SDS PODCAST EPISODE 779: THE TIDYVERSE OF ESSENTIAL R LIBRARIES AND THEIR PYTHON ANALOGUES, WITH DR. HADLEY WICKHAM

Show Notes: http://www.superdatascience.com/779

Jon Krohn: 00:00:00 This is episode number 779 with Dr. Hadley Wickham, Chief Scientist at Posit. Today's episode is brought to you by Intel and HPE Ezmeral Software.

00:00:14 Welcome to the Super Data Science Podcast, the most listened-to podcast in the data science industry. Each week we bring you inspiring people and ideas to help you build a successful career in data science. I'm your host, Jon Krohn. Thanks for joining me today. And now, let's make the complex simple.

00:00:44 Welcome back to the Super Data Science Podcast. We've really been on a roll with our guest lately on this show, and today is no exception with the superstar Dr. Hadley Wickham as my guest. Hadley is Chief Scientist at Posit. He's also adjunct professor of statistics at Stanford University, Rice University, and the University of Auckland. However, he is best known as the creator of the Tidyverse suite of open-source R Libraries for data science, including the essential libraries, dplyr and ggplot2.

00:01:15 On top of that, he's written several seminal books on R programming, including the mega bestselling book "R for Data Science". Our book giveaways on this podcast have proven to be popular, so we're doing one for Hadley's books too. I will personally ship 10 physical copies of Hadley's books to people who comment or reshare the LinkedIn post that I publish about Hadley's episode from my personal LinkedIn account today. Simply mention in your comment or reshare which book you'd like. "R for Data Science", "Mastering Shiny", "Advanced R", "ggplot2", or "R Packages". So there's five options for you there. I'll hold the draw to select the 10 book winners next week, so you have until Sunday, May 5th to get involved with this book contest.

00:01:57 Today's episode will be primarily of interest to hands-on practitioners like data scientists and machine learning engineers. In this episode, Hadley details why the iconic open-source company RStudio rebranded to Posit, the philosophy of his Tidyverse, amusing backstories on its most iconic packages and why the Tidyverse is invaluable for all data scientists to be familiar with. He also talks about the open-source projects he's most excited about today and how you can easily get involved with career-bolstering open-source projects yourself. All right, you ready for this magnificent episode? Let's go.

00:02:35 Hadley, welcome to the Super Data Science Podcast. It is a surreal experience for me to have you here. I have seen you in person. In fact, we have actually seen each other in person, I'm sure of this. And so let me tell you these stories. I didn't tell you this before we started recording, so you're just getting this on air. Circa 2014 in New York at an O'Reilly Strata Hadoop World Conference, you did some kind of hands-on training. It might've been a half-day training. They used the flight times data set, does that all track?

Hadley Wickham: 00:03:12 Yep. I was involved, yeah.

Jon Krohn: 00:03:14 So I did that. I was in the audience for that. And then a couple of years later in 2016 at the Joint Statistical Meetings in Chicago, there was an announcement from RStudio that Hadley Wickham would be at the RStudio booth at this certain window of time. And I walked by a couple of times and we made eye contact and you kind of smiled friendly, but I was too nervous to talk to you. I didn't know what to say. You were at that time and still today one of the most iconic people in data science to me. I was just like, "Well, there's no, what do I say? How do I introduce this?" And so now I finally know what to say. I've got a question to ask you. Hadley. Welcome to the

Super Data Science Podcast. Where in the world are you calling in from?

Hadley Wickham: 00:04:03 I'm calling in from Houston, Texas.

Jon Krohn: 00:04:06 Nice. It is truly such an honor to have you on the show. You were on the show in the past, so four years ago you were on the program and that was specifically episode number 337, but at that time our host was Kirill Eremenko and the timestamp on this is pretty interesting. That episode was published in February 2020, so it was a pre-pandemic world.

Hadley Wickham: 00:04:34 Different time.

Jon Krohn: 00:04:34 Very different time. Exactly. So all of that has passed and we're almost back to normal except that so many data scientists are working from home. All right, so straight into the technical content, Hadley. If there's one word that is most associated with you, it's got to be tidy. In 2014 you wrote a highly cited paper called tidy data and you're also an author of the popular Tidyverse, a collection of packages that share a high level design philosophy. Last but not least, you have been writing a book called Tidy Data Principles, which is to be completed next year in 2025. What does tidy mean in the context of data and programming and yeah, what's the guiding principle? That's definitely not a question that ChatGPT would've asked you.

Hadley Wickham: 00:05:25 Tidy to me is about having things that are kind of well organized and well broken down into little pieces that you can then reassemble, like Legos. I think that's been a motivation for a lot of my work is how do you take some big, maybe kind of vaguely ill-defined problem and then break it down into concrete pieces that you can actually get stuck into and experiment with and play around with and iterate towards a final solution?

Jon Krohn: 00:05:57 Yeah, so in your tidy data paper, you draw parallels between tidy data and the principles of relational databases, specifically Codd's relational algebra. What is Codd's relational algebra? Could you elaborate on how database design can benefit statisticians and data analysts in their work?

Hadley Wickham: 00:06:15 Yeah, so you can go and look up on Wikipedia or something, what Codd's relational algebra actually is, but I can never remember it. But I think, I don't know. It's one of those very precise definitions where every word makes sense individually, but when you string them together in a sentence, it's very hard to understand what it means. But I sort of got into relational data because my dad had done a lot of database design for capturing data about cows in particular and this idea-

Jon Krohn: 00:06:56 About cows?

Hadley Wickham: 00:06:57 About cows, for cattle breeding. And this idea of making sure that each fact, each unique fact is recorded once in a data set rather than having it potentially either split across multiple places or recorded in multiple different ways in different places. And so I think that the ideas of Codd's relational algebra are really important. You don't want to have inconsistencies in your data, but it's really difficult for folks who are not trained in databases and computer science to get the idea of the algebra. And so a lot of that, the idea of the tidy data was how can I frame this in a way that makes more sense to statisticians and data scientists and people working with data?

00:07:49 And to me it's just like you've got a rectangle and all tidy data really is we put the variables in the columns and you put the observations in the rows and you going to wonder, that legitimately took me eight years to figure out? It seems so simple in hindsight, but it's just one of those things that once you figure it out and explain it to

other people, it makes a lot of sense, but it takes a while to get there.

Jon Krohn: 00:08:14 It takes a while. Even for me as my first time using the tidy data principles, which must have been many years ago now, it was probably a decade ago or more. But that first time kind of wrapping your head around shaping the data in this tidy way, it is so different from the way that we're typically taught in university and so therefore that first time doing things in the tidy way exactly as you described, where each piece of information is only replicated once as opposed to having a whole column where the variable name, let's say you have a binary outcome and every row of this giant table is outcome zero or outcome one.

00:08:56 It's so wasteful, especially if that's like a string, it's incredibly inefficient and so popping over to the tidy principles where that goes away is transformative. But it did, I remember the first time trying to meld data. I'm like, what? And even when I saw it the first time, it took me a while to figure out and I was kind like, this isn't right. I felt like I needed to change it into the way that I'm used to even work with it.

Hadley Wickham: 00:09:24 Yeah, it's really interesting. At the moment I'm on the sort of run the program committee for Posit::conf, so we decide on what talks we're going to have and how we're going to arrange them. And so the program committee is mostly data scientists and so the way we think, we do a lot of organizing data in Google Sheets because it's so easy to collaboratively edit. But it's always in a tidy format and what that typically means is you can't actually get a shape, you can't really understand the shape of the whole program without joining three different things together. And it's interesting when I share with my colleagues in marketing who have to then turn this into the website, the way they want to put that in a Google sheet is just so

totally different. There's like merg sells, there's colors. I get it, but just like you could not compute on data in that form, but it's so much easier to look at as a human.

Jon Krohn: 00:10:20 Are you able to, when you're in a data science company like Posit, are you able to kind of dictate marketing, you've got to do things our way, the data scientist way or that's impossible?

