2016

DATA SCIENCE

REPORT

CrowdFlower

INDEX

INTRODUCTION

Our 2016 Data Scientist report is a follow up to last year's effort. Our aim was to survey professional data scientists with different years of experience and fields of expertise to find out not only where they feel their profession is going, but what their day-to-day job is like.

What we found was really interesting. For starters, data scientists spend the most time doing the thing they enjoy doing the least. Yet, they still overwhelmingly love their jobs. We looked at how important data scientists think machine learning will be both for their particular role and the industry in general. We also got wildly varied predictions for how the field will evolve in the next five years.

METHODOLOGY

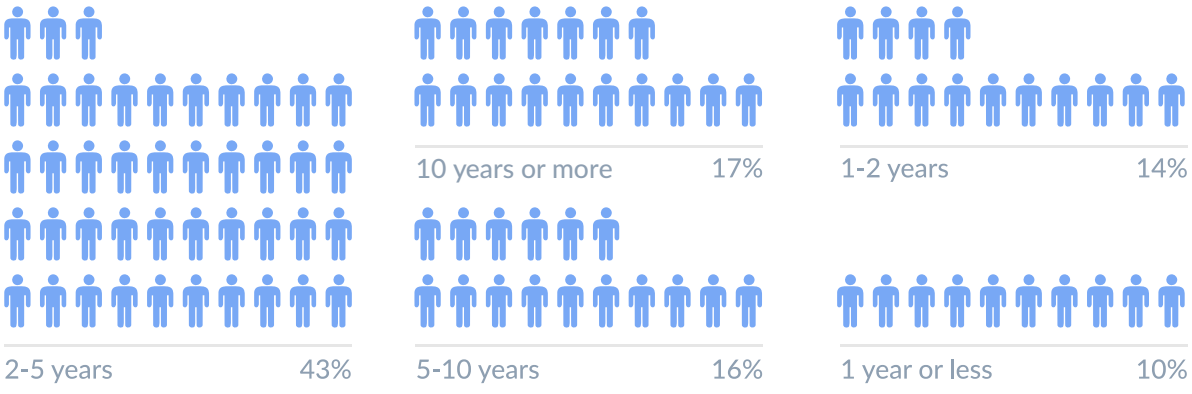
Much like our 2015 Data Science Report, this year's edition comes from real surveys and interviews with real data scientists. We asked them a series of questions about their day-to-day job, what their frustrations are, and so much more. We also ran a few jobs on our platform to look for specific skill sets that employers are looking for right now so data scientists know what skills are most in demand. We've compiled the most interesting trends and replies in the report. We hope you enjoy it.

Who Took the Survey?

Up front, let's break down the sort of data scientists we surveyed. We captured opinions from highlevel CDOs to folks just starting in the field. They had diverse skill sets and varied areas of expertise.

Since data science is still a fairly new discipline–once famously called the sexiest job of the 21st century by D.J Patil in the Harvard Business Review–it's unsurprising that about two thirds of our respondents had been in the field for 5 years or less.

That's not to say we only surveyed data scientists fresh out of college. In fact, our biggest respondent pool had somewhere between two and five years experience.



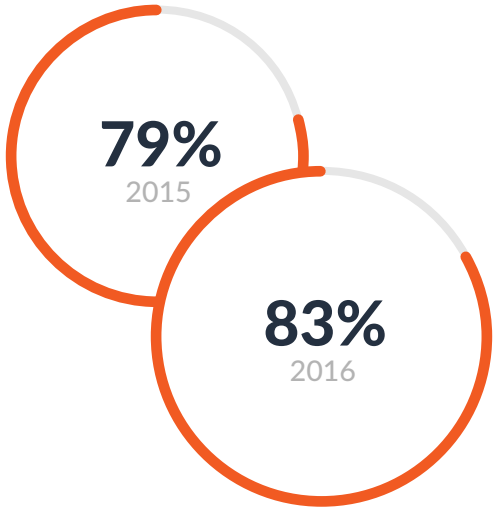
Experience in years of respondents

FINDINGS

There's a Still a Shortage of Data Scientists (And it Might Be Getting Worse)

Last year, we found that 79 percent of respondents said that there were a shortage of data scientists in the field. And while that was staggering, our survey found that in 2016, things might be getting worse.

A full 83% of respondents said there weren't enough data scientists to go around. And with more and more enterprises and organization investing in data this trend is likely to continue.



Respondents who said there weren't enough data scientists to go around

Data Scientists Love Their Jobs

Even though there aren't enough of them to go around, your typical data scientist loves their job.

We asked our respondents to rank how happy they felt in their current position on a simple five point scale. Over a third (35%) of respondents gave their job the highest mark possible. And about half (47%) gave it a 4. In other words? Over 80% of data scientists are really happy at work.

