STATE OF DATA SCIENCE 2022

PAVING THE WAY FOR INNOVATION

ANACONDA.

EXECUTIVE SUMMARY

This year, we conducted our State of Data Science survey to gather demographic information about our community, ascertain how that community works, and collect insights into big questions and trends that are top of mind within the community.

As the impacts of COVID continue to linger and assimilate into our new normal, we decided to move away from covering COVID themes in our report and instead focus on more actionable issues within the data science, machine learning (ML), and artificial intelligence (AI) industries, like open-source security, the talent dilemma, ethics and bias, and more.

In the spirit of democratizing data, we are making the raw data from our 2022 State of Data Science survey available to the public via Anaconda Nucleus.

TABLE OF CONTENTS

METHODOLOGY

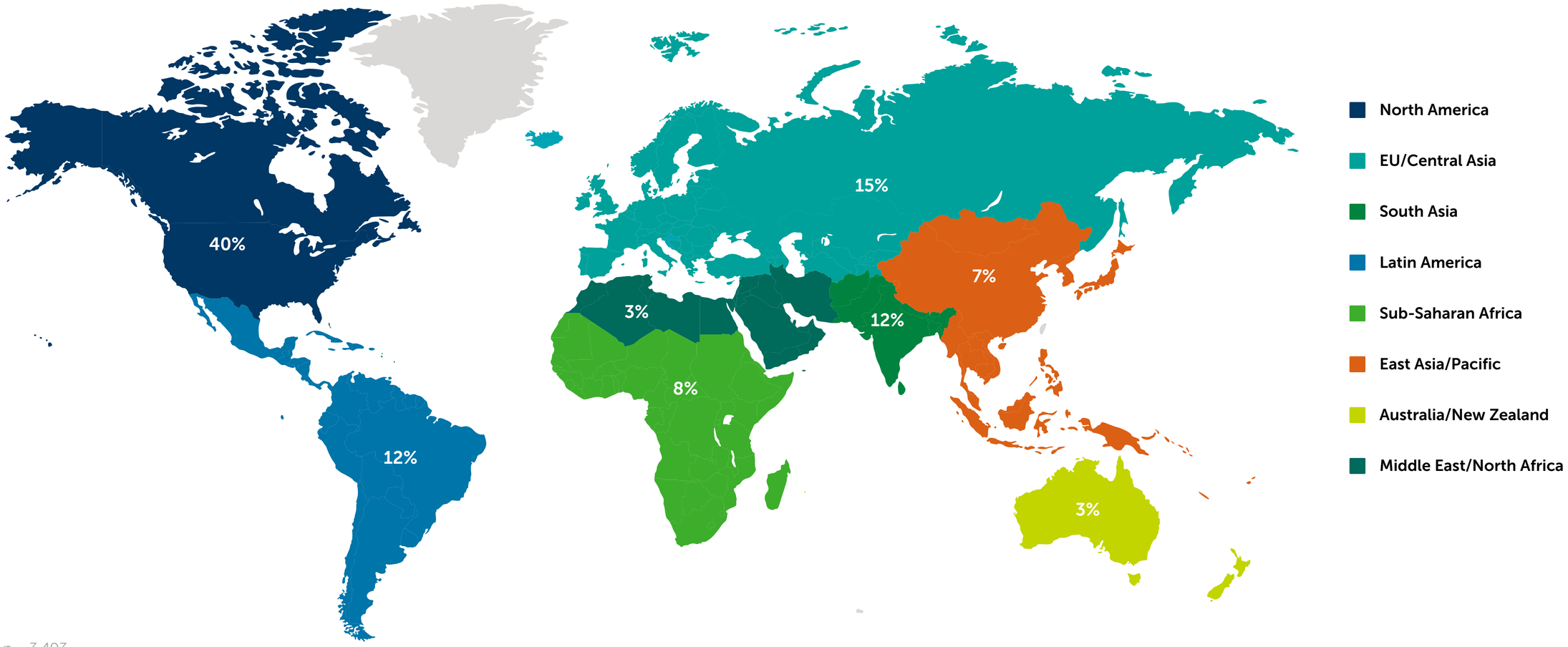
3,493 individuals from 133 countries and regions took part in our online survey conducted from April 25, 2022, to May 14, 2022. Respondents came from the Anaconda email database, Anaconda.org, social media, and other sources. They had the opportunity to participate in a sweepstakes drawing as an incentive for completing the survey, and five winners were selected at random after the survey was complete. The respondents were divided into three separate tracks: students, academics, and those working in commercial environments. Each of these different cohorts was asked some universal questions, while some questions were unique to each cohort's experience. In the report, we indicate whether responses came from the entire set of respondents or from a subset.

Note: All percentages are rounded to the nearest whole percent. Due to rounding, some numbers may not equal 100.

THE FACE OF DATA SCIENCE

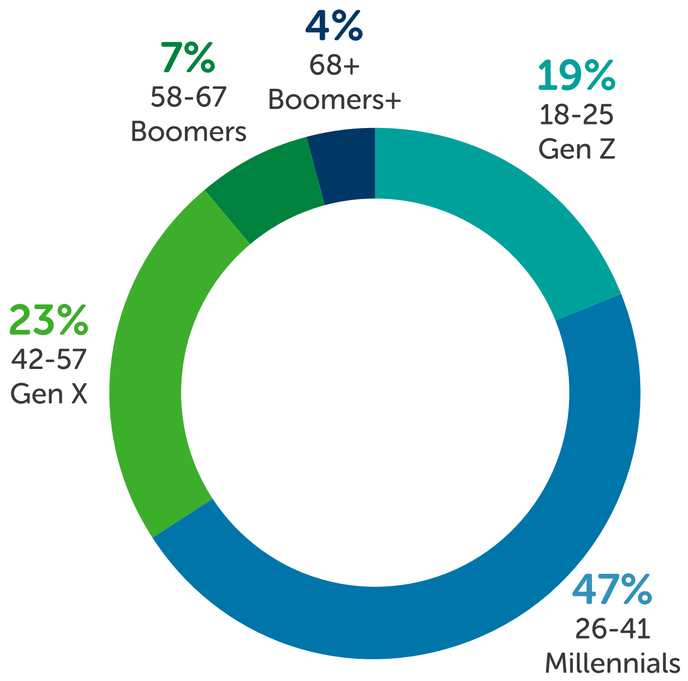
As with years past, we began our survey with a series of demographic questions. Our respondents span a vast range of geographical locations, ages, and job functions, and capturing their demographic info each year offers insight into how the data science community is evolving.

3,493 individuals from 133 countries and regions took part in our 2022 survey.



n = 3,493

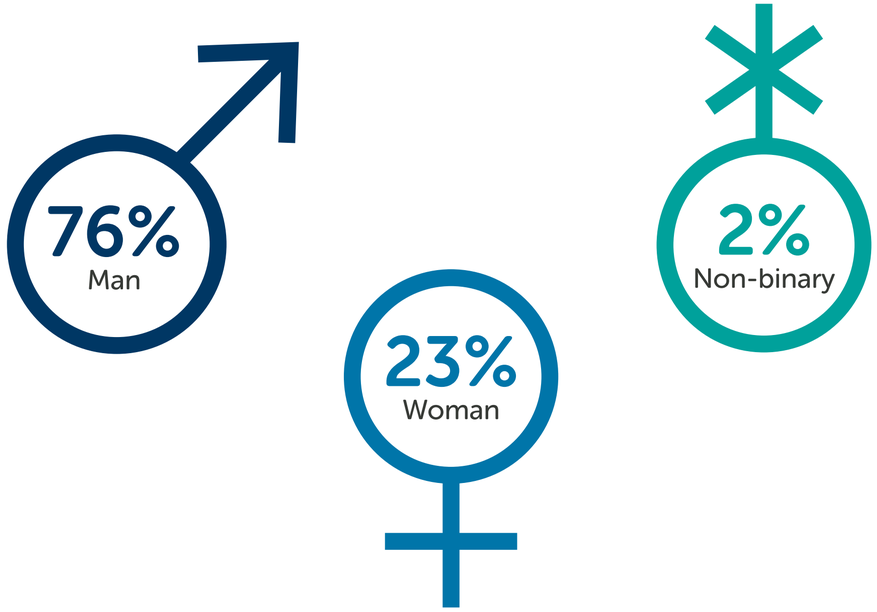
Respondent Age



n = 3,493

Our respondent set skews toward younger generations. Of the 3,493 respondents, 66.54% are either Generation Z (19.47%) or Millennials (47.07%). We saw an increase in Generation X respondents by about 5% compared to 2021. Only 10.48% of respondents are 58 or older.

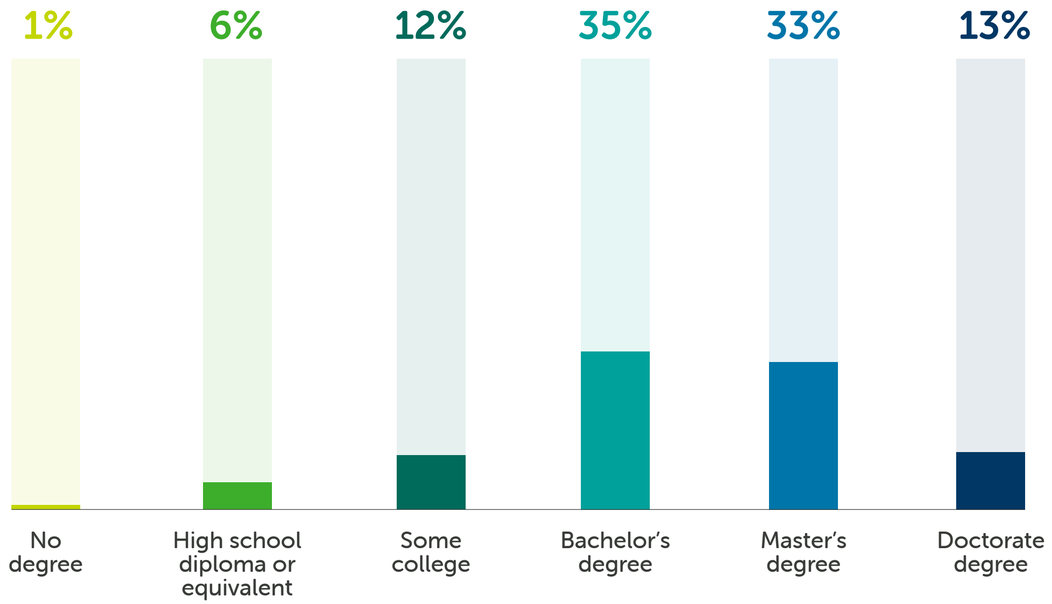
Respondent Gender



n = 3,493

The gender data from our respondents supports what the community continues to see across male-dominated STEM fields. The industry as a whole demonstrates room for improvement when it comes to increasing gender diversity.

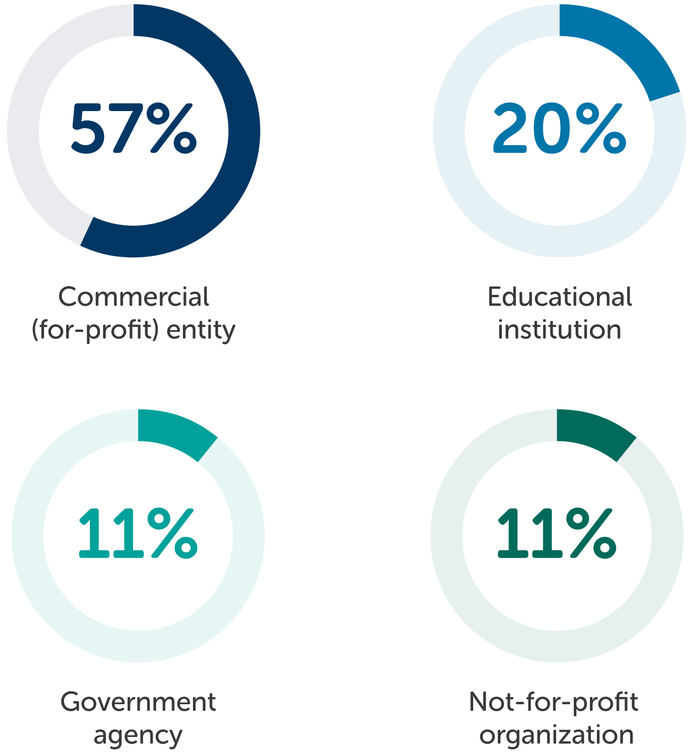
Respondent Education Level



n = 3,493

The majority of our survey respondents are well-educated. 80.71% have at least a college-level degree, and there was about a 12% year-over-year (YoY) increase in the number of respondents who hold advanced degrees. 19.29% of respondents do not hold any degree.

