CONVERSATIONS ON DATA SCIENCE

ROGER D. PENG
HILARY PARKER

Conversations On Data Science

Roger D. Peng and Hilary Parker

This book is for sale at
http://leanpub.com/conversationsondatascience

This version was published on 2016-08-06

Contents

Not So Standard Deviations: The Podcast

Thanks for purchasing this book!

The content in this book is based on episodes of *Not So Standard Deviations*, a podcast that we started in 2015 and continue to publish about every two weeks. On this podcast, we talk about the craft of data science and discuss common issues and problems in analyzing data. We also compare how data science is approached in both academia and industry contexts and discuss the latest industry trends.

You can listen to recent episodes on our SoundCloud page or you can subscribe to it in iTunes or your favorite podcasting app. We are also available on Stitcher and through the Google Play Music store.

If you want to support the podcast directly, you can become a patron at our Patreon page. For $2 per episode, you can get a really cool *Not So Standard Deviations* hex sticker!

If you have any feedback for us about the podcast or the book, or if you have any questions you'd like us to discuss, you can email us at `nssdeviations@gmail.com` or tweet us at @NSSDevations.

Thanks again for purchasing this book! We hope that you enjoy it.

Roger Peng and Hilary Parker

Life as a Data Scientist

Hilary

So what's your story, Roger?

Roger

I'm an associate professor of biostatistics at the Johns Hopkins Bloomberg School of Public Health. And I have been here for a very long time. I've been here for 12 years.

Hilary

Wow.

Roger

I started as a post-doc, then I was an assistant professor, and now I'm an associate professor. And I'd say the job—what do I do on a daily basis? Well, one of the issues of this job is that it changes a lot, right?

What I try to tell students nowadays is that it's a lot like starting your own company. And not that that necessarily helps people because most people don't know what it's like to start your own company. But I imagine it's like starting your own company in the sense that you are out there on your own. It's like starting your own company but with a salary. So if you actually were going to start your own company, you would have nothing, right? Or maybe some investment or whatever. But there would be no salary.

Being an academic is like if you were to start a company and you were paid a salary, so that's nice. But, of course, you can't become a zillionaire as you could if you started a really successful business. So you limit the downside, but you also limit the upside. To me that's being an academic, in a nutshell.

Hilary

But what does your company do? You know what I mean? I think that's definitely something—I think Hilary Mason said something about that, "It's a startup of one." Do you feel like the types of companies you could start are broader than they would be in "the real world?"

Roger

I have no idea. Honestly, I think, probably not. I think there's different kinds of things. You're able to do things that people are not willing to pay for directly. So you don't have to worry about commercial viability, but you do have to worry that *someone* is going to want to pay for it, whether it's the government or a nonprofit or whatever.

There are constraints, obviously. You can't just do whatever you want. I think you do start up on your own and it depends on what field you're in. You have to hustle to get things going. You've got to make a name for yourself. You have to develop a reputation. You have to get students. If you're in basic sciences, you have to build a lab.

I think in statistics, it's not as big, so to speak. It's not very common that you have these 20-person labs in statistics. But you do have to grow yourself a little bit to the point where you can have a number of different things going because you can only do so many things yourself.

Hilary

In this analogy the funders that you're talking about are grants from the government.

Roger

Right. So, in my line of work–I do statistical methods for environmental health problems and these can be populations at a large scale, population health problems, and also clinical studies. The organizations that are interested in paying for this are the National Institutes of Health, the Environmental Protection Agency, and a couple of nonprofit types of institutes and organizations. So you have to work within their broad agenda. NIH has a very broad agenda, so it's not very constraining.

EPA is typically more policy oriented, so they're interested in research that would support that. But there's, generally speaking, a lot of flexibility there. When I started, I got hooked into a group here at Hopkins that was doing this stuff already. I had the advantage that there were people working in this. I wasn't totally on my own. There were people here to work with me, and I could latch on to people's projects from the beginning and have some time to figure things out until I was running my own grants and developing more of my own projects.

That's something I sometimes tell people to look for: If you're going to a new place, if you're looking for a job somewhere, are there established groups there that you can plug into at the beginning, just to get started? It's important to not get sucked into a large group and then you just get lost in there, you know? But it's often good just to start so they can hit the ground running.

Hilary

It's nice to have people to bounce ideas off of. Although, I'm also finding that people have different preferences for that. Some people want to walk in and be the only person doing something. It's definitely a personal preference to some degree.

Roger

I would agree with that. It wasn't my personal preference, but I can see other people want to be doing their own thing from the get-go. And I think a lot of people see academia as the opportunity to do what you want on day one. And I think that's totally fine, but not everyone is ready for that on day one.

Hilary

Were you glad that you did a post-doc?

Roger

I was super glad. I think it was a great for me. I always tell my students to do a post-doc because I think in statistics, statistical training tends to be generic in many ways, and if you're going to focus on a specific area, it's good to have extra time to learn the area. And so, I think a post-doc is really useful just to season yourself a little bit, so to speak.

It's also good to have some time to get out some of your thesis papers, published papers, without having the responsibilities of being a faculty member. I'm seeing that doing a postdoc is a little more common now. Now, when you go on the job market, if you're a fresh grad student competing against someone with a two-year post-doc, it's challenging.

I think a postdoc is a good couple years to have time where no one's going to bother you to serve on a committee or whatever, but at the same time, you get paid better and you have more authority, or autonomy, I should say, to do your own thing. I appreciated doing it, and I think it helped me in the long run.

Hilary

Another thing you just mentioned is these other responsibilities, for example, teaching. How does that fold up into your startup analogy?

Roger

Well, I imagine if you're going to start your own company, you wear many hats, right?

The three main jobs you do are research, teaching, and then miscellaneous. In a lot of places, you won't have to teach maybe for the first year or something like that, but eventually it does get folded in. And one of the things that's really important that I discovered that no one told me about is that it's important to teach the classes to the first year students.

Hilary

Why is that important?

Roger

Because if you don't teach in the first year sequence, then you may not meet them. For example, if you only teach an elective, if they choose not to take your elective, you may never meet some of the students.

If you teach a first year class, then they see you from day one, or from the first year and they know who you are. They know what you do. And if they're interested in what you're doing, they might come and talk to you and say, "Hey, I want to work with you on whatever." But I found out that when I rotated out to another class that I taught that was mostly second and third year students, I didn't really get to meet anyone in the first year, and then by the time they're in the second and third year they've figured out what they want to do already.

Hilary

So it's your recruiting opportunity.

Roger

Yes, in part it's a recruiting opportunity.

It's also different in a public health school because we don't teach undergrads. So the volume of teaching is quite a bit less. It's made up for in terms of the volume of the research. We do a lot more grant writing and that kind of stuff, but we don't teach the thousand-person class, or the two thousand-person class, or whatever, that you might have in a large undergraduate institution.

Hilary

Although at Hopkins there is a 500-person Intro to Biostat class, essentially a full year that all the public health students take. All the masters' of public health and masters' of health science students, I think.

Roger

But even still, it's not as many as you might teach at a large state school or something like that. Also, graduate students just have a different mentality than undergrads do. It's much more professional. They have a better sense of what they want, so it's just a very different type of environment.

Hilary

Yeah, that's funny.

I feel the need to explain that the reason I know what classes are at Hopkins is that I went to Hopkins. So you were a professor when I was a grad student. I worked with Jeff Leek on genomic stuff, and then got interested in tech towards the end of grad school. I was also really interested in teaching, and still am. I really like teaching.

I ended up applying for tech data jobs and went to Etsy as a data analyst. I would describe the work there as, from the stat perspective, it's very much like being an internal consultant. So I specifically am a product analyst. These were all words I had no idea of when I was going into it. All the different roles that a data scientist could take.

There's different teams at Etsy that work on different things like the search experience, or the checkout flow, or the listing page. I'll consult with those teams and do any sort of analysis that they feel they need in order to make product decisions. That can be something like A/B testing. It can be opportunity sizing, figuring out how many people are using a certain feature, figuring out the behavior of people using certain features, things like that. Theses are things that, I think, are pretty common, especially at a startup. A data analyst is obviously, for any company that wants to be data driven, it's a critical person. It's the person who can actually parse and explain the data and tease apart causality and think about problems critically.

It was actually a surprise for me when I got to tech to see how similar it was to my biostat training because product development is very similar to drug development and drug development is something that in a biostat program you learn about, like phase I clinical trials, phase II, etc. It's pretty similar, this idea of building up to a big launch of something and justifying it with various levels of data.

Roger

I actually think that– My personal feeling is that actually a Ph.D. in biostatistics, specifically, is one of the best degrees that you can get today.

Hilary

I agree. That's not just Stockholm syndrome.

Roger

Even better than a Ph.D. in statistics, and the reason I say this is because biostatistics has traditionally, and I think still is, more applied, and it has these traditional connections to pharmaceutical companies and things like that. But I think it offers you this amazing flexibility.

One of the things I—and I don't even know if this is true—but I tell people this anyway. I feel like if you work in industry—at a company, and especially a smaller type of company like Etsy, as opposed to an enormous company like Microsoft—I feel like there's more of a sense that everyone's on the same team and you're working for, broadly speaking, the same purpose. Is that true or false? Am I just making that up?

