CrowdFlower
2015 DATA SCIENTIST REPORT

CrowdFlower

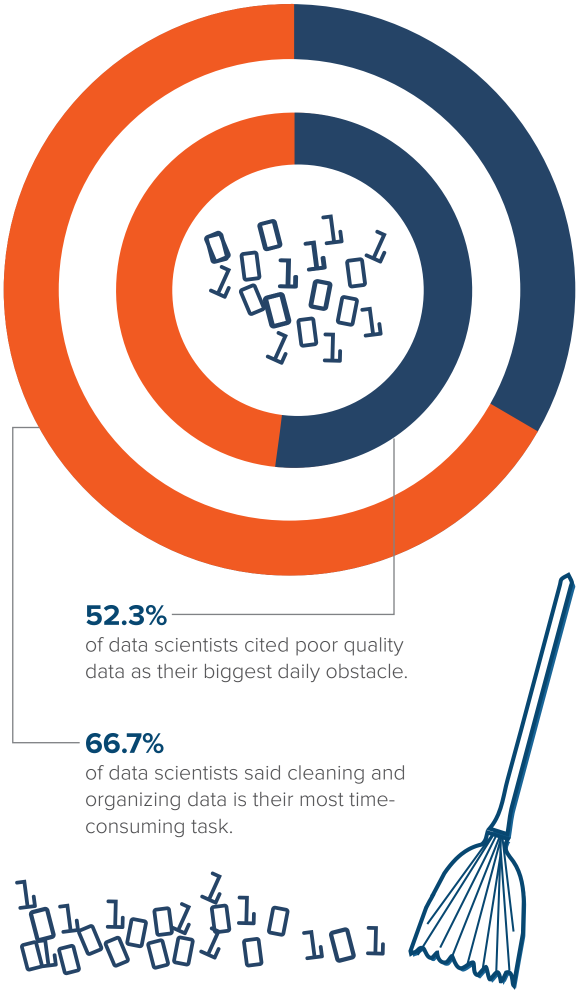
CONTENTS

EXECUTIVE SUMMARY

The volume of data being produced and stored in the world has reached a critical mass. To compete and win, companies must shift their business focus from merely accumulating and maintaining data to making it valuable, rich and usable. Enter the data scientist, who is on the frontlines of the fight to make sense of all this data and capitalize on the rapidly evolving big data landscape.

CrowdFlower surveyed 153 data scientists nationwide between November 2014 and December 2014 to understand the role data scientists play in organizations and whether companies are tapping into the full value these professionals could bring to the table. Several notable findings came to light:

Messy and disorganized data is the number one obstacle holding data scientists back. Cleaning and organizing data is the most time consuming and least interesting part of data scientists' jobs, cited by two-thirds of respondents. This leaves less time for strategic work, with 39.9 percent of respondents reporting they don't have enough time to do analysis. With the amount of information growing exponentially and data scientists remaining in short supply, companies must find better ways to gather, clean and understand the data that drives efforts from predictive modeling to product strategy development as well as business decision-making overall. Simply accumulating more data isn't the answer. It must be enriched.



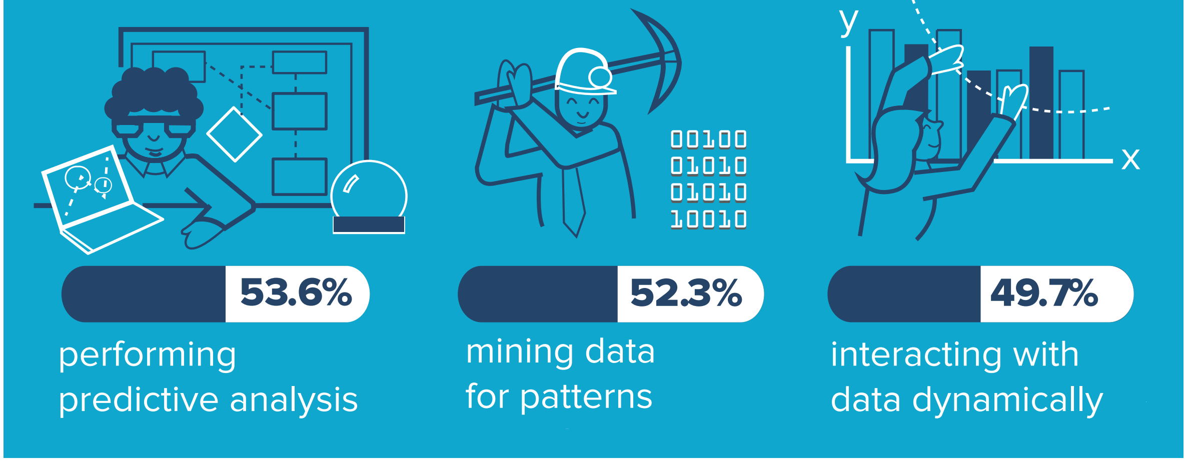
52.3%
of data scientists cited poor quality data as their biggest daily obstacle.

