HighFens Inc

From Data to Insights: Unleashing AI's Power Across Industries

HPE Discover 2024
June 19th, 2024, 1 PM PST
Session: CF6874

- In today's dynamic business landscape, AI has emerged as a game-changer. With the advent of pre-trained Large Language Models (LLM) and Generative AI, organizations across industries are harnessing this transformative technology. But which sectors can truly leverage its power? And how does the technology stack for AI differ from the traditional one?

- Join us as we delve into these critical questions, exploring real-world applications, industry use cases, and the essential components that fuel AI's capabilities. Discover how AI is reshaping the future of business.

Frederic Van Haren
CTO

Name:	HighFens Inc.	https://highfens.com
Founded:	April 2016	
Description:	- Consulting & Services - AI & Gen AI Solutions	
Customers:	- NDA for Vendors (HW & SW) - Enterprises (End-users)	

HighFens Inc. was founded to guide and assist customers in implementing innovative and disruptive technology for the digital world.

The team at HighFens has been working with Big Data, HPC, and AI for over 15 years. We have a track record of introducing innovative solutions to solve complex problems at scale.

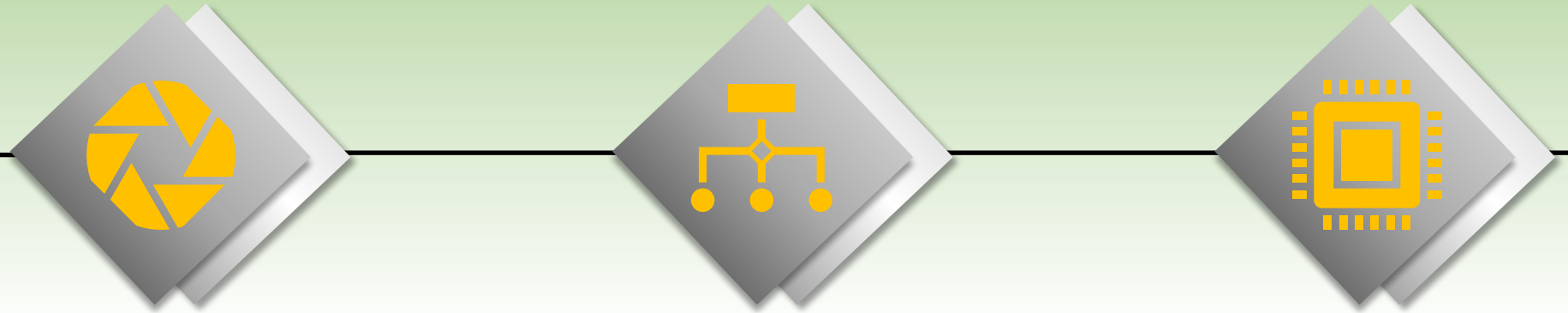
Turn our experience and knowledge to your advantage and achieve your business goals. AI solutions that fit your need based on a scalable & reliable eco-system.

We will cover the fundamentals of AI and explain commonly used terminology. What are the challenges associated with Training and Inference?

Artificial Intelligence

What are the factors that decide what the infrastructure stack should look like? What options do I have? How does software fit into this?

Infrastructure stacks

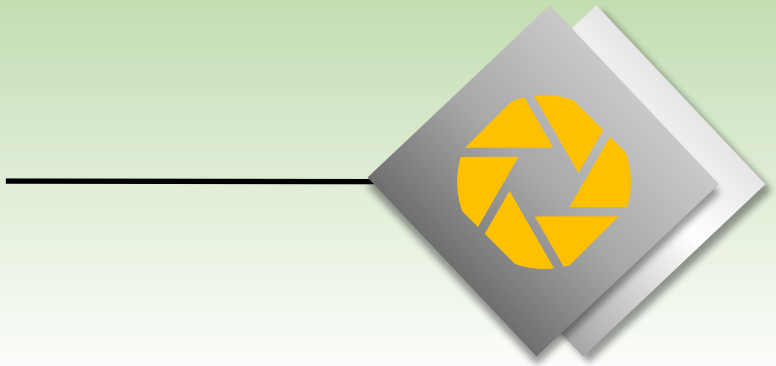


Generative AI

What is Generative AI, and how does it differ from traditional AI? We will explain the components of a Generative AI solution and discuss LLMs and the popularity of RAG.

