data
iku

WHY ENTERPRISES NEED DATA SCIENCE, MACHINE LEARNING, AND AI PLATFORMS

A WHITE PAPER BY DATAIKU

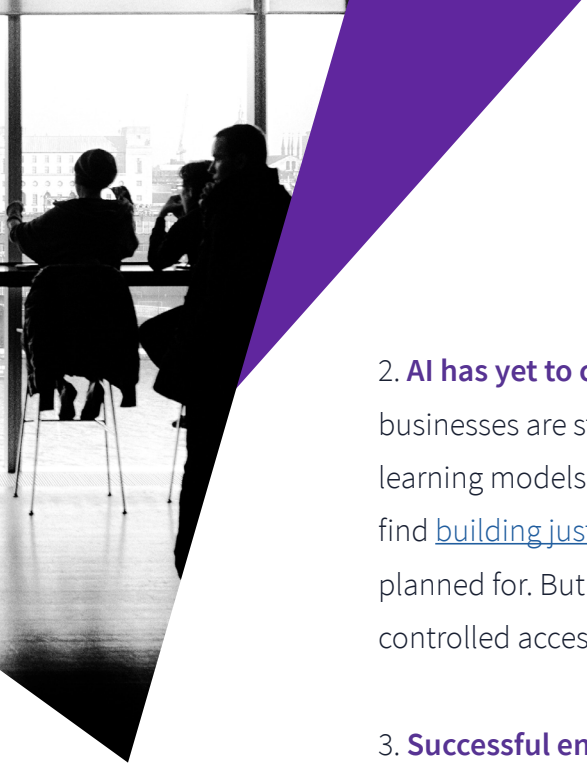
www.dataiku.com

INTRODUCTION

THE STATE OF THE MARKET

In order to go in-depth on what exactly data science and machine learning (ML) tools or platforms are, why companies large and small are moving toward them, and why they matter in the Enterprise AI journey, it's important to take a step back and understand where we are in the larger story of AI, ML, and data science in the context of businesses:

1. **Enterprise AI is at peak hype.** Of course, the media has been talking about consumer AI for years. But since 2018, the spotlight has turned to the enterprise. The number and type of devices sending data is skyrocketing while the cost of storing data continues to decrease, which means most businesses are collecting more data in more types and formats than ever before. And in order to compete and stay relevant among digital startups and other competition, these companies need to be able to use this data to not only drive business decisions, but drive the business itself. Now, everyone is talking about how to make it a reality.



2. **AI has yet to change businesses.** Despite the hype, the reality is that most businesses are struggling to actually leverage data at all, much less build machine learning models or take it one step further into AI systems. For some, it's because they find building just one model is far more expensive and time-consuming that they planned for. But the large majority struggle with basic challenges, like even organizing controlled access to data or efficient data cleaning and wrangling.

