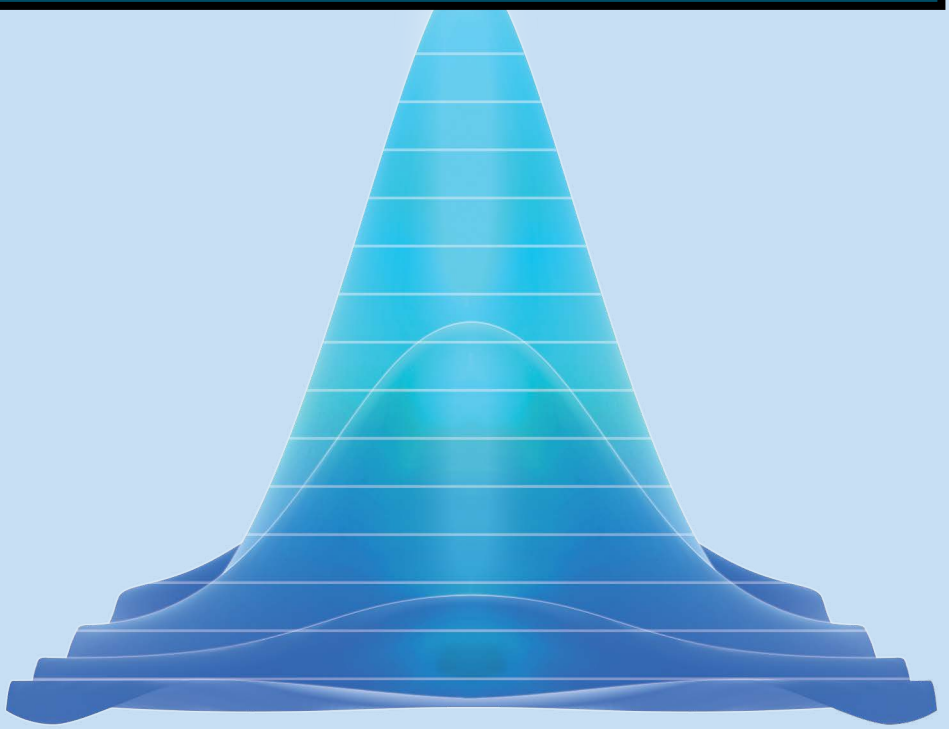
THE

# DATA SCIENCE HANDBOOK

**FREE SAMPLE CHAPTERS**



ADVICE AND INSIGHTS FROM

## 25 AMAZING DATA SCIENTISTS

FOREWORD BY JAKE **KLAMKA**

DJ **Patil,** Hilary **Mason,** Pete **Skomoroch,** Riley **Newman,** Jonathan **Goldman,** Michael **Hochster,** George **Roumeliotis,** Kevin **Novak,** Jace **Kohlmeier,** Chris **Moody,** Erich **Owens,** Luis **Sanchez,** Eithon **Cadag,** Sean **Gourley,** Clare **Corthell,** Diane **Wu,** Joe **Blitzstein,** Josh **Wills,** Bradley **Voytek,** Michelangelo **D'Agostino,** Mike **Dewar,** Kunal **Punera,** William **Chen,** John **Foreman,** Drew **Conway**

BY CARL **SHAN** HENRY **WANG** WILLIAM **CHEN** MAX **SONG**

# DJ PATIL VP of Products at RelateIQ

## *The Importance of Taking Chances and Giving Back*

*DJ Patil is co-coiner of the term 'Data Scientist' and co-author of the Harvard Business Review article: "Data Scientist: Sexiest Job of the 21st Century."*

*Fascinated by math at an early age, DJ completed a B.A. in Mathematics at University of California, San Diego and a PhD in Applied Mathematics at University of Maryland where he studied nonlinear dynamics, chaos theory, and complexity. Before joining the tech world, he did nearly a decade of research in meteorology, and consulted for the Department of Defense and Department of Energy. During his tech career, DJ has worked at eBay as a Principal Architect and Research Scientist, and at LinkedIn as Head of Data Products, where he co-coined the term "Data Scientist" with Jeff Hammerbacher and built one of the premier data science teams. He is now VP of Product at RelateIQ, a next generation, data-driven customer relationship management (CRM) software. Most recently RelateIQ was acquired by Salesforce.com for its novel data science technology.*

*In his interview, DJ talks about the importance of taking chances, seeking accelerations in learning, working on teams, rekindling curiosity, and giving back to the community that invests in you.*

*Additional Reading:*
*1. Commencement Speech on the Importance of Failure*
*2. Harvard Business Review: Sexist Job of the 21st Century*

**DSH:** Something that touched a lot of people from your presentations is your speech on failure. It's surprising to see someone as accomplished as yourself talk about failure. Can you tell us a bit more about that?

**DJ:** Something most people struggle with when starting their career is how they enter the job market correctly. The first role you have places you in a "box" that other people use to infer what skills you have. If you enter as a salesperson you're into sales, if you enter as a media person you're into media, if you enter as a product person you're into products etc. Certain boxes make more sense to transition in or out of than other ones.

The academic box is a tough one because automatically, by definition, you're an academic. The question is: Where do you go from there? How do you jump into a different

box? I think we have a challenge that people and organizations like to hire others like themselves. For example, at Ayasdi (a topological machine learning company) there's a disproportionate amount of mathematicians and a surprising number of topologists.

For most people who come from academia, the first step is that someone has to take a risk on you. Expect that you're going to have to talk to lots and lots of people. It took me 6 months before eBay took a chance on me. Nobody just discovers you at a cafe and says "Hey, by the way you're writing on that piece of napkin, you must be smart!" That's not how it works, you must put yourself in positions where somebody can actually take a risk on you, before they can give you that opportunity.

*Nobody just discovers you at a cafe and says "Hey, by the way you're writing on that piece of napkin, you must be smart!" That's not how it works, you must put yourself in positions where somebody can actually take a risk on you, before they can give you that opportunity.*

And to do that, you must have failed many times, to the point where some people are not willing to take a risk on you. You don't get your lucky break without seeing a lot of people slamming doors in your face. Also, it's not like the way that you describe yourself is staying the same; your description is changing and evolving every time you talk to someone. You are doing data science in that way. You're iterating on how you are presenting yourself and you're trying to figure out what works.

Finally someone takes a chance on you, but once you've found somebody, the question is how do you set yourself up for success once you get in? I think one of the great things about data science is it's ambiguous enough now, so that a lot of people with extra training fit the mold naturally. People say, "Hey, sure you can be a data scientist! Maybe your coding isn't software engineering quality coding, but your ability to learn about a problem and apply these other tools is fantastic."

Nobody in the company actually knows what these tools are supposed to be, so you get to figure it out. It gives you latitude. The book isn't written yet, so it's really exciting.

**DSH:** What would you suggest as the first step to putting yourself out there and figuring out what one should know? How does one first demonstrate one's value?

**DJ:** It first starts by proving you can do something, that you can make something.

I tell every graduate student to do the following exercise: when I was a grad student I went around to my whole department and said, "I want to be a mathematician. When I say

the word mathematician, what does that mean to you? What must every mathematician know?"