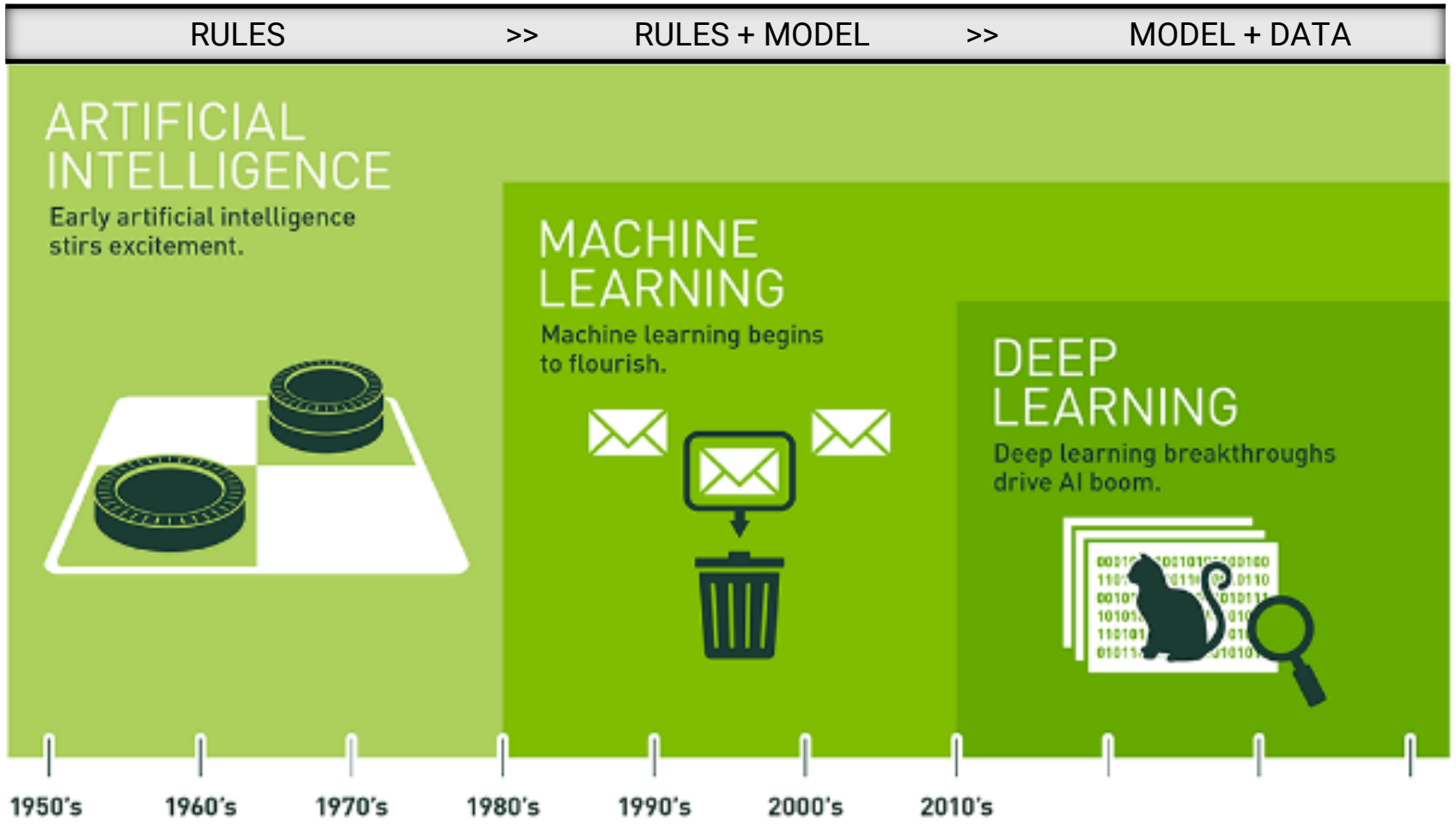
We will cover the fundamentals of AI and explain commonly used terminology. What are the challenges associated with Training and Inference?

Artificial Intelligence



The Evolution of Artificial Intelligence

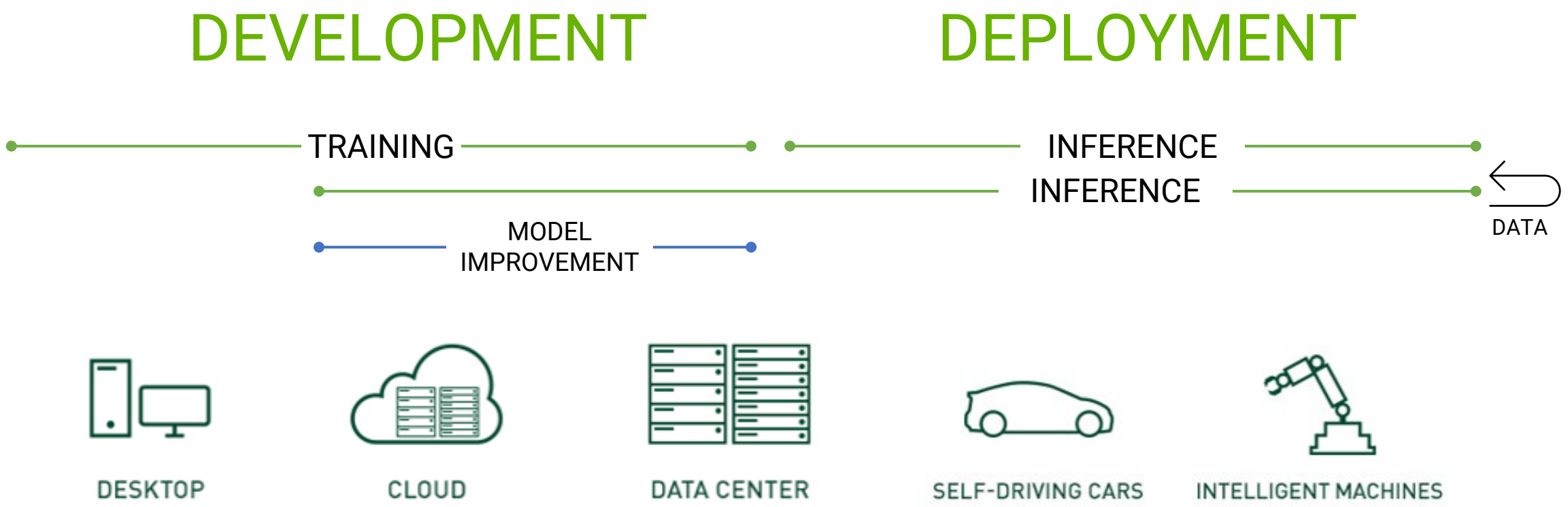
Ability to learn without being explicitly programmed