Hadley Wickham: 00:10:31 It's impossible. I mean I don't think I could dictate it. I could certainly do more to try and persuade people or help them do their jobs better, but that's a lot of work and you are fighting the fact that none of their tools think about data in this way and I don't really want to have to go and create the Tidyverse for marketing and the Tidyverse of finance and every other field. I think we have certainly more penetration of Quarto and those sorts of document generation tools, but there's still quite a high bar for folks who are like, if you're not used to using GitHub, collaborating in this way is pretty tough, even though that final product can be pretty nice. And that's one of the things I hope for the future of Quarto was like these tools for scientific documents that let you mingle text and code, but also can work with Google Docs and multiple people can be contributing to them. You can comment on them, you can share them with the non-technical folks.

Jon Krohn: 00:11:41 Quarto is something that we also have as a whole topic area later for discussion. It is a great tool that I think anybody can be using. Within data science or amongst data analysts particularly, have there been persistent challenges that you faced, like technical hurdles or adoption-resistant, conceptual misunderstanding? Has there been any of that over the years? With tidy data in particular. Sorry.

Hadley Wickham: 00:12:06 Tidy data?

Jon Krohn: 00:12:06 Yeah.

Hadley Wickham: 00:12:09 Not too much. I think there's definitely some areas where there's just such strong conventions for the field of having things that I would think want to be in a single column spread across multiple columns. There's some types of data where just arranging things in non-tidy forms, it's just much, much more memory efficient. But by and large, I think most people have looked at the tidy data framework and even if they don't agree with all the tools or don't use R, still find that framing really, really useful. And I think that thought of let's separate that out a separate step rather than trying to make every other tool do a little bit of tidying just makes life so much easier.

Jon Krohn: 00:12:55 A hundred percent. And for people out there who haven't had the tidy experience, it is absolutely worth wrapping your brain around it because once you do get used to it, everything becomes so much easier and all of the tools in your Tidyverse work so seamlessly together. Yeah, it's like, I don't know. When I was a kid, you'd have some computer video games that were just garbage and buggy and you'd constantly, you'd be able to walk through walls or whatever, and then you get your, for me it was having a Super Nintendo for the first time and nothing ever crashes and everything just works. And the Tidyverse is kind of like that. You get the data formatted in that way and it's just smooth sailing.

Hadley Wickham: 00:13:36 One of the things we've sort of wondered about. If we were going to have a metric that me and my team we're going to try and optimize, I think the metric that I want to optimize is the amount of time you spend in flow state where you are just thinking of stuff you want to do with data. The code just flows out of your fingers and it all just works. And we're certainly still some distance from that perfect utopia, but over time it really feels like the amount of time you can stay in that flow state, stay answering,

asking questions of the data that you actually care about, not like how do I get this thing into this other function so I can actually just do what I want to do?

Jon Krohn: 00:14:21 Ready to master some of the most powerful machine learning tools used in business and in industry? Kirill and Hadelin, who have taught millions of students worldwide to bring you their newest course Machine Learning Level 2. Packed with over six hours of content and hands-on exercises, this course will transform you into an expert in the ultra-popular gradient boosting models, XGBoost, LightGBM and CatBoost. Tackle real-world challenges and gain expertise in ensemble methods, decision trees and advanced techniques for solving complex regression and classification problems available exclusively at superdatascience.com. This course is your key to advancing your machine learning career. Enroll now at superdatascience.com/level2. At superdatascience.com / level and the number 2.

00:15:03 Exactly. On that note, in a recent article about strategies and function design, you have discussed the importance of making strategic choices explicit to users. So hopefully this is ringing a bell. Do you think that this kind of concept of providing a clear presentation of strategies in the way that you're designing a tool or function that actually enhances the user's experience or functionality, maybe allows them to be in that flow state for more of the time that they're programming because they understand the thinking behind the strategic choices in the tools?

Hadley Wickham: 00:15:45 Yeah, I think so. And kind of framing that even a little more broadly, one of the things that me and my teams think about and talk about quite a lot is how much do we want to force you to learn some new concept? That's really, that might really, it's like a better mental model. It's going to help you in the long run, but until you learn what that thing is, the code is going to be a bit of a

mystery for you. And that kind of balance of we want you to learn some new ideas like this idea of tidy data. Clearly there's this pretty clearly a big payoff of getting that concept into your head. And then versus other times, are we just teaching you some kind of technical jargon that's really useful for us, but maybe it's just more junk to fill your brain up with? So that's one of the things we're going to think about a lot. How much do we want to accommodate your existing mental model versus how much we want to give you a new and better mentor model, either possibly against your will.

Jon Krohn: 00:16:54 Amongst all of the libraries that you've developed in the Tidyverse, so there's things like reshape and plyr for shaping the data and being able to have pipelines of data within this tidy framework. Is there any particular library that is near and dear to your heart that you feel like this was one that either, I don't know, conceptually developing it? You must kind of have favorites, right? They're not all equal to you, right?

Hadley Wickham: 00:17:22 Yeah, I mean, I have to say one of my favorites is like dbplyr, which allows you to write R code, dplyr code and then translates that automatically to SQL. And one of the things I love about it is this sort of combination of this really deep technical knowledge of R that you need to make this work in terms of translating the R code. And there's also, there's no way to do it perfectly. You cannot translate perfectly every piece of R code to equivalent SQL. There's also this kind of like how do I carve out the biggest benefit with the smallest amount of work? I think the combination of those two things, something I really enjoy.

Jon Krohn: 00:18:08 Nice.

Hadley Wickham: 00:18:09 Kind of related to that, there's bits of the testthat package. One of the, to test that as a package for doing

unit testing in R. And one of the things I worked on a couple of years ago was this package called Waldo, which is all about concisely describing the difference between two objects. And that's kind of like a similar problem. You've got this deep technical understanding of the language and all the objects and writing C code to iterate through them and then how do you present that to the user in a way that helps them see the differences as easily as possible. So that kind of tension there between this again, programming and human psychology, I just find that really interesting and fun to explore.

Jon Krohn: 00:18:54 Yeah, it must be interesting to think to yourself, "Well there's clearly this need to be addressed and it's an impossible problem to solve perfectly. So how can we get most of the way there in a way that will satisfy most people?" And those are two libraries that I definitely need to spend more time with. I don't think I've used dbplyr or testthat actually despite my love for the Tidyverse.

Hadley Wickham: 00:19:24 It's also just one of those things where I'm like, it just seems like a miracle that it works so well. I think the thing that's fascinating to me is that it really reveals for the kind of core data, this core stuff you do to vectors and data science, whether you're summarizing them or you're filtering them, it's basically the same code in R or Python or SQL or JavaScript. You can express pretty much the same things in every single language in a way that is surprising and interesting to me.

Jon Krohn: 00:20:01 That is cool. Beyond the libraries that we just mentioned in the Tidyverse, one favorite of mine is Shiny because it allows you to so rapidly build interactive web applications for data analysis, especially compared to any other web development framework that I've tried and I don't have experience, much experience developing any kind of web tools, but I can very easily use Shiny to get a web application up and running for people to have a self-

service dashboard that's click and point. Do you want to talk a bit about Shiny? And actually I think something about it is that it's now, and we're going to talk about Posit soon and the change from RStudio, but this is something that works kind of across programming languages now, right?

Hadley Wickham: 00:20:44 Yeah, exactly. So there's now Shiny for R and Shiny for Python, and they're completely separate code bases, but the idea that really unifies them is this idea of reactive programming and the idea of reactive programming I think at its heart is pretty simple. You've got a bunch of inputs to your app, things that people can change, and you've got a bunch of outputs. And what reactive programming does, it just automatically figures out what's the minimal amount of work to do when you change one of the inputs to update the needed outputs.

00:21:16 And again, that's one of these ideas like tidy data that takes a little while to get your head around. It's quite possibly an idea you've never encountered before in programming works a little bit differently to things you might have encountered, but once you get that idea, it just gives you this incredible toolset to create apps where things just work and you don't have to worry about things either like updating too often doing a bunch of needless work and making your app too slow or just failing to update. And so you've got these mysterious bugs in your app where things don't change when you expect them to, which is one of the most frustrating things to try and debug when something doesn't happen. That is fun. Yeah. So yeah, Shiny, really, really, really cool.

