Interview with Xiao-Li Ming

For podcast release
Monday, June 8, 2020

KENNEALLY: Data scientists everywhere today are heeding this sage advice: Don't waste a crisis. The coronavirus pandemic presents them with opportunities to explore important social and scientific questions. Welcome to Copyright Clearance Center's podcast series. I'm Christopher Kenneally for *Beyond the Book*.

In the COVID-19 era, data scientists have the expertise and a professional obligation to play vital roles, says Harvard statistics professor, Xiao-Li Meng, the founding editor in chief of the *Harvard Data Science Review*. Professor Meng asserts that research based on datasets can yield important insights, from the efficacy of virtual learning and the impact of declining air pollution to best practices for vaccine development. The coronavirus crisis, he asserts, amounts to a massive stress test for data science at a critical time in the field's development. Professor Meng joins me now from his Boston area office.

Welcome to *Beyond the Book*, Professor Xiao-Li Meng.

MENG: Thank you so much Chris, for having me here.

KENNEALLY: We look forward to exploring with you the issues that you raise in a special online edition of the *Harvard Data Science Review*. It looks at COVID-19 unprecedented challenges and chances, and you make the point in your editor's note that COVID-19 pushes every system to extremes. It's a point that you're not alone in making. President Obama made the same point recently, that COVID-19 reveals the flaws in our society, in our systems, as much as it shows up the strengths. Can you tell us how that applies to the field of data science?

MENG: Sure, I'm very happy to. As I wrote there, and you mentioned, this is not a new point. Indeed, it is a massive test on virtually almost anything. And I have to say it hit data science particularly hard.

As I wrote in my first editorial for *Harvard Data Science Review*, data science, itself, is what I called an artificial ecosystem since there was many subjects from humanities, such as philosophy. Think about the issue of all the ethical issues, are from social science, for example, thinking about responding behaviors to

questionnaire surveys. And of course from physical and mathematical science, computer science, statistics.

So the stress test has revealed that – or at least confirmed a lot of issues. I want to just talk about one issue, which is that there is a lack of global protocol of data collection, or even understanding of the differences. Yet you have seen that the comparison of countries and systems all the time. But the issue here is because different countries, different systems, they have different definitions, different ways of collecting data. So we're not even talking about comparing apples with oranges, which at least they are still round fruits. We're really talking about comparing apples with bananas or maybe even oranges with footballs. The way I've been thinking about it is that how important it is to have a common global standard. Unfortunately that is likely way too much to ask for now, even forever, to have a common global standard. But what we need is really a converter, very much like electric plugs, no matter where you go, you wanted to have it convert to convert to the local system, so you can at least make things work to avoid being in the dark. This COVID-19 shows clearly that we're far from even having a converter, and I think that's a lot of problems for data science itself.

KENNEALLY: Well, the points you make, Professor Meng, regarding the need for a global standard is especially important because of course this is a global pandemic and the issues that COVID-19 is raising are global issues. It's also opening a window into a whole variety of ways of looking at our planet that we've never had a chance to do before. Tell us about these global case studies that were unthinkable only a few months ago, and some of the kinds of projects that are being undertaken right now.

MENG: Well, thank you for that very timely question. It just happened, it was reported by the *Journal of Nature Climate Change* that a study has estimated that compared to April 2019, the daily average of the human generated CO2, globally, was down by 17% this past April. Now as a statistician I have to say that I need to look into the study to know, understand how to collect the data, how reliable these estimates are. So I won't vouch for the number yet, but it's a good indication of the kind of global studies that are otherwise just impossible, unimaginable.

But just imagine that pre-COVID-19, if someone proposed, say, well, let's shut down the society for months and just see how it affects the pollution, everybody else would be thinking that's just a joke, even to well-intended. But now we're exactly living in such a situation. So this is just one of many examples the study being conducted, or should be conducted.

Let me just mention quite a few. You think about studies of the effectiveness of working from home on a massive scale, thinking about online education versus residential educations. Thinking about the impact – study the impact of these social isolations on mental health, not to mention all the studies comparing globally the similarity and the differences between medical, social, economical, political, you name it, all kinds of systems. So I think anyone who's interested in the betterment of the society, we have a lot of things to do, and people say never waste a crisis, especially where this one comes with so costly life and livelihoods, we should all take advantage of this unprecedented opportunity to study a lot of things otherwise not just possible.

KENNEALLY: The kinds of data science that you've carried on for almost 20 years at Harvard University, they have been confined, if you will, to Harvard Yard. But today, with the COVID-19 crisis, the issues that you're discussing matter to everyone on the planet. And so I wonder about the challenges that you see when communicating data science research to the public. The consumers of data science, the amateur consumers that we are all becoming, I suppose, should be asking questions whenever they're presented with research.

