From Data Governance to AI Governance:
Pillars and Pitfalls

Capgemini

A WHITE PAPER BY CAPGEMINI + DATAIKU

www.dataiku.com

This white paper is brought to you by Dataiku and Capgemini

About the Author

Jon Howells
Lead Data Scientist @ Capgemini

Jon is a lead data scientist at Capgemini with experience in leading teams and developing data products across FMCG, Public Sector & Financial Services. Prior to this, he completed a Masters in Computational Statistics and Machine Learning at University College London.

About Capgemini

A global leader in consulting, technology services and digital transformation, Capgemini works with clients to create and deliver robust capabilities and solutions that drive impact with data. Covering everything from data strategy, information governance, data science and data engineering. Through analytics, machine learning, and AI, Capgemini equips clients with everything they need to differentiate with data.

Introduction: The Dawn of a New Governance Era

Data governance is certainly not a new concept - as long as data has been collected, companies have needed some level of policy and oversight for its management. Yet it largely stayed in the background, as businesses weren't using data at a scale that required data governance to be top of mind.

In the last few years, and certainly in the face of 2020's tumultuous turn of events, data governance has shot to the forefront of discussions both in the media and in the boardroom as businesses take their first steps towards Enterprise AI. Recent increased government involvement in data privacy (e.g. GDPR and CCPA) has no doubt played a part, as have magnified focuses on AI risks and model maintenance in the face of the rapid development of machine learning. Companies are starting to realize that data governance has never really been established in a way to handle the massive shift toward democratized machine learning required in the age of AI. And that with AI comes new governance requirements.

Today, democratization of data science across the enterprise and tools that put data into the hands of the many and not just the elite few (like data scientists or even analysts) means that companies are using more data in more ways than ever before. And that's super valuable; in fact, the businesses that have seen the most success in using data to drive the business take this approach.

But it also presents new challenges - namely that businesses' IT organizations are not able to handle the demands of data democratization, which has created a sort of power struggle between the two sides that slows down overall progression to Enterprise AI. A fundamental shift and organizational change into a new type of data governance, one that enables data use while also protecting the business from risk, is the answer to this challenge and the topic of this white paper.

Here, we'll explore how companies seriously engage in scaling on AI need to enhance their governance approach - with data governance as a cornerstone. While they may require organizational change to achieve, in the long run, it will allow for Enterprise AI at scale that is responsible and sustainable.

Why Governance?

Most enterprises today identify data governance as a very important part of their data strategy, but more often than not, it's because poor data governance is risky. And that's not a bad reason to prioritize it; after all, complying with regulations and avoiding bad actors or security concerns is critical.

However, governance programs aren't just beneficial because they keep the company safe - their effects are much wider:

Save money

- Organizations believe poor data quality is responsible for an average of $15 million per year in losses.[1]
- The cost of security breaches can also be huge; an IBM report[2] estimates the average cost of a data breach to be $3.92 million.
- Robust data governance, including data quality and security, can result in huge savings for a company.

Improve trust

- Governance, when properly implemented, can improve trust in data at all levels of an organization, allowing employees to be more confident in decisions they are making with company data.
- It can also improve trust in the analysis and models produced by data scientists, along with greater accuracy resulting from improved data quality.

Reduce risk

- Robust governance programs can reduce the risk of negative press associated with data breaches or misguided use of data (Cambridge Analytica being a clear example of where this has gone wrong).
- With increased regulation around data, the risk of fines can be incredibly damaging (GDPR being the prime example with fines up to €20 million or 4% of the annual worldwide turnover).

Figure 1: Benefits of Effective Governance

Ultimately, governance isn't about just keeping the company safe; data and AI governance are essential components to bringing the company up to today's standards, turning data and AI systems into a fundamental organizational asset. As we'll see in the next section, this includes wider use of data and democratization across the company.

