SDS PODCAST EPISODE 131 WITH EUGENE DUBOSSARSKY

Kirill: This is episode number 131 with Data Science Thought Leader Eugene Dubossarsky.

(background music plays)

Welcome to the SuperDataScience podcast. My name is Kirill Eremenko, data science coach and lifestyle entrepreneur. And 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.

(background music plays)

Welcome back to the SuperDataScience podcast, ladies and gentlemen. And today I'm ultra excited about the episode because I have the legend of data science, Eugene Dubossarsky on the show. And legend is in no way, not even in the slightest, an overstatement because Eugene is indeed regarded as a thought leader in the space of data science, definitely in Australia, that I can tell you for sure, and I would even go as far as saying across the world, across the globe. Eugene is a person who has started multiple companies and has participated in multiple companies in the space of data science. Just recently he started Advantage Data, it's a world-class data science consulting firm that works with C-level executives and CEOs. He's also the director and principal trainer at Presciient Analytics. He's the Head Founder of Data Science Sydney, and he's also the Chief Data Scientist at Alpha Zetta, a global network of data science consultants.

In short, what I love about Eugene is his approach to data science. It's a really no-BS approach. His philosophy is that

data science should add value to the business and if it's not adding value, then it's not proper data science. And you will feel that sentiment throughout this episode. And in this episode we're actually going to talk about quite a few things. We're going to talk about the purpose of data science, we're going to talk about fake analytics and how to distinguish real data science from fake analytics, how to know if your business is falling into the trap of fake analytics. We're going to talk about large scale data science consulting, and we're going to talk about how executives should think about data science in the years to come and what they can do to get closer into that space.

So this episode is going to be very beneficial for anybody who's looking at the space of data science and wants to see what's going on, and in general wants to pick the brain of a very successful, very prominent thought leader in the space of data science. But also very importantly, this episode is going to give a lot of value specifically to executives, to directors, to business owners, to managers who are looking at their own businesses and their own skills in this space. So all in all, a very exciting episode. Can't wait for you to check it out. And let's get started. Without any further ado, I bring to you Eugene Dubossarsky.

(background music plays)

Welcome ladies and gentlemen. Today I've got a very special guest, a renowned data scientist in Australia, Eugene Dubossarsky, on the show. Eugene, welcome, how are you doing today?

Eugene: Thank you very much, Kirill. Thank you for your welcome, and I am doing very well here on a pleasantly not-too-hot Sydney day. We've just had a massive heatwave, but it's very nice here right now.

Kirill: Yeah, we were talking about it just before the podcast, and it's crazy. You mentioned bats. In Australia, we have bats who live on trees, and they were dying because it was so hot.

Eugene: Apparently thousands of them. Hundreds or thousands of them, from what I read.

Kirill: What was the temperature during the heatwave?

Eugene: The top recorded one, I believe, was 47 degrees.

Kirill: 47 degrees Celsius. Wow, let's see what is that in Fahrenheit. That's 116 Fahrenheit, just for those listening in the US. That's crazy. Yeah, it's been a very different summer this time of the year. We also had this heatwave in Brisbane, I think it was in October. It was pretty hectic.

Eugene: I'm just sitting back and marvelling at your mental arithmetic skills. They're better than mine.

Kirill: (laughs) No, I just went on Google! Definitely, I'm not that good. I have only Fahrenheit a couple of times, so I don't understand it that well either. Ok, well very excited to have you on the podcast. For those who don't know, Eugene -- you're probably going to be best to tell the story yourself, but I'll just give a quick intro. Eugene is a very influential person in the space of data science, and he presents at lots of different conferences. You have your own company, Presciient Analytics, you're an author of an R package, at least one that I know of, maybe more. And you are

constantly invited to events to help companies, even organisations, figure out data science on a strategic level. Does that sum it up about right, or is there anything else that you were involved with in data science that we probably should know before we get started?

Eugene: Probably worth mentioning that a lot of my focus over the last 4 years has been training more that consulting. So Presciient is very much a training business, but I've recently launched with some associates a consulting company called Advantage Data. And I'm also branching out with Advantage Data into the space of data valuation and commercialisation, and doing a fair bit in the startup space as well. Plus I run a number of communities. I run Data Science Sydney, which I'm proud to say now has 4200 members and growing all the time, and hopefully events every week, and a few other things. But probably worth saying that I've been doing this for about 20 years now, commercial analytics, commercial data science, that is.

Kirill: Fantastic. Thank you for the outline. That's very impressive, 4200 members. And I'm sure we'll talk a bit about that. But to get us started, we know now the different areas that you're in, and it's really hard for me to even fathom where to start. What is probably the part about data science that you're most passionate about, given your 20 years of experience?

Eugene: There is the geeky component, where the mathematics, the philosophy, the sheer cognitive, conceptual framework, is just so damn impressive. But it's a completely different universe to the universe of dealing with human beings and

supporting them in their decision making process. Which is surprisingly a different thing again than the, I guess, abstract business goal of just achieving success, you know, of building a model that makes money, or makes the world a better place, or wins a Kaggle competition. So I guess there's a sort of success drive, there's very interesting issues around people and their decision making, and how they want or don't want their decision making augmented and their real incentives, and whether they actually want to make decisions, that they pretend that they're decision makers. And then there's the geeky mathematical stuff. And I love it all. It's all good.

Kirill: Nice, fantastic. Okay, that's a good summary of what data science is and there's lots of elements. Probably the way I'd like to proceed from here, I want to mention how I met you for the first time. I was at a conference in Sydney, where I was working for SunSuper, a pension fund, at the time. And you were presenting. It was more kind of like management – even though I wasn't part of management, but somehow I got onto that conference – more of a management-oriented presentation and conference. And you were talking about data doesn't lie, that data will always tell you what it is. I think your presentation was entitled like, "The truth isn't always pretty," and you were talking about how management often expects something from data, but it shows something else. That's probably the first time I started to think about data in a strategic sense. I know that that's quite a big part of what you do in your role. Could you tell us a bit more about that? What is data in terms of strategy for companies these days?

