Practical Significance | Episode 31—Talking Shop: Introducing the ASA Caucus of Industry Representatives

Donna LaLonde: Well, welcome everyone to *Practical Significance*. It's a little bit hard to believe that we are recording the July podcast, 2023 is flying by, but we have a really special podcast for you today. We have, as our guests, two folks who have been instrumental in helping us establish our new ASA Caucus for Industry Representatives. And we're going to talk about the caucus and some other interesting topics. And so today we have Ginger and Amarjot, and we're going to start by asking you to tell us a little bit about your day job and introduce yourselves before we jump into caucus conversation. So, Ginger, I'll start with you.

Ginger Holt: Great. Thanks for having us today. I am Ginger Holt. I am the technical lead in the infrastructure data science team at Databricks. And for those of you unfamiliar with Databricks, it's a data and AI software company. And our mission is to really help organizations make all their data ready for analytics. empower data science and data-driven decisions across your organization and be able to rapidly adopt machine learning. And we really just want to level the playing field so everyone has access to the same quality of LLM and AI analytics as big tech companies do, along with all the hardware and software infrastructure. So, within infrastructure, data science specifically, we are informing a design and build of our operational infrastructure to have industry-leading efficiency, reliability, and scalability of our products.

Donna LaLonde: Thanks, Ginger, and Amarjot over to you to introduce yourself and tell us about your day job.

Amarjot Kaur: Hi, thanks for having me. My name is Amarjot Kaur and I work for Merck Research Labs. And I'm in the development statistics group where I head the respiratory and immunology therapeutic area. And as part of my job, we get to do a lot of research. We do clinical trials to help patients who are suffering from many different, and sometimes life-debilitating diseases. My role is to be involved in the design of clinical trials as well as the analysis and interpretation of them. And from the day-to-day job I have to wear various hats. During a typical day, I probably go to various meetings where sometimes we have a very technical discussion with statisticians, and on the other hand, sometimes go into meetings where we have totally strategic and planning kind of discussions. And also, as part of my role, I also have to be a cheerleader for my teams and to also focus on development, and make sure that everybody is growing in their roles and doing the best we can, as a team.

Ron Wasserstein: Thanks for those introductions. And I'm just sitting here thinking, Donna, that these people are way too busy to be talking to the likes of us. Let's go ahead and start with you, Ginger. With regards to the Caucus of Industry Reps, can you tell us a bit about the goals of that organization?

Ginger Holt: So, the ASA thinks members in academia and government are well supported. There's a lot of structure and venues for that, but we wanted something more specific for industry statisticians and data scientists. So, we created the ASA Caucus of Industry Representatives, and it's really to provide a

venue and structure to address the unique challenges for industry statisticians and data scientists. So, some of these examples: we're really here to promote the profession of statistics and data science in the private and public sectors, and then assist companies employing those members in industry. And we want to provide a venue for discussion of unique issues for industrial data scientists and statisticians to facilitate interaction between both the private sector and the public sector. Different verticals within industry as a whole, we're data scientists, so we like to collect data, identify, and collect data that are helpful to our employers and to industry in general. And then we also just want to meet as a group, right? So, we have a yearly meeting, and yearly workshops, focusing on specific challenges. LLMs are one example that may require a lot of input from this group in the future.

Ron Wasserstein: Thanks, Ginger. So Amarjot, you've had a lot of different leadership roles within the ASA. What excited you about this one, about being part of the executive committee for the Caucus Industry Reps,

Amarjot Kaur: As Ginger mentioned, just the goals of this caucus, I identify with all of them. Having worked on various ASA committees or sections or other places, this is an opportunity for ASA also to have two-way communication, right? With the Caucus, representatives, and hear about any issues for which ASA might be able to, help provide resources, and it could be part of future initiatives, right? So, so I think it's, it's a win-win situation, for this caucus. And I know that there has always been an emphasis or some kind of, at least the desire to help all sections of the membership of ASA, whether it's in industry or academia or government.

So, this caucus is much broader and brings under the umbrella all the public and private sectors and so on. So, I think this is a good opportunity to create two-way communication of ASA with representatives from a broader group of industries. So, I think that that excites me and it's a structured approach, which is sustainable and would be good.

Donna LaLonde: Ginger and Amarjot. One of the things that I've enjoyed about the meetings with the executive committee and actually the town hall meeting as well that you hosted, was hearing more about what folks view as the most significant challenges facing industry statisticians and data scientists. And, recognizing that that's a lot of different people in lots of different sectors. But we'd love to hear from you a little bit about what you think are the most significant challenges. And maybe Amjad I'll start with you and then go to Ginger.

