2017

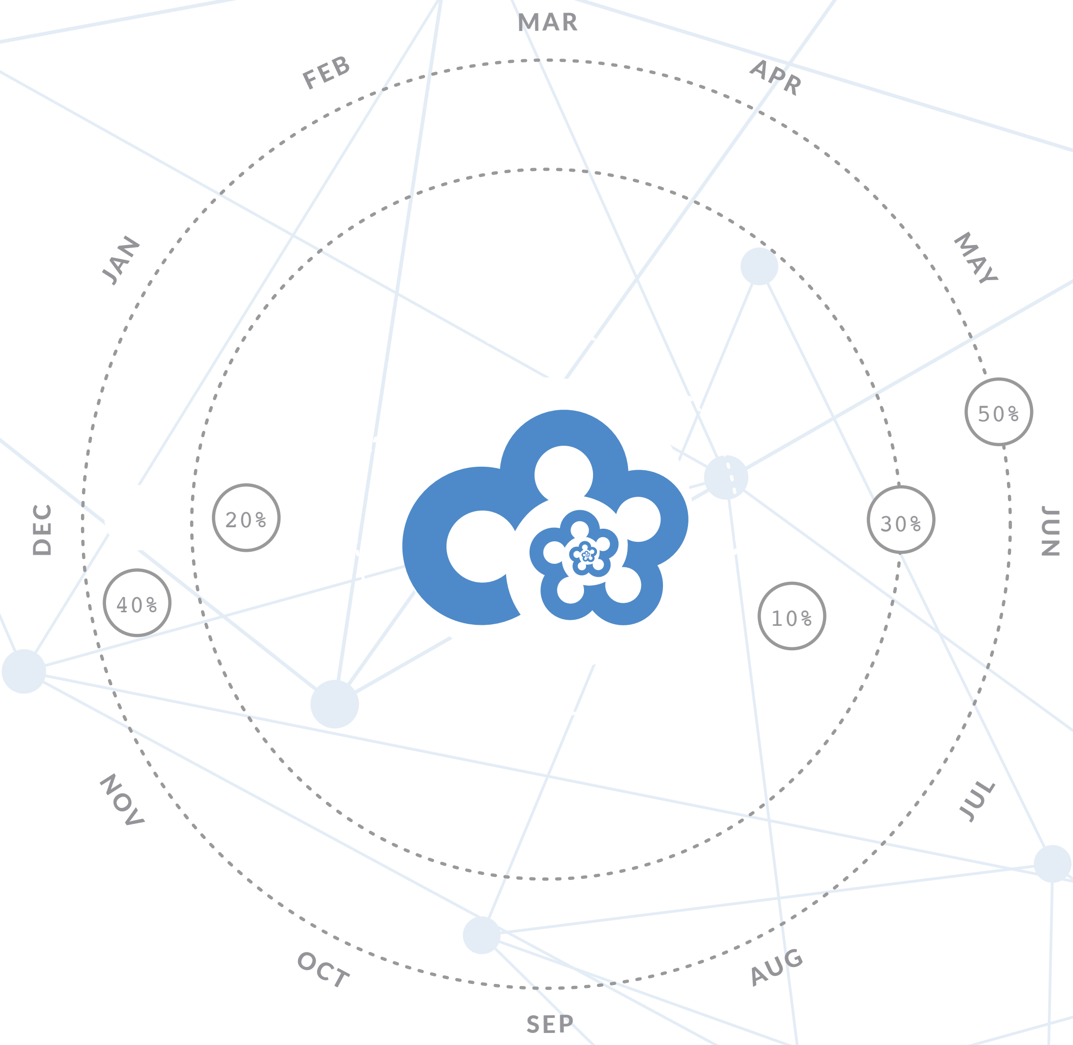
DATA SCIENTIST

REPORT

BROUGHT TO YOU BY

CrowdFlower

DATA SCIENTIST REPORT 2017
OVERVIEW



These past 12 months have been busy in the world of data science. AI and machine learning have become hot topics — not only in tech circles — but in mainstream business conversations. CEOs are asking direct reports to develop AI plans, companies utilizing machine learning are gaining significant competitive advantage, and data scientists are more in demand than ever.

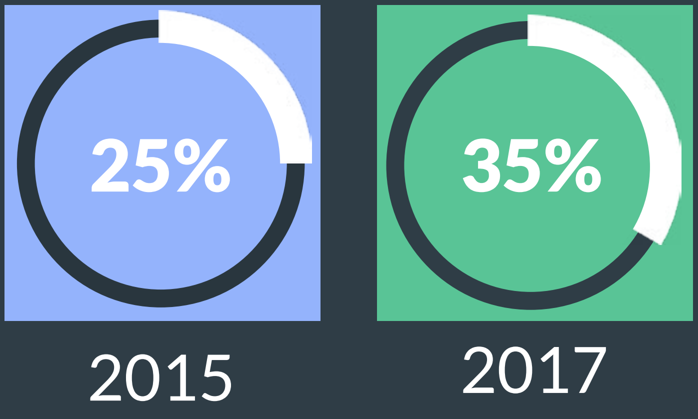
This year's report includes some updates on data from years past, looking at how data scientists spend their time, job satisfaction levels and obstacles to success. We've also included some data on data itself. What kinds of data sources data scientists work with, how much, and where it comes from. As well, in this year's report, we dive into the relationship between data and algorithms. Finally, as AI becomes more pervasive — not just in scientific and technical communities, but into the common vernacular — we asked data scientists to comment on some of the biggest trends in AI, from self-driving cars to concerns about the ethics behind AI and automation.

STATE OF DATA SCIENTISTS

In previous years' reports, we've discussed the dearth of data scientists. Demand still appears to outpace availability with the vast majority of data scientists receiving recruiting calls on a regular basis. Despite a significant misalignment between how data scientists want to spend their time versus how they are actually spending time (yup, still stuck in those 'janitorial tasks') most are happy in their jobs — and happiness appears to be growing year over year.

DATA SCIENCE GROWS UP.

