datacamp

# Insights from Data Leaders

## A DataFramed Digest

datacamp

The past year has been marked by a great acceleration. Achieving digital transformation, becoming data-driven, and scaling data science capabilities are on every organization's mind. According to a 2021 NewVantage Partners survey, 99% of firms are reporting active investment in data science and machine learning. However, fewer than 30% of organizations are experiencing transformational business outcomes as a result of these investments, and only 24% of them claim they have created a data-driven organization.

The path to becoming a data-driven organization is a long and arduous process, as it requires investments in scaling various levers such as infrastructure, people, tools, and more. While navigating this long arduous process, it is crucial to access insights from leading experts at the forefront of data science and organizational data transformation.  This is why we relaunched the DataFramed podcast, where we sit down with industry experts and thought leaders to discuss the latest best practices and applications in data science and organizational change.

In this DataFramed Digest, you will find distilled insights from the most useful episodes of DataFramed. We will share highlights on the role of data science and literacy in preparing cities for emergencies, how to make data science work useful and transformative, the importance of data quality and trust when forging data-driven organizations, and the critical importance of building data cultures.

# Subscribe to the DataFramed Podcast

LISTEN ON Spotify

Listen on Google Podcasts

Listen on Apple Podcasts

# The DataFramed Digest

## Table of contents

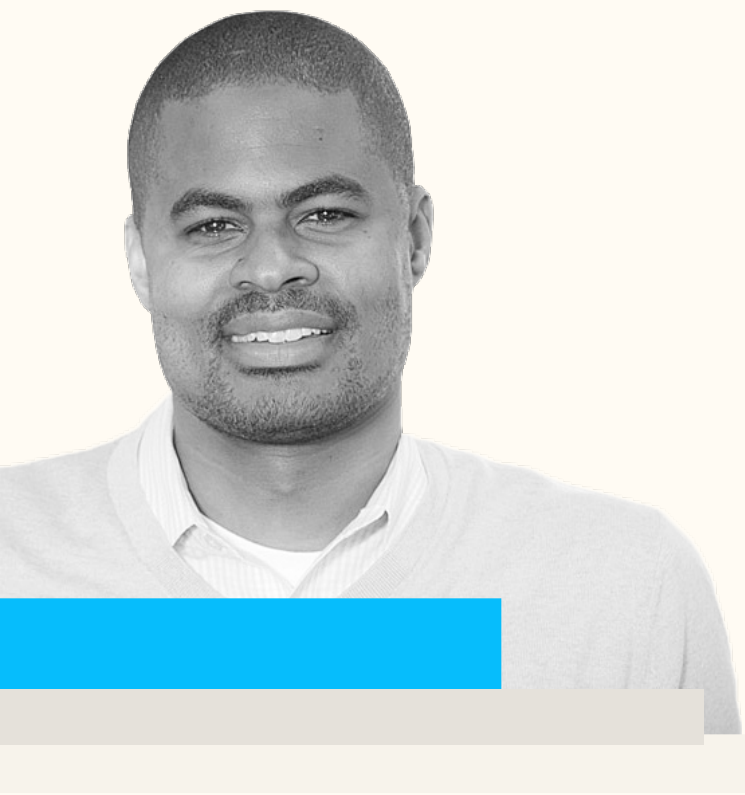
# How data science is creating smart cities

Amen Ra Mashariki is the former Chief Analytics Officer of New York City. He is currently a principal scientist at Nvidia and the Global Director of the Data Lab of the World Resources Institute. Prior to becoming Chief Analytics Officer of New York City, he was the director of the Mayor's Office of Data Analytics in New York. In 2012, Ra Mashariki was one of the 11 individuals appointed by President Obama to the 2012-2013 class of White House Fellows. Immediately after the fellowship, he was appointed the Chief Technology Officer for the Office of Personnel Management. In Creating Smart Cities with Data Science, Ra Mashariki discussed how he has made the city of New York smarter.

