Data Trends and Predictions 2021: The Year of Data Fluency

datacamp

The Great Acceleration

In 2008, Chris Anderson proclaimed in **Wired** the "end of theory," and that the combination of big data and statistical methods would completely revolutionize how we do business, conduct scientific research, and think about the world around us. This concept has since been dubbed the data science revolution (**Milken Institute**).

In a lot of ways, the data science revolution has achieved what it promised. DeepMind was able to use deep learning to reliably uncover how proteins fold —a problem that's eluded biologists for decades, and one whose solution can revolutionize the field of biology (**Nature**). Google and Facebook have entirely revolutionized the ad industry, using data and machine learning to optimize for click-through-rate. OpenAI has developed GPT-3, a natural language generation program that promises next-generation intelligence in the software we use (**New York Times**).

The data science revolution has always made the impossible possible. With data science methodologies and technologies, specialized teams have worked on solving important problems like self-driving cars, algorithmic trading programs, and protein folding. However, the real data science revolution **makes the possible widespread**, where everyone in an organization can understand data, communicate insights from that data, and make more informed decisions with data. It's about creating data fluent organizations and societies, where everyone is equipped with the necessary skills they need to be informed citizens and employees.

The past year has been tumultuous, with many lessons still being revealed today. The Covid-19 crisis has accelerated digital transformation, forcing incumbent organizations to digitize their processes, modernize their business models, enable data access, and upskill their workforce for a data-driven age (**Microsoft**). Digital, data-driven organizations like Zoom, Amazon, and PayPal benefited from the new normal (**Financial Times**) and solidified their places as leaders in their respective markets. The Covid-19 crisis has also proven the need for everyone to be data-fluent, informed citizens (**Dataversity**), as data can be used to inform and misinform us on the state of the pandemic.

The real data science revolution is materializing as we speak. What are the immediate implications? What are the new best practices for organizations to democratize their data and become data fluent in the process?

8 data trends that will drive data fluency

1. Data infrastructure tooling will mature around data democratization
2. Machine learning operations will support deploying models at scale
3. Third-party data will become more accessible than ever
4. The Jupyter ecosystem will further drive data democratization
5. Augmented analytics will catalyze a new age for data fluency
6. Data visualization goes mainstream
7. Upskilling becomes more crucial than ever
8. Data skills will cross over to every academic discipline

datacamp

1

Data infrastructure tooling will mature around data democratization

As organizations looked to modernize their data infrastructure, the amount of investment in data infrastructure startups have more than doubled over the past three years alone (a16z). As such, we've seen the emergence of a plethora of tools in the data engineering stack. These tools have helped organizations take in raw data, ingest and transform them, store them in a centralized location, and produce outputs in the form of descriptive and predictive analytics.

Currently, the data engineering stack is fragmented, with many tools competing over different elements of the data infrastructure. Over the next year and more, we will see the consolidation and standardization of various tools in the data engineering stack. This, alongside cloud adoption, metadata management tools, and centralized data governance platforms, will mean that organizations will deliver data that is **discoverable, reliable, compliant, and actionable** across a variety of end-users.

By 2022, public cloud services will be essential for 90% of data and analytics innovation.

—Gartner

2

Machine learning operations will support deploying models at scale

Research gains in artificial intelligence will continue to fascinate the general public. Indeed, these gains can be transformative, with applications in biology (DeepMind), self-driving cars (Tesla), and engrossing automated narrative fiction (AI Dungeon).

"By 2024, Gartner predicts that 75% of organizations will have moved from experimenting with Artificial Intelligence to operationalizing it."

However, most of the value of machine learning will be driven by existing techniques and methodologies. According to Kaggle's State of Machine Learning and Data Science 2020 survey, the most commonly used machine learning techniques in production are relatively simple algorithms such as Linear or Logistic Regression, and Decisions Trees or Random Forests. The challenge in reaping the benefits of machine learning at scale, however, lies in operationalizing it within the organization.

The past year saw the importance of monitoring models in production, as shifting consumer behavior due to the pandemic fundamentally changed the complexion of the data feeding into models in production (McKinsey). Over the next year and more, organizations will focus on deploying machine learning at scale. This will include the seamless integration of machine learning models within data infrastructure, developing MLOps capabilities and governance models for monitoring models in production, and creating a tighter human-machine feedback loop, resulting in more data consumers interacting with machine learning models to make data-driven decisions.