Hilary

That's definitely true. In the startup analogy, it's not a startup of one when you're at a startup that has more than one person. There is a sense of working together on stuff. That's not to say that there's not disagreements internally. Obviously, any workplace is going to have differing perspectives and people coming in and wanting to do things slightly differently.

I think in academia because every professor is a startup of one, people defend their ideas a lot because there's a lot at stake for that idea. If you have this idea and you're building your entire research career around it, you're not going to be very flexible on that. Whereas, I think within a startup, it's the idea of the startup, and so you have to choose carefully what kind of statistical hills to die on, if you will, because you can't stop progress. It's a really different attitude and people work together.

I found it a lot easier to get help—there's a whole ops team, and they do all the infrastructure like maintaining the servers and things. If I needed help with commands, I have a whole team of experts who I can go to. And they're happy to help. There's definitely a spirit of camaraderie. Not that academia doesn't have that, but there's less incentive.

Roger

I think one thing that's a big misconception with academia is that because you all work for the university, that you're all employees. And yes, you are employees, but it's like if you all worked at an incubator or a co-working space. There might be many different companies in that space. And yes, you're all in the same building, but you don't all work together. You all have an independent mission.

You will work together if it serves the two purposes. But there isn't a natural inclination to work together because everyone's trying to develop their own program and build their own thing. It's not like everyone hates each other. Some universities have these reputations for being competitive and nasty. And I don't sense that. It's not like that.

But there's not the feeling like we're all on the same team, we're all striving to achieve the same goal. We're not shipping the same product or something like that. The last time I worked outside academia was a long time ago, but I did feel like in those jobs that I've had, that especially if you're all working on the same product, there is a camaraderie there. Let's all just do it together, because we want to do something that's really good in the end. And I don't get that quite as often in academia because I think there's—although you do collaborate, it's like yes, you're collaborating, but you're also doing your own thing.

Hilary

When you were talking about that, I was thinking about how it's amazing at Hopkins that there was a shared computing cluster. And there were professors who worked on that full time. That's even unusual and camaraderie like in a way that other—I feel like that's an unusual shared resource. I feel like I've known professors who would have their own cluster. They would have a stack of computers in their office. That's a lot of work to do alone.

Roger

Exactly. I think when we did that, it was novel at the time. I think that model is more common now. But it was new at the time.

Hilary

That was extremely helpful as a grad student to just be able to walk in and be, "Okay, I can submit jobs to this cluster and there were shared resources, and I could ask other students and all those things". When you go to a tech company, especially a more established one, your first week you might be just rotating with engineers learning the basic infrastructure and how to push code and how to file tickets. It's much more like everyone's doing the same thing.

Another thing I was thinking too was that in academia, a big issue is getting scooped, when I need to publish this paper and someone else published something similar. Whereas in industry, tying your ego to a project is seen as a bad thing. The priority is getting the project done, and if someone else can get it done faster and they're more specialized in it, then you should be happy. It's like this balance of personal development versus getting something done for the company. That was a really different attitude as well. That was a mental shift—you can't be possessive over ideas that are shared.

Roger

One question I have for you is, who or what drives the kinds of problems that you end of working on?

Hilary

Yeah, that's a good question. It's driven by the questions that arise. At Etsy, and this will be common for tech companies, besides our data analysis team, there's also a data science team, and that's our machine learning team. They're people who develop products for using data that go on the website, versus I'm analyzing products that other people are building. In the past, I've analyzed machine learning products.

In terms of the stuff that's being developed or the questions that I'm answering, sometimes, they'll think of things when they're doing user research with people using the Etsy app. It's usually like one team is working towards some large goal, so questions arise along the way about how to develop that and what the best things to go after are. And then sometimes we also drive—we'll take time to do independent research projects on various user data to understand what are some big overarching questions that we're interested in. What are, maybe, some things that we've been neglecting that we shouldn't be? The questions come from all over.

You want to be helpful to the teams you work with, so sometimes that means doing something that you don't think is the most interesting question, but it'll help make a decision. You have to be very motivated by helping with decision making in a role like this. Similarly, for the machine learning team, they have to be motivated by building successful products, which may or may not be using the most cutting edge machine learning techniques. It's an area of research, and there's always improvements, and there's deep learning and all these things. But at the end of the day, you have to care about how it's performing and if it's serving the needs of people.

Roger

There's an analogy to that for what I do. When you collaborate with people who are in the various sciences, they have a problem they need to solve. And it may not require you to develop a brand new method that—and publish a JASA paper and do this whole thing. They may need the quick and dirty thing, and the quick and dirty thing may solve 80% of their problems, which is all they need. So you have to balance—but that exhibits the tension between building your own career and helping out your collaborator, I think, because writing that brand new method that you take a year to develop, that helps you out, but it doesn't help them out, necessarily.

Part of the job, I think, is to just figure out how to balance it. And sometimes you can make lemonade in the sense that you can do the quick thing and get a nice paper out of that, and then leverage that into a methods type of paper. And if you're good at this, which I'm not necessarily, but if you're good at this, you get two papers out it. So it sounds very similar actually.

Hilary

Yeah, it's basically the same. I do think, though, I think that speaks to–it's just being an applied statistician in a different context.

That's why sometimes the whole "data scientist is a helpful applied statistician"; that attitude can make me cranky because I'm, "Oh, I actually learned to be a good applied statistician from a biostat department." It is possible to teach these things. It's an interesting time where people have strong opinions about where statisticians have let them down. At the same time, we haven't necessarily–we talked about this in our first podcast where we haven't necessarily done a good job of training.

I specifically worked in genomics, and I took some genomics classes. I took classes on DNA structure and stuff like that. But if you view data science as applied statistics within tech, we don't do any of that training that we think is necessary for applied fields.

It's extremely hard for anyone to teach because you need to have this whole ops team maintaining this Hadoop cluster. We have a whole team that maintains our Hadoop cluster that we use, and so in order to teach a good applied class, you would have to have a cluster. That's getting easier with Amazon web services and stuff, but it still requires people to know about it. And professors are coming from academia and might not have access to these things so it's much harder to teach. And so I think that's where this gulf comes in.

Roger

In terms of what we're able to teach, I think a lot of times we think there are certain things you probably should learn at a job or at a company because if things are changing very quickly or technology is evolving faster, then it might be better to just go wherever you're going and learn specifically what they're doing there. Because if we were to teach it here, it would be more generic and it may not be exactly what you're going to do wherever you're going, and then you just have to learn it there anyway. So I think it's hard for us to teach certain topics that are moving quickly, because we always have this desire to abstract a little bit and generalize in terms of what we teach.

Hilary

You need to be able to pick up skills at a job. I don't know if we always instill people as well as we should with that. I say that because I feel like I've run into a few stat grad students who have some rigidity when it comes to what tools they're using, and like, "Oh. Well, I don't do this." If we want to train people to go into this field, we need to explicitly say that more. You're going to need to learn different languages, and you're going to have to learn a lot more tech stack stuff. And, unfortunately, I feel like stat students have to get interested in that now in grad school, because even for the type of research projects you do within academia, you have to learn how to code somewhat production wise.

Roger

One of the things I was trying to get at with you is that one of the ways that we sell academia, and probably not that successfully, is that you get to choose the problems you work on and there's no one telling you specifically what to do. But it sounded like you have a lot of problems that are coming in and that ultimately you're going have to—I guess you guys prioritize what you end up working on, right?

Hilary

It's definitely a balance of prioritizing the needs of teams versus what we're interested in, but yeah, that's definitely true. And I feel like right now I'm dipping my toes into a research project that I'm really interested in. So yeah, I think that there's—you are limited to the data types.

Working at an e-commerce company, I'm always going to be working with data from the website, right? But I feel like you probably feel like you're tied in to working with environmental data because that's what you've been doing for so long. I don't feel especially limited.

Roger

That's good. It's interesting to hear because I think the fact of the matter is in academia, it's not like one day you're working on astronomy and the next day you're working on genomics or something. It doesn't work that way. I've been working on the same basic type of data for the last ten years. So, in principle, you do have a lot of flexibility in academia, but in reality, if you want to be known for something and you want to have a track record in a certain area, you have to be working on it for quite a while. Practically speaking, you will end up working with very similar data for a decent part of your career, at least until you switch into doing something else.

Hilary

Another thing is that the problems that I work on do generalize to other tech companies. So there's that academic feel to it too, where it's, "Oh, okay. I can see what people at Facebook are doing or people at other places that evaluate products similarly to how I do." So that idea of having an outlet for generalizing results is definitely there too, which is nice.

I really like that aspect of the problem, thinking through like, "Oh, this is how I solved it, and how can you generalize that, or how can you make it easier to apply this method in the future?" Write a little R package internally or something. There's that feeling. So I have not found it to be that different, except for the physical environment, like the co-workers and things.

Roger

One question I had for you is whether you miss teaching at all?