So why is that? It's hard to tell from a simple survey, of course. But judging by the varied responses

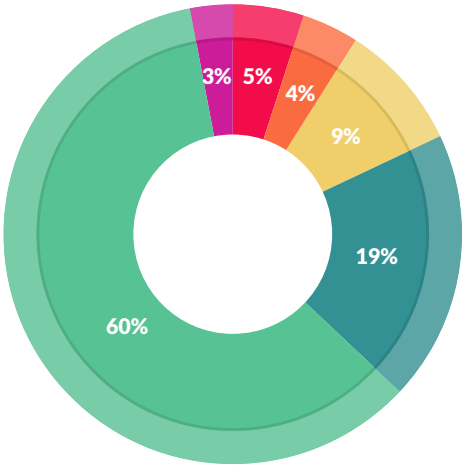
we got about the future of the data science, the most salient takeaway was how excited our respondents were about the evolution of the field. They cited things in their own practice, how they saw their jobs getting more interesting and less repetitive, all while expressing a real and broad enthusiasm about the value of the work in their organization.

As data science becomes more commonplace and simultaneously a bit demystified, we expect this trend to continue as well. After all, last year's respondents were just as excited about their work (about 79% were "satisfied" or better).

Data scientist job satisfaction

How a Data Scientist Spends Their Day

Here's where the popular view of data scientists diverges pretty significantly from reality. Generally, we think of data scientists building algorithms, exploring data, and doing predictive analysis. That's actually not what they spend most of their time doing, however.



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

As you can see from the chart above, 3 out of every 5 data scientists we surveyed actually spend the most time cleaning and organizing data. You may have heard this referred to as "data wrangling" or compared to digital janitor work. Everything from list verification to removing commas to debugging databases–that time adds up and it adds up immensely. Messy data is by far the more time-consuming aspect of the typical data scientist's work flow. And nearly 60% said they simply spent too much time doing it.

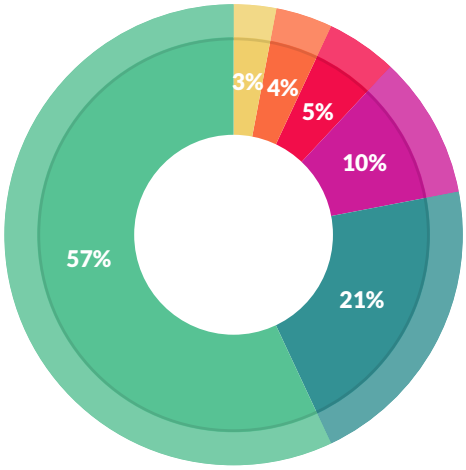
Why That's a Problem

Simply put, data wrangling isn't fun. It takes forever. In fact, a few years back, the New York Times estimated that up to 80% of a data scientist's time is spent doing this sort of work.

Here, it's necessary to point out that data cleaning is incredibly important. You can't do the sort of work data scientists truly enjoy doing with messy data. It needs to be cleaned, labeled, and enriched before you can trust the output.

The problem here is twofold. One: data scientists simply don't like doing this kind of work, and, as mentioned, this kind of work takes up most of their time. We asked our respondents what was the least enjoyable part of their job.

They had this to say:



What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

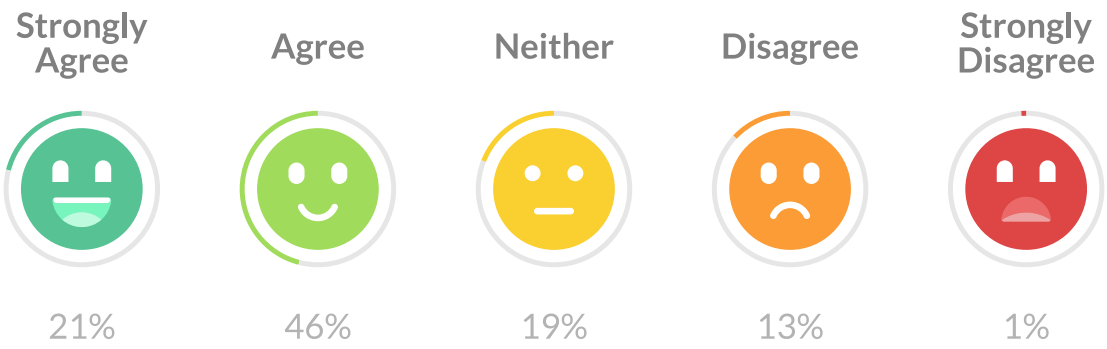
Note how those last two charts mirror each other. The things data scientists do most are the things they enjoy least. Last year, we found that respondents far prefer doing the more creative, interesting parts of their job, things like predictive analysis and mining data for patterns. That's where the real value comes. But again, you simply can't do that work unless the data is properly labeled. And nobody likes labeling data.

