



 snowflake® DataRobot

# FIVE THINGS A DATA SCIENTIST CAN DO TO STAY CURRENT

Tips for evolving in the ever-expanding world of data science



CHAMPION  
GUIDES

# TABLE OF CONTENTS

- 2** Data scientists are in high demand
- 3** Accelerate AI adoption with automation and operationalization
- 3** Tip #1: Use AutoML frameworks to boost productivity
- 4** Tip #2: Adopt MLOps to operationalize AI
- 5** Advocate for better data management and access to new data
- 5** Tip #3: Promote data consolidation for workflow efficiencies
- 5** Tip #4: Unlock your data for greater insights
- 6** Tip #5: Collect more data and encourage more ideas
- 7** Become an automation and data champion
- 8** About Snowflake

# DATA SCIENTISTS ARE IN HIGH DEMAND

Data scientist continues to be one of the hottest jobs in the labor market. Since 2012, the number of different data science roles has increased by over 650%.<sup>1</sup> In 2018, LinkedIn reported that the “data science manager” and “data science specialist” roles have experienced four and five times growth, respectively, between 2012 and 2017.<sup>2</sup>

Any organization that has tried to hire a data scientist in the last five years knows just how hard it can be. And the numbers demonstrate why: In just one year, from 2018 to 2019, demand for data scientists increased 56%.<sup>3</sup> Data scientists have a unique blend of skills that brings together computer science, statistics, modeling, mathematics, and business acumen, which represents a tough skill set to find.

The demands on data scientists are also increasing. Organizations want to inject artificial intelligence (AI) and machine learning (ML) into critical business processes to remain competitive. Data scientists often find themselves with more to do than is humanly possible.

The good news is that technology advancements are providing exciting ways to automate ML and implement better data management techniques. However, data scientists must stay on top of these trends in order to learn best practices and promote their usage.

This ebook provides five tips for how data scientists can stay current while excelling in an ever-changing, AI-driven workplace. Although adopting new technologies and processes can be time-consuming and daunting, by doing so, data scientists will quickly discover more time for the tasks that demonstrate their unique value-add.



<sup>1</sup> [forbes.com/sites/louiscolombus/2017/12/11/linkedin-fastest-growing-jobs-today-are-in-data-science-machine-learning/](https://www.forbes.com/sites/louiscolombus/2017/12/11/linkedin-fastest-growing-jobs-today-are-in-data-science-machine-learning/)

<sup>2</sup> [economicgraph.linkedin.com/research/linkedin-2018-emerging-jobs-report](https://www.economicgraph.linkedin.com/research/linkedin-2018-emerging-jobs-report)

<sup>3</sup> [cio.com/article/3397137/6-ways-to-deal-with-the-great-data-scientist-shortage.html](https://www.cio.com/article/3397137/6-ways-to-deal-with-the-great-data-scientist-shortage.html)

# ACCELERATE AI ADOPTION WITH AUTOMATION AND OPERATIONALIZATION

Recent developments in technology and organizational structure can ease the challenges data scientists face. By adopting automation-first technology and promoting the operationalization of ML models, data scientists can accelerate their own productivity and discover more time for the highest value-add tasks.

## TIP #1: USE AUTOML FRAMEWORKS TO BOOST PRODUCTIVITY

Automated Machine Learning (AutoML) represents an opportunity to transform how data science and ML work together. As the name implies, AutoML platforms automate the tasks associated with developing and deploying ML models.<sup>4</sup>

These platforms standardize and democratize data science best practices, optimize and train data across a multitude of algorithms, and accelerate tasks that are extremely time-consuming and require vast amounts of knowledge, such as feature engineering. AutoML also helps data scientists avoid human-based errors that occur during manual data modeling processes, which ultimately makes the entire process more accurate and less prone to bias.<sup>5</sup>

By automating data science processes and ultimately deploying the most powerful ML models, data scientists can save a significant amount of time and effort during data modeling and they can achieve

faster insights. For example, data scientists typically perform feature engineering and then run a data set against the three to five algorithms they are most familiar with. This process can take more than a month to create a few models. In contrast, AutoML platforms can run the same data set against hundreds of different algorithms in parallel, with unique feature engineering optimized for each algorithm, and complete all the permutations and iterations to create 80 to 100 models automatically in less than half an hour.

