SUPER DATASCIENCE
MAKING THE COMPLEX SIMPLE

SDS PODCAST EPISODE 45 WITH EU JIN LOK

Kirill: This is episode number 45 with Kaggle Grandmaster Eu Jin Lok.

(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)

Hello everybody and welcome to the SuperDataScience podcast. I am so pumped about this episode. Today we have the legend Eu Jin Lok on the line. Eu Jin is a client manager at Deloitte Australia. And even though we worked at the same firm for about two years, we never actually crossed paths. I constantly kept hearing "Eu Jin did this", "Eu Jin achieved this", "Eu Jin finished this project." So lots and lots of great feedback about Eu Jin, but never actually had a chance to cross paths. And finally I met him for the first time today and we had an amazing podcast session together.

And something you need to know about Eu Jin, Eu Jin is super passionate about Kaggle. So this is a person who holds the title of Kaggle Competitions Grandmaster. In order to get that title, this is the highest title you can possibly get on Kaggle, and in order to get it you need to have at least 5 gold medals and a solo gold medal. He has completed 6 – or according to his LinkedIn it's 6, it probably is already more – 6 competitions on Kaggle, including ones where he got first place, second place, he's won $3,000, he's won $30,000 (well

he and his team won $30,000). So lots of incredible accomplishments in that space, and that is exactly what we're going to talk about in this session.

In this session, you'll find out everything you need to know about Kaggle, about the variety of competitions, about the tools that he's used, about the journey that he's been on, and he'll even give you tips on how you can get started with Kaggle competitions, and you'll find out why you need to be doing that as soon as possible.

So a very interesting discussion with a legendary figure in the space of analytics (well at least here in the consulting space in Australia) and I can't wait for you to hear all the amazing insights that Eu Jin had to share. Without further ado, I bring to you Eu Jin Lok, the Grandmaster of Kaggle.

(background music plays)

Hello everybody, welcome to the SuperDataScience podcast. I am super pumped and super excited today because on the line I've got the legendary Eu Jin Lok from Deloitte. Eu Jin, how are you today?

Eu Jin: Very well thank you, Kirill. How are you?

Kirill: I'm well too. Where are you calling in from?

Eu Jin: I'm calling from the office at the moment, actually! Because right now I am actually participating in a general hackathon event. So it runs for 24 hours. So I am very pumped, so this is a very, very nice time to be talking about data science!

Kirill: That's awesome! This is crazy, and I am so excited because I've heard so much about you just working at Deloitte, I kept

hearing "Eu Jin this," " Eu Jin that," "Eu Jin did this," "Eu Jin accomplished that." And because you're in Deloitte Sydney, right?

Eu Jin: I am Deloitte Melbourne, actually.

Kirill: Oh, Deloitte Melbourne, ok. Deloitte Melbourne.

Eu Jin: But I've worked across a couple of Sydney jobs and a couple of Brisbane jobs as well.

Kirill: Yeah!

Eu Jin: So you may have actually heard my name across different states!

Kirill: But yeah, finally great to meet you for the first time! And at the same time, you're in the middle of a hackathon! So what's this hackathon? Tell us more about that.

Eu Jin: So this hackathon thing is a yearly thing that Deloitte organises at around this time of the year, essentially. And what it does is Deloitte wants to be at the forefront of innovation, and this is an event to basically try and generate ideas that could put Deloitte at the forefront compared to the other competitors.

Kirill: Yeah, gotcha. And you said it's like you just finished your working day on a Friday and now you're going to be doing the hackathon for 24 hours. Is that right?

Eu Jin: That's right, yeah. It's going to be a very long day and night for me. I am pumped. When you have the passion and the drive and the determination to do something great, there's really nothing to stop you, as you know, Kirill. When you're

passionate about something, nothing stops you. I'm very pumped to do this.

Kirill: Okay. That's really awesome. I'm excited for you and I hope you pass it. So do you have an idea that you can share with us that you're going to be working on?

Eu Jin: Not sure, actually. I have roughly some idea as to maybe something to do with machine learning or artificial intelligence, but generally the whole point of this event is everyone goes in there and comes out with something new. I suppose we work with everyone to see what sort of ideas we can cross-pollinate and build something great. I suppose the whole idea is a combination of different people from different backgrounds who come together to create a much greater idea. I'll definitely push on the data science front, but I'm pretty sure everyone else will come in and say, "Oh, you know what? We could also do this," and that would create an even greater product or ideas.

Kirill: Yeah, gotcha. Okay, that's really cool. I really hope you have a great time and come up with some cool ideas. But let's get back a bit to your career. So you're at Deloitte right now. What do you do there? What department are you in and what does your work entail?

