[JAMES HAIGHT]: This is the Hadooponomics podcast and I am, of course, your host, James Haight. Good to be back with you here, great week this week. We're really excited to bring on Edd Dumbill to the show. Edd works at Silicon Valley Data Science, one of the most interesting companies out in the Valley in the world of Big Data. You guys are really gonna love his perspective. What I love about Edd is he's done a lot of really interesting things. He started the now massive Strata + Hadoop Conference back when it was a fledgling conference just getting off the ground. But he also spends a lot of time working with blue chip companies, telling them how to bring data into their operations and how to really maximize their investments. So some pretty interesting stuff, and really this episode is a bit more on the practical side. Edd's spent a lot of time hands on, helping people construct good data environments for their companies going forward. And so really, if that's your sort of thing, we have an excellent episode for you today. A really good chance to get some helpful tips and to bring it back into your job and to drive forward your investments.

But also, if it's not your thing, Edd's a bit of a futurist. And so we actually take a pretty good look into what's upcoming in the world of machine learning and artificial intelligence. Some really interesting implications for the future of work and the future of how we interact with machines, and how we will perceive work in the future. So if that's more your cup of tea, we've got a little bit of something for everyone in this episode. Excited to bring it to you. And of course if you wanna follow up, find us at bluehillresearch.com/hadooponomics, it'll show you how to get in touch with Edd, as well as the transcript, and of course the recording of this. Otherwise, find us on iTunes, Stitcher Radio, or find me on Twitter. If you really like what we're doing, go to iTunes, leave us a five star review, we would really appreciate it. And if not, send us a note and we're gonna really try and engage you guys in more conversation. So let us know what you think, what you like, what you don't like, and we'll work it into future episodes.

So with that, send us a message, happy to connect. And that's really it for me, so I'm gonna step aside and let's go straight into the interview with Edd.

Okay, ladies and gentlemen, I'm here with Edd Dumbill. Edd, you are the VP of Strategy at Silicon Valley Data Science, it's a really interesting consultancy out in California. Welcome to the show.

[EDD DUMBILL]: Thanks very much, it's great to be on.

[JAMES]: So Edd, before we dive into all the things we want to talk about, give us some background. Who you are, what do you care about, and we can take it from there.

[EDD]: I've really had a lifetime of being somewhere in between developing software, doing start ups, writing, and educating. So I enjoy being on the edge of emerging technology and translating that into useful things for folks. Most recently I've spent the last few years in the Big Data and data science world, both at O'Reilly, where I helped start the Strata conference, and now it's Silicon Valley Data Science, which is conducting a really interesting experiment, I think, about can you bring data-driven business out of Silicon Valley and into the enterprise.

[JAMES]: Absolutely, and think you hit on something I wanted to touch on at the beginning. I first met you at the Strata Conference this year out in San Jose, and I was introduced to you as the guy who invented the show, which I thought was pretty interesting. And for our listeners, I'm sure a number of them have actually been to the conferences, but the Strata + Hadoop conferences put on by O'Reilly and a number of other players is growing. It's an amazing sort of congregation of folks talking about all things Big Data, and it has been just off on a rocket ship. And I'm wondering if you can just speak to, what's it mean when someone says that you started this conference? [laughs] And sort of take us through the timeline of how this evolved.

[EDD]: Yeah, absolutely, I can say from the beginning that there's a huge team behind the Strata event, and I was lucky enough to be there at the genesis. But the one or two interesting things that I think I helped contribute, at the time I was working at O'Reilly, I was responsible for their conference software, and also I was Chair of the Open Source Convention. And we'd seen something new happen in really what you would consider to be the platform stack, right? You either before were a Microsoft shop, an Oracle shop, or you were sort of an open source LAMP stack shop. And we saw that two things were happening, really. One, that there was a scale out cloud stack emerging that people were developing on, and the second thing is that the role of the data scientist had emerged. The analytics were being used in product situations, not just as an operational concern. But really part of building out products and taking advantage of the network and large amounts of users.