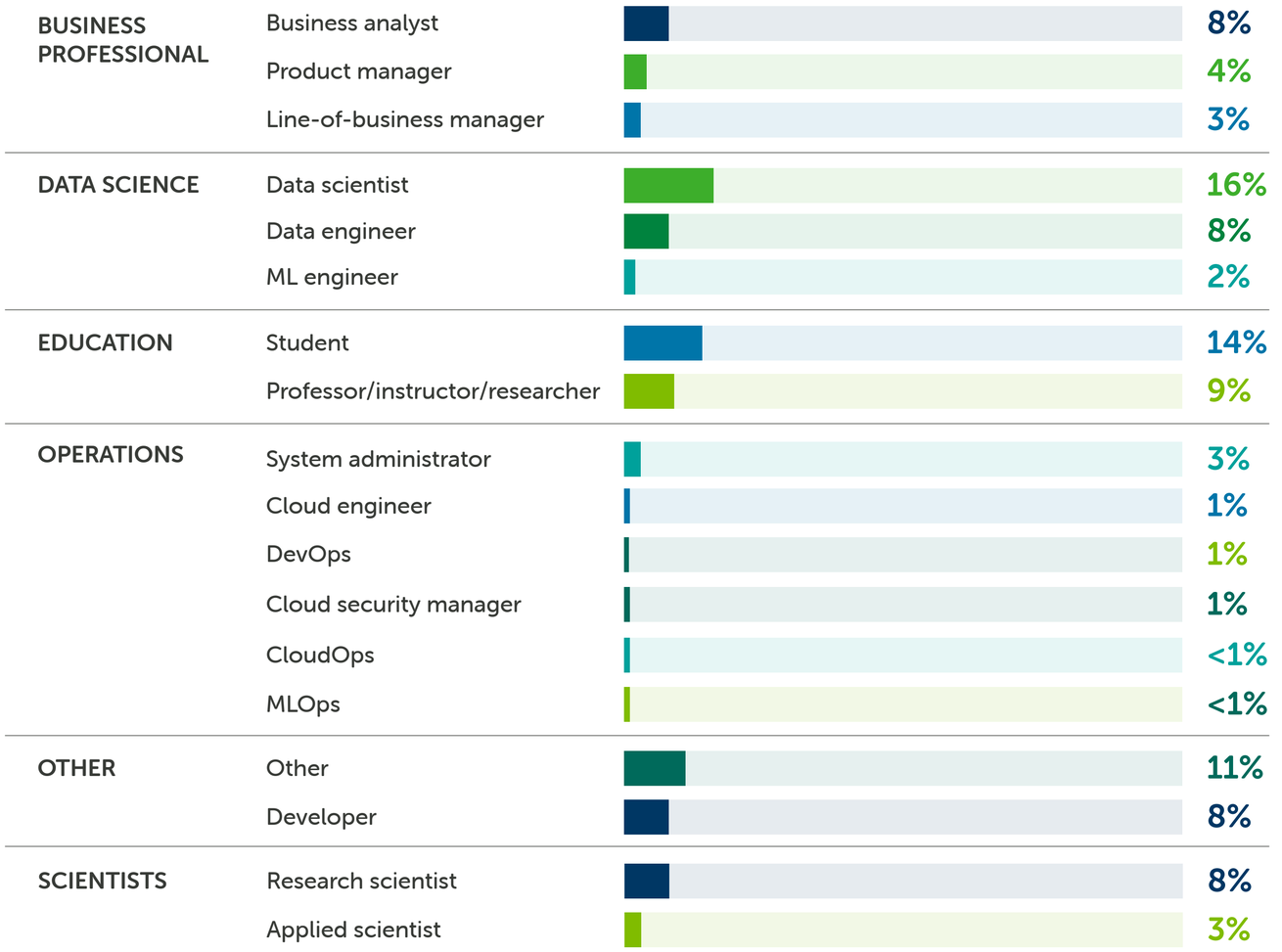
Respondent Company Type



n = 2,924

Respondent Primary Job Function

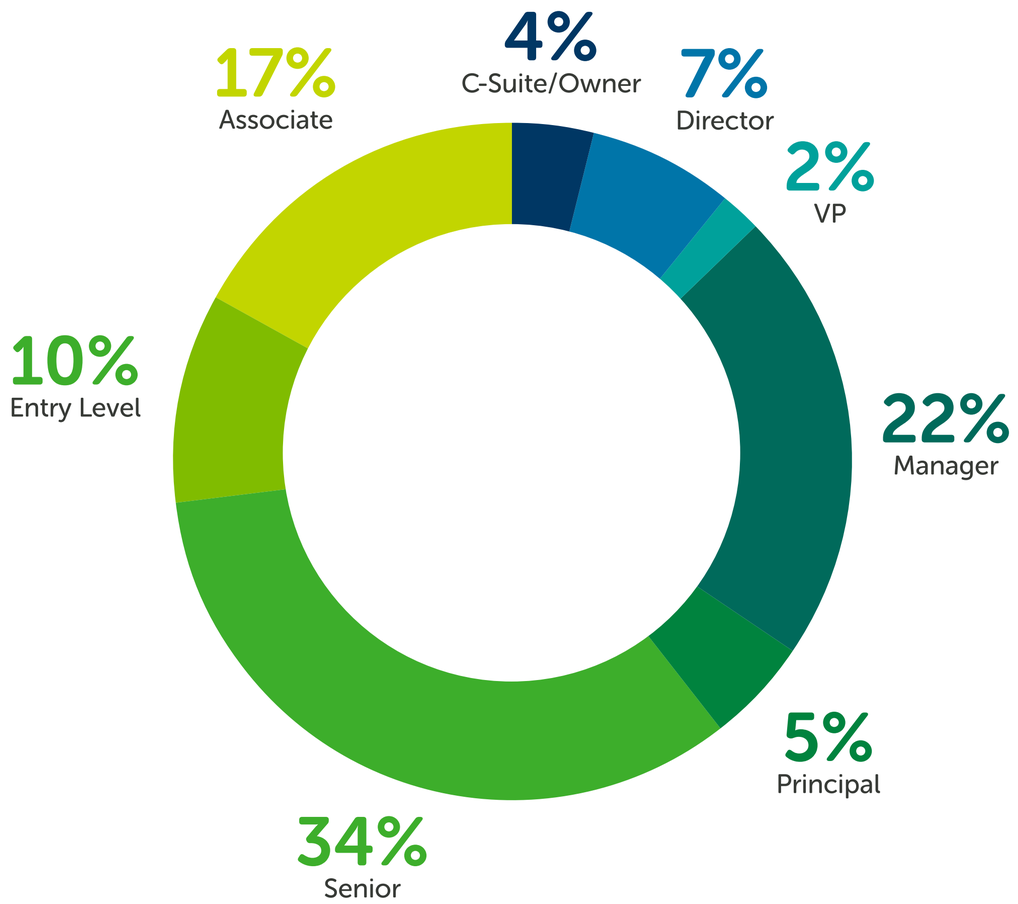
The current professional landscape includes a wide variety of data-focused roles, and there is often overlap between job functions—and in some cases, individual titles apply to multiple tasks. With that in mind, we asked respondents to select the role that best reflects their job function. 16.46% of respondents are data scientists. More data scientists (+5%) and fewer students (-13%) responded this year compared to 2021.



n = 3,493

Respondent Current Job Level

Most (33.93%) of survey respondents hold senior-level positions, a 9% increase from 2021. The percentage of respondents who hold entry-level positions has decreased by 5%, and the percentage of respondents who hold a VP-level or C-suite position has decreased by about 7%.



n = 1,966

DATA PROFESSIONALS AT WORK

Most of our respondents work in commercial environments. We took a deeper dive into their responses to ascertain where data professionals sit within their organizations, how they spend their time, what tools they use, and their most significant challenges.

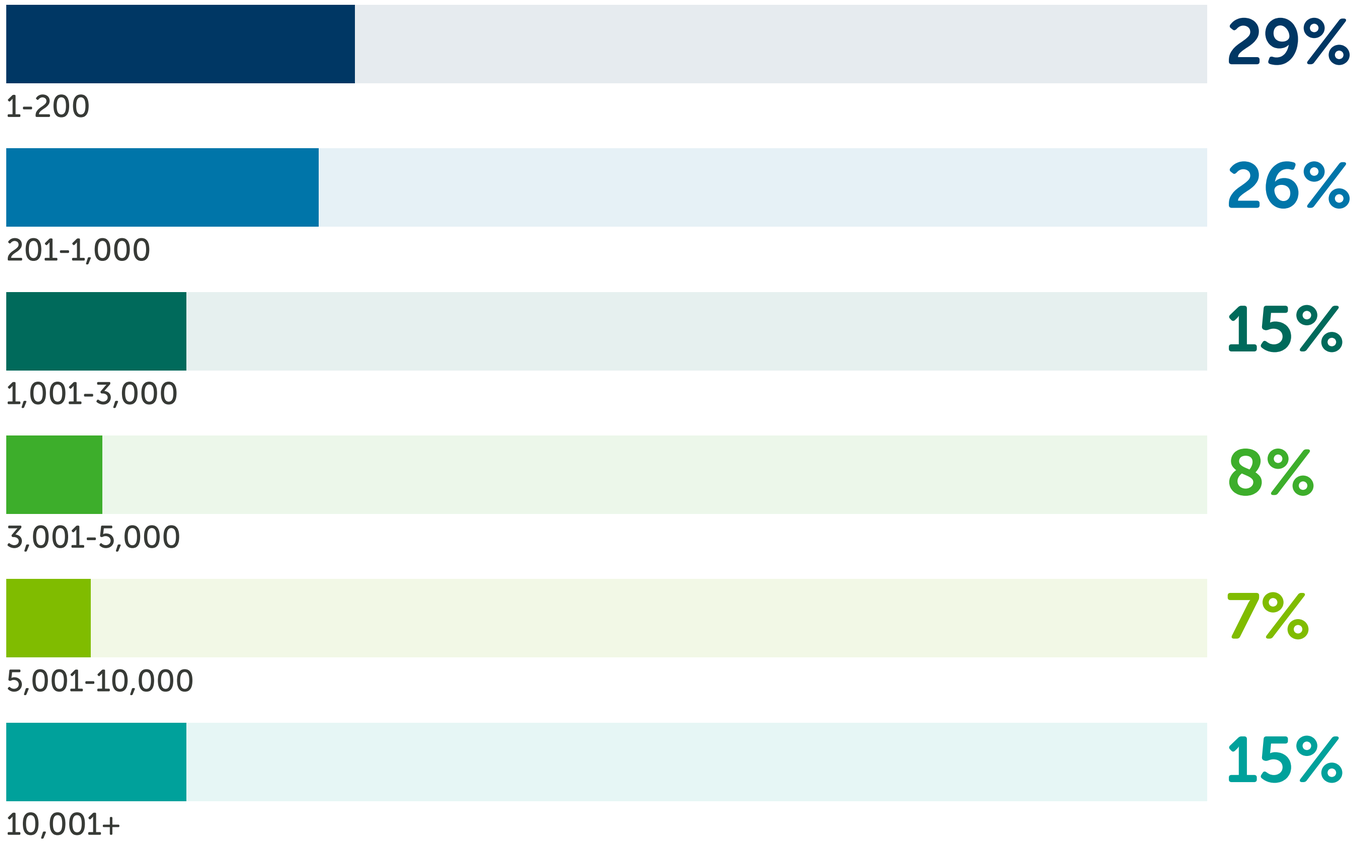
Respondent Industry

Technology	11%	Defense	3%
Finance	9%	Professional services	3%
Consulting	8%	Telecommunications	3%
Healthcare	5%	Insurance	3%
Automotive	5%	Media	2%
Government	5%	Transportation and logistics	2%
Research and development (R&D)	5%	Agriculture	2%
Electronics	5%	Food and beverage	2%
Manufacturing	5%	Other	1%
Engineering	5%	Nonprofit	1%
Construction	4%	Pharmaceutical	1%
Energy	4%	Entertainment	1%
Education	3%	Hospitality	1%
Retail	3%	Utilities	1%

n = 1,966

Companies across a wide variety of sectors—from government to telecommunications to nonprofit to pharmaceutical—rely on data-driven roles. The top five industries represented in our survey are technology, finance, consulting, healthcare, and automotive.

Company Size



n = 1,966

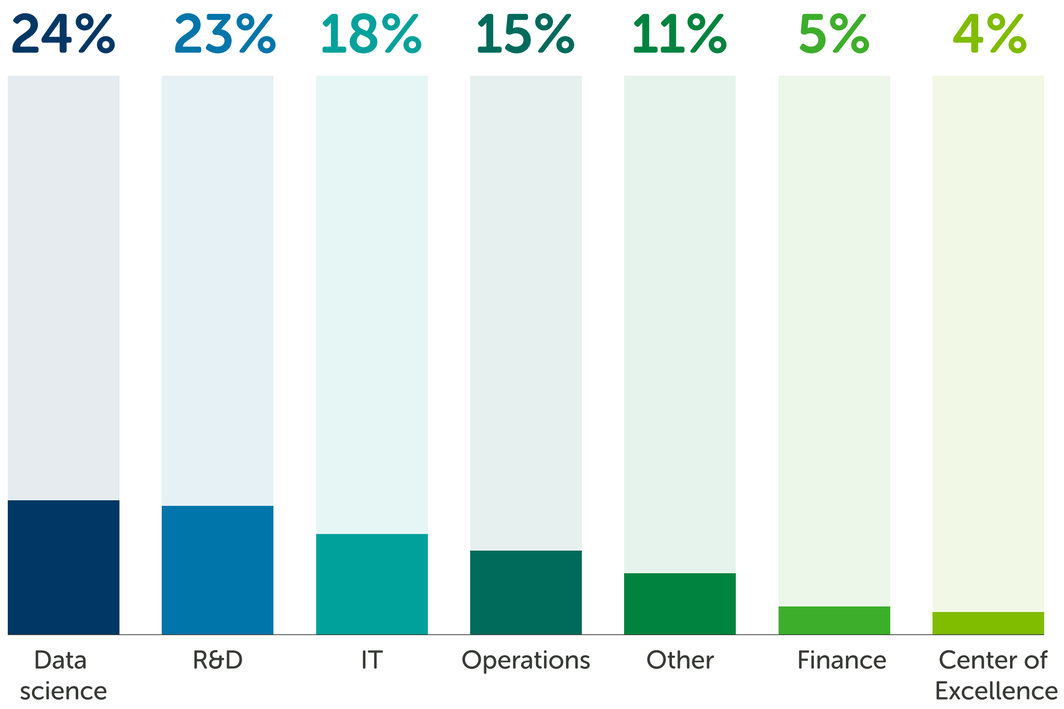
55.14%

of respondents work for companies with 1,000 employees or less.

What department does your role fall under?

Where do data roles fall within organizations? The short answer is: everywhere. The longer answer is, it depends on the specific organization. Sometimes there is an entire data-focused team; other times, data scientists work within other departments like finance or even marketing.

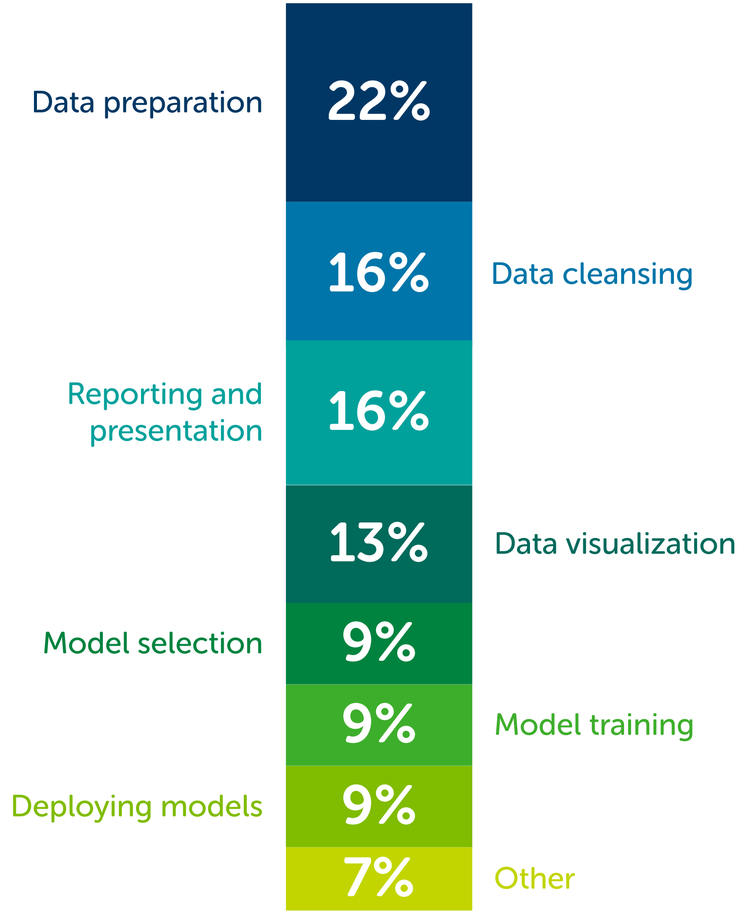
Most of our commercial respondents (24.47%) work within a data science department, while 22.89% work in R&D and 18.01% work in information technology (IT).



n = 1,966

How do data scientists spend their time?

Data professionals spend their time on a variety of tasks that require diverse technical and non-technical skills. Respondents indicated they spend about 37.75% of their time on **data preparation and cleansing**. Beyond preparing and cleaning data, **interpreting results** remains critical. Data visualization (12.99%) and demonstrating data's value through reporting and presentation (16.20%) are essential steps toward making data actionable and providing answers to critical questions. **Working with models** through selection, training, and deployment takes about 26.44% of respondents' time (-8.56% YoY).