## What does “smart city” really mean?

The term “smart city” goes hand in hand with the city’s ability to extract data and apply these insights into strategic initiatives that improve quality of life for citizens. There are two main dimensions to smart cities. The first dimension is how to respond to external forces that affect the city, i.e., being reactive. A prime example of this is Covid-19. City leadership, citizens, and city workers had to adapt their behavior and do things they have never done before. In this case, a smart city has the ability to react in such an efficient, timely, and precise way that it minimizes any damage to infrastructure and to its residents.

The second dimension is by being proactive. Every city carries out strategic planning, for example: How can we lower greenhouse gas emissions and what mechanisms can we put in place to reach the goal of reducing emissions by 50% over the next 30 years? A smart city leverages data science and analytics to ensure the efficient and timely execution of proactive initiatives throughout the city.



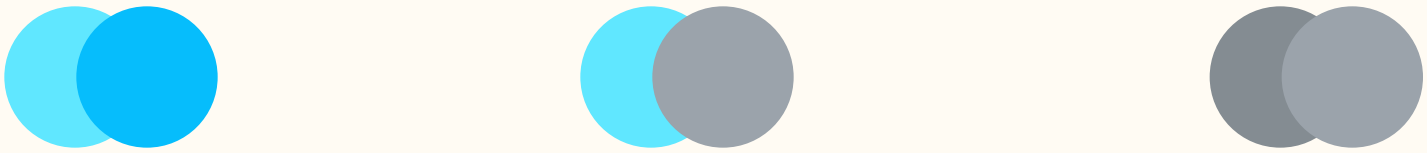
**“When I think about a ‘smart city,’ it means a city that can be reactive where it hasn't been reactive before and one that can be proactive. All in the pursuit of protecting residences and citizens, and growing their quality of life.”**

Amen Ra Mashariki, former Chief Analytics Officer of New York City

Listen to the episode →

Another way to view a smart city is by viewing it through the lens of known knowns, known unknowns, and unknown and unknowns.

## How smart cities should classify data

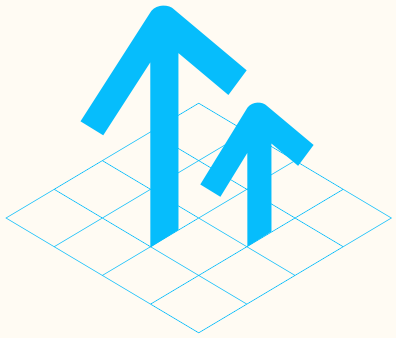


**Known Knowns** | **Known Unknowns** | **Unknown Unknowns**

A smart city is one that has a clear idea about its known knowns. Hence, it is possible to solve those problems quickly and efficiently. The known unknowns are problems that cities know they don't have their arms wrapped around, but can be thoughtful about ways to invest and learn about these unknowns. Finally, for the unknown unknowns, as Ra Mashariki said "Look, there are things that we just don't know.

> **“My office's job was to use data, to burn down the haystack, to make it easier to find the needles. Instead of having inspectors go to 400,000 homes, which would be fairly cumbersome, we say, "Here's a list of 843 landlords that you should reach out to."**

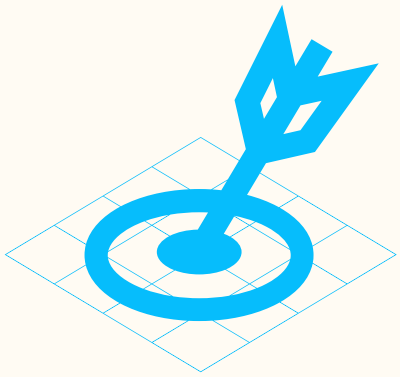
Amen Ra Mashariki former Chief Analytics Officer of New York City



## Reactive example protecting house tenants

One reactive example occurred when the New York City Council found that landlords were raising rent rates in order to get tenants to leave their apartments. This action is not only illegal, but it goes against the human rights and civil rights policies set up in New York City legislation.

