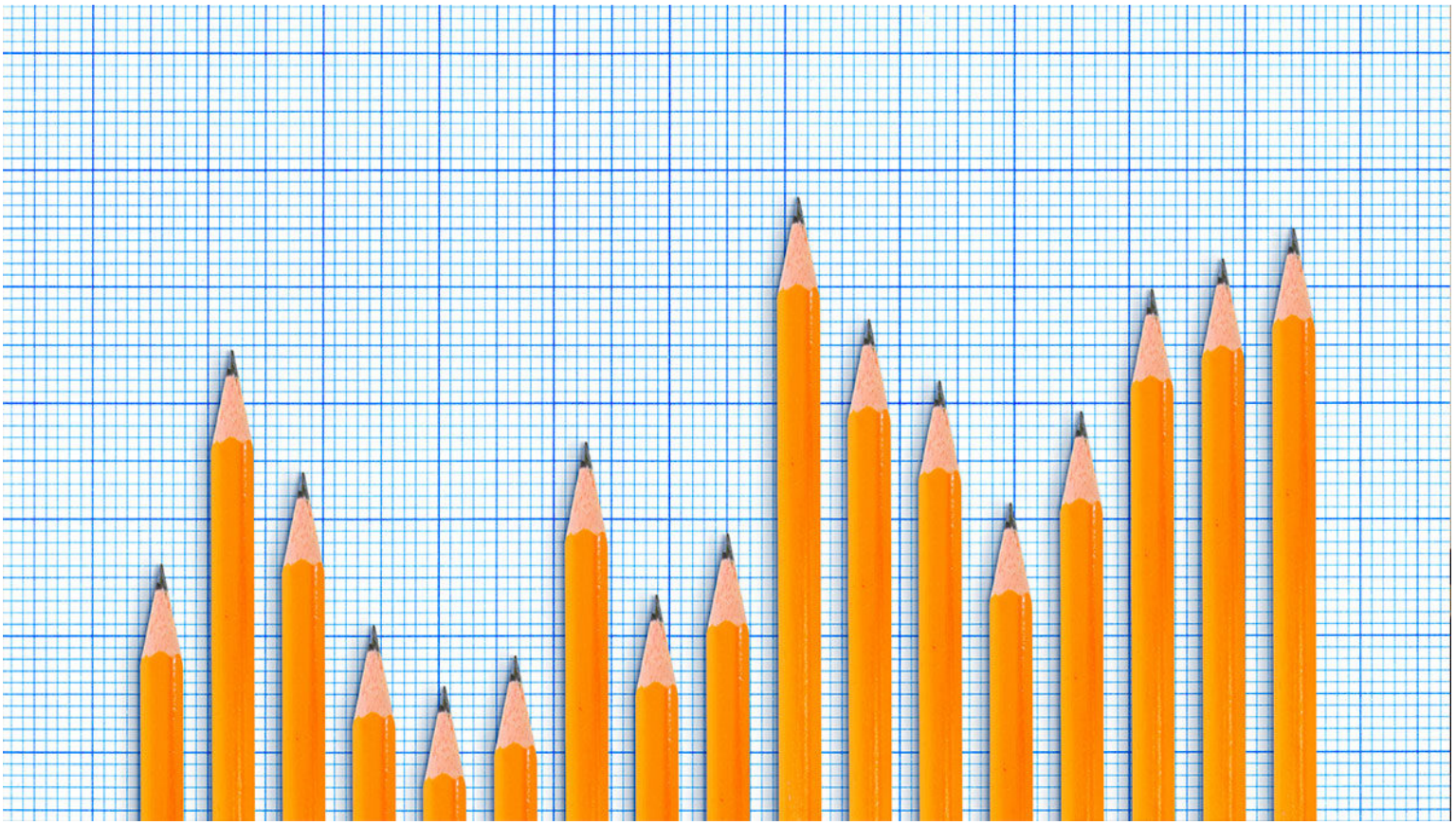
ANALYTICS

What Data Scientists Really Do, According to 35 Data Scientists

by Hugo Bowne-Anderson

AUGUST 15, 2018



BURAKPEKAKCAN/GETTY IMAGES

Modern data science emerged in tech, from optimizing Google search rankings and LinkedIn recommendations to influencing the headlines Buzzfeed editors run. But it's poised to transform all sectors, from retail, telecommunications, and agriculture to health, trucking, and the penal system. Yet the terms "data science" and "data scientist" aren't always easily understood, and are used to describe a wide range of data-related work.

What, exactly, is it that data scientists do? As the host of the DataCamp podcast *DataFramed*, I have had the pleasure of speaking with over 30 data scientists across a wide array of industries and academic disciplines. Among other things, I've asked them about what their jobs entail.

It's true that data science is a varied field. The data scientists I've interviewed approach our conversations from many angles. They describe a wide range of work, including the massive online experimental frameworks for product development at booking.com and Etsy, the methods Buzzfeed uses to implement a multi-armed bandit solution for headline optimization, and the impact machine learning has on business decisions at Airbnb. That last example came during my conversation with Airbnb data scientist Robert Chang. When Chang was at Twitter, that company was focused on growth. Now that he's at Airbnb, Chang works on productionized machine-learning models. Data science can be used in a number of different ways, depending not just on the industry but on the business and its goals.

But despite all the variety, a number of themes have emerged from these conversations. Here's what they are:

What data scientists do. We now know how data science works, at least in the tech industry. First, data scientists lay a solid data foundation in order to perform robust analytics. Then they use online experiments, among other methods, to achieve sustainable growth. Finally, they build machine learning pipelines and personalized data products to better understand their business and customers and to make better decisions. In other words, in tech, data science is about infrastructure, testing, machine learning for decision making, and data products.