Eu Jin: I work in a team called Strategic Capabilities. We are basically an internal function to the whole firm. Our work basically entails delivering on special projects for the firm, strategic projects. So essentially, the vision of the firm overall for Deloitte is that we want to be the most innovative firm compared to our competitors. In order to deliver on those ideas, it will require cutting edge technology and very

special skillsets. Those ideas will get funnelled to our team and we will deliver on those special ideas essentially.

Our team is made up of three core skillsets, essentially: digital, data and design. I specialize in the data area and I work together with the designers to create the best, most beautiful product and also offer the digital guidance to create the most powerful platform that can support the user frontend, and data obviously will be the smarts and the intelligence behind delivering those special projects.

Kirill: Okay, gotcha. Back when I was at Deloitte, you were working in – and correct me if I'm wrong – in the data science division, right? You were facing external clients. And now you moved to a more internal role. Is that right?

Eu Jin: Yes, that's right.

Kirill: Okay. And how are you finding the shift? Is it a bit different to be working predominantly on internal projects?

Eu Jin: Oh, yeah. It is very, very different, actually. And it is a good change, as well. They both have challenges on both ends. Working with clients, you do your best to deliver on what was agreed upon with the client, essentially, but you never really see it through all the way, which is a downside. Whereas working internally, you get to see how your ideas or the work that you've created and delivered actually changes the firm overall. You can see your product gets adopted all of a sudden by the whole firm and all of a sudden people behave in a different way and it makes a positive impact on the firm. So that's really great to see. When you work internally you have that visibility to see how what you do

makes an impact for the firm and for society as well at the end of the day.

Kirill: Yeah, totally. I can totally imagine that. And that was kind of like the main drawback for me. On one hand you learn a lot and you get exposure to so many different industries while you work in consulting, but on the other hand you never see the fruition of your projects. That's kind of like the double-edged sword there. I'm looking through your LinkedIn and this is crazy. Guys listening to this podcast, this person—I'm sorry, man, but I think you're crazy. 1st place – Kaggle competition; 2nd place – Hewlett Foundation; Merck Molecular Activity Challenge; Heritage Health Prize; 2nd place Kaggle competition. How do you do this? I can't imagine. Tell us more about all of these competitions. What's going on there?

Eu Jin: Yeah. So, as you can tell, it has been a passion and hobby of mine, which I'm very, very glad for. You know, because I use Kaggle as a platform to learn and to practice my skill and harness my skill. I just love competing, in Kaggle especially, because traditionally at work you don't have the flexibility to explore different ideas and new technologies, and Kaggle just allows me to express that new idea and new technology. And it harnesses my skills, as well.

At the end of the day, it's a lot of work, basically, and it can only happen with true passion in the industry. I love data science, I love machine learning, I'm very passionate about artificial intelligence. Although my day-to-day is nontechnical because I'm a manager here and I very rarely go into detail, outside of work I make sure that I'm always

on top of the changes of what's new in AI so I use my time outside of work to hone my skills and Kaggle is the platform for me to hone those skills, essentially. I like to think that I'm just lucky to win some of those competitions because when AI was still very new, I was early enough to be in the whole area. I just did more work earlier compared to other competitors. Basically that's all it is.

Kirill: Yeah. That's amazing. I like what you said about making sure your skills are up-to-date. It reminds me of a podcast we had with Yaw Tan. I think it was episode #12. He was a manager and he was managing a big unit in a bank, like a credit scoring unit or something, and he realized that he was slowly over time getting further and further away from the technical side of things. So what he did is he quit his job and he became a consultant to kind of get his technical side of things back on track. But your approach is a bit different. You're still maintaining your managerial position, but in your free time you're doing these Kaggle competitions. I think that's a very valid approach, a very powerful one, to develop both of those types of skills at the same time.

Eu Jin: That's right. When you have a passion for what you do, you do whatever it takes or whatever works to get you ahead eventually. Yeah, many people have different approaches, so whatever works for them, right?

Kirill: Yeah, true. So, can you walk us through a Kaggle competition? What happens at the start? I personally have never done a Kaggle competition. What happens when it's announced and then you see it? Just walk us through the whole step of the journey of a Kaggle competition.

Eu Jin: You should try it, Kirill. (Laughs)

Kirill: I totally should. I would really want to.

Eu Jin: I can tell you that the journey from start to end—actually, I can even tell you the journey from doing the first ever competition, that is the hardest. It is the hardest, it is the biggest inertia to do your first competition, and especially to do your first submission. After that, you just become used to it. I think of it like doing a really big workout. When you go to the gym, let's say you have not done it before, you always get this nervousness as you look around at everyone, and you're afraid that you might be doing the wrong thing or you don't know what you're doing and you have this hesitation and that hesitation pulls you back. And what I noticed from a lot of Kagglers, and also speaking for myself, when I first started, that's the biggest inertia, when people aren't sure what they should do.