66.7%
of data scientists said cleaning and organizing data is their most time-consuming task.

Relief is in sight with new ways to empower data scientists. Despite these challenges, most data scientists are satisfied in their job. Nearly one-third (30.1 percent) even think it's "totally awesome." Yet companies can do more to empower data scientists by making several key changes, including:

- Taking data cleaning duties off their plate and giving them time back to focus on the interesting and strategic tasks of predictive analysis, mining data for patterns and interacting with data dynamically.
- Providing additional resources to acquire all necessary tools to effectively do the job.
- Setting clearer goals and objectives on projects.

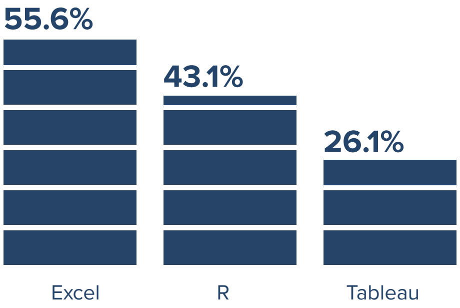
Data scientists are happiest at work while:



Building diverse skills is the key to success as a data scientist today. Contrary to conventional wisdom, earning an advanced degree isn't perceived as the main path for advancement in the data science field. Working with a diverse portfolio of problems, networking and collaborating with fellow data scientists, and gaining strong business acumen are more important than a master's or doctoral degree.

Various traditional, modern and open source technologies comprise data scientists' workflow. The survey found that although Excel is still the most commonly used tool, data scientists also use 47 other tools and languages to do their jobs. Nearly all data scientists use open source software, and tried-and-true open source languages such as R remain major parts of data scientists' toolbox. This report also explores the most in-demand data science skills.

Popular data science tools include:



SURVEY OBJECTIVES

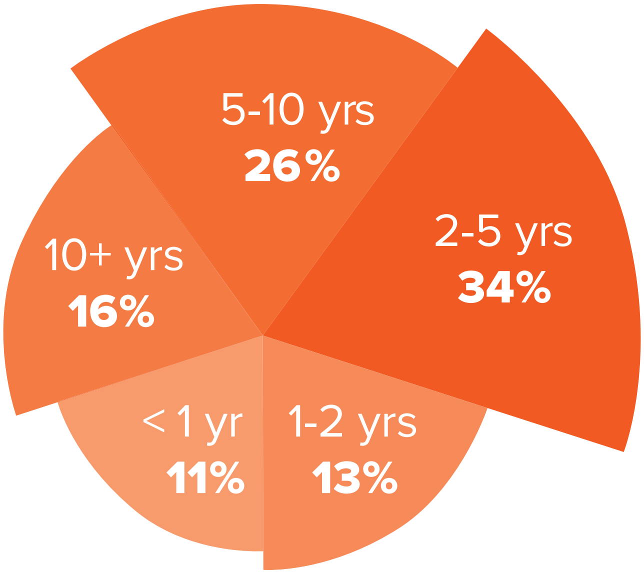
While there is a general industry consensus around the importance of data scientists, there are varying assumptions around how to deploy them to drive measurable business impact, capitalize on the promises of big data, and support their career success.

This survey was designed to uncover what is and isn't working in the field of data science and give organizations visibility into how to build more productive, strategic and happier data science teams. We hope the findings of the CrowdFlower 2015 Data Scientist Report provide actionable insights companies can incorporate into their business plans to drive growth throughout the year and beyond.

SURVEY METHODOLOGY

A total of 153 General Population respondents from CrowdFlower's online research panel completed the survey. Respondents work for companies of varied sizes and sectors, mostly in the United States. All respondents have "data scientist" in their job title or job description on LinkedIn.

