9 best practices every data science leader should follow

ANACONDA.

What's inside

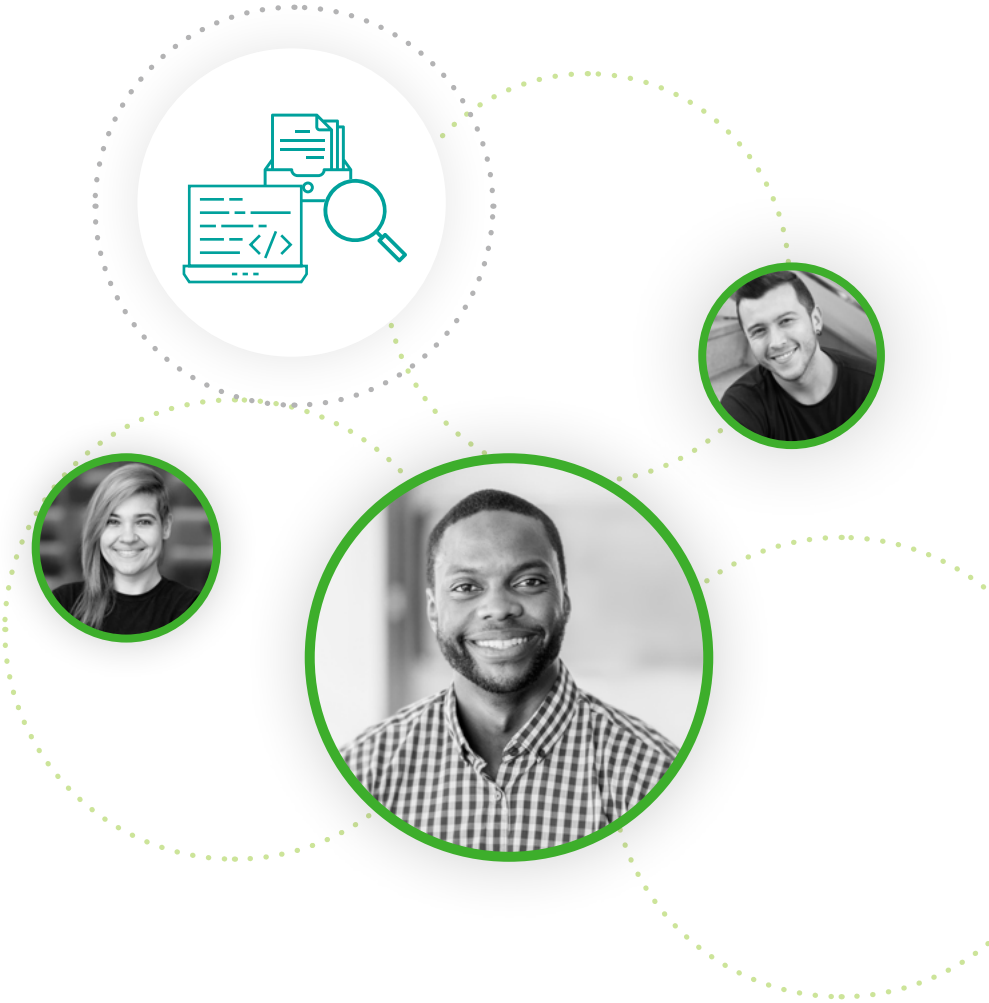
Introduction

Being a data scientist is hard. In addition to the combination of advanced mathematics and coding skills required to do the job, it's a newer role for many organizations, so data scientists are called upon to navigate corporate political landscapes, forage for the right IT resources, and establish new workflows across departments to do their jobs effectively. These best practices will help you be more effective at your job, lead the way for future data scientists, and establish a department that's innovative, productive, and an integral piece of the larger organization.

1 Optimize your use of open source

Data scientists across industries are leveraging the power of open-source data science and machine learning tools. Thousands of dedicated developers, data scientists, and researchers around the world pour their efforts and their genius into these libraries and packages, so it's no wonder enterprises look to open source for innovative solutions.

Because open-source tools are such an important part of the data science technology stack, make sure your hiring criteria reflect this. Data scientists that have already contributed to open-source projects will have a better understanding of how to evaluate and manage open-source tools by looking at code activity, package metadata, release history, and project contributors. They should also understand when and how to make pull requests if packages can be updated, enhanced, or made more secure to meet your organization's needs. In addition to hiring data scientists and developers with open-source expertise, consider working with a vendor that provides support for open-source tools and libraries.



