EBOOK

# 2021 Trends: Where Enterprise AI Is Headed Next

www.dataiku.com

# Introduction

2020 was a year defined by uncertainty. But as we dive into 2021, we look under the hood to gather our learnings from last year. We find growing glimpses of optimism and preparedness for the future — particularly when we observe the changes 2020 ignited. When we look back on modern history, we can note a few commonalities:

According to patent data from 1750 to 1970, there were significant spikes in innovation during the Long Depression of the 1870s and 1880s and the Great Depression of the 1930s.[1]

The Great Recession of 2007-2009 saw the creation of startups like Airbnb, Uber, and Venmo (which have since become massive global companies).[2]

More than three-quarters of executives believe that the 2020 global health crisis will create significant new opportunities for growth, according to McKinsey.[3]

"Throughout the COVID-19 crisis, the majority of organizations have been maintaining or even increasing their investments in artificial intelligence (AI), according to polling results from a Gartner webinar in May 2020."[4]

While conscious to not undercut the lasting impacts of the pandemic on families and businesses across the globe, we take this approach in an effort to help organizations find the opportunities that exist and can arise in 2021 with the implementation and execution of data science and AI at scale. Further, they'll be positioned to react and respond to future periods of crisis and adversity with more clarity and will be able to preserve business continuity, mitigate risk, and establish a basis for postcrisis growth.

Change won't happen overnight, but with the right tools, teams, and tactics, organizations will be better equipped to harness those opportunities and translate them into valuable business advantages (i.e., lower costs, increased revenue, more automated processes). It is our hope that the AI trends outlined here will spur organizations into action, catalyzing enterprise-wide change and a renewed focus on collaboration, agility, and efficiency.

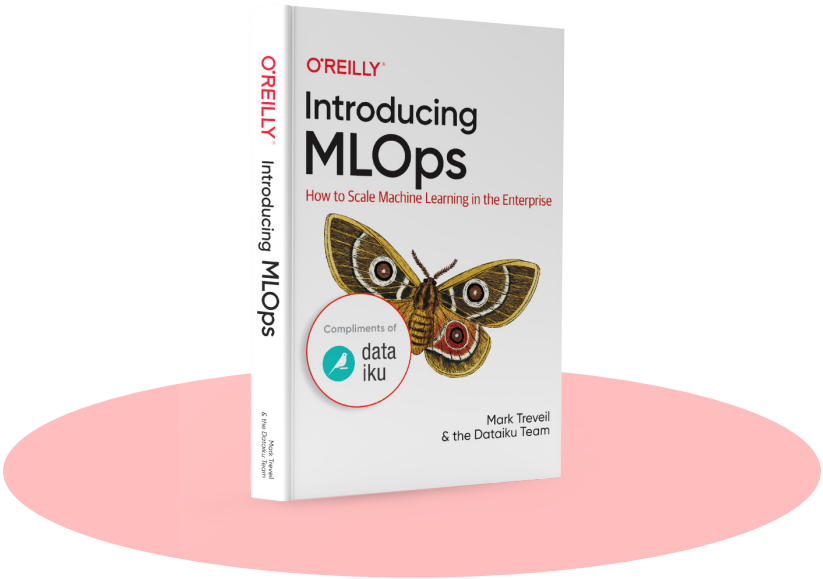
[1] *https://www.theatlantic.com/national/archive/2009/07/innovation-and-economic-crises/20576/*
[2] *https://www.businessinsider.com/successful-companies-started-during-past-us-recessions-2020-4#venmo-also-got-its-start-in-2009-towards-the-end-of-the-great-recession-14*
[3] *https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/innovation-in-a-crisis-why-it-is-more-critical-than-ever*
[4] *Gartner - Debunking Myths and Misconceptions About Artificial Intelligence, 2021, Saniye Alaybeyi, Pieter den Hamer, 10 September 2020*

# MLOps Will Become Even More Critical

According to the O'Reilly book "**Introducing MLOps: How to Scale Machine Learning in the Enterprise**," "Business leaders view the rapid deployment of new systems into production as key to maximizing business value. But this is only true if deployment can be done smoothly and at low risk." Easy to say in theory, but how is this series of steps practically accomplished at scale?

Last year, we predicted that the year 2020 will almost certainly be one where MLOps processes are formalized and staffed as well as bolstered up with tools that make the job smoother and more manageable. Little did we know at the time how true this would actually become based on the events that unfolded throughout the year.

Organizations will take their MLOps foundations (established in recent years) and go a step further to implement detailed processes and requirements around **drift monitoring using MLOps**. Input drift is based on the principle that a model is only going to predict accurately if the data it was trained on is an accurate reflection of the real world. If a comparison of recent requests to a deployed model against the training data shows distinct differences, there is a high likelihood that the model performance is compromised.



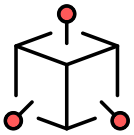
In 2020, the significant drift observed was a result of the global health crisis, a wake-up call to data scientists and their broader teams about the importance of drift monitoring in deployed models. As a result, this new year is bound to include organizations using MLOps to put more structure in place around drift monitoring so that models can be more agile and accurate. In order to be able to rapidly pivot during future crises, organizations with models in production should ask themselves:

- Have the business needs related to the models changed given the circumstances?
- Has the availability, reliability, and relevance of input data changed? If so, are there new data sources that can be used to improve prediction targets?
- Is the granularity of the predictions or of the models (i.e., the time horizon for forecasts) still adequate?
- Should the predictions be consumed in a different way during a period of volatility? For example, if the predictions used to be directly used by an automated system, should a human expert be incorporated at each stage?

