SDS PODCAST EPISODE 675: PANDAS FOR DATA ANALYSIS AND VISUALIZATION

Show Notes: http://www.superdatascience.com/675

Jon Krohn: 00:00:00 This is episode number 675 with Stefanie Molin, bestselling author, as well as software engineer and data scientist at Bloomberg. Today's episode is brought to you by Posit, the open-source data science company, and by AWS Cloud Computing Services.

00:00:18 Welcome to the SuperDataScience 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.

00:00:50 Welcome back to the SuperDataScience Podcast. I'm joined by the super sharp Stefanie Molin today to provide you with an episode jam-packed with practical tips on using the pandas library in Python for data analysis as well as for data visualization. Stefanie is certainly the right expert for covering this topic. Stefanie is the author of the bestselling book, Hands-On Data Analysis with pandas, which is currently in its second edition. She's a sought-after provider of hands-on pandas and data visualization tutorials at top industry conferences. She's a software engineer and data scientist at Bloomberg, the giant data-centric corporation that is ubiquitous throughout the financial industry where she tackles problems revolving around data-wrangling and visualization, as well as building tools for gathering data. She holds a degree in operations research from Columbia University, as well as a master's in computer science with a machine learning specialization from Georgia Tech.

00:01:39 Today's episode is intended primarily for hands-on practitioners like data analysts, data scientists, and machine learning engineers, or anyone that would like to be in a technical data role like these in the future. In this episode, Stefanie details her top tips for wrangling data in pandas, in what particular data visualization circumstances you should use pandas, matplotlib, or seaborn, why everyone who codes, including data

scientists, should develop expertise in Python package creation, as well as contribute to open-source projects, the tech stack she uses in her role at Bloomberg, and the productivity tips she honed by simultaneously working full-time, completing a master's degree and writing a bestselling book. All right, you ready for this excellent episode? Let's go.

00:02:27 Stefanie, welcome to the SuperDataScience Podcast. It's awesome to have you here. Thank you for making the trip to record in person with me in New York. You work in New York, so wasn't too arduous of a journey for you, I hope.

Stefanie Molin: 00:02:39 No, it was not. And thanks for having me.

Jon Krohn: 00:02:41 Yeah, my pleasure. So we met in person at ODSC West in San Francisco in the, I don't know, November or maybe late October, 2022.

Stefanie Molin: 00:02:53 It's going to be like off by, off by one over there.

Jon Krohn: 00:02:55 Yeah, something like it. It was, it's always around Halloween. And we had a lunch. There were a number of people who I think have been on the show. Matt Harrison, I think was there. Maybe Serg Masís, our researcher was there. He, of course, as always has prepared some amazing questions for you. But we ended up chatting for a really long time, and I found you to be an absolutely fascinating person who clearly knew your stuff. And so I asked you to be on the show, and now it's finally all come together. So you are at ODSC West giving a half-day tutorial on data visualization. And in addition to your brilliant tutorials, you're also an author. So I have Stefanie's gigantic book with me here. It's called Hands-on Data Analysis with Pandas. I've got the second edition. She brought me a signed copy here to film with. And our viewers of the YouTube version can see this is really gigantic. The book is almost 800 pages long, and it sounded like Stefanie would've liked to have made it more

than 800 pages, but it becomes impossible for a book to be bound, I guess at that point.

Stefanie Molin: 00:04:04 Yeah, I mean, it, it would've been slightly over, but it's just, you know, we're pushing the limits of, of what can be done with a paperback, I guess.

Jon Krohn: 00:04:09 Yeah. It's bursting with fabulous information on pandas. And because Stefanie is not only data scientist, but also software engineer, there's lots of really practical guidance on how data scientists can be better at writing software. So we're going to talk about that a lot in this episode. But I want to let you know about a generous offer that Stefanie has made. So the first five people to respond to my post about this episode. So on the Tuesday morning New York time, when these episodes come out, every Tuesday and every Friday when episodes come out, I post that morning from my personal LinkedIn account that the episode is live. And in that I will mention that the first five people who comment asking for a copy of Stefanie's book will get a free digital copy. And there are some advantages to a digital copy, aren't there?

Stefanie Molin: 00:05:07 Yes. One of them being, it weighs significantly less.

Jon Krohn: 00:05:10 Much less. You might need to buy an like external storage device to handle all of this content. But yeah, generally lighter, even that, that external hard drive that you carry around to have Steph's book Yeah. Will be lighter than the book itself. So yeah, super generous offer. Thank you very much, Stefanie. So what was your content philosophy when you wrote this book? Which you've, it's clearly been very popular. Second edition came out just two years after the first edition. So, the first edition 2019, second edition 2021. A Korean version came out late 2022, and a Chinese version is expected about a month after this podcast episode airs. So yeah, what was your content philosophy as you set down to write this gigantic book?

Stefanie Molin: 00:05:58 So, I think it was threefold. The first thing was, as I, myself was learning, I was reading a lot of different books, trying to figure out exactly how things worked and-

Jon Krohn: 00:06:08 Yeah. Before you'd even considered writing a book?

Stefanie Molin: 00:06:09 Yeah. This was just me trying to even transition into a data role. That and back way back [inaudible 00:06:14]. And you know, when I got through that process as, as I finally felt like I was able to, you know, apply some of that knowledge and get into a data role, then I felt that it would be nice to be able to give back to the community, to pay it forward, the knowledge that I had as a way to also show to myself, you know, look how far you've come that you've now have all this knowledge but also to do it in a way that I would've liked to see when I had done that. So, you know, the second point was a lot of times the examples were a bit contrived or using random data, and it's not always clear for newbies to see the connections and how you're actually supposed to apply things. So I wanted to address that. I wanted to be the book that I wanted to have back then. And then the third piece is, you would be amazed just how much you learn by having to teach. So you really find where all the holes in your knowledge are, and you figure out how to address them and then get across the message that you need to get.

Jon Krohn: 00:07:11 Nice. Yeah, all of those points resonate with me. From my experience writing Deep Learning Illustrated. It felt like the perfect resource, for learning deep learning as far as I was concerned, didn't exist. So it was nice to be able to patch that up and create what I believe is that resource. And then, yeah, also being able to patch up all the gaps in my knowledge, because when you are having to write something, you really start to learn where the holes are, the things that you kind of that you thought you knew maybe, but that you realized that they were kind of on shaky foundations. So, cool. Very nice to hear that. Thank you for creating this this book for the community. So, it's all about data analysis with pandas. What

advantages does the pandas Library have over other analytics libraries that somebody might have as options in Python?

Stefanie Molin: 00:08:04 Well, I think pandas is, is kind of a, one of the big advantages is that you're going to find it also everywhere. So it's kind of has that like, market share advantage, you know, if you need to be able to use it in one spot and go somewhere else, you're going to find lots of examples, lots of resources, a big community behind it. And it's also, you know, because it's been around for a while, it's well tested. You can trust that it's going to work, it's going to get the job done. And I find it to be especially great if you're just prototyping things. You're just exploring initially. And it's a, it's a great way to get started with your analysis.