2 Institute a security-aware culture

In our 2020 State of Data Science survey, we asked respondents to rate their level of concern about open-source security. Among data scientist respondents, the average level of concern was 2.9 on a 5-point scale. But when data scientists don't monitor for potential threats, vulnerabilities inevitably creep into models over time. **Data science leaders must step up and collaborate with IT and security leaders to take charge of their data science and machine learning pipelines.**

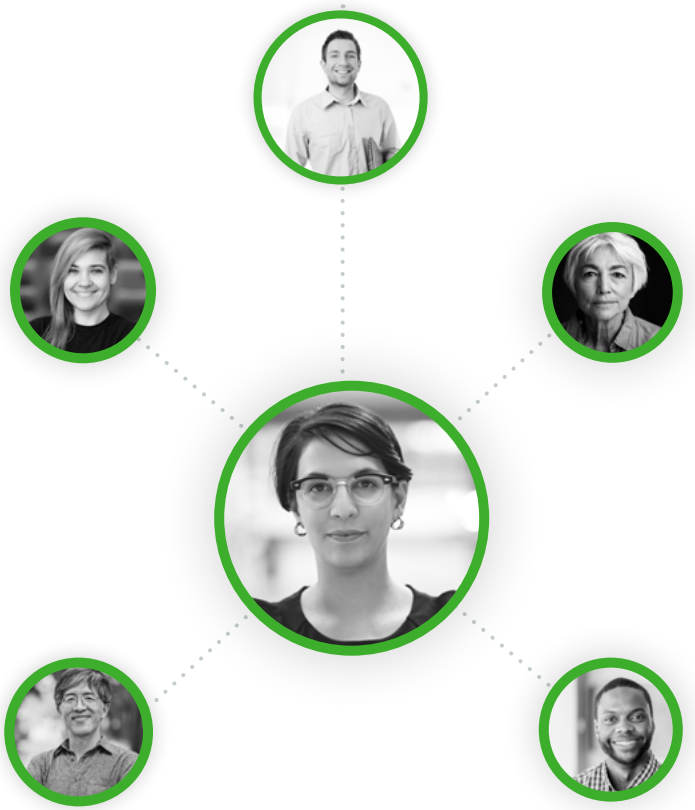
Because these pipelines usually involve the use of open-source libraries, it's important to understand your organization's risk tolerance for open-source software. Learn about Common Vulnerabilities and Exposures (CVEs), how to look for them, and how to monitor your environment for high-risk packages. Ignoring a high CVE score can result in data breaches and unstable applications.

Learn more about how to implement a security policy for your data science team in our Enterprise Guide to a Secure Data Science Pipeline.

3 Devise a team structure that maximizes impact

Many data scientists don't start out on teams, rather they are scattered across the organization and assigned to specific lines of business to solve particular problems. This is usually an effective way to begin implementing data science in an organization, as it's easier to demonstrate business impact with small, focused projects. But over time, data scientists will need to collaborate to develop processes and eliminate redundancy. They will also need to work with IT to understand what it means to put their projects into production, the limits of their resources, and what security standards must be met.

Many organizations have found that a hub and spoke model works best for a data science team. With this model, some data scientists remain within lines of business, while others work in a data science lab or center to help data scientists and analysts across the organization.



Learn about the pros and cons of four different types of data science team structures in our guide, How to Structure a Data Science Team that Clicks with Your Organization.

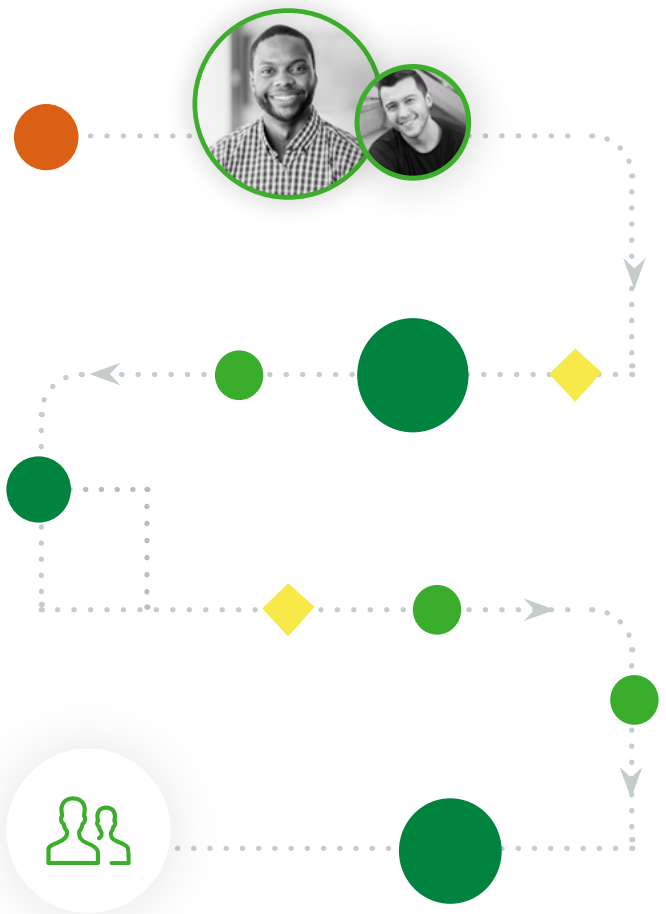
4 Customize presentations by line of business