Source: NVIDIA

End-to-End AI Solutions @ Scale

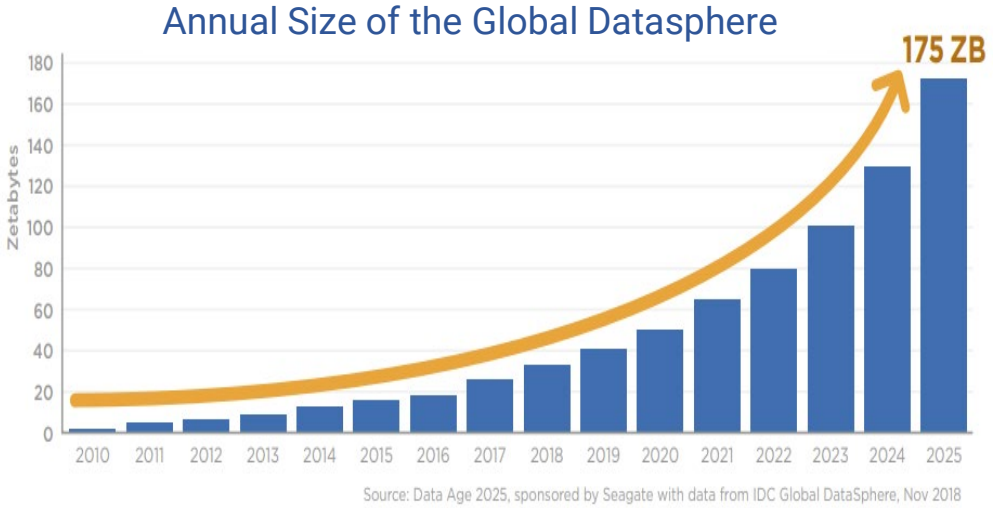
The path to unlock the full potential of AI



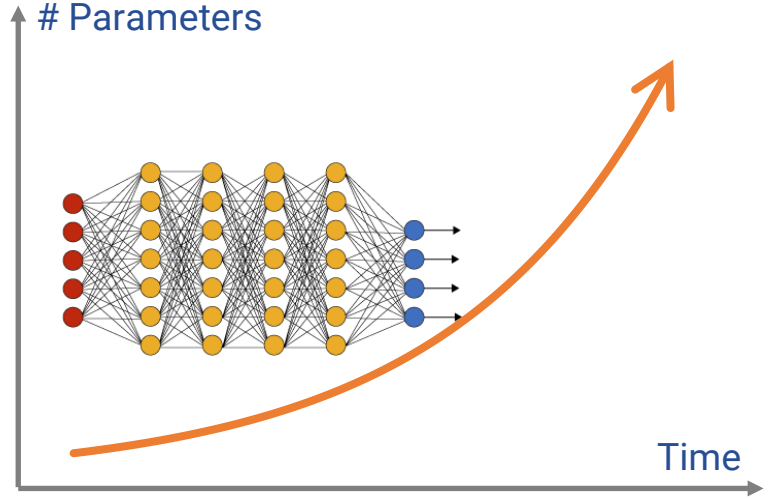
TRAINING - Challenges

Staying relevant is a significant challenge for Enterprises

Data Growth: x2 every 3 years



Complexity: x10 every year



PRODUCTION - Challenges

Half of the AI projects don't make it to Production

54% of AI projects never make it to production[1]

Scaling AI is a significant challenge

24% "Increasing data complexity & data silos"[2]

25% "Lack of infrastructure, tools & AI platforms"[2]

34% "Limited AI skills, expertise or knowledge "[2]

An integrated solution for streamlined development and deployment

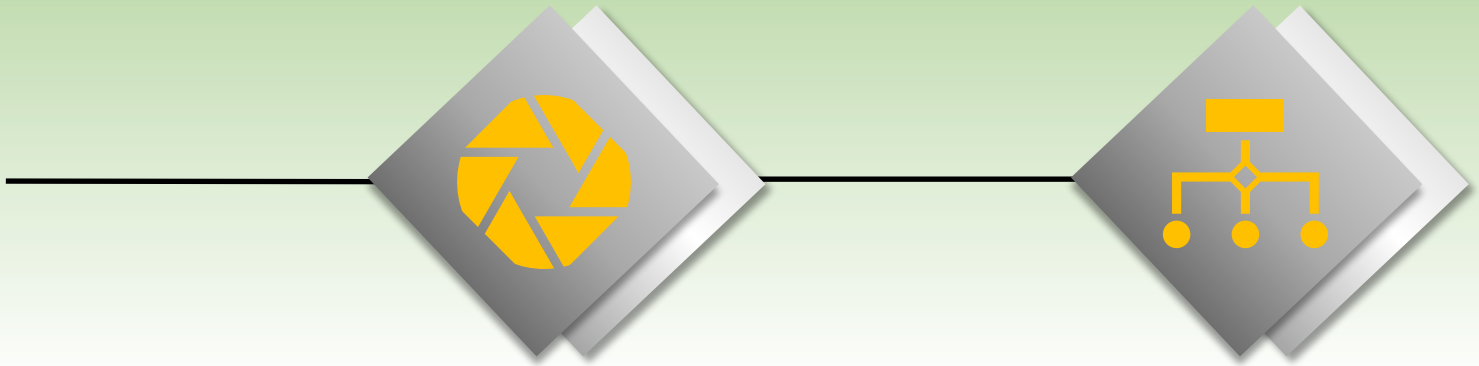
Scalable & Reliable ECO-system

- AI Tools & Frameworks
- Data Management
- Accelerated Infrastructure

Source: 1. Gartner "Executives Think Automation Can Be Applied to Any Business Decision", August 2022
2. IBM "Global AI Adoption 2022", May 2022

What is Generative AI, and how does it differ from traditional AI? We will explain the components of a Generative AI solution and discuss LLMs and the popularity of RAG.