Respondents have worked in data science for varying lengths of time. The breakdown is as follows:



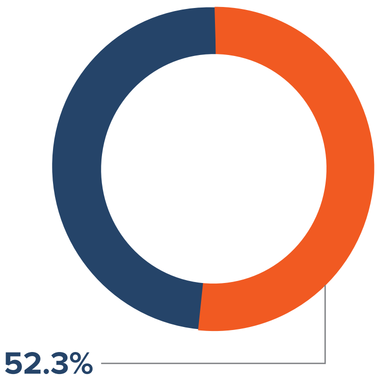
Respondents answered 12 multiple choice questions, three of which included an option to write in a response. The study was fielded from November 2014 to December 2014.

DETAILED FINDINGS

What's Holding Data Scientists Back in 2015

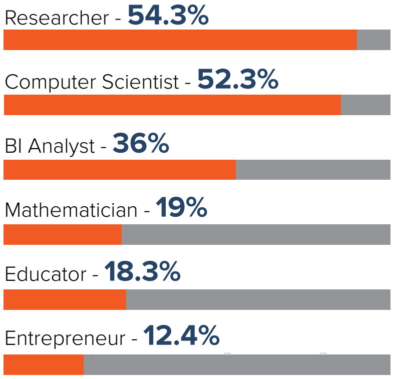
Messy data is a known problem facing data scientists and survey findings demonstrate just how far this challenge reaches. When asked what are the most common obstacles they encounter, the highest percentage of respondents said "spending too much time cleaning data" (57.5 percent), followed by "poor quality data" (52.3 percent).

These results point to the mushrooming data supply and underscore the need to gain data that has relevance and value to drive better business outcomes. Technology limitations, cited by 30.1 percent of respondents, may be part of the hindrance to gaining clean, high-quality data and modeling data effectively.



of data scientists cited poor quality data as their biggest daily obstacle.

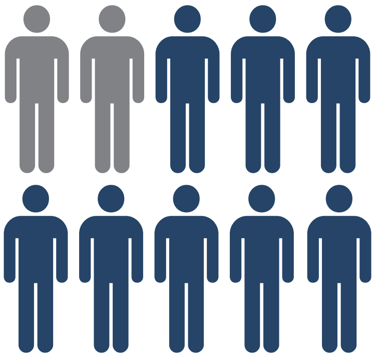
Inside the Day-to-Day Role of a Data Scientist



Data scientists often wear several hats at the same time. Even though they may share the same "data scientist" title, all or part of their job functions vary. When asked to select one or more descriptions that applied to them, most survey respondents identified as researchers (54.3 percent) or computer scientists (52.3 percent). Survey results also show that some data scientists consider themselves to be in business intelligence analyst roles (36 percent), while some identify as mathematicians (19 percent), educators (18.3 percent) and entrepreneurs (12.4 percent).

There's Not Enough of Us: The Data Scientist Shortage

Nearly 80 percent of respondents said there is a shortage of data scientists. This finding correlates strongly with two obstacles uncovered later in the survey: "too much time spent cleaning data" (cited by 57.5 percent of respondents) and "insufficient time to do analysis" (cited by 39.9 percent of respondents). These results suggest that an increase in qualified data scientists would enable companies to balance workload and improve the overall breadth and depth of their data science capabilities.



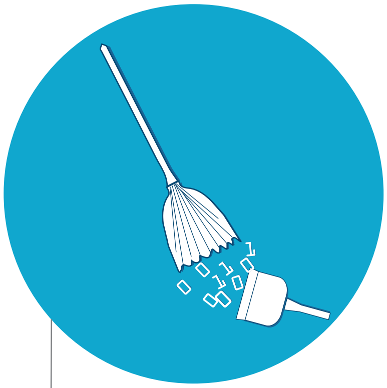
These results also corroborate findings around the high demand for data science skills and the shortage of people who actually have them. Notably, Accenture reported in 2013 that 80 percent of new data scientist jobs created between 2010 and 2011 have not been filled.[1]

Further, McKinsey Global Institute estimates that by 2018, “the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.”[2]

