# THE RIGHT DEEP LEARNING SOLUTION FOR YOUR BUSINESS

NVIDIA

“The McKinsey Global Institute reckons that just applying AI to marketing, sales, and supply chains could create economic value, including profits and efficiencies, of $2.7T over the next 20 years.”

*The Economist*

# WHERE TO START?

Businesses need to assess three key areas before they can determine the right deep learning solution for them:

- **Understand** your capabilities in terms of resources and staffing
- **Look** to relevant use cases that match your situation
- **Leverage** available data and manage your data well

## **UNDERSTAND** YOUR READINESS FOR DEVELOPMENT AND DEPLOYMENT

One of the biggest hurdles for businesses is curating an experienced in-house team with the right background and talents for designing their deep learning solutions. What stage of deep learning development is your AI team in—"crawl," "walk," or "run"?

### CRAWL

Good news! Whether you need an introduction to AI fundamentals or a workshop for your team, the NVIDIA Deep Learning Institute can help. Your team can also leverage discussion forums. Why recreate the wheel when you don't have to?

### WALK

You can take the successful work of others and begin to experiment. Implement their projects and adjust and scale them to fit your data and organization. Leveraging your basic skills, you can apply new convolutional neural network (CNN) models and tuning parameters to improve your accuracy.

### RUN

Your deep learning talent is probably itching to design their own networks and solutions—which means you just need to define the problems you want to solve first. You can focus on data preparation, neural network design, and hyperparameter tuning to develop solutions that provide the highest accuracy and differentiate you from the competition.

## **LOOK** TO RELEVANT USE CASES FOR YOUR BUSINESS.

While use cases vary across industries, the most common ones fall into these categories and are usually associated with the listed neural network:

- Image classification or object detection: convolutional neural network (CNN)
- Time-series predictions: long short-term memory (LSTM)
- Natural language processing: recurrent neural network (RNN)
- Unlabeled data classification and data labeling: autoencoder (AE)
- Anomaly detection: autoencoder (AE)
- Recommender systems: multilayer perceptron (MLP)

Work with your deep learning talent or consultants to identify which use cases best match your organization and desired solutions. Then recreate a successful, already proven method.

> "You can leverage NVIDIA's large deep learning ecosystem to provide quick access to deep learning infrastructure, cloud service providers, learning resources, and external talent."

## **LEVERAGE** ANY DATA YOU HAVE AVAILABLE.

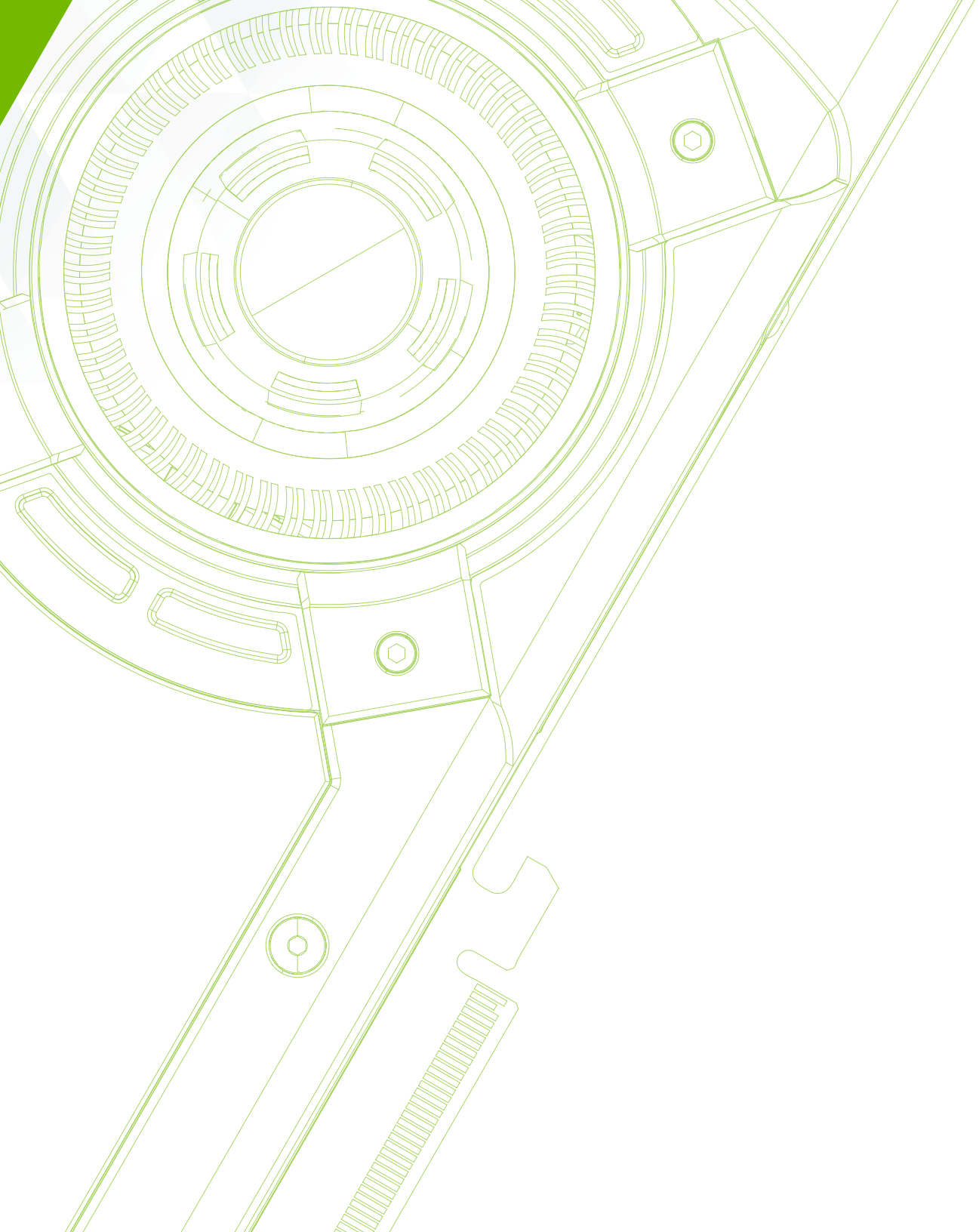
The more data you have, the better. New deep learning networks depend on it, and complex models need more data than simple ones.