Hilary

Well, that's a good question, because I definitely do. But this weekend, for example, we were talking about it earlier, I taught a class with Mark Hansen, who leads the Brown Institute at Columbia. He reached out to me because he wanted to teach coding and R, and do a workshop on machine learning–computational journalism, essentially. And so that was actually really fun.

That was a one-day thing where I went in with David Robinson, my academic uncle (even though he's younger than me), and the two of us went in and it was really cool because we got these journalists who had never coded before to use ggplot and they got pretty far. We were looking at UN data and they were able to do comparison plots of different countries, of the percent that they're voting yes over time.

So there are outlets. You can also obviously teach your co-workers. There's a way to scratch that itch without teaching full time. But it's obviously not the same as building a career around teaching. I don't know in the future what balance I'll have for me. I like teaching. I want to keep doing things like this. But I don't feel like you have to necessarily sacrifice it.

Roger

One of the things that I've been talking about with a couple of people around here recently, and I don't know if this is true where you work, so I'll be curious to hear. But one of the things that, especially in a medical institution, there's a lot of forces driving you as you go on in your career to grow in many different ways. And I think personally, I like working with other people, but I don't like working with a *lot* of other people. But there's a lot of forces driving you to do exactly that.

If you want to be a leader in some way, then you have to run a big program or you have to be an administrator, or you have to have a big research center. Ultimately, you're working with a lot of people. And you have to fight against that. If you just want to be doing research with a small group and just writing papers and doing things, you have to fight against that in a way. What's the path in terms of, for you? I guess at some point you end up being in management or something like that, right?

Hilary

Well, so tech companies usually have two tracks of promotion. This is pretty common where there's something called the IC, which would be an individual contributor. And then there's a management track. You can get promoted within a company while maintaining this IC status. Or you can hop over from one to the other. You can hop and try being a manager.

Some people will hop back and be like, "I don't want to be a manager." But in that way, it's actually nice because I think there's a lot more—I think, realistically, you can't run a company if you don't manage, right? So IC has its limits and then management also has its limits. In some ways, this IC track isn't really available to academics, which is really ironic since academia is supposed to be where you go to think. But every professor is managing their own little company and managing students, and also not trained in any of these things of course.

It's definitely interesting where this idea that you want, especially the IC track, especially with engineering, there's this acknowledgment that there's people who really don't ever want to manage and they just want to work on their projects. There can be technical leadership without actually managing. You can become an engineering manager, or you can become a tech lead on a project.

So you're meeting to guide the project, but then the people that you're meeting with also have an engineering manager who's making sure that they're happy and fulfilled and all these things. There's a lot more infrastructure in place for these different roles, I think. There's not anything wrong with wanting to be a manager or wanting to be an IC.

Roger

There's not a stigma against not being a manager type of person in academia. I just think that, because there's plenty of examples of people who just did killer research for 20 years, or whatever. But I do think that you have to consciously do that.

I think because there's a lot of forces that are driving you towards the administrative, the manager type of role. And it can happen to you without you even noticing it in a way. There's a lot of incentive to be like, "Oh, you should have lots of students and you should have lots of post-docs." And it all sounds good, but then at some point you have four post-docs and three students, and you're like, "Wait a minute. I'm not doing any research."

It can happen to you without you noticing. It's not necessarily that there's one direction that everyone prefers or that the community prefers, rather it just pushes you in that way, and if you don't want to do that, it's fine. But you have to be cautious of it.

Hilary

What I was thinking when you were talking about that is that if you were an IC, you'll always have a manager. You'll always have someone whose job is to help take care of things and keep you relatively protected from internal politics or whatever so that you can just work. That's what an IC professor lacks, and then, I think that's frustrating.

Having a great manager is such a pleasure because then you can just think about the things you want to and they help maximize your ability to do the things you want, and that's their full-time job is to do that and keep you happy. So it's nice. It's definitely a difference. But to be a successful IC professor, you would have to learn to do that yourself.

Roger

I think also, it's different in a more traditional arts and science type of campus. But at a medical institution, like public health, or medicine, or something like that, I think it's generally more hectic, and because it needs a lot of grant writing, there's a lot of going from project to project. It's just a little busier, I think. It's harder to just keep your head down in some ways and just work.

Hilary

No, that makes total sense, yeah. And again, it's inherently a very collaborative field.

I know it's running long, but I was going to say when you were talking about your career, where does the Coursera conglomerate fold in? Your choice to teach these classes and generally be engaged in the community that way, how does that fold in with your academic career?

Roger

We'll find out. It's weird because it all happened so fast. I think the university is still trying to figure out what is the role of all this and how does it play out in terms of the faculties' duties and things like that. In the last, I guess, two or three years that we've been doing this, it's all just been on top of all my other stuff. So I haven't really given anything up in terms of my responsibilities to do the Coursera stuff. It was tough at first when we were figuring out how to do it because we didn't know what we were doing. It was taking a lot of time to build these courses and record all these videos and things like that.

It was painful in the beginning. On the one hand it was exciting because it was really new and no one knew what was going to happen. But on the other hand, it was tough because it was new and no one knew what to make of it. No one at the university knew what to make of it. So it was hard to classify in terms of this is part of your job or it's not. I think ultimately, I don't think it's going to go away.

I think universities, in general, will come around to figuring out how to fit this into the possible duties of a professor, I think. And I think the success that we've had here has made it easier to see that, okay, you can play a role in your portfolio of things that you do. But even still, though–

Hilary

What does success mean? What does it mean for it to be successful?

Roger

That's an excellent question. Well, it's tough to answer because there's–so the people who take those courses, they're not official students. The connection to the university is weaker, to a certain extent. The university is still trying to figure that out in terms of what the goal is, what the vision is, and things like that. I think in some ways we got out ahead of the university just in terms of doing all these courses and it becoming very popular. So now we're trying to figure out, okay, what's the long-term strategy here? It's happening right now, even still, I think. So I don't have any–there's not a complete answer yet, unfortunately.

Hilary

But it sounds like part of it is people liking it and being popular is, obviously, that's what you want.

Roger

Well, different schools have taken different approaches. There are a number of schools who have used it basically as marketing, basically as lead generation. So they want students to enroll in their in-person courses or in their official online courses, so they'll put up half a course on Coursera, just to tease it, and then say if you really want to get the full experience, come to our university.

Then there's other universities who've basically just taken a class that they teach in person and just dump it online, basically unmodified. They're like 12 weeks long, or whatever. It's crazy. And that's it. It's just a platform for them to put their content online. I think in some ways our program was unique because it wasn't a program that we taught in person. It was a brand new program, but it wasn't designed to get people to come to Hopkins.

We actually conceived it as a full program, the Data Science Program. So that was unique at the time, I think. Now, more people are doing it I think. But it was weird for us to do that because then people would say, "Oh, do you teach this at Hopkins?" And we're like, "Well, not really." We don't have an in-person equivalent.

Hilary

It's funny because there's courses on Coursera that I think, "Oh, I wish I'd had that in grad school. That would have been helpful to have before coming here." Especially the reproducibility stuff. That was something I picked up on my own, essentially, and with the help of professors and things. But having a class in it would have been cool.

Roger

A lot of the stuff that we built for the program filtered back in. So we teach it in person now. The whole thing's been very quickly moving, a bit of a moving target at the university level. And so it's a little bit challenging, but it's mostly been interesting for us.

Hilary

Yeah, it seems like it's a successful endeavor, even if success isn't defined.

Roger

Even if we don't know what that means, it feels successful right?

Analyses that Seem Easy

Power and Sample Size Calculations

Roger

I want to talk about analyses that seem easy, but end up being hard. This comes up often for me, but one of the things I want to talk about was the power and sample size calculation, which is the bread and butter of the biostatistician.

Hilary

And also, web company–the ones that are doing things right.

Roger

I'm interested to hear what you have to say about this. In my job, the way it works often is collaborators come to me, and they're designing a study. It might be a clinical trial, it might be an observational study, and they just want the sample size calculation (i.e. how many people they need to enroll in the study). Good collaborators give you a couple weeks' notice, and bad collaborators give you a couple of hours' notice. Often there's some grant deadline that's pending and they need to write this for a grant.

Very often, the presentation of the problem comes off as, "I know this is really easy. Can you just do this really quickly and just give me a number?" Often, the sample size is just determined by the budget so that's fixed, and then they want to know what the power is going to be or the estimated effect size is going to be. So there's three things, "What's the power? What's the effect size that we can detect? And then what's the sample size?" Actually, I least often calculate the sample size. More often, I calculate the effect size.

Hilary

That's the right way to go.

Roger

As Karl Broman once famously said, "The sample size is equal to the cost per sample divided into the budget."

Hilary

Yes, when somebody comes to you for a sample size calculation you say, "Well, how much money do you have?"

Roger

On one end, you might say, "Sample size is a function of the budget." But another way to say it is, "Sample size can help you determine whether the question you're talking about is reasonable or totally infeasible."

If the sample size is 100, then it may be feasible, but if it's 1,000, then it's not, at least in my context.

Hilary

That's my biggest qualm with…I know from having worked in stats, this happens all the time where someone comes to you too late, they've already designed their experiment and haven't decided what they are going to do, and then they ask for a sample size calculation at the end, and you have to say, "Great, you can run this experiment for 87 years. You'll see the effect size that you're interested in." And they're coming so late that you feel like a wet blanket.