The City Council reached out to Amen's office asking for support on how to use data to find landlords who are participating in illegal tenant harassment. Amen's team was able to narrow down a list of possible landlords based on historical harassment data. By leveraging this data, inspectors were able to efficiently target high likelihood cases of tenant harassment.



## Proactive example Empowering small businesses

A more proactive example was called Business Atlas. This project consisted of tying together data across the city to provide free market research information about the city. This way small business owners could use this tool to identify places where they should likely open up businesses. Amen's team combined data on crime, parks, education, census, opening hours, and more. By using this tool, anyone in the city could type an address and determine if their business type would be likely to succeed in the desired area or if there were better location options.

## The data literacy underpinning smart cities

According to Amen, city-wide data literacy is at the core of building smart cities. Whether working with spreadsheets or managing access databases, everyone engages with some kind of data, so everyone needs to know how to leverage it. Data literacy guarantees the normalization of terms and verbiage that creates a common data language across government agencies. Diving deeper, exploratory data analysis would allow individuals to at least identify potential issues with any given dataset.

To move one step forward Amen and his team organized data drills. The hurricane Sandy after-action report revealed a gap in government agencies' ability to share data in the case of an emergency. Data drills, just like fire drills, allow agencies to safely practice data sharing in case of emergencies. As Amen explained, the New York Police Department (NYPD) knows how to work well with the Fire Department in New York City (FDNY), which knows how to work well with the Sanitation Department, and so on. However, when it came to data, agencies didn't have a solid footing on how to share relevant actionable data. During data drills, Amen's team tried to understand what tools are used by each department and how to make them more compatible with each other. Then, each team ran through different emergency scenarios, such that when an emergency does occur each team knows how to share and integrate their data on a city-wide level.

**"You cannot have a truly data-driven government if only a small privileged sect of employees are data experts or data literate."**

Amen Ra Mashariki, former Chief Analytics Officer of New York City

# For greater impact, the work of data scientists needs to be useful

One of the biggest challenges data teams have been facing over the past decade is scaling the number of models they put into production, and how to attribute value to their work.

Dan Becker, CEO of decision.ai and former head of Kaggle Learn; Sergey Fogelson, VP of Data Science at Viacom; and Alessya Visnjic, CEO and co-founder of WhyLabs, came on DataFramed to share their insights on how data science is evolving to address these fundamental challenges.

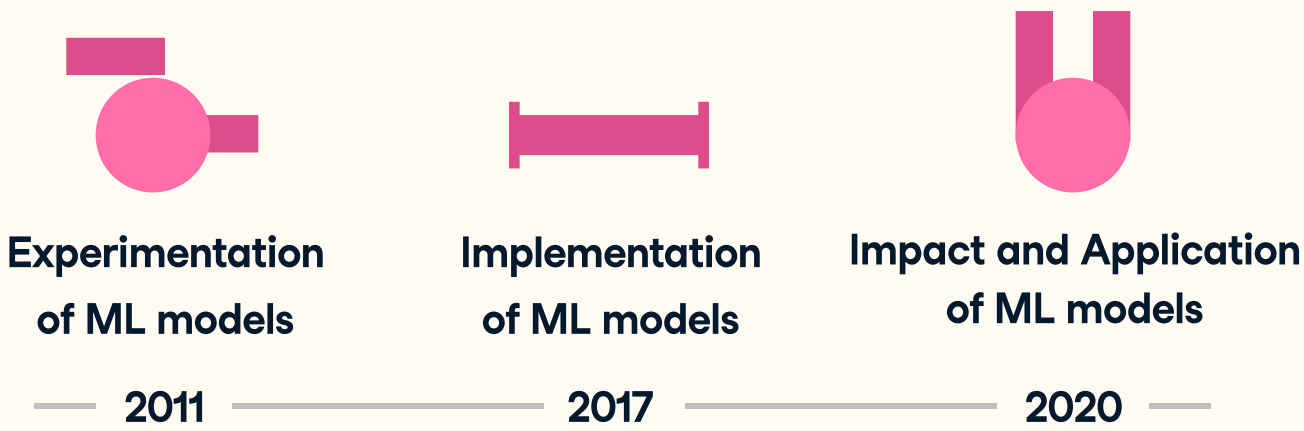
## Entering the era of usefulness in data science

