datacamp

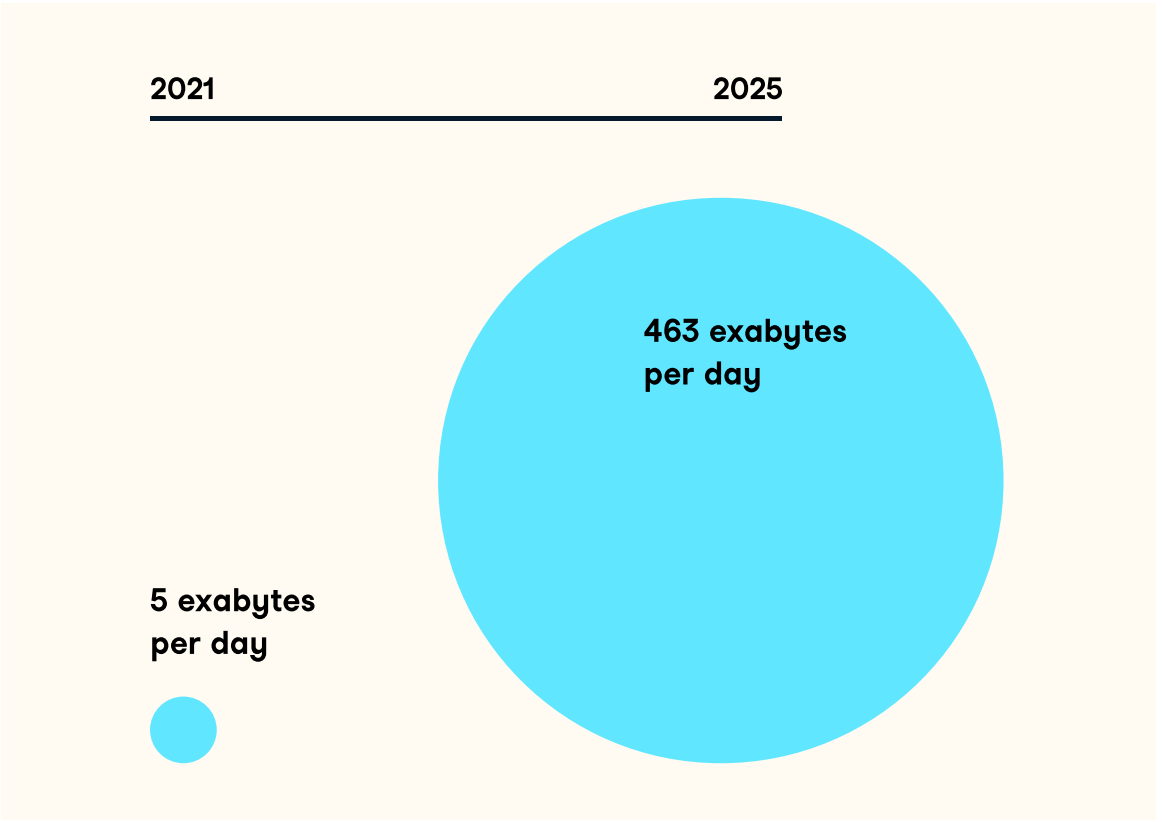
2022 Data trends and predictions

In an age of rapid digitization, data literacy is shaping the dawn of a new era

COVID-19 has led to a massive leap in digital transformation across all industries.

During the pandemic, companies reacted to the dramatic shift of consumers towards digital channels by digitizing their internal operations, supply-chain interactions, and their product offerings. A survey by McKinsey revealed that COVID-19 accelerated the adoption of digital technology by several years, and those changes are here to stay.

Now more than ever, this rapid digitization is leading to an explosion of data. According to Accenture, by 2025, the world will produce 463 exabytes of data in a day. This is 92 times the amount generated daily in 2021, and approximately the size of the entire world's storage capacity in 2009.



Companies sitting on treasure troves of data are feeling the pressure to become data-driven. Not only are data-driven organizations 23 times more likely to acquire customers, but they are also 19 times more likely to be profitable. This pressure will animate much of the stories we'll see in the upcoming decade. Similar to how organizations adapted to the rise of personal computing and the internet over the past 40 years, they will need to adapt to a new era that will be defined by data literacy.

In 2022, companies embarking on data transformation programs in hopes of reaping the rewards of a data-driven culture will find themselves adopting new tools, embracing new technologies, new ways of work, and embracing talent and culture transformation at scale.

