JOURNEY TO DATA SCIENTIST

INTERVIEWS WITH MORE THAN TWENTY AMAZING DATA SCIENTISTS
Journey to Data Scientist

Interviews with More Than Twenty Amazing Data Scientists

Kate Strachnyi
Dear Reader,

First of all, thank you for picking up this book! My name is Kate Strachnyi, and I am excited to share some interesting stories with you.

I’ve recently become obsessed with data science and the processes of applying a scientific method to answer the questions concealed in various forms of data. I was inspired to better understand these methods and ways I could master this incredible field.

In my pursuit of data enlightenment, I signed up for various tutorials and training sessions with sites, such as Lynda.com and Coursera. I started reading blogs, signed up for Kaggle competitions, and attended several meet-ups in my local area (NYC).

Overall, I found the culture within the data science community to be very inviting and welcoming of new members who shared similar levels of inquisitiveness.

I have always been compelled by a desire to derive meaning where otherwise none seemed to exist and quickly learned that the simple task of building a visual representation was an effective way to establish order from the chaotic frenzy of complex data fields.

To put it simply, where most see chaos, I see beauty. A beauty that only unfolds when we’re able to tame those vast complexities and allow for a story to be told. A story without ambiguity and ripe with absolution and truth. A story, by data.

I started a blog called storybydata.com and a YouTube channel, Story by Data, to track my journey to data scientist, as well as to share my knowledge of data science.

I thought it would be really interesting to talk to experts in the field, so I began to speak to data scientists and noted down words of wisdom. I believe that there are others like me out there—people wondering about how to break into the data science field, who would love to hear from experts. Therefore, I decided to write this book!

My goal is to make the data scientist role more accessible to the public. The advice that I have is as follows: don’t be deterred by thinking that there are too many barriers for entry (PhD, and so on). People took various paths to obtain the coveted title of data scientist.

Do you want to be a data scientist? Continue reading to gain amazing insights and career advice from over twenty data scientists from companies such as IBM, Bloomberg, Springboard, and Honeywell. These data scientists come from different industries and countries. One common theme that came out of all interviews is that in order to be successful, you must be innately curious. Never stop asking questions of the data, and never be satisfied with the results.
This book is dedicated to all aspiring data scientists. Stick with it and you’ll get there!
Acknowledgments

Thank you to my mother for watching my two little girls as I juggled my responsibilities of motherhood, a full-time job, obtaining a data science certificate from Johns Hopkins (Coursera), and creating content for my Story by Data blog and YouTube channel. Thank you to my husband for pitching in when I needed to meet my writing goals and for believing in me.

Most of all, thank you to all the interview participants for making time in their busy schedules to have a conversation or correspond in support of this work. Your input has truly made this book what it is.
Data Scientist

A quick definition of a data scientist—a person who possesses several key skills and knowledge, including proficiency in mathematics, statistics, programming, and an understanding of his or her industry/domain. A data scientist must be able to think critically and solve problems that most people wouldn’t dare to attempt. In fact, in a recent article, the Harvard Business Review identifies the profession as the “sexiest job of the 21st century.”

A data scientist typically has some experience in working with large and/or messy data sets, being able to figure out how to address vague problems and come up with interesting insights from data, despite the sometimes-overwhelming volumes of it.

These characteristics are at the core of the data scientist and as you will soon learn from the following interviews, data scientists are innovators who are eager to think out of the box.
Contents

1. Raj Bandyopadhyay: Director of Data Science Education at Springboard
2. Matt Williams: CEO at WD Creative Analytics
3. Hector Alvaro Rojas: Data Scientist at SDP Consulting
4. Mike Tamir: Chief Data Scientist at Intertrust Technologies Corporation
5. Alexander Bessonov: Data Scientist at Bloomberg LP
6. Mansi Gupta: Data Scientist at LinkedIn
7. Kunal Jain: Founder and CEO at Analytics Vidhya
8. Khalifeh Al Jadda, PhD: Lead Data Scientist at CareerBuilder
9. Sergey Bryl: Data Scientist at MacPaw Inc.
10. Diane Reynolds: Chief Data Scientist at IBM
11. Vaibhav Gedigeri: Data Scientist at Honeywell
12. Sean Preusse: Lead Analyst at Westpac Group
13. Andrew Paul Acosta: Data Scientist at Milesius Capital Resources
14. Rosario Silipio: Principal Data Scientist at KNIME
15. Ben Taylor: Data Scientist at HireVue
16. John Bottega: Senior Advisor—Chief Data Officer Forum
17. Steven Tedjamulia: CEO at Predictive Network
18. Jared P. Lander: Chief Data Scientist Lander Analytics
19. David Rubal: Chief Technologist, Data and Analytics, Principal Data Scientist at DLT
20. Renee M. P. Teate: Data Scientist at HelioCampus
21. June Andrews: Data Scientist at Pinterest
22. Armando Viera: Data Scientist at ContextVision AB
23. Charles Givre: Senior Lead Data Scientist—Strategic Innovation Group at Booz Allen Hamilton
Raj Bandyopadhyay  
Director of Data Science Education at Springboard

