DELL Technologies

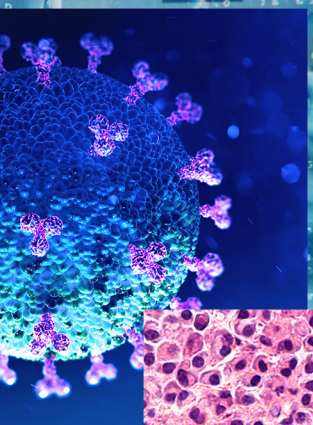
NVIDIA

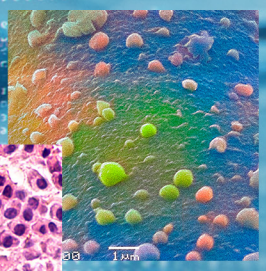
From Capture to Discovery

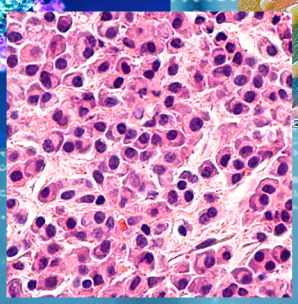
Using Data to Drive Research & Generate Outcomes

A SPONSORED PUBLICATION FROM

GEN Genetic Engineering & Biotechnology News







Accelerate Your Next Discovery

Focus on Outcomes, Not IT

Dell Technologies Validated Designs offer jointly engineered solutions from Dell and NVIDIA. These enable IT infrastructure to be easily deployed for life sciences organizations to help drive innovation and accelerate research. Capture, analyze, and store your clinical and research data on performant, scalable, and secure platforms.

- AI
- NEXT-GENERATION SEQUENCING
- CRYO-EM
- DIGITAL PATHOLOGY
- ADVANCED IMAGING
- DRUG DISCOVERY

NVIDIA. DELL Technologies

TABLE OF CONTENTS

From Capture to Discovery

Using Data to Drive Research & Generate Outcomes

INTRODUCTION

Metamorworks/Getty Images

Data, and the technology used to make sense of it, have driven and enabled many of the recent discoveries in human history. From sequencing the genome to sending astronauts to space to the rapid development of the COVID-19 vaccine, data was imperative in all these discoveries and accomplishments. Data and the findings that can be interpreted from it will continue to drive human progress, but data itself brings its own challenges.

The natural world is generating more data than ever before. With the use of technology such as the Internet of Things (IoT), samples and data points can be captured at rapid speeds from near and far. Organizations are also implementing technology that is rapidly generating massive amounts of 0s and 1s. As research turns to the digitization of workflows such as digital instruments in the field of microscopy and the use of tools such as artificial intelligence (AI), data generation is occurring at accelerated rates. This leaves organizations tasked with capturing, analyzing, managing, recalling, and archiving this data in real time. This poses a data management challenge that affects scientists, researchers, executives, and IT professionals alike. However, life sciences organizations have a unique opportunity as more data means deeper insights, better models, and enhanced research.

In order to continue driving human progress, researchers and scientists will continue to rely on data and technology. Thus, an advanced data management strategy coupled with the right IT infrastructure is imperative and can help accelerate their next discovery. The collection of articles to follow highlights the role of data in life sciences, some of the workloads that are creating this influx of data, and the importance of data management.

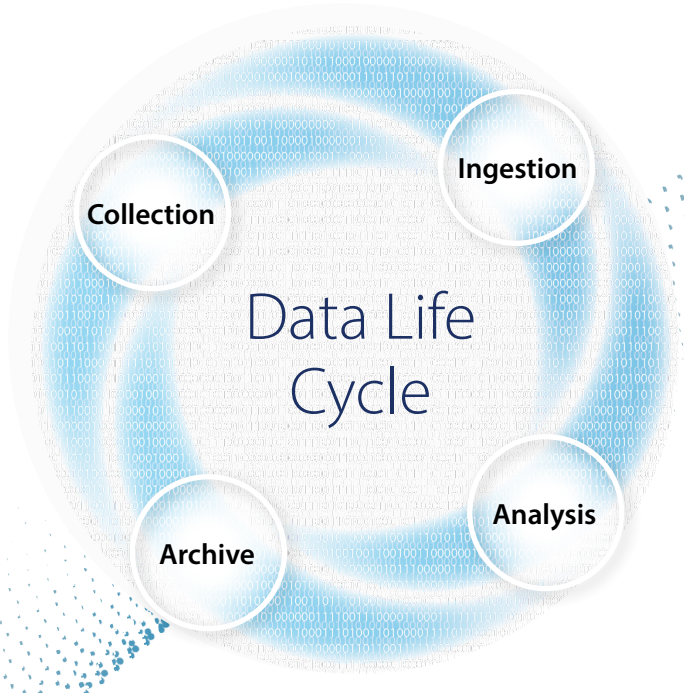
Big Data Sets, Big Challenges, Better Outcomes

The importance of IT infrastructure across the data life cycle

Collecting samples studying groups, and measuring variables may be where data collection begins, but it is just that, the beginning. As life science and research organizations look to drive human progress and make new discoveries, they look to data to unlock new insights, reach conclusions, and achieve great advancements in understanding the natural world. One of the biggest challenges that these organizations face is managing the life cycle of the vast amounts of data from collection and ingestion to analysis and, finally, to archive. The right IT infrastructure enhances researcher's abilities for data-driven breakthroughs.

Organizations continually turn to technology to advance and accelerate findings. This can be seen with technological advancements and the digitization of workflows such as microscopy in the form of digital pathology and cryo-electron microscopy (cryo-EM). These technologies can generate in excess of 20GB per pathology case and anywhere from a single terabyte to ten terabytes with cryo-EM. Another data giant is the genomic sequencer, a common instrument in many labs. One sequencer can create up to 4 TB of digital data per day. Simply put, that's a lot.

Genomics, precision medicine, drug discovery,



and clinical trials, all rely on these instruments and methods. The data generated in these processes are crucial to everyday operations at many institutions in the life sciences industry and this data holds the keys to advancing discoveries. Data generation is just the beginning of the IT complexities. High performance computing (HPC) and efficient network transmission of these large data sets need to complement the capture of this data. An integral part of the solution for effectively managing these data sets is having performant and secure IT infrastructure. This not only can aid in accelerating analysis tasks but also can simplify management of the data through its life cycle.

Greenbutterfly/Getty Images

Most of the generated data in these environments is unstructured. The analysis of the data is like assembling puzzle pieces and requires resources that efficiently store and effectively administer these large data sets. This results in infrastructure throughput requirements that can keep up with initial write demands while also supporting image processing and analysis demands. Take for example the secondary and tertiary analysis phase of next-generation sequencing (NGS), these processes consist of comparing a genome to a DNA reference sequence to generate a list of variants and then interpreting these differences and annotating them. These steps take place after the initial sequence and write have occurred, generating more data and relying on the HPC and data storage to deliver performance for the analysis.