Jon Krohn: 00:08:35 Yeah, very good reasons. And it's, it's fun, easy to use. We had Wes McKinney in episode number 523. He was the original creator of pandas. And he hasn't been working on pandas for several years now. And he described to us in that episode that one of the key reasons behind that was because as data sets have become larger and larger and larger over the years, the kinds of data sets that we're expected to work with as data scientists have grown exponentially larger. And so we're seldom working with data sets on a single machine these days. And so he started working on the Apache Arrow project, so he launched this new project that has a, so in the same way that you say everybody uses pandas, so even this Apache Arrow project, it has a pandas-like syntax because everyday scientist knows how to use it pretty much.

00:09:34 And so the advantage of Arrow though is that it allows you to easily in the backend have multiple devices working simultaneously on the data that you're working on, which is harder in pandas until very recently. So just a month ago at the time of this episode's release, it's fresh news for us at the time of recording is pandas 2.0. And so it's so fresh I don't know very much about it yet, but I do

understand that the, one of the main ideas behind pandas 2.0 is that in the backend, instead of using NumPy, which is optimized just for a single device, it uses West McKinney's Apache Arrow in the backend. So pandas is now is extensible over multiple devices. And so maybe someday in the future, Stefanie will have a third edition of your book in pandas 2.0.

Stefanie Molin: 00:10:24 Maybe.

Jon Krohn: 00:10:25 And the good news is so from the little bit that I do know above pandas 2.0 is that apparently most code will still work the same. So it's supposed to have made, you know, really no differences to the front-end API.

Stefanie Molin: 00:10:42 It's a relief for me, because a big project coming up this weekend is updating all of my content.

Jon Krohn: 00:10:47 Thanks. So let's talk about how we use pandas. It's clearly great for just wrangling data. The data frame structure is native in it. It's actually something that, I don't know, I was an R user before-

Stefanie Molin: 00:11:03 I was too. Yeah.

Jon Krohn: 00:11:04 Yeah. And so data frames were a really convenient way of working with two dimensional tables of data in R because you can have different data types for each column. So if, you know, before I was using R I was using MATLAB, and in MATLAB you had to specify the type, at least at that time of the matrix. And so you could have a float type matrix, and then everything in that matrix had to be float values. But in R it was cool having these data frames where you could have a column that was text to describe what the row is or to be an ID. You could have column names and then each column could be a different data types. You could have a boolean like true-false column and a float column and an integer column and a text column and whatever all mixed together. And so pandas allowed you to have that in Python. But so data-wrangling

is this really core part of working with pandas. Is there any aspect of data-wrangling that you highly recommend to our listeners in pandas that they might not know about?

Stefanie Molin: 00:12:12 I guess a few things come to mind, more than what we initially discussed. So one big thing is chaining. I think the more you use pandas is the more you'll see that just chaining the operations together is going to save you a lot of effort and just make a lot cleaner code. You don't have end up data frame 1, 2, 3, 4, 5, 6, and you don't remember which order you ran them or which one you're working with. Assign method that's going to come big when you, when you're using the chaining. I'm a big fan of that. Most of my code will end up having assigned calls at some point. I feel like people maybe aren't so familiar with that method.

Jon Krohn: 00:12:48 Yeah, I'm not that super familiar with-

Stefanie Molin: 00:12:50 Oh, so-

Jon Krohn: 00:12:50 So like, so the chaining thing I get, so that's actually, so going back to my R days even the dplyr library is great, right? For, for chaining operations together. So then it becomes very easy to look at your code or similar if you can do it in bash with a pip operator where you have operation after operation after operation. And it makes it very easy when you come back to your own code or you're reviewing somebody else's code to see, okay, we have this processing pipeline, for example, or this analytical pipeline where instead of having to name a new data table for each step of the way, you just flow very nicely through all these operations. So it makes for super readable code. So I definitely agree on the chaining as something that everybody should be doing. And Matt Harrison, who we were saying was at that lunch back at ODSC West. He's in episode number 557 of this podcast. It was actually the most listened-to episode in 2022. And a big point that he was making is that everybody should be chaining, in that

episode. It was, the episode was called Effective Pandas and chaining was his like, number one thing that people should be doing. But then you mentioned assign.

Stefanie Molin: 00:13:59 Assign. So when you're chaining, let's say you want to create a new variable that's based on some other columns in there, and you want to create a new column that's just part of the flow. You can create 10, you can overwrite three all-in-one call, and then move on.

Jon Krohn: 00:14:13 Cool.

Stefanie Molin: 00:14:13 And it works really nicely with, if you like, for some reason want spaces or like, let's say you're going to then do a plot right at the end. So chain all the way to a plot and you want to have nice labels. You might want to have space instead of an underscore. And you can actually do that if you just put it in a dictionary and then unpack it as you're going. So that's a neat trick to have up your sleeve, save you some steps.

Jon Krohn: 00:14:34 Cool.

00:15:09 This episode is brought to you by Posit: the open-source data science company. Posit makes the best tools for data scientists who love open source. Period. No matter which language they prefer. Posit's popular RStudio IDE and enterprise products, like Posit Workbench, Connect, and Package Manager, help individuals, teams, and organizations scale R & Python development easily and securely. Produce higher-quality analysis faster with great data science tools. Visit Posit.co—that's P-O-S-I-T dot co—to learn more.

00:15:12 That sounds awesome. You said you had a few, were those the-

Stefanie Molin: 00:15:14 I think the third one maybe is just like plotting that you can plot in pandas. I think sometimes people are surprised that that's there. It definitely feels like a natural

extension. Now I will say that because it is such a high level on top of all these other things, sometimes things aren't quite plugged in yet the right way. I have fixed a couple of those-

Jon Krohn: 00:15:33 Personally.

Stefanie Molin: 00:15:34 Oh, yes. Personally. So, open-source contributions, so if someone else finds them, you have to, it's your responsibility to fix them. Help us all out.

Jon Krohn: 00:15:43 Yeah, we're going to talk a lot later in the episode about the enormous number of open-source contributions you've made to all of the major data science libraries in Python. But in the meantime, let's quickly jump into this visualization thing. So yes, people don't always think of pandas as their go-to visualization library, but it can do a surprising amount. And you, I realize that there's not always your workshops and your book don't necessarily overlap, but I know that your workshops if you do kind of like your full workshop, it ends in visualization.

Stefanie Molin: 00:16:24 Yeah. So I have two workshops. I have a Pandas Workshop that's, you know, getting you started in pandas. So it's almost like if this book had like...

Jon Krohn: 00:16:34 Prequel?

Stefanie Molin: 00:16:35 Kind of, yeah. Like, like that's like your quick whirlwind tour and then you can dive into the book and like really get a deeper understanding.

Jon Krohn: 00:16:42 You don't, you don't cover 800 pages of content in a half-day workshop?

Stefanie Molin: 00:16:45 That feels cruel. I've actually managed to learn how to speak on double speed. So, buckle up everybody. No. So the workshop is more of your beginner workflow, you know, getting started, just seeing kind of what's available. Some examples of commonly used, like very, very

common things getting you to the point where like you understand what a data frame is, you understand how to get data into the data frame, you can wrangle it and at the end, you know, time permitting, we just do a little bit of plotting. And then my second workshop is just on data visualization. So that one I just assume, well, [inaudible 00:17:20] much assume that you know pandas, but like, it's like, here's what our data look like that we're going to plot. Don't worry about how we got here because the focus is how do we take this and actually make a visual out of it. And because pandas is so high level with the plotting, there are sometimes things that you just, you just can't do unless you then go down a layer and actually work with matplotlib directly. So it's very important to be comfortable with that.

