2020 STATE OF DATA SCIENCE

Moving from hype toward maturity

We conducted our annual State of Data Science survey this year with a particular focus on understanding how data science as a discipline is maturing in commercial environments, and how academic institutions are preparing students to help lead the next generation of data scientists.

As data science continues its ascent to prominence and stability, the institutions which rely on it are still developing understanding of how to integrate, support, and leverage it. Even as enterprises rush to gain strategic advantage from data, investing heavily in technology, programs, people, and processes, there remain important challenges - challenges which will never be resolved, but will be areas for ongoing management and attention.

Data scientists are poised to take a unique leadership role in their organizations, as businesses and institutions navigate through challenges including value realization, integration of open-source tools, talent management, and bias mitigation. By helping to drive strategic transformation in these areas, data science practitioners and leaders can help ensure that the discipline achieves its full potential to improve society, consumer experience, and business performance.

ANACONDA®

TABLE OF CONTENTS


Methodology

2,360 people from more than 100 countries participated in our online survey, which was fielded from February 12 to April 20, 2020. Respondents were recruited via social media and the Anaconda email database, were offered the opportunity to participate in a sweepstakes drawing as an incentive to complete the survey, and four winners were selected at random after the survey closed. The respondents were divided into three groups: students, academics, and those working in commercial environments. Each of these cohorts was asked some unique questions, and some questions were presented universally. In this report, we have tried to indicate whether data reflects the responses of the entire set of respondents or a subset. Note: some totals will not sum to 100 due to rounding.

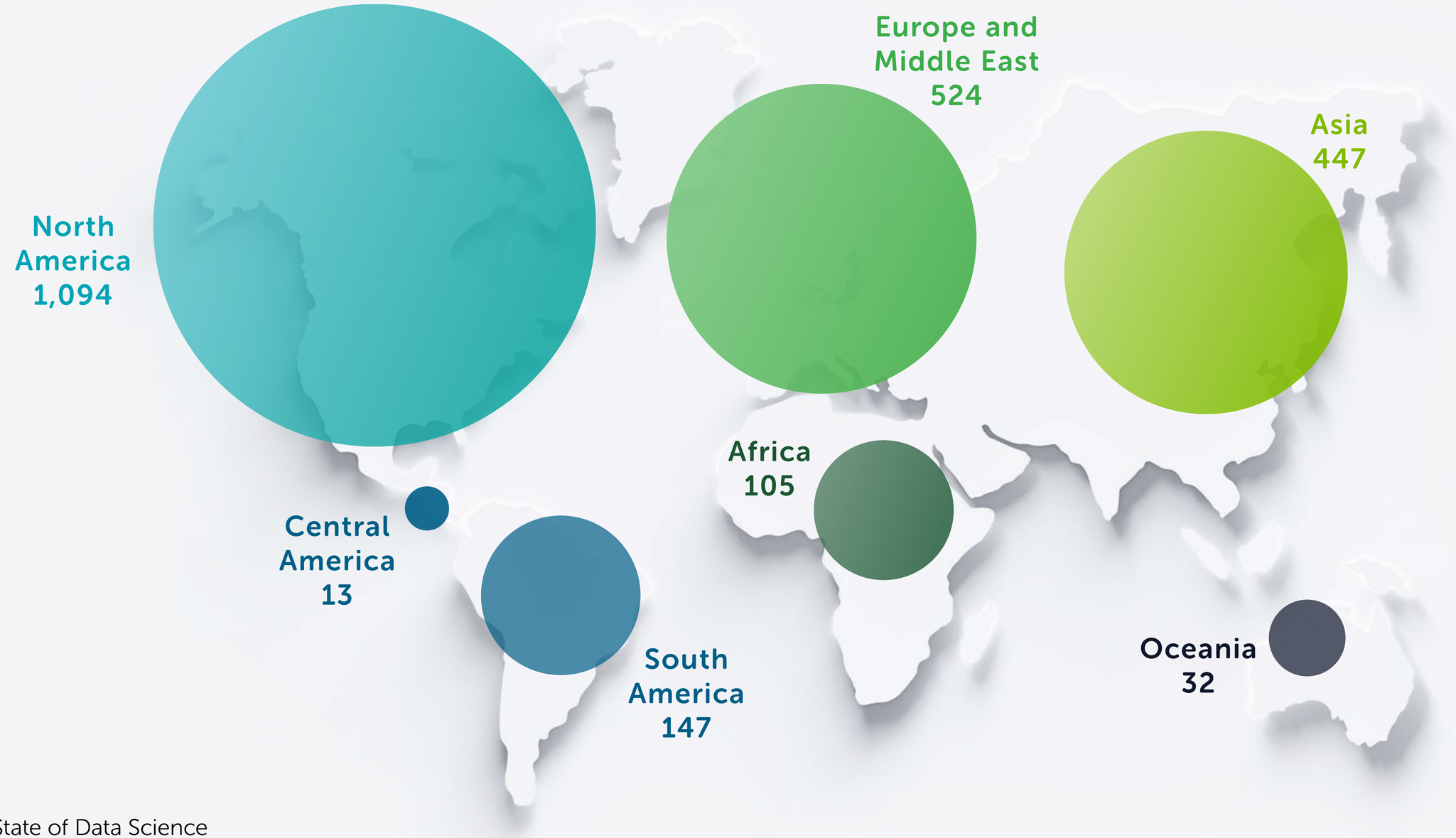



ANACONDA®

THE FACE OF DATA SCIENCE

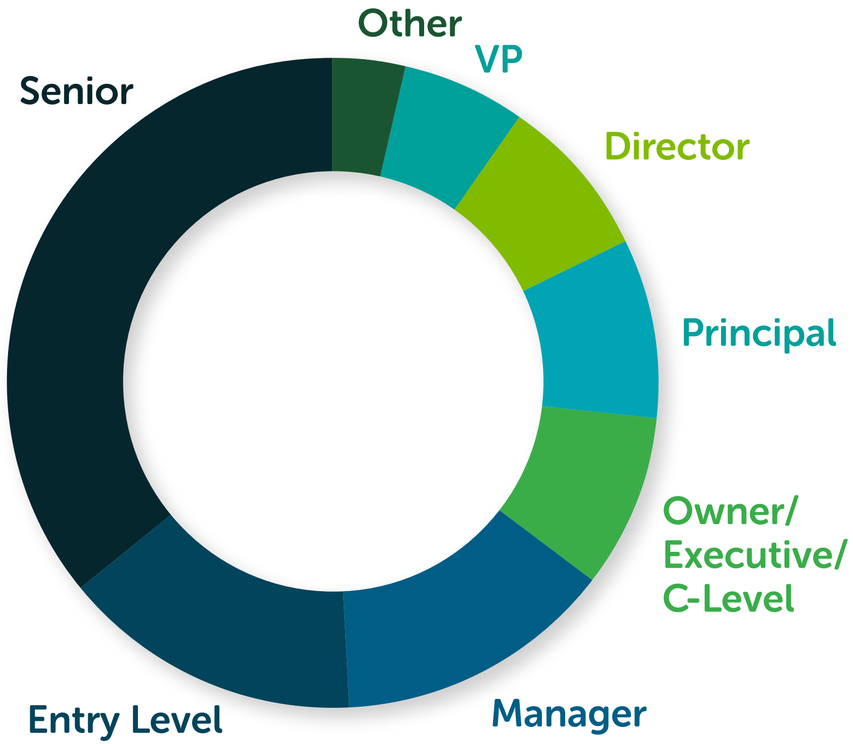
We started this year's survey with a series of questions designed to map the world of data professionals - geographically, in the organization, and even generationally. Data scientists and other data professionals span ages, education levels, and are situated around the globe. A look through the demographics of our respondent set provides an interesting snapshot of today's data science community.

MORE THAN 2,300 PEOPLE FROM 100+ COUNTRIES PARTICIPATED IN OUR ONLINE SURVEY

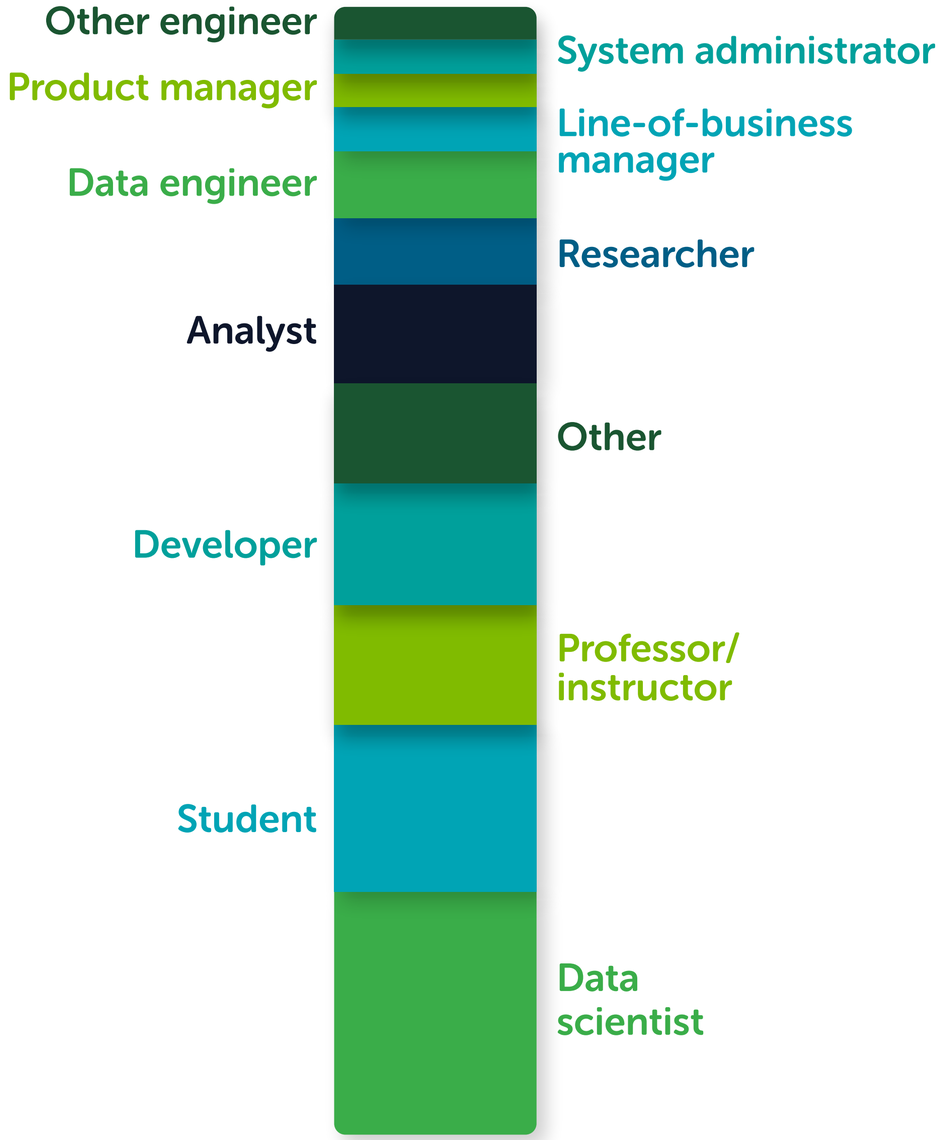


Our respondents trend young and educated. 30% hold a bachelor's degree, 36% a master's, and about 1 in 5 hold a doctorate degree. The remaining 15% have not yet completed a university degree. Nearly half (49%) fall into the Millennial age cohort (23-38), and 9% fall into the 18-22, or Generation Z, cohort. Another 28% fall into Gen X (39-54) and the remainder (14%) are over the age of 55.