The use of HPC and the supporting infrastructure also opens the door to machine learning and AI technologies. By turning to digitized workflows researchers can now feed data generated and collected by modern instruments into these algorithms. These technologies can help speed analysis, automate processes, and accelerate discoveries for life sciences organizations. Some of the uses of

AI include identifying abnormalities and patterns samples in images and creating models.

Once the analysis has been finished and the findings recorded, data is sent to the archive. In many situations, data is capital and needs to be secured like any other property. In these compute-intensive life sciences environments, data storage is responsible for feeding the data on demand, while also keeping it safe during long-term retention. As organizations continue to generate terabytes a day, the need for scalable platforms that can tier performance is necessary to complete the data life cycle.

Getting the right infrastructure in place can be a challenge as needs and requirements can vary across different organizations and use cases. End-to-end solution planning is required to ensure the right technology is selected from the start to support research and discovery. This is where Dell Technologies and NVIDIA can help.

As a trusted IT partner, Dell Technologies offers the infrastructure with the performance, scalability, and reliability needed for these processing-intensive workloads. The expansive portfolio includes servers, enterprise-class unstructured data storage platforms, hyperconverged infrastructure, data protection, and networking in addition to workstations, PCs, and high-resolution monitors. Dell Technologies and NVIDIA have the unique capability of supporting data from collection to discovery with technology that ranges from edge devices to multicloud connectivity for data analysis and GPU utilization.

Dell also offers jointly engineered solutions with NVIDIA that combine the HPC, networking, storage, and software that are ideal for life science applications, including technology stacks specifically configured for NGS and AI deployments. These validated and tested architectures, coupled with NVIDIA Clara™ software, enable life science organizations to confidently move forward with research and development projects.

Advancements being made by life science organizations are truly phenomenal, and the future offers great promise to improve the lives of people across the globe. Dell Technologies strives to be a trusted and strategic partner to organizations focusing on advancements in healthcare, life sciences, and pharmaceutical research. From edge, to core, to cloud, Dell Technologies offers the technology to support your data across the lifecycle. ■

Biopharma Is Going Digital … Bit by Bit

Digital manufacturing elements are being adopted piecemeal, mostly by the largest and the newest firms, without full connectivity across the supply chain

Biopharmaceutical companies are embracing Industry 4.0 ideas and using digital technologies to revolutionize manufacturing and improve the quality of medicines. Well, at least a few of the larger firms are. Others will take longer, say experts.

Going digital takes resources. Setting up the infrastructure needed to gather data and feed it back to analytical and modeling systems is a complex process that takes time and costs money, says Caterina Minelli, PhD, principal research scientist at the National Physical Laboratory, the national measurement standards laboratory for the United Kingdom.

"Digital manufacturing approaches [encompass technologies such as] artificial intelligence, big data, cloud-based computing, and the Internet of Things (IoT), as well as real-time predictive modeling, robotics, and automation," she continues. "Some aspects of these elements can be combined to deliver digital connectivity across the entire supply and manufacturing chain, enabling data to be connected and allowing information to flow through the chain and be easily retrieved.

"This connectivity requires a significant data infrastructure underpinned by standardized data ontology, terminology, and instrument interfaces. It also requires the ability to manage big sets of data."

As a result, biopharma's move toward digital manufacturing has been led by larger companies that have the capacity to invest. GlaxoSmithKline (GSK), for example, has been working to digitalize its operations for several years. GSK started using digital technology to enhance interactions with patients and consumers. The company is also interested in using digital technology to optimize manufacturing operations.

GSK recently indicated that it is working with Siemens and Atos, two of the world's leading expert companies in digital transformation and

GlaxoSmithKline

technology, to realize the "vaccine factory of the future." GSK's vaccine factory will incorporate a digital twin. (A digital twin is a complete and real-time simulation of an entire manufacturing process. By providing computational models that let scientists test and modify experimental parameters, a digital twin can improve process development more efficiently.)

GSK expects to deploy digital twins more generally. "Traditional process development methods are still being used, but our ambition is to progressively implement digital twins," says Sandrine Dessoy, digital innovation lead, GSK. "The development of digital twins is an investment not only for R&D, but also for manufacturing and quality control activities."

DIGITAL TWINS

GSK's approach is to work with industrial data management experts to build digital models of candidate processes and then to try to predict how they will perform before they are scaled up and implemented on the factory floor.

"We started with a proof-of-concept project in collaboration with Atos and Siemens to determine how a digital twin for a vaccine production process should be defined, built, and demonstrated," Dessoy detailed. "The next

Encouraged by the success of the digital twins project at GSK Vaccines Wavre, GSK looks forward to implementing digital twins technology throughout vaccine development and production. Early in a vaccine project, high-throughput experimentation in combination with a digital twin could produce data that would allow theories to be tested. Further down the line, a digital twin could drastically reduce real experimentation, conserving materials and reducing energy consumption.

step was to move from the minimum viable product (MVP) to a robust, scalable future-ready platform—this was the main challenge."

MVP is the technical term for a system that functions based on the bare minimum of components. According to Dessoy, the "minimum number" for GSK's digital twin turned out to be quite large.

"Many components, such as fast hybrid models, sensors, process analytical technology, automation, and data streaming and modeling platforms, had to be integrated while a fast data flow, from equipment to model and to automation, was ensured," she added. "The solution needed to be GMP compliant and flexible enough to be implemented in R&D and manufacturing with a diverse range of equipment."

Despite these challenges, the digital twin project has been a success. It has given GSK's vaccine manufacturing team more options, and according to Dessoy, the digital twin approach is being applied to well-established and newly

introduced production processes alike.

"We have been running two approaches in parallel," she relates. "One approach involves the development of a complete digital twin for new vaccines; the other, the implementation of a digital twin for monitoring of lifecycle projects.

"The complete digital twin will control the process and will be transferred to GMP and production alongside the vaccine project. The lifecycle digital twin follows a simpler approach: feed the model with years of data to better monitor the process in real time and detect deviation, but with no regulatory impact. This simpler digital twin is already delivering value in production and increasing capacity."

BIOTECHNOLOGY BOOST

So, for companies like GSK that have the requisite resources and time to invest in development, digital manufacturing is becoming a reality. However, for many smaller or less deep-pocketed companies, the adoption of digital technologies is taking place more slowly.

"Different organizations and market subsegments are adopting different elements of the digitalization approach," explains Minelli, who works with manufacturers from a range of industries on the analysis of uncertainty and variance in digital systems. "There are many who have developed and optimized processes around digital notebooks and wearable technologies like augmented and virtual reality headsets for training and maintenance purposes.

