SDS PODCAST EPISODE 253: SOLVING PROBLEMS WITH DATA SCIENCE & UBER

Kirill Eremenko: This is episode number 253, with Associate Professor at the University of California San Diego, Bradley Voytek.

Kirill Eremenko: Welcome to the Super Data Science 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.

Kirill Eremenko: This episode is brought to you by my very own book, Confident Data Skills. This is not your average data science book. This is a holistic view of data science with a lot of practical applications. The whole five steps of the data science process are covered, from asking the question, to data preparation, to analysis, to visualization, and presentation. Plus, you get career tips arranging from how to approach interviews, get mentors, and master soft skills in the workplace.

Kirill Eremenko: This book contains over 18 case studies of real world applications of data science. It comes off algorithms such as random forest, K-nearest neighbors, Naive Bayes, logistic regression, K-means clustering, Thompson sampling, and more.

Kirill Eremenko: However, the best part is yet to come. The best part is that this book has absolutely zero code. So how can a data science book have zero code? Well, easy. We focus on the intuition behind the data science algorithms, so you actually understand them, so you feel them through, and their practical applications. You get

plenty of case studies, plenty of examples of them being applied.

Kirill Eremenko: And the code is something that you can pick up very easily once you understand how these things work, and the benefit of that is that you don't have to sit in front of a computer to read this book. You can read this book on a train, on a plane, on a park bench, in your bed before going to sleep. It's that simple, even though it covers very interesting and sometimes advanced topics at the same time.

Kirill Eremenko: And check this out, I'm very proud to announce that with dozens of five star reviews on Amazon and Goodreads, this book is even used at UCSD, University of California San Diego, to teach one of their data science courses. So if you pick up Confident Data Skills, you'll be in good company.

Kirill Eremenko: So to sum up, if you're looking for an exciting and thought provoking book on data science, you can get your copy of Confident Data Skills today on Amazon. It's a purple book, it's hard to miss, and once you get your copy on Amazon, make sure to head on over to www.confidentdataskills.com, where you can redeem some additional bonuses and goodies just for buying the book. Make sure not to forget that step. It's absolutely free. It's included with your purchase of the book, but you do need to let us know that you bought it.

Kirill Eremenko: So once again, the book is called Confident Data Skills, and the website is confidentdataskills.com. Thanks for checking it out, and I'm sure you'll enjoy it.

Kirill Eremenko: Welcome back to the SuperDataScience podcast, ladies and gentlemen. Super excited to have you back here on the show. Today we have an amazing guest, Bradley Voytek, joining us for this episode. So what you need to know about Bradley is that he was the first data scientist, and the person to kickstart data science at, wait for it, Uber. You heard it right, at the famous ride sharing company Uber, back when they were just starting out in 2013. So you'll hear a lot about that journey on this podcast, and what Bradley learned, and what lessons he can share with you today.

Kirill Eremenko: Also, you will hear about what it's like to hire a team, how to build a data science team at a startup, or at any company for that matter, what kind of challenges you will be faced with, and how to overcome them. We'll talk about automation resisting data science skills. There's a lot of fear that data scientists might actually fall victim to the process of automation and might not be needed anymore. Well, in this podcast, you will find out what skills Bradley identifies that can help you to resist automation.

Kirill Eremenko: And finally, we'll talk more about Bradley's work at the University of California San Diego, where he is an Associate Professor in the Department of Cognitive Science, the neurosciences graduate program, and the Halıcıoğlu Data Science Institute at UC San Diego. And this is super exciting, because UC San Diego is actually spearheading the world of data science education, and specifically, the HDSI, or the Halıcıoğlu

Data Science Institute, is advocating for data science to be recognized as a separate science.

Kirill Eremenko: So we'll talk a lot about Bradley's work there, and specifically, you'll learn four very valuable philosophical points, which you can make in any conversation to argue for data science being a separate field. So a very exciting podcast packed with lots of value.

Kirill Eremenko: Can't wait for us to get started, and just before we do, I wanted to read out our fan of the week, and this one goes to Marin Jorian, and I hope I'm pronouncing the name correctly, who said, "A great podcast for future data scientists. This is a great podcast for those who want to pursue a career as a data scientist. I listened to every single episode of this series, and it feels like I have gained a huge amount of experience from the interviews. Highly recommend it."

Kirill Eremenko: Thank you very much, Marin, for this recommendation, and for those of you who haven't yet rated this podcast, head on over to iTunes or your podcast app and leave us a review. We'd really appreciate it, and I personally love reading your comments. And on that note, let's dive straight into it. Without further ado, I bring to you Bradley Voytek, Associate Professor at UC San Diego.

Kirill Eremenko: Welcome back to the SuperDataScience podcast, ladies and gentlemen. Super excited to have you back here on the show, because I've got a very, very cool person here with me today, Bradley Voytek from San Diego. Bradley, how are you going today?

Bradley Voytek: Going okay, it's not too bad. I wish I was back in San Diego right now. It's sunny and 75, or 24 C, I guess, at home. But I'm up in the Bay area, where it's rainy and cold right now.

Kirill Eremenko: Yeah, you said you were helping out with your parents, your in-laws, with cleaning up after a flood. How's it been going?

Bradley Voytek: It's okay, we're doing all right. They're okay. They were prepared for it, so we've been scraping mud off of the walls, and taking down all the walls, and cleaning everything up, and disinfecting everything. But hopefully everything will dry out, and we can hang everything back up.

Kirill Eremenko: Yeah, man.

Bradley Voytek: Not too bad.

Kirill Eremenko: Crazy. Does it flood often there?

Bradley Voytek: Like this, like once every 20 to 30 years. This is one of the worst ones they've had in a long time. So that's why we're up here helping out.

Kirill Eremenko: Well, hopefully everything goes well and it doesn't happen for a long time.

Bradley Voytek: Yeah.

Kirill Eremenko: Bradley, super excited to have you on the show today. Been a huge fan of yours for a couple years since … I think since the time when you were working for Uber, or when you had just left Uber. Yeah, tell us a bit about your whole career. What's been happening since then, since you were at Uber, and in general, how did

you get there in the first place, like where has your career taken you?

Bradley Voytek: Yeah, I've had a little bit of a strange career. The Uber stuff was in some ways very short, and in some ways felt very long. So, it's actually kind of a funny little small world sort of story that got me involved with Uber. So back in 2010, I had just finished my PhD, and I was supposed to move to Germany. I had a job offer to do post doctoral research in Germany, which is like the next stage. If you wanna stay in academia, in the sciences, after you finish your PhD, you usually go on to do a post doctoral fellowship somewhere.

Bradley Voytek: And my wife and I went out there, and I gave this talk, and I met with the group, and we decided we didn't wanna move out of California, all the way to Germany. And so, I was kinda without a job. I had sort of managed to piece together a small job for a little bit, and was complaining to a good friend of mine about not really sure what I wanted to do next. I'd just done my PhD in neuroscience, I really loved neuroscience research.

