



# DATA SCIENCE BEST PRACTICES FOR HEALTHCARE AND LIFE SCIENCES



CHAMPION GUIDES

EBOOK

# TABLE OF CONTENTS

- 3** Introduction
- 4** 1. Define business goals
- 5** 2. Get your data in one place
- 6** 3. Prioritize data security and governance
- 7** 4. Create an effective data science organization
- 8** 5. Use the right tools
- 9** 6. Secure a data cloud platform designed for collaboration
- 10** 7. Avoid vague data semantics
- 11** 8. Practice model validation
- 12** 9. Leverage predictive analytics
- 13** About Snowflake



# INTRODUCTION

Data science has the potential to transform nearly every industry, and perhaps no other sector stands to benefit more than healthcare. Yet for many reasons—including the sensitive nature of health data and its associated security needs—healthcare has been slower than other industries to reap the benefits of cloud-based ways of working, let alone multi-cloud data warehousing. Yet there are so many untapped benefits for the industry if and when it fully embraces digitization. Regardless, effective and efficient data science approaches require a fast, easily scalable solution/technology made possible by moving to the cloud.

The good news is that things appear to be shifting. According to [Accenture](#), “[A few] years ago, healthcare organizations were experimenting in the public cloud. Now their focus is on migrating legacy technology to the cloud and transforming data and applications.” But, as [reported by Cloud News](#), there are significant challenges ahead, including data integration, application mobility, and, of course, rising costs.

To help healthcare and life sciences organizations prepare for a more data science-centered approach, here are nine best practices designed to help ease your journey and address any nagging misgivings.

# 1. DEFINE BUSINESS GOALS

As with any technology adoption, “why” a company is planning to use this technology is just as important—if not more—than the “how.” It is crucial for organizations of any kind, including those in healthcare and life sciences, to clearly define the problem they are trying to solve and how they’ll do it. The last thing companies want is to **deploy models that fail to address** their goals or serve their needs.

Among the questions IT decision-makers might ask to ensure a successful data science project are:

- Are the company's data problems well defined?
- How will we measure success and impact?
- How invested and confident are stakeholders that this project will provide a measurable return on investment?

To help answer some of these questions, a well-established **CoE** focused on data and analytics is highly beneficial. A central hub that houses data and analytics knowledge and initiatives can disseminate knowledge throughout an organization. This can accomplish two things at once: 1) showcase the value of the insights available through the use of data cloud technology, while 2) helping to develop a data-first culture that can serve as a model for the entire company.

But above all, a dedicated data-centric CoE should help a company develop deeper expertise that can be translated to other departments over time.



## 2. GET YOUR DATA IN ONE PLACE

Siloed data is a common challenge for many industries, **particularly security conscious-industries like healthcare and life sciences.** Yet, having all of the right information in one place—in a way that allows decision-makers to actually use it—is the only way that data science can fulfill its potential. For example, a more streamlined data organization can help doctors share patient information more seamlessly, reducing paperwork, bureaucracy, and the potential for errors.

On a macro level, data that is centralized and usable can enable healthcare and life sciences organizations to spot trends, identify challenges, reduce inefficiencies, and become more proactive and dynamic enterprises. This should result in markedly improved healthcare for the general population, while allowing for advanced research and innovation among healthcare practitioners.

The challenge is that many organizations, especially those that predate the software and computing era, have collected decades' worth of data in a wide variety of formats. For example, general practitioners collect basic patient health information, specialists maintain their own separate records, and lab networks gather bloodwork data. Meanwhile, a range of divisions within hospitals collect chart information, radiology images, and so on, while insurance companies separately house records for medical procedures and costs.

**This highly fragmented environment represents a massive challenge for today's healthcare leaders.**

The continued surge of new data sources - from artificial intelligence, wearable technology, IoT, analytics, and others - is also an ever-increasing challenge for life sciences organizations that need to leverage all of their data in near real-time to power better clinical and business decision-making and outcomes.

That's why consolidating data through a platform that supports structured data (preferably in relational tables), semi-structured data from a variety of sources, and unstructured data such as image files and PDF documents is so critical. Thankfully, a growing number of modern cloud data platforms are designed to easily make sense of large and diverse volumes of data.

### 3. PRIORITIZE DATA SECURITY AND GOVERNANCE

Protecting a company's data is of utmost importance for nearly every industry, but it can be a matter of life and death for healthcare organizations. Cybercriminals exploit this reality for financial gain, making the industry a target. Last year alone, healthcare organizations experienced security breaches that impacted **45 million people**. While life science organizations are predicted to lose USD 642 billion globally due to direct cyber-attacks by 2024, according to an Accenture report.

