***Practical Significance* | Episode 36: Let's Get Real World With Brian Tarran.**

Donna LaLonde: Well, welcome, everyone. We are finding it difficult to believe that it is 2024 and we are about to begin another season of *Practical Significance*. But we couldn't be more excited about beginning the season with our guest, Brian Tarran, to talk about *Real World Data Science*, an initiative of the Royal Statistical Society, or RSS, which the ASA has recently joined as a partner. We'll get to *Real World Data Science*, but we want to start by learning a little bit more about you, Brian. So, if you could share a little bit about your background and your career. And, personally, I want to know what this Umbre de Metal and video gamer stuff that I read about on X?

Brian Tarran: I'll explain that one. There is a story behind that. But the basic version is I really love heavy metal music when I was at college, so at about eighteen, I played in a terrible, terrible, noisy band. My friends and I are always talking about metal music, and I'm sure this is the title of a song by a heavy metal band from Brazil or something like that. I believe it might be Sepultura.

I can't remember. I was Googling it but couldn't find anything. But anyway, that's kind of the origin of that. The video gamer thing is pretty self-explanatory. I like playing video games and I make it my goal to the playing of video games as a 40-year-old man – acceptable because my dad tells me that I should have stopped years ago, but I still love them. But this is not what we're here to talk about. I could do a whole podcast about video games. So, my background is, that I'm a journalist. So, ASA members will hopefully know me from my work on *Significance* magazine. That was between 2014 and 2021.

Prior to that, I worked in the market research industry. I was editing a trade publication, if you like, for market researchers. So that's how I became interested in statistics. We got to meet some statisticians as part of our reporting on market research, and that's what led me to *Significance* magazine, the RSS, and the ASA. I'm not a statistician or a data scientist, but I like to think that ten years of working with statisticians and data scientists has taught me a little bit about how data scientists and statisticians approach problems and think about problems, and I hope it's made me a little smarter from reading your work.

Donna LaLonde: I believe it has taught you a lot, and we would definitely welcome you into the fold. But to get down to the business of the podcast, tell us about *Real World Data Science*. What was the motivation for getting it started, and where do you see it going in the future?

Brian Tarran: Yes, of course. Well, I mean, a bit of background to *Real World Data Science* - it's an online publication, a new digital publication that we have developed. The first version of it went online in October 2022, and the site as it is now, was fully launched in March 2023. It's a publication that aims to speak to all data scientists, whether they are just starting out as students, whether they've been in the

industry for a couple of years, leading teams, or teaching the next generation, whatever it might be, we really wanted to focus on, as should be evident from the name, real-world applications of data science.

What motivated it from the RSS side, was a data science task force that the president of the RSS at the time, Sylvia Richardson, had set up. That brought together people from academia and industry and the RSS fellowship to think about what role should the RSS play in the data science conversation. What can the RSS as an organization, contribute to and support our members who work as data scientists?

And also, how can we kind of advocate for the important place and the important role that statistics plays in the data science toolkit? The group of people that I worked with from that task force were predominantly statisticians who worked in industry, in data science roles in data science teams, and in organizations. And as I say, they were keen to stress that a lot of what is reported on about data science now, a lot of what you see is kind of driven by marketing hype, a company blog post, complicated sounding solutions to issues.

And the goal of that is essentially to sell companies things. It's like the whole big data hype. It's about what we can sell you to make you feel like you're doing data science. And our data science professionals, they were saying, “We think we really need to shed some light on what data science actually looks like end to end.” What is the process, how it begins with a conversation with the internal client or the external client, whatever it might be about. What is the question that you want answered? What's the problem? You want to solve, and then the steps that you walk through that process till you come at an output, an outcome, some insight, advice, whatever, that you can deliver to the client, to demystify what data science actually is. I believe there was a sense that anyone could do data science, right? It's just about pulling packages off of GitHub and chucking some data into it.

But statisticians will tell you, knowing how to deal with data, knowing the right way to approach a particular problem, knowing that there isn't always, in fact, very rarely a right way, it's just about an informed, sensible way of doing things that's crucial. So, trying to introduce some of those concepts through our case studies, through our discussions, through our explorations of key data science ideas and methodologies and techniques, you name it. So that's what we're building up to.

