Scott Tranter, and Alex Alduncin on [learning about public understanding of exponential growth](#) is an excellent example. As the authors rightly conclude, “One often laments that the general population does not have the time or resources to understand the data science behind social scientific concepts—this is a rare scenario in which most now have both.”

Whether meeting challenges or leveraging opportunities, data scientists have both the expertise and professional obligation to play vital roles. This call to action is highlighted and well summarized by the two panel discussions featured in this special issue. The first features [eminent biostatisticians and epidemiologists](#), and the second, [leading economists, sociologists and operation researchers](#). As summarized by the second panel, “Data science will be useful in estimating prevalence and in analyzing the ongoing natural experiment on pandemic response, in which different counties, states and countries are taking different approaches and implementing these at different stages of community penetrance.”

Data scientists are vital for a very simple reason. None of the aforementioned investigations can be meaningfully carried out without quantitative investigations, whether in the forms of risk and uncertainty assessments, data collection, processing, and analysis for reaching experimental conclusions, or effective data visualization and communication to maximize impact. Executing these tasks, however, becomes extremely daunting in the context of a global public health crisis. They require swift, unprecedented, and concerted efforts from the data science community, as I have already witnessed and experienced from organizing this special issue. We find ourselves in a perfect storm.

Perfect Storm

Carrying out a serious data science project is never easy; any data scientist who has gotten their hands dirty could prattle on for hours, if not days, about their knotty experiences. The typical sources of frustration and difficulty come from the so-called ‘Five Vs of Big Data’: Variability, Variety, Veracity, Velocity, and Volume. It’s fortuitous that ‘Variability’ comes first in their alphabetical ordering (and perhaps even more so considering that, as a ‘meta-variability’ reminder, the lists of ‘Vs’ one can find online have a large variability, from 3Vs to 42Vs!). As I discussed in my [previous editorial](#), broadly speaking, data science is about discerning which parts of the variation, or variability, are signal and which are noise. In the context of understanding the COVID-19 pandemic, the grand challenge is to be able to make such distinctions from a large variety of extremely low quality data, and we must make the call *now*. That is, the first four Vs are interacting in the most challenging way imaginable: dissecting variations in a large variety of data is a daunting task even when they are completely trustworthy and time constraints are not an issue. (Ironically, the fifth V, Volume, a popular indication of ‘Big Data,’ is scarcely an issue for COVID-19.) But COVID-19 data are anything but trustworthy for the purposes of societal and population-level planning, whether it comes to estimating the size of infected populations, the public understanding of risk, or the prediction of social and economic impact. We also must race against the spread of the virus—waiting for an encompassing plan next week may result in more lives lost than adopting a reasonable partial solution today.

COVID-19 thus is also a massive stress test on data scientists, because it creates a perfect storm for data science. We are given extremely dark data, bafflingly complex problems, very little time, an enormous number of affected stakeholders, and life-or-death consequences. The phrase ‘unprecedented challenges’ is often used with this or that theatrical intention, but in the current case it does not even capture the trying tasks we data

scientists have taken upon ourselves. In my three-decade involvement in editorial work, from serving as a reviewer to Editor-in-Chief, this is the first time I have received messages from the authors such as, “This has been an intense effort (one that almost brought me down, actually, though alas I prevailed),” or “students and I met at 8 p.m. every evening for two hours in the past 10 days!” or “The speed of things is scary and I am worried about the stress this mad rush is putting us through.”

The mad rush contrasts with the “intellectual’s otherwise punctilious rhythms,” in the words of an author featured in this special issue. The ‘punctilious rhythms’ of scholarly work are typically the result of the time that serious researchers and scholars take to contemplate, compare, contrast, confirm, etc., even or especially when they apparently have achieved their goals. Such rhythms are particularly crucial for data scientists, because it is too easy to ‘get a result’ in the data science space. Dump any data in some statistical software or machine learning package, and chances are that one will get some output, and even more likely that one can then come up with some ‘intuitions’ to rationalize the results. With so much at stake for the current pandemic, and with so many stakeholders anxiously waiting for almost any kind of information, the potential dangers posed by such ‘results’ simply cannot be overstated.

