Get with the Program(ming): Accessible Data Science, with Dr. Mine Cetinkaya-Rundel(S02E03)

Not Another Science Podcast

March 17th, 2021

Introduction

[Helena Cornu] I find that I have a tendency to swallow my words, especially when I say “Not Another Science Podcast”

[Tom] Welcome back to *Not Another Science Podcast?!*, I’m Tom

[Helena] And I’m Helena.

[Tom] For episode four of the series we are speaking to Dr Mine Çetinkaya-Rundel, a senior lecturer in statistics and data science in the school of maths here at the University of Edinburgh. Mine is a huge advocate for accessibility when it comes to data and statistics, and uses the latest research on teaching combined with open-source learning techniques to improve student outcomes and boost the representation of women and other minority groups.

[Helena] To this end, Mine has created an open-source data science course for both teachers and students, called Data Science in a Box, co-authored three free online introductory statistics textbooks, and she also runs the massively popular Statistics with R course on Coursera. We spoke about all of these things and about what the future holds for teaching and learning in the aftermath of the COVID-19 pandemic. It was a really cool conversation, so let's jump in!

[Helena] Before we start, this podcast is sponsored by Greiner Bio-One, supplying laboratory, diagnostic and medical products to research institutions, higher education, the NHS and others across the UK. For details of the full product range, visit www.gbo.com.
And now, on with the show!

Cue the bongos

Main

[Mine Cetinkaya-Rundel] My name is Mine Cetinkaya-Rundel. I am a senior lecturer in the School of Maths and Statistics and Data Science. My primary focus has been teaching introductory data science.

I was at Duke University previously and that's a course that I developed there that got quite popular and I, you know, I can only claim so much credit for that, because I think it's that the field has gotten quite popular and the tooling for doing data science has gotten a lot more user friendly, which allows us to bring that content more easily into the introductory classroom.

I also have an industry affiliation with R Studio, which is a company that builds a software through which you can run R, Python and a bunch of other stuff as well, though I work mostly on the R side of things. But I'm on the education team there, so my efforts there are also very kind of similar to my academic efforts. So the idea is getting R to more people and making it easier to learn and develop resources.

[Helena] Yeah, I definitely feel like data science is one of those buzzwords that you hear more and more at the moment. How would you define data science for someone who doesn't know anything about that?

[Mine] It is doing stuff with data at its core. I feel like I've used the term ‘extracting meaning from data’, but sometimes there's no meaning to extract, so sometimes you're just doing stuff with it until you give up and say, well, maybe there wasn't a signal here necessarily, maybe it was all noise. But it's with the aim of extracting some sort of insight from the data.

What that often entails is having a feel for the right type of question to ask, either stumbling upon the right data source for answering that question or building a system by which you collect that data. If you're practicing data scientist, a lot of what you do is designing systems by which you can collect the type of data that will help you answer the type of question that you want to answer.

And then oftentimes cleaning that data. I think however well you build that system, there is always things that you want to clean and tidy up. And then there's the aspect that feels a bit more like traditional statistics, which is more modeling the data and actually making some insights based off of that.

[Tom] So kind of the data science and the statistics go hand in hand. You kind of need both if you're if you're looking to draw some meaning.

[Mine] Absolutely you need both. I think you need both at the beginning stage where you're designing a process by which to collect data. You should have a good sense of how sampling works. What's a representative sample? How do you design an experiment to collect some data?

So those are very foundational statistical topics and then also once the data has been processed and then shaped to be modeled, then the models that you're using, you know, have their roots in statistics theory oftentimes. I think how data science perhaps is a bit more different is this additional focus on wrangling and tidying data. At least from a curricular perspective, these used to be the things we didn't use to teach students much, assuming they'll figure it out as needed, but actually knowing how to do that more efficiently and in a more principled way, in a reproducible way, is something incredibly worthy of teaching.

And then also the focus on communication of it. I mean, I think there's presenting results as you found them, and there's really thinking critically about who is my audience and how do I best present the results for them.

[Tom] I was just going to ask how you actually ended up working in this area in the first place, you know?

[Mine] I studied actuarial science as an undergraduate, I think because my mum thought that would be interesting. I really have like no real reason for going into that initially.

