



DATA SCIENTIST REPORT 2018



INTRODUCTION



Figure Eight has been taking the pulse of the data science community for quite a while now. But a lot has changed since our original Data Science Report in 2015 (run back when we were still known as CrowdFlower). Machine learning projects are multiplying and more and more data is required to power them. Data science and machine learning jobs are LinkedIn's faster growing jobs. And the internet is creating 2.5 quintillion bytes of data each day to power all of it.

Another thing that's changed since 2015? The community is grappling with more ethical issues than ever before. Data privacy, of course, has always been a paramount concern. But as AI is increasingly used to make big decisions like medical diagnoses and courtroom sentencing, these ethical considerations require careful debate. Understanding what those involved think about the technology they're pioneering felt important. In fact, we asked 500 ethical professionals—like doctors, clergy, and police

officers—how they felt about AI and contrast their opinions with our core constituency of data scientists near the end of the report.

Without further ado, let's get to our findings.



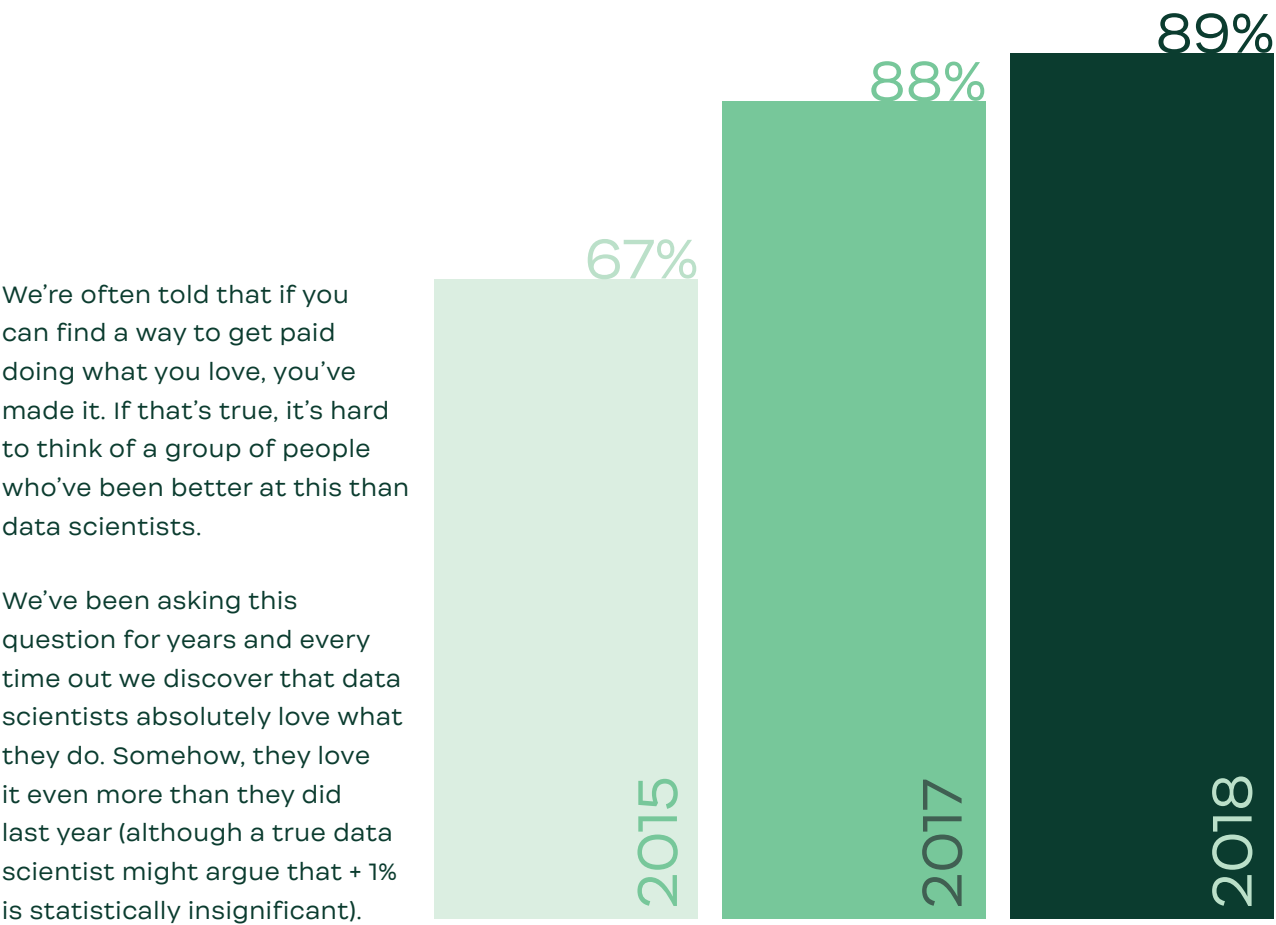
DATA SCIENTISTS

DON'T LIKE THEIR

JOB, THEY LOVE IT

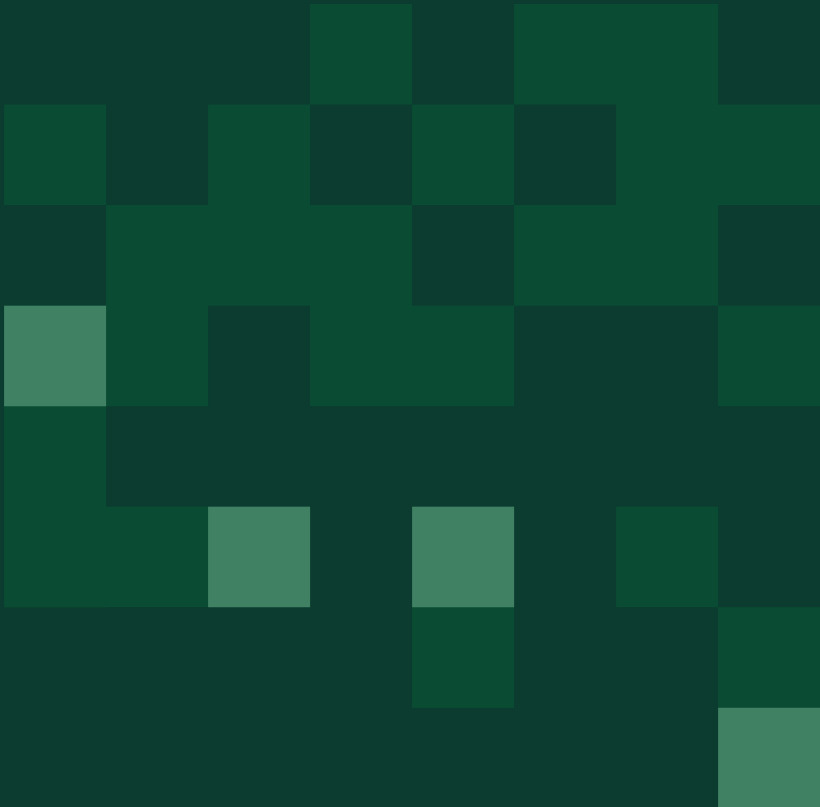


DATA SCIENTISTS WHO ARE “HAPPY” OR “VERY HAPPY”