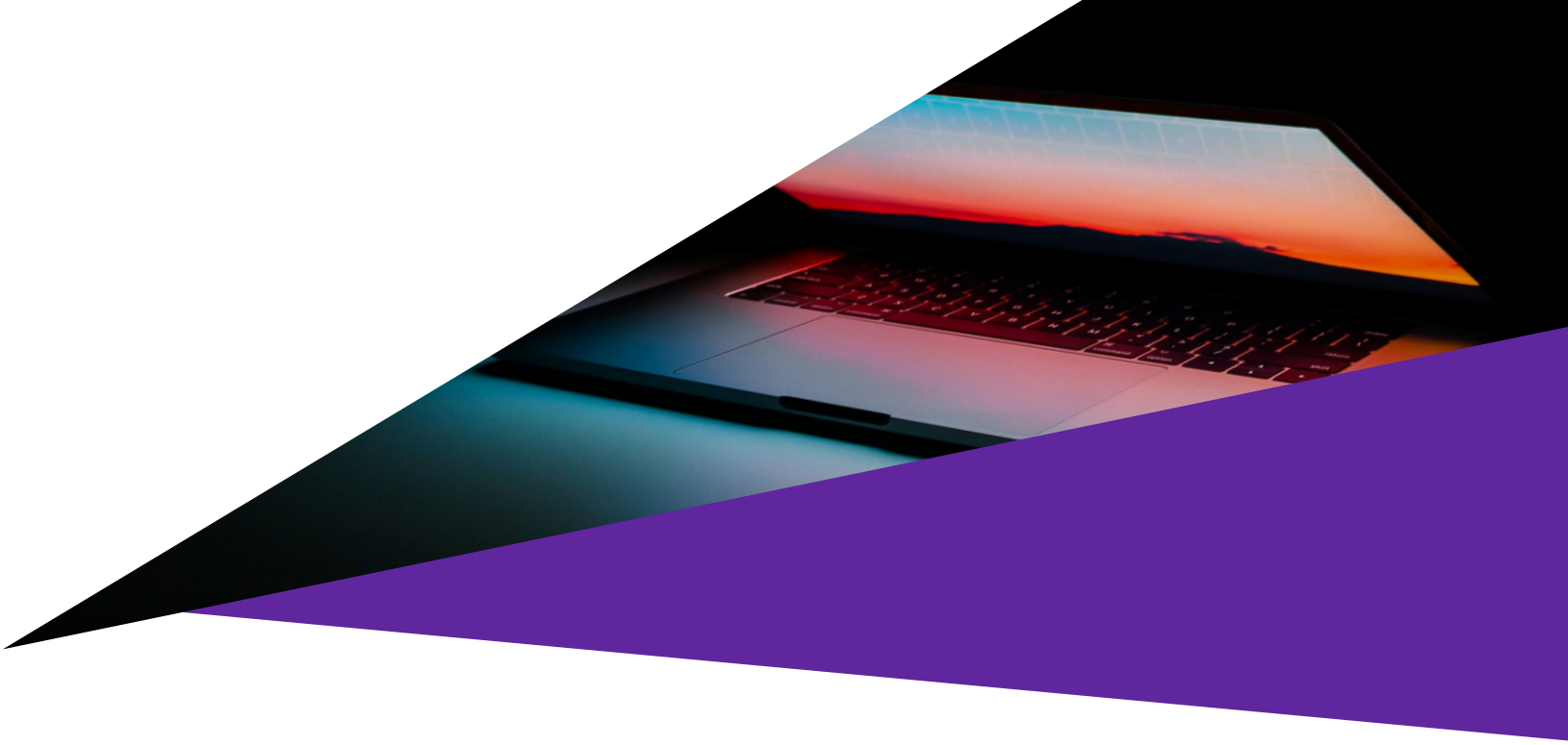
3. **Successful enterprises have democratized.** Those companies that have managed to make progress toward Enterprise AI - like Pfizer and GE Aviation - have realized that it's not one ML model that will make the difference; it's hundreds or thousands. And that means scaling up data efforts in a big way that will require everyone at the company to be involved. Enter: democratization. In August 2018, Gartner identified Democratized AI as one of the five emerging trends in their Hype Cycle for Emerging Technologies. Since then, Dataiku has seen the word "democratization" creep into the lexicon of AI-hopefuls everywhere, from the media to the board room. And to be sure, it's an essential piece of the puzzle when it comes to understanding why data science and machine learning (ML) platforms.

DATA TEAM HIRING: CRITICAL, BUT NOT ENOUGH

Hiring for data roles is at an all-time high. Already in 2019, according to job listing data, data scientist is the hottest profession out there. And though statistics on Chief Data Offers (CDOs) differ, some put the figures as high as 100-fold growth in the role over the past decade.

Hiring data experts is a critical component to a robust Enterprise AI strategy; however, hiring alone doesn't guarantee results, and it isn't a reason not to invest in data science and ML platforms. For one thing, hiring data scientists is expensive - often prohibitively so - and they're only getting more so as their demand grows (plus, spoiler alert: the fancy ones with the PhDs tend to be unavailable, because 80 percent of them are taken by Google).

The reality is that when the goal is going from producing one ML model a year to tens, hundreds, or even thousands, a data team isn't enough because it still leaves a large swath of employees doing day-to-day work without the ability to leverage data. In other words, without democratization, the effect of a data team - even the best one made up of the top data scientists - would be limited.



As a response to this, some companies have decided to leverage their data team as sort of an internal contractor, working for lines of business or internal groups to complete projects as needed. Even with this model, the data team will need tools that allow them to scale up, working faster, reusing parts of projects where they can, and (of course) ensuring that all work is properly documented and traceable - more on this later. A central data team that is contracted out can be a good short-term solution, but it tends to be a first step or stage; the longer-term model of reference is to train masses of non-data people to be data people.