[1] https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/
[2] https://www.ibm.com/security/data-breach

AI Governance, Defined

Traditionally, data governance includes the policies, roles, standards, and metrics to continuously improve the use of information that ultimately enables a company to achieve its business goals. Data governance ensures the quality and security of an organization's data by clearly defining who is responsible for what data as well as what actions they can take (using what methods).

With the rise of data science, machine learning, and AI, the opportunities for leveraging the mass amounts of data at the company's disposal have exploded, and it's tempting to think that existing data governance strategies are sufficient to sustain this increased activity. Surely it's possible to get data to data scientists and analysts as quickly as possible via a data lake, and they can wrangle it to the needs of the business?

But this thinking is flawed; in fact, the need for data governance is greater than ever as organizations worldwide make more decisions with more data. Companies without effective governance and quality controls at the top are effectively kicking the can down the road, so to speak, for the analysts, data scientists, and business users to deal with — repeatedly, and in inconsistent ways. This ultimately leads to a lack of trust at every stage of the data pipeline. If people across an organization do not trust the data, they can't possibly confidently and accurately make the right decisions.

Historically, IT organizations have addressed and been ultimately responsible for data governance. But as businesses move into the age of data democratization (where stewardship, access, and data ownership become larger questions), those IT teams have often been put in the position — incorrectly — of also taking responsibility for information governance pieces that should really be owned by business teams.

Does Your Organization Have Processes in Place to Ensure Data Projects Are Built Using Quality and Trusted Data?

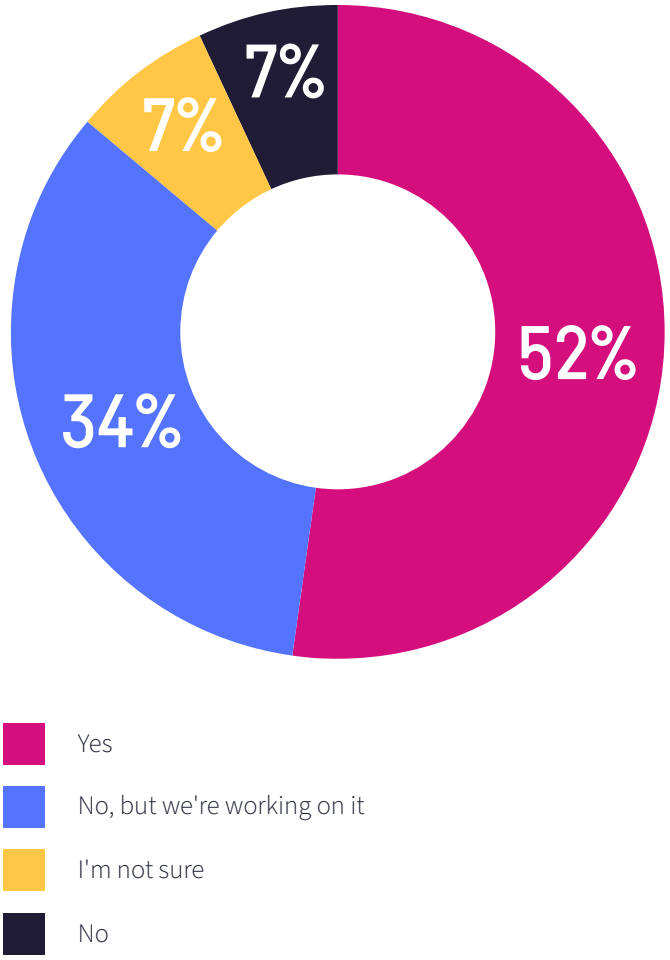


Figure 2: Processes for Quality, Trusted Data; Source: **Dataiku AI Impact Survey**

Why? Because the skill sets for each of these governance components are different. Those responsible for data governance will have expertise in data architecture, privacy, integration, and modeling. However, those on the information governance side should be business experts — they know:

- What the data is
- Where the data comes from
- How and why the data is valuable to the business
- How the data can be used in different business context
- How the data ultimately should be used, which in turn, is the crux of a good Responsible AI strategy (this is critically important and will be discussed in more detail in another section)

In short, data governance needs to be a collaboration between IT and business stakeholders.