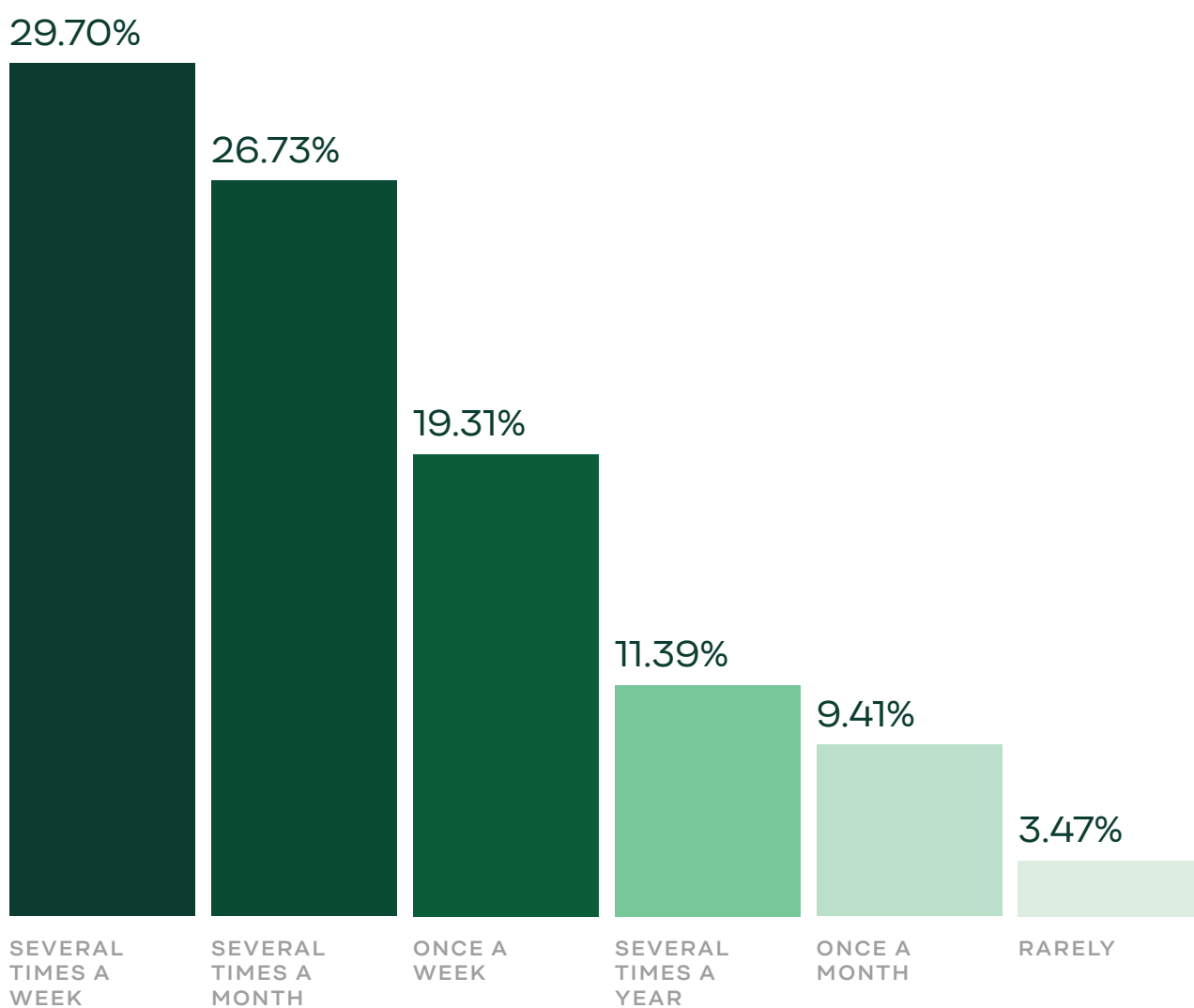



IF YOU LOVE YOUR DATA SCIENTIST, DO NOT SET THEM FREE



Data (and data science) inspire a lot of breathless headlines. Peter Norvig famously outlined the Unreasonable Effectiveness of Data. The Harvard Business Review dubbed data scientist as the sexiest job of the 21st

century. The Economist jumped on the “data is the new oil” bandwagon a year back. We’d bet plenty of you remember “Big Data” being a thing people realized might suddenly be useful.





But it's worth remembering it wasn't always like this. After all, some of the same companies that now hoard and guard proprietary data didn't track and save user interactions ten or fifteen years ago. The explosion of cheap server space and the realization of what could be done with copious amounts of information helped make all those headlines in the last paragraph a reality.

With so much data to process, and with so much of it going towards business processes and initiatives that create real value for their organizations, it's no wonder that data scientists are in such high demand. When we asked data scientist how often they get contacted about new gigs, the results speak for themselves.

Data scientists are front and center in all of this. And when we asked them how often they get contacted about new gigs, the results speak for themselves:

Nearly 50% of data scientists get contacted at least once a week about a new job opportunity, with about 30% getting contacted several times each week. 85% get contacted at least once a month.

In other words, data science talent is still in big demand. So if you've got good ones in your organization, keep them happy. They've certainly got options.

HOW OFTEN DO YOU GET CONTACTED FOR NEW JOB OPPORTUNITIES?



WHAT HOLDS DATA SCIENTISTS BACK? THE DATA, NOT THE SCIENCE

The thing about data scientists is they're greedy. Not in a bad way, mind you. We know plenty and they give fabulous holiday gifts. But no matter how much data they have, chances are, they need more.

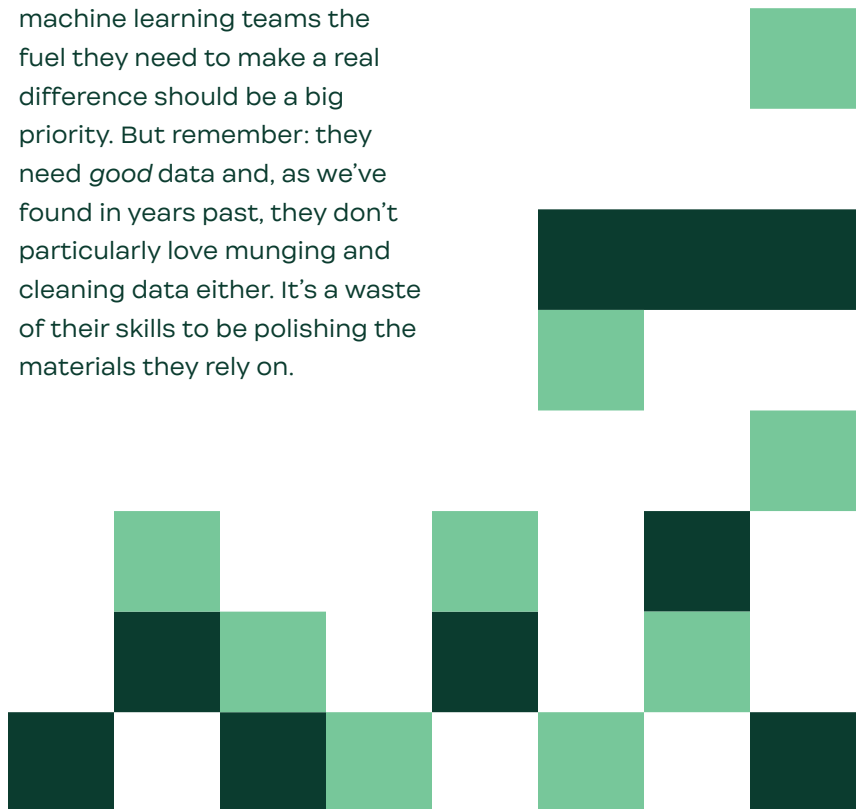
We've been running this survey for a few years now and this is *always* the biggest challenge for the community. In fact, last year, about 50% of data scientists cited this as one of their top three snags in their day-to-day. This year, that number has increased even more to 55% cited it as their biggest.

55%

CITED QUALITY/QUANTITY
OF TRAINING DATA AS BEING
THEIR BIGGEST CHALLENGE

Here's what data pros know: high quantities of high quality data are what build accurate models and inform smart decisions. And the more of it you have, the more confident your model will be.

Anything your organization can do to give your data and machine learning teams the fuel they need to make a real difference should be a big priority. But remember: they need *good* data and, as we've found in years past, they don't particularly love munging and cleaning data either. It's a waste of their skills to be polishing the materials they rely on.

