How do I become a data scientist?

Foundational skills to focus on for a career in data science

ANACONDA®

What’s Inside

Introduction

According to LinkedIn, there has been a 37% annual growth in hiring for data scientist jobs between 2015-2019. The profession has topped their Emerging Jobs list for three years running, and the platform sees this level of rapid growth in data science roles across all industries. Businesses and organizations of all kinds now understand the value of good data and those that can wrangle, make decisions from, and communicate it.

Seeing these numbers, it can be hard to imagine that the role of data scientist didn't exist when many of those currently in the field were born. It wasn't until the 2000s when data science as it is known today was acknowledged as an emerging discipline. Because it is a recently developed profession, there is not one clear path to become a data scientist. You couldn't even get a degree in data science itself until very recently. While more and more universities are offering data science-specific degrees, it is more common for data scientists to have taken a less linear route on their career path.

While individuals may get into this field through different journeys, they share foundational technical and soft skills. To dig deeper into this, we spoke with five Anaconda employees in data science and related roles about their background.

Albert DeFusco
Data Scientist,
Product

Matthew Brock
Principal Engineer

Martin Durant
Software & Data
Engineer

Sophia Yang
Data Scientist

Michael Grant
Vice President,
Services

For those thinking of becoming a data scientist or those who want to transition into the field from another, **this guide will cover the skills needed to be successful and share stories and guidance from those currently in the field.**

What should you study in school?

There is no right answer to this question. While there are common degrees, it is also possible to transition into data science without a background in STEM. That being said, you do need to have an appreciation and a love for math and science.

In a field that evolves so quickly, ongoing education is a necessity. Many data scientists have their master's degree and their PhD in a more specialized area than their undergraduate education. While there are other ways to learn technical skills, formal education teaches independent thinking, focuses on foundational skills and research, and was generally viewed as extremely valuable by those we interviewed.

The most common undergraduate degrees:

- Computer Science
- Mathematics
- Engineering
- Data Science (becoming more common as universities offer programs)

Outside of these, natural sciences and economics are also relevant fields of study.

> “ Advanced degrees are certainly not prerequisites for being able to do the work. But being able to get a job is a different matter. If you can show you have the skills, then the degree is not the most important thing. Having said that, the proportion of people who get degrees is very high, and the proportion of those who go on to get a master's or more is much higher than it's ever been. So you will be up against people who have the same experience as you but also have a degree.
> That competition makes it hard to stand out.

Martin Durant, Software & Data Engineer

What if you don’t have a background in STEM? Do you need continued education?

The great thing about data science is that many skills you would learn at a university can be self-taught, learned through alternative training, or gained through real world experience. While an undergraduate degree in a relevant field is not a prerequisite, it is important to demonstrate relevant training. If you are interested in data science, you must show initiative and dedication to learning the technical skills listed in the next section. Fortunately, the rise of online training and alternative education has made it easy to find self-serve resources or specialized courses to fit your needs.

These alternative trainings seem to be most appropriate for beginners or those who are experienced but want to learn a new, specific skill. A bootcamp or other condensed training in more advanced areas like machine learning will be more valuable if you have a basic understanding of mathematics and programming first.

> “
>
> I’ve taken courses through Coursera and a few others. Classes in school go more in depth on the fundamentals. Online courses are short and fast; you can learn something practical, like a tool or a language, really quickly. If you’re new to data science or want to learn the basics of a new skill, these online courses are worthwhile to get started. I still learn new things all the time with online training, like through MIT OpenCourseWare.

Sophia Yang, Data Scientist

> I did take two different machine learning and AI courses at Coursera and Udacity. It was valuable to hear expert perspective on the field, and it also confirmed that I was a good fit for the field because of my background in mathematics. I definitely would recommend taking these courses, but you will get more value out of them with a foundation in mathematics. I would warn that there is a danger of choosing a piece of software or tool and running it at your problem without really understanding the mathematics underneath the hood. When that software doesn't produce the results you expect, you should know why.

Michael Grant, Vice President, Services

The following organizations offer a variety of training and certification courses across different skills within the field of data science:

- Kaggle
- Udemy
- Coursera
- Udacity
- General Assembly (and other, similar bootcamps)

In addition to these third parties, tech companies have begun offering their own relevant training and certifications.

