SDS PODCAST
EPISODE 3
WITH
DR. WILSON POK

Kirill: This is episode number three with Nanophysics PHD turned data scientists Wilson Pok.

Welcome to the Super Data Science Podcast. My name is Kirill Eremenko, Data Science Coach and lifestyle entrepreneur. Each week we bring you inspiring people and ideas to help you build your successful career in Data Science. Thanks for being here today and now let's make the complex simple.

Welcome everybody to the Super Data Science podcast. Super pumped to have you here on board today. This podcast is all about interviewing the top data scientists in the world to help you build your successful career in data science.

Today our guest is Wilson Pok. Wilson is a friend of mine who I met back at Deloitte when I just started. Only worked for a few weeks together but Wilson's such an approachable guy. He's one of those people that are so open, so outgoing and at the same time very modest that we got along very well and stayed in touch.

Now he's in Sydney. He's working at a consulting firm called Ambiata. Before that, he worked at a large bank. He's got some varied experience both in consulting and in the industry. But what's fascinating is Wilson actually has a PhD in Nanophysics. He went through all that research, through all that education academia and in the end, he still decided to move to Data Science.

As you'll see in the podcast we actually discussed that it's not an uncommon thing for people to move from Science and specifically Physics. But as you saw, from one of our

previous podcast, podcast number one where Reuben Kogel moved from being a chemical engineer into data science.

It's not uncommon for people to move from science into data science, so from physics and chemistry and other sciences into data science.

We'll learn why in this episode or at least you'll get our opinions on that matter. That's one of the interesting things that we discuss in this episode.

Also, you will learn how Wilson, now given his background approaches business problems. It's quite different to your standard approach because problem is that they teach you back at University. It's more of a statistical base approach.

Specifically, when we talk about Bayesian inference and randomize control trials. Quite interesting approach where you don't think of episode problems as a single outcome like having a single outcome but you rather think of it as having a range of outcomes. That's what Bayesian inference is all about so we'll have each other about that.

Then you'll learn a bit more about Wilson does on a daily basis and that's isolating effects of marketing companies. If you're doing data science for marketing or within a company to assist with marketing then you'll definitely find a lot of this information useful to help you better deliver valuable insights.

Also, were going to mention how to drive change into business. Because it's always challenging once you find something using data on how to communicate but more of how to actually drive business change to help people re-educate themselves and focus on other things when you find

that certain parts of their jobs are not actually bringing any value to the business.

It's always a challenging question, psychological question, and it's got a lot of emotions involved in it. Wilson will give you his take on that now that he's working consulting. He does that on a daily basis. Definitely, we can learn a lot from there.

For the techies is out there for those of you who want to get into the nitty gritty of the Data Science. Wilson's going to go into a lot of detail on R modeling libraries that he uses in his day-to-day job. He's got it all down pass. You'll see he'll name a couple. Some of that even I haven't heard off before. It'll be very interesting to see what libraries he uses and how he goes about his modeling.

Also, if you're a Python fan you will see that Wilson works with Python as well and he'll name a very surprising Python IDE. Before getting it away, it's something that we discuss in our courses. It's not your first guest that you would go to for a large consulting firm.

So, definitely lots of valuable insights in this podcast. Can't wait for you to hear our conversation and let's jump straight into it. I introduce to you Dr. Wilson Pok.

Hello, everybody. Welcome to this podcast. Today with us today we've got Wilson Pok. Wilson, welcome to the podcast.

Wilson: Thanks, Kirill.

Kirill: Wilson is a great friend of mine. We used to work together. Actually, we only worked together for just a few days or a few weeks back when I was a graduate at Deloitte. Then

Wilson moved down to Sydney so I'm super excited to catch up.

Wilson, tell us a little bit about yourself. What do you do currently?

Wilson: Yes. I'm at Data sciences at a company called Ambiata. It's basically a data science company. What we do is we take lot of the data that the companies have and we inject it into our platform and mainly do marketing interventions. We do a lot of experiments on designs to try to work find out what works and what doesn't. But to sum it up I basically say that we apply the scientific methods to business problems.

Kirill: Okay, very interesting. It's like a consulting type of firm.

Wilson: Yes. It's consulting.

Kirill: You've had quite a bit of career. The last time I remember we worked at Deloitte together then you moved on to Westpac, which is one of the big four banks in Australia. Are you still with them?

Wilson: No. Actually I left Deloitte. I went to Westpac for a bit did some predictive models there. Some, what I thought was basic stuff but I think Westpac at that stage was still struggling a little bit with their data issues. Since then I've moved on done to Ambiata where their focus is purely on the data and so I've been able to use the tools that I want to use and we cover more interesting problems so to say.

Kirill: Awesome. We'll definitely get back to your challenges and wins at Ambiata and Westpac in a second. But what I would like to ask because you're the first person on this podcast that actually holds a PhD and has done extensive research. Can you tell us a bit more about your background? What did

you study and how did you get to this level of academic achievement?