MLOps is a set of practices that combine machine learning, data engineering and DevOps skills to effectively deploy and monitor models in production.

3

Third-party data will become more accessible than ever

As organizations start making the most of their data for analytics and machine learning, they will need to begin leveraging third-party datasets to refine their models and analytics outputs. In the past, utilizing third-party datasets securely and scalably was hindered by time-consuming steps such as licensing, standardizing, and consolidating third-party data. On the flip side, data providers looking to share their data failed to capture all the value from it as there were no centralized services to manage data delivery across various users.

The advent of data exchanges and marketplaces aims to solve that. Services such as AWS Data Exchange, Snowflake Data Marketplace, and more have made it easier than ever for organizations to access third-party data and monetize their data. Over the next year and more, expect more organizations to leverage data marketplaces and exchanges and third-party data to become more trusted and actionable than ever.

By 2022, 35% of large organizations will be either sellers or buyers of data via formal online data marketplaces.

—Gartner

4

The Jupyter ecosystem will further drive data democratization

Over the past decade, Jupyter Notebooks have become a staple of the data scientist's toolkit. The notebook interface has streamlined the data science workflow, allowing data professionals to quickly prototype, explore insights, and share data narratives across their organizations (Netflix).

The next generation of Jupyter-powered data science interactive development environments will further drive data democratization. We've already seen solutions like Google Colab integrate collaboration features into their notebooks. naas.ai is making it simple to create data pipelines with notebooks. Mode is streamlining the switch from SQL to R or Python in a notebook environment and making analysis easily accessible. The next generation of notebooks will lower the barrier to entry to working with data, making it easier than ever for all members of the organization to access and build on top of insights, build data narratives, access data lineage, and facilitate collaboration between technical and non-technical experts alike.

74% of Data Scientists use Jupyter-based IDEs as their primary development environment.

—Kaggle

5

Augmented analytics will catalyze a new age for data fluency

As organizations look to become data fluent, expect business intelligence tools to become supercharged with augmented analytics capabilities. Augmented analytics can be defined as a business intelligence approach that uses natural language processing, graph analytics, and machine learning to extract data insights and stories automatically. We've already seen this feature emerge in business intelligence tools such as **Tableau**, **PowerBI**, and **Qlik**, and it will only get better in 2021 and beyond. This future of business intelligence means that actionable and effective data insights will become ubiquitous across the organization.

Conversational interfaces will further lower the barrier to entry for business intelligence adoption and empower data consumers across the organization to ask relevant questions about their data using simple english natural language queries, with no coding or pointing and clicking involved. Furthermore, graph analytics and machine learning will let business intelligence tools push insights a user may not have specifically asked for, as more than half of business users don't have a strong understanding of how other teams' data can be useful for them (**IDC**).

The global market for business intelligence will grow from USD 23.1 billion in 2020 to USD 33.3 billion by 2025.

—**Markets and Markets**

6 Data visualization goes mainstream

In the past year, the power of data visualization was put on full display. Data stories on the Covid-19 pandemic and the US elections became the most important ones of the year. The general public readily consumed visualizations from Our World in Data, Financial Times, The Economist, and more. Moreover, organizations are tapping into their data to tell the world their story, from Shopify's real-time tracking of shipments on Black Friday to Spotify's Wrapped.

The use and understanding of data visualizations have gone mainstream and will continue to grow as we become informed citizens, consumers, and decision makers. Business intelligence tools will continue to democratize data visualization for the organization, and data literacy skills, including the need to understand visualizations, will become critical for every employee.

> As we look toward 2021 and beyond, data literacy skills will be foundational to every role at every level of an organization, and will be critical in determining who survives and thrives as leaders going forward.
>
> —Andrew Beers, CTO at Tableau

7

Upskilling becomes more crucial than ever

The digital and data fluency skills gap is at the heart of the divide between incumbents and leaders. The PwC 2020 Global CEO Survey cited that 74% of CEOs are concerned with the lack of critical skills in their organization, with 32% of CEOs believing the skills gap is the biggest threat to their organizations. Leaders are realizing that without examining data insights, organizations won't be able to accelerate their digital transformation initiatives.