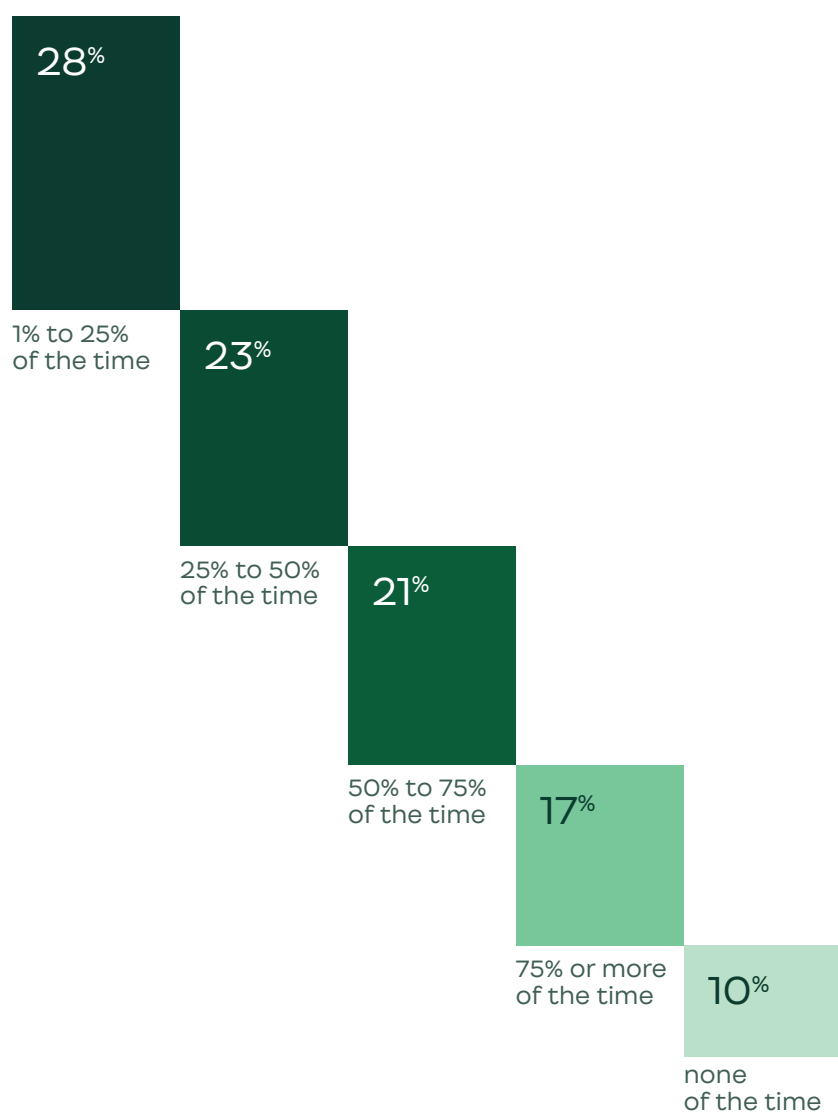


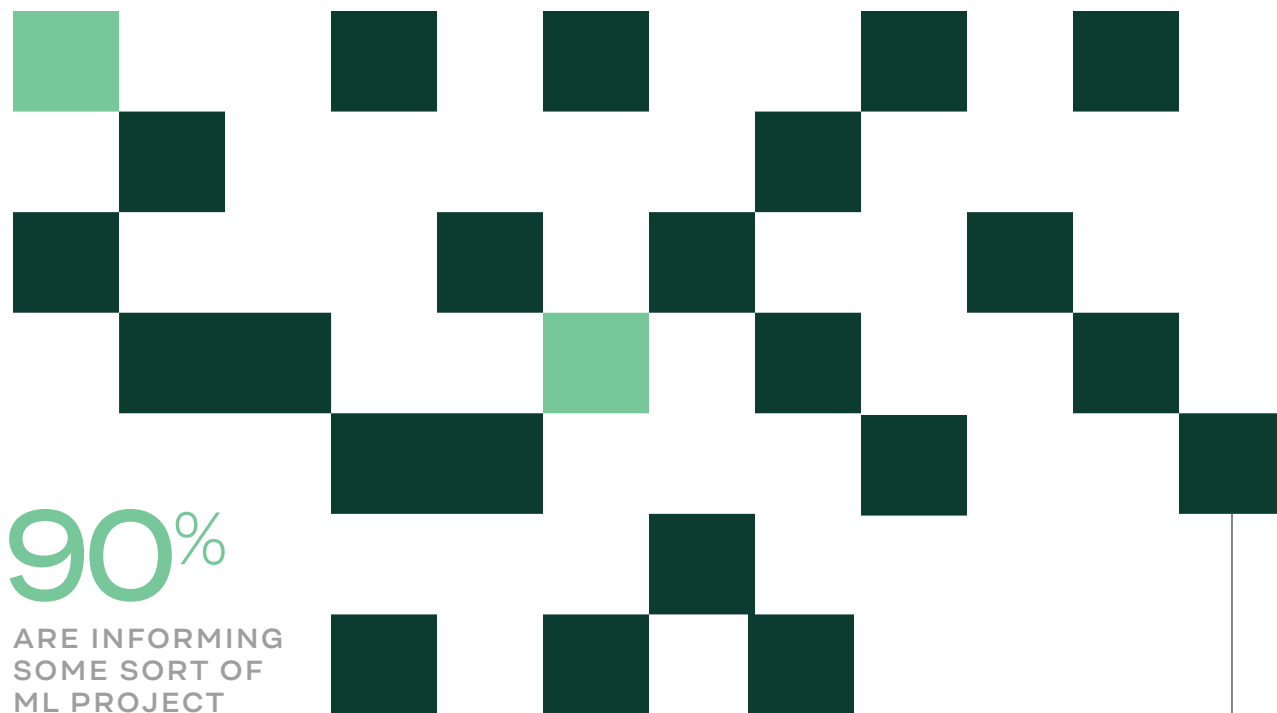
FEED THE MACHINE

LEARNING

One question we'd never asked before was what exactly data scientists do with all that data anyway. But as our platform has grown and we've been able to lift the hood on a lot of projects, more and more of the data that comes through Figure Eight directly informs AI and machine learning projects. So we thought we'd just come right out and ask: what percentage of their work is used for AI?

PERCENTAGE OF WORK USED FOR AI

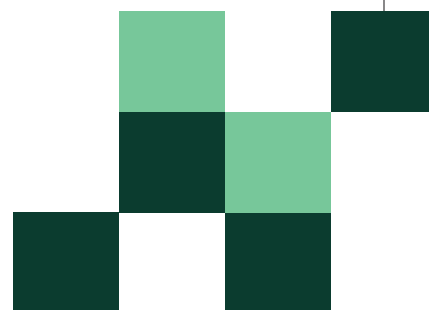




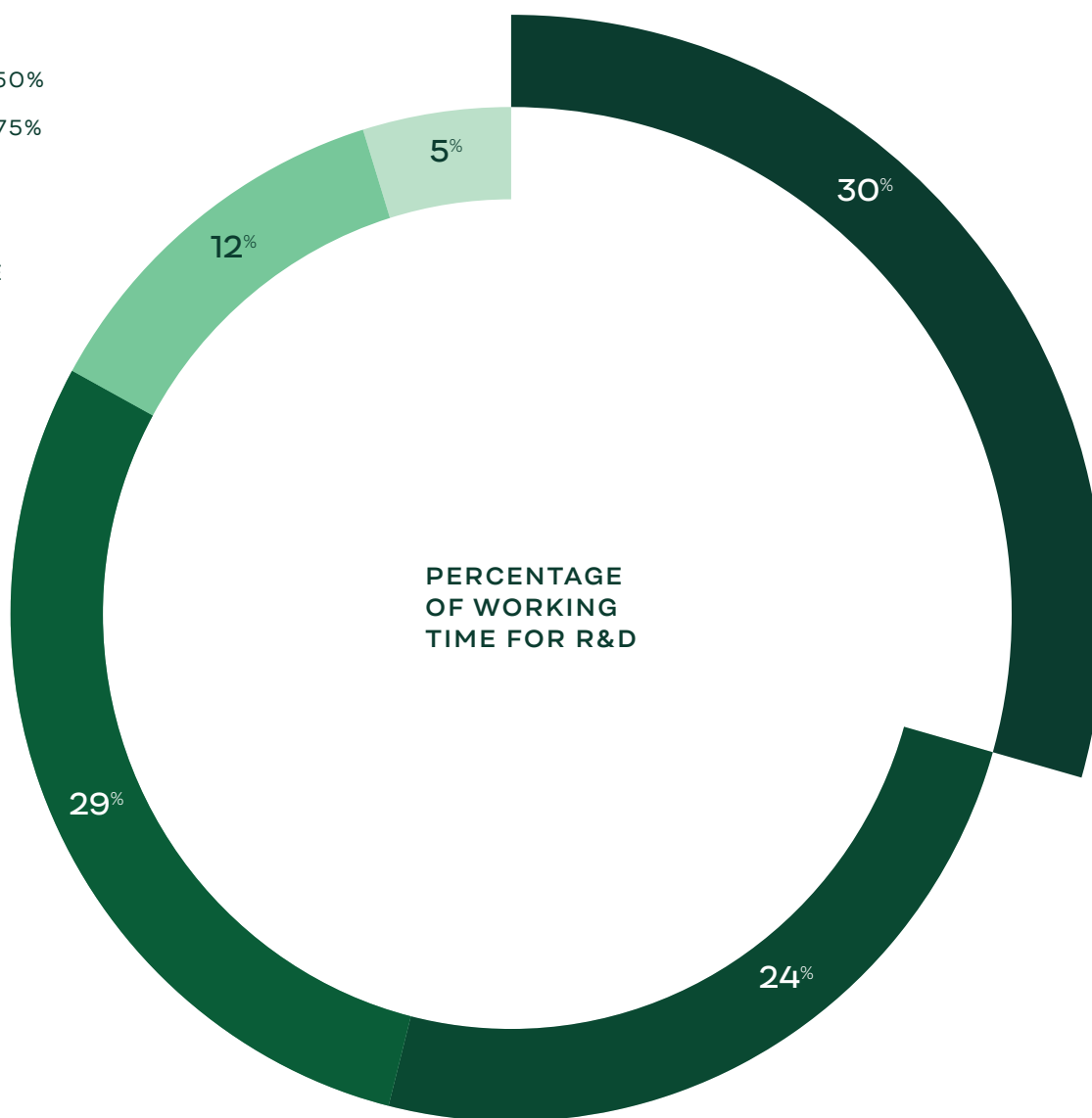
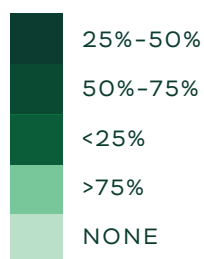
While about 1 in 10 data scientists say none of their work informs AI projects, nearly 40% say a *majority* of their work does.

With massive increases in investment for these exact types of initiatives, we're excited to see where this

number goes next year. But it's undoubtedly a positive development. Instead of being tasked with cleaning log files, data scientists are seeing their work inform cutting edge solutions across their organizations. No wonder they're so happy.



HOW MUCH OF YOUR
WORKING TIME GOES
TO R&D (AS OPPOSED
TO PRODUCTION)?



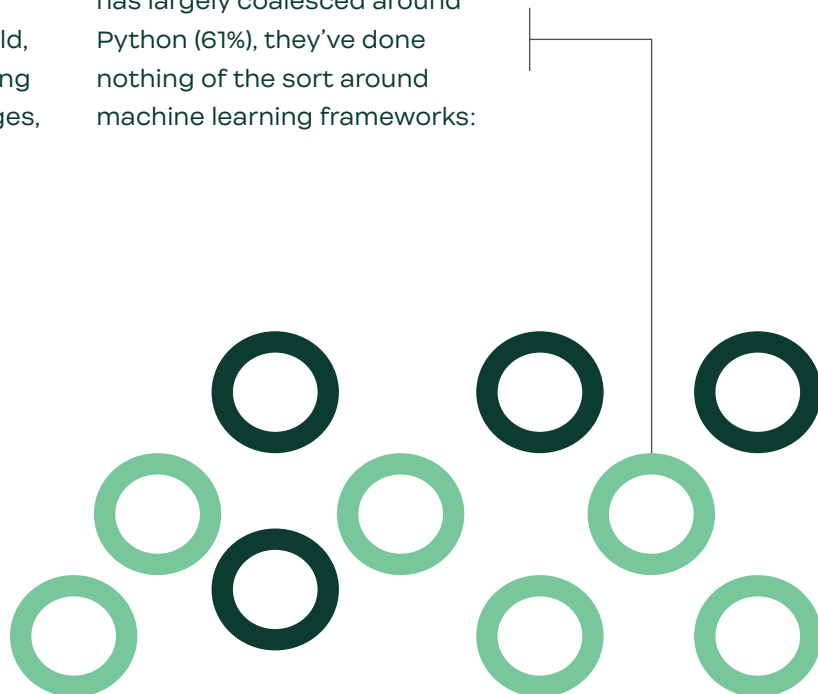
THE TOOLS BEHIND THE TALENT

Back in 2015, one of our more interesting findings centered around the tools data scientists used. It wasn't that Excel was still a crucial part of their day-to-day but rather that the breadth of tools and approaches they used was very wide. In fact, the folks at Partially Derivative picked up on this fact in a podcast episode they titled "Keep Data Science Weird."