Eugene: All right, so here is something that I find surprising. The bit I find surprising is that everyone else finds it surprising. So when I say to people there's only one purpose to data science, and indeed data analytics in general if you want to say data analytics is bigger than data science—I'm going to use the two terms almost interchangeably. I think the boundary is very fuzzy and the exception is negligible. But there's only one purpose to data science, and that is to support decisions. And more specifically, to make better decisions. That should be something no one can argue with. Would you agree with that?

Kirill: Yeah, I can agree with that.

Eugene: Well, the first thing I find is, the moment I say that, people start putting up hands and suggesting why data science may be other things. And inevitably, either what they're thinking of is really just another case of supporting decisions, which is fine, people have all sorts of distinctions. But worse, a lot of people are engaged in activity in their jobs. You know, they're being paid money and they're building careers on activities that aren't actually in support of decisions. They actually serve other purposes. Now, in some cases those purposes are not entirely useless, but they're still not what I would call analytics and data science because it doesn't support decision making.

Also, what is more interesting, a lot of people only start to look at their job and the data science they do in the context of decision making after this is spelled out for them, and only then realize that perhaps they're not doing data science at all. Because a lot of data science out there is not in the

service of decision making. What's it in the service of is a very good question, and perhaps we'll leave that as an open one for now. We can explore it and that can be a half-hour conversation and you might want to move on to other things.

Kirill: Yeah. Can you give us an example of the first thing you mentioned, when people are doing something that's not in the support of a decision, like they're doing data science that's not in support of a decision and that's not completely useless, I guess.

Eugene: Okay. So, I'll give you an example that's not completely useless and I'll have to add one additional constraint to the sentence. And you will say, "How is that a constraint?" You will see in a second. The constraint is data science is only data science when it supports decisions for that organization. Now, how does that restriction work? Well, in the context of that restriction, think about all the compliance analytics that people are doing out there, you know, Basel III or any other regulatory regime, I guess SEC in the United States, compliance analytics that people do.

It's an enormous amount of analysis, there's predictive modelling, there's advanced multivariate statistics, there's copulas, there's value at risk calculations. It's very complex stuff going on, but at the end of the day the objective is not to produce something that is going to help the company manoeuvre its way through the decisions it needs to make to avoid risk and find opportunity. Rather, it's done to make a regulator go away. And the objective is not to do it as well as you can because your competitors might be doing it better.

The objective is to do it well enough to make the regulator happy.

Now, that may support the regulator's decision making in some cases. We can debate whether it does or not – sometimes I guess it does, but it sure as heck is not done with the primary or indeed any purpose of decision support for the organization. So that's compliance analytics. Now, I'm not saying compliance analytics is useless, but I'm saying there's something qualitatively different between compliance analytics on the one hand and the sort of analytics you do when you're supporting decisions in an organization. But there are worse things, which are completely useless, and I see a lot of those.

Kirill: Any examples of that?

Eugene: Sure. You have a certain kind of powerful individual who makes a career for themselves by being associated with the hot trending topics. They realize that this data analytics stuff is a hot trending topic, so they're powerful and they mandate a data analytics function to the creator because they're that powerful. Now, that data analytics function doesn't actually have a purpose, it doesn't have a mandate. And it's not clear what, if any, decisions it's meant to support because no one in the process of mandating these functions said, "We have these decisions that we need to make better. This is what better looks like. This is what worse looks like. And this is the kind of input we would like to improve these decisions."

Rather, it's "This advanced analytics stuff looks really, really cool. We'd like to be known for doing it," and first order of

business is to hire someone usually not super qualified in data science, I might say, to be the manager—not hire, maybe promote internally. Hire a bunch of data scientists, usually not give them the support or the resources they need to do their jobs – and there's a lot to say about that. By the way, for me, one of the red flags that tells me that a useless operation is going on is when you've got expensive people, but they're not given the data, the hardware, the access to open source tools, indeed any of the support they need. And the most important support of all is having a customer, having someone who actually makes decisions from the analysis you produce. And when a function doesn't have that and the manager of the function is basically a salesperson whose primary job is to find someone in the organization to care, the analytics function is effectively a consulting function that's constantly looking for customers. That tells me that its primary purpose is not to support decisions.

Kirill: Interesting. Very interesting philosophy. And what would you say to our listeners, how would you tell them to self-reflect and look at their roles and understand how that fits into what you just described? Like, are they doing data science in a useful way or are they falling into the pitfall of one of those other examples that you gave just now?

Eugene: We can do that. What you may want to do at some point in this interview is also to convince people that it's in their interests to try to do the real stuff. I mean, ideology and rah-rah data science aside, I think that as the industry matures, the sort of people who've optimize their careers for pretend data science effectively are going to be in for a rude shock.

Now, what do you look for in a functional data science team? This also relates to something else which I'll bring up, which is I get a lot of people coming for advice, mentoring, and one of the things that comes up again and again is things that I tell them to ask prospective employers and things to look for with prospective employers.

So, here is Eugene's tell-tale sign that something might be wrong. Now, it's not a 100% correct—I'd say it's about 98% correct if confirmed. And if it's not confirmed, there may still be problems. But here's Eugene's first tell-tale sign. If you've got a data science function where enormous expense is being spent on a data science team—and hey, we're expensive. There's a few data scientists, there's management, there's support, etc., but the hardware that the data scientists get to use is effectively the same hardware as any other resource in that organization, it's not in data science. And the extreme case I tend to quote, which is now probably a little bit dated, is just a 2GB RAM laptop, and they also have no access to the cloud. Obviously if you have access to the cloud, the power of your desktop isn't that important.

But when they have absolutely tiny, inadequate laptops to work with and no cloud access, you know that something is very wrong. And what's more interesting is when you ask them why this is the case and they start um-ing and ah-ing about oh it's political, or it's sensitive, or we have a bureaucratic process, you have to ask yourself, how is it that someone can command the enormous budget to hire people at $150,000-$200,000 a head, and yet not have the interest to cut through a little bit of political red tape to give

this team the resources to do what they actually need. And how can the state of affairs be in existence for months or years? Because very often it is.