Bradley Voytek: And this friend of mine was one of my closest friends from high school. I used to actually, looking back on it ironically, drive him to work everyday, or to school everyday. I was the one friend who had a reliable car enough to pick up everybody, and so I drove everybody to school everyday in high school. And this guy ended up dropping out of college and undergrad, moved around, and he ended up in the Bay area doing software development, and ended up working at Uber,

back when Uber was still called Uber Cab, and it was like a four person startup.

Kirill Eremenko: Wow.

Bradley Voytek: And I was complaining to him, at a … I had some friends over for dinner, and he's like, hey, you kinda do like, data stuff, right, for your neuroscience research? You should come check out this company that I started working for, this startup. And so, he and I kept talking about it, and I met with the CEO at the time of Uber, Travis Kalanick, and we had lunch.

Bradley Voytek: And he was telling me about sort of division of the company, and I started talking to him about the data that they're collecting. They have the phones, and so, they knew where the drivers were located, and where people were requesting rides from. And I was like, it'd be pretty cool to see how people move around. At the time it was only in San Francisco. And so, I was like, it'd be cool to see how people move around San Francisco, right, like what are the areas that are most popular, where are the users, what times of day, and that kinda stuff.

Bradley Voytek: And so, I started initially, back when it was a really small company, my job was essentially just to try and convince people that Uber was a company worth working for, that they had interesting data analytics problems. 'cause at the time, in 2011, the hottest things of the Bay were obviously Google, Facebook, Twitter was really taking off. So you had these companies that are like, everybody wanted to work for.

And so, trying to convince people to be a data scientist at a ...

Kirill Eremenko: A taxi company.

Bradley Voytek: Taxi company, yeah, exactly. This was before ride sharing was even a phrase. People would be like, why do you care? This doesn't sound interesting at all. And so, I had this mix of doing some data analytics, but doing a lot of public speaking at various tech events, trying to convince people Uber had interesting data problems. And yeah, they were some really fun, interesting times. But it was like, I just kinda at this great position, though. My wife was pregnant with our first kid, and so I told them, okay, I was gonna work for them full time until we had the baby, and then I was gonna quit.

Bradley Voytek: I helped them build out their data team, then I was gonna leave. And so I did, but we re-negotiated so that I would stay on as basically a consultant, and continue evangelizing at various tech things for Uber, talking about the data problems that they had, and trying to help them hire data science teams. And I did that off and on, until I finally started my own research lab at UC San Diego in the neurosciences. And at that point, I couldn't do it anymore, so I quit.

Bradley Voytek: But it was a super interesting experience, to get to meet all these amazing data analysts in the Bay area at the time where data science really just started becoming a thing. It was an amazing experience.

Kirill Eremenko: Wow, that's so ... I can't believe ... I didn't know this. I can't believe I'm talking to the, virtually the

grandfather of data science at Uber. That's insane. That is so cool, man. So, was the goal of these conferences that you were attending, and speeches you were giving, was the goal to actually find data scientists for the Uber team?

Bradley Voytek: Yeah, that was a lot of … Part of it was just to talk about how interesting some of the problems were. I can give you an example of one of the blog posts. One of the last things I had done for Uber was write this public facing blog post that was an explanation about the behind the scenes technical side of things, which was, AB testing is very common in a lot of industries. You have whatever version that the product is right now is the A version of the product, and you wanna make some perturbation to the design of that product, and so you create a B version of that product, and you send that out to some sub selection of your users or clients, and then you see, is the B version performing better on whatever your KPIs, your key performance indices, than your A version. If so, you're gonna adopt it.

Bradley Voytek: The problem with Uber is, you've got the A version of the product, and at the time we were talking about, how does the dispatch algorithm work. And so, by that I mean like, you open up the phone. If you're a driver, one of the drivers for the company, and you open up the phone, and you've got the app going, and then you get a ping that says, hey, there's somebody in need of a ride.

Bradley Voytek: When you as a rider open up the app and request a car, which driver gets that ping first? And so, how do

you determine who that gets sent to first? And it turns out there's obviously many different algorithms you could use to determine who would get that ... which driver would get notified first. And the problem is, if you wanna make a change to that algorithm, you can't really AB test it.

Bradley Voytek: In AB testing what you do is you send it out to some subset of your users, and all of those tests are independent. But if you make a change to the algorithm to a subset of your riders and drivers, that's gonna influence what happens to the rest of the system going forward. Everything is interconnected, 'cause the driver that would've gone to you now went to somebody else, and you don't know if the change that you made, any changes that you observe in the behavior of the system, and whatever the KPIs are, are due to the fact of the change that you made, or just some really complicated nonlinear downstream effects.

Bradley Voytek: And so, one of the last projects I'd done for Uber was, I tried to develop a rudimentary simulation system, which is, you can create an agent based model for people who are a little bit more on the technical side, where you simulate the city of interest. So let's say, San Francisco, which is seven by seven, pretty small city, relatively speaking, in terms of geography. And you know where all the roads are, because you have open roadmaps, and things like that.

Bradley Voytek: And you know historically, given Uber's past data, at any given moment in time, what's the probability of somebody opening up the app and requesting a ride from one part of the city, or the other part of the city,

or the other part of the city. And at any given moment, where drivers are around that city, and how do they move around the roads, and things like that.

Bradley Voytek: And so, one of the last projects was building this simulation framework where you say, okay, we're gonna now simulate like a million iterations of the city going forward for a week, using dispatch algorithm A versus dispatch algorithm B, and see in simulation which one performs better. And then you do a bunch of different iterations of different kinds of dispatch algorithms simulated forward to help narrow down computational models of which algorithm might be best for a city.

Bradley Voytek: And then one you have that, you can start to do the slower real world testing of figuring out which algorithm is performing better in the real world. But that was a really fun project, 'cause I'd never done anything like that before. And I had a lot of free rein to do those kinds of out of the world sort of projects of, hey, I have an idea, what if we simulate San Francisco, or New York, or Chicago. And that was super fun. I really enjoyed those kinds of projects.

Kirill Eremenko: Wow, that's so cool. So basically in the simulation, you replaced that algorithm for all the users, is that correct?

Bradley Voytek: Yeah, we don't actually have any … it's all statistically. You're just drawing from samples of, what's the probability of anybody coming in the system here, or here, or here, here, at this time, or at this time, or at

this time, or at this time. 'cause there's certain neighborhoods that are more popular than others, so-

Kirill Eremenko: Oh, yeah, yeah, but I mean like the AB tests that you were running. You wanted to replace algorithm A with algorithm B. In the simulation, you would do it not for a subset, but for the whole population.

Bradley Voytek: Yeah, exactly, right. 'cause the subsets, like I said, because all of these sort of, the way that everything is interconnected, it's really hard to figure out if your experiment is working the way they're intending, or if there are some crazy nonlinear effects that you can't predict downstream. So it's like chaos.

Kirill Eremenko: Yeah, you can't isolate the effect if you don't know what's gonna ...

Bradley Voytek: Exactly.