Their point? Because data science was still a young field, there wasn't an overwhelming consensus on what languages,

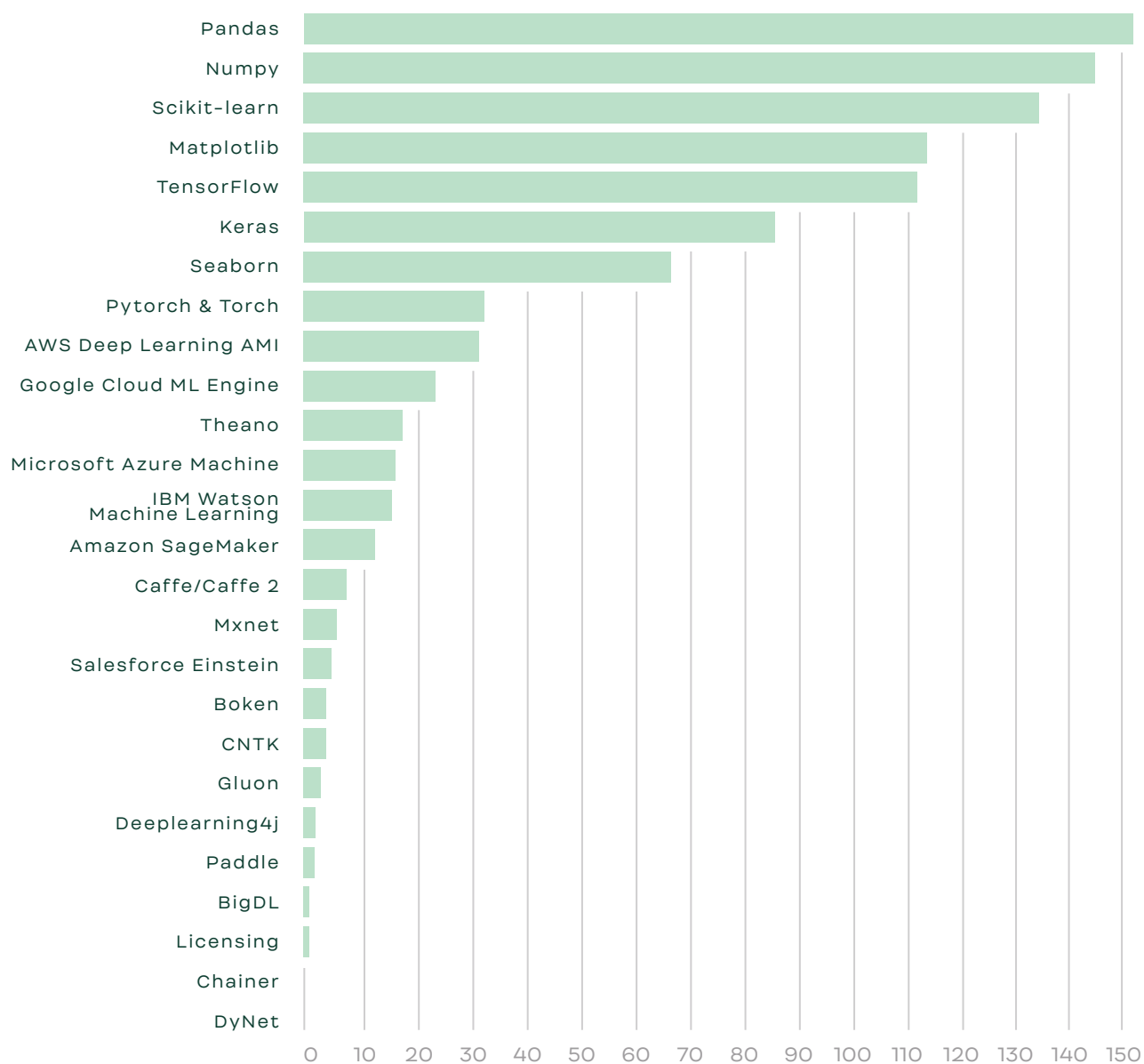
tools, and frameworks were best. Data scientists got to stay creative, finding the specific suite of tactics that worked best for their particular projects.

You could argue that machine learning is in a similar place right now. There isn't a codified set of strategies but a wide range of approaches to solve what were once intractable problems. While the community has largely coalesced around Python (61%), they've done nothing of the sort around machine learning frameworks:





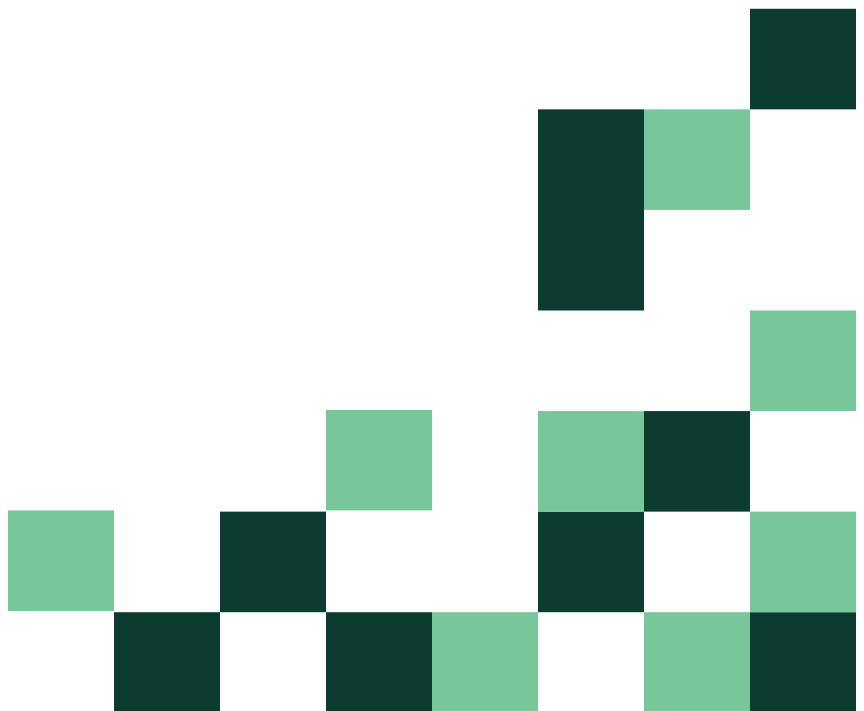
POPULAR MACHINE LEARNING FRAMEWORKS



What really jumps out here is the preponderance of open source tools. Pandas and Numpy have been around for a while. Same goes for Scikit-learn and Matplotlib. TensorFlow's origins are Google, but that was open-sourced too.

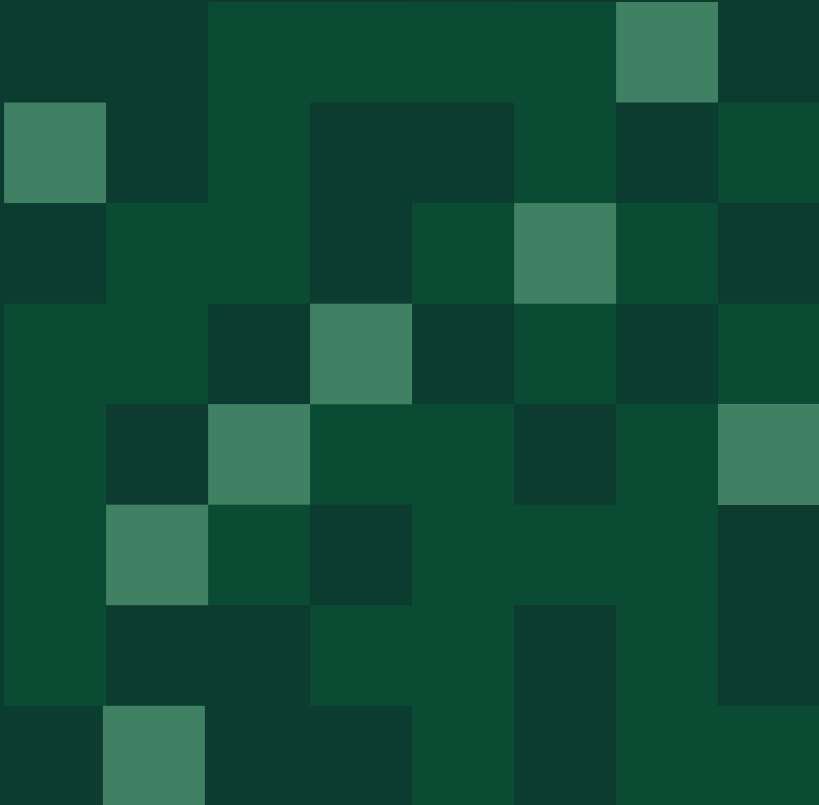
Notably, a lot of the more white label solutions are further down the list. This isn't to judge quality mind you—some of these tools are tremendous—but it does speak to a particular fondness for open-sourced, community-driven software in

the community. And since a lot of these frameworks have been around for quite some time, early adopters are intimately familiar with them. It'll take time, effort, performance (and a little marketing budget) to unseat those open source frameworks.