So there's definitely something going on, and typically when O'Reilly sees something going on it looks to make friends and understand what's really happening. So we got a few folks together, names now that have been in the scene for a long time, and had a little think tank, and got together, and at the end of the day brought a conference to life. Now that first one, 1,300 people in Santa Clara, which, in the context of many thousands that are turning up to the Javits Center in New York, now, it seems small. But actually 1,300 people for a conference launch for O'Reilly was almost double or triple what you might actually expect the first time out of the gate. So we knew that something's happening, and we knew that when all the data scientists turned up to Strata, they said, well, I've found my people, these people are doing the same kind of things that I am, which was profoundly exciting. And since then, as well as the phenomenon of data science, which was originally the really unifying thing about Strata, Big Data as an industry, and the large number of events that have flowed into it, it's really taken off. And now you're seeing the events, as I said, of many thousands of people, trade show thing that takes over the center, the convention center and the Javits Convention Center in New York each year.

[JAMES]: Yeah, absolutely, and I was actually at the one at the Javits Center in New York as well, and to say it's overwhelming with the amount of people and things going on there is probably a bit of an understatement. It's sort of a really interesting spectacle to go and see.

So part of what I want to get into, right, is on this podcast we always try and talk about how can folks actually get value out of their Big Data investments. And certainly some folks have been very successful and others have been a little disillusioned because they spent a whole lot of money and they've spent a couple years, and they're still developing science projects. I wanted to sort of put that in the context of your purview of watching sort of this community evolve, and folks being able to talk to each other, and finding their people, as you say. Can you talk to this narrative of some people have not had success with their Big Data investments, and what's changing, or what needs to change in order for that to actually be profitable for them in the future?

[EDD]: Absolutely, and I think the most important thing to realize with the Big Data world is, as much as it is with any technology is, it really isn't a case of you buy it and magic things happen. And we identified this right at the beginning at Strata, when we had a thought that really likened analytics to a big flat box of magic that a CEO would say, here, analytics is for me! And then wonderful results would appear, right?

[JAMES]: [laughs]

[EDD]: You have to understand a little bit about how to actually interface the exploration and exploitation of your data into solving a business problem. So right now, this is probably manifested most clearly in the trend of talking about data lakes, with large repositories where you can unify data that was previously siloed. That is a good thing, right? But there's no way in which, if you invest several million dollars in a year in that, putting the data together, that that will result in immediate business benefit. And if the wrong people are leading it, which is to say, if the right people aren't interested in it, then it's not gonna get joined into business, it's not going to be created with an eye to the problems it solves. And I think that's the most important thing. You can get value straight away if you create your first use case to solve a problem you actually have. And this is the pattern if you look at successful adoptions of Big Data. This is absolutely the pattern they've had. And even before the trend we have now of it into an organization is a perfect example of that, when the use of analytics, data science, in one department took off, the neighboring departments said, I want some of that, can you teach me, and so on and so forth. So there's two ways it can go, it can go bottom up or top down. But the absolute imperative is that it has to be joined with something that the business actually needs to do.

[JAMES]: Okay, so you talk about platforms and I mean, we've seen the rise of some pretty big name companies, Hortonworks, Cloudera. We actually had the Chief Security Officer of Cloudera on the show. These guys are making tons of noise and they are really pushing forward this idea of becoming this platform. What are they building for? You mentioned everyone used to just be Oracle, SAP, Microsoft shop, right? But things are changing, and what has changed and where are we going?

[EDD]: I think it's really difficult to imagine that coming in through Hadoop as your entryway is gonna lead you to become a platform company that has widespread adoption. I think that the Hadoop companies are in a really tricky space, they're racing against time to integrate vertically to be immediately more useful to their customers. And they're racing against people who already have a lot of sway and a lot of budget analyst customers, frankly. So looking at things like how more difficult is it for Cloudera or Hortonworks to establish and develop a community and enterprise application frameworks such as Microsoft or Oracle than it has for Microsoft or Oracle to figure out how Big Data integrates into what they've already got.