Roger

It's always horrible to deliver that kind of news because the question that you focused on for weeks is totally infeasible.

I feel like the less-prepared collaborators come to you and say, "Everything's done. We just need you to stamp this number and say that you approve." The better collaborators, rather than saying, "Hey, I need a number," or, "I need a power calculation," or, "I need an effect size," they say, "We need to have a conversation about the science and we need to know what's feasible and what's not." At a high level, you need to know in terms of order-of-magnitude-type of questions, in terms of feasible or not feasible. Then when you know what's feasible, you can figure out, "How can we optimize so that we ask a question or we look at the effect that we can get the most juice of in terms of the budget and how much we can afford to do." Do you have the same experience?

Hilary

I think one thing that's sort of interesting is that, because experiments are such an ideal, I feel like we lose sight of the fact that they are *the* gold standard, (and the best thing you can do if you want to find causation) and that there are alternatives out there. So if something's infeasible, it doesn't mean you can't do *any* science whatsoever. You can still do analysis. There are statistical methods to deal with imperfect setups. I've found that conversation is hard to train people to have. They'll either come to you and say, "We want this number, and if this number doesn't work, then let's throw the whole thing out."

I feel like the conversation that I wish I had more that I don't is, in many companies, you can train engineers and product managers to understand that experiments are important and sample size is an important aspect of doing the right type of experiment. But if the sample size calculation comes out and it's a very low-traffic page or something, and you get to this situation where you can't get something perfect, there are still so many options in-between doing nothing and doing an A/B test or an experiment.

I find it hard to get people to have that conversation at the right moment in time. I think it's sort of the flip side of teaching people to do the perfect thing is that there's sort of less area for gray or in-between, and that's somewhere where, I think, in academic science or statistics is maybe more of an understanding of that.

Roger

I have to think back on that. One thing I've found—and maybe academia is a little unusual in this way because it's supposed to be doing things that are kind of "different"—but I've never had a totally routine power calculation. I feel like every time someone's needed one or I work with someone, there's always been a couple of things that make this problem unique, you know?

Hilary

I think that is this difference I was talking about. In a tech company, if you're changing something on a website and then you decide to change something else a week later, you can follow the exact same procedure. In a tech company, it's easier to make things standardized, which is why I think experiments have sort of flourished in that environment.

Roger

It's easier to control things.

Hilary

It's so much easier. And then traffic is "cheap", or it's easier to get sample size. Samples are very cheap to acquire, usually. It's one of the things where it's, "Why not just throw the gold standard at it. It's relatively straightforward and cheap to get. Experiments are cheap." Whereas in medicine or in academia, in most academic applications, samples are much more expensive to acquire and so there's all these methods to take account for that. I kept running into this problem where if you can't do the perfect thing, people don't understand that there are statistical methods for the imperfect approach.

There's a lot of work being done on causal inference, for example, people at Facebook are looking at that problem right now. That's something where that is almost a niche thing in tech companies and the standard is these perfectly-implemented experiments. It's sort of the flip problem from your research in environmental health—you can't run an experiment on temperatures.

Roger

No, it's definitely not the norm.

Hilary

You run into a different set of problems when you try to apply experiments at scale. It's been interesting to see the flip side, and it's all driven by this cost of samples.

Roger

Now that I think about what you're saying, it's seems a little weird that in academia, I feel like all the time, people are scrapping whatever data they can find because running an experiment on people is just so expensive. So they just kind of gather whatever data they can get, and then you run into all these problems in terms of the analysis and drawing causal conclusions, whereas you guys have all the data coming in from wherever. You're more conditioned to do the controlled experiment in a non-academic environment.

Hilary

That is true. So conditioned to do that that there's not necessarily an understanding that this is this gold standard, perfect way of analyzing causal inference and that there are other options available. That's been super interesting for me.

Roger

It's just a function of money, obviously, but I would have thought that the thinking would have been the other way around.

Hilary

It runs into its own set of problems because there's sort of an all-or-nothing attitude. You either just look at traffic and understand things, or you can do an experiment and understand whether or not it was causal. But then doing something in-between, like propensity score matching or something like that, isn't an option. I feel like when you look at web experiment or web analysis, it's one pole or the other and there's not as much in-between except from these large tech companies.

Roger

I actually kind of have a nerve-wracking power calculation.

Hilary

Do tell.

Roger

I was on one of these multi-center clinical trials and they had an interim analysis. Normally in a clinical trial, you're not allowed to look at the data, or at least the outcome data, until it's over because otherwise, that might bias your implementation of the study. This is a five-year study and in year three or so, there's an interim analysis planned where we would look at the data half way through and determine whether or not the experiment was having an effect. Usually, the idea is that if it's having this dramatic effect, then you would stop the study because then the control group is not getting the treatment, and it's not really ethical.

One of the issues that also can come up is if there's no effect, then you have to determine whether or not continuing the study will allow you to observe an effect. It's conditional power for futility. If you think there's going to be no effect, will gathering more data allow you to see the effect? If gathering the other half of the sample won't allow you to see it, then you can draw this kind of idea that it's futile. Anyways, I feel like I've never had so much riding on one calculation, you know?

Hilary

Yeah.

Roger

It's not like they make the decision based on that one thing that I say, because there's other data that they look at, too. But it was kind of interesting to go through that process for the first time.

Hilary

Were you dusting off the derivation of the different…

Roger

I was like…yeah, “I really better get this calculation right this time, unlike all those other times.”

Hilary

I find power calculations deceptively hard because no one ever talks about what it’s powered to, like you’re detecting an X% change in the effect. There’s that concept of the difference that you’re powered to observe 80% of the time. I find that part of the reason why it’s deceptively hard because the moment you start actually digging into it, if you’re discussing it with someone, there’s all these things that they weren’t expecting to even come up, like judgment calls you made about the false positive and false negative rate.

Roger

In my experience, it’s very difficult for people to say in most cases what a practical effect size is.

Hilary

That’s a very common question that comes up at Etsy, “What effect size should we be seeing? And everyone comes to the statistician thinking, “This is a statistics question.” It’s not.

Another one that comes up a lot is what correlation means “something”, and I’m like, “Trust me, if you’re in physics, correlation of 0.8 would be terrible, but if you’re in social science, that would be amazing.”

Roger

I’ve found that people will usually have some upper or lower bound. If I just say something ridiculous like, “What about a 90% change,” they’ll be like, “Oh, no, no, that’s too big.” I usually try to bracket the interval somewhat, and then if I can get it within the reasonable amount, I’ll just take some middle value, and they’ll be okay. I know it’s hard because for a lot of health-type questions or biological questions, you just don’t know. The human body is just too complicated. So you don’t really know what a meaningful change is going to be. And so you just have to take a stab at it. In that case, as a statistician, I’ll just make up the number. That’s why the effect size is always the easiest thing to calculate because no one ever knows what it is.

Hilary

Especially if someone’s saying, “We have X dollars to spend.” Realistically, we could never have more than 100 people in this study. Then that makes it easy. I will do that here, too. I will say, “Okay, there are certain constraints on the type of experiment we run. I’m going to present you with the maximum detectable effect size that you’re going to see. So manage your expectations accordingly.”

A/B Testing

Roger

The other thing that I thought sounded like it was straightforward to do, but probably actually hard, is A/B testing or what we call clinical trials. I think people understand that clinical trials are hard because you've got to enroll people or whatnot. But A/B testing in a tech company, it seems like it would be easy. You've got data coming in all over the place. You're testing two things. What could be so hard about that?

Hilary

I think there's a sort of attitude that it's very easy. A lot of data scientists will have this attitude that it's not that interesting of a problem. I think it's super interesting. I also come from biostat so that makes sense. But it seems like it's very simple, "Oh, just count on one side versus the other and then you're done." I think the complexity comes from the data-generating process.

There's all sorts of judgment calls you make, like what constitutes traffic, a visit. If you have visits, and if you have multiple people visiting multiple times within one experiment, if they're always in the same variant, then you have correlation within the visits, right? There's things like that that. The web traffic gets so oversimplified before you even see it that the A/B tests always get results. You'll be able to do a simple calculation like a proportions test, but then you won't be able to explain strange patterns in the results. When you start digging, it becomes a horror show.

Roger

When you were talking about that just now, I already started to sweat.

Hilary

The number of times I've said this personally, that the i.i.d. (independent, identically distributed data) assumption for proportions tests is almost always violated, either the independence, or the identically distributed part. It's almost always violated when you're doing A/B testing. Then it's a question of scaling–how much the broken assumptions are affecting it. T-tests can be robust, but it is not as simple as it sounds.

You have to choose when someone shows up what variant to show it. So you have to do that via a pseudo-random process. Even that, I'm sure you're kind of immediately thinking, "That is not perfect," right? So even the way that you're bucketing people isn't perfect, and then those get rolled up into visits, which is an imperfect kind of arbitrary thing. The way visits are defined by Google, that I think a lot of companies use, is it has to be some action on the site. So some event is generated, and the time between the events is less than 30 minutes. So once two events happen that are more than 30 minutes apart, those are 2 separate visits. So it could be someone who got up and had lunch and sat back down and was on the website. It's a very imperfect.

