STATE OF DATA SCIENCE 2023

AI TAKES CENTER STAGE

ANACONDA.

Executive Summary

For the sixth consecutive year, Anaconda conducted our State of Data Science survey to surface insights about the demographics of the data science community, use cases across industries, and trends related to artificial intelligence (AI). This year's survey revealed insights about the topic on everyone's minds: generative AI and how it is changing the way data scientists work. The survey also addressed the use of open-source software (OSS), security methods and challenges, and how organizations are dealing with rapid innovation and upskilling. Our report highlights key insights related to these areas and what they could mean for the future of data science in a world increasingly driven by the predictive capabilities of machine learning.

As always, we're democratizing data and publicly sharing the complete raw data from our 2023 State of Data Science via GitHub so you can explore it for yourself.

Table of Contents

Methodology

2,414 individuals representing 126 countries took part in our online survey, which was conducted from June 16, 2023 to July 10, 2023. Respondents came from the Anaconda email database, Anaconda.com, social media, and other sources, and respondents were invited to participate in a sweepstakes drawing as an incentive for completing the survey. Five winners were selected at random after the survey was complete.

Respondents were divided into four different tracks: IT workers, data science practitioners, students, and researchers or university professors. All respondents were asked the same demographic questions, but some questions were unique to each track. In the report, we indicate whether responses come from the entirety of the respondents or from a specific subset. All responses are self-reported.

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

Four Themes

We identified four themes from our survey, and our report goes into detail about these areas.

1

The **data science community** is a global and significantly diverse group of well-educated innovators and builders. Those reporting identification with the LGBTQIA+ community is higher than the U.S. national average.

2

Organizations and data professionals are actively pursuing the use of **generative AI**, particularly large language models (LLMs). Workers have concerns about AI and job loss.

3

Security and alignment with IT and security teams remain ongoing challenges. The majority use open-source software. IT workers are not confident in their abilities to identify and remediate open-source vulnerabilities. Few professors discuss security in courses.

4

Data practitioners are confident they are delivering measurable **business value**, mostly by saving time and reducing costs. Data prep and cleaning continue to be challenges. Organizations are hiring for AI roles.

The Data Science Community

As we have in previous years, we began our survey with demographic questions to determine the diversity in geographical locations, age, experience, job functions, and identity. We're eager to track and understand the continuously evolving ecosystem of the data science community. We will continue to promote international inclusion and strive to represent diverse backgrounds, experiences, and locations.

2,414 individuals representing 126 countries took part in our 2023 survey. The majority of respondents come from the United States, with India being the next most reported country for respondents. While not every respondent chose to disclose their race identification, the majority (58%) of those who did share identify as white, indicating a need to continue efforts to diversify the talent pool in data science.

Our respondent set skews toward younger generations. Half are millennials. Only 12% of respondents are 58 or older. This is in line with tech careers generally leaning younger and with millennials now being in their 30s.

The gender identity data from our respondents supports what the community continues to see across male-dominated STEM fields. We see some improvement in gender diversity, as 23% of respondents in the 2022 State of Data Science survey identified as female, compared with 29% of respondents in 2023.

For the first time, we asked respondents about their sexual orientation and identity. While the United States average for LGBTQIA+ individuals hovers around 7%, o**ne-third (33%) of global survey respondents identify as part of the queer community**. This larger-than-standard margin is a sign of the diversity within the data science community.

We divided the survey into four tracks according to job function. **The majority of respondents work as a data science practitioner or in a related job, have a college-level degree, and come from commercial organizations**. This year, a higher percentage of respondents are from government agencies (15%) and non-profit organizations (14%) than we saw in 2022, when there were 11% in each of those organization types.

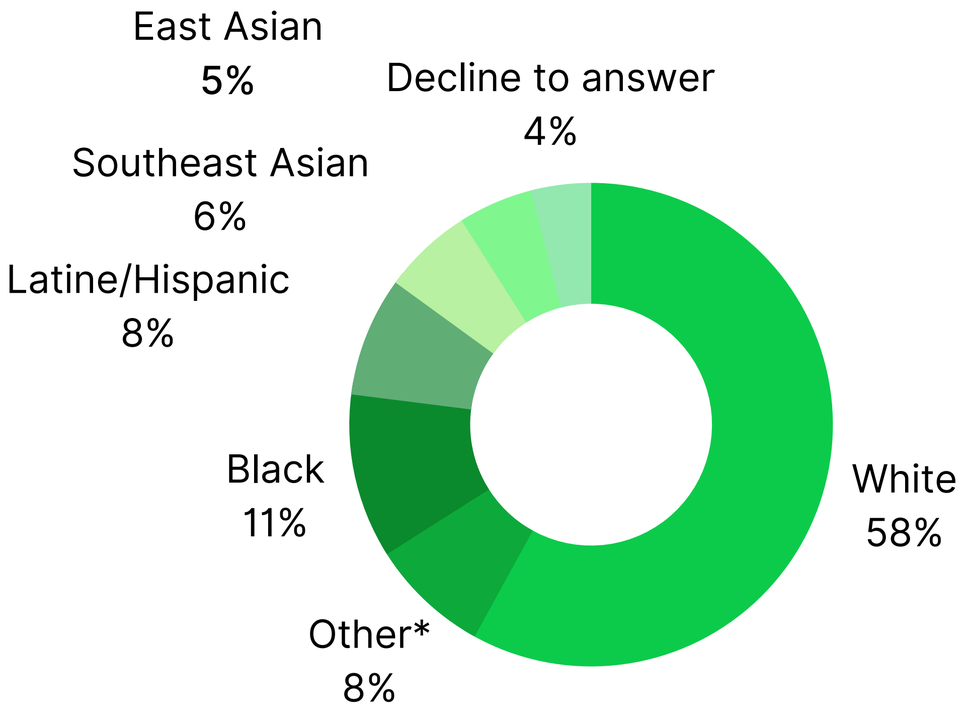
Respondent Location

North America - **49%**
South Asia - **9%**
Latin America - **8%**
Sub-Saharan Africa - **6%**

East Asia/Pacific - **4%**
Australia/New Zealand - **3%**
Middle East/North Africa - **3%**
EU/Central Asia - **18%**