MENG: That's an excellent question. Let me say that, as a data scientist, started as statisticians, the work we do were always rooted in applications we're working on. Work on problems with astronomers and psychiatrists, psychologists. But you are absolutely correct, now we have a much broader audience to communicate to. For me, the central challenge in this kind of environment is how do you communicate honestly the large uncertainties in your study, and at the same time, ensure the public can trust and have confidence in what we report, and our ability to get the best out of this worst data or the situation. This has happened quite often that when you present things – because as a scientist, we tend to present things in a very nuanced way. We talk about all kid of uncertainties, all kinds of caveats, but sometimes that projects a sense in some general public view is we're just not sure what we're doing. Why do we trust these experts? They don't know what they're doing.

And so I think the key challenge here is how do we help the population understand the sources of these large uncertainties? They're not due to scientists ability to analyze the data, but rather because these data themselves are of very low quality. And how do you explain all these complexities in simple ways that people say, OK, we understand. You guys doing the best and that's the best that we can understand. The best situation actually still was lots of uncertainty and that's exactly where we're at now. We have so much uncertainty, far more than most people realized

probably in their life because we have so many things that are hanging on, those things we just don't know, but we have to make some decisions.

KENNEALLY: And do you feel that data scientists have a responsibility to open up their notebooks and reveal their methods?

MENG: Oh, I think absolutely. Other than sometime you do run into confidentiality issues, like if you do medical studies you obviously want to protect a patient's privacy. But other than that, I think we should be as open as possible for at least two purposes. One is that it's the way to gain public trust. How you know people usually trust you more when you just tell them, here's what I've done. Now I just say look at what I – here are the results. But the second is really to also really help educate and inspire future generations of aspiring data scientists, youngsters to coming to study data science because it's fascinating. When you look at what we do, it's just so complicated, but we still have to make some sense of out of it. There's mathematics, there's statistics, there's computer science, there's philosophy, there are social science, behavior, all kinds of studies into it because – Let me put in a very simple term. What data scientists is an educated guess. They try to guess the best of the data we will provide. So you can see how much fun and how much frustration we have. So I definitely want to open this as much as possible for education purpose, for general trust purpose, all kinds of purposes.

KENNEALLY: And we are months into this pandemic, and as you say the data gathered on COVID-19 is admittedly of low quality. At least it's early quality, put it that way. But why does that matter, and maybe why shouldn't it matter nevertheless? We have a real need to get working and to find some solutions to these problems.

MENG: Absolutely. This is related to what I just said. When you're trying to make educated guess, you're trying to do the best out of what you have, but you still have to understand what you have could be really, really low quality. In fact, one thing we do is try to avoid ourselves being misled. Now I would say that you put it mildly, saying it's a low quality, early quality. I'm going to say some of the data are of extremely low quality, and I can actually quantify that. The reason it's extremely low quality is because of the selective nature how these data were collected.

Now some of the selective natures are not in anybody's control. We're all trying to save lives, whatever we can do, we do it. Other selective natures actually well intended. For example, we do know that we tend to test more of those people who are showing the symptoms or more likely to get infected. I think that's for very good medical and ethical reasons. We should test those people first, especially

when we don't have enough tests available. But if someone just use these data to estimate the actual infection rate in the population, they would grossly overestimate the infection rate because you are only testing people more likely going to have this terrible disease.

As I wrote in my editorial for this special issue, that we currently tend to have good understanding, the low quality data tend to lead to low quality findings. That's well understood. What is much less understood and even appreciated is how low it can go, quantitatively. As an example given in my editorial, if you test at this moment 10,000 people in New York State in this selective fashion as we are currently implementing it, for the purpose of estimating the state-wide infection rate, the 10,000 selected to test is equivalent to about 20 random tests. So you have a 99.8% of reduction of the sample size. That kind of drastic reduction is one of the few things that most people still don't understand. I surprise myself after the (inaudible) that's a couple of years ago.

So that's the kind of thing that we need to understand because when you see that kind of low data quality you realize that you will do things very differently. You at least will not project that kind of strong confidence as if I have 10,000 cases, therefore the rate is pretty sure. You would know you could be far away from the truth. In fact we have reports now, people find these estimated rates off by 50 times, 80 times, so I'm not surprised by these reports because I know the data quality was extremely low to start with.

KENNEALLY: You talk about this as a moment that's a stress test for just about everything, and it is including scholarly publications. You and your editorial board have to manage your responsibility to, on the one hand, maintain scientific rigor, and on the other hand, work under these extreme time constraints so that we can begin to see the research as early as possible, and so that perhaps may have impact on others working on these issues around the world. So what did you learn in that process about the state of scholarly publishing today?

MENG: That's an extremely timely question. I learned a tremendous amount, this I have to say. I have been involved in editorial publishing, both for myself and to help others in the last 30 years, this is by far the most stressful time. Let me give you just a little bit of how we do what we do, and what kind of time scale we were working on.