> *"The Self-Service Data Program at GE Aviation was born out of a conversation in a conference room. The idea was that you would never be able to hire enough data professionals to meet the data demands of the business, so instead, why not turn the business into data professionals. Taking that premise we started to define what self-service meant for us and how it would work."*
>
> - Jon Tudor, Senior Manager of Self-Service Engineering and Analytics | GE Aviation
> (Source: GE Aviation: From Data Silos to Self-Service)

OPEN SOURCE: CRITICAL, BUT NOT ENOUGH

There's no question that open source technologies in data science and machine learning are state-of-the-art and that organizations have to adopt them to be dynamic and future-minded. In fact, the bleeding edge of data science algorithms and architecture is only about six months ahead of what is being open sourced, whether directly by these companies or via original development or reverse engineering.

Take, for example, TensorFlow, which is a library for building and training neural networks - there is simply nothing like it on the market right now. It was open sourced by Google (specifically by their Google Brain project) in late 2015, which means that anyone using it is using the same standard that Google is using for their neural network development.

In addition to being on the bleeding edge of technological developments, using open source makes it easier to hire and onboard a team. Not only are prospective data scientists interested in growing their skills with the technologies that will be the most widespread in the future, but also there is less of a

learning curve if they can continue to work with tools they know and love instead of being forced to learn an entirely different (and usually less innovative) system.

It's important to remember, though, that keeping up with that rapid pace of change is difficult for enterprise-sized organizations. These latest innovations are usually highly technical, so without some sort of packaging or abstraction layers that make the innovations more accessible, it's very difficult to keep everybody in the organization on board and working together. A business might technically adopt the open source tool, but only a small number of people will be able to work with it. Not to mention that governance can be a huge challenge if everyone is working with open source tools on their local machines without a way to have work centrally accessible and auditable.

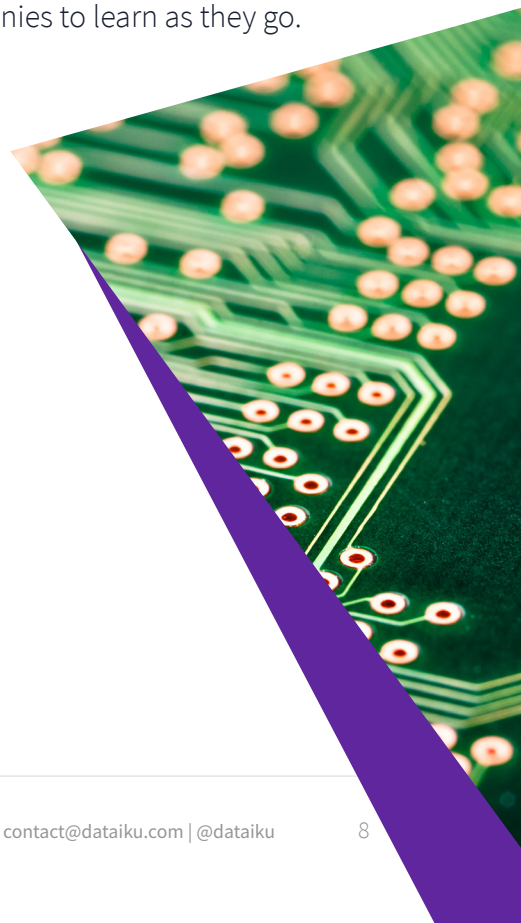
Data science and ML platforms have the advantage of being usable right out of the box so that teams can start analyzing data pretty much from day one. With open source tools, you need to assemble a lot of the parts by hand, so to speak, and as anyone who's ever done a DIY project can attest to, it's often much easier in theory than in practice. Choosing a data science and ML platform wisely (meaning one that is flexible and allows for the incorporation and continued use of open source) can allow the best of both worlds in the enterprise: cutting-edge open source technology and accessible, governable, control over data projects.

ENTER: DATA SCIENCE AND ML PLATFORMS

In their most basic form, data science and ML platforms are tools to enable Enterprise AI by allowing people within the organization to:

- Use data to **produce predictive analytics (or machine learning)** solutions.
- Scale by providing **transparency and reproducibility** throughout the team and within a project.
- Access all data and work together on data projects in a central location (i.e., **collaboration and governance**).