But once they start doing it and over time—maybe the first couple of competitions that they do, they might not do so well but after a while they get the hang of it. And once they get the hang of it they start to feel the high – not the high, but the adrenaline, it becomes really exciting and that excitement drives that passion and that passion keeps driving you to compete. So it brings out the competitive edge in you. That's when you start to see yourself learning in leaps and bounds.

Essentially, it's that drive, that competition, that actually really makes you learn a lot, essentially. When I first started, my first couple of competitions, I didn't do really well. I remember my first competition, I did horribly. But after that

I saw a very steep improvement and then I started winning competitions. I just think of it as—I wasn't afraid to do what's wrong. I was willing to learn from my mistakes and I just went into deep end and that gave me an edge over everyone else, I suppose, back then.

So here I am now. These days—because I haven't been in a Kaggle competition for a while, a lot of people have overtaken me and it all comes down to the same factor, you know, it's all about practice and a lot of hard work. So, the more practice and hard work you put in, the better you are going to be, essentially.

Kirill: All right, gotcha. So I'm looking through your competitions. For instance, the Heritage Health Prize – is that another Kaggle competition?

Eu Jin: Yes. All of them are Kaggle competitions, basically.

Kirill: All right. So, here you forecasted year ahead hospital stays for patients in the Heritage Network with an average error of 0.46417 days. So you were able to predict how long people would stay at the hospital even before they got to the hospital?

Eu Jin: Yes, that's right.

Kirill: Can you tell us a bit more about that? That sounds mind-blowing, it sounds like magic to me. Can you tell us how you did that? How did you go about it? What algorithm did you use? What tools did you use?

Eu Jin: Oh, yes, sure, if I can remember, that is! (Laughs) It's been a few years. Essentially, we were given a dataset that was quite heavily anonymized and we had to predict basically

whether a patient was going to be hospitalized in the next couple of months or so and how long for. So we had to predict basically, are they going to be in the hospital for a day, or are they going to be in the hospital for 15 days.

We were basically given a dataset to learn off which basically will have a patient and the type of sickness or disease that they had and a bit of demographic information. I think it was very heavily anonymized but we had some information on that, you know, learning from that particular dataset. We then applied the pattern and see for a new set of patients in the next year or so, how likely are they going to be in the hospital for a day, or two days, or 15 days, essentially.

Kirill: Okay. And what algorithm did you use for that?

Eu Jin: Back then we did an ensemble of many different algorithms. We looked at linear regression, we looked at Random Forests, and gradient boosting as well. Back then it was really, really advanced. But we didn't look into any deep learning models at all, because at that time it hadn't really kicked off yet. But we looked at pretty much most of the traditional supervised learning models that you could think of two years ago. And once we had all of them, we ensembled them and we submitted our predictions based on an ensemble of various different models. It was up to 50 or 60 different models. It was a lot, basically. I don't even remember. It could be more than that, actually. 60 models is probably just a conservative estimate.

Kirill: Wow! That's crazy. And is there a risk of overfitting when you're really using so many models just to tailor to that specific dataset?

Eu Jin: If you do it properly, ensembles don't generally overfit and that's the beauty of it. It's basically ensuring that when you use different models and you train on different parts of the dataset using different data modelling architectures, it just ensures that at the end of the day everything would just average out. Yeah, ensembling is a great way of not overfitting, if done properly.

Kirill: Okay. I'm assuming you were using Python for that.

Eu Jin: I was using a combination of R and Python.

Kirill: R and Python? Interesting.

Eu Jin: Yeah, that's right. Because surprisingly enough, what actually won't come as a surprise to a lot of people, basically Python gradient boosting is quite different to R's gradient boosting. Actually, Random Forest is the greatest one. R's Random Forest is vastly different from Python's Random Forest. So if you build off the exact same exact dataset, the exact same parameters, one in Python and one in Random Forest, you get some pretty interesting differences in the results.

Kirill: Interesting. Why do you think that is?

Eu Jin: I think it's just the way it's being coded. It all comes down to the author and the way they've actually coded it. I suppose random seed could be one factor of course, but I think it's just the way different authors wrote the models, essentially. I could be wrong, but what I can tell you is there are definitely differences in those two. I will say that Python is a little bit better generally. It varies, but if you have to go with one, Python is generally the safest option to go with.

Kirill: Okay. If you don't mind, I'm going to pick a few more out of your Kaggle competitions. Is that cool with you? They're just so interesting.

Eu Jin: Yeah, sure. Go for it!

Kirill: Okay. How about the one you took 1st place, 'Give Me Some Credit'? You won a prize of $3,000 and it was a credit scoring model. Can you tell us a bit about that? What was the challenge there and how did you go about it?