As such, here are nine data science trends and predictions that you can expect in 2022.

Contents

datacamp

01 Organizations accelerate culture transformation programs

Companies hoping to collect dividends from their wealth of data might be disappointed when their effort to become data-driven falls flat. According to the New Vantage Partners 2021 Survey, the lack of data culture is the culprit that impedes a large majority of organizations from becoming data-driven.

29%

of organizations are experiencing transformational outcomes with data science and AI

99%

of respondents who are not experiencing transformational outcomes cite lack of culture and skills as the biggest impediment for change

NewVantage Partners CXO Survey 2021

Defined as "the collective behaviors and beliefs of people who value, practice, and encourage the use of data to improve decision-making", data culture is the basis to capture opportunities and value from a company's data.

Building an organization-wide data culture takes time and discipline. It involves setting up proper data governance, investing in data literacy programs, and training or sourcing talent with specialized data skill sets. More importantly, it requires data to be integrated seamlessly into business operations and processes so that data-driven decision-making becomes the norm.

The onus of building a data culture is on the organization's Chief Data Officer (CDO), who is in the driver's seat of the culture transformation program. The CDO leads by example, starting a virtuous cycle of data-driven decision-making that propagates through the organization.

This year, we highlighted how Allianz Benelux is investing in data literacy and culture transformation programs. As a strong data culture becomes ingrained, data-driven decision-making takes over gut-based guesswork. Similarly, Gulf Bank created a data ambassador program to propagate data-driven decision-making throughout different departments.

In 2022, expect organizations to double-down on these culture transformation programs, by creating dedicated roles and departments for culture transformation, and by doubling down on supporting levers that help scale a data culture.

> **Building a data culture is not an option; it is business-critical."**

Building Data Cultures Webinar

Sudaman Thoppan Mohanchandralal
Regional Chief Data & Analytics Officer
Allianz Benelux

02 Organizations will scale data governance

Today, bad data costs $3.1 trillion per year to the US economy alone, lending truth to the adage of "garbage in, garbage out".

Data quality is the ingredient to establish the data trust needed for making data-driven decisions. Organizations are further internalizing the importance of data quality an will look to further establish proper data governance, which involves managing the availability, usability, integrity, and security of its data.

- The size of the data governance market will experience

  5X growth by 2025

  Data Goverance Market - Growth, trends, COVID-19 impact, and forecasts (2021 - 2026)

Data governance is key to ensuring that internal stakeholders can access quality data that is **compliant, actionable, and usable**. Conversely, the need for data governance grows in tandem with the demand for self-serve analytics.

Today, companies measure data quality on its accuracy, completeness, validity, and timeliness. As the data grows in volume and complexity, data practitioners find it ever more challenging to monitor these metrics and maintain high-quality data pipelines in real-time.

As such, expect the solidification and emergence of new domains and categories such as Data observability, which address that exact pain point. Simply put, data observability is a collection of technologies that identify, troubleshoot, and resolve data issues in near real-time. With multiple start-ups (e.g., Monte Carlo, Databand, and Observe.ai) offering data observability solutions, implementing real-time observability is now possible.

Data observability allows data teams to detect system-level changes to the data set and potentially catch data quality issues as early as possible. This translates to reduced data downtime and higher data quality.