Roger

And 30 minutes, you just made that up, right?

Hilary

Google made it up at some point

Roger

And that's like a 0.05. It's never going to change, right?

Hilary

Exactly. Google is the Fisher because they're the ones that just made this arbitrary decision that has affected the field tremendously. Because the reason why companies will use that definition is because most companies will have started with some version of Google Analytics on their site, and then they're like, "Well, we want it to be apples to apples with this old system so let's just continue to use the same definition, and that makes thing simpler."

Roger

It seems like every company probably starts with Google Analytics, right?

Hilary

Google makes it very easy so a lot of people use it. I had Google Analytics on my academic site, actually, might still. I haven't looked at it in like a really long time. But they make it so easy–just paste in code, and you get it.

Roger

I know. It's a little too easy. And also, a lot of service providers now, it's all integrated. So you just type in your ID and it just goes.

Hilary

When I'm talking about this problem, I always say, "Have we given people just enough rope to hang themselves with." If you are giving people these very simple, digestible statistical testing, and they don't have any idea or they've never been educated about all of these caveats that are going to make this testing so simple, that makes it really hard to untangle when things start to look weird, which invariably will happen when you have these weird, complex data structures that might end up breaking.

I think it's a really cool field. I'm definitely focusing more and more on experiments, and I think it's a fun place to be a statistician because you're thinking about the same problem in a different way with a different set of constraints. But the most frustrating part is that there's this attitude, "Oh, it's just A/B testing. That's easy."

Roger

So one thing…I will just reveal my total ignorance of this area, but I'm going to say it anyway. One thing that you said that kind of like a light bulb went off is when you said when the people come into the site and you have to assign them to a group. And let's say for the sake of argument, there's two groups, right? And I always just figure, "Well, there's some random number generation process going on in the background and you flip a coin and you get heads or tails." But you said, I guess, you hash something that's unique to their visit and then…

Hilary

It becomes deterministic for the visit.

Roger

It’s always deterministic, ultimately, but I guess it occurred to me that you have all these instances and because you have so many people coming at the same time, right? I guess it would be hard to synchronize a random number generator across all these instances, right? I was thinking, “Why do you hash something? Why didn’t you just have some coin-flipping in the background?”

Hilary

Yeah, that’s the issue.

Roger

I’m sure it’s more efficient, too, to take some piece of data from the user and turn it into their assignment.

Hilary

Yeah. I mean, the thing that I have learned from working at a tech company is that whatever a statistician would do, some data engineer has done something much more efficient but much more complicated-sounding to simulate what the statistician says.

Roger

I imagine, in many cases, that’s a function just like the realities of the load and whatever. I think what Christopher Volinsky said that when he came to AT&T, I think Daryl Pregibon told him basically, “Everything that you’ve learned, forget about it.” It’s a version of that.

Evidence-based Data Analysis and Automation

Roger

I recently had this epiphany and tell me if you agree. As a statistician, I teach a lot of classes and things, and we have this habit of giving people a bunch of choices, “You can do this, you can do that, you can do regression, you can use a smoother, you can do whatever.” There are five different models that you can choose or strategies that you could implement.

But we often don’t tell people how you choose between those types of things. It’s almost like we explicitly don’t do it. You look at any statistics class, at least any statistics class that I’ve taught, it’s always “Here is a series of methods and you could use this or you could do that, and they are all kind of good, but there’s no rationale for choosing one over the other.” I guess my thinking around that is the reason is because in many cases we don’t have one, and I feel like that’s a fundamental gap in data analysis. Do you agree with that?

Hilary

I do. Because it’s one of the biggest things I had to learn going into a more applied job. I don’t want to throw the field of statistics under the bus or anything. Good statisticians do this also, but it is just like this idea that something could be theoretically correct, but practically not making a big difference. I feel like when I taught Intro Stats at Hopkins I would definitely emphasize that you can have a statistically significant result, but it's practically not that important. But then we never discuss that beyond hypothesis testing, at least in my intro material, because you could use that rationale for choosing different models. It's like, "Okay, what's the practical difference of this more theoretically correct model?" I feel like that has been something that I would like to adapt to.

Roger

I think we have this aversion to cookbook-ery, which may be a good aversion, but we may have gone too far in the other direction. I think what really hammered this home for me was I was teaching our biostatistical methods class, and I'm teaching the last term of this year-long sequence. It was the second to last lecture and a student came up to me, who had done well in the class, and she's says, "I feel I learned all these different things, but I still don't know what to do."

She was being very honest with me. I appreciated that because I realized that "You know what, you're totally right. I haven't told you what to do." It was the second to last lecture, and so at the last lecture, I tried to come up with "Here is what you do," a list of things, and it was extremely hard. First of all, it went against my natural training, and also it wasn't like I knew the answer to that question.

It's not like no data has ever been analyzed. People are analyzing data every minute of the day, but the larger process of analyzing data, as opposed to the very narrowlyfocused specifics of fitting a regression model is not very well understood.

Hilary

Now that I think about it, there are times when we do accept cookbook-ery, if you will, because I'm thinking about anova testing at work and that's the example where it's, "You first do this. You compare all the groups and then, dependant on the result of that, you then take the second step like doing the pairwise comparison." So it's one of the few times where I thought, "Yep, this is the acceptable workflow," when it's true that it does not need to be the workflow. I think there are good reasons for it, but there are good reasons for other workflows, too.

Roger

And it's not really true, because, for example, if you fit a regression model, the fundamental operation in regression is inverting a matrix. But we don't talk about the 10 different ways to invert a matrix. We just say, "You use the `solve` function or whatever. You just invert that matrix." You just assume that there is some perfect way to do it and then you just do it, and it works 99.9% of the time.

So we don't have to be a numerical analyst, but if you talk to a numerical analyst, they will tell you there are 15 ways to invert that matrix, and they will tell you all the details of all the different ways, and the strengths and weaknesses of all of them. We allow *that* to be automatic, but we are averse to having our own field become automatic.

Hilary

That's exactly it. And I'm totally open to the idea that this might not be the state of the field, but I feel sometimes some applied computer scientists are at that next step that you are talking about. They are just kind of like, "This statistical method is fine and I'm much more interested in the cross-validated error." But then, I naturally say, "You need to think carefully about these different choices!" That's just the training and the mindset you get hammered into you.

Roger

I think it's a fear of being automated out of existence. It's the "robots taking over the world" kind of fear I think, but there is always something. Every time you automate something there is something new that you don't understand.

I think another issue is that the kind of automation that I'm talking about is at a larger scale. For example, how you do deal with missing data and then fit a model? That's a very common task–you're always going to have missing data. The problem with automating that kind of thing is that it's not something that you can very easily analyze with traditional tools like math or whatever, and so you want to understand the properties of doing one procedure to fill in missing data and then fitting a linear model, for example. The properties of that are kind of difficult to understand without either resorting to simulations or something like that. The kind of traditional ways that we study methods don't really work as well when you start looking at the bigger problem, I think.

Hilary

You can create models in order to understand a process versus creating models for prediction. Those each have different sets of validation criteria, but then that can be automated. It's true that the problems themselves can be categorized fairly easily. It is funny because now I'm thinking about this, I guess the one time we violate this is when statisticians teach hypothesis testing. That's the one time we say, "Okay, the data has to be in this format and then we are okay with you doing this analysis framework and automating it."

Roger

I think my last point on this issue is getting the data to the point where you can do the automated thing, it's a huge part of data analysis. There are a lot of important things that happen in that process, and that could have a big impact on your results or your decisions. I think we ignore that part because, from an academic standpoint, it's very difficult to analyze that part of the analysis, which could often be 90% of the work of the analysis. There is a part that we set up the data to do the hypothesis test.

Hilary

It's an art, but it's true. I'll say that to people. If someone has no experience with statistics and they are a little experienced with math, I'll be very quick to point out that this is not math. It's true that statistics theory is math, but applied statistics is very much not math. It's not objective, the way that we look at problems and the choices we make. Those are subjective calls. We can quantify. We can try to do as objective a job, but there isn't a theory around creating a compelling narrative with data.

Team Communication

Roger

If you work on a team, what kinds of tools do you use to communicate and work together? For example, there's tools like Slack and there's HipChat and other things like that. And I'm actually curious, I'm curious to ask you this question. Because I'm wondering what you're used to and what you've used in the past.

Hilary

I feel like this is a topic I could go on literally for hours.

Roger

All right, I'll start the clock.

Hilary

My initial reaction to this is just that this is not a solved problem yet and I don't think many teams have this mastered because there's some level of—this goes back to the language wars. Or not even wars, but just people having different preferences. Getting everyone on the same page with a collaboration tool is almost as hard as those other things, in my opinion. Especially because the work is usually really individual.

What I started doing recently is using GitHub Flow, even when I'm working alone. The idea with that is that you create a new repository for every project. And then, and I really like this, for every to-do that you have, you might make out a list of, "I want to run linear regression, I want to do this visualization," whatever. You make those GitHub issues, and then you go down the line and accomplish each of the problems and close the issue.