Wilson: My PhD was in slightly different field, it was in Physics. I studied Nano scales silicon devices but to be honest I think since I've left academia, the story is I didn't see an academic career for myself. I wanted to try other things so I went through industrial roles.

But to be honest since I've been in business I've actually run into a bunch of other Physics PhD as well. It seems like it's a common path that a lot of PhDs in quantitative fields say they still seemed to pop up in the data science world. I've been running into them. It's pretty good.

Kirill: Yes. It's definitely. I agree with that. Myself, even though I don't have a PhD I did do a bachelors of Physics. Mine was also nanotechnology and laser Physics. Yes, it is kind of like I see people, like minded people popping up in the field of data science.

Why would you say that it's like— what was your reason behind leaving academia? Why did you decide to move into something different?

Wilson: With academia, it's very much about depth. You go into very, very deep field, technology into very narrow fields. I wanted to try other things. I wanted to try my hands at other programs and so I felt that consulting type projects would be more interesting.

But the other point is that actually I think that there is a space for people with quantitative background to come into business because to be honest the traditional people with business degrees they're not so strong in quantitative fields.

There's kind of room for people with them, science degrees or engineering degrees to go into business. It's just requires you to think about things in a different way. But, yes, I think it's a huge opportunity.

Kirill: Yes, definitely I agree with that statement that in academia or physics even, you have to become very narrow in your focus. You kind of at some point realize that if I'm going to do this, I'm going to be doing the same thing for the rest of my life and it has to be the perfect, you know what I mean, it has to be the perfect match for you to agree to such sacrifice.

Wilson: Yes.

Kirill: And, often of the case it's not, oftentimes it's not. It's always a great thing that you can always move in to business and consulting with all these skills that are very transferrable.

Wilson: Yes, that's right.

Kirill: Just on that, what do you feel is a skill or I don't know, maybe, some knowledge that you developed during your PhD because that is a long time. You're doing it for several years like five or six years.

Wilson: Yes, it's five years. I think I've been a bit of a new perspective now. When I was trying to make the move from industry, I'm sorry, from academia to industry, I didn't have to spend a long time trying to argue the case what will be the transferable skills.

But now that I being in the industry for several years now, I would look back and I would say that the thing that stands out the most is, as being the most transferrable or most useful was basically being clear with definitions. Being able

to define a problem and define the premise of the problem and define the primary problem clearly. I think that that in traditional and in my conventional businesses environments it doesn't happen enough. There's a lot of buzz words, a lot of very concepts but when you come from a quantitative of background you do have to nail definitions down really, really clearly. I think that has helped me in a lot of projects that I've been on.

Kirill: That is very interesting because not so often is that problem is brought up that usually people talk about doing the analysis as a very big deal and then presenting the insights the final part as also big deal but what you're outlining here is the preliminary effects is when you're finding out how to describe the problem.

Wilson: Yes, that's right. All the analysis, programming, working with data that's all important. I assume that anyone with any science degree already has that. It's what I didn't realize was that that's being to define things so clearly. I didn't know that that was so useful in business and it really is.

The other thing I would talk about is something that happens, or at least in science and not so much in business, is that in science we talk about uncertainty a lot. We have error bars around predictions. We don't describe things by single numbers. Often, we describe things by distribution and that kind of thinking comes naturally to someone in a Science field. A lot of things are uncertain but you can quantify your level of uncertainty.

In business that doesn't happen at all. In business it's traditionally that you report on a single number for the profit of a single quarter, it's a single month, of how the conversion

rate of a marketing campaign, it's usually single number. It's not really reported as a distribution or the range of uncertainty. That doesn't really happen so much.

I find that that's one of the things that is important to challenge. This is one of the things where a science-based way of thinking comes in. It's completely different to beyond traditional business way of thinking which is all about hard numbers, it's very certain. Whereas if you come from a science background you know there is a level of noise and uncertainty that is always present whether you acknowledge it or not. I think that that's the other aspect that I didn't think would be important but it turns out that is an important from my background.

Kirill: That's fantastic. That's a very solid observation that indeed in business it's different. Just then thinking about our listeners because most of them are probably not come from a science background as you, how would you recommend developing that kind of mentality, that kind of thinking about business problems in terms of not just hard numbers but what you described, the distribution approach and the error range?

Wilson: One of the ways to get into this and it's also been a learning experience for me is looking into your listeners, you've probably heard of Nate Silver, he has a book called the Signal and the Noise. It's a good introduction to this, to the concept of let's called Bayesian influence.

When you look into Bayesian probabilities and Bayesian ways of doing statistics it starts to introduce you to this idea that you never have the final answer. You have approximations to the final answer or to the truth but you

can never be, like completely certain, you have to condition your statements on you know you have the 95% credibility, things like that. I would actually suggest looking into looking up on Bayesian analysis as a way to understand the uncertainty.