Kirill Eremenko: Yeah, gotcha. Wow, very, very cool. And, yeah, not many data scientists get such free rein, as you mentioned, to do crazy experimentations. Would you say, right now, probably almost anybody would say that Uber is not a taxi company, it's a data company, right? That's their power, that they have access to all this data, and that they can use it.

Bradley Voytek: Oh, totally, yeah.

Kirill Eremenko: When was the tipping point, for a company like Uber, when they realized that, hey, the value we're creating is not just the infrastructure, and the partnerships with drivers, and the connections with riders, but actually the data that we can use. When was that tipping point, and did change the trajectory of the

company, or was it going there in that direction from the very start?

Bradley Voytek: I think that it was a pretty data driven company from the very start. Although, one of the early things that we had done was try and figure out, where to me it became obvious. I don't know, maybe it was obvious to everyone else already beforehand, but to me it became obvious when I was looking at just what is the power gained in terms of having historical knowledge about statistical trends of where people get rides to and from.

Bradley Voytek: I don't know if people remember, but prior to ride sharing services, the only way to get from point A to point B was either drive yourself, public transportation, or getting a cab, or having a friend drive you. Getting a cab meant a street hail, and if you're in a part of the city that doesn't have a street hail, what you would do is you'd call a taxi company ahead of time, and then you had a bunch of taxi company phone numbers saved. And you'd say, "Hey, tomorrow morning I need a ride to the airport. Could somebody pick me up at five in the morning?"

Bradley Voytek: And whether or not somebody shows up to your house at five in the morning to pick you up was a roll of a dice. It was not a guarantee. And it was frustrating. Unless you were willing to chip in a huge amount of money to get like a black car service, where they would absolutely guarantee that you could have somebody at a certain place at a certain time, but it'll cost like five times more. There was a lot of uncertainty, in terms of trying to get a ride from one place to another.

Bradley Voytek: And so, really, it became pretty obvious, like if you're collecting the data, and you know, on average, on a Sunday morning at 5 a.m., there are still people that want rides, because people are maybe trying to get to the airport on Sunday morning. And then you can look at the data historically and you can say, well, look, there's still a demand at five in the morning on a Sunday. Obviously there's gonna be demand at like … You see these peaks of demand at like, Fridays, Saturdays, at like … I guess it would be Saturday and Sunday technically, two in the morning, in San Francisco when the bars close.

Bradley Voytek: There's huge amounts of demand at those times, so you can be pretty sure you're gonna need a lot of people on the system, willing to give rides at two in the morning. But you also know approximately what parts of the city those drivers should be in. Like, the distribution of bars and restaurants around a city is not uniformly distributed, and especially hot spots.

Bradley Voytek: And so, it became really early on, obviously that you can get a huge amount of value, and that value comes from reducing uncertainty of the user. When you can see that car coming to you on the map, and you know the driver's name and license plate number, and it gives you an estimated time of arrival, and you can see it all happening in real time, that's a huge bit of peace of mind, if you are in a rush to get somewhere, and you're worried about whether or not you're gonna get to the airport on time, and whether or not your driver is even gonna show up at all.

Bradley Voytek: And so, just having access to that kinda data is tremendously powerful, and comforting. It's a good case for users. It saves a lot of time and uncertainty for you. I haven't proselytized like this in a long time. I haven't done this in a really long time. Everything I talk about now is my UC San Diego research. So it's pretty funny to like, put on my old hat of talking about why data analysis at Uber is interesting.

Kirill Eremenko: That is so fascinating, and I can't wait to get to talking about UCSD. I'm sure that's gonna be a whole new conversation. So I wanted to talk to you one more thing about your time at Uber. What was the-

Bradley Voytek: Yeah, of course. I'm not trying to guide the conversation. It's just that it's funny for me to be doing this. I haven't done this in a long time.

Kirill Eremenko: Yeah, I can imagine. What would you say, at Uber, was your biggest learning? Like, you came in there with a lot of experience in data, and like a very developed background in the area, especially in academia. What would you say your biggest learning in terms of data science was during your time at Uber?

Bradley Voytek: Oh, you know what? That's a good question. I actually didn't come in with a huge amount of data knowledge, which is ironic. I did a ton of learning on the job. Even to this day, I still tell people, I teach data science to undergraduates. I'm pretty well versed in all of these topics. But even to this day, if I'm talking to … very rarely if I do any kind of consulting, I tell people, look, if you're looking for somebody who's a deep learning, or a machine learning, or a AI expert, I am not your

person. If you're looking for an amazing program, software developer, I'm good, I'm just, I'm not that. I am not the best at these things.

Bradley Voytek: Like, what I learned really well at Uber is trying to figure out how to … and I guess this comes from the science side of my background too, how do you run an experiment, and often times a data driven experiment, to answer the questions that are important to you or your business, and how do you use the data to help guide you to those important questions in the first place.

Bradley Voytek: I feel like there's a very strong synthesis. There's like a synergetic relationship between using the data to come to data driven principles or to do data driven decision making. But then, circling back, and then running actual experiments, to test the data driven hypotheses, then you keep iterating like that. And developing an experiment is non trivial and very hard, because there are many other factors that could be hidden to you, that you are not accounting or controlling for, that may be influencing your results.

Bradley Voytek: So really understanding where do your data come from, what can your data tell you, how can you run an experiment to better test your hypotheses, and how can you then also use that to generate better data going forward. It's like that whole ecosystem is really what I feel like I learned, that was totally and completely new, that was outside of what I had really learned from my time doing a PhD.

Kirill Eremenko: Wow, that's such a cool way of looking at indeed an ecosystem of asking, using data to make data driven decisions, but also setting up experiments to get more data, and so forth and so on. Very cool. What would you recommend to somebody who's new to this concept? What would you say the one biggest tip is, in this whole process of ecosystem? What's like one pitfall to avoid, or one useful hack that you've learned from this?

Bradley Voytek: To be totally honest, I would say, I see this in the undergraduates that I teach, who are getting degrees in data science. Don't focus just on the technical, mechanical, computational side of things. So many students want to learn machine learning, to go into data science. And to be totally honest, I feel like for most data science jobs out there, 80 to 90% of the machine learning that you're doing is gonna be some variant of a regression. And that's pretty straightforward, right?

Bradley Voytek: Like, everybody gets into it 'cause they wanna do some really cool, sophisticated AI, deep learning, machine learning stuff. But you'll be carried very far if you have a good grounding in your algebra, and understanding of the general linear model. But what that means is, I think that what people should be focusing time on is, to be totally honest, social sciences too.

Bradley Voytek: Programs like demography have been thinking for a very long time, how do you collect and analyze data from a huge number of people. You think about the US census, which is trying to collect information on all 300 million plus US citizens, where do they live, and

all this kind of stuff, and how old are they, in order to get … That's how the US democracy works. That's how we determine how many congressional representatives every district gets.

Bradley Voytek: So, the social sciences I think are under-leveraged amongst data scientists, because they know how to run experiments that are very large and complicated, and they also think about what are the data that we're collecting, and how do we keep those data private and secure. And so, not just the technical, computational, mathematical side of things, which is critical, obviously, but don't discount the fact of, where do our data come from, what are the processes that generate our data, like the political and social processes, not just the technical processes. And then, how do we actually run experiments. So maybe go talk to some of your colleagues over in the social sciences.

