LEADING THE DATA RACE

The trends driving the future of data science

Featuring contributions from leading thinkers including:

Kirk Borne
Principal Data Scientist and Executive Advisor, Booz Allen Hamilton

Jim Duarte
Principal, JL Duarte & Associates

Meta S Brown
President, A4A Brown

Todd Mostak
Chief Executive Officer and Co-founder, OmniSci

CONTENTS

INTRODUCTION

LEADING THE DATA RACE

Pressure to convert massive volumes of data into real-time actionable insights has triggered a data race—and if you're not in it, you're already losing

Data is everywhere. Most businesses today are making efforts to capitalize on the vast potential of an ever-increasing range of untapped data sources to produce actionable business insights.

But this is about more than getting the right answers; it's about getting ahead of the pace of decision making. Speed is everything. In an age of connected devices and automation, decisions increasingly need to be made in real time.

The result is a data race. Business and IT leaders are arming themselves with cognitive technologies such as data science, artificial intelligence, machine learning and deep learning to push new limits and create a tangible competitive advantage. Whether improving efficiencies, optimizing production processes, or understanding individual customer behaviors, these tools and technologies will unlock a wealth of benefits.

The possibilities for innovation have never been greater: fraud detection in finance, predictive maintenance in manufacturing, sentiment analysis in media and entertainment and improving safety in chemicals are just a few examples of how emerging use cases are going to change the face of every industry.

This report identifies three key trends that will enable teams to push further, faster, in this race. From a shift to edge computing through to the advent of machine learning as a service, and the emergence of accelerated GPU technology, these trends are shaping new possibilities to drive speed, productivity and innovation in data science and machine learning.

To avoid falling behind, Data Strategists and IT leaders need to quickly adapt their data strategies to capitalize on the opportunities these trends are creating today.

CHAPTER 1

POSSIBILITIES AT THE EDGE: PUTTING INTELLIGENCE WHERE YOUR DATA IS

The future of data analytics is shifting from the cloud to the network edge to drive real-time decision making

POSSIBILITIES AT THE EDGE: PUTTING INTELLIGENCE WHERE YOUR DATA IS

The major technology trend of the past decade has been the mass migration to the cloud. The economies of scale and breadth of online services have meant that organizations of all sizes have adopted cloud services for a variety of IT functions, to such an extent that modern approaches to building and running applications are now described as "cloud native."

But, for businesses that want to stay ahead in the data race, centralizing everything inside massive cloud data centers is becoming limiting. The arrival of 5G networks and a boom in connected devices as part of the Industrial Internet of Things (IIoT) will produce vast quantities of real-time data—all of which will need to be rapidly analyzed to inform timely business decisions. In a world of emerging technologies and powerful new analytics models, speed is as critical as accuracy—and in this world, the cloud is going to fall short.

According to Gartner,[1] while only about 10 percent of enterprise-generated data is created and processed outside a traditional data center or cloud, this figure is expected to soar to 75 percent by 2025. Santhosh Rao, Gartner's Senior Research Director, concludes organizations are therefore going to have to consider a decentralized approach: "As the volume and velocity of data increases, so too does the inefficiency of streaming all this information to a cloud or data center for processing."

75%

The proportion of enterprise-generated data that will be created and processed outside a traditional data center by 2025 according to Gartner.[1]

This means making a potentially game-changing shift: away from the cloud towards edge computing.

Oliver Schabenberger, Executive Vice President and Chief Technology Officer at analytics firm SAS, argues the edge should be the starting point for enterprise organizations. This is because everything generating data outside of a data center and connected to the Internet is at the edge.[2]

"That includes appliances, machines, automobiles, streetlights, smart devices in the home, locomotives, pets and healthcare equipment," he says.

For data scientists, shifting intelligence to the point of collection opens up a new world of possibilities. For starters, it presents the opportunity to finally realize the potential of IIoT and use connected devices to collect lots of different data types and learn from it without having to sort it first. This allows data scientists to capture insights from things like wind turbines or doors or streetlights, without knowing what they are looking for.

But more immediately it raises the tantalizing prospect of three major benefits: faster response, greater scalability due to processing being distributed around the network, and cost savings by minimizing the bandwidth used. All of this adds up to being able to push new boundaries in analytics and do more, faster. →

> "The ability to do intelligence or knowledge discovery at the point of data collection is critical in many applications now."
>
> **Kirk Borne,** Principal Data Scientist and Executive Advisor, Booz Allen Hamilton

THE NEED FOR SPEED

Of all the potential benefits of edge analytics, the speed of data processing is the most pressing. Applications such as streaming media, IIoT, and VR and AR applications all require large amounts of data to be delivered with very low latency. If a sub-second response time is required, then waiting for a request to the cloud is not a viable option. As a result, the potential of some of the most exciting emerging technologies can only be truly realized at the edge.

Kirk Borne, Principal Data Scientist and Executive Advisor at Booz Allen Hamilton, says the race to keep up with the pace of data accumulation is driving the shift to the edge in data science. "The ability to do intelligence or knowledge discovery at that point of data collection is actually critical in so many applications now. You no longer have the luxury of bringing data back to your business center and spending a year analyzing it."

This is particularly true with the example of self-driving cars. Autonomous test vehicles operated by Waymo (formerly Google's self-driving car project) have been estimated to generate between 11TB and 152TB of sensor data every day.[3] And sending data back and forth across a network will take at least 150–200 milliseconds,[4] assuming the connection is reliable enough.

Edge devices are key to ensuring the vehicles take up their place as the ultimate IIoT devices, communicating with the environment around them and, crucially, with other vehicles on the road to make split-second decisions. There is no margin for error or lag when reacting to ever-changing road conditions. Pausing for instructions from a cloud server remotely analyzing data is simply not an option for self-driving vehicles. Only the edge can deliver the rapid decision making required to make autonomous cars a safe, viable prospect.

Less glamorous examples are no less critical. For instance, a control system operating an industrial machine will need to stop immediately if a human is too close.[5] The smallest lag can be the difference between life and death.