Where do we go next? Key for us is to continue to develop the platform, and we're really delighted to have the ASA on board as a partner. I had great fun working with ASA colleagues on *Significance* for the best part of ten years, and I believe making sure that we reach a broader audience, making sure that we are speaking to the international data science community is really key for us, and also having a voice in the artificial intelligence conversation.

It's something that I know in the UK, as in the US, there's a lot of activity at the moment: data statistics. Data science plays an obvious role in artificial intelligence because the artificial intelligence we're talking about primarily are models built on, driven by data, and understanding that data, understanding how those models are built. Statisticians and data scientists have a perspective on that and how best to use those models. So, we really want to make sure that we're part of that conversation.

Ron Wasserstein: So, Brian, let's drill down just a little bit more. Talk a little bit about the experience of building the website and what you took away from doing that.

Brian Tarran: It was a learning experience for me. I gave a talk at a conference in September in Chicago, telling a bit about my journey, and there is a video online and we're going to hopefully share it on *Real*

World Data Science. Now, for me, as a journalist, I had no experience building websites. I was involved in website projects in the sense of saying, I believe it should look a bit like this and the headlines should go there and be in that point size or whatever, but actually building something, I'd never had the opportunity to do that before. And when I was doing the research around what *Real World Data Science* could be, there was a kind of sense really, that I wanted to make sure that anything that we built was being built for the audience of potential contributors. Now, with eight years on that magazine, I spent a lot of time annoying contributors because they deliver these lovely, well-set-out articles in LaTex format. And I'd say, "No, I'm sorry, we're just a Microsoft Word and Adobe InDesign type of organization." So, I thought, well, there are tools that data scientists are using to produce really beautiful outputs, reports, books, websites, whatever. What are these tools? How do I use these tools? How can we build something from that?

Literally, I believe two people have mentioned Quarto. I hadn't ever heard of it before, but the documentation is so easy, helps you to upskill or feel like you're upskilling quickly, and they give you just enough to get you started and get you excited about the potential, but not enough that you could do everything straight away. Right? You have to start digging into stack exchange, stack overflow threads, putting callouts for help on GitHub pages and repos, all that sort of stuff. And the whole community, that whole open-source community, is just so engaged and so willing to help and share and support. That just kind of inspired me to keep going.

And now we've got to the point where I believe we've built a platform where data scientists who use these tools should feel empowered to be able to write and contribute and understand how we're putting this together. It's not built using publishing tools for publishing people like me. It's built using data science tools for data scientists. And we hope that that will provide an additional kind of layer of engagement for people.

Ron Wasserstein: That is super cool. Thanks, Brian. So how can our listeners get involved?

Brian Tarran: There are lots of ways. So, at the moment, people tend to just drop us an email and say, "I've got an article idea," and we'll discuss it back and forth over email.

Where we want to get to, though, is having people use the full functionality of the site, of Quarto, and of the GitHub infrastructure that we use. So, the site is run off of GitHub pages. People can technically take a copy of our repository. So, they'll clone the repository. They can create an article within the correct article directory, design it, put the features that they want in it, and then essentially send that to us via what they call on GitHub, a pull request.

Say basically, "Hey, I've built this thing for your site. Take a look, what do you think?" And then we can have the kind of review process back and forth via GitHub. It's all transparent out in the open. People can see the conversations; others can contribute if they feel they've got something to add. That's where we'd like to get to.

So if people are listening and they think, I've worked on data science projects, I really want to help people to understand what data science looks like in a real-world context, trying to solve real-world problems, then please do go to the GitHub repository, clone it, run the site on your machine locally, build what you want, build what you think is cool, what people would want to read and share, and then let us

know about it and do it that way. I believe that would be taking full advantage of what we've built and the potential.

Donna LaLonde: So, Brian, I wanted to follow up on the contributions because one of the pieces that I believe is interesting on this site is the case studies. I wondered if you wanted to say a little bit more and if you are interested in the submission of more case studies. Are there particular areas that you'd love to see case studies in?

Brian Tarran: Yes. I wouldn't want to limit people to say, we only want case studies on this sector of this industry or whatever it is. We're keen to see the full range of what's possible out there with data science. Building out the case study resource is what's going to be the sort of flagship element of the site if you like because there are a lot of hypothetical examples out in the data science publications of people giving examples, you can do this with this, and here's the package I've built so you can replicate the work yourself. What we want to do is kind of make sure that people understand what data science looks like in context. That was what came out of our research. That's what people were telling us.