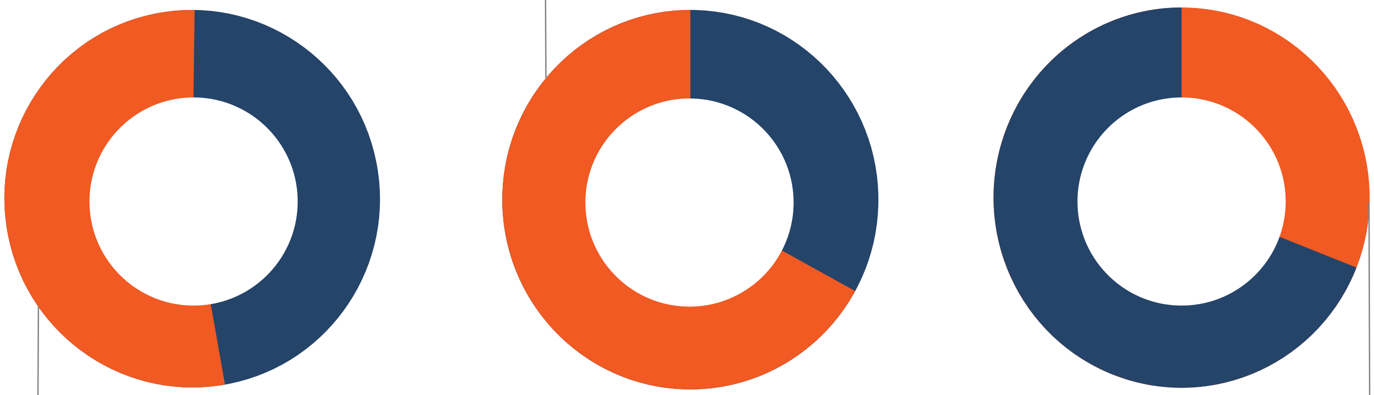
Most Time-Consuming Tasks for Data Scientists

When asked to cite up to two of their most time-consuming tasks, the highest percentage of respondents (66.7 percent) said “cleaning and organizing data.” This finding correlates strongly with the most common obstacle respondents reported earlier in the survey—“too much time spent cleaning data,” cited by 57.5 percent of respondents.

“Collecting data sets” was cited by 52.9 percent of respondents as one of their most time-consuming tasks, coming in a close second to cleaning and organizing data. Offloading these time-devouring tasks from data scientists’ plates represents a significant opportunity for companies to gain efficiencies and give data scientists more time for the strategic work they also actually enjoy doing.



66.7%
said cleaning and organizing data is one of their most time-consuming tasks



52.9%
said collecting data sets is one of their most time-consuming tasks

30.7 %
listed mining for patterns in data among their most time-consuming tasks

[1] Accenture, “The Team Solution to the Data Scientist Shortage,” 2013
[2] McKinsey Global Institute, “Big Data: The Next Frontier for Innovation, Competition, and Productivity,” 2011

What Data Scientists Love and Hate About Their Job

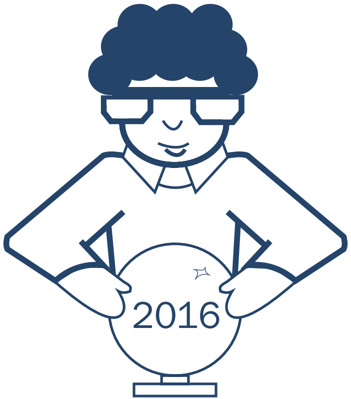
Job Satisfaction

Nearly 79 percent of respondents are satisfied in their job, with almost one-third (30.1 percent) finding it "totally awesome." This finding indicates that data scientists are generally happy in their position, despite the challenges they face.

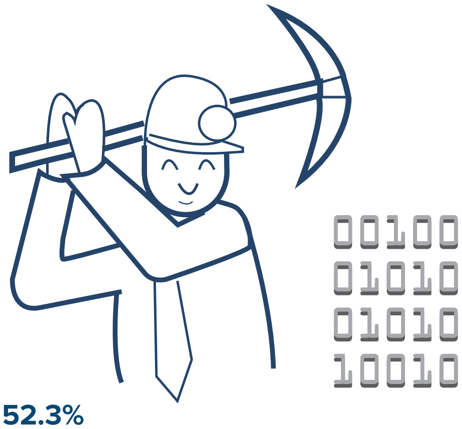
At the same time, the survey suggests ways data scientists can find greater satisfaction, as described later in this report in the "How Organizations Can Empower Their Data Science Teams" and "How to Empower Data Scientists" sections.