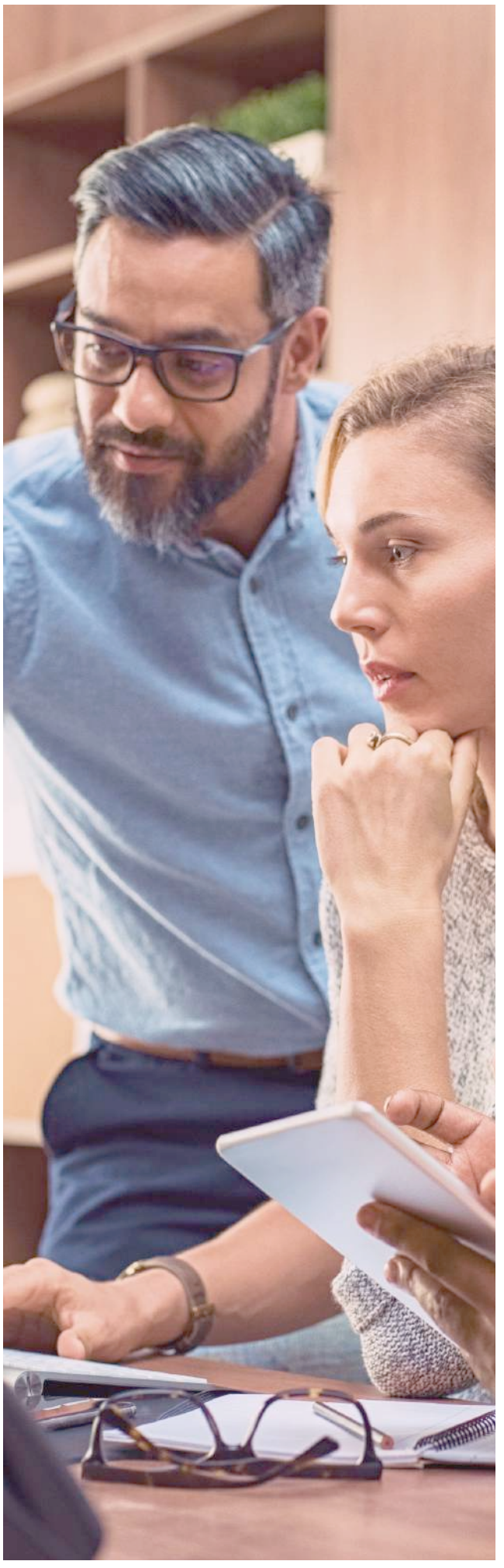
As companies seek to amass more data and manage their data quality at scale, data observability will continue to gain momentum in 2022.

“The number one challenge for companies that want to become truly data-driven is to build the trust in data... for companies that truly want to become data-driven, data quality and data observability have to be a top priority.”

[Creating Trust in Data with Data Observability]

Barr Moses

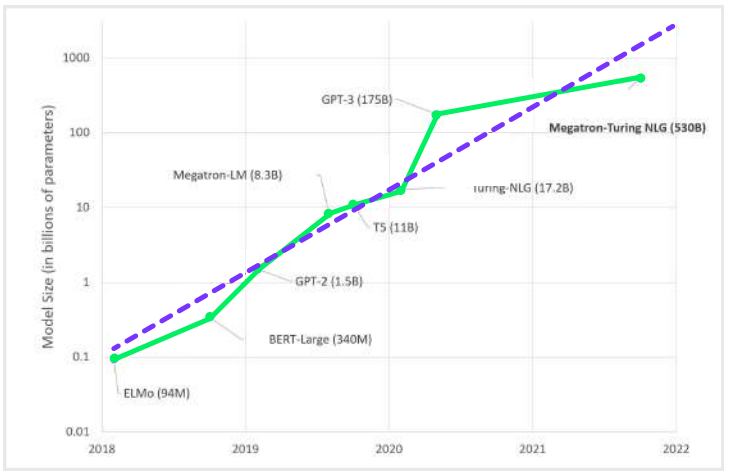
CEO, and co-founder
Monte Carlo



03 NLP ushers in a new generation of low-code data tools

As evidenced by the rapid development of large language models and the rise of startups such as hugging face, the past few years saw the NLP space progress tremendously.

Over the next year and more, the increasing scalability of such models will generate an arms race for the biggest and most effective large language model. For example, 2020's GPT-3 was hailed as the largest and most powerful generative language model to date, only to be recently out-sized by the Megatron-Turing NLG 530B model, which is almost four times bigger than GPT-3.



Large language models get larger over the years

As large language models grow, they develop groundbreaking capabilities with transformative impacts on science, society, and organizations. For example, GPT-3 and Megatron-Turing NLG demonstrated their ability to perform tasks that they are not explicitly trained on and generate alternative types of texts, including computer code and guitar tabs. It is only a matter of time before the next large language model pushes the boundary of what natural language processing can do.

Arguably the most impactful use-case of large language models for organizations will be the rise of NLP-based low-code and no-code tools.

For instance, the integration of GPT-3 into Microsoft's Power App will empower non-technical users to build an app using conversational language. OpenAI's codex will increase the productivity of data scientists and empower them to focus on the data problems they're trying to solve, and not on boilerplate code. As no-code and low-code tools lower the barriers to coding, expect to see a rising influx of citizen developers and citizen data scientists within organizations.