n=2,414

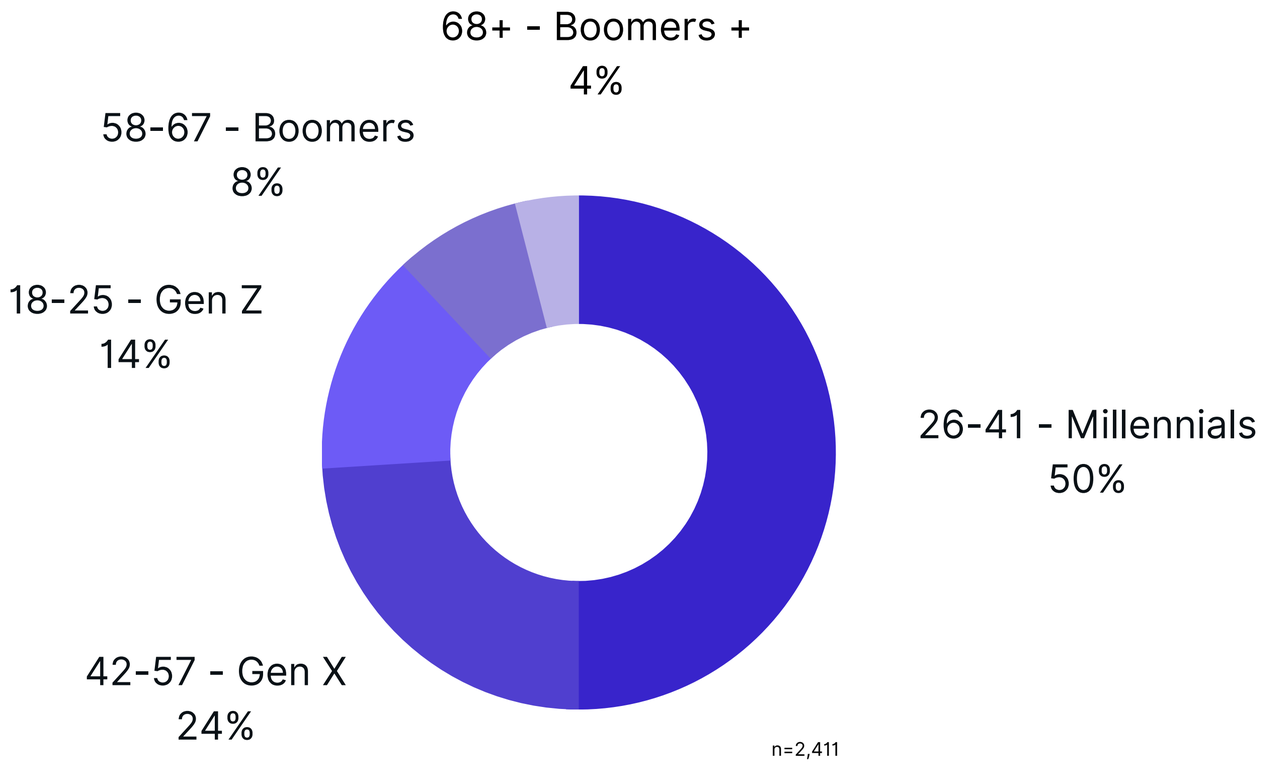
Respondent Race



n=2,133

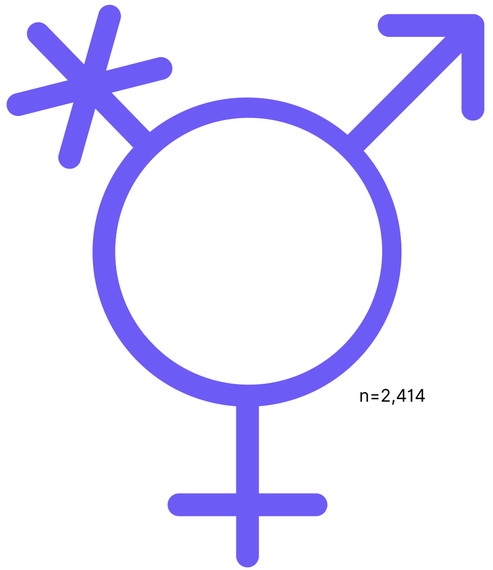
*including mixed race and indigenous

Respondent Age

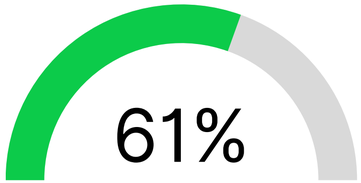


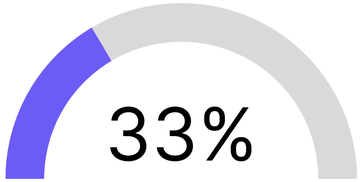
n=2,411

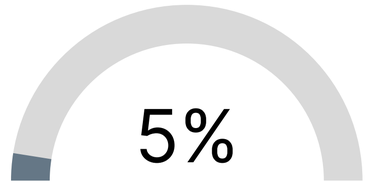
Respondent Gender



Respondent Sexual Orientation



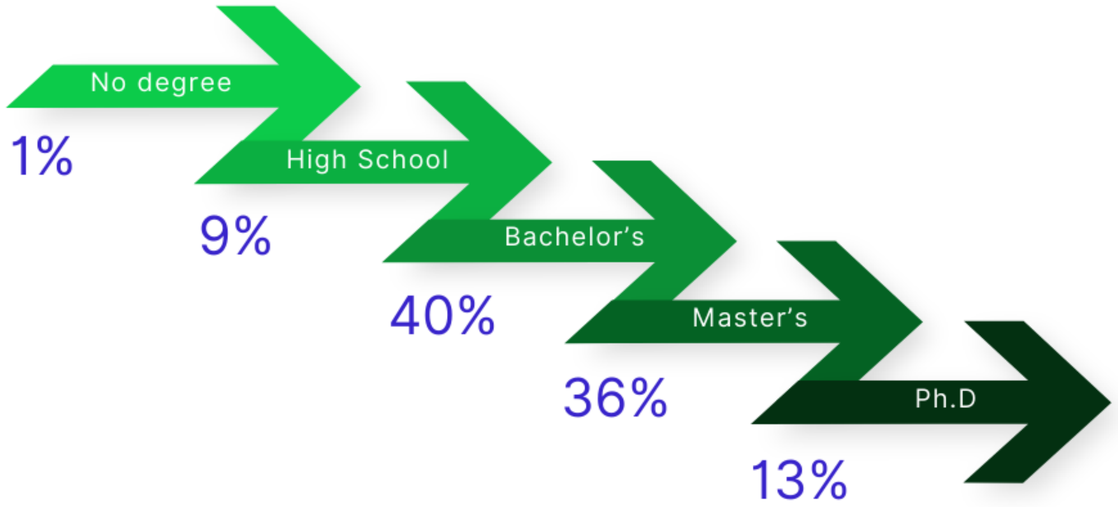




*LGBTQIA+ stands for lesbian, gay, bisexual, trans, queer or questioning, intersex, asexual, and all other queer identities