The edge also presents opportunities to improve safety by using real-time data in aerospace. "Airplane engine manufacturers want to have some sort of reliable heavy-duty compute on the plane itself," says Todd Mostak, CEO and Co-founder of analytics firm, OmniSci. "Because if there is a problem, [at present] there may not always be the opportunity to send back however many terabytes the plane generates in flight, and people want to do some analysis and anomaly flagging actually on the plane."

Even where the immediacy of decision-making is less critical, the cumulative business risk of delay can quickly mount up. For example, supermarket chains trying to avoid store waste may turn to the edge to ensure they can provide store managers with real-time insight into the operational performance and temperatures of their fridges across multiple store locations, so they can take prompt action to protect goods in the event of a failure.[6] →

152TB

The amount of sensor data reportedly collected every day by some of Waymo's self-driving test cars.[3]

SHAPING CUSTOMER EXPERIENCES

In the Media and Entertainment industry the edge also enables businesses to meet growing demand for improved, more personalized customer experiences. Data scientists are under increasing pressure to translate user data into valuable insights faster through recommendation engines and sentiment analysis. This is the reason leaders like Netflix, Uber and Amazon are choosing to develop their business models around the edge.

High-speed 5G networks are only expected to ramp up user expectations for flawless customer experiences on the go. With some network operators estimating performance to be anything from 10 times faster up to 1,000 times faster than current networks, users will be hungry for anytime experiences such as augmented reality or streaming of 4K video to mobile devices—and this demands speedier delivery of data that only the edge can provide.

THE NUMBERS GAME

One implication of this shift in focus is that the hardware requirements for edge computing are going to change. Vendors are responding to the appetite for innovation at the edge by moving from simple embedded devices and microcontrollers towards more powerful systems such as workstations or even micro data centers.[7]

For businesses handling vast amounts of IIoT data, the edge should enable a way to combat the cost of transmitting data back to a data center or the cloud. This is a point made by Seagate's Jeff Nygaard, Executive Vice President and Head of Operations, Products and Technology.

"It's not free to move data through a pipe from endpoint to edge to cloud; it costs money to send data through that pipeline. The idea that you should really only be moving data if you need to move the data is something you should be thinking about," he says.[8]

Edge computing allows for data to be filtered and processed before it is sent to the cloud, reducing the volume of data transmitted and the network costs that would otherwise be incurred. This opens up a world of possibilities for businesses to realize the full potential of IIoT devices to obtain previously unviable data sources for business-critical insights in real time.

150-200

The minimum number of milliseconds it is estimated data takes to travel across a cellular network, assuming that there is a strong network connection.[28]

As an example, an oil rig operating in a remote location in the ocean may have thousands of sensors producing large amounts of data, most of which may be redundant. There is unlikely to be a reliable internet connection, and streaming all of this information back to base via a satellite data connection would be →

EXPERT VIEW:

The future of data science at the edge

Kirk Borne
Principal Data Scientist and Executive Advisor, Booz Allen Hamilton

"To manage fast moving data of all kinds of variety you need edge devices that are at the point of data collection—and are specific to the type of data collection —whether it's imaging data or stream, like cyber network data, whatever is at the edge. The ability to do intelligence or knowledge discovery at that point of data collection is critical in so many applications now."

Jared Dame,
Director of AI and Data Science, Z by HP

"Edge includes mobile workstations and devices like the Jetson Nano and the Intel Movidius chips. All of those things will be out on the edge collecting, filtering and making data so that the cloud isn't the endpoint for everything. This means you're not hung up waiting for all the data to be processed up to the cloud in order to make a business decision."

> **"Edge computing has an exciting future, extending way beyond applications such as IoT, self-drive vehicles and industrial automation."**
>
> **Darren Seymour-Russell,**
> Head of Data Science, Mudano.

prohibitively expensive. Instead, a local edge computing system could compile the data and send daily reports to a central data center or cloud for long-term storage, so only the important data gets transmitted.[9]

Edge computing is thus about implementing a more distributed architecture that is key to delivering scalability, since data processing takes place closest to the source of the data, making it more feasible for business users to gain real-time insights from the data being gathered. For IT leaders and Data Strategists, this enhances the decision-making process by avoiding the risk of network failures and delays that might impact performance with a centralized architecture.[10]

But edge and cloud are not mutually exclusive. A new hybrid model is emerging where the edge and cloud work together. A factory environment could be one example; in a smart factory, data is not funnelled directly to the cloud, as valuable factory floor context may be lost in the transition. Instead, filtered data is sent to the cloud during off-peak hours when timeliness is no longer a concern.[11]

POSSIBILITIES FOR DATA SCIENCE AT THE EDGE

Jim Duarte, Data Scientist and Principal at LJ Duarte and Associates, believes the shift to the edge will be enthusiastically embraced by data scientists—especially by engineers who place predictive models at the source of the control.

"More and better edge applications will be possible as better analytics create better models and the scalability of technology will put it in more locations," he predicts. "Complex workflows will have better controls to minimize constraints for improving optimization. More robust edge analytics will be created as controllers can be scaled down and be put in hostile environments because their smaller size will enhance the ability to protect them and keep them functioning longer."

For data scientists, performing sophisticated data analysis at the edge will enable prediction-based outcomes that will empower businesses to make greater efficiencies and cost savings.

As an example, consider a large piece of expensive industrial machinery, such as an earth mover. Machine learning could be used to identify and map the relationships between the machine's sub-assemblies and components, then monitor those components and run simulations to predict their future state.[12]

Such a model would not only be able to predict future failures in the machine, but also the probable time of failure. The calculations to do this would simply have been too complex in the past. With this capability, the company would be able to schedule maintenance and avoid unplanned downtime, as well as make spares inventory much more efficient.

DRIVING BETTER BUSINESS DECISIONS

Edge analytics opens up a whole new realm of possibilities in machine learning applications. In financial services, for example, data scientists can train algorithms to assess in real time if a transaction is unusual for a particular customer, and request verification or even just block it if there is at least a 95 percent probability of it being fraudulent.[13] →

Machine learning algorithms also fit perfectly with the underwriting tasks that are so common in finance and insurance, performing underwriting and credit-scoring tasks that can help employees make speedier and more accurate decisions.