Amarjot Kaur: I believe we are in a data revolution, right? There is so much data, we have both opportunities and challenges. And I think the one thing is for us to be prepared, right? To synthesize this large amount of information in a meaningful way. You can always do a lot of things with large data, but the question is, are we answering the right questions, right? Do we understand what we are doing? So I think that preparedness and that awareness of synthesizing all this. it's an opportunity because it's a lot of information and to make a decision – robust decision making, it's important to incorporate all that information in our analysis and how best to do it in a meaningful way. This continues to be something I think we need to pay attention to so that we are not lost, in the translation, but also just being aware of the fundamentals of the analysis and so on. So, I think that's one challenge, and I don't think we need to be very conscious of it. And also, I think the automation is great. Automation of tools that a lot of people can just press a button and get the answer, but in the process, I just sometimes worry if we lose, what exactly we are getting out of these automated tools. This is something to be aware of and make sure that we don't lose sight of that.

Donna LaLonde: Ginger, what do you think about the challenges?

Ginger Holt: Yeah, I definitely agree with Amarjot. I've been working in forecasting, basically my whole career in many different domains. And the themes I will also apply not just to forecasting, but any sub-discipline within statistics or machine learning. So, kind of what I'll talk about there can be applied there as well. But in forecasting that the fundamental challenges are really the same everywhere, I think forecasting is pretty ubiquitous. And for context, I was in academia for a few years, but otherwise, I've been mostly at very large companies like BP, HP, Walmart, Facebook, and now Databricks, but in all these places really, we needed unified forecasts across the company. So, standardized data sets, standardized methodologies, and having one source of truth for all decision-makers, whether that be in finance, sales, marketing, or whatever, make forecasts consistent across granularities.

When you compare forecasts at different levels, making sure that they add up is consistent. we want as much accuracy as possible without overfitting. We want explainability for the answers we provide, right? So, we can make actionable decisions. And then scaling, Amarjot mentioned tooling a lot. We want to scale the forecasting process or any process by building generalizable tooling there. And Donna mentioned this town hall we had a few weeks ago now, but we did collect some data there, for the participants and we found we're targeting professionals as well as students, for this caucus. And so, from the data collected during that meeting, we found that professionals really are interested in increasing collaborations across industries. So, I think having these discussions and sharing knowledge across biomedical and financial manufacturing, all these different sectors would be very useful.

And then just knowledge sharing amongst the community of industry data scientists and statisticians. So that was a major thing. And then also continuing education of technical skills as we know, just like in the past six months, like the amount of growth in LLM capabilities that we've seen, there's just such a need for people to learn and learn more about what's been happening. And then students as well. So, students are really interested in learning more about the career paths and journeys of statisticians and data scientists, like among different industries and kind of crossover, right? Being in one field and then moving over to more of a data science role from a non-technical role. So, I think people are very interested in that. The caucus is going to be focusing on these issues, right? And, looking for ways to provide additional support to our community. I'm anticipating a question about LLM, so I'll expand on significant challenges in that domain a little bit.

Ron Wasserstein: Well, here comes that question. You can't get around discussions of ChatGPT and large language models. They're everywhere. Earlier this week I was at a conference on risks and opportunities for artificial intelligence in biopharmaceutical medicine. It's being talked about everywhere. In fact, you mentioned asking the right questions, and at lunch at that conference, a colleague was telling us that the ChatGPT was able to tell him how he could have asked better questions, which I thought was both awesome and creepy at the same time, which maybe sums up how I feel about large language models right now. But in any case, for both of you and Ginger, let's just start with you. What are your thoughts about these models and what changes do you see on the horizon as a result?

Ginger Holt: Well, fortunately, I'm at Databricks, so we have a lot of customers using our platform, right? So, we actually just recently published a report on the status of data and AI in 2023. I think that's the name of the report. But I think we can safely say that we are in the golden age of data and AI. So, we believe that AI is really going to usher in the next generation of product and software innovation. So, we're already seeing this playing out in the market. So, this report was published based on data from

9,000 of our customers. And so, three main takeaways from that report: the first one is that companies are adopting machine learning and LLMs at a rapid pace. So, NLP is really dominating those use cases with an accelerated focus on LLM specifically. The second thing is that really open source is winning in today's data and AI markets. So, 8 out of 10 of our most widely adopted data and AI products are based on open source. So, I think that's good to know. And I don't know if you've seen, there's a lot of leaderboards out there ranking different LLM products and they rank them on accuracy and categorize them based on open source or not open source, and then if they're allowed for commercial use or not. And so, if you look at those leaderboards, basically the biggest takeaway is that open source is catching up really quickly. And so, companies are really seeing the benefit of keeping their data private, not giving their data to open AI or other LLMs and training their own models based on their specific use cases, their own data, and their domain.