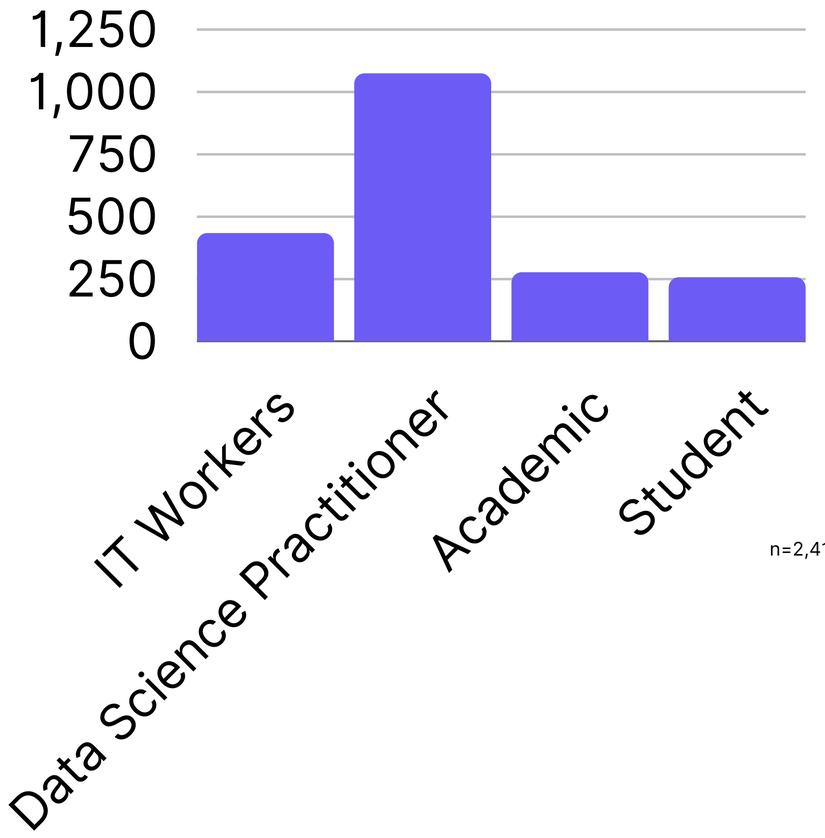
n=2,414

Respondent Education



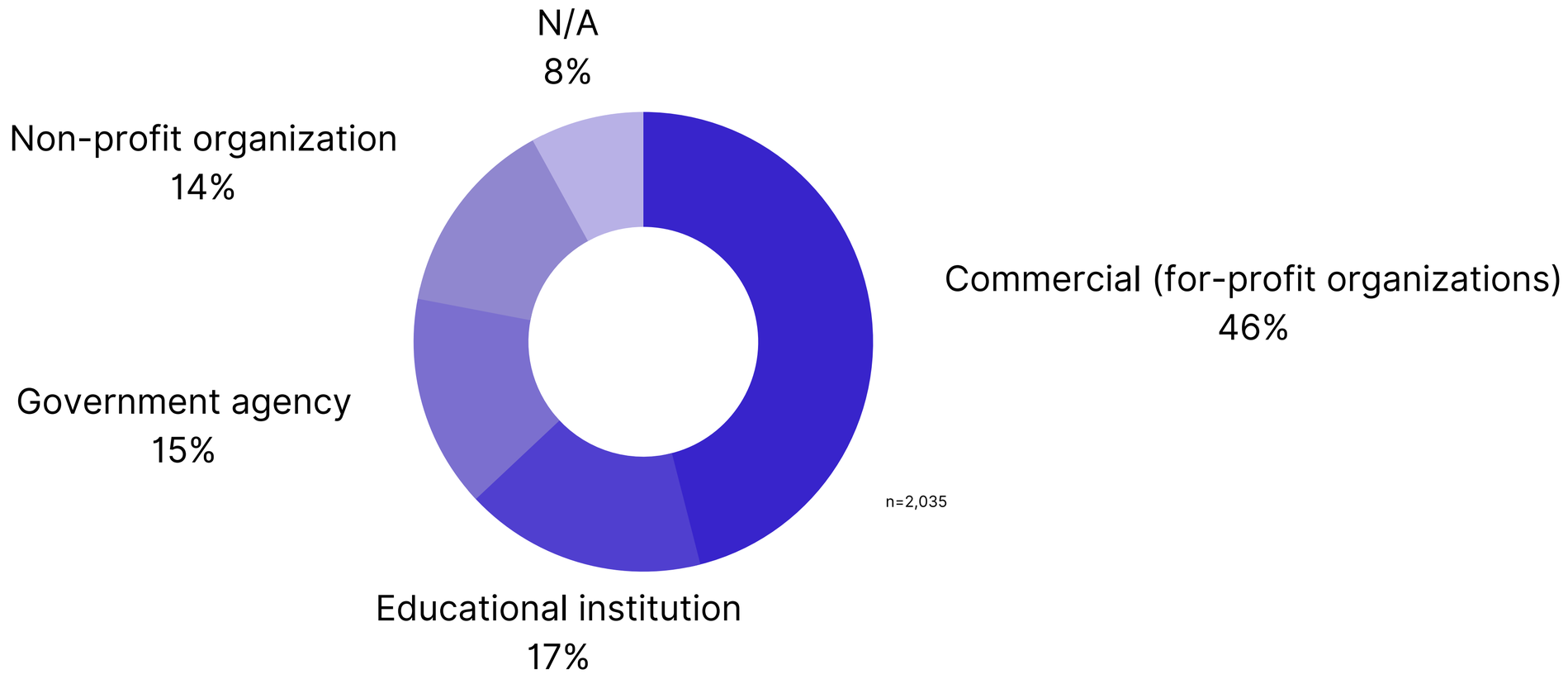
n=2,106

Primary Job Function



n=2,414

Organization Type



Generative Artificial Intelligence (AI)

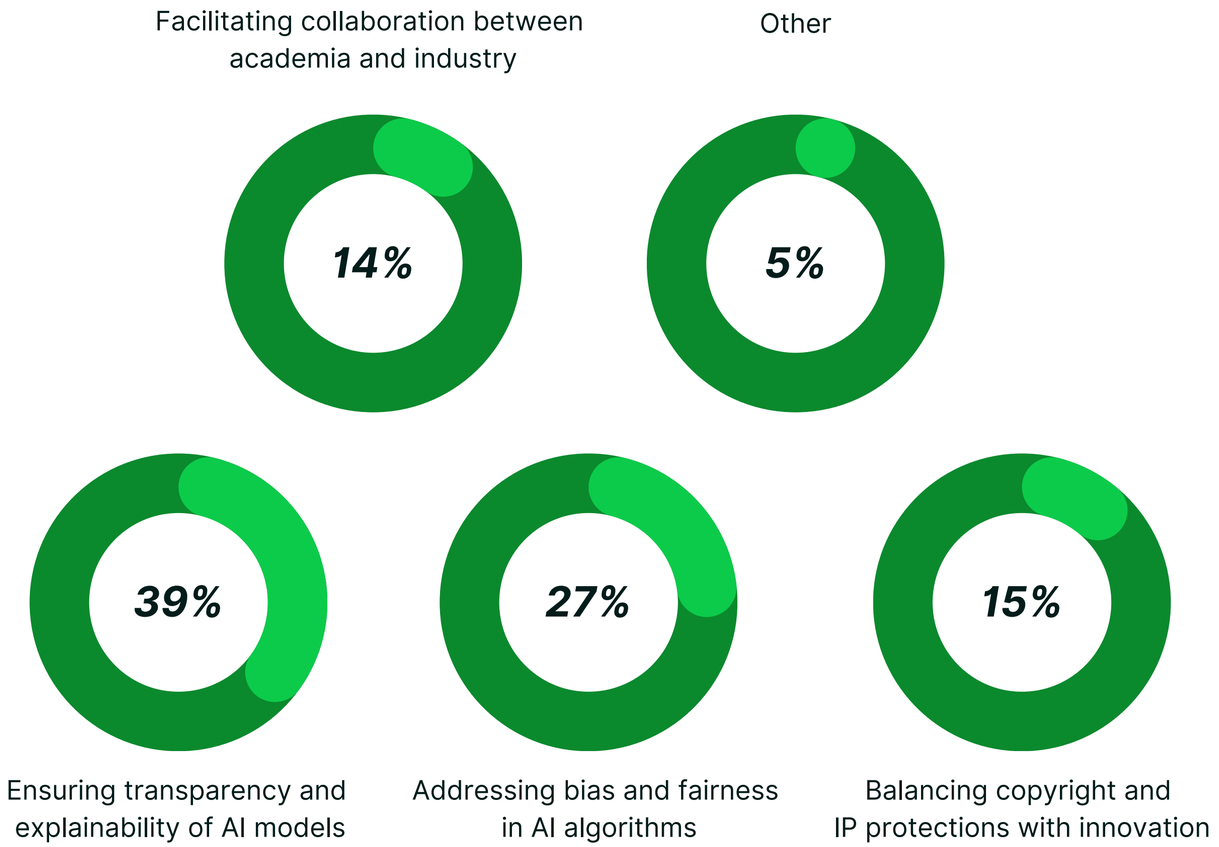
Generative AI has been in the news for some time now, with the introduction of high-performing large language models (LLMs). In our data science practitioner track, **40% of respondents say their companies are working on internal generative AI tools, such as LLMs.** While there are many conversations about the ethics and sustainability of generative AI, we focused our questions on how generative AI is being used and its relationship to the talent landscape, emerging job titles, and concern around job loss.

The majority (63%) of data science practitioners say they're using generative AI the same amount or more this year compared to 2022. Respondents who report using these tools and techniques in their work most commonly use them for content creation (e.g., text or image generation) and data cleaning, visualization, and analysis. Less common use cases include automating tasks and writing and debugging code.