And I think, actually, there are other trends going on right now, Big Data is not the only one, the only transformation. You look up and down the IT stack, you're looking at cloud and containerization actually being an important set of technologies that's coming in. So if I gave you my predictions for who's a big IT platform going forward, you have to add in AWS and Google on top of the Microsoft, and Oracle, and SAP actually. Because they are part of the larger move towards the cloud, which gives that opportunity for a shift and integration with their services. It's difficult to be in a pure, horizontal layer and have long term sustainability, especially when there is this rising tide of open source. No sooner can you think that this thing you'll build on top of Hadoop that will make it better, and you try and sell it, then three of your competitors have got together to try to create an open source equivalent, and they've kind of let the air out of what you're doing. So it's a very competitive space. So in this context of being able to run experiments, being able to move faster and change faster, this foundation of infrastructure we have with cloud, DevOps, and open source that allows an organization to really summon infrastructure much more quickly and much more cheaply. And also get rid of it as well, where you're not talking about CapEx, you're talking about OpEx, clearly you've not been stuck with a legacy of investments that you need to optimize.

And then there's one layer above it, right? These other trends or other movements that we've heard of over the last ten years, which is how to build on top of it. So we understand now about platforms and APIs and internal services. A very flexible way to build. Instead of building monolithic applications you start to build robust services that can be reused and composed when you have, let's say, a web service, as we used to call them, I guess we probably should have another name for them now, on top of your customer's database, let's say. And all the clients of that hit that service rather than go down to the database. Then you're free to evolve that scheme of the database much more quickly. There's not a lot fragile, one-to-many connections going in.

So that's one thing, and then, how do you develop on top of that? So we've had in the last couple decades the Agile software movement, which I know is a lot of, to some people, every time you get some technique it becomes gospel. I would just say it presents ways of working that allow you to flex more easily with the business needs. You talk about small cross-functional teams very heavily integrated with what the business needs. So then we've got infrastructure which we can summon, ways of working both in organizations and in software architecture on top of that. And then finally on top of that you can put data analytics. Once you have all this stuff automated and flexing, you need a way to steer the ship, to figure out where you should go, and that's the point of the big science. To look for patterns in your organization once it's been digitized in this way.

[JAMES]: So one of the things, right, and maybe we should've talked about this more at the outset, but you guys at Silicon Valley Data Science, you are working with hundreds of companies, right? These are blue chip firms that you're helping to build up their Big Data strategy, and this is the pattern that you're seeing across them? Is that what we should understand?

[EDD]: Yeah, absolutely, when we go in and help them build or recommend, we realize that every organization needs to be building towards a data platform. That is to say, unifying its data infrastructure, making it easy to have that experimental approach. Because the thing is that we don't know what we don't know [laughs], so to speak. The Donald Rumsfeld unknown unknowns. Working with data is not the same as building a house, right? It's more like an archaeological dig where you're not entirely sure what you will find and what will be the prize thing that you'll exploit. It's more like software development in the product department sense half the time. And the way of working has to be different, and you certainly can't have large investments hanging off the end of each notion. We all know, in company settings, there's this two year pilot project which has just never died, it's a zombie system that's never quite delivered, because somebody's attached a professional, political status to keeping this thing going or they don't want to lose the budget. So at the same time we're talking about all the technical stuff we're talking about an organizational attitude as well that companies need to understand. That the nature of failure and to fail fast and fail often, and it's finding those good results is a positive contribution because it tells you where not to go, and as long as you're moving quickly, and haven't got too much invested in those dead ends, then it's a good thing.

[JAMES]: Yeah, and what you're really talking about, right, is rapidly iterating because the business circumstances are gonna change, and what you're trying doesn't necessarily align. And it seems like, and I actually had a conversation with one of your colleagues, and he put it really well. He said, look, data science is a probabilistic exercise. You can't just solve things in that left to right fashion that we would all love to, that makes it really easy to build applications. You have to try things and you have to be wrong before you're going to be able to get it right. And so what you build from a decision making engine in your company has to be able to support this experimentation. And it has to be able to support getting things wrong a lot and quickly, and then finding what works, and finally bringing it into production after the fact. And I thought that was just a really succinct way of sort of summing up the challenges that a lot of companies have.