- O'Reilly Live Training
- IBM Data Science Professional Certificate (through Coursera)
- Google Professional Machine Learning Engineer
- Tableau Desktop Specialist
- Oracle Business Intelligence Training and Certification

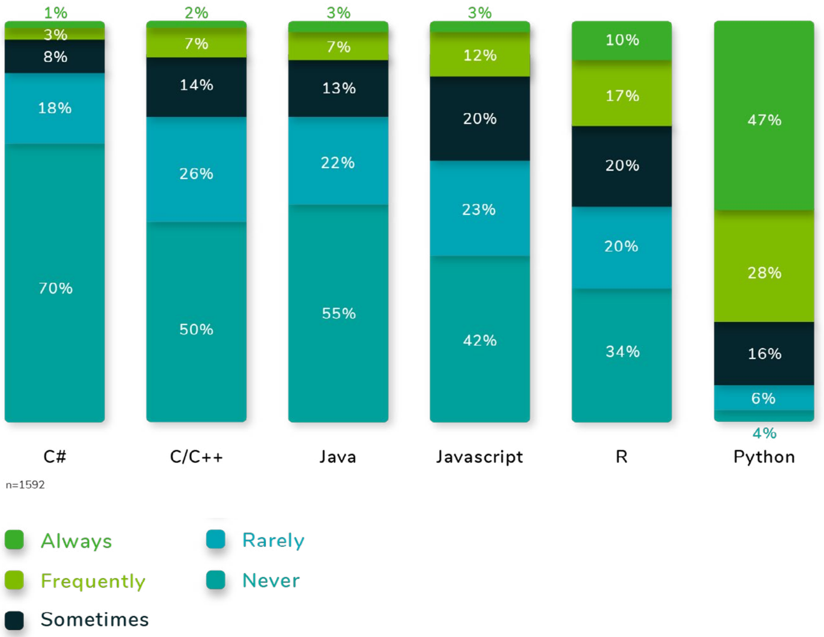
Many universities also offer online versions of their individual courses. Based on the role or career path within data science that you would like, find the education and training that best matches your needs. In addition to courses, data scientists can also learn from their peers through networking in professional groups or at conferences and by participating in open-source communities, which we will discuss further on.

What technical skills are important in data science?

Data science involves multiple disciplines. In our annual report on the state of the industry, we noticed that there are gaps between what enterprises need and what universities teach students. 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.

While data scientists don't need to have deep level mastery of each of these, they should have the fundamentals. From there, you can choose one or a few to specialize in.

HOW OFTEN DO YOU USE THE FOLLOWING LANGUAGES?



n=1592

ENTERPRISES REPORT THAT THEY ARE MISSING KEY SKILLS THAT STUDENTS DON'T REPORT THEY ARE LEARNING, AND UNIVERSITIES DON'T REPORT THEY ARE TEACHING

What students learn	
Python	85%
ML	55%
Data Viz	49%
Probability & statistics	43%
Deep learning	42%

n= 346

What universities offer	
Python	72%
Probability & statistics	59%
ML	54%
Data viz	53%
Advanced mathematics	49%

n=238

What enterprises lack	
Big data mgmt	39%
Advanced mathematics	36%
Deep learning	31%
Engineering skills	26%
ML	25%

n=1216

Programming

Programming allows you to transform data into actionable insights; it is fundamental to the discipline. In our study, Python and R were the most commonly used programming languages. If you're choosing between one, we recommend mastering Python and then adding R to your repertoire.

Though other programming and tools have gained popularity, it is also still expected that data scientists can write and execute queries in SQL. Many of the organizations mentioned in the previous sections offer training on these languages.

> Just because you can get the job done, doesn't mean you're a good programmer. There is a lot to it. For example, version control -- this isn't part of a programming class but is something that is absolutely essential to understand if you ever want to work for a company. The same could be said for documentation, testing, and other things that surround programming. You will become a better programmer by exercising those skills. [...] I think it's worth the effort to become an all around good programmer, even if you don't think that will be the main part of your job.