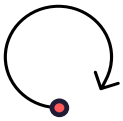
The way organizations approach model drift will depend on whether it is possible to get immediate or delayed feedback. In both cases, the efforts related to model monitoring should increase during and after periods of fluctuation and change given the high risk of model drift:

| Situation | Example | Drift Detection Mechanism |
|---|---|---|
| **Immediate feedback** | A recommender system in an e-commerce setting | Direct comparison of the prediction and the new measurement |
| **Delayed feedback** | A forecast for next month's sales | Detection of a potential change of the input data distribution or of the predictions distribution<br><br>Comparison of the prediction and the new measurement when the latter is available |

Organizations won't stop there. Aside from using MLOps for the short term to address model drift during economic turmoil or periods of drastic change, teams will also likely look to implement MLOps practices for the long term in an effort to more effectively scale their machine learning efforts. Being able to transition from one or a handful of models in production to tens, hundreds, or even thousands that have tangible business value and can adapt on a dime will require key facets of an MLOps practice, including:

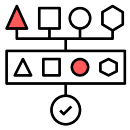


**Model input drift detection** that looks at the recent data the model has had to score and statistically compares it with the data on which the model was evaluated



**Easier creation of validation feedback loops** — such as with Evaluation Recipes in Dataiku — to compute the true performance of an existing model against a new validation dataset

**Automated retraining** and redeployment



**Dashboard interfaces** that foster collaborative modeling of global pipelines

***"AI experimentation, a key part of MLOps, will become more strategic.*** *Experimentation takes place throughout the entire model development process — usually every important decision or assumption comes with at least some experiment or previous research to justify those decisions. Experimentation can take many shapes, from building full-fledged predictive ML models to doing statistical tests or charting data.*

*Trying all combinations of every possible hyperparameter, feature handling, etc., quickly becomes untraceable. Therefore, we'll begin to see organizations define a time and/or computation budget for experiments as well as an acceptability threshold for usefulness of the model."*

***- Florian Douetteau***

When scaling AI projects in 2021, organizations should not attempt to predict the future, but rather build for resilience. It's impossible to predict what the future of technology will look like as it is ever-evolving, so organizations need to be able to keep pace with the changes and swap in the most innovative solutions possible. This isn't to say that organizations should drop their existing technologies whenever a new product comes to market, but rather align themselves with the products that will help them most accurately solve their bespoke business challenges and set them up for success during future periods of uncertainty.

The rise of model drift monitoring and MLOps as a whole over the course of 2020 is a great sign of the data science and machine learning industry growing in maturity, as it demonstrates more and more machine learning models are being deployed in production every day. It also shows that teams are taking ownership of making sure they have a robust plan in place for standardizing and managing the entire machine learning lifecycle.

**Moving forward, MLOps is no longer a nice-to-have**, but rather a prerequisite for scaling data science and machine learning at the enterprise level, and doing so in a way that doesn't put the business at risk. These practices help organizations bridge the gap between the technical and business teams, creating transparency among both parties with regard to what models are deployed in production and what their impacts are on the business.

# Teams Will Need to Infuse Agility Amidst a Post-Pandemic Environment

*According to Gartner, "Resetting everything is an easy phrase but difficult to implement when you want to move forward. Our greatest expectation is that companies will seize the opportunity to find new ways of solving problems that recognize technology is best suited to helping us do better in every aspect of life … [Organizations or CIOs] need to look at what technology investment will help them in their mission to reduce costs while also driving their digital journey toward growth." [5]*

In 2021, the use of AI for sustained resilience will be underscored, particularly with regard to empowering every team and employee to work with data to improve their business output. The challenges we observed in 2020 will remain in 2021 for teams that don't have a collaborative data science platform.

## 1. ACCESS TO SYSTEMS:

Whether accessing the various data sources or the computational capabilities, doing so in a remote setting can be challenging. One way that organizations can adapt during and after a crisis is establishing a thoughtful approach to people, process, technology alignment and, namely, investing in those initiatives via a collaborative data science tool. Businesses need to be able to pivot in future times of volatility, so they'll need to be equipped with the skills and AI systems to adapt over time.

According to McKinsey, "Crises … are watershed moments for companies to evolve. Successfully managing a business model shift first requires determining which aspects of the model have been impaired and are unlikely to return."[6] While remote and hybrid work environments may have been challenging to adjust to at first, they allow businesses to adapt without interruption to any open data science and machine learning projects (and thus AI strategy or progress overall).

[5] *Gartner - Gartner's Top Strategic Predictions for 2021 and Beyond: Resetting Everything, Daryl Plummer, Janelle Hill, Rita Sallam, Gene Alvarez, Frances Karamouzis, Svetlana Sicular, Mike McGuire, Kyle Rees, Rajesh Kandaswamy, Magnus Revang, Anthony Mullen, Nick Jones, Jennifer Irwin, Emily Potosky, Emily Rose McRae, Dave Aron, Todd Yamasaki, Don Scheibenreif, 20 November 2020*

[6] *https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/innovation-in-a-crisis-why-it-is-more-critical-than-ever*

A platform like Dataiku enables remote work by allowing people of all job functions to access data and work together on projects in a central location, facilitating good data governance practices and enterprise-wide collaboration.

