comet

How to Scale Today’s Data Science Initiatives

How Data-Driven Organizations Leverage People, Processes, and Technology for Success

Introduction

Companies are exploring machine learning (ML) as a tool to automate and refine data analysis, predict customer churn, get ahead of evolving malware, make recommendations to customers, detect fraud, automate document creation, and perform many other tasks quickly and accurately. And according to a McKinsey & Company survey, 50 percent of companies have adopted artificial intelligence (AI) in at least one business function, with adoption highest for service operations, product or service development, and marketing and sales.

However, while ML and AI get a lot of the headlines—and rightfully so—they are themselves part of a broader discipline: data science. Simply put, the goal of data science is to deliver business value from data. ML leverages data science techniques and models to help computers learn from data without having to be programmed to do so. Companies needn't try to jump straight to ML or AI to obtain operational value from their data. In fact, it's possible to gain excellent insights from a report or a spreadsheet, if the data scientists have done their jobs right.

> ***While ML and AI get a lot of the headlines—and rightfully so—they are themselves part of a broader discipline: data science. Simply put, the goal of data science is to deliver business value from data.***

For many organizations, a key use case for data science is leveraging data to support business decision-making. And one of the biggest challenges is democratizing access to data so decision-makers across the organization have access to a single source of truth in which to base conclusions and actions. Even on a limited scale, that can be a tall order for most organizations. To deliver true ROI on data science, machine learning, and AI investments, organizations need the ability to scale data-driven initiatives and deployment across the enterprise—across projects, teams, stakeholders, and infrastructure.

It's easy to create a big vision for the possibilities of leveraging ML and other data-driven initiatives, but scaling that vision across the organization doesn't happen overnight. While many organizations succeed at collecting and analyzing data, many more experience challenges when going beyond their pilot projects—primarily in making data consistent and accessible across the entire enterprise, including projects, teams, stakeholders, and infrastructure. That's why it's important to start small and ensure that your processes support your scaling goals.

The purpose of this eBook is to help you lay a solid foundation for successful data programs that scale beyond a single project, department, or workgroup.

To get started, first find your North Star

One common way data science initiatives veer off course relates to a lack of clarity around the strategic business goal—and not having key metrics in place against which to measure that goal. Many organizations want their first initiatives to make a big splash and show improvements across the board. However, the best place to start is to identify one "North Star" metric that directly affects a key business driver, such as increased revenue. The North Star should be the measure that helps you gauge the effectiveness of your experiment.

The North Star will depend on your industry, your organization, and its goals, and it may change as your organization's goals change. Let's say you're in the business of delivering a SaaS tax preparation application. Your working hypothesis is that increasing the speed at which users can complete their tasks will make your service more attractive to new customers and therefore increase revenue. The North Star metric is, then, the average time it takes for users to file their tax returns.

> ***Many organizations want their first initiatives to make a big splash and show improvements across the board. However, the best place to start is to identify one "North Star" metric that directly affects a key business driver.***

Once you've identified a North Star metric, the next step is to verify if optimizing for this North Star metric will affect the business KPI that you care about. In our example above, the goal is to optimize the speed at which users can complete their tasks in order to increase subscriptions and revenue. Your first set of tests need to establish whether you've selected the correct North Star metric to begin with. You might find that optimizing this metric doesn't do anything at all for your business objective.

To adopt data-driven practices and ML, you will need systems and processes that allow you to effectively run and communicate experiments. The nature of these experiments will vary across business functions. Teams that are only interested in how product changes affect their KPI will need to run experiments like A/B tests and analyze the resulting data from these experiments. Teams interested in using ML will have to run experiments related to their models, as well as how the model outputs affect the North Star metric and business objective.

Getting started involves:

- Identify a KPI of interest, such as revenue
- Select a "North Star metric" that affects the KPI
- Verify that the North Star metric actually affects the KPI

Time to scale up, but there's a problem or two

When it comes time to scale your data initiatives beyond an initial project or small group, organizations often experience problems. If you don't have a good process for defining metrics, collecting data, and ensuring data integrity, it will be extremely difficult to get good results from your data program.

Optimizing your North Star metric generally happens naturally, but it can quickly get out of hand. The answer to one question can lead to 10 more questions because, in our tax example, there are myriad factors that affect the user experience (UX)—page load time, placement of controls and action links, availability of help, even the color and fonts of UI elements. By varying these UX elements, you can find new metrics that directly affect the original North Star metric. As the number of experiments grows, you may find you need a different view of the data or data not previously collected, or you may be faced with a brand-new initiative with additional data needs. All of these may require a change in the data collection process to get the data you need.