Martin Durant, Software & Data Engineer

Mathematics

Like programming, mathematics underpins the field of data science. A general understanding of statistics, probability, algebra, and calculus is required. 36% of survey respondents from enterprises said their organization was lacking advanced mathematics skills. Honing these skills could help prospective data scientists differentiate themselves and be better equipped to grasp more complex skills.

> To me, there's a hierarchy between these technical skills. For example, machine learning almost has mathematics and programming as a prerequisite. You're not going to get as far as you should in machine learning without the foundation of math and programming.

Michael Grant, Vice President, Services

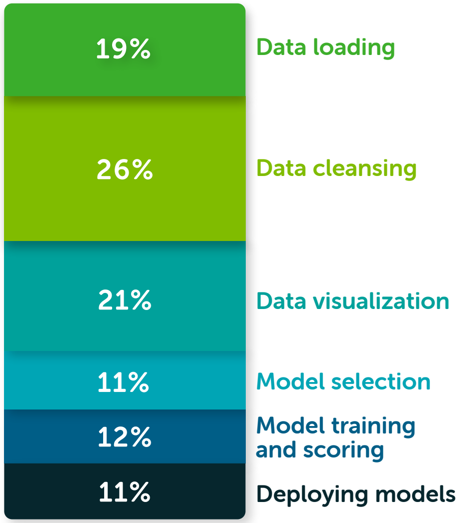
Data engineering

Our data also showed that data wrangling takes up a large part of data scientists' day — almost half of their time is spent on the combined tasks of data loading and cleansing. Considering that data wrangling must happen before data can be used or processed for a project, this is an important but overly time consuming task.

Knowledge in data engineering can help data scientists and their organizations reduce time to value. As mentioned, big data management was the top requested skill from enterprises in our survey. While true data science roles may not need to do their own data engineering or database management, they need to be a good partner to those they're working with who do.

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%



Data visualization

Similarly with database management skills, 24% of survey respondents said that their data science team lacked data visualization skills. Moreover, only 49% of the students who took the survey said that they were being taught data visualization in school.

Visualization is extremely important for being able to represent data effectively. While many data scientists have a basic understanding of visualization, many do not consider themselves experts because the field is too large to master. Data visualization is becoming an extremely valued skill — those that can specialize may stand out from other data scientists.

Machine learning

If working with big data is part of your goal, you will need to dive deeper into machine learning. Machine learning techniques allow you to automatically utilize information in order to make predictions or decisions without being explicitly programmed to do so. Data scientists should be familiar with techniques like supervised machine learning, decision trees, logistic regression, neural networks, etc. As mentioned, it is recommended to have a foundation in math and programming first.

> Producing visualizations is easy, but doing it in a way that means something is much harder. It's hard to teach, hard to learn. You need to understand your audience, and you need to understand your data -- separately from each other. Then, you can use those two together in a magical way to produce something to tie the audience to the data. Being skilled at data visualization is a good way to get a job. How can you show yourself to the world? A good, understandable visualization captures attention.

Martin Durant, Software & Data Engineer

What soft skills will help you excel?

For data scientists, soft skills are just as important as technical acumen. Data scientists can often be isolated within an organization, and the technical nature of their work can be difficult for others to understand.

One concerning finding in our study was that 40% of professionals said they “almost never” or “only sometimes” can effectively demonstrate the business impact of data science within their organization. Without being able to show the impact, data science projects may not ever get off the ground — and those that do run the risk of being deprioritized or canceled. It is critical for data scientists to build relationships with technical and non-technical colleagues alike and be able to clearly communicate their work and its value.

Ability to self-learn

Every person we spoke with acknowledged that at least some of their skills had been self-taught. In a field that is quickly maturing and where there is fluidity between role definitions, data scientists (and those in related roles) have to constantly upskill. While formal and alternative education solves for a portion of this, there is often not an opportunity to fit in training in advance of needing to use a skill — skills are learned on the job.

> Oftentimes, you are assigned a task that you have not encountered before, so you have to learn something very quickly that is likely very technical. You have to be able to understand and implement new skills and be able to explain them to others.