Healthcare is another field where analytics driven by machine learning is expected to make a huge impact by driving significant efficiencies in terms of time and cost—but also, crucially, by saving lives. From helping pathologists make a quicker and more accurate diagnosis from medical images, to identifying patients that might benefit from new types of treatments or therapies, healthcare professionals now regard AI as the future of healthcare.

"Today when you go to hospital and have an MRI, the data is there immediately, but it takes probably a week until you get the result," says Amit Marathe, Head of Software and AI at HP. "I see healthcare shifting towards more real-time analytics."

One future application example might be a hypothetical emergency medical responder (EMR) system that could run predictive algorithms at the same time as a doctor is examining the patient. In this vision, the system would display the real-time diagnosis, pathology results and treatment options, as well as each option's potential effectiveness.[14]

> **"Today when you go to hospital and have an MRI, it takes a week until you get the result. I see healthcare shifting towards more real-time analytics."**
>
> **Amit Marathe,** Head of Software and AI at HP

THE RISE OF INTELLIGENT DEVICES

These advances are propelling AI outside of the data center and into devices and machines we use in our work and our everyday lives.[15] Deloitte predicts a coming "era of pervasive intelligence" that will be marked by a proliferation of AI-powered smart devices able to recognize and react to sights, sounds and other patterns. Increasingly, data scientists will be able to train machines to learn from experiences, adapt to changing situations, and predict outcomes.

"Edge computing has an exciting future, and we see this extending way beyond the 'traditional' applications such as IIoT, self-drive vehicles and industrial automation and into areas in which a personalized user experience is key," predicts Darren Seymour-Russell, Head of Data Science at data consultancy, Mudano.

For example, customers might be offered services or concessions tailored individually for them during secure banking interactions. With their intelligence embedded rather than living in the cloud, such devices will enable all kinds of applications that require instantaneous response and robust performance even when connectivity is poor or not available.

OUTLOOK

Data is increasingly being generated at the edge rather than the data center, and a real-time response is needed in many applications, so that processing the data also needs to happen at the edge. Businesses need to evaluate how and where data is collected and used in their organization, and whether performing analytics at the edge of the network might significantly boost performance while avoiding data congestion. However, edge computing does not replace the cloud, but extends its reach, providing a local point of presence and reducing bandwidth costs.

KEY TAKE AWAYS

- Processing at the edge cuts costs by reducing costs associated with network traffic and latency
- Analytics at the edge is vital where a rapid response is required, or bandwidth is limited
- There is a noticeable shift in trend from pure cloud strategy to a hybrid strategy that includes edge + cloud

CHAPTER 2

MACHINE LEARNING AS A SERVICE: AI FOR ALL

Access to an emerging range of subscription AI tools will transform data science

MACHINE LEARNING AS A SERVICE: AI FOR ALL

Once upon a time, the data that most businesses had to work with was mostly structured and small in size. This meant that it was relatively easy for it to be analyzed using simple business intelligence (BI) tools.

Today, this is no longer the case. Much of the data that organizations are mining is unstructured or semi-structured, and the trend is growing such that more than 80 percent of corporate data is expected to be unstructured by 2020.[16]

On top of this, the rate at which this data is being created is expected to increase at such an extent that IDC predicts the global datasphere will grow from 33 zettabytes (ZB) in 2018 to 175 ZB by 2025.[17]

Simple BI tools are no longer capable of handling this huge volume and variety of data, so more advanced analytical tools and algorithms are required to get the kind of meaningful, actionable insights that businesses need. To keep pace with demand for insights that can drive quicker, better decision making, data scientists are looking to artificial intelligence (AI), machine learning (ML) and cognitive computing technologies to take analytics to the next level.

No organization can afford to fall behind. As a result, IDC predicts that worldwide spending on AI and cognitive computing will reach $77.6 billion in 2022, more than three times that in 2018,[18] while the total global business value derived from AI is forecast to reach $3.9 trillion within the same timeframe, according to Gartner.[19]

THE ADVENT OF MLAAS

But getting started with AI techniques such as ML is challenging. Not only does this require highly specific, high-demand skill sets, it may also call for specialized IT infrastructure and software tools—not to mention a sound data strategy. All this adds up to a significant up-front investment that can be cost-prohibitive for many businesses.

In response to this challenge, vendors have begun offering Machine Learning as a Service (MLaaS).

As the name suggests, MLaaS is a subscription-based model that offers access to AI tools, in the same way that many business applications are now offered in a software-as-a-service (SaaS) model. These AI services can range from developer tools to data pre-processing and model training, through to fully-trained ready-to-use models that can be accessed through an API and integrated into business workflows.

The advent of MLaaS means that, instead of investing in creating their own AI resources, organizations will be able to turn to vendors for an easier, lower-cost ecosystem of offerings that can be customized to their needs. This heralds a new era for data science—one in which AI tools become easier to use and more accessible to a broader range of companies and roles within organizations.

LOWERING THE BARRIERS TO ENTRY

The implications of the shift to MLaaS are huge. For starters, MLaaS will democratize access to cognitive computing, by making it more cost-effective for organizations of all sizes to experiment with ML and analytics tools. →

$3.9 TRILLION

Expected total global business value derived from AI by 2022, according to Gartner[19]

175 ZB

The predicted size of the global datasphere by 2025 according to IDC[17]

> “As you take away a lot of the unnecessary expense and manpower that’s needed... they’ll invest more in getting results.”
>
> **Jake Gardner,** Enterprise Account Executive, Domo

“When it’s a service, you take the whole platform administration piece away from the enterprise,” explains Jake Gardner, Enterprise Account Executive at Domo. “That cuts down a lot of the headache and cost that traditionally surround data science and data repositories, and as you take away a lot of the unnecessary expense and manpower that’s needed for the companies to do it, they’ll invest more in getting the results that they want through the data science. That’s going to drive a lot of innovation in the industries.”

MLaaS will also enable smaller organizations to scale up at the pace of demand.

“The greatest benefit of service technologies for analytics is flexibility,” states Meta S. Brown, President of data consultancy A4A Brown. “These [MLaaS] offerings allow teams to ramp up computing resources quickly, or to use a lot of computing power for occasional needs, without the obligation to support those resources at all times.”