Extremely Dark Data

The vast majority of COVID-19-related studies published or posted up to now have been conducted far more thoughtfully than mere exercises of software-driven number crunching. Yet even thoughtful work can be dangerous when investigators do not take into account the extremely low quality of most of the available COVID-19 data. Building elaborated epidemiological or mathematical models without taking into account the weaknesses of the data generating mechanism is still statistically unprincipled, because data quality fundamentally trumps everything else. A vivid demonstration of this was provided by Walter Dempsey, [in a tweet he posted on April 5, 2020 \(@walthdempsey\)](#), using an identity I published in 2018 in *Annals of Applied Statistics*. (I generally avoid citing references in my editorials other than the ones being featured in *HDSR*, but the source of the identity is needed here because the mathematical proof it contains compels us to accept the seemingly impossible fact below.)

Assuming merely a half-percent data-selection correlation (between testing positive and being tested), Dempsey shows that selectively testing 10,000 people for SARS-CoV-2 in New York State is statistically equivalent to testing about 20 individuals randomly for the purpose of estimating the positive rate in New York State. This is a 99.8% loss of sample size, a quantification and qualification of the term ‘extremely low quality’ or ‘extremely dark’ data, since we have very little reliable data to carry out the usual statistical procedures that would otherwise rely on the input data forming a representative sample of our target population. Indeed, as Dempsey emphasizes, the half-percent data-selection correlation is likely an underestimate for COVID-19 testing because the selective testing is not induced by individual behaviors but explicitly imposed by government and health organizations due to the shortages of available tests.

Selective testing is necessary for medical purposes and for saving lives when there are not enough tests available. But studies that use reported positive cases from such samples without adjustment for estimated population infection rates would be seriously misleading in general, and the more selective testing we do, the more liable we are to mislead ourselves. This is because, without taking into account that the usable data size (i.e., the effective sample size) is so small, we would be drawing on the grossly inflated sample size to measure and assess uncertainty (for instance, in determining the margin of errors). This would lead to a host of problems, for example, extremely narrow confidence intervals, almost certainly missing the target. That is, we would be trapped by the ‘Big Data Paradox’: the more data, the surer we fool ourselves, as discussed in the aforementioned 2018 article.

To make matters far worse, selective testing is just one of many data quality problems for analyzing COVID-19 data, and it is the easiest or, at least, the most visible one to deal with. For anyone who wishes to gain an overall picture of the troubled COVID-19 data landscape, the article on [identifying biases in estimating COVID-19 case fatality rate \(CFR\)](#) by Anastasios Nikolas Angelopoulos, Reese Pathak, Rohit Varma, and Michael I. Jordan in this special issue is an essential read. Their graphical model, with over a dozen nodes, depicts vividly the nature of the many-headed beast of COVID-19 data complications. They categorize these complications approximately into five sources of bias: under-ascertainment of mild cases, time lags, interventions, group characteristics, and imperfect reporting and attribution. They also provide an extensive (but by no means exhaustive) discussion of the magnitude and direction of the biases caused by these sources. They rightly emphasize that the total error of any estimator, however perfect for addressing some of these biases, cannot be determined without seriously improving the data quality (e.g., via random testing) because these biases may go in opposite directions, leading to partial cancellations.

As an example of the change of direction in the bias, we may note that before the symptoms and risks of COVID-19 were widely understood and well-popularized, deaths due to COVID-19 were likely to be misclassified, and postmortem tests were rarely performed or even possible. But by now, if someone tested positive for COVID-19 and then died, it is much more likely that the death would be counted toward COVID-19 mortality, even though that person could have died in the same period from causes other than those related to the infection. This is especially the case for members of vulnerable populations. Whereas death attribution in the presence of competing risks is never a trivial matter, the task of properly assessing the excessive mortality rate due to COVID-19 is of great importance for local and national planning. For example, it is essential for figuring out when and how to end the lock-down—a situation which can, in itself, have devastating consequences (e.g., for those who need timely medical attention for non-COVID-19-related conditions). Unfortunately, modeling COVID-19 dynamics with competing risks has not been “at the forefront of what the epidemiological modelers are considering,” as pointed out in the aforementioned first panel conversation. This conversation touches upon a wide range of testing-related issues (such as PCR verses serologic tests) as well as other topics like vaccines and treatments. The panelists also broach other research-based challenges,

emphasizing, for instance, the importance of collecting better data and using more comprehensive models for assessing the impact of COVID-19 on health care and social and economic systems.