[EDD]: Yeah, absolutely they do. And this is why some necessarily won't win. I said before, it can happen bottom up or top down, but when you're talking about changing the decision making abilities of a company as a whole, you have to have executives buy in. So one of the things we're particularly interested in is the job title of the chief data officer. And anytime a job title gets invented it's usually to fix something [laughs] that didn't really work too well before, or couldn't be possible in the current structure. And in many of the cases where a chief data officer exists, it's because it's an imperative to unify silos and departments. To unify business and IT, and bring everything together with a mandate from the top.

[JAMES]: Can you talk to us about where you think the future of that role's going?

[EDD]: Early on it's actually the regulated industries that had created this role, because even if your organization can't see its way to unifying silos to create business benefits, when the government comes asking for information you really have to be able to do it. Otherwise you're out of business. So there's a carrot and there's a stick, and unsurprisingly, you see a lot of sticks in the early business of this. I know of a lot of chief data officer roles in government settings, where again, there's an imperative to cut costs to provide better service to the citizens.

So yes, I think there is definitely situations where creating it puts a highlight on a thing that needs to happen. Now that's not to say, there's a number of organizations that have a chief digital officer. You see it a lot, especially in Europe, for instance, where the move to understand online, to make sense of apps, and the web, and so on, it makes sense to put all that in a sort of chief digital officer thing so the organization can learn the competency without having to change absolutely everything. And then it forwards, it spreads, it factors in in the same way that the rise of the influence of the CIO in the last 20 years has. But I think the key factor, whether we're talking digital or we're talking data, is that these executives are strategic allies to the business. They're not service centers, right? The worst kind of arrangement you can have is a siloed black box of analysts and data scientists who get requests shouted over the wall and then have to throw reports back over the wall later on. It's really interesting that that is not strategic alignment, that's just being a service center.

[JAMES]: [laughs] Really it all comes down to people in so many ways, right? If you want to have a transformational data science team, they need to be challenged by new work. They need to be constantly stimulated and be able to solve really difficult problems. That's why consulting firms, like Bain, McKinsey, BCG, they have the smartest people in the world because that type of work attracts really smart, talented people, and they want to be around that. And if you're locking people up in a black box just so you can march people around and say hey, look at my PhDs in the closet, that's just a recipe for disaster.

So what I want to sort of transition into is, we're talking about, effectively, how to organize your company to make better decisions, right? And be more agile, and that's sort of what you mentioned earlier that it all comes down to. There are a number of really exciting innovations that are augmenting our ability to do this. Particularly artificial intelligence, machine learning, the ability to automate just about anything. You track this closely. I'm curious, what are you seeing? Because this, to me, is an area of almost limitless potential.

[EDD]: There's certainly a lot of excitement around now with AI in the popular media and in general, starting off with IBM's Watson winning Jeopardy, and more recently the Google computer AlphaGo winning, I think, four out of five Go games against the champion. And this captured the imagination because we're starting to think, can computers act like humans? And of course, [laughs] the truth is equally exciting but a little more prosaic. So what, in particular, we have the ability to take advantage of now is machine learning frameworks that only the Googles and the Microsofts have been able to use hitherto. They're just making these things available, and we're starting to learn a lot more about the application of deep learning neural networks than in the past. So let me kind of zoom in on exactly the kind of thing that they are good for, and really it's a lot to do with recognition. One of the things that is very hard to have a computer do is to understand everything about, even in a relatively small domain. We've been trying to do this for years with things like expert systems and so on. And this is partly responsible for what they call the AI winter, right? Every time something AI works, they stop calling it AI. Although it never seems to really work all the way. There is the notion of the AI winter where we're all skeptical that it will never work. All of which is to say that there are some contacts there of recognition where good enough is good enough. And this is what is so exciting about machine learning stuff, that we can now make a machine iterate over hundreds of thousands of photographs and look for things that are likely to be people, or everybody wants it likely to be cats.