Jon Krohn: 00:17:42 Nice. And I do have specific questions about visualizations, matplotlib, why you might use pandas matplotlib or other libraries in certain circumstances. But before we get to that, I want to quickly, you weren't expecting me to ask you about this, but when we were preparing for this episode, I learned something really interesting about you in terms of how you prepare for workshops. So first of all, tell us about how, depending on the projector-

Stefanie Molin: 00:18:05 Oh yes.

Jon Krohn: 00:18:07 You render different slides. So fill us in on that. And then you also filled us in on some really cool stuff you're doing with jQuery.

Stefanie Molin: 00:18:15 Yes.

Jon Krohn: 00:18:15 For interactivity in your presentation. So I think our technical listeners will probably find both of those Interesting.

Stefanie Molin: 00:18:20 Okay. And this is not in support or against jQuery, just to clear that up. So I think, so I started getting, I started

presenting at conferences I guess 2021 and they were all virtual. And as someone who was not comfortable public speaking at all, that was great. Right. But there are some interesting challenges then when you take that on the road in person, right? So when you're sharing your screen, you know the specs, you know, everything fits on the screen, it's the right size, everyone can read it because if they're not, if they're standing too far away from their laptop, they can fix that, right? With a physical room, not so much. So the challenge I had with the seconds so I was speaking at Python last year and they sent a very detailed email with a specs of their projector and the minimum sizes and contrast everything should be and I'm using reveal.js. So, if you create a notebook, a Jupyter notebook, you can actually define which cells should be slides and like subslides and fragments. And then from there you can just export what you have into a slideshow.

Jon Krohn: 00:19:30 You make your slideshows in Jupyter Notebooks.

Stefanie Molin: 00:19:32 Yes.

Jon Krohn: 00:19:33 Wow. And then reveal.js is the, is what allows you to do that alchemy.

Stefanie Molin: 00:19:40 Well, so it's a little bit more involved in that. So basically like baked into Jupyter Notebooks. So if you've ever done like the file what is it, like Save As or Download As?

Jon Krohn: 00:19:49 Yeah.

Stefanie Molin: 00:19:49 There's an option in there for reveal.js slides. There's a bunch of options. And so it's through MB Convert, if you're familiar with that?

Jon Krohn: 00:19:56 Yeah, yeah, yeah.

Stefanie Molin: 00:19:57 So then there's templates and you have to be careful because periodically you update Jupyter, you get a new

template. Maybe you don't like the CSS that got changed. Something's now too big. Header 1 is massive. That happened to me right before a presentation. Header 1 just like absolutely massive and then like, this doesn't fit on your screen. And so I spent probably a month trying to come up with some way to alter the slides that would work on a bigger screen while also still having the ones that work on the smaller screen.

00:20:28 And so you end up having to have a lot of challenges where like, now you need, like, images have to be SVG. Because if you're showing a PNG that gets created with like 100 DPI by default, and you try to blow that up, that's going to look really bad on the high-resolution projector, right? So part of that then is taking the template and, you know, flexing your CSS muscles and finding everything that you want to change that you want to make sure is yours, you know. Then I have like a lot of hacks in the template itself using jQuery. So some of the things are to create like, different named pages. So if I want this page to be like, easily referenced, I can say like, oh, this is section 1 and then I can map that to a Key 1. Or this is talking about the assign method I can have Slash assigned and then it goes right there and that's automatically computed based on me just adding a tag to that cell. And then the final thing, which I'm really excited about, will be released by the time this episode airs is having some more speaker tools, I'll say. So I'm going to have like an exercise timer built-in using jQuery things just to give me a warning or heads up. Like, you know, we're nearing time.

Jon Krohn: 00:21:43 And, and that happens with like hotkeys?

Stefanie Molin: 00:21:45 Yes.

Jon Krohn: 00:21:45 Yeah. And so I suspect, well, I mean maybe you're not going to do it because it's jQuery and you don't feel a hundred percent.

Stefanie Molin: 00:21:54 It's very hacky. All of it's very hacky.

Jon Krohn: 00:21:55 Yeah. But do you think this is going to be open-source maybe someday?

Stefanie Molin: 00:21:58 I'd actually do want to, because ... so I have two workshops I mentioned, right? So that means anytime I figure out something, oh, this is great, I'm going to add this. It's prototyped on whichever one I happen to be working on at the moment. And then it has to be poured over. So it just so happened that I was working on the pandas one. So it's been sticking on like it's been on a branch on my personal computer for a while, I haven't pushed it up. This weekend, hopefully.

Jon Krohn: 00:22:19 Nice. Yeah. Push it up to GitHub.

Stefanie Molin: 00:22:21 Yes. But I want to have an open-source, because I plan on doing more workshops on different things. Not necessarily data, essentially I'm thinking about one on like pytest. I've started exploring that a bit. And so having something almost like a workshop and knit where it then brings in everything I need. Because there's a lot of this I think is like, this is the standard, like of how I want it to look. And I want it to be consistent, right? I actually have something kind of similar, but just for myself. Is like, I have an intro slide that I show at any session that I give. So when people come in, they know where to find the content. And towards the end of last year, I just made like a small little wrapper around that. So it just, it's like a shell script that asks me, where are you presenting? And I typed that and it says, work, which workshop are you presenting? And I say one or two and it just creates a slide for me. And then I have just have to show that. so.

Jon Krohn: 00:23:14 Nice.

Stefanie Molin: 00:23:16 I'm all about automating everything I possibly can.

Jon Krohn: 00:23:19 You sound like a software developer. Yeah.

Stefanie Molin: 00:23:22 What a coincidence.

Jon Krohn: 00:23:23 Very cool. All right. So we'll have to look out for that library from you in the future. Yeah, so that, you know, we can also be automating aspects of our presentation creation. Nice. So before I was forcing you to talk about all this technical content on automating presentations, which was super interesting, we were about to get into visualizations. So, visualizations play an essential role in data analysis, especially if you're doing some exploratory data analysis. You've never seen a data set before. It can be helpful to plot some things out. You can see things right away. In your book you say it's easier to find patterns and visualizations, it's more work to arrive at the same conclusion by looking at numbers of tables. And this is a really obvious thing to me, I think, I think it's easy to agree with that statement when you're just looking at tables of data.

00:24:18 It isn't always obvious the patterns emerge, but once you plot things it can be super obvious that there's some, some you know, relationships between variables or the shape of distributions can be interesting or outliers can appear. You just learn so much about your data from visualizations. So, visualizations are actually so important to you, you've been having a lot of fun lately with your Data Morph library, which shows some kind of interesting quirks about summary statistics and visualizing 2D data. So, I'll be sure to include a link to the Data Morph GitHub project that you have made. And so people can see a visually, but so for example, at least at the time of recording at the Read me at the top of it for the Data Morph project, there is a 2D plot, a scatter plot of a panda, and it turns into a 2D scatter plot of a star, and then it can turn back into a panda.

00:25:29 And the dots just kind of all gradually move. But you have these summary statistics on the screen about the mean of the points on the X axis as well as the Y axis, about the standard deviation of the points on the X axis

and the Y axis, and even the correlation between X and Y for the points and all of those things to two decimal point precision stay exactly fixed as you go from panda face to star face, back to panda face. It's fascinating to watch happen. Why did you think to do this and what's the point of it?