The issue of economic and social impact is the focus of the second panel conversation, which explores some long-term, even potentially permanent, implications (e.g., on telemedicine, or increased working from home). The second panel also echoes the first panel in calling for better estimates for epidemiological modeling: “for this we need disease and antibody testing on representative samples of the population.” This sentiment was further stressed by another article in the special issue. [Ray et al.](#) document the heroic efforts of a data science team to help the government of India to decide on the duration of the country’s lockdown period. These authors emphasize the “pivotal role of increased testing, reliable and transparent data, proper uncertainty quantification, accurate interpretation of forecasting models, reproducible data science methods, and tools that can enable data-driven policymaking during a pandemic.”

Timely Impact and Time-Honored Rigor

All the challenges discussed above condense into a sort of meta-challenge for (data) scientists: how best to maintain scientific rigor under extreme time constraints? What I have observed during the process of organizing this special issue is beyond heartwarming, and it gives me great hope that collectively we can have a timely impact while preserving time-honored intellectual rigor. Authors, editors, and reviewers all put in the kind of efforts that, during normal times, would be considered unachievable or even unimaginable. Specifically, given the time-sensitivity of the issue at hand, *HDSR* invited reviewers to complete their reviews within one week, regardless of how technical a submission was. While one week might be considered as normal pace for some fields, for data science journals it is virtually unheard-of. Considering the constraints on everyone, especially the experts we needed, the board members and I initially sent review requests to more people than needed in anticipation of decline. To our pleasant surprise, nearly everyone invited responded right away, and almost all of them provided reports within a week. Many reports were of great depth and posed challenging questions, from quality of data to choices of investigative approaches or models, and from effective communication of findings to careful considerations of policy implications.

This, in turn, puts extraordinary pressures on the authors, with some articles receiving a half dozen detailed reports in one week. Whereas some authors characterized the collection of the reports as being ‘excessive’—and it is by any normal standard both in terms of the quantities and asks—authors put in even more extraordinary effort in addressing *all* comments, and they did so within 1-2 weeks. Frankly, I was not expecting that all the authors would be in a position to address all the comments, given their ‘excess’ and the time constraints. I was therefore prepared to serve as a ‘mediator’ between some authors and reviewers, since if there was ever a particular time to stress the aphorism ‘let not the perfect be the enemy of good,’ this would be it. To my great surprise (and relief), my mediation skills (such as they are) were not needed at all. The authors did such a thorough job that the reviewers were deeply impressed, and some of them apparently were as pleasantly surprised as I was.

For example, David Leslie’s article on [tackling COVID-19 through responsible AI innovation](#) tripled its content and now has over 200 references. One reviewer reacted to this expansion: “It is much more grounded in evidence, has a solid thread of argument—improved by orders of magnitude.” Another remarked, “I am not sure if I have ever seen such a big difference between two paper drafts.” This is by far the longest article in *HDSR* so far, and while length is not a good measure for substance, Leslie’s article is truly a treasure for anyone cares about ethical issues in AI and data science. It even comes with a fascinating appendix on ‘The Normative Dimension of Modern Scientific Advancement,’ depicting the normative-historical perspective the author took in explaining “how the successful development of a particular set of inclusive and consensus-based social practices of rational problem-solving carried out in the face of insuperable contingency has relied upon a corresponding release of the moral-practical potentials for cognitive humility, mutual responsibility, egalitarian reciprocity, individual autonomy, and unbounded social solidarity.”

The article by Angelopoulos, Pathak, Varma, and Jordan on estimating case fatality rate (CFR) went from addressing only the issue of time lag to providing a much more comprehensive roadmap for identifying all kinds of biases in the COVID-19 data and their likely magnitudes and directions, earning the praise that, “it’s a *significant* improvement from the previous version” (emphasis is original). Ray et. al’s article received six anonymous review reports, and the authors provided a 33-page point-by-point response, the most substantial one I have ever seen, as echoed by another reviewer: “The effort the authors put in to the review was herculean. One of the most comprehensive response to reviewers I have ever seen.” With the permission of the authors and reviewers, we will post the 33-page response online, together with some other great comments and responses, demonstrating, at least partially, the tremendous efforts made by authors, reviewers, and editorial members. More importantly, these comments and responses should help readers, especially future data scientists, to understand better the intense, under-the-hood processes of producing and publishing rigorous data science through dialogue-driven scholarly exchange, critical evaluation, revision and expansion. In particular, they illustrate how experts push each other vigorously to seek bigger pictures, deeper contemplation, broader impact, clearer acknowledgment of limitations—all integrated components of scholarly work and, more broadly, of intellectual pursuits of the highest caliber.