The data fluency skills gap is also at the heart of the divide between those who will thrive in the data-driven economy and those left behind (PwC). As countries grapple with mass unemployment and the aftershock of the Covid-19 crisis (The Hill), expect governments and companies to undergo serious upskilling and reskilling programs for 2021 and for many years to come.

By 2022, a third of G2000 companies will have formal data literacy improvement initiatives in place.
—IDC

8

Data skills will cross over to every academic discipline

As organizations realize the need to become data-driven, data fluency skills will become a mandate across every academic discipline. Over the past few years, universities have started incorporating data fluency programs across various disciplines and majors to prepare the workforce of tomorrow.

For example, the University of South Florida Muma College of Business created a citizen data scientist program to equip all its graduates with the data skills necessary to succeed in tomorrow's workplace. In 2020, DataCamp witnessed a tripling of learners taking advantage of the DataCamp for the Classroom program. This democratization of data skills will also start taking place in primary and secondary schooling, as academic institutions of all levels are looking to provide the foundational data skills needed to thrive in the data economy. Over the course of next year and more, expect academic institutions to further invest in data fluency skills in academic programs of all types.

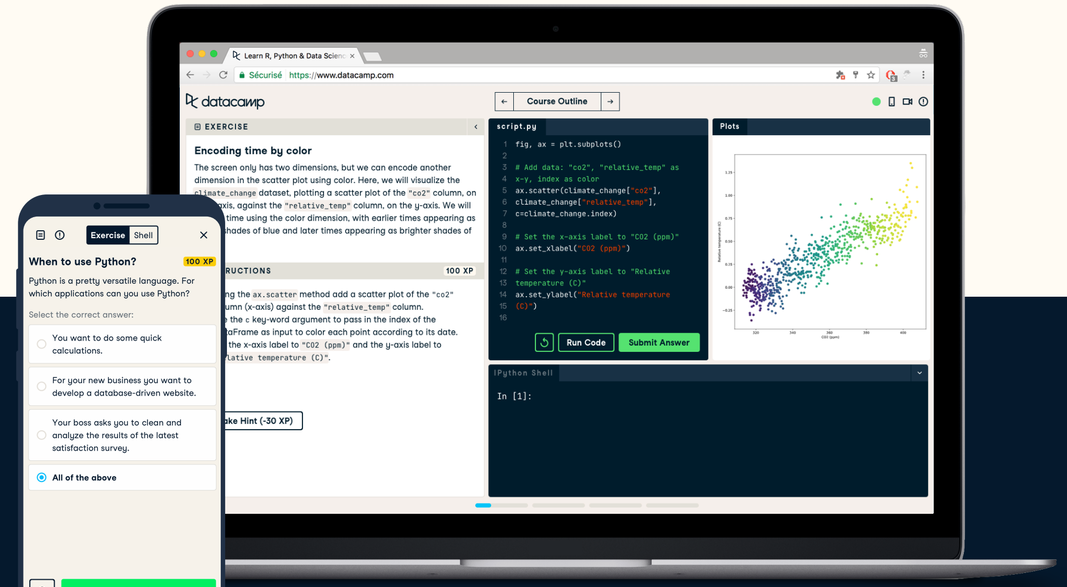
> By 2024, 80% of secondary schools will offer curricula targeting specific digital skills (such as coding or cloud technology) designed for post-high-school jobs, or for jump-starting tertiary learning opportunities.
>
> —Gartner

DataCamp's mission

The real data science revolution is unfolding in front of our eyes, and our mission has never been more important. We aim to democratize data science education by building the best platform to learn and teach data skills, and to make data fluency accessible to millions of people and businesses around the world. This is why seven million learners and more than 1,600 organizations—including Google, Intel, HSBC, eBay, PayPal, Uber, and more—use DataCamp to assess, learn, practice, and apply their data skills. Start your data fluency journey today—visit datacamp.com/groups/business.

"Data is the core of a business today. Yet most companies only analyze a fraction of their data, and do so inefficiently. Many relegate data science knowledge to a small group within the company. Consequently, they face an enormous skill gap that they can't hire their way out of. This runs counter to the data transformation initiatives that most companies are going through today. Democratizing data skills and making entire organizations data fluent is where we come in."

Martijn Theuwissen, DataCamp COO



clear as data