In From Predictions to Decisions, Dan Becker discussed the evolution of machine learning in organizations, and how we're quitely entering a third era of machine learning within organizations. He discussed how about eight years ago, data scientists were hired for the sake of experimentation and learning what were the possibilities of machine learning within organizations. Back then, it did not matter if models got deployed or not and whether they provided value for the organization. Fast forward to 2017, and the goal then was to ensure models get deployed within business processes. Getting these models deployed requires skills both on the data engineering side and on the interpersonal side, as getting buy-in from executives is still a challenge most data teams need to overcome. During the last few years, data teams began adapting to growing expectations of organizations, where every model deployed needed to be validated through the lens of the metrics organizations care about, like revenue, retention, profit, and loss, and so on and so forth. He discussed how the growing intersections of decision sciences and machine learning, can help organizations optimize their model outputs to business decisions.

**"There's a demand more than ever for the work of data scientists to make an impact on the metrics businesses care about."**

Dan Becker, CEO of decision.ai

Listen to the episode →




## Data science has come a long way in producing value, but there's still work to be done

In the past and present of data science, Sergey Fogelson, Vice President of Data Science at Viacom, joined us to discuss the evolution of data science in the past decade. He explained the great gains the data tooling space has offered data teams to become more efficient and cost-effective. For example, earlier in his career, he described how data teams would work with large amounts of flat files, and spend large amounts of time retrieving data, whereas now most structured data can be efficiently queried from a data warehouse within minutes.

According to Fogelson, however, the most important advancement data teams have witnessed over the past few years is what he calls the "skunkworks revolution," mainly driven by orchestration frameworks like Airflow and Luigi. These tools have enabled the rapid development of data pipelines, which has allowed data teams to automate data jobs and increase the velocity of your data workflows.

Despite the great gains, however, there is still a long way for data science to mature into the value-maximizing function it's been promised to be.

**If only a small fraction of your people use SQL as well as they use Excel, the entire organization would benefit immensely."**

Sergey Fogelson, VP of Data Science at Viacom

Listen to the episode →

The largest places for improvement can be considered from two angles, technical operationalization, and organizational culture and skills. Even though orchestration and pipelining frameworks have radically operationalized data science workflows, they are still brittle in many aspects. There are still gains to be made when seamlessly updating different components of the machine learning pipeline without having to retrain new models from scratch.

What's arguably the more important area of improvement is shifting organizational culture and skills for widespread adoption of data science. Executive buy-in on data transformation initiatives is crucial for the adoption of data science throughout the organization. This is even more crucial for organizations founded before 1995, which were not founded with the internet in mind as the primary engine of value creation. More importantly, democratizing data skills is a key component for enabling everyone to make more data-driven decisions in faster time frames. Sergey argues that there are a lot of data experts within organizations that leverage Excel's VBA and long formula scripting to manipulate data. In their own rights, these people are experts in manipulating data, and with some SQL skills, they can radically open up the array of decisions they can answer with data quickly.

# Operationalizing machine learning with MLOps

In our Data Trends and Predictions 2021 report, we argued that the rise of MLOps will support the further deployment of models at scale. As the year moves by, the conversation around MLOps, DataOps, AIOps,, and ModelOps is only growing.

