Practical Significance | Episode 40: How Can I Help You Today?

Donna LaLonde: Well, welcome, everyone, to the April edition of *Practical Significance*. April, as I'm sure you all are aware, is Mathematics and Statistics Awareness Month. So, what better way for us to celebrate mathematics and statistics than to talk about a topic that we all should be aware of, which is artificial intelligence?

We are delighted to have three guests with us, and we are going to start, as we do by tradition, by asking our guests to introduce themselves and tell us a little bit about their day jobs. And so, Tian, I'll start with you, then we'll move to Mark, and then we'll go to Hong to so Tian.

Tian Zheng: Hi. Thanks for inviting me to be here. My name is Tian Zheng. I'm currently a professor of statistics at Columbia, and for the past five years, I've been the chair of the department. So, at Columbia, in addition to running a department, I also participate in a lot of collaboration with applying scientists in multiple disciplines. And it would be an understatement that every discipline is being flooded with excitement in AI. So, I'm super excited about what we will talk about today.

Mark Glickman: Hi, I'm Mark Glickman. I'm a senior lecturer on statistics at Harvard University. I've been at Harvard in this role since 2016. I'm also a senior statistician at the Center for Healthcare Organization and Implementation Research, which is a VA Center of Innovation. I guess the reason that I was invited to join this prestigious group is that I am also the chair of the ASA Committee on Data Science and Artificial Intelligence, which is a committee that just started in 2022. It's a new, young, vibrant committee, and I am delighted to be chairing it. And I'm also excited to be talking about some issues at the intersection of statistics and AI in this conversation as well.

Donna LaLonde: And Hongtu. Please tell us about what you do for your day job.

Hongtu Zhu: I'm a professor of biostatistics, and actually, I have a joint permit for multiple departments, including statistics, computer science, radiology, and genetics. And before that, I worked in a tech company for three years as a geoscientist. That's how I got into AI and also statistics and data science, in general. I have been at UNC since 2006. I work on multiple projects. I work closely with a computer scientist and a radiologist, and also people from other departments in the medical school.

Ron Wasserstein: Well, we're sure delighted to have the three of you on the podcast today, and we have a million questions we're going to ask you, just a small subset of the questions that we have. So, I'm going to start by asking you how has, or how should the field of statistics evolve to meet the challenges of AI development and application. And I'll just call on you one by one. Mark let's start with you this time.

Mark Glickman: The area of AI recently, in my view, has been an area that statisticians have been paying pretty close attention to. I think what has been happening over the past, say, ten years or so, is that there have been these pretty incredible successes in AI. The methods that underlie a lot of these procedures, particularly in neural nets and deep learning, are a little mysterious. On the one hand, there is a pretty solid foundation for a lot of these methods, but on the other hand, it doesn't seem to correspond to things that statisticians have been thinking very much about. I think one of the burgeoning areas in statistics and how it's connected to AI is just understanding the underpinnings of a lot of these methods in AI, which just seem to work.

So, there are certainly a number of statisticians who are doing some amazing work at the forefront of understanding a lot of the underpinnings of these mostly deep learning methods that are just driving technology in very impressive ways. I'll also say that the other area that statisticians are involved with, this whole area of AI, not so much in a direct developmental way, is helping to be able to evaluate AI methods and work with essentially the output of AI methods to develop approaches that have maybe a little bit more familiarity among statisticians.

And just as one quick example, the whole area of conformal inference is a really big area that's attracting a lot of attention among statisticians, mainly because conformal inference basically starts with what could be a black box, and then starts making predictive inference without really even knowing very much about what generates observations. And it's a very powerful tool that statisticians are jumping onto. And that's a good example of an area in which statisticians continue making some very important progress in parallel with the development of AI.

Ron Wasserstein: Thanks, Mark. So Hongtu, what say you?

Hongtu Zhu: For the field of statistics to move forward in the AI era, there are four things we need to consider. The first one is that the current curriculum needs to be modernized to fulfill the evolving demands of modern data science. Our curriculum was mainly developed almost 30 or 40 years ago with a little bit of modification, particularly in areas such as engineering capabilities, practical data analysis appearances, and proficiency in data mining techniques. And the second thing is the existing evaluation systems need some changes.