We asked each survey participant to highlight their biggest concern when it comes to the use of generative AI. **The majority (39%) cite the need for transparency and explainability in AI models.** Bias and fairness in AI algorithms (27%) also are cited, along with facilitating collaboration between academia and industry (14%), and balancing copyright and IP protections with innovation (15%).

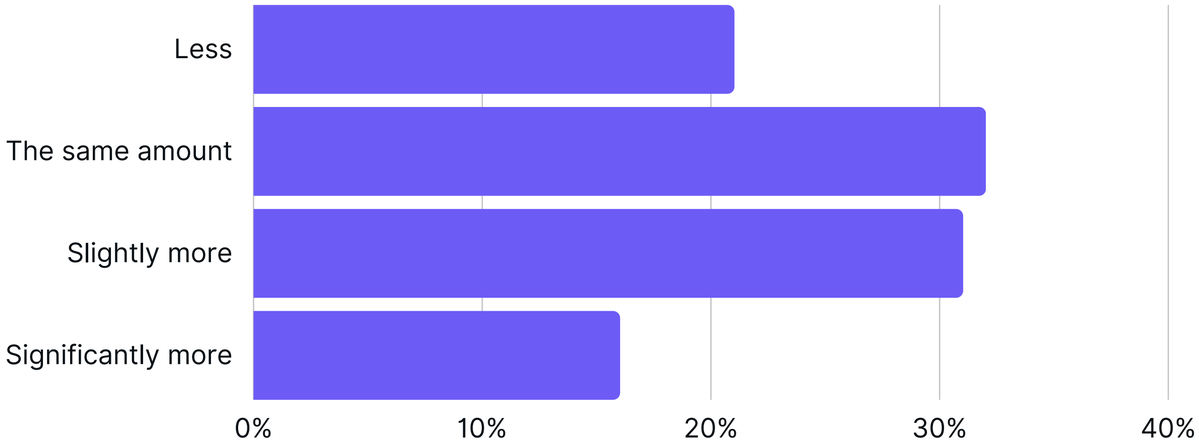
Between our IT and data science practitioner tracks, about **half of respondents report worrying about losing their job to generative AI.** The majority (65%) of IT workers and nearly half (45%) of data science practitioners are concerned about job security. The majority of organizations are providing upskilling pathways for learning new AI tools and technologies: a clear signal that organizations want to help their employees learn so they can retain and grow top talent.

What is your biggest concern regarding regulations for generative AI?



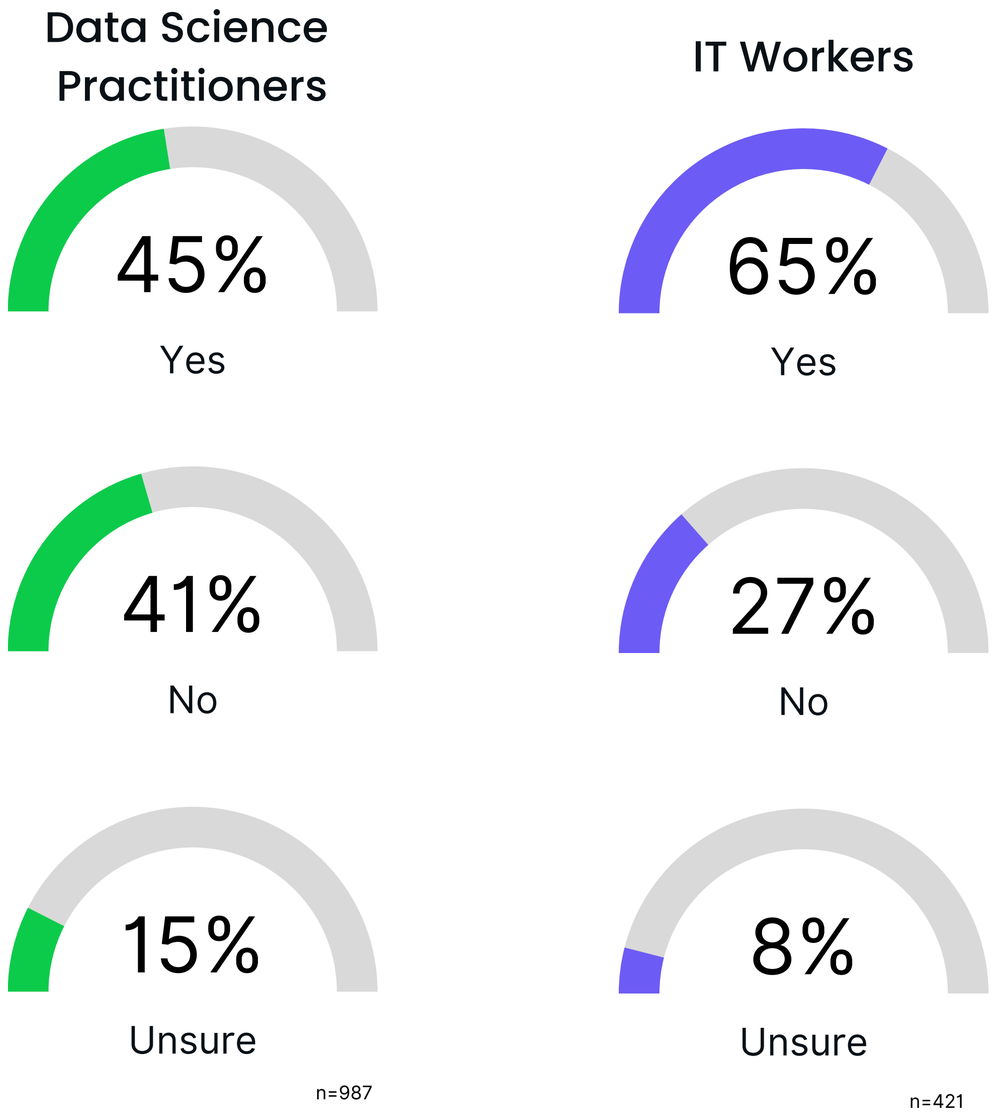
n=1,904

In your work, how much more time have you spent this year on generative AI techniques such as generative adversarial networks (GANs), variational autoencoders (VAEs), or transformer models compared to last year?



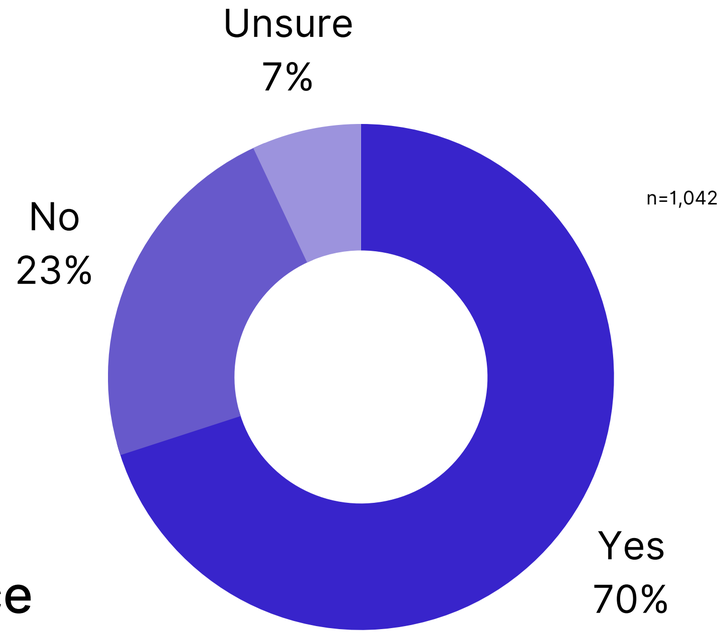
n=1,056

Do you feel your job is threatened by the rise of generative AI tools?

