SUPER DATASCIENCE
MAKING THE COMPLEX SIMPLE

SDS PODCAST EPISODE 57 WITH RYAN COMPTON

Kirill: This is episode number 57 with Head of Applied Machine Learning Ryan Compton.

(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)

Hey guys, welcome back to the SuperDataScience podcast. Super excited about today's episode. We've got Ryan Compton from Clarifai on the show, and Ryan is the Head of Applied Machine Learning. So something you know need to know about Clarifai is that it's a company that helps people and organisations to use deep neural networks to understand images better, to understand what's going on in the image, to tag images, to classify images. So definitely check out their website, www.clarifai.com if you're in front of a computer right now because it just gives you some really cool examples, like right on the front page, I'm there right now, you can just click all these different images and it shows you how it's classifying them.

And that is exactly what Ryan works with. So he's the person behind a lot of their work, and a lot of what their algorithms are doing, and in this podcast, that's what we're going to be talking about. So we're going to learn what Ryan's role is all about, how convolutional neural networks are used to classify images, and why. So this podcast is

more about the applied side of things, why artificial intelligence is used in the world, what are the real use case scenarios, and how artificial intelligence is helping organisations deliver better services and products.

And of course, in this podcast, we're also going to talk about careers and Ryan will give you some of his tips on how you can break into the space of machine learning and artificial intelligence. So I'm very excited about the chat ahead and without further ado, I bring to you Ryan Compton of Clarifai. (background music plays)

Welcome everybody to the SuperDataScience podcast. Today I've got a very exciting guest on the show, Ryan Compton, who is the Head of Applied Machine Learning at Clarifai. Ryan, welcome to the show.

Ryan: Thank you.

Kirill: So tell us a bit. Where are you calling in from?

Ryan: I'm calling in from Clarifai Incorporated. Our headquarters is in Chelsea, in Manhattan.

Kirill: Ok, awesome. And it was really interesting how we met, right? We met at the ODSC conference and you were doing a talk.

Ryan: That's right. I was giving a talk of what computers look at when they look at nudity. It's a great talk about the NSFW filter that I helped to build.

Kirill: Yeah, definitely it was. And yeah, it was interesting because I had to select the sessions I was going to go to on the app before, and I was like, "woah, somebody's talking about how

computers look at nudity? That's like how -- I don't know, it could have gone two ways. It could have been completely boring and just very straightforward, but it went the opposite way. You were talking about how convolutional neural networks, like the deep learning algorithms, look at nudity. And that was pretty cool. What was your impression? How did the audience take your talk?

Ryan: You know, I get a lot of positive feedback from that talk. I think the point I always wanted to get across, independent of nudity, I think is a really interesting and important point. And what that is is that if you look at the way people used to build computer vision algorithms back 20 years ago, everybody would have to design features by hand. So if you had to make a nudity filter in 1996, you would do something where you look for skin, you look for stick figures, you look for body parts and human anatomy shapes that you understand, that you know. When you design them by hand, it works ok, but it's also tiresome, and it's also limited by how good of an artist you are, or how creative you are, or how much you know about human anatomy.

On the other hand, when you tell a neural network to do the same problem, all you need is a ton of data. A ton of data that's not safe for work, and a ton of data that's safe for work. You give it to the algorithm, and then it figures out on its own what kind of things do you want to moderate. What kind of things does the computer think actually make up nudity. And we can visualise those features with a process called a de-convolutional network. And in the talk, what I do is I show some visualisations of what the machine learned when it learned how to filter out nude photos, and you can

find that it learns things which look a lot like various parts of human anatomy in a lot of detail, much more than any scientist or researcher would ever draw on their own by hand.

Kirill: Yeah, totally. And I remember the navel kept coming up in your talk, that when a computer sees the navel, it's like a big chance that that is nudity, right?

Ryan: Yeah, that's actually one of the things that's really interesting about the work of data science when you're doing artificial intelligence. It's that you play a game that you want to build a machine learning classifier. But the way that you build it is by engineering the data sets rather than engineering the classifier very much. Or the features. And what happened when I built the first nudity model, and the one that I used in the talk, is that I took a whole lot of nude photos, and I put that into one category. And then I took a whole lot of safe photos and put that into the other category.

Kirill: When you say you took, you didn't actually take them. You got them from somewhere?

Ryan: Yes. I trained AI on photos I didn't actually take. Loads of photos, and I was able to acquire them some how. Right. So the classifier, every nude photo, would have a belly button exposed. And I didn't intend for that to happen, but when I showed the classifier all of these things, and I told them "this is a nude photo", unintentionally it learned that belly buttons are important. And I was able to use a deconvolutional network to discover that. It was the same with red lips. All of the photos of women, it would think if the lips are very red, then that means it's probably nudity.

