A Stroll Down Sesame Street: A Conversation with Dr. Peter Licari

By: Nicholas Lepore



[1]

D*r. Peter Licari wears many hats on this spinning rock we both occupy. He's a behavioral data scientist, an avid D&D player, a catcher of Pokémon, a debater of the minutia, a contemplator, and a dad. He is as talented at video games as he is at servicing the data intelligence needs of the major data firm, Morning Consult LLC. He often describes himself as a student of the social world, with a passion for learning and advancing the field of data science in both a personal and professional capacity while also keeping himself grounded by spending his devoted time with his wife and the brain behind the PhD, his 2-year-old daughter Rose.*

Nick: I must admit, after Rose denied my request for an interview I was absolutely crushed. Although, I am happy that she connected us and gave me the second-best option.

Peter: You know what, I get it. I understand why she's the first choice. She is the brains behind the operation. She sat on my lap as I did my dissertation. She's a fluent R programmer and I'm just a poser. She codes in base R, data table, dplyr and all that jazz. She's fluent across the dialects and I'm just happy to be here.

Nevertheless, let's have some fun and start with Cookie Monster. Hobbes, Locke, Montesquieu, Rousseau, and even Sesame Street have debated the question of ethics and what separates right from wrong for decades. Research firms have unprecedented, and arguably unfiltered access to individuals. What responsibility, if any, do organizations have to protect the data they collect and use? And what does 'protecting' data look like in practice?

So, for one, my daughter is a huge fan of Sesame Street, and I am as well, and this question holds up very well. But nonetheless, practitioners must remember, and it's difficult, that there's a psychological distance when looking at rows in a spreadsheet and you don't necessarily imagine that these rows are measures of people's lives. Now you don't need many of these measures to be able to triangulate and identify these individuals. For example, if you have the right five pieces of data, you can essentially single out any one person in United States.

So as survey practitioners we have responsibilities to protect this. For example, we need to make sure the data is, for one, anonymized as best we can. That if we say responses are going to be kept confidential, that we keep them confidential.

You discussed anonymity as a means of protection for respondents. Can you talk further what it means to anonymize data?

Absolutely. Zooming out, many practitioners, like folks in the Census, will inject random noise into variables and different measures. They do this so that when you look at something from an aggregate level, you'll have the correct information. However, if you zone in and try to look at any individual person, you wouldn't necessarily be able to say who is Jane Doe or Joe Shmoe or whomever because it's been manipulated through the inclusion of random noise.

---

[1] The New Yorker: Daily Cartoon: Thursday, October 14, Digital, "Cookie Monster" :By Adam Katzenstein— *https://www.newyorker.com/cartoons/daily-cartoon/thursday-october-14th-cookie-supply-chain*

Now I want to flag that this practice is not without controversy. Some practitioners have found that and argue that there can be distortions. It's very difficult to be able to maintain averages across all levels of potential scope. Maintaining averages while injecting random noise is quite a challenge.

Can you say a bit more about injecting random noise? What does that look like in your day to day?



[2]

Of course! The intuition here is that randomness is something that by the fact that it is random it is inherently uncorrelated with everything. Not perfectly of course because there's no such thing as a perfect zero correlation. But the idea of injecting random error is to be able to provide little anonymity into people. For example, imagine you have a one to hundred scale, and you want to inject a certain amount of error to it. You might say that your error injecting function is going to be something that has a mean of .0 error with a standard deviation of 1 or 10 or whatever. What that'll do is it will pick a random number and be like, "Oh okay, I'm going to adjust these numbers randomly." And now you no longer have this "perfect" record of what people said.

Really insightful. All this talk about anonymity and protecting individual data has me thinking about whether people really know what they're handing over when they answer a survey. I'm curious on your take on whether individuals, under this idea of "voluntarily," give researchers permission to access this information about them and/ or build these advance machine learning models about people like them.

It's a really interesting question and frankly a relatively new question we as researchers are beginning to grapple with. From a legal perspective people are "covered" because they "agree" *[Peter at this point is throwing covered and agreed into air quotes as talks]* to the 500 or so pages of terms and services and then like magic we become inoculated from legal repercussions. But what's interesting though is that people, to a certain extent, don't necessarily know all that they've agreed to. And as we've discussed at length, this is largely a symptom of today's tech. I don't think that's necessarily something you're going to see change coming out big tech and it's likely going to be a policy intervention to deviate to what I feel is the status quo among these firms.

In terms of survey research, we don't force people or lock them into a room and say, "you must answer this survey or else." Nevertheless, I'm not sure that not forcing people to take a survey constitutes full consent if they don't know what these measures are going to do.

For example, I'm thinking about a lot of machine learning algorithms where people might be answering funny questions here and there about X or Y topic. But if people were fully informed that that data that is being collected is being used to create an algorithm that is difficult to audit and has the potential to create biased outcomes then I'm not confident people would be so forthcoming.