Generative AI

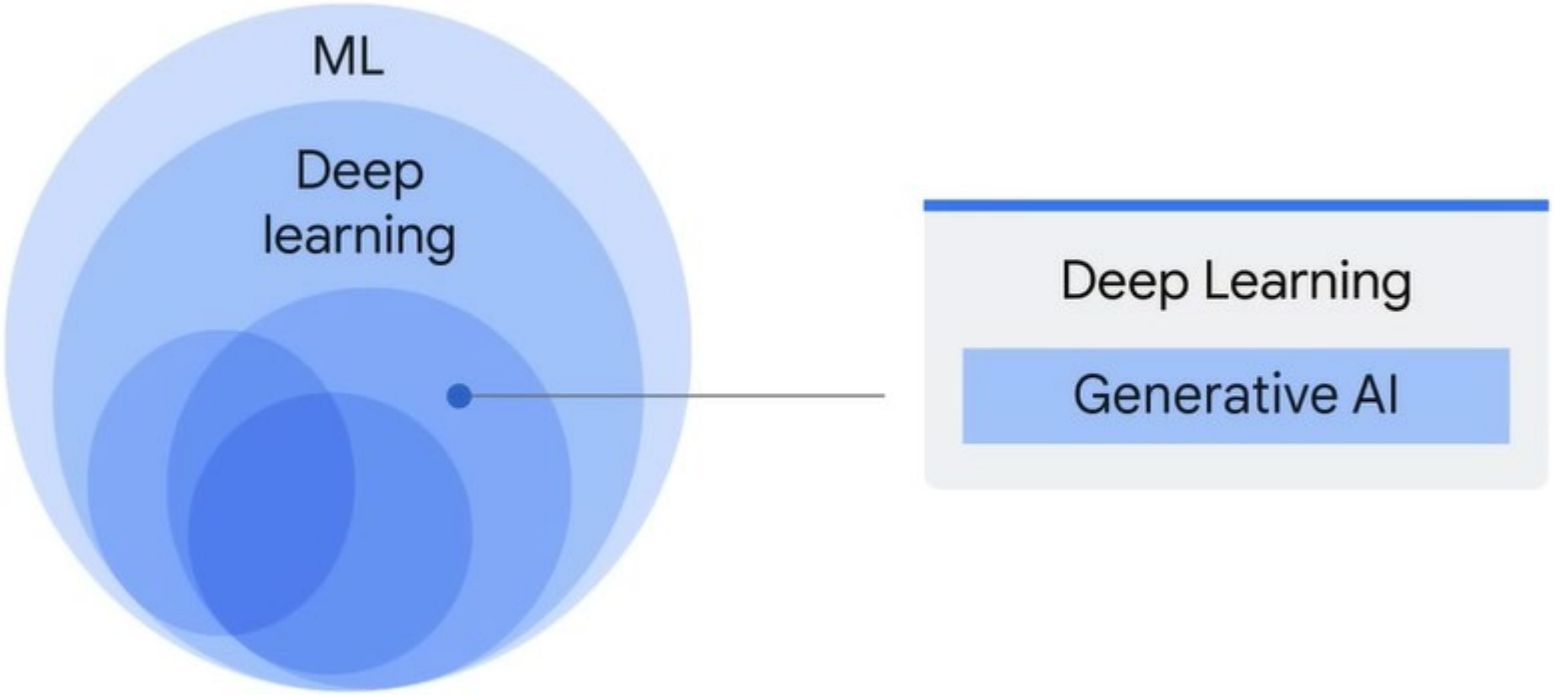


Generative AI (Gen AI)

Use existing content (text, image, video, audio) to create new content

Generative AI and large language models (LLMs) are a subset of Deep Learning

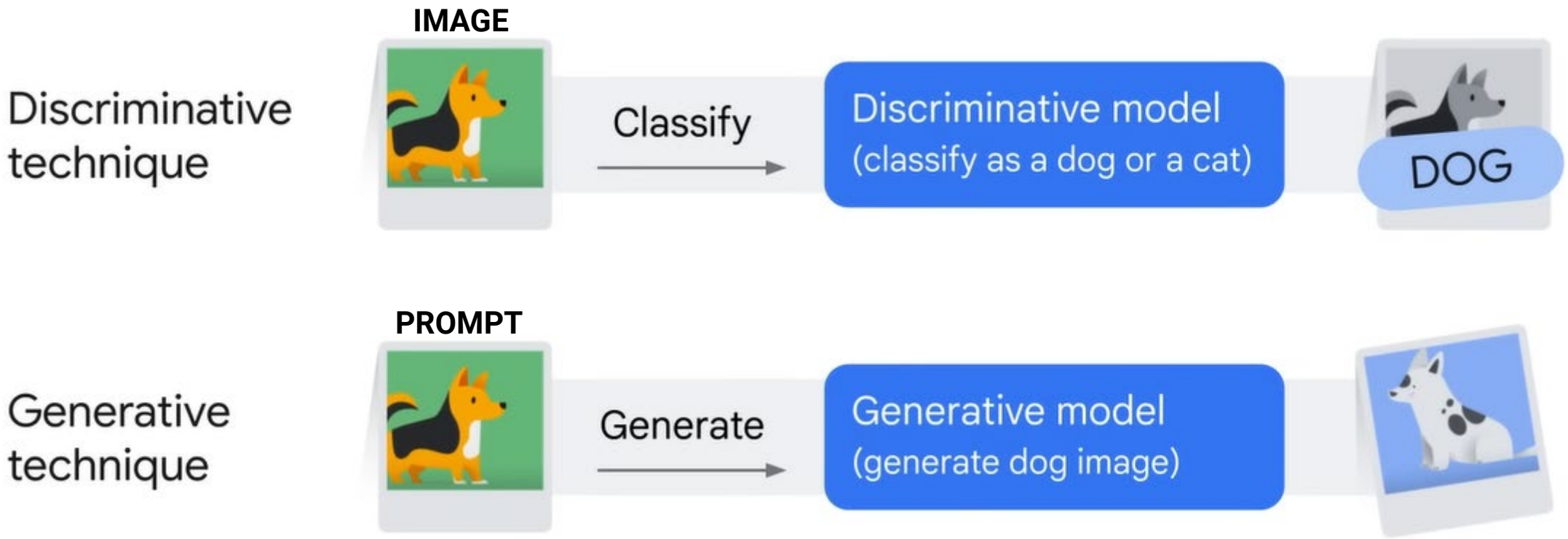
Pre-trained LLMs are available from OpenAI, Google, Meta,…



Source: Google

How is Generative AI different?