And these are side effects of what it was going for, but it would have never known unless it had a deconvolutional network.

And then what I can do is I can iteratively redesign my data sets to have photos of red lips where people are wearing clothes. Or photos of belly buttons which are kind of safe for work. And then I can sort of balance out those side effects to kind of really help it zone in on what I'm trying to identify.

Kirill: Gotcha, gotcha. So you kind of like don't allow it to cheat and really identify true nudity, not just belly buttons and red lips.

Ryan: Right. Right.

Kirill: And that's really fascinating, and I'd love to dive a bit more into deconvolutional neural networks. But before we do that, can you give us a bit of a description of what is the application of your work? What kind of organisations use your company's services and what is the benefit of being able to – even in this example, to train up neural networks to discover nudity and to discover not-nudity. So of course it's not just a hobby, there's a huge industry behind this.

Ryan: Right, absolutely. So our company doesn't just do nudity, we have a lot of different customers from a lot of different verticals. I would say that we have two types of customers that really play a big role in the use of our services. One type of customer is interested in image retrieval, so the product that we build at Clarifai is a neural network as a service. You give us an image, we run it through a neural network and we give you back tags which understand what concepts are present in the image. And stock photography companies

like 500px, they'll use us to help improve search on their site by enriching their image data with tags that are generated by a neural network and then they can improve search, they can do better image retrieval problems.

On the other hand, we have people who want to solve image removal. They care a lot more about taking images that have a particular concept and throwing those away. This is the kind of thing you would see in the NSFW filter. We also work in moderation. Sometimes people have photos of guns or drugs or various types of violence. We have ways to remove that as well, and what we do is we help those customers moderate their sites and that plays a big role in helping us build our products, is understanding if people want to do removal or if they want to do retrieval. And it helps us pick a point as to whether or not we care about false positives or if we care about false negatives. It's a whole different type of framework, but that's the type of uses that we have.

Kirill: Gotcha. That's really fascinating. And I really like the example you showed in the talk where you have a baby camera and — can you talk a little bit more about that, because I don't think I can describe it as well as you did, about the baby camera situation?

Ryan: Oh, right. So in addition to the pre-packaged models that we build, we also give our users the ability to customize their model, to identify images, to meet concepts that they define on their own. So one of the things that I did once we built this custom training project is, independent of our company, I had a Nest Cam. My Nest Cam worked okay, but it had a motion sensor on it and the motion sensor would fire every

time something would move. So my son, he was sleeping in the crib, and every time somebody would open the door or walk by the door, I would get this e-mail that says, "Motion detected in the baby's room," and it was totally uninteresting because I just don't care that somebody walked by the baby's room wearing a towel.

What I wanted to do was get a better way to kind of understand, is the baby asleep, is the baby sitting up, is the baby standing, and kind of have a nice baby monitor to give me an alert if he wakes up. He's going to cry anyway, but I thought it would be a fun project. So what I did was I took a Raspberry Pi, I pointed a camera at the crib, and then I trained a classifier to kind of understand the baby is not in the crib, the baby is asleep or the baby is standing up. And every half second or so it takes a photo, it classifies it as standing or sleeping or missing, and then I can get an alert if I want to. I usually don't need that, but I can get an alert and I'll know that he stood up and he woke up. Or that he went to sleep. That's one of the things that we did that I can do with our custom training solution.

Kirill: And is that training solution available to anybody or is it paid subscription?

Ryan: You have to pay — I think our public pricing has about a third of a penny per image to classify. Yeah, the whole thing is an API and you can use a UI to train models, and it's open. We deployed it and we launched it and we have a lot of people using it.

Kirill: Awesome. Where can our listeners find it if they'd like to explore it more?

Ryan: I think the best way to find it is our website. It's clarifai.com and then you click around and you'll find everything that we've built.

Kirill: Gotcha. Awesome. That's really cool and I think that's a very interesting use case with the baby and definitely all the things that you're doing in the company are very exciting. Tell us a bit more about your background. How did you come to be the Head of Applied Machine Learning in a company like this and doing such exciting work?

Ryan: Yeah. Before I worked here, I was doing a lot of work in data mining. I had a postdoc at Howard Hughes Laboratory. We did lots of big data and optimization and various types of work. And even before that, I did a PhD in image processing at UCLA. I was in the math department and all of the work I was doing as a graduate student was optimization to solve problems which are often related to image processing.

The skillsets that I was developing, they carried over a lot into my future work as a postdoc or into my future work in industry now, because I'm really still solving math problems with linear algebra and optimization. But one of the things that I really sort of took away from my graduate education that I find myself reusing all the time whether it be as an applied machine learning researcher in industry or in academia, is this kind of situation where you're working on a problem, the problem isn't defined very well, nothing that you do works, nothing that you do is likely going to work, and you just kind of have a mind-set where you say, "That's okay."