RESPONDENT JOB LEVEL



RESPONDENT JOB ROLE



Data scientists represent the largest portion of our respondents, but our survey drew responses from a number of roles and seniority levels.

RESPONDENT INDUSTRY

Technology	15%	Academic	4%
Consulting	9%	Manufacturing	3%
Healthcare	9%	Education	3%
Finance	7%	Insurance	3%
Energy	6%	Retail	3%
Banking	4%	Telecommunications	2%
Government	4%	Automotive	2%
Engineering	4%		

n=939

77% of our respondents who work in commercial environments do so in these 15 industries.

Data professionals do their work everywhere, in institutions of higher learning, nonprofits, and commercial entities. About 15% of our respondents self-identified as students. Of the remaining 85%, 1,232 work in commercial organizations, 478 in academic institutions, 203 in not-for-profits, and 176 in government agencies. Virtually every industry is represented in our respondent set—our commercial respondents work in thirty-three industries, ranging from entertainment and recreation to technology, finance, healthcare, and energy.

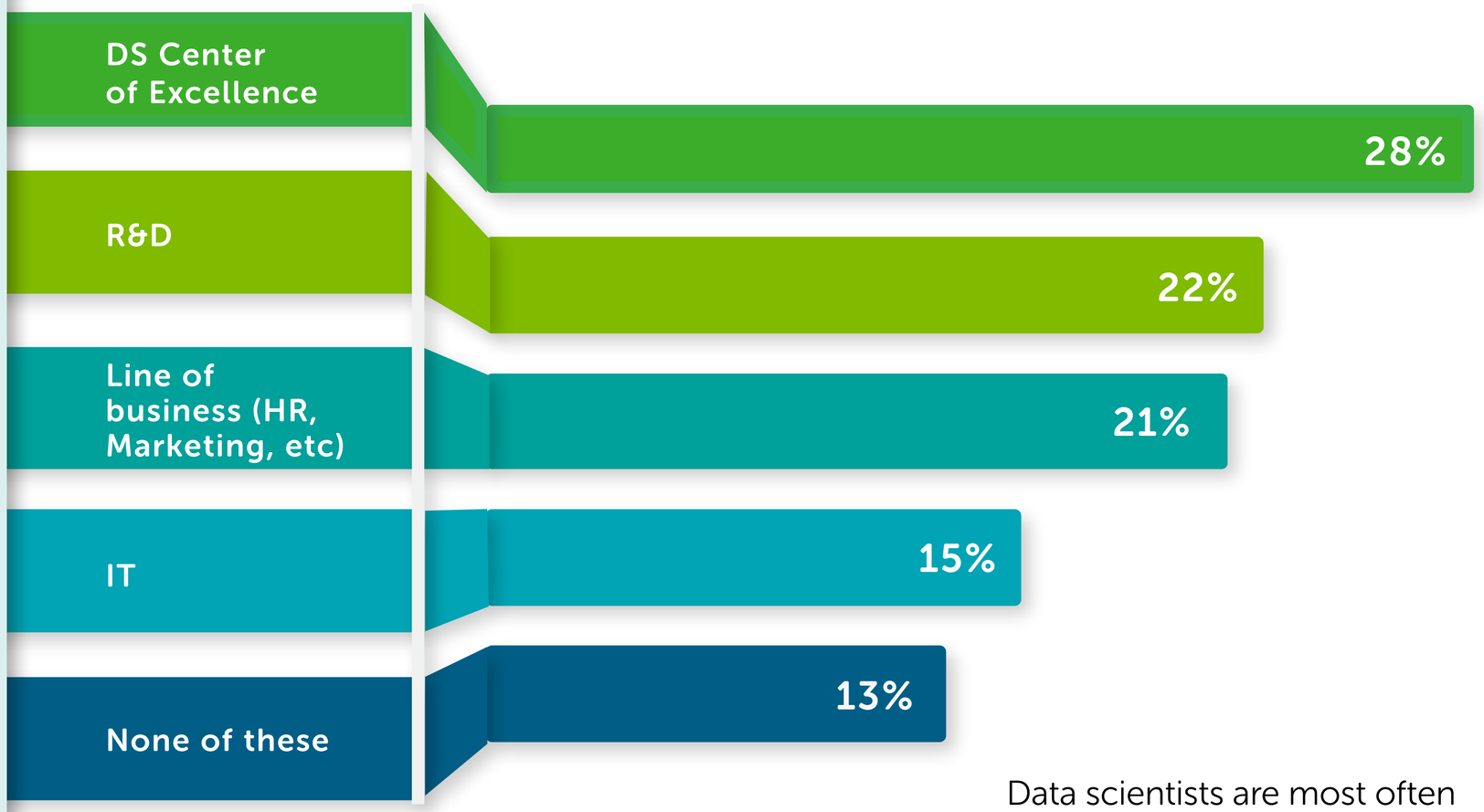
DATA PROFESSIONALS AT WORK

The majority of our respondents (59%) work in commercial environments. We took a closer look at those environments, to find out where data professionals sit in the organization, how they spend their time, and what tools they use. Another question on our minds: what are their greatest challenges?

Where do data scientists fit in the organization?

There is no one-size-fits-all approach to data science team structure. About one in five data scientists work in a variety of departments, but 28% are stationed in a centralized data team or Center of Excellence today. As data science continues its ascent to a strategic discipline in many organizations, we expect larger organizations to establish a Data Science Center of Excellence to maximize the business impact from data science and provide professionals an opportunity to cross-train in various departments. In the survey, organizations with more than 10,000 employees were most likely to have already deployed this model.

WHAT DEPARTMENT DOES YOUR ROLE FALL UNDER?