If you're going to build something, this is what we want to see. We want to understand how data science is actually used within organizations, what the process is that people go through, and how the work that's done feeds into organizational decisions, product development, whatever it might be. So whatever sector people are working in, if they think there's something they could share and talk about, then I encourage them to get in touch.

The big area of data science that we don't know enough about really is what goes on in industry. I imagine it's going to be difficult for that. You're working for a big, fast-moving consumer goods company to be able to talk completely openly and honestly about the work that you do because of all the confidentiality that there will be around that work.

So, there will be a certain element of kind of hypothetical examples, but in that sort of context, we're kind of already starved of information about what data science actually looks like in those contexts. Just knowing that there's a kind of hypothetical example that's grounded in a reality I believe will be helpful, especially to students, in terms of helping them to understand how I apply what I'm learning in my course when I'm out in the real world. It's really valuable, I believe, for educators as well to understand what is expected of data scientists in the workplace, because that helps them to craft their courses and to make sure that when people are leaving education, they're leaving with the right skill sets and the right approaches. That's the value of case studies. But we don't limit it to that.

We're open to people writing articles about big ideas in data science, and viewpoint pieces on the whole area of AI and data ethics. We even had a nice tutorial up over the holiday season on how you can use R and ggplot two to build Christmas cards, which was really fun and showed the lighter side of *Real World Data Science*. But I believe as Nicola Renny, the author of that, explained some of these skills, once you learn how to build Christmas cards, it opens up your eyes to, I can use what I've learned there to do something with my outputs in generative art. Data art can also be really practically helpful when you're back at your desk.

Ron Wasserstein: So, Brian, as I was listening to you talk about all that, I started thinking how many words you were using that you couldn't have imagined that you would be saying when you started your journalism career.

Brian Tarran: Yes, well, quite a lot. Because don't tell anybody this, but I kind of got into journalism so I wouldn't have to deal with numbers. I always liked math. I studied it at college. They're not American colleges. This is in the UK. It's the school you go to after you finish your secondary education. But I lacked confidence or I lacked teachers who could instill that kind of confidence or inspire me to pursue statistics, data, and math as a career. And one of the things I keep reflecting on now, like having had this opportunity to develop *Real World Data Science*, to start using some of these tools like Quarto, messing around with bits of R and Python coding to try out little things. I just feel like I wish I could have my time again. I could go back and start down that path.

I know it's not too late. I'm only forty-one. I could decide to retrain, but at the same time, I love what I do. I just wish I could take the skills I've got for building websites and with things like Quarto and maybe just push it a little bit further, do a bit more of the analysis. Just have a bit more confidence to know what I'm doing.

Donna LaLonde: Brian, I suspect that's in your future.

Brian Tarran: Well, yes, it might be. So, my wife is just finishing up a degree course, a part-time degree course. So, when she gets her evenings back, maybe it's my turn to lock the door of the office and just stay up until the early hours, catching up on reading and study and all that sort of stuff.

Donna LaLonde: I wanted to bring you back, though, to your journalism beginnings and ask you a question about public journalism. What are you most excited about and what are you most worried about? And especially in this world of AI and large language models and fake videos.

Brian Tarran: And all that, if someone like me, who's a bit further on in their career, seeing these tools and what's possible and being excited by the potential, I can only imagine what it must be like being a kind of journalism student at the age of 16 or 17, and seeing the kind of work that is published in the *New York Times*, in the *Financial Times*, in the *Economists*. These are amazing visual data outputs.

Some of the analysis that's done as well, the original analysis that these organizations are doing, and just the way they kind of blend journalism, data, art and weave it all together into these interactive, compelling stories, which is what's exciting. It's exciting to think that the journalist's toolkit is now more than just a notepad, a pen, and a Dictaphone. I would be surprised if most journalists now are not being exposed in some way early on to what you can do with open data and code. And that has a lot of potential, a lot of opportunity for good, informed work to come out of that.

The worry though, is, as you say, AI and generative AI. There was a report that came out in the UK, it was this week, looking at the potential impact of AI technologies on jobs and comms. Communication was an area that was flagged. So as a young person now, would I want to get into journalism if I thought that in a few years, a Chat GPT-like model was just going to be doing the work that I would have done? I hope people aren't dissuaded, because originality of thought is something that these machines still can't do, can't come up with ideas that haven't already appeared in their training data. I would hope that people do still pursue that path.