- **Source external data:** If you don't have your own data, you can turn to external sources to increase your capabilities. Look at integrating public sources like social media, the news, or the weather. You can also leverage publicly trained models (e.g., look for Tensorflow models on GitHub or search for "model zoo" examples for other frameworks). Many sites also offer helpful datasets, including industry- and use case-specific datasets.

- **Equip your team with a strong analytics platform:** With large amounts of data, more complex networks, and the iterative nature of development, you'll want to ensure you have a powerful analytics platform to run your training on.

- **Invest in powerful hardware for faster data analysis:** The faster you can process data, the faster you can pull insights. Investing in powerful hardware optimized with the right software only makes things easier for your data science team and expedites your company's AI adoption.

The biggest barriers to building networks in-house are usually capability, time, and budget, and not every company has in-house talent, data, or platforms. Fortunately, you can leverage NVIDIA's large deep learning ecosystem to provide quick access to deep learning infrastructure, cloud service providers, learning resources, and external talent.



# THE RIGHT BUSINESS SOLUTION

How do you pick the right deep learning platform? From researchers and startups to enterprises, NVIDIA offers a portfolio of solutions optimized for specific use cases, making deep learning accessible and affordable for everyone.

## WORKSTATIONS

Get access to high performance computing (HPC) for deep learning from the convenience of your workspace. Deskside solutions are ideal for researchers, startups, and those engaged in productive experimentation.