Let's say my first issue is run a linear regression on a data set. Then I would create a new branch of the project and add in the code for the linear regression. And then make a pull request to the initial project, to the master branch of the project saying, "Okay, I accomplished the linear regression and here's the code," and then merge that back into master. And then continue doing that for each of the issues.

It's a little overwrought, one might say, and it's definitely a lot of extra work. But the reason I really like that is that, A) I really like the idea of checking off issues and having all the code contained in the pull request, and then B) I think that opens it up to collaboration really easily. So you can imagine you make 50 issues and you're like, "I'm going to work on the linear regression, you're going to work on something else," some other aspect of the project. You could be working on those simultaneously, like do the code review for each other.

I came to this after we had a few big group projects at Etsy that communication was definitely an issue because some people were merging straight to master, some people were doing things locally, and, yeah, it just got really confusing and there wasn't a consistent way of communicating with each other.

So I do think this is the ideal solution, but I also don't know how people…I'm not sure if that's really feasible with the type of work that it is. It's easy to get people on a software project all on the same page with Git commits and a certain GitHub Flow or whatever, but I think it's a little harder when you're talking about data science projects.

Roger

I have one quick question. Just so I understand, this is basically just using GitHub to manage the communication, right? It's not a separate piece of software, just so we're clear.

Hilary

No, no, no. Yeah. And GitHub Flow, it's there. That's something that GitHub—I don't know if they defined it, but they have a lot of material online explaining it much more eloquently than I just did about the exact Flow. If you tweet out about this, people will be like, "Oh, GitHub Flow, we don't use that for this esoteric reason." There's a lot of debate, especially from software developers, on this. I'm sure many of our listeners are software developers who are having reactions to me saying GitHub Flow right now.

But that being said, I feel like it's a good starting point and there's a lot of education materials out there for it. But I never had success—I have to be totally clear that I never even worked with two people on this. Maybe once I had a collaborator, and that was a software engineer who wanted to try R and was able to fork a repository and do some changes and then submit a pull request.

I don't think I've ever successfully collaborated with other data scientists on using GitHub Flow. But it was designed for collaboration, so I can't believe it wouldn't work. You know what I mean?

Roger

Well, it's kind of designed for large-scale distributed projects.

Hilary

Exactly.

Roger

I'm curious then how do you communicate with people then? You just e-mail?

Hilary

Well, no. So you can tag people. Ideally what would happen is that you would create a pull request, "Oh, I added linear regression." And then you can tag your collaborators and say, "@Roger, can you look this over and say that it's okay?" And then you would get an e-mail, or however you set up your GitHub notifications, but you would be notified. And then you could go in and you could write, "Looks great, Hilary. As usual, A+ work."

Roger

Because that's how we communicate.

Hilary

There's not a ton of infrastructure there, but you could have a conversation in line with the code. You'll have a record of that conversation and—so I think that this isn't how I've seen it happen a lot, but I think borrowing on software developing tools is really the only way to solve this sustainably.

Because things like Slack, I'm sort of of the camp that immediate chat stuff is not great. It's hard to archive the conversation and it's distracting and it's hard to make sure that everyone is on the same page. Not that I don't like Slack as a tool, but just for this type of work I feel like having it archived is better. And e-mail, I just hate e-mail for stuff like this.

Roger

I have to say I've come around on e-mail.

Hilary

Really?

Roger

I've gone back and forth so many times now.

Actually I learned something new here, I did not realize that—I had heard of this GitHub Flow thing, but I did not realize how people used it for project coordination.

But I have used, I think, every kind of project management/collaboration tool under the sun. And it is all a nightmare.

Hilary

Yeah.

Roger

So there's Basecamp, there's Slack, there's...I think I've used Asana. It always starts off with a great deal of enthusiasm, and then in 100% of experiences it just dies off.

Hilary

It's true. No, actually for me GitHub Flow has been one of the only things that didn't die off. Because it keeps the to-dos very linked to the project itself.

And it's not extra work because it's just part of the workflow. You know what I mean? It's not a separate tool where you're like, "Okay, I did this and now I have to walk over here and check off that I did it." In my head, as I'm saying this, I know that there's a lot of criticisms of GitHub, too. And again, I'm not saying this specific implementation is the perfect one, but just working with tools that make the in-line collaboration more easy, if that makes sense.

I was going to ask you how you communicate with your students.

Roger

It's a nightmare, I have no system. I basically e-mail.

One of the issues that I have is that often my students and postdocs work on different things. So it would not be logical to have the project that student A is working on and the project that student B is working on in the same Git repository. Those things should be separate. Right?

Hilary

Yeah.

Roger

So I can communicate with student A and I can communicate with student B, but student A and student B don't have a means to communicate with each other. Right? Because they're working on different projects. So unless they're sitting next to each other, which sometimes is the case, they don't really have a means. It's sometimes useful for students to communicate with each other even if they aren't working on the exact same project because there are maybe things in common, like they might have R questions or whatever.

I was actually talking to someone over the weekend actually about how–so this person had like nine students. Which is, in my opinion, insane. But anyway.

He uses Slack and the idea is that the students can kind of talk to each other. Because in that case the students are not sitting all in the same place. And the students can kind of talk to each other if they have questions, and that way they don't have to go to meet with him to ask an R question or a project-related question.

Hilary

That's a great idea. And in terms of kind of the camaraderie that you need–this almost goes back to what we were talking about, was it last week when we were talking about distributed teams versus central teams?

Having students who can't talk to each other would be almost like having people embedded within project teams rather than a centralized team. And again, I think that it's important...if you want people to be able to ask questions about R and figure out—get code review help and things like that, I definitely think that having a shared communications space is key for that.

Roger

And that way, in something like Slack, the little bits can be saved, and so you can kind of refer back. And also, I think it's weird because you're training the students to be independent, meaning separate in some sense, right? But you do want them to talk to each other, of course, in your lab setting. So I think you want both.

Hilary

I don't think those are mutually exclusive. I don't think that communicating is necessarily not being—in fact, that's kind of the key to independence, is being able to figure out things on your own by leaning on a support network.

Roger

I feel like with academic projects, it's not like you're all working to build my software. You need to be able to stake out your own piece. And so there may not—and I think I may be more extreme than others, but I feel like the projects that my students and postdocs have worked on have generally had very little overlap, except for maybe the topic area.

So I don't know. I haven't used Slack. I'm thinking about it though for future purposes. But I use it for other things, but not for working with students.

Hilary

It's just part of this immediacy culture, which is, I don't know. I don't have well-formed thoughts. I know a lot of people have well-formed thoughts on it, but I am not one of those people yet. I have to say, I used to be much more into it—we had our own kind of IRC implementation at Etsy and I used to have it open all the time, and then I started to close it. Because it's just—you get distracted.

Roger

The funny thing about GitHub is that I feel like I kind of…my initial exposure to it was quite some time ago, I think before a lot of these kinds of features were implemented. And so, just like with anything, your picture of a tool is a snapshot of when you first started using it.

So when you say things like this and other people talk about it, I'm always a little bit surprised. I'm like, "Oh, I didn't realize it worked that way."

Hilary

The curse of being the early adopter, the beta tester.

As I said, I kind of am hesitant. I totally recognize that someone who's a software engineer will have so many more nuanced opinions about the various version control and collaboration tools than I have. Because kind of that whole system I just described works really well if the feedback you get is just, “Good work, looks good,” or, “Make this small change.” If it’s like, “No, you need a complete overhaul and half the code is going to change,” I think that makes it much more harder to get the detailed comments and being able to do in-line comments and everything. There are issues like that that could end up being really important for your particular use case.

Whereas, for mine, because I was always just committing…I was doing pull requests to myself. So I was never going to be like, “This is terrible, go back and restart.” And so for mine it was just sort of a way of organizing and documenting for other people, and having other people lightly tuned in.

It really depends on the nature of the collaboration. Nothing is going to replace–at some point it’s just trying to mind meld with the other person. Making sure that you’re both on the same page and thinking about the exact same things. And nothing is going to replace sitting next to each other and talking and having regular meetings. No tool is going to make that happen.

Roger

You can’t replace communication.

Origin Story

Roger

I think we've talked about the whole #rcatladies thing, but I don't think I know the origin story really.

Hilary

I don't think many people know the origin story, but it is all documented on Twitter. It all started on Twitter—not this last useR! but the one before that, so it would have been 2014, the one that was at UCLA.

The back story is that I have this friend Sandy. We actually met when we were applying to grad schools. We did the tour at University of Pennsylvania together. She ended up going there. I ended up not, but then we became Facebook friends, and this was back in earlier days of Facebook, and we just kind of kept in touch over the years and our mutual love of cats emerged in this online friendship. It became clear that we are cat-lady friends. Her name is Sandy Griffith. She works at Flatiron Health, and she has a PhD in biostatistics.

This friendship emerged, and we started to see each other at various statistical conferences. One time she came and visited me after I came to Etsy. Through this, she decided to move to New York and have a similar career to mine, at a startup. Actually, that time she came to visit New York, she convinced me to go to useR!, and then I also feel like I convinced her to move to New York, so it was a productive visit.