For graduate school I had that for many years, things not working and things being hard. It used to be hard for me to be in that type of environment and now I'm just used it. You know, I was a data scientist for a while here, now I'm leading our applied machine learning group. All of the work I'm doing has a fairly large degree of uncertainty as to "Is this even possible?" And going through graduate school, going through a postdoc, going through situations where you never know that what you're doing is possible really helps prepare you for that type of work. I'd say my background was graduate school, math PhD, but the thing that is the most important skill I learned is, "Nothing is going to work and that's okay."

Kirill: (Laughs) Gotcha. But what inspired you to kind of move from math into machine learning and data science? It's a bit of a different field. Like, you could have ended up being a math lecturer somewhere or a statistician. Why data science?

Ryan: Well, it is definitely a different field. I think one of the things that really kind of inspired me here is deep learning wasn't a huge thing when I was a student. But when I first walked into Clarifai's office, Matt showed me a demo of custom training and he showed us a bunch of images. He clicked on five or six images. They were flags. Some of them were flags of the United States, some of them were flags of China, some of them were flags of Mexico. There's a big array of images with different flags. He clicked on like four or five flags and said, "Okay, I'm going to train a classifier." He trains a classifier and then all of a sudden it understands American flags versus Chinese flags versus Mexican flags. And he did

that with almost no work at all, with no understanding of features, just "Here's some data," and the machine understands it. The amount of time that it went from an idea for what a computer vision algorithm should do to actually having it built and deployed and ready for use, I just couldn't believe it. I thought it was impossible. I thought there was no way anybody could make something like this work, and then I saw it work. I was totally amazed and very inspired when I saw that demonstration and I was just thinking, "Okay, this is the stuff I have to do. This is the stuff where if I don't do this, this is going to replace me with a machine." So I figured that's what I have to do, that's what I have to learn and so far it's been very exciting.

Kirill: That's awesome. Yeah, that's pretty cool. I can see how that could go down. Like, when you see something that so captivates you, you can't stop but understand how it works and also be part of it. I totally appreciate that feeling. All right, so we've talked a bit about why you got into data science. Let's go back to when we're talking about convolutional and deconvolutional neural networks. Can you start off by telling us a bit more about the tools that you use on a daily basis? I'm assuming it's maybe Pi Torch or Tensor Flow and things like that. Can you share a bit of insights there?

Ryan: Yeah. Actually, we definitely use a ton of Python. In terms of network architectures, the company I work at, we predate Torch and TensorFlow and Pi Torch by some time. TensorFlow was released when I was an employee. So there was actually a framework that was built up internally called Strite that the CEO Matt and his colleague David had

worked on when they were graduate students. It does neural network operations, it does convolutions, it does pooling, it does all of these things, but it's just something that was built in-house.

We use that extensively and it's very important to us. And there's other times when we look at stuff like Tensor Flow. We'll look at the other ones, but for me personally, my role in team is not a research scientist. I don't design architectures, I really design datasets to work with architectures that our research team is great at designing.

So for me, the tools I use the most aren't neural network frameworks. I actually try to make datasets, and the way I do that is often with Hadoop. So I'm much more well-acquainted with HDFS or MapReduce or those types of tools because we have hundreds and hundreds of terabytes of images and I have to sift through those and sort through those and make sure I can build a dataset that works really well on the images that I care about.

Kirill: Okay. That's really interesting. Let's talk a bit more about that. Why is building datasets so important? You know, if you ask somebody, they're thinking in your work, the most important thing would be the neural network side of things and things like that. But why are images and building the datasets such an important part?

Ryan: Right. So there's a lot of work that can be done with publicly available standard datasets. And if I just wanted to build a neural network on top of ImageNet or on top of Caltech 101, MS-Coco, these things exist and they're great. But because you don't have a ton of data that's unique when you look at

publicly available data, you're in a bit of trouble because a lot of people can compete directly with you, just head on, because you use the same data to accomplish similar tasks. That puts us at a bit of a disadvantage. So we focus a lot on making sure we have good datasets to solve the problems we need to solve.

Additionally, if you think of a problem like NSFW, there is no public academic dataset for nudity filtering. It's just not something that people who work in academia are distributing and thinking about. So we know that there's an industry need for that, but there's no datasets that are off-the-shelf readymade that you could just use that would solve the problem for you. So I often do have to build those datasets and I have to build them to the best of my ability and I have to understand if I have any biases when I'm thinking about what should go into each category. I have to understand if I'm getting data that's actually clean. There's a lot of stuff I have to do. I have to use humans-in-the-loop. We do a lot of things with active learning to make sure that we have human computation collecting the right data for us. I would say those are the main reasons I think about it, either because it gives us an advantage or because it gives us something that would not be possible from public datasets.