But I believe there is a danger. We're both in the UK and in the US coming up to a big election year. I was at an event last week where they were talking about the danger of generative AI in terms of it polluting the discussion, the political discussion, and the debate. We've always had problems with misinformation and disinformation. The way that the speakers were describing it was that generative AI has supercharged the ability to create and disseminate misinformation and disinformation. And that's a worry because there are still no good ways of determining whether something is genuine or not.

We had that problem as well before Chat GPT, right? People would still read reports on questionable sites and take them as truth. So, it's not that AI has created a new problem, it's just made that problem more severe and made it more convincing. It can be hard, and I believe it's going to be a worry to see what happens, and we'll struggle probably, to know whether it impacts the actual outcomes of the election. But just the way it might affect the tone of the debate and the discussion, I believe, is a concern.

Donna LaLonde: Thanks for that. And it strikes me that if listeners are interested in exploring that more, that might make a nice contribution for *Real World Data Science*.

Brian Tarran: Absolutely. Yes. So, I put a short blog post up on the website. When I came back from that event, it was Tech UK's Digital Ethics Summit. But we'd love to explore that theme in more detail. Some of the potential solutions. Some of these speakers were talking about content watermarking as a way of solving this issue. We've looked before at whether you can kind of use statistical analysis to take the differences between human-written text and AI-written text. The field might have evolved since that post was published. So, if people have gotten more up-to-date information that they'd like to discuss. I believe that would be really interesting. But generally, just getting that issue out there, and flagging this as a problem is something that I believe the platform could do.

If we can motivate other data scientists and other statisticians to talk about this issue, it will make people aware of the problem and maybe make them slightly more critical every time that they read something just to stop and ask that question. How do I know that I can trust that? And again, this goes back to what I said earlier about working with statisticians. For ten years I haven't learned how to do statistics, but I believe I've learned when I'm presented with information and data, the statistician's first response is always, how do I know that? Where's that data coming from? And I believe that's the kind of mindset we all need to be in right now.

Donna LaLonde: Absolutely, asking the good questions. So, I'm going to ask a final question to wrap up. This has been a fantastic conversation, but I always like to know what's on our guest TBR. Reading, listening to, watching. I don't know very much about video games, but I'll say playing.

Brian Tarran: Well, yes, I'll do all of them then. So, playing. I've been going through a game called Cyberpunk 2077, which with this conversation around AI, is a future in which AI has basically taken over and we're all part man, part machine. Really a great game, but also a really good narrative, lots of musings, philosophical musings on the nature of humanity. I'm reading the Oppenheimer biography. I saw the film over the summer and working my way through *American Prometheus* is the title. And for what I'm watching with the sad news in mid-December of the death of Andre Brower, the actor who played Captain Holt on *Brooklyn 99*. A few years ago, for Christmas, I got the entire box set of *Homicide Life on the Streets* and I haven't yet watched it. So, in tribute to Andre, I will be cracking on with that. I'm told it's excellent.

Donna LaLonde: That sounds fantastic. A great way to spend a little bit of downtime. Well, Brian, thanks so much for joining us, and we definitely look forward to seeing how the partnership on *Real World Data Science* grows. And now, by tradition, I will turn it over to my colleague Ron for Ron's top ten.

Ron Wasserstein: Thank you, Donna. We've been at the podcast for a while now and it is a good time to take stock of how we are doing. Here are the top ten signs the *Practical Significance* podcast is successful.

10 We have a steady number of sponsors. (Zero, but it's steady.)

09 We are one of the top statistics podcasts in Moldova.

08 Listenership has tripled since the podcast began. We have nine listeners now.

07 Donna never gets any complaints about the podcast content.

06 Our budget still allows for a top-ten list. Plans to reduce it to the top eight were scrapped.

05 Plans to reduce to one podcast host (you can guess which one) were also scrapped.

04 We have a rating of 5 on Spotify. Just one rater—but that person loves us!

03 Some guests have even been willing to make a second appearance.

02 Our podcast is THE favorite on my iPhone.

And the number one sign the *Practical Significance* podcast is a success:

01 Welcome to our fourth year, baby!