I did it, and the answers I got were all different. What the hell was I supposed to do? No one had a clear definition of what a mathematician is! But I thought, there must be some underlying basis. Of course, there's a common denominator that many people came from. I said, okay, there seem to be about three or four different segmentations. The segmentation I thought was the most important was the segmentation that gave you the best optionality to change if it ended up being a bad idea.

As a result of that, I took a lot of differential equations classes, and a bunch of probability classes, even though that wasn't my thing. I audited classes, I knew how to code, I was learning a lot about physics — I did everything I could that was going to translate to something that I could do more broadly.

Many people who come out of academia are very one-dimensional. They haven't proven that they can make anything, all they've proven is that they can study something that nobody (except maybe their advisor and their advisor's past two students) cares about. That's a mistake in my opinion. During that time, you can solve that hard PhD caliber problem AND develop other skills.

*It first starts by proving you can do something, that you can make something.*

For example, aside from your time in the lab, you can be out interacting with people, going to lectures that add value, attending hackathons, learning how to build things. It's the same reason that we don't tell someone, "First, you have to do research and then you learn to give a talk." These things happen together. One amplifies the other.

So my argument is that people right now don't know how to make things. And once you make it, you must also be able to tell the story, to create a narrative around why you made it.

With that comes the other thing that most academics are not good at. They like to tell you, rather than listen to you, so they don't actually listen to the problem. In academia, the first thing you do is sit at your desk and then close the door. There's no door anywhere in Silicon Valley; you're out on the open floor. These people are very much culture shocked when people tell them, "No you must be working, collaborating, engaging, fighting, debating, rather than hiding behind the desk and the door."

I think that's just lacking in the training, and where academia fails people. They don't

get a chance to work in teams; they don't work in groups.

Undergrad education, however is undergoing some radical transformations. We're seeing that shift if you just compare the amount of hackathons, collaboration, team projects that exist today versus a few years ago. It's really about getting people trained and ready for the work force. The Masters students do some of that as well but the PhDs do not. I think it's because many academics are interested in training replicas of themselves rather than doing what's right for society and giving people the optionality as individuals to make choices.

**DSH:** How does collaboration change from academic graduate programs to working in industry?

**DJ:** People make a mistake by forgetting that data science is a team sport. People might point to people like me or Hammerbacher or Hilary or Peter Norvig and they say, oh look at these people! It's false, it's totally false, there's not one single data scientist that does it all on their own. data science is a team sport, somebody has to bring the data together, somebody has to move it, someone needs to analyse it, someone needs to be there to bounce ideas around.

*People make a mistake by forgetting that data science is a team sport.*

Jeff couldn't have done this without the rest of the infrastructure team at Facebook, the team he helped put together. There are dozens and dozens of people that I could not have done it without, and that's true for everyone! Because it's a bit like academia, people see data scientists as solo hunters. That's a false representation, largely because of media and the way things get interpreted.

**DSH:** Do you think there's going to be this evolution of people in data science who work for a few years, then take those skills and then apply them to all sorts of different problem domains, like in civics, education and health care?

**DJ:** I think it's the beginning of a trend. I hope it becomes one. Datakind is one of the first examples of that, and so is data science for Social Good. One of the ones that's personally close to my heart is something called Crisis Text Line. It comes out of DoSomething.org — they started this really clever texting campaign as a suicide prevention hotline and the result is we started getting these text messages that were just heart wrenching.

There were calls that said "I've been raped by my father," "I'm going to cut myself," "I'm going to take pills," really just tragic stuff. Most teens nowadays do not interact by voice - calling is tough but texting is easy. The amount of information that is going back and

forth between people who need help and people who can provide help through Crisis Text Line is astonishing.

How do we do it? How does it happen? There are some very clever data scientists there who are drawn to working on this because of its mission, which is to help teens in crisis. There's a bunch of technology that is allowing us to do things that couldn't be done five, six years ago because you'd need this big heavyweight technology that cost a lot of money. Today, you can just spin up your favorite technology stack and get going.

These guys are doing phenomenal work. They are literally saving lives. The sophistication that I see from such a small organization in terms of their dashboards rivals some of the much bigger, well-funded types of places. This is because they're good at it. They have access to the technology, they have the brain power. We have people jumping in who want to help, and we're seeing this as not just a data science thing but as a generational thing where all technologists are willing to help each other as long as it's for a great mission.

Jennifer Aaker just wrote about this in a *New York Times* op-ed piece — that the millennial generation is much more mission driven. What defines happiness for them is the ability to help others. I think that there is a fundamental shift happening. In my generation it's ruled by empathy. In your generation, it's about compassion. The difference between empathy and compassion is big. Empathy is understanding the pain. Compassion is about taking away the pain away from others, it's about solving the problem. That small subtle shift is the difference between a data scientist that can tell you what the graph is doing versus telling you what action you need to do from the insight. That's a force multiplier by definition.

**DSH:** Compassion is also critical for designing beautiful and intuitive products, by solving the pain of the user. Is that how you chose to work in product, as the embodiment of data?

**DJ:** I think the first thing that people don't recognize is that there are a number of people who have started very hard things who also have very deep technical backgrounds.

Take Fry's Electronics for example. John Fry, the founder, is a mathematician. He built a whole castle for one of the mathematical associations out in Morgan Hill, that's how much of patron of the arts he is for them. Then you can look at Reed Hastings of Netflix, he's a mathematician. My father and his generation, all of the old Silicon Valley crew were all hardcore scientists. I think it just goes on to show - you look in these odd places and you see things you would not have guessed.

I think there's two roles that have been interesting to me in companies: the first is you're starting something from scratch and the second is you're in product. Why those two roles? If you start the company you're in product by definition, and if you're in product you're making. It's about physically making something. Then the question is, how do you make? There's a lot of ways and weapons you can use to your advantage. People say there is market assessment, you can do this detailed market assessment, you can identify a gap in the market right there and hit it.

There's marketing products, where you build something and put a lot of whizbang marketing, and the marketing does phenomenally. There are engineering products which are just wow — you can say this is just so well engineered, this is phenomenal, nobody can understand it, but it's great, pure, raw engineering. There is designing products, creating something beautifully. And then, there's data.

The type of person I like best is the one who has two strong suits in these domains, not just one. Mine, personally, are user experience (UX) and data. Why user experience and data? Most people say you have to be one or the other, and that didn't make sense to me because the best ways to solve data problems are often with UX. Sometimes, you can be very clever with a UX problem by surfacing data in a very unique way.

*Because of the pace at which the world changes, the only way to prepare yourself is by having that dynamic range.*

For example, People You May Know (a viral feature at LinkedIn that connected the social graph between professionals) solved a design problem through data. You would join the site, and it would recommend people to you as you onboard on the website. But People You May Know feels creepy if the results are too good, even it it was just a natural result of an algorithm called triangle closing. They'd ask, "How do you know that? I just met this person!" To fix this, you could say something like "You both know Jake." Then it's obvious. It's a very simplistic design element that fixes the data problem. My belief is that by bringing any two elements together, it's no longer a world of one.