And it usually goes hand in hand with other things. Like, they don't really have access to good data either. And usually any engineering resources are either entirely absent and PhD level data scientists are expected to be doing all of that data munging, which in many cases is fine, but if they're freshly out of university, in many cases they don't even have those skills. They don't know how to use enterprise data systems to extract data. And it's probably not the best use of their time. Or worse, the engineering function is in a different silo, which is incidentally also the silo that's meant to be providing data and often isn't.

And then there's the resourcing issue. Usually when you have a situation like this, the sponsors have a very clear idea on software and the software is usually provided by some big name vendor and probably costs about as much or more than the human resources. And often this is not entirely to the data scientists' liking and not a terribly good decision. But yeah, apart from getting some big name vendor technology and hiring a bunch of expensive data scientists, there is a bottleneck with resourcing. That's usually a sign that something is very wrong.

Kirill: Okay. That's some great examples. I can totally agree with that. I've been in situations where you don't have the right tools for the job even though the team is being built and usually that doesn't go that well. I'd like to touch on what you mentioned before. We can see now, with your tell-tale

signs, how to look at these types of things. What would you say to that same question that you asked? You already started answering this question, but why is it important to be doing data science rather than pretend data science?

Eugene: Well, I actually think that the age of pretend data science is not going to last for very long. It's a temporary state due to the state of relative affluence in some Western countries at the moment, a state of relative immaturity in the managerial class. All these things are changing. A little bit more maturity, a little bit more competitive pressure, and I think data science will have to get a whole lot more real. My view is I don't know when there is going to be a big economic crisis in Australia or if there's going to be a crisis, but all I can say is we're world record holders in sustained economic growth of 26 years. And I'm inclined to believe that there's at least some support for the argument that what comes up must come down.

I'm not saying I know when or if the Australian economy will come down, but I will say that I strongly believe that when it does or if it does come down, a lot of the pretend data scientists are going to lose their jobs because they're not actually doing anything people understand and they're costing a lot of money. And I think with economic crisis, overnight we're going to have a very different data science industry.

Kirill: Okay, gotcha. I guess our listeners can really tell from the start of this conversation that you're a very straightforward person and you speak the truth as it is. And when you do consulting and you go into organizations and you see this

happening – for example, you see pretend data science or you see things that are happening not in the way that is sustainable or is going to grow into a proper data science practice in the future. Can you tell us how you approach the conversations with management? I'm just curious to understand how do you tell them that this is happening in their organization?

Eugene: Well, the truth is that the sort of people who hire me are not into this pretend stuff. I've deliberately made myself too unlikable to the pretend side of the market. You're saying I'm very straightforward. That polarizes my market. I'm sure there'll be listeners who are going, "I'll never hire this guy." Good, fine. Also, I'm very expensive. Being very expensive is good because if people actually don't understand the value of data science other than as CV fodder, then they see it as a commodity. And as a commodity, I'm very expensive.

But if they have real problems to solve or they find themselves between that rock and hard place of needing to do something to make better decisions and needing to act on the data they have but having no idea where to start and being honest about having no idea, then they talk to me. Now, what does happen sometimes is that I will have conversations such as I described with prospects and I come across these situations and I guess we very quickly both realize that we're probably better off not doing business. I also come across situations, much more commonly, where data science, for better or for worse, is now huge. So it doesn't exist in one little place in an organization; it exists in all sorts of places and there will be silos and sometimes there will be warring silos. Sometimes, in fact—not my

employer, but some other silo of the organization may have some of these issues and they would benefit from me pointing it out, which I can do quite honestly and provide strategies for appropriate manoeuvring in that context.

Kirill: Okay. And I think this is a good transition. Actually, just to comment on that, I really agree with that approach. I really think, for anybody who is looking to do consulting in the space of data science, that they can save a lot of headache. When people know that you're expensive and people know that you are not going to sugarcoat it, you'll say it as it is, you will only get the people that you want to work with. You won't have to adapt to the wrong type of client. I think that's a very smart approach.

Eugene: And I have to say, shout-out to my clients, they're fantastic people and they're terrific to work with. What this selection process does, it doesn't just give you the right people, but it gives you the right way of working with them. You can manage their expectations perfectly.

Kirill: Fantastic. On that note, can you tell us a bit more about your whole consulting business? Let's say I'm a client, I come to Eugene and I need something done. What exactly, what kind of services do you provide or how is your day-to-day structure? You go into a business and then what happens?

Eugene: Okay. There's what I was doing yesterday and there's what I'm doing today. More precisely and less metaphorically, there's what I was doing until about three months ago and then there's what I'm doing now. Until three months ago, it was just me at Presciient, which is primarily a training

business, and I would do consulting but with a very particular client, which I still do. I call it true consulting because it's not a contracting job. I'm not being a casual employee of the company. I'm not applying cookie cutter [indecipherable 26:01]. What I'm being is an advisor. I do it on a retainer basis and usually the way it works is that my clients buy me in four-week blocks which they renew on a rolling basis, and they have access to my time for one day's worth a week. And because many of my clients are not in Sydney, sometimes I might not see them for a month or so. So it's not all face-to-face, but it's quality, not quantity.

In one case, I think I effectively did a whole—well, I added value that was needed to be added in about three hours, and that was effectively a month's work, the rest was detail. So there's this advisory thing where I may touch data, I may cut a bit of code, but it's very much experimental. And generally I'm helping people do things like start analytics functions or launch analytics projects or figure out strategically how to position analytics, how to promote analytics, how to build a team, how to also improve what they're doing.

So there's the strategic side, but there's also the geeky technical side where sometimes it's a matter of, "We know exactly what we're doing. We're building a predictive model. Here is our out of sample accuracy rate. We want to improve it." So you go from being like a Big Four management consultant to being like a Kaggle competitor.