As well as being a cost-effective way of bringing ML tools into an organization, MLaaS should make them more accessible to a wider range of roles and skill sets.

“I think that as data science and ML get faster and easier to use, they are going to be used by a broader swathe of organizations and people, even analysts are going to be expected to have a basic understanding of how to leverage data science and ML techniques,” predicts Todd Mostak, CEO and Co-founder of OmniSci. “ML and data science will increasingly be embedded in ordinary business tools that people don’t typically associate with advanced data science.”

He points out this trend is already happening with tools like Tableau used as a popular BI product for Excel so that people can spots anomalies in spreadsheets or in time series can find out the cause. “The user may never know they are doing data science—they just want answers.”

This could mean data is used in new and unexpected ways →

> “The greatest benefit of service technologies for analytics is flexibility.”
>
> **Meta S. Brown**, President, A4A Brown

EXPERT VIEW:

The future of Machine Learning as a Service

Jake Gardner
Enterprise Account Executive at Domo

"As MLaaS grows it will enable a new style of data scientist—not your statistician or PhD, or even master's and math-type person, but your average everyday user, your line manager, your analysts who might have more of a marketing focus. Everybody will be able to take this and be able to use it. It'll just be so easy and so cheap that your marketing department will be doing machine learning to understand its impact on the sales and inventory, and how that maximizes ROI."

Amit Marathe
Head of Software and AI at HP

"In the future, all businesses will be able to use MLaaS and self-driving data science platforms to generate real-time visual insights to save lives, improve productivity, increase profitability and transform their business like never before. Using the power of data, businesses will be able to disrupt themselves, as opposed to waiting for somebody else to disrupt them."

by more organizations and more people—and that better decisions are made through every managerial layer.

A CATALYST FOR INNOVATION

All of this means that MLaaS is expected to unleash a new wave of innovation.

To date, the data race has been dominated by organizations with pockets deep enough to invest heavily in AI. These businesses are already enjoying wins across a wide range of industries. For instance, in financial services ML is being used to detect fraud and identify money-laundering behaviors. In a recent survey by Refinitiv of 450 financial professionals, comprising executives and data scientists, 90 percent of respondents indicated that they had deployed ML in one or more departments, and three-quarters believed it now represents a core part of their business strategy.[20]

MLaaS will level the playing field, enabling organizations with more limited resources that may be lagging behind to play catch-up and incorporate ML into their workflows.

For example, businesses will be able to use self-service data science platforms to generate real-time visual insights and improve productivity. These will enable data scientists to provide insights about individual customer behaviors to such an extent that businesses can produce just the right quantity of highly customized goods and services for each customer. This could drive significant production efficiencies and significant market advantage.

Sectors such as healthcare could benefit from customization using individual patient vital signs and medical history. For example, research firm, Emerj, envisages using an ML application with an agent (such as Amazon Alexa) to adjust a patient's dose of pain killers or antibiotics by tracking data about their blood, diet, sleep and stress.[21]

Alternatively, ML might be deployed to automate repetitive tasks, such as search and information retrieval, or sorting products into various categories. Businesses could quickly cut administrative costs, increasing efficiency and freeing up staff for higher-value tasks.

Kirk Borne, Principal Data Scientist and Executive Advisor at Booz Allen Hamilton, believes that MLaaS will evolve to be able to advise businesses on what problems they should actually solve with their data.

"One prediction for the future is that we're going to be applying our data science and these algorithms to our data science itself. So, the data science is going to bring about that meta level of data discovery—that is: 'Here's what you should look at,' 'Here's what you should pay attention to,' or 'Here's the trend or pattern that needs your attention right now.'"

NEXT-GENERATION AI SERVICES

This evolution of MLaaS services is likely to be rapid. According to Research and Markets, the MLaaS market is expected to expand with a compound annual growth rate of over 43 percent for the period 2019 to 2024. It →

MLaaS could be especially beneficial in low-tech sectors such as agriculture and oil and gas

Borne goes further still, predicting that organizations will be able to pick and choose the functions they need, just like today's users can browse an online store for smartphone apps.

"I believe they [MLaaS services] will be like commodity markets around algorithms," he says.

Borne cites the example of chatbots, which are already available via APIs and enable a small company to have almost a full customer service department, even with few staff.

"If someone calls your company for some service or with questions about things, the first line of answering can come through the chatbot, which has been fed your FAQs (frequently asked questions). So, the chatbot will know how to answer the FAQs, which is usually about 80 percent of the work of the customer service representative, anyway."

Service providers are already expanding beyond just offering ML developer platforms to offer a feast of specific ready-made functions, which can be accessed via APIs and integrated into an organization's workflow to solve numerous specific problems.

Google, for example, has Google ML Engine for experienced data scientists, while in beta is Cloud AutoML, described as a suite of ML products aimed at those with limited ML expertise.[21]

As the MLaaS eco-system evolves, so too will the sophistication and complexity of the capabilities. Services are expected to extend beyond predictive analytics, data transformations and visualizations, and move into facial recognition and natural language processing too.

In today's market, services like Amazon Comprehend offer the ability to find company names in analyst reports, or identify negative reviews[23]—but in the future, natural language processing systems will be able to offer more human-like understanding of speech.

"I think that natural language processing will eventually get to a point where it will be able to determine where a person is from based on how they use words and how they construct a sentence, as well as slang—possibly even down to the particular state or →

> **"I believe they [MLaaS services] will be like commodity markets around algorithms."**
>
> **Kirk Borne,** Principal Data Scientist and Executive Advisor, Booz Allen Hamilton

43%

The expected compound annual growth rate for MLaaS services from 2019 to 2024 according to Research and Markets[22]

believes the MLaaS model will dominate the AI market, with users able to choose from a wide variety of solutions focused on different business needs.[22]

"There's going to be a fork in the road where you're going to have two very clear categories," predicts Amit Marathe, Head of Software and AI at HP. "One is going to be people who want to have ultimate control and build everything themselves. The other is agriculture, oil and gas, and anything else that's non-tech—those people don't have the manpower to assemble teams to go and build everything from scratch."

region," says Jared Dame, Director of AI and Data Science at Z by HP.