I have been published in many papers, in machine learning conferences and journals. Based on my observations, we needed to speed up the review process in our field and make our papers more acceptable to the data scientists and the practitioners and not make them too mathematical. When people use this method, they don't want to see a lot of math notation inside. They tell me the story and how to use it in practice.

Another thing I'm thinking about is I have been attending neuropsych meetings for years and our conferences are too conservative in the sense of giving opportunities to the young rising stars and active researchers. Because in neuropsychology, on the first day of the tutorial, all these lectures, as graduate students, and also assistant professors, we need to think about that. I think that's important because the whole field is growing very fast. The third thing is the systems need to include all the data science-related journals and conferences, not just focus on all the statistical journals because there are so many new journals and conference proceedings, we need to open to our field.

Another thing I think is very important is to encourage and promote a greater portion of the statisticians across various study sections. And actually, we need to train them for effective communication and equitable contribution. This is super important and in particular all these study sections in AIHA, we need more professional citations to participate in all these panels

This is very important; I think about the field and also, it's highly beneficial. We need to get involved in the private sector to invest in our research and also in product development because we can do the same thing. We use that to develop all these pipelines, protocols for the FDA, all these things.

Ron Wasserstein: So, thanks, and thanks for just suggesting, you know, small incremental changes. So those are all big deals that will take some cultural change. So, thanks for raising those issues. Tian?

Tian Zheng: I just want to build on what Mark and Hongtu said. Mark covers a lot of exciting research trajectories we are observing, and Hongtu is commenting on some infrastructure changes needed. So, I'm going to answer Ron's question about cultural shift. I think as a discipline, as a community, we have come a long way since 15 years ago when big data first started drawing global attention. We have to give the credit that the statistics community has made changes over the last 15 years. The way we teach, the way we run our department, the way we recognize and ascend our word. We have been recognizing and promoting computing-intensive, applied, data-intensive applications, machine learning, and research more than we used to. And then we have to take this moment; and recognize we have done that. So, these computing machine learning and data-intensive applications are three very important pillars of AI. These are not changing.

We were moving in the right direction and we definitely should continue to move in that direction. We have thought leaders putting out, like being used veritable data science textbook just came out and David Donelfell published a data singularity paper calling for a cultural shift for statistics. All these are in the right direction. However, I think then how did this change happen? This change happened because we felt that data science is moving faster than statistics and we need to catch up. We had a slow start when big data first happened, and we and a lot of departments came together and took on initiatives and make change happen.

But over the past two to three years, there has been a feeling that we are falling behind again. We're being left out in some of the AI conversations. This is because AI is moving at a much faster speed. A few technological breakthroughs in recent years enabled so many more fields to embrace AI computing data than in the previous big data era. And now the whole research horizon is broader and more exciting. So, therefore, what needs to happen for the citizen community is to recognize we are doing the right thing by shifting in these three directions. However, the wind is blowing stronger and we need to continue what we're doing well, but with a more proactive effort to accelerate and encourage more energetic participation. And in addition, there's one thing AI is different from big data. Big data calls for collaboration.

Statisticians can contribute effectively by being a very effective collaborator. AI involves workflow and systems. For us to contribute meaningfully to AI, at least some of our statisticians need to engage in the end-to-end AI workflow to bring back important problems for the whole community. If no one in statistics is willing to be the pioneer to go into AI and lead AI applications, then our company will not be able to keep up speed with AI research. So that's what I think is a culture shift we need to make, continue what we do and build faster.

Ron Wasserstein: Thank you for that, Tien. And thank you all three for these great answers. How's my head spinning a little bit? Speaking of head spinning, I now want to ask you what you think are the biggest challenges for statistics in the future of AI. You've touched on this a little bit, but Hongtu, let's start with you. What do you see as the biggest challenges?