“Often people tend to underestimate the business and communication aspects of data science.”

Path to Data Science

I have an educational background in very traditional computer science. A bachelor’s degree in computer science and a master’s degree and PhD in computer science. I have worked in programming languages and compilers though it’s very traditional computer science stuff. Graduated with a PhD in 2008 and started working as a developer at a start-up company in Atlanta. After working for three years, I realized that I wanted to do some data science software development while staying on in technology.

Considered something like management consulting, but I realized that I didn’t want to; as much as I was interested in business, I wanted to remain close to technology, and I was not really interested in dedicating one hundred hours a week in consulting. This was the time Coursera launched its first course on machine learning. I took the course as it was kind of cool, and wanted to learn more about it. I started taking online courses, and began to change jobs, taking up jobs with more data engineering and data science roles. Eventually, ended up as the chief data scientist at a company called Pindrop Security, which is a security start up. I was the first and only data scientist. Eventually, over a period of two years, I built an entire data science team.

Later, I moved to San Francisco where my wife was working, so I quit my job in Atlanta. I was already mentoring for Springboard, so I took a full-time role with them to maintain and develop data science courses.

Selecting Candidates for Data Science Team

Pindrop Security is basically a machine-learning system used to detect fraud. In such a company, especially when it is a cyber-security company where the core product is based on data science, obviously, the technical knowledge has to be very high. The people you select, whether for transformation, internship, or on a contract basis have to be very good.

I hired three data scientists. One of them was a PhD in machine learning, so it was a very theoretical aspect for the group. The other two people had PhDs in aerospace engineering. One of them especially had experience in handling data as she had worked on data-related projects.

The third person had a master’s in aerospace engineering and also an MBA—a very good combination of both technical knowledge and business skills.

The team had a broad set of skills with no fixed background. All of these people were very technically competent and had both engineering and math backgrounds. When I was hiring and interviewing people, there were two issues that I really found especially with fresh graduates: even those with data science degrees didn’t have much experience with real-world data. So basically, though many of them had done projects as part of their courses, the projects were on data that was already cleaned up or data that was more toy data sets. They didn’t have the experience of working on a data science project from scratch, figuring out how to collect the data, how to clean the data, what kind of data to collect and how difficult was the trade off in collecting...
different kinds of data. One criterion that I had was to look for candidates with experience in real-world data.

The other issue that I found in fresh graduates was that they weren’t able to explain their projects to anybody outside the technical field; outside the data science field. Working as a data scientist in industry, whether it’s in a small or in a big company, is not like an engineering role. It means as an engineer you work with the engineering team and do not need to explain to anyone outside your team or to the management. However, whether you are a senior data scientist or a junior scientist manager, you have to interact with people in the sense that you are typically working with the product manager or the sales personnel or the marketing personnel: somebody who is outside your technical team and may not know anything about the work you do. What that means is you have to have really good communication skills to make them understand.

**An Understanding of Business**

You must be able to understand what business marketing is, what the salesperson is saying, and what their problem is. You must also be able to translate this into a data analysis problem. Then when you perform any kind of analysis that gives results, you have to explain it to them in the language they understand. You cannot say “90 percent true detection and 5 percent false positives”, etc. You have to explain it and translate it into how much money they can make or how much more efficient the system can become.

This type of communication skill was really hard to find. So that was what I really tested for. I typically would ask them to take a project that they were working on or they had worked on, and explain it to someone in the room who was not a data scientist. I would take on the role of a management executive or a business manager or a product manager and they would have to explain it to me. What I really looked for in fresh graduates and people with experience was the ability to communicate what they had done.

**Working with Messy Data—Springboard**

At Springboard, we have data science courses that are data driven, and we also emphasize from the beginning that students have to work on a capstone project, because the guidelines for the actual projects for both mentors and students is that it has to be an interesting and realistic project.