A4A's Meta Brown is particularly excited by the possibilities of machine listening services.

"Nobody's talking about it, but the potential is tremendous," she says. "For example, if you've ever used YouTube search, you know it's very difficult to find what you want unless the keywords you use match the title. Video and audio content search could be vastly improved if machines understood all of the content."

AN ERA OF COLLABORATIVE DATA SCIENCE

One of the barriers to broader adoption of ML and data science in general is the enormous complexity that users have to come to grips with. Dame says that automated ML systems will be required to get over this hurdle. He predicts that so-called "low-code" development platforms, which support rapid application development with minimal hand-coding, will allow anyone to create a working solution.

"I see the future with automated ML is low-code environments being able to produce automated ML, or automated deep learning as well, for the general consumer. So someone who is an artist, who has no computer background, will be able to go in, build an algorithm on a web tool that could be a cloud-based utility or on a local machine, and then publish it as a service to help sell their wares or do something."

A more flexible, open-access approach to ML will have significant implications on the role of the data scientist and team structures.

For instance, Jim Duarte, Data Scientist and Principal at LJ Duarte and Associates, argues that data strategy needs to extend beyond IT/OT and that more subject matter experts (SMEs) from various disciplines need to be involved. As MLaaS makes more tools available to a wider group of roles, the need for a more collaborative approach will only grow.

"The biggest opportunities are for management to become educated beyond their belief that everything data belongs to IT/OT," he says. "When IT/OT is given the budget for 'data' and expected to make decisions on storage, cleansing, formatting, accessibility, as well as analytics, in a vacuum, then there will be areas that will be unable to make good decisions from data. One example would be choosing a ML tool. IT/OT has little experience with advanced analytics."

MLaaS will make more tools available to →

> **"The biggest opportunities are for management to become educated beyond their belief that everything data belongs to IT/OT."**
>
> **Jim Duarte,** Principal, JL Duarte & Associates

data scientists and allow them to focus on higher-value tasks. Although the role of some data scientists may shift, Duarte does not see the need for them diminishing.

"Data scientists hold the keys to the whole data picture, from collecting it to presenting it to decision makers," he says. "What excites me is the larger acceptance of analytics for decision making. The key is to position the analytics among the decision makers at all levels."

SHAPING THE FUTURE

No longer will data science be the preserve of larger, better-resourced businesses. MLaaS will be an increasingly valuable tool for data science teams at organizations of all sizes to obtain vital data insights and predictions across most industries. A rapidly growing MLaaS ecosystem is set to enable integration into business workflows through APIs, while the technology will become more accessible, putting the power of ML into the hands of more of the workforce.

As MLaaS makes more tools available to a wider group of roles, it should also drive innovation outside of traditional data science roles. Organizations that can broaden data strategy beyond their IT team and put analytics at the disposal of decision makers at all levels will see the biggest gains.

"I think it is going to become a big differentiator for those organizations that are able to become data driven, and there will be the haves and the have-nots," says Mostak. "It will become a market advantage to use data science because there is so much data at your disposal these days."

KEY TAKE AWAYS

- MLaaS is democratizing access to ML, making it less complex and costly to use
- An ecosystem of MLaaS tools will enable easier integration into business workflows
- Easier-to-use MLaaS tools will lead to broader use within an organization for non-specialists
- MLaaS will allow data scientists to focus on higher-value tasks

CHAPTER 3

ACCELERATED GPUs: THE POWER TO PUSH LIMITS

A new generation of GPUs and GPU-accelerated software is set to unleash a wave of productivity and innovation

ACCELERATED GPUS: THE POWER TO PUSH LIMITS

Many organizations are fighting an uphill battle to handle the scale and velocity of data they are harvesting.

Whether training machine learning (ML) models or crunching large volumes of data, analysts and data scientists on the front line often experience frustration as their effectiveness is hindered by the limitations of compute power.

Unfortunately, these challenges are becoming more acute. On the one hand, the pressure to extract real-time value from connected devices and big data is mounting. And on the other, the sheer volume of data being created is burgeoning. According to IDC, worldwide data creation will increase to 175 zettabytes by 2025—a tenfold increase on that produced during 2017.[24]

Furthermore, as data scientists increasingly use ML models to capture insights, many data tasks are growing increasingly complex, making training and deployment more difficult. This is putting a strain on the traditional approach to analytics led by CPUs (central processing units) and fueling demand for hardware that will perform ahead of the pace of innovation.

In short, data scientists need greater compute power. And this is driving a major shift to a new era of accelerated hardware—one that is will be defined by the emergence of a new generation of ultra-powerful GPUs (graphics processing units) and GPU accelerated software. This move opens up a new realm of possibilities for data scientists by super-charging productivity, speeding up workflows and unlocking the full potential of ML to accelerate analytics.

The shift to this new era is already well underway. The GPU market alone is predicted to surpass revenue of over $80 billion by 2024, an increase of more than 31 percent from 2018—by which time worldwide GPU industry shipments are anticipated to reach 121,000 thousand units, according to a recent Global Market Insights report.[25] As the rate of adoption increases, so too will the advances in what can be achieved. →

$80 BILLION

The value that the GPU market is predicted to exceed by 2024 according to Global Market Insights.[25]