Kirill Eremenko: Gotcha. So, would you say that's an important or was an important part of your job at Uber, not just doing the analytics part of things, but actually talking to other members of the team, and seeing what insider knowledge they have about a certain process that's happening in the business?

Bradley Voytek: Yeah, I feel like that was a pretty hard learned lesson for me in some ways, early on at Uber. So, Uber used to only be in San Francisco, and then there was a lot of conversation about where does Uber go next, what cities does Uber go into next. And we would try and do some data driven decision making. I remember I was analyzing … This is really early on. We're trying to figure out what would be the best city for Uber to go

into next. How do you define best? What do you mean by best?

Bradley Voytek: Because best is really critical. The next cities that you go into, it's gonna use resources from the company, in order to spin up a new team in whatever city you're going into. So it's best, the city that would give you the most worldwide advertising, like to help spread the word that Uber exists, is the one that would … the people in the city would most need, therefore and catch on quickly. Like, how do you define best?

Bradley Voytek: And so, that was a big question. And so, the way that we approached it would be, I remember looking at, what is the problem that Uber is solving? We used to call it the last mile problem, which is the public transportation in the San Francisco Bay area is pretty damn good. But it can get very frustrating for people to get to that like, finish that last mile, where you like, walk to a train station, get on the train, and get to where you needed to go. But then you've got another half mile to a mile to go from wherever that train gets off to your work or your home. And if it's a rainy day, not everybody wants to walk uphill in the rain through San Francisco for that last mile. The buses get super crowded, right?

Bradley Voytek: And so, what we started to do was I started analyzing data, looking at, okay, well, let's look at some of the major US cities, and figure out from various data sources, where are the parts of the city where most people live. Like, show me the population density for any given city, show me the public transportation that work for those cities, and show me where bars, and

restaurants, and night life are in those cities. Which cities have the biggest delta between where those public transportation stops are, and the major hubs of where people live and where people go to hang out. The cities that had the biggest average delta between that would maybe be a good place to go. Does that make sense?

Kirill Eremenko: Yeah, yeah.

Bradley Voytek: So, not constraining yourself to … Uber didn't have data from other cities, so we didn't know which cities Uber would work best in, but we were trying to figure out what were the problems that Uber's solving, and then what other data sources exist, that we could get access to, that would help us try and come to a data driven decision, even though we don't have the data ourselves.

Bradley Voytek: But then it became very obvious, as we started talking about this, we're like, oh, look, New York City's a really good place. But it's ignoring the social aspect of New York City, where people are like, you can't go into New York at Uber. Do you know how easy it is to get a cab in Manhattan? You just walk out onto the street and raise your hand, then a cab will pull over. And so, nobody's gonna wanna use Uber and stand in a corner for a random car to show up, when you could just raise your hand and get a car.

Bradley Voytek: But then, it got even more complicated, 'cause you talk to some of your friends, who were people of color. You have friends who are black that can't get a cab to stop for them at two in the morning, because that's just the

way the societal structure is in the United States, where you have some people who are racist, and they're less likely to pull … cabbies might be less likely to pick up a black person in two in the morning, if they're a white person.

Bradley Voytek: So then you have all of these other complicated factors, of like, what do you mean by best? Like, is it trying to reduce this racist tendency? Is it trying to solve this last mile problem? Is it trying to find the cities where you can … X, Y, and Z, right? There's so many ways of defining best. And then it becomes a really hard balancing problem, of how do you take in all of those factors into account.

Bradley Voytek: So anyway, that was an incredible learning experience, 'cause I'm like, oh, I had never thought about how easy it is to get a cab. I haven't spent much time in New York City. Maybe this isn't the best place to try and go to next. Maybe it is, I don't know, right? Like, how do you make these decisions?

Kirill Eremenko: So it's a balance, as you said, it's learning to, as a data scientist, learning to balance the insights that you can get from the data, but also leveraging domain knowledge, social, and I don't know, other types of considerations that might not be so evident from the data.

Bradley Voytek: Exactly, yup.

Kirill Eremenko: Awesome, fantastic. Well, thanks so much. That's a great excursion into the world of Uber. At this stage, I wanna move away a little bit from Uber, and switch gears to talk about your current work, because it's

extremely exciting, if not even more exciting than what you were doing back there.

Kirill Eremenko: You're an Associate Professor at UCSD, which is the University of California San Diego. So yeah, tell us a bit about that. You're in the Department of Cognitive Science and Neuroscience, is that correct?

Bradley Voytek: Yeah. So, I'm in the Department of Cognitive Science, which is an eclectic department. Actually, Geoff Hinton, who just won the Turing Award, a very famous deep learning pioneer, he was one of the early members of the Cognitive Science Department at UC San Diego. He did his post doctoral research there with the founding members of the department.

Bradley Voytek: And that department sorta grew out of this blend of psychology and neuroscience researchers in the 1980s, that were also sort of getting pretty adapted, doing things like neural networks, and realizing, on the computer science and mathematics side of things, people working on neural networks were trying to answer very similar questions that some of these neuroscientists and psychologists were trying to answer. And so, they got together, and started incorporating a lot of different, pretty cool, at the time radical ideas, and they founded this department.

Bradley Voytek: And so, I got hired into cognitive science, and I'm part of the neurosciences program here. My lab actually does mostly neuroscience research. So we try and figure out how do different parts of the brain talk to each other, and how do the 86 billion brain cells or so

that we have communicate in this really messy, biological, noisy environment in our brains.

Bradley Voytek: And then I also helped start this new ... That's the Halıcıoğlu Data Science Institute at UC San Diego, which we launched almost exactly one year ago, like one year and a couple weeks ago.

Kirill Eremenko: Wow, congrats.

Bradley Voytek: Yeah, it's really cool. That's a huge institute.

Kirill Eremenko: What does it do?

Bradley Voytek: Yeah, so it's ... When I started teaching at UC San Diego in 2014, I started teaching at data science class, introduction to data science, which was like the first intro to data science class at UCSD. And that class got crazy popular. By the third or fourth time that I ran that class, we're on a quarter system, so we do three quarters of teaching per year. By the third or fourth time I offered it, it was maxed out at like 400 or 500 students taking my class.

Kirill Eremenko: Wow.

Bradley Voytek: And so, by the time that started really growing, we teamed up, my department, cognitive science, we teamed up with computer science and the department of math here, and we started putting together a undergrad data science major. And so that major launched a little over a year ago, and that major has 550 undergrads enrolled already. It's like a top 10 major on campus. And once that major started coming together, one of our computer science lecturers here on campus, this guy, Taner Halicioglu, he teaches a

couple of classes in the computer science department. He was a UCSD undergrad. After he finished his undergrad, he went on, and he ended up being the first non founding employee at Facebook.

Kirill Eremenko: Oh, wow.