Eu Jin: That one was a relatively small competition, actually. Credit scoring is essentially very popular for the banking industry or finance industry, basically trying to predict people who are likely to default on their loan, whether it's personal loan or home loan or whatever. Basically they asked for the characteristics in terms of someone who is likely to default on their loan. It could be a factor of their salary; it could be a factor of how many other home loan products that they have, and so on and so forth.

We were given a dataset of different types of customers who have defaulted and what are their characteristics versus people who haven't defaulted and what are their characteristics. So we created a model. There wasn't a lot of features, maybe less than 30, essentially. It was a very simple dataset and it was a classification problem, which is basically whether someone defaults or not.

I suppose what really allowed us to win 1st place in that competition was because we were able to — based on 30 different features, we created an additional 60 based on the original 30, so feature extraction was a key component in winning that competition. When it comes down to it, the

models were very basic. You had a traditional Random Forest and gradient boosting was pretty much the only thing that we used there. We didn't do a lot of ensembling, but we did a lot of work in feature extraction. So that was our critical success factor, if you like.

Kirill: Awesome! And can you tell us a bit more about feature extraction? Given some features, how do you come up with new ones?

Eu Jin: It could be as simple as taking a ratio of one number against the other. This is going to challenge my memory. So, let's say, for example, you have an information around the total savings of that person as one feature and then another feature could be the amount of loan that they have for this particular product, a home loan, and another loan, say personal loan, right? If you just feed those three information in the model, if you compare to taking ratios, those three plus ratios of — if you take the personal loan and divide it by your total savings, that gives you another bit of information.

And yes, although those two bits of information already exist and you're really just creating additional information out of what's already available, that actually does help the model and a lot of people don't realize this. They think "Oh, since we already have features in there, creating additional ones is not going to change anything," but actually it does help the model.

Kirill: Totally. Or if you have the balance, sometimes for monetary units it's a good idea to take the logarithm of the monetary unit. It normalizes it out because on a log scale, there are

certain proportions that can be visible in terms of people's wealth rather than on a linear scale.

Eu Jin: Exactly. And actually, another good one, for example, is dates. A lot of people just discard dates as not being important, but actually dates contain information around which particular weekday it is and whether it's on a weekend, which month it is – Christmas could be very different from normal June. When you don't extract that information out, the model doesn't know, so you kind of have to help the model figure out by extracting which one is the weekday, whether it's a weekend, and which particular month. So, that's how we were able to extract a lot of additional value or information out of what was given, essentially.

Kirill: Fantastic. And this one, the big one – 2nd place, The Hewlett Foundation, automated essay scoring. Guys, I hope everyone is sitting down for this because Eu Jin and his team in this composition took second place and got a prize of US$30,000 for a Kaggle competition for just having fun, basically. Tell us more about that. That sounds like an awesome project.

Eu Jin: Yeah. That was actually a lot of fun. It was also a lot of work, but back then it was totally the second biggest competition after the hospital one that we spoke about earlier. That competition probably took most of my time doing it because it's in the area of text analytics or natural language processing, essentially. Basically what that competition was about is basically trying to predict an algorithm that is able to automatically grade high school essays. It was very, very

innovative and very cutting edge at that time. It was really exciting.

Before even jumping on the competition alone, we sort of ensembled a team where we knew people who had great performance in that space before. Yeah, we went in as a 5-man team from all across the world, including William Cukierski, who is now currently working for Kaggle as a Head of Data Science. He's Head of Competition, actually. It was a great experience. I got to learn a lot from all the other team members who had very different specialized skills. With a combined effort, we were able to get second place. I'm very, very proud of that achievement.

Kirill: That's awesome. Tell us about what tools did you use and what algorithms.

Eu Jin: All sorts. I suppose the two key ones were R and Python. At that time, deep learning hasn't taken up yet, so we were doing a lot of LSA – latent semantic analysis, as well as latent Dirichlet allocation, so LDA for short. Basically what it is, it's a topic modelling which basically tries to extract the semantic part of a sentence. In an essay, there might be a lot of words in there, but ultimately they form a theme. There could be a topic around sport, a topic around social wellbeing, or a topic around different commercial things. So being able to extract those hidden entities and then predict whether that person has actually scored well or not on that particular topic is basically what it boils down to in the competition.

Kirill: Thank you, that's great. I'm going to just pick out one more. So the most recent one that you participated in: 2nd place –

Kaggle competition, 'Acquire Valued Shoppers Challenge'. So this has something to do with brands and discounts, what kind of discounts shops give their clients. Can you tell us a bit more about that challenge?

Eu Jin: Yeah, that challenge, again, is very different. It's a market basket analysis. The problem entails trying to predict whether the customer is likely going to be purchasing this particular product based on what they have purchased in the past. In the data science industry, it's well known as market basket analysis. That competition revolves around that area, essentially.