Kirill: Gotcha. And that's where your deconvolutional neural network comes in where you can understand if the way you picked the images did have a bias which you unintentionally introduced.

Ryan: That's right. That's actually something that I have used to discover my biases. And in the talk I talk about how I used it to find navels are not safe for work.

Kirill: Gotcha. So we mentioned a couple of terms. I would like to clarify these for the benefit of our listeners. Let's start with ImageNet and Caltech 101. What are those and could you maybe describe those in a bit of detail?

Ryan: Yeah. ImageNet is a competition which has a lot of attention recently in the media. There's a dataset that's pretty standard and what happens is every year all of the industries and all the graduate students who want to compete, they look at this database called ImageNet, it's organized according to the WordNet hierarchy – only the nouns – and you get hundreds and thousands of images for each item in WordNet. Everybody can just go download this dataset and when they're done, they can compete and see who has the best algorithm to recognize the WordNet hierarchy on the ImageNet database.

Caltech 101 is another database that has pictures of objects belonging to 101 categories, about 40 to 800 images per category. Anybody who wants it can just go download it. It's been around since 2003 and it's been used for a huge number of research papers.

Kirill: And the value of those is immense, right? It's not just datasets that appeared out of nowhere. People sat there for days and even months in a row just categorizing these images and putting them into an archive, right?

Ryan: Oh, yeah, absolutely. These things are extremely hard to build, so my role in my company, actually—I'm not really the

person who would be competing in the ImageNet competition. The role that my team or that data science plays in an AI corporation is more setting up a competition that the people who design the architectures would be competing on. A lot of work goes into building these datasets. Caltech 101 was built by Fei-Fei Li. They're not easy things to build and they're extremely valuable when you build them well.

Kirill: Yeah, gotcha. Okay, that's very interesting. That kind of transitions a bit into what we discussed before the podcast, the role of a data scientist in an organization that chooses artificial intelligence. Now I can see how that's a topic that's so close to you and how it has its own space of being a data scientist in an organization specifically focused on artificial intelligence. Let's talk a bit more about that. What are your thoughts and what are your experiences of being that data scientist in an AI organization?

Ryan: Right. That was actually something that I thought about a lot when I first started at Clarifai, is I joined the company, I had never been a data scientist before, and the boss said they don't know what a data scientist is supposed to do. And I was not entirely sure. So I had to kind of take a guess at it, but the guess that I eventually came up with is that I think the role of data science can be helpful to artificial intelligence if you think of designing datasets in order to improve machine learning models.

You know, 7 or 8 months later I found out that lots of data scientists are doing things like customer churn prediction. There's lots of other stuff with dashboards that maybe they

were expecting, but people have been very happy with the definition I came up with because artificial intelligence, so much of the feature learning is done as a function of the data that you have available that if you don't have somebody who's full-time understanding how to build and predict with a model or how to evaluate a model or who has some good intuition for what type of machine learning classifiers are even feasible to attempt, I think you're at a loss. These kind of people are very important and I think the role of the data scientist in a corporation should be to generate the data that feeds in to the artificial intelligence to make it as effective as possible.

Kirill: That's really cool. That's kind of like large scale of what we very often hear in data science. Like, you have a data science project and if that's 100% of your time and effort, then 80% of your time and effort goes into the data prep, data cleaning, data gathering, data preparation part. And only 20% is the analysis. What you're describing here is exactly that but in a huge scale. Your whole role is about data preparation, is about making sure that the algorithms that are then trained on the data, that they are being trained on good data that will give valid insights rather than just garbage in/garbage out type of scenario.

Ryan: Right, absolutely.

Kirill: So how big is your team? How many people are involved in this operation of preparing the data for the training of algorithms?

Ryan: Sure. My team — we have a number of people at various levels of expertise who are making sure that our datasets are

ready and clean and good for production use. Our engineers — we have about five who are solid engineers who understand machine learning and understand how to work with large datasets. And then additionally we also have a number of people that are sort of more entry level, and aren't programmers, that are just curating our data and telling us, like, "In this dataset you gave us, you didn't know it but there is a ton of bellybuttons."

So sometimes DeconvNets is a lot of work just to find something that I was talking about. Sometimes you just have to look at lots of pictures to see it. And then of course, I don't know if you want to count offshore or outsourced people who are annotating images for us for like half a cent per click. We use that a lot too. So the number of humans employed to do this is fairly large if you count them, but otherwise I would say there's about 10 people who are full-time cleaning and curating.