Another way to say this is, how do you create versatility? How do you make people with dynamic range, which is the ability to be useful in many different contexts? The assumption is our careers are naturally changing at a faster rate than we've ever seen them change before. Look at the pace at which things are being disrupted. It's astonishing. When I first got here eBay was the crazy place to be and now they're on a turnaround. Yahoo went from being the mammoth place to now attempting a turnaround. We've had companies that just totally disappeared.

I see a spectrum of billion dollar companies coming and going. We're seeing something very radical happening. Think about Microsoft. Who wouldn't have killed for a role in Microsoft ten years ago? It was a no brainer. But not anymore.

Because of the pace at which the world changes, the only way to prepare yourself is by having that dynamic range. I think what we're realizing also is that different things give you different elements of dynamic range. Right now data is one of those because it's so scarce. People are getting the fact that this is happening. It gives a disproportionate advantage to those who are data savvy.

**DSH:** You mentioned earlier that when you were looking to become a mathematician you picked a path that optimized for optionality. As a data scientist, what type of skills should one be building to expand or broaden their versatility?

**DJ:** I think what data gives you is a unique excuse to interact with many different functions of a business. As a result, you tend to be more in the center and that means you get to understand what lots of different functions are, what other people do, how you can interact with them. In other words, you're constantly in the fight rather than being relegated to the bench. So you get a lot of time on the field. That's what changes things.

*One of the first things I tell new data scientists when they get into the organization is that they better be the first ones in the building and the last ones out.*

The part here I think people often miss is that they don't know how much work this is. Take an example from RelateIQ. I'm in the product role (although they say I'm supposed to be the head of product here, I think of these things as team sports and that we're all in it together), and I work over a hundred hours a week easily. If I had more time I'd go for longer hours. I think one of the things that people don't recognize is how much net time you just have to put in. It doesn't matter how old you are or how good you are, you have to put in your time.

You're not putting in your time because of some mythical ten thousand hours thing (I don't buy that argument at all, I think it's false because it assumes linear serial learning rather than parallelized learning that accelerates). You put in your time because you can learn a lot more about disparate things that fit into the puzzle together. It's like a stew, it only becomes good if it's been simmering for long time.

One of the first things I tell new data scientists when they get into the organization is that they better be the first ones in the building and the last ones out. If that means four hours of sleep, get used to it. It's going to be that way for the first six months, probably a year plus.

That's how you accelerate on the learning curve. Once you get in there, you're in the conversations. You want to be in those conversations where people are suffering at two in the morning. You're worn down. They are worn down. All your emotional barriers come down and now you're really bonding. There's a reason they put Navy Seals through training hell. They don't put them in hell during their first firefight. You go into a firefight completely unprepared and you die. You make them bond before the firefight so you can rely on each other and increase their probability of survival in the firefight. It's not about bonding during the firefight, it's about bonding before.

That's what I would say about the people you talked to at any of the good data places. They've been working 10x harder than most places, because it is do or die. As a result, they have learned through many iterations. That's what makes them good.

**DSH:** What can you do on a day-to-day basis that can make you a good data scientist?

**DJ:** I don't think we know. I don't think we have enough data on it. I don't think there's enough clarity on what works well and what doesn't work well. I think you can definitely say some things increase the probability of personal success. That's not just about data science, it's about listening hard, being a good team player, picking up trash, making sure balls don't get dropped, taking things off people's plates, being there for the team rather than as an individual, and focusing on delivering value for somebody or something.

*If you watch kids running around a track, and the parents want to leave, the kids always answer, "One more! One more!" You watch an adult run laps, and they are thinking, "How many more do I have to do?"*

When you do that, you have a customer (could be internal, external, anybody). I think that's what gives you the lift. Besides the usual skills, the other thing that's really important is the ability to make, storytell, and create narratives. Also, never losing the feeling of passion and curiosity.

I think people that go into academia early, go in with passion. You know that moment when you hear a lecture about something, and you're saying, "Wow! That was mind blowing!" That moment on campus when you're saying, "Holy crap, I never saw it coming." Why do we lose that?

Here is a similar analogy. If you watch kids running around a track, and the parents want to leave, the kids always answer, "One more! One more!" You watch an adult run laps, and they are thinking, "How many more do I have to do?" You count down the minutes to the workout, instead of saying, "Wow, that was awesome!"

I feel that once you flip from one to the other you've lost something inherently. You have to really fight hard to fill your day with things that are going to invigorate you on those fronts. One more conversation, one more fight, one more thing. When you find those environments, that's rare. When you're around people who are constantly inspiring you with tidbits of information, I feel like that's when you're lucky.

**DSH:** Is all learning the same? What value can you bring as a young data scientist to people who have more knowledge than yourself?

**DJ:** There's a difference between knowledge and wisdom. I think that's one of the classic challenges with academia. You can take a high school kid who can build an app better than a person with a doctorate who works in algorithms, and it's because of their knowledge of the app ecosystem. Wisdom also goes the other way: if you're working on a very hard academic problem, you can look at it and say, "That's going to be $O(n^2)$".

*I'm a firm believer in the apprentice model*

I was very fortunate when I was at eBay, as I happened to get inserted in a team where there was a lot of wisdom. Even though eBay was moving very slowly in things we were doing, I was around a lot of people who had a disproportionate amount of wisdom, so I was the stupidest guy with the least amount of tours of duty. But at the same time, I was able to add value because I saw things in ways that they had never seen. So we had to figure out where that wisdom aligned and where it didn't.

The other side of that was at LinkedIn, when you're on that exponential curve trajectory with a company. People say, "Well you were only at the company for three plus years," but I happened to be there when it grew from couple hundred to a couple thousand people. Being in a place where you see that crazy trajectory is what gives you wisdom, and that's the type of thing that I think compounds massively.

**DSH:** Many young people today are confronted with this problem related to knowledge and wisdom. They have to decide: Do they do what they're deeply passionate about in the field they care most about? Or do they do the route that provides them with the most immediate amount of growth? Do they go compound the knowledge of skills, or do they build wisdom in that domain?

**DJ:** It's a good and classic conundrum. I've gone with it as a non-linear approach: you go where the world takes you. The way I think about it is, wherever you go, make sure you're around the best people in the world.

I'm a firm believer in the apprentice model, I was very fortunate that I got to train with

people like James Yorke who coined with the term "chaos theory." I was around Sergey Brin's dad. I was around some really amazing people and their conversations are some of the most critical pieces of input in my life, I think I feel very grateful and fortunate to be around these people. Being around people like Reid Hoffman, Jeff Weiner is what makes you good and that gives you wisdom.

So for that tradeoff, if you're going to be around somebody that's phenomenal at Google, great! If you're going to be around someone super phenomenal in the education system, great! Just make sure whatever you are doing, you're accelerating massively. The derivative of your momentum better be changing fast in the positive direction. It's all about derivatives.

**DSH:** What do you think about risk taking, and defining oneself?

**DJ:** Everyone needs  to chart their own destiny. The only I thing I think is for certain is that as an individual, you get to ask the questions, and by asking the questions and interpreting the answers, you decide the narrative that is appropriate for you. If the narrative is wrong, it's your narrative to change. If you don't like what you're doing, you get to change it.