In addition to the sensitive nature of healthcare and clinical data, the industry faces a myriad of regulations for non-sensitive data collection and storage. Not only do companies need to navigate the European Union's General Data Protection Regulation (GDPR), but they must also comply with Health Insurance Portability and Accountability Act (HIPAA) laws. Data governance is inherently complex, especially in this case, when sources likely vary in type, location, and rate of change.

That's why protecting data is at the top of any healthcare or life sciences organization's list when making decisions about how to apply data science and which partners to choose for storage and warehousing.

Healthcare IT leaders must consider many factors when evaluating partners. For example, platforms should include access-control measures and encryption by default. Partners also need to enable companies to directly control authentication and the use or sharing of login credentials.

In addition, security concerns may extend beyond a company's owned-and-operated products to applications built on top of its software. In those cases, the same level of security controls must apply to any outside vendors or tools.

Of course, the promise of data science is that once a company's information is better organized and centralized, department leaders can use it to gain and share insights throughout the organization. Security needs must be balanced with use. The best data platforms make it easy to securely access and share sensitive data.



## 4. CREATE AN EFFECTIVE DATA SCIENCE ORGANIZATION

The right tools and technology will only take an organization so far. Healthcare and life sciences companies must also build strong data science teams and organize them in a way that promotes effectiveness.

Finding and installing top data science talent is rarely easy. The best way to start is for healthcare and life sciences organizations to answer a few key questions: “What is the current state of our company’s data?” and “What do we really need data scientists to do?” This fundamental assessment can help steer companies toward better staffing decisions and accept that **some talent will be a better fit than others**, based on the companies’ unique challenges.

Once companies have the most qualified people in place, they need to position their new hires for success. It’s often best to begin with a relatively simple challenge to “get a quick win” while working through any potential kinks. It’s also helpful to track progress through a metrics dashboard that provides relevant data for all stakeholders. This provides individuals with clearly defined benchmarks while gaining buy-in from the C-suite.



## 5. USE THE RIGHT TOOLS

Data science challenges are unique, and each company and team will have different needs and ways of measuring impact. It's crucial to identify the best tools that fit a team's skill set, composition, and goals. Often, the best way to ensure a company is using the right tools is to provide options; some companies will seek out a range of best-in-class data providers and cloud vendors to obtain the deepest and most valuable insights; others will look for a provider that can deliver the smallest set of tools that suit their unique needs.

Of course, a long list of potential partners can be overwhelming. When making partnership decisions, companies should seek partners that have been validated by trusted third parties and have proven industry track records. Regardless of which tools a healthcare or life sciences company chooses, it's important that tools can be integrated to avoid exacerbating their data siloing problem rather than solving it.

In addition to breadth, life science and healthcare organizations need solutions that are agile, scalable, and flexible along their data science journey, as they may still be figuring out their needs and the level of service or sophistication required.



## 6. SECURE A DATA CLOUD PLATFORM DESIGNED FOR COLLABORATION

Change is constant in nearly every industry. When a company adopts and implements a cloud-based, data-centric set of solutions, they are never “done” with their data science mission. Market dynamics, regulations, consumer attitudes, technology, and other variables have the potential to shift dramatically.

For example, the health care industry saw its data collection needs suddenly surge at the outset of the pandemic. At the same time, life sciences organizations needed to dramatically accelerate their work, and demonstrate agility and flexibility as the world sought more rapid developments of treatments and vaccines.

In this new era, the expectations have been raised for both groups, who will be expected to operate in a far more data-driven, high-speed fashion to more quickly and effectively address changing population health needs.

When a company decides to entrust its valued data with a partner, it's crucial that the partner is designed to allow for shifting workflows and can work across clouds and regions. There is great benefit in cloud data platforms designed with both pricing and usage flexibility built in. Another factor to consider is consistency in security governance offerings because usage volume or needs can change monthly or annually. The best partners are also typically able to strike a balance between accuracy and explainability.

Ideally, the right data cloud platform is designed to streamline a company's data science requirements over time, including machine learning, AI, and predictive application development for any level of analytic workload.



## 7. AVOID VAGUE DATA SEMANTICS

When a data science professional is looking to “train” a new model, the choices made up front regarding what features to emphasize and what data points are most important to extract, can make or break the process. Data scientists can easily misinterpret a column’s meaning in a table if the columns are incorrectly labeled, for example statistical data and health survey data. Misinterpreting data is a common problem that can lead to algorithmic bias and erroneous predictions. **Data preparation during modeling** can solve this problem.