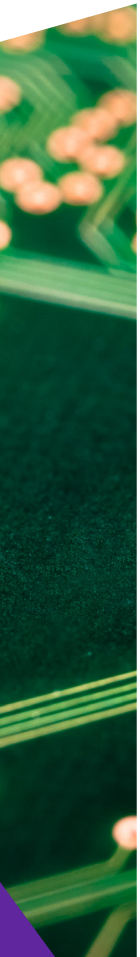
Ultimately, data science and ML platforms are about time. That means time savings in all parts of the processes (from connecting to data to building ML models to deployment), of course. But it's also about easing the burden of getting started in AI and allowing businesses to dive in and get started now - not waiting for years until technologies sure up and the world of AI becomes more clear (because spoiler alert: that may never happen). Getting started on the AI journey is intimidating, but data science and ML platforms can ease that burden and provide a framework that allow companies to learn as they go.



THE HOW

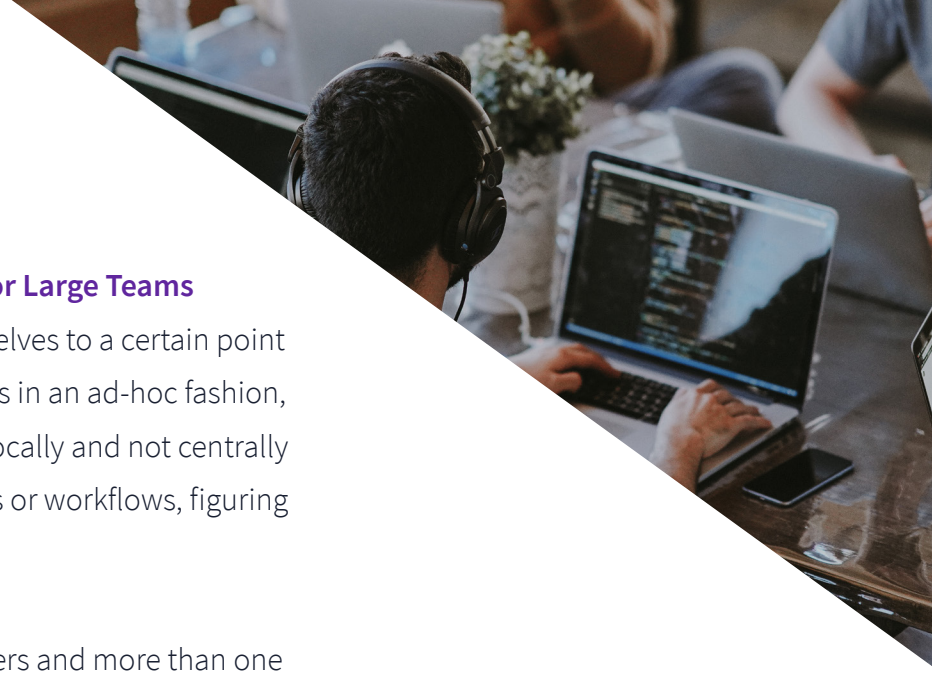
Data science and ML platforms allow for the scalability, flexibility, and control required to thrive in the era of Enterprise AI because they provide a framework for:

- **Collaboration:** A way for additional staff working with data, many of whom will be non-coders, to contribute to data projects along with data scientists (or IT and data engineers).
- **Data governance:** Clear workflows and a way for team leaders to monitor those workflows and data projects.
- **Efficiency:** Finding small ways to save time throughout the data-to-insights process gets companies to business value faster.
- **Automation:** A specific type of efficiency is the growing field of AutoML, which is expanding to automation throughout the data pipeline to alleviate inefficiencies and free up staff time.
- **Operationalization:** Efficient means to deploy data projects into production quickly and safely.
- **Self-Service Analytics:** A system by which non-data professional from different lines of business can access and work with data in a controlled environment.



"Expert data scientists, citizen data scientists and application developers require professional capabilities for building, deploying and managing analytical models."

- The 2019 Gartner Magic Quadrant for Data Science and Machine-Learning Platforms



THE WHY

Ad-Hoc Methodology is Unsustainable for Large Teams

Small teams can potentially sustain themselves to a certain point by working on data, ML, or larger AI projects in an ad-hoc fashion, meaning team members store their work locally and not centrally and don't have any reproducible processes or workflows, figuring things out along the way.

But with more than just a few team members and more than one project, this becomes unruly quickly. Any business with any hope of doing Enterprise AI needs a central place where everyone involved with data can do all of their work, from accessing data to deploying a model into a production environment. Allowing employees - whether directly on the data team or not - to work ad hoc without a central tool from which to work is like a construction team trying to build a skyscraper without a central set of blueprints.