Kirill: That's a huge team for data cleaning and it kind of reinforces the notion how important this part is. Have you ever had situations, not through the fault of your own, but just generally, maybe stories you've heard where algorithms have been trained up on data which wasn't well prepared and then that led to wrong insights?

Ryan: Yes. Definitely there's a lot of mistakes. I can say that with the NSFW model that I was discussing, there's that problem with navels. It gets false positives on things that you don't expect. Some of the kinds of things that really get wrong because the dataset is missing is something like an artistic nude. So, Robert Mapplethorpe photographs, these are often

very explicit but because they're taken by an artistic photographer, something about them just doesn't click with the net and you can miss those. So I didn't take into account as much of that as I probably could have and as a result, the model was wrong. Definitely my own biases as to where I got data play a large part into how much the algorithm is going to generalize.

Kirill: Gotcha. That kind of makes me wonder—as humans, we rarely make mistakes. It's still possible to mistake something for another thing and we've all seen those pictures where there's like a duck but it looks like a rabbit, or rabbit which looks like a duck, that type of thing. But at the same time, it sounds like these convolutional neural networks, or these algorithms in general, make way more mistakes, it's much harder for them to actually understand what they're looking for. When do you think the time will be when machines will be as good as us in seeing, for instance, nudity and understanding what it is, and not making that mistake which you just described?

Ryan: To be honest, I think we're already there. The part that we haven't yet reached is this extra layer of content or context, because sometimes nudity isn't necessarily inappropriate or "not safe for work". The idea of "not safe for work" is very context-dependent on your work. Whereas the concept of nudity, that's just some pixels. And the machines are really good at that but sometimes extra context about "Is this appropriate? Is this artistic? Is this something that I think is okay to show on the front page of The Guardian?"

It always depends a little bit more on what humans think about because humans understand a lot more about culture and a lot more about context than the machines do. So moderating photos which might be upsetting for reasons other than nudity is very difficult. And this is one of the things which I think we still have a long way to go in computer vision. It might even be outside the scope of what computer vision can do without extra information about text or some other kind of data that explains to you the context of the photograph rather than just the pixels.

Kirill: Yeah, gotcha. So it's kind of like that image that you showed in your talk where there's a statue of Buddha which looks like it's pinching an airplane which is flying way behind, just because of the way the photographer was positioned. To a human, that makes a lot of sense. It's so funny or it's pretty cool because it's a plane and it's a statue of Buddha, they're completely miles apart but the way you put the camera you got it as if Buddha is pinching the airplane, whereas there's no way a machine will understand that without that extra layer of intelligence behind it.

Ryan: Absolutely. I think this is a major problem also for things like content moderation. If you think of a picture like the death of Aylan Kurdi – that was the Syrian refugee that drowned – a photo like that is very upsetting, but you won't understand it without additional context. The Buddha pinching the airplane is very interesting, too. A human who sees that, there's a lot of things that trigger in a human's brain, like, "Oh, this is a fun photo. This looks like it shouldn't be happening, but it is." But the algorithm only

sees pixels. It doesn't get that extra stuff that makes it interesting.

Kirill: Gotcha. Okay, thanks for the rundown. By the way, people who are listening to this — guys, Ryan has a blog. Ryan, you don't mind if I share it?

Ryan: Go nuts!

Kirill: Yeah, it's ryancompton.net (surprise surprise ".net" like convolutional net!) and there's lots of pictures there. Like, know it's probably hard to imagine all these things that we're talking about on the show, but if you want to check out some pictures and there's even a talk, like the actual video of Ryan's talk there, you can definitely check it out there. Now I'd like to move on to a bit of more general topics about data science. What's your biggest challenge that you've ever had as a data scientist?

Ryan: The biggest challenge I think I've had — I would say at my current role, sort of defining what I should do in an AI company, that was the vaguest and most daunting thing that happened. When I first saw how successful neural networks were, I had a lot of doubt that I would be able to contribute anything to the people who can actually build and deploy this type of technology. It just seemed to be way ahead of what I thought was possible and I didn't know what I could do. But once I got a better understanding of the technology, I thought, "Okay, there is a real need here for having a good understanding of datasets and how to design them effectively." I would say figuring that out was probably the biggest challenge.

Kirill: Gotcha. That's very powerful. And what would your recommendation or advice be to people who are in a similar situation? Because not everybody is going to come up to the same conclusions that their role is to support a function by preparing data. That's very relevant to what you're doing, but other people might have other scenarios where they come into an organization and, as with what happened with you, their boss tells them, "Hey, look, you're a data scientist, but I have no idea what you should be doing. Go figure it out." What should they look into? How should they think about this situation?"