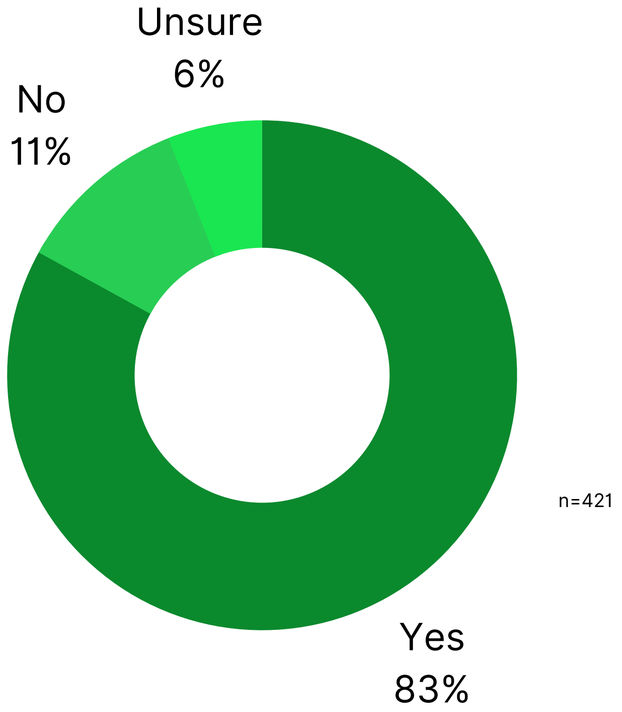


Does your company have set upskilling pathways to support employees learning new AI technologies on the job?

Data Science Practitioner



IT Workers



Open-Source Software and Security

The majority of data science practitioners and IT workers use open-source software in their workflows and at their organizations. For data science and machine learning, open-source software refers to software repositories, packages, libraries, and platforms that are freely available to the public, allowing users to access, use, modify, and distribute the source code.

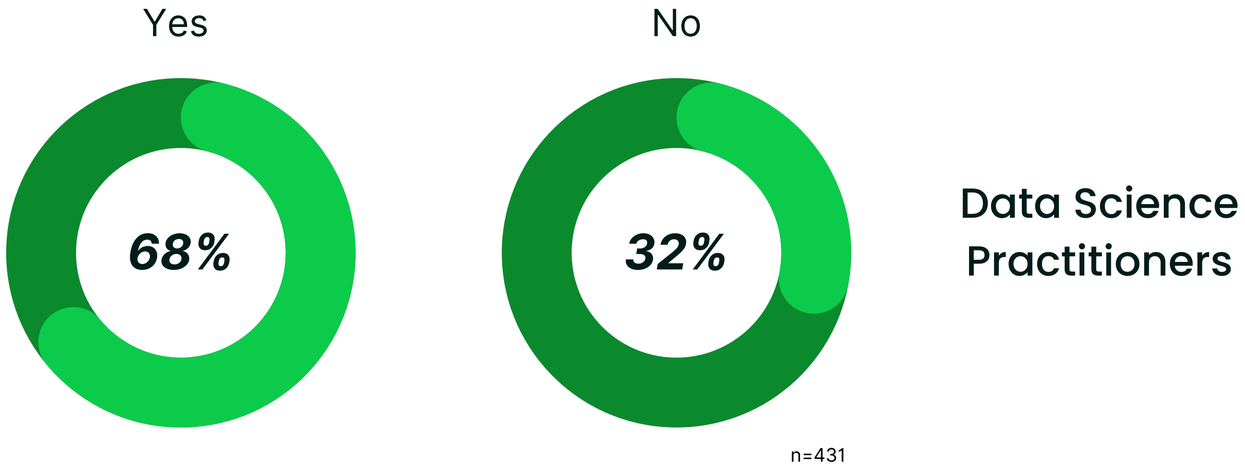
While many organizations use OSS, the protocols for ensuring security can be laborious and time-consuming. Many IT workers cite manual checks as the go-to form of security, and of those, 80% report they or their teams spend 25-50% of their time on these checks. **Only 18% of IT workers feel confident in their abilities to identify and remediate vulnerabilities associated with OSS.**

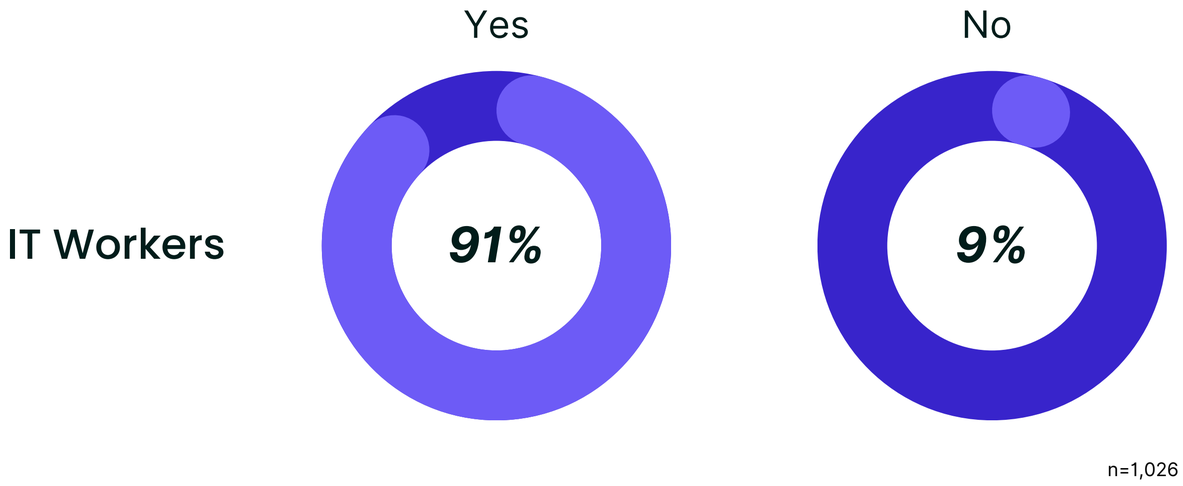
Cyberattacks are on the rise, and security and alignment with IT and security teams remain ongoing challenges in data science. Last year, about a third (34%) of 2022 State of Data Science respondents cited IT and information security (InfoSec) standards as the biggest impediment to moving commercial models into production.

Security awareness and action on vulnerabilities seems to be improving. In this year's survey, 54% of all respondents are aware of the National Institute of Standards and Technology's (NIST) National Vulnerability Database. **Of those working with a U.S.-based company, 50% report their organizations have strict security protocols.**