Sophia Yang, Data Scientist

> A data scientist typically works outside of the domain of what a company does. You need to learn enough about the specifics of the domain that you're in to be able to understand how to effectively solve your business' problems and to communicate with stakeholders. When ramping into a new industry, talk with as many people as you can and ask questions, so you that you can learn faster.

Martin Durant, Software & Data Engineer

> Hiring managers can't reasonably expect someone to have exactly the set of skills they're looking for in a job -- the data science industry isn't old enough. What do I look for instead?
> I look for general aptitude in mathematics and programming and a certain demonstration of grit and persistence. If you can show evidence that you have the confidence and general skills to tackle problems you may not have encountered before, then that tells me you can fill in gaps in your abilities in real time.

Michael Grant, Vice President, Services

Related to this skill is the ability to rapidly bring yourself up to speed on any given industry. From marketing analytics to healthcare to food services, data scientist roles can be found in any industry. If you take a position at a new company, you must be able to learn the business, and learn it fast.

Individuals going down this career path should be insatiably curious, driven to improve, and willing to put in the work to teach themselves new concepts.

Relationship building

As mentioned, data scientists can become isolated within their own projects or area of an organization. Relationship building, both with other data scientists and non-technical colleagues, is necessary for data science initiatives to be successful.

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, and 28% are stationed in a centralized data team or Center of Excellence (COE). While we expect more organizations to establish COEs for data science over time, for now, data scientists within an organization will have to make an effort to connect with each other and share best practices.

In addition to the support of your data science colleagues, successful data science projects require the buy in, feedback, and resources from people in different areas of the business.

When working on cross-functional projects, data scientists should spend as much time understanding stakeholders and their needs as you do understanding your data — this is a step that many undervalue. Rather than taking orders and fulfilling requests, turn the process into a dialogue. Having a thorough understanding of how stakeholders will be using your data will help you be a better partner.

> Collaboration and negotiation -- effective team play -- is something that a lot of people who are highly technical struggle with and is extraordinarily important. For me, it has been extremely important because I work directly with customers. I am always having to effectively negotiate, before and after the sale. Even if I wasn't working with customers, I still have to work with my team members and people in other departments. I need their support, and they need mine. Navigating this can go a long way.

Michael Grant, Vice President, Services

> Your ability to manage a good working relationship with product owners is critical, so that you can build the best product possible and everybody wins. I considered it my responsibility to give engineering and product management clear guidance on what was possible, less possible, and just completely impossible. It was important to be able to manage expectations. Sometimes, when I wasn't sure, I would offer to do a 1-2 week proof of concept to flesh it out.

Matthew Brock, Principal Engineer

Communication

Clear, frequent communication is critical to showing the impact of your work. When asked about relevant soft skills, every person we spoke with mentioned its importance. In our survey, nearly a quarter of respondents said that the data science/machine learning area of their organization lacked communication skills.

As a data professional, it is your responsibility to function as a translator, helping others understand the data and how it should be interpreted and used. When it comes to communicating about your data science projects, it is essential to be able to connect your work to larger business goals, or else those you are communicating to will have trouble prioritizing or supporting your project.

Effective communication requires tailoring your communications to your audience: use metrics that your stakeholders care about and language they understand. Use stories, analogies, effective data visualizations, applications, and dashboards to represent your data. Practice explaining complex data or topics with non-technical audiences to refine your skills.

> Data science roles have a somewhat unique position where in some cases the rationalization and explanation are just as important (and possibly more) than delivering the technical solution.