This is where scaling data-driven initiatives gets tricky. As you try to allow larger groups of people to access the data to initiate more experiments across additional business functions, you run into new challenges. For example, having a unified strategy may become difficult because teams tend to operate with their own set of assumptions and hypotheses. And there can easily be a lack of communication (or miscommunication) about how teams receive data. At some point, everyone in the organization that's drawing on the same data for insights needs visibility into a shared understanding of how data is sourced, transformed, and distributed within the organization.

Next, we'll explore how the right team structure, infrastructure, and tools affect scalability.

Organize your team

Effective data science starts with the strategic organization of people. The basic structure for data science teams can be centralized, decentralized, or a combination of both. Smaller organizations will naturally adopt a centralized structure. For larger companies or start-ups expecting big growth, the ideal structure is less clear.

> ***Effective data science starts with the strategic organization of people.***

In a centralized model, all or nearly all data professionals work in one team to service multiple business groups. The advantage of a centralized team is that you can assemble all the required skill sets needed to develop and scale data initiatives in one place. These include data validation, modeling, insight gathering, and visualization, along with the engineering resources to clean and organize data. In addition, you can have dedicated resources to put a model into production and optimize it for performance.

One disadvantage of a centralized team is that it can spread support thin and not reach every part of the enterprise. Some business unit managers may not feel that their needs are appropriately prioritized, making it difficult for them to act on time-sensitive issues. In addition, with data scientists removed from the day-to-day activities of the business units, they may not fully gain the deep domain knowledge they need to be effective.

Decentralized teams do their work directly with a business unit, making them an integral part of the business unit team. Decentralized teams tend to be agile and responsive, and by working alongside business analysts and operations, data scientists can quickly become immersed in the required domain knowledge. Business unit managers are happy because their work is prioritized.

The downside to decentralized teams is that silos of talent and knowledge are inevitable. Data professionals may feel isolated from their counterparts elsewhere in the organization, with less access to the expertise they need to vet ideas and share what they are learning, leading to job dissatisfaction. Decentralized data scientists may be less aware of insights and models developed by other teams. Finally, if decentralized data scientists lack ready access to engineers, they can be at a disadvantage when it comes to putting a model into production.

> ***For most companies, full centralization or decentralization of data science resources is inefficient. A hybrid team model can address this.***

Hybrid teams can be an effective way to gain many of the advantages of centralized and decentralized teams with fewer drawbacks. In a hybrid team structure, the data science team is centralized but has small teams dedicated to serving individual business units. Another hybrid approach is to have a centralized data science center of excellence where a leadership team supports data scientists in the field to encourage best practices and facilitate knowledge sharing across departments.

Either way, a hybrid structure helps ensure data professionals get the domain knowledge they need to serve individual business units, and it gives business units a better shot at having access to data science expertise. The smaller teams also have ready access to any skill sets or expertise required for meeting the needs of a business unit.

Develop a shared understanding of the data

As your data-driven organization grows in sophistication, you can expect to have concurrent projects in development, testing, and production. Scaling to this level requires a shared mindset in the business units and at the C-level around using data to drive decisions. All parts of the enterprise need to understand the outcomes they want to test for, know how to identify and measure key metrics that reflect those outcomes, run experiments to validate assumptions, and feel confident that measurements are accurate. This shared understanding will help drive buy-in from management, departments, and individuals across the enterprise, which is essential for successful data-driven organizations. Otherwise, you're simply collecting data—you're not leveraging it to move the needle in the right direction.

> ***This shared understanding will help drive buy-in from management, departments, and individuals across the enterprise, which is essential for successful data-driven organizations. Otherwise, you're simply collecting data—you're not leveraging it to move the needle in the right direction.***

Experimentation helps measure where you are today, measure the delta from where you want to be, and gives data science teams the insight to course correct. Organizations must nurture a culture of experimentation and willingness to test, and, yes, to fail without getting discouraged. If you don't get the outcome you expect, either you're not measuring correctly or your hypothesis is wrong. Knowing which of these factors is driving the unexpected results is crucial to success when scaling initiatives.

Select the right infrastructure

Infrastructure can be an enabler or bottleneck to scaling data-driven initiatives. The basic choices are on-premises, in the cloud, or a hybrid of both. Which direction you go largely depends on your industry and your organization.