04 L&D becomes a fabric of company culture

According to a 2021 PwC survey, 74% of CEOs are concerned by the lack of key skills within their organizations. This is especially telling in an age where health crises, uncertain economic growth, and technological changes are driving systematic uncertainty for organizations. Learning and development is proving to be the silver bullet to the lack of key skills—with the World Economic Forum predicting a 38% of additional global GDP gained from upskilling by 2030.

Over the past year, the shift to digital drastically changed the views of executives and leaders on learning and development. As learning programs helped employees stay productive at home during the pandemic, many C-suite began to recognize the value of L&D. As a result, the percentage of L&D departments with leadership reporting to the C-Suite climbed from 24% at the start of the pandemic (March 2020), to 63% a year later (March 2021).

As the world settles into the new normal of remote work and the need for culture and skill transformation rises, L&D budgets will be dedicated to creating vibrant learning ecosystems that provide effective learning and communities of practice to drive the era of data literacy.

As such, look for companies to scale their upskilling and reskilling programs by building internal data science skill academies that drive a learning culture within the organization.

> **“As we’re entering the year, organizations are facing the need for the biggest upskilling and reskilling programs ever.”**
>
> Learning and Development in for the Data-Driven Age

Marcus Robertson
Global Curriculum Lead
NatWest Group

05 MLOps will continue to mature within organizations

MLOps entered the vernacular of the machine learning community in 2019 and gained tremendous traction in the years since, and rightfully so. Creating proof-of-concept with machine learning today is simpler than ever, but 87% of models never make it into production. Companies wanting to extract value from machine learning cannot do so at scale without production-level AI systems. This is where MLOps comes into play.

MLOps is a set of practices that combine machine learning, data engineering, and DevOps skillset. More broadly, it is a combination of tooling, culture, and processes that help ensure machine learning models continuously deliver the experience that they were designed to deliver. It involves standardized processes that automate machine learning deployment workflows, from data management to model training and deployment. For instance, Uber has successfully implemented MLOps principles to provide real-time predictions at scale.

> **“Data scientists will have to think about their models in post-production since that’s when machine learning starts generating value.”**
>
> Operationalizing Machine Learning with MLOps

Alessya Visjnic
CEO and Founder
WhyLabs

A report by Cognilytics estimated that the MLOps market will be valued at $126.1 billion by 2025. The MLOps market is currently dominated by startups, which have collectively raised $3.4 billion. KubeFlow, Algorithmia, and MLFlow are a few examples of MLOps tools at the frontier of scaling enterprise-level AI.

In the next year and beyond, the MLOps stack will grow in maturity and become more standardized, and more data science teams will begin to adopt MLOps for deploying, managing, and monitoring their machine learning models in production.

06 Responsible AI becomes more operationalized

A 2021 investigation by the Markup reveals that 80% of black mortgage applications to be unfairly discriminated against due to bias in AI algorithms.

Separately, an internal study by Twitter in 2021 reveals that its existing image-cropping algorithm favors men over women. These are but two examples of AI learning and amplifying existing human biases in 2021 alone.

Companies seeking to extract value from AI must ensure that their implementation of AI remains fair and responsible. Companies that fail to do so are doing themselves a disservice with massive damage to their brands and reputation. More importantly, not only can biases in AI become a PR disaster—but they also reproduce existing prejudices and exacerbate inequality.

Regulators globally are paying more attention to AI systems. The European Union is the first governmental body to issue a draft of comprehensive AI regulations. Starting 2020, any automated decision-making delivered to the Canadian federal government must go through a thorough impact assessment. It is only a matter of time before other nations would follow suit and mandate that any use of AI is responsible.

That is why companies are increasingly adopting and operationalizing principles of Responsible AI. It ensures that AI remains fair, interpretable, privacy-preserving, and secure. Google, Microsoft, and Capgemini published their AI Code of Ethics and implemented steps towards such goals. Companies looking to operationalize principles of Responsible AI should look into adopting frameworks, tools, and processes. An example of a framework is PwC's Responsible AI Toolkit, which addresses the various dimensions of Responsible AI, including governance, privacy, security, safety, and robustness, among others.

We need to be thinking about ethics, risk identification, and accountability. We need to make sure that we not only enjoy the benefits of AI, but also remain in control, understand what can go wrong, and how best to deal with it with responsible AI."

The Future of Responsible AI

Maria Luciana Axente

Responsible AI and AI for Good Lead
PwC UK

07 The rise of the data mesh

Today, the data lake is one of the most widely adopted data architectures. Yet, its shortcomings have prompted the rise of the data mesh.