Hongtu Zhu: I think in terms of. In my opinion, the biggest challenge to our community is. There are several things. Why is the a diminishing pool of tenant students? I think because nowadays the CS and the grammar many this kind of tenant students, that's the number one. The second thing is funding opportunities. There are a lot of funding opportunities for AI-related research.

I think there's much more because both the government and federal government and also private sectors invested donors in that. Okay. That's the basic thing that we're facing. To me, that's the big issue. Number three is the emergence of new areas in the private sector. The new domains with no statistics involved. Okay. I think that's the three things I worry a lot about every day. Okay. That would be a big problem for us. I was just talking. I think that these big challenges are big enough.

Ron Wasserstein: Thank you for that, Tian, how about you?

Tian Zheng: So, we have a lot of potential. I often say that no one is going to deny that AI needs statistics. Statistics should be central to AI to address a lot of important problems. Recently I have heard several seminars where I see AI as a new way for us to think about what a model is, like expanding the definition of the model is creating new ways of quantifying uncertainty. And recently I think that they're creating new ways for us to think about probability distributions through all the diffusion models. So, these are opportunities.

For example, this week, NSF just released a new “call for grants” for AI-assisted mathematical research. So obviously AI is expanding the toolbox of mathematicians. So, shouldn't AI be part of our inquiry too? In addition, AI can also be part of our toolbox. So, in my view, one of the biggest threats to our discipline now is a talent pipeline, faculty development, and research resources.

Like Hongtu said, if we're not proactively revising our curriculum to enable our students and young researchers to take better advantage of AI in their toolbox, and to be able to develop a deeper appreciation and understanding of the AI system so that they can make their research relevant, then all these “call for proposals” where the national resources for AI research will not have too much of involvement of statisticians. And then in a few years, they will recognize that this is still central to AI research.

But our pipeline has already dried up. So that is what I see as the biggest challenge for statistics. Of course, there are intellectual challenges, like problems, you know, what is the probability distribution for complex objects that AI can help us create? That's the kind of open challenge one can think of. But we do need to worry for the next generation, whether we can continue to have the generation of bright and excited young researchers identify as decision-makers, learn about the central foundational principle of statistics, and at the same time be able to embrace research in AI.

Ron Wasserstein: Thank you, Tien. So, Mark, let's wind up this question with you.

Mark Glickman: So, without duplicating what Hongtu and Tian said, I will reemphasize that one of my concerns has to do with a lot of the tensions that have been going on over the years between statisticians and typically computer scientists in the area of machine learning and deep learning in particular, where all this development has been done in a way where the statisticians have been struggling to be at the table. And I think going forward, one of the biggest concerns is that statisticians may if they're not proactive enough, be left behind. And basically, the world of AI is going to try to operate and continue its development without statisticians at the table.

I think it's important to recognize that statisticians do have value and try to make this clear to the computer science community, in particular the ones who are mainly behind the AI movement, because if we don't, then we're going to run into some real issues. It's also important, in making this case, that AI, while incredibly impressive, doesn't necessarily solve all problems.

Even fairly basic statistical approaches are not only perfectly adequate but often are superior to AI approaches to solving problems. I was just talking to a colleague recently who's in the area of sports analytics. He has a lot of pressure on having all these predictive models being constructed through deep neural nets. At the end of the day, he ends up implementing a logistic regression, and it outperforms some of these deep-learning models. So, there's plenty of room to perform fairly straightforward applications of basically classical methods, and those are going to be perfectly fine.

I'll mention one other thing, which is that I think one of the areas that statisticians can be much more vocal about is that we have the expertise on uncertainty propagation, which most leaders in the field of AI are maybe not paying quite as much attention to, in part because they're working with so much data that maybe they don't have to worry as much about uncertainty of any inferences that they make. But that's so ingrained in the statistician mindset, and I think as things proceed, where we're working with enormous data sets that are being used for very personalized kinds of applications, like personalized medicine or personalized education, then you do need to start worrying about the level of uncertainty in your conclusions, and that's something that statisticians can very much help with.

Donna LaLonde: So, Mark, what are you most excited about?

