SDS PODCAST

EPISODE 780:

HOW TO BECOME A DATA SCIENTIST, WITH DR. ADAM ROSS NELSON

Jon:	00:06	This is episode number 780 with Adam Ross Nelson, author of How to Become a Data Scientist.
	00:19	Welcome back to the Super Data Science Podcast. Today's Five-Minute Friday episode was filmed live with a guest, the author, Adam Ross Nelson. We filmed it at the Open Data Science Conference East, ODSC East, which was held in Boston in late April. Adam is the author of the books How to Become a Data Scientist and Confident Data Science. He teaches statistics at the University of Wisconsin-Madison, where he obtained both his PhD and JD, his doctor of law. Today's episode will be of interest to anyone who'd like to land a new data science role. In it, Adam provides his top piece of advice for data science job seekers. Ready? Let's jump right into our conversation.
	00:58	I'm here with Adam Ross Nelson at ODSC East. Adam is the author of How to Become a Data Scientist, and he was here giving a talk on how to become a data scientist. Thought we'd do a quick Five-Minute Friday on his favorite chapter, the most valuable chapter from your book, Adam, for our listeners on how they can become a data scientist.
Adam:	01:19	Well, thank you so much for having me. And yeah, one of my favorite chapters in the book How to Become a Data Scientist: A Guide for Established Professionals, it's going to be chapter seven, and it's the chapter on portfolio projects. If you've studied this topic at any length, you know that transitioning into data science, one of the key strategies is having a project that you can show a recruiter, show a hiring manager, or multiple projects potentially, show coworkers that you can do data science. And in that chapter, I talk about a data science portfolio, what it is, and there's 10 portfolio ideas that you can build, specific ideas about different projects that you can build. For example, one of my favorite projects in the chapter is going to be building your own data set. And I'm

not necessarily talking about building a fictional data set. That's a great project idea too.

02:18 This one I'm talking about collect data, collect original data, document the process used to collect that data, why you collected the data, how you collected the data, engineer that data, engineer features, make the data available for others to use. Fantastic way to show your skills. Another one is going to be put natural language processing into production. And one of the nice things about this project idea is you don't have to train a model. There are so many available pre-trained models for natural language processing, specifically sentiment analysis. You can take the APIs from AWS, Google, or Azure, and I also recommend taking a look at spaCy and Natural Language Toolkit. This is a project that can really showcase your ability to put a project into production. You don't have to worry about the training. You can take someone else's model and just put that into production.

03:15 Oh, bonus: all of the major platforms have quick start guides for those APIs. They service tutorials for how to do this. Follow those tutorials, make it relevant to your own specific use case, one that you care about and that you're passionate about and make it interesting. And then there's another project idea that I've been kicking around for quite a while now but I came up with pretty much after this book published. Find a data set, go through the data augmentation process with a data set that somebody else has out there. Again, write about your methods. How did you augment, why did you augment, make the augmented data available. And this is a fantastic way to showcase your skills in data science. And then one of the things I like about this, again, is for better or worse, it doesn't involve necessarily building a model, but it does demonstrate super important skills in the data science field.

Jon:	04:14	I love these ideas, Adam, and I'm glad that this is your favorite chapter on how to become a data scientist because for me as well, when people come to me... This happened to me yesterday at the conference. Yesterday I was giving a full day training on open source LLMs, and a gentleman came up to me afterward and said, "There's so much. Where do I start? How do I manage to understand all of this? What's the path to do that?" And I said, "Don't worry about it. Just do a project. Just pick an LLM project." And like you're describing, having your own data, something that's of interest to you. So either collecting it, like you said in the first instance, or augmenting in the second instance. Maybe that's something that you could talk about a little bit more, is just what does it mean to augment a data set?
Adam:	05:02	Oh, yeah, that's a good point. To augment a data set, one of the best use cases for this would be if you have photographs, for example, and you're trying to build some computer vision. You might have 1,000 images, but you really want 10,000 images. So the classic most traditional approaches to augmenting that data set, and essentially what we're trying to do, we're trying to create additional observations for training. You're going to look at the original images and you're going to apply a rotation or a flip or a flop, a random zoom or a random crop, and what this does is it creates natural variation.
Jon:	05:41	[inaudible 00:05:42].
Adam:	05:42	What's that?
Jon:	05:42	A flip, a flop, a zoom, a crop.
Adam:	05:44	Exactly. It's fun. You can start to have fun with it.
Jon:	05:47	Yeah.

Adam:	05:48	The other one is for text data. One of my favorites is augment text data, so just unstructured text data by translating it through three or four languages. For example, one of the data sets that I work with quite a lot is bank complaints. It's a fictional data set, customers from a bank and they have some sort of complaint about a product. To augment that, I would translate the English into German, from German maybe to French, from French to Spanish, and then Spanish back to English, and the new English version is going to be different. It's going to be a variation on the original. And in that process, I've doubled the size of my data set, I've doubled the number of observations. So a project like this is really accessible for a lot of folks at all levels of the field.
Jon:	06:40	Fantastic, Adam. Well, thank you for taking us through your favorite chapter of your great book, How to Become a Data Scientist. And yeah, as I already said, I couldn't agree more with that being the most important thing. Having your own project demonstrates to interviewers that you have the drive to actually do something yourself on something you're interested in as opposed to just saying, "I would like a high-paying job," coming into the interview. And then it gives the interviewer lots of things to talk about. You can publish this to your own GitHub repo and people can explore it. Bonus points if you create some simple interactive UI, probably, around it. You could use Shiny or you could use... There's lots of other-
Adam:	07:20	I think a simple Flask front-end can be a really good place to start for that. Pythonanywhere.com by the folks over at Anaconda is a great platform for building a simple front-end.
Jon:	07:31	Nice. Yeah, so you can crack out your laptop for the interviewer to play with your data science project. And I am confident that more times than not, your interviewer will be super impressed and you'll get a job offer.

Adam:	07:45	Yes.
Jon:	07:46	Nice. All right, thanks, Adam, for taking the time.
Adam:	07:47	Thank you for having me.
Jon:	07:48	I [inaudible 00:07:49] this busy ODSC conference and yeah, we'll catch you again on air sometime soon.
Adam:	07:53	See you again soon.
Jon:	07:54	All right, that's it for today's practical episode on leveraging real world, self-driven data science projects to land a new data science role. If you enjoyed it, consider supporting the show by sharing, reviewing, or subscribing, but most importantly, just keep on listening. And until next time, keep on rocking it out there. I'm looking forward to enjoying another round of the Super Data Science Podcast with you very soon.