Kirill: Okay, that's a very good recommendation. I think one of our guests on the podcast has already recommended that book as well, The Signal and The Noise by Nate Silver. And Nate Silver has a website as well, very, very interesting one.

Wilson: Yes. The blog is really good. It's actually of the websites that have really cool visualizations that are good at explaining concepts. I think it's called setosa.io.

Kirill: Ha! Setosa like virginica setosa and stuff like that. Like what do they called the Iris data set for the Fisher's Iris data set that's where they come from.

I haven't seen setosa.io but I definitely will put the links that we've discussed in the show in the show notes so that everybody can get to them later. That's some great advice. Now moving on to your current work. Can you tell us a bit more? You've told us about Ambiata and the type of consulting that you do there. Can you tell us a bit more about some more specific recent projects that you've been working on if you can disclose that information of course?

Wilson: The main types of projects that we try to run now, essentially we run randomized control trials. What we try to do is identify causal effects, which is not something that's typically common in business.

Typically, they run a marketing campaign and did the conversion rate, go up or down. It depends on a lot of things. It might have been the intervention, it might not have been.

What we try to do is we try to be very clear in isolating the effects of certain things. What we found is that the gold standard of establishing causal effects is to run randomized control trials.

The kinds of successes that we've had would be things like along the lines of like running experiments, demonstrating that a certain marketing campaign has no effect or only has an effect that's within the level of noise. As a result we can say that you safely turn off that marketing activity and not have any difference with sales.

The other things that what we would typically do is we do uplift modeling. We send out an intervention, which could be like a direct mail or email or whatever marketing intervention. We would identify the people who are most affected by that intervention.

Not the people who are going to buy the product anyway but the people who were directly affected by the intervention. That kind of uplift modeling is also a type of project that we focus more on.

Kirill: That's very interesting. Basically, the first type that you just describe the last companies significantly cut cost. If there's activity that they're performing, they don't need to perform, they just cut it off. Basically your analysis you'll pay for itself. On the other hand if you don't need to cut the cost sale, they'll know the quantitative effect of that marketing activity.

Wilson: That's right. A lot of— we found that of a lot of marketing activity has been that like there are a lot of opinions, a lot of knowledge that accumulated but it's not clear what that knowledge is based on. It does seem like it's based on data.

We find that if we ask there is the question on what would happen if we switched everything off. It's kind of a destructive question because a lot of companies that have these huge marketing departments, they have to keep active, they have to keep in their running campaigns.

But what we try to do when we first interpret it, what the company is to step back and to actually measure how effective these campaigns really are.

We find that we are almost no regrets in how we measure effectiveness than normally are on their own.

Kirill: Yes, I love that about data science. I've personally worked at a company and I've worked with companies and I've seen companies where there are such massive giant machines. There are so many people that are working there. They have all these archaic processes that have established. They're like entrenched.

People just come to work and sit there for like six hours and then does it work for an hour. It's just like it makes me sad that people could be spending their lives doing better things and instead they are doing these archaic and just old school, old fashion marketing processes or whatever other type of processes. You guys are acting as a disruptor to this old fashion way of thinking.

Wilson: Yes. It's hard to change all these large companies. You can't just go and change one marketing campaign. You have to do

it all at once and you say, “Everyone you have to all play by the same rules. You have to all respect this universe to control. You will have to follow the same measurement standards.”

You can't just go to one particular company and will just try this one because they’ll say, “Why are you measuring me with this truthful matrix whereas everyone else can…”

Kirill: Yes, get's some slack.

Wilson: Yes. I'm kind of have to go to the head of marketing and look at the way you transform the way you do your marketing and just one campaign at a time.

Kirill: Yes. A lot of people think of it because I've also kind of, not that I've initiated anything of that scale as you described but you even told me to people about things like they often take it in a very adverse effect because they think they're going to lose their jobs out of it which is not always the case.

In many situations if somebody's been off to company for a while, the company is more likely to keep and all they have to do is to kind of up up skill and take some courses, learn some new skills, and evolve the way they approach their work.

There if you're open minded you should be excited about learning new things. That's the way I think about it.

Wilson: Yes, absolutely. Also, we found that a lot of companies that we've worked with. You’ll talked to the analysts at the ground level. They'll agree like, they’ll know that, it’s a lot of…

Kirill: It's a useless process.

Wilson: Well, it's certainly more complicated than it needs to be and asking the question, well what would happen if the process is, was much simpler or even random. What would actually happen?I think it's more why they accepted that on official level than unofficial level.

Kirill: Definitely. And, most of the time in companies like that the top level managers they just don't have the time to do everything. They can't keep track or everything.

What would your— I just would have to ask you because I find myself in a situation like that where I'm a data scientist in a company and I see a process that is useless. I see people doing it day in and day out this certain part of their role, which can be completely taken out of the equation and nothing will change.

My gut feeling is that I know that's the case. What steps would you recommend to somebody like in my position, somebody who's not consulting from external but from within the company, can see that something is not necessary?