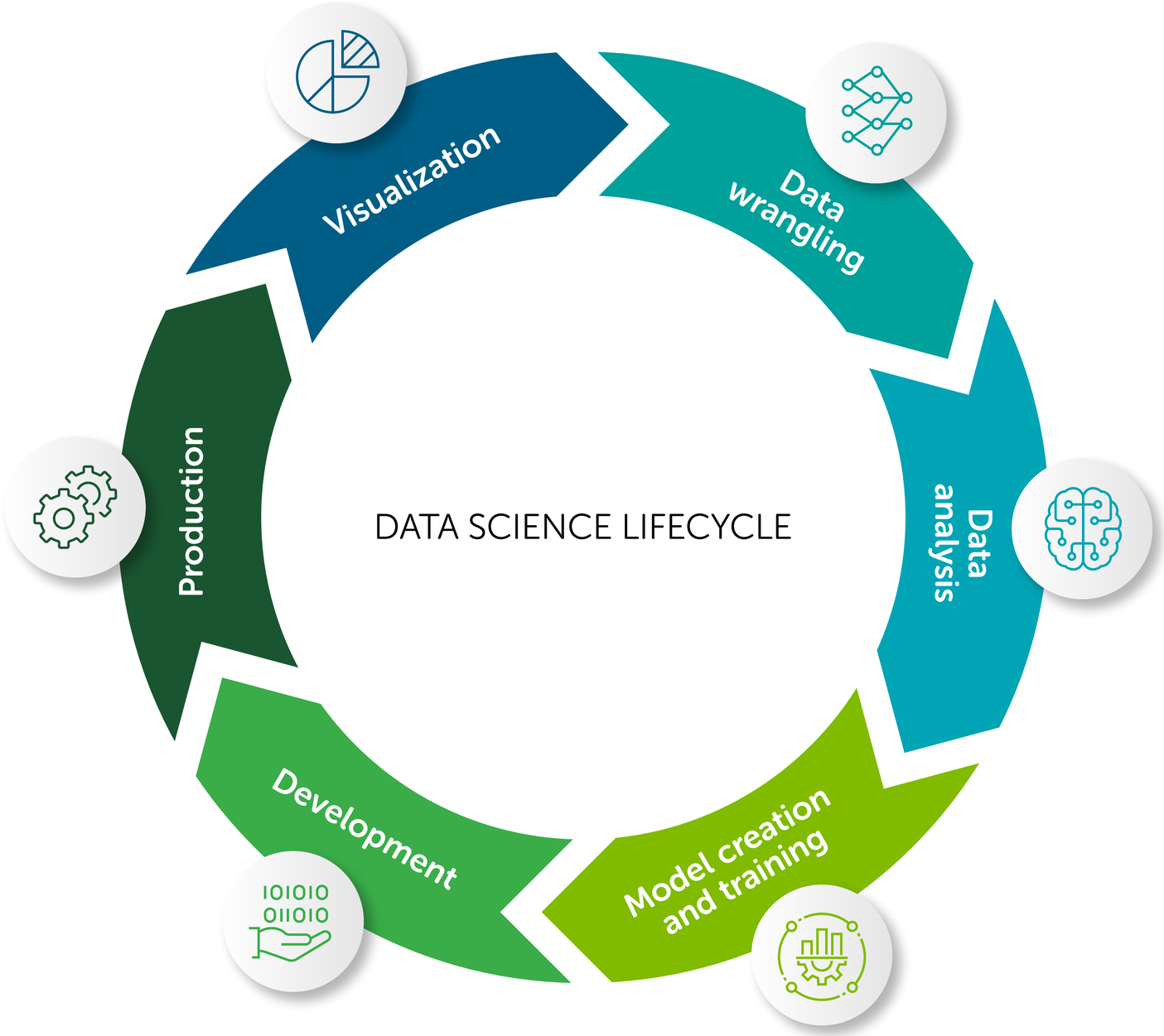


n=522

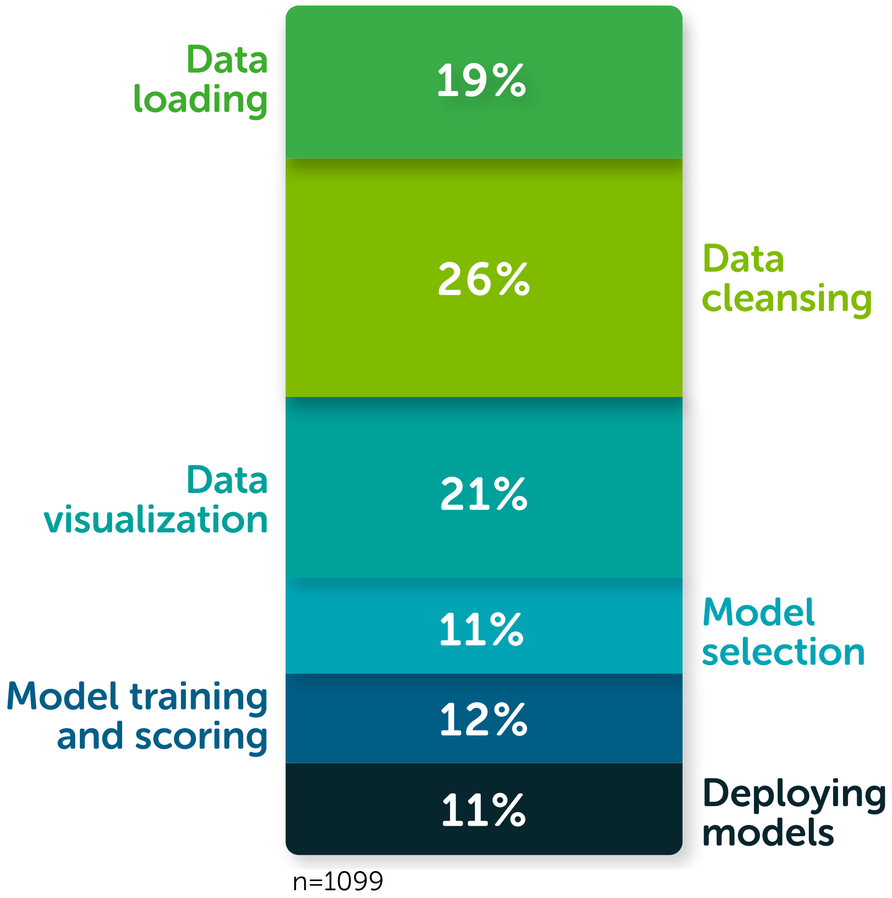
Data scientists are most often working in a Data Science COE or R&D organization.

A day in the life

Data scientists spend their time across a wide range of tasks that require a diverse skill set. Today's data scientist is a mathematician and a modeling expert who needs to understand everything that may impact their workflows. From data preparation, passing through visualization best practices and model training, and ending with DevOps knowledge to achieve proper deployment, a data scientist must have a good handle on all the components that may impact their analyses.



THINKING ABOUT YOUR CURRENT JOB, HOW MUCH OF YOUR TIME IS SPENT IN EACH OF THE FOLLOWING TASKS? (PLEASE ASSIGN A PERCENTAGE; TOTAL MUST ADD UP TO 100%.)

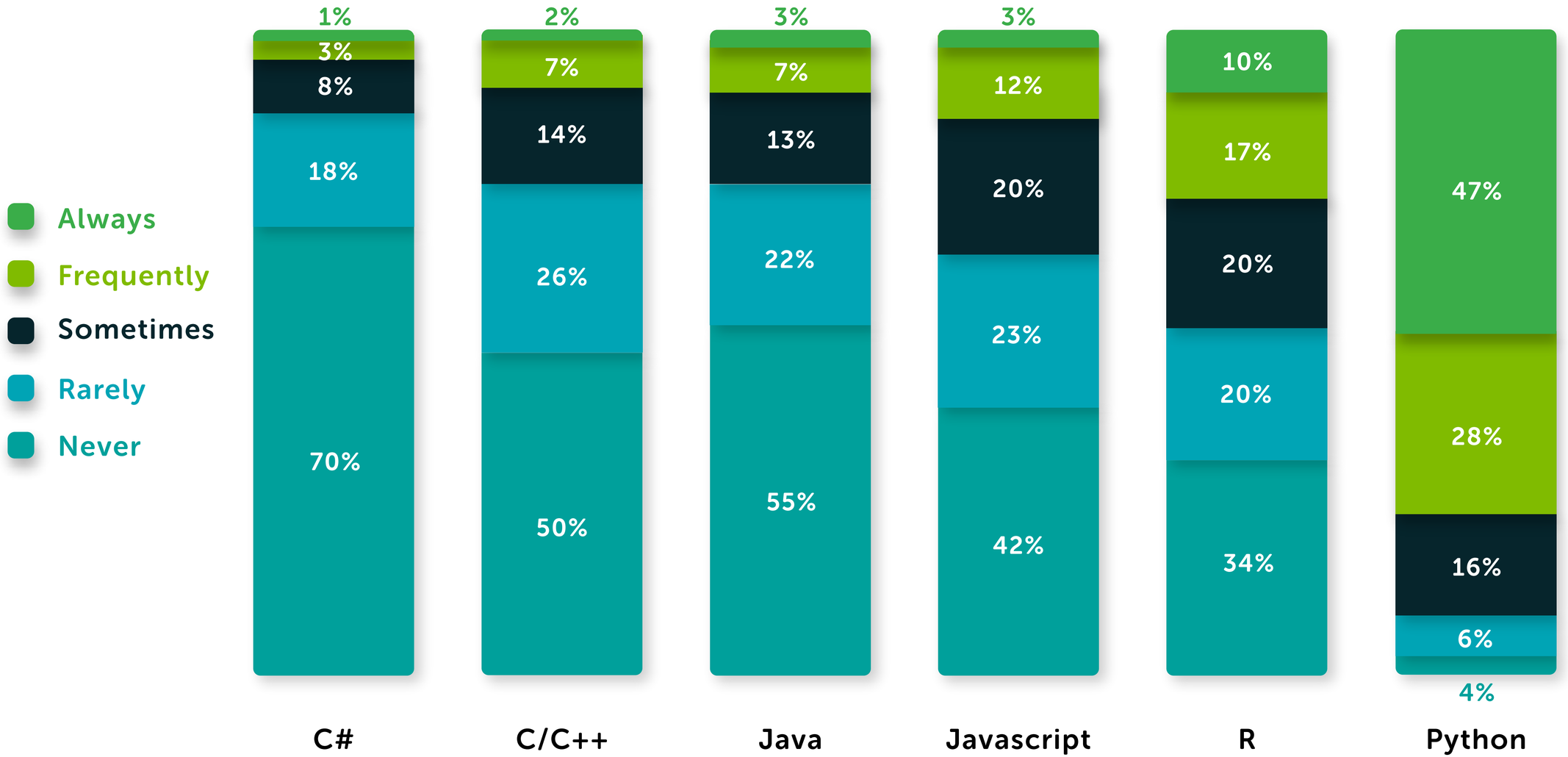


For most respondents, data management tasks still consume a disproportionate portion of work time.

We were disappointed, if not surprised, to see that data wrangling still takes the lion's share of time in a typical data professional's day. Our respondents reported that almost half of their time is spent on the combined tasks of data loading and cleansing. Data visualization tasks come second, taking about 21% of time. Modeling tasks consume the remaining third of a data professional's time, with selection comprising 11%, training and scoring 12%, and deployment 11%.

Data preparation and cleansing takes valuable time away from real data science work and has a negative impact on overall job satisfaction. This efficiency gap presents an opportunity for the industry to work on solutions to this problem, as one has yet to emerge.

HOW OFTEN DO YOU USE THE FOLLOWING LANGUAGES?



n=1592

While languages like C++ and Javascript are still in the mix,
Python leads the pack for data science usage among our respondents.

The data science toolkit

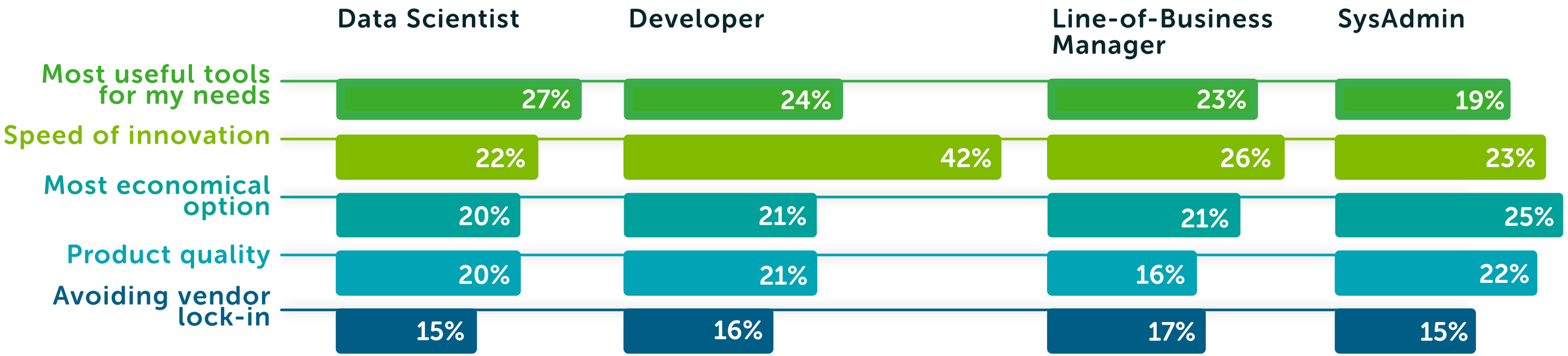
In addition to the variety of skills required of data scientists, the discipline also demands that practitioners have fluency in a wide range of tools and technologies, from data management to visualization tools, familiarity with open-source packages and libraries and even DevOps. Our respondents lean on a diverse set of tools and platforms to get their jobs done, drawing from a mix of open-source and proprietary solutions and multiple programming languages.

Python is the most commonly-used language among our respondents, with 75% of respondents using it frequently or always in their roles. R takes second place with 26% of respondents relying on it regularly.

Given the nature of this survey sample and Anaconda's widespread popularity, Anaconda was cited as the most commonly-used platform among those working for commercial organizations, followed by RStudio, which is also used by 44% of Anaconda users. But enterprises are using a number of tools and platforms to deliver on their data strategy, including a mix of proprietary, open-source, and hybrid solutions. We hope to see expanded collaboration among industry players to ensure interoperability and harmonization among different tools.

Respondents cited multiple benefits of using open-source technologies. When asked to assign a proportional value to each of five commonly-cited benefits of open-source software, "most suitable tool for my needs" and "speed of innovation" took the most points. We did note that developers assigned comparatively high value to the speed of innovation coming from open-source solutions, assigning on average 42% of the total value to this attribute and an additional 24% to its suitability for their needs. In contrast, data scientists take a more balanced view of these two attributes, assigning 27% of the value to utility for their needs and 22% to the rate of innovation.