Generally, it's faster and easier to scale in the cloud, as scaling on-premises often requires the purchase, installation, and maintenance of a new system while paying for support. However, there are considerations, such as regulatory compliance, that may affect your decision to use (or not use) the cloud. For industries like banking and insurance, the personal nature of the data and stringent security and access control requirements generally dictate an on-premises solution. However, if you want to leverage the cloud for scalability—whether public or private—you need to ensure that personal data is handled in a compliant manner.

Scale-up with the right processes and tools

It's easy to get excited about a new tool, especially one that helps get insight from data. But the capabilities and ease-of-use of a data analytics platform aren't nearly as important as having the right processes in place. Before selecting analytics software, it's imperative to standardize the way you collect and organize data. Lack of standards can lead to serious difficulties when scaling your data initiative. While different parts of the organization need different views of the data, the format and integrity of the data itself must be consistent.

> ***Before selecting analytics software, it's imperative to standardize the way you collect and organize data. Lack of standards can lead to serious difficulties when scaling your data initiative.***

It's much better to have solid processes and a mediocre analytics tool than to invest in the most comprehensive software, only to discover your insights aren't consistent across projects, teams, and stakeholders. With the right data processes in place, you can select any data analytics software that fits your budget and needs.

First, look for a solution that supports your use cases. The best way to test this is to run a trial with a representative user base. The data analytics software should also be easy to set up and deploy—even non-technical users should be able to start using the software within an hour or two. Anything that's complex for users will slow adoption and therefore your ability to scale quickly.

If you're processing a lot of data, the software should have the capacity to handle the expected volume of data with plenty of room to grow—the goal is to enable you to scale up. If your team relies heavily on experimentation and you have models that drive business impact, you'll want to know the performance limits of the tool. For example, can it tackle creating visualizations for thousands or hundreds of thousands of data points?

Ensure that the software you select has the analysis and visualization features you need. It should create the reports, charts, and graphics your organization needs to gain the insights it requires.

Finally, to use your analytics software effectively, you'll need access to adequate documentation, a strong customer support system, or both. Some tools have vendor-sponsored or independent user communities that share helpful information, best practices, and pro tips. If you get stuck on a problem, you want to be sure you have a support network to fall back on for help.

Scaling up requires the right process *and* the right tools

- **Standardize data collection and organization**
- **Solid processes are more crucial than expensive analytics packages**
- **Look for solution(s) that:**
  - **Support your use cases**
  - **Are easy to set up and use**
  - **Can handle the volume of data you expect**
  - **Has strong analytics and visualization**
  - **Provides strong support**

Scaling deployment and continuous improvement

Finally, you'll need a process for testing and experimentation and a process for deployment. Using a purpose-built platform to design and run data experiments will help organizations ensure that business decisions are based on consistent, reliable analysis.

For example, when running experiments to evaluate new metrics, visualizations, or models, it's important to log the data about the experiments in a centralized store. Logs will come in handy when, in the process of scaling, you encounter problems. A good data science platform will handle the logging and will also provide tools to visualize and organize the data to deliver the insight you're looking for.

After running a series of experiments, you'll want to share the results with others. Your data science platform should make it easy to share your work to make it more visible, which is crucial to scalability.

> ***Above all, a data science platform designed for an efficient workflow will enable all the above without custom coding for data collection, experimentation, visualization, reporting, and sharing. Easing that workflow allows your data scientists to spend more time getting the organization to insights through experimentation rather than waiting for programmers to deliver custom code.***

Conclusion

To get real insight from data across projects, teams, and stakeholders—insights that deliver real business value—organizations must be able to scale data initiatives to as many business decision-makers as possible. Getting to that point is as much about processes and culture as it is about tools and technologies.

The ability to scale requires a combination of an appropriate data science team structure, solid processes for standardized data collection and storage, a culture of experimentation, a flexible infrastructure—and selecting software tools that facilitate every aspect of your data science initiatives. Only then can you turn data into insights at scale across the entire organization.

READY TO GET YOUR DATA SCIENCE AND ML INITIATIVES OFF THE GROUND FASTER?

Comet enables data scientists and teams to track, compare, explain and optimize experiments and models across the model's entire lifecycle, from training to production. With just two lines of code, you can start building better models today.

Contact Comet to learn more.

comet

Comet provides a self-hosted and cloud-based meta machine learning platform allowing data scientists and teams to track, compare, explain and optimize experiments and models.

Backed by thousands of users and multiple Fortune 100 companies, Comet provides insights and data to build better, more accurate AI models while improving productivity, collaboration and visibility across teams.

Learn more at www.comet.ml