Kirill: So it's like a recommender system, like on Amazon, right? They look at what you've bought and what are you going to buy next.

Eu Jin: Yeah. Long story short, that's basically essentially it. For that competition, a lot of work was going into data. You mentioned recommender systems. So, a large part of the competition involved looking at recommender networks, essentially. It was a lot of work. It was also a combination of different themes and again, we grouped up and we combined different models to get the best outcomes. Yeah, that competition was really, really exciting as well.

Kirill: Awesome. A lot of the times you said it was a lot of work. Can you give us a rough estimate for these two projects, those big ones that you had? How long did it take you to complete this Kaggle competition?

Eu Jin: The essay competition took maybe three months, and it was very steady, every day about 2 hours to invest time to actually work through it. That competition, we tried to be

very disciplined in how we approached it. The acquired valued shoppers was slightly different in the sense that I started out going solo first, so the first couple of early stages of the competition were very intense and I found a point where I actually achieved the maximum potential. I ran out of ideas, essentially. I dried out. There was a month to go and we sort of teamed up with another group to combine our efforts. That combination of efforts helped boost our rankings up to second place. So at the tail end, it was a lighter load in the sense that it was all about how do we best bland our models or our solutions together to get the best possible outcome? It is ensembling components at the end of it, essentially.

Kirill: Okay, gotcha. So, throughout the competition, you create the model and then you run it and you compare it to the target that the client wants? Or do you compare it to other people's results? How do you know if you do well or not?

Eu Jin: Generally you create your own validation set, which basically just means you were given a dataset and you take a representative sample and you don't train your model on it. So as you're building your model, you always validate it against that set which we call the validation set, and then you start making your predictions and then you see how well you do on the public leader board, which is also another sample of the test set.

So that's sort of how you gauge how well you're doing, to ensure that you don't overfit, essentially. And through time as well, you become more experienced around what actions you take, which will result in whether you will overfit or not.

It's little steps, basically. You need to make sure that you take a correct representative sample and that you build your model in this way and make sure you validate it and then you sort of see how close you are to the validation set and then you make your submissions. And be very confident that sometimes the leader board can be misrepresented, so it all comes down to experience and how well you know your data ultimately.

Kirill: All right. But do the clients normally say, "We want 95% accuracy"? Do they set your target, or is it just whoever has the best result wins?

Eu Jin: Yeah, whoever has the best result wins ultimately. The client will generally take the top three, or whichever positions have a prize, they will take the solutions from those ones. And it all comes down to whoever does well in the actual private leader board. The private leader board is essentially the true reflection of the actual rankings.

Kirill: Okay, gotcha. And the leader board is updated as the competition goes along? So it's not just something you see at the very end?

Eu Jin: There's a public leader board that gets updated as you make submissions, but at the end of the competition there's a private leader board, and that's when the final ranking will be judged.

Kirill: Gotcha. And tell us a bit more about working in a team. A lot of these projects, you worked with at least one other person or even in a team of four or five people. What are your tips for people finding teammates to work together?

Eu Jin: Working in a team is always good because you can bounce ideas and share your approach and divide up the workload. But the thing about working in a team at the very early stages is that you don't get the variety of the approach because you know what the other person is doing. That means that you are biased in terms of what approach you can take. But when you don't work in a team, you are forced to find your own solution first and then later on, once you join up with another person you realize, "Oh, that person actually has a very different approach." And that blending, that combination, actually creates a lot of boost to your actual score, to your solutions.

My advice is usually try to go solo first and once you really hit a point where you can't go any further, then start to team up with people. And in terms of who to team up with, anyone that you feel has a slightly different approach to you. So do not pick someone that you know is probably going to do the same thing as you. If you are using R and you are using Random Forest and gradient boosting all the time, do not pick someone who will go with that approach. Go with someone who might be using Python and deep learning. That way you get a lot of boost in your result and ideas in the pool, essentially.

Kirill: Yeah, some more variety to your project.

Eu Jin: That's right, exactly.

Kirill: That's a good tip. Do you have any more tips? For instance, one question I would have is, people who are just starting out, should they aim for the bigger competitions to get that more intense experience and more intense working practice

on the datasets, or should they go for smaller competitions just to get like a gradual start into Kaggle? What would your idea be there, and do you have any other tips for people who after this podcast are going to run and look at Kaggle and see if it's for them?

Eu Jin: That largely does come down to personality as well. Some people prefer to start safe, and once they find their bearings and then go to the bigger competitions that sort of swim in the bigger pool, essentially. If these people are more comfortable with that approach, then I would suggest they do the starter competitions first. Digit recognizer is a bit complex, but do like the Titanic competitions and all the tutorial ones to start to get a grasp as to how the competition works. And then after that, jump into the bigger pool, essentially.