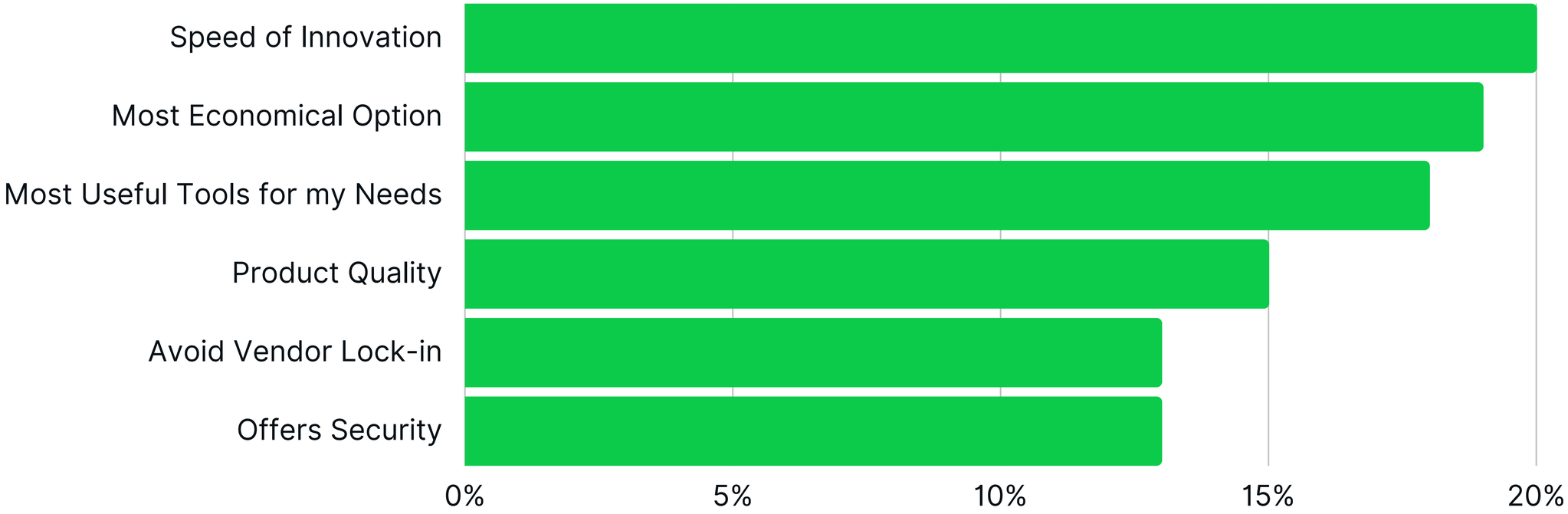
Few college and university professors (16%) identify OSS security as a frequent topic of discussion in class, with 34% saying they rarely or never discuss security. The discrepancy between employee security concerns and academic focus may indicate the need for expanded educational opportunities about OSS security.

Do you use open-source software in your current workflow?



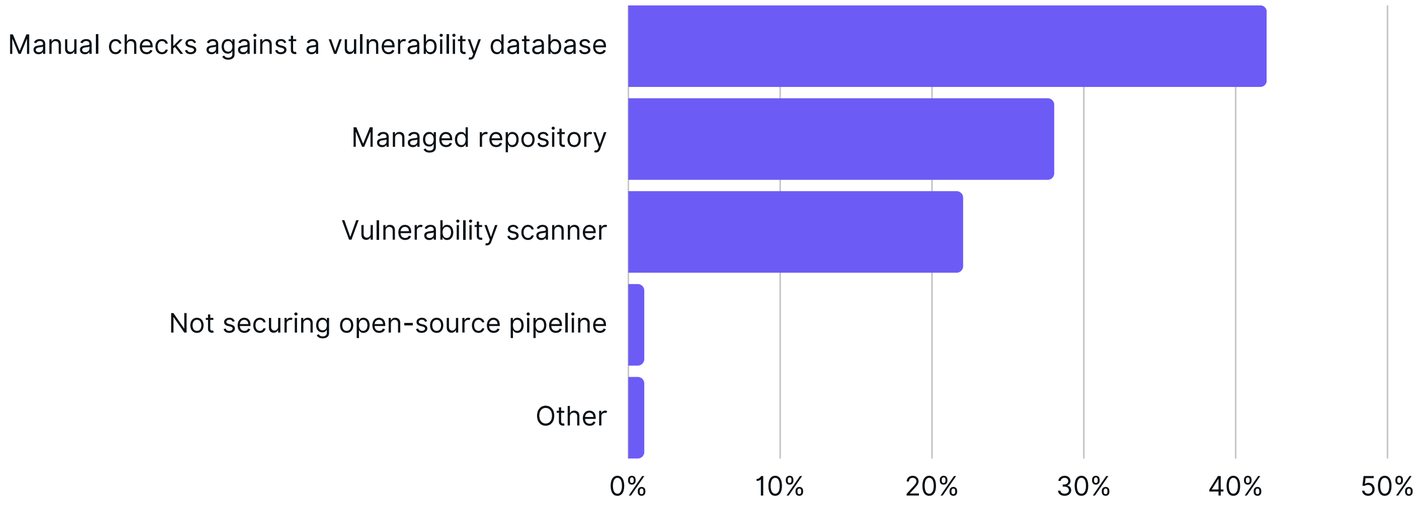


What do you value most about open-source software? (Respondents assigned a percentage to each feature)



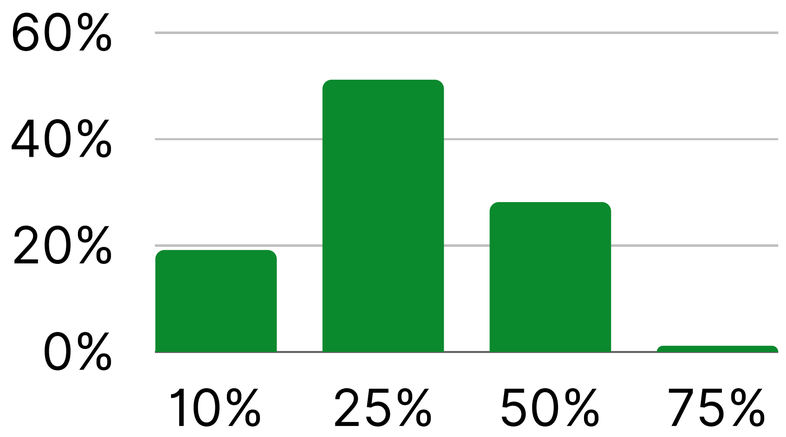
n=421

How does your company ensure the open-source packages used for data science and machine learning are secure and meet enterprise security standards?



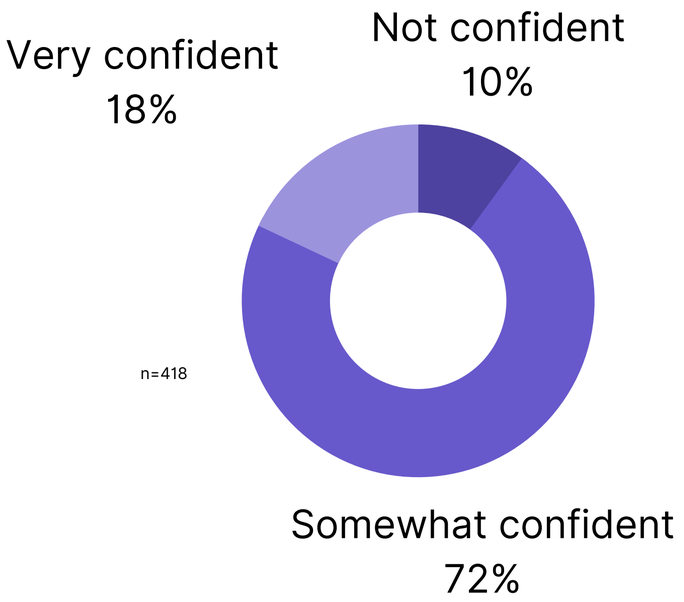
n=431

If you chose manual checks, how much of your time/your team's time is spent on these checks?