"However, true digital connectivity across the supply chain is still in various stages of development, with a range of new standards and digital tools required to fully realize the benefits available through a complete shift to digital control."

This view is shared by Richard Vellacott, CEO of BiologIC Technologies, an industrial software and systems developer. He maintains that for mid-sized biopharma firms, the transition from analog production is a daunting prospect: "Biopharma companies are incrementally adopting Industry 3.0 digital manufacturing but will never achieve full-stack Industry 4.0 digital adoption because they have too much interest in defending the status quo."

Vellacott allows that the situation is different for emerging biotech companies. These companies are building their manufacturing infrastructure from the ground up, so they are more likely to embrace digitalization.

According to Vellacott, the shift to Industry 4.0 will occur in biomanufacturing much like it occurred in other kinds of manufacturing: "Transformative digital manufacturing—because it is so disruptive to incumbents and existing supply chains—will emerge from new companies that have fundamentally reimagined the power of digitalization in what we might call the Kodak moment for therapy."

Vellacott suggests that the move to full-stack digital manufacturing may be accelerated in light of recent events. "The pandemic has painted a clear picture in the public eye of how slow, expensive, and inflexible our classical biomanufacturing processes are even in response to urgent threats," he observes. "The industry has to focus on order-of-magnitude improvements. Only full-stack Industry 4.0 digital manufacturing can make order-of-magnitude improvements to therapy cost, the availability of rapid on-demand manufacturing at the point of need, and elastic manufacturing capacity."

"If we achieve this, we will vastly increase accessibility for patients. The benefits will be transformative. They will include mitigation of pandemics, pervasive adoption of personalized therapies, and far more sophisticated combination therapies."

Vellacott says that technological solutions are being developed with emerging companies in mind. Indeed, he points out that such solutions are being developed by his own firm.

"BiologIC's biocomputer is an example of a full-stack digital manufacturing approach," he declares. "It is based on highly integrated designs enabled by pioneering innovations in additive manufacturing and novel methods of operation, sensing, and data intelligence. Key features include software configurability to enable elastic capacity, digital simulation to allow for novel process development, and machine learning to optimize manufacturing strategies for complex bioproducts."

PANDEMIC IMPETUS

Like Vellacott, Minelli is of the opinion that the coronavirus pandemic is bound to accelerate the biopharma industry's adoption of digital technology. "The pandemic," she explains, "has necessitated an increased pace of adoption of digitalization solutions across biopharma manufacturing." She suggests that when companies integrate different systems to solve immediate problems, they set precedents for collaborations of larger scope.

"The pandemic has pushed the pharmaceutical sector workforce to adopt digital methods of communication," she adds. "This, in turn, has forced manufacturers to ensure that the employees have access to data and analysis tools remotely. Finally, remote access has accelerated the adoption of digital manufacturing techniques. Employees can be productive off the shop floor and spend time generating insights about manufacturing processes."

REGULATORY INNOVATION

Digital technologies will play a greater role in drug production whether they are pioneered by large biopharma companies or emerging biotech firms or both. In any case, regulators will need to prepare for the change.

Gorodenkoff/Getty Images

"Inevitably, technological revolutions happen at a pace far quicker than the response time of regulators and governments, and they will have to respond after the fact," Vellacott states. "Digital companies cannot slow their pace. Instead, they have to navigate the environment with new strategies that fundamentally assure therapy efficacy and patient safety while leading the public debate based on the need to radically reduce therapy costs and increase their accessibility.

"Forward-thinking governments should invest in Industry 4.0 digital manufacturing technologies if they wish to develop sovereign manufacturing control and supply chain resilience. The regulatory environment is most in flux with the newest cell and gene therapies. I believe that the Food and Drug Administration in the United States and the Medicines and Healthcare products Regulatory Agency in the United Kingdom are responding in the best way they can while always ensuring quality for patients." ■

Data Management Plan Vital for Digital Manufacturing Focus

Use of data in drug manufacturing has increased in recent years, says Phil Braun, director of client solutions at NECI. He adds that the emergence of new product types, the focus on rare diseases, and technology advances are driving the change.

"The influx of new therapeutic modalities and small market/rare disease, aka non-blockbuster, drives a relook at the business as a whole," he notes. "The shift to smaller manufacturing volumes, flexible facilities, innovation in compliance approaches, rapid release paradigms are causing upstart organizations to consider digital out of the gate, aka 'greenfield digital.' This is matched by advances in technology (cloud, big data lessons learned for

Since 2016 Moderna has had the goal of being the first 'all digital biotech.' Moderna's digital strategy was integrel to the speed in which they developed their COVID-19 vaccine.

Puddin Tain/Creative Commons

Laurence Dutton/Getty Images

enterprise data context) that enable digital."

He cites COVID-19 vaccine developer, Moderna, as an example of a firm that has adopted a digital strategy, explaining the approach helped speed up development activities.

"There are 'digital unicorns' like Moderna which since 2016 had had the goal of being the first 'all digital biotech,'" continues Braun. What began early in their R&D, digital innovation has permeated the culture and carried through to native digital clinical manufacturing.

"As they near the potential to commercialize for their COVID-19 vaccine, the commitment to digital has fueled their ability to start-up their new vaccine product candidate," says Braun. "The timeframe from gene sequence to process development to clinical manufacturing is completely different than traditional biopharma. The role of digital is compelling in this case."

DIGITAL MONSTERS

For companies looking to digitize manufacturing operations piecemeal over a longer period, the process is likely to be more challenging and complex.

"Traditional biopharma has grown up like a 'digital Frankenstein'—siloed decision making along traditional lines separating operations like R&D, clinical trials, commercialization, and quality control has created islands of automation, data, and analytics," says Braun.

For such firms, Braun explains, digitization requires consistent data contextualization and management practices across the whole enterprise to enable process analytics, supply chain

visibility, and predictive quality assessment.

"This is a mammoth challenge and investment, given the starting points in the 'islands' across the enterprise," he points out. And the connectivity that does result to tie the data together creates definite challenges in cybersecurity and performance."

Amgen is a good example of a firm that successfully pooled data to create a "data lake." Although the process was costly, according to Braun, "we understand that the investment was very, very significant."

COMMERCIAL FOCUS

Another advantage of adopting a digital manufacturing strategy early is that it makes it easier to plan for commercial-scale production.