Stefanie Molin: 00:26:02 So, it actually started with my workshop, the Pandas Workshop. So, I feel like when you're just having someone or just introducing someone to pandas and you spend an hour and a half talking about wrangling, and that's really where pandas shines, people maybe who are new to the announcement, then well, why do I need to visualize anything? I have these statistics. And I feel like that's also, you know, plotting is significantly harder than just calculating the statistics. So there is a tendency to like, I'll just do what's easy, you know, this is a quick way to explain it, and it's, it's very dangerous. And so I see data more as a, it's a teaching tool. It's not, it's not my idea. This, this idea of having data sets that have the same [inaudible 00:26:47] but look very different goes back.

00:26:50 There's examples maybe we'll put links to, I think it's actually in the book too, Anscombe's quartet, which is a set of four data sets and I think there's maybe 10 points in each. But it shows how like you have like a line where there's just one point off that's an outlier and that can have the same summary statistics as line that looks very, very different. And so in 2017 researchers at Autodesk made a thing, well, they take, they use a Simulated annealing and they take a dinosaur scatter plot, and then they morph that into other images. And so I was thinking, can I make like a fun thing here at like a panda converting into another thing? And so I spent a weekend diving into like the very research-y code and trying to get it work for something else because it was clearly like hard-coded, like this works for the dinosaur.

00:27:38 And I think at maybe at some point they had plans to work more on that, but I don't think it happened. It, it

was a challenge. So, I got it to work and then I was really excited about it. I had my image and then I was like, but this would be so nice if other people could also just play around with it just to see how this kind of thing works. And you know, it's a fun gimmick to have for presentations, right? So I spent like the last three months or so building this. But it's been a really insightful experience for me also on the software engineering side. So like, this was a lot of refactoring code, thinking about ways that, you know, I can get this to work for any data set. So how could I take in a data set and determine where to draw the circle that it should morph into?

00:28:24 So, you know, you take the means of each direction, right? But then when you're, you may be making a bullseye. So there's two circles now. How do you decide what the radius should be and how do you vary them? Things that are like the star where, how do you decide where to place it? It's all ratios and it's dependent on the data.

Jon Krohn: 00:28:40 Yeah.

Stefanie Molin: 00:28:40 And so there's a lot of like interesting thought process on that.

Jon Krohn: 00:28:43 Yeah.

Stefanie Molin: 00:28:43 And then also figuring out how to do like a very clean test suite, which most of the time you won't see in the wild because you're just doing it as you go. Having a chance to come in on something that's small and contained and then build that up. And the same thing with doing docs and publishing a package. And it's all stuff that you don't necessarily do or see when you're working internally at a company. You have your own company's processes, you're working on that way and you don't necessarily have the know-how of how that happens in the wild. And so it's, it's been a very enriching experience and I've

learned a lot and I've brought a lot back to my work as well from the experience.

Jon Krohn: 00:29:19 Nice.

00:29:19 Are you stuck between optimizing latency and lowering your inference costs as you build your generative AI applications? Find out why more ML developers are moving toward AWS Trainium and Inferentia to build and serve their Large Language Models. You can save up to 50% on training costs with AWS Trainium chips and up to 40% on inference costs with AWS Inferentia chips. Trainium and Inferentia will help you achieve higher performance, lower costs, and be more sustainable. Check out the links in the show notes to learn more. All right, now back to our show.

00:29:58 Yeah. So it's useful for your visualization teaching because it's a data visualization, clearly. It's useful for your other kind of teaching because it shows this statistical effect, and how lots of different shapes, including in this case, animations of shapes, morphing, data morphing between different shapes and there's like a dozen or so different kinds of shapes that you have out of the box that work-

Stefanie Molin: 00:30:25 More to come.

Jon Krohn: 00:30:26 And more to come. And you can contribute-

Stefanie Molin: 00:30:29 Yes, yes.

Jon Krohn: 00:30:29 Your own as well now because this is, this is a live GitHub project and yeah, it will have a link to it in the show notes. And so yeah, so as a statistical concept it shows how so many different kinds of graphs, so many different kinds of shapes can have the same mean along both axis, the same standard deviation along both axes and the same correlation. So that's super interesting. And you use something called simulated annealing in order to

allow this animation, this morphing to happen. What is, what's simulated annealing?

Stefanie Molin: 00:30:59 So, simulated annealing is a, an AI technique. So you start out, and I think it is, easiest to me to always think about this in terms of like particle movements, I guess. So if you think about something when it's in a gaseous state, things are moving very, very fast. So you can even see in the visual, which people who are listening can go and, and look. But at that point you're, you're more willing to accept points that maybe move farther than you would like from your target shape. And then over time you're kind of going more into the solid state in that sense. So the temperature, you're actually using temperature, so that is decreasing over time. And so over time you're, you're more strict with what you'll allow. So initially you get like bigger movements and then it kind of like slowly will converge onto the shape that you want.

Jon Krohn: 00:31:45 So temperature is a variable-

Stefanie Molin: 00:31:48 Yes.

Jon Krohn: 00:31:49 In the simulated annealing, and as temperature increases the particles, the dots in-

Stefanie Molin: 00:31:53 Temperature, you decrease the temperature. So you start high and then you decrease it down.

Jon Krohn: 00:31:57 Right, right, right, yeah. But so just kind of generally speaking, if you, if you increase the temperature, the particles move more.

Stefanie Molin: 00:32:02 Yes.

Jon Krohn: 00:32:03 Just like heat in actual particles in real life.

Stefanie Molin: 00:32:07 Yeah, yeah, yeah. So, it's more like at higher temperatures you're more willing to accept something that's like bad, right? So maybe you move the point in the

wrong direction. But the reason you need this is because otherwise, you can fall very easily into a local minimum. So you want to make sure you're, you're properly optimizing. So it's kind of, it's like a hill-climbing algorithm.

Jon Krohn: 00:32:27 Cool. Very interesting. So check it out Data Morph package. It illustrates how, yeah, it's staggering. All these different kinds of shapes that have these fixed means, standard deviation, correlation. And you can watch in real time as you morph between these shapes with all those summary statistics being fixed. And so yeah, so the reason why I brought that whole project up is because clearly you're really into visualizations and it also gave our listeners a taste of these software development principles that you instill to your students and in-person as well as in your book. So things like test cases and just creating open-source software, well documented.

Stefanie Molin: 00:33:12 Object-oriented programming. That'll make it easier. If anyone is out of shape, it'll now be much easier than it would've been in the past because of that like small modular reusable code.

Jon Krohn: 00:33:21 Nice.

Stefanie Molin: 00:33:21 Very important.

Jon Krohn: 00:33:23 So back to the visualizations and the way that you talk about them in your book. You provide an overview of different plots available in pandas and matplotlib. And then in chapter 6 you introduce seaborn for relatively advanced plotting. So what are the kinds of things that people might want to do in pandas directly? When would you want to get into matplotlib? And then in what situations should you be considering using these relatively advanced plotting techniques in seaborn?

Stefanie Molin: 00:33:59 So I think it kind of depends on what your data looks like. So if you're working with wide-format data, then you want

to just prototype something, then pandas is definitely going to be a very natural fit because it's just part of your chaining, no need to reshape the data. But when you want to do something beyond that initial just plot, show the things, or maybe there's certain things that you can't necessarily pass down, like removing the spines off of the plot, then you'll have to reach for matplotlib. And so I think it's, it's very nice to have the [inaudible 00:34:28] dependence because it's the higher level you just get a feel for what goes where and how things plug in. And then you can go into matplotlib and see exactly how things are happening under the hood and where you can plug in or where you can tweak and refine what you, what you're looking for.