Discriminative vs. **Generative**



Source: Google

OpenAI GPT models over time (complexity)

Building useful base/foundation models requires significant investments (and time)

GPT Version	1 (2018)	2 (2019)	3.5 (2022)	4 (2023)
# of parameters	117 Million	1.5 Billion	175 Billion	1.76 Trillion
Compared to v1	x1	x13	X1.5k	X15k

GPT 4 (model training)

25,000 - A100 GPUs

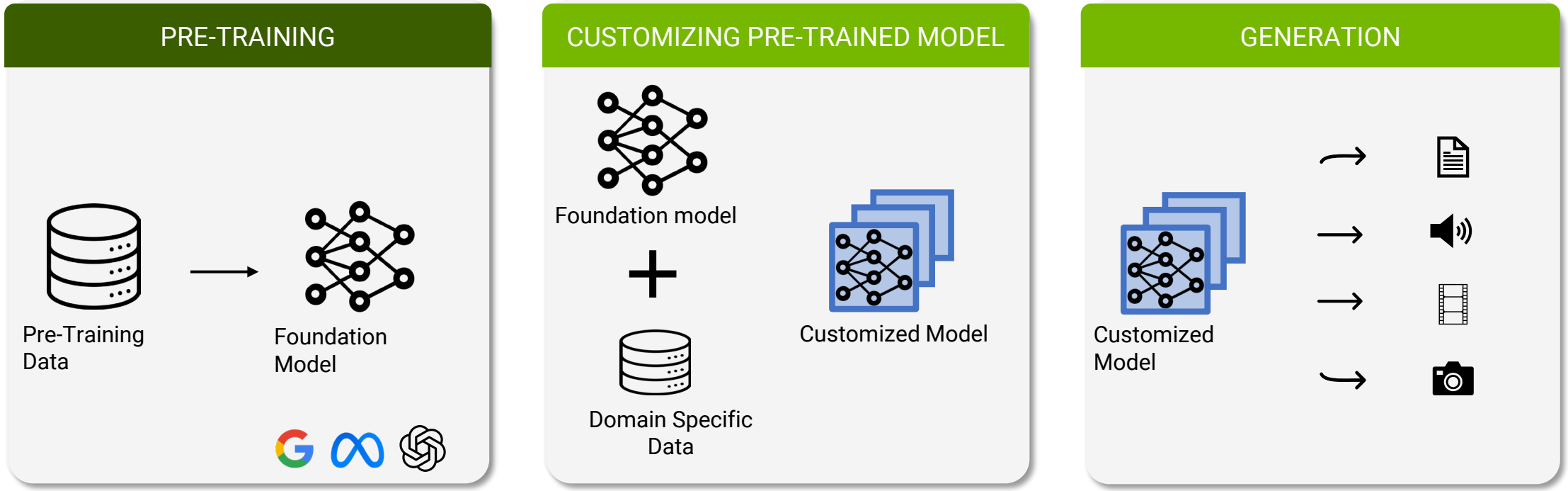
90 days

1 - A100 GPU

6,164 years

Generative AI (simplified)

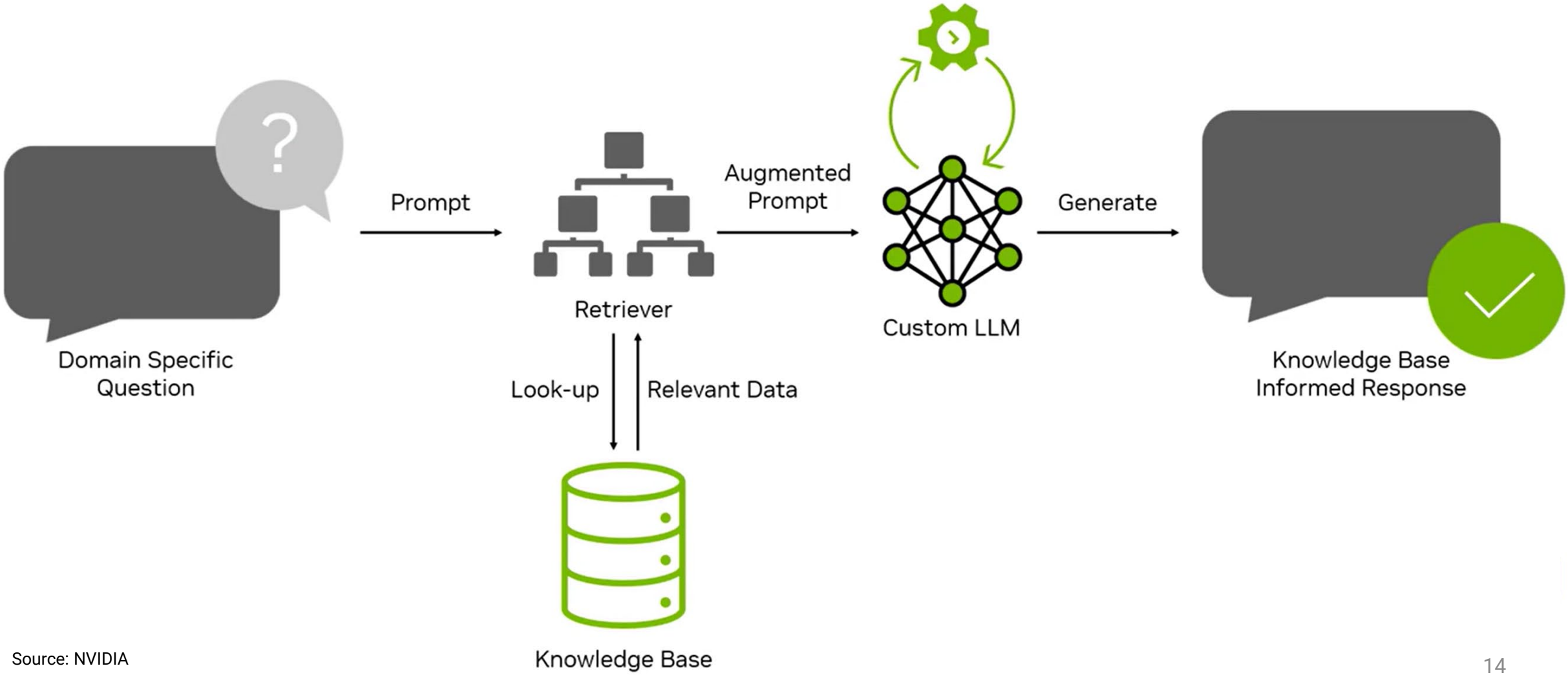
Pre-trained models are the foundation of Generative AI solutions