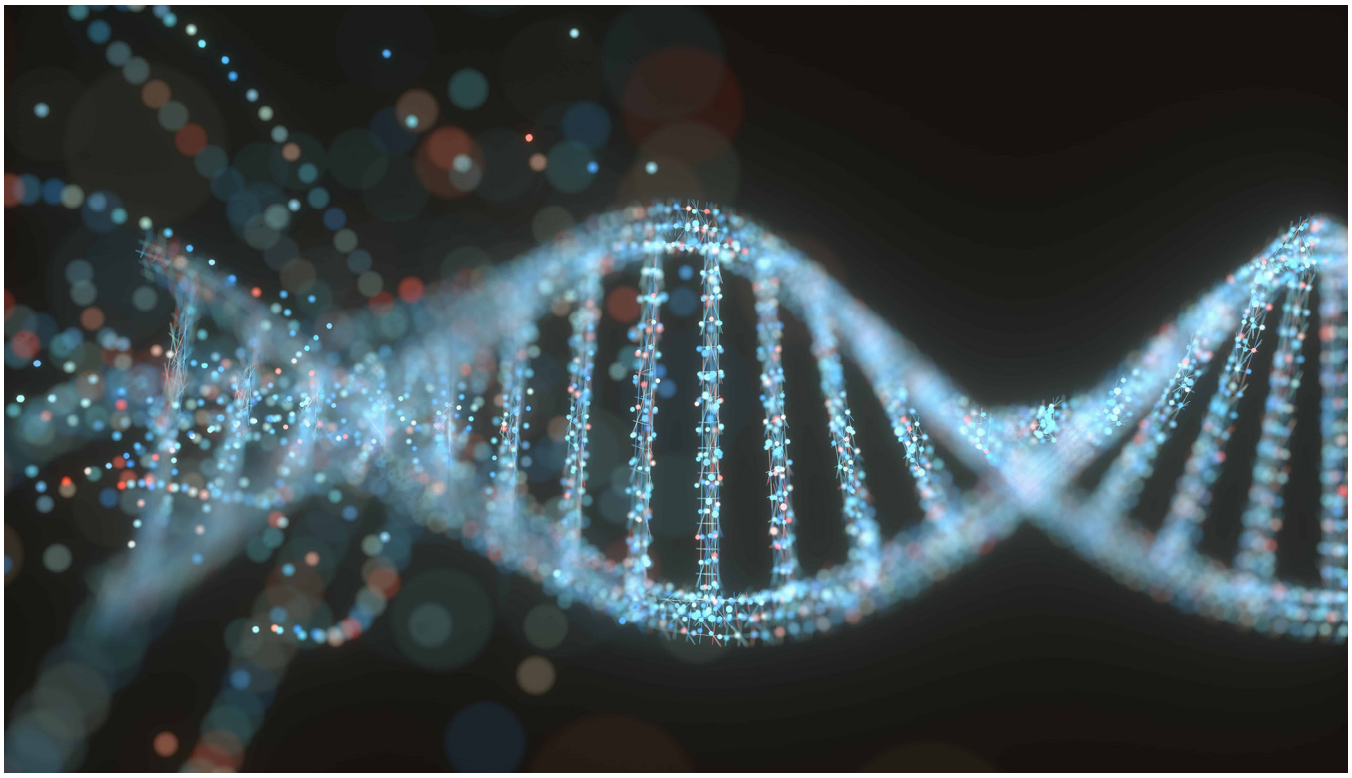
Traditional process development tends to focus including technologies and sensors required to measure that overall goals have been achieved according to Braun, who says systems that would make commercial production more efficient are often overlooked.

"Many bioprocess decisions are driven early in the PD lab or in pilot operation," according to Braun. "The goal there is process endpoints supported by data but not necessarily scalability. Therefore, the automation and digital decisions for future manufacturing, even at the basic connect and collect level are often underserved as the initial decision points don't include digital endpoints, but rather process endpoints."

Fortunately, this potential shortcoming has been recognized by the bioprocessing technology sector, Braun says.

"We see equipment OEMs understanding this challenge more and more and looking to automation digital experts as part of their teams to better capitalize on the outcomes promised by holistic digital design," points out Braun.

"One of our big roles is to be the glue amongst the disparate systems that enter a clinical commercial facility such the objectives of multiple manufacturing stakeholders are met regardless of the unit op starting points." ■



Ktsdesign/Science Photo Library/Getty Images

Addressing Key Data Issues Arising from Next-Generation Sequencing

As the use of next-generation sequencing (NGS) increases in the life sciences industry, and the practice is further adopted in healthcare for personalized medicine, data generation is growing at an accelerated rate. NGS has the ability to create massive data sets. With each analysis averaging 120 GB/genome, a single sequencer can generate up to 4 TB of data per day. That data increases when looking for variants or comparing genomic data from tumors. As many facilities employ multiple sequencers to support workloads, the data generated has the potential to grow exponentially.

Then consider that sequencing the genome is only part of the process: analysis, annotation, and analytics also generate data and add to the overall IT requirements of a sequencing environment.

Why is IT infrastructure so important?

Big data sets pose big questions. Where should the data be stored? How should the data be analyzed? How can analysis be accelerated? The answer is that IT infrastructure can help. To make

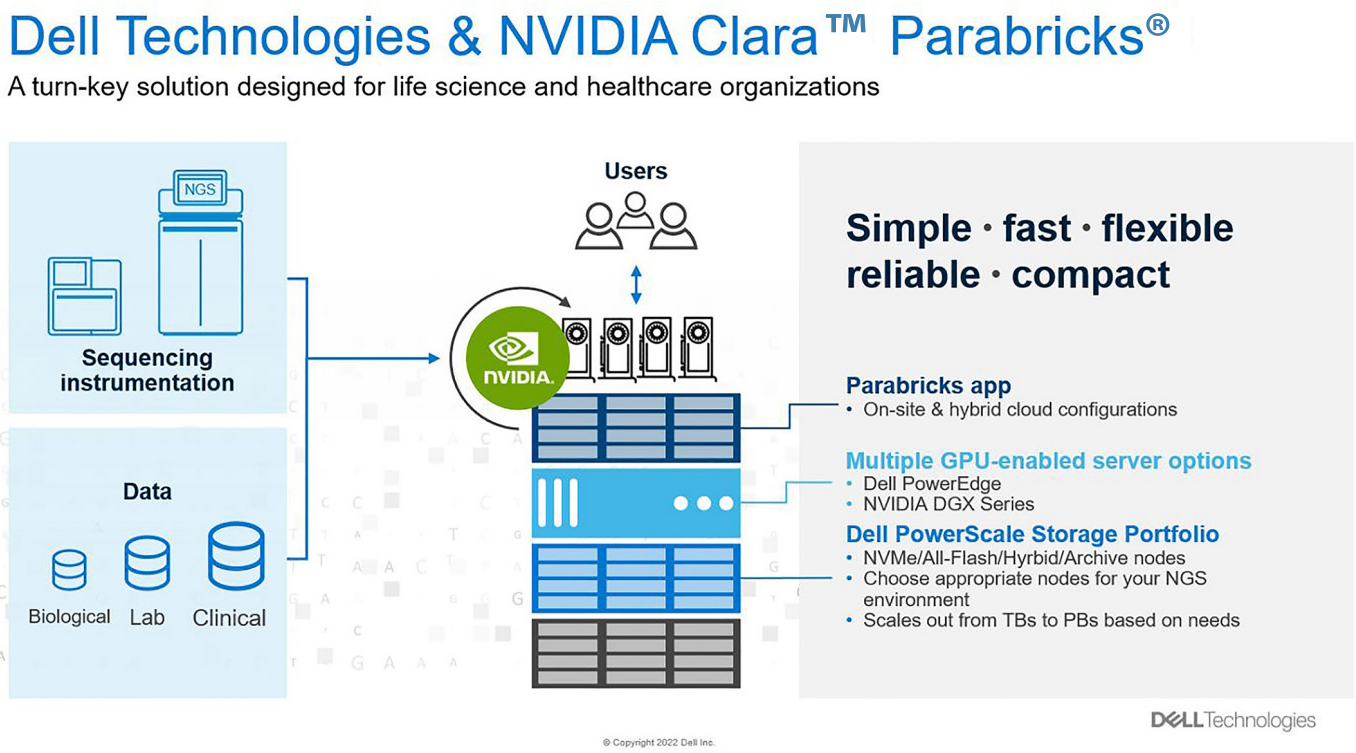
informed decisions, speed analysis, and draw clinical findings, an organization needs an IT environment that can keep up.