In Operationalizing Machine Learning with MLOps, Alessya Visnjic, CEO and co-founder of WhyLabs, joined DataFramed to discuss how implementing MLOps enables organizations to start scaling and extracting more value from data science. Drawing inspiration from DevOps in software engineering, MLOps can be considered as a set of tools, practices, techniques, and culture that ensures reliable and scalable deployment of machine learning systems.

DevOps is all about continuous integration, delivery, and deployment of software systems. MLOps extends this philosophy to machine learning systems, which need to continuously deliver the experience and predictions they were designed to deliver.

**“Data scientists need to think about their models in post-production because only once the model is in production is when it starts generating value.”**

Alessya Visnjic, CEO and co-founder of WhyLabs

Listen to the episode →

Commonly, organizations looked at deploying machine learning models as the last step in a data project. This mindset deprioritizes the importance of maintaining, monitoring, and refining models post-production. However, Visnjic says that looking at what happens to your model post-production is pretty much the most important step, as that's when it starts providing value.

For this shift to happen, Alessya recommends organizations need to further strengthen their data literacy and culture. This will enable everyone involved in the machine learning value chain to start looking at machine learning solutions in a data-centric fashion that enables them to critically improve and iterate on models post-production.

According to Visnjic, organizations should start their MLOps journey as early as possible, especially when they start adopting or building machine learning technology. Before implementing these models, organizations should consider five different pillars that would guide their blueprint: reproducibility, transparency, robustness, quality, and ownership. Afterwards, organizations must chart their own roadmap by looking at their current DevOps team and asking themselves: What are all of the activities, processes, and mechanisms that this organization already implements for traditional software, and then extend their thinking to include data and to include machine learning non-deterministic probabilistic applications.

# Trust in data will make or break your data transformation ambitions

Barr Moses, CEO, and co-founder of Monte Carlo discussed the importance of data quality and how data observability creates trust in data throughout the organization. Barr argues that some companies approach data transformation and become data-driven as a surface-level one-off project. This is typically seen in organizations that have tried to bridge the data maturity gap by merely hiring data scientists and investing in tools. According to Barr, becoming data-driven is a transformational process where organizations require both a mindset, organizational, and cultural shift that need to be supported by technology investments.

There are two main use cases for leveraging value with data within the organization. Companies use data to drive the development of products and to make data-driven decisions at scale. So, how do we actually create a data-driven culture? It starts with collecting and storing data and making sure everyone is equipped with the skills and access to make data-driven decisions.

This extends beyond the data team and requires everyone within the organization to collaborate and work with the data team to be part of the organization's data transformation.

**"Becoming truly data-driven is when marketing, sales, customer success, and product are all very strong customers of the data organization and work hand in hand with them to make these decisions and to power the business. And that's really when you see the competitive advantage of becoming data-driven."**

Barr Moses, CEO and co-founder of Monte Carlo

Listen to the episode →

## How to make sure you can trust the health of your data

When scaling data science within the organization, it is really important to maintain an alignment between the data and executive team, which comes down to data literacy. However, data literacy does not mean that everyone must learn SQL or R. Data literacy means empowering each team within the company to engage and draw insights from data to help achieve the organization's goals.

A second factor is data observability, which is the ability to determine the health of the data by observing its output. Thinking about observability in the context of software engineering, there has been an emerging need to manage infrastructure downtime. Tools such as New Relic or App Dynamics have been developed to help software engineers manage the health of their applications.

In relation to data science, there are no tools to manage the health of your data. However, we still keep ourselves accountable to make sure we have reliable and trusted data. To achieve this, Barr broke down data observability into five core pillars to allow data teams to maintain a strong sense of the data's health.

### Five pillars of data quality

**Freshness**
Timeliness of the data

**Volume**
Amount of data acquired

**Distribution**
Gives you insights about the data's accepted range

**Schema**
Monitoring changes in the data

**Lineage**
Upstreams and downstream dependencies of the data

By combining these five pillars, organizations can be confident about the health of the data, making it easy to understand and quantify the impact of data quality with specific business goals. Hence, best-in-class data teams are not only about great pipelines and infrastructure, but also high-quality data management and governance.