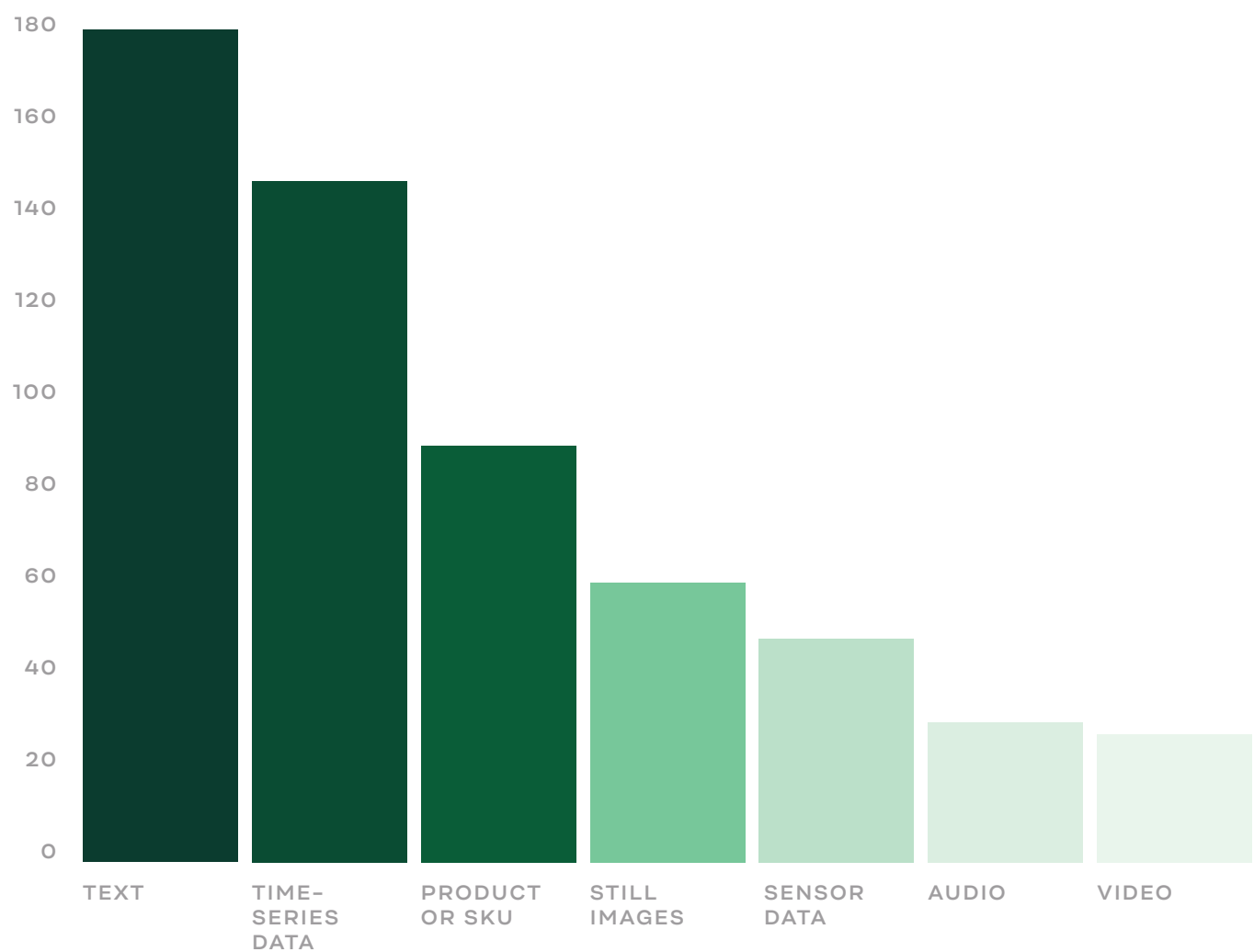
THE DATA THAT DATA SCIENTISTS ARE WORKING WITH IN 2018



With all the media focus on sexy machine learning use cases like autonomous vehicles and home assistants, it's worth remembering that most data scientists aren't actually working with LIDAR or audio utterance data.

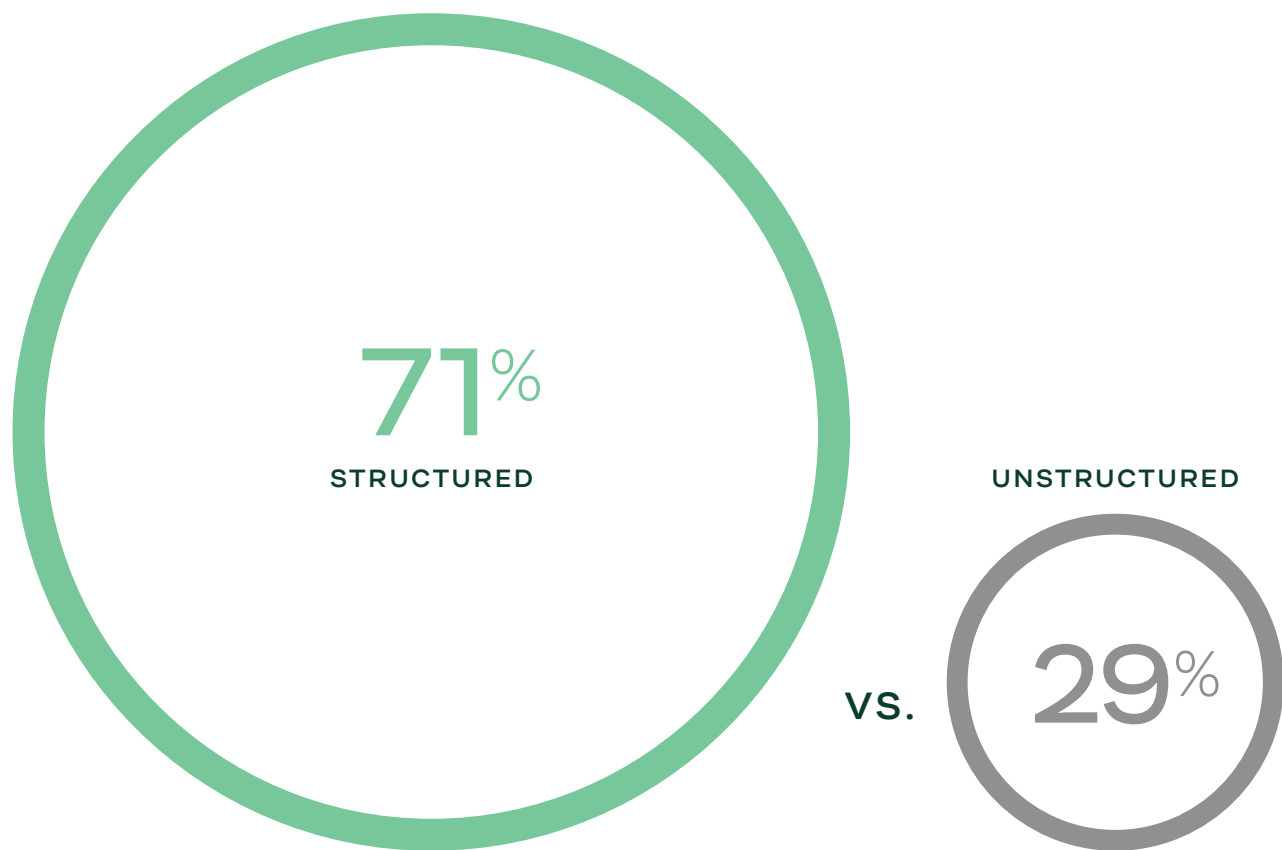
We asked data scientists to select all the types of data they frequently work with and found that, yes, text and time-series data remain the most common ingredients for their projects. Sensor, audio, and video data were the least used, respectively, while about a fourth of data scientists were working with still images.

COMMONLY USED DATA TYPES





DO YOU WORK PRIMARILY
WITH STRUCTURED OR
UNSTRUCTURED DATA?



ETHICS 101

The ethical issues around both building and deploying AI have been magnified in recent years. We've seen countless examples of algorithmic bias in sub-disciplines like facial recognition, employment application review, audio assistants, and more. Just last year, the Supreme Court had the opportunity to adjudicate a matter involving algorithmic sentencing (*Loomis v. Wisconsin*). And while they court chose not to hear that case, it's a safe assumption that legal precedents will be set around machine learning in the coming decade.