16,000

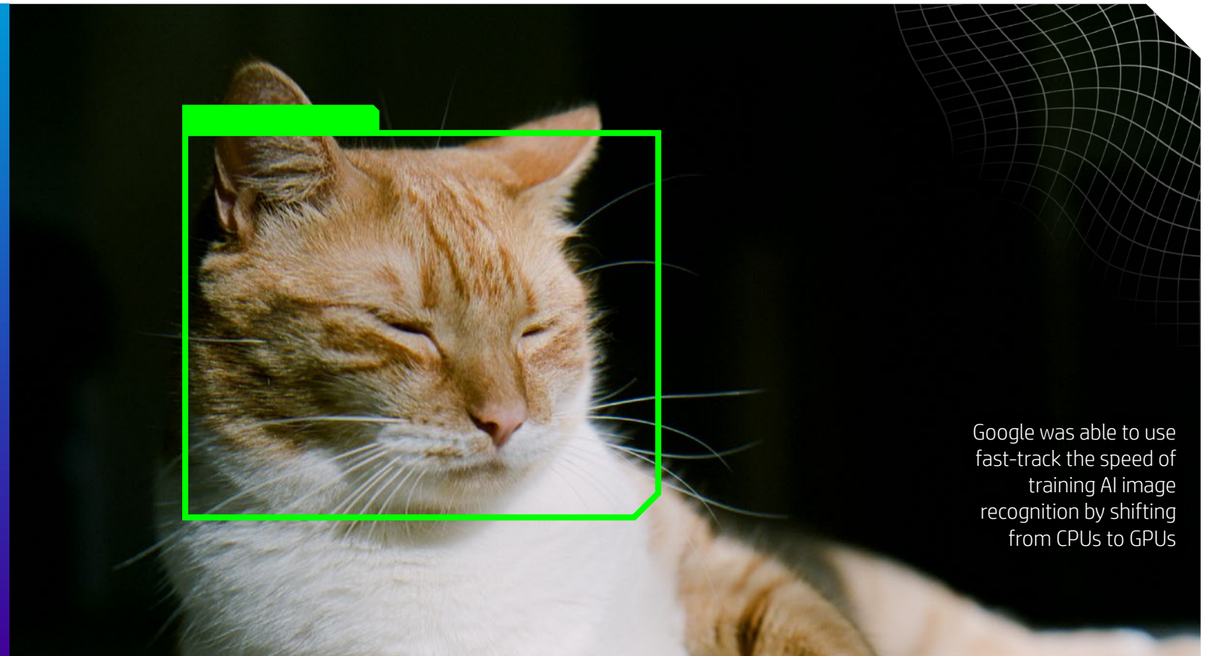
The number of CPUs NVIDIA was able to match in performance with just 48 GPUs for Google's image recognition system. [27]

FAST-TRACKING DATA WORKFLOWS

Much of this trend is being largely driven by the adoption of AI (artificial intelligence), ML innovations in sectors such as healthcare and automotive, and the rise of internet of things (IoT). This is because GPU-accelerated analytics offer teams a way of dealing with the volume and velocity of data associated with big data, enabling data scientists to uncover critical insights faster.

"In the early days of this revolution people were talking about managing your data at the speed of business," says Kirk Borne, Principal Data Scientist and Executive Advisor at Booz Allen Hamilton. "I think that expression needs to be inverted now. We need business at the speed of data."

GPUs are a response to this challenge.



Google was able to use fast-track the speed of training AI image recognition by shifting from CPUs to GPUs

Having started out as accelerators to offload the burden of graphics processing from the CPU in games, GPU architecture proved equally effective for accelerating data science workloads.

They have relatively simple cores, optimized for floating-point throughput. But there are a large number of them—thousands in high-end chips—so they can process a large set of identical computations in parallel.[26] In contrast, CPUs are general purpose and have just a few complex cores optimized for sequential processing of application logic. This means GPUs have a clear advantage when analyzing huge data sets, where the same calculations need to be performed on all of the data.

For example, when Google started improving its AI systems for image recognition it was using 16,000 CPUs to train an AI to recognize photos of cats. By working with GPU maker NVIDIA, it achieved roughly the same performance with just 48 GPUs.[27]

A big reason for the shift to GPUs is that the needs of data scientists and the types of data they are handling is changing. This presents particular challenges for CPUs. For instance, according to software firm OmniSci, around 80 percent of the data created today contains location-time (spatiotemporal) data which is compute-intensive because it requires rapid analyzing at a granular level. This is difficult to achieve using traditional indexing and pre-aggregation techniques, so most mainstream CPU-led business intelligence and analytics systems struggle to cope with spatiotemporal datasets above low volumes.

Also, when handling large volumes of data, traditional CPU architectures require large hardware footprints and consume considerable resources and time in 'wrangling' of data. In contrast, GPU-accelerated data analytics avoid lower-value tasks such as down-sampling, indexing and cubing by 'ingesting' the entire dataset. The parallelism provided by the GPUs means queries can be evaluated in real time without relying on pre-computation.

According to Todd Mostak, CEO and Co-founder of OmniSci, this means data scientists will be able to work "at the speed of curiosity."

"I'm most excited about the agility and the iteration possibilities: where you have →

> **"The productivity yields and gains for, not just the data scientist, but also the organization as a whole, are going to be immense."**
>
> **Todd Mostak**, CEO and Co-founder, OmniSci

1000x

Increase in GPU performance that NVIDIA expects to see by 2025.[34]

Logistics firm, UPS, has been using GPU-accelerated analytics to improve delivery efficiency

an idea you don't have to go to IT and wait two weeks to get some rolled up version of your data. You can use the GPUs to query the data, perhaps even churn the records, pull back, interactively visualize and find what you're looking for—the good data, the bad data—and feed it into the ML pipeline, which instead of taking hours or days comes back in minutes. The productivity yields and gains for, not just the data scientist, but also the organization as a whole, are going to be immense."

ACCELERATING INNOVATION

By enabling real-time analytics and freeing up data scientists to spend more time on high-value tasks, advances in GPU technology are also expected to be a catalyst for innovation.

"I think using GPUs for more advanced data science is going to be really big," states Jake Gardner, Enterprise Account Executive at Domo. "It enables more and more people to have reliable access to the hardware that you need to be able to do more of these advanced types of applications like neural networks."

Already, there are lots of early wins. For instance, in logistics, companies like UPS are using GPU-accelerated analytics to extract value from big data supplied by customers, drivers and vehicles. The firm has created a proprietary tool that uses advanced algorithms to find the optimal routes for delivery trucks in real time, responding to weather conditions and accidents. The impact was hundreds of millions of dollars saved and an improved customer experience.

Many sectors are ripe for change. In finance, capital market trading firms face challenges around market risk reporting, compounded by new regulatory drivers such as Fundamental Review of the Trading Book (FRTB) reporting. GPU-accelerated database technology could vastly improve processing times for their interactive data and big data challenges. A recent trial by CitiHub Consulting with a tier one investment bank found that GPUs outperformed a CPU configuration from 2.5x for simple queries to more than 400x with complex ones.[28]

Similarly, in insurance, companies currently spend hundreds of millions of dollars per year to process data using CPUs to model actuarial decisions. There are easy wins to be had shifting to a GPU-led approach. Elsewhere in the sector, start-ups like Ravin AI is already automating the time-consuming process of vehicle inspections so that car insurers, dealers and rental agencies can use drivers' mobile phone images, via GPU-powered algorithms, to evaluate any damage in real time.