A traditional data governance program oversees a range of activities, including data security, reference and master data management, data quality, data architecture, and metadata management (see Figure 3).

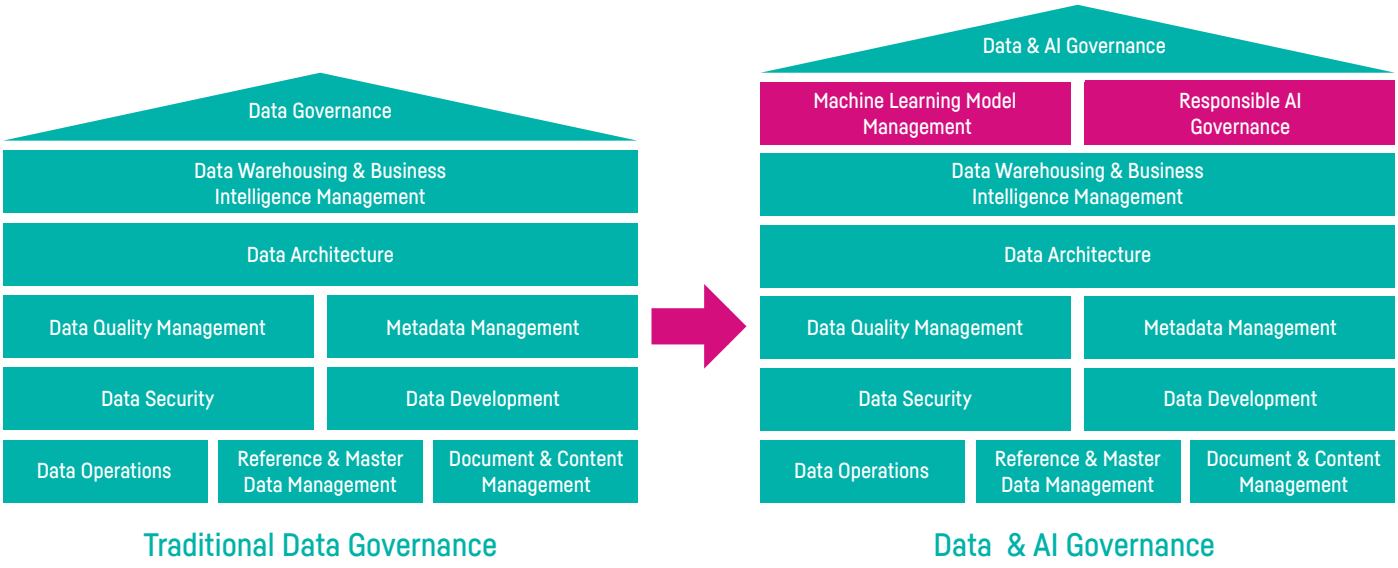


Figure 3: Moving from Traditional Data Governance to Data & AI Governance

Now with growing adoption of data science, machine learning, and AI, there are new components that should also sit under the data governance umbrella (see the right side of Figure 2). These are namely machine learning model management and Responsible AI governance, both of which will be unpacked in further detail here.