Bradley Voytek: And so, he made a ton of money from that, and he donated $75 million to start this institute, which is the Halicioglu Data Science Institute. And so, that institute runs the undergrad data science major, but it's also a collection of faculty from the different departments all over campus, that are working together towards basically making the argument that data science is an independent research field, different from just computer science and just statistics.

Bradley Voytek: It's kinda interesting historically, the first computer science department in the United States was at Perdue University. And if you look back at some of the arguments that people were making about starting this computer science program, and people were like, we already have electrical engineering, mathematical and philosophical logic, and that's all computers are. Why do we need a computer science? Computers aren't the science, they're the thing we use to do science.

Bradley Voytek: And I feel that really parallels data science right now. People are like, we don't need data science, we already have statistics and computer engineering. Data is not a science, it's something we collect for science. And so, the data science institute is actually taking a pretty strong stance that that is not the case, that data

science is actually an independent, interesting area of research in it of itself.

Bradley Voytek: And so, we've got a huge number of professors already involved, ranging from my department of cognitive science and neuroscience, computer scientists, mathematicians, political scientists. We have the Institute for Practical Ethics here, which is pretty involved, and that's a collection of people in the humanities and philosophy. So it's a pretty broad region institute. And there's a ton of really cool, really cool new stuff happening, even in just this first year.

Kirill Eremenko: Wow, fantastic. Well, first of all, congratulations. It's been probably a crazy ride in this first year, very, very exciting. And I find that about San Diego, it's kinda like, I really am excited about it, but at the same time, it's a mystery to me why data science is … San Diego is such a hub for data science. So, there's a reason why we run our data science conference called DataScienceGo in San Diego, because there is so much going on that space in the city.

Kirill Eremenko: Just one of the examples is the way how fast, how rapidly San Diego is working towards becoming a smart city, full of this data driven technology, with sensors, and a lot of automation. Then another example is all the amazing students, and lecturers, and courses that are happening in the space of data science and applied data science in San Diego. What would you say, is there any specific reason how has it historically happened that San Diego is such a hub for data science, or maybe you have the perspective from

at least UCSD, like why is UCSD so focused on data science these days?

Bradley Voytek: I think it kinda grows out of the history of the university. So UC San Diego is a pretty young university. I think it's like, early 1960s is when it started. So it doesn't quite have the same historical legacy as some of the other bigger tech universities. Like MIT and Stanford have been around since the 1800s, right? But, UCSD grew out of part of the Scripps Institution of Oceanography, which is also a beautiful campus. It's like right on the ocean. It's an amazing place in San Diego. In La Jolla, technically.

Bradley Voytek: But, these oceanographers have been working with huge amounts of data for a very, very long time. It's like early climate science, right? Climate modeling, weather modeling, or just like huge data repositories. And when UC San Diego got started, it became just this biotech behemoth. So you have a huge number of biotech companies that are headquartered in San Diego.

Bradley Voytek: And so, you've got this one arm, which is weather and climate modeling, which has been steeped in really complex modeling, and mathematics, and data collection for a long time. And then you've got the biosciences, which are also incredibly data rich, right? Just like, one single person's genome is just an incredible amount of data generated, right? Like, how do you begin analyzing and doing data mining, and all these kinds of things.

Bradley Voytek: And so I think part of it is that, then the other part comes from, a little bit of the technical legacy of the city also. You have companies like Qualcomm headquartered there. Teradata just moved their headquarters to San Diego. So you've got not quite the same tech hub as Silicon Valley, of course, but that mix of the biotech, with the physical tech, with the data driven research legacy that San Diego has, from Scripps Institution of Oceanography, I think that gives it a nice ... It's a right blend of topics that lends itself to the data sciences really nicely.

Bradley Voytek: And part of being a young university also makes it a little bit more flexible, I think. The leadership of the university is a little bit more willing to take chances on new endeavors, like creating this big institute. So, yeah, it's an interesting, interesting time and place to be. It honestly kinda feels like ... I get a similar vibe as I got when I first moved to the Bay area in 2004, I guess, where there were just like so many interesting things happening at the same time, and you can kinda feel something coming together. I get a little bit of that same vibe here in 2019 in San Diego, which is pretty cool. It's nice to go through another round of this.

Kirill Eremenko: Yeah, I totally agree. And I think, so I heard this interesting opinion, and the credits go to Tristan Blake, who was also on the podcast before. I'd love to get your thoughts on this, that the difference between San Diego and San Francisco is that in San Francisco ... Both amazing places, great, have lots of merits, there's fantastic people in both areas. But in terms of the setup, it's little different in the sense that San

Francisco is more fintech, whereas in San Diego is more kind of applied tech, applied data science, like you mentioned, to different types of research within areas of research, whether it's oceanography, whether it's medical sciences, and so on.

Kirill Eremenko: And what that carries is that in fintech, it's more of a zero sum game, in the sense that a lot of these companies are competing with each other, whereas in San Diego, and I felt this as well, that people are ... go out of their way to help each other out, to share knowledge, share experiences. There's a sense of kind of a camaraderie, in terms of businesses moving forward and leveraging data science in that process. Would you agree with that?

Bradley Voytek: Yeah, I think there's like a different speed too, right? Like, fintech and the startup culture, both of those are so fast. And biotech is slow. Experiments take a long time, regulations take a long time. And similarly, climate and weather modeling are also pretty slow. There's a different pace. And so, I think ... And also, I think coming along with that, there's like a different pace of the culture of cities, right?

Bradley Voytek: For example, it's not at all uncommon, when I pull into the parking lot at UC San Diego in the morning, to see professors' cars with wetsuits draped over the driver side mirror, 'cause they went surfing before coming into work, and their suits are drying out, and they're getting ready for the day. It's just like a totally different pace at which people operate. And maybe that's not as exciting. Like if you are young and going for the fast paced thing, it's a little bit slower. But I think coming

along with a little bit of that, let's take a step back and think about what is it that we're trying to solve, there's a lot of positive lifestyle benefits that come along with that.

Bradley Voytek: And like, I just gave a talk a Strata in San Francisco. It's part of the reason why I'm up here. And the talk was about how do we educate undergraduate data scientists in ethics and data privacy. And I was mentioning, I showed some of the course materials, and I'm like, hey, look, my entire course is up on GitHub, every lecture, all of the slides that I give, here are all the assignments that I give the students, here's what the final projects look like, here's the syllabus, everything is open on GitHub. And a professor from another university came up to me afterward, and they're like, I can't believe you just made everything open, and I'm like, wait, why wouldn't I?

Bradley Voytek: And they're like, well, it just takes so much time to put these courses together. And I was like, right, so why don't I just give that to somebody else, and put it on GitHub, and if somebody sees an error in my slides or something, then they can just do a pull request. Like, I don't … What's the point of keeping this to myself, right?

Bradley Voytek: So there's this huge funky collaborative spirit that to me is surprising that it's not universal. And what that leads to is … So I'm the director of the undergraduate data science scholarship program here, and one of the student groups is proposing a collaboration. They're a dance maker, and they wanna figure out how do dance troupes learn how to dance, like how do they learn to

actually coordinate their choreographed movements. And so, they're working with a computer science lab, and ... What's the other group? One of the computational neuroscience groups, to do full motion body capture of dancers learning these dance routines, and then analyzing that full motion body capture over time, to figure out when do people learn how to synchronize with the music and synchronize with each other.