Albert DeFusco, Data Scientist, Product

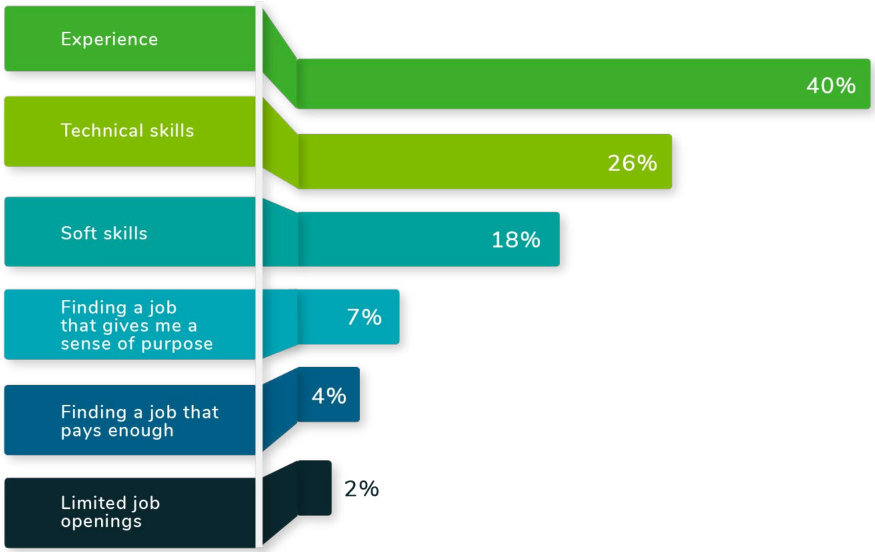
> A critical soft skill is the ability to summarize and advocate for your work to people who might not have the technical knowledge to digest it at the same level. This skill is often underappreciated and can be especially challenging in an international industry. Anyone who can effectively communicate with their colleagues is going to have a leg up over everyone else.

Michael Grant, Vice President, Services

How do you get real world experience?

Education is great, but companies that are hiring for data science roles will be looking for real world application of these skills. This can be a barrier to entry for those just getting started in their career or those who want to pivot into data science from another background. Students that participated in our survey reported that their biggest obstacle to finding a job was experience (41% of respondents). However, real world experience is not necessarily synonymous with full-time employment as a data scientist. There are other ways to show that you can put your training into practice.

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.

How did you transition into data science?

"In my postdoc, I had done a lot of programming to get the job done. That's initially what Python and all the tools were to me. They were a way to get from data to results to publishable articles. I realized I could do something with that because it was so useful in its own merit. I had always tried to produce good code for my own sake, but I noticed that this was not that common of a trait out in the world.
The combination of somebody who understood data, understood the statistics of science, and could actually put these things into code — that's what set me up to be in data science."

Martin Durant,
Software & Data Engineer

"I had been looking around for jobs in the software engineering space for about a year with not much luck. It wasn't easy for me to market myself to those positions from my background in academia and my daily responsibilities. But then I went through a completely left field approach to the industry anyways by coming through training and education. Anaconda gave me an interesting position to teach basic programming and general scientific computing, which I had demonstrable experience in. This gave me the unique opportunity to learn more about data science on the job."

Albert DeFusco,
Data Scientist, Product

"Transitioning to a data science role was a natural fit. I had a basic understanding of data science approaches from exposure in academia and my own experimenting.
It was that in combination with what the specific role required. In this case, it was building a machine learning malware detection product. I checked off all the other boxes in regards to my experience: product engineering, cybersecurity understanding, etc. The role was 90% data engineering and 10% data science.
A lot of my skills in software engineering were ready to rock, and it was really sprinkling a little data science application on top. I was able to learn from the data science team around me; they gave me pointers to fill in the gaps along the way."

Matthew Brock,
Principal Engineer

"I did psychology in college, and then I did a master's in statistics and a PhD in educational psychology. Psychology is actually all about data — before I went into the field, I didn't know that. All I did in grad school was collect data, work with data, find patterns within data, and use statistical models to generate insights from data.
I just fell in love with working with the data, and I wanted to continue doing that after I graduated."

Sophia Yang,
Data Scientist

"I was technically oriented from a young age. I went into college focused on electrical engineering but with a heavy focus on computer hardware and software. My ambition was to become a professor and an academic in electrical engineering. Then, the internet boom happened when I was in grad school, and my plans changed. I discovered I was really good at marrying software development and advanced mathematics together. There are a lot of flavors to that, but it just so happened that my training meshed with machine learning and artificial intelligence well."