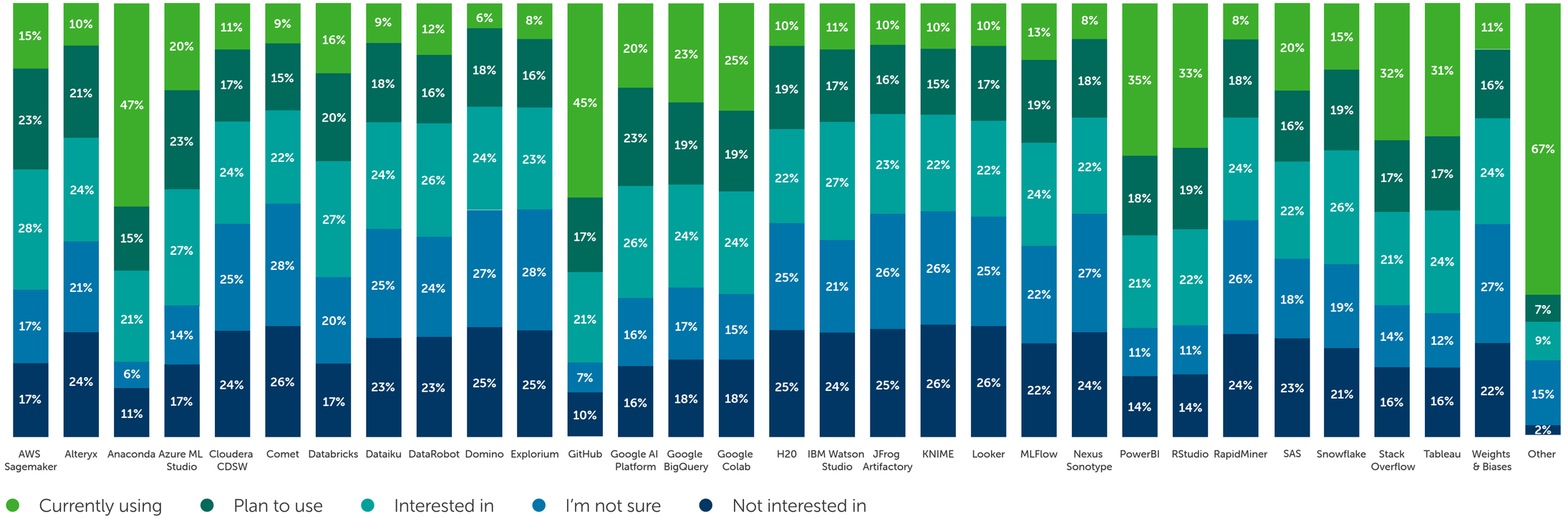
n = 1,966

We asked our respondents how much time they spend on the above tasks, and for each item they entered a number reflecting the percentage of time spent relative to the other options. This is the average of the reported percentages.

Data Science and ML Measures and Tools

Which of the following tools are being used within your organization?

Given our survey sample, it's no surprise that 46.83% of commercial respondents indicated their organizations currently use Anaconda. Other popular tools that organizations are currently using include GitHub (44.94%), RStudio* (33.33%), Stack Overflow (31.57%), and Tableau (30.65%).



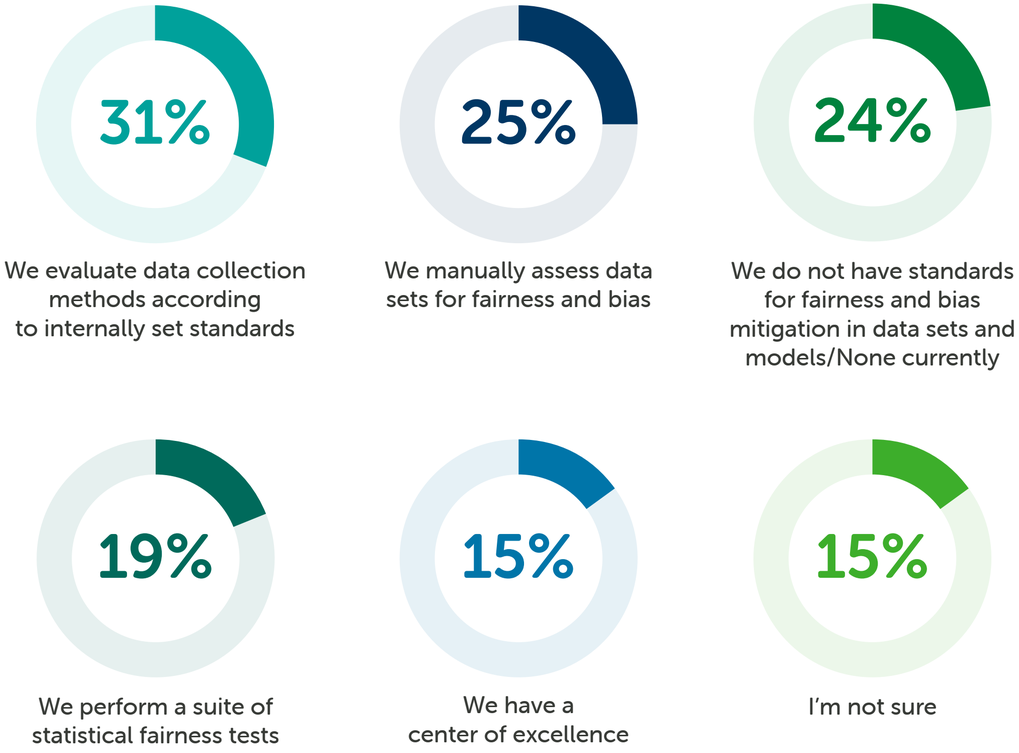
*Note that after our survey concluded, RStudio was rebranded as Posit.
n = 1,373

Data Science and ML Measures and Tools

What kinds of measures or tools do you or your institution use to ensure fairness and mitigate bias in data sets and models?

In our 2021 State of Data Science report, about 40% of survey respondents indicated that their organizations had implemented or planned to implement steps to ensure fairness and mitigate bias over the following 12 months.

This year, we wanted to dig into the specific steps organizations are now taking to ensure fairness and mitigate bias. The most common step is **evaluating data collection methods according to internally-set standards** (30.61%), with the second-most-common step being **manually assessing data sets for fairness and bias** (24.84%). 23.64% of respondents indicated that their organizations do not have standards surrounding/have not implemented measures or tools to address fairness and bias mitigation in data sets and models, and 14.89% aren't sure about their organizations' efforts.



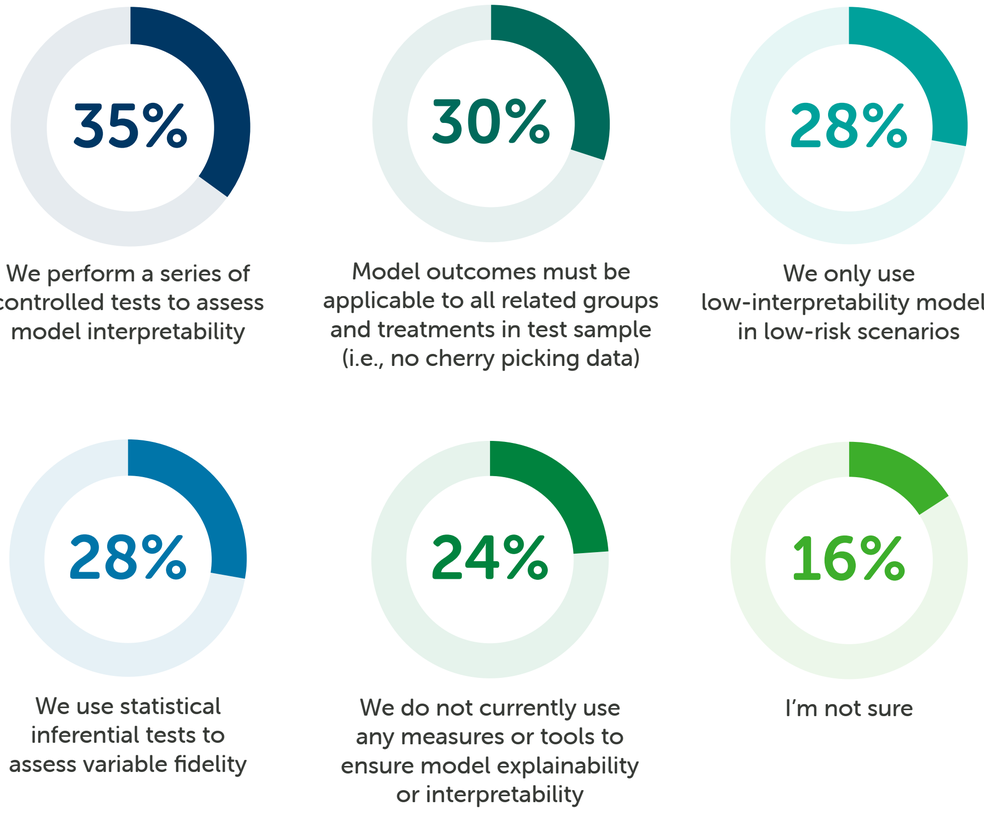
n = 1,578

Data Science and ML Measures and Tools

What kinds of measures or tools do you or your institution use to address model explainability and interpretability?

In our 2021 State of Data Science report, about 41% of survey respondents indicated that their organizations had implemented or planned to implement steps to address model explainability over the following 12 months.

This year, we wanted to dig into the specific steps organizations are now taking to address model explainability. The most common step is **performing a series of controlled tests to assess model interpretability** (35.36%), with the second-most-common step being ensuring **model outcomes are applicable to all related groups and treatments in test samples (i.e., no cherry picking data)** (30.23%). 23.76% of respondents indicated that their organizations do not use any measures or tools to ensure model explainability or interpretability, and 16.41% aren't sure about their organizations' efforts.



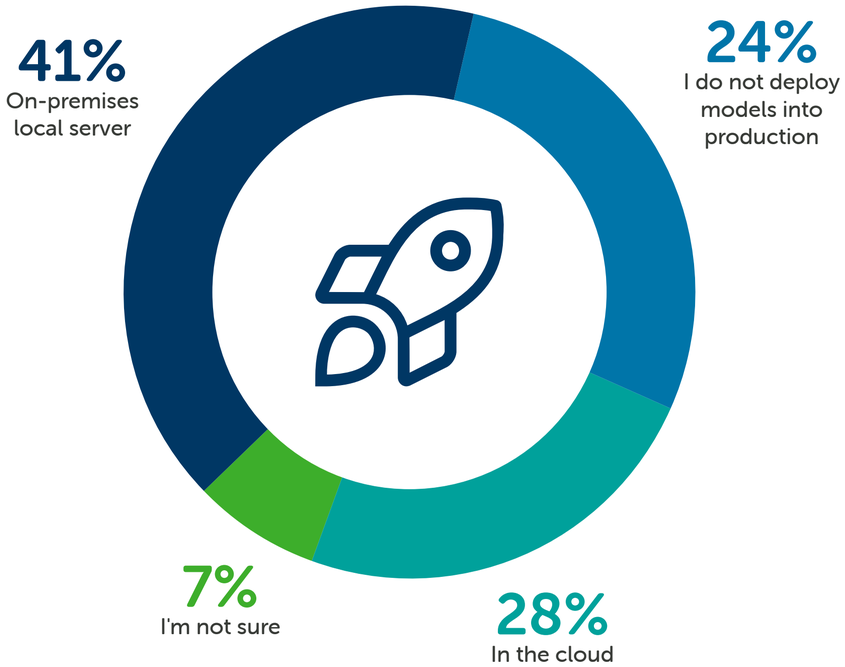
n = 1,578

Getting to Production

Organizations that deploy ML models and other data science outputs to power business functions and products enjoy a competitive advantage that leans on the value data scientists provide. While getting models to production can be one of the most rewarding functions of a data professional, it can also present unique challenges.

Where do you deploy models into production?

Only 23.70% of commercial respondents are not deploying models into production, which means the majority (69.20%) are deploying models into production (7.10% aren't sure), typically via an on-premises local server (41.32%) or the cloud (27.88%).

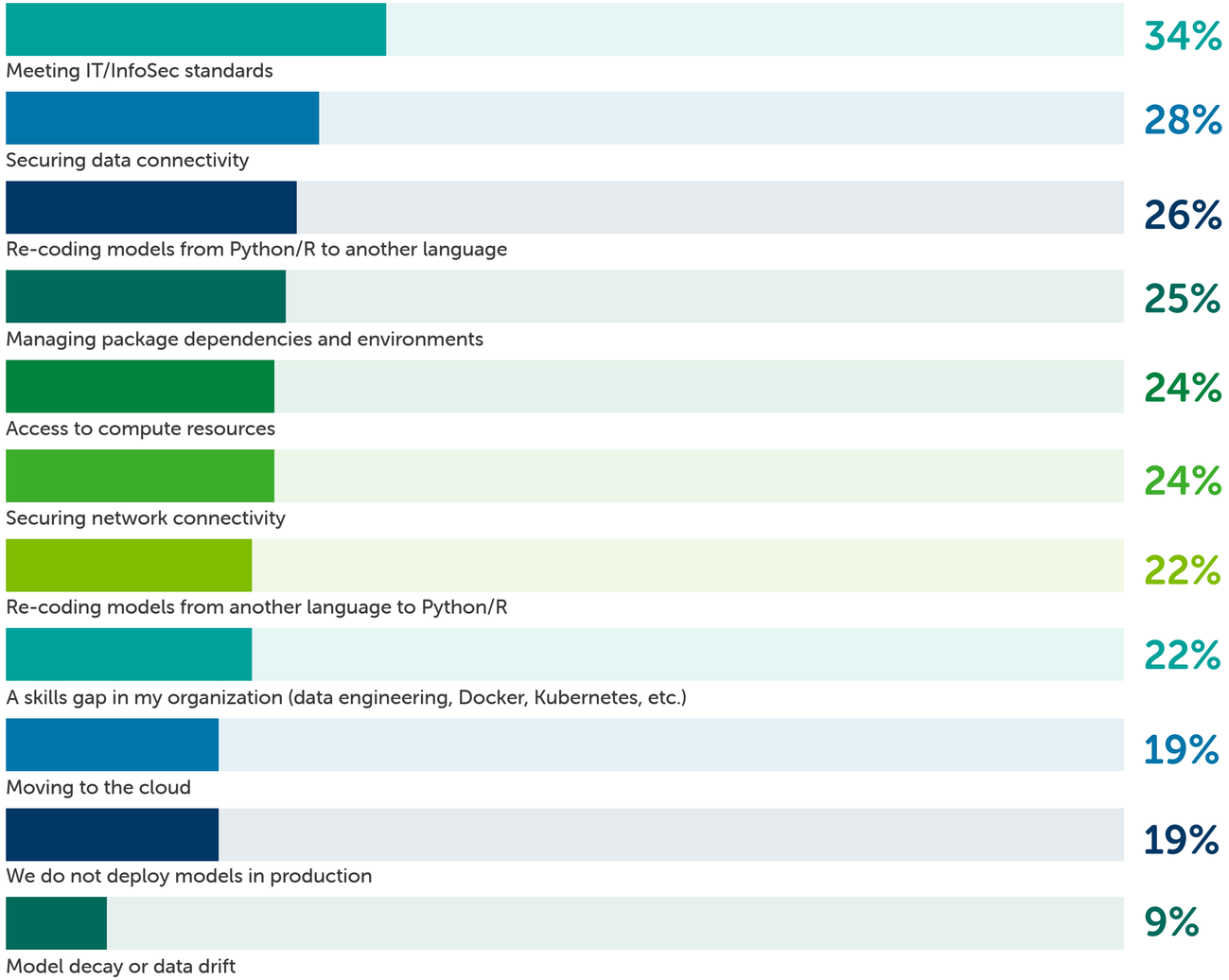


n = 1,578

Getting to Production

What roadblocks do you face when moving models to a production environment?