How would you go about first of all, getting some factual evidence that this is not necessary and then second, how would you go about communicating it to the stakeholders to change things?

What we found is that it's certainly hard to go and then say, "Look we don't think **[0:20:14]** generating a lot of activity and we don't think it's working efficiently."

The more effective thing is to ask the question and see what the evidence says in basically about an experiment. It's

certainly hard if environment is it's volume-driven that needs to be a certain amount of activity per month.

It's about asking the question. There's a lot of I think **[0:20:38]** out there in terms of basically how effective running **[0:20:42]** in terms of establishing what's effective and what's not.

I think I can probably send you a link on the trials at Facebook **[0:20:52]** Facebook once. You would think that Facebook has tons and tons of data that they would be able to predict almost anything. It turns out that even for them like they need to run the experiment **[0:21:04]** establish the truth.

I guess my advice would be to where you can find evidence and collect it and build your case from there.

Because if you let the evidence speak for itself, you don't try to say, "Look I'm going to argue one way or the other." You say, well let's just ask the question "What data would we need to answer that question," and collect the data.

People are willing to listen. They'll listen to you. I'm sure there are always various **[0:21:29]** incentives and things going on. But I think if you have open-minded people listening to you and the evidence is compelling, then they'll see it.

Kirill: Definitely and we'd love to see that link on Facebook trials. That sounds very interesting and definitely **[0:21:46]** the show notes as well.

And speaking of literature, were you there when Giam Swiegers, the ex-CEO of Deloitte sent everybody a book it was called "Who Moved My Cheese?" Or **[0:21:58]?**

Wilson: **[0:21:59]** what was the story about?

Kirill: The story was like the company was undergoing some massive transformations in order to meet the KPI set for Australia, for the Deloitte division of Australia.

And because of all these changes like a lot of departments are merging, people would have to move different roles and so on, and just to preempt all of the panic that was going to happen to Giam Swiegers bought every single one of our employees.

Out of 6,000 people, he sends 6,000 books like **[0:22:30]** you found them in your letter books and they were all the same book, "Who Moved My Cheese?"

And it's about change. It's about how to adapt to change and not resent it but actually take advantage of when things are changing in life. It's a great book, very short one but it's a nice gesture like that. Maybe that's something could be other companies that size could do as well.

You mentioned the Randomized controlled trial several times. Could you go into a bit more detail how to set up a randomized controlled trial and what it involves please?

Wilson: Sure. It's basically exactly the same as what happens in medical trials when they test out a new drug which finds to establish cause and effects.

The way to establish with this drug cause this person to be cured or whether this marketing intervention cause this person to buy something.

You have to essentially create – the way to answer the question is to have two universes way. In one universe, the

person gets the intervention and the other person doesn't get the intervention.

Now we can't do that so the next best thing is we simulate that by having two very, very similar populations. In one population, we apply the intervention. We send them an email or we send them some marketing material. And the other population, which is our control, we don't.

And so right there, that's just step one and we're not targeting this role model. There's nothing, it's just random selection but you make sure that those two populations are as similar to each other as you can.

Kirill: Sorry to drop so it's kind of like an A/B test.

Wilson: It's basically an A/B test. It's a simple concept which is surprisingly hard to implement in these large organizations which have existing marketing campaigns going on, different people getting tied to different things but it's important.

It's important because it allows you to definitively – well as definitively as you can establish that something cause something else. Typically, you don't have that level of control.

Like in economics people, you use natural experiment state they look at. For example in the U.S., **[0:24:38]** certain states and not other states. What effect did that have in the populations and the companies that we worked with, we have these large populations and we can run experiments to see what works and what doesn't work.

But you mentioned A/B testing and so that happens much more in the online space. It's good to that in the online

because you have much higher volumes, you have much quicker feedback times.

Typically in the companies that we worked with, you have a longer feedback cycles or it takes longer for results to come back in and you don't have this high volume. There are some other challenges associated with doing it in the real world.

Kirill: That's quite an interesting **[0:25:19].** Basically an A/B test would involve splitting your whole population and you're showing some of them one thing and showing some of them **[0:25:26]** like the website version.

Whereas for randomized controlled trial, you just keep doing what you're doing normally. But then you synthetically create two samples from your population and then you tests something on each one.

Wilson: Yeah. I think the terminologies maybe from different areas but it's basically the same thing.

Kirill: That's interesting. Basically next time somebody's coming up to their boss and saying they ran an A/B test, instead of saying that, they can sound a bit more sophisticated and say they ran a randomized controlled trial.

Wilson: It's important. The A/B testing would be like you have two candidate website designs and you're trying both of them. Traditionally in a randomized controlled trial, you're testing doing something versus not doing something.

Kirill: Definitely. So kind of like more of a champion challenger kind of situation. You have an existing way of doing things and you want to see if changing that will **[0:26:22]** a better effect.