I mean that's a huge problem right now with machine learning models right now. They're dealing with systemic bias in terms of the training data, the output, and even the results. There's bias at multiple levels and it can truly affect historically marginalized groups in scary ways. If people knew this, like truly knew this, I'm not sure if everyone would be equally as comfortable.

---

[2] Condé Nast: Cartoon: September 2, 2019, *Digital*: "Ugh Construction" By Jason Adam Katzenstein—*https://condenaststore.com/featured/ugh-construction-jason-adam-katzenstein.html*



©The New Yorker Collection 1993 Peter Steiner
From cartoonbank.com. All rights reserved.

"With machine learning, nobody knows you're a dog."

[3]

Diving a little deeper, or I guess zooming out from machine learning a little. What tools have you learned to prioritize or remember that behind the numbers, the percent, and/ or the models that there is a human being behind them and not get lost in the numbers?

It's not a tool in the sense that it's a software that you can download.

You mean there's no R package for ethics?

[Laughter…]

There's no R package that says, "OH! By the way these are people." Honestly, maybe there should be though. You load it in and it just randomly interjects that into your environment:

BY THE WAY THESE ARE PEOPLE!!!

That would be really fun to do though, and we'll chat more about that. In all seriousness though, there's a mental toolkit. I think that you have to stay empathetic. Personally, and other people will have different answers to this question, but I come at this question from a spiritual angle. I post on by Twitter bio that I'm a "Bad Quaker," although, I do try to do better.

One of the testimonies of Quakerism, and really, it's a quality, to lead with a certain level of respect. And that's really important to me. As a data scientist, part of living that testimony out is always remember that there are people behind the data. It's an essential cognitive frame to have and you must be deliberate. It's easy to go and run means, and averages, and standard deviations, and build models with the type of a keypad. It's even easier given the speed in which you can do it now, which makes it harder to be deliberate. Numbers create a psychological distance. But you must deliberately remember that these are people. It is a choice to remember, and as professionals and stewards of the field, we must pause to remember. We must stop for a moment, always.

Thank you for sharing that, truly. I have one final question for you and it's about the field. As you know, anybody can have access to data, access to R, publish on Medium, etc. Unlike other fields like law or medicine, there's a review board that exists that can disbar someone or take away someone's medical license for malpractice. Do you feel that that is the future of our field? Do you feel there should be some governing body to oversee the field of data science and that the field should be more regulated than it currently is?

Oh boy…

[Shared Laughter...]

That is a very tough final question. Do I feel that the field should be more regulated than it currently is? Yes... but the devil is in the details. It is so tricky because at the end of the day, what are we doing… we're analyzing things. Anybody can use R and anybody in theory should be able to. Everybody should have the tools to investigate the world and to generate knowledge and to learn.

To try to stop that is an incredibly treacherous way to go. However, I do think there needs to be culpability and responsibility. Although, I do worry that as soon as start instituting stringent requirements that we only further the issues in our field. This is a predominantly male, old, cis, white field. When we institute additional barriers, we are making it more difficult to access the mode to

[3] The New Yorker: Daily Cartoon: Adapted from July 5, 1993 Print, "On the Internet, nobody knows you're a dog." :By Peter Steiner

investigate the world. I do believe that people join the field with generally good intentions.

But the road to Hell is paved with good intentions.

We need to make sure that we're not closing the door behind us to groups who have historically been precluded from this space. The field needs diverse perspectives. If there is a capital T Truth, then it is unknowable by individual people. We need additional insights from people to approximate truth the best we can. To close the door behind us, it just doesn't sit right with me.

I will say, however, there could be models that work. The one I wrestle with is certification as a platform not as a requirement. This allows someone to still go out and write a Medium post on a topic they're interested in and people can still go out and dissect it the way they normally would. However, if someone did the same thing with an imprint or stamp of a governing body or society that establishes their credentials then people might recognize that this other person is someone who has been trained, gone through a rigorous process similar to law or medical school and has the weight of certification.

I do, however, feel that law and medicine should be distinguished from us. Those fields are not so closely related to the ability to learn about the world the way statistics and data science are.

To summarize, the certification would need to be about a platform and not a requirement. And there should be a governing body that professional punishes individuals who publish, lie, or act unethically in their platform. Ultimately, however, the aim needs to parse out the details and ensure we elevate voices and open doors into the field rather than gatekeep it.



[4]

Peter, as mentioned, wears many hats. Including, which was not mentioned before, the hat of a "Bad Quaker." However, among all the hats he might wear during his day, I can confidently say that he wears the hat of a pioneer. Many people will come and go in the field of data science but not many will leave it better than when they found it… Peter will and already has. He has a gentle soul that takes caution when it comes to someone's heart, and he leads his life with patience and empathy. This field is complex with infinite moving parts. Data science needs more people like Peter, who sweat the details, examine the field from an empathetic and critical lens, while always remaining skeptical and remembering… that the Devil is in the details.

[4] "The Devil Is In the Details" By: by Mike Jory on 11/24/2015