Helena laughs

[Mine] I've always liked math, but I was a little lost in college to be honest. And then I worked as an actuarial consultant, and part of that work I really enjoyed, which was actually working a lot with data like there was a huge aspect of data cleaning that we did. One of my big focuses right now is reproducibility, and thinking back I was doing a lot of like implementing reproducible workflows to our work as best as I can with the limited tooling that I knew about. So it turns out that was an interest that developed during that, but as an actuary, you also take exams. And I used to find them a little bit stressful to have to study for after work. And one day I decided if I have to study anyway, I wonder if I should study a little more generally, get a degree that could be more applicable than just in one area. And then decide what I want to do after that.

So I applied for some PhD programs in mostly statistics 'cause I really like that aspect of working with data and then it was during my PhD that you know, I more realized I enjoy this field, but I particularly enjoy teaching.

[Helena] So you created an introductory Open Access data science course called Data Science in the Box. Was this is an area that you thought was lacking? Was there a gap that you identified? What was your reason for creating this specific course?

[Mine] I have always put my course materials out there openly with like an open license and I put them on a public webpage and so other people could find them.

But the thing is, sometimes it's… You know when you're teaching a course, it's hard to disentangle the core specific logistics from the actual content because you shouldn't disentangle them. For the students you're teaching, those should all go in Harmony, but for somebody else looking to see "How is this course being taught? Can I borrow materials from them?" it's helpful for them to actually see a version of it without the semester specific details or the institution specific details.

A lot of people had told me they were using my materials, but I knew that I wasn't making it super easy for them to do so beyond just posting things as is. And then on the other hand, I personally love learning from others who are teaching similar material, whether it be in, you know, conference talks or like a webinar they give or just a happenstance they say: "Oh, I use this dataset for teaching this particular topic," but oftentimes you get these anecdotes, or you're not sure What did the students learn prior to that? What was the prerequisite knowledge? What did they do right after that? So you learn about this one like interesting example, and then, you know, the day comes that you're going to teach that same example, and it's like 11:00 PM and you're thinking: “I heard someone say something cool, I'll start exploring and get ready for my 9:00 AM class,” and you realize you don't have enough information to go off of. So this was also meant as… for instructors who might already have their own curriculum, and want to just borrow maybe one or two activities or exercises or datasets so they can see in a full curriculum where they sit. So they can have a good sense of prereqs and what comes after.

And then I think the third aspect of it is to be helpful to educators who are tasked with building a new curriculum who might have nothing to start with. Probably have great ideas, but maybe nothing written down to start with.So if they want to kind of adopt it as a whole, it's there. I know that a small number of people have done that, were adopting the whole thing, and by the time they're teaching the course for the second time, obviously they're tweaking things to, you know, make them work, better for themselves. And a large number of people use bits and pieces, and that's kind of basically what I was thinking would happen, but I'm happy to see that it's helping people on both ends.

[Tom] You've mentioned previously that there’s people who are taking the course that you were kind of surprised, that there’s PhD students or people like that, who maybe should have already had this training? Why is it important to teach students about these things kind of earlier?

[Mine] There are two reasons why it's important to teach and teach this material, or at least one of them is. Some of the computational tooling we teach in this course is actually really not easy to learn. The earlier you're exposed, the more time you have, you know, to digest this material and grow into your own workflow, which may or may not be what we taught in the class, but one that you feel comfortable with.

The other is traditionally, as I said in especially statistics, it's been that teaching the computing part and also the nitty gritty of things like getting data in, storing a clean version of your data, or doing these reproducibly writing code that's legible, these are not things that have gotten a lot of air time in class. They were all things we expected from students in a way, but we just assumed they should figure that out and we'll teach them the important parts, the theory. I think they're equally important, so this is not to say the theory isn’t important, but in any statistics curriculum, the theory classes are already there. You know, maybe they should be revamped, but there there. There's no like airtime lacking for them. But there was airtime lacking for the: “Let's work with data. Let's take a look at the data and try to clean it and not just do it haphazardly, but do it in a principled way.” Because each of us, you know, anyone who identifies as a statistician, probably has their own workflow for what they do when a collaborator emails them a data set, there's probably a particular set of steps they go through that they have settled into. It's just some of those steps people have spent painstaking number of hours figuring out, probably have made lots of mistakes that they might not be willing to admit, and the goal is to try to reduce that friction, to make it easier for them to settle into their personal workflows and also make them aware of standards out there.

[Tom] I guess, it sounds a bit like it's kind of getting the students to just do it, and 'cause that's such a good way of just learning is going in and making those mistakes yourselves.

[Mine] Yeah, yeah exactly.

[Helena] Is that… So from your experience of teaching, what are some of the key things that you need in a course for it to be effective and engaging for students? And actually, does that differ between in person teaching an online teaching?