*If the narrative is wrong, it's your narrative to change. If you don't like what you're doing, you get to change it.*

It may be ugly, maybe hard or painful but the best thing is when you're younger, you get to take crazy swings at bats that you don't get to take later on. I couldn't do half the stuff I was doing before, and I'm very envious of people who get to. And that's a part of life, there's the flip side of when you do have family, or responsibilities, that you're paying for that next generation. Your parents put a lot on the line to try to stay in a town with great schools, and they may not have taken the risk that they would've normally taken to do these things.

That's part of the angle by which you play. It's also the angle which is the difference between what it means as an individual and team player. Sometimes you can't do the things that you want to do. It's one of the reasons I've become less technical. Take someone like Monica Rogati or Peter Skomoroch, two amazing data scientists and engineers at LinkedIn. What's a better use of my time? Taking a road block out of their way or me spending time debugging or coding something on my own?

In the role I have, in the position and what was expected of me, my job was to remove hurdles from people, my job was to construct the narrative to give other people runway to execute, their job was to execute and they did a hell of a good job at it.

**DSH:** You have talked about your research as a way to give back to the public that invested in you. Is there an aspect of the world that you feel like could really use the talent and skills of data scientists to improve it for the better?

**DJ:** I think we're starting to see elements of it. The Crisis Text Line is a huge one. That's why I put a lot of my time and energy into that one. But there are so many others: national security, basic education, government, Code for America. I think about our environment, understanding weather, understanding those elements, I would love to see us tackle harder problems there.

*Only work on simple things; simple things become hard, hard things become intractable.*

It's hard to figure out how you can get involved in these things, they make it intentionally closed off. And that's one of the cool things about data, it is a vehicle to open things up. I fell into working on weather because the data was available and I said to myself, "I can do this!" As a result, you could say I was being a data scientist very early on by downloading all this crazy data and taking over the computers in the department. The data allowed me to become an expert in the weather, not because I spent years studying it, because I was playing around and that gave me the motivation to spend years studying it.

**DSH:** From rekindling curiosity, to exploring data, to exploring available venues, it seems like a common thread in your life is about maximizing your exposure to different opportunities. How do you choose what happens next?

**DJ:** You go where the barrier of entry is low. I don't like working on things where it's hard. My PhD advisor gave me a great lesson — he said only work on simple things; simple things become hard, hard things become intractable.

**DSH:** So work on simple things?

**DJ:** Just simple things.

# CLARE CORTHELL Data Scientist at Mattermark

## *Creating Your Own Data Science Curriculum*

*After graduating from Stanford, Clare Corthell embarked on a self-crafted journey to acquire the knowledge and skills to understand and analyze macro-behavioral trends. One thing led to another, and her collection of resources turned into the Open Source Data Science Masters - a curriculum of online courses, books and other resources that one could use to learn the mathematical and programming foundation crucial to a data scientist.*

*Clare took a risky move by crafting her own degree program, outside of traditional educational institutions. She faced skepticism of a self-taught individual in a job that is typically inhabited by PhDs, but also found a community of supportive colleagues.*

*Overcoming these challenges, Clare completed her Open Source Data Science Masters and found herself as a data scientist at Mattermark, a venture-backed data startup working with large datasets to help professional investors quantify and discover signals of potentially high-growth companies.*

*Additional Reading:*
*1. Open Source Data Science Masters*

**DSH:** What was your background, before you began the Open Source Data Science Masters and before your role at Mattermark?

**Clare:** I'm a product person and an entrepreneur. I fell in love with startups long before I attended Stanford, where I designed a degree in a then-obscure program called *Science, Technology & Society*. You get to marry two engineering tracks, so I ended up designing a degree in product design and digital development, which then got me started working on product with early stage companies.

Before the OSDSM (Open Source Data Science Masters), I was designing and prototyping products for an early stage education technology company in Germany. Designing from user anecdotes alone became difficult when you only pull from anecdotes, so I started digging deeper into analytics and customer profiling. I started thinking about observing meta-trends among users instead of studying their behavior with a clipboard from behind a one-way window. What if I just ran several tests on two different prototypes? Then we would have data to tell us which one to develop! But as with many European startups, the company didn't get funded, so I had a few weeks to think about how this

new perspective fit in. On a long layover in Barcelona, I ordered an espresso and wrote down the technical skills I would need to dissect meta-trends and understand user data. That list laid out 6 months of full-time work, after which I'd really be able to do some damage. This became the Open Source Data Science Masters.

*I started thinking about observing meta-trends among users instead of studying their behavior with a clipboard from behind a one-way window. What if I just ran several tests on two different prototypes? Then we would have data to tell us which one to develop!*

As with any story, it is now retrospectively clear that I would secretly fall in love with an applied statistics class I cheekily called "Exceltastic." We worked with Bayes' Theorem and Markov Chains in the business context, figuring out things like how many cars can pass through two toll booths per hour. Everyone else sulked and moaned through munging spreadsheets while I harbored a dirty secret: I loved Excel models! Even so, I didn't know when my toll booth throughput calculations would be demanded of me, nor what class logically comes next. It took getting into industry to shed light on the value of keeping metrics. Things like my Exceltastic class don't seem to fit into an overarching puzzle, but we believe they shape our path. That's the power of confirmation bias. One of my favorite designers has this phrase that he prints in various media: "Everything I do always comes back to me." I've always found that fitting.

**DSH:** What is the Open Source Data Science Masters? What does its curriculum look like?

**Clare:** It's a collection of open-source resources that help a programmer acquire the skills necessary to be a competent entry-level data scientist. The first version included introductory linear algebra, statistics, databases, algorithms, graph analysis, data mining, natural language processing, and machine learning. I wrote the curriculum for myself, then I realized that people all over the internet were asking for it, so I published it on GitHub.

In August, I opened the curriculum for pull requests on GitHub. Without feedback it's difficult to know whether you've covered the right things. Further, it was an effort to get feedback on the idea of an institution-free degree, a kind of home-school for advanced degrees. The internet was astonishingly supportive and excited — and that excitement is addictive. It makes you want to be more transparent, and to become part of other peoples' wonder in learning new things.

**DSH:** How did you get started with the Open Source Data Science Masters?

**Clare:** I knew that a traditional Masters program would take at least the next three years of my life, but even more importantly it wouldn't focus on what is core to the profession I wanted to enter. I knew what I wanted and I was willing to take the risk of a non-institution education.

*It took getting into industry to shed light on the value of keeping metrics. Things like my Exceltastic class don't seem to fit into an overarching puzzle, but we believe they shape our path.*

I set out for the curriculum to take 6 months to complete (March - August 2013), with a small project at the end and various programming mini-projects focusing on scraping, modeling, and analysis. It was amazing how difficult it was to manage myself. School gives you this structure that you don't have to question or design, which you don't really see until you have to manage your own curriculum and deadlines. There's a lot of product management that goes into an educational track like the OSDSM. I'm grateful to all the people who supported me and helped me throughout, even if they didn't quite understand the strange and uncharted waters I was braving to get there.

**DSH:** How did you find the resources?

**Clare:** I reverse-engineered most of it from job descriptions that interested me. This meant companies I believed would grow quickly and provide the most opportunity: mid-stage startups, 100-200 people, existing data science teams and reverence for the methodology. I didn't want to be the lone wolf and knew I needed mentorship.