Jon Krohn: 00:22:06 It allows you to spin up basically that Super Nintendo game that I was just describing that just kind of works like you think it should. People don't walk through walls accidentally as they're using your dashboard that you developed in literally minutes. It's cool.

Hadley Wickham: 00:22:20 Yeah, it's funny. I remember talking to Joe Cheng who wrote Shiny very early on and I was like, "Joe, you think R users want to make websites? You can use Ruby for that. Use PHP for that. Why on earth would a data scientist want to make a website?" And now it's so obvious you, because don't want to give decision makers in your organization just like a PDF. You want to give them a little interactive app. And there's just been so many examples of people just really impressing their bosses with Shiny because you can whip up something in a couple of hours that looks like a polished app, does exactly what you want.

00:23:03 I remember a very early phone call from a Shiny user saying we saved him a quarter of a million dollars because instead of going and finding a contractor to implement a web app and a dashboard, he just did it himself over a weekend. And not only is there that cost and time benefit, but also that if you as a data scientist can do it yourself, you don't have to try and communicate to someone else exactly what you want. That is tough.

Jon Krohn: 00:23:34 Exactly.

Hadley Wickham: 00:23:34 Working with other people. It is tough.

Jon Krohn: 00:23:36 It is. Well, and this also allows you to make changes yourself. If you notice an issue or a user complains to you, you can just go in and fix it as opposed to needing to be a middle man. A middle person, I guess.

Hadley Wickham: 00:23:50 Yeah, I think one of the interesting things about dashboards is if your dashboard is successful, people are going to demand changes to it very, very quickly. If you have a really, really good dashboard, that means there's going to be two or three execs in your company who now want to make a bunch of little tweaks to it. And if that's some weeks-long process where you've got to figure it out,

then communicate to some web engineering team, that just kills the whole thing.

Jon Krohn: 00:24:18 And I think with how often executives think they want a dashboard and then relate it to how often they actually use it, that is another strong point for using Shiny because that way you're not wasting weeks or months developing a dashboard. You're wasting hours a day.

Hadley Wickham: 00:24:38 Exactly. I mean, just in general, that whole iteration, the more you can do to increase your iteration speed, the more effective it makes you. Because again, it's so hard to predict in advance what's the thing that's going to be valuable. There's definitely a lot to be said to just trying out a ton of things and seeing what sticks rather than doing a bunch of upfront planning and just hoping desperately that you've got a really good mental model of the world and your idea works.

Jon Krohn: 00:25:05 So we are going to talk about, as I already mentioned, the Posit name change. We'll end up talking about Python a bit. For our listeners, if there are listeners out there who don't already use R, why should they be using it? For me, I can actually give one example, which is for me, for data visualizations, I still find I can do things way more quickly, have much more fun making visualizations in R, and get exactly what I want. There had been in the past attempts to create a ggplot style Python library, but the one that I had been using became deprecated and harder and harder to use. It never had all the functionality of your ggplot2 anyway. Anyway, so that's my big example. I don't know if you have big examples of why people might want to use R still today.

Hadley Wickham: 00:25:56 Yeah, I mean on the topic of ggplot2 specifically, I think the best Python equivalent is plotnine, and that's actually by a developer Hassan Kibirige that we've been sponsoring at RStudio, at Posit I think. And I think that's

that's the best possible realization of ggplot you can get in Python. But I think there's things about the design of R language that just make certain tasks much easier and more natural to express in R code than you'll ever be able to do in Python. And I think that comes down to at the heart of it, like R is more of a special purpose programming language. It's designed from the ground up to support statistics and data science, and I think that has a lot of benefits, particularly if you've never programmed before. I think you can get up and running in R, using R to do data science, you can do that without learning a ton of programming, get up and running pretty quickly.

00:26:59 And then there's just sort of things about the language that other languages look at R and they're like, "Oh my god, that's a terrible idea or that makes me want to throw up in my mouth." But there's just so many things that are just so well-placed to support interactive data science, where you really want that fast and fluid cycle where you're trying things out. That obviously bends to maybe a little bit of weaknesses on the kind of like, now I've got this thing, I just want to do the same thing again and again and again and again. R tends to be a little bit magical. It tries to kind of guess a little bit more of what you want and that's great when you're working interactively and it guesses correctly. It's not so great when you're working on a server somewhere else and it guesses the wrong thing. But just R, like everything about R I think makes it such a fluid environment for really exploring your data, digging into it, figuring out what's going on.

Jon Krohn: 00:28:00 The hybrid cloud promises freedom, but for AI and analytics, it can feel like juggling chainsaws. Data silos holds you back, tool sets clash, and managing resources across environments becomes a nightmare. This is where HPE Ezmeral Software steps in. On May 2nd, join a free

webinar powered by Intel Xeon Scalable Processors and HP Enterprise designed for data science professionals facing these hybrid cloud AI challenges. You'll learn how to empower AI model building and training with seamless access to global data sets, leverage built-in connectors and a curated open-source ecosystem to focus on innovation and trade in Frankenstein infrastructure for consistent environments that simplify resource management and accelerate any AI journey. See the show notes for the link to this free webinar on May 2nd.

00:28:44 Speaking of differences between R and Python, I seem to remember, and you can correct me if I'm wrong about this, but I feel like you have a famous tweet from years ago where somebody says something like, and it must've been a famous poster themselves that you responded to and I can't remember, it might've been like Wes McKinney or somebody like that saying that one of the advantages of Python is that it's faster than R, and then you have this super famous reply of, "What is that? And I will make it faster." Do you know what I'm talking about?

Hadley Wickham: 00:29:20 I don't, but I know I've seen things like that in the past.

Jon Krohn: 00:29:26 Yeah, it's a misperception because Python isn't actually that fast itself. I mean whole languages like Julia have come up to be faster than Python.

Hadley Wickham: 00:29:37 Yeah, I think one of the reasons often the biggest, you have the worst arguments with your family and not with strangers. With people who are so similar to you, you tend to have more friction than the people are really different. I think because R and Python are actually really close together in the spectrum of programming languages, it's so easy to see all of the little things that look weird to you as opposed to looking at some programming language that's miles away and it just looks, it's totally different. I just think that, I don't know, I think there's something to

that because we're close, you can see these little differences. And certainly when I see things in Python that people are like, "Wow, that's really cool." I'm like, "Challenge accepted. I'll make that better in R."

Jon Krohn: 00:30:26 Yeah, exactly. So let's dig into that a bit now. For 11 years, you've been the chief scientist at Posit, makers of open-source software for data science, scientific research, technical communication communities. Many R users will know Posit as the makers of RStudio, a full-featured integrated development environment, IDE for R, which I myself have been using for as long as I can remember. Basically, as long as I have been typing, I have been using RStudio. And RStudio, as you actually kind of let slip earlier in this episode when you were talking about Joe Cheng, I think. Oh no, no, no. You're talking about plotnine and how he was like, "RStudio is supporting. Wait, no Posit." And so that's two years ago. The company name has changed to Posit and from a distance, I mean, I don't even think it's from a distance, I think this is explicitly related to how Posit is now supporting more than just R. Is that right?

Hadley Wickham: 00:31:24 Yeah, yeah. I mean the goal of RStudio and now Posit has always been to be this kind of like a company with a long-term vision. We talk internally about this sort of idea of a hundred-year company. And when you think about a company like that, obviously no programming language is going to be around in a hundred years' time. When we started with R, that's something that's near and dear to many of our hearts that will always be, but we also wanted in the name of our company to embrace that there are now other languages and there will be even more languages in the future.

00:32:01 And I kind of think about this as the Burlington Coat Factory problem. I don't know if you know Burlington Coat Factory, but we have a lot of ads with them on

television. But for a long time they were like, "No, it's like Burlington Coat factory. It's not just coats." And for us to go into customers and say, "Buy RStudio, it's not just R", like it's hard to make that story. So really, really wanted to say, "Hey, for a long time now our products have supported not just R, but Python and Julia and other tools. We don't want to lock ourselves into this. We're going to be R forever regardless of what happens with the rest of the world." So renaming to Posit, rebranding to Posit was really about saying we're in this for the long haul and we care about data scientists regardless of what tool they're using.