This is especially critical as businesses move along their AI maturity journey and data projects underpin increasingly essential activities for the company, making any interruptions to those projects potentially devastating.

## 2. COLLABORATION WITHIN TEAMS:

Without the physical in-office proximity, individuals can become siloed in the execution of their data projects. Aligning people, processes, and technology is a matter of time, particularly as companies navigate their unique market and understand the ever-evolving dynamics in a post-pandemic setting. From a people perspective, the year 2020 more or less normalized remote and hybrid working styles, making collaboration — beginning within teams — even more critical for agility and efficiency.

## 3. COLLABORATION ACROSS TEAMS:

Data projects require buy-in and validation from business teams and also require data engineering and other teams to help with operationalization. On an organization-wide scale, there needs to be a desire to learn from data, collaborate internally within and across teams, and govern data and machine learning — a process that takes the right skills and organizational structure to progressively automate business processes without losing the ever-valuable human-in-the-loop element. Real AI transformation goes beyond hiring and retention and more so has to do with giving everyone a seat at the table, involving more than just data scientists and analysts in data initiatives.

In 2021, organizations should set a goal to reach a point (gradually over time) where data and analytics are deeply ingrained in the company's culture, i.e., scaling to the point of thousands of data projects, hundreds of thousands of datasets, and thousands of people participating in some stage of that process.

## 4. REUSE OVER TIME:

Capitalizing on past projects is key to maintaining productivity and reducing duplicate work. The lack of in-person discussions can limit this ability. Addressing larger, high-priority use cases while also providing the tools to expand those to lots of other smaller use cases by reusing bits and pieces eliminates the need to reinvent the wheel with data cleaning and prep, operationalization, monitoring and — in doing all of that — ensures that data scientists are spending their time on high-value tasks.

In the pharmaceutical space, for example, organizations can build a model for a specific disease in order to predict when a patient is ready for a specific kind of drug therapy. Once the first model is built, it can be scaled and repurposed for treatment of another disease using the same drug therapy as the first disease, enabling teams to avoid having to start from scratch, reducing the time to model completion, and injecting efficiency throughout the process.

Ultimately, organizations need to bring enhanced levels of focus, speed, and agility to their data projects amidst the aftermath of the health crisis, which will, in turn, help them unearth new sources of business value.

# Organizations Will Shift From "What Is Responsible AI?" to "How Can We Implement Responsible AI?"

Up until now, a lot of the conversations around the topic of Responsible AI have been "We haven't thought about this yet" or "How can we think about it and acknowledge the harms and impacts that AI can have on the world?" Teams might be determining how Responsible AI differs across job functions (data scientist versus an analyst, for example), agreeing on and establishing a framework for their organization's ethical rules, and putting checklists into place for Responsible AI across the AI pipeline.

In 2021, we believe we'll see more organizations put this research and work into practice. There's no longer a need to convince people that this is the way to go, as they've already gotten there. Now, it's going to be a matter of bringing organizations the expertise to implement the responsible use of AI across their existing and future use cases.

**According to David Ryan Polgar, tech ethicist and founder of All Tech is Human:**

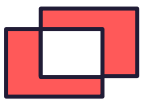
*"There is a good reason why working in Responsible AI will grow in popularity. Similar to the rise of data scientists which correlated with companies 'wrestling with information that comes in varieties and volumes never encountered before,'[7] companies are hiring responsible technologists now that they are wrestling with thorny societal impacts of artificial intelligence. The growing influence of AI on our daily lives and society at large necessitates a more thoughtful approach with its development and deployment."*

[7] *https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century*

We've reached a defining moment where Responsible AI is recognized as critical to help address the societal impacts of AI head on — organizations are creating bespoke Responsible AI roles to lead the effort from start to finish and putting processes in place to ensure the AI systems they use are human-centered and grounded in explainability, interpretability, and fairness. When scaling use cases, it is important for organizations to acknowledge that, while there is tremendous business value and potential, there is a risk of negative impact.

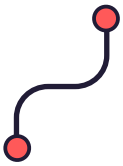
The continued rise of MLOps will help bring Responsible AI the attention it deserves, as organizations will need strong MLOps principles to practice Responsible AI and vice versa. Organizations will move beyond awareness and understanding of the realities of Responsible AI and work to broaden the scope of people participating in the creation of AI tools and systems in a diverse and inclusive way. Dataiku's vision of Responsible AI involves:

**Explainability** to ensure we don't have impenetrable black-box models but rather models that can explain themselves and tell us what's happening in order to present the results and allow individuals to say it's wrong, causes harm, or involves a risk of harm



**Creation of a culture of transparency**, where everyone involved in the data-to-insights process is involved in carrying out Responsible AI, ideally vis-à-vis proper, centralized tools

**Global governance of data and AI**



**Fairness throughout the machine learning pipeline**, from data ingestion and transformation to building models toward an explicitly stated goal to implementation and continuous monitoring

# Diversity, Equity, and Inclusion (DEI) Will Take the Spotlight for Organizations in Pursuit of AI

In addition to the global health crisis, 2020 also brought long overdue dialogue about racial injustice and inequity to the forefront, which — for the data science community — spiraled into a renewed focus on increasing diversity and making data science a more inclusive field on the whole.