While the term 'data scientist' is relatively new, the high demand and popularity has already resulted in many more budding data scientists. Newcomers abound. In 2015, 25% of data scientists had been in their roles for less than 2 years. Two years later, that number has increased to 35%, a clear indication of many new data science graduates and the now 551 colleges worldwide offering degrees in data science.[1]

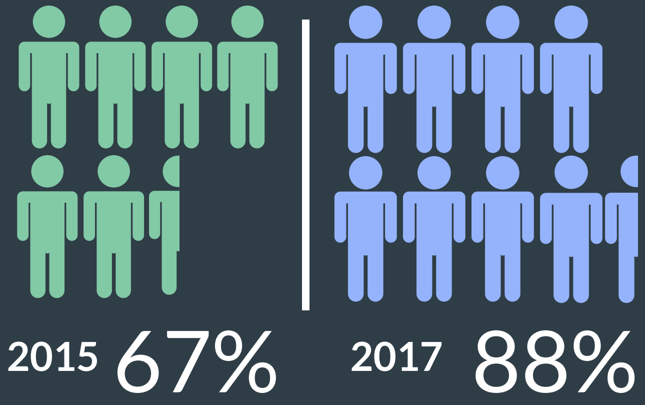


Data scientists in their role for less than 2 years

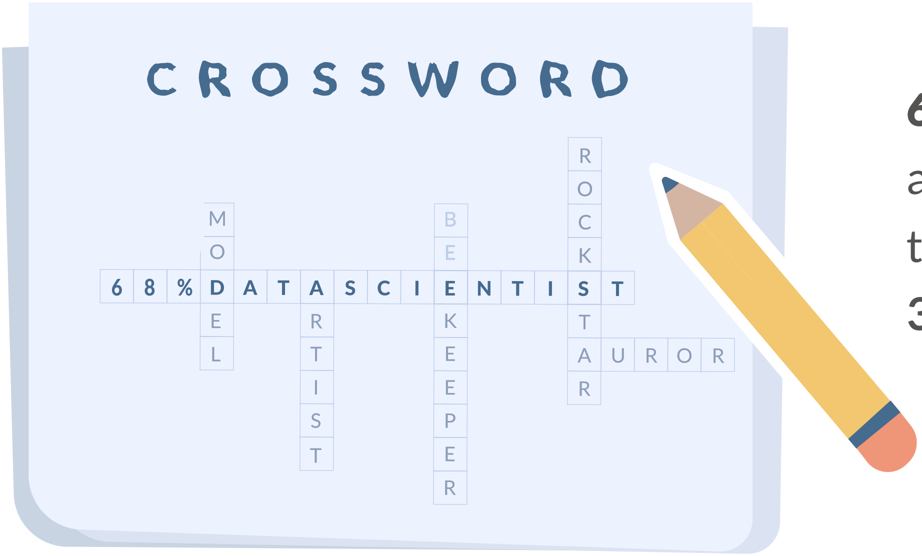
It could be all of that youthful optimism, but overall happiness amongst data scientists is growing as well with those claiming to be 'Happy or Very Happy' in their roles increasing over 20 percentage points in the past 2 years.

But hey, who wouldn't be happy working in what Harvard Business Review dubbed (in 2012), 'the sexiest job of the 21st century'? While the distinction has been referenced somewhat exhaustively — even appearing as the answer to a New York Times crossword puzzle clue earlier this year — 64% of data scientists agree that they are working in this century's sexiest job. The remaining 36% provided us with an array of answers as to what might be deemed 'sexier' with responses ranging from movie star to astronaut to researcher, model, fashion designer, artist, beekeeper, rockstar, auror, and even one data scientist dreaming about leaving data behind to become 'Lady Gaga's dresser'.

Data scientists who are **Happy or Very Happy**



[1] Source: http://datascience.community/colleges



64% OF DATA SCIENTISTS agree that they are working in this century's sexiest job **(but 3% would rather be rockstars)**

WHAT KEEPS **DATA SCIENTISTS HAPPY?**

(and why aren't they doing more of it?)

Still, employers shouldn't take that happiness for granted. Data scientists remain in high demand. Nearly 90% of data scientists (89%) are contacted at least once a month for new job opportunities, over 50% are contacted on a weekly basis, and 30% report being contacted several times a week.

2365
Sent Messages
34
Drafts
Flagged
Trash
Downloading

89%
Data scientists contacted at least once a month for new job opportunities

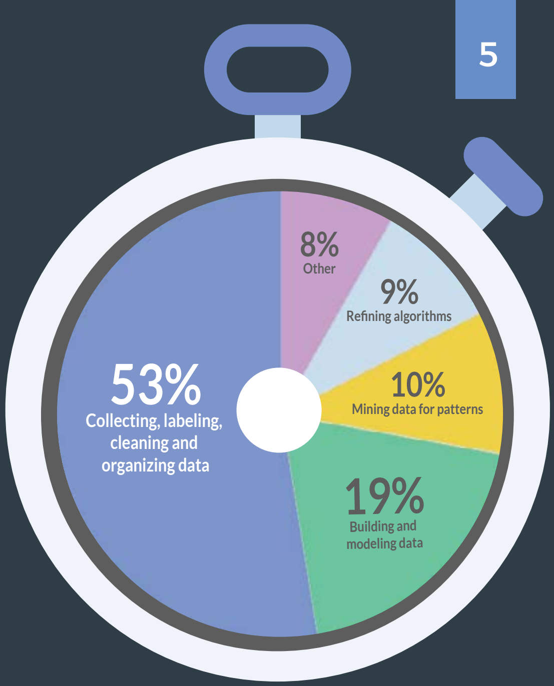
50%
contacted on a weekly basis

30%
report being contacted several times a week

WHAT KEEPS **DATA SCIENTISTS** HAPPY?

(and why aren't they doing more of it?)

What activity takes up most of your time?