WE WOULD LIKE TO KNOW WHAT YOU VALUE MOST ABOUT OPEN-SOURCE TECHNOLOGY. (PLEASE ASSIGN A PERCENTAGE TO EACH FEATURE BASED ON IMPORTANCE)



n=1592

Developers and line-of-business managers assigned higher value to open-source's speed of innovation, compared with data scientists, who valued its utility for their purposes, and system administrators who valued its economy most highly.

Getting to production

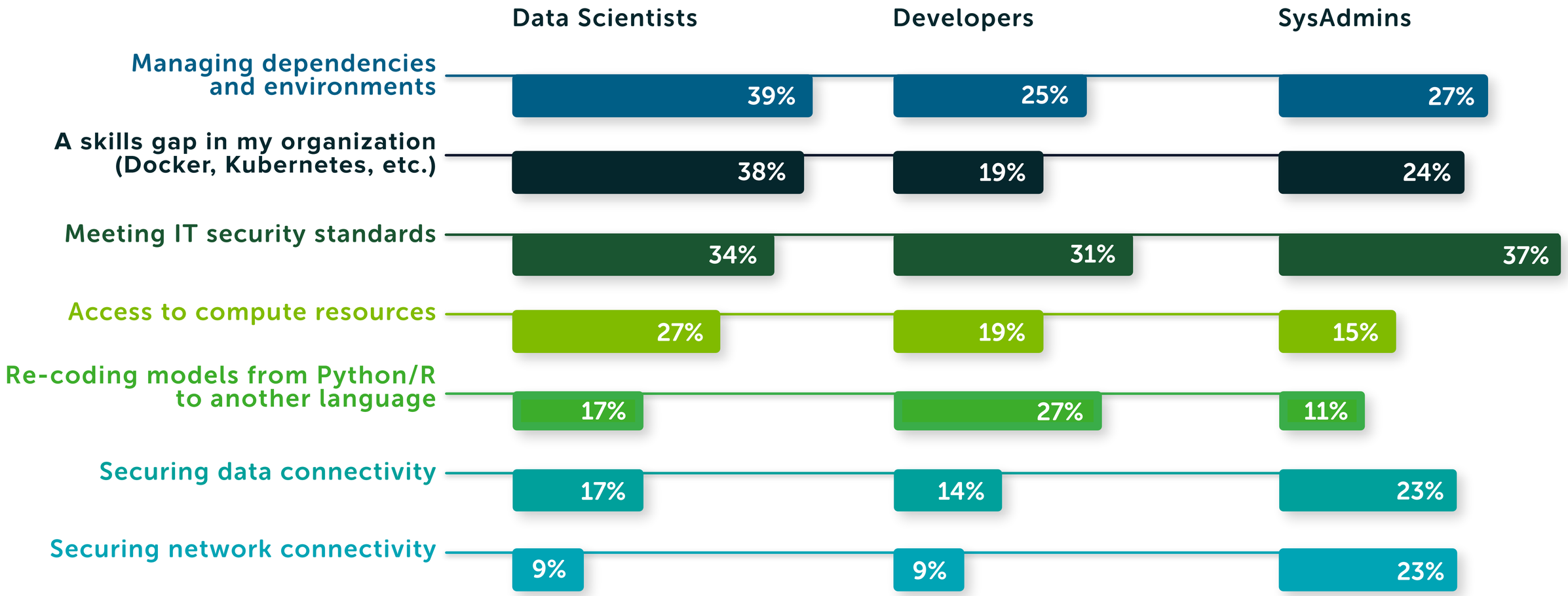
Data's potential value in an enterprise extends well beyond gathering a few business insights and refreshing canned dashboards to monitor KPIs. Competitive advantage from data science lies in deploying machine learning models and other data science outputs to power other business functions and products, and delivering this value is certainly one of the most satisfying aspects of a data scientist's work. But getting to production is often fraught with challenges outside the control of data professionals.

Respondents working in commercial environments cited three top roadblocks to production:

1. Managing dependencies and environments
2. Lack of skills in the organization to deploy models to production
3. Meeting IT security challenges

An interesting finding emerges when we compare responses from those in data scientist roles with those in other roles: data scientists see the biggest roadblock to production as managing dependencies and environments, while developers and system administrators cite meeting IT security standards as their biggest blocker. Meanwhile, developers are most sensitized to the re-coding requirements often involved in productionizing data science output. Enterprises have an opportunity to improve efficiency by driving cross-functional visibility to these challenges and identifying solutions to them.

WHAT ROADBLOCKS DO YOU FACE WHEN MOVING YOUR MODELS TO A PRODUCTION ENVIRONMENT?



Data scientists, developers and system administrators differ on their top production blockers.

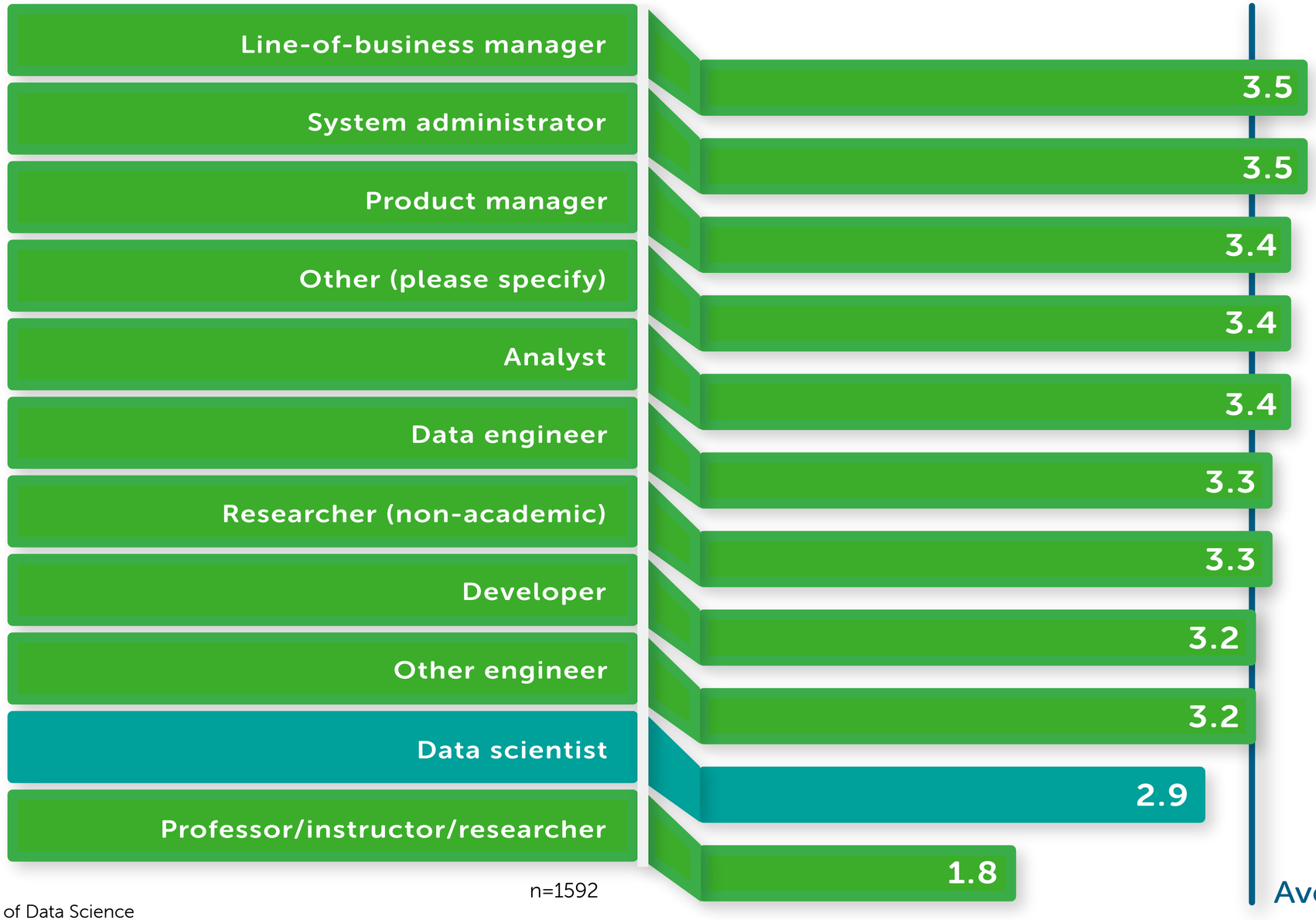
Managing security challenges

As with any development pipeline, data science comes with inherent security management challenges. Compared to proprietary software, open source's very nature enables its contributors and maintainers to catch and patch vulnerabilities quickly. Nevertheless, software security issues are simply a fact of life, and managing them will always consume resources in any organization.

Across our sample, people in different roles bring a different perspective on open source and security. Of note, respondents who cited their profession as professor/instructor/researcher had the lowest level of concern about open-source vulnerability management. On the one hand, this may be because this respondent set is closest to efforts to correct vulnerabilities in open-source tools. On the other, it may reflect a gap in university data science curricula, in which students do not gain sufficient understanding about security and vulnerability management to prepare them for commercial environments.

System administrators and LOB managers reported the highest level of concern about managing security vulnerabilities. This finding aligns with the previous question, in which system administrators cited meeting security standards as a key production roadblock. Business leaders would naturally also be sensitized to this issue, given both its role in slowing deployment of models and overall business risks presented by security issues.

ON A SCALE OF 1 TO 5, HOW CONCERNED ARE YOU ABOUT MANAGING SECURITY AND VULNERABILITIES IN OPEN-SOURCE TOOLS?