People tend to frown on centering the goals of the classroom on applicability in the real world, but a classic liberal educational approach in a technical career pivot won't serve you. This is a technical vocational degree, so the goal was very concrete. I should be employable and employed on a data science (or Analytics Engineering) team after completing the curriculum.

There was another realization that coalesced very quickly: the act of designing from insights of single users does not scale. I was also hankering for something more technically and algorithmically challenging. I'd bought this book before I moved to Germany, *Programming Collective Intelligence*. I just bought it, I really had no reason to. When I first opened it up, I understood next to nothing. But I carried it with me in Germany, and every time I opened it, something new jumped out and I understood more about scaling user insight. The book became my cornerstone, how I measured my progress. It's a bible for Data Scientists.

I also used the following resources/websites:

- **Quora**: This is a great resource for the Valley — it's truly navel-gazing, but if that's what you're doing, it's useful. People like DJ have answered questions about what a Data Scientist does on a daily basis. You can start to discern the technical capacities that are required of you, mathematical foundations that are necessary, and so forth.
- **Blogs:** Zipfian Academy, a data science bootcamp, had a blog. They had a great post on the resources they saw as core to becoming a data scientist: A Practical Intro to Data Science
- **Coursera:** I'm Coursera's biggest fanboy. They're part of this quietly-brewing educational revolution, which will soon be less quiet. My story is a tremor before the earthquake, I'm just waiting for the ground to start shaking.

**DSH:** How much math (probability, statistics, ML) did you try to learn? How much math do you think a data scientist needs to know?

**Clare:** You don't have to know everything. That's why I've tried to keep the curriculum so tightly focused on its goal. Programmers are great at "just-in-time" learning because it's impossible to know everything. That's a great trait. If you have a core set of competencies and understand how to "debug" problems and learn what you need to solve them, you can do damage. And naturally, you improve over time by recognizing new problems as chunks of old problems you've already seen and solved.

*The internet was astonishingly supportive and excited — and that excitement is addictive. It makes you want to be more transparent, and to become part of other peoples' wonder in learning new things.*

So much of this curriculum is abstract, and that's where people get scared. People are scared of math because it's not applied in our education system. But those scary elements of math and abstraction diminish with concrete examples and conversations with others. I had a few phone-a-friend lifelines, and I ate up Khan Academy and Coursera videos. There's something magical about how much more communicative spoken English can be, especially when you can rewind and digest a concept for the second, third, or even fourth time. You can always talk through a problem with someone else, even if they're not an expert. Talking through things is synonymous with debugging. One of my mentors calls this "the rubber ducky method," because if you talk a problem through to a plastic duck, sometimes you start to find the holes in your assumptions. Then you can plug them up.

If you think about people as having different levels of competency in these different realms, it doesn't take long to understand that working as a team allows you to stack

your respective skills on top of one another. Having specialties among the team is really essential to getting things done in a small organization. I was lucky enough to join a company where I get mentorship in verticals where I'm middling or even an amateur. It's amazing to learn with other people. Finding a job where you have mentors and training is essential to continue to grow and improve. And if you're not improving and growing, you're dead in the water. So that's a long-winded way of saying: Working with other people is essential to working with more complex concepts and systems. Rome wasn't built by some guy, and probably not at a weekend hackathon.

**DSH:** What would you do differently if you could redo the Masters?

**Clare:** As Patient Zero of a new type of internet-based institution-free education, I didn't know what to expect. It was impossible to know how I would be judged and whether I would benefit from my experiment. This type of ambiguity usually makes people extremely uncomfortable. It's like leaving a six-year-old in the library by herself instead of putting her in class with a teacher. What is she going to do? Pull a bunch of books onto the floor and see how high she can stack them? Watch birds at the window and think about how wings work? Or is she going to find something interesting and gather books that will help her form her own ideas about the world?

*You don't have to know everything. That's why I've tried to keep the curriculum so tightly focused on its goal. Programmers are great at "just-in-time" learning because it's impossible to know everything.*

I knew that it would be a risk, but I took a leap of faith and left myself alone in the library. In the end, the greatest reward didn't come from the curriculum, it came from what taking a risk demonstrated about me. It led me to a tribe that respected the risk I had taken, and valued the grit that it required to follow through. Many people were displeased that I let myself into the library without an adult. But I'm not interested in taking the recommended path and clinging to a librarian. I have no interest in small ambition.

**DSH:** What's the difference between data science job descriptions & day-to-day role at Mattermark?

**Clare:** Our CEO Danielle was once asked how many data scientists we have at Mattermark. We're all data scientists, she thought — we all use, manipulate, and analyze data on a daily basis to make our customers happier and more profitable. We even all write SQL! That's not something you see every day at a company, but it's essential when you're building and selling a data product. I build products as an engineer, anything from fitting

clustering algorithms, building automated analyses, designing UIs, acquiring new data — it's a startup. It's all hands on deck.

It's not clear that data science is a job title to stay yet. For example, do we know if growth hacking is a subset of data science? We don't. There will always be a top-level salary for a person who can turn chaos into insights. That won't change. Data Scientist is a title we'll continue to use while we figure it out.

**DSH:** What could someone in school, or otherwise without too much background in industry learn from your experience?

**Clare:** The ability to evolve my own career with a self-designed curriculum begins to outline the immense cracks in the foundation of higher education*. The deconstruction of this system was very long in coming, but it's happening now. The lesson is the following: if you take initiative and acquire skills that increment your value, the market is able and willing to reward you.

*The ability to evolve my own career with a self-designed curriculum begins to outline the immense cracks in the foundation of higher education*

Though people continue to believe and espouse old patterns of education and success, these patterns do not represent requirements or insurance. The lack of any stamp of approval is a false barrier. There are no rules.

It's important to understand the behavior of the market and institutions with regard to your career. When breaking out of the patterns of success, know that people will judge you differently than others who have followed the rules.

There are two very discrete things that I learned: The market is requiring people to perform tryouts for jobs instead of interviews, and most companies don't hire for your potential future value.

**Tryouts as Interviews:** The economy has set a very high bar for people coming into a new profession. Job descriptions always describe a requirement for previous experience, which is paradoxical because you need experience to get it. Don't let that scare you, not for a minute. Pull on your bootstraps and get in the door by giving yourself that experience — design and execute on a project that demonstrates your ability to self-lead. Demonstrate that you can take an undefined problem and design a solution. It will give you the confidence, the skills, and the background to merit everything from the first interview to the salary you negotiate.

Even more concretely, work with a non-profit organization (or another organization that doesn't have the economic power to hire programmers or data scientists) to create a project that is meaningful for the organization and also shows off your skills. It's a great way to do demonstrative and meaningful work while also aiding an organization that could use your help, and likely has problems people are paying attention to solving. Win-win.

**Current Value vs Potential:** Look for companies that will hire you for your potential. It's important to be upfront about your grit, self-sufficiency, and ability to hit the ground running. Luckily, with disciplines like data science, the market is on your side. Sometimes companies can spring for a Junior Data Scientist and invest in your growth, which is really what you wanted from the beginning.