Most Interesting Tasks

When asked to choose up to two tasks that they find most interesting, respondents cited predictive analysis (53.6 percent) and mining data for patterns (52.3 percent). Of note, in another part of the survey, respondents also cited these same tasks as the least time consuming. These findings suggest that data scientists are most interested in tasks that activate curiosity and innovation and are core to their skill set and value.



53.6%
of data scientists selected predictive analysis as one of their most interesting tasks.



52.3%
cited mining data for patterns as one of their most interesting tasks.

Least Interesting Tasks

Messy data emerged a third time in the survey when respondents were asked what tasks they find least interesting. The No. 1 response to the question asking what data scientists find to be the least interesting task: cleaning and organizing data, cited by 66 percent of respondents. **In aggregate, survey results demonstrate that messy data is A) the biggest obstacle data scientists face, B) the most time-consuming task and C) the least interesting task.**

The second least interesting task reported was collecting data sets, cited by 51 percent of respondents. This finding correlates with the results revealed in the question about the most time-consuming tasks, indicating that time-consuming tasks are, perhaps unsurprisingly, perceived to be the least interesting.

Data Scientists’ Favorite Tools and Technologies

Respondents named a total of 48 technologies that they use in their daily workflow. (Ten of these technologies were provided to respondents in a list from which to choose. Respondents cited 38 additional technologies through the write-in response option.) Excel was cited the most (by 55.6 percent of respondents). Python, another time-tested technology, was mentioned in 22 write-in responses. These results suggest that traditional tools form the backbone from which many data scientists execute other tasks that require more purpose-built technologies.

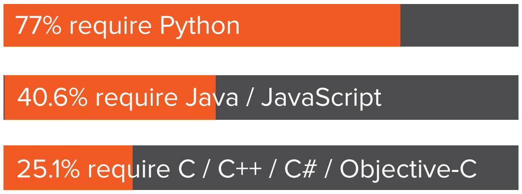
Open source software is also a frequent component of data scientists’ workflow, as reported by 98 percent of respondents. This finding correlates with another result from the survey: the open source language R was identified as the second most used technology, with 43.1 percent of respondents citing it as a tool in their toolbox. These results demonstrate data scientists’ affinity for open source software.

In-Demand Data Science Skills

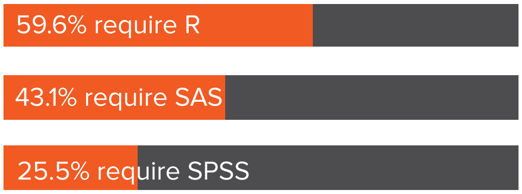
To gain a deeper understanding of the data science skills that are in most demand by organizations, CrowdFlower turned to its own data enrichment platform to collect and analyze 1,024 LinkedIn job postings for data scientist positions worldwide.

CrowdFlower found that the two top skills companies are looking for in a data scientist are programming and coding, seen in 55.3 percent of job postings, and statistical tools, seen in 52.1 percent of job postings. CrowdFlower then dissected each of these top skills to understand the most in-demand sub-skills within them. Results are shown in the graphs below.

Among jobs that require programming and coding:



Among jobs that require statistical tools:



All of these six sub-skills, except C / C++ / C# / Objective-C, were reported in CrowdFlower’s data scientist survey itself as tools that respondents use. This confirms that companies rely heavily on data scientists to conduct programming, coding and statistical work.

CrowdFlower’s data scientist survey also reveals that respondents use tools that don’t require programming and coding, such as Excel and visualization tools, underscoring the breadth of technologies with which data scientists must be comfortable.