Source: NVIDIA

Retrieval-Augmented-Generation (RAG)

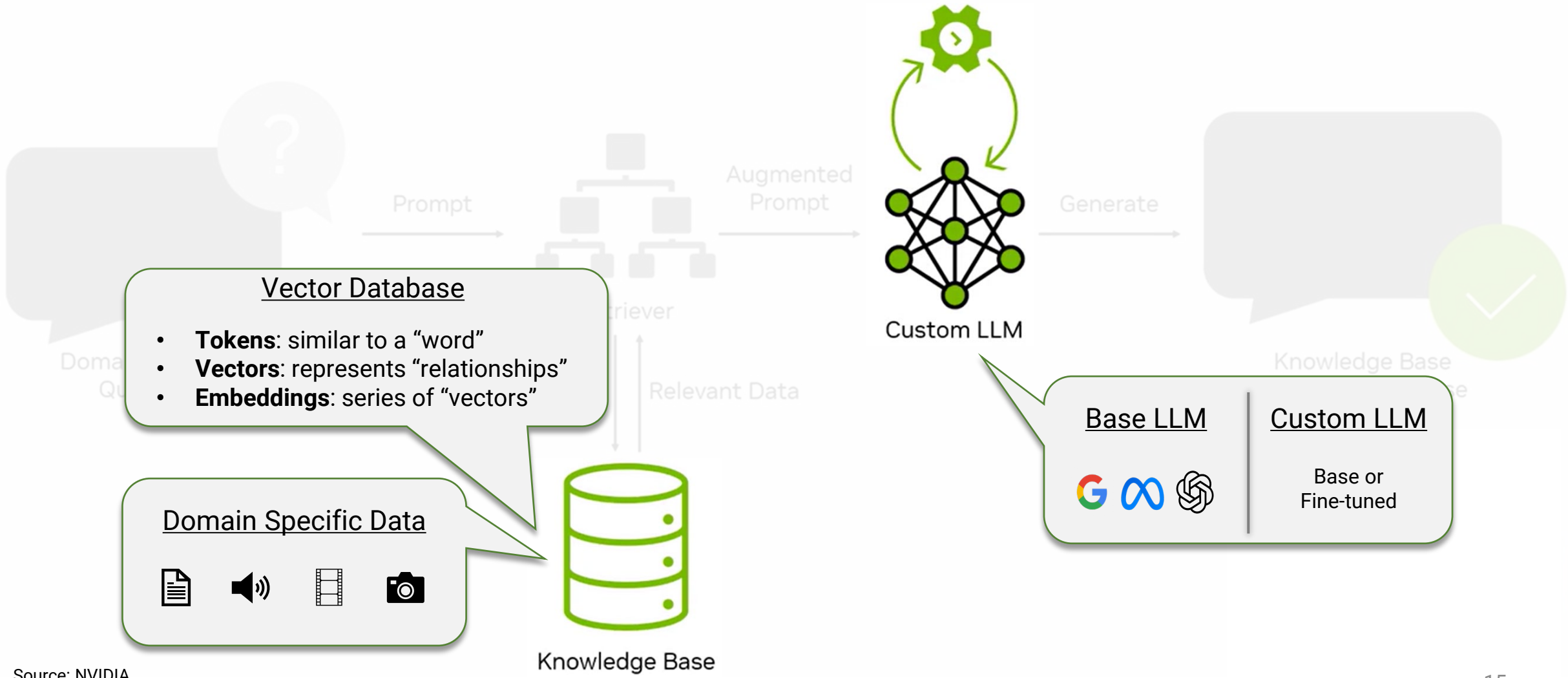
Enhance an LLM with a specific domain or knowledge base WITHOUT retraining



Source: NVIDIA

Retrieval-Augmented-Generation (RAG)

Enhance an LLM with a specific domain or knowledge base WITHOUT retraining



Source: NVIDIA

Conversational AI for Lego sets

Complete Lego catalog (latest), where to buy, cost, availability

Source: Lego

Conversational AI for Lego sets

Complete Lego catalog (latest), where to buy, cost, availability

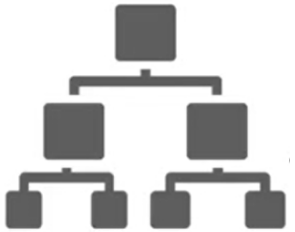
1 PROMPT

Where can I buy them and are they in stock?