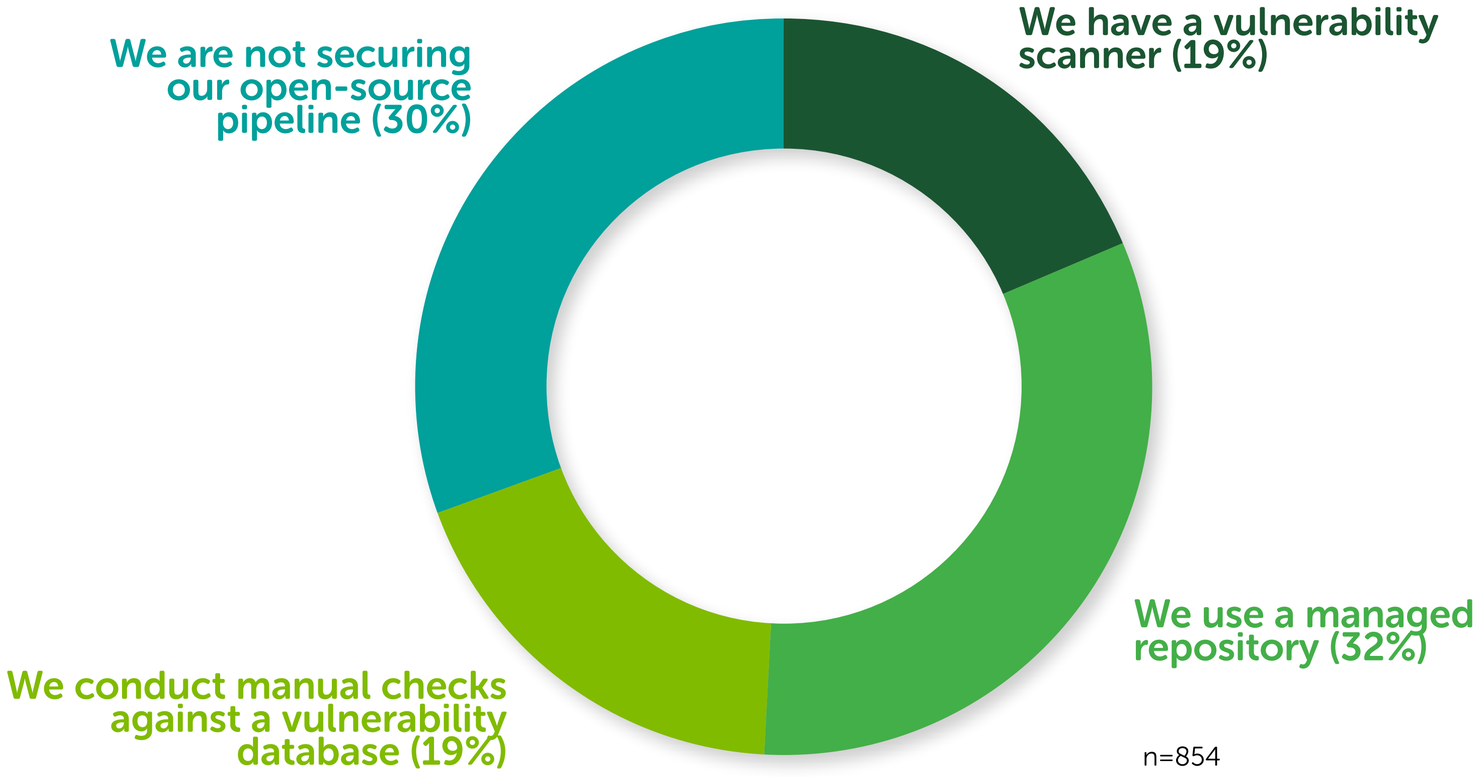


1 = not at all concerned
5 = extremely concerned

n=1592

People in different roles bring a different perspective on open source and security.

HOW DO YOU ENSURE THAT OPEN-SOURCE PACKAGES USED FOR DATA SCIENCE AND MACHINE LEARNING MEET ENTERPRISE SECURITY STANDARDS?



n=854

There are a variety of approaches to securing open source, but enterprises should take note of gaps in their policies and procedures. A concerning 30% of respondents who had knowledge of their company's security practices stated that their organization does not have any mechanism in place to secure open-source data science. Given the prevalence of open-source software in production workflows, this creates vulnerabilities and risks that can deliver far-reaching negative impacts.

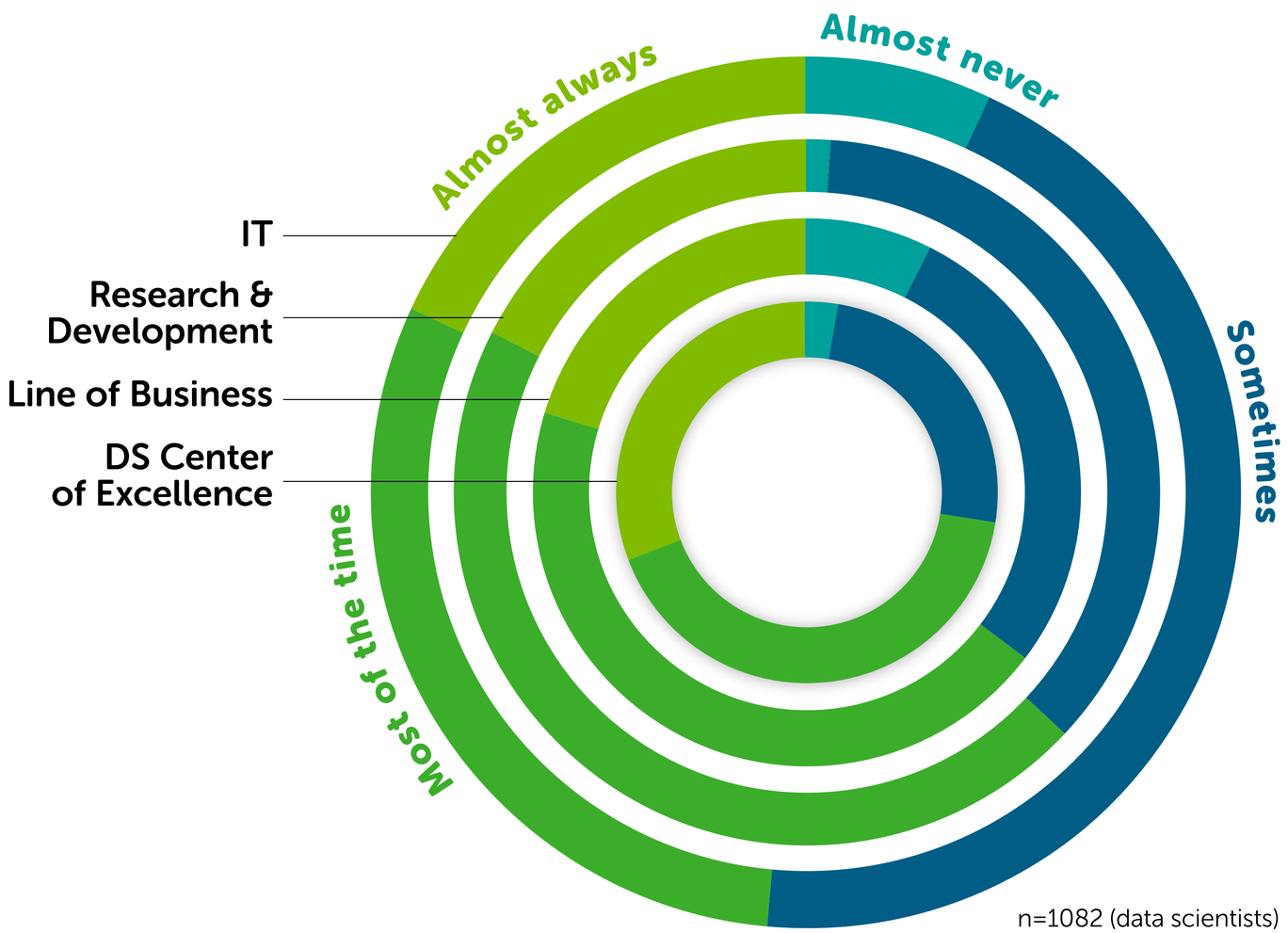
Among those who know their employer's approach to security,
30% report that they have no mechanism in place to secure open-source pipelines.

Demonstrating value and job satisfaction

Data science teams come in many different shapes and sizes. We compared four different approaches to data science in the enterprise: integrated with R&D, integrated with IT, distributed into the line of business, and a single Data Science Center of Excellence (COE). These organizational models seem to have some impact on job satisfaction as well as a data professional's ability to demonstrate the business impact of their work.

Overall, 48% of respondents report that they can demonstrate business impact often or most of the time. For data scientists, the ability to demonstrate impact correlates strongly with where their team is situated in the organization.

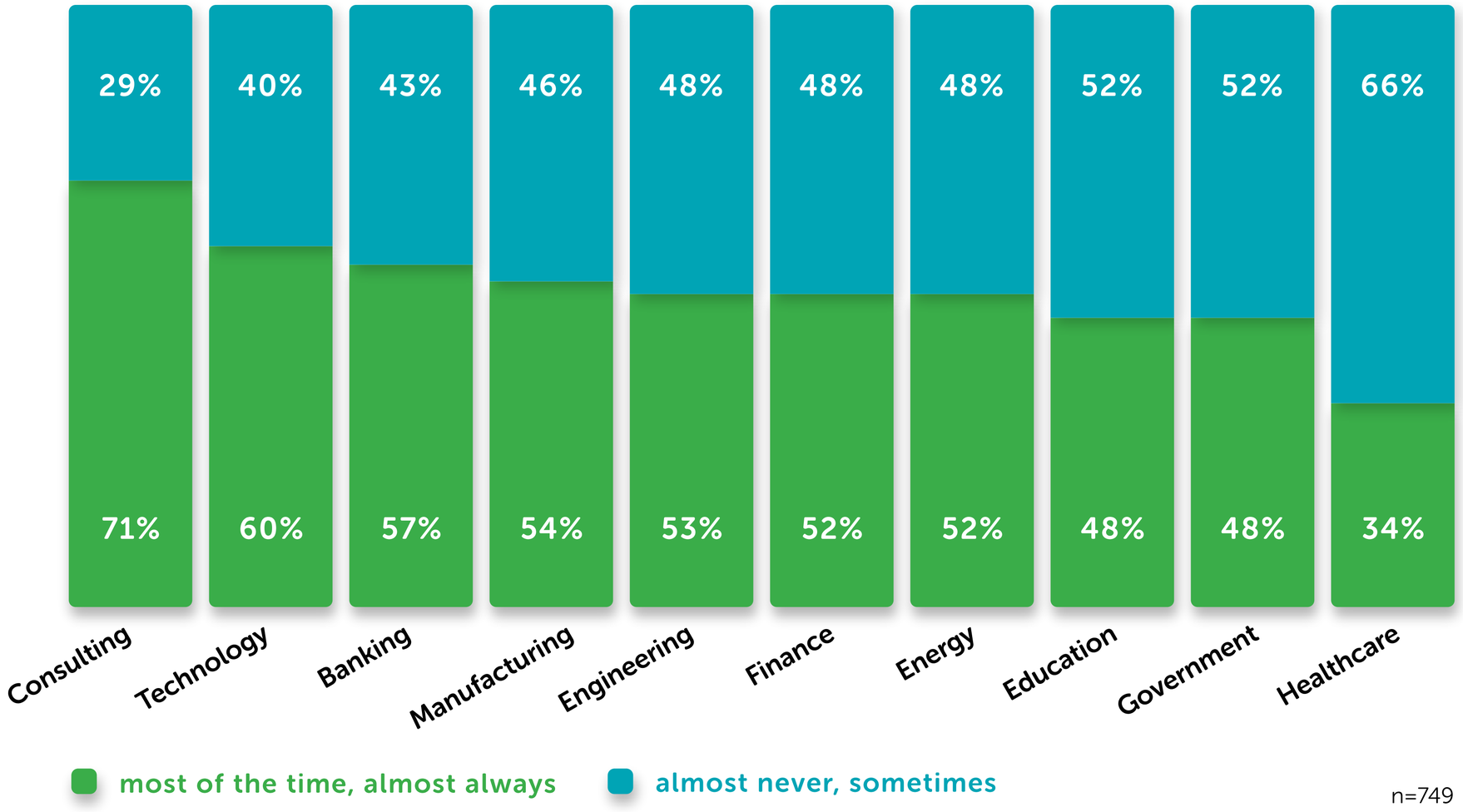
IN YOUR OPINION, HOW EFFECTIVE IS YOUR TEAM AT DEMONSTRATING THE IMPACT DATA SCIENCE HAS ON YOUR COMPANY'S BUSINESS OUTCOMES?



n=1082 (data scientists)

Professionals working in a COE are most likely to report that they can often demonstrate the business impact of their work, while those in IT organizations report the most difficulty doing so.