Creating a data dictionary can help businesses avoid headaches resulting from poor data semantics. A data dictionary provides context and the proper meanings of different columns in a database. Companies can also create a feature library. Since data scientists often create new features from the existing ones to improve the predictive ability of machine learning models, it is crucial to store a record of newly engineered features for future reusability.

## 8. PRACTICE MODEL VALIDATION

The more that life sciences and healthcare organizations embrace data science, the more likely they are to incorporate machine learning and AI products to maximize their data-driven approach. These are not “set it and forget it” products today. Rather, these tools require a certain amount of real-world testing and training to **ensure they are working properly and generating the expected results.** As **The Wall Street Journal** explained, “There is no shortage of hype about artificial intelligence and big data, but models are the source of the real power behind these tools.”

That’s where model validation comes in, which allows companies to make adjustments and mitigate risks when utilizing such powerful computing technology. This process can ensure that these models correctly employ data they have not seen before and can be relied upon for future use. Among the key reasons companies need to validate their models is to challenge and **test their assumptions, for example, whether a particular drug will generate the desired outcome in a target patient population.** Beyond that, they need to be clear on what they’re trying to learn, what success looks like, and most importantly, that they are inputting the right data. “The most crucial component of an effective machine learning model is a unique and diverse dataset, combined with a unique model of reinforcement learning,” reported **Forbes**.

When evaluating potential partners, companies must ask questions to ensure that the data platforms they are considering support validation and can take action to refine models when issues arise. Among the key questions to ask are:

- **How do you support restoring data that has been transformed or changed?**
- **How can I compare data today vs. a prior point in time?**

Besides choosing the right technology, companies should take every measure to **ensure that the data they use in testing is clean and validated.** Flawed data can lead to flawed testing. In some cases, overfitting can result in models that perform well during the validation process but end up failing in the real world.

## 9. LEVERAGE PREDICTIVE ANALYTICS

Healthcare and life sciences organizations have been collecting some form of patient and research data for decades, but that doesn't mean they've been actually leveraging this information to help drive better patient and clinical outcomes. That's especially true in healthcare, where most organizations have focused on maintaining electronic health records, missing opportunities to use data to uncover patterns and trends that are likely to manifest in the future.

Today, the most sophisticated data platforms are designed to help companies shift from descriptive to predictive analytics. Instead of viewing a dashboard that tells a CEO what happened in the past, these tools can predict and analyze future outcomes and conditions. This may dramatically change how companies operate and anticipate disruptions. As a result, it is incumbent upon technology decision-makers to drill into a potential partner's predictive capabilities. Ultimately, healthcare and life sciences organizations should aspire to ascend the analytics maturity curve to a place where decisions can be optimized over time.

Speaking of maturity curve, machine learning is still in its earliest stages of development. Because of its nascency, companies and industries are adopting this tech at varying levels and speeds, leading to growing fragmentation. Therefore, as the market promises to sort itself out over time, it's important to ensure that various tools and products feature interoperability at their core.

According to Forrester, most companies bought into machine learning's promises but are struggling to operationalize it. The sheer number of new solutions in the market makes that even more challenging. Many businesses "have identified lack of collaboration, AI skills, and cohesive technology toolchains as their top challenges in expanding ML capabilities." The more that companies are able to keep their options open and foster collaboration and cohesion among ML partners, the more effective at maximizing the use of data.

Clearly, there are a number of factors to consider when using data science to create better business and clinical outcomes. For the healthcare and life sciences industries, these are not minor considerations given its unique inherent risks and regulatory requirements. That said, the payoff for them is the ability to harness their data and apply it across the many aspects of the business is immeasurable. The best way to get started is to turn to experienced experts in this arena who are well-equipped in this highly specialized and nuanced field to address an organization's needs and unlock its full potential safely and securely. For more information on how you can use the Snowflake Healthcare & Life Sciences Data Cloud to unlock the power of data, visit [Snowflake for Healthcare Care & Life Sciences](#).



# ABOUT SNOWFLAKE

Organizations use Snowflake's Data Cloud to unite siloed data, discover and securely share data, and execute diverse analytic workloads across multiple clouds and geographies. Organizations, including 510 of the 2022 Forbes Global 2000 as of July 31, 2022, use Snowflake Data Cloud to power their businesses.

Learn more: [snowflake.com](https://www.snowflake.com)



© 2022 Snowflake Inc. All rights reserved. Snowflake, the Snowflake logo, and all other Snowflake product, feature and service names mentioned herein are registered trademarks or trademarks of Snowflake Inc. in the United States and other countries. All other brand names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective holder(s). Snowflake may not be associated with, or be sponsored or endorsed by, any such holder(s).