So, that's what we're seeing. And the biggest challenge here is the hallucination issue. AI is sociopathic, right? It can lie to your face without feeling bad. So, we really need to figure out how to prevent this from happening. So people have been actively researching this, right? improving the diversity of your training data, eliminating inherent biases that may be present, developing better regularization techniques, employing adversarial training and reinforcement learning – things like that. So, I think that's a big problem to solve before we have more reliability, but I think definitely it's going to change the way we work and add efficiency to what we do.

Ron Wasserstein: Thanks Ginger. Amarjot, what are you thinking?

Amarjot Kaur: Yeah, Ginger, gave nice opportunities and challenges in this, so I would agree with you. Ron, like what you said, it's both exciting and scary, and sometimes creepy too. So, there are values to using these AI tools. I can see in our drug development, for example, in the discovery phase, of identifying what molecules to take forward and so on. And, and that's a huge, undertaking, right? So, AI and machine learning, those kinds of tools could be useful to finding the most promising candidates, in a more precise and efficient way. So definitely there are a lot of other applications where you can see the value of it. And I think the challenging part is, of course, because there has to be a human element to it.

Whatever is coming out of AI, right? It can have the hallucination effect like Ginger mentioned, and it could give answers, could be totally wrong, right? So, you can just rely on it like that, just because it came based on a lot of information. So, one parallel – I always think of it like a navigation system in our cars, right? I mean, it's extremely useful too, right? I mean, we have so much dependence on it these days, but at the same time, you also have to look at the signs on these roads, right? Sometimes the navigation, if you don't pay any attention to it, just follow that. You may just fall off the cliff or get into some ocean, right? So, therefore, I think it's in a similar way. I mean I think there's a lot of unknowns here in this. I mean, if you look at the news or read about it, I mean, some people are really worried about it, where it is going and what's going to happen, right? Because of the unknowns associated with it. But again, opportunities are there as well. So, that's how things move forward right there, there's always a hesitation, but hopefully, there would be some kind of boundaries around which people can function and make the best of it.

Ron Wasserstein: Thanks. Thinking about the GPS for a moment, I was cruising through the northern part of New York City, going through the Bronx on my way back from this conference, and it gets really confusing there. At one point the GPS basically seemed to suggest that I was just on my own for a few minutes, and good luck, it would help me out again when it could. So, I was thankful for the actual signs.

So, Amarjot, let me come back to you, and then we'll swing over to Ginger with this question, which is basically about career advice. One of the things we love to do on the *Practical Significance* podcast is to get our guests to share experiences and advice with our listeners. So, what would you say is your best career advice?

Amarjot Kaur: Stay curious, right? Stay curious and be adaptable to change because things are changing. I mean, life is never a straight line, right? So, and also, I think, as I said earlier, before jumping into finding solutions or analytical solutions, first try to understand the question, right? What is it that we are trying to answer and how we can best answer it? It doesn't always have to be so complicated, right? So, curiosity and adaptability, I think, are important. And I think what also is important is the ownership of what we do. Whatever task is assigned, no matter what kind of task it is at any level, right? We should own it in a way that when we are working on it, we should think beyond what the ask is. Because I think as statisticians and data scientists, we are the ones closest to the data.   And the way we can kind of find unexpected things through the data is, is something we should pay attention to that, right? So, try to see what more we can do to make, um, the whole process efficient and find solutions, from that data. And then, my advice to everyone is to volunteer and stay professionally active in various statistical roles with ASA or other organizations, because, to me, I think personally that that has been a very fulfilling part of my career. And, and I think that's something that helps to develop and diversify your worldview.

Ron Wasserstein: Thanks, Amarjot. Ginger, what kind of advice would you have?

Ginger Holt: Yeah, I mean, I echo a couple of things that Amarjot said about being adaptable to change. I think applying that to your career, I would say take a model, predictive-control type of approach like planning a career for the long term and then letting three to six months pass and then do another pass, right? Based on the new information you have, how am I going to iterate that, right? Based on my new interests, based on new developments in the field, and new knowledge that you may have. So, being intentional about setting a timeline, in order to do that iteration on career planning, and have a policy of having research heroes, right, that you follow on Google Scholar and then you keep up to date with advancements. And then also it helps thinking about these people, right? About keeping your standards high, imagining those people looking over your shoulder as you're doing your work and delivering analysis, right? It's like what would they think of this work? Staying active in conferences is another thing to just keep learning. That would be my advice.

Ron Wasserstein: Thanks, that's great advice from both of you. I was just thinking as you were talking about this, that I guess I'm realizing that maybe I'm just getting to the point in life where I'm completing the training set and about ready to work on the test set. We'll see how that goes and I'll report back.