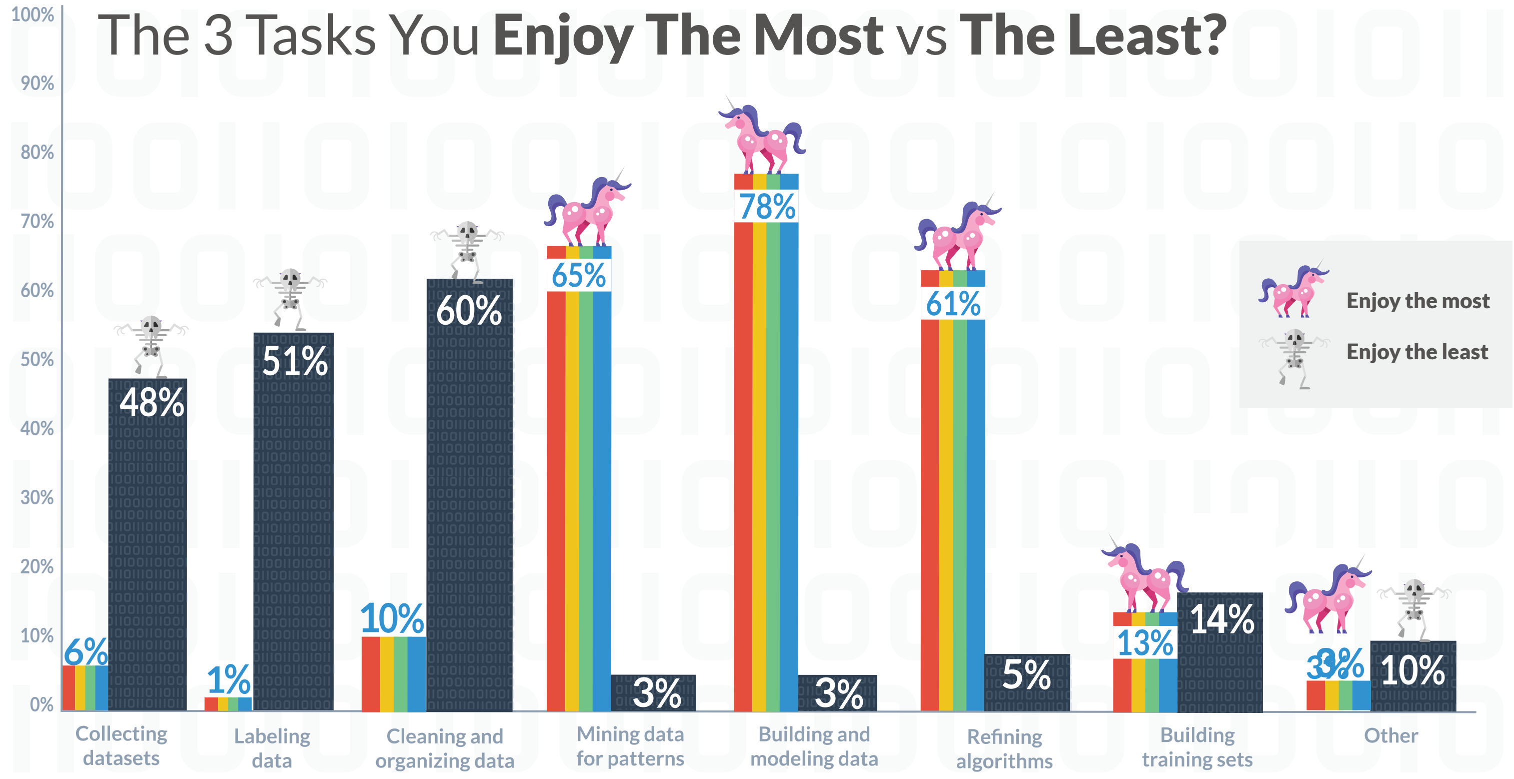
In what appears to be instinctively the inverse of optimal (and a continued trend from previous years), data scientists are spending an inordinate amount of time on the tasks they dislike the most and little time on activities they enjoy.

Data scientists are happiest building and modeling data, mining data for patterns and refining algorithms. These three more cerebral tasks rank nearly 8 times higher in popularity amongst data scientists than more 'janitorial tasks' yet a mere 19% of data scientists report spending most of their time on the top ranked activity — 'Building and Modeling Data'.

As with previous years, 'Janitorial Tasks' rate distinctively low on data scientists list of preferred tasks. A whopping 60% list 'Cleaning and Organizing Data' as one of their least 3 favorite tasks, 51% complain about 'Labeling Data' and 48% placed 'Collecting Datasets' as one of their top 3 dreaded ways to spend time.

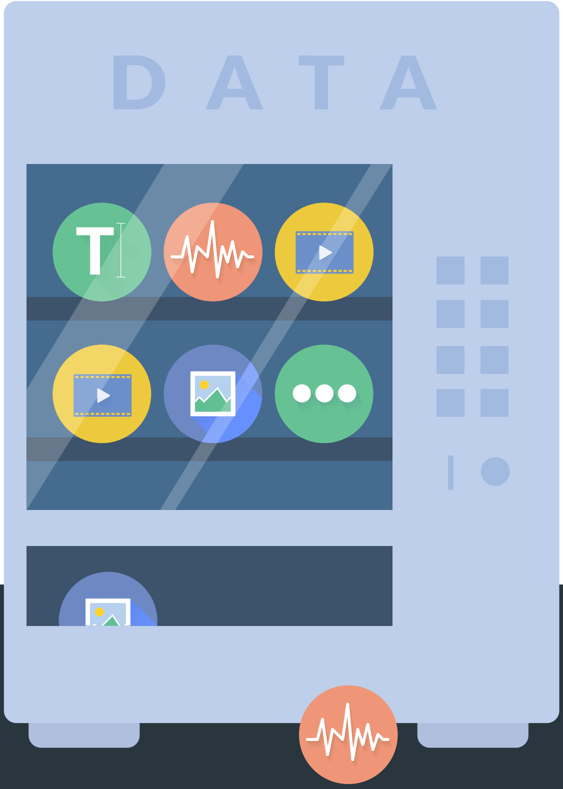
Conversely, these same data 'Janitorial Tasks' take up the most time. Fifty-three percent of our data scientist respondents are spending the most time on the tasks they dislike the most with 45% spending most of their time on the overall least favorite task: 'Cleaning and Organizing Data'.