Machine Learning Model Management

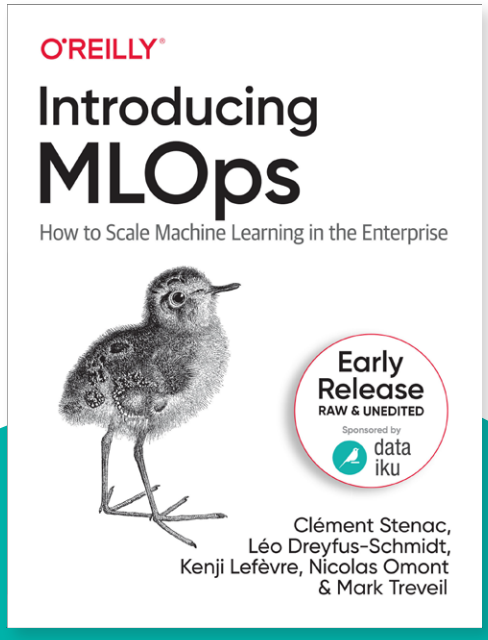
Just as the use of data is governed by a data governance program, the development and use of machine learning models in production requires clear, unambiguous policies, roles, standards, and metrics.

A robust machine learning model management program would aim to answer questions such as:

- Who is responsible for the performance and maintenance of production machine learning models?
- How are machine learning models updated and/or refreshed to account for model drift (deterioration in the model's performance)?
- What performance metrics are measured when developing and selecting models, and what level of performance is acceptable to the business?
- How are models monitored over time to detect model deterioration or unexpected, anomalous data and predictions?
- How are models audited, and are they explainable to those outside of the team developing them?

It's worth noting that machine learning model management will play an especially important role in AI governance strategies in 2020 and beyond as businesses leverage Enterprise AI to both recover from and develop systems to better adapt to future economic change.

Go Further



Responsible AI Governance

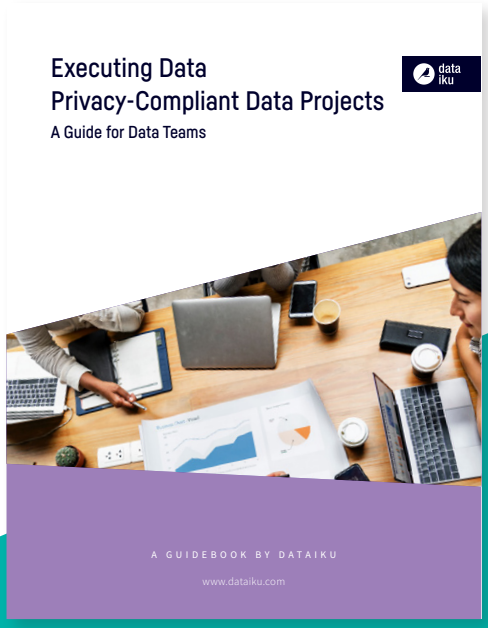
The second new aspect for a modern governance strategy is the oversight and policies around Responsible AI. While it has certainly been at the center of media attention as well as public debate, Responsible AI has also at the same time been somewhat overlooked when it comes to incorporating it concretely as part of governance programs.

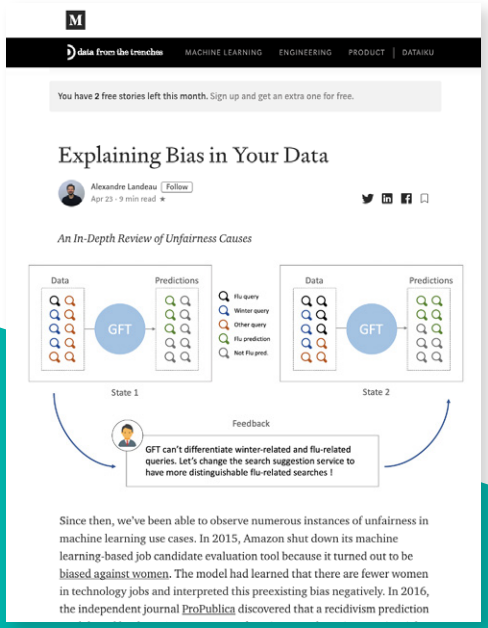
Perhaps because data science is referred to as just that — a science — there is a perception among some that AI is intrinsically objective; that is, that its recommendations, forecasts, or any other output of a machine learning model isn't subject to individuals' biases. If this were the case, then the question of responsibility would be irrelevant to AI - an algorithm would simply be an indisputable representation of reality.