Whether you choose to have a centralized data science team or specialists working in lines of business, your team should understand how to speak the language of the business units they work with. For example, one data scientist might understand marketing metrics and KPIs while another might understand manufacturing KPIs.

It's essential that common terms and acronyms are used in presentations with their respective lines of business. This will help you establish common ground and help everyone understand whether a data science project was successful and to what degree. It's also a good idea to begin working with lines of business by building custom dashboards that serve their unique needs. Then, refer to these dashboard metrics on a regular basis in joint meetings as new data science project goals are discussed and the effects of decision-making based on model output are evaluated.

5 Bring IT and developers into the POC phase

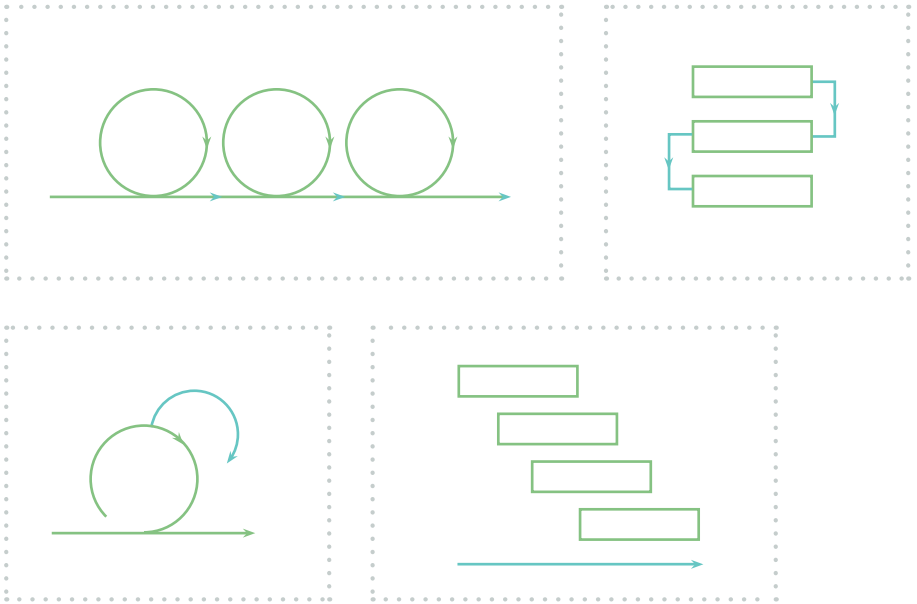
Nearly half of data science projects never make it to production. **One way to help ensure models finally make it into the hands of end-users and bring value to the business is to involve IT and software developers early in the process.** Bring them in to ensure security protocols are met early on. By evaluating software components that will be used to build a model, data will be managed securely and continue to be managed securely once a model is in production. IT teams can also help secure the right infrastructure for model training and production, and developers will help ensure a better end-user experience for the final product.



6 Establish a workflow

Because data science is a new function in many organizations, custom workflows must be established. Many organizations are turning to software development for a workflow model, using Agile principles and Scrum methods for data science output. But this doesn't always work for research-intensive data science. As data scientists know, the steps needed to arrive at the final goal are not always clear. Research and data exploration can yield results that drastically shift the course of a project.