Ryan: Right. The other kinds of things that I would say really matter and I think really become valuable skills is just having a good intuition for when machine learning is going to work and when you can extract value from the data that is available to you in whatever particular situation. Oftentimes it doesn't work and building an intuition just by trying lots of models, trying datasets you never even thought would make sense, you can often find things that do work and once you can extract something where machine learning provides a delta, that delta is really big. So anything you can do to experiment, to get your hands on more data, to get your hands on different modalities of data, is really going to help you build your intuition and build your abilities as a machine learning researcher or a data scientist.

Kirill: Gotcha. It sounds like you're almost equating data scientist to machine learning researcher. Do you think there's other parts of data science that are set aside from machine learning or, in your view, is it that machine learning and data science are pretty much the same thing?

Ryan: I would say that actually data scientist is typically an industry term. whereas "machine learning researcher" you often hear just in other circles, like academics. One of the things that comes along with being in industry, and this is important for academia too, but something that I think happens a lot more often in industry is that you need to be able to effectively communicate your results to an audience which is very much from a different background, very unfamiliar, doesn't understand anything that went into what you're building and doesn't understand how to quantify the value of what you're doing.

So being able to effectively communicate as efficiently as possible what you've done for a customer or within the organization, being able to do that I think goes a long way as a data scientist. Because if you want to work on a big project that's going to take 6 months of machine learning research, you do often need to get a lot of buy-in from your company or get a customer to let you work on a prospective project that might not actually get out. So being able to communicate effectively is something that I think is very important. It goes with data scientist, but not necessarily with machine learning researcher as strongly.

Kirill: Okay, thank you. All right, next question: What is a recent win that you can share with us, something that you've done in your role that you're proud of, some challenge that you've overcome?

Ryan: Actually, I would say one of the great things about my current role is, when I've built models, the models have been used a lot. So I would say that several million requests per

day will go through a model that I've created and I can actually meet the people who are using it. So there is companies that use the technology that we build and sometimes I go meet them and they say, "I was using your model to moderate my photos and I've seen 90% of the content that I usually would have to do by hand just completely disappear. And then the other 10% I can now do by hand very effectively." So I'd say a real recent win is just I went to a customer – I can't say who it is – and I sat there with the person who is actually using the results from the model and they liked it.

Kirill: That's always good. But at the same time, does it sound to you like the work you're doing is getting the world a bit closer to disrupting a lot of jobs, a lot of people's manual work and how do you feel about that?

Ryan: Yeah, my goal is actually — I want to automate my job. I'm working very hard towards automating my job.

Kirill: Why? Why would you do that? (Laughs)

Ryan: Well, I feel like the closer I get to automating my job, I can start to go do something else that might be more sophisticated. And when I'm doing that, it is a little bit challenging because I know if I'm not doing that, someone else is trying to. And it's definitely going to require a shift. I think a lot of the stuff that I've learned in the past, when I was doing an image processing PhD, there was a lot of things that I would work on and my friends would work on, very sophisticated research, and now people are using neural nets and they get results that are as good or better with just a few days of messing around, than you know,

some of the most impressive PhD theses from a few years back.

So the people who wrote those theses, they're not still working on those problems. They had to make a shift to do something new. And the work I'm doing now, I think it's going to make a lot of things obsolete. And it's going to make technical jobs obsolete. I don't work in self-driving cars, but I notice a big concern over there. But definitely the rate of change in terms of skillsets one must develop in order to stay current in the field, it's going up and it's not just low skilled workers, it's actually people in our line of work, too. They need to learn things that don't get replaced by a machine. And if it looks like what you're doing is potentially not replaceable by a machine, you should try to replace it by a machine and see how far you get because you might be able to do it.

Kirill: Gotcha. That's very interesting because it kind of ties in with what you said – or just the general sense I developed is that you got a PhD in mathematics and then you applied that knowledge to be successful in your career. What about people who don't have PhDs, who are not looking to get a PhD? Given what you just said about the increasing complexity of this space, and many other spaces, that you really have to be very knowledgeable to keep up with the automation process and be ahead of the curve, what is your recommendation to those people? How do people without PhDs and this in-depth knowledge of mathematics or machine learning, how do they break into this field and how do they develop a career for themselves here?

Ryan: Right. The thing that I took away from my PhD that was more important than any specific thing that I learned was just the mindset of being able to try lots of different stuff and not be scared to try ideas that look like they're not going to work. This stuff I did in my PhD, the concrete results I produced did not see a whole lot of use and I think they could be replaced with a neural network quite easily.

So the actual things that I learned, those were not as important as just kind of the ability to think critically and scrutinize assumptions and then try stuff out and then adapt fast to new research directions. So even without a PhD, I believe that skillset is obtainable, it's just something that I think either you will get naturally or you'll get through lots of trial and error and solving lots of problems.