But for people who are really passionate and have that competitive edge to it and are willing to really work hard and actually race to the top, go into the deep end straight away. You will learn a lot if you do that. And obviously, some people like that because they aren't afraid and they are willing to just jump in the deep end. If they fail, they fail. They don't really care and they would do it again. I know some people who have actually done that and they just realize that they've jumped into the big pool too soon, and they realize "That's not for me" and it hampers their confidence. I have seen people experiencing that, so that's why I would say. Those two approaches are probably the most suitable ones, depending on which personality you are.

Kirill: Yeah, gotcha. I'm looking at your LinkedIn now, and just for everybody listening out there, winning in a Kaggle competition, or even participating in one—like, here you have one where you took 14th place—it looks really good on your LinkedIn. So, people who are considering, if you're looking for jobs, this is such a good place to get started. Jump into Kaggle just so that you can then show it off on your LinkedIn or on your resume or CV so people see that you are really interested in these topics.

For you, Eu Jin, I wanted to ask you — obviously you're pretty comfortable doing what you're doing at Deloitte. If somebody is not looking for a job, if somebody has a job that they love and so on, what benefit does doing Kaggle competitions bring to your actual day-to-day role?

Eu Jin: If you work in the space of data science, doing Kaggle competitions sharpens your skills and gives you access to a lot of projects that require those specialized skills. And I say this because if you are very early in your careers, obviously you come out of university and if you have been through a degree in computer science or data analytics, you are still working on a non-commercial – you are still working with a toy dataset, let's put it this way. In a commercial world, it's very different. There's a lot of other challenges that you have to think about from a commercial perspective, but a lot of it comes down to — if you hone your skills in Kaggle competitions, you get better recognized and you have actual proven skills in terms of dealing with real-life data and actually have been through that process of quick iteration. You are more likely to be recognized as well by the firm if

you work for large corporations, so you definitely have a lot of cut through there.

And it also says a lot about your aptitude and your attitude as well. The fact that you spend time to sharpen your skills outside of work just means to the person who doesn't really know you well in the firm that you have the right attitude and you are willing to go the extra mile. Doing Kaggle competitions is a proof that you have the technical ability to actually do it. That's what I've seen, at least in Deloitte anyway, that a lot of these large firms, that's how they see new, very fresh new people in the industry.

Kirill: Thank you. That's some great advice. And I really love that excurse into Kaggle, but now I want to go back a bit and talk about you. Can you tell us a bit, how did you start? I see that you studied a Masters of Econometrics at the Monash University and that's, as I can imagine, that's quite different to data science. How did you make that transition from econometrics to data science and why?

Eu Jin: Yeah, so from an educational background, I actually – when I was at university, like a lot of people, I wasn't quite sure what I wanted to do. Back then I actually thought I wanted to do marketing so I did my degree in marketing first and then did a Masters in Econometrics only because I knew I wanted to do marketing but I knew I was good at math and I was good at stats, so I specialized in that area. Later through my career, as I started working in the industry, I started in marketing first and then I started to realize that actually I didn't quite like marketing, but I really loved the data. I really loved what the data says about the customers, I really

like what the data can actually bring – so much information about customers.

Marketing in the traditional sense is – everything is "airy fairy," so that just started to drive towards more specialized data analytics. Basically I learned everything from scratch. I come from a background where I don't know programming. I did not program in R or Python back then. When I came to Deloitte, I did not know a single programming skill at all, not even SQL back then. Not even SQL.

Kirill: Wow! You should have done your preparation, man. SQL takes like two days to learn. (Laughs)

Eu Jin: It took me three months to learn it. I had to take the hard road, which is a lot of work, and really, really be disciplined and learn the language. I'm still not super great at it these days because I wanted to focus more on Python and R, the exciting stuff, which is your deep learning, your predictive modelling. I wanted to focus on that area essentially. But yeah, if you have the passion, I think nothing really stops you from doing what you need to do to get where you need to get to.

Kirill: Okay. That's very impressive, actually. So you started learning the programming languages – SQL, R, Python – and pretty much right away, you jumped into Kaggle competition. It's not like you had five years of experience before you did the first Kaggle.

Eu Jin: Exactly. I went straight into Kaggle as a learning platform. That's what I did.

Kirill: Wonderful. I hope people listening to this are going to take some notes from that because that's very impressive. Like you say, it's probably a great learning experience to learn on real world examples.

Eu Jin: That's right. And right now, I'm actually doing an actual competition on Kaggle. After taking about a year's break, I'm back into Kaggle again doing one of the new competitions for text mining. And the reason why I jumped into that is because I knew it would push me to learn deep learning or recurrent neural networks. I have no idea. I have only just recently come to know about that and I knew the only way for me to really accelerate my learning is if I actually did the competition. That would really drive me to learn and practice and actually get exposure in my free time. I knew that if I didn't do a Kaggle competition, I was going to be a bit lazy and it would take me a lot longer to get there. So I kind of use Kaggle as an accelerated platform for learning.