The 3 Tasks You **Enjoy The Most** vs **The Least?**

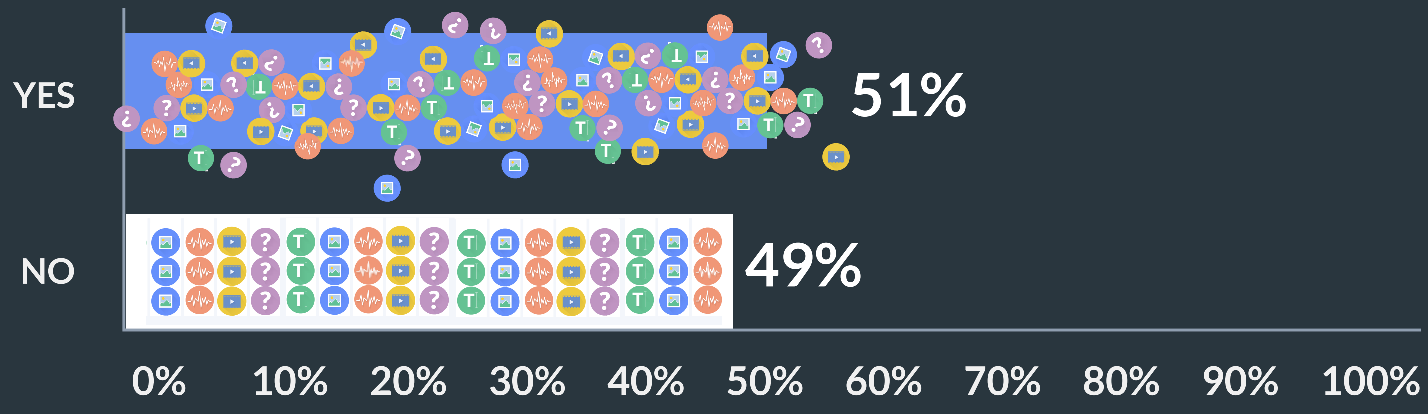


MORE DATA
ON THE DATA

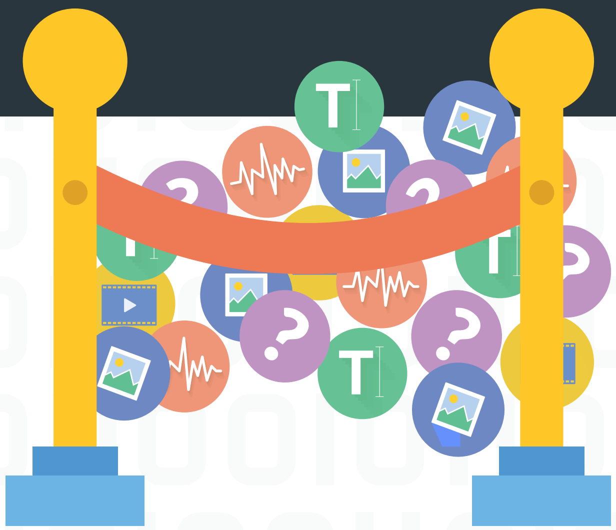
At the heart of any data scientist's job is the data. In this year's survey we decided to take a deeper look at the data itself: how data scientists feel about it, obtain it, categorize it, and how much of it there is. In 2017, data scientists are looking at more data than ever before, a bulk of which is unstructured data in various formats, such as text and images. However, 'Access to Quality Data' was cited as the #1 roadblock to success for data scientists with 50% ranking it within the top 3 obstacles to achieving their goals.



Does a signifcant amount of your work involve UNSTRUCTURED DATA?



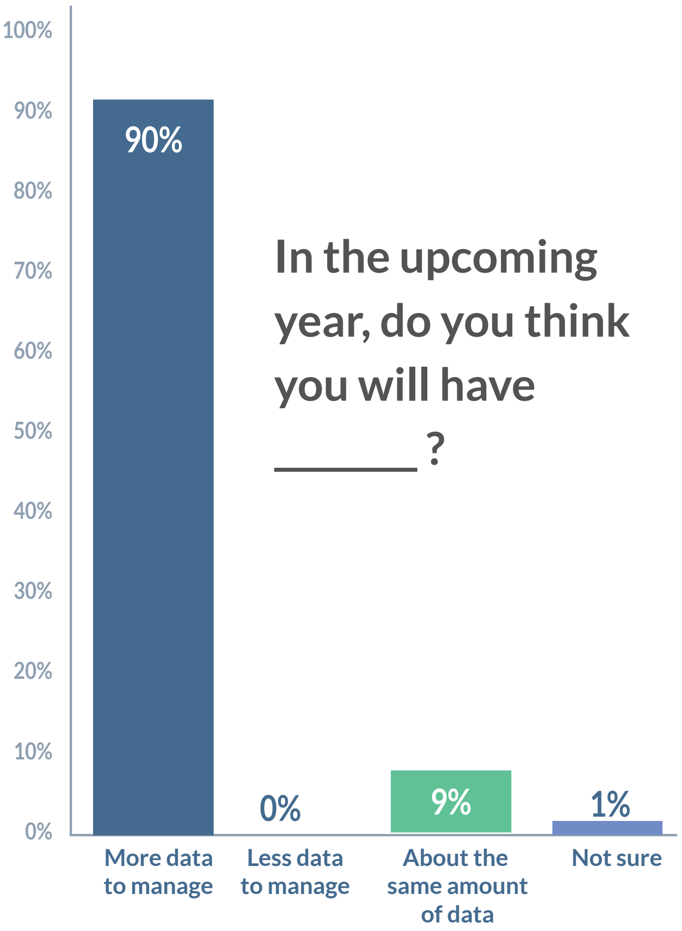
'Access to quality data' was cited as the **#1 roadblock to success** for AI initiatives.



A DELUGE OF DATA

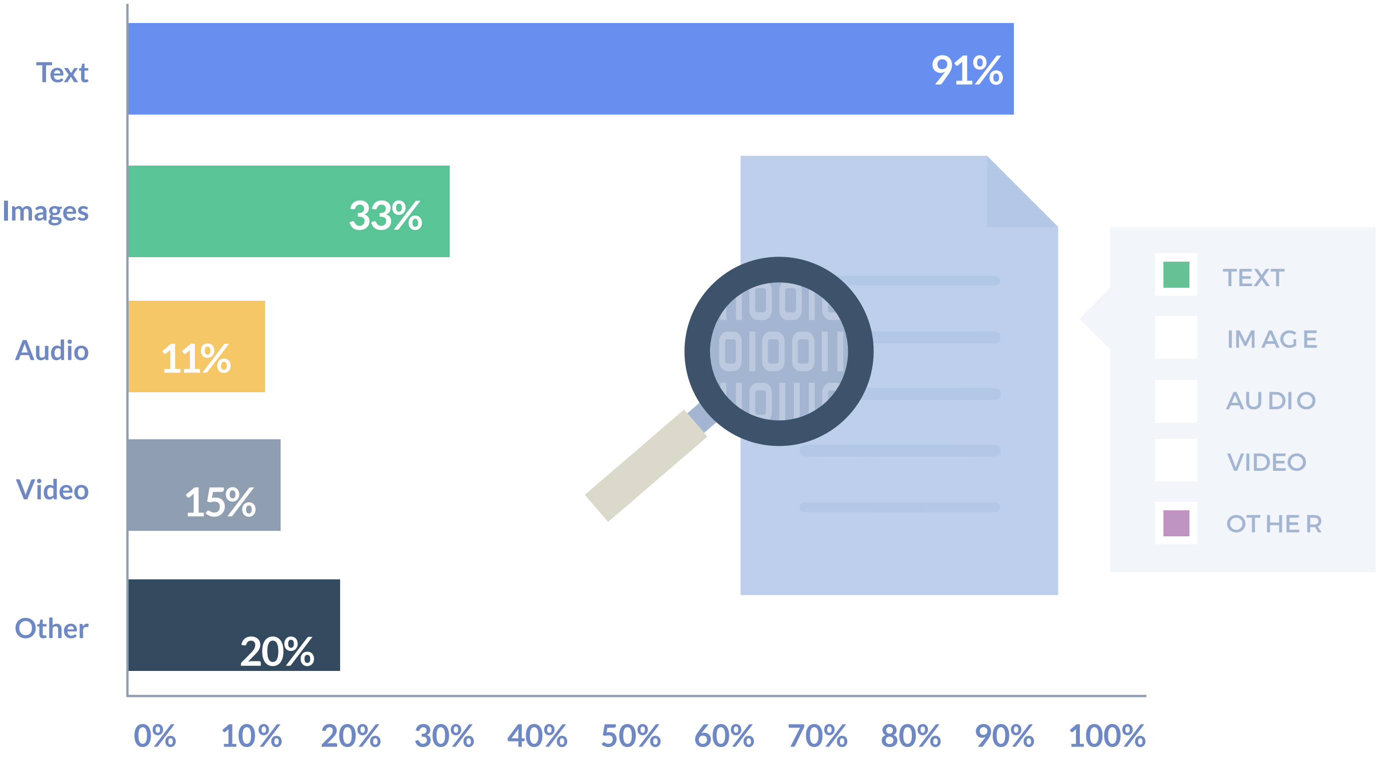
The sheer amount of data is certainly not the issue. Ninety percent of survey respondents predict they will have more data to contend with in 2017 and ZERO percent believe the amount of data will go down. One challenge is certainly the amount of unstructured data not only available, but critical to the success of multiple projects. Slightly more than half of our data scientists (51%) are spending a significant amount of time working with unstructured datasets.