*Talk with people who can recognize hustle and grit, and not necessarily those who are looking to match a pattern drawn from your previous experience.*

Everyone will tell you this, but I work on product so I'll underline it even more strongly: Learn to write production-level code. The more technical you are, the more valuable you are. Being able to write production code makes you imminently hirable and trainable.

*[*NB: Don't think for a minute that I don't believe in the tenets of a true liberal education - quite the contrary. I continue to read philosophy and history, in part because we cannot draw fully upon the knowledge of man without doing so. These are essential elements to being a purposed, ethical, and effective person - but they don't directly accelerate a career. The true liberal education has nothing to do with market forces, and never should. Higher Education as it exists today and Liberal Education should be held as wholly uniquely-motivated institutions.]*

**DSH:** How was your self-taught path to becoming a data scientist received by company recruiters? What advice would you share with entrepreneurial individuals who are interested in the field?

**Clare:** Talk with people who can recognize hustle and grit, and not necessarily those who are looking to match a pattern drawn from your previous experience. Often, these kinds of people run startups.

Recruiters gave me a very real response: They didn't see my course of self-study as legitimate. It's hard to give yourself a stamp of approval and be taken seriously. I wouldn't recommend that just anyone do what I did — it will take a while for autodidactism to

become more accepted, and maybe it will never be a primary pattern. But maybe people like me can help expose this as a viable way to advance professionally. I know that great companies like Coursera will continue to innovate on these new forms of education, keep quality high, and democratize access.

**tl;dr**

If you want to get to the next level, wherever your next level may be, it's possible to pave your own road that leads you there. It's a monstrously tough road, but it's your road.

# MICHELANGELO D'AGOSTINO

## Senior Data Scientist at Civis Analytics

## *The Election as Physical Science*

*Prior to working in data science, Michelangelo was an undergrad at Harvard in physics. He finished his PhD in astrophysics from Berkeley, and developed a love of working collaboratively on hard problems with other people while analyzing neutrino data for the IceCube experiment.*

*Michelangelo started his data science journey as a senior analyst in digital analytics with the 2012 Obama re-election campaign, where he assisted in optimizing the campaign's email fundraising efforts and analyzed social media data. Afterwards, he worked as Lead Data Scientist at Braintree before he rejoined many of his former colleagues from the Obama reelection team at Civis Analytics. At Civis Analytics, Michelangelo works on statistical models and writes software for data analysis.*

*Michelangelo has travelled to the South Pole and has written about science and technology for* The Economist.

*In his interview, Michelangelo shares his story and offers practical advice on transitioning from a PhD into data science. He also shares his thoughts on data science for social good.*

*Additional Reading:*
*1. Physics Today: A physicist reshapes his career*
*2. Quora: How do I apply data science for social good*

**DSH:** Can you talk about your career from undergrad to PhD? How did you transition to data science and data analysis, and what were your various roles afterwards?

**Michelangelo:** My career has been strange. Sometimes, it feels like a random walk, but it's more of a greedy algorithm because every time I've had a choice of what to do, I think about what seems like the best opportunity, the most interesting thing for me to do. It's worked out really well even though there hasn't been this overarching plan.

I started as a Harvard undergrad and studied physics. I always really loved physics, but I also really loved doing other things outside of physics. So, I took tons of literature and history classes when I was an undergrad. I liked working in the lab and doing the research stuff, but I always had lots of different interests.

I graduated, and I wasn't sure if I wanted to go to grad school because I wanted to get a job. Looking back on it now, I wish data science existed when I was an undergrad. I really loved quantitative research. I loved the stuff I did in the lab, but it felt a little distant to me. It didn't feel connected to the world. I think that's what always made me a little hesitant about research, but there was not really a path for technical people when I graduated from college to do things outside of academia. You could go work in finance, and tons of people I know worked on Wall Street. But outside of that, there wasn't a clear thing to do.

*My career has been strange. Sometimes, it feels like a random walk, but it's more of a greedy algorithm.*

I took a fellowship to go teach physics at a boarding school in England for a year because I also really loved teaching. It was a great way to experience teaching physics, learn about high school kids and what they're like, and travel around Europe. I really enjoyed it, and I could really see myself teaching for a long time, but I started to apply to grad school because I knew that teaching would always be there. I liked it and I knew I could go do that, but if I wanted to go back to grad school, I felt like I had to do that relatively quickly before I got too old and just too tired to go to grad school.

I started grad school at UC Berkeley in physics, and I enjoyed the classroom aspect of it. I started doing research in condensed matter physics. I enjoyed that, too, but I was basically in a second sub-basement. I was doing this condensed matter research, and I had this feeling that it was detached from the outside world. Also, it was really solitary.

I made a transition and switched research fields to particle physics and astrophysics. I did my PhD on a neutrino physics experiment that is located at the South Pole. It's called *IceCube*, and it's operating now. We basically buried sensors in the polar ice cap to measure cosmic neutrinos. It was a transition for me because, all of a sudden, I was working with a couple hundred people all around the world, half in Europe, half in the US in all these different time zones. It felt like I wasn't working on something by myself. I was working on really interesting problems with other smart people and doing really hard work. I think that was what kept me in grad school — knowing that I was working with other smart people in a collaborative environment.

I found out that it suited my personality a lot better than solitary research, and that was when I was introduced to everything I know about data science. That's when I learned most of the statistical techniques and most of the computer programming I know, and it was when I started using machine learning techniques. Basically, the common thing in particle physics now is you have a big detector, and there are tons of things happening in your detector, the vast majority of which you don't care about and are not trying to

study. But you also have something happening in your detector that you care about.

The whole game is trying to figure out what is signal and what is background. These detectors are so complicated, and the signal-to-noise ratios are so low, that techniques from computer science and machine learning have really infiltrated physics. It's interesting because the old school professors don't like machine learning. You go to these seminars with old guys from the 1960s, and they ask aggressive questions and give you these looks. They don't like these techniques. They just think they're black boxes. But for the younger generations, they are common tools to do the most sensitive analysis of the thing you care about.

*I started learning R and just took any chance I could get to learn — like going to meet-ups and other things that one does to learn these things out of the classroom, messing with data sets and going to hackathons.*

That was how I was introduced to machine learning. For my thesis research, I used a lot of neural networks to do pattern recognition for a particular kind of neutrino signal in the detector that we cared about. I found that I liked programming and statistical work and machine learning a lot more than I liked lab research.

That was how I was introduced to this field, and I finished my PhD. I did a post-doctorate for a year in neutrino physics, and this was when the term data science first came out and people started talking about it. I started reading lots of blogs about the field, realizing that this is something I wanted to do and had the right skills for.

I started messing around with Kaggle when Kaggle first came around. I started learning R and just took any chance I could get to learn — like going to meet-ups and other things that one does to learn these things out of the classroom, messing with data sets and going to hackathons.

One day, as a post-doc, I was in my office randomly reading KDNuggets, which is a blog for learning data science stuff. They posted an ad for the Obama campaign looking for scientists, statisticians, and computer scientists to go work for the campaign. It seemed like an intriguing opportunity for me. I had never worked in politics before, but I had always been interested in it. Because I had been reading a lot about data science and making that transition, it seemed like a good opportunity to test out if I was any good at that stuff and if this work was interesting to me in a low pressure environment as it was only a year. I didn't have to quit my post-doc. But in reality it was actually the opposite of a low pressure environment being there.