As I said, I like both words. What's changed is, with Advantage Data, I'm now part of the team and now we're doing implementation/execution projects as well. For

example, one of the things Advantage Data has done just now is developed a rather innovative credit scoring system for a loans business. And we're also putting together a comprehensive clinical research and analysis framework and performing analysis for a medicinal substances company. So there it's a matter of guiding our existing resources, reviewing the work, managing by exception and of course interacting with the client, which is the most important thing of all.

Kirill: Interesting. So Advantage Data, is it correct that you're more on the product side of things, analytics products?

Eugene: No. Well, that's not out of scope. I think product is the furthest thing from where I've been. I probably want to say a little bit more about analytics as decision support versus analytics as a product. I sure have a lot to say about that, but Advantage Data is everything is in scope. Product is in scope and I guess it's implementation, deployment that's in scope. It's not just high level advisory.

Kirill: Okay, I understand. So you have a team. If you're able to disclose this, how many people do you have on your team?

Eugene: How long is a piece of string? I think there's four of us that are full-time and sort of a growing cloud of part-timers that sort of coalesce into full-time.

Kirill: Interesting. So why this shift? Why did you decide to make this transition three months ago?

Eugene: The opportunity just came to me, to us. I was working on a number of projects with the core team members. We found that we worked together very well and we found that a

number of opportunities were coming our way so we went and grabbed it. I should also say there's another thing that I'm involved in, which is I'm now Chief Data Scientist of a global consulting firm called AlphaZetta which is in 27 countries and has 330 consultants and just had a very interesting and fun conference in Bali. They're a very interesting outfit because all AlphaZetta consultants are independent. And something else that I've noticed is that the best data scientists go off on their own. So, this is a consulting firm that leverages globally the capabilities of the elite independent analytics professionals, data scientists. And I'm very proud to be nominated their Chief Data Scientist.

Kirill: That's really cool. Congratulations. That's a huge thing. And how do you manage all these three things at the same time? You're not just in the space of consulting. You do a lot more for the community and so on. Where do you find the time?

Eugene: I think a lot has to do with the fact that the training is something that I can, I guess, do on a dime now. I'm not letting go of all my courses, all my machine learning courses I'm still teaching myself. The language courses, the R and Python courses, are now being taught by other people. And managing and running training is something I can do without it consuming full or even half of my time. And with consulting and with the communities and everything else, the key is to be working with good people. And to only do the things that I do well that others might not be able to do. So just add the value that I add, which isn't really measured in hours. That's not how I work. But managing by exception,

reviewing, giving key input when it's needed – somehow it all works.

Kirill: Gotcha. What does managing by exception mean? You've mentioned this a couple of times now.

Eugene: What I mean is, don't micromanage and don't dictate the same process over and over again. Trust that you're working with smart people and trust that they've got it handled when straightforward things are happening. They come to you when they have a problem or you come to them when you see that they've done something wrong.

Kirill: Gotcha. So managing when there is an exception rather than all the time?

Eugene: Yeah, outlier detection.

Kirill: Okay. Moving back a little bit to the consulting that you've been doing in Presciient up until three months ago, what is your view on executives and data science? We're seeing a huge shift just in the world towards data. Do you think that it's important for executives to be proficient in the space of data science?

Eugene: Okay, what you're really asking about is a topic that I like to talk about called data literacy. So let's have an analogy. You see, I think there's a data literacy or more generally logical literacy revolution coming. It hasn't happened yet, but I think it's coming. My analogy is computer literacy. With computer literacy, the amount of things that you and I can do with computers—indeed, you and I are not good examples. The amount of things just about every human being we pass by on the street every day can do with a

computer these days would make them look like a computer whiz in the eyes of someone 30 years ago. Just in terms of what people can do on a smartphone now. The computer literacy of people who use a smartphone is mindboggling and someone 30 years ago would have seen that as being a computer specialist.

At the same time, computer specialists haven't gone away. 30 years later, we have probably more IT people than we've ever had, right? Except everybody is also computer literate. Similarly, I think the executive of the future will need to be much more literate about data and just be a much clearer and more methodical thinker than many are today. And the problem is, we've talked about this issue of fake data science. The reason fake data science happens is because it's much easier for people to get excited about things than it is for them to understand those things, and that includes understanding how they're meant to benefit from those things or what their role is with those things. So, I think executives are getting more literate slowly, I think they need to get a whole lot more literate. And the question is, is the existing executive class going to become adequately educated or will they be replaced? In fact, will entire companies be replaced by smarter companies?

Kirill: Gotcha. So existing executives, what can they do to get more literate? Do they need to take a course? Do they need to read a book? What's the best approach?

Eugene: I think it's just a matter of get involved. I think courses are great. I think books are good. I think getting mentoring is key. I think the big gap, whether it's an executive or an

actual aspiring data scientist, the big missing piece is good mentoring, as in work with someone who actually knows the stuff, have them review what you do, give you advice, and suggest next steps. I think it's vital for execs to also think about what problems they're solving first rather than think about what topic is exciting first. And if I may throw one more thing into the mix, I really want the executive class to stop thinking of this whole data space as an IT space. That's probably the biggest stumbling block they've got.

Kirill: Where would you put data science in an organization?

Eugene: That's a good question and the answer, as always, is more difficult than you'd like. Because the answer is, I would put it in a part of the organization that doesn't currently exist. And that's intelligence. So in the military, they have intelligence. They have a function whose job is not to do projects, not to do business as usual, not to follow process, just to keep the organization's eyes open. So in an organization that genuinely wants to avoid risk and seek opportunity, you need that function and that's where analytics lies.

And to be a bit friendlier to existing structure, I'd say it should be in the strategy function. The only reason I'm hesitant to say that is because I don't think a lot of strategy functions actually do strategy in the sense of supporting strategic decision making.

Kirill: Gotcha. Okay, that's a very apt way of putting it. I've been in organizations where—at Deloitte it was of course consulting, external thing, but then in SunSuper it was more in the marketing space at the time, I don't know what it is like

now. Yeah, I've seen it in different scenarios. And then speaking of strategy, what is your definition of analytics strategy or data strategy?