To ensure that the rate of secondary analysis keeps pace with the rate of raw NGS data generation, organizations should—at a minimum—ensure they have sufficient computing and storage resources matched to the output capacity for a fleet of sequencing instruments. Without this, the organization risks an analysis backlog.

An organization's archive capacity requirements will vary according to organizational goals and processes, but that capacity is typically determined by the organization size and type, data access types, frequency of access, retention periods, and intended use—for example, research or clinical use.

What is Dell's part in genomic sequencing?

As an IT vendor, Dell Technologies is a trusted and strategic partner to many companies in the healthcare, life sciences, and pharmaceutical industries. Dell offers not only the individual pieces of the IT puzzle, but we also offer a trusted and validated design for genomics that leverages Dell PowerEdge servers with NVIDIA® Ampere GPUs, NVIDIA Clara™ Parabricks® software, Dell PowerSwitch networking, and Dell PowerScale storage. This architecture combines IT resources required for various forms of genomic data analysis in a compact, easily scalable solution. A



typical solution capable of processing 20 human genomes per day (50 X coverage).

Who are your customers? How do you service your customers?

Dell Technologies is a trusted partner to our customers. We are uniquely positioned to offer true end-to-end solutions with monitors, PCs, HPC, data storage, and data protection. We service large and small life sciences companies including pharmaceutical, med/tech, food processing, and environmental, as well as higher-ed, private, and government research institutions, in addition to many healthcare customers who have started their transformation to delivering personalized medicine.

What strategic relationships does Dell have that enable the company to be a leader in the space?

Dell Technologies has an extensive partner ecosystem that better allows us to service our customers. In the field of genomics, we have strong relationships with NVIDIA, PetaGene, and Vyasa Analytics.

- NVIDIA uses Clara Parabricks, a GPU-accelerated computational genomics application framework that can greatly accelerate analysis.
- PetaGene specializes in genomic data compression, offering up to 60–90% reduction, while remaining lossless with transparent readbacks.
- Vyasa Analytics offers highly scalable deep learning software and analytics that permit organizations to ask complex questions across large-scale data sets to gain critical insights for better decisions.

What other workloads does Dell focus on in the life sciences?

In addition to sequencing, we focus on image file and object management workloads including digital pathology, cryo-EM, medical imaging, and other data-intensive tasks. Dell can support any computational workload in the life sciences space with industry subject matter experts and end-to-end solutions from the edge to the core to the cloud. ■

Brief

Validated Designs for HPC Life Sciences, a jointly-engineered solution from Dell Technologies & NVIDIA, helps accelerate analysis for next-generation sequencing.

Making Advanced Microscopy Possible with Dell Technologies & NVIDIA

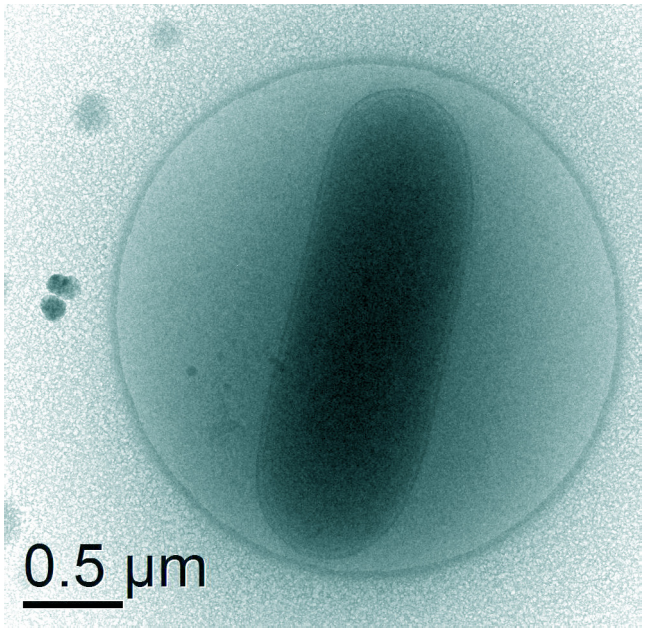
How biomolecules function and interact is fundamental to understanding diseases, developing new drugs, and administering medical treatments

Cryogenic electron microscopy (cryo-EM) is an imaging method that allows direct observation of proteins in native and near-native states without dyes or fixatives, giving researchers the ability to study cellular structures, viruses, and protein complexes in molecular detail.

This reconstruction of three-dimensional, near-atomic resolution structures of biomolecules often requires thousands of images and complex computing, making it difficult to deliver high-resolution structures.

Cryo-EM can help meet this challenge. Its success will be shaped by wider adoption, increasing data sizes, maturing of the market, and the rise of deep learning.

Cryo-transmission electron microscope image of Shewanella oneidensis MR-1.