Mark Glickman: Probably like many other people, I woke up one morning in late 2022 with the news that ChatGPT was made available to the public. And I guess the reason I paid attention was I was reading an article that was entirely written by ChatGPT, and then I was told, ”Yes, this is something you could play around with.” Even though I was certainly paying reasonable attention to what was going on in the world of AI, I was stunned at what some of these generative AI algorithms could do using large language models.

I'll start by just simply saying that I'm very excited about the promise of generative AI and a lot of what it can do specifically in the area these days. I mean, I regularly use generative AI to help me with my writing, at least at minimum as a way of proofreading my writing, because the language models are so good that you can pretty much guarantee that your writing is going to improve, also even for coding tasks.

So, if I need to implement something quickly, or I just don't want to spend a half-hour writing something up, I'll just, in a sense, explain what I need to have coded up, and then I'll just get my answer immediately. So, it's a huge, huge time saver. And I'll mention the one other area that I'm impressed by and I think has an enormous potential impact is really in the area of enhancing the creative process.

So, for example, I'm a musician, and just being able to use generative AI to help inform writing music or even just writing stories or text is such an enormous help and a time saver and not a replacement for generating stuff that comes out of my head, but something that can be very helpful as an assistive tool. And so, some of these ideas are just incredibly helpful.

Donna LaLonde: Tian, I'll go to you. What are you most excited about?

Tian Zheng: I'm a data person, so I'm generally excited about cool and fun data sets people put out there. It used to be very hard to find a collaborator who was willing to share data because AI is so exciting and I have never been approved by so many collaborators. Just saying, oh, I have data, I have interviews. I heard AI and ChatGPT can help us analyze the data. So, we want to talk about this. So, I got very excited about this willingness to collaborate.

In addition, I also recognize that every time you collaborate with a new discipline, you need to learn how they organize their data, and what a special format their data is in. But nowadays AI has done such a good job of educating people, saying for you to use the AI tool that there are the following data formats you have to comply with. So, our collaborators have also been educating themselves in machine learning AI. So, the data are becoming better, better in format now that they have less bias, but at least they come in a little bit better format than YouTube. In addition, the kind of tools Mark mentioned, such as ChatGPT, also make collaboration easier in a way that I used to ask a lot of questions to my collaborator about some simple definitions.

I sometimes feel bad about bothering them too much with some very fundamental questions, say in climate science or epidemiology. Now I have ChatGPT as this tutor, a very patient tutor, and I can just ask anything and then really develop an appreciation of a background knowledge very quickly. I'm most excited about the richness of the opportunity available to us, available to nearly anyone in statistics who wants to embrace collaboration.

Donna LaLonde: Great, Hongtu, we'll end with you. What are you most excited about? [

Hongtu Zhu: First, the thing is, we can modify and adapt many AI tools to our projects and expand our spectrum in various applications.

They give a lot of opportunities and I do a lot of image analysis. AI choice is pretty useful for many tasks. It's revolutionized the field. Also, nowadays I do a lot of NLP types of things because when you have open AI, you don't need to collaborate with NLP research. I can do all the things because that gives us opportunities. The second thing I'm thinking about is integrating AI with many existing statistical methods for further method development. I combine the neural network with quantile regressions for many kinds of projects, and in the tech company in particular, related to the treatment effects and also distribution or reinforced learning type of things, because you can, look under the problems much clearer from that aspect.

And number three is we can improve many more complex theoretical problems because AI, basically the tools remaining developed for all these kinds of pattern recognition problems. There's a lot of successful challenges behind it. I see these challenges. At the same time, I think that's an opportunity for us, right? Let's create new models and new kinds of scenarios. We need to work on it and it makes me excited.

And a lot of things we can improve our education programs by using AI tools actually in our universities. And we have these kinds of, you know, all kinds of committees from the, from the provost office to the dean's office, from the department. We are thinking about basically how to integrate all these AI tools with the current research and also our education programs. You think about all the kids' elementary school, they'll be talking about the tools, they use for educational purposes.

I'm excited about these things. I don't feel that threatened. Other than that, I want to embrace AI tools.