Eugene: Okay, there are two aspects and they're completely different. There's the strategy of analytics and there's strategy from analytics. They strategy of analytics is, at a high level, what do we need to do to make analytics happen? And the strategy from analytics is, at a high level, what do we need to do now that we have access to all of this analytics? What does the organization need to do to open its eyes? And then there's what does the organization need to do now that its eyes are open?

Kirill: Gotcha, okay.

Eugene: So that's the first question. And they're both important questions. So, how do we—of the two, the most important one of course is what we do once their eyes are open, what does the organization do? So, what are the decisions that the CEO, the board, the CFO, the actual C-suite, not the pretend C-suite, actually needs to make? How can those decisions be better, not easier? This is important. Not how are these decisions made easier. They've got to make them harder because you've got more to think about. But how do you as a result make better decisions because you know stuff that you didn't know before? That's the strategy aspect of analytics. Does that make sense?

Kirill: Yeah. I'm just thinking of something that you said at that presentation, that conference. Correct me if I'm wrong, but you've had a very specific view on the term actionable insights. I'd love to revive that conversation.

Eugene: Okay. So, this follows on from what I've just been saying. Interestingly, I've been a bit of a hypocrite, but sometimes I use actionable insights not in the same negative context. But what gets me is that a lot of—like I said, there's a lot of pretend analytics, as in the analytics isn't used for any decision making at all. We're now talking about something slightly more benign, which is actually being used in decision making, but there's pathologies in how it's used.

One thing I've come across is executives saying, "We want actionable insights." And you go, "Well, what's an actionable insight?" And what you realize they mean is, "Don't give me stuff that I have to think about. Don't give me half-digested insights that are just going to make my work more difficult. Give me what I should do." And I'm thinking, "Well, that's great. You want a machine that's going to generate this organization's strategic decisions." That's a bit far-fetched, frankly it's silly, but okay, let's say you want that. Let's say you get that. Let's say you're an executive who gets those actionable insights that tell them what to do. I have one question: What's the role of the executive now?

Kirill: Exactly. Why do we need you, right?

Eugene: Yeah. So not only – it betrays the level of cognitive mediocrity, it betrays a disinterest or inability in thinking and synthesizing information, which by the way, along with data literacy, is something that's going to be a huge trend with executives. Executives are going to be information-hungry geeks. I don't think the executive of the future will be the kind that doesn't like thinking, that doesn't like new information. But a lot of current executives don't like

thinking and they don't like new information. And in that case, if it's not an actionable insight, meaning pre-digested, premade decision, they're not interested. Well, the future doesn't really have a place for them.

Kirill: Yeah, I agree with you. The current or more traditional type of executive that we're seeing now is very different to what I can see happening in 10 years from now.

Eugene: Let's dwell on that word, 'executive.' It's a very interesting word, isn't it? Because it implies that the important skill of a powerful senior person is their ability to execute, their ability to make things happen. The interesting thing, I think, over the last 10 years I think we've had a very interesting trend and I think Eric Ries of The Lean Startup might have put this very eloquently. Unfortunately I don't remember his exact quote, but it was something like this: that it's getting easier and easier to do stuff. We're getting better and better at doing stuff. We're getting better and better at building things. We have better technologies, better methodologies, for actually achieving objectives. It's becoming cheaper, it's becoming easier, it's becoming faster. If we want to get something done, it's relatively cheaper, easier and faster than it was in the past.

But we're living in a much more complex, much more rapidly changing, and much more competitive world. And therefore, the thing that's becoming more and more difficult is knowing what to do. So doing things is getting easier, but knowing what to do is getting harder. So, my point is that being an executive, being someone who makes things happen, I think it's going to be a relatively low-level thing in

the future. But being a good decision maker is going to be the elite skill. And I'm hoping that we have a different word for important person that isn't 'executive,' because I think it's their decision making skill that's going to matter more than making things happen skill.

Kirill: Okay. And you mentioned that ideally, if they got this machine that can help them spit out these actionable insights, what's the point of the executive? We in some way are getting closer to that. What are your thoughts on companies like DataRobot, for example?

Eugene: Okay, something that unfortunately gets swept under the hood—by the way, I love DataRobot, I love their product. A friend of mine has recently joined their team. Xavier Conort, who is their Chief Data Scientist is also a friend and an amazing data scientist. Shout-out to Xavier. What I will say is that I'm talking about analytics now, we're talking about analytics now and interchangeably discussing at least two very different things, one of which is what you might call strategic analytics, which is analytics supporting decisions the CEO might make. And those are based on insights, they're qualitative, they're complex, each decision is different, each decision is very valuable, and they're relatively rare. These decisions aren't made every second by that executive. They're made maybe every few weeks or every few months – hopefully more frequently than every few months – but they're not made every second.

On the other hand, you have what I might call operation analytics, of which predictive modelling, supervised machine learning, is a perfect example as it's applied, which is it's

usually applied not to replace the decisions of the CEO, it's used to replace the decisions of possibly the most junior people in the organization, the sort of people who might have been manually deciding who to send campaign information, who to target in a campaign or—okay, maybe not the most junior, but relatively junior, like underwriters in an insurance company. And every decision is exactly the same, it's do we or don't we give this person an insurance policy? Every decision is relatively inexpensive. We're talking about hundreds of millions of dollars in total, but every decision only concerns a few thousand dollars.

So, there's operational analytics which is very, very frequent, very small, very similar, very junior, and there is strategic analytics which is very big, each one is very different, each one is very complex and it supports the decisions of the most senior people in the organization. Now, I'm not seeing machine learning doing much to support the decision making of the very senior people. You mentioned DataRobot. My understanding of DataRobot is it's a way of automating the machine learning process, so it will be great for a retention model or it will be great for credit scoring, for most other the things. I won't say you can't use it to run a company, but perhaps I'm not aware of those sorts of applications for supervised machine learning.

Kirill: Okay. So that component that you were talking about, the intelligence component inside an organization that's driven by data, that is still relevant and that's not going to be replaced anytime soon?