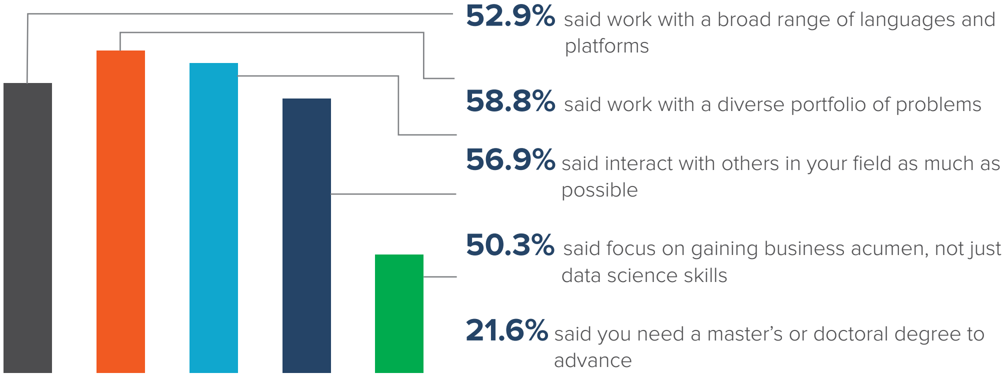
Technologies Data Scientists Use

- Alteryx
- Cassandra
- Clojure
- Cloudera Enterprise
- D3
- Elasticsearch
- Excel
- GraphLab (now Dato)
- Hadoop
- Hortonworks Hadoop Infrastructure
- IBM PureData System
- IBM SPSS Modeler
- In-house technology
- Java
- JavaScript
- Julia
- Kafka
- Looker
- Medidata Rave
- MongoDB
- Neo4j
- NLTK
- NumPy
- Orange
- pandas
- Pentaho
- Pig
- PostGIS
- PostgreSQL
- Python
- R
- Redis
- Redshift
- SAP BusinessObjects
- SAS
- SAS Business Intelligence
- SAS Visual Analytics
- Scala
- scikit-learn
- SciPy
- Spark
- SPSS
- SQL
- Stata
- Storm
- Tableau
- Vertica
- Vowpal Wabbit

Advice to the Next Generation of Data Scientists

Survey results indicate that developing a diverse background is important for breaking into the data science field. When asked what advice they'd give new industry professionals, the highest percentage of respondents said "work with a diverse portfolio of problems" (58.8 percent).

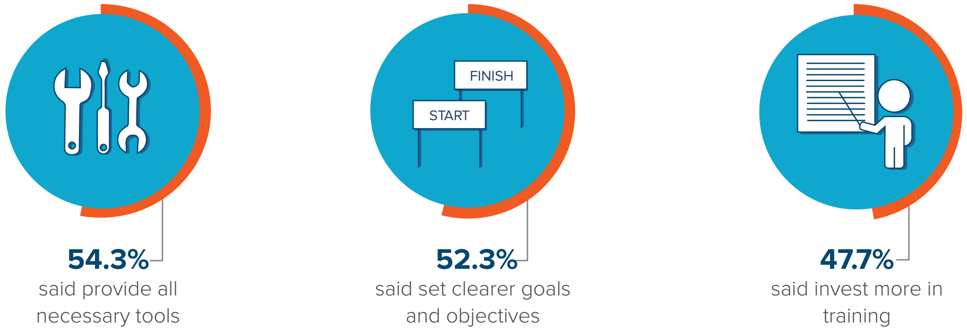
In addition, 56.9 percent of respondents advised to "interact with others in your field as much as possible" and 50.3 percent said to "focus on gaining business acumen, not just data science skills," while only 21.6 percent said "you need a master's or doctoral degree to advance." These results suggest that hands-on, real-world experience and practical knowledge can take data scientists much further in their careers than an advanced degree. Data scientists' advice was as follows:



How Organizations Can Empower Their Data Science Teams

Survey results indicate several ways organizations can create stronger data science teams by empowering data scientists. Top solutions were "to acquire all necessary tools to effectively do the job" (cited by 54.3 percent of respondents), followed closely by "set clearer goals and objectives on projects" (cited by 52.3 percent of respondents).

Also high on data scientists' list was "invest more in training and development to help team members continually grow their capabilities" (cited by 47.7 percent of respondents). This correlates with the advice to new data scientists that respondents provided in the survey: "focus on gaining business acumen, not just data science skills."

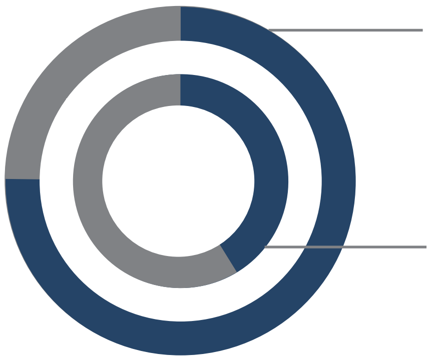


TOP TAKEAWAYS

What It’s Like to Be a Data Scientist

The survey results suggest several key qualities about what drives data scientists in their profession and what a typical day is like for them:

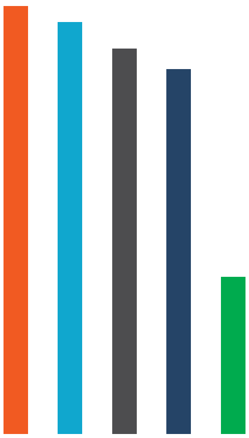
Data scientists are in the field because they’re passionate about their work, not necessarily because it’s a hot industry. Most respondents entered the data science field before it hit the mainstream in 2012. More than 75 percent of respondents have been working in data science for 2+ years, with 41.2 percent having been in the field for 5+ years.



75.2% of respondents have been working in data science for 2+ years

41.2% of respondents have been working in data science for 5+ years

Data scientists recognize the opportunity to add value by connecting the dots between data and business outcomes. Some of the top career success factors that respondents reported include “work with a diverse portfolio of problems” (58.8 percent) and “focus on gaining business acumen, not just data science skills” (50.3 percent).



● **58.8%** said work with a diverse portfolio of problems

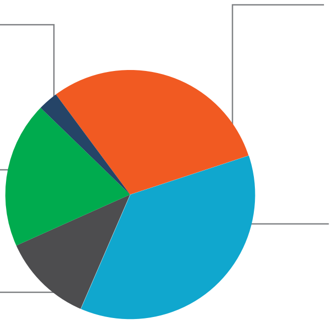
● **56.9%** said interact with others in your field as much as possible

● **52.9%** of respondents recommended working with a broad range of languages and platforms

● **50.3%** said focus on gaining business acumen, not just data science skills

● **21.6%** said you need a master’s or doctoral degree to advance

Data scientists are generally satisfied in their jobs. Nearly 79 percent of respondents are satisfied in their job, with 30.1 percent finding it “totally awesome.”



2.6% said “I want to find something else to do”

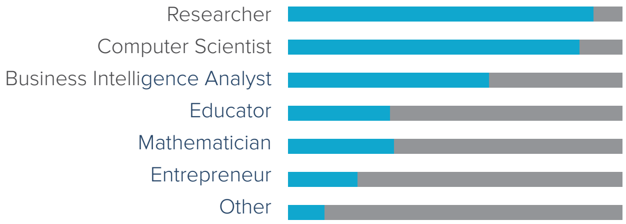
19% said “I wish it was better”

11.6% said “It’s just okay”

30.1% of respondents said their data science role is “totally awesome”

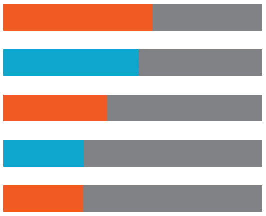
36.6% said “It’s good”

Data scientists identify themselves as holding multiple roles. Most respondents identified as researchers (54.3 percent) or computer scientists (52.3 percent). Other roles with which they identified include business intelligence analysts (36 percent), mathematicians (19 percent), educators (18.3 percent) and entrepreneurs (12.4 percent).



The Challenges Facing Data Scientists Today

When asked explicitly about the most common obstacles they encounter, respondents said:



Too much time spent cleaning data (57.5 percent)

Poor quality data (52.3 percent)

Insufficient time to do analysis (39.9 percent)

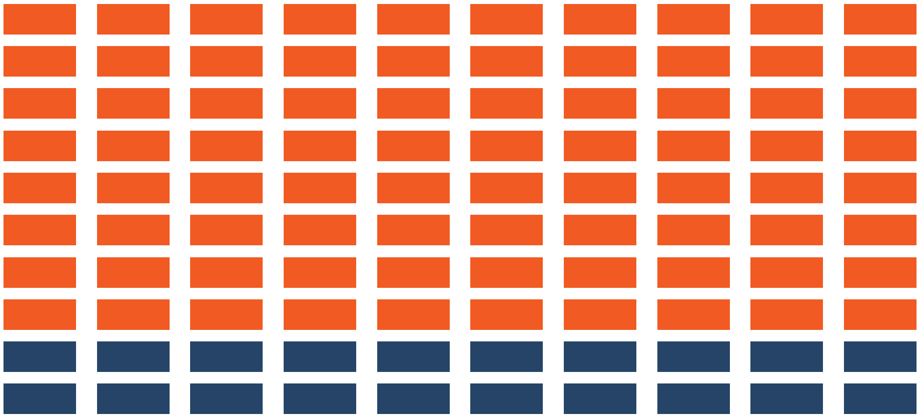
Limited ability to model collected data effectively (30.7 percent)

Limitations of technology and tools at my disposal (30.1 percent)

The survey results also suggest several other challenges facing data scientists beyond messy data, including:

Lack of team members to help turn data into valuable insights. Nearly 80 percent of respondents believe there's a shortage of data scientists, which indicates a lack of professionals to gather, clean and enrich data to drive better business decision making.