Jon Krohn: 00:32:52 One of my favorite things that you can do really well, thanks to the dplyr library that you led development of is piping. And so you can extremely easily have functions passive, just like if people are familiar with Unix programming pipes there. Where you have output from one function goes the input to a next function and prior to me discovering dplyr, which was probably around 2010, if that makes sense. Prior to that I would have so many variables in my workspace. It was just such a pain to keep them all straight and you just end up in these weird situations, where should I be investing time thinking about the name of this intermediate variable? Am I going to use this later or should I just name it like intermediate variable 15 and have really ugly code?

00:33:47 And so piping gets rid of all that where you can read the flows like a sentence. You're like, okay, this pre-processing step happens, then this next, and you can just see it so easily. It makes it so elegant to read. Do you think we'll get to a point where, and I have used some kinds of piping attempts in Python, but my experience of that has never been, and I guess it's been a few years since I've tried, but it seems like it's never been as smooth or as easy as with R. And maybe that's related to what you were talking about earlier with data visualization.

Hadley Wickham: 00:34:22 Yeah, the native equivalent of piping in Python is method chaining. If you're using Pandas, you do something dot something.

Jon Krohn: 00:34:35 Yeah Panda's-.

Hadley Wickham: 00:34:37 Dot something. But the big difference between method chaining and the pipe is in method chaining, all of those methods have to come from the same class. They have to live in the same library, the same package, whereas with piping, they can come from any package. And I think the thing that's really interesting about that is that has meant Python has tended to have these fewer bigger packages like Pandas and Scikit-learn, Matplotlib, kind of everything in order to work with method chaining, everything has to be glommed into this one giant package.

00:35:15 Where with R, because you can combine things from different packages, the equivalent of Pandas is kind of like dplyr and tidyr and readr and a bunch of other things. It's way easier to add extensions to ggplot2 than Matplotlib that work exactly the same way because you can just combine them with different pieces. So I think that's just one of these interesting subtle differences in language design that lead to fairly big impacts on the user experience and almost even how the community has to work together and form.

Jon Krohn: 00:35:51 Yeah, it makes perfect sense and it's actually, your explanation is so simple for how that's happened and that had kind of escaped my attention as to why it worked so well in R. When you were last on this podcast four years ago, you said that you wanted to marry the Python and R languages. Four years on, how do you assess the progress made in achieving this dream, especially through projects like Apache Arrow?

Hadley Wickham: 00:36:18 Yeah, I think we've come a long way and Arrow has made a big difference in just being able to seamlessly move data from one platform to another, one programming language to another. And then kind of coupled with that, I think the other technologies that's really interesting is DuckDB. You can use DuckDB from R, you can use it from Python and you don't have to have a database file. You have a directory full of Parquet files and it means that people are using the same kinds of tools, just expressing them in the language that they feel most comfortable with.

00:37:03 Another sort of a similar thing is Keras and a lot of the machine learning toolkits and Python, the reason that they are fast is not because Python is fast, it's because you express those high-level ideas in Python and then they get compiled down to some low-level machine code. And that's why packages like the Keras package for R, which is maintained by one of my colleagues at Posit, Tomasz Kalinowski, it does the same thing. You express these ideas in R rather than Python, but then it gets compiled down to machine code using exactly the same toolkit. So I think that we're just going to continue to see more and more of that. R is not fast, Python is not particularly fast. What is fast is people really caring about stuff and Rust and C, and then you write a more user-friendly interface on top of that, the programming languages the data scientists use every day.

Jon Krohn: 00:38:03 Yep, yep. Makes perfect sense. And those libraries that you mentioned there are super cool in addition to the Arrow that I mentioned, speaking of Wes McKinney, and we actually have, so we have a whole episode about that. Let me quickly look up the number here. So back in episode number 523, we had Wes McKinney on and he talks about the Apache Arrow project at length, really cool one. And the other projects you mentioned there, DuckDB as well as Keras for R. Yeah, super cool invaluable

packages that people should be trying out on the show for sure.

Hadley Wickham: 00:38:42 Yeah, Arrow's also been top of mind for me lately because I've been working on some enhancements for the bigrquery package, which allows you to get data from BigQuery. And previously the only way you could get the way that the BigQuery package used is downloading the data as JSON. And JSON is a great interchange format, but it is horrendously inefficient if you're sending data frames of data. And so thanks to some folks who've been working on the bigrquery storage package, which talk to the Google API using Arrow instead of JSON, you can download data like an order or sometimes two orders of magnitude faster because you're using a data format specifically designed for the type of data that data scientists care about. And it's kind of nice to see that in practice and that kind of dream of Arrow of, "You've got data over here and you want to get it over here, let's make it as easy as possible."

Jon Krohn: 00:39:45 Speaking of multiple languages and interchanging between them, do you personally encourage teams to be multilingual or do you often write in multiple different languages, maybe even within the same project? How do you think about that in your own?

Hadley Wickham: 00:40:02 I'm a hundred percent R. And as far as I can tell, I probably always will be, that's my job. If there's something that I could do better in Python, I will write an R package so that I don't have to. But that's not the reality of most people's lives and most people's jobs. And I think most people tend to be like 90% R or 90% Python. But in general, it's just better to be pragmatic. If there is something that's way easier to do in another language, you can learn the basics of R, you can learn the basics of Python pretty quickly so you can use that tool.

00:40:41 And I think that's one area where I think generative AI is really interesting, just being able to generate code in another language quickly. And I've been using it quite a bit to generate JavaScript. I do the occasional web thing and I really like it. I know enough about JavaScript that I can look at what it produces and say that looks right, but for me to manually figure out would just take so much longer. And so I think it's really interesting to think about how that's going to affect programming languages. If it's really easy to translate between them, maybe this means the barriers between them are going to erode a little bit more.

Jon Krohn: 00:41:28 I think That's right. It's amazing that we've gone this far into the episode without talking about generative AI yet. That's kind of a nice refreshing and actually looking at all the topics we have lined up. I don't think any of it touched on GenAI, which is kind of crazy today. But as you were talking earlier about how in a hundred years how Posit is aiming to be a hundred-year company and we think about what programming languages will there be a hundred years from now? I was just, that was batting around in my head, bouncing around in my head as you were speaking, and I was thinking how I wonder if a hundred years from now anybody will be programming? Because I wonder if just natural language expression of things will be so powerful or I wonder if we'll be working at all. I don't know.

Hadley Wickham: 00:42:10 It's just going to be like a Mad Max hell scape. I think the other thing that's really interesting to me though is if people are really going to be using generative AI LLMs for programming a lot, what does that mean for new programming languages which are not going to have any training data available for them? That seems like that's going to raise the barrier to new languages even further. Yeah, it's also just what happens to Stack Overflow and are we, this idea of poisoning the well, people are stopping

using Stack Overflow, which is fine in the short term, but where's all the training data going to come from in the future? I don't know. It's exciting and scary.

Jon Krohn: 00:42:57 Yeah, it is exciting and scary. I have this perhaps completely unfounded intuition that somehow it's, and there's people way smarter than me who have spent a lot of time thinking about this who could easily crush what I'm about to just say. You might even do it right now. But somehow I have this feeling based on how quickly issues like hallucinations have been stamped out, the jump between GPT-4.5 and GPT-4 with how much less it's hallucinating. I have this completely unscientific, uneducated intuition that somehow we're going to be fine on this front, that we're not going to end up having a complete... There's a specific term for this. You probably know what is this. It's this concept of the degradation as we only have synthetic data and then GenAI.

Hadley Wickham: 00:43:56 Beating it into itself. Yeah, I don't know.

Jon Krohn: 00:43:57 Yeah, there's a specific term, and I've used it on this podcast before, if somebody can send me a message on social media and be like, "You bonehead, two weeks ago you talked about this with this guest that's an expert in it." But yeah, that concept, it's something like degeneration, that kind of sound, that kind of vibe. But I don't know. I somehow feel like we're going to figure it out that even though people will probably contribute to Stack Overflow less and less that the quality of generation that comes out. And I guess part of maybe what gives me confidence is my recent experience with Claude 3 from Anthropic, which I have found so amazing. I went overnight from using GPT-4 all the time every day to using Claude 3 because everything is so sharp. And the interesting thing about that is they do tons of synthetic data generation as well as reinforcement learning with AI feedback as opposed to human feedback. And so it seems

to me like going in that direction is going to be okay. But anyway, I've gotten really off piece.