IN YOUR OPINION, HOW EFFECTIVE IS YOUR TEAM AT DEMONSTRATING THE IMPACT DATA SCIENCE HAS ON YOUR COMPANY'S BUSINESS OUTCOMES?



n=749

Like organizational structure, a data professional's choice of industry may also determine the degree to which they feel they can demonstrate the business impact of their work. Consulting, technology and banking have the highest rates of reported ability to demonstrate value; data professionals in healthcare environments reported being able to do so only one-third of the time.

Data professionals in consulting, tech, and banking most often report success in demonstrating business impact from their work.

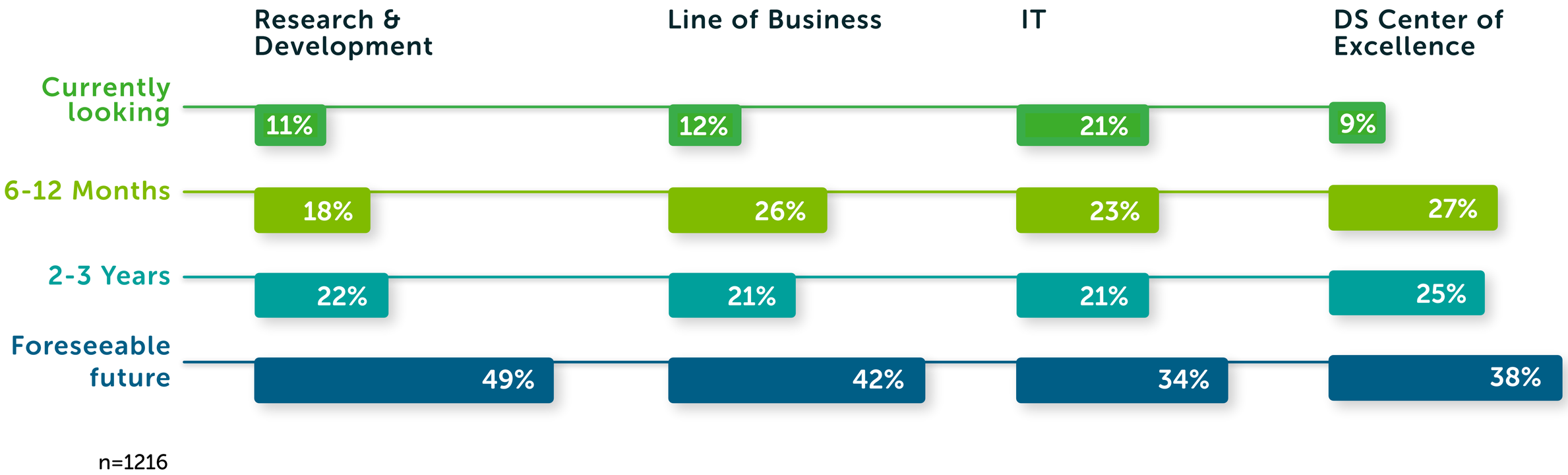
Job satisfaction is correlated with organizational structure among our respondents. Data professionals working in R&D organizations report the longest planned tenure with their current employers, followed by those working in an LOB. Given that data professionals working in IT organizations also report frustrations in demonstrating business impacts from their work, perhaps it is no surprise that only 34% of them plan a lengthy tenure with their current employer.

Across all of the different departments, there is potential for a high rate of employee churn in the 1-2 year horizon. Given the well-understood talent shortage in this profession and the need for data scientists to develop a strong understanding of the environments in which they work to add value, organizations should identify and invest in high-impact programs to drive retention among data professionals.

ONLY 34%

OF DATA PROFESSIONALS WORKING IN IT ORGANIZATIONS PLAN A LENGTHY TENURE WITH THEIR CURRENT EMPLOYER

HOW LONG DO YOU PLAN TO STAY WITH YOUR CURRENT EMPLOYER?



n=1216

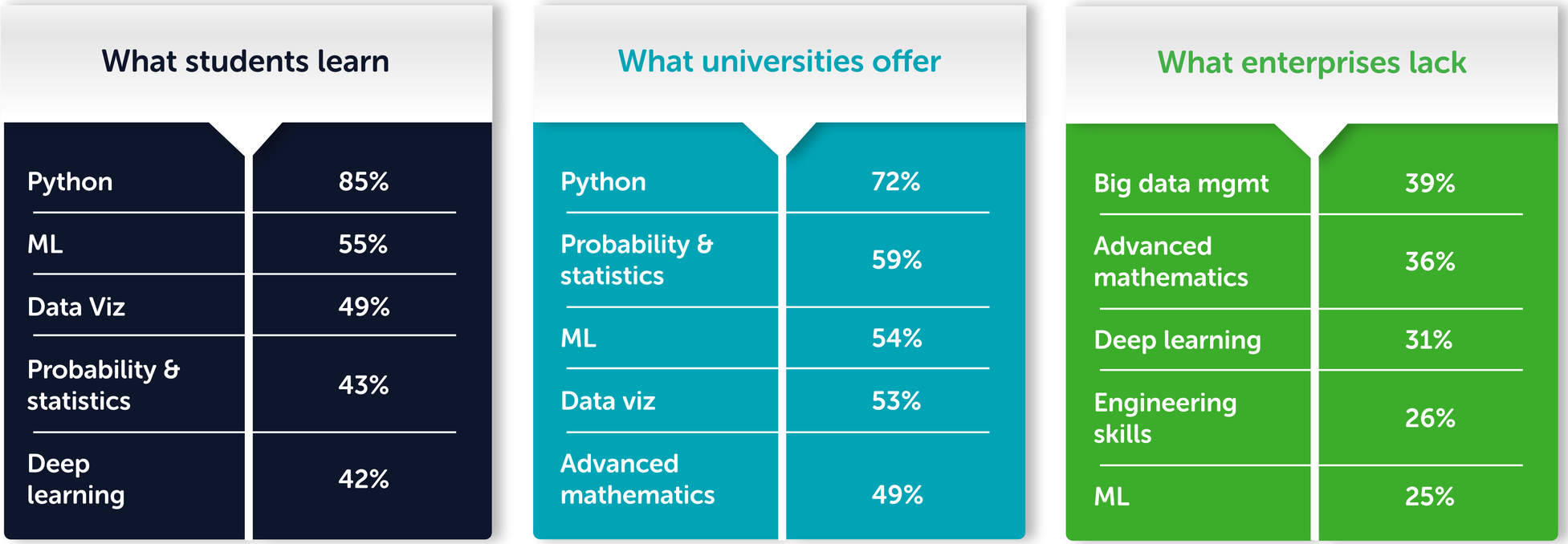
IT organizations should take note: fully 44% of data professionals situated in your organization plan to seek employment elsewhere within the next year.

THE SKILLS GAP

As universities and other institutions add or expand their data science degree programs, we wondered if data science graduates are ready for the workplace, and how they are experiencing the job market for early-career hires.

Do universities and other institutions adequately prepare data scientists?

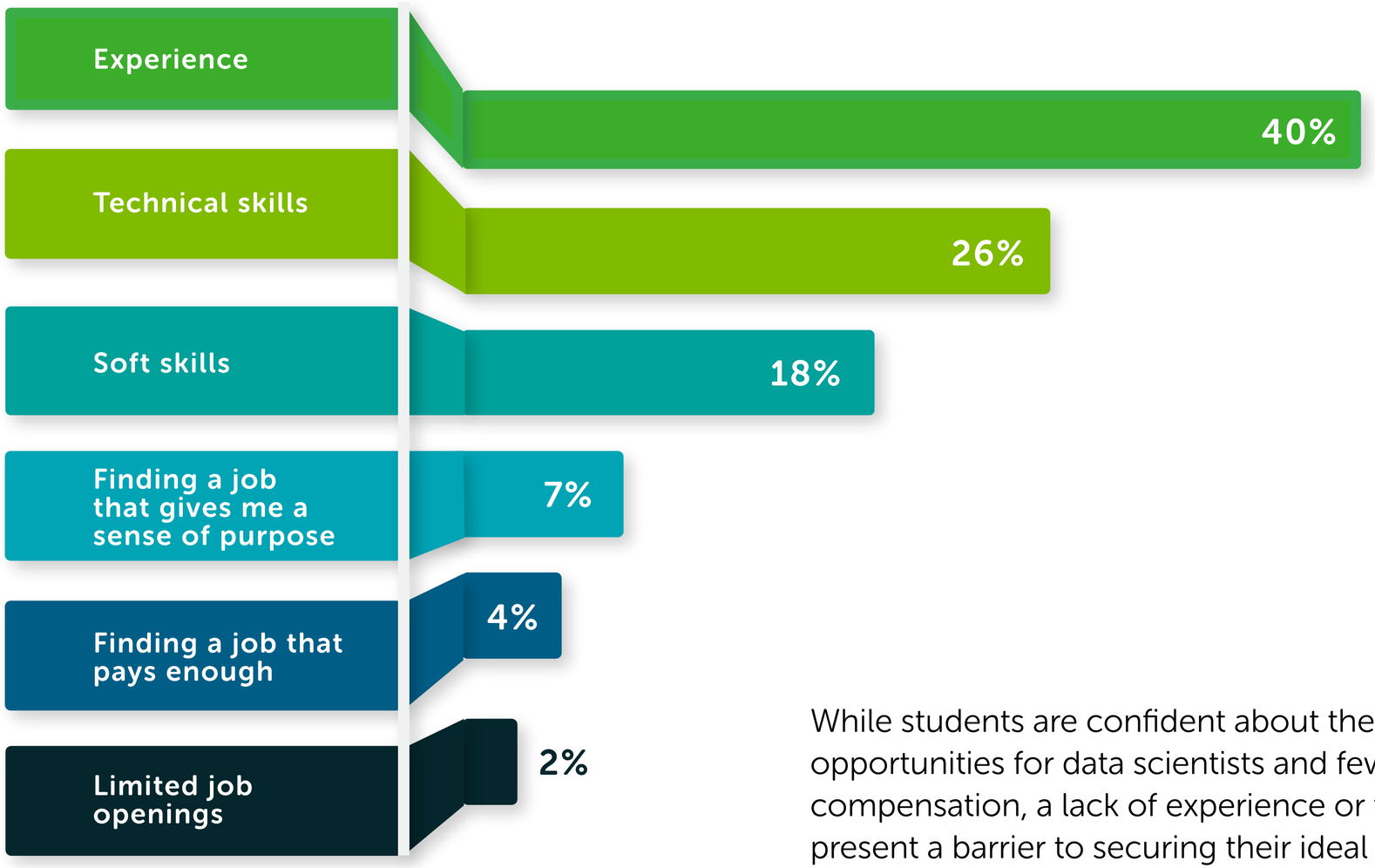
Our study indicates that there are gaps between what enterprises need and what institutions teach. Two of the most frequently-cited skills gaps among respondents working in enterprise environments - big data management (38% of respondents) and engineering skills (26%) - do not rank in the top 10 skills offered in university programs. We asked students what they're learning, universities what they're teaching, and enterprises what skills they are lacking.