The result is a phenomenal boost in productivity. Automation means more time to focus on complex business problems and creative solutions rather than manual modeling. Once the automatic modeling is done, data scientists can use their business intuition to improve or tweak hyperparameters of models as they see fit. Better yet, they can transform or combine model features or find third-party data to supplement the training data sets (which is an area where machines currently fall short).

As new algorithms with better capabilities and performance are created and refined, AutoML platforms provide data scientists with quick access to cutting edge algorithms without requiring them to study and master each new algorithm. The most up-to-date best practices are automatically built into the AutoML platforms, with guardrails in place so novice or citizen data scientists can't forget to complete a critical step in the process.

<sup>4</sup> [infoworld.com/article/3430788/automated-machine-learning-or-automl-explained.html](http://infoworld.com/article/3430788/automated-machine-learning-or-automl-explained.html)

<sup>5</sup> [heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f](http://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f)

## TIP #2: ADOPT MLOPS TO OPERATIONALIZE AI

Global spending on AI and ML technologies is predicted to hit \$37.5 billion in 2019 and accelerate to \$97.9 billion by 2023, with a compound annual growth rate of 28.4%.<sup>6</sup>

However, this accelerated spending will be for naught if organizations cannot operationalize these technologies. According to Gartner, only 47% of ML models actually go into production, which represents a huge gap between data science resources and actual business value and impact.<sup>7</sup>

Simply put, AI requires resources beyond data scientists in order to be effective. Operations teams must be involved to help with critical operational and production tasks such as monitoring, alerting, upkeep, and compliance. The practice of DevOps provides a model for how collaboration between two teams (in this case, development and operations) accelerates production and minimizes errors.

The same holds true for AI and ML. Referred to as MLOps, or machine learning operations, this new practice promotes the joint management of the ML data pipeline by bridging the gap between data scientists and operations teams.

By building a collaborative partnership between these two teams, those who are in the appropriate roles can manage tasks that fall naturally within

their realm. For example, ML operations teams can oversee the systematization, regulation, usage, and deployment of AI in production. They can monitor how models are performing and alert data scientists to data drift of key inputs to the model. This type of support frees data scientists to focus on business issues and deal with data preparation and ML models to provide faster insights.

To ensure success with MLOps, it's important to establish rules of engagement between the teams, set measurable and joint KPIs, and determine the best ways to comply with regulations, especially around

consumer data.<sup>8</sup> Neither team can, or should, manage AI usage alone; organizations will achieve better results when the two teams marry their knowledge and strengths to operationalize the process.



<sup>6</sup> [https://www.idc.com/getdoc.jsp?containerId=IDC\\_P33198](https://www.idc.com/getdoc.jsp?containerId=IDC_P33198)

<sup>7</sup> [informationweek.com/big-data/ai-machine-learning/getting-machine-learning-into-production-mlops/d/d-id/1335050](https://informationweek.com/big-data/ai-machine-learning/getting-machine-learning-into-production-mlops/d/d-id/1335050)

<sup>8</sup> [medium.com/@ODSC/what-are-mlops-and-why-does-it-matter-8cff060d4067](https://medium.com/@ODSC/what-are-mlops-and-why-does-it-matter-8cff060d4067)

# ADVOCATE FOR BETTER DATA MANAGEMENT AND ACCESS TO NEW DATA

Cleaning and organizing data takes up an inordinate part of a data scientist's day. Worst of all, 78% of data scientists view it as the least enjoyable part of the job.<sup>9</sup>

While wrangling data is a natural component of the role, it doesn't need to be so tedious. With advancements in cloud-built data platforms and new capabilities around secure data sharing, data management options exist that decrease the time required to discover and normalize data and increase the opportunities for access to new data sources.

## TIP #3: PROMOTE DATA CONSOLIDATION FOR WORKFLOW EFFICIENCIES

Obtaining data is the first thing data scientists do. Remarkably, this fundamental requirement points directly to two of the biggest data challenges faced today: Data generated by disparate sources is often stored in separate silos and data exists in formats that are challenging to combine.

As a result, data scientists must undertake the time-intensive task of discovering and gaining access to data before pulling it together and cleaning and normalizing it to create a consistent and manageable data set. Not only is this process inefficient but it also becomes an ongoing issue when more data must be retrieved to complete a model.

Because of the daily challenges with this data quagmire, data scientists are the perfect spokespersons for data consolidation. By encouraging organizations to invest in a data platform where all data (structured and semi-structured) can be consolidated properly and made accessible from a single source, workflow efficiencies can be achieved not only for data scientists but also for data analysts, business users, executives, and even auditors.