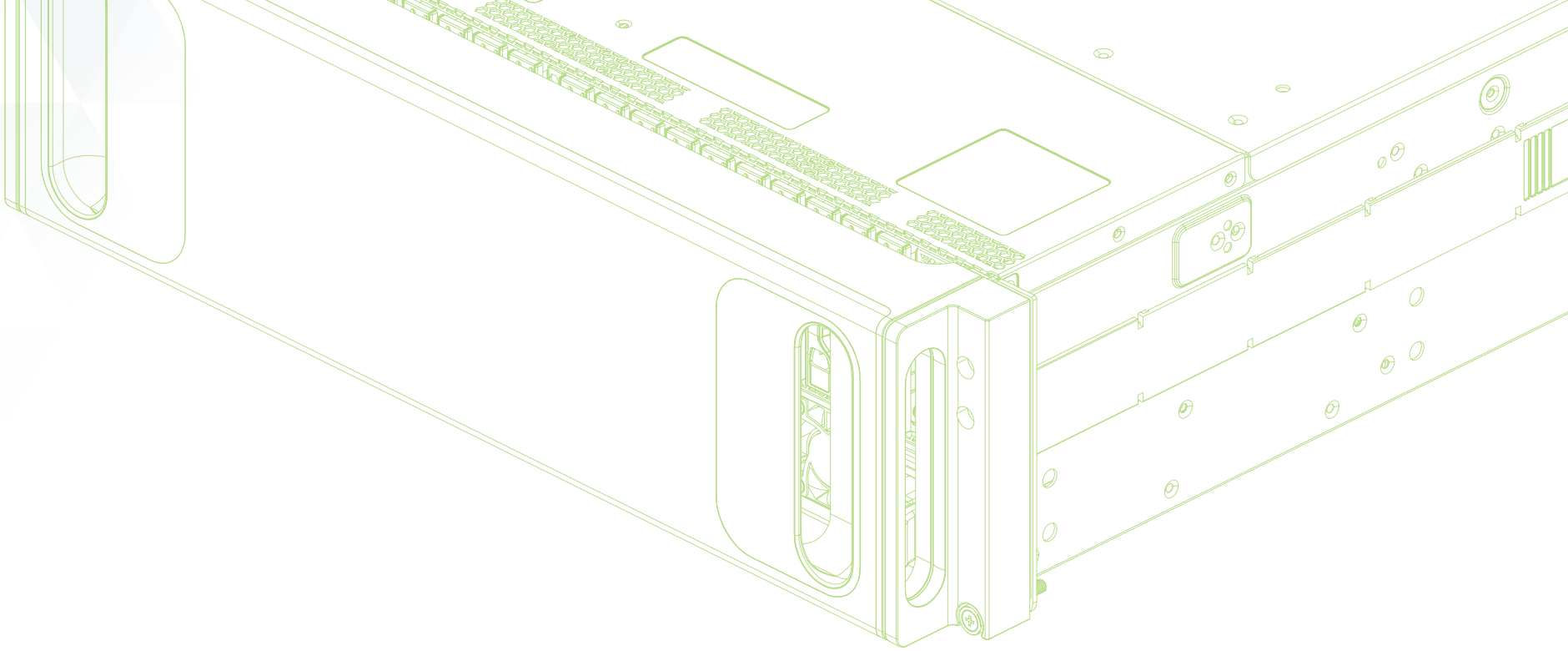
- **NVIDIA® DGX™ Station** is the world's first purpose-built AI workstation. Powered by four NVIDIA Tesla® V100 GPUs, it delivers 500 teraFLOPS (TFLOPS) of deep learning performance—the equivalent of hundreds of traditional servers—conveniently packaged in a workstation form factor built on NVIDIA NVLink™ technology and the NVIDIA GPU Cloud Deep Learning Software Stack. An integrated hardware and software solution, its full-stack optimization delivers maximized performance and is backed by NVIDIA's enterprise-grade support to keep you up and running.

- **Other Deskside Products:**

  > **NVIDIA TITAN V** is the most powerful graphics card created for the PC, driven by NVIDIA Volta™ and delivering new levels of desktop performance.

  > **NVIDIA Quadro® GV100** is powered by NVIDIA Volta and reinvents the workstation to meet the demands of next-generation AI.

Learn more >



## DATA CENTER

GPU-accelerated data centers deliver breakthrough performance with fewer servers, less floor space and power consumed, and greater efficiency, resulting in faster insights with dramatically lower costs.

- **NVIDIA DGX-1™** is the essential tool of AI research and development. Packing the power of over 800 CPUs, it delivers one petaFLOPS (PFLOPS) of AI performance in a single node. Built on eight NVIDIA Tesla V100 GPUs, configured in an NVLink topology, and architected for proven multi-GPU and multi-node scale, DGX-1 enables organizations to effortlessly scale their AI-powered applications and services.

- **NVIDIA DGX-2** is the world's largest GPU. The first 2-PFLOPS system with 16 fully interconnected GPUs and the revolutionary NVIDIA NVSwitch™ technology, DGX-2 enables unprecedented compute density and scale in the data center. By exploiting greater model and data parallelism, DGX-2 can solve the biggest AI challenges that demand the largest datasets and the most complex deep neural networks.

- **NVIDIA HGX-2** fuses high performance computing and AI computing into the world's largest GPU platform. Integrated by NVIDIA OEM and ODM partners into powerful servers optimized for AI training and inference, HGX-2 is powered by 16 Tesla V100 GPUs and NVSwitch to provide 2 PFLOPS of compute and 0.5 terabyte (TB) of unified memory. It replaces 300 CPUs and saves significant cost and space in the data center.

- **GPU-Accelerated Server Platforms** represents diverse classes of servers that deliver optimal performance for a broad array of accelerated workloads, from AI training and inference to supercomputing and virtual desktop infrastructure (VDI) applications. This classification recommends the optimal mix of GPUs, CPUs, and interconnects for diverse training (HGX-T), inference (HGX-I), and supercomputing (SCX) applications.