Weather forecasting is an example of a notoriously difficult area due to the huge amount of processing power required to model future conditions from weather data. GPU acceleration will play a major role in boosting processing to deliver more accurate forecasting, enabling power companies to more accurately predict electricity demand in different areas, and farmers to better prepare for dramatic shifts in weather.[29]

There are numerous areas in healthcare where GPU acceleration of data science reaps dividends. Analysts looking to uncover hidden patterns and correlations between biological information and effectiveness of pharmaceuticals are seeking more powerful tools beyond legacy medical analysis software.[30] →

In oil and gas exploration, traditional geophysical analysis software struggles to analyze and visualize the large volumes of data needed to determine borehole viability. GPU-accelerated analytics could use a wealth of real-time petrophysical data to enable better visualization capabilities to inform new drilling opportunities.[31]

There are also ways to disrupt data science processes themselves. According to data scientist Gregory Piatetsky-Shapiro, Founder and President of KDnuggets, GPU-accelerated analytics will enable data scientists to crunch through huge amounts of data and enable vastly more sophisticated AIs that may eventually completely automate many data science processes.[32]

OPTIMIZING FOR AI

NVIDIA in particular is driving innovation in GPU acceleration. The firm effectively pre-empted the current boom by more than a decade through the development of software to allow its GPUs to process the millions of minuscule computations that data science workloads require. The company has also been optimizing its GPUs for deep learning, adding functions to make training and inferencing deep neural networks faster and more energy efficient.

To take advantage of the power of its GPU technology, NVIDIA is driving the data science software ecosystem with NVIDIA RAPIDS, a collection of NVIDIA GPU-accelerated open source libraries and APIs for accelerating end-to-end data science including deep learning, machine learning, and data analytics. The software ecosystem is critical to unlocking the power of GPUs.

This is an ongoing process, since datasets and ML models are going to get larger and more complex over time, which means that GPUs need to keep getting faster and more powerful to keep up.

"Next-generation AI hardware solutions will need to be both more powerful and more cost efficient to meet the needs of the sophisticated training models that are increasingly being used in edge applications," says Chris Nicol, Co-founder and Chief Technology Officer of Wave Computing.[33]

NVIDIA only sees progress in GPUs ramping up. It predicts that new GPU architectures for graphics and AI will continue the increase in compute power such that we can expect to see a thousand times greater performance by 2025.[34]

STRIKING A HARDWARE BALANCE

It is also important to realize that AI and ML accelerators such as GPUs do not replace CPUs. In the vast majority of cases, CPUs are still required to handle the application logic, while the GPU provides the heavy lifting where required. For this reason, the combination of CPUs with GPUs will deliver the best value of system performance, price and power.

"Ideally, it would be a combination of both—CPU cores as well as GPU cores," says Amit Marathe, Head of Software and AI at HP. "Almost every team I know of does data science and machine learning in parallel, you would want to combine the two, put CPU and GPU accelerators in one machine, and provide that as a hybrid solution for all the teams."

Mostak agrees, saying that it should not be seen as a GPU →

> **"Next-generation AI hardware solutions will need to be both more powerful and more cost efficient for the models being used in edge applications."**
>
> **Chris Nicol,** Co-founder and Chief Technology Officer, Wave Computing

EXPERT VIEW:

Accelerating business decisions with GPUs

Darren Seymour-Russell
Head of Data Science, Mudano

"Computing power is key to deriving insights, and hence advantage, not available to a competitor. So, from a financial services analytics perspective, we see the demand for GPU-accelerated deep learning platforms increasing ever upwards."

Todd Mostak
CEO and Co-founder, OmniSci

"I think that GPUs are going to dramatically transform the field of data science because people will actually come to expect it can be done interactively in real time, and when the data scientist doesn't have to go get a cup of coffee or even sleep before getting to the next step in the machine learning feedback loop, then I think we'll be able to move dramatically faster."

> **"The GPU will evolve, because right now it's just good for doing everything generally. In five years we'll have a specific design for finance and another will be biomedical research."**
>
> **Jared Dame,** Director of AI and Data Science, Z by HP

vs CPU choice. "I think you'll see a convergence in compute where more and more people are going to embrace hybrid compute scenarios, where you're going to have some of the workload running on GPUs and some of it running on CPUs," he says.

NVIDIA has introduced a new class of professional workstation—the data science workstation. This platform combines the latest Quadro GPUs, which now include Tensor cores for accelerating AI workloads, with a complete GPU-accelerated software stack to provide an integrated hardware and software solution for data science. The data science workstation gets data science projects up and running quickly, eliminating the time-consuming task of building and maintaining the multi-application and multi-library software installations required for data science workflows.[35]

According to NVIDIA's Director of Global Business Development, Geoffrey Levene, having such a "personal sandbox" for data work is a real boon for data scientists. "They are finding they can do a week's work in one day with GPU-accelerated workflows."

POWERING THE FUTURE

Next-generation GPUs are likely to be the dominant accelerator for the near future. In the next few years, we may see the GPU market diversify into models optimized for specific markets. "I think what we will see in the future is actually an increase in specialized hardware and software for specialization of AI technologies, streamlining it for various segments and verticals," says Jared Dame, Director of AI and Data Science at Z by HP. "Finance will have its own hardware/software combination that will streamline its particular workflows and then biomedical sciences will have a slightly different one, and so on between all different segments.

"Take the GPU for example. It will evolve because right now it's just good for doing everything generally. And in five years we'll look at that piece of tech and it will actually have 10 iterations of it. One will be a specific design for finance and then another will be biomedical research, one will be security, visual image recognition, natural language processing."

As businesses compete to attract and retain data scientists and analysts, the need to equip talent with the best hardware has never been clearer. And as analytics capabilities continue to improve, so will competition to get the right insights faster.