n= 346 n=238 n=1216

Enterprises report that they are missing key skills that students don't report they are learning, and universities don't report they are teaching.

IN YOUR OPINION, WHAT IS THE BIGGEST OBSTACLE TO OBTAINING YOUR IDEAL DATA SCIENCE JOB?



n=346

While students are confident about the number of opportunities for data scientists and few of them worry about compensation, a lack of experience or technical skills can present a barrier to securing their ideal role.

Early-career data scientists and the job market

As they turn their attention to the job market, students that participated in our survey report their biggest obstacles to finding a job are experience (41% of respondents) and technical skills (27%). Strong internship and practicum programs can address these gaps, and universities should ensure that these programs go beyond providing a resume enhancement and hands-on-keyboard technical skills. Good internships also prepare students for the nuanced challenges faced by a data professional in an enterprise: serving as a "data translator," demonstrating business impact from their work, and influencing colleagues cross-functionally to address production roadblocks and secure access to resources. Partnerships between universities and employers have the potential to surface curriculum gaps such as those noted above.

Considerations for a more complete educational experience

As data science's role in enterprises continues to gain prominence, two more findings from our study prompt concern. Only 15% of universities responded that they are offering training in ethics, and business knowledge is also only covered in 15% of data science programs. These two skill sets, already important components of a data professional's work, will become critical as data science becomes more core to business outcomes. We recommend that universities increase the prominence of these two subjects in their curricula, threading key themes into standard coursework and practical experiences, as well as explicitly requiring the completion of dedicated courses in each topic. Without sufficient grounding in business fundamentals and ethical considerations, no data scientist is fully prepared for their career.

Ethics and business knowledge are already important components of a data professional's work, and will become even more critical as data science becomes increasingly core to business outcomes.

THE BIG QUESTIONS

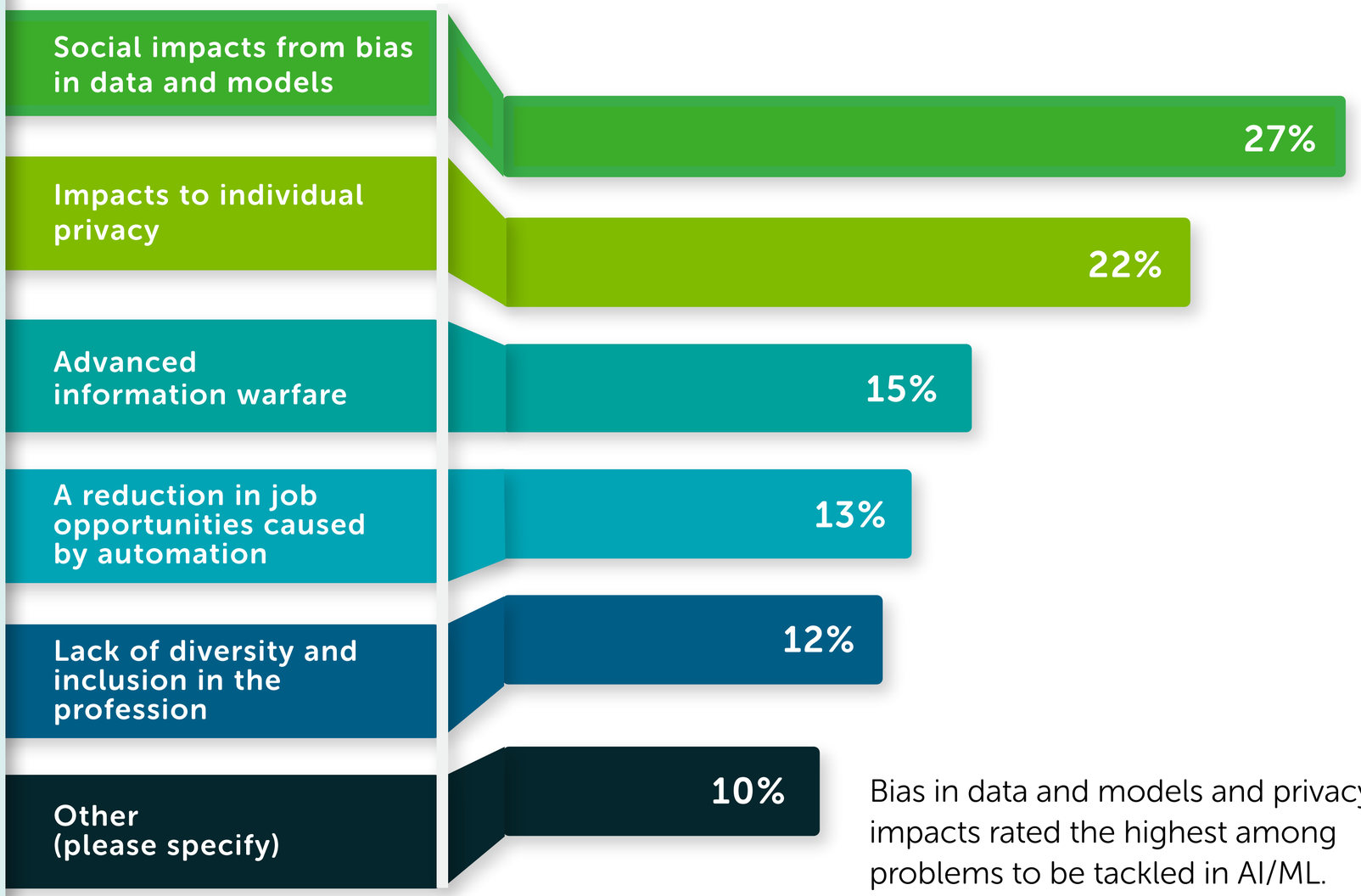
We wanted to know what worries the community of data professionals who responded to our survey. Given the increasingly influential role of ML and AI in business and society, data professionals grapple with big questions.

Issues of concern

We asked respondents what they consider to be the biggest issue to tackle in AI and ML. The largest portion of respondents (27%) cited social impacts that stem from bias in data and models, and 21% cited impacts to individual privacy.

Fully 10% of respondents to this question selected "other" and offered suggestions outside of the provided response set. The most commonly-mentioned themes: organizational/ business understanding of ML and AI, hype about the potential of these technologies, non-bias related social impacts such as the environmental impact of computing and wealth concentration effects from AI and ML, and skills gaps for real-world data science work.

WHAT DO YOU THINK IS THE BIGGEST PROBLEM TO TACKLE IN THE AI/ML AREA TODAY?



n=1592

Bias in data and models and privacy impacts rated the highest among problems to be tackled in AI/ML.

The impact of modern data science on day-to-day business practices, politics, and society is growing rapidly. Important and complex questions of ethics, responsibility, and fairness should be on the minds of every data scientist, business leader, and academic.

There are no simple answers to these questions; rather, their consideration should be a constant thread informing data science work. Enterprises should treat ethics, explainability, and fairness as strategic risk vectors and treat them with commensurate attention and care, but we have concerns about the data professional workforce's ability to do so today. Only 15% of instructors and professors that responded to our survey said they are teaching AI/ML ethics, and only 18% of student respondents say they are learning AI/ML ethics.