Models Need to be Monitored and Managed

The biggest difference between developing traditional software and developing machine learning models is maintenance. For the most part, software is written once and doesn't have to be continually maintained - it will generally continue to work over time. However, machine learning models are developed, put in production, and then must be monitored and tweaked until performance is optimal.

Even once performance is optimal, model performance can still shift over time as data (and the people producing it) changes. This is quite a different approach, especially for companies that are used to putting software in production. And it's easy to see how issues with sustainability could eventually cause - or exacerbate - problems with ML model bias. In fact, the two are deeply intertwined, and

ignoring both can be devastating to a business's data science efforts, especially when magnified by the scaling up of efforts. All of these reasons point to having a platform that can help manage model monitoring and management.

AI Needs to Be Responsible

Children are taught from a young age that subjects like science and math are all objective, which means that inherently, people believe that data science is as well - that it's black and white, an exact discipline with only one way to reach a "correct" solution, independent of who builds it. Yet we've known for a long time that this is not the case and that it is entirely possible to use data science techniques (and, thus, create AI systems) that do things, well… wrong.

In 2015, for example, both Google Photos and Flickr received backlash after mislabeling African Americans as gorillas. Even as recently as last year, it appears that Google, at least, may not have completely resolved the issue, proving just how difficult it is to get AI right. Though neither company ever broke down the details of the problem, it was probably a question of bias in the dataset and/or a misspecified loss function (that is, the method by which machine learning algorithms actually learn). A centralized system to which everyone has access to test for and prevent against irresponsible AI (or, on a smaller scale, ML models and even predictive analytics projects) is critical.

Governance is Getting Tricker

With the amount of data being collected today, data security (especially in certain industries like finance) is critical. Without a central place to access and work with data that has proper user controls, data could be stored across different individuals' laptops. And if an employee or contractor leaves the company, the risks increase not only because they could still have access to sensitive data, but because they could take their work with them and leave the team to start from scratch, unsure of what the person was working on.

On top of these issues, today's enterprise is plagued by shadow IT; that is, the idea that for years, different departments have invested in all kinds of different technologies and are accessing and using data in their own ways. So much so that even IT teams today don't have a centralized view of who is using what, how. It's an issue that becomes dangerously magnified as AI efforts scale and points to the need for governance at a wider and more fundamental scale across all lines of business in the enterprise.

Reproducibility of Data Projects and Processes Required to Scale

Nothing is more inefficient than unnecessarily repeating the same processes over and over. This applies to both repeating processes within a project (like data preparation) again and again as well as repeating the same process across projects or - worse - inadvertently repeating entire projects if the team gets large but doesn't have insight into each other's work.

And no business is immune to this risk - in fact, this problem can become exponentially worse in large enterprises with bigger teams and more disconnect between them. In order to scale efficiently, data

teams need a tool that helps reduce repeated work and ensures that work between members of the team hasn't already been done before.

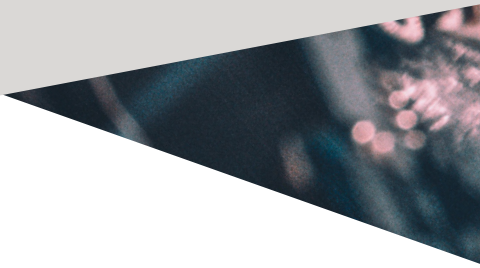
Leverage Data Analysts to Augment Data Scientists' Work

Today, data scientist is one of the most in-demand positions. This means that data scientists can be both (1) difficult to find and attract and (2) expensive to hire and retain. This combination means that in order to scale data efforts to work toward Enterprise AI, it will inevitably need to be filled out with business or data analysts.

In order for the two types of staff to work together effectively, they need a central environment from which to work. Analysts also tend to work differently than data scientists, adept in spreadsheets and possibly SQL but usually not coding. Having a tool that allows each profile to leverage the tools with which (s)he is most comfortable allows for the efficiency to scale data efforts to any size.

Need to Create Models that Work in Production

nvesting in predictive analytics and data science means ensuring that data teams are productive and see projects through to completion (i.e., production) - otherwise known as operationalization. Without an API-based tool that allows for easy deployment, data teams likely will have to hand off models to an IT team who then will have to re-code it. This step can take lots of time and resources and be a huge barrier to implementing data projects that truly impact the business in meaningful ways. With a tool that makes it seamless, data teams can easily have an impact, monitor, tweak, and continue to make improvements that positively impact the bottom line.