So then we went to the useR! conference. Anyone who has been at those conferences—it's mostly men. I think any developer conference is going to be that way, even more so than a statistics conference or an analytics conference.

Roger

Ironic that I've never been to useR!.

Hilary

It's super fun. Those are the most fun conferences and the most productive in terms of just me learning a bunch of new stuff and getting inspired by people implementing things. It was super cool, so I highly recommend.

Sandy and I were both there and we were having fun. But when you're a woman in a situation like that you kind of have two options: you can either try to blend in or try to stand out. You can imagine the path that we decided.

Roger

Which one did you do, yeah?

Hilary

I wonder….

That's a whole different discussion about women and tech issues, but the immediate consequence of this was that there was this guy Romain Francois. He is a developer on `dplyr` amongst other packages, and he gave a presentation. I actually wasn't even there, but he gave his presentation, and one of the slides in his presentation was of him as a toddler holding a kitten. Sandy and I immediately fave'd this tweet within a minute. Then I took a screenshot. Someone took a picture of this slide and tweeted out "Oh, Romain was into cats before it was cool."

Then we both fave'd it and then I took a screen shot of the fact that Sandy and I have fave'd this ridiculous slide and I was like, "Sandy, I'm not surprised that the two people who fave'd this tweet in this conference are you and me," Then I replied, "I'm hereby declaring that we are the #rcatladies." And then it just took off from there.

People were into it. Romain was into it, the guy who sent the tweet, Karthik Ram, he was into it. It just became a fun conference thing where we are like, "Oh, let's play this up." We both genuinely love cats and so we were enthusiastic to tweet photos of cats or whatever. But then it just took on a life of its own after the conference.

Roger

It kind of did, right? Yeah. It went on Twitter and then...

Hilary

It's just genuinely a fun thing. Sandy and I are both bon vivants, if you will, we like just having these jokey things. It's just something that we thought was fun to revive, but then it was also this feeling of "Okay, yeah, I don't feel like I personally totally fit in at an R conference." So it's just a way to say, "Okay, I'm going to make this my own a little bit." Then it was really cool, because then we started to get tweets from—there would be women who would tweet from an R Meetup and be like, "Where are my other #rcatladies? I'm the only woman here."

It was cool because it became this kind of subversive, different culture, compared to the dominant R culture. To this day, people will tweet out a picture of their cat sitting on their laptops while they are coding in R. I mean, you did choose a cat for the cover of one of your books.

Roger

That's right. Well, that was my cat.

Hilary

You should know that #rcatladies is all inclusive. You don't have to be a lady. You don't have to own cats.

Roger

I don't know if I am included, but I feel included.

Hilary

You're definitely included. There have also been some spinoff groups. There's like the #RDogFellas…

Roger

I did see that, yes.

Hilary

And then there is like RPoodlePeople. One of my friends tweeted that out, so there's whole spinoff groups.

Roger

That I have not seen.

Hilary

It is just fun. It's nice to have things that are fun aren't traditional programming stuff.

Not that there is anything wrong with the traditional programming stuff. I think being at that conference…I mean, this is not to say that anyone at that conference wasn't welcoming. Everyone was happy that we were there, but no matter what, the dominant culture comes out. So it was just fun to have little group of people who are just doing something different.

Base Graphics vs. ggplot2

Roger

Jeff Leek about a month or so ago wrote a post on Simply Statistics titled, “Why I don’t use ggplot”, and he talked about a bunch of reasons why he uses base plots only.

Hilary

I’m going to add that that was in response to our podcast episode where we made some comment about him not using ggplot.

Roger

I saw your tweet about that and I totally forgot.

Hilary

How can you forget? This is our shining moment. I meant it when I said, “This is how Rush Limbaugh started.”

Roger

I was going to ask you what that tweet was about. Now it totally makes sense.

Hilary

We made some comment when Jenny was here, saying, oh, ha-ha, or you could use Base Graphics like Jeff, and so then he wrote this. He was like, "All right, you all have been teasing me long enough."

Roger

David Robinson wrote a response post in his blog explaining why ggplot is better than Base plot. Is that an accurate way to summarize that?

Hilary

Let's just say this post was probably on the backburner for quite some time. He needed Jeff to state his case so that he could respond.

Roger

And then only a couple of days ago, or maybe a week ago, Nathan Yau of FlowingData wrote a post comparing ggplot and Base graphics. And I guess he also said that he mostly uses Base graphics for the things that he does. And then chaos ensued, right?

Hilary

Another thing that happened after that was that Hadley finally responded. He'd been chiming in, but he finally was like, "If you want to draw pictures, base graphics is better, but most people don't want to draw pictures. They want to visualize quickly. And I think that's a fair point.

Roger

I just wanted to add that Ben Casselman, who's the chief economics editor for FiveThirtyEight went on a little tweet storm about ggplot also, and mentioned both Nathan's and Jeff's post too.

Hilary

His was really good, too.

Roger

So I have a lot to say about this. Maybe I won't say it all, but I think one of the things I thought was interesting about Nathan's post is that depends on how you make the comparison.

If you just take any given plot, let's say a box plot here, a base plot or a box plot with ggplot are for the most part going to be kind of the same.

Hilary

I totally agree. The line by line code comparison is not where the strength is. ggplot isn't trying to be efficient and use as few lines as possible.

Roger

Exactly. So for any given thing, there's probably not much to compare. It's interesting because I hesitate to say this, but I'll say it anyway. I think it's very similar to the conversation that you and I had about how to convince someone that method A is better than method B if method A and B give the same answer.

Hilary

hat's a great point.

Roger

It's one of those things where ggplot and base plot can produce the same plot, and for this given application they're basically the same, so how do you convince someone that one is better than the other?

And I think the answer is you can't really. Just like if the mean and the median produce the same number, it's impossible to convince someone that maybe the median's going to be better than the mean or vice versa, right?

Hilary

I hadn't even thought about that, but that's a perfect analogy.

Roger

The argument for ggplot in many ways is what *could have happened*, you know? If you're going to be doing exploratory analysis with a bunch of plots, and you need to make a bunch of plots, then the kind of abstraction and the structure of ggplots allows you to do that faster. Wouldn't you say that's roughly the argument?

Hilary

Absolutely.

Roger

This is what I hesitate to say, but I'll say it anyway. It's kind of like the Frequentists versus Bayesians argument actually. I'm totally serious.

Hilary

Do go on.

Roger

I'm so glad you humor me. Because if you think the Bayesian angle minus the prior is really focused on a given data set - it's conditional on the data. Integrated with a prior, what do you believe?

That kind of philosophy is not really concerned with what might have happened, whereas obviously the frequentist approach is, across many different replications what could have happened? That's the hard part about confidence intervals. You have to think about what might have happened across many different replications.

And so the ggplot argument is very much along the lines of, yes, we're given plots. It is what it is, but if you think about what might have happened across many plots, it's a much more powerful tool, whereas with base plots, if you're making all kinds of different plots, then you're struggling with code and you're hacking code all the time, right? But if you're focused on making just the one thing, then it's harder to make that comparison.

Hilary

I think that was a great analogy. Good work.

Roger

I mean, it's probably crazy. I'm going to give it that.

Hilary

I thought Nathan Yau's post was really interesting because he did not realize what he was treading into. He had a follow-up that said, "I just want to use plot. Leave me alone." And I was like, "Oh." That was when I realized this is how it started, this is Rush Limbaugh.

But he is obviously brilliant. He has something going on in his head that most people don't. He's able to hone in on exactly the visualization that will matter. He even tweeted a picture about how to create a graphic and he said, "Oh, first step is drawing it." So he's thinking about the whole problem in such a different way than what ggplot is optimized for. It's generalizing across a different use case completely.

Roger

I hate to characterize what he does, but I feel like a lot of what Nathan Yau does is he takes a data set and an idea about how to visualize it and really goes to town.

If you're focused on one problem like that and what the best way to visualize it is, then it almost doesn't matter what the tool is. You're going to optimize it until you get the best thing that you can.

Whereas if you're doing lots of different things, maybe not across different data sets, but if you're looking at many different angles of the data set, then that's a different use case.

Hilary

ggplot just lets you bang through fifty different visualizations very, very fast. The thing I could never get away from is that having the theoretical framework is just so powerful, because it allows you to not even think of what the code is. You're just like, oh, I want to add a histogram and it's + geom_hist to get our histogram. It's so easy - the code and the thought process mirror each other so well that you don't even have to look up code. You can just do it.

When you're so honed in on one data set, you're not thinking about the whole process. And again, for that very specific use case of someone wanting to look at the data in a bunch of different ways from a bunch of different angles before deciding what to do permanently.

Roger

Right. Anyway, that was my take on that.

Hilary

That's a great take and I think I agree with it. My take was a little bit more psychoanalytical if you will.

Roger

Really?