Courtesy of Pacific Northwest National Laboratory

MASSIVE DATASETS

The scope and complexity of cryo-EM data have greatly increased with advancements in automation and visual technology. Cameras with higher sensitivity capture images at faster frame rates. With improved sample preparation, automation for data acquisition, and instrument uptimes, the requirements for data processing and computing continue to increase. For example, within a typical experiment, it often takes 1,000–8,000 images, captured from 4–8 terabytes (TBs) of raw image data, to generate high-resolution, single-particle maps. Some CryoEM labs can easily generate

about 15 TB of data per day and need to store that data for long periods. Understandably, labs can quickly run into capacity and computational challenges.

Over the past few years, almost all compute-intensive steps in single-particle workflows have been ported to take advantage of GPU processors, which shorten processing times dramatically. To keep up with increasing data sizes, cryo-EM applications need to be optimized for higher-end GPUs. Dell PowerEdge servers that harness the capabilities of NVIDIA GPUs can help to keep up with this data-intensive workflow. Having a reliable high performance computing (HPC) solution in place means data is captured and available for analysis. By leveraging jointly engineered solutions from Dell and NVIDIA, your IT infrastructure can keep up with the demands from your microscopy technology. When it comes to data storage that offers the throughput necessary for analysis, choosing flexible, easily expandable solutions like Dell PowerScale is key to remaining focused on research rather than IT infrastructure.

THE MATURING MARKET

Traditionally, processing cryo-EM image data to uncover protein structures and create high-resolution 3D maps requires expert intervention, prior structural knowledge, and weeks of calculations on expensive computer clusters. As it becomes mainstream, cryo-EM is fostering demand for commercial-grade, non-expert software. These software solutions involve using algorithms to automate specialized and time-intensive tasks.

Within leading pharmaceutical companies' structure-based drug-design pipeline, for example, they use the cryoSPARC software suite.

The cryoSPARC platform utilizes NVIDIA GPUs to enable automated, high-quality, and high-throughput structure discovery of proteins, viruses, and molecular complexes for research and drug discovery.

INFUSION OF DEEP LEARNING

Selecting individual protein particles in cryo-EM micrographs is an important step in the single-particle analysis. It's challenging to identify the particles due to a low signal-to-noise ratio and the tremendous variations that occur in biological macromolecular complexes. By leveraging positive-unlabeled learning, a small number of example protein projections can train a neural network to detect proteins of any size or shape. Topaz, an open-source application with this capability, detects significantly more particles than other software methods when tested. Powered by NVIDIA GPUs, it drastically cuts down the amount of data that needs to be manually labeled.

HPC IN THE CRYO-EM WORKFLOW

Cryo-EM methods are opening up opportunities to explore the complexity of macromolecular structures in previously inconceivable ways. From early-phase research systems to large data centers, GPU-accelerated HPC is enabling end-to-end workflow acceleration. By optimizing key workloads for data acquisition and single-particle reconstruction, GPUs continue to deliver paths to scientific and healthcare breakthroughs. With GPU-based computing and deep learning, advancements in cryo-EM will increase its reliability and output and, ultimately, its adoption and success. Layering this compute capability with powerful, flexible and secure storage infrastructure that can take advantage of cloud-based workloads will continue to align research requirements with storage needs and maintain the pace of adoptions. As life sciences organizations continue to develop and invest in advanced, data-intensive workflows, having powerful technology from partners such as Dell Technologies and NVIDIA will become imperative. ■

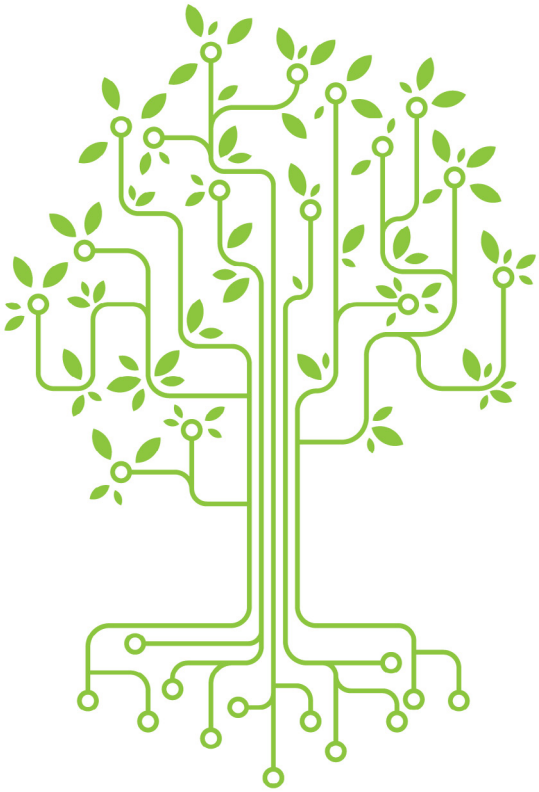
Building Data Ecosystems for Modern Biomedical Research

To derive value from fast-growing data collections, research organizations are using high-performance platforms, exploring standards, and prioritizing FAIRness

Research in the life sciences, biotechnology, and biomedicine is entering a period of disruption in how scientific data are collected and analyzed. This disruption is having widespread impacts since data have always been central to research. Consider how the typical research project progresses: a problem is defined; a study is conceived (and hopefully funded); experiments are run; data are collected, analyzed, interpreted, and used to support conclusions; and conclusions are communicated to select audiences or the broader scientific public.

In research, data connect the experiment and the generation of knowledge. Increasing the speed of that process is more important than ever. Quality of life can be improved, and lives saved as a result—as the COVID-19 pandemic has illustrated so vividly.

Wrangling data has become more challenging because of the enormous growth in the production of scientific data. Disruptive scientific innovation requires organizations to transform these petabytes of data into a strategic asset by making them findable, accessible, interoperable, and reusable (FAIR). Organizations that build effective scientific data ecosystems to harness the knowledge inherent in their data will pioneer



rajitha2r/Getty Images

new discoveries the fastest.

For the past 20 years, biomedical research has been utilizing laboratory technologies that generate unprecedented amounts of data in a short time. This increased rate of data generation has been primarily driven by the pace of innovation in laboratory technology. These data are produced by instruments such as high-throughput genomic sequencers, next-generation fluorescent microscopes (like lattice light-sheet microscopes), cryo-electron microscopes, flow cytometers, and a host of other imaging, resonance, and data collection technologies.