5

2 RETRIEVER

Tokenizer + Vectorizer

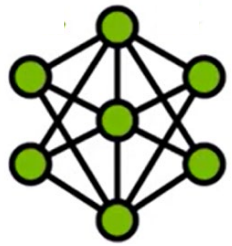


Look-up

Relevant Data

3 CONVERSATIONAL LLM

Augmented Prompt



Generate Answer

4 ANSWER

Amazon...in stock...
Princess Leia...$195...
Darth Vader...$125...
Mandalorian...$175
Place an order?

6

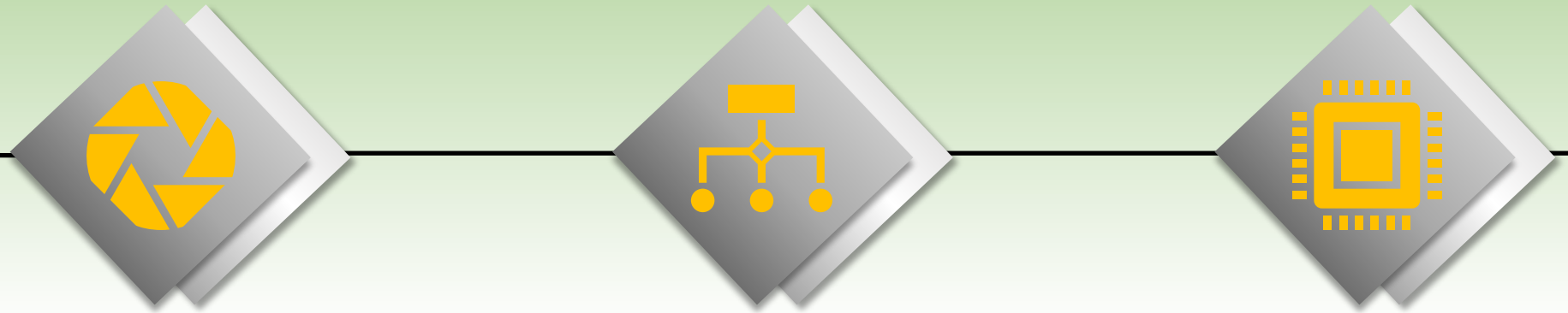
Tokenizer + Vectorizer

Price Availability

amazon

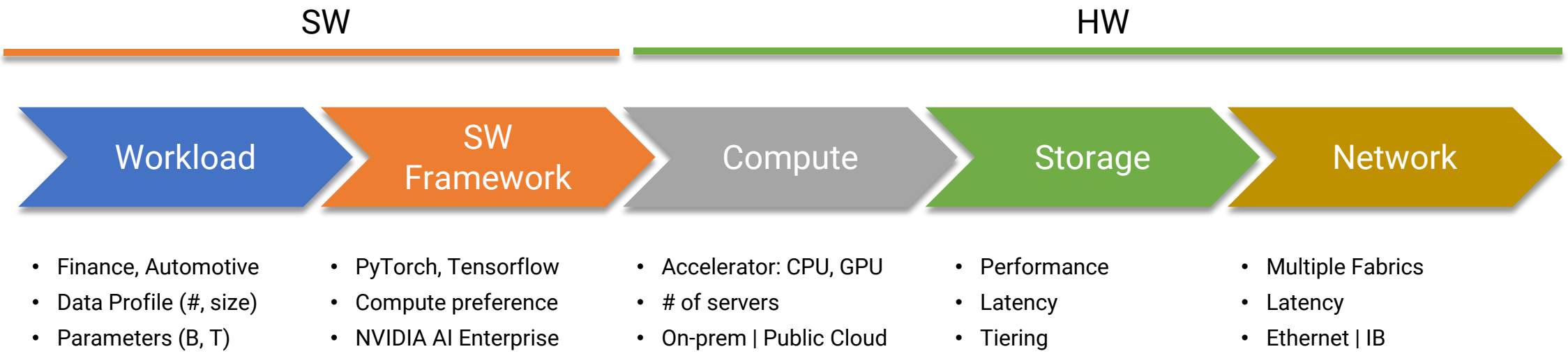
What are the factors that decide what the infrastructure stack should look like? What options do I have? How does software fit into this?

Infrastructure stacks



Making AI Infrastructure decisions

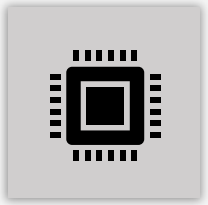
The workloads (SW) drive the infrastructure (HW) decisions



Traditional Compute Stacks

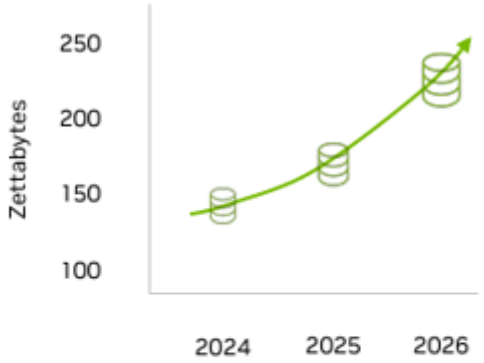
CPU-based solutions can’t keep pace with the demands of AI challenges

CPU (HPC)



- Several Cores
- Low Latency
- Serial processing

Challenges (CPU)