This misconception is extremely dangerous not only because it is false, but also because it tends to create a false sense of comfort, diluting team and individual responsibility when it comes to AI projects. Governance around Responsible AI should aim to address this misconception, answering questions such as:

- What data is being chosen to train models, and does this data have pre-existing bias in and of itself?
- What are the protected characteristics that should be omitted from the model training process (such as ethnicity, gender, age, religion, etc.)?
- How do we account for and mitigate model bias and unfairness against certain groups?
- How do we respect the data privacy of our customers, employees, users, and citizens?
- How long can we legitimately retain data beyond its original intended use?
- Are the means by which we collect and store data in line not only with regulatory standards, but with our own company's standards?

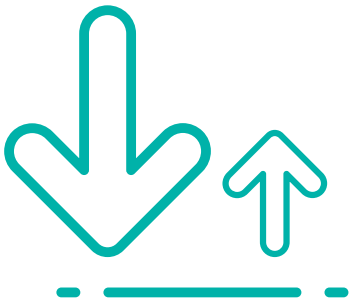
Go Further





Five Keys to Defining a Successful AI Governance Strategy

1. A Top-Down And Bottom-Up Strategy



Every AI governance program needs executive sponsorship. Without strong support from leadership, it is unlikely a company will make the right changes (which — full transparency — are often difficult changes) to improve data security, quality, and management.

At the same time, individual teams have to take collective responsibility for the data they manage and the analysis they produce. There needs to be a culture of continuous improvement and ownership of data issues. This bottom-up approach can only be achieved in tandem with top-down communications and recognition of teams that have made real improvements and can serve as an example to the rest of the organization.

2. Balance Between Governance and Enablement

Governance shouldn't be a blocker to innovation; rather, it should enable and support innovation. That means in many cases, teams need to make distinctions between proof-of-concepts, self-service data initiatives, and industrialized data products, as well as the governance needs surrounding each. Space needs to be given for exploration and experimentation, but teams also need to make a clear decision about when self-service projects or proof-of-concepts should have the funding, testing, and assurance to become an industrialized, operationalized solution.

3. Quality at its Heart

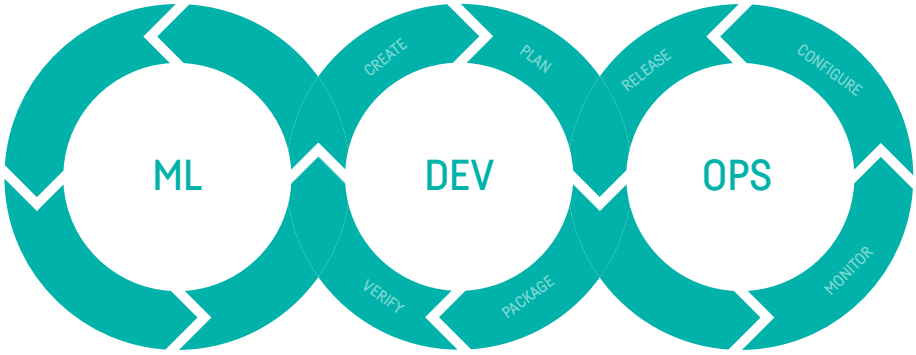
In many companies, data products produced by data science and business intelligence teams have not had the same commitment to quality as traditional software development (through movements such as extreme programming and software craftsmanship). In many ways, this arose because five to ten years ago, data science was still a relatively new discipline, and practitioners were mostly working in experimental environments, not pushing to production.

So while data science used to be the wild west, today, its adoption and importance has grown so much that standards of quality applied to software development need to be reapplied. Not only does the quality of the data itself matter now more than ever, but also data products need to have the same high standards of quality — through code review, testing and continuous integration/ continuous development (CI/CD) — that traditional software does if the insights are to be trusted and adopted by the business at scale.