Additionally, a single data location protects organizations against the use of incomplete data sets or bad information. When error-riddled or redundant data is corrected, everyone in the organization benefits from access to cleaner data. It's also easier for teams to communicate when date definitions are consistent and commonly understood.

## TIP #4: UNLOCK YOUR DATA FOR GREATER INSIGHTS

Many organizations today have the equivalent of an iceberg of data: Only a small subset of data is visible, while most data exists "underwater" and therefore remains unexamined and unused. It's important for data scientists to ensure their organizations process as much of their data as possible.



<sup>9</sup> visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower\_DataScienceReport\_2016.pdf



Invisible data is especially prevalent for organizations that use data lakes for long-term storage. Using a data lake for long-term storage puts your data lake at risk of turning into a data swamp, leaving data scientists with a mass of disorganized, uncurated, unmanaged data that is outdated and unsuitable for use in production.

In fact, the best way to ensure light is shed on “dark data” is to make data easily discoverable, accessible, and usable. In some cases, this may mean foregoing typical data governance, metadata, and automation processes and marking the data as “experimental” or “raw.”<sup>10</sup>

Chances are, there's real value in huge data sets. Companies can benefit by setting up self-service data refinery platforms where employees can refine data and share it with each other, further refining and sharing their joint insights.

Sharing data with trusted partners, suppliers, customers, and vendors is another way to derive value from internal data. For organizations that are interested in monetizing data, marketplaces are a natural option, because other companies may be willing to license and pay for anonymized, aggregate data.

### **TIP #5: COLLECT MORE DATA AND ENCOURAGE MORE IDEAS**

The flip side of sharing data is bringing new data in. Organizations rarely collect all the data they need for analysis, and data scientists should not feel constrained to use only the data they have within their own virtual walls.

Data gathering should be part of every organization's data strategy, and data scientists are in the perfect position to push open new data “doors.” A huge value-add for any organization is a data scientist who thinks big and broadens horizons by asking, “What other data could I use?”

Mature data companies such as Netflix and Uber are more likely to have data scientists who are accustomed to this mindset or even to have dedicated “data hunters.” “More data from more sources” is a refrain that every data scientist should adopt. There's no harm in reaching out to other companies and asking if they would share or exchange their data or, better yet, telling them that you want to buy their data.

Of course, data scientists don't need to lead the charge alone. Another idea gaining momentum is the use of a center of excellence (COE) for data, analytics, and data science.<sup>11</sup> By centralizing resources, organizations create efficiencies and discover new ways for teams to collaborate around data, analytics, and business questions. This type of setup brings cognitive diversity to the conversation, which helps to push boundaries and unveil new areas for exploration. It also helps establish consistent best practices to help new users get onboarded quickly and with confidence.

<sup>10</sup> [information-age.com/data-swamp-data-lake-123481597/](https://www.information-age.com/data-swamp-data-lake-123481597/)

<sup>11</sup> [cio.com/article/3391579/analytics-success-starts-with-a-center-of-excellence.html](https://www.cio.com/article/3391579/analytics-success-starts-with-a-center-of-excellence.html)

# BECOME AN AUTOMATION AND DATA CHAMPION

The future for data science is full of opportunity. Data scientists represent many of the brightest and most curious minds at work across dozens of industries, which demonstrates why demand is high and many of the most business-critical challenges fall under their purview.

With this momentum, now is the perfect time for data scientists to take the lead around AutoML, MLOps, and best practices for data management and usage. Every organization wants and needs to extend its competitive advantage and improve its ability to deliver data-driven insights, and data scientists are the natural champions for promoting increased efficiency and faster results.

Find out more at [Snowflake.com](https://www.snowflake.com) and [DataRobot.com](https://www.data-robot.com).





## ABOUT SNOWFLAKE

Snowflake's cloud data platform shatters the barriers that have prevented organizations of all sizes from unleashing the true value from their data. More than 2,000 customers deploy Snowflake to advance their businesses beyond what was once possible by deriving all the insights from all their data by all their business users. Snowflake equips organizations with a single, integrated platform that offers the only data warehouse built for the cloud; instant, secure, and governed access to their entire network of data; and a core architecture to enable many types of data workloads, including a single platform for developing modern data applications. Snowflake: Data without limits. Find out more at [snowflake.com](https://www.snowflake.com).