Do Data Scientists Have What They Need?

With a shortage of data scientists out there in the world, we wanted to find out if they thought they were properly supported in their job. After all, when you need more data scientists, you'll often find a single person doing the work of several.

Mostly, they have access to the tools they need. Broadly, this means tools, applications, and programs. We asked our respondents to agree or disagree with the following statement: I have access to the tools I need to do my job effectively.

Here's what they said:



It's notable that only 14% of respondents felt they were being held back by their tools. That evidences that, while there may not be enough data scientists, their organizations are committed to giving them the best possible chance at success. And that's never a bad thing.

We wanted to learn a bit more though. We asked our respondents what areas of support they wished their employer offered but doesn't. It should probably not come as a surprise that about a quarter of the respondents wanted a larger team. A shortage of data scientists generally predicts something like that. But what they wanted most of all was more support and direction from their management or executive team (27%).

The Top 10 In-Demand Data Science Skills

Data Science, as a field, is still evolving. Which is to say that what's a best practice today might be replaced by a better practice tomorrow. We looked at nearly 4,000 data science job postings on LinkedIn to find out what skills organizations wanted from their new hires. We ran those job postings through the CrowdFlower platform and had our contributors mark which skills showed up in which jobs.

Here are the top 10 in-demand skills for data scientists:

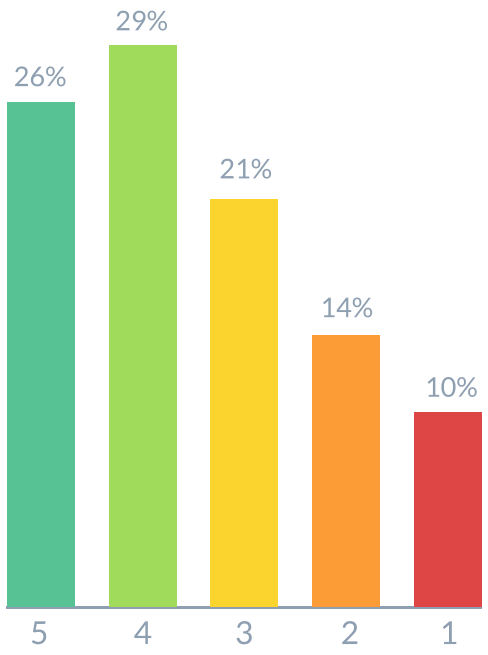
Skills	Job skill appears in	% of jobs with skill
SQL	1987	56%
Hadoop	1713	49%
Python	1367	39%
Java	1287	36%
R	1120	32%
Hive	1099	31%
Mapreduce	768	22%
NoSQL	657	18%
Pig	561	16%
SAS	560	16%

What's Next for Data Science?

What's next, to put it simply, is machine learning. ML has already been adopted in some way, shape, or form, by most of the world's biggest companies, and with big players in the tech space like Google, Microsoft, Amazon, IBM, and Facebook open-sourcing their machine learning tools, the momentum is there for massive advancements.

We wanted to know if our respondents are also focused on machine learning in the next year. We asked our them to rank how important ML was for their organization on a scale of 1-5.

In other words: really crucial. Over half our respondents noted machine learning had significant importance for their companies and their departments, while only one in ten marked that it wasn't very important at all. We expect that 10% to shrink even further next year.



Scientists rating the importance of machine learning

CONCLUSION

As more and more organizations adopt data as a key driver of decision making, the importance of streamlined, well-oiled data science teams is going to remain paramount. But the current status quo probably isn't sustainable. On the one hand, we see a shortage of data scientists while on the other, they're spending too much time cleaning and munging data. This is time that could be much better served doing predictive analysis and building out machine learning practices.

That's not to say that cleaning and labeling data isn't important, of course. Analysis on bad data is a garbage-in, garbage-out sort of scenario. Rather, organizations that want to get the most of their data should aim to fix the problems their teams have now. They should talk to them and find out exactly what takes up their time. By mitigating the effort their teams spend doing janitorial data work, they'll be able to empower their teams to do the valuable tasks that data scientists actually enjoy doing.

About CrowdFlower

CrowdFlower's people-powered data enrichment platform helps data scientists train algorithms to consistently provide the most relevant search results for ecommerce websites. It fills in the gaps in your data by adding product descriptions, IDs, image tags, and other metadata to give you cleaner, more complete data. It can also handle the most intricate product categorization, giving you the kind of advanced taxonomies your product search needs. Leverage the power of the world's largest on-demand workforce, and take your ecommerce search experience to the next level.