But There is No Free Lunch

No matter how hard we work and how desperately we desire a particular outcome, we must also understand and accept the fact that there will always be limits that we cannot surmount. Understanding limitations and boundaries is a key telltale marker of the difference between experts and novices. For example, there is a mad dash around the globe right now to create vaccines for COVID-19 by people and entities with various motivations, capabilities and resources. Yet the vast majority experts, including those participants on the first panel, are telling us that we are at least 18 months away from seeing vaccines available for general public. There are at least three broad lessons we can draw from this reality, especially for the data science community, most broadly conceived.

The most obvious (yet often sidelined) lesson is that scientific processes and products must be approached with a due respect for their own internal rules of development and maturation. Admittedly, forces such as willpower, passion, and fear all have their critical equatorial roles in the human ecosystem. For example, when we want or avoid something badly, our will power will drive us to put in maximal effort and prioritize measures of attainment or avoidance to the extreme—the current global lock down is a perfect illustration. Yet there will be boundaries that we simply need to respect, regardless how omnipotent we may believe ourselves to be. In vaccine development, a key limitation is our theoretical understanding of how a vaccine will actually work in the wild when released into the general population, and hence we must conduct the time-consuming and diligent clinical trials, a point I shall discuss more shortly.

For the data science community, the limitation of the data is one fundamental boundary we must respect. Anything beyond the boundary requires a potentially hazardous leap of faith. There is simply no free-lunch. Any time that we think we are enjoying a free lunch, we deceive ourselves either because we didn't realize that we actually have paid for it (e.g., in the form of uncheckable model assumptions) or that someone else has paid for it (e.g., using prior information built on a previous study). The aforementioned article by Ray et al. illustrates vividly how one can use model assumptions and previous studies (e.g., on earlier SARS outbreaks) to circumvent the limitation of the COVID-19 data. Explicitly recognizing assumptions made and information borrowed is essential for assessing the reliability and validity of our results, because they depend critically on how reasonable our assumptions are, and how relevant the borrowed information is.

The second lesson is the critical importance of theory, or at least theoretical guidelines, especially when we need answers under stringent time constraints. As mentioned above, a key component of the lengthy waiting period for vaccines is the clinical trial. Clinical trials are very expensive and time-consuming, and the success rates for vaccines are among the highest when compared to trials in other disease areas such as cancer, as demonstrated by Andrew W. Lo, Kien Wei Siah and Chi Heem Wong in their timely article on [estimating probabilities of success of vaccines and other anti-infective medications](#). Nevertheless, they are still typically below 40% even for the most resourceful vaccine developers. Clinical trials are currently the gold standard for establishing the effectiveness of preventative and therapeutic medications. But this very fact reflects the fundamental limitation of our theoretical understanding of how a medication will work in a general population, despite the tremendous progress we have made in biomedical, life, and behavioral sciences. Though this is a major bottleneck in our quest for the speedy development of any vaccine, it is also—to mix metaphors—an essential backstop that prevents an overstretching of the limits of research capabilities and human understanding. We must test vaccines out empirically and with established and reliable clinical methods. There is simply no short cut in this process, if we want to conduct our trials scientifically, statistically, and ethically.

For the data science community, these kinds of bottlenecks perhaps are best understood by comparing how an expert versus a novice approaches modelling. Equipped with good background knowledge and theoretical understanding, experts would not need to waste time on trying out inferior models or methods; but novices,

who have to rely on trial and error to discover what works, have no such recourse to experienced judgment. No statistician would (knowingly) fit an ordinary linear regression to a binary outcome, but a novice might or even might try to defend such a choice (as happened in a local data science seminar several months ago).

A less trivial example is the potential use of black-box algorithms to discover patterns that relate to the etiology of a disease like COVID-19. Using opaque algorithmic techniques may be necessary in areas we know very little about, but that's exactly because we don't yet have sufficient theoretical understanding of these areas. Pattern-learning empirically in complex situations requires a lot of data to do well, but in the case of a pandemic, we hope to have as few 'training samples' as possible. With lives and livelihoods at stake, we neither have the time nor the desire to try and err with unreliable quick-fixes where the problem itself calls for the sensible pursuit of skilled data science. Sound theoretical understanding and insight is essential for timely, quality decision and actions. Hence we should push for more theoretical knowledge in everything we do, especially in areas where solely empirical investigations have been prevalent.