[Mine] Attention span is always a hard thing, and sometimes I hear of this mentioned, as you know, this generation can't focus on anything, but I don't think that's true. I also can't focus on a 50 minute conference talk like the whole time, unless there's something that actually grabs my attention.

Making it engaging for the students, there are two aspects to it. One of them is using examples that they can relate to, which I think is the hardest part of this job because as I age, I don't know really what people relate to. So that becomes harder and harder, but I think using examples that is interesting to them is one thing to grab their attention and the other one is, you know, not just lecturing at them for a long time, but instead getting them involved in doing some sort of activity. And students, rightfully so, don't like it when that feels like busywork. It should be something meaningful. So trying to build a class in a way where you stop for a little bit, have them work on something that maybe kind of unlocks the next topic for them.

And the way we kind of try to address this with the shift to online teaching is by making videos that are shorter in length. The other thing in terms of both in person and online teaching, especially teaching computing, I think, is it's very useful to do live coding in front of students because it kind of slows you down. I feel like if you're solving like a proof, nobody just wants to see the whole proof, they want you to talk through how you work through it. So to me, live coding is very similar and you inevitably make mistakes. And then when you make mistakes, you can narrate how you recover from them. It both, I feel like, makes the instructor human as opposed to like someone who knows it all, like gets it right the first time they try, 'cause that's not possible. And then also the narration of how you recover from a mistake, I think is equally valuable as just them seeing what the right answer is, so they learn something about the process.

So one of the things we did with the shift to online teaching since we were doing asynchronous videos, is I added these live coding sessions that were optional for the students, but many of them attended. And I would pick a data set. There's this wonderful project called Tidy Tuesday. So each Monday there's a dataset released and folks work on it for doing data visualizations or summarizations and tweet about them on Tuesday. So that's why it's called tidy Tuesday. But we would usually pick one of those datasets for that week, and then work through it and I asked the students to kind of put in the Zoom chat for me what they wanted me to do. And at some point it turned into a game where they tried to find things they thought I may not know how to, so I would struggle.

Helena laughs

But it seemed like for those students who joined, they were really enjoying that because it was almost like a game. In that particular instance, I felt like it was actually a lot easier to keep them engaged because it was online. Because if it was in person, maybe the two people in the front row would ever talk to me, I think. But when it's online, it's a lot easier to just chat and be like: “Do this! Do that,” where I think half those students would probably not speak up in class.

[Tom] Are there other other positives that have come from online teaching that you think will kind of stick around after this?

[Mine] I do think that attendance in office hours has increased. And this is already something I knew because like here at the University I teach mostly year one students, but my office is in Kings building so no one wants to come out there. So when you hold office hours, they don’t want to come. So even last year I was holding Zoom office hours because they didn’t want to take a bus to go see me to ask one question on the homework, which I understand.

So I think that's another thing that might add flexibility to our lives, like it is nice to do certain things in person for sure, but if a group of people want to ask one-off homework questions, I don't think that should add like 20 minutes of travel time to their day.

I really hope video making doesn't stick 'cause it like just drains my soul.

They laugh

But there's no reason why anything that happened in class can't be recorded in this day and age. There may be students who can't make it that day, there may be students choosing to join remotely, but from a production perspective it is so much more time consuming to get those videos right.

If I had to teach again, I’d be like: “Well, maybe I don't need to update that example,” but I am fairly certain if I am teaching in class, I would update the example.

[Tom] And I suppose as well, that's not ideal from a teacher’s standpoint, but I guess as well from a learners standpoint, you kind of want to feel like the teacher is as engaged as you are, and you want to feel kind of that validation, I suppose.

[Mine] Yeah! In class there's a lot of nodding and eye contact like these are important visual cues for someone to be like: “Did what I say just makes sense?” One of the things that make marking worth it is the student interaction. If you take that away from me, I don't know how you do the really difficult parts of your job.

[Helena] Nodders are the best people. When you give a presentation and you see someone just go. Yeah, it just. Yeah, thank you to those people. *She laughs*

[Mine] Yeah, absolutely absolutely.

[Tom] I'm a sympathy nodder, so even if I don't understand it, just because I want to make the person feel OK, I'm like yes.

[Helena] For me, that's been one of the hard things about presenting or talking to a group virtually it's just most people tend to have their cameras off, or their cameras are really tiny, or you can't see them because of the way that you're presenting whatever, and so it's just sort of you in isolation and you don't, yeah, you don't get that social response which would normally be a huge part of presenting in front of people. It’s less stressful, because then I don't see them.