Great strides are being made in industries other than tech. I spoke with Ben Skrainka, a data scientist at Convoy, about how that company is leveraging data science to revolutionize the North American trucking industry. Sandy Griffith of Flatiron Health told us about the impact data science has begun to have on cancer research. Drew Conway and I discussed his company Alluvium, which "uses machine learning and artificial intelligence to turn massive data streams produced by industrial operations into insights." Mike Tamir, now head of self-driving at Uber, discussed working with Takt to facilitate Fortune 500 companies' leveraging data science, including his work on Starbucks' recommendation systems. This non-exhaustive list illustrates data-science revolutions across a multitude of verticals.

It isn't all just the promise of self-driving cars and artificial general intelligence. Many of my guests are skeptical not only of the fetishization of artificial general intelligence by the mainstream media (including headlines such as VentureBeat's "An AI god will emerge by 2042 and write its own bible. Will you worship it?"), but also of the buzz around machine learning and deep learning. Sure, machine learning and deep learning are powerful techniques with important applications, but, as with all buzz terms, a healthy skepticism is in order. Nearly all of my guests understand that working data scientists make their daily bread and butter through data collection and data cleaning; building dashboards and reports; data visualization; statistical inference; communicating results to key stakeholders; and convincing decision makers of their results.

The skills data scientists need are evolving (and experience with deep learning isn't the most important one). In a conversation with Jonathan Nolis, a data science leader in the Seattle area who helps Fortune 500 companies, we posed the question, "Which skill is more important for a data scientist: the ability to use the most sophisticated deep learning models, or the ability to make good PowerPoint slides?" He made a case for the latter, since communicating results remains a critical part of data work.

Another recurring theme is that these skills, so necessary today, are likely to change on a relatively short timescale. As we're seeing rapid developments in both the open-source ecosystem of tools available to do data science and in the commercial, productized data-science tools, we're also seeing increasing automation of a lot of data-science drudgery, such as data

cleaning and data preparation. It has been a common trope that 80% of a data scientist's valuable time is spent simply finding, cleaning, and organizing data, leaving only 20% to actually perform analysis.

But this is unlikely to last. These days even a great deal of machine learning and deep learning is being automated, as we learned when we dedicated an episode to automated machine learning, and heard from Randal Olson, lead data scientist at Life Epigenetics.

One result of this rapid change is that the vast majority of my guests tell us that the key skills for data scientists are not the abilities to build and use deep-learning infrastructures. Instead they are the abilities to learn on the fly and to communicate well in order to answer business questions, explaining complex results to nontechnical stakeholders. Aspiring data scientists, then, should focus less on techniques than on questions. New techniques come and go, but critical thinking and quantitative, domain-specific skills will remain in demand.

Specialization is becoming more important. While there is no well-defined career path for data scientists, and little support for junior data scientists, we are starting to see some forms of specialization. Emily Robinson described the difference between Type A and Type B data scientists: "Type A is the analysis — sort of a traditional statistician — and Type B is building machine learning models."

Jonathan Nolis breaks data science down into three components: (1) business intelligence, which is essentially about "taking data that the company has and getting it in front of the right people" in the form of dashboards, reports, and emails; (2) decision science, which is about "taking data and using it to help a company make a decision"; and (3) machine learning, which is about "how can we take data science models and put them continuously into production." Although many working data scientists are currently generalists and do all three, we are seeing distinct career paths emerging, as in the case of machine learning engineers.

Ethics is among the field's biggest challenges. You may gather that the profession offers its practitioners a great deal of uncertainty. When I asked Hilary Mason in our first episode if any other major challenges face the data science community, she said, "Do you think that imprecise ethics, no standards of practice, and a lack of consistent vocabulary are not enough challenges for us today?"

All three are essential points, and the first two in particular are front of mind for nearly every *DataFramed* guest. At a time when so many of our interactions with the world are dictated by algorithms developed by data scientists, what role does ethics play? As Omoju Miller, the senior machine learning data scientist at GitHub, said in our interview:

> We need to have that ethical understanding, we need to have that training, and we need to have something akin to a Hippocratic oath. And we need to actually have proper licenses so that if you actually do something unethical, perhaps you have some kind of penalty, or disbarment, or some kind of recourse, something to say this is not what we want to do as an industry, and then figure out ways to remediate people who go off the rails and do things because people just aren't trained and they don't know.

A recurring theme is the serious, harmful, and unethical consequences that data science can have, such as the COMPAS Recidivism Risk Score that has been "used across the country to predict future criminals" and is "biased against blacks," according to ProPublica.

We're approaching a consensus that ethical standards need to come from within data science itself, as well as from legislators, grassroots movements, and other stakeholders. Part of this movement involves a reemphasis on interpretability in models, as opposed to black-box models. That is, we need to build models that can explain why they make the predictions they make. Deep learning models are great at a lot of things, but they are infamously uninterpretable. Many dedicated, intelligent researchers, developers, and data scientists are making headway here with work such as Lime, a project aimed at explaining what machine learning models are doing.

The data science revolution across industries and society at large has just begun. Whether the title of data scientist will remain the "sexiest job of the 21st century," will become more specialized, or will become a set of skills that most working professionals are simply required to have is unclear. As Hilary Mason told me: "Will we even have data science in 10 years? I remember a world where we didn't, and it wouldn't surprise me if the title goes the way of 'webmaster.'"

Hugo Bowne-Anderson, Ph.D., is a data scientist and educator at DataCamp, as well as the host of the podcast "DataFramed." @hugobowne

---