Hadley Wickham: 00:45:08 Yeah, I think the thing that makes me less optimistic is my Tesla and this promise of self-driving cars, which just doesn't seem to be getting any closer. Because every, I don't know, 50% of the time that I pull into our garage, it thinks the random collection of tools on the wall is like a semi. So I'm just like, and that's clearly something that's been very much hyped and a bunch of money has been... It's just going to be interesting to see. We're clearly in this explosive growth and is it just going to flatten off or is it going to keep going? Is it going to get steeper? Who knows?

Jon Krohn: 00:45:47 It's interesting. And Hadley, I don't know if you knew this or not, but that's called full self-driving. It can fully self-drive. I mean-

Hadley Wickham: 00:45:57 I don't know if everyone got it, but we got the full self-driving for free for a month. So I've been trying it out and it's just sometimes it's great, but a lot of times we're like, oh my God, this is scary. I don't understand how people get into these accidents with the self-driving because I'm like, there's no way I would trust this without keeping, my hands are white-knuckle on the steering wheel because I don't trust it. It's interesting.

Jon Krohn: 00:46:24 Since the start of April, I've been offering my machine learning foundations curriculum live online via a series of 14 training sessions within the O'Reilly platform. My curriculum provides all the foundational mathematical knowledge you need to understand contemporary machine learning applications, including deep learning, LLMs and AI in general. The linear algebra classes are wrapping up soon, but my calculus, probability statistics and computer science classes are still to come. The first two calculus sessions are available for registration now.

We've got the links for you in the show notes, and those will cover all the essential calculus you need for machine learning. Calculus level 1 will be on May 22nd, Calculus level 2will be on June 5th, and registration will open soon for Calculus levels 3 and 4, which will be on June 26th and July 10th. If you don't already have access to O'Reilly, you can get a free 30-day trial via our special code, which is also in the show notes.

00:47:15 Well, I guess back on topic, the last thing that we were, when we were on pieced and the flow I was in, we were talking about multilingual usage. We don't need to have a long discussion of this, but I suspect Hadley that some of our listeners out there are doing their data analysis primarily in Excel. And Python was recently integrated into Excel. Do you think that in the future Microsoft might support R too in the Excel kind of environment?

Hadley Wickham: 00:47:45 Probably not. And I have a vague... No, I don't. I have a vague sense that someone told me the way that it actually works and the way the Python works in Excel is pretty unpleasant and pretty hacky. But I think it's true that probably 90% of data science is still done in Excel, and that's just a huge audience of people going from a no code environment to a code first environment. I think so many people will benefit from that, and it's definitely starting to code for the first time is scary and it can feel intimidating. I think that's another place where large language models are going to come into play and be really interesting, but just the more we can do to help people switch from that point and click environment to something that's more reproducible, something where you've got code that other people can read and critique. That's really, really valuable.

Jon Krohn: 00:48:45 So maybe we won't be going down the Excel route, but another big innovation that has actually happened recently is R7, which I heard about for the first time doing

the research for this episode or reading Serge, our researchers research for the episode. Do you want to tell us about R7 and the problems that it's aimed to address?

Hadley Wickham: 00:49:08 Yeah, so we actually renamed it to S7 relatively recently.

Jon Krohn: 00:49:12 Oh, really?

Hadley Wickham: 00:49:15 It's called S7, so it's called S7 because there are two. This is a lot of historical minutiae, but the language that came before R was called S, and S was the kind of introduced object-oriented programming and S version three and S version four. So in R, there are two types of object-oriented programming, S3 and S4, the chief types. And the idea of S7 AKA R7 was to try and add those two things together and get the best of both worlds. So S3 is really just a very lightweight set of conventions. It's not like object-oriented programming in any other language, basically. It's very, very lightweight. S4 swings too far in the other direction. It's very formal. There's a lot of boilerplate. It's quite complicated. Things can go wrong in weird ways.

00:50:12 So the idea of S7 was really to try and find the sweet spot in between them, take the nice features that S4 had, add them on top of S3 in a backward compatible way so that we can hopefully switch. Hopefully we're not just adding another object-oriented programming style to R, but we're actually supplanting S3 and S4 over time because everything you can do in those two, you can do in S7, and you can do it more easily and there's better documentation and tooling and all that kind of stuff.

Jon Krohn: 00:50:42 For our listeners who don't have a computer science background, what does it mean for a language to be object-oriented and that it can have these kinds of grades from lightweight like you were describing with S3?

Hadley Wickham: 00:50:56 Yeah. I don't know. Yeah, objects a new program with them. I don't know. And it's especially weird in R because when you're using R, you're not really aware that you're using object-oriented programming. Unlike in Python, where I think you're much more aware of that you have objects and you call methods on those objects. Object-oriented programming is much, much less important in R as a data scientist. I think you benefit from it because packages use it. So I think that the main benefit to you as a data scientist is not that you are going to be writing S7 code, but the packages that you use are going to, and they're going to be able to write code faster and more correctly from the get-go. So hopefully more like a general uplift of developer productivity and are probably not going to affect data scientists day-today that much.

Jon Krohn: 00:51:56 All right, cool. So speaking of switching between different levels of object-oriented programming, you've already mentioned on the show how Posit has an ambition to build a company, a suite of tools that could last a hundred years. What kinds of principles or philosophies do you think are critical to creating a legacy that lasts that long in technology?

Hadley Wickham: 00:52:18 Yeah, that's a good question and I don't think we know for sure. But one of the things that make us different as a company is that we are a public benefit corp, a PBC or a B Corp rather than being an LLC. And what that means is kind of fundamentally baked into our charter is that our sole goal is not to optimize shareholder revenue, which is kind of the classic LLC model. We explicitly consider other stakeholders like the community and our employees as what we are trying to do as a company. And I think that is pretty special because we can say legitimately, we don't want to make products that lock you in. We want to make products that, we want to help you, and we're going to sell you products that are going to help you do your job, and you're going to hopefully pay us money for those

because they save you time, they allow you to do things that you couldn't otherwise do, but we're not just about their money.

00:53:24 We really care about your life as a data scientist, we want to build tools for you. We want to build open-source tools for people that don't have a bunch of money. We want to improve academia. So I think that that's part of the mission is we are not optimizing for short-term profit. We can say we are going to take a longer view. And so part of that is also we are not a VC driven company. We don't have to explode in either a good way or a bad way in three years' time where our investors want to get money out of so that we can kind of say, "Hey, we've got the time, we've got the financial stability and the kind of vision to look out a little bit longer in the future and do the right thing rather than the expedient things possible."

Jon Krohn: 00:54:17 It is a super cool company. There are so many amazing people who work at Posit, so many of the biggest names in data science work alongside you at Posit. And on top of that, I know some of the investors that have invested in Posit and they're so excited about the mission that you guys have. It's an incredible company and it would not surprise me if my-

Hadley Wickham: 00:54:41 Other thing that I think that's been really neat is there's also a bunch of people at Posit that you don't see from the outside who are fantastic. In finance and sales and people ops who are explicitly, we couldn't afford them usually, but they're drawn to this mission. They've done their time in corporate America. They don't just want to work at a company that cares about itself and is optimizing its bottom line, but they want to work at a company that is trying to do good in the world in some way. And that's really, really cool.

Jon Krohn: 00:55:19 No doubt. Switching gears a bit here from Posit in general, back to a topic that I assured our listeners we would get back to later on in the episode is you created ggplot2, which is to this day the most popular data visualization library for R. It's had nearly 140 million downloads at the time of recording, which is crazy. And you have an associated book ggplot2: Elegant Graphics for Data Analysis, which is super popular. And so my understanding from using ggplot2 now for over a decade, if I remember this correctly, you can tell me what I get wrong here or elaborate on it a bit, but it's based on a grammar of graphics principle, which I guess was a book, and that didn't necessarily have anything to do with programming. It was just a way of expressing visual information. And you picked that up and you converted it into a programming framework.