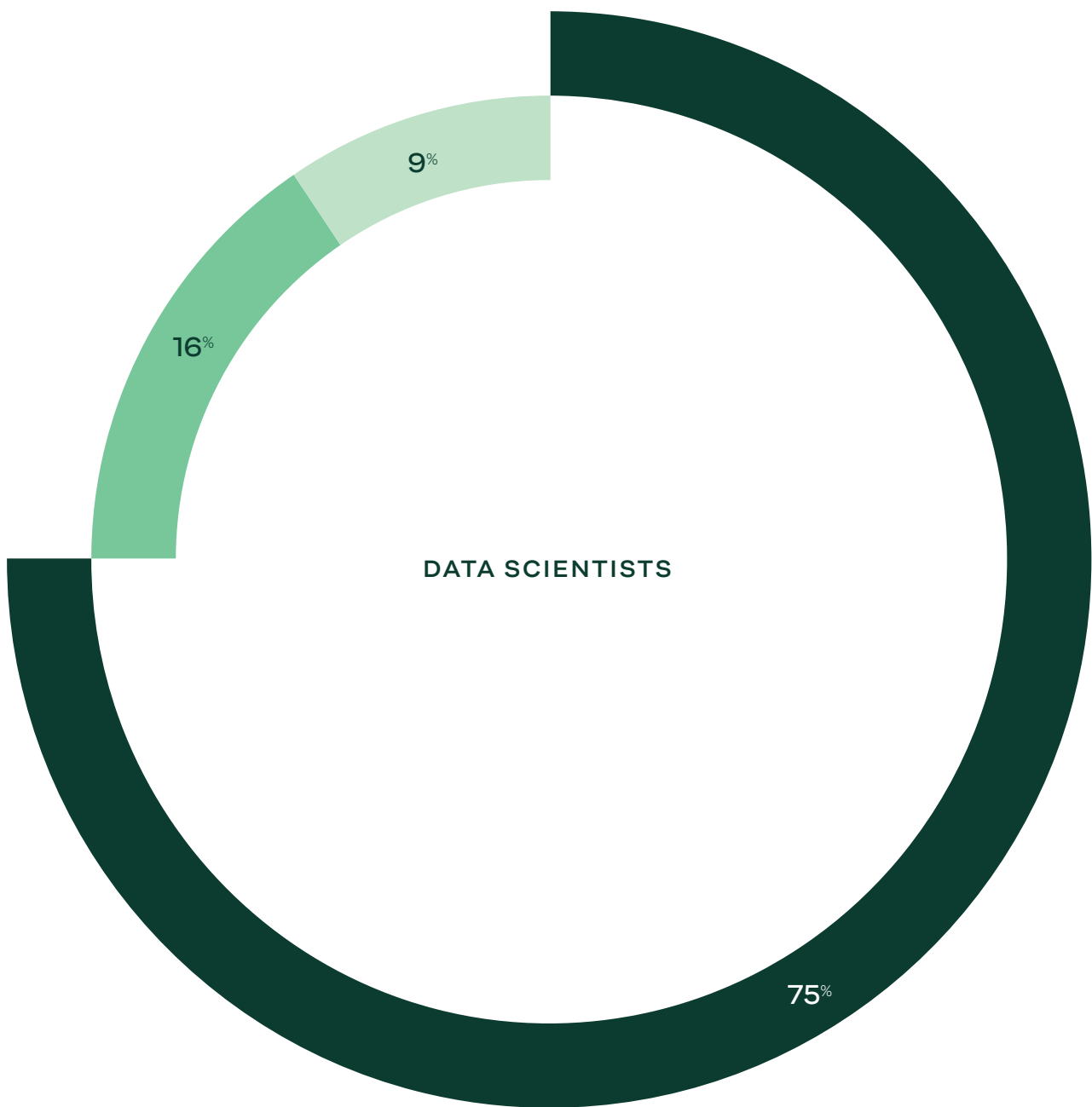
These are the sort of ethical issues we were most concerned with in our survey. Not the long-term, science-fiction-tinged ethical issues around

agency, general intelligence, or defining the boundaries of consciousness, but real-world issues the field—and the public—are grappling with right now.

Keep in mind that we also ran a survey capturing the opinions of ethical professionals like doctors, clergy, and law enforcement. We'll be contrasting the viewpoints of data scientist with those professionals throughout this section.

For starters, data scientists are really optimistic about AI in general.

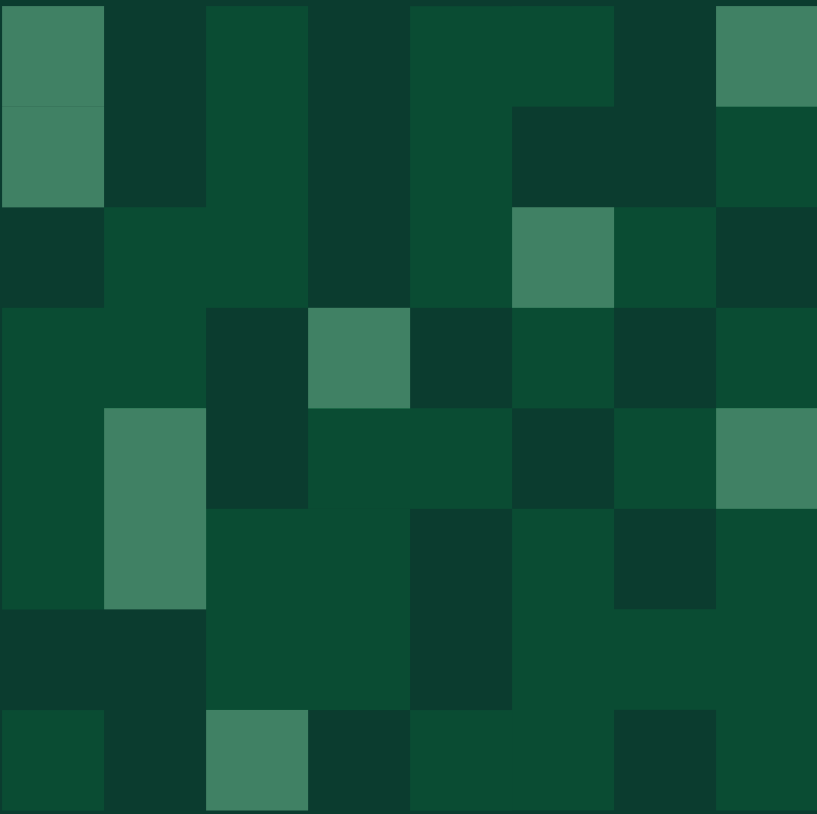
While both groups felt that there would be more attendant good than ill, the big outlier here is that ethical professionals are fairly blasé about the potential changes AI will bring to society. Which makes sense. We know that data scientists are far more intimately involved in the space than, say, a judge might be. They're more informed about and more invested in this technology, so are less likely to feel that work won't make much of a difference at all.





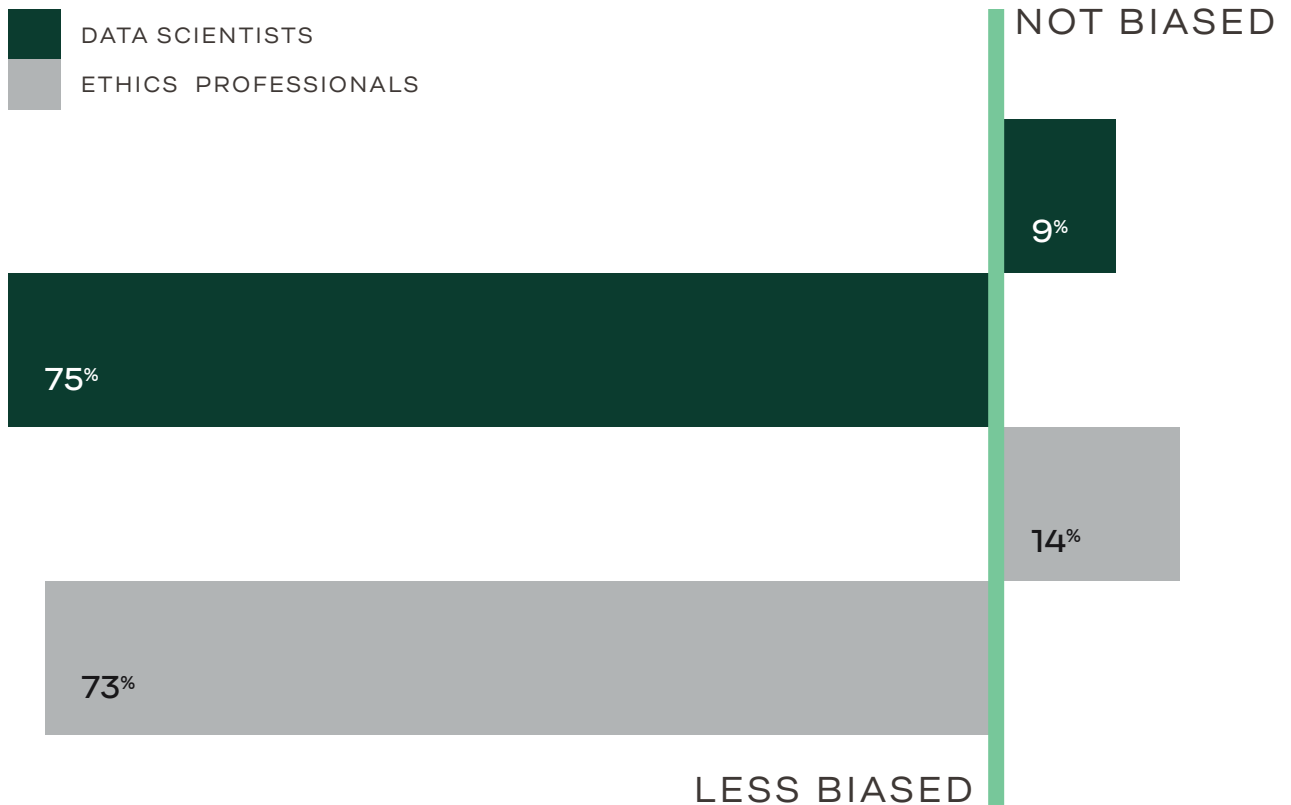


ARE WE STILL IN DENIAL ABOUT ALGORITHMIC BIAS?



We mentioned a handful of fairly well-known cases of algorithmic bias in the last section. In fact, a recent MIT Technology Review cautioned that “biased algorithms are everywhere and no one seems

to care.” But when we asked both data scientists and ethical professionals if they believed AI was more or less biased than people, our respondents thought otherwise:



Now, we understand that comparing technological bias to human bias is inherently a slippery idea. It really depends on how optimistic your view of human nature is and whether you blame algorithmic bias on human programmers, data, or some other, vaguely ineffable reason. But it was interesting to see how many responses coalesced around both “less” biased and, especially, “not biased at all.” After all, we have copious examples of algorithmic bias actually happening.