Donna LaLonde: I love the idea of research heroes, that is really great, and the notion of them looking over your shoulder. And Amarjot I have to say that I've been thinking a lot about how you engage in what you do and how you value what you do in the grand scheme of things. And so, I love your advice on owning, right? Feeling like you own your job and so that you have that sort of obligation to do your best. This has just been fantastic. So, I say at the end of the podcast that we always like to ask, but in fair disclosure, (it's probably because Donna really likes to add to her podcast listening and music and reading list). So, I would love to hear from both of you. What's on your to-be-read list, your playlist, what you're watching, any and all of the above. And Ginger, I'll start with you.

Ginger Holt: Okay, cool. Well, I'll kind of nerd down a little bit. <laugh>. I recently caught up basically on the whole causal inference topic. So, I read *Causal Inference for Statistics, and Social and Biomedical Sciences*. I found it very well written, very easy to understand, and just genius in its simplicity because the authors really derived the whole notion from fundamental statistics that I already knew. So, I would recommend that. And then kind of on the personal front, I have a five and six-year-old, so for the past really five years I've been reading a lot of parenting and child development books, just to try not to screw my kids up so much <laugh> but, I continue to be interested in the brain development of children, like as they age, right? How much logic they have, what they can understand. And I found that pretty interesting. And then in general, the books that I like to read are about data analysis and applied economics, leadership, and then organizational psychology, and especially books that kind of combine those topics together. I just really find those very interesting to read.

Donna LaLonde: Ginger, I'm actually thinking we need to start a Caucus of Industry Representatives reading group, yeah?

Ginger Holt: Yep. <laugh>.

Donna LaLonde: And so that would be so much fun, but Amarjot that what's on your listening, reading, watching lists?

Amarjot Kaur: Well, my list is not as nerdy as Ginger's because <laugh>, I read mostly the newspaper and the *New York Times*. I try to get through the op-eds on that. And then I think on Tuesdays I look forward to the science times and the food section is another important read. I have a lot of cuttings from there. During COVID, our library started this virtual book reading club that my husband and I joined and that forced me to read all these fictional books. I realized how interesting they were because a lot of work of fiction is actually based on some reality, right? And then of course the author makes it and takes it wherever they want to take it, but they do some research. And I think one of those which I really enjoyed was *A Gentleman in Moscow*, which was a book I think was very well written. From a literary point of view, it was very nice, and it takes you back to a lot of the history of the Russian Revolution and so on. But it's all interesting as it happens all in a hotel, very interesting.

And more recently, I read *The Midnight Library*, which I also enjoyed a lot. Now I don't get so much time to read big, thick books, but my husband is an avid reader. He makes recommendations and he makes sure that I read it. So, which is good. I enjoy reading a lot, but I think mostly day to day it's the newspaper.

Donna LaLonde: Well, Amarjot, I just have to share with you that I am also a *New York Times* fan and I actually still get the *Times* in print on Sundays, and that makes Sundays worthwhile.

Amarjot Kaur: Yeah, yeah, I, we get it every day in print.

Donna LaLonde: I succumb to this sort of digital version on Monday through Saturday, but there is something just wonderful about holding the book review in my hands. Well, this has been super fascinating and wonderful to have a chance to talk with you, and I will let our listeners know that we will put a link for the application for the Caucus of Industry Reps, and we have a microsite that we are growing. We'll put that in the show notes. Now for another *Practical Significance* tradition. I will turn it over to Ron for his "Top 10."

Ron Wasserstein: Thank you. Donna. JSM is just around the corner and some of you may be thinking about the t-shirt you would like to rock at JSM. Statisticians should always have a nice "fit." Yes, the *Practical Significance* podcast is always ready to help. So, we bring you suggestions for what slogans you could flex at JSM. All these ideas come from folks who responded to a tweet from me asking for suggestions. Thanks to Felicity, Venkat, Seth, Lisa, Aaron Cable, and Eric for the items chosen. Here are the "The Top Ten Statistics and Data Science T-Shirt Slogans," for you to flex at JSM 2023!

#10 Keep Calm and Be Significant
#9   With great power comes significant responsibility.
#8   Data is My Superpower. What's yours?
#7   Yes, I know you hated your stats class in college.
#6   What's your research question?
#5   I'm consistent, I'm efficient, and I'm almost sure.
#4   "R" you ready?
#3   Found the problem. It's μ.
#2   A data scientist is a machine for turning coffee into models.
#1 p = .051

Thanks again to all who contributed t-shirt slogan ideas. Next month I will share with you the "Top 10 Rejected T-shirt Slogans." Thanks for listening and hope to see you at the Joint Statistical Meetings. We'll continue the conversation on this podcast next month.