Bradley Voytek: And I'm like, that's a damn student ... That's like computer science doing full motion body capture stuff, and some of the computational neuroscience people are also collecting heart rate measures and things like that. That's a crazy collaboration. I love that kind of thing. Those are the kinds of things that you try and facilitate when you have a really big collaborative environment. You just sorta step back and see what interesting ideas students and people here, researchers can come up with, right? And that only stems from that comfortability ... comfort, I guess, with collaboration.

Kirill Eremenko: Very, very interesting stuff, and I think that also very well resembles the general nature of data science. Like if you look online, if you look at data scientists online, generally, they help each other out, right? There's a lot of sharing, and stuff like that. And that's really cool that there's, in San Diego and in UCSD, this same thing is happening.

Bradley Voytek: And you look at like ... Oh, sorry, I was gonna say, you look at like, the foundational bits of software for data science, modern data science, Python and R are just

critical components, right? And those are both opensource programming languages. And so, data science is inextricable from opensource software development and this sort of development communities, I think. You can't separate them.

Bradley Voytek: Which, for all of current and all of the ill that entails. It's a complicated thing, but I think modern data science, it really comes about because of these sorts of vibrations.

Kirill Eremenko: Yeah, for sure. Definitely. Very exciting. Very excited for you that you're there. I wanted to ask you, so, you mentioned that one of the things that you guys are pushing forward is that data science is a separate field of study. What argument would you make? I think this would be interesting for our listeners to know. When they're having these discussions or debates with somebody, what's a good argument you would make to support that view?

Bradley Voytek: I can make one that's on the technical side, and one that's on wearing my cognitive science side of things. So, on the cognitive science side of things, I think, the what is data science and how is data science unique argument that I give is almost a philosophical one of, how is it that we can record numbers about the world around us, and learn fundamental laws about the way that the universe works?

Bradley Voytek: By observing the motion of the planets in the night sky, and writing down where and when they are, and how fast they travel, and all of the secondary metrics that you can get from that, of like, acceleration and

things like that, just through doing that, observation and recording data points, you can infer fundamental laws of the universe. That's the bigger philosophical side of things. Like, how is it that just recording data from the world around us tells us something about the way that the way around us works, which I find fascinating. 'cause then once you do that, you can then make predictions about the future, and that's an amazing thing, that writing numbers down allows you to predict the future.

Bradley Voytek: The more technical side of things is, in computer science there's this idea of big ol' notation, computational complexity, like what is the runtime of an algorithm? Is it linear, does the algorithm get … does it scale linearly, in performance with the amount of data or numbers that scale logarithmically, or what have you? And I find that there is … I've been sort of programmed to this idea of, there are certain classes of data driven problems that, as you add more data, your ability to predict gets a little bit better, but there are other classes of problems, where it seems like you add a little bit of data, and your ability to predict gets a lot better.

Bradley Voytek: And I like to give the example of automatic machine translation. So translating languages has been like this holy grail for linguistics for decades. These amazingly brilliant people have been arguing for a very long time about, what are the language universals? Are there certain aspects of language and grammar that are universal across all languages, and which ones are specific? Can we identify those universals and

hard code them into computers, and can we then, in order to do translation, do we have to encode a bunch of conditionals that are specific for each language?

Bradley Voytek: And these arguments are happening, and they continue to happen, but while all this is happening, Google comes along, and they release Google Translate. And they're like, actually, if you just throw deep learning at enough data, you get a pretty good translation automatically. People make fun of Google Translate, because it makes some kind of silly errors sometimes, but it's pretty damn good, and it's kind of like magical. When you look at, I can pull out my phone and have a conversation with somebody who speaks not a bit of English, and I speak not a bit of their language, but I can pull out my phone, and they could talk into my phone, it'll translate it for me, and vice versa. It's a little slow, but like, that's the stuff out of sci-fi.

Bradley Voytek: But that didn't work until Google had huge amounts of data. They needed just enormous amounts of data to throw at the problem, before it finally got to work. And you look at some of these computer vision and deep learning problems, where it's like, the best algorithms do, like arbitrarily, like say, 95% accuracy. And then, as we get exponentially more data, they get up to like, 99% accuracy. And you have exponentially more data couple years later, and it's 99.2% accuracy. Like, there are certain problems that like, you throw exponentially more data at them, and you eke out just a little bit better performance, but there are other problems that

you throw just a little bit more data, and they do tremendously better.

Bradley Voytek: So I think about this idea of like, data complexity. Like, how much data do you need in order to come to an accurate prediction, given the amount of data? Like, what classes of problems are linear data problems versus exponential data problems, and things like that. And I guess the final aspect, like what I think is like the art of data science that I find the most interesting, is integrating heterogeneous data sets.

Bradley Voytek: So you think about, imagine you're a data scientist working at Facebook, I don't know. And you need to predict, what's the probability that some user is gonna click on any given ad. And that number, the output that you want is a number between zero and one, the probability that they are gonna click on that ad. What do you theoretically have access to as a data scientist working there?

Bradley Voytek: You have access to the freeform text that they write in their status updates, as well as when and from where they wrote those status updates, and from what kind of device. You also have access to maybe their social network. Who are they friends with? How strong are each individual friend connection? Maybe they include a hyperlink in their status updates. Those hyperlinks are themselves graphs that have some information, right? So if I post a link to, I don't know, the New York Times in my Facebook status updates, I'm more likely to post the link probably to like, Washington Post than I am to like, Fox News or something, right?

Bradley Voytek: And then you have maybe access to their photos, that you do computer vision on, and then you extract text information about what's in those photos, how many people. So you have all this different kinds of data. You've got image data, freeform text data, social graph, you have some demographic information. How do you get all those different kinds of data, text, image, geography, temporal data, and come to a probability that they're gonna click on the ad? Those are very different kinds of data, and so you're gonna have to statistically integrate them in some way to come to a probability of ad clicking.

Bradley Voytek: It's non trivial, and that's not just a statistics problem, in my opinion, and that's not just a computer science problem. It's sort of a new space of, how do you integrate different kinds of data, in order to do math on them. And I find that to be … So those are like the big different kinds of domains, in my opinion, of what makes data science relatively unique.

Kirill Eremenko: That's very cool. So, to sum up, and correct me if my understanding is wrong, we got three main pillars. The first pillar which you described was that it's phenomenal and philosophical, actually, that we can observe the world, write down some data, and then use that data to describe how the world works, and in fact make predictions about the future. When you think about it deeply, it is a philosophical notion, that is not trivial that through just numbers, we can extract all these insights about the world, about the laws that govern it. So that's the philosophical component of this field.