4. Model Management

As machine learning and deep learning models become more widespread in the decisions made across industries, model management is becoming a key factor in any AI Governance strategy. This is especially true today as the economic climate shifts, causing massive changes in underlying data and models that degrade or drift more quickly.

Continuous monitoring, model refreshes and testing are needed to ensure the performance of models meet the needs of the business. To this end, MLOps is an attempt to take the best of DevOps processes from software development and apply them to data science.



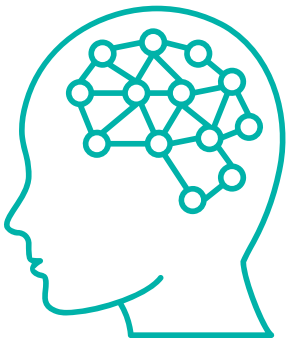
Experiment
Data Acquisition
Business Understanding
Initial Modeling

Develop
Modeling + Testing
Continuous Integration
Continuous Deployment

Operate
Continuous Delivery
Data Feedback Loop
System + Model Monitoring

Figure 4 - MLOps

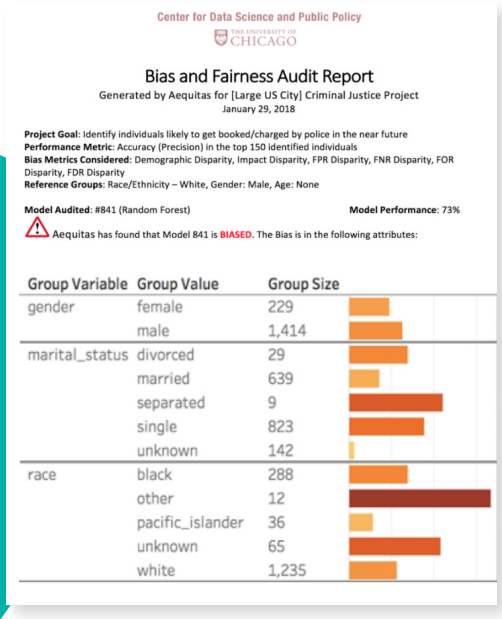
5. Transparency and Responsible AI



Even if, per the third component, data scientists write tidy code and adhere to high quality standards, they are still giving away a certain level of control to complex algorithms. In other words, it's not just about quality of data or code, but making sure that models do what they're intended to do.

There is growing scrutiny on decisions made by machine learning models, and rightly so. Models are making decisions that impact many people's lives every day, so understanding the implications of the decisions they make and making the models explainable is essential (both for the people impacted and the companies producing them).

Open source toolkits such as Aequitas[3], developed by the University of Chicago, make it simpler for machine learning developers, analysts, and policymakers to understand the types of bias that machine learning models bring.



Center for Data Science and Public Policy
THE UNIVERSITY OF CHICAGO

Bias and Fairness Audit Report

Generated by Aequitas for [Large US City] Criminal Justice Project
January 29, 2018

Project Goal: Identify individuals likely to get booked/charged by police in the near future
Performance Metric: Accuracy (Precision) in the top 150 identified individuals
Bias Metrics Considered: Demographic Disparity, Impact Disparity, FPR Disparity, FNR Disparity, FOR Disparity, FDR Disparity
Reference Groups: Race/Ethnicity – White, Gender: Male, Age: None

Model Audited: #841 (Random Forest) **Model Performance**: 73%

Aequitas has found that Model 841 is **BIASED**. The Bias is in the following attributes:

Group Variable	Group Value	Group Size
gender	female	229
	male	1,414
marital_status	divorced	29
	married	639
	separated	9
	single	823
	unknown	142
race	black	288
	other	12
	pacific_islander	36
	unknown	65
	white	1,235

Figure 5 - Example Aequitas Bias and Fairness Audit Report

Similarly, Dataiku can compute and display subpopulation analyses as a part of its end-to-end data science, machine learning, and AI platform offering, which help assess if models behave identically across subpopulations. Of course, in any case, it's always up to the individual organization to establish what's considered fair (or not) for its particular use case.