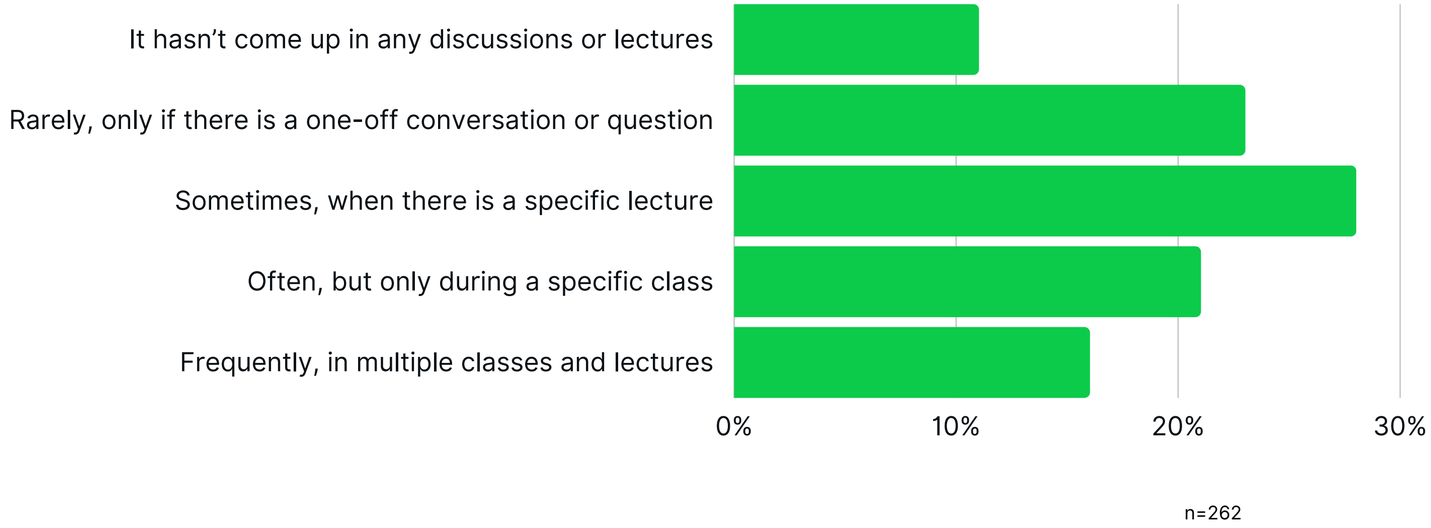


n=170

How confident are you in your ability to identify and remediate vulnerabilities associated with open-source software?



For instructors: How often are topics related to open-source security taught in classes or lectures?



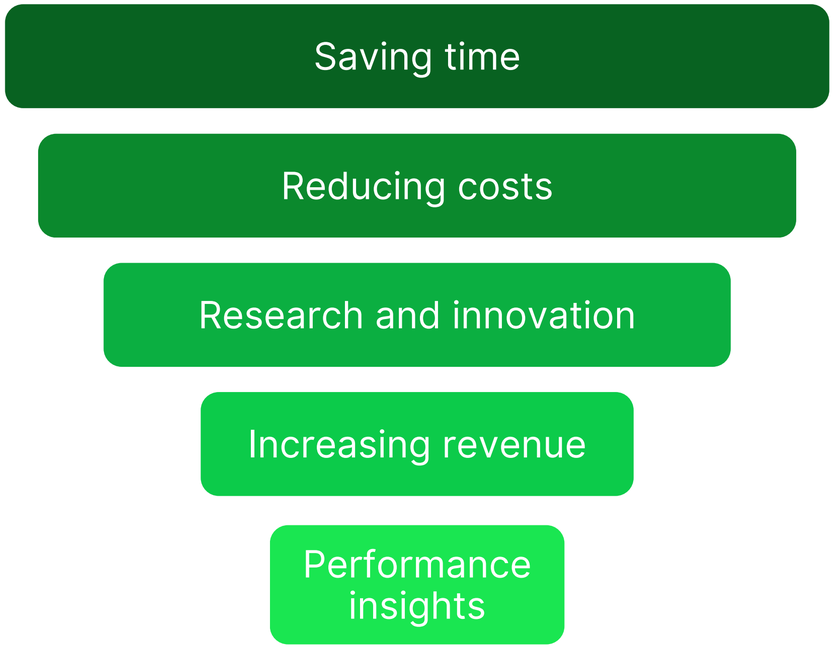
Business Value with Data Science

While artificial intelligence and machine learning persist as tactics to speed the generation of business insights, data science practitioners are answering the call with innovative technologies, in-house model production environments, and money-saving efforts. We asked our practitioner respondents to share how their **teams are impacting the bottom lines of their businesses** and to share more about their process for driving value for their organizations.

The most common ways that data science practitioners report helping deliver value are by saving time, reducing costs, and aiding research and innovation. **The majority (82%) of respondents are confident in their ability to deliver business value with data.** This strong majority may indicate that data scientists and practitioners have clear use cases and have been armed with the skills they need to handle a number of different tasks and challenges. This also may speak to practitioners' ability to explain business requirements, research, and model outcomes to executives.

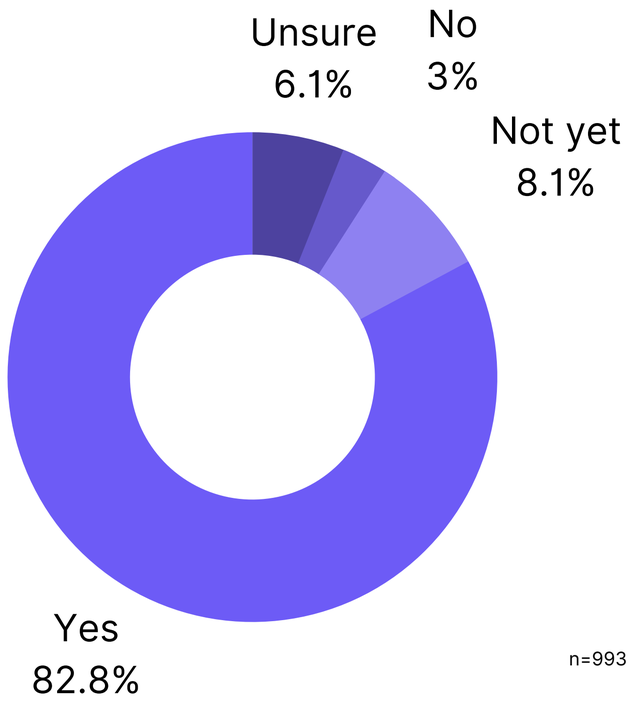
As the need for data-informed decision making grows, we've seen a **change in job titles and descriptions**. It seems that businesses want to use AI to improve operations or the products and services they offer to the market. A number of new jobs have joined traditional data analyst and ML engineer roles in hiring.

How do you and your team deliver value to your organization using data? (Responses ranked by frequency)