Of course, *why* this happens is the real issue here. And oftentimes, it’s not the inherent nature of, say, a deep learning model, but the data that *powers* that model. The bias is latent and unintentional. But it does exist. Solving it requires real effort, asking the right questions, annotating data conscientiously, collecting data to repair bias, iterating on models, and showing empathy for end-users (among others).

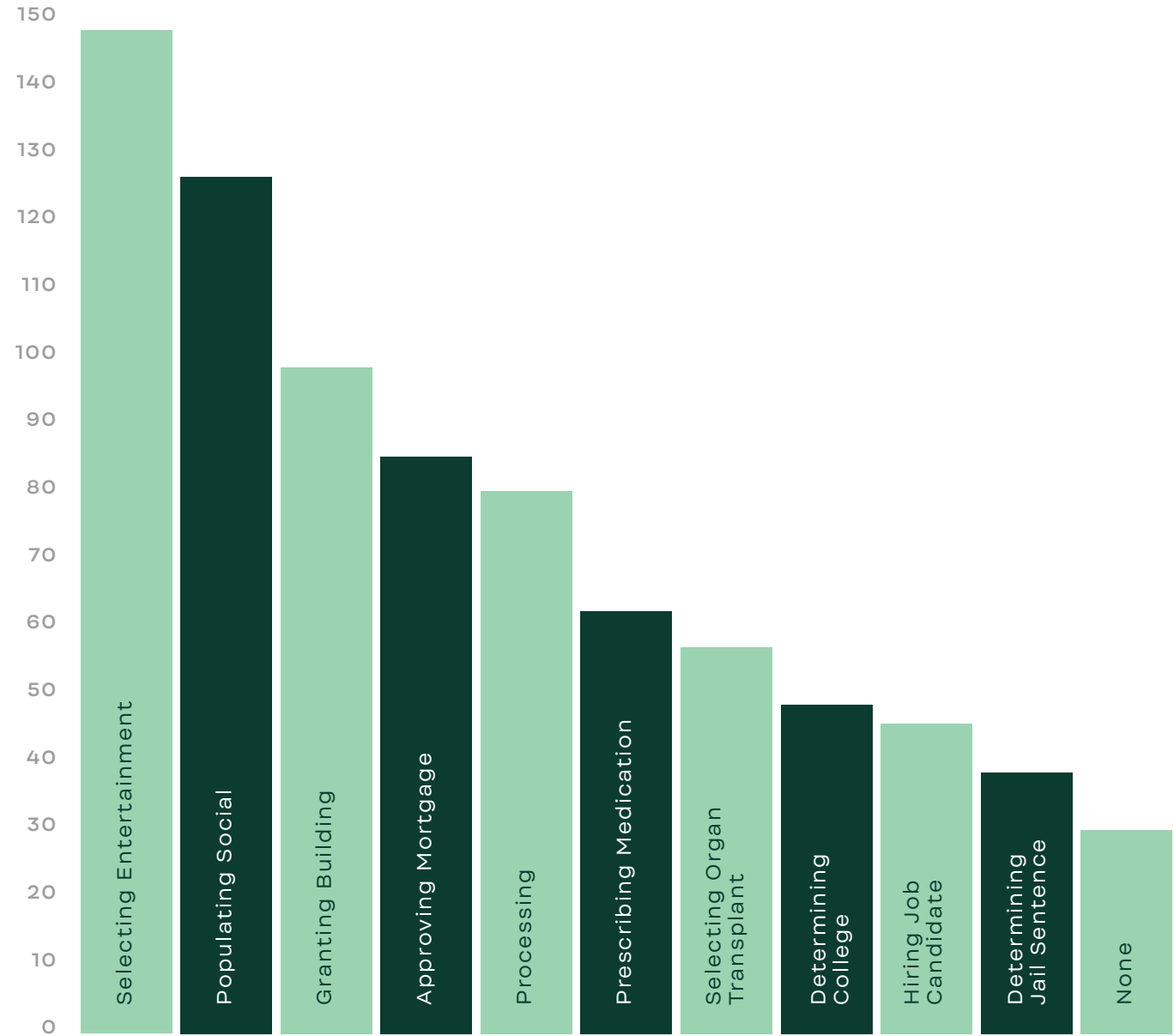
WHAT ABOUT REAL-WORLD AI IMPLEMENTATIONS?

Most internet users deal with machine learning on a daily basis. Product and entertainment recommendations, search engines, new feeds, you name it: all are increasingly being powered by ML.

In fact, let's start there. Because when it comes to decision-making, most data scientists have no problem with the scenarios AI's already sewn in. The more crucial the situation, the less likely they're comfortable:

How much of these opinions have to do with the seriousness of the decision versus the fact we've already seen success applying AI to those less vital scenarios isn't something we can necessarily say for sure. What we can say is that implementing AI across all sectors of society isn't something the community has a big appetite for. When the people behind the technology are preaching a slower, more sober approach, we'd all do well to sit back and listen.

APPROPRIATENESS OF AI MAKIKNG DECISION, SCENARIOS





TO RELEASE

OR NOT TO

RELEASE,

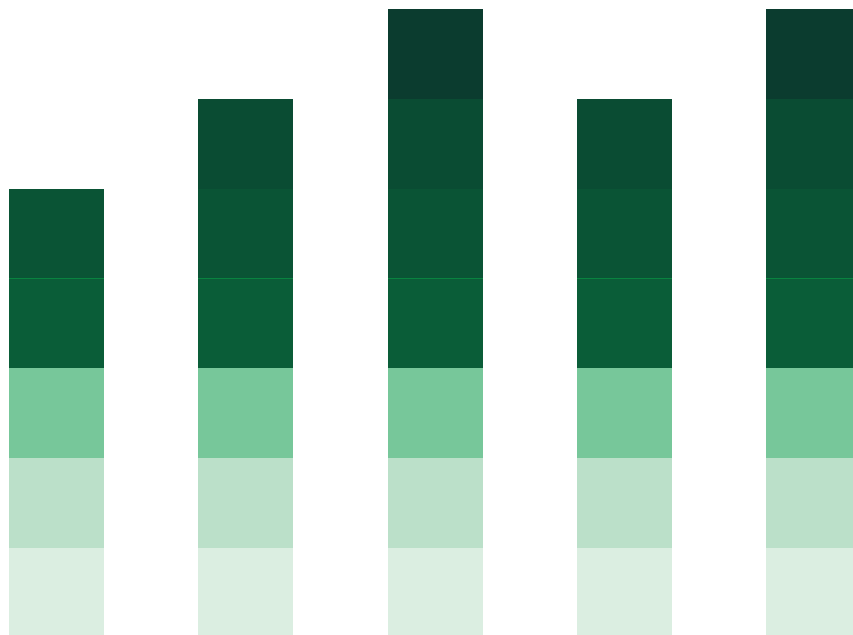
THAT IS THE

QUESTION



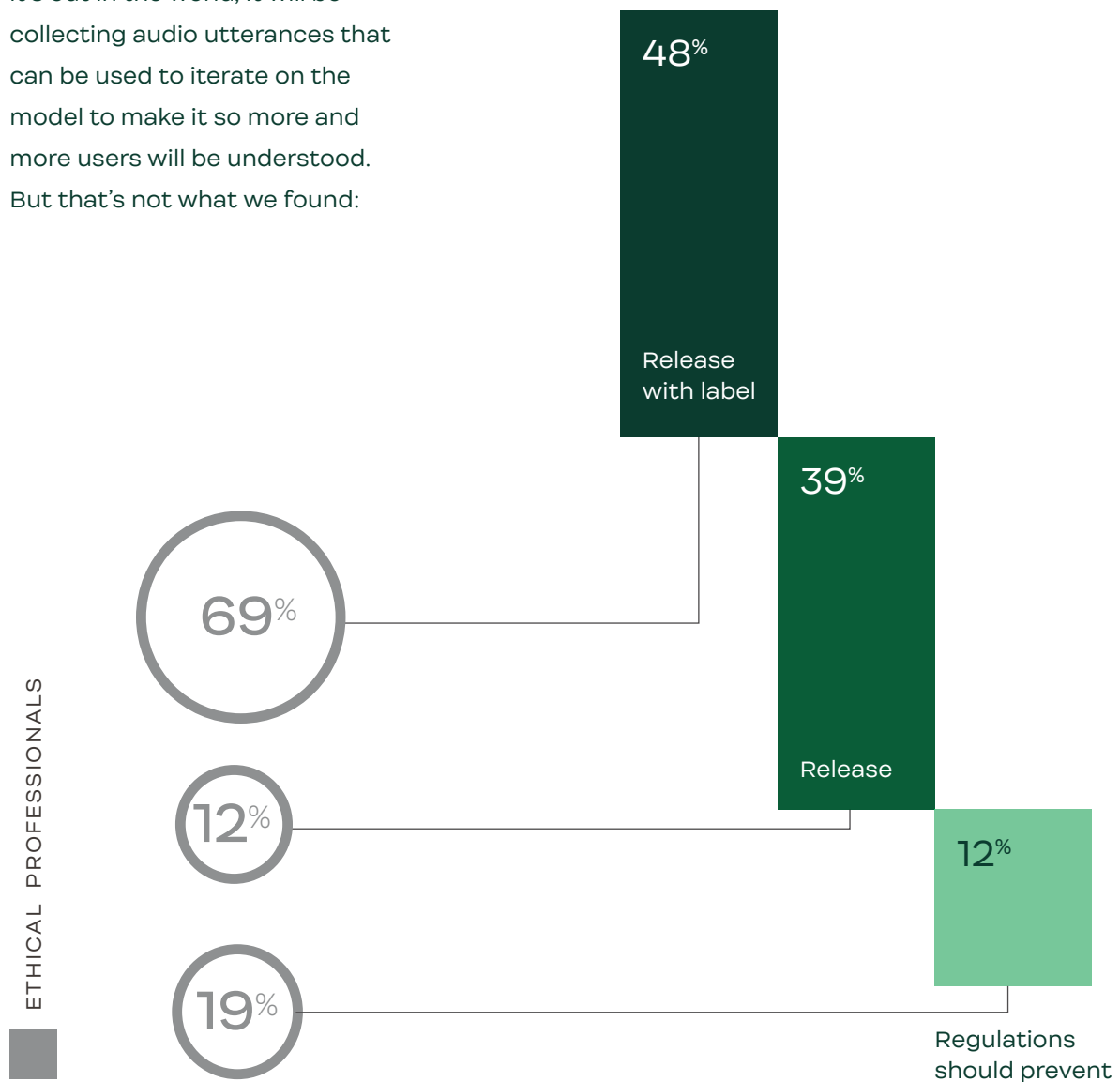
Audio interfaces are becoming more and more prevalent with each passing day. Comscore predicts that 50% of all searches will be voice searches by 2020 and there are already roughly a billion voice searches each and every month. But even the most advanced voice-activated assistants struggle with regular, everyday speech. This is especially true when the speakers themselves aren't native speakers, have regional accents or dialects, or code switch between languages.

