Formerly Beyond the Book

VELOCITY OF
content

A Copyright Clearance Center Podcast

The Devil In the Data
2020 Year-in-review

Special guests

- **Xiao-Li Meng, editor-in-chief, *Harvard Data Science Review***
- **David Fajgenbaum, author of *Chasing My Cure: A Doctor's Race to Turn Hope into Action***
- **Tatiana Khayrullina, Outsell, Inc.**

KENNEALLY: Welcome to Copyright Clearance Center's podcast series. I'm Christopher Kenneally.

In the final weeks of the year, this program is looking back at the past twelve months.

2020 has earned notoriety for a series of natural catastrophes and human tragedies – most profoundly, for the global pandemic of the COVID-19 virus.

Scientists and researchers, however, are heeding this advice: Don't waste a crisis.

The pandemic and its attendant challenges present opportunities to explore important questions in medicine and other sciences. In publishing, economic and professional disruption is opening new paths for information access.

In this edition of a three-part review for 2020, we learn why "data analyst" may be the year's hot job title.

Formerly Beyond the Book

VELOCITY OF content

A Copyright Clearance Center Podcast

In the COVID-19 era, 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.

Harvard statistics professor, **Xiao-Li Meng**, the founding editor in chief of the *Harvard Data Science Review*, says data scientists have the expertise and a professional obligation to play vital roles at this critical moment. The coronavirus crisis, he asserts, amounts to a massive stress test for data science at a critical time in the field's development.

KENNEALLY: You make the point 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: 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: 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.

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: Cytokine storms may sound like an unusual meteorological phenomenon, but these storms are medical ones. Normally, cytokines regulate the human body's immune system. When attacked by an infection, though, cytokines can be released in excessive amounts, leading to organ failure. Such storms are one of the most devastating effects of COVID-19.

An unusual and innovative laboratory at the Penn Orphan Disease Center in Philadelphia has turned its attention to the novel coronavirus that causes COVID-19. Leading that effort is a groundbreaking physician, scientist, disease-hunter, and bestselling author, David Fajgenbaum.

One of the youngest individuals ever appointed to the faculty at Penn Medicine, David Fajgenbaum is author of *Chasing My Cure: A Doctor's Race to Turn Hope into Action*.

While in medical school, Dr. Fajgenbaum spent months hospitalized in critical condition from idiopathic multicentric Castleman disease, an extremely rare disorder of the lymph nodes. The physician eventually sought a cure himself, spearheading a fresh approach to research and discovering a treatment that has put him into extended remission.

KENNEALLY: Your laboratory – it sounds like it's a rather different kind of laboratory than the one we might imagine if we close our eyes and think of test tubes and beakers and Bunsen burners. What your material is, as far as you're investigating, is data.

FAJGENBAUM: That's right. We have what we call a translational laboratory. Half of our lab does, just as you said, kind of the traditional beakers and experiments in dishes, we work with animal models. And then the other half of the lab does computational research, where we actually – the experiments we do are on computers.

Given the situation we're in right now with this virus and concerns about not practicing social distancing, those sort of concerns, we've actually directed the computational side of our lab towards COVID-19. So it's the folks who don't run laboratory experiments who are really focused on COVID-19 in my lab. Thankfully, there are a number of other labs out there that are running the experiments on samples of infected cells of model systems that are infected with COVID-19.

We have a team of folks that are going through all of the published data – all of the published case reports, case series, clinical trials of any drug that's ever been tried against COVID-19 to first just categorize what's being given, and then most importantly, understand what's working and what's not working. Because we knew that it would take literally years for just a few people to go through the 2,500 published studies, we enlisted an army of volunteers. So it's actually a total of 30 people went through 2,500 papers in 12 days and extracted out every single data point on every drug that's ever been used against COVID-19, and we're in the middle of running the analyses on those data right now.

KENNEALLY: So you're really crowdsourcing. This is an approach that takes advantage of the power of networking – not just the technology side, but the human side.

FAJGENBAUM: You're exactly right. You know, can you think of another cause, other than COVID-19, that would get people rallied together to want to spend their nights and weekends and hours fighting this? It's amazing the human spirit and so many people coming together and saying, I'm going to spend any moment of time that I have to help out here.

KENNEALLY: This approach you're undertaking, Dr. Fajgenbaum, it sounds like you know what you're doing, because indeed you do know what you're doing. (laughter) Your own experience, a remarkable experience which you tell the tale of in *Chasing My Cure*, was to suffer this rather dramatic disorder and then decide that the person to cure it was yourself.

FAJGENBAUM: That's right. I was diagnosed with this awful disease, Castleman disease, while I was a third-year medical student, and I was so sick that I had my last rites read to me, because the doctors didn't think I would survive. Fortunately, I survived that first episode, but I went on to have four more deadly months-long relapses where no one thought I would survive. Thankfully, chemotherapy saved my life. But it made me realize that if I wanted to live and if I wanted to maybe have a family one day, that I would need to identify a drug that could save my life.