Kirill: Gotcha. I'm so glad you mentioned that because a lot of the time I hear this notion that "I don't have a PhD or a Master's or even a Bachelor's." People get kind of put off by that and they think, "Maybe I should do more education. Maybe I should go back to university and study a little bit more about machine learning or data science or mathematics."

But I totally agree with you that it's all about the critical thinking. That's what you must be focusing on. And whether you developed it at university, or whether you developed it through trying to start a business, or whether you developed it through playing chess, or whatever way you develop that critical thinking, that's going to benefit you the most in your aspirations of becoming a data scientist and that's what you should focus on. If anything, find other ways to develop that critical thinking along your path as you go into the space of

data science, not necessarily through going back to university or getting a PhD. That's one way of doing it, but definitely not a compulsory way of doing it.

Okay, so that's a pretty cool answer. Thanks a lot for that. I've got an interesting question for you. This might be a bit linked to what you mentioned before, but nevertheless, what is your one most favourite thing about being a data scientist, that one thing that makes you wake up in the morning with a smile when you go to work and have fun?

Ryan: Yeah, absolutely. I think that the amount of creativity and the amount of autonomy that you get in the job is fantastic. It's almost like being a painter. You show up with a blank canvas and you want to paint it. You show up with a bunch of unstructured data and a blank file and you have to figure out what's valuable, what's interesting. "Nobody knows. Maybe I can figure it out." I really enjoy the amount of autonomy and the creativity involved in pulling out valuable data and machine learning methods from a complete mess. I really enjoy it.

Kirill: Fantastic. I love that answer. I don't think I've ever met anybody as excited about data preparation as you are. It's just probably because of the way you've approached it that you've turned it into something fun and exciting that you enjoy doing and it's creative even. I think more people should approach it like that. That it's not just a tedious task that you have to do every time before you do your analysis. You've got to be creative about it and it can actually turn into fun and that's when you will achieve the maximum results.

Ryan: Yeah, definitely.

Kirill: Okay. So, the big question, and a very interesting question to get especially somebody's opinion on this because you are in such an advanced space in artificial intelligence, you're a data scientist, not just in a firm that does services or creates products, but you're actually in a firm that does artificial intelligence. So the question is, where do you think the field of data science is going and what should our listeners prepare for to kind of anticipate what's coming in the future so that their current, their careers are blooming and they've got lots of job opportunities coming forward?

Ryan: Yeah, I actually think there's going to be sort of two places where you see a ton of growth. One of them is going to be automation. What we're going to find is that as a lot of things become tried and true, and we start to hone in on methods that work across many different datasets like XGBoost, you're going to start to see that a lot of the tasks that data scientists do are going to get automated by other data scientists. So this kind of place that we've been in for a long time where data science was like nothing, that all sort of grew very quickly. I think that also parts of it will go away very quickly because they will be able to be automated.

On the other hand, there's this huge push in artificial intelligence that not only is going to make it easier to automate some stuff, it's also going to open up the doors to what's possible. Because what I used to think was very important, or what I used to think I was really good at, was understanding, "Here's a bunch of data, here's machine learning algorithms that I know. I know the intersection of

when machine learning is going to work." But then as you see deep learning and AI get better and better and better every six months, the amount of things that are possible just keeps expanding. And then adapting what's automatable to what's not yet possible and what's barely possible, I think the future data scientist is always going to be taking the stuff that's just barely possible and then turn it into something that's automatable.

Something like transfer learning with neural networks, nobody even knew about it a few years ago and now it's something that I do all the time almost automatically. That's kind of where I think it's going. I think the new things that are possible from AI are going to be automated by data scientists.

Kirill: Gotcha. That's a really cool view on things, that people need to be open-minded to find these places where they can automate stuff. What would you say is the most important skill for people to have going into this new era?

Ryan: One of the most important skills, aside from being able to try new stuff without worrying if it's going to work, I would say one of the things that's very important is understanding the right way to use human computation. When you have a problem that you're trying to solve, it's really informative to get a good understanding of when you want to use humans-in-the-loop, when you want to go use something like Mechanical Turk to get better results.

And I think that I see this lacking in a lot of people that I meet who work with data. A lot of people kind of neglect the fact that when you have something that's not quite possible,

you can often get a lot of work from humans to check that and figure out how well things are working, is the problem just badly defined, or is the machine learning algorithm not working? So one of the skills that I think is really important to me, which isn't always emphasized as much as it should be, is humans-in-the-loop to compute things and to figure out how good you can actually do.

Kirill: Gotcha. So, human-in-the-loop basically means that instead of something being fully automated the whole cycle, you've got humans doing some part of the job. For instance, with your navel example, you've got humans classifying some of those images or looking for those navels to make sure that there's less bias, something that the machine wouldn't necessarily pick up and highlight as a problem, it would actually instead use that to its advantage to kind of cheat the problem.