By using GPUs, these same companies can save millions on computational costs and achieve more accurate results. Keeping up the pace of accelerator innovation will lead to even greater advances.

OUTLOOK

GPU technology has already given a turbo boost for data scientists looking to develop ML models to find better solutions to problems. In the meantime, businesses that want to stay ahead of rivals need their teams to have the right hardware to make the right decisions.

KEY TAKE AWAYS

- GPUs are optimized for the kind of calculations used in ML workloads
- Advances in GPUs will drive performance gains for the next five years
- Businesses need to ensure they have the right hardware to drive decision making
- Having sufficiently powerful hardware will help retain talented data scientists

CONCLUSION

DECISION TIME

Data Strategists and IT leaders need to reconsider their data strategies to compete in the data race

So, where is your business in the data race?

The leaders are now focused on maximizing their data capital and using it to optimize processes or drive new business models. Meanwhile, the laggards are losing opportunities as they struggle to understand what data assets they have and how to make use of them. Wherever you are, now is the right time to pause and consider your data strategy in light of the key trends outlined in this report.

As most data will soon be generated at the edge of the network, businesses need to understand the kind of processes happening at their physical locations, and what kind of analytics must be carried out for a real-time response.

With the growing number of machine learning (ML) services, access to such tools will be open to everyone. While this may change the role of data scientists, they will continue to be crucial to shaping businesses' data strategies and leading the way in the data race. In a 'low-code environment' where ML applications are accessible to all, data science should perform a broader, better integrated function across a breadth of business areas far beyond IT/OT or data science.

To compete at the pace of decision making, data scientists need tools that optimize workflows. Without the compute power to embrace these emerging technology opportunities, they will be disadvantaged from the outset. For tasks such as exploratory analysis of huge datasets, this will require GPU acceleration. Businesses therefore need to invest in hardware that is up to the task.

For those who respond to these trends, the emerging possibilities in data science are endless. As big data meets compute power, analytics work is set to be transformed in organizations big and small. When the previously unimaginable becomes the everyday—that's when you'll know you're winning the data race.

Find out more about how your business can lead the data race

LEARN MORE

Sources

CHAPTER 1 - Expanded data science at the edge

1 https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders
2 https://www.sas.com/en_us/insights/articles/data-management/the-future-of-iot-on-the-edge.html
3 https://www.tuxera.com/blog/autonomous-and-adas-test-cars-produce-over-11-tb-of-data-per-day/
4 https://medium.com/@lance.eliot/edge-computing-is-big-booster-for-autonomous-self-driving-cars-c36537fcef5f
5 https://blog.bosch-si.com/bosch-iot-suite/why-edge-computing-for-iot/
6 https://www.information-age.com/data-intelligence-edge-analytics-iot-123480944/
7 https://www.sas.com/en_us/insights/articles/data-management/the-future-of-iot-on-the-edge.html
8 https://blog.seagate.com/enterprises/idc-and-seagate-what-enterprises-must-know-about-edge-computing/
9 https://www.networkworld.com/article/3224893/what-is-edge-computing-and-how-it-s-changing-the-network.html
10 https://www.networkworld.com/article/3331978/edge-computing-best-practices.html

CHAPTER 2 - Machine learning as a service

11 https://www.smartindustry.com/blog/smart-industry-connect/ai-in-the-smart-factory-a-new-approach-for-smart-industry/
12 https://www.accenture.com/us-en/_acnmedia/pdf-50/accenture-edge-analytics-mobility.pdf
13 https://towardsdatascience.com/machine-learning-in-finance-why-what-how-d524a2357b56?gi=ed60b99facf9
14 https://www.healthcatalyst.com/clinical-applications-of-machine-learning-in-healthcare
15 https://www2.deloitte.com/us/en/insights/focus/signals-for-strategists/pervasive-intelligence-smart-machines.html
16 https://www.edureka.co/blog/what-is-data-science/
17 https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf
18 https://www.idc.com/getdoc.jsp?containerId=prUS44291818
19 https://www.gartner.com/en/newsroom/press-releases/2018-04-25-gartner-says-global-artificial-intelligence-business-value-to-reach-1-point-2-trillion-in-2018
20 https://www.refinitiv.com/perspectives/ai-digitalization/the-machine-learning-trends-transforming-finance/
21 https://cloud.google.com/automl/
22 https://www.globenewswire.com/news-release/2019/07/15/1882687/0/en/Machine-Learning-as-a-Service-MLaaS-Market-Outlook-to-2024-by-Application-Organization-Size-End-user-and-Geography.html
23 https://aws.amazon.com/comprehend/

Sources

CHAPTER 3 - Accelerated GPUs

24 https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

25 https://www.globenewswire.com/news-release/2019/01/29/1706699/0/en/Graphic-Processing-Unit-GPU-Market-to-cross-80bn-by-2024-Global-Market-Insights-Inc.html

26 https://www.datascience.com/blog/cpu-gpu-machine-learning

27 https://www.wired.co.uk/article/nvidia-artificial-intelligence-gpu

28 https://www.citihub.com/insights/whitepapers/gpu-accelerated-databases-addressing-frtb-risk-results-reporting-and-other-performance-at-scale-challenges-in-financial-services/

29 https://www.hpcwire.com/2019/01/09/ibm-global-weather-forecasting-system-gpus/

30 https://www.omnisci.com/solutions/use-case/clinical-trial-analysis

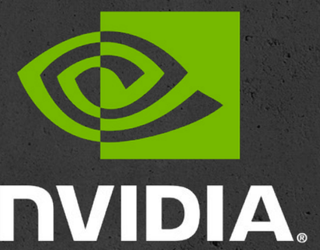
31 https://www.omnisci.com/solutions/use-case/well-logging-formation-evaluation

32 https://blogs.thomsonreuters.com/answerson/future-of-data-science

33 https://emerj.com/partner-content/artificial-intelligence-hardware-adopt-first/

34 https://www.nextbigfuture.com/2017/06/moore-law-is-dead-but-gpu-will-get-1000x-faster-by-2025.html

35 https://www.digitalengineering247.com/article/the-rise-of-data-science-workstations


NVIDIA