Noelle Silver, Head of Instruction, Data Science, Analytics, and Full Stack Web Development at HackerU with experience at Amazon, Microsoft, and NPR, summarizes the issue nicely. She says,

*"The reality is that when we train machine learning models with a bunch of data, it's going to make predictions based on that data. If that data comes from a room of people that look the same, talk the same, act the same, they're all friends — it's not a bad scenario. In the moment, you feel like things are good. No one is really seeing any problems; you don't feel any friction. It's very misleading, especially in AI. So you go to market.*

*The problem, though, is not everyone looks like you, talks like you, or thinks like you. So even though you found a community of people that built this software that thinks the same, as soon as you go to market and someone other than that starts using it, they start to feel that friction."*

As evidenced, DEI goes far beyond just hiring and retention to close the data science talent gap. It involves diversifying the builders and developers of AI tools (establishing diversity of thought, if you will) in order to make sure the output can ultimately help a diverse range of people is just as important.

In 2021, companies will look to include people who are representative of those who will use the algorithms if they want to truly reduce bias and foster diversity. While most training datasets have been developed against a small percentage of the population, companies will now look to consider expanding their scope to design training datasets that are all-inclusive. The more inclusive the group building the AI and working with the data science tools and datasets, the less risk for bias and other negative impacts.

The notion of inclusive engineering will become more mainstream to support these diversity efforts. In order to ensure diversity is baked into their AI strategy, companies must set aside the time and resources to practice inclusive engineering. This includes, but isn't limited to, doing whatever it takes to collect and use diverse datasets. This will help companies create an experience that welcomes more people to the field, from education and hiring practices to building inclusivity-driven teams and upskilling for retention.

Just like Responsible AI is tightly intertwined with our anticipated trend about MLOps becoming even more critical, it is also highly related to DEI in the scope of data science best practices. Responsible AI goes beyond ethics and includes intentionality and accountability, which are outlined below:

| | |
|---|---|
| **Intentionality** | Ensuring that models are designed and behave in ways aligned with their purpose. This includes assurance that data used for AI projects comes from compliant and unbiased sources plus a collaborative approach to AI projects that ensures multiple checks and balances on potential model bias.<br><br>Intentionality also includes explainability, meaning the results of AI systems should be explainable by humans (ideally, not just the humans that created the system). |
| **Accountability** | Centrally controlling, managing, and auditing the Enterprise AI effort — no shadow IT[4]! Accountability is about having an overall view of which teams are using what data, how, and in which models.<br><br>It also includes the need for trust that data is reliable and being collected in accordance with regulations as well as a centralized understanding of which models are used for what business processes. This is closely tied to traceability — if something goes wrong, is it easy to find where in the pipeline it happened? |
| **+ underpinned by a human-centered philosophy** | |

A report from the Berkeley Institute for Data Science on best practices for fostering diversity and inclusion in data science says, “Having diverse and inclusive workplaces is also crucial when discussing related issues like biases in applications of data science, which can reinforce existing inequalities.”[8] The two go hand in hand. Just like the goal with Responsible AI is to fully align deliverables with intentions, ensure outputs are fit for purpose, and ensure resilience across all dimensions to support AI initiatives, organizations should design their DEI practices — both internal and external — according to an explicitly stated goal, with a plan in place to track progress, and in a way that is sustainable for the future.

[8] *https://www.researchgate.net/publication/330832444_Best_Practices_for_Fostering_Diversity_and_Inclusion_in_Data_Science_Report_from_the_BIDS_Best_Practices_in_Data_Science_Series*

# The Continued Rise of Edge Computing (and Why It Matters to the Enterprise)

The Internet of Things (IoT) offers a myriad of benefits to the enterprise, such as monitoring business processes, improving customer experiences, increasing productivity, and saving time and money. Edge computing, which puts the information processing closer to the people producing and consuming it, has gained traction recently with the ability to deploy powerful machine learning models on many cheap and constrained devices — a trend that we'll continue to see as long as IoT growth persists (and that can be transformative for IoT-savvy industries such as manufacturing, transportation, utilities, and healthcare).

By putting the computation and data storage closer to the devices data is being gathered from, one avoids latency issues that can hurt an application's performance. Perhaps the greatest benefit of edge computing, though, is the ability for organizations to process and store data more rapidly, allowing for the use of more real-time applications. With an edge computing model, for example, an algorithm can run locally on an edge server (or on an IoT device like a smartphone).

With the significant volumes of IoT data, though, comes the high cost of moving and storing data, difficulty in detecting anomalies and creating visualizations from massive datasets, and using conventional training sets large enough to provide reliable models on extreme sets. Edge AI, which we believe will continue to grow in popularity and enter mainstream adoption in 2021, means that machine learning algorithms are processed locally on a hardware device, without requiring a connection. It uses data from the device and analyzes it to produce real-time insights in milliseconds, providing added levels of security as the risk of data breaches during transit is much lower.