00:34:45 And then seaborn, if you have, instead of the wide format data, the long format data or some kind of mix. So very good example, and where I will typically reach for seaborn quickly, is if you have a column that maybe you want to color the data by, that's going to be very painful in pandas. Because you'll have to reshape everything and it might not actually necessarily be that easy to do for what the visualization you're going for. And seaborn makes that very easy. One of the examples I actually have in my data visualization workshop is we take Stack Overflow questions that were tagged with a set of data, well yeah, data libraries and Python. And we show their growth in the questions that are tagged like that over time. And what's interesting is if you look at the questions that were tagged with seaborn, but were actually before, like they were created before seaborn was ever created, you see people will go back later and then tag it seaborn once there's a better way to do something. And one of those, the first one I think I have, or one I call attention to at least, is about coloring the data. So, you can find some pretty nasty-looking code that you'll have to do and looping just to get it in matplotlib and it becomes a trivial one-liner in seaborn.

Jon Krohn: 00:35:59 I also, I find that seaborn is just prettier.

Stefanie Molin: 00:36:01 It is very nice. So that comes from the style sheets though. So you can use with matplotlib, the seaborn style sheet. They have like cross boarded it, so.

Jon Krohn: 00:36:10 Got it. All right. Nice. So pandas makes it very easy to get started on plotting, matplotlib allows for a bit more advanced control over the plots. So for example, if you want to move the tick marks, that kind of thing, you can do that matplotlib, and then seaborn, we have even more flexibility, particularly with respect to things like colors. Working with colors is easy. And I think, yeah, in general aesthetics are very, it's even the defaults in seaborn ... like that, especially like the defaults in seaborn the same kind of plot relative to matplotlib, it's, it's just, it tends to look really pretty out of box.

Stefanie Molin: 00:36:46 Yes.

Jon Krohn: 00:36:46 And before we started recording, you mentioned to me something about matplotlib, and I can't remember what this was, but there's something that you're like, there's something that people should know about matplotlib that most people don't.

Stefanie Molin: 00:37:00 Ticker.

Jon Krohn: 00:37:01 Ticker?

Stefanie Molin: 00:37:01 The ticker module in matplotlib. I always have this, it's in the book, it's in my trainings. So there are oftentimes where you're maybe visualizing a quantity that maybe has a set of units or let's say percentages, and you have now zero to one, or do you multiply by a hundred? Or if you wanted to normalize based on a constant factor, you can actually just use the ticker and do all that for you. So you simply, you import ticker, obviously, and then you just hook it up to the axis that you're, you're working with. So you can say in this case on this plot that I'm working with, the Y axis is a percent. So you pass it the percent formatter and you tell it what's the base. So you just say

1, that means they're all, they're already percentages. So it'll multiply by whatever factor it needs to and [inaudible 00:37:47] on the percent for you. And so you have already a huge improvement in how the plot looks. And then within ticker you have, you know, various kinds of ways to format the ticks and also to place the ticks. So maybe you're working with data that's clearly, strictly integers, but maybe your scale is small enough that matplotlib shows you fractional units and you don't want that there. That's easy to correct with ticker.

Jon Krohn: 00:38:13 Nice. So the ticker module makes it easier to format and place ticks in your plots, ultimately making it easier for you to convey some specific concept to your viewer.

Stefanie Molin: 00:38:24 Yes.

Jon Krohn: 00:38:24 Perfect. Nice, all right. Thank you for all those visualization tips. Another thing that you are obviously expert at, and you cover a lot in your book is statistics. So, we even talked about this a little bit. So, you know, summary statistics like means, standard deviation, correlation, these are the kinds of concepts that people know. But statistics goes a lot deeper. There's a lot of frequentist, Bayesian techniques to be analyzing data that, you know, we can solve a lot of the same problems especially if they're on relatively small data sets with statistical approaches as opposed to machine learning methods. But statistical approaches, they can have really useful things with them that we rarely get in machine learning, like p-values, confidence values that allow us to have some an estimate of how much we should trust say an experimental result.

00:39:19 Could be an AB test on your web platform where you could say, oh, you know, it is statistically significant this difference between user behavior in case A and case B. And so therefore we can feel comfortable moving ahead with this product decision or in more serious things like, this seems like this drug works and we should give it to

patients. AB testing and saving lives, almost exactly the same [crosstalk 00:39:47]. So statistics hugely useful, and probably when people are getting started in data science, there are so many different things for them to tackle. There's the programming things like just learning pandas, there's learning the theory of statistics and understanding the importance of that. So what do you think, if someone's getting started with data analysis, do you think that they should dive more into statistics first or dive more into the programming aspects like understanding how pandas works?

Stefanie Molin: 00:40:19 I think that depends on, you know, first of all person's learning style. Maybe where they're more interested? You might be more interested on the coding side and more interested on the statistics side. You don't want to get discouraged, so maybe if you're really interested in statistics and the coding scares you a little bit, maybe do some of the statistics get really invested and interested. And then you can take a step back and say, okay, how do these statistics map to different methods in pandas? And then maybe the reverse too, if statistics is a bit overwhelming and you want to approach it from the code side I think it's just important kind of just to do what works for you and stay positive about it.

Jon Krohn: 00:40:55 Yeah. And I think it's great to do them together. For me personally, I love learning the stuff hands-on as I can imagine somebody who wrote a book called Hands-on Data Analysis, would agree. And so yeah, for example, I created a Statistics For Machine Learning course and I am publishing that. I published the first few lectures on YouTube like a year ago. And then I've been completely overwhelmed with work. And we'll get back to that, I promise those of you who are following on YouTube or on my Udemy course but people can get it from O'Reilly, it's like an eight-hour course and it's, and everything is done in Jupyter Notebooks in large part with pandas because it's such an easy way to just play around with things and say, oh, like what if I change this parameter? Or what if I

change this data set? And it just makes it so easy and fun.

Stefanie Molin: 00:41:43 I think one thing you have to be a little careful about in the panda side just came to mind is just how different libraries in the ecosystem treat like standard deviation, whether it's sample population and-

Jon Krohn: 00:41:54 Right.

Stefanie Molin: 00:41:54 It's not always obvious what's there. So yes, read, read the docs before you dive in and get confused about why math answers don't match.

Jon Krohn: 00:42:01 That is something that so we end up doing that in my course. So there's like, there's a degrees of freedom parameter-

Stefanie Molin: 00:42:09 Yeah.

Jon Krohn: 00:42:09 That I remember. I can't remember which way it goes, but it was like, yeah, if I calculate the standard deviation in NumPy using the default parameters, if I want to get the exact same number in pandas, then I need to specify these degrees of freedom to be slightly different.

Stefanie Molin: 00:42:24 Yeah.

Jon Krohn: 00:42:24 So yeah, that's a really good catch. Yeah. So in addition to statistical analysis in your book, you also specifically dig into financial analysis. So you dedicate an entire chapter to financial analysis using the Stock Analysis Package that you built. So can you explain the motivation behind creating this package and how it can help users?