Indeed, many of the studies on assessing the current pandemic were put together quickly, thanks partially to the theoretical epidemiological framework of the SEIR model and its various extensions; see the aforementioned articles by Ray et al., or the article by Shomesh Chaudhuri, Andrew W. Lo, Danying Xiao, and Qingyang Xu on [Bayesian adaptive clinical trials for anti-infective therapeutics during epidemic outbreaks](#). These trials have the potential to speed up the drug approval processes but with the theoretically understood and controlled trade-offs between false positive and false negative errors. At the same time, the popular theory for the SEIR model is a deterministic one, which fails to capture all the uncertainties in our data (especially because of their extremely low quality) or in our modeling configurations (e.g., the basic reproduction number). This in turn forces us to conduct sensitivity studies to check how robust our results are given these uncertainties and configurations. All these endeavors take time and resources (but these labor-intensive commitments should not be prohibitive), hence the more well-developed theory we have, the faster we can make progresses.

I also want to emphasize that by 'theory,' I do not mean merely mathematical or scientific theory. David Leslie's article, mentioned previously, demonstrates the crucial role that philosophical, ethical and social theories can play in guiding responsible AI innovation for tackling COVID-19. An interdisciplinary and holistic approach that integrates many different streams of theoretical understanding can only enrich the insights of the data science community and bolster its contributions to confronting the complex set of clinical, epidemiological, and socioeconomic problems that surround the pandemic. In a broader sense, as data scientists, we can greatly sharpen, expand, and strengthen the many valuable perspectives that already comprise the cross-disciplinary character of data science by reflectively integrating them with even more from the humanities and the social sciences. This is a considerable challenge but also an exciting opportunity that still lies ahead of us.

The third lesson is the importance of having well-defined metrics for assessing and evaluating success and failure. All clinical trials have or should have clearly pre-specified end-points. Whereas the nature of these

endpoints may vary, they must be predetermined to prevent cherry-picking, or over-fitting, to use a data science term. A recent conversation with a leading physician fighting the COVID-19 pandemic reminded me how important such pre-specifications are to medical experts. An almost last-minute switch of the end-point of a recent COVID-19 treatment trial from preventing fatality to shortening recovering time for the survivors made this physician and his colleagues highly skeptical about the treatment's efficacy. They should, of course, have trepidation. When so much is at stake, there are all sorts of motivations and incentives that seduce us to make premature or otherwise misguided declarations of success. But regardless of whether our motivations are noble (e.g., saving those in desperate need) or evil (e.g., making a quick profit irrespective the risks of harming human life), how a treatment actually works is not at the command of our wishes. Whatever time it takes to understand a treatment and to ensure it works will simply have to be taken. Any attempt to push for faster 'success' via manipulation of processes or products—other than efforts to improve the trial object (treatment, vaccine, method, etc) itself—is both unethical and dangerous.

The fact that the medical community is taking time to develop COVID-19 vaccines by following well-established scientific protocols should give us more confidence in their ultimate effectiveness, despite the painful waiting period. On a smaller and more minor scale, the same can be said about the greatly revised articles in this special issue. Whereas the revision process may have been painful or even brutal for some authors, the end products that we see in this special issue are likely among the most trustworthy and informative data science work on COVID-19 available now and in the near future.

I'd like to conclude this special editorial by offering my heartfelt thanks to all the people who have made this special issue possible, and my best wishes to everyone for maintaining the most sound mind (and sleep) during this globally stressful time. I also want to invite everyone to check out this special issue on a weekly basis, as we roll out more COVID-19-pertaining articles, for as long as it takes. But let's also hope (and pray) that it will not take too long! Until then, stay connected, albeit not physically.

Acknowledgments

I am deeply grateful to comments and edits from authors, especially David Leslie and Andrew Lo. As with all my previous editorials, Susanne Smith did a marvelous job to prevent my Chinglish from overflowing, and for the lossless compression.

Disclosure Statement

Xiao-Li Meng has no financial or non-financial disclosures to share for this editorial.

©2021 Xiao-Li Meng. This editorial is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the editorial.