According to a report from Omdia, AI edge devices will increase from 161.4 million units in 2018 to 2.6 billion units worldwide annually by 2025, with top AI-enabled edge devices including mobile phones, smart speakers, PCs and tablets, head-mounted displays, and drones.[9] In practice, edge AI uses the compute on these devices to train, load, and inference machine learning models closer to the user.

In a future release of Dataiku DSS, users will be able to run real-time analytics on time-sensitive (often IoT) streaming data, so be on the lookout for that! Across all industries, this will enable data analysts and business analysts to access streaming data easily and use it in existing solutions and workflows.

*[9] https://www.businesswire.com/news/home/20180913005354/en/Artificial-Intelligence-Edge-Device-Shipments-to-Reach-2.6-Billion-Units-Annually-by-2025-According-to-Tractica#:~:text=Tractica%20forecasts%20that%20AI%20edge,units%20worldwide%20annually%20by%202025*

# Cloud Architecture as the New Normal

> **"By 2022, public cloud services will be essential for 90% of data and analytics innovation."[10]**

During the onset and continuation of the global health crisis in 2020, many organizations with on-premise architecture needed support quickly with tasks such as expanding their memory, supporting a high volume of users, versioning machine learning models, and sustaining an entirely virtual workforce. We observed that many Dataiku customers moved to the cloud in order to adapt to the new environment and compete in the next wave of data science and we believe we'll continue to see this transition throughout 2021.

To note, though, many organizations using the public cloud are using more than one cloud service provider. A hybrid cloud solution combines on-premise data centers and private cloud with one or more public cloud services, with proprietary software enabling communication between each distinct service. This approach allows businesses to have more flexibility to move workloads between cloud solutions as needs and costs fluctuate.

A hybrid cloud solution can also be relevant for organizations that need to offer services in private data centers and via an on-cloud subscription. This would allow them to build web applications and services or machine learning models and use them both on-premise and on-cloud, as well as leverage the hybrid architecture to maintain communication between applications or data flow between the cloud and on-premise infrastructures.

[10] *Gartner - Top 10 Trends in Data and Analytics for 2020, Rita Sallam, Svetlana Sicular, Pieter den Hamer, Austin Kronz, W. Roy Schulte, Erick Brethenoux, Alys Woodward, Stephen Emmott, Ehtisham Zaidi, Donald Feinberg, Mark Beyer, Rick Greenwald, Carlie Idoine, Henry Cook, Guido De Simoni, Eric Hunter, Adam Ronthal, Bettina Tratz-Ryan, Nick Heudecker, Jim Hare, Lydia Clougherty Jones, 11 May 2020*

Dataiku, which is considered a cloud-native platform because it has integrated compute execution into native hosted services from all major public cloud providers, can be helpful for seamless and secure hybrid cloud deployment. Cloud enables organizations to leverage elastic resources for AI computing, which, in turn, enables them to scale computing resources up or down according to usage and provide a unified way to run in-memory, machine learning, and Spark jobs on Kubernetes.

The cloud can help deliver stronger user experiences and improve overall business efficiencies. Particularly in a postcrisis environment, organizations will want to realize the scale associated with leveraging the elasticity of cloud resources. Dataiku's strategic partnership with Snowflake enables organizations to quickly build, deploy, and monitor all types of data science projects at scale, combining Snowflake's highly scalable computational power and processing flexibility with Dataiku's machine learning and model management capabilities.

# Machine Learning Will Be Woven Into More BI and Analytics Roles and Tools

As machine learning and AI techniques continue to evolve and gain adoption, automating tasks like data preparation, machine learning model creation behind the dashboard, and even insight discovery, the future of BI looks very different from what many organizations do now.

According to Gartner, data and analytics leaders need to "evaluate, on a regular basis, your existing analytics and BI tools and innovative startups offering new augmented and NLP-driven user experiences beyond the predefined dashboard."[11] Over the last several years, BI has given way to data science, machine learning, and AI, as organizations aimed to level up their analytics maturity, establish data-driven processes, and identify tools to help usher in these wide-scale changes (namely, data science platforms).

However, it is what has happened recently that we believe will be more prominent in 2021 and beyond — the integration of machine learning in more BI and analytics roles and tools. With the troves of quality data available today (paired with the many avenues of consumption), organically there's bound to be changes in the way BI and analytics are used across the enterprise. BI is on pace to further move away from traditional look-back analysis (taking past data to shape future decisions) and toward even more sophisticated predictive and prescriptive analytics.