Stefanie Molin: 00:42:48 So I think there's a couple pieces to that. So the first thing is there's a lot that can be done with analyzing finance. There's tons of metrics that can be calculated and for the most part some of them can be explained easily. And I think it's something that everyone at least

can understand, you know, why you would want to analyze that or like the different aspects of what you would want to be looking for. And then I chose to make a package, and actually I didn't say this before, but this, when I had the idea for this chapter, I think I was on vacation and I had to ask for an extra two weeks I think, to work on it because I was like, oh my God, I have the best idea. I'm going to make a package. And it's going to show people how they can make reusable modular code they can share within their team, that just to show how you can be a person who's working with data analyst, a data scientist, but still understand what it takes to write good code.

00:43:47 Because that's only going to help you later when you revisit it, understanding what you did, why you did it, being able to explain it to people, other people to benefit from it. And so the way I actually structured the package is so you can see a bunch of different concepts. So there's like a static class and like why it would make sense to do something like that, why you would need to source some information in the init for like reading things in. And so it tries to hit on a bunch of different concepts to give you a good breadth of knowledge of how you would structure, how you would make a package, you know, how you would build classes, how you would approach the problem. And it's meant to be something that people can fork and then build upon and like, oh, I want to add this other method. I want to see how this works. So kind of like a Lego starter kit, if you will.

Jon Krohn: 00:44:31 Cool. That's great. Yeah. So the purpose of this package, maybe wasn't primarily about giving somebody a financial analysis library, it was more about providing the Lego blocks.

Stefanie Molin: 00:44:43 Yeah.

Jon Krohn: 00:44:43 The building blocks for understanding how to build great open-source software packages that are shared in GitHub

and that you know, follow the standards, the highest standards that software developers expect and that maybe a lot of data analysis data analysts and data scientists aren't familiar with already.

Stefanie Molin: 00:45:02 Yeah.

Jon Krohn: 00:45:03 Cool. And you took a three-week vacation to do that.

Stefanie Molin: 00:45:06 No, no. And it wasn't taking the vacation to do it. I think I had the idea while I was on the vacations. I think it was kind of like I had to finish before the vacation, then I had my time and then I was like, I'm, this idea's too good and I'm not going to make the deadline. I have to say something. But I think, I think it was, it was worth it because I get a lot of-

Jon Krohn: 00:45:25 The book deadline?

Stefanie Molin: 00:45:26 The book deadline.

Jon Krohn: 00:45:26 Ah, I see.

Stefanie Molin: 00:45:27 So I get a lot of comments.

Jon Krohn: 00:45:29 Sorry, I thought you were asking for two more weeks off of work.

Stefanie Molin: 00:45:31 Oh-

Jon Krohn: 00:45:32 To work on it. You were asking for a two-week extension.

Stefanie Molin: 00:45:34 Well, because I was only working out on the weekends, so, you know, you imagine two weeks it's just like another two, [inaudible 00:45:39] few days. So it's just not much. But I, but I frequently get like comments when people say that, you know, they like the book and what were your favorite chapters? It's almost always that is one of them.

Jon Krohn: 00:45:48 Oh.

Stefanie Molin: 00:45:48 Just seeing how you could, it's not so much like, because other stuff is going to be available in a lot of spots. Like, I can learn how to use different aspects of the library once they've seen it the first time, they can console other things. But seeing how you would structure things like in practice and that's like a very hands-on chapter and that's a big thing that people highlight.

Jon Krohn: 00:46:10 Sweet. Yeah, that sounds super useful. It sounds like something that I should personally be brushing up on. The data scientists on my team would probably appreciate it, and not to mention the software developers at our company. So clearly open-source is an important thing to you. We've touched on it a number of times in this episode, including just now with this topic of this financial analysis package and how people find it so useful for understanding how to create these open-source packages. So you've contributed to many of the key open-source libraries that data scientists use, including pandas, scikit-learn, matplotlib, seaborn, and NumPy. So what drew you to maintain all these open-source tools, the data scientists know and love? Yeah. What, what galvanized you to do that in the first place?

Stefanie Molin: 00:46:59 So I think being a software engineer that's, it's kind of my nature. If I see something that's not working or an example of like, this doesn't make sense and you try to figure out how to actually use it can I improve the documentation? Can I fix the bug that's in there waiting for it to be fixed versus fixing it yourself? It's, you know, you're going to wait a lot longer and then you can just get it fixed very fast. And it's also, you know, part of wanting to be part of that community, right? So you, you're using that stuff, you're benefiting from it and the whole culture of open-source is to then, you know, pay it forward and give back. You see something's wrong, you go fix it and everyone wins.

Jon Krohn: 00:47:36 Yes.

Stefanie Molin: 00:47:37 And it's also just a fun experience to explore. You learn more about the library when you have to dig into it. How am I going to add this functionality? How does this thing work?

Jon Krohn: 00:47:46 Yeah.

Stefanie Molin: 00:47:46 Even the documentation examples, it's like you can find what they call meta issues [inaudible 00:47:51] they'll list a library like, oh we need document, we need examples for all of these methods. Just pick one. And I actually picked some and NumPy has one right now on masked arrays and I picked one in there and it ended up in Data Morph because I'm like, this is exactly what I needed. So you can, you can easily find things that you have no idea were there, but just because you [inaudible 00:48:13], oh, I'm going to help them make an example for here, you have to figure out how it works, why would you use it and make an example.

Jon Krohn: 00:48:17 Yeah. So it's one of those selfishly benevolent things that people do, like writing a book, where you're like, I'm going to do this great thing for society, but simultaneously you're learning so much more in-depth that you become a much better expert yourself.

Stefanie Molin: 00:48:32 And then you do sometimes get praised. So.

Jon Krohn: 00:48:33 Yes.

Stefanie Molin: 00:48:33 One of the things I added while procrastinating for one of my final, or maybe it was a midterm exam, my master's degree, I was scrolling through the seaborn issues and I saw one where they wanted to add horizontal and vertical reference lines to plot grids. And this was like at this point in time, like a year or so old, and I was like, I can do this. And so put off the study and I quickly implemented that and it's something that I always highlight in my workshop that after an hour in, everyone can make that contribution. It's just a matter of understanding the

underlying library. And you know, right when that came out, there was a, it was a tweet, the, I don't remember who it was from now, I could show you later, but, but there was a tweet a guy who runs a blog. He was like, oh, this great rough line functionality is now in seaborn and

Jon Krohn: 00:49:22 Nice.

Stefanie Molin: 00:49:22 Then the creator tagged me. He is like, oh thanks Stefanie. And like, that, that was an amazing experience too.

Jon Krohn: 00:49:28 Yeah. That must be super rewarding. Particularly if you are using a library like one of these libraries that most data scientists use and that you've contributed to. You come across this issue that everybody may or may not be aware that they're experiencing. And then you fix that and you can see downstream when that gets integrated into the major release. Yep. And it's fixed. Like knowing that you did that yourself, that must be satisfying.

Stefanie Molin: 00:49:53 Oh, it's amazing. And it's very nice because everyone is, is wants contributions. They're going to thank you. If they can give you a shout out, you'll get a shout out. So it's, it's a very rewarding experience.

Jon Krohn: 00:50:04 Sweet. So it might be mind-blowing to our listeners that with all of the open-source work that you do and all the writing you do and all the teaching you do, that none of that is your day job. So you do all of this stuff on the side.

Stefanie Molin: 00:50:19 Yes.