79.7%
of respondents felt there is a shortage of data scientists

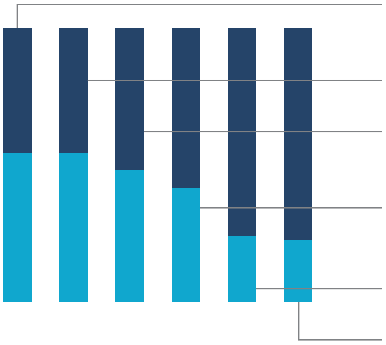


Imbalanced workload.
Two-thirds (66.7 percent) of respondents said cleaning and organizing data is among their top two most time-consuming tasks. This leaves less time for strategic work, with 39.9 percent of respondents reporting they don't have enough time to do analysis.

Fuzzy project goals.
When asked how to empower data scientists, most respondents (52.3 percent) said their organization needs to set clearer goals and objectives on projects.

How to Empower Data Scientists

When asked explicitly about how to empower the data science team, respondents wanted their companies to:



54.3% - Provide additional resources to acquire all necessary tools to effectively do the job

52.3% - Set clearer goals and objectives on projects

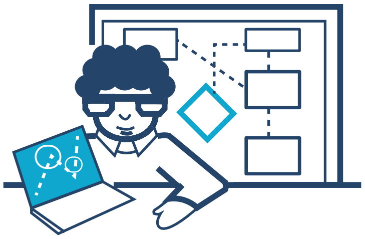
41.2% - Allow for realistic timing expectations for completion of projects

47.7% - Invest more in training and development to help team members continually grow their capabilities

23.5% - Increase headcount on the data science team

22.2% - Offload data cleaning responsibilities from the data science team

Survey results indicate another key way to empower data scientists: letting them focus on interesting and strategic tasks. The top three most interesting tasks respondents listed were predictive analysis, mining data for patterns and interacting with data dynamically.



Predictive analysis, in particular, also imposes some of the smallest demands on data scientists' time, with only 11.1 percent of respondents reporting it as a time-consuming task. These results suggest that data scientists are not doing the type of work they are most passionate about—extracting usable, rich data to help organizations predict trends and behavior patterns.

CONCLUDING THOUGHTS

For decades, companies have been amassing and storing data. While this has paved the way for the big data era, companies are now awash in a glut of data, yet are experiencing a poverty of actionable insights and intelligence.

The status quo is not sustainable for companies that want to cash in on the promise of big data. *They must shift from simply collecting big data to acquiring and using* ***rich data****, i.e., data that is complete, accurate and fully deduplicated.*

Getting there means giving data scientists more time and resources to focus on the strategic and analytical aspects of their job. They must be able to hang up their data wrangler and data janitor hats and be empowered to live up to their full strategic potential as providers of actionable data insights and business decision support.

Many companies have invested significantly in their big data infrastructure and are betting on data to solve a lot of their challenges. Optimally armed data scientists are critical to these efforts, making it imperative to set them up for success in both their careers and their roles within organizations.

LEARN MORE ABOUT CROWDFLOWER

Founded in 2009, CrowdFlower is the leading data enrichment platform for data scientists. Its quality-control technology is the most accurate and fastest way to collect, label and clean data from an on-demand workforce. The platform automates the management of these online contributors to tackle tasks that require human intelligence—like search relevance tuning, data categorization, image annotation, metadata creation, sentiment analysis, transcription and de-duplication.

Backed by Trinity Ventures, Bessemer Venture Partners, Harmony Partners and Canvas Venture Fund, CrowdFlower has over 150 customers, including Unilever, Autodesk, eBay, Edelman, YP.com and VoiceBox.

For more information, visit www.crowdflower.com or connect with us on Facebook, Twitter, LinkedIn or Google+.