Ryan: Right. That's a good example. I think anything about active learning is a really good skill to understand and to think about.

Kirill: And that other service you mentioned is Mechanical Turk. Could you elaborate a bit more on that?

Ryan: Yeah. Mechanical Turk is an Amazon product. I think it has one of the best product names I've heard. There is this story, a long time ago, there was this Turk who was a robot that was mechanical and it would play chess. This is like 100 years ago and people would go play chess with the Turk robot and the Turk robot would beat people at the game of chess. And they thought there was an intelligent robot made out of wood and pulleys 100 years ago, but the story was

that actually underneath this robot was a little man under the table who was moving all the chess pieces so everybody was tricked into thinking it was a machine, but it was actually a human underneath there that was playing the game.

So Mechanical Turk is a product where they've just crowdsourced a ton—like, you can pay for humans to do anything you want. You go into the website and you pay them half a cent per click on these images and then people will go and they'll click on your images and you can get 10,000 images labelled in an afternoon. So it's a very useful way to think of your problems when you're not sure how well something is going to work or maybe you're not sure that the problem is well-defined, to see how far you can get with humans-in-the-loop.

Kirill: That's fantastic. That's definitely something I'm going to put into my list of tools that I can use. It's very often that you come across tasks like that. Even somebody doing their homework assignment for university might have a task where they might need 1,000 people clicking on stuff or surveying 1,000 people on something. So that's a pretty cool tool. That kind of brings us to the end. Thank you so much for coming on the show. It was very exciting to hear from you and hear all the amazing work that you're doing. Where can our listeners find you or contact you or follow you if they'd like to learn a bit more about your career?

Ryan: Yeah, I think you can follow me on — I have a blog, ryancompton.net, it's on Feedly and whatnot. And you can

also find me on Twitter @rycpt. I'm reasonably active, but not too active.

Kirill: Gotcha. LinkedIn?

Ryan: I guess my name is Ryan Compton, but there's a lot of us. One of them is in New York and works at Clarifai. There's only one of that.

Kirill: Okay, gotcha. All right. Thank you so much. We'll definitely include those in the show notes. And one final question for you today: What is your one favourite book that you can recommend to our listeners to become better data scientists?

Ryan: Yeah, there's actually a book that I really enjoy. It's called "The Craft of Scientific Writing" by Michael Alley. It's not going to teach you data science, but it has a ton of good examples of bad scientific writing, as well as how you could correct them in order to make these things denser, better at communicating technical information with less words. I think it's great for anybody who is writing a research paper, but I think it's also great for anybody who needs to communicate technical results to non-technical people.

Kirill: Fantastic. Thanks a lot. So that's "The Craft of Scientific Writing" by Michael Alley. And I can totally stand by that. Data science is not just about the technical skills. It's also about the communication skills. Once again, Ryan, thank you so much for coming on the show. I think so many people are going to get value and insights from all of the amazing things that you shared today.

Ryan: Thank you for having me.

Kirill: So there you have it. That was Ryan Compton, the Head of Applied Machine Learning at Clarifai. I really hope you enjoyed the podcast. There were some very interesting insights that Ryan shared into how artificial intelligence is actually applied in the real world and also how you can break into the space of artificial intelligence and data science in artificial intelligence, how you can structure your career there.

My personal favourite part was when Ryan spoke about the difference, the distinction between data scientist and artificial intelligence. For me personally, it's always been the case that artificial intelligence is kind of like the continuation of data science, it's the next step where data is can take, to go into data science, machine learning, artificial intelligence, and you develop your career that way. But in Ryan's view, you draw a line between the two. You have data scientists and then you have artificial intelligence experts and those are kind of a bit separate and their designations are separate. And whether you agree with that or not, the interesting thing is that it is a view of the world that has helped Ryan achieve success in his career and build amazing products, amazing tools for the world to use and be very happy with his job. So maybe there's something to learn from that and maybe that's a view that you can consider in your own career.

So there we go. Definitely check out the Clarifai website. I'm there right now and it's very interesting to click around here and see what they have. You can actually get their API for free and play around with images there. So if you're developing any products or if you know how to work with

APIs, that could be very helpful to get a view on where this is all going. So it's www.clarifai.com. But other than that, there's just some really interesting images that you can play around with and see how the artificial intelligence is classifying that.

Also, you can get the show notes for this episode at www.superdatascience.com/57. There we will share a link to Ryan's blog, Ryan's Twitter, Ryan's LinkedIn, so definitely hit him up, connect with him and see how his career progresses and what he gets up to in the future. And on that note, I hope you enjoy today's podcast. I can't wait to see you next time. And until then, happy analyzing.