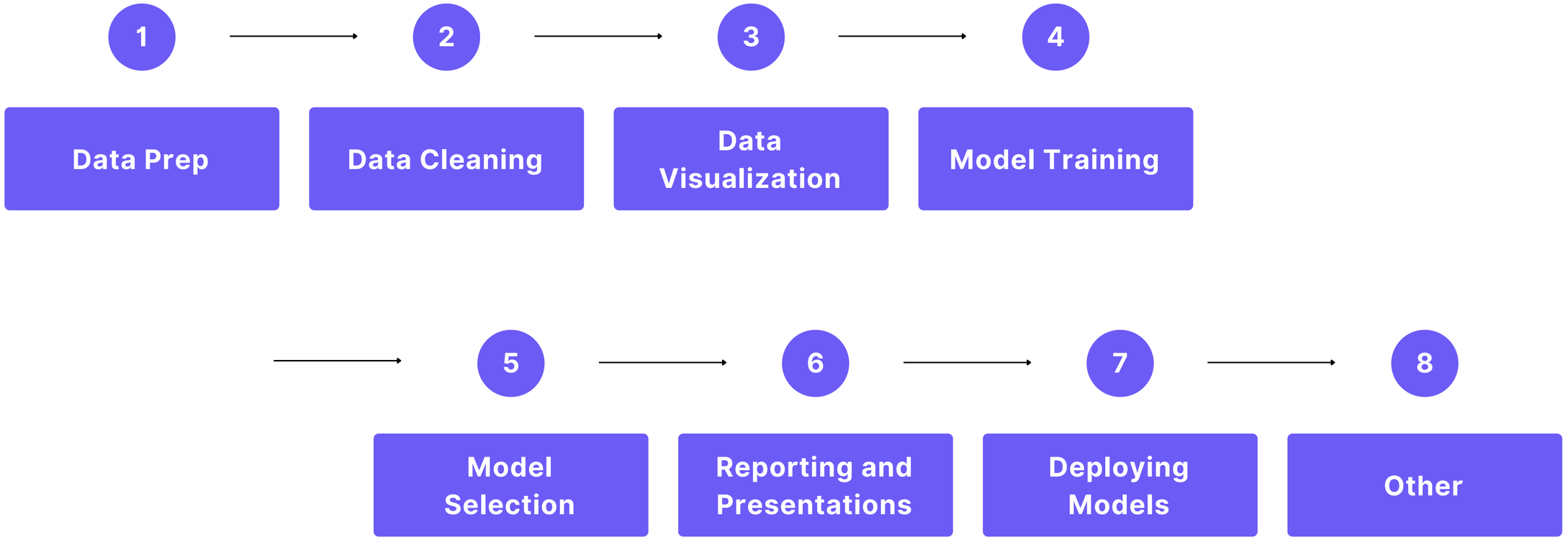
n=993

Do you feel these data initiatives are delivering measurable value?



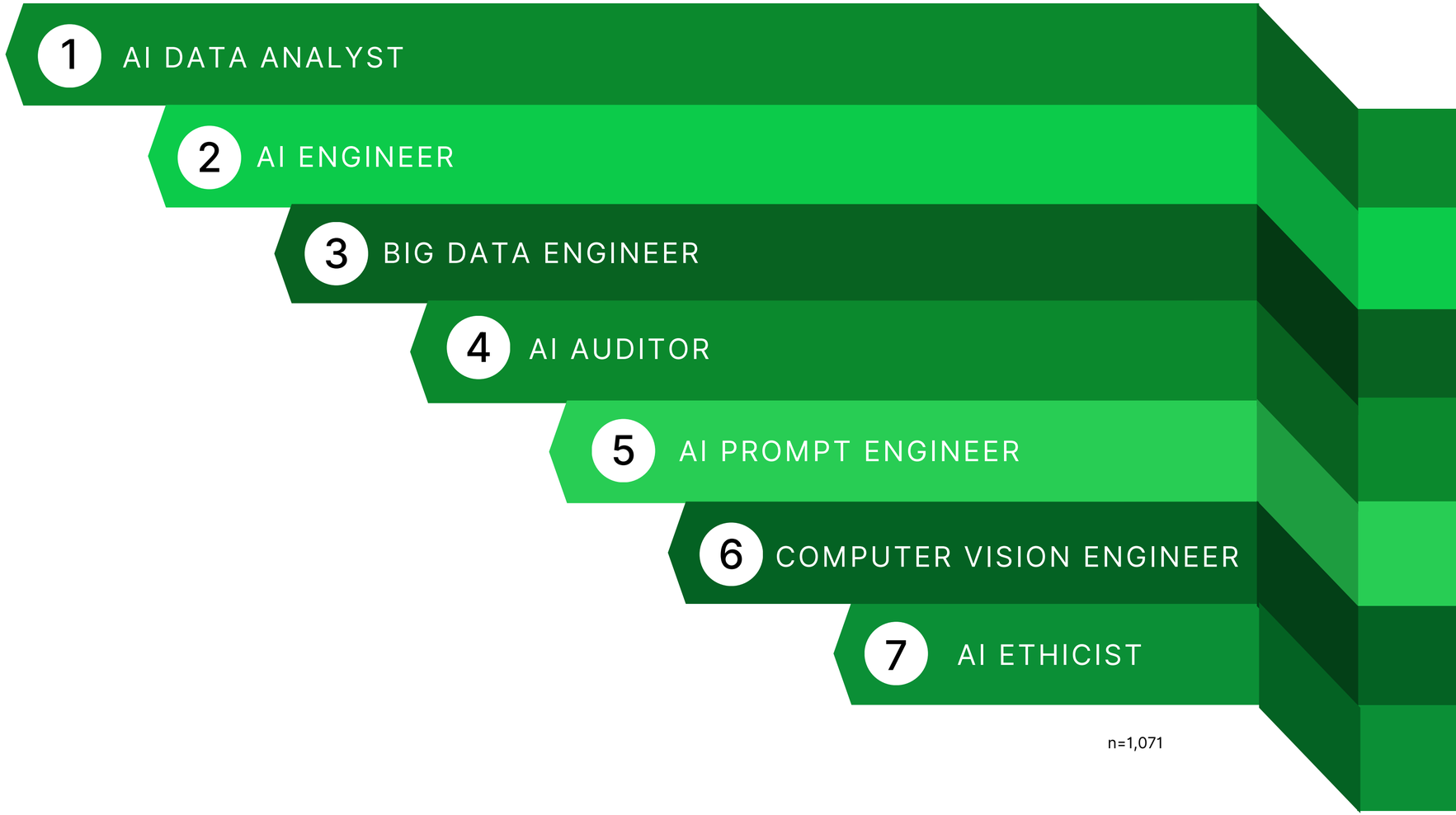
n=993

Thinking about your current role, what tasks are most time consuming? (Responses ranked from most to least time consuming)



n=1,071

Has your company already hired or plan on hiring any of the following roles? (Responses ranked by frequency)



n=1,071

Key Takeaways

The **data science community** is a global and significantly diverse group of well-educated innovators and builders. For the first time, we asked respondents whether they identified as LGBTQIA+, and the percentage of those reporting ‘yes’ is higher than the United States average. However, the majority of respondents in our survey are white men. As the community is making strides in diversity, we’ll need to continue our efforts to ensure our industry remains an open, diverse, and inclusive space for doing science with data and building machine learning solutions.

Generative AI is already changing how data scientists work. While a majority of workers worry about losing their jobs to AI, we’ve seen a balanced response of companies offering upskilling pathways with learning. This leads us to consider that while automation and some content creation may require fewer human resources, companies will look to those same employees to adapt, take on new challenges, and learn new tools to keep up with the frenetic pace of advancements in machine learning.

The majority of data practitioners and IT workers are using open-source software. With cyber-attacks on the rise, there is a growing need for **security** at all stages of the data science lifecycle. However, IT workers are not confident in their abilities to identify and remediate open-source vulnerabilities, and too few college and university professors are discussing security in their courses. Data practitioners and the security teams who support them should review and consider adoption of emerging security frameworks and governance programs.

Key Takeaways Continued

Data practitioners are confident in their abilities to drive measurable **business value** with data, and they report their work saves time and reduces costs. Practitioners continue to struggle with data preparation and cleaning. Increasing ethical concerns around generative AI have given rise to new roles in business and academic settings. Students continue to learn about bias, ethics, and responsible AI in traditional programs and in online courses.

At Anaconda, we will continue to champion data science, AI, and the practitioners and professionals who build, maintain, and use the open-source packages, libraries, and repositories that make insights, intelligence, and innovation possible. We are committed to providing centralized, secure access to thousands of Python and R repositories, packages, and libraries. We will continue to steward open-source projects that make it easier for you to innovate, build, and deploy effective solutions in your field.

About Anaconda

With more than 40 million users, Anaconda is the world's most popular data science platform and the foundation of modern AI development. 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.

To learn more, visit our website, create your free Anaconda account, attend an event, and follow us online.