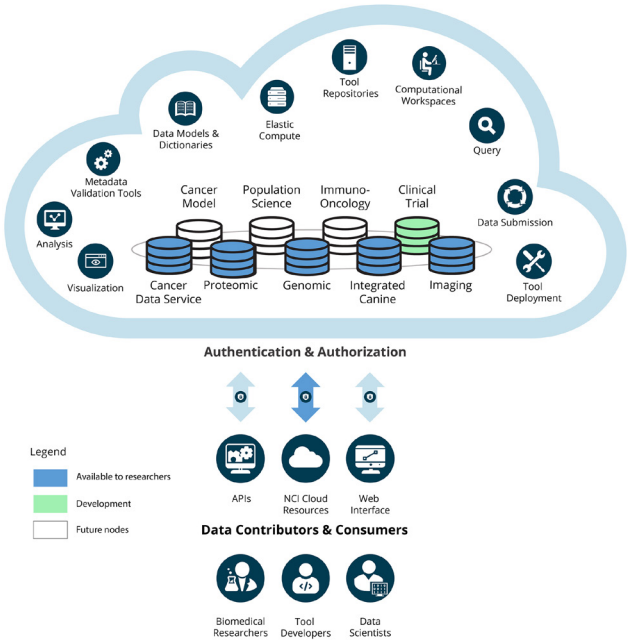
By applying these technologies, researchers hope to dig deeper into the problems that plague humanity, find new disease treatments, and realize exciting concepts such as precision and personalized medicine. These technologies are inspiring researchers to launch masterful studies and collect data that may, upon analysis, help us chip away at the mystery of life.

This explosion of data collection led to the information age and the coining of ill-defined buzz terms like "big data." An unintended consequence of this change was that information technology (IT) personnel in life sciences organizations were caught off guard. They suddenly had to support massive amounts of data (a petabyte in 2012, hundreds of petabytes today) without the necessary budgets, skill sets, or support systems.

Prior to this data tsunami, IT groups mostly supported document storage, databases, email, web, security, and printers. Now that the data deluge is upon us, aspects of the hard-earned expertise of the high-performance computing (HPC, also known as supercomputing) community are slowly being adopted to help them adapt.

Life sciences organizations began to invest in scientific computing by building modest to large HPC systems, installing large storage systems, and working to better connect laboratories to data centers so that data could flow more easily. These adaptations took many years and occurred at varying rates per organization. Ultimately, though, the data center and advanced computing technol-

NCI Cancer Research Data Commons (CRDC)



The NCI Cancer Research Data Commons (CRDC) is a cloud-based data science infrastructure that connects data sets with analytics tools to allow users to share, integrate, analyze, and visualize cancer research data to drive scientific discovery.

ogies became as integral to life sciences research as a microscope or a next-generation sequencer.

Modern research projects cannot be done without an advanced technology infrastructure. HPC, storage systems, high-speed networks, and public cloud environments have become essential lab tools, not merely devices for IT to operate and maintain.

DROWNING IN DATA

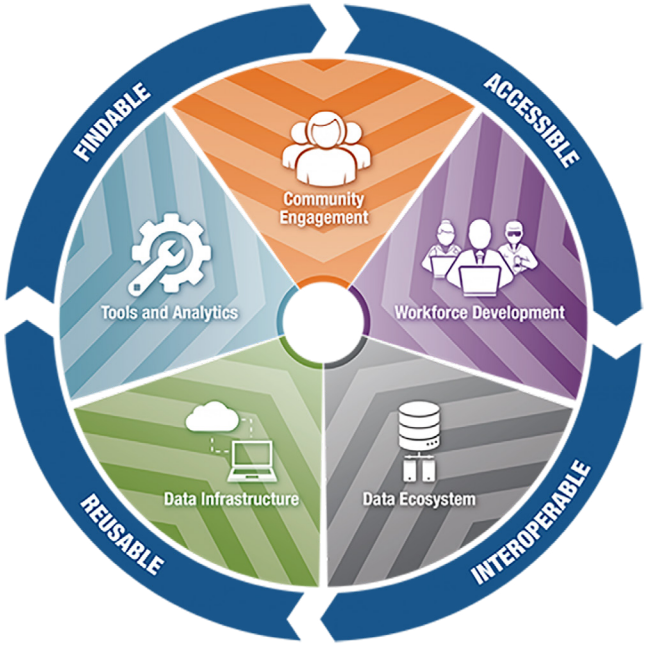
Today, the scientific community confronts a data landscape that is not just more expansive, but also more varied. There are now vast repositories of scientific data and organizations creating every conceivable type of data architecture (i.e., data lakes, oceans, fogs, and islands), culminating in the growth of data commons as a fundamental scientific data architecture. With innovations in data science and bioinformatics, it may be possible to start discussing how the current pace of data accumulation can be maintained. One possibility is the development and adoption of common data standards for biomedical data.

However, most researchers in our field are still creating their own data formats and metadata assignments while putting their data wherever they think is best for their research. Without effective standards, high-value data are being spread across every data storage medium imaginable, including portable disks that end up being shoved into drawers. IT is unable to keep up with hardware acquisition, and the data are piling up stochastically.

The backlog of data has led most biomedical organizations to turn to public cloud providers (such as Amazon Web Services, Google Cloud, and Microsoft Azure) to alleviate their on-premises logjams. Public clouds have also proven highly beneficial for collaborative data analytics by placing data and computational resources in close proximity outside of the security restrictions local enterprise networks place at their borders.

However, the sprawl of data, the backlog of data analysis, and the difficulty of combining multiple datasets for more detailed studies have led to the realization that collecting data alone is not useful. To give value to the data, it needs to be analyzed, interpreted, and converted into knowledge for the community to consume. This realization has led the life sciences community to the end of the information age and into the analytics age.

Much of the world has turned to artificial intelligence (AI), specifically machine learning (ML) and deep neural networks, to solve the problem of interpreting large and potentially unstructured datasets. The hope is that by creating inference models that represent the data, it will become possible to perform analyses more quickly and to assign meaning more easily. This methodology, which has been hyped by hardware and software vendors for the last few years through marketing



The Office of Data Science Strategy (ODSS) leads implementation of the NIH Strategic Plan for Data Science through scientific, technical, and operational collaboration with the institutes, centers, and offices that comprise NIH

campaigns, had the positive consequence (perhaps unintended) of driving a desperate field to explore its viability and has resulted in several important innovations in data science.

Unfortunately, there are still several issues facing the life sciences community. First, without unified data standards and common approaches to data governance, data will never become FAIR—an end goal for the community to efficiently utilize public data with any other research project. Additionally, well-curated data tagged with common and actionable metadata that clearly define what the data represent are needed to create the necessary datasets to train the ML models. If the data aren't clear, ML models will not be useful.

Additionally, despite the claims by the industry, deep learning is not the magic bullet algorithm that will save everyone from their data sprawl. It is helpful only in certain situations and only when data is well curated.

Once the life sciences community gets a better hold on its data, starts progressing toward unified data standards, and accomplishes FAIRness to a meaningful degree, we'll begin to understand the actual value of the collected data. We'll make more informed choices about which data to keep and what qualifies as intermediate or lower value items in storage schemes. If we can slow the exponential growth of data, start managing the backlog, and work toward common data platforms (data commons are a good start), we'll establish true scientific data ecosystems across the industry.