That's very cool. And looking at your LinkedIn, which is mind-blowing by the way, your work at Westpac with **[0:26:32]** modelling and custom analytics and visualizations.

Can you tell us a bit more about that especially their type of logistic regressions, classification trees and even neural networks that you've worked with?

Wilson: Back in day is really interesting because you have some pretty detail information about people's processing behavior and incomes. But it was mainly what we found was that in that space some very simple models where are already quite good.

A lot of times it was just looking at what just trying to which were the good targets for cross-selling or upselling and retentions. Just getting some basic models on large populations which data sets. Mainly those are the main goal there.

But also I found that it was also not just that. A lot of those educational as well in terms of trying to explain to people **[0:27:26]** done traditional marketing.

I have a pet background in many segmentation, explaining them the concept of how model works and how model can be used to target in their personalized leads for their marketing campaigns.

I found that traditional marketing people have these lot of knowledge but it's kind of **[0:27:44]** but it's not really – it's hard to quantify that knowledge.

Well you don't want to do is say that "We're going to **[0:27:50].** We're going to replace the entire marketing team." That's not true.

There is prior knowledge that traditional marketing has. It's just a matter of how we bend that risk, the database techniques that we would use.

Yes, a lot of it was educational in terms of explaining how models work. For people with data science background, it's kind of obvious how you would use modeling.

But if you start telling your people you don't have that background and it's more of traditional like they won't quite grasp but they'll see as model as just a way to get insights so that they can learn and do their targeting. But there's a lot of education around that and communication.

Kirill: **[0:28:34]** found that once you deliver some sort of model, especially when it affects people's workflow and whether it's a negative or a positive way, if you go back to those people and then present it to them like you are not in the complex terms that it is but just more like layman terms, they really appreciate it.

Even though they don't understand the mathematics behind it, they understand why the company is now changing its approach to doing things.

That way, not only you're fulfilling these people's curiosity and helping them better transition to the new type of work, but also you're creating data advocates among them.

Even though they don't understand these things, they see that they work and they want more of them. So they then starts coming to you with more questions and more ideas about data and how that can help other processes can be improved. That's a very powerful thing.

Wilson: That's right. To be honest, it's not really their job to understand it either. But as long as they're **[0:29:29],** as long as they're various checks are in place and they're reassured that the model works and that is based on sound principles, then they're happy to use it.

Ultimately what matters is what gets measured in the end. What's important is whatever goal they define, whatever KPI is, they will measure either the traditional approach or the more basic approach. They'll measure both on the same metric and whichever one works and that's the one that they go with.

Kirill: That's definitely a good way to look at it. Speaking of modelling, this one thing I was curious about in a one company where I worked, we had models.

Those models were created long before I came there and actually by an external consulting party. They kind of like implemented them and the job was over, done and they forgot about them.

They are running there in sequel every month and I think **[0:30:25]** unsupervised for 18 months, just one of those retention type of model so what are the likelihood of somebody, the churn likelihood is.

And so when I looked at them, actually no it was just me, it **[0:30:41]** pivotal another consulting party and we looked at them and they were completely gone. They deteriorated.

I wanted to ask you what is your experience with model deterioration and what's your advice on preventing model deterioration and looking after existing models?

Wilson: It depends on the industry. Certainly some industries deteriorate. They change more quickly than others. Some industries are more robust. It depends on the data.

But certainly obviously there are things you can put in place in terms of continuous scoring and continually monitoring model performance.

I guess what's the most important thing is in the problem definition stage, you define what metric you want to measure yourself against. Once that is established and you continue to score the model, obviously you would have that hold out control set as well.

It's only at occasion like the companies who purchase these models, they need to be aware of that. The models do degrade overtime. I don't know some of the consulting company came in and sold them on this idea and said, "Okay the model's done. Pay us and get out of here." But it's in everyone's interests to know that models degrade overtime. You have to retrain them, data changes.

I think **[0:31:51]** completely change the data warehouse and so the models are completely obsolete. Absolutely it's important that people they're checks in place to make sure that models are fulfilling their purpose.

The other comment I would make is that it's hard for any single person to **[0:32:10].** These data pipelines, they're so long and they're so complex that no one person has a full view of everything.

No one has full control over the entire pipeline. They may have control over the part of it and so it's very easy to fall into the trap, to assume that someone else's taking over the

problem, to assume that someone else's checking that the model is valid when no one is doing that.

This is, I guess, just one of the challenges in working with large organizations that I'm sure you've experienced as well.

Kirill: For sure. Especially in a new field such as data science where something new is introduced into the company and then nobody knows whose responsibility that is. It's very important to outline those responsibilities at the very start. Especially if we've got listeners who are on the managerial or a division supervisor, it's very important to outline the new assets that are created in this data pipelines, how they're maintained going forward so that things like model degradation don't occur.

Then moving on to the tools. What are the tools of your trade, Wilson? What do you use on a daily basis at your control at Ambiata?