So just as you said, the steps we're taking right now against COVID-19 are right out of the same exact playbook that we used to figure out how to save my life. My book's called *Chasing My Cure*, but really we should have probably titled it *Chasing Our Cures*, because thankfully the drug that I identified that's saving my life, we're also giving it to other patients, and it's saving other patients' lives as well. This sort of process of dissecting a disease, searching for drugs that could be repurposed, and then tracking how they work, running clinical trials – I'm only alive today and being able to talk to you because of that playbook. And we hope that utilizing the exact same approach for COVID-19 is going to be really powerful.

There was a really important moment in my own battle against Castleman disease. That was after my fourth relapse, I learned there were no more drugs in development and that if I didn't get involved in research, that no one was going to identify a drug that could save my life. There was this moment where I went from hoping that things would work out and hoping that someone would figure out a drug to saying, I want to turn my hope into action. I'm actually going to take action in fighting back against this disease by conducting research and doing the work myself.

I was devastated – and still am devastated – about what's happening with COVID-19. And I thought to myself, I really hope some lab out there goes through all of the FDA-approved drugs and understands what drugs may work, what drugs may not work. I really hope a lab out there puts together an immune map to track how all of these data fit together so that other scientists can build upon it. And then I thought to myself, you know, I've realized in my own fight that if I'm going to hope for it, if I'm going to pray for it, then I need to do something about it. So I turned my hope that someone would do it into action, and then I was able to rally my team and a number of volunteers from the Castleman Disease Collaborative Network to get involved in fighting against this disease.

KENNEALLY: Well, you mentioned playbooks, and playbooks are something you also know well in addition to research. You were a quarterback at Georgetown. And I wonder, David Fajgenbaum, whether that's what it feels like right now. You're sort of getting back into the pocket, getting ready to throw that pass to the researchers who can take it further downfield.

FAJGENBAUM: I think that's a really great analogy. And I think science is such a team sport. There are very few team sports that require such coordination and collaboration as football does. You need everyone to play a particular role and to be a part of a bigger effort.

Research is a similar sort of thing, where not one of us – no single researcher or physician – can do this alone. It absolutely requires a team. And I think that the approach of taking a collaborative approach and also making data open source is so critical. Because anyone who thinks they can do it on their own is wrong. It absolutely requires all of us working together. And it's not just scientists working together.

As you said earlier, we've got people that are non-scientists – we call them citizen scientists – who are part of this crowdsourcing effort to pull the data together. And we have people who are practicing social distancing to make life a little bit easier and others who are helping to get PPE. So it doesn't matter what your background is. I think it's really important, to use that same analogy, to say we can all be a part of this game.

KENNEALLY: Where are you right now – at the kitchen table, seated by a desk in the spare room? Maybe you're hiding from the family in the garage, or else you found yourself a quiet corner in a damp cellar. It doesn't matter, really. You and I and many more around the world are at work – virtually, digitally, online.

With the world in lockdown during the COVID-19 pandemic, the offices and classrooms where we once would gather and collaborate are now off limits.

COVID-19 has impacted every aspect of our lives, while rewriting the rules for business, academia, technology, publishing, and the media. In May, Outsell, Inc., analyst Tatiana Khayrullina told a CCC Town Hall that even in the early stages of the crisis, publishers should be preparing for the touchdown after the lockdown.

KHAYRULLINA: There's a lot of emphasis on artificial intelligence being put to good use. However, it has also become clear that where in the past, we would think that, OK, take a dataset and attach it to an algorithm and it should be good enough, what publishers are essentially teaching the market right now is that using just one dataset, just one source of data, just one source of content, will probably not provide the insights that the researchers are looking for, that the market is looking for. And the market is learning this and learning this quickly.

So in the future, it will be really hard to compete, even with a nice and fast and modern artificial intelligence tool which is only using one set of data, one set of content, so we'll see a lot more collaboration between information providers.

I would like to also mention another set of insights that is also coming from this current experience. We have seen many publishers report an increased traffic through their websites because their content is now free and available. I would suggest that publishers can learn a lot from the essentially mountain of data they are now sitting on that they didn't have access to before. That data can provide interesting insights if you ask the right questions. And the type of questions that can be asked is, well, how big is my entire market – my addressable market? Now that my product is free, who's coming to look at it? Who's coming to use it? Are there segments in this audience that I never suspected would be interested? Should I rethink my market segmentation? Am I in the right geography? Do I see a lot of traffic coming from different regions? So on, so forth.

It's the user data that publishers are getting in exchange for making their products free that is really useful to mine for insights and to fine-tune the strategy in the months to come.

KENNEALLY: In 2020, we have responded to COVID-19 with a set of conditions and expectations that constitute the essentials of "the new normal" – social distancing; online learning; mask wearing and hand washing. All the while, we hear about vaccine trials and worry over the latest coronavirus case numbers.

A public conversation grounded in scientific data has grown commonplace. The tales the data tell may be dark or bright. But it is not up to the data to choose – it is up to us.

This podcast series is brought to you each week 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.

I’m Christopher Kenneally. Thanks for listening. Best wishes for the coming year.