Jon Krohn: 00:50:19 So you work at Bloomberg, which is one of the world's leading providers of financial data if not the leading global provider, I think is certainly the biggest brand in that space, as a financial data provider. And your role there, as we've alluded to, combines both software engineering and data science. So you research and develop solutions to help improve and automate Bloomberg's information

security processes using data and machine learning. Given that it's information security, my expectation is that you cannot tell us very much about this job, on a podcast. Please list the most vulnerable... No, but if you could, if, if there's anything about the role that you can tell us, you know, the kinds of at a high-level software libraries that, that you're using, or maybe things that have been open-source from work there. I don't know if there's anything you can share us, share with us. I'm sure it'd be interesting for our listeners to hear what, what's involved with being a software engineering and data scientist at Bloomberg.

Stefanie Molin: 00:51:24 Yeah, so my team, we're a small team, so it's very scrappy and, you know, you're going to be learning lots of new things before going there. I didn't know how to do like the infrastructure side of things, so I learned how to like, set up cloud machines and like configure them and get them working with like the full stack.

Jon Krohn: 00:51:40 DevOps.

Stefanie Molin: 00:51:42 Yeah, Apache. And then moving into, you know, Python is our backend and then we have React as our front-end and you know, also in interfacing with the database. So it's the full stack. I also do some D3 work. So one thing I can share is, I guess, last year I was working on a what I call an explorable sankey. So in D3, if you were just to throw your data at it and make me a sankey, it's like, sure, but if you give it too much data, then it crashes your browser. So, not so, not so great. So what I decided to do, and this actually came as an idea from a coworker who I was creating the visualizations for, said, it'd be great if you could just click through and explore it. And it was something that it, you know, it sounds like a cool idea, like a pipe dream, right?

00:52:34 But you know, at the time I was you know, doing my master's and so I guess like just more thinking about the data structure side, it happened to be like a data

structure course, and that was kind of what cracked it open for me. I was like, wait, actually this is, this is definitely possible and it's just creative use of the data structure. And so what I am able to do now is you can give it a ton of data and it knows, okay, I'm going to show three levels of this sankey, but I need to do a roll-up here and group this into other. And then if I want to see what's in there, I can maybe hover over it to get a peek or click into it. And that's a way that now someone can explore their data visually rather than having to like mine through rows and rows of textual information.

Jon Krohn: 00:53:14 That's also, it's a perfect case study what you just described, of how despite us moving to this world where algorithms like GPT-4 can write great code, and there's, you know, I have people commenting on YouTube videos for the podcast saying, is there any point in me getting started in a data analyst job or data science job now that there's all this automation and what you've said, there just is a perfect case in point on one of the kinds of things that is not going to go away for the foreseeable future, which is understanding the fundamentals of our field. So, you know, I've created content on linear algebra calculus, probability theories, statistics, and data structures and algorithms for, in my case, machine learning specifically. So all this, you know, this is content that I've released. It's all available in O'Reilly today.

00:54:05 And we'll eventually all be on YouTube. The linear algebra and calculus already is, and the probability theory stuff should start coming out again soon. And eventually I'll get through all that content and it will all be publicly available for free. But the, so you gave this perfect example of how your understanding of algorithms and data structures from the formal master's that you were pursuing in machine learning provided you with this, oh, this thing that seemed like a pipe dream is actually just a, we just need to understand the data structure and we can make this thing happen in real life.

Stefanie Molin: 00:54:40 It's actually just clever manipulation of that. Like, how can I take, you know, maybe this massive graph and then like selectively group bits that act like this is a subgraph, but you're not actually showing them. And it, that's really, it's amazing how when you have these breakthroughs, like, it's like, oh, that was actually so simple, but it's the thinking outside the box that, like you said, I think is, that's always going to be something that, that we have the advantage for.

Jon Krohn: 00:55:04 Yep. So yeah, so now that our listeners kind of have this full perspective of all the things that you've done or are doing, so you work full-time in what I'm sure is a demanding role at Bloomberg. You have written two editions of your book, you do teaching, and recently you completed a master's in computer science from Georgia Tech with a machine learning specialization, doing all that at the same time. And I would also, I would something that I've recommended on episodes in the past, but haven't in a while. I haven't personally done the Georgia Master's course, but when people come to me and they already have an undergraduate background and they're, they're wondering like how they can be advancing themselves professionally as a data scientist, that is my number one go-to recommendation. If they're like, if you want to do a serious, like, multi-year program it's relatively affordable. It is super rigorous. And yeah, so obviously I highly recommend it. What drew you to pick that program?

Stefanie Molin: 00:56:11 Well, I needed one online for obvious reasons. Scheduling. I just liked the approach. I liked for me that it was more of a software engineering focus. So I did the CS program. And I liked that it was approaching the ML side, but with heavy, heavy CS concepts. I found that like the other programs that I looked into that were on the ML side were a lot more maybe on the math and the concepts, which wasn't really going to get me to where I wanted to go. And also, you know, this role at Bloomberg was actually my first software engineering role, really. I was kind of doing

like some hybrid thing in a previous job, but it was all like self-taught. So what I loved about the rigor and the course selection, was that I was like growing at double speed while pursuing, you know, working and then and doing the master's. I learned a tremendous amount from it. I would definitely recommend that program to anyone considering it.

Jon Krohn: 00:57:09 Nice. And going back a little further into your history, you have a undergrad in Operations Research from Columbia University, from the engineering school there, and you are not the only guest that's recently been on that has that kind of background. So Josh Wills in episode number 665 also came from that kind of background and it provided him with a tremendous foundation for, like you, doing a ton of extremely valuable work both for the public as well as for private companies for fast-growing tech companies across software engineering and data science. So what is it about, what is Operations Research to recap for our listeners and why was that so useful to you in everything you've been able to do in your career so far?

Stefanie Molin: 00:58:00 So I'm going to rewind actually. So I went into school for chemical engineering and after one semester I realized that that was not going to happen. And at the time had never taken a computer science course at all. And then I had to find a major that didn't require more chem, and that was Operations Research. And then had to take computer science and that was very scary for me. But what I really liked about the Operations Research program is that it's not like you're focusing just on one thing, really, like chem or bio or econ. You get a breadth of a lot of things. And so you're, you're, it's almost like a problem-solving degree, an optimization, right? So you get the econ background, you get the stats background, you get the coding background, you get like simulation, like all these things that, like, very helpful on understanding the business, understanding the data, and being able to work hands-on with things. So there was a point where after I had graduated and I was in the workforce and I

started getting really into coding again and thinking, oh, maybe I should have done a computer science degree. But now I've come to the point where I'm like, no, that was definitely the right degree because it really set me up for a breadth of options. And just having those different areas where you can draw upon is tremendously helpful, I think.

Jon Krohn: 00:59:23 Nice. Yeah. Super cool. It sounds like a fascinating and useful background kind of thing that, you know, I wish I'd known about back when I was doing chemistry. So, all right how do you do it? How do you, when you have all these things going on, when you are working full-time at Bloomberg, when you are writing your book and doing a master's, how, like, how do you do that? What would your productivity tips be for us?