Wilson: I use R and Python. I do some **[0:33:31]** that's mainly because the data sets that we work don't basically takes files. We don't actually access databases directly ourselves.

Within that, in terms of – I guess I'll probably use **[0:33:43]** every day. That's the plugin package per hour. In terms of modeling itself, I'd use carets. That's the modelling package for our –

Kirill: How do you spell that?

Wilson: Caret – c. a. r. e t. I think there's another equivalent in Python as well. Here in Ambiata there's a bunch of us that use R and a bunch that use Python.

Every time I mentioned there's **[0:34:09]** one of my colleagues mentioned there's an equivalent packaging in place. In terms of modeling up in terms of building the models. We've been using a lot of Vowpal Wabbit. It's one of the new models.

Kirill: Vowpal Wabbit. How do you spell that?

Wilson: It's v-o-w-p-a-l w-a-b-b-i-t. It's basically a very fast algorithm for building new models. When you have like large data sets that we work, it's handy to have something that pretty much just work very quickly. You're not sort of limited by the size of your data.

The other thing that we would use typically some XG boost.

Kirill: What's that for?

Wilson: That would be for building – Vowpal Wabbit is more for medium models. Whereas for XGBoost, we probably use that more for our three-based models that time that we train.

But in terms like modelling that's probably get much more sophisticated than that. The hard part is always getting the data and getting the data in the right shape.

Kirill: Definitely. It sounds like you totally know your tools – caret, Vowpal Wabbit and XGBoost. It sounds like you got it all covered there.

Wilson: We don't use too many. We're trying to make the problem fit with tools that we have. When you have tools that are good enough that general purpose enough, you can apply them to a large number of problems without changing too much.

Kirill: Definitely and the question that I always ask when somebody mentions both R and Python, "Which one of the two do you prefer?"

Wilson: I have been using R, mainly because I like to use ggplot2. I know there's apparently there's a Python version of ggplot but I never use it.

But in terms of data manipulation, I would probably use Python more. I would use the Pandas package in Python much more. I think eventually I think as the package is for Python start to mature, I'm like start doing more and more in Python. And just in terms of data visualization, I end up using ggplot2 because I'm much more comfortable with that.

Kirill: I agree. You can't be ggplot2 and I had a look at the ggplot version for Python, it's very juvenile state. It's not very tested. That's not anywhere near to ggplot2 and R.

Just my curiosity, why in your opinion is the Pandas package in Python better than the integrated data frame functionality in R?

Wilson: I think there was some speed. There was some benchmarking. That was done I think Pandas was it came out slightly faster.

Kirill: Yes. I kind of heard some similar research as well.

Wilson: Only because I think that the Python codec itself is neater but visually it's neater as well.

Kirill: Definitely.

Wilson: Other than that, I'm not particularly strong, I'm not basically an advocate with one over the other. It's just a matter of preference, I think.

Kirill: Fair enough. For R, I'm assuming you use the RStudio for Python. What ID do you use there?

Wilson: The tough one that we use, we end up using IPython. We have this IPython Notebooks that we use. I don't have so much experience with them but the little experience I have with IPython Notebooks, it looks very good.

Kirill: That's awesome because now there are actually called Jupyter Notebooks, right? Because they have implemented. You can put in Julia, Python and R.

The best part is like I get a lot of students come to me, I won't say a lot but occasionally I get the question because I have a course on Python and I use a Jupyter Notebooks as the development environment to both. And Jupyter Notebooks, they're kind of like in browser modes. I don't know. Maybe you might have a different –

Wilson: Same here.

Kirill: Like they're in browser mode and students asked me is this just for training and then like in real world do you use Python in a different kind of thing like RStudio has its own but no, like it's awesome this example day to day basis you use Jupyter or IPython Notebooks which are in browser mode and you are able to analyze data, perform data manipulation and drive insights

Wilson: Yes, because we're doing the same tasks, we're trying to– we will have the data set, we don't know anything about it. We have to understand it and response right away. These Notebooks are good for that

Kirill: I find it— I haven't used them workplace environment heard that they're good for collaborations so somebody can just

save it somebody else can open and it's easy to move around like that

Wilson: Exactly.

Kirill: That's very cool. Can you tell us as a data scientist you always run in to lots of challenges on a daily basis and different, some are similar, some are different, what has been for you the biggest challenge ever as a data scientist?

Wilson: I think the biggest challenge, it's never the data. The data is messing and the cleaning and all that. The bigger challenge is always the interactions with these larger organizations. It's navigating the— like for example, trying to find out where the data is and some people have different conclusion of the data. It's navigating the maze on getting your hands on the data in the first place.

Kirill: Definitely.

Wilson: I think that's a common thing for any consulting company but definitely getting your hands on that data but also, like I mentioned, these pipelines are so long, like getting the data, being sure that it's as clean or as reliable as you can get it to be and then taking you to through the whole process and doing the data and equation on your side on building the model and then the scoring. That whole program is so well.