Michael Grant,
Vice President, Services

Practicum programs and internships

Participation in internship, practicum programs, and research positions can address this gap for students and early career professionals. When in school, whether undergraduate or graduate, look for relevant internships or participate in offered practicum or research programs that will allow you to practice the skills you're learning in classes. Internships aren't just for undergraduate students either — even if you're already working professionally, taking an internship or pursuing additional education is still a great option and shows that you're committed to pursuing a new career path. When evaluating other educational programs, ensure they include project components to take your training beyond just theory.

Personal projects

Securing employment can be difficult to do, and some aspects are beyond your control. The best way to show your experience with real world application is to contribute to or create relevant projects. Once you have learned the foundational skills in the previous section, try your hand at a few data science projects.

The easiest way to do this is to get involved with open-source communities. Contributing to open-source projects is a great way to hone new skills, make connections with others, and give back to the community. There are a number of open-source communities and projects to choose from. At Anaconda, we are proud to distribute and contribute to a variety of open-source projects.

> My graduate research assistant gig at Los Alamos National Laboratory gave me relevant data infrastructure, data engineering, and high performance computing experience. This was 2005 or 2006 before data science really came into vogue. At the time, the Roadrunner supercomputer was at the top of the charts, and I was lucky enough to be part of that group and run some cool experiments. To be able to see that side of a laboratory and interact with the cluster was a great experience and gave me a sense of complexity, scale, and criticality. It was also a lot of fun.

Matthew Brock, Principal Engineer

If you're interested in open-source projects but aren't sure where to begin, start with existing projects' documentation. Before you get to the point where you're contributing code, you can help improve documentation by finding typos, filling in gaps, and suggesting other improvements. Then, you can try contributing code. Contributing code for open-source projects forces you to write good code because other people will be reviewing and seeing it. You have to write tests, and you have to prove that what you did actually works. All of these things are excellent training in the long run.

In addition to working on others' projects, you can create your own, although this is more challenging. If you're creating an open-source project, it can be hard to get others to work on it organically without a lot of exposure.

If you need data to work with, you can get your hands on a data set through a variety of sources. A number of government agencies, including the U.S. Census Bureau and the FBI, have publicly available data online. Similarly, major companies like Walmart, Google, and more also share some of their proprietary data as well. A quick search online for public data sets will pull up more than enough resources to get started.

> Open source is really easy. There are thousands of projects that are active. Project maintainers will respond to a question or proposed code within a day with helpful feedback in an adult, inclusive way. You can't really find that anywhere else. When you propose code or raise an issue, so long as you are reasonable, you will find a very positive welcoming attitude in open-source communities.
>
> Data science, machine learning, etc. have been born in and pushed forward by open source. It's good to be part of that. When you are applying for a job, you can actually point to something -- it's all public. You can say, 'These are the real, material changes that I've made to these particular projects. These projects are important to the data science and machine learning world.' That's excellent proof to show that you really know what you're talking about.

Martin Durant, Software & Data Engineer

Full-time experience

The end goal is full-time employment as a data scientist. But there are other, related roles that will get you closer to this ultimate goal. Consider roles in engineering, operations, data analysis, or business intelligence — roles that allow you to work with a business' data. When possible, work on projects within these roles that allow you to flex the technical and soft skills required for data science roles.

In addition to roles that are related to data science, you can also apply data science techniques to a non-technical role. For example, if you work in marketing and have taken a few data engineering or programming classes on the side, see if you can apply what you've learned to your existing work. You can then showcase these projects on your resume and in interviews.

> Different companies define data-related roles differently. Different companies have different names for the same role. In my experience, in some cases, there can be a great amount of overlap between data analyst and data scientist roles, for example.

Sophia Yang, Data Scientist

> Ideally, it's more of an evolution than a major jump. You can crawl before you walk before you run. Don't start with the most advanced AI deep learning algorithms. Start out with good data engineering practices, visualization, and reporting before tackling basic machine learning. Find ways to incrementally layer on data literacy and see how far you can go.

Michael Grant, Vice President, Services

Conclusion

Data science is an exciting, constantly evolving field — and one with endless points of entry. No one single path leads to this career. Instead, skills that can be learned through a combination of education and practical experience are the foundation. For those with the right mix of technical acumen and ambition, the data science field is open for exploration.

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.