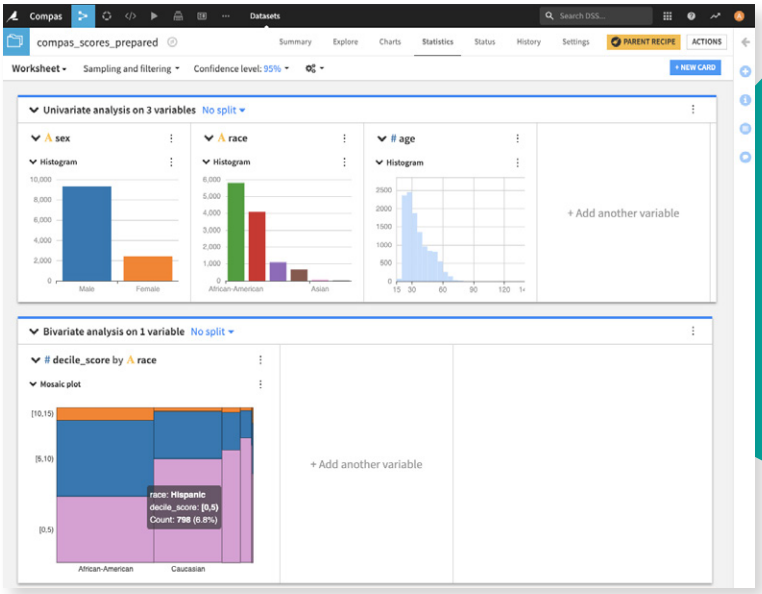


Figure 6 - Subpopulation Analysis in Dataiku

[3] http://www.datasciencepublicpolicy.org/projects/aequitas/

Data & AI Governance Pitfalls

Data and AI governance isn't easy; as mentioned in the introduction, these programs require coordination, discipline, and organizational change, all of which become even more challenging the larger the enterprise. What's more, their success is a question not just of successful processes, but a transformation of people and technology as well.

That is why despite the clear importance and tangible benefit of having an effective AI governance program, there are several pitfalls that organizations can fall into along the way that might hamper efforts:

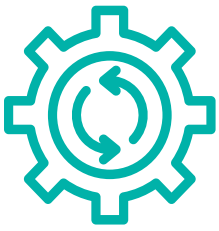
People

Lack of Senior Sponsorship

A governance program without senior sponsorship means policies without "teeth," so to speak. Data scientists, analysts, and business people will often revert to the status quo if there isn't top down castigation when data governance policies aren't adheared to and recognition for when positive steps are taken to improve data governance.

Culture

If there isn't a culture of ownership and commitment to improving the use and exploitation of data throughout the organization, it is very difficult for a data governance strategy to be effective. As the saying goes, "Culture eats strategy for breakfast." Part of the answer often comes back to senior sponsorship as well as communication and tooling.



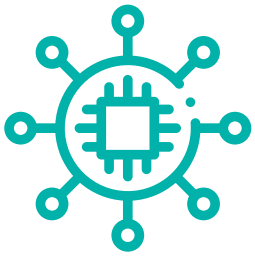
Processes

Poor Communication

A lack of clear and widespread communication around data governance policies, standards, roles, and metrics can lead to a data governance program being ineffective. If employees aren't aware or educated around what the policies and standards are, then how can they do their best to implement them?

Lack of Training & Education Resources

Training and education is a hugely important piece of good data and AI governance. It not only ensures that everyone is aware of policies but also can help explain practically why governance matters. Whether through webinars, e-learning, online documentation, mass emails, or videos, initial and continuing education should be a piece of the puzzle.