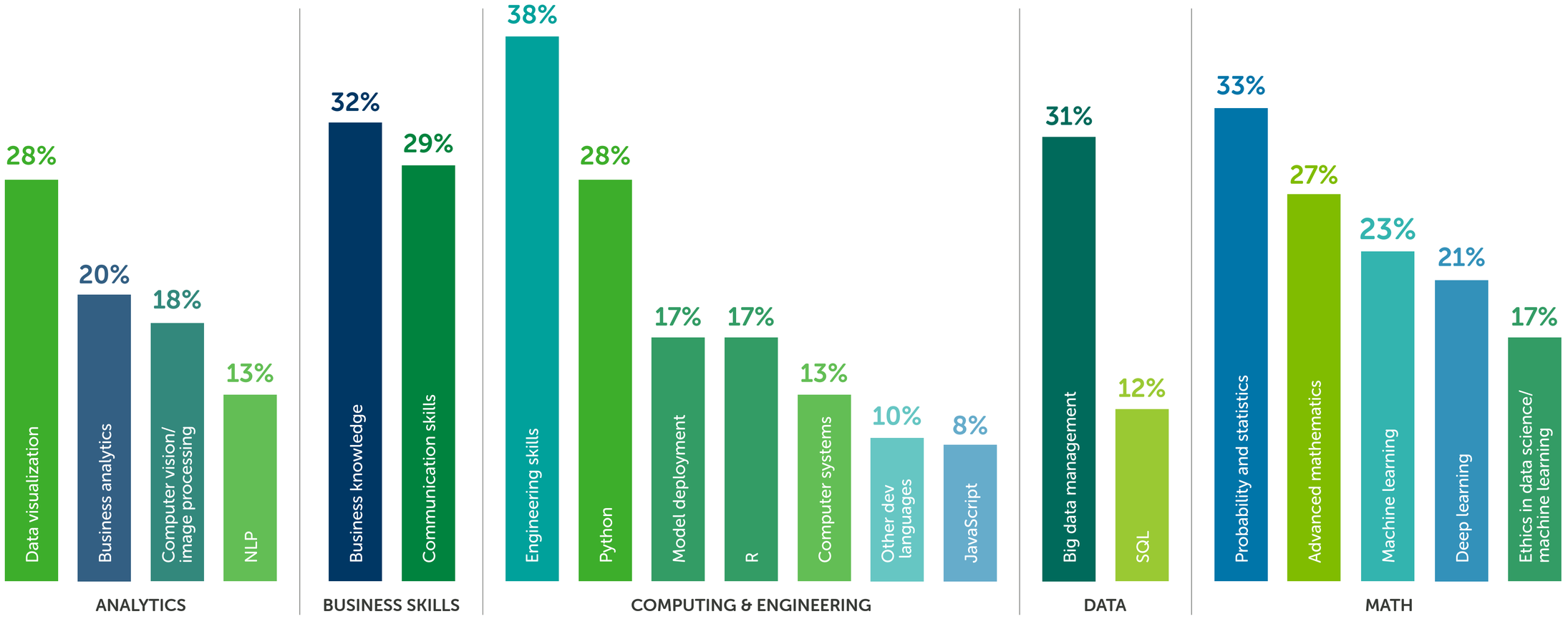
Most (41.32% of) commercial respondents deploy models into production via an on-premises local server. The top roadblocks respondents face when moving models to a production environment include meeting IT/InfoSec standards (33.88%), securing data connectivity (28.45%), and re-coding models from Python/R to another language (26.12%).



n = 1,455

Skills Gaps and Continued Learning

In your opinion, what are the most important skills/areas of expertise missing in the data science/ML area of your organization?

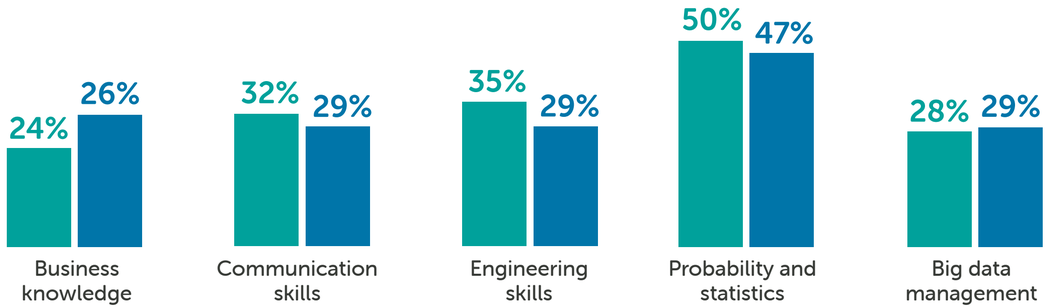


n = 1,443

Skills Gaps and Continued Learning

In your opinion, what are the top five most important skills or areas of expertise?

According to respondents, the top five most important skills/areas of expertise missing in the data science/ML areas of their organizations are engineering skills (38.12%), probability and statistics (33.26%), business knowledge (32.22%), communication skills (30.56%), and big data management (29.24%). But are educational institutions teaching these skills, and are students learning them? Let's take a look:



■ Educator respondents:
What topics, tools, or skills is your institution teaching students of data science and machine learning?

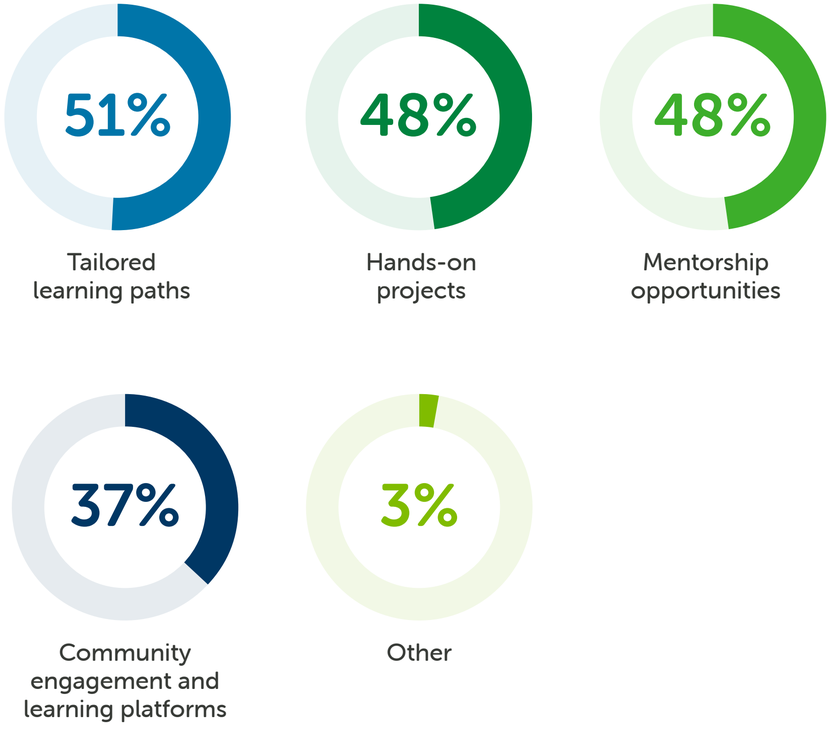
■ Student respondents:
What topics, tools, or skills are covered in your courses in preparation for entering the data science/ML field?

n = 517 and n = 407

There seems to be a fairly close correlation between the number of educational institutions teaching engineering skills, probability and statistics, business knowledge, communication skills, and big data management and the number of students learning about these topics. Of the topics, probability and statistics are covered and learned most frequently, while business knowledge is covered and learned least frequently. Overall, there is opportunity for both teachers and students alike to place greater emphasis on these skills in order to better prepare students to enter the workforce.

What tools and resources do you feel are lacking for data scientists who want to learn and develop their skills?

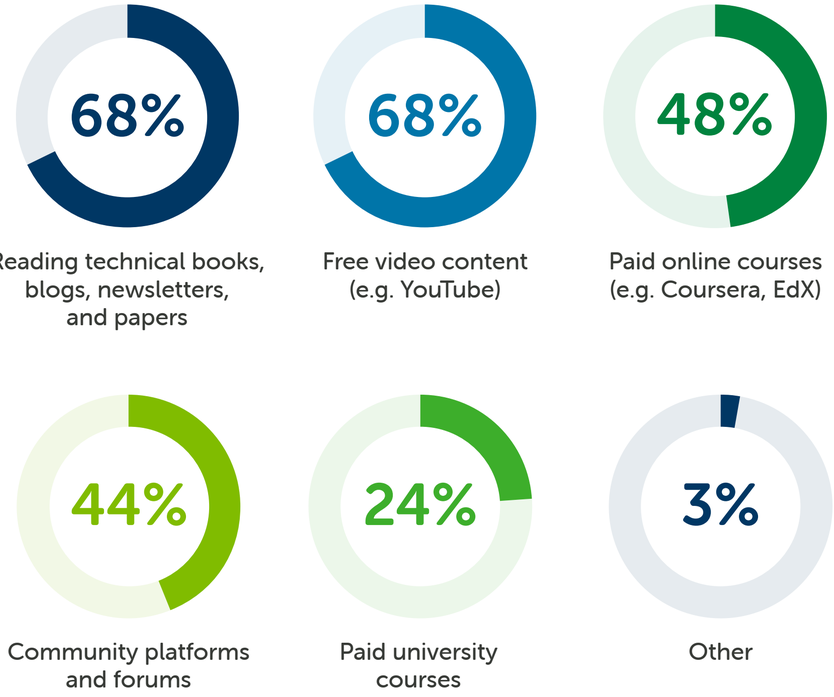
Tailored learning paths (51.30%), hands-on projects (48.33%), and mentorship opportunities (47.63%) are the top tools and resources respondents feel are lacking for data scientists who want to learn and develop their skills.



n = 2,154

How do you typically learn about new tools and topics relevant to your role?

Most respondents typically learn about new tools and topics relevant to their roles by reading technical books, blogs, newsletters, and papers (68.25%) or via free video content (67.97%).

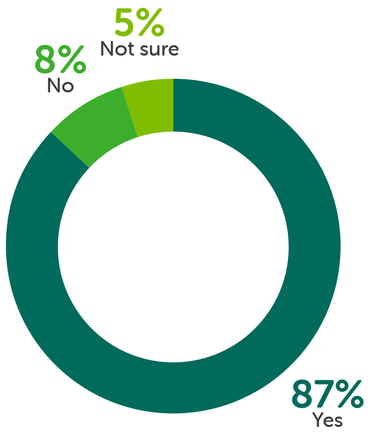


n = 2,154

ENTERPRISE ADOPTION OF OPEN SOURCE

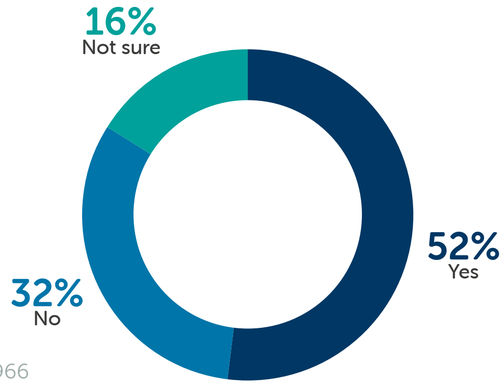
Using and contributing to open-source software (OSS) sets the most innovative organizations apart and often saves them significant time and resources. As opposed to purchasing tools from a vendor or building them in-house, the crowdsourced OSS model leverages multiple minds at work to accelerate projects that could otherwise take years.

Does your organization allow the use of open-source software?



n = 1,966

Does your employer encourage you and your team to contribute to open-source projects?

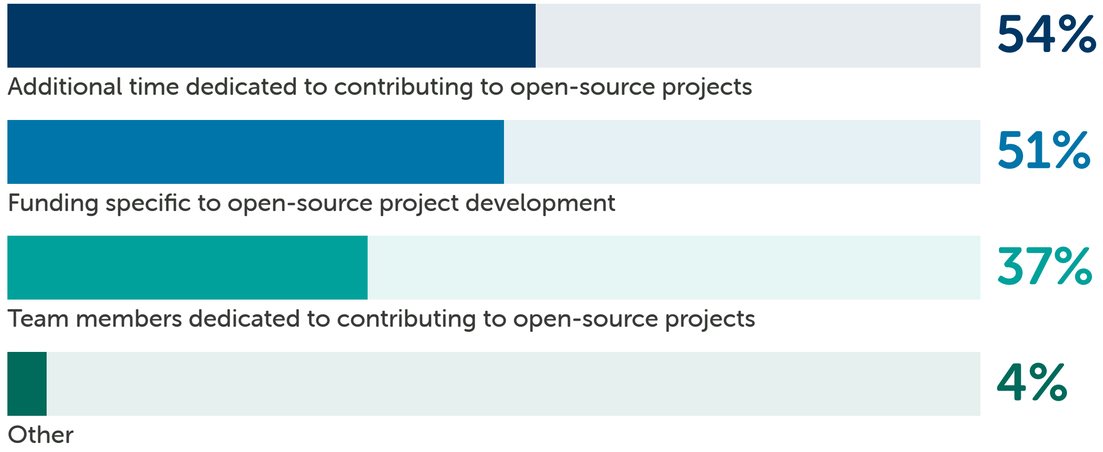


n = 1,966

How does your employer empower you and your team to contribute to open source?