Hadley Wickham: 00:56:14 Yep, yep. I mean, The Grammar of Graphics, the book by Lee Wilkinson really felt like it was a book written for me because I read it and I was like, "Wow, this is how I want to be able to think about visualizations and describe visualizations with code." And that's what I attempted to do, turn the book into something that you could actually use for free with open-source software. If you go back and read the book, there's plenty of deviations and places where maybe ggplot2 didn't capture the full spirit of it, but overall, I think it's a pretty high fidelity implementation.

00:56:53 And just that idea of pulling apart these pieces, not thinking about this list of graphics like a pie chart of line chart, a bar chart, but thinking about these components that underlie the graphic, like the data and the scales and the geoms and the stats, just a really, really powerful way to think about visualizations that allows you to create the thing that's tailored specifically for your problem that maybe no one has ever created before.

Jon Krohn: 00:57:17 And so following on from this already very popular ggplot2 package. How do you envision the future of data communication, data publication? Particularly with the development of tools like Quarto, which you mentioned earlier, but we didn't really talk about it in detail. So maybe kind of describe the kind of key features of Quarto and how that fits into this whole data visualization perspective.

Hadley Wickham: 00:57:40 Yeah, I mean, Quarto at its heart is really a tool for communication, particularly scientific and technical communication because what it allows you to do is integrate your code and its results with your thoughts about the results. So you can write text and then integrate it with your results. And of course, as opposed to having a Word document where you may be copying and pasting plots, with Quarto, you end up with this document that is a hundred percent reproducible. That as the data changes, you click a button, it reruns all the code, recombines everything together, and you have this high fidelity document.

00:58:20 And there's a bunch of cool things about Quarto compared to its kind of predecessor R Markdown, Quarto works really well with Jupyter Notebooks as well. So if you've got a Jupyter Notebook, which is a fantastic tool for you as a data scientist and your colleagues who are data scientists, but if you're sharing that with other people, you might not want to share all the details. Quarto will help you turn that into a polished, it could be a polished PowerPoint presentation, polished PDF, a polished website, a polished book. But something that's really aimed at not just showing the mechanics, which of course you can show if you want, but the thinking behind it, the explaining what's going on with the data.

Jon Krohn: 00:58:59 That's super cool. And I was a big user of R Markdown. I haven't used Quarto yet myself, and it sounds like I need

to be, particularly for that element that you just described there, of being able to hide things. I have a number of times in my corporate life or in my commercial life created Jupyter Notebooks in situations where for whatever reason, say a Shiny tool wasn't the right thing, there was something about what I needed to be building where it made sense to me to create it in Python, in a Jupyter Notebook. And then I hand it off to somebody who doesn't program and they have to see all kinds of things about the Notebook that are probably confusing to see. And so to be able to hide those things with Quarto sounds really cool.

Hadley Wickham: 00:59:50 And because Quarto is built on top of Pandoc, you can produce pretty much any type of output you can imagine, which is really, really cool. So all the books I write today, they're all written with Quarto. I use, I make a free and open-source website that everyone can read for free and then it also creates a PDF that the publisher gets and turns into a physical book. You can turn it into an EPUB as well. It's just really nice to have that single source that you can then produce multiple versions of.

Jon Krohn: 01:00:23 Very cool. Throughout this episode we have been talking about, not just in the context of Quarto, but in lots of items over this episode, particularly the name change from RStudio to Posit. We've been talking about communities beyond the R community. With the rebranding to Posit and the expanded mission, how does the company engage with and foster beyond just the R community that RStudio was already easily the best known brand within?

Hadley Wickham: 01:00:51 That is a very good question and something I think we're trying to figure out too, and that we've had it easy for so long. I think it's fair to say RStudio is beloved in R community, everyone knows who we are and most people like us. And now we're coming into the Python

community, which is much, much bigger. And now we're a small fish in a very, very big pond and we're trying to figure out what we should be doing.

01:01:24 And then on top of that, since we promote in the last five years or 10 years since we really had to promote stuff, new things, it's all changed. We don't use Twitter anymore. It's now should we have a TikTok account? Just how are you going to reach out to a developer community has changed pretty profoundly from when we were really promoting the Tidyverse and promoting Shiny and promoting R Markdown. So I think with something we are trying to learn, doing the things that I think make sense everywhere, carefully listening to the community and trying to fix their pain points and just trying to figure out what's new, how do we find the people where they are and reach out to them.

Jon Krohn: 01:02:16 Yeah, sounds sensible enough. Tricky question. I realize you're not in the marketing department.

Hadley Wickham: 01:02:22 Yeah, I mean I get involved in a bunch of discussions though it's interesting and it's much, much harder. It feels much harder to me than it used to be and I don't know whether that's because it is actually harder or just like I have kind of got out, I'm out of touch with the youth of today, so I'm not sure what we should be doing.

Jon Krohn: 01:02:41 Yeah, it's hard to know, but a funny kind of thought that came to mind for me just now and I don't know why it took me so long for this to come in. But earlier in the episode, one of the first things that we were talking about was you made this comparison of how with R and Python them being so similar, it's kind of like a family dynamic where because you're so similar, you end up having these heated exchanges. And so that tipped me off your thought just now about how you have a sister Charlotte who is

also a statistician, she's a longtime university statistics professor like you are, and she even works at Posit, right?

Hadley Wickham: 01:03:19 Yep. That's correct.

Jon Krohn: 01:03:22 So what are your arguments like now? Let's hash out your last argument with Charlotte on air. I'll play her.

Hadley Wickham: 01:03:35 I would say we mostly try not to talk about work stuff that much. It's funny though, to work with the same company as you.

Jon Krohn: 01:03:48 But how did that kind of evolve? I don't actually know who's the older one, who's the younger one?

Hadley Wickham: 01:03:54 Yeah, I'm the oldest by three years. We ended up having, we did a few courses at college together. I originally started off in medical school, so I did a three-year degree, a bachelor of human biology and then decided I didn't like that, so I went back and did a bachelor of science. So we ended up kind of overlapping and I still remember there was some classes after lunch that I'd routinely fall asleep and had to rely on Charlotte's notes. But I think we both kind of got into statistics of data science through process of elimination. I started in medical school, she started in physics and both eventually ended up in a place where we both did our undergrad to the University of Auckland, the home of R. We were really exposed to a bunch of R there and I think that's really what got us both into it.

Jon Krohn: 01:04:50 That makes a lot of sense. And I don't know, it is kind of interesting. I have a sister who's two years younger as well, and we also kind of converged in roughly the same place from completely different backgrounds and it's interesting how that can happen. I think-

Hadley Wickham: 01:05:05 The genetics is. I don't know it was nature, nature or nurture, but it's, yeah, interesting .

Jon Krohn: 01:05:10 For me, this kind of work, working in data science, working in AI, I think it's just the most exciting that everyone should be doing. And so why isn't that a career choice that everyone's making? On that note of this being obviously the best job, you had actually mentioned when you were on this podcast last time that you thought everyone should learn to code. Maybe GenAI has changed that a bit, maybe people don't need to as much. But on the note of coding, do you have advice for people who would like to be contributing to the open-source R community, maybe even particularly to the Tidyverse?

Hadley Wickham: 01:05:50 Yeah, we're actually, after Posit::conf this year in Seattle on August 15th, we're going to be doing a Tidyverse developer day, which is like if you can make it to Seattle on that day, that is the best way to learn how to contribute because we're going to have about 60 folks in the room along with 20 helpers. We just help people get started with their first pull request on GitHub, learn how to document stuff with the Roxygen2 and the whole R packages workflow. That's just a really fun day.

01:06:27 If you can't make it to that like most of the people in the world, I think the key is to find somewhere to start small where you're still making a meaningful contribution. And I think one way to do that, particularly for anything I contribute to is I am terrible at proofreading stuff. So there's almost always certainly small typos and grammatical errors that you can fix in my work and that help, I think that does genuinely help people and it's just a great way to get started.