There's two points where the data comes and the data goes out and making sure that you have it as robust as you can to you know, like they'll be changes that they'll come up. There'll be things like they'll change the definitions of one of the columns of something and they won't tell you. Like because you don't have control over the whole pipeline, I guess making things full tolerant…

Kirill: Yes.

Wilson: It's very odd. You can't expect everyone to be perfect but mistakes will happen. Making your project full tolerant to that, that is very hard.

Kirill: Yes, it's a lot of effort. Like were looking back at good old days sometimes this stuff just goes down right.

Wilson: Yes.

Kirill: Somebody says the data set and all the column are in French.

Wilson: Yes. You get like a sushi and…

Kirill: Yes. Like you say somebody like deletes a column or adds a column half way through the project.

Wilson: Yes.

Kirill: Yes. You got a data set in a weird format and it's so big that you can't open it. You don't know what to do. It just the easiest stuff that goes down.

Wilson: Yes. It's making sure that there are certain standards to interface standards that adhere with us. That's part of the most important and mind-heart is part of the job.

Kirill: Okay, fair enough. Speaking of your work at Ambiata, what would you say is one of your most recent biggest wins that you can share of us. Something that you or maybe as part of a team or by yourself you have achieved and that you're very proud of as a data scientist.

Wilson: One of biggest achievements that we've had is being able to— in just terabytes and terabytes data of website data, web log data that's not in clean format. It's these very messy

jason files, the tags are all completely jumbled and there's no system to work. We've been able to ingest that and we're also ingesting it from mini organization and being able to turn that into useful data.

That's data that basically no one else at that company was able to work with but we've got a really good team of engineers in our company so they were able to ingest that. As daytime is were actually able to work with that. As result show internal aspects of the company they've never seen before.

Being able to work with a bunch of really talented engineers who can extract the messy, messy web data that was really a big win for us.

Kirill: Yes, that sounds like a big undertaking. When you say web data is user activity data or some other type of data?

Wilson: Yes. Basically, grew went to which website lair, what articles they did they read, how long did they spend there, what kind of articles do they read, stuff like that.

A lot of that you can combine it and work out. You can use it to basically build models to find out. If I noted this person is going to do this, then who else is going to do that. Yes, it's a big piece of work.

Kirill: Yes, Kind of like classification on one side and probably you'll be able to do some a recommender type of system on the other side.

Wilson: Exactly.

Kirill: But it's surprising because by now you think you'd be all handled by Google Analytics or some other data point creation tool.

Wilson: Yes, there are a lot of companies vying for space but it's being able to do this at scale and also being able to do it quickly. It takes a lot of work to get that right.

Kirill: Fair enough. Just out of curiously, when you as a team go into a company that consulting for how who big are your teams normally?

Wilson: It does depend on the nature of the project but typically the way the project starts out is we do some simple work that it's simple to us but for whatever reason is they were able to look at the data in a different way that they have never been able to. We start out very small with basically one or two of us just getting some basic insights, including demo slides that serves to put business.

One, it gets our clients comfortable with working with us. Now you see that we can deliver a useful work to them. But on the other hand, number two is that it also gets us familiar with the data and gives us an opportunity to explore and work out what the problem is. Once we understand that, once we feel that we've got a good hand all over the data, we can start pursuing more interesting problems as well.

Kirill: Yes, that's interesting, good approach. I think that's a good idea to get the client comfortable first and then scale up.

Alright and another interesting question I have is what's your most or one most favorite thing about being a data scientist.

Wilson: My favorite part of being a data scientist is that you can let the data speaks for itself. You don't have any biases. You don't have any agenda to push. You just stay little cautious.

The genre I push is to say, "Alright. Look the lucky evidence speaks for itself." That's actually very powerful because you can cut through a lot of opinion, you can cut through a lot of beliefs that have formed in an organization over time. Yeah, very elegantly cultivate that and say, "This is what the evidence, what the data says."

Kirill: That's fantastic! I also loved that about data science that in a week there's like consulting.

Consulting is like top dollar approach, data science bottom approach. But in consulting, often times consultants don't like you know, they're hired by somebody so they often times have to say something to please that party that hired them in order to keep the job and that is often expected of them. Whereas with data, whether you like it or not, data is just going to tell the truth and the truth can be ugly at times but it's just going to tell them the truth.

Wilson: Yes.

Kirill: And there's no way around it, right?

Wilson: Yeah, the truth can definitely be ugly and it can burn us as well. But the point is that if you don't expose the truth and someone else will so you might as well get to it first.

Kirill: Exactly and it's for the best. Fantastic! We're nearing our end of this podcast or interview.

I was just wondering, from all of your experience, from your PhD, from your work at a bank, from your work in

consulting, from your work in the industry, you've seen everything about data science like I can possibly imagine. Where do you think this field is going? Where do you think the field of data science is going to be in five or even ten years from now?