WHAT SHOULD THE IDEAL PLATFORM PROVIDE?

While the above represents a general definition of a data science platform, of course, different tools provide a different array of features. For optimal productivity and efficiency on a data team, there are specific features to look for.

TEAM AND STAFFING

Ideal Data Science or ML Platform Feature

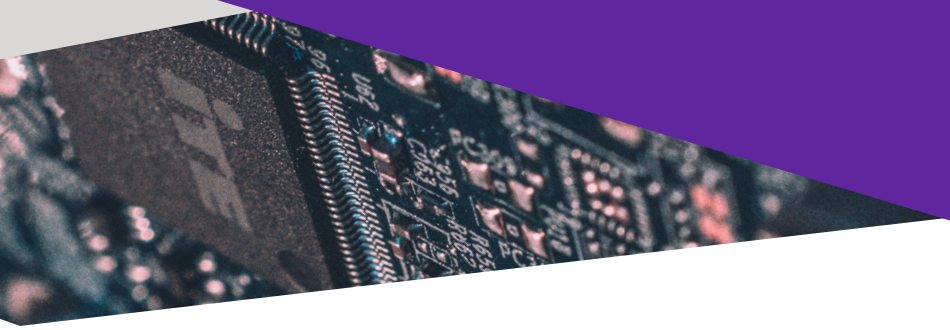
- Allows contributions via code or through visual interface for analysts.
- Facilitates faster data prep and allows analysts to do data preparation.
- Allows multiple profiles to work together on different project components.

To Solve This Challenge

- Data analysts can't contribute to data projects in meaningful ways.
- Data preparation takes too long, takes up valuable data science resources.
- No collaboration among or between data analysts/data scientists/business teams.

For These Benefits

- Cost savings in getting the right profiles to do the right tasks to match their skills.
- Data scientists reserved for higher-level, more interesting tasks instead of data prep; better retention of talent.
- Faster, more scalable process to go from raw data to prediction.



TECHNOLOGY

Ideal Data Science or ML Platform Feature

Robust access controls and monitoring.	Data- and tool-agnostic integration.	One tool that handles everything from raw data to deployment.

To Solve This Challenge

Lack of centralized data access and user controls.	Difficulty connecting to and combining siloed data sources; difficulty orchestrating together a wide variety of tools being used throughout the data-to-insights process.	Use of different tools/platforms for data preparation, model development, and model deployment.

For These Benefits

- Team can use the latest and best data science technologies without being constrained to a tool.
- No loss of productivity on in-progress projects when team members leave.
- Instant tool adoption since each team member can use languages and techniques with which (s)he's familiar.
- Overall centralization and organization allows for multiple data projects to happen simultaneously (better scalability).

OPERATIONS

Ideal Data Science or ML Platform Feature

- API-driven model deployment.
- Allows for model deployment as a reproducible package.
- Reusability and replicability.

Data Team Challenge This Feature Addresses

- Inability to quickly put models in production
- Inability to make changes to models already in production.
- Inability to easily share best practices across team members and between teams.