For example, Dataiku's partnership with Tableau aims to move the needle on data democratization and help organizations share data stories across the business. In this new age where machine learning is highly connected to BI and analytics tools, analysts can:

- Explore data using predictive analytics and generate insights that go beyond traditional BI and analytics, such as analyzing data quality for machine learning
- Easily share the results with the rest of the organization through dashboards or visualizations to prove data-readiness for advanced machine learning
- Work on analytics and AI use cases under one roof
- Learn new skills that allow them to be involved in machine learning projects, from data access and prep to AutoML (which is a win win, as it also helps the data science team leverage the analyst's business and domain knowledge in the data science workflow)

[11] *Gartner - Top 10 Trends in Data and Analytics for 2020, Rita Sallam, Svetlana Sicular, Pieter den Hamer, Austin Kronz, W. Roy Schulte, Erick Brethenoux, Alys Woodward, Stephen Emmott, Ehtisham Zaidi, Donald Feinberg, Mark Beyer, Rick Greenwald, Carlie Idoine, Henry Cook, Guido De Simoni, Eric Hunter, Adam Ronthal, Bettina Tratz-Ryan, Nick Heudecker, Jim Hare, Lydia Clougherty Jones, 11 May 2020*

# Democratized Data Quality Will Play a Key Role in Data 4.0

*"Data quality is potentially the single most important factor in success. I say that because if you don't have accurate data, nothing else works. A lack of quality data is probably the single biggest reason that organizations fail in their data efforts."*

These words from Jeff McMillan, Chief Data and Analytics Officer at Morgan Stanley, underscore the critical importance for data quality across all stages of data science and machine learning projects.

But where does this commentary fit into the evolution of data? Data 1.0 involved the use of data in specific business applications like payroll automation, Data 2.0 used data to support enterprise-wide business processes (i.e., supply chain logistics), and Data 3.0 identified data as the valuable organizational asset that will truly transform business models and processes. It is Data 4.0, though, that we have shifted towards, as organizations across the globe have realized that scale, automation, and trust can truly only be achieved with AI and machine learning capabilities.[12]

While there are other facets of data management that will see explosive growth in Data 4.0 (such as improved data labeling techniques and more acute focus on data privacy), it is data quality that will become a higher priority for many companies in 2021, as they aim to streamline it across people, processes, and technologies. According to Gartner, "By 2022, 60% of organizations will leverage machine learning-enabled data quality technology for suggestions to reduce manual tasks for data quality improvement."[13] We anticipate a rise in organizations aiming to democratize data quality — such as soda.io, for example — dedicated to identifying data issues, alerting the proper teams, and helping them diagnose and manage data quality issues.

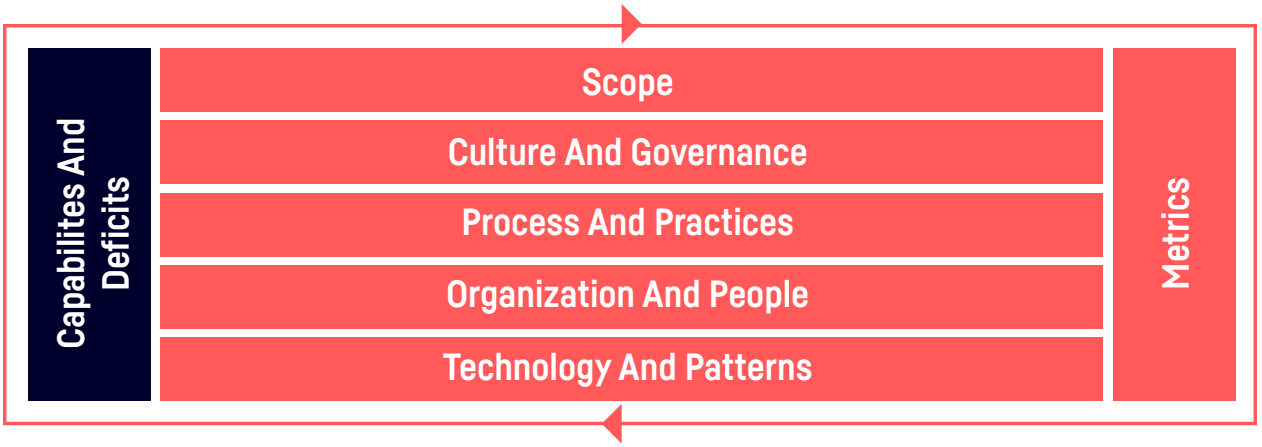
[12] *https://blogs.informatica.com/2020/05/20/data-4-0-the-soul-of-digital-transformation/*

[13] *2020, Gartner - Build a Data Quality Operating Model to Drive Data Quality Assurance - Melody Chien, Saul Judah, Ankush Jain, 29 January 2020*

Organizations across all industries face a wide range of challenges when it comes to data quality, including but not limited to unlabeled data, poorly labeled data, inconsistent or disorganized data, an inundation of data sources, a lack of tools to adequately address data quality concerns, and process bottlenecks. Many of these problems stem from a lack of data governance within the greater organization, illustrating the need for a sound data governance strategy that allows existing data storage systems to be connected in a centralized, controlled environment and provides enterprise-level security.

Without undercutting the importance of data quality, it is important to note that it is just one part of an overarching governance strategy and one part of the holistic data science and machine learning lifecycle. Given that it is so tightly woven into these different processes, it's clear that there is no one tool that easily addresses it. While AI tools and processes can certainly help usher in defining improvements to data quality, it needs to go beyond that and really be part of an organization's end-to-end strategy. For example, it's not just something to look at when building a new machine learning model, but something to consider as part of MLOps. Organizations need to understand the source, quality, and appropriateness of the data and automate data quality checks for compliance.