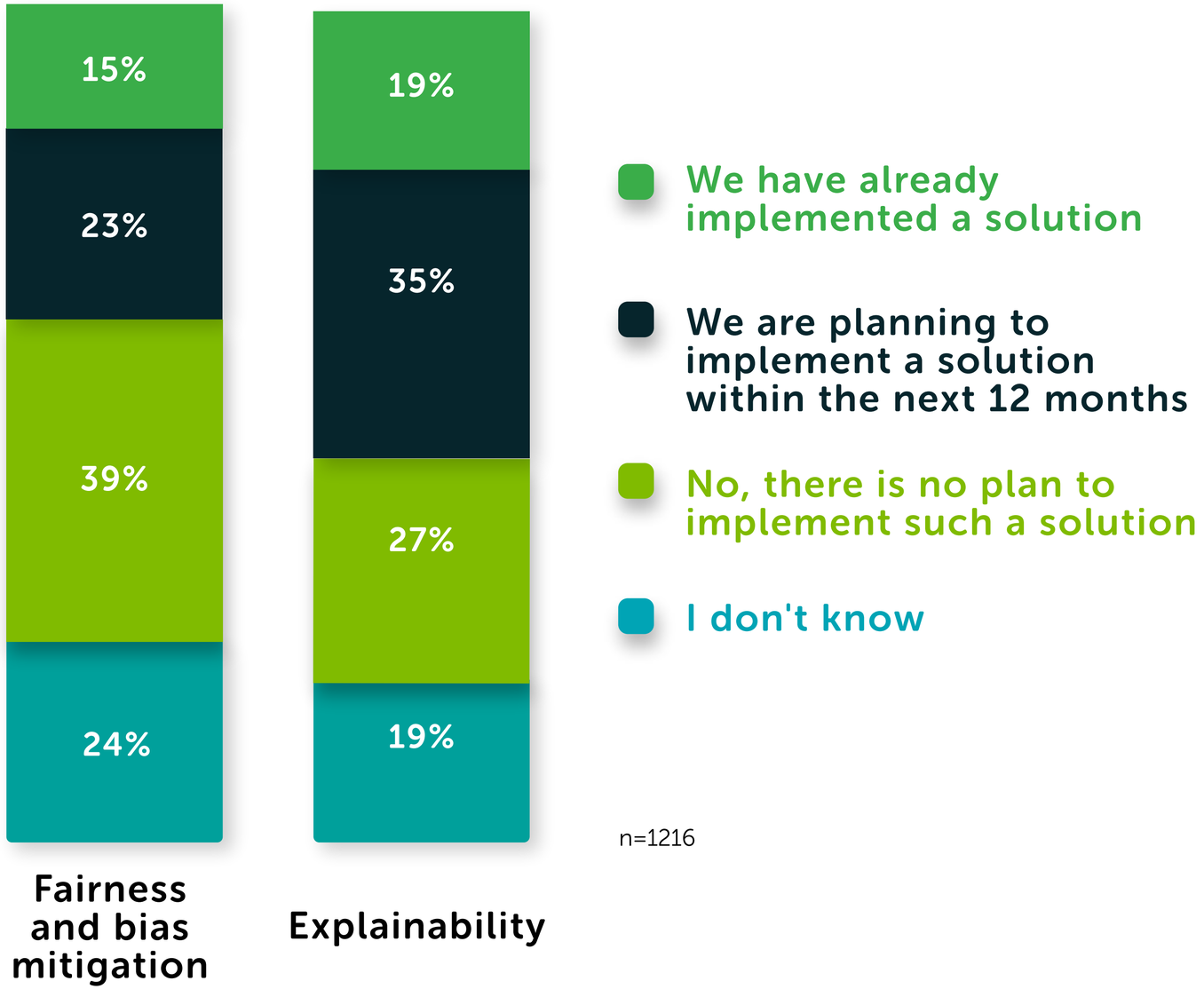
Fairness & explainability

Tools addressing both fairness and explainability for ML models have begun to emerge and may expand the solution set for these problems. Fairness tools measure bias in models and data sets, while explainability tools help data scientists explain their models' decisions. They can be used to explain decisions made by "black box" models or to help build and train "glass box" models.

Only 15% of respondents indicated that their organization has already implemented a fairness solution, and only 19% said they have an explainability solution in place. Looking ahead, more organizations (35%) are planning to implement explainability tools than fairness tools (23%).

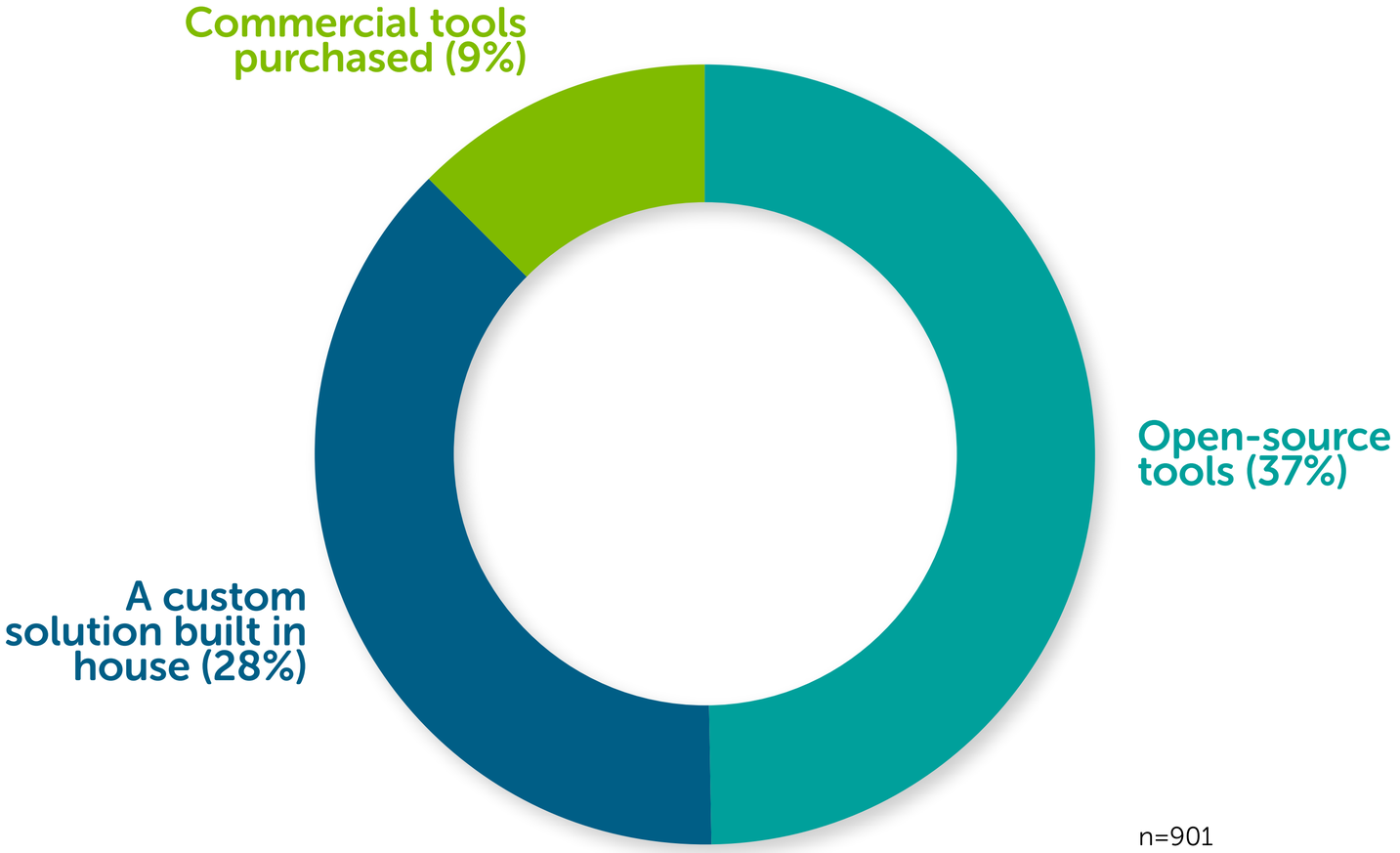
Adoption of fairness and explainability tools is emerging, but many organizations are holding off on plans to implement these solutions.

IS YOUR DATA SCIENCE TEAM WORKING ON SOLUTIONS TO ADDRESS DATA BIAS OR MODEL EXPLAINABILITY?



n=1216

REGARDING EXPLAINABILITY AND FAIRNESS, WHAT KIND OF SOLUTIONS HAVE YOU IMPLEMENTED OR DO YOU PLAN TO IMPLEMENT?



Regarding explainability and fairness, what kind of solutions have you implemented or do you plan to implement?

The embrace of open-source solutions is unsurprising given the innovation currently happening in this space. We are particularly excited about the promise of initiatives like deon, FairLearn, and AI Fairness 360 to address bias, as well as InterpretML, and LIME in the explainability space.

Questions about fairness and interpretability are more important than ever. As AI uses data to make more impactful, life-changing decisions (hiring, judicial sentencing, and credit approval, to name a few), humans must ensure that these decisions are as free of bias as possible and that they are explainable to those who are affected.

LOOKING AHEAD

Four themes for organizational focus

Today, the data science discipline is finding its identity, but the journey to maturity is ongoing. We expect that the next 2-3 years will continue data science's trajectory towards becoming a strategic business function across a wider range of industries than today, but that institutions and enterprises will face continued growing pains in the process. Leaders will find their attention increasingly focused on four themes:

1 Value realization

Commercial organizations still struggle to reach meaningful insights and value from data science. Data analysis work done in isolation can offer important insights, but falls short of the discipline's full potential to transform industries, improve society, or offer competitive advantage. Getting data science outputs into production will become increasingly important, requiring leaders and data scientists alike to remove barriers to deployment and data scientists to learn to communicate the value of their work.

With the newfound prominence of epidemiology and other data sciences in the wake of the COVID-19 pandemic, and the use of data analysis and visualization in studies of racial injustice and police violence, the value of data analysis has become clear to a wider audience than ever before. This may continue to raise the profile of the discipline and its importance in a wide range of industries.

2 Integrating open-source technology

Given the rapid acceleration of activity in the data science discipline, the wide range of tools in use by data scientists, and the sensitivity of many data sources, the integration of open-source tools into existing security and governance procedures will continue to be a prominent concern on the minds of IT and business leaders.

As developers did in the past, data scientists will challenge existing security processes with demand for innovative tools and prolific use of open-source libraries. Data professional leaders and IT professionals will find themselves at the table time and time again as organizations adapt to new requirements to securely support data science. Organizations should take a proactive approach to integrating open-source solutions into the development pipeline, ensuring that data scientists do not have to use their preferred tools outside of the policy boundary.

3 Talent development, acquisition, and retention

Much has been written about the data science talent shortage. Universities are increasingly adding and expanding data science degree programs, numerous other organizations have created professional reskilling and upskilling offerings, and organizations are competing for talent with aggressive compensation and perks.

The talent issue has numerous dimensions, and the most successful organizations will address it from multiple angles. First, educational institutions and employers should collaborate closely to ensure that new graduates come into the workforce with the knowledge and experience required to succeed, and conversely, that enterprises are ready to embrace the innovation and fresh approach to analysis that they will bring to the table.

Second, employers should look beyond compensation and design holistic data science talent retention strategies focused on helping data scientists gain experience articulating the value of their work, giving them opportunities to continue to grow their skills, and ensuring they do not toil in obscurity.

Finally, employers should take seriously the business value of cross-training. In the short term, organizations can cross-train and upskill domain expert employees in modern data science methodologies and tools to improve outcomes. But over the long term, data scientists should be trained across multiple domains to continue their professional development and increase the value of their contributions.

4 Bias mitigation and explainability

Of all the trends identified in our study, we find the slow progress to address bias and fairness, and to make machine learning explainable the most concerning. While these two issues are distinct, they are interrelated, and both pose important questions for society, industry, and academia.

We were troubled to find that 39% of enterprises do not have a plan to implement solutions for bias mitigation, and that 27% have no plans to tackle explainability. Above and beyond the ethical concerns at play, a failure to proactively address these areas poses strategic risk to enterprises and institutions across competitive, financial, and even legal dimensions.

We see an opportunity for data professionals to exert leadership within their organizations and drive change. Doing so will increase the discipline's stature in the organizations which depend on it, and more importantly, it will bring the innovation and problem solving for which the profession is known to address critical problems impacting society.

About Anaconda

With more than 20 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.