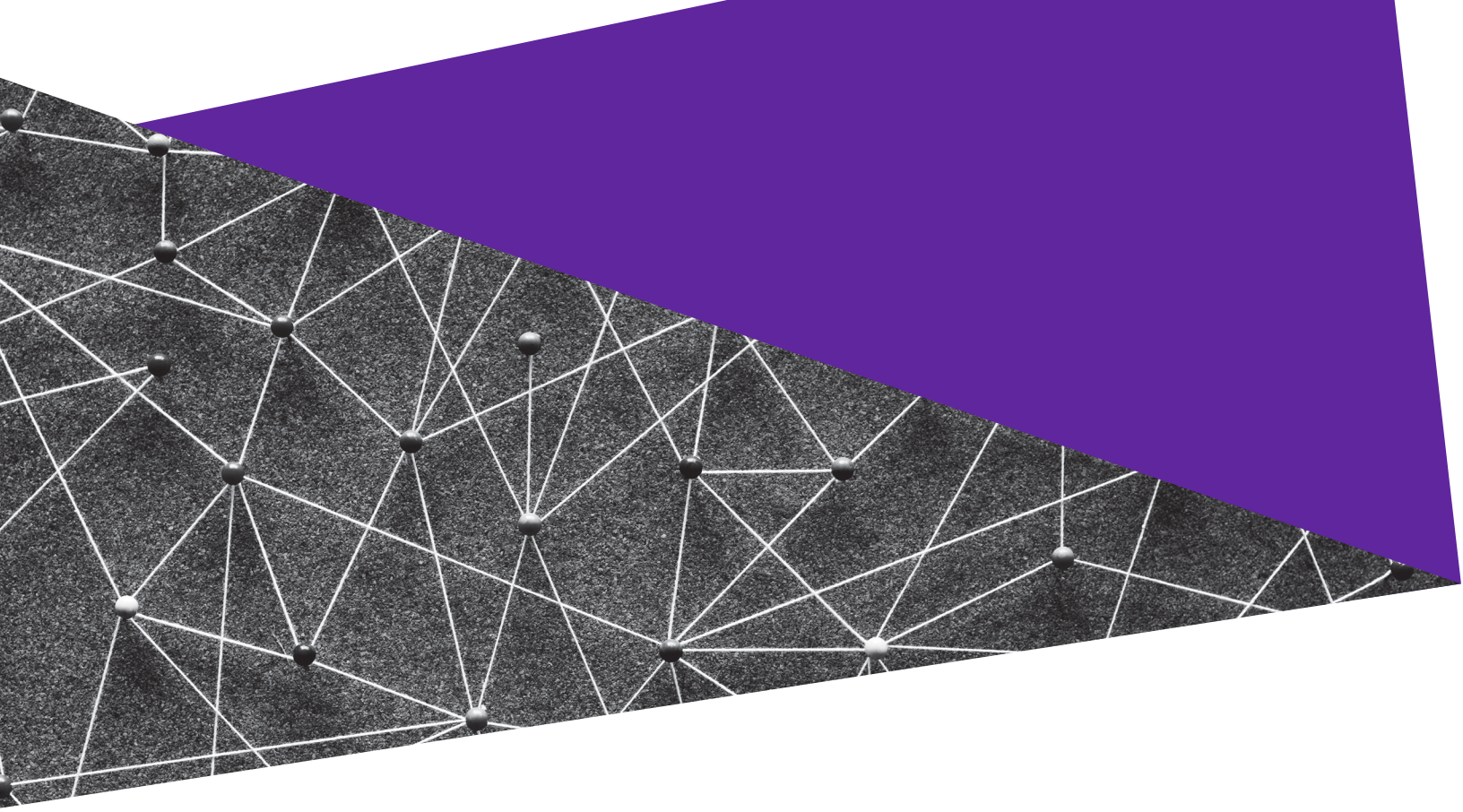
For These Benefits

Business sees more ROI faster on data projects, meaning more resources and better funding for the team, allowing it to grow and take on more exciting projects.

ADDRESSING THE FEAR OF VENDOR LOCK-IN

Understandably, some enterprises are reluctant to commit to data science and ML platforms that might lock them in to a certain system or specific technologies - this is especially the case if they have been burned in the past by cumbersome, expensive systems that hindered the very data efforts they were supposed to accelerate.

The key to mitigating this fear is to choose a data science and ML platform that not only is built to handle all parts of the data-to-insights process (to avoid having to augment abilities with further tools later on), but also to choose one that is completely flexible, open, and innovative when it comes to technology integrations.



One smaller detail to look for to ensure that the company doesn't get locked in to technologies that hinder their overall growth in Enterprise AI is to ensure models can be exported so that should the business change directions later, all work is not lost.

But more broadly, ask questions about not only the ability of the potential platform to be integrated with all current technologies (programming languages, ML model libraries that data scientists like to use, and data storage systems), but about the vision of the company. It should be wide enough such that any new technologies the company may want to invest in the future can be easily integrated with the platform later on due to the vendor's interest in staying open and cutting-edge.

THE DIFFERENCE WITH DATAIKU

Dataiku is the centralized data platform that moves businesses along their data journey from analytics at scale to enterprise AI. By providing a common ground for data experts and explorers, a repository of best practices, shortcuts to machine learning and AI deployment/management, and a centralized, controlled environment, Dataiku is the catalyst for data-powered companies.

What sets Dataiku apart?

- **It is truly built for everyone:** Democratization has been Dataiku's mission since its founding in 2013. It is fundamentally designed for and used by all kinds of professionals (whether they are IT, data scientists, data engineers, software engineers, business people, managers, analysts, and more) at companies large and small around the world. Dataiku pioneered the idea of horizontal vs vertical collaboration when it comes to data efforts - that is, people working together with others who have roughly the same skills, toolsets, training, and day-to-day responsibilities vs. people from across teams working together who might have vastly different responsibilities. And the platform handles both types of collaboration with ease with a complete suite of features that enable communication and allow both technical and non-technical staff work with data their way.