Kirill Eremenko: The second pillar is the complexity up there, that how much data do you have, how much data do you need for some problems. Different problems require different ones there. So this is probably like any science. Chemistry works with molecules and atoms, physics works with laws of nature, math works with numbers. Any science has a fundamental substance or fundamental asset that it works with, and data science, it is data, and the fact that data complexity can vary is one of the characteristics of data. And had it been just a constant, like data is always this thing, then that's not as interesting. But here, the fact that it's varying, and that affects all the problems that we're solving, so there is some sort of underlying substance of it there that we need to work with, and that can describe this field of science. So that's pillar two.

Kirill Eremenko: And pillar number three is that there's an art to data science, right? All these sciences, they all have an element of creativity or an exploration art to them, and data science is no different. You need to know or establish ways to integrate data together to solve different problems, and it's not trivial, it's not linear. It requires some sort of human input. And so, when we put all those three pillars together, the philosophical, the substance, and the artistic, that is sufficient reasoning to separate data science into a field of its own.

Bradley Voytek: I believe so, yeah. And of course, I guess if you're gonna throw in one more, there's the social aspect, right? Like, where did the data come from, who's generating the data, how are they being used, and how

does it influence policy and decision making, and culture, and all that kind of stuff, right? So yeah, they're absolutely right. You nailed it.

Kirill Eremenko: No, you nailed it. I just summed it up. So that's really cool. So, guys, if you're ever having a conversation and somebody's arguing that data science is not a separate field, it shouldn't be, there you go. There's some ammunition for those discussions.

Kirill Eremenko: The other thing I wanted to chat to you about, we got about 10 more minutes on this podcast, teaching data science. So, you've been teaching data science at the university. You mentioned your current class is about … or the enrollment right now is about like 500 people in this program of data science. What have you noticed about … Any insights you can share with us about teaching data science?

Kirill Eremenko: For instance, one of the things that pops into my mind is, what are some of the common things that students struggle with, and how do they get over them in the course of learning data science?

Bradley Voytek: The biggest thing I see students struggle with is data intuition, and algorithm intuition.

Kirill Eremenko: Yes, that's my favorite thing, favorite topic.

Bradley Voytek: When I give talks about this, there's a slide I included, it's a question that somebody put on Stack Overflow. And the student's question was something along the line of, okay, look, I know how to do PCA, and I can calculate eigenvalues and eigenvectors like a machine, but I don't know why I'm doing it or when I should be

doing it. Like, even though I understand all of the math, I understand the algorithm perfectly, I don't understand why I need to be doing this, and which problems are suited for this. I forget the exact wording, it's a little bit more wordy than that.

Bradley Voytek: But that really like, that is what I see. You get these students that come in as first year undergraduates, and they've taken AP statistics, and AP calculus, and AP computer science, and they just know all of these mechanical side of things. But then, you put them in a machine learning class, where they get the MNIST data set or something like that, which is a really nice, clean, well described test data set that lots of people use, and they'll throw different machine learning algorithms, import scikit-learn and Python, and throw some stuff at it, right?

Bradley Voytek: But that's all mechanical. They're not engaging with the ideas. And I feel like that's the value of … that universities still provide, over and above what you could just do from self teaching online, and through online classes, is this idea of a broad education about why are you using this algorithm. Why does this algorithm work? When does it fail? When does it not fail? What do you do if you have messy, noisy data.

Bradley Voytek: And so the way I describe it is, I'm trying to figure out how to teach students automation-resistant data science skills. There's this big concern that I hear amongst certain data scientists, that like, in five years, most of their job is gonna be automated out of existence. You're just gonna be able to import the data cleaning package from Python, and just churn your

data through that, and then import all data science models, and run automatically through all of that, and then you'll be done.

Bradley Voytek: And so, yeah, people really need to have the strong technical background. They need the math, they need the programming. But they also need the, what are those automation resistant skills, and how do you teach those to undergraduates. And so, that's really what I've been focusing on in the last year.

Kirill Eremenko: Wow, that's so cool. Don't leave us hanging, what are those automation resisting data science skills?

Bradley Voytek: Okay, all right. So, let's see. I've talked a little bit about this idea of how do you integrate these different kinds of data sets. You're talking about, you've got, for any given user, or account, or whatever, you've got geographic data, textual data. What is the appropriate dimension across which to integrate those data? You can integrate data temporally, so number of words they use per hour, and from where they post and when, and how their social network looks over time. Or you can integrate geographically. What kinds of posts they write when they're in point A versus point B, or whatever. So that's one thing I've already talked a little bit about.

Bradley Voytek: The other one would be, how do you foster a data first thinking. So teaching students, how do you use data to solve a problem? Like, you're gonna create a company, like in Uber, right? What data should you be collecting from day one, in order to make your product better? And what data are you allowed to collect? What

data should you collect? What data should you not collect?

Bradley Voytek: And then, also just generally this idea of data literacy, and intuition, and creativity, which is, we've talked about also a couple of minutes ago. Data communication and visualization, so how you actually use data to tell the story that you are trying to tell, accurately, without lying. And that's important, right, because most practicing data scientists, at the end of the day, they have to convince somebody else, who is making the product and implementing their solutions, they have to convince them that the data are telling them something that is truthful and accurate.

Bradley Voytek: And so, those four things, of how do you integrate different kinds of data, how do you foster this data first thinking, where you can use data to solve problems creatively and collaboratively, even if you don't have access to the data yourself, even if your own team isn't generating the data. So going back to the problem I gave about Uber, which city does it launch in next. Well, it's solving this last mile problem. So take a data set looking at where do people live, take another data set of, where are public transportation stops. Then take another data set of, where are bars and restaurants located.

Bradley Voytek: Then using and analyzing those data to try and figure out what are the cities that have the biggest distances between public transportation, and where people live, and where people play, basically, hang out. So how do you teach that? How do you teach that kind of data creativity? And then data literacy, intuition, and then

the visualization sides. So, that's what we try and do. It's not easy, to try and figure out how to build the classes around being creative with data. But I think we've got a pretty good batch of undergraduates coming out of this program in the next year or two, when they start to graduate. I think they're pretty clever.

Kirill Eremenko: That's fantastic, that's fantastic. So, it almost seems like in the data science project life cycle, where you've got to like, identify the problem, get and prepare the data, run the algorithm, visualize the insights, and then present the insights. In those five steps, it almost seems like the only ones that can be automated are number two, get and prepare the data, and number three, run the algorithms, build the models.

Kirill Eremenko: Whereas preparing for the problem, like asking the right questions, finding out, getting the domain knowledge, and things like that. And then on the tail end, where you have visualizing, making sure that you're visualizing without including any false information or making it even biased towards something, and also the presentation part, the data scientist will still be required there.

Bradley Voytek: And a big part of that, this is a hot topic in the general field right now, is explainable models. Deep learning can do some pretty amazing thing, but how do you interrogate what is happening under the hood, in order to come to some human level understanding of why some model's working better than another one, right? So that still requires the human level of intervention in a number of ways, right?