**CRITICAL COMPONENTS FOR DATA QUALITY OPERATING MODEL**




***Source:*** *2020, Gartner - Build a Data Quality Operating Model to Drive Data Quality Assurance - Melody Chien, Saul Judah, Ankush Jain, 29 January 2020*

# Companies Are Implementing True Self-Service Analytics Initiatives (and Will Continue to)

At Dataiku, we've been advocating for data democratization since our inception. For years we've been saying that to be a data-powered organization, everyone — no matter what their role or team is — should have appropriate access to the data they need to do their jobs (and do them more efficiently and effectively) and make impactful decisions based on that data.

While we have witnessed a gradual transition across our customer base to a more self-service model when it comes to data access (and therefore giving more people a seat at the table and enabling them to find creative ways to leverage data), it was not until 2020 that we observed organizations starting to wholly understand and implement a self-service analytics (SSA) vision.

We're seeing a wave of change (and know we will continue to in 2021 and beyond) as companies do not want to limit data and AI initiatives to a small team. We firmly believe that we'll continue to witness the explosion of data democratization (and, in turn, more projects in production and more high-value business results) in the form of companies launching SSA while they scale AI. This last bit is the key differentiator — not only are organizations grasping and implementing SSA, but they are going beyond it to operationalize their data projects to drive real business value.

Below, we'll illustrate how Dataiku customers have embraced this change, making strides to integrate SSA into their core business strategies.

## GE AVIATION

Initially, GE Aviation had difficulty scaling their data efforts due to a myriad of reasons — siloed data, no central repository for data projects, no project reuse which led to time lost, and no central vision for using data. They built The Digital League, a cross-functional team (with leaders from supply chain, finance, IT, and engineering lines of business) to firm up a central vision and one physical location for their data projects.

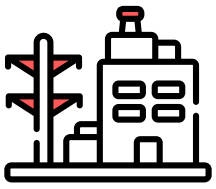
After that, GE Aviation saw success but knew there was still work to do with regard to tooling, data access and security, and streamlining the link between self-service and actually putting data into the hands of everyone. They realized they needed to scale data initiatives beyond The Digital League.

To support this growth and democratization, GE Aviation established two teams:

- The self-service data team, responsible for things like user enablement, tool administration (including Dataiku), usage monitoring, process development, and deployment of data projects
- The database admin team, responsible for ensuring data products adhere to governance policies, data used in deployed projects is used appropriately, and helping with user support

Known as Self-Service Data (SSD), GE Aviation's data initiative encompasses both SSA and operationalization for both lines of business and IT users. It allows them to use real-time data at scale to make better and faster decisions across engineering (to redesign parts and build jet engines more efficiently), supply chain (to get better insights into their shop floors and supply chain), finance (to understand key metrics like cost and cash), thve commercial group (to transform sensor data into analytics services), and more. The self-service initiative is viewed as ongoing and one that requires support and continuous improvements to achieve success.

## MULTINATIONAL ELECTRICAL SYSTEMS COMPANY



This company, which provides services across aerospace, defense, transportation, and security, was focused on long-term projects that might not realize their potential for 10 to 15 years. Resultantly, they had a gap of opportunity to drive business value from data science and machine learning initiatives and decided to create an internal group (with a Hub and Spoke operating model) to help the lines of business accelerate their innovation and technology adoption, specifically as it relates to machine learning.

The approach allowed groups like banking and payment, enterprise and cybersecurity, and IoT to embark on more sophisticated analytics, machine learning, and AI projects independently. Like that of GE Aviation, the process wasn't a "set it and forget it" one that was created overnight and never revisited — rather, it involved a calculated approach of starting small, showing value to the business and gaining buy-in early, and building on those successes to scale.

After doing that, solidifying the value with a second use case, and establishing a formal process for project evaluation and prioritizing new use cases, the team made sure to create infrastructure and processes that allowed the sub-groups to take on more use cases on their own, enabling them to have access to the data they needed for SSA and, ultimately, operationalization.

## MULTINATIONAL TELECOMMUNICATIONS COMPANY

This company, a leading provider of information and communication technology to service providers, measures the performance of its customer service representatives by their ability to successfully resolve customer support issues (i.e., they need either customer or colleague approval to close any support ticket).

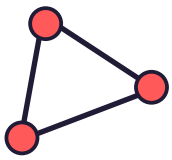
They want their customer service representatives to be valuable to their customers, while simultaneously protecting their intellectual property. For example, they're working to change their mindset and approach to share data-driven figures rather than explanations and health scores rather than groups of numbers. One specific case, though, was an intermittent issue that had been open for six months.

Traditionally, the customer service team would use Excel to bring together multiple data points and create data visualizations to try and identify the root cause of the support issue. It was not unusual for them to spend three to four days doing this. Upon using Dataiku, the customer service representative responsible for the outstanding issue (who had little data science experience) spent three hours analyzing the data to understand the potential causes of the support issue, presented the details to the account team, and received buy-in on the analysis. The representative was then able to close the support case.

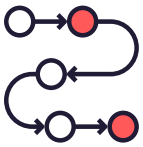
What previously took the team three to four days in Excel (without any business outcome) was resolved in four hours with Dataiku (and an improved business outcome). Moving forward, the team is beginning to focus on how many similar processes they have within their worldwide support function in order to scale this improved time to insights.