According to Gartner, Inc. a research and advisory firm, unstructured video and image data, derived from the proliferation of cameras and sensors, is expected to exceed 80% of all internet traffic by 2019 and, by 2020, 95% of video/image content will never be viewed by humans but will have been analyzed by machines.[1] This significant uptick in visual data was reflected in our survey responses as well. While it's no surprise that almost all respondents are working with text data, a good portion of data scientists are utilizing images (33%) and video (15%) as well.



In the upcoming year, do you think you will have ______?

[1] Gartner Innovation Insight for Video/Image Analytics 2016, Nick Ingelbrecht and Melissa Davis, September 22, 2016

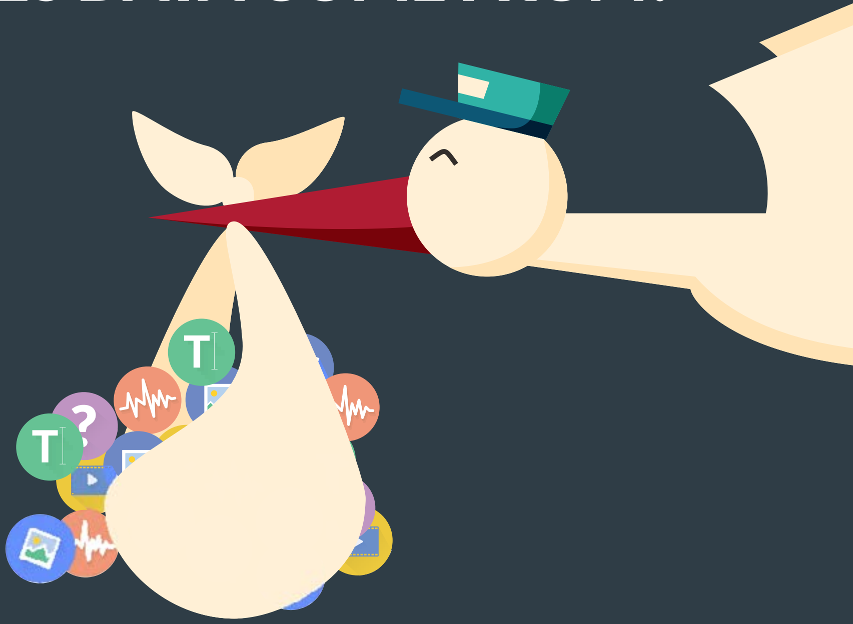


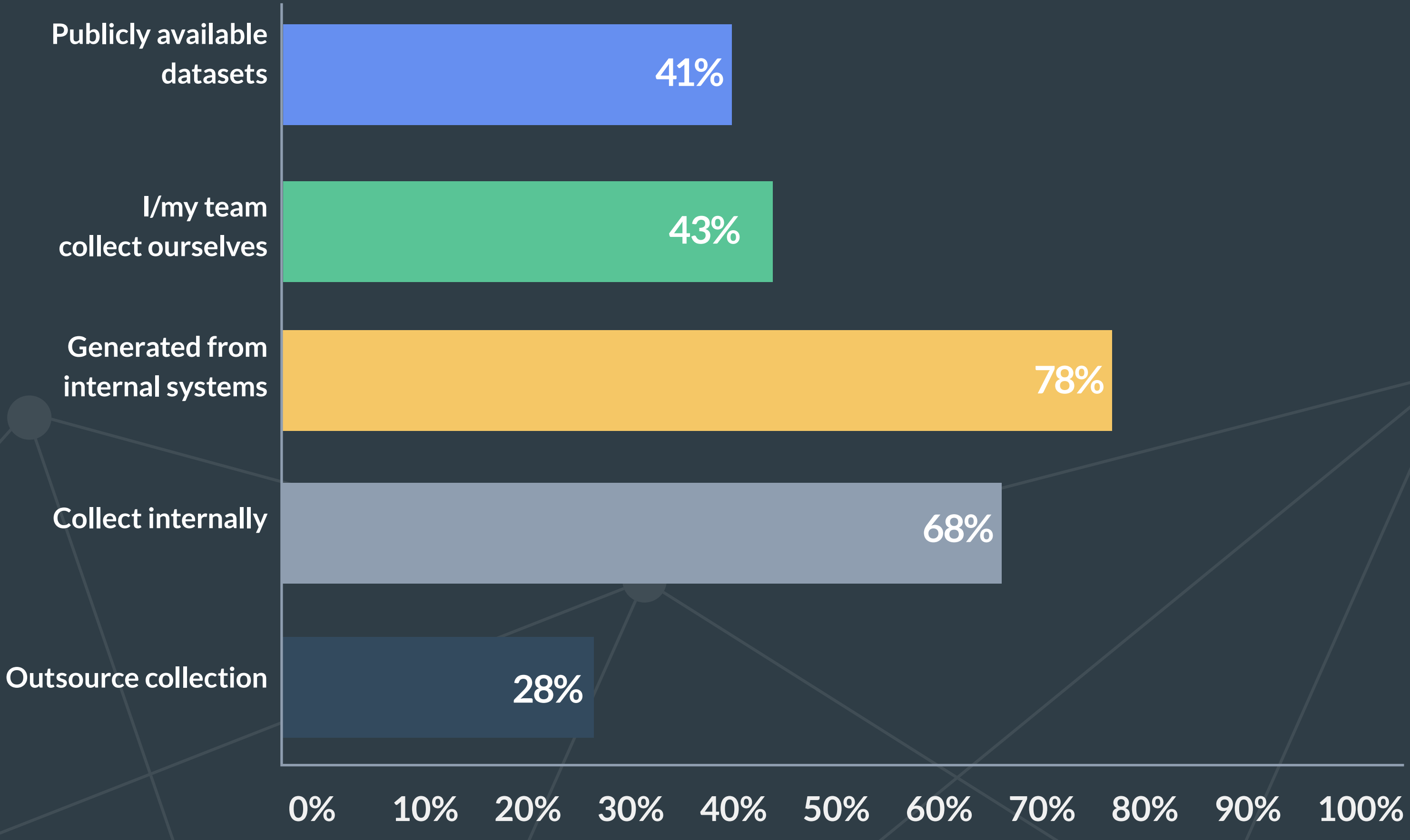
QUALITY OVER QUANTITY

While there's no shortage of data, access to quality data is definitely an issue. Specifically when it comes to AI projects, 51% of respondents listed issues related to quality data ('getting good training data' or 'improving the quality of your training dataset') as the biggest bottleneck to successfully completing projects.

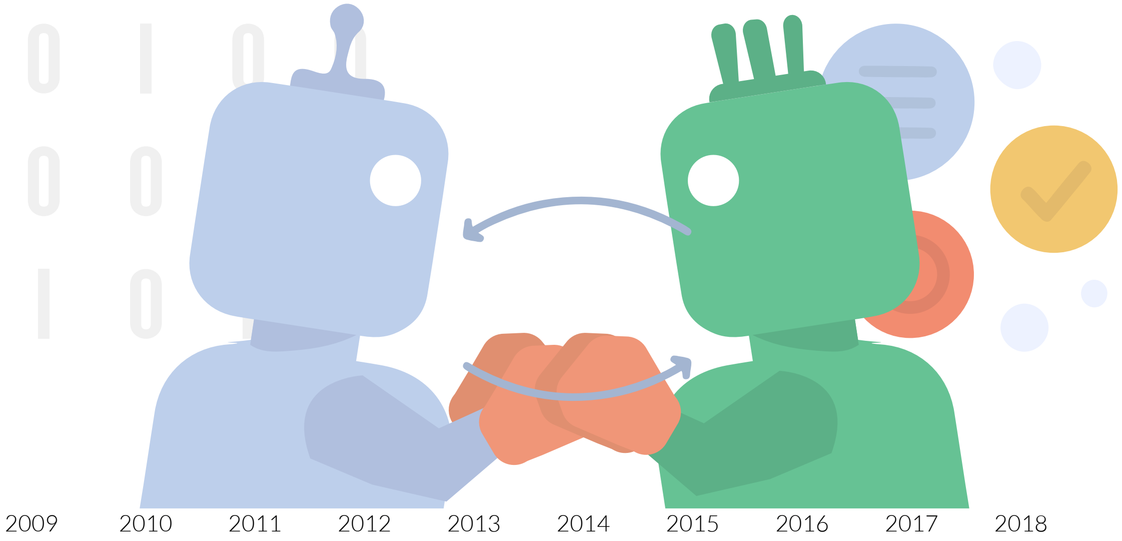
MOMMY, WHERE DOES DATA COME FROM?