Eugene: Oh, absolutely not. Look, there's a long conversation to have even about what human components—I don't think we'll ever be replaced in the context of operational analytics, but admittedly, with things like DataRobot, it's much more highly automated than it used to be. I actually don't think that even in operational analytics automation removes the need for data scientists.

I think the Random Forest algorithm was basically automated machine learning compared to everything that came before it, but what's interesting with Random Forest is that it didn't reduce the number of data scientists – if anything, it increased it. I think data scientists are just needed to make much more complex decisions about well, what's our target variable, what problem are we trying to solve, what's our success criterion. And still, Kaggle is still being won not by automated algorithms, but by people. People who actually understand the subject matter and can therefore do manual and highly specific feature engineering. I don't think that's going away just yet.

Kirill: Okay. Shifting gears a little bit, there's a famous quote by Andrew Ng that AI is the new electricity, AI is going to become prolific. Proliferation of AI means that it's going to be everywhere, every business is going to have it. Do you think this is the case? And if so, how fast do you think AI is going to take over businesses?

Eugene: Well, I'm still trying to work out what people mean by AI. I've heard so many definitions of what AI is and what it isn't, especially vis-à-vis machine learning, which is apparently not the same as deep learning, which comes as a surprise to

me. Let's take electricity. Imagine it's 1900 and I told you that electricity is going to take over businesses in the future, because in a sense it has, right?

Kirill: Yes.

Eugene: But in another sense, it's taken over aspects of the business and not others. Somehow people are still not using electric pens and people are still not wearing electric shirts and the carpets aren't electric. Not everything is electric just because electricity has taken over the business. So, one interesting question with AI, however you define AI, is what is AI going to take over in the business? Now, I don't dispute Andrew Ng's point that AI is going to become ubiquitous in businesses and much more prolific than it is now.

We could also trivially say it's already happened because everybody in the business has a smartphone and every smartphone has a Siri or an Alexa or—what's Google's assistant called, I forget. I've got an iPhone. That's a kind of trivial way for that to be true, but not terribly interesting, right? So I'd like to understand in what way will AI take over businesses and I'd like to temper that comment by addressing this implicit expectation a lot of people have that AI is AGI, that we already have artificial general intelligence. I think a lot of laypeople are under this misunderstanding that AI is basically as smart as us, it's creative, it can do anything. And it's just not, it's not that smart. Not yet, and I don't think it will be for a while.

Kirill: And what is your view? Do you think data science is something that will morph into artificial intelligence the

more it becomes ubiquitous in businesses, or it's going to be two separate elements?

Eugene: I think you're talking about brands rather than practices. About 10 years ago, I was doing data mining and then one day I was doing data science. I can tell you that I didn't notice a real change. It's just a brand. And similarly, whatever I'm talking about with AI, in some cases it's a different thing. When I use Siri, I'm using AI and it's not the same as me doing machine learning. But are we going to need professionals whose job it is, as human beings, to infer things from data that machines can't automatically? Absolutely. And is the existence of more complex, more highly automated, more intelligent tools for obtaining insights from data and advantage from data going to make more people more rather than less necessary? Yes, I think it's going to make good people more necessary.

Kirill: Okay. My question then is, people who are doing data science now, who are studying data science, SQL, Tableau, R, Python and so on, is it a good idea for them to start looking into things like Keras, like TensorFlow, PyTorch and so on, to start going towards that rising trend of artificial intelligence? Is that where their job is going to be 5-10 years from now or is it going to stay data science as data science?

Eugene: Okay, so by your definition, AI is deep learning and deep learning is not data science?

Kirill: I would say for most people who do data science now, I don't think they apply deep learning in that space.

Eugene: Okay, that's interesting because in my experience, I've seen people apply that distinction, I've also seen people apply a

distinction between AI and deep learning. And I find that for some people AI means chatbots, which needn't mean deep learning or any machine learning at all. Let's leave out distinctions aside. Let's just address your question, which is, should people be learning deep learning? Perhaps. The interesting thing for me with deep learning is it's very powerful, certainly very powerful.

But it strikes me as something where I don't know that it's as ubiquitously applicable as more traditional machine learning is in the sense that the sort of problems that deep learning is really good at solving, like image problems, speech problems, text problems, you generally only need to solve them once. I know that's not always the case, I know that there are lots of innovative solutions. And in the cases of startups in particular, there's lots of room for innovation with deep learning tools.

But the sort of things that most white collar organizations, insurance companies, banks, telcos, et cetera need to solve, you know, retention, credit scoring, fraud detection, I find that the other stuff is more than adequate and often better.

Kirill: Interesting. So your comment then would be that deep learning is not a necessary skill right now for somebody to rush into?

Eugene: Well, here's the thing. Are we talking about what's going to get you employed or are we talking about what's going to keep you employed in 5 years' time? Right now, it's a very hot skill. There's no excuse not to get some skill with deep learning. The question is, should it be your only skill? And

should it be the thing that you invest a ridiculous amount of time in, or do you just get literate with it?

Yeah, you definitely don't want it to be your only skill. And the other thing about deep learning, deep learning is great if you're building models for accuracy. Not all data science is about accuracy. Indeed, not all data science is about supervised machine learning. The other thing I'll say is that if you want to be a data scientist 10 years from now, I think the key skill is going to be statistics. I think if you're investing your time in learning IT frameworks and packages and treating it basically as an IT activity, you're not going to have the elite skill of actually understanding the data. Statistics is that skill and in the future I think it will matter.

Kirill: Very interesting. Why do you think that?

Eugene: Well, the data is telling you something. We keep talking about insights. The data has a story to tell. Statistics is basically a language that lets the data tell its story. It tells you the story of uncertainty in the data and the story of complexity and relationships in the data. Now, if you're just interested in pressing a button and creating a machine that's going to give you a certain degree of accuracy for a predictive model, maybe you don't need to know that stuff. But the moment anything goes wrong with the way your problem is formulated or something goes wrong with your data or you've got a completely different task where your job is to actually try to understand something, God forbid, rather than just achieve an accuracy score, lacking statistical skill, you're going to get very lost. You can produce all the pretty visualizations in the world, but you

won't really understand what they're saying. Or worse, you'll think you understand and you might miss very important things or misinterpret it.