[Tom] You don't see the confusion.

[Helena] *She laughs.* I can just imagine that everyone's perfectly understanding

One of the things that I saw that you worked on as well which I thought was incredible, was using active learning techniques to try and increase the retention of women and minorities and in STEM subjects. Could you tell us a bit about that?

[Mine] This was both kind of a practical approach and also a research project I worked on. The overall goal was to look at what are pedagogical approaches that might help kind of retention of underrepresented minorities in classrooms. And when I say classrooms were mostly focusing on large introductory classrooms where it might not be as straightforward as in a smaller classroom setting for everyone to be taking part equally. One of the experiments we did was using clickers, which are these gadgets that you can kind of respond with, but they're anonymous. It's not like when you ask students a question and you expect them to raise their hand to participate, because then it becomes very visible to everybody else. So we try to do this experiment with these teaching sessions that happened outside of the classroom. So we had a volunteer group of students who, in one of the sessions we did a pure lecture and then in the other one, we did these clickers and then we asked these self efficacy questions that were kind of validated measures from education literature. And we did indeed see, you know, higher average values reported in terms of self efficacy. Which doesn't necessarily mean there is definitely a causal pathway for retention, but there is literature that suggests that if those self efficacy feelings are higher, that that is associated with retention.

[Tom] Could you explain just quickly what self efficacy means? Is it kind of the feeling of, that you belong in that space?

[Mine] I would say it's closest to self confidence, but not just like generic confidence per se, but the feeling of “I can achieve this.” So there are various aspects of it where you can think about either your background coming into it or for example vicarious learning, you see others who look like you who are doing the teaching so you can see yourself in their situation.

So there are like various aspects of it and then there are kind of validated questions from literature, that we used as part of the survey. We simplified them a little bit, but we used this part of the survey at the end of these experimental sessions.

There were a couple of reasons why we worked on this project. One was this thinking about how can we build kind of educational experiments. It was myself and a colleague who was an economist who were working on this, and neither of us had really worked on experimental stuff for education. And we worked on this project with a group of undergraduate students. So they were involved with designing the experiment, running the experimental sessions, collecting data, cleaning it, analyzing it and stuff. So we were able to engage undergraduates in the whole kind of spectrum of this research project, which I thought was like a side added bonus to working on it.

[Helena] What are some of the main things that you've identified that are really important for increasing retention rates?

[Mine] In terms of increasing self efficacy, one of the things that we did identify was this kind of opportunity to participate. Many educators do like involving students in their teaching sessions by asking them questions like: what do you think? But usually what happens, especially in large courses, is over the course of a semester, a handful of students will emerge to be the people that give answers to like every question and the rest of the class. And the rest of the class might be engaged, but they're not necessarily actively engaged, and so tools like either clickers or it could be some other pedagogical intervention, like: “OK, I'm going to stop lecturing now. You work on a mini exercise with people around you.” So where people are genuinely actively doing stuff creates an opportunity for them to participate.

[Tom] I was actually wondering because you run the course on Coursera, that kind of massive data science course. How do you keep people engaged in things like that? 'cause obviously there's less of that direct feedback.

[Mine] Yeah, that that is a good question, and that's a hard question. The Coursera course is, in a way, a very traditional online course in the sense that the videos have been pre-recorded and they don't adapt to students learning the videos themselves, we try to make them engaging in the sense that like there's some visual stuff happening in every single video that might keep the audience's attention, but it in itself does not necessarily promote like active engagement with the material, except each of the videos have these like questions. So just as I would have clicker questions, it actually pauses the video and makes you answer a question. When I try to take Coursera courses I really like those things, 'cause you know if I was kind of dozing off, I can refocus and kind of continue on.

But the other thing is, each week's material is accompanied by a computational lab that they have to do something in R applying that material. So that's I think really where students start putting into practice what they have learned. And that's the bit that tends to generate lots of online conversation in the discussion forums. And a benefit of having had that course around for years is that now there's a large community of folks who have taken that course, so the discussion forum is quite self sustaining in that way, and so there's always lots of chatter, an activity there and folks are helping each other out, which has been really, really nice to see.

[Tom] When you mentioned the big kind of community that had grown around the course, it reminded me of the work you're doing with R ladies and kind of some of the cool workshops and talks that you guys put on. And that's kind of a similar thing, I guess? You're creating this community where people feel kind of safe and welcome to come and practice data science and statistics.