Bradley Voytek: Because ultimately, at the end of the day, we don't want just a model that explains the world, we want a model where we understand what the different variables in that model represent. It does no good to show somebody, F equals MA, without telling them what F, M, and A stand for, right? Otherwise, you're just saying this number equals the product of two other numbers, and that's not helpful at all, right? Like, F, M, and A are things that humans understand. They have meaning. And so, it's one thing to have a mathematical model that perfectly predicts the future, which is useful, but as humans, that's unsatisfying. We want to understand what do those variables represent, and how can we interpret them and leverage that information to do something better.

Kirill Eremenko: Yeah, yeah, totally. Totally agree with you. Bradley, one more question before we wrap up. From all the things that you've seen, the things that you've taught, the work that you've done, where do you think the world's going in terms of data science, and what should our listeners, and in fact, what should your students that are graduating in the next couple of years, what kind of world are they going into? What's the world for data scientists gonna look like?

Bradley Voytek: I think the next stage … Okay, I try and be optimistic. I'm of the opinion that there's a lot of, it's the great power comes with great responsibility, and I think this is … The future holds a lot of good, positive possibility. And so, in order to get there, we gotta try and figure out how to teach the right things. And I think the future of data science is coming to tools that allow

people to visualize data that's important to them, that they want on demand.

Bradley Voytek: I'll give an example from my field of neuroscience. There's a tool that my lab is currently working on trying to build, which says, well, we have these massive databases that tell us what parts of the brain … which genes are expressed, how strong, and in which parts of the brain. You have 20 something thousand different genes that are expressed in the brain and are expressed differentially across the human brain.

Bradley Voytek: We know, from another set of millions of studies that have been conducted in neuroscience over the last 100 years, which parts of the brain are associated with what. So I point at any given part of the brain and I say, there's an 87% chance that that's a visual part of the brain. And here's another part of the brain. That's a 17% chance this is the language part of the brain. Alzheimer's, memory, attention, whatever, right?

Bradley Voytek: And then you have another database that says, here are all the connections between all the different brain areas. What's the probability that this brain area is connected to that brain area, and that brain area is connected to that brain area, with physical connections between synapses, between neurons. And these all exist in these very different data sets, and databases, and data formats.

Bradley Voytek: And I think the power and the future of data science is going to be bringing together these disparate data sets that exist in different domains and integrating them in

ways that are greater than the sum of the parts. And I think that's gonna be the case all over the place, geographic and geospatial data, brain based data, temporal data, textual data. You're gonna start seeing people integrating and creating tools for integrating disparate data sets, that allow for better, more informed data driven decision making, that will answer questions that are important to us, like helping people cure disease, reduce pollution, increase the quality of life, and so on.

Kirill Eremenko: Fantastic, fantastic. Thank you. That's a very bright future. And Bradley, thank you for coming on the show and sharing all these insights. It's been a really cool, cool chat. Very, very thoroughly enjoyed it so far. It's amazing.

Bradley Voytek: There's nothing a professor likes more than just bloviating himself. Thank you giving me the outlet to do that.

Kirill Eremenko: That's awesome. Bradley, before I let you go, could you please share with our listeners, where can they get in touch, contact you, follow you, and find out more perhaps about the courses that you're teaching at UCSD, and other work you're doing?

Bradley Voytek: Yeah. So my website is voyteklab.com, V-O-Y-T-E-K-L-A-B. I am also, my lab is on GitHub, and we have a lot of the … like the neuroscience related stuff that we do, as well as we curate a number of open data, open tool repositories, resources, as well as, I think my classes that I teach are maybe on there or discoverable through there. And that's github.com/voytekresearch,

all one word, all lower case. And then also on Twitter, @bradleyvoytek, all lower case, all one word.

Kirill Eremenko: Fantastic, and is it okay to connect with you on LinkedIn?

Bradley Voytek: Yeah, and on LinkedIn, I'm also just Bradley Voytek, same as my Twitter.

Kirill Eremenko: Awesome. Guys, make sure to get in touch and follow Brad, and one more thing. I've got one last question for you. What's a book that you can recommend to our listeners, to help them in their careers, or maybe even in their lives?

Bradley Voytek: Oh, wow. Can I give one actual useful book and one book that maybe gave me an idea of proto data science thinking?

Kirill Eremenko: Yeah, sure.

Bradley Voytek: So the useful book that I really like is, there is a book, Data Science From Scratch, which I don't know if anybody's ever recommended that to you before, but that's from Joel Grus, which is just, that's super handy. I use it to teach my undergraduates sometimes. It's a very easy to use book. And then-

Kirill Eremenko: I've met Joel Grus. He's a really cool guy.

Bradley Voytek: Yeah, I keep trying to get him to come down to San Diego and guest lecture. I should reach out to him again. And the other one is, honestly, the Asimov, Isaac Asimov Foundation trilogy. There's this idea of psycho-historians, which are people that are integrating all this information about people and

behavior, and able to predict that the intergalactic empire, or I guess intragalactic empire is about to collapse. And in anticipation of that collapse, they start to collect all of the world's information into this encyclopedia, so that they can reduce the severity of that collapse.

Bradley Voytek: And so, that's like … Looking back on having read that as a kid, I'm like, yeah, they're just collecting all of this data, and they're able to predict the future, by collecting data about trillions and trillions of different people across the galaxy. That's kind of this proto-modern data science viewpoint, right? But it's all … In the book, it's all geared toward doing something good and useful, and I like that.

Kirill Eremenko: Fantastic. Thanks for sharing. So, Data Science From Scratch by Joel Grus, and the trilogy from Isaac Asimov. Okay. Well, once again, Bradley, thanks so much. It's been a huge pleasure having you on the show. And I'll be in San Diego. Looking forward to catching up in person. Definitely I feel very excited about that.

Bradley Voytek: Yeah, we should. Let me know when you're in town.

Kirill Eremenko: For sure.

Bradley Voytek: Thank you for having me on. I appreciate it. It's been fun.

Kirill Eremenko: So there you have it. That was Bradley Voytek, Associate Professor at UC San Diego, and one of the first data scientists at Uber. I hope you enjoyed this conversation as much as I did, and personally for me,

the most valuable takeaway were probably those four points that Bradley shared about how you can argue for data science to be recognized as a separate field. I think a lot of us have these conversations quite often, and sometimes it's hard to articulate why data science is actually a science, and should be recognized equally to physics, or mathematics, and chemistry, and I'm glad that we went into this conversation with somebody with so much experience and knowledge in the space, such as Bradley, on this podcast. And I hope that it was helpful for you as well.

Kirill Eremenko: And on that note, you can find all of the materials that we talked about on this episode in the show notes, which are at www.superdatascience.com/253. That's superdatascience.com/253. There you'll find a transcript for this episode, any materials that were mentioned, plus all of the social links to Bradley's profile, so make sure to follow him there, follow his career, and see what other new exciting things he gets up to along the way.

Kirill Eremenko: And if you enjoyed this episode, make sure to spread the love, send it to somebody who you think might enjoy it as well, and might learn from the things that Bradley had to share from his exciting career journey. And on that note, thank you so much for being here. Can't wait to see you back here next time, and until then, happy analyzing.