01:07:01 You can learn the mechanics of how do I do a pull request, how do I build the documentation for a package, how do I check a package without having to also master all of the technical skills at the same time. So I think documentation's a great place to start. The other place is look at issues, see if you can help the other person get a

little bit further or maybe even help them just produce a useful reproducible example so that when the developer of the package does come by, they've got something really clean and crisp that illustrates the problem. And that's just a great way, I think looking at someone else's problem and trying to recreate it yourself is just, I don't know, going to be a great way to improve your own programming skills too.

Jon Krohn: 01:07:46 This is a complete change of topic now, but it is just where I've gotten to and kind of our topics to cover. We basically covered all of the technical stuff that I wanted to, and so I don't have a good transition, audience. My apologies to this next topic. It's just a complete jarring topic change, but on your website you have listed, and we'll be sure to put these in the show notes, you have Wickham Family Recipes, and you also have a really nicely designed page of cocktails, which could almost be printed as a menu in a restaurant. It's very nicely done.

Hadley Wickham: 01:08:24 I do actually, I don't think I have it anyone near, but I have actually turned those into a book as well.

Jon Krohn: 01:08:31 I'm not surprised. It's 161 cocktail recipes curated by yourself. You can browse by ingredients or browse by tags so you can click through and say, "Wow, I really like banana liqueur" and find all the cocktails with that in it. So that's a really fun resource. I'm sure there are some listeners out there that will be checking that out. I got that in the show notes for you. And so that's a fun thing to check out.

Hadley Wickham: 01:08:57 I'd say one of the first places where I was really blown away by, let's bring it back to GenAI again, is that that cocktails website is powered by a YAML file that has all the cocktails and the ingredients in, and I was really impressed when RStudio enabled Copilot integration. It just enabled it for every file including the YAML files. So

now I can just type in title and the name of the cocktail and a good 75% of the time it completes the ingredients using exactly the same YAML syntax that I use for every other cocktail. And I was just like, "Wow, that's pretty neat. That's pretty cool."

Jon Krohn: 01:09:38 That is very cool. So my question for you is, given that you obviously have interests beyond just programming, just work, how do you optimize your routine or your flows? We talked about getting into a flow state earlier in this episode, so it sounds like something that's important to you. We also, we pulled out from a video that you made a long time ago, actually it's just seven months ago on the Posit PBC YouTube channel. You gave us a hint that key to your routine is prioritizing creativity and productive work first. I don't know if you want to elaborate on that.

Hadley Wickham: 01:10:24 Yeah, I think one of the things we've tried to make part of the culture that RStudio, now Posit, is this idea of deep work and that it's okay to carve out three or four hours or more from your schedule to just be able to invest in some project that you need to have that level of focus on. I still manage to do that. I try and keep Tuesdays and Thursdays free from meetings so I can focus on programming challenges. Recently, it feels like I've just had a lot more meetings and more kind of manager stuff, which I don't love, it's important. It's not something that amps me up, but I know it's important and I do it.

01:11:17 And that's sort of something I think I'm really hoping to get back to more in the coming weeks is just carving out that time, finding some big technical problems to get my teeth into. One of the things I have been thinking about quite a lot and trying to learn more about is R in production or data science and production. Generally just trying to understand what does that mean really? And then once I've turned solidly understood that really thinking systematically how can we make that easier for

folks using both of our open-source products and with our pro products. And I'm excited to kind of start digging into some of those technical challenges.

Jon Krohn: 01:12:02 Yeah, certainly that is another classic quip from Python developers is that you don't use R in production, but of course you can. And so yeah, lots to dig into there, I'm sure. Beyond that, what else are you excited about in this fast evolving data science field? Are there particular challenges or paradigms that are kind of up and coming that you're really excited about?

Hadley Wickham: 01:12:29 Don't know if I'd say I'm excited about it, but interested in it is these ideas around data contracts and now like most folks or many data scientists are working hand in hand with data engineers who are giving them, that's where they get their data from. But often that data is not produced primarily for data scientists. It's produced for some other business purpose and then the data engineers change that data in order to optimize that other data, that other business purpose. And then all of the data scientists' dashboards and models just break with confusing and bewildering error messages.

01:13:09 That just seems like an interesting kind of intersection of technical and people problems. It definitely seems to be something like a lot of companies are struggling with right now. I don't think I want, I don't want to solve that, but I want someone to solve it so I can be like, "Oh, okay, this is a good framework for thinking about it and then maybe I'll copy some of the tooling and art to make that a bit easier." I think that's interesting. Generally, I think just this whole, I'm not super invested in it, but just curious about our IMS and generative AI and how is that going to, particularly how is that going to affect learning R the first time? Learning data science? And teaching it too, because it fundamentally changes what questions you can put on an assignment and expect what questions a student can

answer without having to put in any intellectual work of their own.

01:14:10 And I'm very much always in the favor of lean into that. If now there's a question that a computer can answer, we need to find new questions. I think that's just really, that's interesting and it's not like, I don't know I'm... I don't think I'm going to be fitting my own models or training models or really understanding the math behind them, but just purely how does this change the way we think about teaching data science I think is really interesting and how can we build this into our tools? What are the resources we should be thinking about? All that kind of stuff. Interesting, exciting and scary.

Jon Krohn: 01:14:52 I met a professor at the University of Illinois a couple of nights ago at the time of recording. So Data Universe is a new conference that was just held for the first time in New York this week at the time of recording. And I met this University of Illinois professor at a dinner. He was sitting next to me and he was doing some really cool things with GenAI in education. So he had created a whole bunch of documents. He was describing how there's this problem increasingly with getting corporations to share real data for students to work with. And so he created a fake company called Nile, which was supposed to be like Amazon and generated dozens of business documents and data tables and then had all of that built into a chat interface so that students could use that as a key stakeholder that they're working with in the business and create data products for. I thought that was cool.

Hadley Wickham: 01:15:57 Yeah, yeah. So many possibilities and it's going to take a while just to figure out how to use them effectively.

Jon Krohn: 01:16:03 For sure. All right, so that's the end of my questions. We've got actually a crazy number. I'm going to have to curate them. My apologies to listeners who asked so

many questions we had, I'm pretty sure a record number of questions for Hadley, which is cool. It shows how much of an impact you have in the community in general, Hadley. Yeah, lots of big-name guests, but I've never seen anything like this, so I'm going to pick a few. First of all, Austin Ogilvie, thank you for connecting me with Hadley. He said that he's looking forward to this episode in one of his comments on LinkedIn, and I should have mentioned it at the outset that it was thanks to Austin for connecting us that you are here. Really appreciate that. Me sheepishly making eye contact 10 years ago and not being willing to actually talk to you at the Joint Statistical Meetings didn't cut the mustard.

01:16:58 So here's one from Matias Baudino. Matias is a longtime listener. He's a BI analyst at a company called Brain Technology. And this question, it ties into a lot of the R and Python stuff we've been talking about here. He's wondering if there is a combination and you're probably, I don't know if this is something you would think about. So is there a combination of Python libraries out there that could give you equivalent functionality to the Tidyverse or is there nothing out there like that?

Hadley Wickham: 01:17:30 I will tell you that if such a thing existed, I would call it the Pidyverse. That os far as I've got. I mean, I think there are two packages that kind of close to the major pieces of the Tidyverse and that's plotnine for ggplot2 and Siuba that's very similar to dplyr.

Jon Krohn: 01:17:54 How do you spell that?

Hadley Wickham: 01:17:55 S-I-U-B-A. Plotnine developed by Hassan, who we mentioned earlier, and Siuba is developed by Michael Chow, who's also at Posit now. So I don't think, we don't have any plans to try and make the Pidyverse for Python. I think Python's just a different environment. It's much more diverse than R, it's not clear to me that having this

kind of really centralized packages would work, but like me and my team remain a hundred percent focused on R. So I'm always happy to talk to people working on stuff about in this area of Python and give my 2 cents, but not something that I'll be working on personally.

Jon Krohn: 01:18:40 And we had Doug McLean, who's a lead data scientist at Tesco Bank, also a long time listener who's asked many questions of guests in the past. Thank you for all those, Doug. He wanted to make sure that we talked about Shiny and deploying apps in R as well as in Python, and he wanted to make sure that we talked about how much better R Markdown or now Quarto is relative to Jupyter Notebooks. So you get that he wrote "Yuck" next to Jupyter Notebooks.