Kirill: Okay, that's a good point. What kind of statistics are we talking about?

Eugene: Well, I'd start with the basics. I think the most elite skills, the most crème de la crème statistical skills, statistical skills that I'm aspiring to acquire at the moment, and I can't say I've acquired, the things that I find in the most elite data scientists are the sort of things I think econometricians these days have, and that is Bayesian statistics, a really good intuitive understanding of Bayesian statistics and being able to do predictive modelling using the Bayesian approach rather than the classical approach.

And also inferring causality. So, the various statistical or econometric AI epidemiological techniques, for however imperfectly, inferring not just that correlation occurred but, in fact, that A caused B. Because if you're making decisions, you want to know that when you do A, B is going to happen. Especially if you're doing something like running a country, running an economy.

Kirill: Yeah. Like they say, correlation doesn't imply causation, right?

Eugene: That's right. And when you're making very big, very costly, very risky decisions, you want to know that there is a causation there. You want to know how the causation works. And at a more operational level, next there's the action systems. They're basically causal impact machines. If you don't understand causal impact, although admittedly

that's done with a much more easier way of growing it, which is if you're doing it properly with randomized controlled trials.

Kirill: Can you elaborate on that a bit more, causal impact? What do you mean by that?

Eugene: Well, you've got a bunch of data. You've got a bunch of things that happen, but you can see that when certain things happen, other things happen too. Or when certain things happen, other things happen shortly after them, right? But as you said, correlation doesn't imply causation. What that means is we can't therefore know that when we do the first thing, the second thing will happen. It may be there's a third thing that causes both of them, we just don't know.

But there are statistical methods that help us, at least to some extent, at least in some circumstances, unravel that and give us a bit more assurance that when we do A, B will happen with a certain probability. That there is a relationship between the action and the outcome that's causal, that A causes B, not just that A and B happen together all the time.

Kirill: Okay, very interesting. An interesting perspective, I don't think I've heard that one on the podcast before. Statistics – specifically Bayesian and causation statistics. Okay, I've got a couple of rapid-fire questions for you. Ready?

Eugene: Sure. Shoot.

Kirill: Okay. What technique do you use most commonly?

Eugene: Again, in what context? If we're talking machine learning and supervised machine learning, I'd say Random Forest is my number one go-to tool because it's so easy to use as a sort of a can opener to try to understand what's going on in the data. And even though it's not one of the most interpretable methods, there are ways of inferring interpretability, but I find that it's powerful as a simple supervised machine learning method, but also powerful as an outlier detection method. There's a video online on the Data Science Sydney channel where you can see how you can use Random Forests for outlier detection to detect mistakes in the mapping of an electricity network, for example. So yeah, Random Forest is probably the number one favourite.

Kirill: Gotcha. Next one, what's your favourite software tool or programming language?

Eugene: I'm an R junkie. I've got nothing against Python and precisely because of deep learning I am getting more and more familiar with Python, and I have a Python course in the Presciient courses suite. But yeah, R is pretty much my go-to tool for everything.

Kirill: And this Presciient courses suite, tell us a bit more about that. These are online on-demand courses or is this something that you teach in person?

Eugene: It's something I teach. They're face-to-face courses. The flagship course is basically like a machine learning 101 where you don't do any coding, you use a point and click tool, but it's to learn the most important concepts in

machine learning, to get people intuition, a feeling for the most important aspects of machine learning.

Kirill: Intuition is so important.

Eugene: It's also hard to get. It's very hard to get in this space.

Kirill: This is Sydney-based, right?

Eugene: No, it's all over Australia. It's been taught in Singapore as well, and I'm hoping this year to take it to New Zealand.

Kirill: Nice.

Eugene: And if there's demand from anywhere in the world, I'll go and teach it.

Kirill: Fantastic. Okay, we'll get to the links at the end of the podcast. Next question: What's the biggest challenge you've ever had as a data scientist?

Eugene: I think the biggest challenge I ever had as a data scientist was to figure out what the heck to do with my life given my views of my role and the industry as I've described it. It was basically how do I structure a satisfying and a lucrative career in data science, which I love, while I feel like the industry is still too immature to appreciate the good stuff and do it properly. But I think I'm on top of that now.

Kirill: Gotcha. Yeah, I know, I can totally agree with that.

Eugene: All it took was three hours of soul-searching and a very nice café in the old city of Chiang Mai.

Kirill: Nice. What I like about the way you structured yours is that you stayed true to yourself. If you see something is fake analytics, you won't hesitate to tell someone it's fake

analytics. If you don't like people asking for actionable insights, you will tell them, "Actionable insights is not what you want."

Eugene: Well, maybe not that rudely or maybe not to their face, but I think usually it's their employees that don't want to be there that need to know this.

Kirill: Okay. What is your one most favourite thing about data science? We kind of touched on this question at the start. You mentioned there's lots of areas that you love about data science.

Eugene: I just find that when data science is in an appropriately sort of competitive, challenging context, for me it's my creative medium, it's where I get to create. It's where I get to use everything I know and heck, it's still not good enough, but sports betting, financial trading, Kaggle competitions, they're the most fun things in the world. I wish I had time for Kaggle competitions. I think they're the most fun things to do. I'm not a top athlete, I'm not a great artist. This is what I do. This is what I'm given to think about.

Kirill: I totally agree. Also, this one we kind of touched on as well, but maybe a summary would be great. Where do you think the field of data science is going and what should our listeners look into to prepare for the future that's coming?

Eugene: Well, I think there's going to be a general data literacy, which means a more demanding clientele in the buying side of the market. Something I'm betting on very heavily at the moment that I should have mentioned before is quantum computing. Presciient has a quantum computing course coming out and there will be more quantum computing

offerings and material from Presciient and Advantage Data in the near future.