Kirill: Gotcha. It pushes you further. And interesting you mentioned deep learning because a couple of times you were describing the previous Kaggle competitions, you said that deep learning wasn't in the game, it hadn't picked up yet. And now that deep learning is here and you're actually doing this competition, what difference do you feel? Do you feel that it's going to give you more power to solve these challenges?

Eu Jin: Yes. That's my point of view, which is that deep learning is pretty game changing in the industry right now. It is still at an immature stage, and when I say immature, I mean a lot of people are still very unfamiliar to what it does and what it

can do and its potential. Even understanding how it works, not a lot of people know that even, because it's so new. But it's actually so game changing, especially in the health care right now. A lot of people are saying this, and I'm going to repeat the same thing that everyone says, which is it will revolutionize the industry. It will revolutionize a lot of different industries, not just data science.

Prior to this, I didn't actually know much about deep learning and I actually forced myself to learn it the hard way, which is to learn it from the ground up, understand how it works, what it can do and make sure that I'm in the forefront of the industry. Basically it's all about making sure you get ahead of the cycle and be up-to-date to the changes in technology and on the bleeding edge of technology, essentially. That's my point of view.

Kirill: Yeah. I totally agree with that. I can resonate with that sentiment. And remembering something that Andrew Ng, the head of research or artificial intelligence at Baidu and also the founder of Coursera keeps saying these days, he keeps saying that AI is the new electricity, basically. You know, how electricity changed everything for every industry, same thing with artificial intelligence. It just goes into any industry and completely, like you say, revolutionizes it. So there are definitely big changes coming up ahead.

Eu Jin: Definitely.

Kirill: Yeah. I wanted to ask you—you seem to be really passionate and driven about data science and this whole space of machine learning and competitions. What or who has been your biggest influence? What keeps pushing you forward?

What keeps you being hungry for more knowledge and just not ever stopping?

Eu Jin: I think passion is definitely one. I just like data out of sheer curiosity, and being able to extract a lot of information and insights from data. It's definitely something that I really enjoy. I especially love to tell a story to someone and say, "This is what the data has told us and this is what it can do." So being able to wow the crowd to say "Wow! That is such a cool…" Being able to do that is game changing. So I really, really like being in that sort of environment. It's hard to find the right words to describe it but it's passion, basically.

And over time, I've also looked up to different people who were really key in terms of my development throughout the journey. Andrew Ng is one of them. He's not just smart and he's not just a genius – he thinks in a bigger perspective as well. He thinks about how AI can change the world. He thinks big and he is thinking about transforming the industry. And I think you need that passion, not just "I just love the technical stuff and that's about it." But you also need a vision. That vision is the one that actually really, really drives you and makes you believe in what you do is the right thing. And if you like what you're doing right now, it's even better. You know, if you're doing the right thing for the greater good that you believe in and you actually enjoy doing it, that's the ultimate combination, if you like.

Kirill: Gotcha. So what's your vision? I know you said you want to contribute to changing the world of artificial intelligence and

stuff, but do you have a specific thing that you want to change in the world?

Eu Jin: Well, I don't know if I have the ability to change the world. (Laughs)

Kirill: Of course you do!

Eu Jin: I'd say my own goal in terms of where I see myself in the next 5-10 years is being able to change the mindset of people to start to see the power of artificial intelligence and what data can do, either within my own professional networks or within the firm. If I can achieve that within the firm, I would retire if I was able to do that. Basically, this is my ambition, which is that in the commercial industry right now, here in Australia, a lot of people are still unaware of what data can truly do and I want to show to everyone that with the current technological changes that we have now and the abundance of data, there's so much more we can do not just for making lots of profit, but also for the community's good. That's where I will hopefully be able to be, being part of that journey.

Kirill: Gotcha. Just on that, I have this feeling that you will never retire because this space is changing so much. It's so exciting, right? You're just going to continuously see more and more cool stuff coming up. I think you'll be changing the world up to your late 60s or 70s. Yeah, prepare yourself for that.

Eu Jin: I'll just do lots of Kaggle competitions. Maybe I'll do that. (Laughs)

Kirill: You know how right now in your Kaggle you have the 'competitions grandmaster'? You'll be like 'competitions retired veteran' or something. (Laughs)

Eu Jin: That's right. Get a Medal of Honour to accommodate that too!

Kirill: For sure. One important question I wanted to ask you, and I think I've been asking this quite a bit on the podcast because it's just been a trending concern in the world that as artificial intelligence, as deep learning progresses, and it's really trending now and you see things like Google DeepMind and you see computers beating people not only in chess but in Go, like AlphaGo won the Go championship recently, you see self-driving cars. You see all these things. Soon computers are going to be winning your Kaggle competitions before you know it.