# The path to building data cultures

A common theme found throughout this digest is how data culture underpins extracting value from data science.

Sudaman Thoppan Mohanchandralal, Regional Chief Data and Analytics Officer at Allianz Benelux, joined DataFramed to discuss the impact of building a data culture within an organization. He explained his best practices for building data cultures and how data culture can be broken down into malleable organizational habits.

In any organization, habits define the culture. Sudaman explained that building data culture starts with the reinforcement of habits within the organization.  Encouraging data-related habits and routines will transform the organization's culture into a data culture.

By breaking down a habit into its main elements, we are able to understand how to build data habits within our organization. A habit always starts with a cue or trigger, then a routine, and finally, you receive a reward. All decisions in an organization are instigated by a specific reward and a trigger. However, Mohanchandralal explains that the routine to obtain the reward can be changed to improve the organization's habits to a more data-driven culture.

**"Data culture is not just an option to succeed in data analytics initiatives, it is business-critical."**

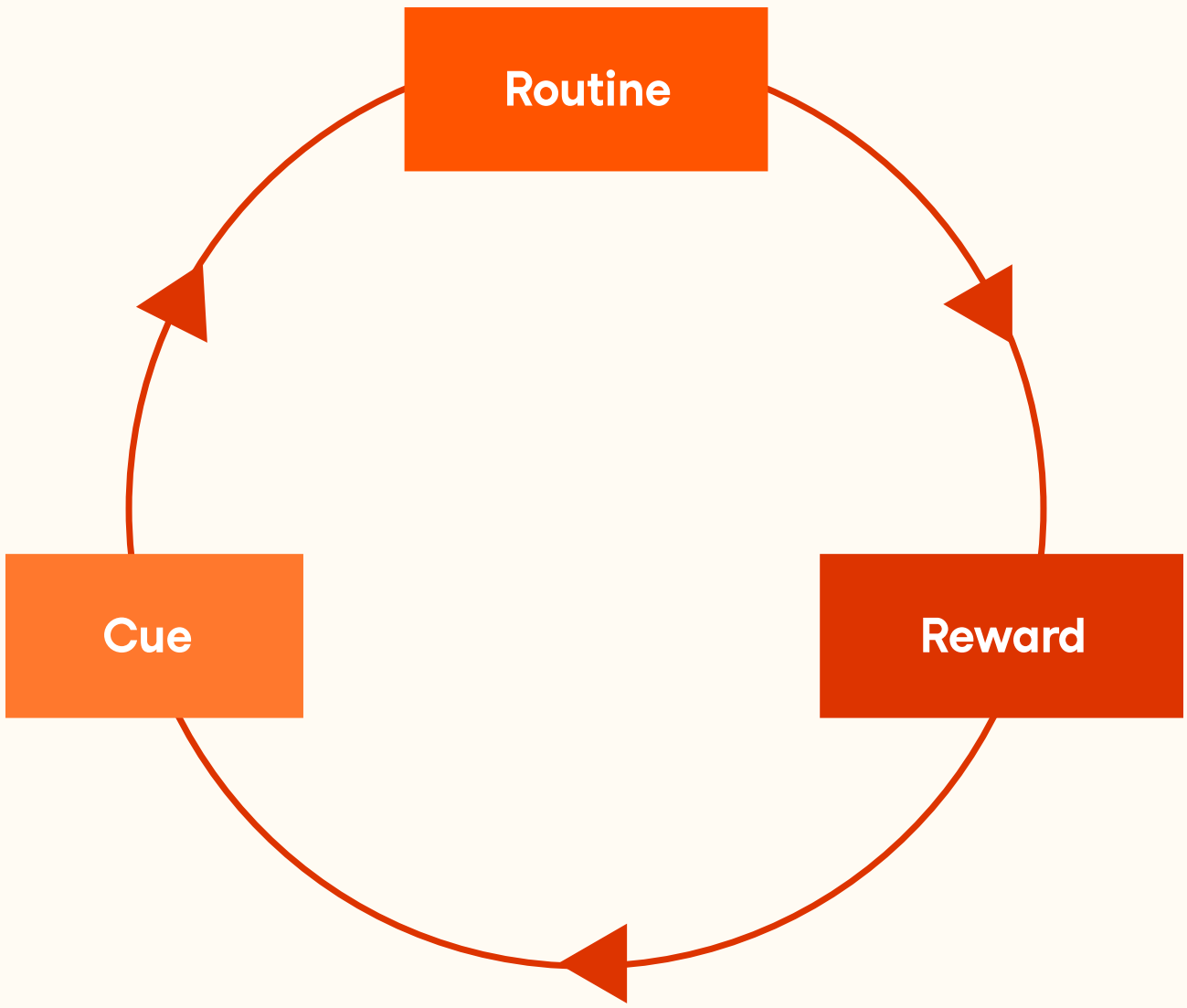
Sudaman Thoppan Mohanchandralal, Regional Chief Data and Analytics Officer at Allianz Benelux