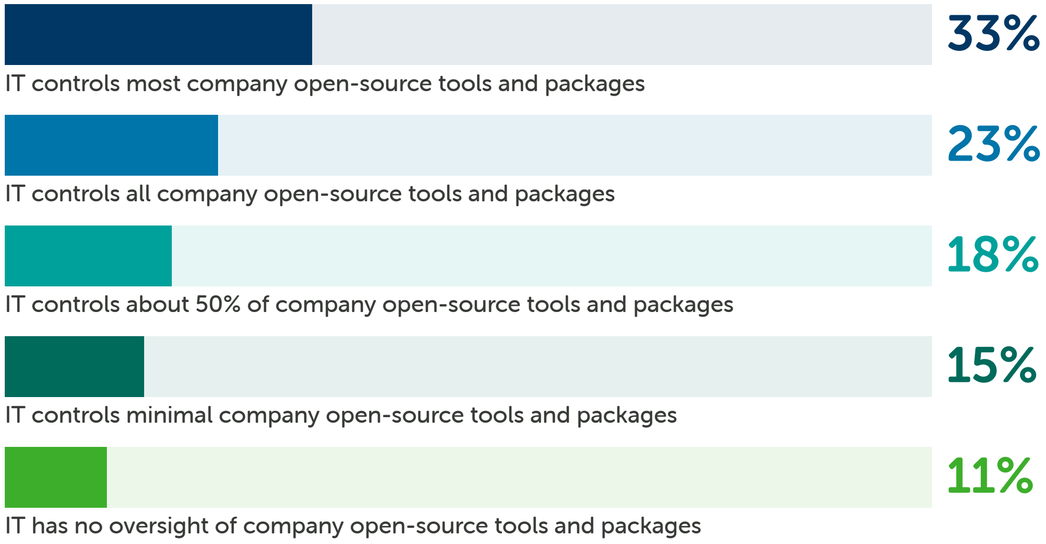
In 2021, 65% of commercial respondents said their teams were encouraged to contribute to open-source projects, and the majority of those respondents (54%) said that their employers were empowering them to contribute to open source through an **increase in funding related to open-source project development.**

This year, just 51.99% of commercial respondents said their teams are encouraged to contribute to open-source projects—about a 13% YoY decrease, perhaps due to security concerns. The majority of these respondents (54.04%) said that their employers are empowering them to contribute to open source through **additional time dedicated to contributing to open-source projects.**



n = 890

How much control do IT users have over the open-source tools and packages your company uses?



n = 1,526

Within the subset of commercial respondents whose organizations allow the use of open-source software, 88.99% indicated that IT controls company open-source tools and packages to some extent, with 56.49% indicating that IT controls most or all company open-source tools and packages.

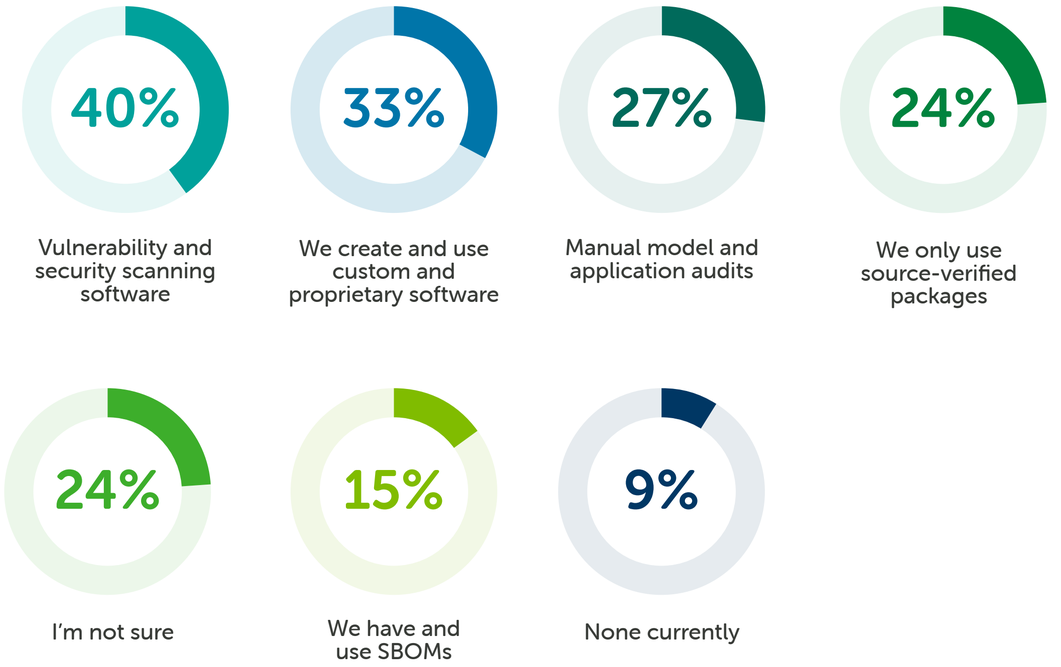
Securing the Software Supply Chain and Open-Source Pipeline

Like any software, open source comes with inherent security challenges. Multiple individuals are often involved in OSS creation, maintenance, and evolution, which can increase risk. On the bright side, it can also allow for vulnerabilities to be caught and patched more quickly.

It's certainly possible to reap the benefits of open-source software while being mindful of pipeline management and security. In this section, we'll look at practices and sentiments surrounding software supply chain and open-source security.

When asked about how organizations that use OSS ensure their supply chains are secure and meet enterprise security standards, 40.39% of respondents say they use vulnerability and security scanning software, 32.76% create and use custom and proprietary software, and 27.48% do manual model and application audits. Only 8.70% are not securing their open-source supply chains, and 23.69% aren't sure.

How does your organization ensure its open-source supply chains are secure and meet enterprise security standards?

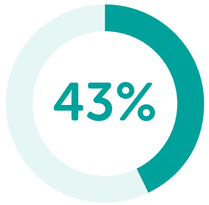


n = 1,874

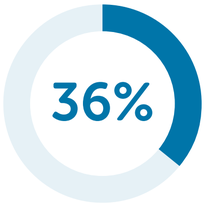
Securing the Software Supply Chain and Open-Source Pipeline

Securing the Open-Source Pipeline

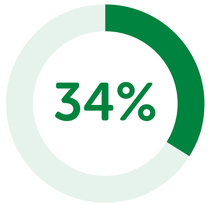
When asked about how organizations that use OSS ensure their data science and ML packages are secure and meet enterprise security standards, 43.03% of respondents say they use a managed repository, 35.76% use a vulnerability scanner (+5.76% YoY), and 34.13% do manual checks against a vulnerability database. 19.17% are not securing their open-source pipelines (-5.83% YoY), and 23.11% aren't sure.



We use a managed repository



We use a vulnerability scanner



Manual checks against a vulnerability database

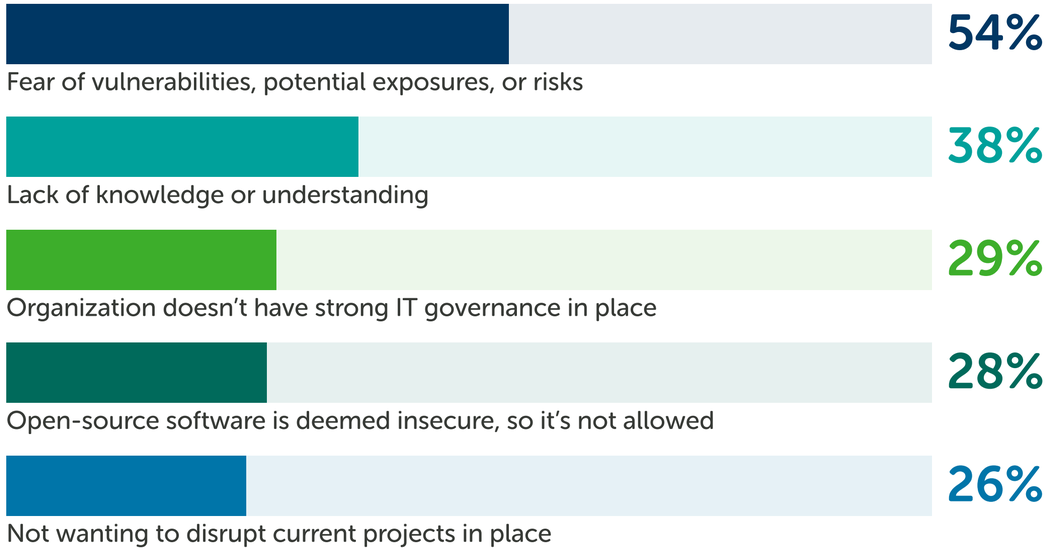
We are not securing our open-source pipeline

I don't know if we're securely downloading open-source packages

n = 1,471

Securing the Software Supply Chain and Open-Source Pipeline

What roadblocks are preventing your organization from leveraging open source?



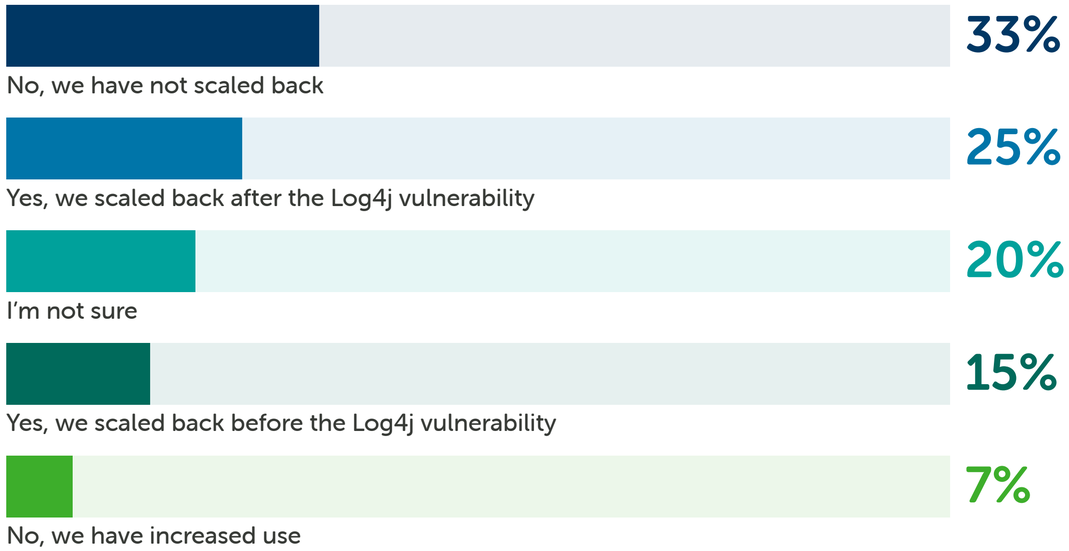
n = 145

According to our respondents, the majority of organizations are using open-source software. But of the 7.73% of respondents whose organizations are not, the biggest reason why is **fear of vulnerabilities, potential exposures, or risks** (54.48%) (+13.48 YoY).

Earlier, we called out the approximate 13% YoY decrease in the number of commercial respondents who said their teams are encouraged to contribute to open-source projects. It's possible that the 13.48% increase in the fear of vulnerabilities, potential exposures, or risks is related to this change.

Securing the Software Supply Chain and Open-Source Pipeline

Has your organization scaled back the use of open-source software in the past year due to concerns around security?



n = 1,526

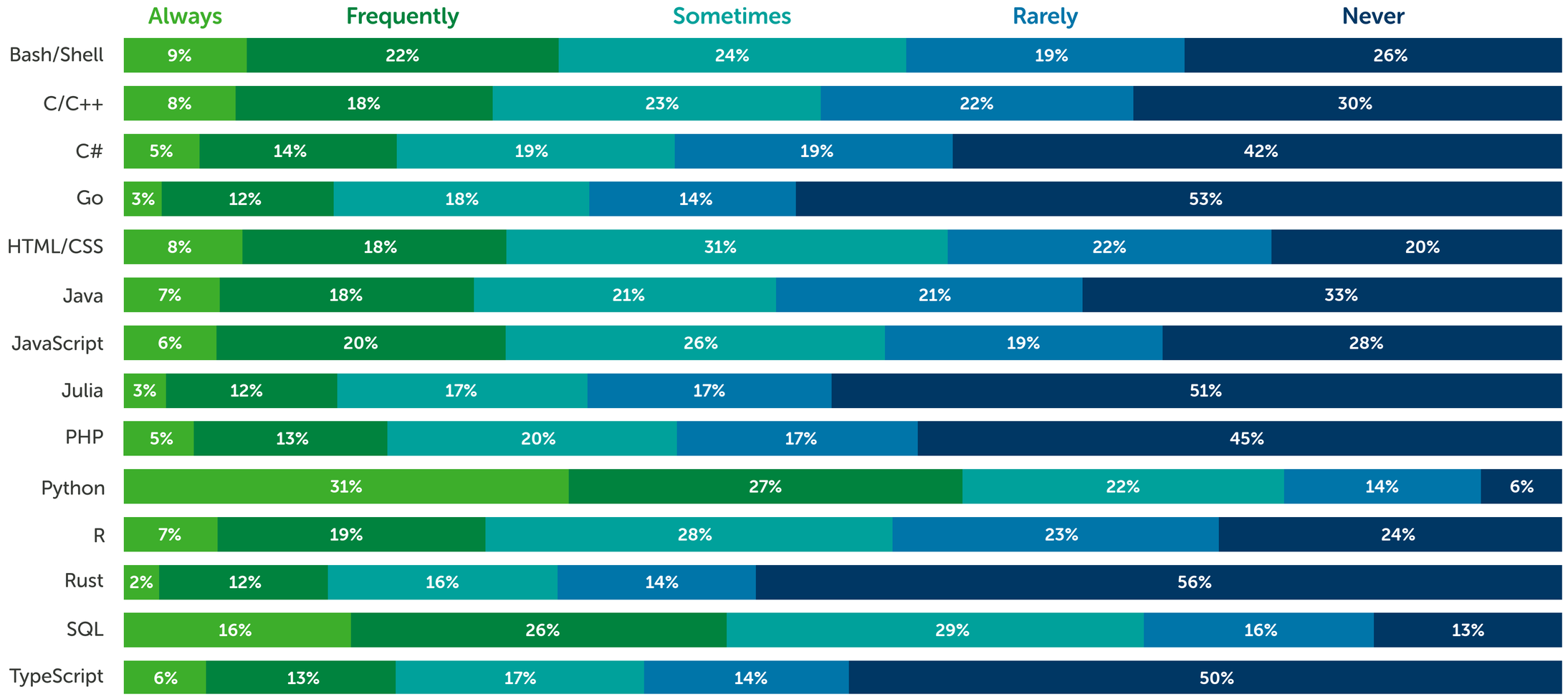
The Log4j incident in late 2021 was a disruptive and far-reaching example of an open-source security breach. 24.90% of commercial respondents indicated their organizations scaled back their open-source software usage after the incident.

32.96% of respondents indicated their organizations have not scaled back their open-source software usage due to concerns around security. This speaks to the value of open source; organizations don't want to miss out on the affordability, flexibility, and limitless technological advancement opportunities affiliated with OSS.

POPULARITY OF PYTHON

Python continues to be the preferred programming language among data scientists, researchers, students, and professionals worldwide.

How often do you use the following languages?



n = 2,274

How often do you use the following languages? (cont.)

The majority (58.26%) of survey respondents indicated they always or frequently use Python, and only 5.72% of respondents never use Python, making it the most popular choice out of all of the survey options—just like last year.

And why *wouldn't* Python be the most popular programming language? It's an accessible language that makes a wide variety of programming-driven tasks possible. There are hundreds of thousands of Python packages available, making Python applicable to many use cases. As such, Python is often used by non-programmers and students in addition to programming experts, and it makes a great teaching language for AI and ML. In fact, 61.32% of our academic-track respondents indicated their institutions are teaching Python to students of data science and machine learning, and 81.08% of student-track respondents indicated Python is covered in their courses.