So the question is, are you in any form concerned that artificial intelligence is not going to only do good, but it's also going to do evil for this world? Evil in the sense of evil for humans, that it might see us as a threat, it might see us as something that's not necessary on this planet?

Eu Jin: That's a very good question. I never really thought of the evil side of the AI. I suppose, artificial intelligence is still manmade. And anything that is manmade, people could also use for evil intentions, I suppose, for a lack of a better word. We have to be very aware of the powers that we have because at the end of the day, it all comes down to the person. We have to be aware that what we are doing here with the powers of AI, you know, we have to use it for the

good. It can be used for bad things as well, and we have to make sure that we try our best to not allow that.

I don't even know how we can prevent that, but I think it comes down to education and making people aware that we have a responsibility to ensure that we are doing it for the good. Things like self-driving cars, we are doing this because it helps people who are disabled or who have less capability of travelling from one place to another for whatever reason. That now allows them to go places that they have never been before. That should be our driving goal. That should be what we aim to for the betterment of society and we just have to keep nailing that message to everyone to make sure that we do the right thing.

Kirill: Gotcha. Okay, understood. I think that was pretty much all the questions I had. It's been great. Thanks a lot for coming on the show and sharing all this knowledge. For the people who've listened to this and got inspired by you, how can they follow you, get to know more about your career, and maybe even get in touch?

Eu Jin: You can definitely add me on LinkedIn. I'm more than happy to answer questions or give advice to people who would like to follow the same journey that I've been through, especially for people who have gone through a very similar journey and want to know how I've actually gotten through it. I'm more than happy to provide advice. Just hook me up on LinkedIn.

Kirill: Gotcha. Cool. We will add your LinkedIn profile on the show notes. And one final question today: What is your one book that you can recommend to our listeners to help them become better at data and data science?

Eu Jin: There is definitely a lot of books out there and it's depending on the amount of time that you can spare. I generally don't read a lot of books. I do a lot more actual practice. But if there's one book I would recommend, it's "Dear Data." "Dear Data" has a very interesting story behind it. It's basically about two authors. One lives in the U.K. and one lives in the U.S., if I'm not wrong, and basically they send each other postcards. Each week they make it so that they use data and also data visualization as the means of communicating what they have done during the week.

Every week it's a very different theme or topic to it and it's just really creative and it's just great to see how two people can connect just by postcard, how they can build that relationship, that special relationship just by using data. The creativity behind it is just amazing. It's probably not in the area of building algorithms or advanced machine learning techniques, but at the end of the day it's still data and it's just a beautiful story in there. It's been really great to read. Actually, I'm still reading it. I'm only in week three and I'm really enjoying the story so far. I read it when I go to bed and it just puts me in a very good mood to have a good night's sleep.

Kirill: Fantastic. Thanks a lot. And it's totally cool that it's not about algorithms. I think it's important sometimes to mix things up with a different type of book and I think you do enough of the algorithms anyway through Kaggle competitions. Yeah, thanks for the advice. It sounds like a really interesting read – "Dear Data". Guys, check it out. Again, thanks so much, Eu Jin, for coming on the show and sharing this wealth of knowledge and telling us all about

your experience with Kaggle. I'm sure that so many people are going to get inspired by this.

Eu Jin: Thank you. My pleasure.

Kirill: So there you have it. I hope you enjoyed this podcast super saturated with knowledge. Even I told Eu Jin after the podcast — I couldn't believe how quickly time flew by. This hour just flew by. I had so many other questions that I wanted to ask about his career, about his work at Deloitte and all these other things, but at the same time this podcast is already full of value and I really hope that this has put into perspective what Kaggle is and what these competitions are all about.

Hopefully now that you've heard it all, you are going to go and at least check out the Kaggle website, at least have a go at their practice datasets to see if this is something that you'd be interested in pursuing further. And as you can see from this episode, it's a huge benefit to your career whether you are looking for a job or whether you already have a career which you're happy with. It's a wonderful addition to help you practice, to help you develop new skills and to really be on the, as Eu Jin put it, "bleeding edge of technology" to know the best methods and techniques that are out there.

For me personally, the biggest takeaway of course was all this Kaggle knowledge, all this experience that Eu Jin was able to share with us today. It was truly phenomenal. And as always, you can get the show notes for this episode at www.superdatascience.com/45 where you'll be able to connect with Eu Jin on LinkedIn. We will put up all the

resources mentioned in this episode as well as the transcript for today's show.

And if you did enjoy this episode, then help us spread the word about the show by leaving us a review on iTunes. That would really help us out. And on that note, I'm really thankful for you spending this hour with us today. I really hope that you felt as if you were a part of our conversation and I can't wait to see you next time. Until then, happy analysing.