Wilson: It's hard to say because like to be honest, I've never been really comfortable with the phrase "data science" because it's not a well defined term and everyone has a different definition of it and all that and there's a lot hyphen buzz passwords. My fear has always been that after the hype, we'll get disillusioned and rejected.

I think what's more important is not like data science the field itself, but data literacy. I think is more important that everyone has some basic. Not everyone needs to be an expert in this field but they have some basic understanding of what it means to look at the data and see the distribution and be able to come up with an intelligent question about it. Not just the following the same mode, just getting numbers out of Excel and doing the same old basic calculations that has always been done. I think having data literacy at the managerial level is important. I think that's what's going to happen.

I also think in terms of data visualization, I think people are going to be expecting more. People want to sit through plenty slides of the same thing over and over again. They want to be engaged and they don't want to be engaged with what they're being presented with. I think data visualization is highly important for that.

Kirill: Fantastic! I totally agree. Data visualization is getting more and more popular. You just need to look at Tableau Public

and see how many people are jumping on top of that every single day.

Wilson: Yes.

Kirill: Also with the data literacy, you can actually feel when you come into a company you can actually feel how people are starting to realize this that they know that they need to learn data and they want to because that's where the world's going.

Wilson: Yes.

Kirill: It's not data lives. It's not just in the workplace. It's definitely where everyone is going.

Your background is obviously PhD and you had a lot of involvement in Academia. If somebody doesn't have a PhD, do you think they can still get into the space?

Wilson: Yes, absolutely. Like to be honest, there're so many free online resources out there that it's just matter of dedicating the time to learn it. There're plenty of resources like learning how to program link, learning basic statistics.

It's not easy but it's an investment that you just have to spend some time in it if you really feel it's important. Yes, I think you certainly don't need a PhD for it. It's learning how to think like a scientist, that certainly helps but I think that anyone with a curious mind can do that.

Kirill: Fantastic! Love it. That's a great segway into our closing questions. Do you have any career aspirations that you could share with our listeners or any inspirations that could push our analysts and data scientists further to better their careers?

Wilson: Like for me personally, I never really studied Bayesian Statistics in university so I feel like I need to definitely read more on that and learn more about that. Again, there's a package called Our Stand which is good for Bayesian Analysis. I still need to learn how to use that properly.

In terms of inspirations, I think there's a lot of— I recently found out that Facebook has an insights blog. There're some cool papers that Facebook occasionally publishes. Yes, there're some interesting feeds going on there as well.

Kirill: Okay, definitely we're going to include that in the show notes and link to the Facebook since I'll be interested to check that out as well.

If our listeners would like to learn more about Wilson Pok, where can they find you? Where they can learn more about your career?

Wilson: Yes. Probably the easiest way is to just search for me on LinkedIn or you could go to my company's website, which is Ambiata.com. That's probably the simplest way.

Kirill: Awesome! So Ambiata is a Sydney based company.

Wilson: Yes, that's right.

Kirill: Awesome! Guys who are listening to this, if you ever need some consulting or data science work or consulting in the space of data science, go to Ambiata.com and write them an email and ask specifically for Dr. Wilson Pok to make sure you get the right person. I'm sure Wilson will sort you out and give you the best possible advice.

Our final question for today, your one favorite book about data science that can help our listeners become better at this profession?

Wilson: The obvious one is Elements of Statistical Learning, which is like the bible data science. You probably heard that one. But I'll actually name a couple more that are not so obvious.

One is the Visual Display of Quantitative Information by Edward Tufte. Like I said visualization is very important. I think it's not even a data science. It's more of a graphic design book but it has principles and how to make your data visualizations clear and impactful.

The other one that I would recommend it's more of a marketing book but it's called How Brands Grow by Byron Sharp. It's basically like a proper marketing science book. This guy is a professor from Adelaide. He's written this book about applying scientific methods to marketing. I'm going through it now and it's really good.

Kirill: Well, fantastic! Well there you go folks. Not just one book, Wilson recommended us three books that you can look into: Elements of Statistical Learning, Visual Display of Quantitative Information, and How Brands Grow. Check those out, sounds some like some interesting reads.

Wilson, thank you very much for coming unto the show and sharing your knowledge and wisdom with us.

Wilson: My pleasure. Thanks Kirill.

Kirill: Alright, take care.

There you have it and thank you so much for being here today, really appreciate you. That was Dr. Wilson Pok. Hope you picked up lots of valuable information from here. I definitely learned a thing or two.

This was episode number three of the Super DataScience Podcast. We're ramping things up. I'm super excited that this is picking up. Lots of people are starting to follow us. Lots of people are starting to listen in.

If you know anybody who might be interested, don't forget to share this episode. It's Superdatascience.com/3 and that's exactly as well where you can get the show notes for today. So head on over there and check out the show notes for this episode.

Also, while you're on the website, make sure to leave a comment in the comment section below this episode. We would love to hear from you and know what you thought of today's podcast. I hope to see you next time. Until then, happy analyzing.