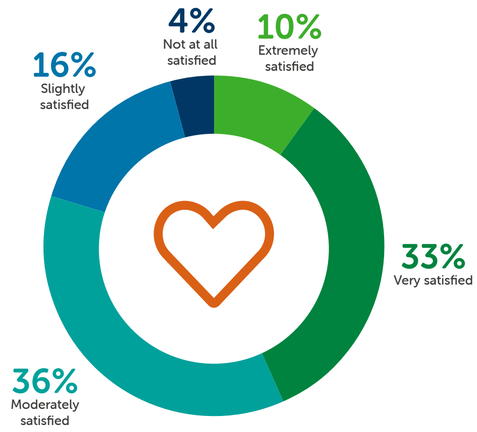
Anaconda pioneered the use of Python for data science, and in the past year we've taken big steps toward our goal of democratizing data science and advancing Python accessibility. We developed PyScript, a framework that allows users to create rich Python applications in the browser, and acquired PythonAnywhere, a cloud-based Python development and hosting environment that simplifies the web development process and empowers teams to write programs from any modern web browser.

DATA JOBS AND THE FUTURE OF WORK

According to the Bureau of Labor Statistics, "employment of computer and information research scientists is projected to grow 22 percent from 2020 to 2030, much faster than the average for all occupations." Considering this increasing demand for data talent alongside the current shortage of said talent, organizations must prioritize employee needs to attract and retain employees and minimize churn.

Job Satisfaction

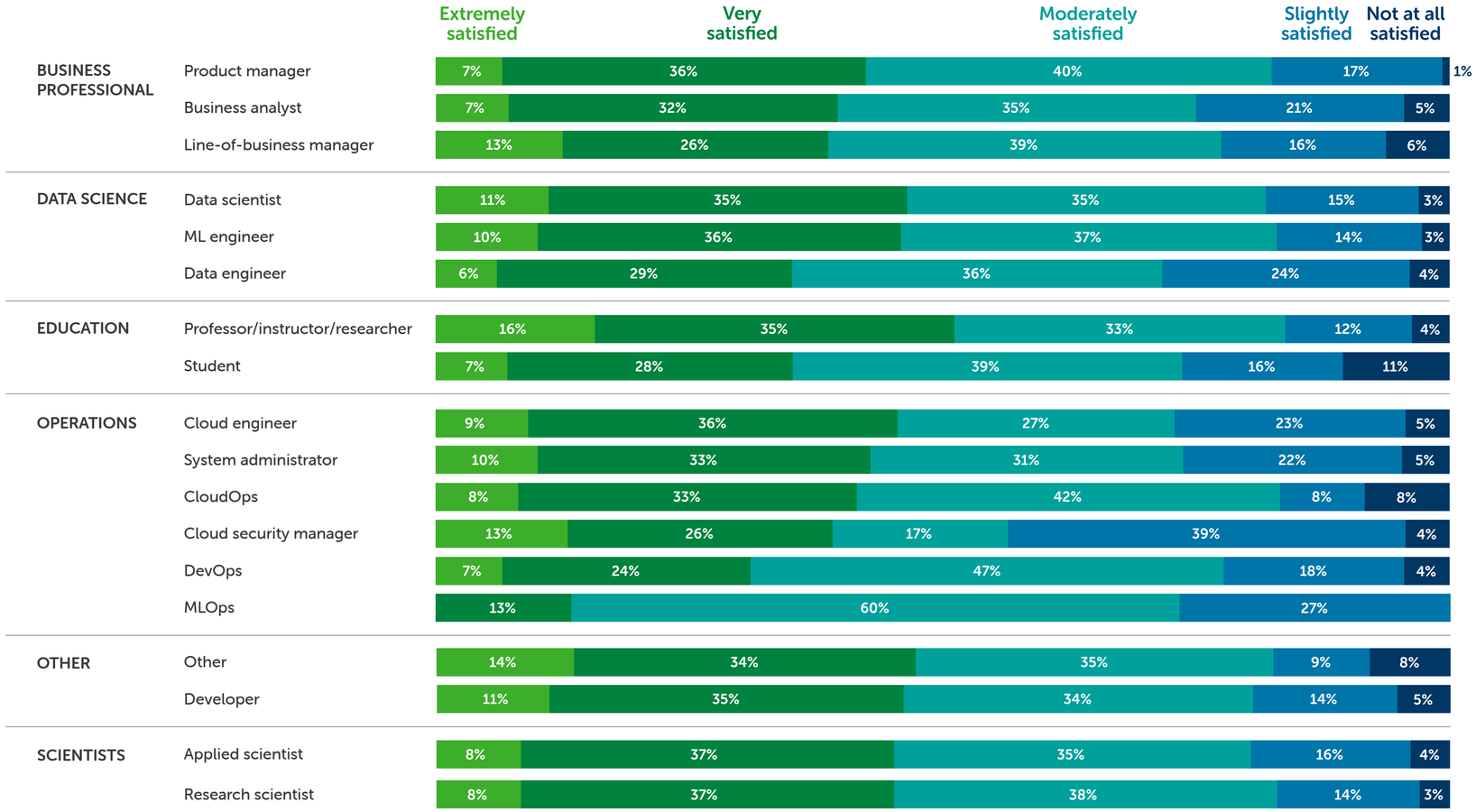
How would you rate your job satisfaction in your current role?



n = 2,893

The majority (69.27%) of survey respondents are moderately or very satisfied with their jobs. Only 4.42% are not at all satisfied, and only 10.30% are extremely satisfied.

Job Satisfaction by Role



n = 2,893

Job Satisfaction

What would cause you to leave your current employer for a new job?

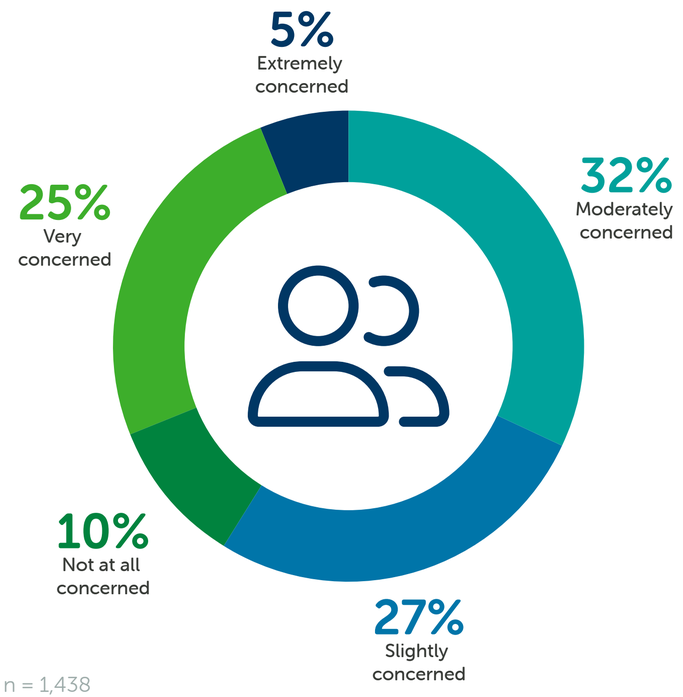
Respondents were asked to select the top option besides pay/benefits.

n = 2,581

31.64% of respondents indicated that they would be most motivated to leave their current employers for more responsibility/opportunity for career advancement.

The Talent Shortage

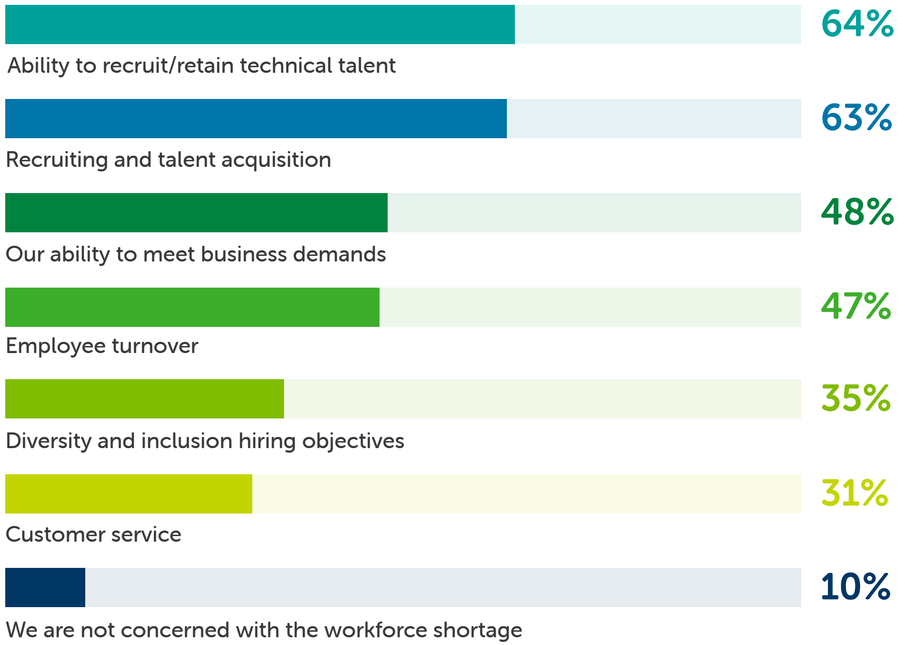
Attracting and retaining employees has challenged the technology industry more than ever before. How concerned is your organization about the potential impact of a talent shortage?



n = 1,438

Most (62.51% of) commercial respondents indicated that their organizations are at least moderately concerned about the potential impact of a talent shortage. Only 10.43% indicated that their organizations are not at all concerned.

What are you most concerned about with regard to the current workforce talent shortage?



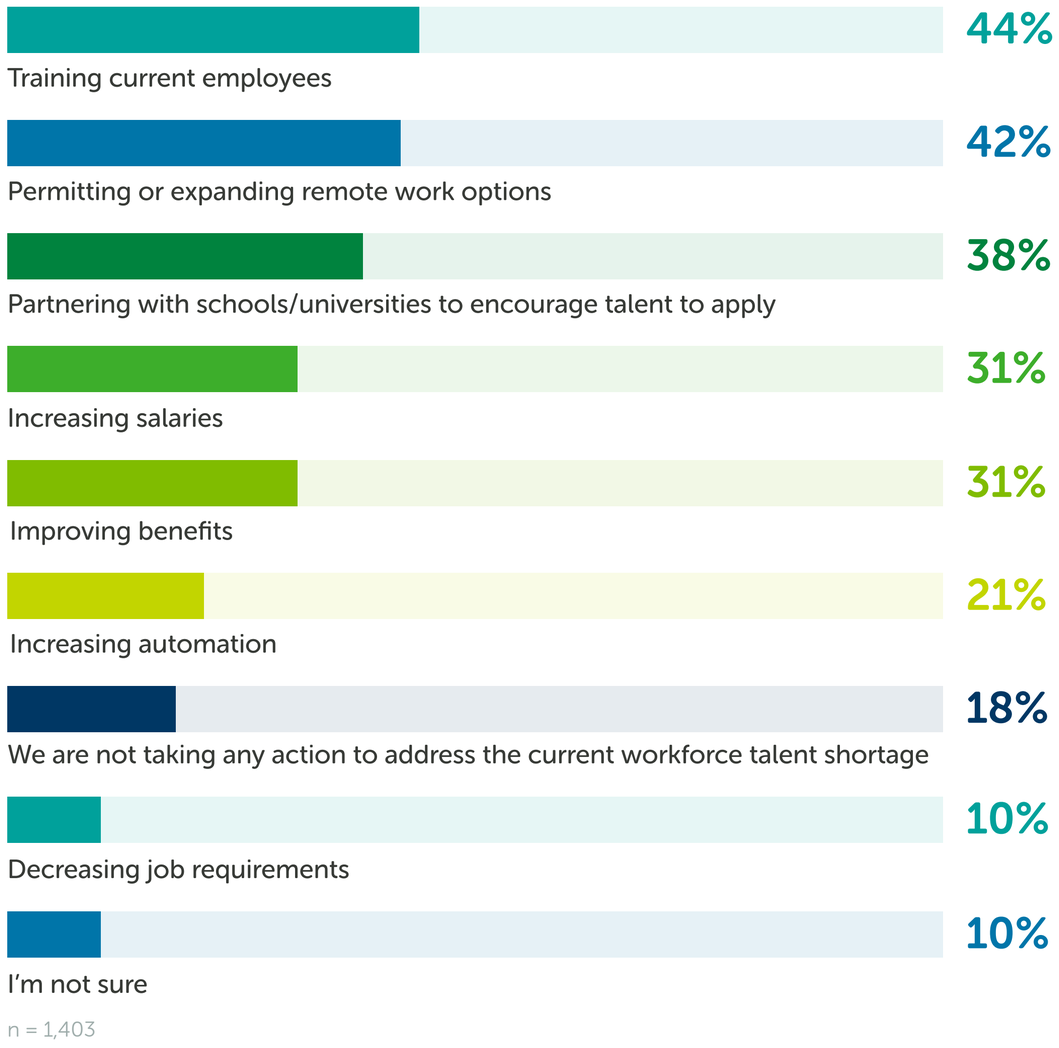
n = 1,241

Commercial respondents indicated that their top concerns surrounding the current workforce talent shortage are the ability to recruit/retain technical talent (63.66%) and more general recruiting and talent acquisition (62.77%). Respondents seem to be least concerned about diversity and inclusion hiring objectives (34.65%) and customer service (30.94%).

The Talent Shortage

Which of the following steps is your organization taking to address the current workforce talent shortage?