Stefanie Molin: 00:59:55 Well, first of all, stop sleeping. No, so for me it really became down to it. You have to accept that there's no way you're going to get all the stuff done and it becomes a exercise in prioritization. So at any given point, there's going to be things that like absolutely have to be done now. Things that maybe can wait slight a little bit. Things... also you have to realize, is this really necessary? Do I need to spend another three hours to get like another one point up on what this is or are those three hours better spent on tackling this other thing that hasn't gotten any attention in a while? Okay, so, you know, in, in summary, so for me it's very important having the knowledge of what, you know, the levels of prioritization, what really needs to be done right now, what can wait and then structuring, and I'm a big fan of to-do list, so structuring that maybe sub-dividing things.

01:00:48 And I went through several periods. I'm like, what's the best way to do my to-do list? And at one point I had like a school to-do list and at work to-do list and they're like another to-do list. And then you find that maybe you're not picking the right things off the right piles at the right time. But there's going to be some level where it's just

going to be, you have to figure out what works for you. Are you a list person? Are you a reminder person? But definitely the prioritization and what needs to be done right now versus what can be done tomorrow. And not trying to do everything. It's very important.

Jon Krohn: 01:01:18 Nice. Yeah. Prioritize, prioritize, prioritize. So given how much you have accomplished in your career already, certainly, big things are ahead in the rest of your career. So what are you hoping to look back on perhaps when you retire many decades from now?

Stefanie Molin: 01:01:39 Well one thing I know that I'm going to look back very fondly is the fact that I went and got started in the public speaking angle of it.

Jon Krohn: 01:01:51 Right.

Stefanie Molin: 01:01:51 So that was something that... I'm very introverted, so doing that the first time was very scary.

Jon Krohn: 01:01:57 I would genuinely never guess that from this interview experience.

Stefanie Molin: 01:02:01 Okay, that's good. But I mean, it's, it was something that I like I had, after writing the book, thought like I had seen other people you know, in my circle that had spoken at conferences and I thought, you know, one day I'll do that. It wasn't something I was actively pursuing. Because again, that kind of a thing is scary the first time you do it. And I did get an invitation during the pandemic from ODSC to a virtually. And I discovered that I really enjoyed, like, I enjoyed making the content. That was something I knew already. But I enjoyed like actually connecting with the people and doing it even though it was virtual and it's not really the same and through three times of doing it and being incredibly nervous, those three times, I then got to the point where I felt like I could do this in person.

	01:02:50	And then when I did do it in person, it was so enriching because I really felt like I was making a difference. I was connecting with those people. And then the feedback afterwards, it was just like, this was like, I absolutely want to keep doing this and I want to do this because I feel like I'm, I'm making a difference and I'm helping people. So that's, that's definitely one thing. And I think kind of just looking back at the impact, right? So like the fact that there's a book and it helps people and the open-source contributions and, and the fact that I've, you know, made a name for myself.
Jon Krohn:	01:03:24	Nice. That's a great answer. So, beyond your book do you have a book recommendation for us?
Stefanie Molin:	01:03:36	So one of the things I promised myself when I finished my master's degree, I wrote this down on a plane to motivate myself to study for my final, was my freedom list, all the things that I was going to do when I finally graduated. One of them was reading because I was an avid reader once upon a time - fiction. And I've made it through one book since and then I started Data Morph and I haven't read anything. That book was The Hunger Games prequel, which I did enjoy. I like The Hunger Games.
Jon Krohn:	01:04:05	Cool. Before they were hungry?
Stefanie Molin:	01:04:07	Before they were hungry. No, they were still hungry. That was good. A book I have to read in my to-read pile is Brag Better, which I hear very good things about. So I am looking forward to-
Jon Krohn:	01:04:19	Nice.
Stefanie Molin:	01:04:19	To digging into that.
Jon Krohn:	01:04:21	Cool. So it's about kind of just being comfortable representing yourself publicly or something.
Stefanie Molin:	01:04:26	Yeah. And I feel like that it's-

Jon Krohn:	01:04:27	In meetings.
Stefanie Molin:	01:04:27	In meetings, yeah. And I feel like learning or just doing the public speaking in general has made me a lot better and a lot more comfortable in meetings and it's definitely like skill transfer there. It's more about like kind of your elevator pitch or like how do you talk about yourself and it's, you know, it's more, I think it's focused towards women because there's studies that show that you, you feel a little uncomfortable and I know I do. And so it's just, you know reframing your, your thoughts. At work, actually, we had, I got a free copy signed by the author. We had an event at work and she spoke about the book and she was, she was great. So I'm, I'm looking forward to digging into that.
Jon Krohn:	01:05:08	Nice. That sounds like a great recommendation. And then for people who want to get more recommendations for you and hear what you're up to, your latest after this episode, how should people follow your work?
Stefanie Molin:	01:05:20	So GitHub, obviously for code work and the workshops and the book repos, all on GitHub. I'm also on LinkedIn and Twitter.
Jon Krohn:	01:05:31	Nice. Well Stefanie, thank you so much for an awesome, highly informative episode. Thank you for making the trek down to film with me in person. It's been a ton of fun and we're going to have to catch up with you again sometime in the future, I hope.
Stefanie Molin:	01:05:44	Sounds good. Thanks again for having me, Jon.
Jon Krohn:	01:05:52	Thoroughly enjoyed hanging out with Stefanie today and learning so much from her. In today's episode, she filled this in on how chaining and assing-ing are her favorite pandas data-wrangling tricks. How pandas is great for creating plots quickly, but matplotlib allows for more flexibility, particularly through the ticker module that she highlighted. And seaborn is best for handling colors and

creating aesthetically pleasing visuals with minimal effort. She also talked about how her Data Morph Python library enables you to gradually transform a 2D scatter plot from one elaborate shape to another elaborate shape without impacting the axes' means, standard deviations, or correlation. She talked about how package creation is useful for data scientists to learn how to create excellent shareable code. How at Bloomberg, she uses a Python backend, a React front-end, and D3 for visualizations, and how ruthless prioritization is the key to her remarkable productivity.

01:06:45 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 Stefanie's social media profiles, as well as my own social media profiles at superdatascience.com/675. That's superdatascience.com/675. I encourage you to let me know your thoughts on this episode directly by tagging me in public posts or comments on LinkedIn, Twitter, or YouTube. Your feedback is invaluable for helping us shape future episodes of the show. And if you'd like to engage with Stefanie and me in person as opposed to just through social media, we'd love to meet you in real life at the Open Data Science Conference East, ODSC East, which is coming up next week in Boston from May 9th to May 11th. I'll be doing two half-day tutorials. One, we'll introduce deep learning with hands-on downloads in PyTorch and TensorFlow, and the other, which is brand new, we'll be fine-tuning, deploying, and commercializing with large language models including GPT-4. In addition to these two formal events, I'll also just be hanging around and grabbing beers and chatting with folks. It'd be so fun to see you there.

01:07:48 All right. Thanks to my colleagues at Nebula for supporting me while I create content like this SuperDataScience podcast episode for you. And thanks of course to Ivana, Mario, Natalie, Serg, Sylvia, Zara and Kirill on the SuperDataScience team for producing

another excellent episode for us today. For enabling that super team to create this free podcast for you, we are deeply grateful to our sponsors whom I've hand selected as partners because I expect the products to be genuinely of interest to you. Please consider supporting this free show by checking out our sponsors' links, which you can find 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. Finally, thanks of course to you for listening. It's because you listen that I'm here. Until next time my friend, keep on rocking it out there and I'm looking forward to enjoying another round of the SuperDataScience podcast with you very soon.