The data lake is centralized, has highly coupled pipeline architectures, and is operated by siloed data engineers. Such centralized architecture cannot meet the needs of enterprises with rich business domains and diverse data sources. Not only that, the high degree of coupling between various stages of a data pipeline results in bottlenecked existing infrastructures struggling to keep up with new data sources. Moreover, data engineers are often disconnected from users.

That is why some companies are moving away from data lakes in favor of data mesh. A relatively new concept coined by Zhamak Dehghani, a data mesh is a data architecture that addresses the limitations of existing data lakes. A monolithic data infrastructure has a central data lake, responsible for the consumption, storage, transformation, and output of all data. In contrast, a data mesh has distributed "data products" – each handled by a cross-functional team of data engineers and product owners.

Organizations that implement data mesh can benefit from the higher speed of data delivery, stronger data governance, and greater business domain agility. This is especially true for large enterprises with varied data sources, large data teams, and rich data domains.

Data teams at Zalando, Intuit, and JPMorgan Chase have already implemented the data mesh architecture. As the benefits of a data mesh become apparent, we expect more enterprises to experiment with the data mesh architecture. Since implementing a distributed data mesh requires a paradigm shift from existing architectures, such a change will be gradual over 2022 and beyond.

08 New generation of tooling will improve the data team's productivity

As data scientists become more sought after and the modern data stack evolves, we'll see a new generation of tools that will increase the data team's productivity. These tools can eliminate the need for manual work, freeing up their time to perform higher-value non-routine tasks like data pre-processing, feature engineering, and model deployment.

AutoML Tools

Many machine learning projects share similar processes of hyperparameter tuning and model selection. Recognizing that these tasks are repetitive and time-consuming, creators of AutoML tools promise to automate these tasks efficiently.

Data science teams are spoilt for choice for AutoML tools, each with its strengths. While some like H2O AutoML and Python's Auto sklearn are focused on modeling traditional machine learning algorithms, others like Auto PyTorch allow for tuning deep learning architectures.

Data science collaboration tools

Data science teams lament the difficulty in collaborating across the data science workflow. It is often challenging for multiple data scientists to collaboratively write code when performing data exploration and ML modeling. As a result, data teams need to manually handoff code, which can be error-prone and time-consuming.

A new wave of data science collaboration tools addresses these pain points. For instance, Databricks offers a unified platform to collaboratively run analytics workloads across the data science workflow, and DataCamp Workspace will allow data scientists to collaborate asynchronously in real-time in the future.

Synthetic data generation tools

On the other hand, the data-centric AI popularized by Andrew Ng ushered in a new generation of tools for improving data quality. Most notably, synthetic data generation tools are built to generate labeled data at scale. Not only does synthetic data hold the promise to eliminate the need for manual data labeling or data collection, but it can also remove biases that arise with real-world data.

“With AutoML, a lot of the problems that data scientists work on are likely going to be streamlined. Therefore to remain competitive, data teams need to improve their skillset to be ahead of the curve so that the value that they can bring is more than just `.fit() .predict()`”

Preventing Fraud and Boosting eCommerce with Data Science

Elad Cohen

VP of Data Science
Riskified

09 The talent crunch and flexible work will broaden and improve the search for data talent

The Great Resignation saw 4.3 million Americans quit their jobs in August 2021 alone. This problem is particularly pressing for the tech industry, where resignations rose by 4.5% since last year.

As employees leave their jobs in droves, employers face a formidable task to retain and hire new employees, including data talents.

Employers are doing what they can to stop the mass exodus. They are offering not just better compensation, but also greater flexibility to work remotely. Apple, Google, Facebook, and Amazon are some examples of tech companies that have delayed their return-to-office till 2022. As the pandemic stretches on, it is clear that remote work is here to stay, according to Laura Boudreau, a Columbia University economics professor.

As working-from-home becomes more prevalent, companies are becoming less restricted to geographical boundaries in hiring talents. Companies seized the opportunity to widen their talent pool during the pandemic, as evident from the 280% increase in remote job postings on LinkedIn since March 2020.

In response, 46% of remote workers are planning to relocate in 2021, according to a survey by Microsoft. Against the backdrop of remote work and distributed teams in 2022, we foresee that companies will prioritize **skills over zip codes** in their hiring policies.

datacamp

Want to succeed in the era of data literacy?

Bridge your team's data literacy gap and become more data-driven.

Explore DataCamp for Business