I applied. I had this interview, and I got an offer. The funny thing was I assumed that since it was a political campaign, they would pay me so little money that there was no chance I would be able to accept the job. It turned out that it was basically the same as my post-doc salary, which shows you how well-paid academics are.

*The funny thing was I assumed that since it was a political campaign, they would pay me so little money that there was no chance I would be able to accept the job. It turned out that it was basically the same as my post-doc salary, which shows you how well-paid academics are.*

I took the job. I started in November 2011 and worked through election day. It was a transformative experience for me in a couple of ways. First, I realized that a lot of the things I was doing were not dissimilar to the things I was doing in physics. I spent tons and tons of time writing Python code to grab data from APIs (Application Programming Interfaces - the way one programmatically interacts with another application or data stream) or to scrape data. It was a lot like writing data acquisition code in physics. I was doing statistical stuff in R rather than the packages we used in high energy physics, but I was still building statistical models, predictive models. Instead of particle physics models, I was building models to predict how much money a fundraising email was going to make from its early returns. If we sent an email asking people to drive to a neighboring state to canvas, who found the people most likely to respond favorably to that email so that we could just target those people.

I realized that the techniques of working with data and understanding statistics and being able to visualize something and tell a story about it — these were all things that I learned in physics and that carried over really well into this data science context.

We can talk more later about the campaign if you're interested, but we did lots of modeling, randomized experiments, e-mail fundraising optimization. It was an amazing experience. It was actually the first time I felt the technical skills I had could be used to affect the world, to work towards something that could affect the world positively. That was really cool.

Then, I thought briefly about going back to finish my post-doc afterwards, but I decided that I really liked working in data science more than I liked working in research. It was like the feeling I had when I switched to astrophysics. I like working with people a lot more than I like working by myself. I like to work on things that have more impact. You see a lot more of it in industry, in data science, than you do in research. I like the pace a lot more. I think research can often be very slow, especially particle physics. It takes 10 years to build an experiment now. You have to have a monastic personality to be a

physicist nowadays.

I found the pace suited me better, and the work was actually just as interesting or more interesting to me than a lot of the stuff I was doing in physics. That's how I ended up where I am. After the campaign, I went to a startup in Chicago called Braintree, which does credit card processing for other startups like Uber, Airbnb, Github, and a lot of other growing tech companies. I started the data team there, and it was a really interesting introduction to the world of startups. Then, Braintree ended up getting acquired by Paypal in the fall, and for reasons mostly unrelated to that, I decided to make a switch and come to work with old campaign colleagues at a startup called Civis Analytics that spun out of the analytics shop on the Obama campaign.

At Civis we're doing really interesting data work for a lot of political clients and campaigns like we did on the Obama campaign, but we're also working with some interesting non-profit and corporate clients. We're really trying to do a lot of individual level predictive stuff like we did on the campaign, focused on political and social good work.

That's my story in a nutshell.

**DSH:** You mentioned that some of the most useful things you did during your time as a PhD were working on hackathons or working on Kaggle or data sets and working with people. Do you have more to add to that? What was the most useful part of being a post-doc and PhD student for your later data analysis/data science career?

*I always tell students that I think the most useful skill you learn in grad school is how to teach yourself stuff and how to figure out things that you don't know. That's one thing. The second thing is to be stubborn and beat your head on a problem until you make progress. It's really those two things.*

**Michelangelo:** I always tell students that I think the most useful skill you learn in grad school is how to teach yourself stuff and how to figure out things that you don't know. That's one thing. The second thing is to be stubborn and beat your head on a problem until you make progress. It's really those two things.

I feel like grad school gave me confidence. Physicists tend to be a pretty arrogant bunch. They think they can learn anything, but that was the lesson I learned in grad school. I don't know every programming language in the world, but I'm confident that if I spend a few months, I could pick up a new programming language or pick up some new infrastructure tool or modeling technique. I can teach myself those things. I can go out there and read academic papers, read software manuals, and teach myself the tools I

need to get the job done. I think that's pretty common across grad school fields. Most of the things you learn you don't learn in the classroom. You learn by completing a project and teaching yourself things. In data science, that's a crucial skill because it's a quickly growing field and it encompasses a ton of things. You can't finish a degree and know all the things you need to know to be a data scientist. You have to be willing to constantly teach yourself new techniques.

That was one of the things I learned in grad school. The other is the ability to work on a hard problem for a long time and figure out how to push through and not be frustrated when something doesn't work, because things just don't work most of the time. You just have to keep trying and keep having faith that you can get a project to work in the end. Even if you try many, many things that don't work, you can find all the bugs, all the mistakes in your reasoning and logic and push through to a working solution in the end.

*And the final thing is that it really helps to have experience working with data.*

Having confidence in yourself is another thing. I think that working on a really hard problem like in grad school can help you learn that. And then there are just the technical things like learning how to program, running on large computer clusters. Those are the things that I think are really helpful from grad school, but the advice I give to grad students is: if you feel like you want to leave grad school and do something else, keep that in mind when picking which tools and techniques you use for a dissertation. If you can write your dissertation in Python rather than some obscure language like FORTRAN, it's probably going to be better for you. Try to be as marketable as possible with the things you learn when you're doing your PhD.

And the final thing is that it really helps to have experience working with data. The only way to learn how to work with data is to actually work with data. You can read about it, and people can teach you techniques, but until you've actually dealt with a nasty data set that has a formatting issue or other problems, you don't really appreciate what it's like when you have to merge a bunch of data sets together or make a bunch of graphs to sanity check something and all of a sudden nothing makes sense in your distributions and you have to figure out what's going on. Having that experience makes you a better data analyst.

**DSH:** So far, you've given a lot of advice for graduate students, for example working with more common tools or working with data. Can you expand on that because you are a physicist-turned-data-scientist? What is your advice for other physics PhD students or other physicists who are transitioning to data science?

**Michelangelo:** My advice is to recognize the skills that you have. In terms of actually mechanically making that transition, there are lots of ways people can learn more about the field and demonstrate their interest. From a hiring perspective, when I talk to PhD students who say they want to be data scientists, I become skeptical if they haven't taken any active steps. "Hey, I participated in these Coursera courses or these Kaggle competitions. I've gone to the Open Government Meetup and have done these data visualizations." Things like that demonstrate that you can work on problems outside your academic specialty, and they show that you really have initiative. They also show that you can teach yourself new things.

The worst thing is when people present the physics job market as terrible, and they say that's why they want to get a data science job. You don't want to hire someone like that. You want to hire someone who's going into data science because of what it's like, because they want to work on data in the real world. You want it to be a positive thing rather than a negative reason that they're leaving physics.

*I'm excited about future applications where data science is going to be seen as a positive force.*

To be honest, it's not a terrible reason to want to leave academia because there's no job, because you're lonely, because you're working on a tiny, tiny problem. Those are good reasons to leave academia, but from a practical standpoint, when you're presenting yourself to other people, I think you should focus on the positive reasons that you're excited to do something else rather than why you're negative about what you're doing. All those things are true, and all those things are reasons why I also personally decided to leave.