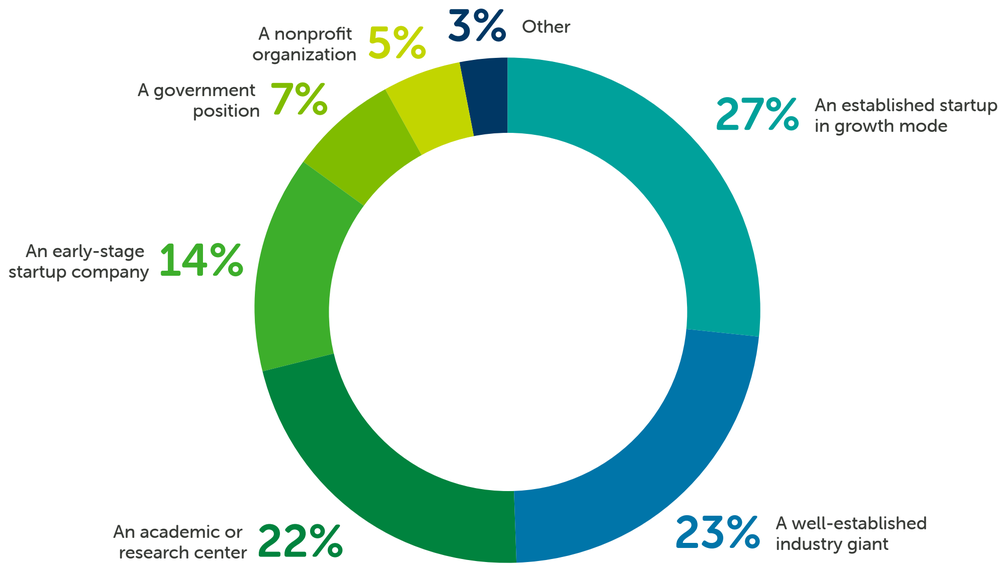
Training current employees is the most prevalent step (44.05%) that organizations are taking to address the current workforce talent shortage, with permitting or expanding remote work options close behind (42.27%).



n = 1,403

The Future Workforce

Where Students Hope to Work

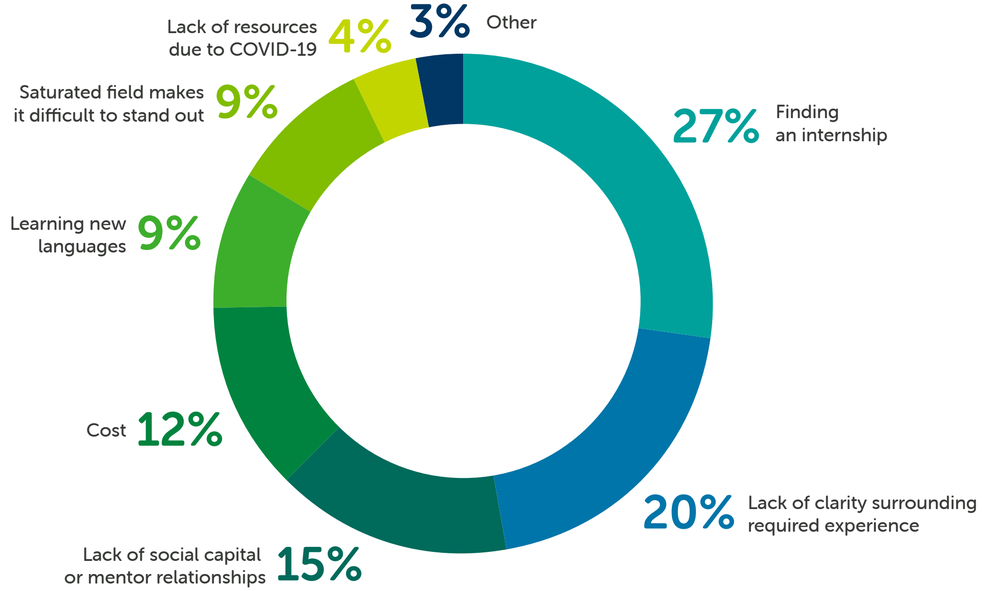


n = 407

26.78% of student respondents are hoping to work for an established startup in growth mode upon completing their programs. Tied for second choice are well-established industry giants (22.60%) and academic or research centers (22.11%).

The smallest percentages of respondents (6.88% and 4.91%, respectively) hope to work for a government organization or nonprofit.

Our student respondents shared their biggest obstacles to obtaining experience required for a career in data science or related fields.



n = 407

26.54% of student respondents cited finding an internship as the biggest obstacle to obtaining required experience, with the second-biggest obstacle being a lack of clarity surrounding required experience (19.90%).

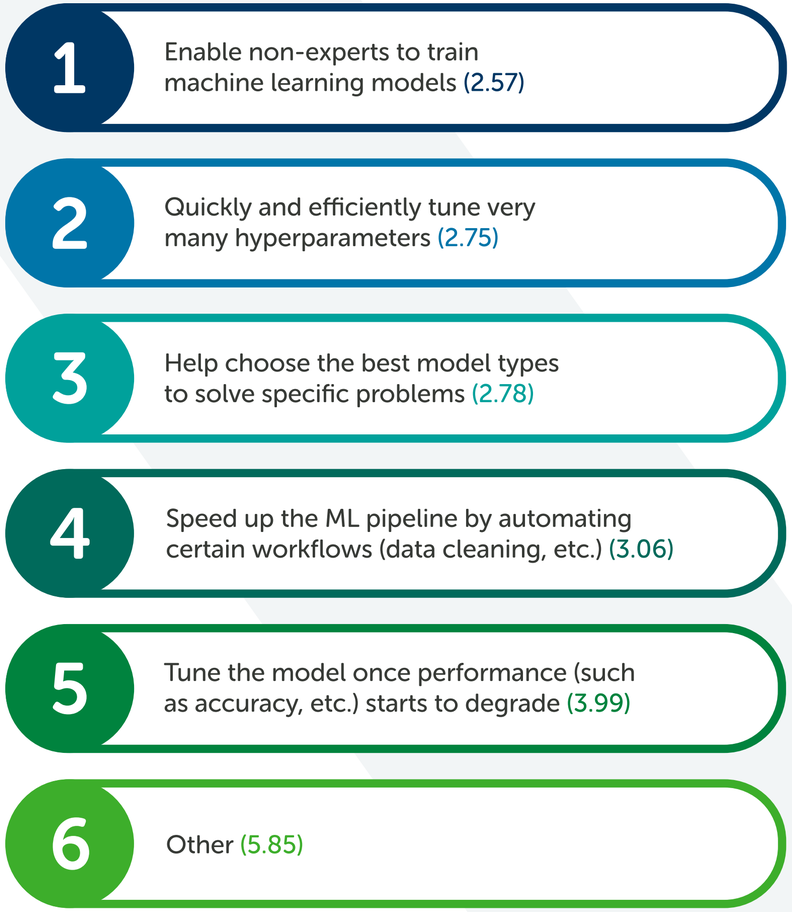
BIG QUESTIONS AND TRENDS

Data professionals are grappling with big questions that will likely impact how the industry evolves. With one eye on the future, we seek to unpack respondent sentiments surrounding trends, opportunities, and perceived blockers to growth.

AutoML

Last year, 55% of survey respondents indicated that they hoped to see more automation and AutoML in data science. This year, we wanted to dig into what respondents think an AutoML tool should do for data scientists.

On average, enabling non-experts to train machine learning models is the #1 task non-student respondents think an AutoML tool should do for data scientists.

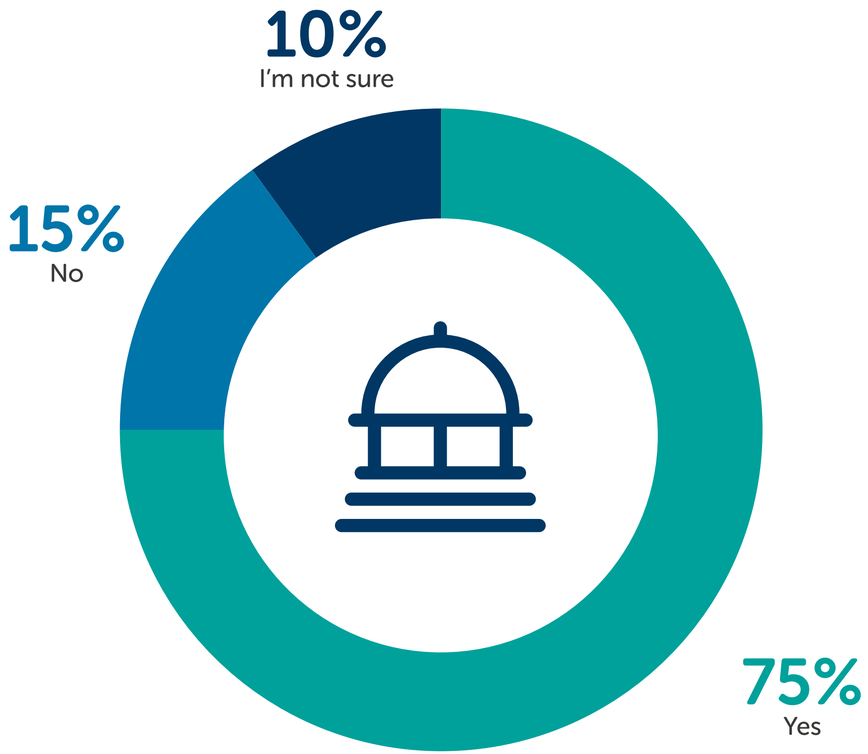


We asked respondents to drag and rank the options from most to least important, with the first being most important.

n = 2,042

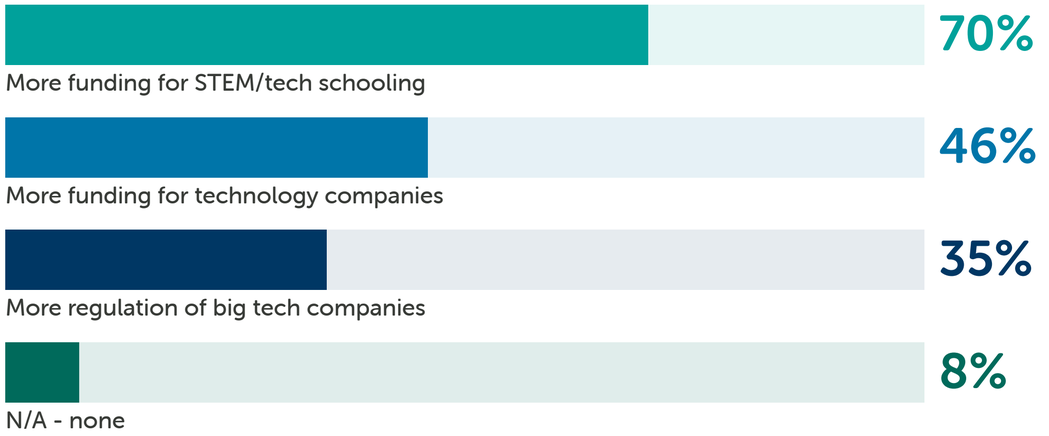
Government Involvement in Tech

Do you believe the government should play a larger role in strengthening technological innovation and manufacturing in your country?



n = 1,443

Which of the following governmental incentives do you support?

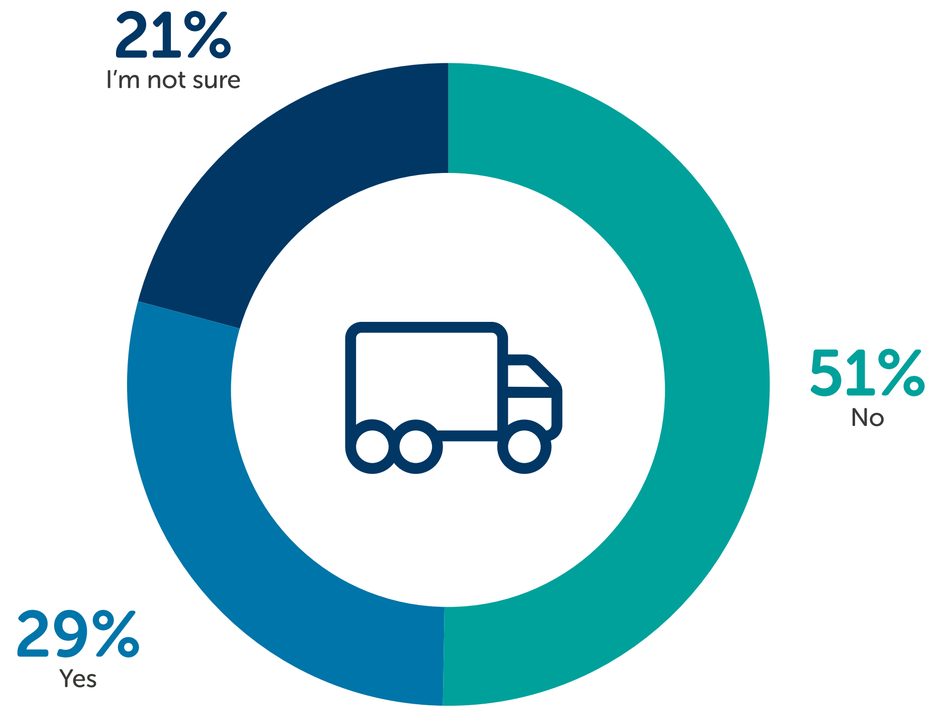


n = 1,337

A vast majority (74.98%) of respondents think the government should play a larger role in strengthening technological innovation and manufacturing. **More funding for STEM/tech schooling** is the incentive that the most respondents (69.78% of the respondents who comprise the aforementioned majority) support.

Challenges to Innovation

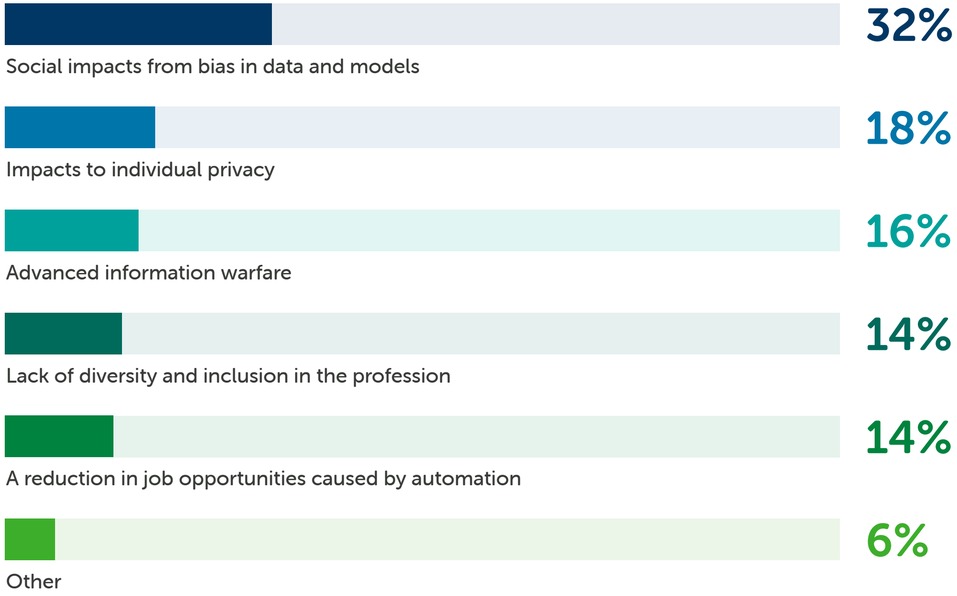
Have supply chain disruption problems such as the ongoing chip shortage impacted your access to computing resources?



n = 2,133

Fortunately, most (50.59% of) respondents indicated that supply chain disruption problems such as the ongoing chip shortage have not impacted their access to computing resources.

What do you think is the biggest problem in the data science/AI/ML space today?



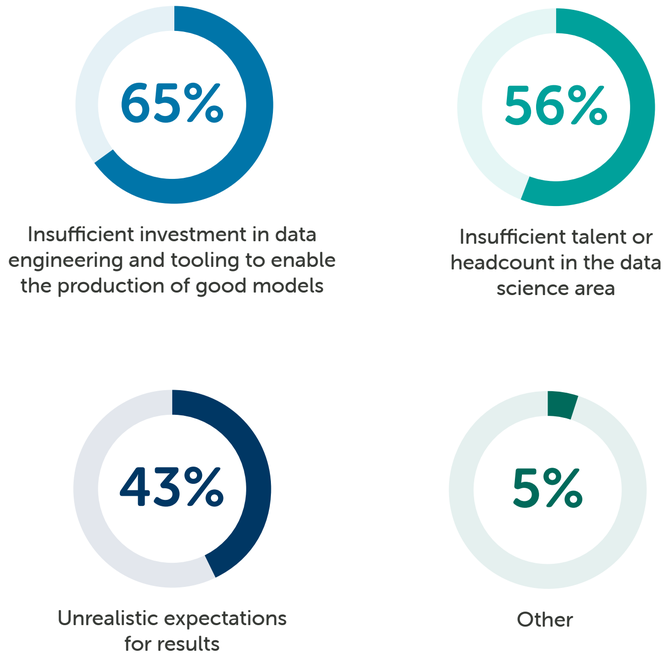
n = 2,154

31.94% of respondents feel that **social impacts from bias in data and models** is the biggest problem in the data science/AI/ML space today.