What this means is that you start by coming up with a client. The idea is that even before you start working on the project you know this is the problem that I am going to solve and this is the client. For example, one of my students did a project where he would predict a landslide and the client would be the department of transportation who had to decide where to build roads.

We try to make projects as realistic as possible. We emphasize that students collect real-world data and it’s not simulated data or collecting data from cases that are already clean. Then any kind of analysis they have to do themselves and come up with the presentation and report targeted at that client. So that’s the approach that we take at Springboard to ensure that the students have the experience of the real-world data that helps them to learn how to clean up the abstract data and also have the experience of communicating with real-world clients.

**Mentors at Springboard**
All our mentors are active data scientists. There are people who work as data scientists as part
timers or as contract employees, who bring industry perspective and can tell you what industry
standards are basically.

When you hear a mentor telling you why you are doing something, you gain confidence in what
you are doing and it helps you when you talk to other people who are experienced in data science
and are looking for a job or some information in the field.

**Spiked Number of Interest in Data Science**

Interest and registration is definitely going up; recently, we launched a new program called data
science career track—which is also called premium offer and it combines data science with job
guarantee and career support. The number of applications for that was way more than we
expected.

**Interesting Project at Pindrop**

At a project that I worked on in Pindrop Security, they had a system that analyses the audio of
calls coming into a call center, which determines if the call might be fraudulent.

However, as part of being a data scientist especially in a company like that you probably have to
speak about what are the data sources that you can analyze or that can add more color to the
problem or what are the other methods that you can use.

There was this interesting problem that came to us from a client that was from banks, credit card
companies, and other such companies. They said they had some interesting observations in their
interactive voice response or IVR, which actually means automatic menus like press 1 to do this
or to do that.

What they realized was that their IVR or what they suspected was actually that IVR had been
attacked by the robot for code and it was trying out different combinations, in an attempt to guess
combinations of account numbers and pins intending to extract customer information from that.
When you have a bot, you can do millions of transactions a second. You can try several pins.

Imagine if there’s a breach like what happened at Target or a bunch of credit card numbers or
account numbers have been stolen and then they already have the pin. For those accounts, we can
log in and get information about the accounts, which can be used in future attacks.

That’s what they suspected; however, they didn’t have a way to verify it and they also had no
way to actually detect it. What they did was that they gave us data numbers for six months of
IVR. I was the only data scientist at that point. I promised to do it. We took on that project and it
took me about three months to cover the initial prototype and six months to come up with
something that I could confidently say worked. Basically, I came up with a machine-learning
solution to be able to verify the hypothesis. And there actually was that kind of pattern. The
question we needed to solve was how to detect repeated attack patterns and how to separate that
pattern from the regular users. It’s difficult to distinguish between someone who forgot their pin
number and was trying different combinations and someone who was attempting fraud. These are
interesting questions that came up. We got a provisional patent on this.

**Approach**
I looked at the menu tree to learn what was the path followed for each call in the menu tree; then I extracted the features and added some labeled data initially about which calls are fraud and which calls are by the customers. Then I labeled certain calls as fraudulent or certain activities as fraudulent. And they have to come up with that classification and tell me how I will identify fraud.

**Advice to Aspiring Data Scientists**

People wonder if they should have a master’s degree or what credentials are important for the role. What I tell them is that what is most important for employers is having a good portfolio.

Projects show your technical competence and your communication skills. GitHub is important because it is the place for data scientists to show off their work. If you have a project on GitHub you will probably have documentation and write up for those projects, you will have a presentation and so on. That’s what you show your potential employer and the potential employer or the data scientist will look at that and will say this person knows his or her stuff, how to communicate, how to present things and so on.

Ultimately, having a good portfolio is far more important than what kind of credentials or resume you have in data science even right now. I have noticed that when you work with a mentor where you give that kind of portfolio, you will be able to create projects that you are good at and add it to your portfolio. But you can do that on your own; ultimately what you want in your portfolio are realistic projects, which display your technical and communication skills.

**Data Science Is a Combination of Skills**

Data science is a combination of skills. It’s an amalgamation of being a programmer, an algorithmic or a theoretical person, a statistical person, and a business person. Often people tend to underestimate the business and communication aspects of data science. They focus on the programming aspects: which language should I learn, or which algorithm should I learn. The communication and basics of business aspects are almost equally important to make or break your career in data science.

I urge people who want to take up data science not to underestimate the business and communication requirement in data science.