The other piece of advice I always give to job seekers is when people talk about data science jobs, it can mean so many different things. At each different place, when they're talking about hiring a data scientist, that can mean something totally different. In some places, they just want someone who can run SQL queries and numbers for every report. In other places, they want people who are actually going to build data infrastructure. In other places, they want some people who are going to build predictive statistical models and design experiments. In some places they want the unicorn that can do all that stuff. So it's really important to ask a lot of questions and figure out what a company really wants when they want a data person. What would someone actually want in that role? Are there other people currently working as data scientists at the company? What are they doing? Is there an engineering team? Is there a product team?

**DSH:** You mentioned earlier about working with people and making a large impact. What about the future of data science excites you the most? What are some of the

positive reasons that you would give to graduate students on why data science has greener pastures?

**Michelangelo:** I'll leave out all the sociological reasons that I already talked about, why it's more enjoyable to work in a collaborative, fast paced environment, and to see the impact of your work. In academia, you don't have clients. In physics, I always felt that we had to beg people for money to do what we wanted to do, and that may still be true in data science, depending on your company. But most of the time, there are people who are interested in the output of what you're doing and really appreciate those skills.

I also think the work is exciting. It's incredibly exciting, and it's still in an early phase. I'm not going to go into the cliché of how much data we're collecting and how all of these organizations are collecting more and more data. Many people have talked more eloquently than I can about that. But it's true. Organizations have tons and tons of data, and they don't necessarily know what to do with it. They're starting to think about what to do with it, and they need help from people like us to actually do that work.

*There are lots of people who are writing tutorials explaining different techniques and different projects they've worked on. None of that existed when I was younger, and it's awesome that you can go out, get that stuff, and get an idea of what's going on.*

This is the reason that I came to Civis. I'm really excited about future data science applications that people are going to look at and think are benefiting society in a positive way. Like working with non-profits to use their data in smarter ways, or working with all the data that cities are releasing now.

Opening up public data is great, but there are not many cities that are using their data in a real, predictive way right now. New York has done some really interesting predictive things. Chicago has released a lot of data, but Chicago hasn't done a lot of interesting analytics with its data as a city. They just release data to the community and hope the community will do it.

I'm excited about future applications where data science is going to be seen as a positive force for good, because honestly I'm a little worried. Right now, a lot of the applications we have with data have to do with targeted advertising, cookie collection, online optimization of ad click rates, etc. That's great, but I'm worried that, at some point, there's going to be a backlash against collecting more and more data about people. I'm hoping that before that happens or when that happens, there are enough positive counter examples of ways data is being used to benefit people and society that it can

prevent some of that backlash.

I wish I could talk about this a little more specifically with the clients we're working with at Civis, because that's something we're really focused on. Before I started here, one of the big engagements we had was with the College Board, the folks who administer the SAT exam. We spent a very long time working with their data and helping them build models to understand which kids weren't going to colleges and universities commensurate with their abilities. Could we predict that? If so, what are the implications for designing interventions to help those high school students? I'm hoping that we'll have more and more examples of data science work like that, work that people feel good about rather than just seeing that companies are trying to collect data from them.

Also, there's the summer Data Science for Social Good Fellowship that one of the data scientists from the campaign started in Chicago last summer and that I was a volunteer mentor for. Some of the projects we worked on addressed really interesting social impact problems, and I'm hoping there will be more and more of these kinds of applications in the future. That's what excites me about the future of data science.

**DSH:** When we spoke to Jace from Khan Academy, it was inspiring to see him apply his knowledge from quantitative finance to education. How can we encourage more of this in the nascent data science community?

*I do worry that there's a little bit of hype, but it's undeniable that there's a very solid grain of truth to the whole data science thing.*

**Michelangelo:** I think people really want to do more and more of this kind of work. I think about this a lot. My wife is a lawyer, and almost all lawyers do some amount of pro bono work in a given year. I think it would be awesome if we could get some engineers and computer scientists and data people to do a certain amount of pro bono hours every year working with a non-profit. A lot of people are already doing that as a volunteer thing, but if we could institutionalize that, I think that would really be awesome for the field.

**DSH:** Your background as a science teacher and as a writer is different than most of the other people we've interviewed. As a science teacher and writer, how is data science doing on the PR side? What is data science missing on the teaching and writing side?

**Michelangelo:** I forgot to mention that earlier. I was briefly a science journalist. I took a summer off and worked at *The Economist* and wrote about science and technology. I freelanced for them for a while after I was in London for that summer. I actually think teaching and writing have helped me become a better data scientist because a lot of what

I do is interact with my colleagues on a daily basis. I teach them new things all the time. They teach me things. We sit in meetings and look at graphs and talk through algorithms and techniques, and we ask each other questions. We explain things to each other, and you tell a story about the data. That's very similar to the things you do in a classroom when you're teaching people something. It's very similar. When you're writing about science, you try to simplify things and explain them to people. Those skills have been useful for me as a data scientist.

As a field, I think data science is doing a pretty good job. There are so many people who are blogging about their work and telling stories about their work. There are lots of people who are writing tutorials explaining different techniques and different projects they've worked on. None of that existed when I was younger, and it's awesome that you can go out, get that stuff, and get an idea of what's going on. I think that is really awesome.

**DSH:** Sometimes, when we talk with people who come from an academic background, they are suspicious of data science. They think of it as a fad. I'm thinking about the hype that might be behind that or how some people react to it. What would you say to somebody who thinks it's a fad?

**Michelangelo:** First of all, I think it's a valid concern. I do worry that data science is being super hyped up right now. Not by the people that are doing it or who know what it's really about, but there are lots of companies who want to sell people things. A few journalists write an article on something, and everyone else feels like they need to write an article too. Then, it becomes this big giant thing.

*This is why I'm excited about more positive examples of data applications. I think the more positive examples of data science that we have, the more it will help counteract a lot of the hype.*

I do worry that there's a little bit of hype, but it's undeniable that there's a very solid grain of truth to the whole data science thing. We do have lots and lots of data, and we're collecting more every day. I can't imagine that companies and organizations are going to want to be less efficient in the future about how they reach out to people, about how they optimize their own operations. I think that trend is going to continue, and they're going to want people to help them analyze that data. The skills you need to do that just don't come from a single discipline like statistics or computer science. They have all the interdisciplinary aspects of what people call data science.

This is why I'm excited about more positive examples of data applications. I think the more positive examples of data science that we have, the more it will help counteract a

lot of the hype. I think all the hype has not just been around data science but about tools. Everyone's been talking about Hadoop. Hadoop is great, but it's a tool. It's not the most important thing in the world, and not every organization needs to have a giant Hadoop cluster, but, with the hype, these organizations are like, "If you're not running a Hadoop cluster, you're not doing anything interesting with your data."

The term big data makes me want to throw up because it's become an overused, overhyped thing. To me, it's not the amount of data you have. It's what you do with the data you have and how you apply it to problems and what interesting things you're doing with it. That's so much more important.

I actually don't think that anything we did on the campaign, when you talk to someone from Silicon Valley, counts as big data. We didn't have a petabyte of data, but it was what we did with it, how we were changing the organization and the practices of the campaign that was really important.