Technology

Lack of a Central Repository

A centralized, controlled environment from which all data work happens makes data and AI governance infinitely simpler. Data science, machine learning, and AI platforms can be a basis for this environment, and essential features include — at a minimum — contextualized documentation, clear delineation between projects, task organization, change management, rollback, monitoring, and enterprise-level security.

Conclusion & Next Steps

There are a few conclusions we hope you take away from this whitepaper:

1. Firstly, traditional data governance, and all the areas underneath it, are still important. Whether that be data quality, master data management or data security.
2. Secondly, data science, machine learning, and AI have added new aspects to the data governance picture that necessitate an expansion of focus and application.
3. Finally, organizations need the right sponsorship, investment, culture and communication to make sure a data governance programme is effective and leads to continuous improvement across the organization.

What does all of this mean, practically, for the enterprise? At a minimum, companies should leverage process tracking to manage governance, answering questions like "Why did we do it this way?" and "Did we do the checks and balances we were supposed to do?" Longer-term, companies should consider technology that strikes the balance between enablement and governance.

For follow-up reading, we recommend **Why Enterprises Need Data Science, Machine Learning, and AI Platforms**. Or for another high-level look at trends on how 200 CTOs, CIOs, and other IT leaders are managing systems (including governance) in the age of AI, **Trends in Enterprise Data Architecture and Model Deployment**.

References

AI Governance: A Research Agenda (University of Oxford)
https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf
Perspectives on Issues in AI Governance (Google)
https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf
Model Artificial Intelligence - Governance Framework (Singapore Digital)
https://ai.bsa.org/wp-content/uploads/2019/09/Model-AI-Framework-First-Edition.pdf

data iku

Your Path to **Enterprise AI**

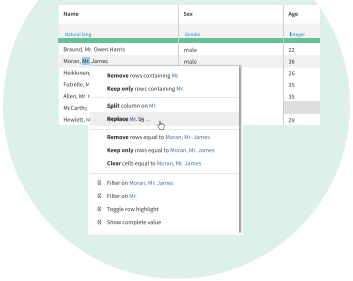
Dataiku is one of the world's leading AI and machine learning platforms, supporting agility in organizations' data efforts via collaborative, elastic, and responsible AI, all at enterprise scale. Hundreds of companies use Dataiku to underpin their essential business operations and ensure they stay relevant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

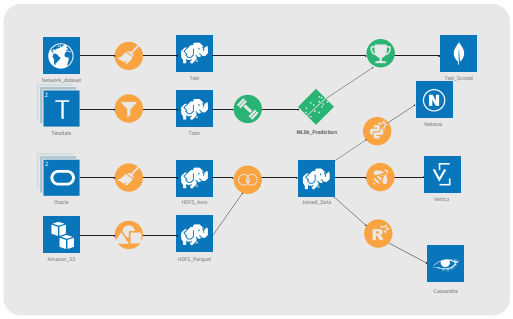
300+
CUSTOMERS

30,000+
ACTIVE USERS

*data scientists, analysts, engineers, & more

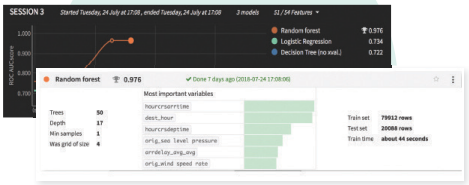
1. Clean & Wrangle



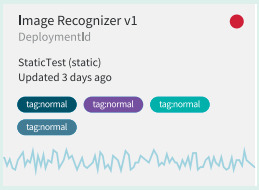


5. Monitor & Adjust

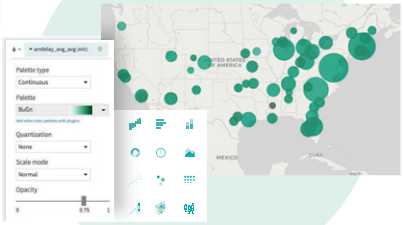
2. Build + Apply Machine Learning



4. Deploy to production



3. Mining & Visualization



WHITE PAPER

www.dataiku.com