Ron Wasserstein: There are a lot of exciting things to think about in that respect. We're going to slide into the career advice portion of the podcast, or what it might be this time, the Biff Tannen portion of the podcast. And you have to be of a certain age and like certain movies to recognize that reference. But for the rest of you, what we're going to be asking here, now that you know what you know from your experiences and the things that we've talked about today, and as you've been thinking about the future, what advice would you go back and give to yourself as a student to prepare for what people are coming in for today? Because I'm sure your students are asking this question as well. So, what would your current self tell your past self? Tian, we'll start with you on that.

Tian Zheng: This is an interesting question. I was pretty adventurous when I was a student. I did a lot of things that other students wouldn't have done. For example, when I was working on my dissertation, I went and took a class from the genetic department on computational biology and sequencing, which involved exam projects, with my friends from the biology department.

And I went to seminar time in other departments. And so, I had a lot of fun. If I travel back in time, I would tell her, to have fun there. You'll enjoy it and encourage her to keep doing what she was doing because I don't believe there was something that I could have tried and I decided not to. I think being adventurous would be my advice to my past self.

Ron Wasserstein: That's great advice, and it stands up for a lot of things. But I have to say, when I think about my student self and working on my dissertation, taking a class for fun to learn some extra things, probably wasn't something I was dialed into at that point, but it's good advice and it's served you very well. Mark, what do you think?

Mark Glickman: I'll sort of take the opposite tack from Chien and say, going back to my student self, I probably would have advised that guy just to have a little less fun. I prioritized making sure that I had a fun time during my graduate school years. Maybe on reflection, I should have hunkered down a little bit more. But anyway, it all had a happy ending. I think in terms of the material that I might have chosen to focus on. I suppose had I known then what I know now, I probably would have veered a little bit more in the direction of paying attention to computing.

I mean, I graduated with my PhD in 1993. And that was essentially the years that Monte Carlo Mark of chain became pretty popular. I was living in a Bayesian world at the time, so it was right at everybody's doorstep. I certainly was a heavy user at the time, but I kind of wish I got a little bit more into the computational details and then just generally follow computation. That said, for students now, I guess I would probably end up giving somewhat conflicting advice, I suppose because on the one hand, I do feel strongly that students should have a pretty strong background in basic statistical principles, and so I wouldn't ignore that at all.

On the other hand, at the same time, if students can certainly be aware of this whole world of machine learning, AI, and even more so data science, where it's important to be somewhat conversant in the science that you're trying to apply a lot of these quantitative methods to, that's an important piece of the puzzle. So, seeing the entire data science pipeline is, I think, an essential piece to being a well-trained quantitative member of society in a very detailed sense.

I certainly do advise plenty of students on what would benefit their career goals, and usually, I do tell them that you do need to pay pretty close attention to statistical principles. But at the same time, it's worthwhile to be pretty well aware of what's going on in the AI world, because that's basically where a lot of the activity is going on.

Ron Wasserstein: Thanks, Mark. It does seem like things turned out okay for you, despite how it may have been back in the day. So, Hongtu, what sort of advice would you like to pass along?

Hongtu Zhu: I have very complicated career paths, but in general, I feel like the most valuable things you need to do are applications, very complicated applications, and really inspire you to develop the methods and also do theory theoretical development. Some people think I'm against the theory. I'm not. I always tell my students that they need to understand the data and understand the problems, and then they're doing method development. And that's how I train my students these days, compared with the students made ten years ago?

I always throw them into a concrete project, and you do all the data analysis, then you understand the data and then send all the methods the problem of the method, and they start to think about how they develop a new method, how to develop a new theory behind it. That's something I always do these days; I change the way I'm mentoring my students.

I also want to emphasize that yesterday I watched a video and they talked about the people working the OpenAI and they said everyone is in the OpenAI. They need to know how to process the data. And then second, they need to code all the methods, need to do the method of research, and also you need to make it a data-centric product. Nowadays you need to know almost everything. It's not just, I just know how to prove the theory. That's not enough. The modern data scientists, well, we will.