Hadley Wickham: 01:19:09 I mean, Jupyter Notebooks are great as a tool for analysis. They're not great as a tool to share with decision makers in your organization, and Quarto now really fills that gap. You can have that Jupyter Notebook and then relatively easily turn it into a polished PDF or Word document or HTML files. Really worth looking into if you use Jupyter.

Jon Krohn: 01:19:33 Moving along from my LinkedIn post that you were coming up to my Twitter post, someone named EconFella wrote, "Just tell Hadley that we'd miss him on Twitter. This place sucks now." And interestingly, originally Twitter had actually hidden that it said, "This post may contain offensive language. Are you sure you want to view it?" And so it's funny, there's somebody's written some code somewhere that someone says, Twitter sucks.

Hadley Wickham: 01:19:57 Yeah, hides it.

Jon Krohn: 01:20:00 Yeah, Twitter misses you, but I totally understand why you would move on. I don't even call it by what it's actually called these days.

Hadley Wickham: 01:20:06 Yeah, no, that's still a angry and annoyed by that whole thing and there's no obvious successor in start. I'm on Fosstodon, which I don't know, it is only going to appeal to the most technical of technical people. I found it hard enough to get up and going on it. I'm on LinkedIn, which I still cringe whenever I use LinkedIn, but it's fine. And then I don't know, Threads just feels like switching, just switching one billionaire to another billionaire. So I've heard good things about BlueSky and I know people have been encouraging me to check it out, so maybe that will be something I try, but it's just like another social media platform.

Jon Krohn: 01:20:54 That seems to be the way to probably go as a replacement for Twitter. It seems to have some momentum, but we will have your Fosstodon account linked to in the show notes for people who get that up and running to join you in there.

Hadley Wickham: 01:21:07 People who have figured out, you're very welcome to message me on that.

Jon Krohn: 01:21:11 So back into a couple other questions that we haven't already covered in the show. This is an interesting technical one from Lalush. So does Posit still have plans of developing an online collaborative editing platform for Quarto similar to the role that Overleaf has in the latex ecosystem?

Hadley Wickham: 01:21:28 Yes, I'd say we definitely have plans. This is part of the vision, we want data scientists to be able to use Quarto to communicate with decision makers, non-technical folks so they can comment, they can change the writing. I think that's still some distance away on our plan, but absolutely we really want to make that, we really want to make Quarto facilitate the communication between technical and non-technical folks as well.

Jon Krohn:	01:21:59	And a bit of a quibble here. So somebody here says that ggplot2, that name is at odds with other names in the Tidyverse. Do you think you'll ever rename or rebrand ggplot2?
Hadley Wickham:	01:22:11	No, it's a weird name, but it's like it's too late to change it now.
Jon Krohn:	01:22:17	I like it. It makes perfect sense to me. Grammar of graphics plot two. Bruno Rodrigues wants to know what your top three movies of all time are.
Hadley Wickham:	01:22:32	Oh God, I hate listing favorite things. I don't know what movies.
Jon Krohn:	01:22:38	Just give us one. Just give us one.
Hadley Wickham:	01:22:42	I mean I really enjoyed the Barbie movie lately.
Jon Krohn:	01:22:45	The Barbie movie?
Hadley Wickham:	01:22:46	Yeah.
Jon Krohn:	01:22:46	There you go. For some reason I was actually thinking about today as I was falling off into a nap. For some reason I was thinking about the Barbie film and I was thinking it is odd that I haven't seen that. It must be a really good one.
Hadley Wickham:	01:22:59	It's pretty good. Yeah, it's pretty fun. Yeah, I'd recommend it.
Jon Krohn:	01:23:03	We've talked about Wes McKinney on this show already, Hadley. What can we expect from you two teaming up now that Wes is at Posit, can you give us any kind of hint into what you might be working on? Maybe Apache Arrow kind of functionality with R or the Tidyverse?

Hadley Wickham: 01:23:23 I wondered if this question was going to come up and I can't give you any hints, but I can tell you it's pretty cool.

Jon Krohn: 01:23:30 Well that's as exciting as it gets. That's awesome. Thank you so much for answering my questions and all of the audience questions as well. Hadley, quickly before I let you go, I know now that you hate listing favorite things and so you don't need to frame this as a favorite, but we always ask guests for a book recommendation, and so it doesn't need to be your favorite book, just any a book recommendation that isn't your own.

Hadley Wickham: 01:23:56 Thanks for that caveat. I was just trying to think. I mean, the book that I don't know that comes to mind the most really, which I love is Gideon the Ninth, which I would describe as it's like lesbian space necromancers.

Jon Krohn: 01:24:10 Okay.

Hadley Wickham: 01:24:12 It's a really good book. I like it. And it's written by a New Zealand author.

Jon Krohn: 01:24:20 Nice. That sounds fun. And so that one's fiction or nonfiction.

Hadley Wickham: 01:24:23 Yeah, it's fiction.

Jon Krohn: 01:24:26 Awesome. And in order to follow you after this episode, you already gave us some of the ways to do that, so I'll have the Fosstodon link in the show notes. I'll have your LinkedIn profile in the show notes. Any other ways that people should be following you?

Hadley Wickham: 01:24:40 And just keeping up with the Tidyverse blog. That's where we post everything about new releases. Cool upcoming stuff.

Jon Krohn: 01:24:47 Awesome. All right, Hadley, thank you so much for taking the time. It has been such an honor to get so much of

your time to be able to speak to a personal idol of mine in this space. What you've done for the community is unreal and that you continue to do. You have really made an unbelievable difference to so many people letting us get into the flow of programming, dig out some really cool insights and make a really big impact with data. So thank you so much, Hadley.

Hadley Wickham: 01:25:15 Thanks, Jon. I really appreciate it.

Jon Krohn: 01:25:24 So cool to be able to chat with a personal idol of mine. Hope you enjoyed my conversation with Hadley Wickham too. In today's episode, Hadley filled this in on how the goal of the Tidyverse developers is to allow data scientists to spend as much time as possible in a flow state. He also let us know how two of his favorite R packages are dbplyr for converting R to SQL and testthat for unit testing. He also talked about how plotnine allows for ggplot2 style data visualizations in Python and Siuba allows for dplyr-like data analysis in Python while Arrow, DuckDB and Keras for R, allow R and Python to be more closely married than ever. He also talked about how Quarto allows natural language and code to be easily blended together into beautiful documents.

01:26:09 As always, you can get all the show notes including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for Hadley's social media profiles, as well as my own at superdatascience.com/779. And if you live in the New York area and would like to engage with me in person as opposed to just on social media on May 17th, I'll be hosting a panel live at the New York R conference that will feature iconic open-source community members Drew Conway, JD Long, Soumya Kalra and Jared Lander. There's always pizza and beers afterward so we can catch up over a cold one. Then Hadley Wickham and other huge names like Andrew Gelman, Hilary Mason, Wes McKinney

and Sean Taylor will be there too. It's so crazy. The New York R Conference is one not to miss regardless of what programming language you do data science in.

01:26:54 All right, thanks to my colleagues at Nebula for supporting me while I create content like this Super Data Science episode for you. And thanks of course to Ivana, Mario, Natalie, Serg, Sylvia, Zara and Kirill on the Super Data Science team for producing another magnificent episode for us today. For enabling that super team to create this free podcast for you, we are deeply grateful to our sponsors. You can support the show by checking out our sponsor's links, which are in the show notes. And if you yourself are interested in sponsoring an episode, you can get the details on how by making your way to jonkrohn.com/podcast.

01:27:24 Otherwise, share this episode if you loved it with someone who might like it. Review the episode online, that I guess somehow helps us build traction on the show. Subscribe if you're not a subscriber already, of course, but most importantly, just keep on tuning in. I'm so grateful to have you listening and I hope I can continue to make episodes you love for years and years to come. Until next time, keep on rocking out there and I'm looking forward to enjoying another round of the Super Data Science Podcast with you very soon.