DIGITAL TRANSFORMATION

The process of working toward a well-established and functional scientific data ecosystem is called "digital transformation," another buzz term. For digital transformation to be successful, the scientific data ecosystem must be designed with a holistic approach aligned with the organization's scientific mission with advanced technology at its core.

Our company, BioTeam, has been working with large organizations for the last several years building up digital transformation strategies and platforms. Starting with the National Institutes

of Health, we've been working with the Office of Data Science Strategy (ODSS) facilitating collaboration with the Institutes and Centers at NIH to form a large data ecosystem out of the many data repositories that NIH funds and maintains.

We've also been working with large pharmaceutical companies to make better use of their data by improving their IT infrastructure to support science, creating collaborative scientific computing environments in the cloud, and building data commons customized to the needs of the organization. The most public of these efforts has been the development of an internal data commons for Bristol Meyers Squibb (BMS) Research and Early Development using the open source Gen3 framework (maintained by the University of Chicago) as the foundation for the system.

Most biotechnology and pharmaceutical companies that we work with are going through various stages of this transformation. Academic institutions and federal science agencies are all either planning or already working to implement some or part of a digital transformation strategy.

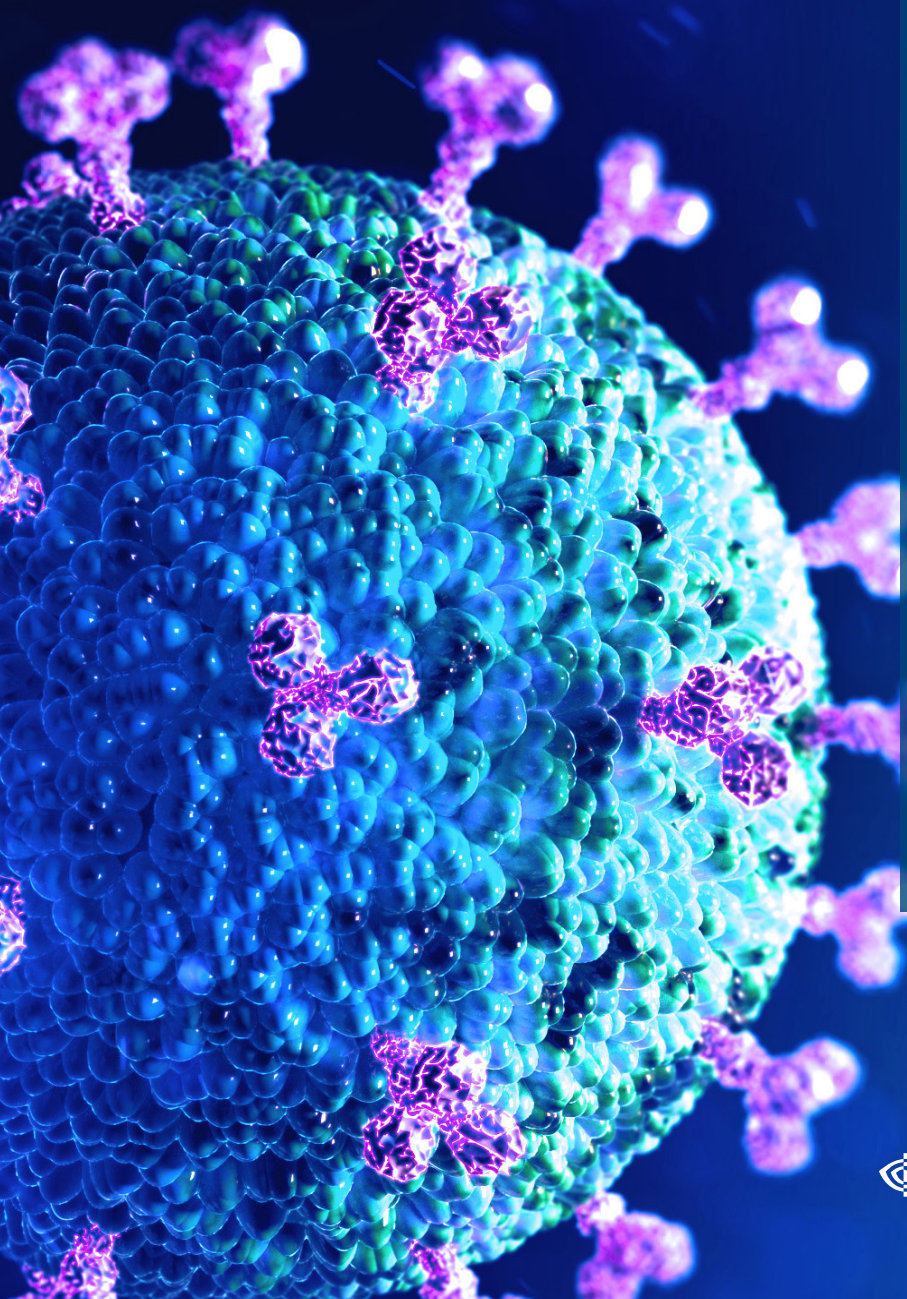
The establishment of productive scientific data ecosystems is within our reach, but it will require unprecedented collaboration. To help foster community-wide agreement, global governing bodies may need to offer incentives and put enforcement mechanisms in place. As terrible as the COVID-19 pandemic has been, it taught us a valuable lesson about the value of collaboration. It proved the need for this kind of coordination as researchers and scientists around the world attempted to rapidly work together to mount a response against this novel and devastating disease.

The aforementioned barriers proved to be profoundly challenging to overcome during the pandemic response, even with public clouds, supercomputing centers, and other organizations donating and prioritizing use of their resources to anyone working on the problem. The lack of an established scientific data ecosystem has drastically curtailed our progress. Nonetheless, this is a truly exciting time in our field as we transition into the analytics age. Let's work together. Let's change the culture in the life sciences, biotechnology, and biomedical research. And let's build lasting scientific data ecosystems that will drive our understanding of life on Earth to the next level. ■

Ari E. Berman, PhD, is CEO of BioTeam, a bio-IT consultancy firm that has been serving the life sciences/biotechnology community since 2002.

From Collection to Discovery

Your Trusted IT Infrastructure Partners



Life sciences organizations need high-performance, scalable, and secure infrastructure

to keep up with demanding workloads that drive discovery. Dell Technologies & NVIDIA understand these needs and requirements with industry expertise and validated IT solutions. From drug discovery to precision medicine and everything in between, we are here to help.

SERVERS

NETWORKING

STORAGE

VALIDATED DESIGNS

GPUS

MULTICLOUD

MONITORS & WORKSTATIONS

NVIDIA. DELL Technologies