What Is AI Computing?

AI computing is the work of machine learning systems and software, sifting through mountains of data to reveal useful insights and generate new capabilities.

Author: Rick Merritt

The abacus, sextant, slide rule and computer. Mathematical instruments mark the history of human progress.

They’ve enabled trade and helped navigate oceans, and advanced understanding and quality of life.

The latest tool propelling science and industry is AI computing.

AI computing is the math-intensive process of calculating machine learning algorithms, typically using accelerated systems and software. It can extract fresh insights from massive datasets, learning new skills along the way.

It’s the most transformational technology of our time because we live in a data-centric era, and AI computing can find patterns no human could.

For example, American Express uses AI computing to detect fraud in billions of annual credit card transactions. Doctors use it to find tumors , finding tiny anomalies in mountains of medical images.

Before getting into the many use cases for AI computing, let’s explore how it works.

First, users, often data scientists, curate and prepare datasets, a stage called extract/transform/load, or ETL. This work can now be accelerated on NVIDIA GPUs with Apache Spark 3.0 , one of the most popular open source engines for mining big data.

Second, data scientists choose or design AI models that best suit their applications.

Some companies design and train their own models from the ground up because they are pioneering a new field or seeking a competitive advantage. This process requires some expertise and potentially an AI supercomputer, capabilities NVIDIA offers.

Many companies choose pretrained AI models they can customize as needed for their applications. NVIDIA provides dozens of pretrained models and tools for customizing them on NGC , a portal for software, services, and support.

Third, companies sift their data through their models. This key step, called inference , is where AI delivers actionable insights.

The three-step process involves hard work, but there’s help available, so everyone can use AI computing.

For example, NVIDIA TAO Toolkit can collapse the three steps into one using transfer learning , a way of tailoring an existing AI model for a new application without needing a large dataset. In addition, NVIDIA LaunchPad gives users hands-on training in deploying models for a wide variety of use cases.

AI models are called neural networks because they’re inspired by the web-like connections in the human brain.

If you slice into one of these AI models, it might look like a mathematical lasagna, made up of layers of linear algebra equations. One of the most popular forms of AI is called deep learning because it uses many layers.

If you zoom in, you’d see each layer is made up of stacks of equations. Each represents the likelihood that one piece of data is related to another.

AI computing multiplies together every stack of equations in every layer to find patterns. It's a huge job that requires highly parallel processors sharing massive amounts of data on fast computer networks.

GPUs are the de facto engines of AI computing.

NVIDIA debuted the first GPU in 1999 to render 3D images for video games, a job that required massively parallel calculations.

GPU computing soon spread to use in graphics servers for blockbuster movies. Scientists and researchers packed GPUs into the world's largest supercomputers to study everything from the chemistry of tiny molecules to the astrophysics of distant galaxies.

When AI computing emerged more than a decade ago, researchers were quick to embrace NVIDIA's programmable platform for parallel processing. The video below celebrates this brief history of the GPU.

The idea of artificial intelligence goes back at least as far as Alan Turing, the British mathematician who helped crack coded messages during WWII.

"What we want is a machine that can learn from experience," Turing said in a 1947 lecture in London.

Acknowledging his insights, NVIDIA named one of its computing architectures for him.

Turing's vision became a reality in 2012 when researchers developed AI models that could recognize images faster and more accurately than humans could. Results from the ImageNet competition also greatly accelerated progress in computer vision .

Today, companies such as Landing AI, founded by machine learning luminary Andrew Ng, are applying AI and computer vision to make manufacturing more efficient . And AI is bringing human-like vision to sports , smart cities and more.

AI computing made huge inroads in natural language processing after the invention of the transformer model in 2017. It debuted a machine-learning technique called "attention" that can capture context in sequential data like text and speech.

Today, conversational AI is widespread. It parses sentences users type into search boxes. It reads text messages when you're driving, and lets you dictate responses.

These large language models are also finding applications in drug discovery, translation , chatbots , software development, call center automation and more .

Users in many, often unexpected, areas are feeling the power of AI computing.

The latest video games achieve new levels of realism thanks to real-time ray tracing and NVIDIA DLSS , which uses AI to deliver ultra-smooth game play on the GeForce RTX platform.

That's just the start. The emerging field of neural graphics will speed the creation of virtual worlds to populate the metaverse , the 3D evolution of the internet.

To kickstart that work, NVIDIA released several neural graphics tools in August.

Car makers are embracing AI computing to deliver a smoother, safer driving experience and deliver smart infotainment capabilities for passengers.

Mercedes-Benz is working with NVIDIA to develop software-defined vehicles. Its upcoming fleets will deliver intelligent and automated driving capabilities powered by an NVIDIA DRIVE Orin centralized computer. The systems will be tested and validated in the data center using DRIVE Sim software, built on NVIDIA Omniverse, to ensure they can safely handle all types of scenarios.

At CES, the automaker announced it will also use Omniverse to design and plan manufacturing and assembly facilities at its sites worldwide.

BMW Group is also among many companies creating AI-enabled digital twins of factories in NVIDIA Omniverse , making plants more efficient. It’s an approach also adopted by consumer giants such as PepsiCo for its logistic centers as shown in the video below.

Inside factories and warehouses, autonomous robots further enhance efficiency in manufacturing and logistics. Many are powered by the NVIDIA Jetson edge AI platform and trained with AI in simulations and digital twins using NVIDIA Isaac Sim .

In 2022, even tractors and lawn mowers became autonomous with AI.

In December, Monarch Tractor , a startup based in Livermore, Calif., released an AI-powered electric vehicle to bring automation to agriculture. In May, Scythe , based in Boulder, Colo., debuted its M.52 (below), an autonomous electric lawn mower packing eight cameras and more than a dozen sensors.

The number and variety of use cases for AI computing are staggering.

Cybersecurity software detects phishing and other network threats faster with AI-based techniques like digital fingerprinting .

In healthcare, researchers broke a record in January 2022 sequencing a whole genome in well under eight hours thanks to AI computing. Their work (described in the video below) could lead to cures for rare genetic diseases.

AI computing is at work in banks, retail shops and post offices. It’s used in telecom, transport and energy networks, too.

For example, the video below shows how Siemens Gamesa is using AI models to simulate wind farms and boost energy production.

As today’s AI computing techniques find new applications, researchers are inventing newer and more powerful methods.

Another powerful class of neural networks, diffusion models, became popular in 2022 because they could turn text descriptions into fascinating images. Researchers expect these models will be applied to many uses, further expanding the horizon for AI computing.

Original URL: https://blogs.nvidia.com/blog/2023/01/20/what-is-ai-computing/