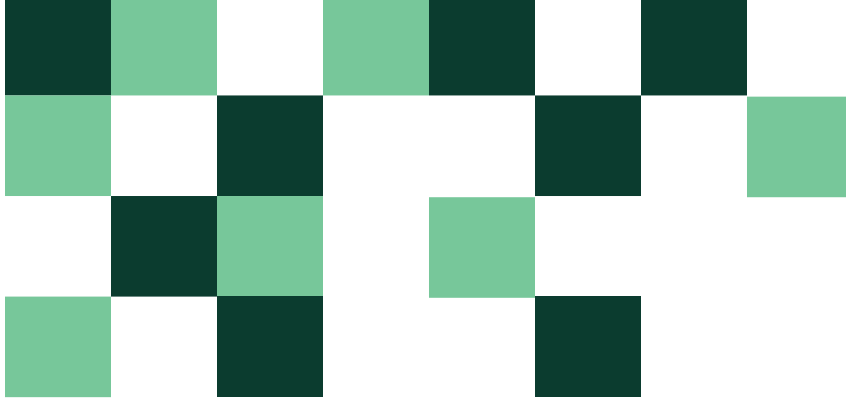
We asked data scientists about this exact issue. Specifically, we wanted to know what should be done about releasing a home assistant that couldn't understand accents or dialects. Should the company just release the product? Should they have to include a warning label that notes certain people may not be able to use the product? Or should there be regulations in place preventing the release of a product in this state.



Frankly, we expected the community would want to release the product as is. After all, once it's out in the world, it will be collecting audio utterances that can be used to iterate on the model to make it so more and more users will be understood. But that's not what we found:

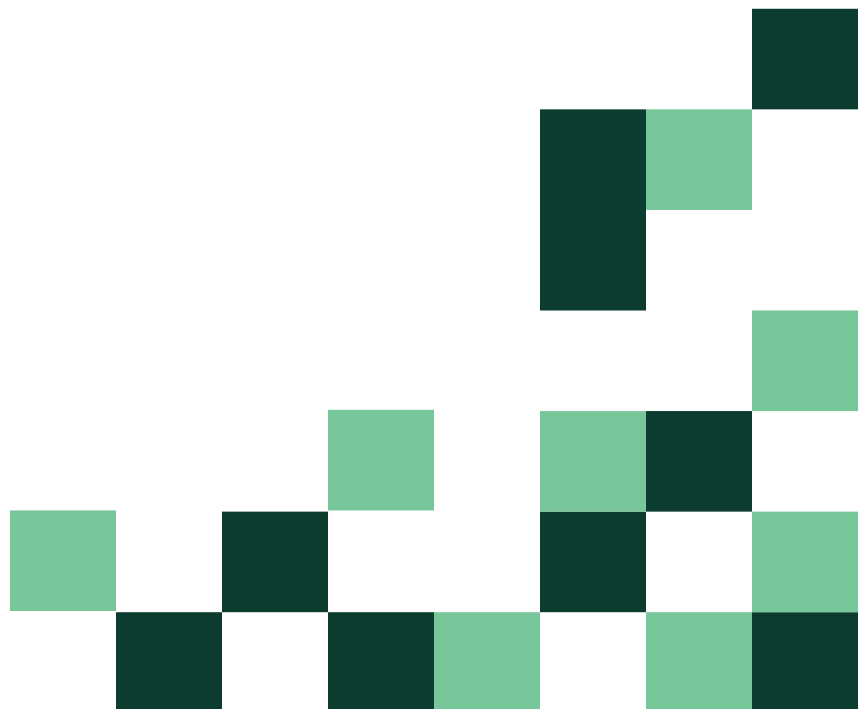
TO RELEASE OR NOT TO RELEASE, THAT IS THE QUESTION





While we were a bit surprised by this finding, it dovetails fairly well with the previous section. The data science community is cautious about these implementations. They're in

favor of transparency. And when you think about the community's fondness for open source platforms and open data, a lot of this really starts to make sense.





DRIVING HOME A BIG DIFFERENCE

The last state-of-AI we posed to both our ethical professionals and data scientists was a simple one: if a self-driving car was proven statistically safer than your average human driver, would you rather drive yourself or use an autonomous vehicle?

For most of our survey, both groups tracked pretty similarly. They largely felt that AI was a force for good, that products should have labels informing who they work for and who they don't, and were more comfortable with AI-driven product recommendations than AI-driven mortgage approvals or judicial sentencing.

They were polar opposites here:

Members of the data science community certainly know more about the work that's going into autonomous vehicle technology than your local clergy, but we certainly weren't anticipating opposite reactions. Why were those reactions so completely different? It's tough to say. But at the very least, if you're working on self-driving cars, you certainly know who to market to first.

WOULD YOU RATHER DRIVE YOURSELF OR USE AN AUTONOMOUS VEHICLE



DATA SCIENTISTS

ETHICS PROFESSIONALS



SURVEY INFORMATION

For this year's report, we surveyed 240 data scientists over Survey Monkey via email and at conference booths. To

get copies of older versions of this report, dating back to 2015, please head to the resource center on figure-eight.com.

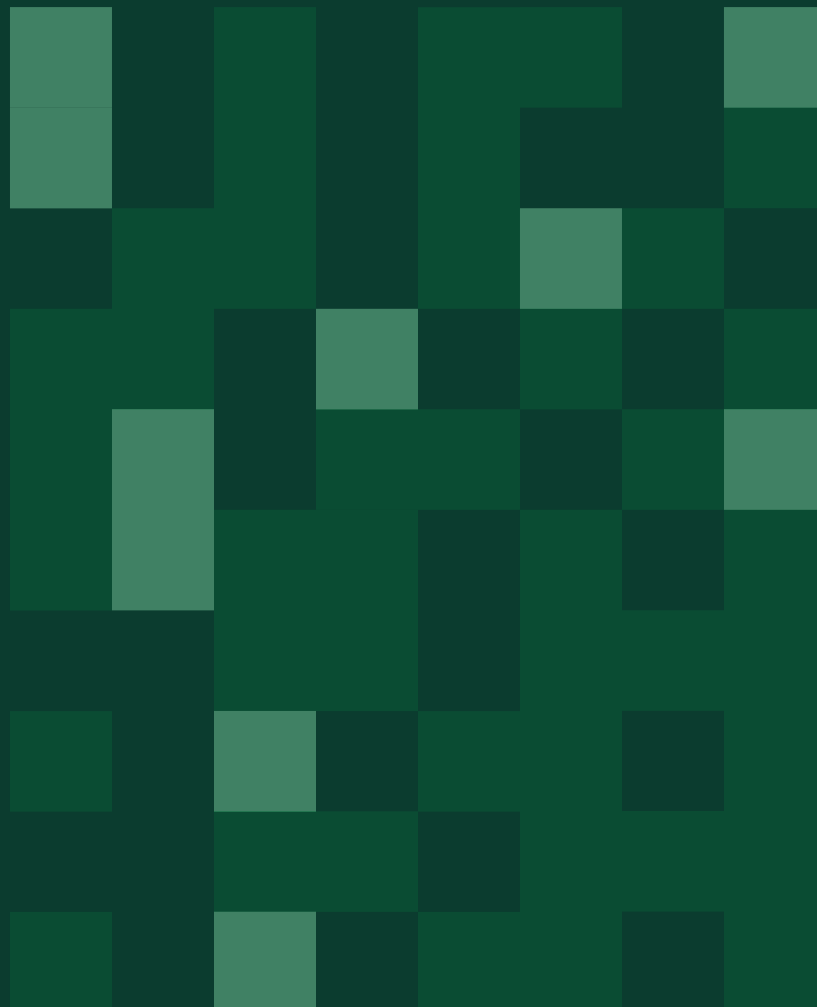




Figure Eight is the essential human-in-the-loop AI platform for data science teams. Figure Eight helps customers generate high quality customized training data for their machine learning initiatives, or automate a business process with easy-to-deploy models and integrated human-in-the-loop workflows. The Figure Eight 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, Figure Eight serves data science teams at Fortune 500 and fast-growing data-driven organizations across a wide variety of industries.

figure-eight.com