[JAMES]: [laughs] Yeah, and I think you bring up an interesting point, too. I remember hearing Facebook's AI chat bot, which is part human, part AI, every conversation that they have is just training the AI engine in the background to eventually replace the people. So that way it can just be a fully automated bot that will answer your questions, and so on and so forth. So it's a really interesting way that, in using the products, you're making them better, and training them.

So what I wanna do now is take a look at, you sort of made the case for AI, and what's happening in machine learning, and where we're going. But you're working with a whole bunch of companies. How are people actually implementing this in their Big Data? Tell our audience how this actually applies to their Big Data investments.

[EDD]: Yeah, well, I'll go to a simple one, and again, it is in a sense remarkably simple, but the business impact was large. We worked with edmunds.com, which is an online database of cars, sort of automobile details, and you go there and you research the car you're about to buy, and so on. And they clearly make a lot of their money through advertising as people research cars, and they can target those ads. Now it may not surprise you to know that the way that people get information from the automobile manufacturers is by that global standard of communication, the PDF. And what they had was rooms full of editors, basically, translating from the terms that were in the PDFs through to an ontology distro of a database on their website. Because of course, there is no universal language with which everybody talks about cars. In fact, every auto manufacturer calls features something slightly different in an attempt to sound cool and differentiate themselves in the marketplace. With all wheel drive, 4 wheel drive, xDrive, whatever you want to call it. And so you've kind of got this human recognition step that has to happen, say, ah, yeah, well I know what that is, it maps to this other thing that I know, it's 4 wheel drive on the car, let's check that off. And they can clearly read this off what is, essentially, paper documents. So one of the bits of work we did with them was not only to deploy image recognition over these PDFs to bring out the text, OCR actually, technology which has been around forever. But then also to use machine learning to do this mapping from its technology. And so you've still got humans, ultimately, managing the 5% of exceptions where things aren't quite known. So you're talking about a process that in the case of the prototypical auto, which is the Ford F-150, I think it has the most configurations of any auto in the world, where it could take two to three weeks to actually transcribe that, and they reduced it to just a day or so. And of course, for a company that's driven by advertising, when do people want to look up details of a car? Well, when it's released. So they're losing the most valuable window, essentially. So for them, it had a real business impact. It freed the editors up to do something higher skilled and more productive. So we're not taking away jobs, in this case. Actually we're helping people do work for which they're more suited than before. And also that time to market's much reduced. It's things like that, right? It these cognitive tasks that the people do. Think about the intelligence services, how often you might want to have people or security firms, they're looking for things that you couldn't recognize because your brain cannot. Clearly in video streams, again, using machine learning, it's a way of just cutting down a lot of the every day tasks and having humans manage the exceptions and situations where more intelligence or more discretion is required.

[JAMES]: It strikes me, the ability to just parse through video streams and just gigantic feeds of unstructured data, whether that's video or audio, and so on and so forth. That as an incredible time saver, and I think there's a lot of untapped data and insights, and I'm curious if you agree with that.

[EDD]: Yeah, there is. Every time we learn to do something. Scale is fun, right? Every time we learn to do something at scale it also has positive benefits when you don't have scale, which is why you should never think oh, that technology is great for Google and Facebook, but it isn't gonna be useful to me. Because it very often is. And one of the inaccuracies of the information age that we live in is that our ability to create information has far outstripped our ability to parse and understand it. I always have a very simple example of this, which is, imagine that you lived in the 1950s but still got the amount of incoming mail you do through your email today. It wouldn’t be hard, looking back, to say well, you'd probably have two personal assistants helping you sort through it and parse it. You'd have a staff, right? But right now we just have ourselves and we're dying under this wealth of information. And frankly there's not been enough care put into these products to help humans be human and not be framed clerks at a computer. So this is one of the things that makes me really, really happy about machine intelligence, is that we can now use all this power. The last ten years we've been spending our time having prettier graphics, and chrome finishes, and nice rounded corners instead of helping people really deal with information better. But now we can put this processing power to kind of righting that, having the computer do a lot more of the cognitive tasks and have the people focused on the right things. So as far as the two points you mentioned, there's a humanizing effect on the technology here, it just has to be better. We cannot sustain or cope with the amount of information we're able to create in this way much longer.