[Mine] Yeah, R Ladies is a fantastic, fantastic initiative and a huge organization now and I have like a slightest, tiniest bit of role, I feel like, in it. But folks have kind of, you know, founded R ladies and now is an international organization and there are many people who are kind of at the leadership of R ladies as well as at various levels of volunteering that really make this engine work.

There's a really active Slack community with very kind of welcoming threads for asking R questions, but then also career questions and whatnot. And the goal is always to create, yeah, this safe and welcoming space. And there are many, many R ladies chapters around the world. So I was involved with one of the organisers for R ladies RTP when I was living in North Carolina. And here in Edinburgh, I'm a Co-organizer with Karima Rivera who's a PhD student at the University. And we try to kind of hold these monthly talks, but beyond that we've also done things like, you know, Tidy

Tuesday weeks for example. So one of the members {...} ran these for a while where people would just like get together on Zoom and work on one of these Tidy Tuesday challenges.

And the idea is to just provide a space for people to talk about things, and I think it has benefits for the audience, obviously, whose getting to hear about these… either learn something new or hear about somebody's interesting project, but then also provide an opportunity for the speakers to have a venue to talk about what they want to talk about or what they want to teach.

So, for example, recently we've had a few PhD students from Stats come and talk about projects they've been working on, and it's been so lovely to hear, and I have to say, as somebody who moved here kind of recently, it was so nice to have that to plug into right away and whether I was an organizer or not is kind of beside the point. I think it's just like having something like that, so I often see on the R Ladies Slack, you know, people get into a PhD program or like move from one country to another and they will like go into that channel and be like: "Hey I'm moving here, like do we have in R ladies chapter there?" And I think that's such a nice thing, to be able to be plugged into.

And beyond that, the R ladies organisation, there's a lot of work for just getting women and underrepresented minority voices out there for other R events happening. Like one of the nice things that's happening right now is a group of folks are providing abstract review for a Big R conference coming up. So I think the initiatives are really endless and each of them tend to be quite effective and are really nice for community building. So I'm very happy to be a part of it and I'm really, really glad that this is really taken off and people are enjoying it. I certainly am.

[Helena] Is there anything that you would like to plug if you'd like to advertise something that you're working on, maybe?

[Mine] We have this Data Fest competition coming out, which is like a weekend long undergraduate competition. It's opened to all undergraduate students from University of Edinburgh and Harriet Watt as competitors, and if you're a graduate student or a researcher or staff, we also have like long heroes as mentors, so there's a role for everyone.

Outro music starts

So if folks are interested, I would say come join us! But otherwise thank you very much for having me on.

Outro

[Tom] Thank you so much to Mine for taking the time out of her busy schedule of pioneering data science and statistics to talk to us, we had an absolute blast! You can find her on Twitter at @minebocek, that’s m-i-n-e-b-o-c-e-k, and if you’re interested in either of her online courses, you can find Data Science in a Box at datasciencebox.org, or head to Coursera and search for Statistics with R.

[Helena] This podcast is brought to you by the Edinburgh University Science Magazine. In each episode we’ll explore fascinating themes and ideas, talk to awesome researchers about their work, and find out about the science being done by our very own staff and students here at the university.

[Tom] If you have any feedback for us, or if you’d like to get in touch with a question or suggestion, you can reach us on our Facebook page, Edinburgh University Science Media, or at our twitter, @eusci, that’s @e-u-s-c-i. You can also drop us an email at eusci.podcast@gmail.com, and you can find the show notes and the latest issue of the magazine at eusci.org.uk. If you would like to be featured on the podcast, please get in touch, and keep an eye on our social media for more information.

[Helena] This episode was hosted by me, Tom Edwick, and my partner in crime, Helena Cornu. The podcast manager is Alix Bailie. The podcast logo was designed by EUSci chief editor, Apple Chew, and the awesome podcast episode art was designed by Heather Jones, our social media and marketing genius. The intro and outro themes are edited from music by Kevin McLeod.

[Tom] Thank you for listening, and until next time, [Helena] Keep it science.

Post-outro shenanigans

[Mine] That is one of my 4 cats who love zoom calls. The other three, no one has ever seen, I feel like, in all this year of pandemic, but Dorian Gray likes to join in on calls.

[Helena] I love it.

[Tom] I love that name.

[Helena] That’s so good, brilliant.

[Tom] To the listeners who haven't seen it, a tail just brushed across the screen and Mine smoothly took the cat, placed it gently down. That is very, well rehearsed.

[Helena] Yeah, absolutely.