As a first step, we took a look at the most popular sources of data for data scientists. While the majority of data scientists utilize data generated from internal systems (78%), over half of them get data from at least 3 different sources including manual internal collection, publicly available datasets, and outsourcing. Finally, while 48% list collecting data as one of their 3 least favorite tasks, 43% of data scientists are doing just that — collecting data themselves.





TRAINING DATA VS. ALGORITHMS

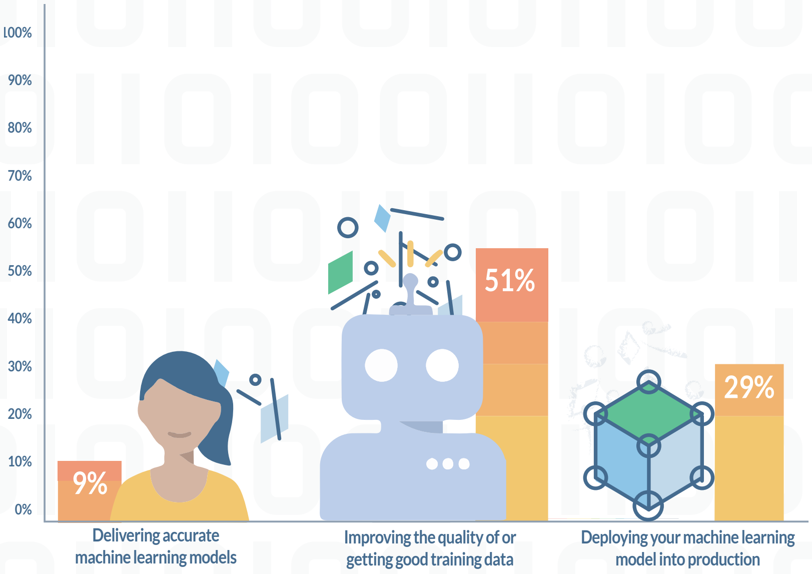


In 2016, with so much focus and fanfare on the promise of AI, the concept of 'algorithms' made its way out of ivory towers and into the common vernacular. You didn't have to be a mathematician to know that the mighty algorithm helped predict NBA champions, estimate crop yields, or predict the results of elections. In this season of AI, algorithms took their place as the belle of the ball.

On the contrary, media seemed to place relatively little emphasis on training data, choosing instead to glorify an almost mythical notion that algorithms magically process huge amounts of data. The reality — it's all about the data. When asked to identify the biggest bottleneck in successfully completing AI projects, over half the respondents named issues related to training data 'Getting good quality training data or improving the training dataset' while less than 10% identified the machine learning code as the biggest bottleneck. Another 30% stumble when trying to deploy their machine learning model into production.