Ron Wasserstein: Thanks, for doing for that. That's great advice for students and also advice for faculty as they are starting as well.

Donna LaLonde: End by tradition with a question that says, in your limited free time, but in your free time, what are you reading? Listening to? Watching? I'm always interested in new books, new podcasts, or new movies. So Hongtu, I'll start with you.

Hongtu Zhu: Yes, that's like a tough question these days. I only read the book with my son. He's a sixth grader, and because he has some difficulties with reading, we read page by page. I also watch the latest news on the TV every day.

Donna LaLonde: Mark, what are you reading? Listening to? Watching all of the above.

Mark Glickman: I haven't been keeping up as much with fun things to do on the side. I'll mention that I'm a huge Michael Connelly fan. He started as a mystery writer and has become more of a police procedural writer, but he is a very good writer and very exciting.

And so, I have his last book waiting to be read. I tend to listen to a lot of music on my devices, so I often just like listening to plenty of Beatles and also Fountains of Wayne. In terms of catching up on video-type entertainment, I've been kind of going back to old movies and shows that I haven't seen, that I probably should have seen.

So, I've started going back to the classics, including seeing the movie *The Third Man*. I haven't seen *North by Northwest*. So, I've been starting to catch up on some of these really old classic movies. And then I've just decided that I'm going to go through the show *Modern Family*, which I'm enjoying quite a lot. That's a very, very funny show.

Donna LaLonde: That's great Tian reading, listening to, watching.

Tian Zheng: Yes. So, it's amazing that we wish to have more time to do all these things. I do enjoy reading. So, one book I would recommend to the audience of this podcast is a book called *Klara and the Sun* by Kajo Ishiguro. He's a Nobel Literacy award-winning author and his more famous work never let me go.

So, both books are Sci-Fi books talking about the future world where technology is interfering with normal life and *Klara and the Sun* is about when we live in a society where humans are living with robots, and the book is written from the perspective of a robot and is written in very beautiful but simple language so Hungto, if you are looking for the next book for your son, I highly recommend it. It's not a heavy read, but it's very beautifully written on the listening part. I put on an audiobook whenever I cook so it's one of killing two birds with one stone.

So, the current book I'm listening to *is My Name Is Barbara* by Barbara Streisand and she reads the book herself. Highly recommend. She even sings in the book. She also has a very adventurous personality. She talks about her decisions to try new things, which I can relate to very much when listening. A book I plan to read next is *How Data Happened*, written by two colleagues of mine at Columbia. Matt Jones and Chris Wiggins are talking about the history of data and I have heard their presentation many times. I know a lot of those spoilers from the book, but I am still looking forward to reading the book. I recently watched the show *The Good Place*, a show inspired by philosophy. So, I found that to be interesting and stimulating and at the same time very funny.

Donna LaLonde: Well, that's great. I appreciate all of these recommendations. Ron, I would ask you, except I know you're reading Izzy's newest book. Ron's daughter has a new book out that's getting rave reviews, so I'll just toss in Izzy Wasserstein's *These Fragile Graces, This Fugitive Heart*. And, with that, we want to thank our guests and we will conclude with Ron's top ten.

Ron Wasserstein: Thank you, Donna. Now for something completely different for this month's top ten list. As you know, an anagram is a word, phrase, or name formed by rearranging the letters of another, such as 'name,' formed from the word 'mean.' Always interested in stimulating the brains of our listeners, *Practical Significance* offers, the "Top Ten Anagrams of Statistical Terms." The solutions are presented at the bottom of the page. Give them a go!

Words

10 Anger
09 Maples
08 Cave rain
07 Asiatic stint
06 Map tearer
05 Ascetic dean
04 Arctic religion
03 Ego rinsers
02 Mainly polo
01 Dork lawman

Solutions

10 Range; 09 Sample; 08 Variance; 07 Statistician; 06 Parameter; 05 Data science; 04 Critical region; 03 Regression; 02 Polynomial; 01 Random walk

Thanks to Inge's Anagram Generator (https://ingesanagram.com/) for assistance.