This finding is particularly interesting when examined alongside the fact that when it comes to the aforementioned talent shortage, only 34.65% of commercial respondents view diversity and inclusion hiring objectives as a top concern. Diversity and inclusion staffing practices—or the lack thereof—can have a direct impact on fairness and bias in data and models.

Challenges to Innovation

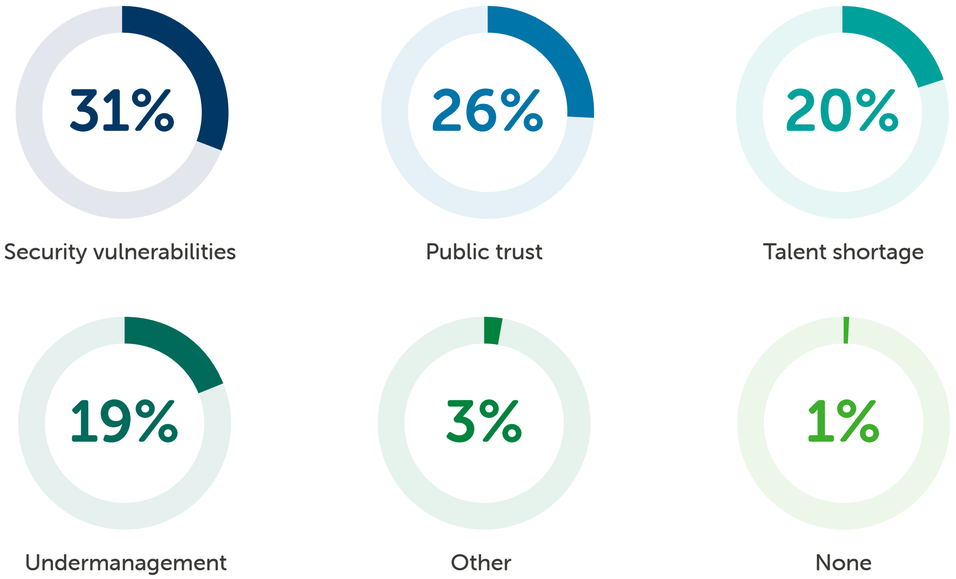
What do you think are the biggest barriers to the successful enterprise adoption of data science?



n = 1,428

Most (65.34% of) commercial respondents think one of the biggest barriers to the successful enterprise adoption of data science is insufficient investment in data engineering and tooling to enable the production of good models.

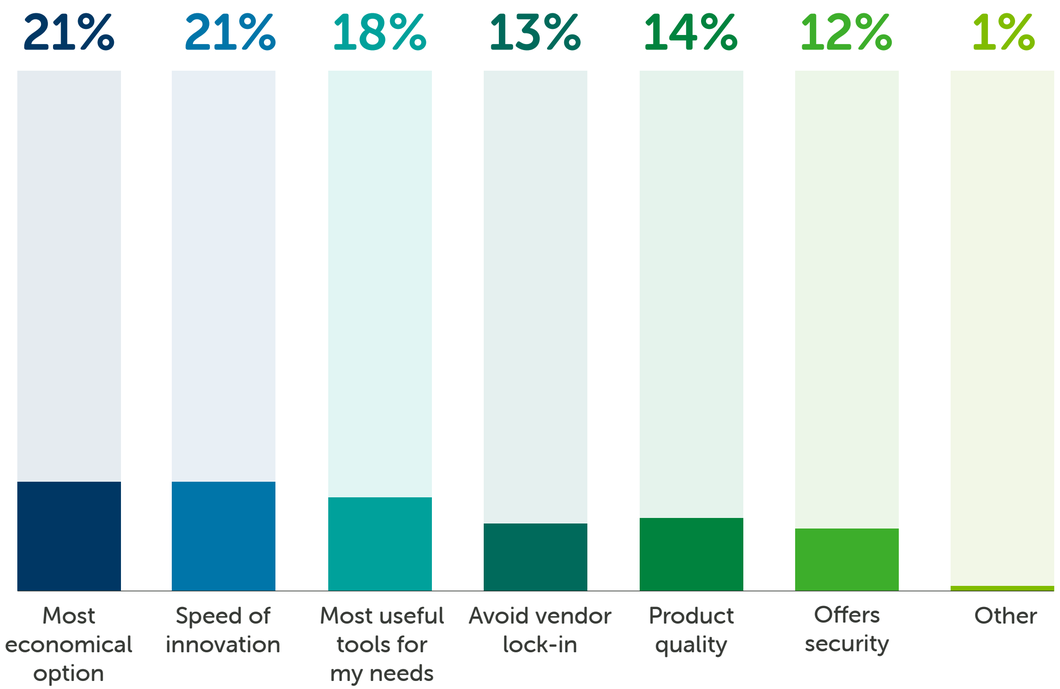
What do you believe is the biggest challenge in the open-source community today?



n = 2,133

Most respondents (31.08%) think the biggest challenge in the open-source community today comes down to security vulnerabilities, which may very well be linked to the issue of public trust (which 25.74% think is the biggest challenge).

What do you value most about open-source technology?



n = 2,154

In spite of any challenges affiliated with open source, survey respondents recognize an array of OSS benefits, with affordability (20.84%) and speed of innovation (20.54%) just about tying for most valued.

KEY TAKEAWAYS AND REFLECTIONS

Past, Present, and Future

1 IN PROFESSIONAL ENVIRONMENTS, OPEN-SOURCE SECURITY IS TOP OF MIND.

40.23% of commercial-track respondents indicated that their organizations scaled back their open-source software usage in the past year due to concerns around security, and most respondents (31.08%) selected "security vulnerabilities" as the biggest challenge in the open-source community today. As noted, organizations are leveraging a variety of measures and tools in their efforts to ensure their open-source supply chains and packages are secure and meet enterprise security standards.

It's no surprise that open-source security is so top of mind, given some of the incidents that have troubled the industry over the last year. There was the aforementioned Log4j incident, followed by the rise of protestware, spurred by economic sanctions the U.S. imposed on Russia in response to the conflict in Ukraine. In March, President Biden made a statement urging organizations to "strengthen the cybersecurity and resilience of the critical services and technologies on which Americans rely." Shortly thereafter, we at Anaconda informed our community of the efforts we've made—and continue to make—to protect commercial users from cybersecurity threats.

The issue of open-source security is likely to remain at the forefront of the data science, AI, and ML industries for the foreseeable future. While 34.80% of academic-track respondents indicated that topics related to open-source security are taught frequently (in multiple classes and lectures) or often (but only during a specific class), 25.75% indicated that such topics are taught only sometimes (when there is a specific lecture), and 23.34% indicated open-source security hasn't come up in any discussions or lectures. Of our student-track respondents, 25.55% indicated that topics related to open-source security are taught frequently (in multiple classes and lectures) or often (but only during a specific class), and 32.92% indicated open-source security hasn't come up in any discussions or lectures. This data points to a discrepancy between the weight of open-source security in the commercial sector and the weight the issue is given in academic environments.

40%

of commercial-track respondents indicated that their organizations scaled back their open-source software usage in the past year due to concerns around security, and most respondents (31.08%) selected "security vulnerabilities" as the biggest challenge in the open-source community today.

2 ORGANIZATIONS WILL CONTINUE TO GRAPPLE WITH THE TALENT DILEMMA.

Over the last couple of years, organizations attempting to scale their data science efforts and accelerate technology advancements and AI/ML adoption have increasingly noticed, if not suffered from, a lack of data science talent. As stated, 62.51% of commercial-track survey respondents indicated that their organizations are at least moderately concerned about the potential impact of a talent shortage, with the biggest concerns surrounding the ability to recruit/retain technical talent and more general recruiting and talent acquisition. 56.09% of commercial-track respondents feel that insufficient talent or headcount in the data science area is one of the biggest barriers to the successful enterprise adoption of data science.

As the data science, AI, and ML presence continues to grow in businesses across industries, organizations need to get creative when it comes to attracting and retaining talent and optimizing their workforces. Training existing employees and permitting or expanding remote work options are the most popular steps that organizations are taking. Organizations will want to monitor both skills gaps and the tools and resources available for continued learning as they continue to upskill their workforces, and academic institutions and students will want to make a note of these gaps and opportunities as they prepare and become jobseekers.

63%

of commercial-track survey respondents indicated that their organizations are at least moderately concerned about the potential impact of a talent shortage.

3 ETHICS, BIAS, AND REGULATION NEED MORE ATTENTION.

In our 2021 report, we proposed that "preventing bias and developing ethical data science [was] critical," and this year's data upholds that truth. As stated above, 31.94% of survey respondents feel that social impacts from bias in data and models is (still) the biggest problem in the data science/AI/ML space today.

To reiterate: While many commercial-track respondents indicated that their organizations are taking measures or using tools to ensure fairness and mitigate bias in data sets and models (most commonly, evaluating data collection methods according to internally-set standards and manually assessing data sets for fairness and bias), 23.64% indicated that their organizations do not have standards surrounding/have not implemented measures or tools to address fairness and bias, and 14.89% aren't sure.

What's more, just 18.76% of academic-track respondents indicated that their institutions are teaching ethics in data science/ML to students, and just 20.15% of student respondents indicated that ethics in data science/ML is covered in their courses in preparation for entering the field.

Plus, only 23.14% of academic respondents and 21.38% of students said that bias in AI/ML/data science is taught frequently in classes or lectures, and 39.44% of academic respondents and 35.62% (-9.38% YoY) of students responded that it's rarely or never taught. While this latter percentage and the YoY decrease it reflects is a step in the right direction, there is certainly room for a greater emphasis on ethics and bias in the educational sphere.

In a 2022 predictions blog post, we touched on regulation and governance around ethical data and AI through government interventions and wide-scale standards. Here we are several months later, and essentially three out of four (74.98%) commercial-track survey respondents feel the government should play a larger role in strengthening technological innovation and manufacturing, with more funding for STEM tech schooling being the incentive that the most respondents (69.78% of the 74.98%) support. Perhaps said schooling should incorporate more topics surrounding ethics and bias in AI/ML/data science.

32%

of survey respondents feel that social impacts from bias in data and models is (still) the biggest problem in the data science/AI/ML space today.

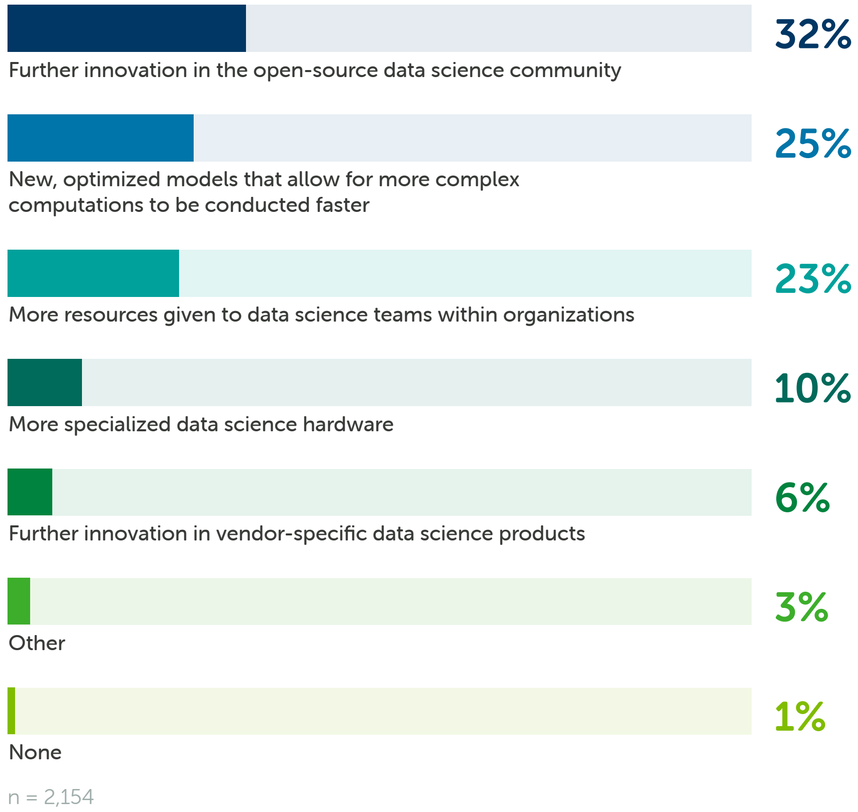
4 THE DATA SCIENCE, AI, AND ML COMMUNITIES ARE READY FOR MORE INNOVATION.

The #1 advancement that survey respondents are hoping to see from the data science industry in the coming year is further innovation in the open-source data science community (32.40%). New, optimized models that allow for more complex computations to be conducted faster (24.98%) and more resources given to data science teams within organizations (23.21%) come in second and third, respectively.

To help support the successful open-source innovation that the data science community wants to see, organizations will need to address the issues and challenges listed above, including open-source security, the talent dilemma, and ethics and bias. Educational institutions and students would do well to monitor these issues and adjust learning paths to reflect and prepare for the commercial open-source landscape. Organizations will also need to encourage teams to contribute to open-source projects; again, only 51.99% of commercial respondents said their teams are encouraged to do so—about a 13% decrease compared to 2021. Surely continued efforts to mitigate the aforementioned issues will make organizations progressively more comfortable with such encouragement.

Anaconda remains committed to identifying and tempering challenges that inhibit open-source innovation and providing products, features, and resources to power open-source advancements, streamline workflows, and instill confidence in OSS among all who rely on it. We continue to proudly distribute and contribute to a variety of open-source projects, and we continue to direct a portion of our revenue to the open-source community through the Anaconda Dividend Program.

What are you most hoping to see from the data science industry this year?



n = 2,154

2022 STATE OF DATA SCIENCE REPORT

If you enjoyed this report, we encourage you to:

- Keep an eye on our blog for more news and thought leadership content.
- Tune in to *Numerically Speaking: The Anaconda Podcast* on your favorite podcast app.
- Create a free Anaconda Nucleus account to access educational resources and connect with other members of the data science community.
- Engage with us at popular industry conferences and events.
- Follow us on social media!

ABOUT ANACONDA

With more than 30 million users, Anaconda is the world's most popular data science platform and the foundation of modern machine learning. We pioneered the use of Python for data science, champion its vibrant community, and continue to steward open-source projects that make tomorrow's innovations possible. Our enterprise-grade solutions enable corporate, research, and academic institutions around the world to harness the power of open-source for competitive advantage, groundbreaking research, and a better world.

Visit https://www.anaconda.com to learn more.

ANACONDA.