Listen to the episode →

Mohanchandralal mentioned that “the golden rule of habit change is to keep the cue, provide the same reward, but insert a new routine.” While evaluating the decision-making habits at Allianz Benelux, the Sales and Distribution Team realized that data was not as heavily used within the broker steering process. In this case, the routine was steering the broker, the reward was the business, and the cue was the timeframe and seasonality. The team was able to improve the cue and reward by implementing more data-driven decisions within the broker steering.

In order to effectively change the routine and empower your team to achieve a high level of data culture, your organization must keep in mind three key elements:

## The habit cycle



## Key elements for a successful data culture

**Common language**

Combine data literacy and subject matter expertise.

**Data governance**

Determine responsibilities for data ownership within the organization.

**Continuous measuring**

Implement a continuous learning loop for constant culture reassessment.

Organizations should keep in mind that achieving a data culture within the organization is a continuous and iterative process. It is important to continuously reassess the state of the culture across different teams. Nevertheless, before the organization decides to allocate time, effort, and money into building a data culture, it is important to consider the two main hurdles that teams usually face during the transition.

### Hurdle #1

## Prioritization

Managers must know where does data culture is within the list of priorities, as organizations always have limited resources

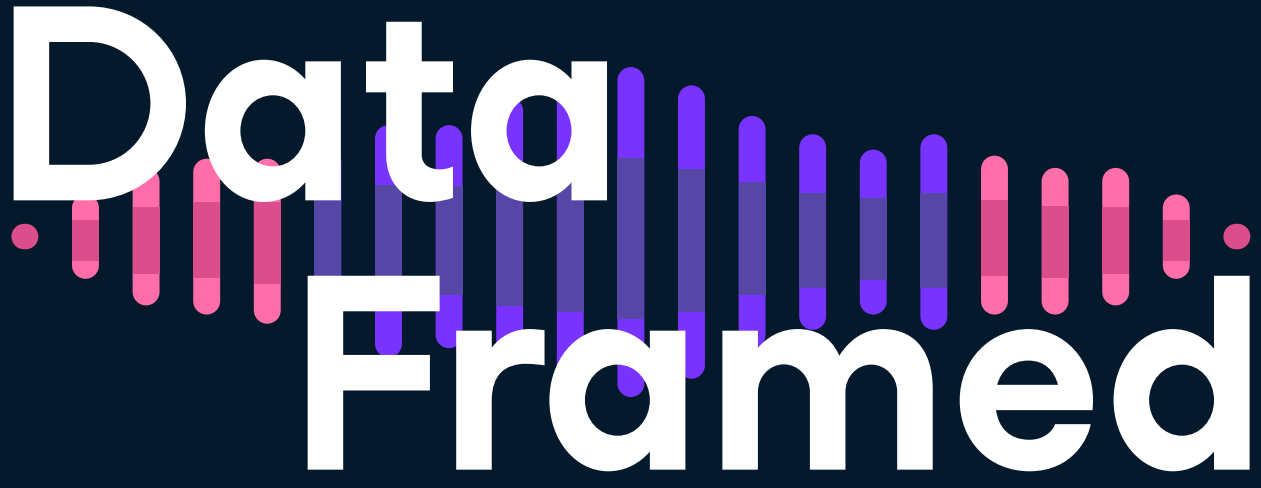
### Hurdle #2

## Sustained investments

Managers must consider the transition time and maintain consistent allocation of time and resources

datacamp

# Subscribe to DataFramed and never miss out




LISTEN ON Spotify

Listen on Google Podcasts

Listen on Apple Podcasts

# Additional resources and tools

We are working hard to make sure we build a data literate world! We've created several resources that you may find useful in your journey to becoming more data-driven

## Webinars

- The Learning Leader's Guide to Data Fluency
- Developing an AI Literate Nation
- Democratizing Data in Government Agencies

## White Papers

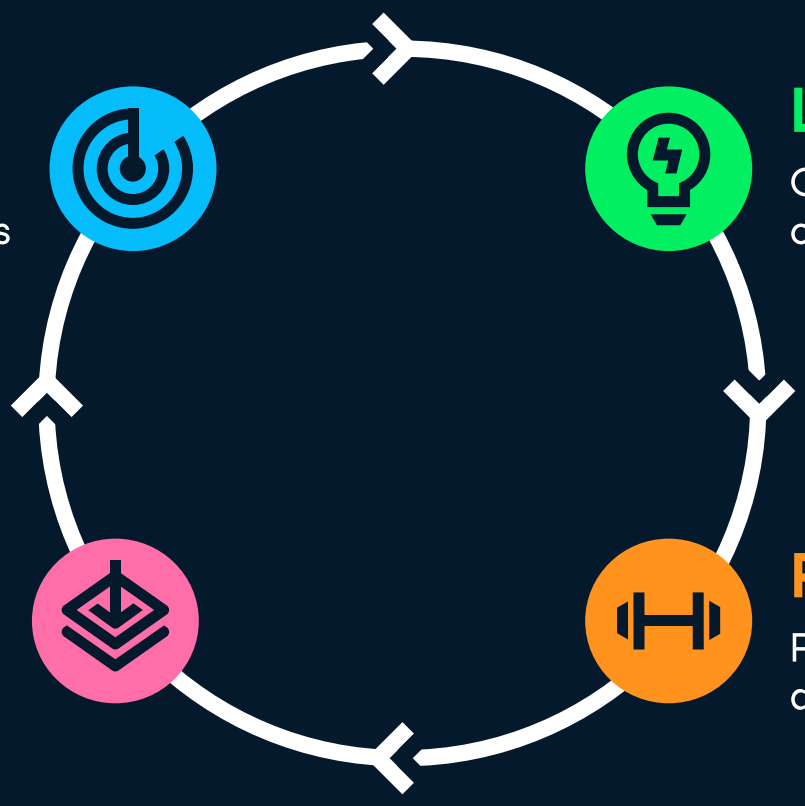
- Data Literacy for Responsible AI
- Data Leader's Guide to Upskilling
- Your Organization's Guide to Data Maturity

## Blogs

- Four Ways Your Team Can Start Leveraging Data Science
- How To Manage AI Projects Effectively
- How to Build a Winning Data Team

# Upskilling with DataCamp

DataCamp's proven learning methodology for learning open source data science has helped many organizations to become data-driven.




## 80% of the Fortune 100 are using DataCamp

Request  a personalized demo from one of our team members! Otherwise, if you feel adventurous you can get started right now.

**Request a Demo** | **Get Started**