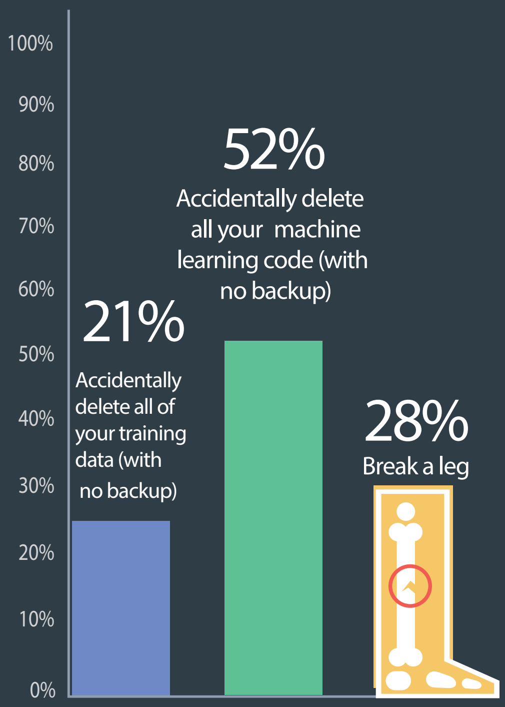
In this year's report, we wanted to see how data scientists feel. Testing our own hypothesis that while algorithms are front in center in data scientists' minds, it's really quality training data that hold the golden key to the success of so many projects. Our survey attacked the question of 'training data verse algorithm' from multiple angles and no matter how we asked, it's clear that data scientists hold their training data sets dear — in some cases more than intact limbs.

What is your **biggest bottleneck** in running successful AI/machine learning projects?

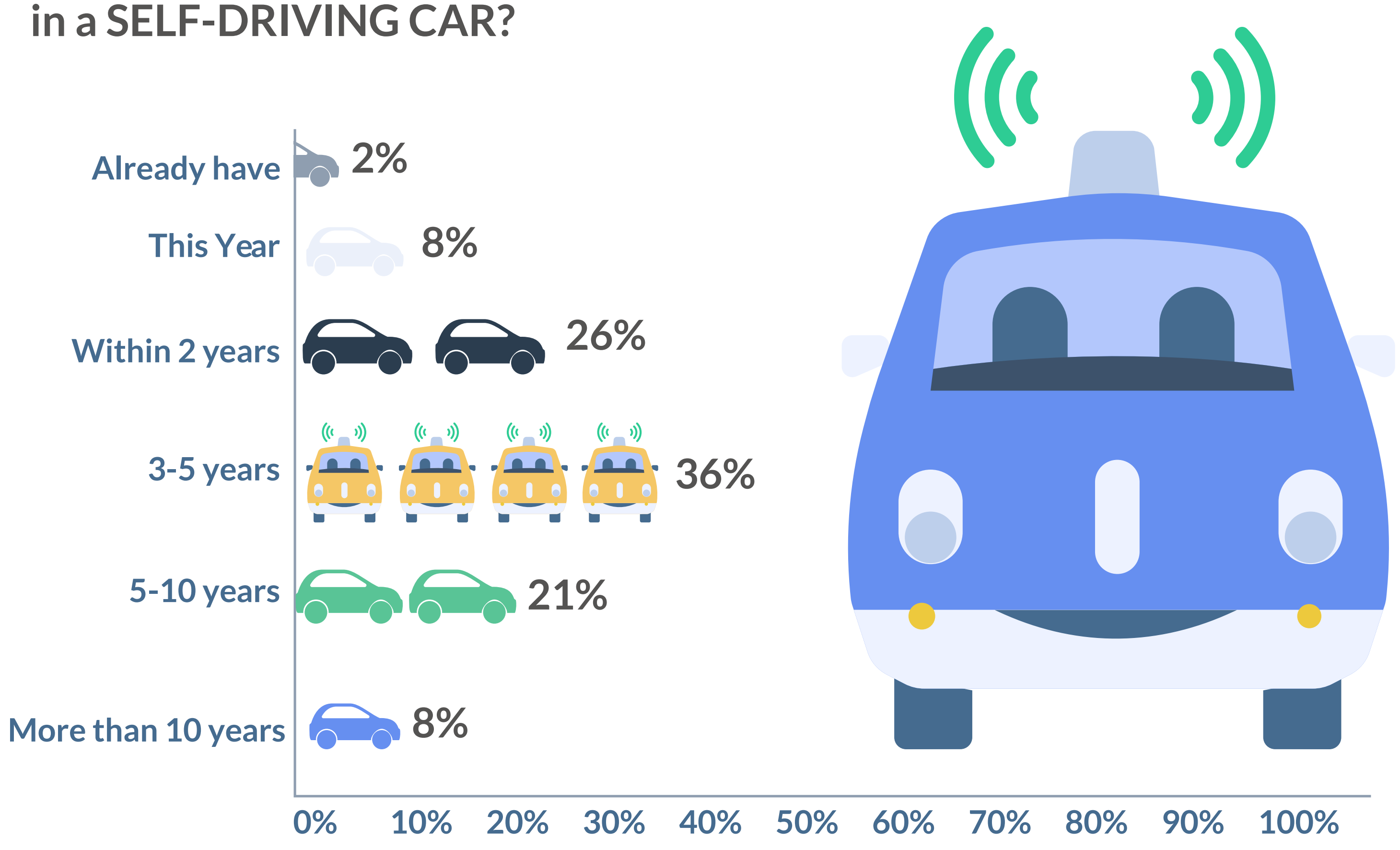


WOULD YOU RATHER....

Perhaps more telling, we not only pitted algorithms against training data but against limb integrity of data scientists. When asked whether they would rather delete their machine learning code, delete their training data or break a leg, slightly more than half (52%) of the respondents opted to sacrifice their algorithms. But when it came to limb versus data, the limbs lose. More data scientists would rather break a leg than accidentally delete their training data.



When do you think you'll first ride in a SELF-DRIVING CAR?



WHICH CAME FIRST, **THE TRAINING DATA** OR **THE ALGORITHM?**

The reality is, of course, that Training Data and Algorithms are independent parts of an iterative-looped process. Like their metaphorical chicken/egg cousin, training data/algorithms are impossible without the other. In our data science-esque take on the age-old chicken/egg riddle, we asked 'Which came first, the training data or the algorithm?' and received responses only true data scientists could muster. Below, some of our favorites:

Depends on our definition of algorithm. Many of the algorithms we use today have their roots in work that far predates the data we work with. Those early algorithms came into being though to analyze data of the era. Least Squares and its ilk were used to analyze astronomical data for example. While one could say that the algorithms typically build upon pre-existing mathematics - the application of that to the particular problem of inferring structure in data is a genuine advancement. I would thus say that the training data predates the algorithm. In some sense a similar phenomena occurs in mathematics where conjectures can drive progress - conjectures often arise as a handful (perhaps large) number of examples that beg for generalization.

Somebody then said, how do I make sense of all this data?

The algorithm is the idea of what would be possible once the data becomes available.

The truth of the relationships in the data that the algorithm finds existed before we found it.

Without data, algorithms are useless, like a meal with only forks and spoons but no food.

In theoretical papers, data is often simulated to discuss the power and feature-richness of an algorithm. For example, Artifical Neural Network algorithms were introduced in the 1940s where the concept of database and computing were still in its infancy. The mathematical layout was laid out even without the "big data", hence it is always algorithms.