[JAMES]: Yeah, I totally agree. So okay, pull out your crystal ball. We talked about sort of the immediate applications, but let's take a look at two, three, five, ten years down the road. What's this intersection of Big Data and artificial intelligence, machine learning, look like? What are you excited about?

[EDD]: The thing I'm most excited about is the idea that I won't be staring at a black rectangle for my job in ten years' time.

[JAMES]: [laughs]

[EDD]: I genuinely hope we're not. This is one of those temporary affordances, like all of the things we've been doing with the Windows, the interface, and so on. Apps on our phones are just these temporary things that have made life better for a while but aren't fundamentally a very satisfying way of existing. So I'm looking for, I think, human-computer interaction to benefit most deeply from this. And very excited about the idea that you can talk to the computer. But that only works if you can talk to computers however you want. I don't know about you, but I find it ridiculous trying to find the right form of words to ask Siri to do exactly the right thing for me still at this stage. It's very exciting but it's of limited use, you kind of still need to learn how to talk like a robot in order to make a robot listen to you.

So looking to that stage where computers won't turn you into a computer. This is the biggest thing, you have to understand how a computer works in order to be able to work it. So if you could do this we'll actually understand how humans work a lot better, for some values of understanding, or at least appear to be understanding. So I think this is going to be the big change, that we'll cease to think of a computer at all in a sense, more that we're interacting with agents on top of that information fabric that really help us be a lot more efficient and take a lot of the needless busy work out of the way.

[JAMES]: So you hit on a lot of the topics that are gonna fundamentally change how we work. And of course, if you think about how we interact with data, when we can start asking questions in exactly the form of how we ask them, rather than needing to query things, and know about how computers work to structure our questions, and our queries, and our code to do it. I think that's a pretty amazing, fundamental shift that we have the opportunity to potentially achieve in the not-so-distant future.

[EDD]: Well I think so. I said before that the difficultly for AI has always been how problematic it is to encode domains. The big data sort of solves half of this problem, right? I was a huge fan of the semantic web, where you have a decentralized vocabulary where it doesn't matter that you called it an automobile and I called it a car. We could've established a rule that said these were the same things, and we could both carry on using our language and so reason with each others data. So the power of the semantic web is so enticing. The difficulty is, who writes all this code in the first place, right? Metadata is so hard. People still can't put title pages in HTML pages properly, and that's one piece of metadata. [laughs] It is very difficult. But essentially, machine learning is statistical method of allowing a computer to apply the metadata in a lot of scenarios through inference, right? Now we're in a world where, if you go into your Google Photos and type cat, pictures of cats will appear. Who did that? Well a computer did that, not a person. So all those folks that have dedicated tagging their albums and whatnot, for goodness knows how long, they don't have to bother, just plug everything into search. In other words, a lot of the encoding inside of domains can now be done by computer. So it becomes a lot more practical to imagine that lots of domains, and certainly over the next ten years, specialized domains, can be made available to AI so we will actually be able to have the scenario where we're talking to the computer and they know that when we are in vacation booking month. There's a particular language you use, and a particular number of concerns and things you want to talk about, and it won't be a spin the robot and it happens to cover all the bases. The computer will know what bases it needs to cover just like any great travel agent would.

[JAMES]: Yeah, it's some incredible, some exciting possibilities for the future. So Edd, as we wrap up here, I'm curious. You talked about a lot of interesting stuff, you guys do a lot of interesting work. For our audience out there, if they want to get in touch with you or find out more, where do they go and what should they check out?

[EDD]: I think the best place to go to for SVDS is our website, svds.com. If you want to catch up with me, I'm Edd, which is @edd on Twitter, and also email at edd@svds.com.

[JAMES]: Edd, it's been a pleasure having you on the show, really appreciate you taking the time here.

[EDD]: Thanks so much, it's been a great conversation.