Like I said, I think statistics is key. I think getting a good grounding in statistics is the most important thing of all. I find people, mentorees come and say, "Which data science Masters course should I do?" and I say, "Do a Masters in statistics." And once again, Bayesian statistics, in particular causal impact analysis – you can't go wrong with those. I'm still wondering what sort of machine learning tools and what sort of mathematical tools are going to come to the forefront now that quantum computing is on the verge of becoming a mature technology.

So yeah, I'd say statistics, statistics, statistics. I'd say do Kaggle competitions, especially if you're a beginner, just do them incessantly. If your question is should you learn Python or R, it's the wrong question in two ways. One is, the language is the least important thing, focus on the methods. And two is, you should look at both of them and you should probably have a look at Julia as well. I think curiosity, entrepreneurship and rigor – which is a very rare combination – those personality traits, developing those personality traits is going to be key.

I think being experimental—this one is a real problem for a lot of people. One of the ways in which data science is so unlike IT is data science is about experimenting, it's about trying things and failing, it's about not knowing what you're going to find until you get there. And that's a huge challenge for a lot of people both from the executive side of the world and the IT side of the world. So being comfortable with

uncertainty and being comfortable with uncertainty about your own career. No one knows what a data science career path looks like. No one knows what we're going to be like in the future. So it's having a certain courage, a certain comfort. How's that?

Kirill: Fantastic! I got mesmerized just listening to that. Great career advice and future advice. I think we're going to wrap it up on that. I think that's a good thing for our listeners to ponder about and for everybody to ponder about. Eugene, how can our listeners get in touch and find out about all your products and follow you or maybe if there's somebody who needs some analytics advice, strategic advice, find out more about the services that you provide?

Eugene: People are very welcome to reach out and people always do. You'll usually find I'm a friendly guy. If you're anywhere around Australia, we can probably catch up for coffee in one of my businesses or your town sometime soon. If you're in Sydney, this can happen pretty much straight away. If you want to see the courses, go to the Presciient website – I'll give you the URL, Kirill, so you can put it up. My LinkedIn profile is a really good way to get in touch with me or just to send me a message. So you can find Eugene Dubossarsky on LinkedIn and just connect with me. What else? If you're interested in what Advantage Data does, just get in touch through one of those channels. I'll provide my phone number as well, you're welcome to call. Is there anything else? Yeah, I've got a Twitter handle, it's not easy to remember. It's @cargomoose, but I'll provide that as well.

Kirill: Cargo moose, where did that come from?

Eugene: That's a long story. My nickname is Moose, and one of my favourite topics is cargo cults. Fake data science is a cargo cult.

Kirill: (Laughs) Okay, I think we shouldn't go down that rabbit hole right now, but yeah. Okay, we'll definitely include all of the links in the show notes and probably even more because there's a lot of things you're working on including Meetup groups. We didn't have a chance to talk about them, but I highly encourage people to check them out. It sounds like a really cool way to connect and get in touch with other data scientists in the space. I just have one more question for you today. What is a book that you can recommend to our listeners?

Eugene: I think it depends on people's level and it depends on their interest. In terms of interest I think we can restrict ourselves to machine learning, supervised machine learning and so forth. I think if they have enough grounding in statistics, you can't beat "Elements of Statistical Learning" by Hastie, Tibshirani and Friedman. That's like the Bible of machine learning. So, for anybody with sufficient amount of maths, what you do is you get that book and you read Chapter Two, and if you can't read Chapter Two that probably tells you you need to learn more stats and maths.

If you can't read the stats and maths, two of those authors and another person published a book called "Introduction to Statistical Learning," which is much more code-based and a bit less mathematical, and it's R code. So, "Elements of Statistical Learning" is the Bible. "Introduction to Statistical Learning" is something else to try.

And if you're an absolute beginner, the books by Graham Williams, particularly his first book. Graham Williams is the current Director of Data Science at Microsoft for Asia Pacific, and I do apologize to your listeners and to Graham for not remembering the name of his first book, but the tagline I think is "From Rattle to R." He wrote a GUI, a point and click tool for machine learning called Rattle that's part of R, it's an R package.

Kirill: Is it called "Data Mining with Rattle and R"?

Eugene: That's the one, yes. Thank you.

Kirill: Not that I've read it. I just looked it up right now.

Eugene: That's one I recommend to raw beginners.

Kirill: Okay, cool. Thank you so much. We're going to recap this a little. So, "Data Mining with Rattle and R" for absolute beginners in statistics; "Intro to Statistical Learning" for those who know a little bit; and "Elements of Statistical Learning" is the one that you need to master to be an ultra-prepared data scientist, prepared for the future. Thank you, Eugene, for coming on the show. It's been a huge pleasure to chat. I know there's so much more. I hope to meet you in person very soon and maybe we will have another podcast sometime soon to cover the topics that we didn't cover off today.

Eugene: Thank you very much, Kirill. It's been a real pleasure.

Kirill: So there you have it. That was Eugene Dubossarsky and I hope now you can see why I refer to Eugene as legend of data science. You can tell that his thinking, his philosophy about this field is so unique that inevitably it influences

people to rethink their approaches and together that creates a whole movement and that's partially how data science is driven forward, through movements like that.

Personally my favourite part of this podcast was the whole philosophy that Eugene has that you could feel throughout the podcast. When we were talking about fake analytics or data science consulting on a large scale and so on, he is not afraid to say the truth about data to executives. On one hand, that shows people the true situation of things in their business and then they can decide whether to do something about it or not. On the other hand, it also helps Eugene avoid clients who would rather be ignorant and rather not accept the truth or have their own version of the truth because he sees no point in working with people or companies like that. He would rather spend his time doing things that actually interest him, where he can make significant impact on the business.

So there we go, that was my favourite part. I'd love to know what yours was. As always, you can get all of the links mentioned in the show at the show notes, which are at www.superdatascience.com/131. There you can also find the link to Eugene's LinkedIn profile and other places where you can find him and follow him. I highly encourage getting in touch, especially if you're a business owner or an executive looking for an experienced, a seasoned data science consultant. And on that note we are going to wrap up this episode. Thank you so much for being here today. I look forward to seeing you back here next time. Until then, happy analysing.