# Conclusion

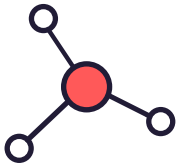
To truly extract value from their data science, machine learning, and AI investments, organizations need to embed AI methodology into the core of not only their data strategy, but their holistic business model and processes. With speed, agility, and scalability front of mind as they bounce back from a turbulent 2020, organizations are focused on figuring out new ways to ignite and expedite "business value generation" — which beyond just being a buzzword phrase, really involves a few main elements:



**Commitment to the democratization of data** throughout the enterprise (and, in turn, tools that are responsible, governable, and free of unintended bias)

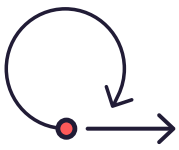


**Streamlined implementation and processes** (automating and reusing whenever possible)



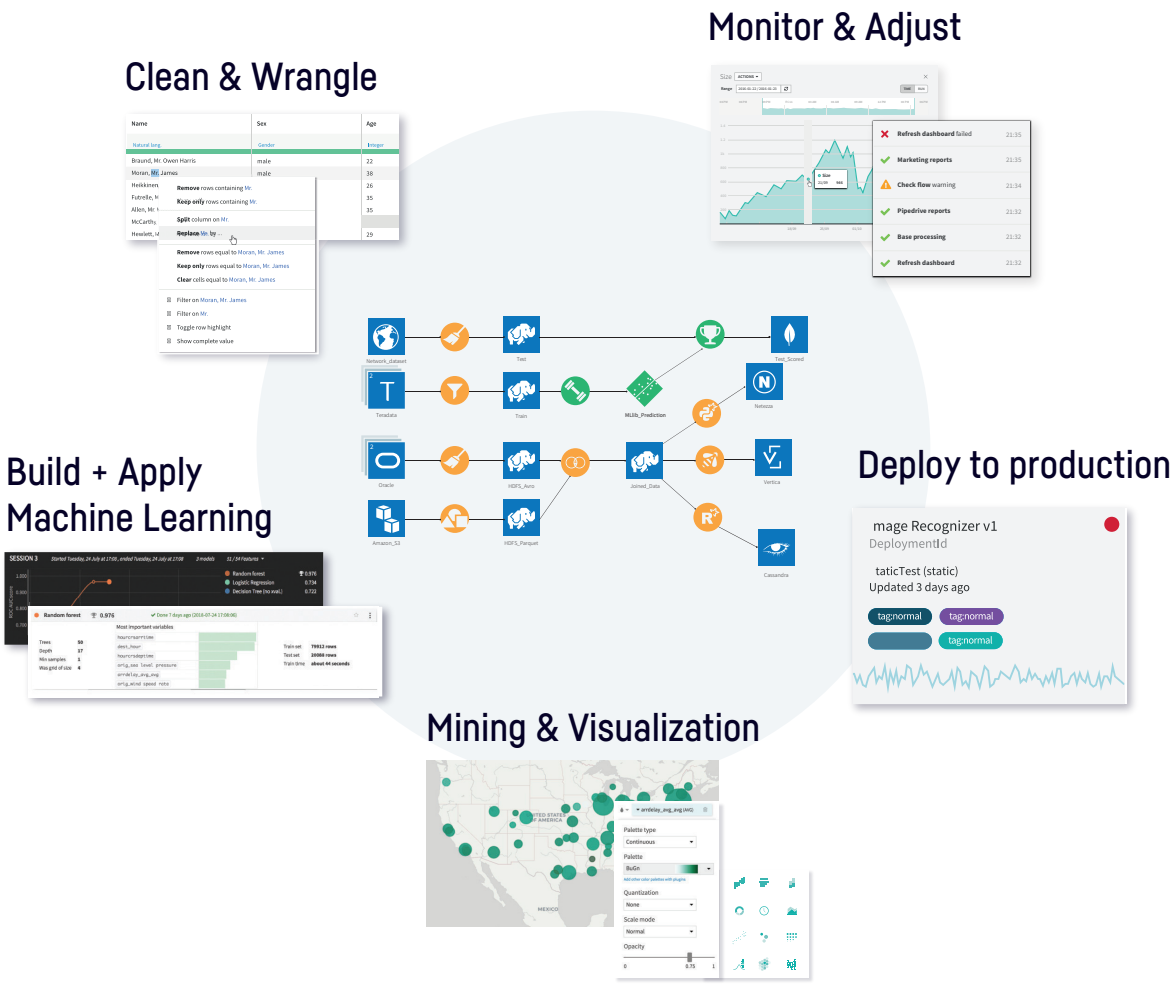
**Having a centralized repository for all data efforts**, accelerating insight extraction

**Continuing to bridge technical and domain expertise**, in order to rally the entire organization around the common goal of faster data-driven insights



**Inserting agility and elasticity** to easily monitor and adjust models as needed in times of economic flux

# Your Path to Enterprise AI




**400+**
**CUSTOMERS**

**40,000+**
**ACTIVE USERS***
*data scientists, analysts, engineers, & more

Dataiku is one of the world's leading AI and machine learning platforms, supporting agility in organizations' data efforts via collaborative, elastic, and responsible AI, all at enterprise scale. Hundreds of companies use Dataiku to underpin their essential business operations and ensure they stay relevant in a changing world.

E B O O K

www.dataiku.com