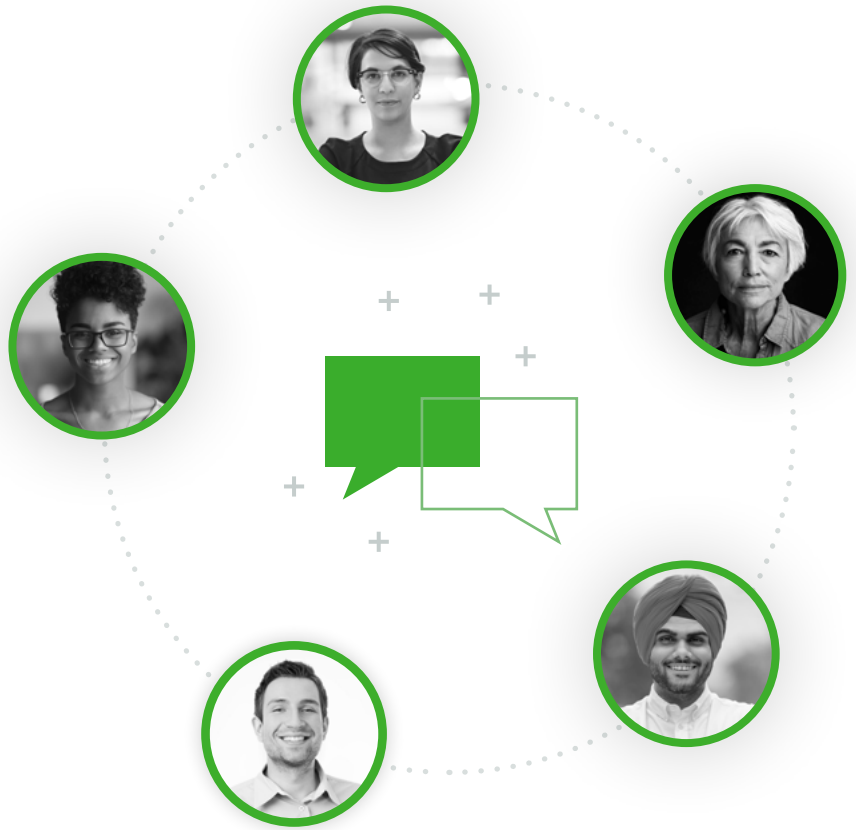
With this in mind, data scientists can still adopt an Agile methodology (especially helpful for data science projects that become web applications) and tweak it to suit their goals and processes.



Keep in mind that projects frequently revert to previous stages and new deliverables can be added in each stage, so keep deadlines soft to allow for changes in course as projects unfold.

7 Encourage collaboration

Whatever the team structure model you choose, encourage collaboration among data scientists and between data scientists and other line-of-business managers. Data scientists scattered across the organization should meet regularly to discuss processes, tools, and projects. Data scientists who are centralized should meet regularly with business managers. Not only does collaboration limit redundancy, but it also makes for a more rewarding work environment. Through regular communication, data scientists will learn more quickly, grow their skill set, make a better case for resources they need, and provide more value to the organization overall.



8 Implement tools that mitigate bias and maximize fairness

Data science and machine learning are increasingly used to help make decisions that impact people's lives through credit scoring, job and college applicant scoring, and even potential healthcare outcomes. When implemented thoughtfully, machine learning can improve human decision-making and reduce racial disparity. On the other hand, when machine learning models are implemented without regard for bias or fairness, they can enforce and exacerbate human biases. This is why ethics should be top of mind for every data science team.

The most important steps data scientists can take are to understand biases in their data and understand how their models make decisions.

Fortunately, several new open-source tools are available to help data scientists do this:

Fairlearn

FairLearn is a Python package that helps data scientists evaluate fairness in models and training data. InterpretML provides insight into the decision-making of black box models and also provides data scientists with the tools to train "glass box" models.

LIME

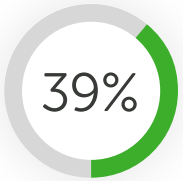
LIME is a PyPI package that is used to explain individual predictions for text classifiers.

Learn more about these tools in Anaconda's Guide to Open-Source Tools and Libraries for Data Science and Machine Learning.

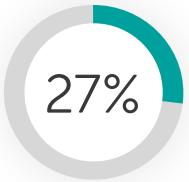
9 Be a voice for ethics and explainability

Of all the trends identified in our 2020 State of Data Science study, we found a lack of action around ethics and explainability to be the most concerning. While these two issues are distinct, they are interrelated, and both pose important questions for society, industry, and academia.

We found that:



of enterprises do not have a plan to **implement solutions for bias mitigation**



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.

So, how can you get started?

Explore the talent within your organization and seek out ethics professionals in other areas. You may find an ethics attorney, managers in ethics and compliance, or even a professional ethicist. Work with others who are ethics-minded to host workshops or seminars that involve senior leadership. Explain the risks and negative impacts of ignoring bias in data and models, and lead the charge for your department and organization.

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.