The world is full of stimuli and information in which we find or employ labels and patterns. That we care about that and want to predict things from it all is secondary.

Sometimes you have an algorithm in search of data but that's still usually inspired by a real problem and, thus, data comes first.

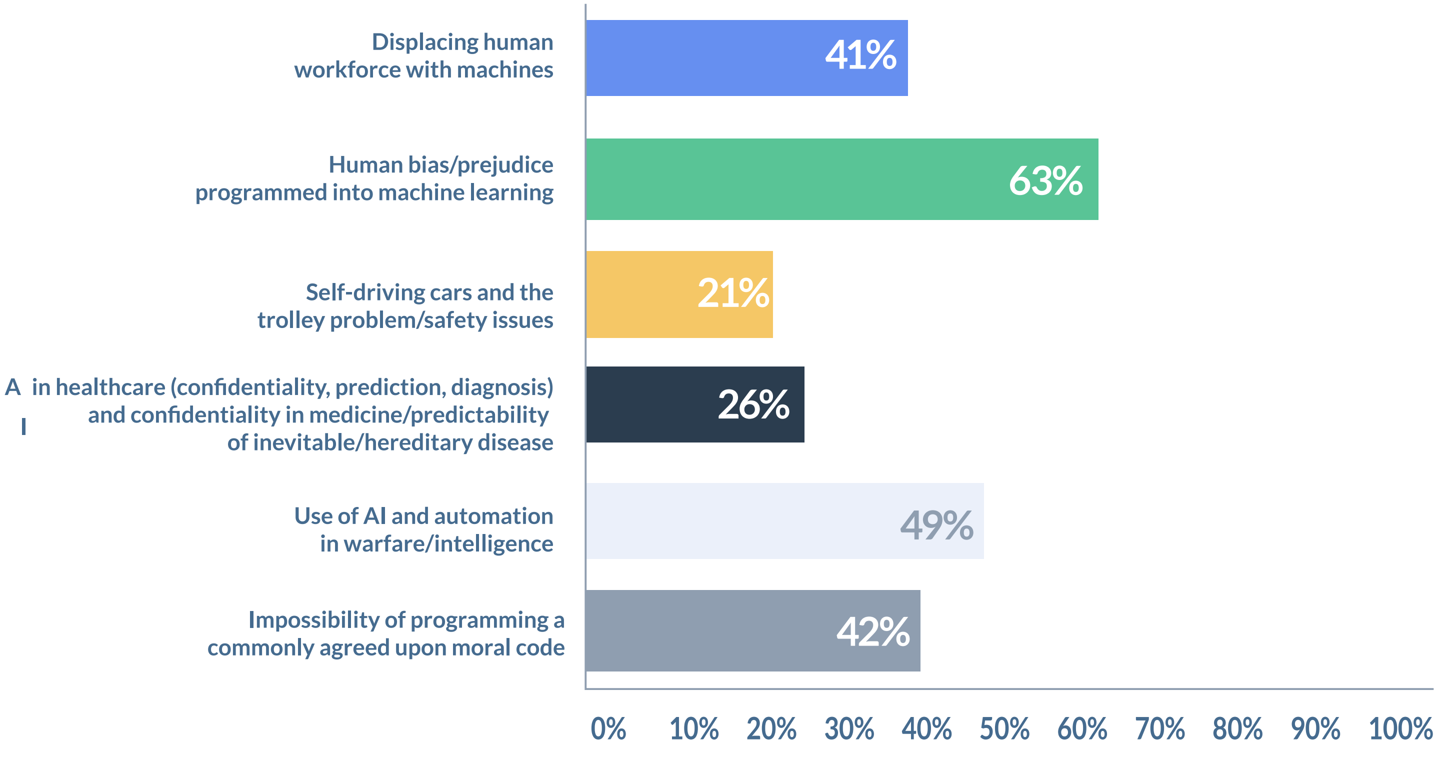
“Technology is giving life the potential to flourish like never before ...or to self-destruct”

MOTTO OF THE FUTURE OF LIFE INSTITUTE

Read any article on AI (and there is no shortage) and shortly behind, you’ll likely find mention of ethical issues. From the White House to the Wall Street Journal to the World Economic Forum, the question of how we program the future is one of the most critical issues facing not just data scientists but society as a whole. In perhaps the most important question in this year’s survey, we asked, “Which of the following do you personally think might be issues regarding ethics and AI?”

The programming of human bias/prejudice into machine learning is the biggest concern of data scientists today with 63% of respondents expressing concern on this specific issue. Implicit in this response, the importance of integrity in training datasets.

The use of AI and automation in warfare/intelligence is a major concern of half of data scientists. Unease on the displacement of human workforces and the impossibility of programming a commonly agreed upon moral code also ranked high on the radar of ethical issues for data scientists tallying in at 41% and 42% respectively.



SUMMARY

In summary, if 2016 was the year of the algorithm, we're proclaiming 2017 the year of data —training data to be precise. Data scientists are spending more than half their time labeling and creating it, they value it over machine learning code (and unbroken legs), it's decidedly determined to come 'before the algorithm' and- most importantly — its integrity is key to providing unbiased models as AI starts to drive our future. Despite the fact that zero percent of data scientists predict they'll be dealing with less data in 2017, quality levels are less predictable and lack of access to high quality training data is the single biggest reason AI projects fail. Given the massive proliferation of AI projects in virtually every sector across the globe, data scientists must work to offload routine work and streamline processes in the face of increasing data, increasing AI projects and a continued shortage of those with the necessary skills.

METHODOLOGY

For this year's survey, CrowdFlower surveyed 179 data scientists globally representing a balance of company ranging in size from <100 to 10,000+. A variety of industries were represented as well, with a slightly weighted emphasis on 'technology' — representing 40% of respondents. The survey was conducted in February and March of 2017.

CrowdFlower

www.crowdflower.com

About CrowdFlower
CrowdFlower is the essential human-in-the-loop AI platform for data science teams. CrowdFlower helps customers generate high quality customized training data for their machine learning initiatives, or automate a business process with easy-to-deploy models and integrated human-in-the-loop workflows. The CrowdFlower software platform supports a wide range of use cases including self-driving cars, intelligent personal assistants, medical image labeling, content categorization, customer support ticket classification, social data insight, CRM data enrichment, product categorization, and search relevance.

Headquartered in San Francisco and backed by Canvas Venture Fund, Trinity Ventures, and Microsoft Ventures, CrowdFlower serves data science teams at Fortune 500 and fast-growing data-driven organizations across a wide variety of industries.

For more information, visit www.crowdflower.com.