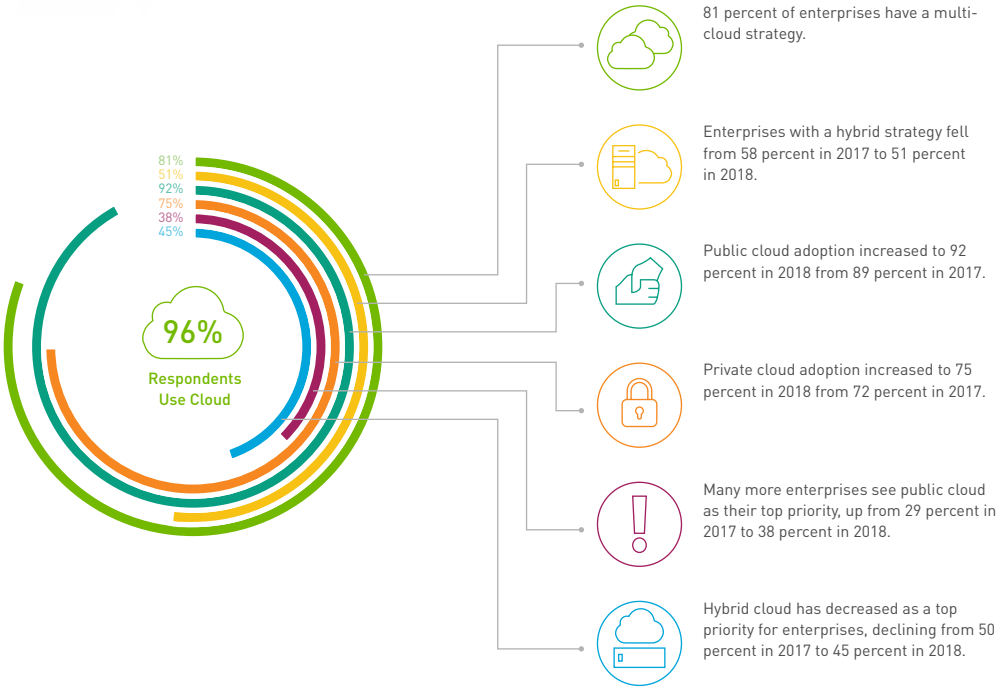
Learn more >

Cloud Computing Trends

## 2018 STATE OF THE CLOUD SURVEY

In January 2018, RightScale conducted its seventh annual State of the Cloud Survey of the latest cloud computing trends, with a focus on infrastructure as a service and platform as a service.

The survey asked 997 IT professionals about their adoption of cloud infrastructure and related technologies. Fifty-three percent of the respondents represented enterprises with more than 1,000 employees. The margin of error is 3.08 percent.




## CLOUD

Cloud computing has democratized the data center and transformed the way businesses operate. Thanks to cloud service providers (CSPs), GPU cloud computing enables businesses to start testing and development quickly without huge investment in IT and technology. Summarized below are three common CSP services:

- **Deep Learning Application Programming Interfaces (DL APIs)** can be leveraged by developers with introductory deep learning capabilities, as the specific functions exposed by these deep learning APIs can be integrated into applications. Examples are image recognition or translation APIs across services such as AWS, Microsoft Cognitive Services, Google Cloud, and IBM Watson.

- **Deep Learning Platforms as a Service (DL PaaS)** are middleware-integrated stacks that expose essential deep learning and data science techniques. They offer more control than APIs, providing users with knowledge of simple and effective tools to build tailored solutions. Examples are AWS SageMaker, Microsoft Azure Machine Learning Studio, IBM Watson, Google Cloud ML Engine, and AutoML.

- **Deep Learning Infrastructures as a Service (DL IaaS)** are designed to deliver end-to-end customizability for the underlying compute platforms and are best suited for users with strong deep learning capabilities. They provide virtualized infrastructure resources that are performance-optimized for deep neural network (DNN) workloads. Examples are AWS EC2 P3, Microsoft NC series, Google Cloud GPUs, Alibaba gn5 and gn4 instances, Baidu GPU offerings, and IBM Cloud bare-metal and virtual GPU offerings.

Find a CSP partner >

# DEEP LEARNING EVERYWHERE, FOR EVERYONE

Whichever solution you choose, you can get all the deep learning software you need from NVIDIA GPU Cloud (NGC)—for free. NGC provides simple access to a comprehensive catalog of GPU-optimized software tools for deep learning and HPC that take full advantage of NVIDIA GPUs on the desktop, in the data center, and in the cloud.

- **Discover GPU-Accelerated Containers.** Choose from a comprehensive catalog of GPU-accelerated containers, including NVIDIA-optimized deep learning software, third-party-managed HPC applications, NVIDIA HPC visualization tools, and partner applications.
- **Innovate in Minutes, Not Weeks.** Access ready-to-run, GPU-optimized containers without the complexity normally associated with software setup.
- **Stay Up to Date.** NGC deep learning containers are optimized and updated monthly to deliver maximum performance on NVIDIA GPUs. Other NGC containers provide easy access to GPU-accelerated software releases.

Explore NVIDIA GPU Cloud >

## RESOURCES

> Explore industry-specific use cases.

> Get access to onsite or online deep learning workshops through the NVIDIA Deep Learning Institute.

> Learn more about Tesla V100, the core of AI.