When I receive an article normally for *Harvard Data Science Review*, we send out the review and usually we ask that the review come back in two to three months, depends on how technical the article is. In this case, I'm requesting one week

turnaround time. Interestingly, because we were worried about not enough busy data scientists would be able to submit their report within one week, so we ask more than we need. We double ask, we ask six people. What happened is I think COVID-19 has motivated people so well, we typically get all six back, and there are extensive comments, very rigorous, very substantial. And then they also were incredibly impressed but also very overwhelmed because now they need to address six sets of responses within a very short time period because it's in their interest, in the journal interest to publish those things as fast as possible, at the same time control the quality. They all work so hard, and I can never see how hard they worked before, that they return to me within one to two weeks, a with the response that surprised me and surprised a lot of the reviewers because they didn't expect that they would be able to answer all the questions.

I give you two examples. One paper comes back with a 33 page response to the reviewers. The other comes back with a 22 page response to the reviewers. These responses themselves are longer than the article. It shows how they addressed the reviewers comments, what things they disagree. So I thought these were fantastic reports so I had made the decision and with the reviewer and the author's support, we are going to actually publish these responses themselves, post online, of course anonymously so the author do not know who are the reviewers. I think that's a way to showcase to – also the general public how rigorously we are doing science and also have a peek of how the sausage is made. And so that's one way we're trying to keep the quality high with this very timely dissemination by having the teams and the people working extremely hard. I am extremely impressed and pleased with this process.

Let me also mention that there are other issues about scholarly publishing that I realize that we need to resolve. For example, for us, we currently from a paper get a final version to my office and to publish. That will take about a couple of week's time because we need to do the technical copyediting, copyediting, proofreading, and post online, double proofreading. Now I've been asking my team, is there some way that we can speed up the process? Because we are working in a weekly basis now. The authors are working so hard to save a day or two. If we actually have to take two weeks, just on the production side just doesn't seem the right thing to do. But on the other hand, we're very limited by our resources.

So the question here is just in publication, how do we shorten during this time these production time to make things disseminate as fast as possible. Of course we post what they call the (inaudible) version to other things to get things out. But still this pushes us to think harder about how to make our production face as efficient as possible.

KENNEALLY: You've given us a pretty balanced picture of data science in 2020, Professor Meng, on the one hand, conceding at least when it comes to COVID-19 that the data quality is not what you would like to see, but it's what you've got. And on the other hand, you've imposed a pretty rigorous regime on the review and the publication of these materials. So I have to ask you as a way to end our discussion about your own confidence level, today, in data science and in the way that it can address the challenge of COVID-19. We are hearing on a daily basis about a possible development of a vaccine, and that itself is having its own kind of stress test because vaccine development isn't something that happens overnight, but people want it to. So how do you feel, sum it up for us, what's your state of mind when it comes to the efficacy that the position of data science today, 2020?

MENG: Thank you for that great question, and I'm going to give you a very balanced answer, as intended. I think we're doing whatever we can, and everyone's working extremely hard. That's where I'm getting my confidence at this moment. So when I'm saying that my confidence at this moment is built upon a collective will power, if you may. But I have to say, that's not an entirely comfortable position to be in because we need to have more reliable systems in place. For example when I talk about the global data common, global data protocol. Too, also a kind of independent global data science community that we're aware of all kinds of political and societal implication landscape that nevertheless can maintain their scientific independence, especially under time pressure when everybody wants to have answer, it's very easy to just say, OK, here's what we do, and take it. But I think we need also a system in place. We also need to have a lot more effective communications to the policymakers as we talk about, but not only locally but on a global scale, which is a much harder one.

So my conclusion is that will power is essential, that's what we rely on now, but if we rely too much on it, it will not be sustainable. Sooner or later we would all get burned out. So I think what we're trying to do now is now understand how do we do data science in this extreme time scale, which is such a massive test, it reveals a lot of issues, a lot of problems, but I think it motivates us to think much more deeply and also broadly about setting up a lot of data commons, all these systems to make things as efficient as possible, at the same time relying on our will power to move through this incredibly difficult time and contribute whatever we should and what we can.

KENNEALLY: Harvard Statistics Professor Xiao-Li Meng, and founding editor in chief of the *Harvard Data Science Review*. Thank you for joining me today on *Beyond the Book*.

MENG: Thank you so much for having me again, and thank you for giving me a chance to talk about all these issues, as well as, I have to say, promote the *Harvard Data Science Review*, it's completely free online, so please check out and read through, especially the special issue. Thank you.

KENNEALLY: *Beyond the Book* is produced by Copyright Clearance Center. Our co-producer and recording engineer is Jeremy Brieske of Burst Marketing. Subscribe to the program wherever you go for podcasts, and follow us on Twitter and Facebook. The complete *Beyond the Book* podcast archive is available at Beyondthebook.com. I'm Christopher Kenneally. Thanks for listening and join us again soon on CCC's *Beyond the Book*.

END OF FILE