- **It is end-to-end:** Many tools and platforms today say they are end-to-end, but they actually only handle one or two parts of the data process, so inevitably, the business needs to purchase other tools to round out the Enterprise AI strategy (not to mention find a way to cobble these tools together and make the data workflow seamless between them). But Dataiku allows everything data-related to happen in one single tool with one simple and unified UI - from connecting to data to ETL to model creation to operationalization to model monitoring in production. And it does this no matter the enterprise's underlying data architecture, industry, or use case.
- **It isn't black box:** Dataiku, both as a product and a company, pushes the idea of responsible AI. That means that it's not a black box where you put data in one end and get results from a model out the other. It provides complete transparency into what data is being used where, by whom, and in which models. This ensures that as regulations and standards around data privacy and AI ethics continue to strengthen, Dataiku will support the transition and the business in achieving compliance.

Customers across retail, e-commerce, health care, finance, transportation, the public sector, manufacturing, pharmaceuticals, and more use Dataiku to power self-service analytics while also ensuring the operationalization of machine learning models in production. By removing roadblocks, Dataiku ensures more opportunity for business-impacting models and creative solutions, allowing teams to work faster and smarter.

CONCLUSION

In short, data science and machine learning platforms are the underlying framework that allow companies to scale and be more productive when it comes to data initiatives, paving the path to Enterprise AI. They should allow for easy (but controlled) access to data necessary to complete complex data projects and initiatives, keep all work centralized (and thus reproducible), and facilitate critical collaboration not only among similar profiles but between them (data scientist, business/data analyst, IT, etc.).

Perhaps most importantly, data science and ML platforms open up the door to true data innovation when teams don't have to spend precious time on administrative, organizational, or repeated tasks. The reality is, in the age of AI, businesses of any size can't afford to work without a data science platform that enables and elevates not just their data science team, but the entire company to the highest level of data competence for the greatest possible impact.

data iku

Your Path to Enterprise AI

Dataiku is the centralized data platform that moves businesses along their data journey from analytics at scale to enterprise AI. Data-powered businesses use Dataiku to power self-service analytics while also ensuring the operationalization of machine learning models in production.

20,000+

ACTIVE-USERS

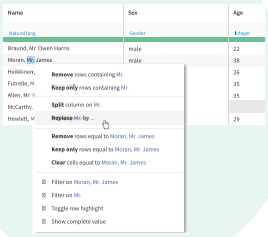
*data scientists, analysts, engineers, & more

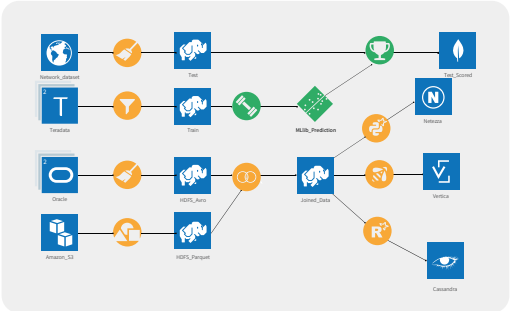
200+

CUSTOMERS

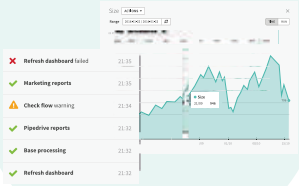
BNP PARIBAS · Santander · paloalto NETWORKS · PREMERA BLUE CROSS · FOX NETWORKS GROUP · Unilever · NXP · SEPHORA

1. Clean & Wrangle



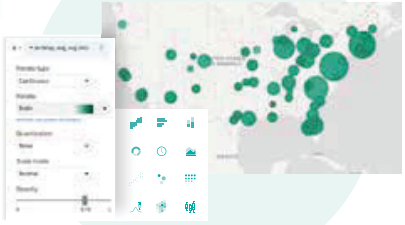


5. Monitor & Adjust

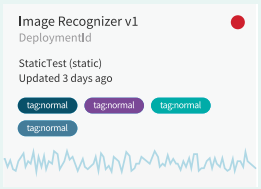


2. Build + Apply Machine Learning

3. Mining & Visualization



4. Deploy to production



WHITE PAPER

www.dataiku.com