Hilary

People think about problems so differently. And I am the type of person that whenever I was taking math classes as I kid, I really liked seeing proofs. I really liked having the theory. I almost always am deriving whatever is actually going on from the theory in order to remind myself what to do. Even taking the mean, I'll be like, oh yeah, I want to take the sum and then I want to divide by the number. And I re-derive that every time.

And I know some of my coworkers just don't think that way, and they're very productive. And I have no idea how their brains work, but they're able to hack things together and remember things better in that way. And so the people who don't see the advantages of ggplot maybe are also the people that don't necessarily think the way I do. Does that make sense?

Roger

Are you alleging that there are people that don't think like you?

Hilary

Let me be clear that I'm not saying that the people who don't care about ggplot don't care about theory. The amount of efficiency I gained from the theoretical framework and ggplot… I just can't believe that it was as big for the people who think, "I don't see a difference." You know what I mean?

Roger

I agree with you in the sense that people approach problems differently. I am probably more on the side of less abstraction and you're on the side of slightly more abstraction?

Hilary

I need to have something simple to remember and then I'll get to the complicated part. But I can't just remember the complicated part.

Roger

I totally understand that. I think personally, I focus on the complicated part. I do appreciate having a simple overarching idea or framework to remember. I guess I don't always gravitate towards that.

Hilary

Well, I also think you've been pretty middle of the road with the ggplot. You're not so passionate about it.

Roger

I can't say I'm passionate about either one.

Hilary

So this reflects exactly. You're saying, yeah, I see it, and it helps, but whatever, whereas I just think back to plots I made.

In fact, this was in my draft blog post response to this, where my first line was literally going to be, "I have thought really hard about how to build this bike shed, so let me tell you exactly."

Thinking back to code and this blog post I wrote back in grad school about the name Hilary, I can make that plot…It took me hours. I was already pretty good at R at that point but it still took me hours to draw those plots, whereas now I can do it in my head in three seconds using ggplot.

I feel like the people who don't feel as strongly may not have that efficiency gain.

Roger

Or they may be the kinds of people who sit and think for a long time about every plot. And you can say what you want about whether that's good or bad, but I think some people, and I don't think I'm necessarily one of these people, just take more time on each plot, so the efficiency gain is not as obvious.

Hilary

That makes perfect sense. I think the field dictates that to some degree. In industry, you're making so many line charts over time and ggplot just makes those so easy. And you want to do some smoothing to get rid of the seasonality of the data. And so it's just a no-brainer, right?

Roger

I think the pace can often determine what you can and can't use, or what you can and can't do.

Do you have anything else to say about this controversial issue?

Hilary

Only an amusing anecdote. I brought up that I had been at a wedding this summer where the bride was a statistician. I get really upset if I feel like I've upset someone, so I thought I had upset her because she was anti-ggplot and I was like, "I love it."

So I saw her just a few days ago and she was charmed by the episode, so that's why I feel ready to talk about it.

Roger

She forgave you for ruining her wedding?

Hilary

No. She was thankful that it had happened. She was like, "I wouldn't have had it any other way."

Roger

Oh, well that's great to hear.

Hilary

I feel like the ggplot/Base R debate is actually canonical bike-shedding. Everyone has thought a lot about it, and it's fun to talk about. No one gets that upset, so it is what it is.

Roger

I think it’s more harmless than the Bayesian/Frequentist debate.

Hilary

That one gets nasty.

Roger

For some reason it gets very nasty.

Free Advertising

Hilary

Did you want to do free advertising?

Roger

I have a good one.

Hilary

You should definitely go first, then.

Roger

This is not cat-related, unfortunately.

So, I'm not a vegan, and never have been, but one of my favorite food writers is this guy J. Kenji Lopez-Alt. Have you heard of him?

Hilary

No.

Roger

So he used to work for this publication called Cook's Illustrated, which I used to subscribe to, and it's one of the first, if not the first, science-based cooking publications.

Now everyone does that, so he works for a website called Serious Eats, and he writes a food column, and he has a new book out called The Food Lab.

One of the cool things he does is every year is he takes an entire month, and he just eats vegan. And then every recipe that he publishes for that month is a vegan recipe. I've come across an amazing number of really good ideas through this.

Hilary

So you're starting to eat vegan more often?

Roger

Not really, but I guess if, by doing it at all, is more often, then yes. But anyway, I just think it's a cool set of recipes, and I've made quite a few of them.

Hilary

That's awesome.

Roger

The other thing I'll just say is the recipes are all vegan, but almost none of them are healthy. So that should be made clear.

Hilary

It's funny, two of my closest friends, neither of whom, they're not in overlapping Venn diagram of my friends—they're from completely different friend groups, let's put it that way—and they're both vegan. So I've eaten a ton of vegan food. It's really good done right.

He's done it for, I think, for three years so you can go to the back, the archive, and check out the older recipes.

Hilary

That's awesome. I'll definitely do that.

Roger

Anyway, that's my thing.

Hilary

Cool. All right, I have a thing, too.

At one point, there was a hashtag on Twitter that was called #fieldworkfail, and it was a bunch of biologists—the type of biologist that goes out into the field and collects samples of whatever.

Roger

Yeah, like real work.

Hilary

Yeah, real scientists. There was some sort of Twitter story at some point where people were all tweeting, you know, “Here are bad things that happened.”

Some artist has started illustrating some of the tweets. And so there’s just these cute cartoons that are illustrations of–for example, one was “Accidentally glued myself to a crocodile while attaching a radio transmitter.” Then there’s this cartoon of a woman with her hands glued to a crocodile.

And they’re really good. It’s like a professional cartoonist.

Roger

Is it, the cartoonist decided to just do it?

Hilary

Yeah, like, I have no idea why, but it’s pretty cool. I’m trying to find… Jim Jourdane. Cartoonist illustrator living in France. I thought it was very funny.

Roger

All right, I guess you just have to search on Twitter for these, right?

Hilary

I did not see the original Twitter storm, although if I had been paying attention, because I follow a lot of real scientists on Twitter, but yeah, someone just retweeted it. It was a photo of one of the cartoons, so I was like, “That’s funny,” And then realized that it was brilliant, and that it was a lot of work that someone was putting into it, so I was pretty impressed.

Roger

When you said “real scientists,” did you mean to distinguish them from data scientists?

Hilary

I actually meant to distinguish them from statisticians. But I actually kind of feel bad bashing data scientists, because data scientists are at least trying to do science in a field where most people don’t do science, whereas statisticians are…

They’re the ones not doing the field work, like crunching the numbers for the real scientists. So it’s funny how I feel differently on that. I’m always trying to bash statisticians who I feel camaraderie with, of course.

Roger

Okay, well I feel much better about that.

Hilary

To be clear, I am bashing you.

Roger

Understood. That’s going to be the topic of another episode.

Hilary

I mean, you just sit there and crunch environmental data.

Roger

I know.

Hilary

There's someone out there, measuring temperatures.

Roger

There are whole days that go by where I don't even go outside. It's a tough job, downloading all that data.

I feel like those people who are out in the swamps, collecting water samples or whatever, they're not going to get scooped, right?

Hilary

I think they worry about that.

Roger

I'm sure they do.

Hilary

It's a totally different. I feel like the type of person who is drawn to that work is very, very different than me, and I still consider myself a scientist, but yeah, I've no interest in sleeping in a tent, collecting samples, getting stung by mosquitoes, all of that.

Roger

Actually, for people who do malaria work, I was talking to some guy who does field work in malaria. And basically, the way that they study these mosquitoes is you just have to let them bite you, and then they collect the mosquitoes.

Hilary

Oh, because they die after they…right? Don't they?

Roger

Well, you have to get them before they die. I don't know how they preserve them. So the basically stick their legs out in the middle of the night, roll up their pants, and then shine lights on their legs, and when they see the mosquito land, they suck them up in a straw, and they spit them out in a cup. Because otherwise, you can't really collect them. You can't kill them, I think is what it comes down to.

Hilary

Are they immunized…can you be fully immunized against malaria?

Roger

No, well, so there's prophylaxis, you can take drugs before you go into a malaria area.

Hilary

I see.

Roger

But they're only so much effective. He was telling me that, basically, if you do research in the area, basically, everyone gets malaria in this area.

Hilary

What?

Roger

Yeah. You just have to get it, basically.

Hilary

Never going to be a field scientist. We even ate lunch outside the other day on the grass, and I was like, "Ugh."

About the Authors

Roger D. Peng is a Professor of Biostatistics at the Johns Hopkins Bloomberg School of Public Health. He is also a co-founder of the Johns Hopkins Data Science Specialization, the Simply Statistics blog where he writes about statistics for the general public, and the *Not So Standard Deviations* podcast. He is the recipient of the 2016 Mortimer Spiegelman Award from the American Public Health Association, which honors a statistician who has made outstanding contributions to health statistics. Roger can be found on Twitter and GitHub at @rdpeng.

Hilary Parker is a Data Scientist at Stitch Fix and co-founder of the *Not So Standard Deviations* podcast. She focuses on R, experimentation, and rigorous analysis development methods such as reproducibility. Formerly a Senior Data Analyst at Etsy, she received a PhD in Biostatistics from the Johns Hopkins Bloomberg School of Public Health. Hilary can be found on Twitter at @hspter.