Data Growth

221 ZB of data by 2026

Scaling Compute

Moore’s law has slowed

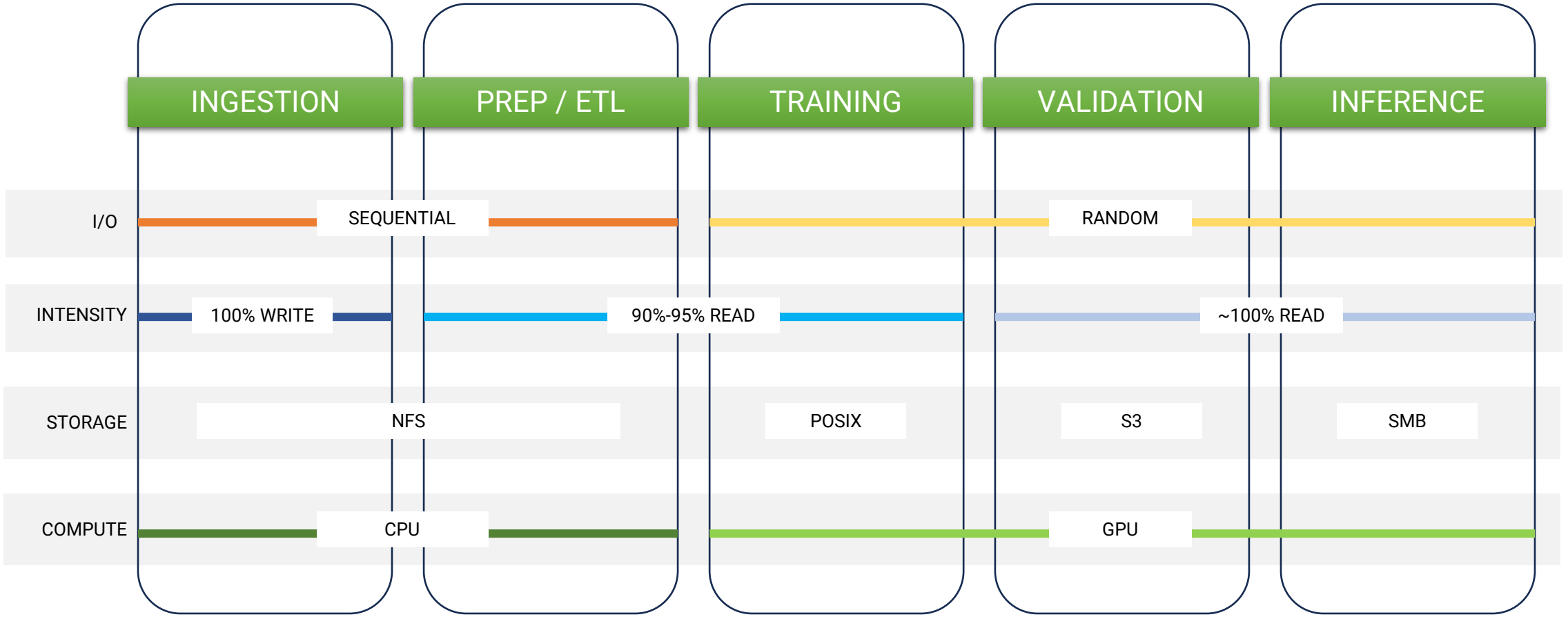
GPU (AI)-Acceleration

- Many Cores
- High throughput
- Parallel processing

Sources: IDC, Global DataSphere Forecast, 2022-2026: Enterprise Organizations Driving Most of the Growth, May 2022
https://www.energy.gov/eere/buildings/data-centers-and-servers

Data Pipelines–Storage & Compute Challenges

Pipelines have varying requirements – one size does not fit all (sample)



Software Stack – ML Platform

An ever-growing number of options with frequent updates

Data

Processing

RAY, HIVE, dask, trino, RAPIDS, Apache Spark

Pipelines

Versioning, Labelling

DVC, LFS, Pachyderm, scale, DELTA LAKE, Labelbox, annotell

Data Sources

LOG, CSV, aws, Azure, SQL, snowflake, CLOUDERA, databricks, lustre, MAPR, S3

Training

Collaboration

GitHub, JFrog

Experiments

Determined AI, mlflow

Scheduling

kubernetes, slurm

Environment

Evaluation

TensorBoard, gradio

Compute

intel, NVIDIA, AMD, habana An Intel Company, cerebras, Qualcomm

Inference

Monitoring

Prometheus, Grafana

Optimization

tvm, deci

Testing

great_expectations, Buildkite

Observability

ALIBI EXPLAIN, truera, modzy, fiddler

Bias, Robustness

ALIBI DETECT, TROJ.AI

Serving, Rollout

TensorFlow Extended, ONNX, TORCHSERVE, BENTOML, SELDON, KServe, TensorRT

Source: HPE

How do you get started?

In short, let us help you achieve your AI goals with our experience

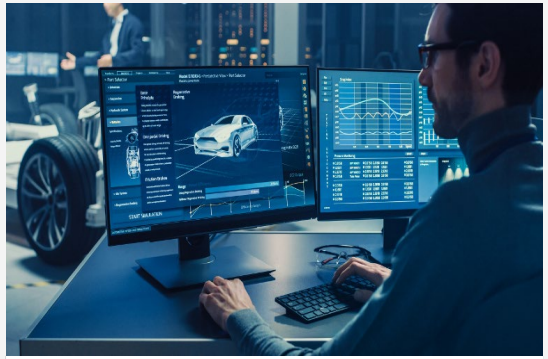
Consulting & Services

We provide organizations with scalable, reliable, cost-effective solutions for old and new infrastructure needs, including scoping, planning, system design, and management support.

With decades of experience in the IT and, more specifically, HPC / Big Data / AI world, we deliver various services to your business, be it physical, virtual, in a Cloud, or hybrid solutions.

+

AI Solutions



We deliver complete AI Solutions to customers who want a comprehensive AI solution and take full advantage of today's innovative technology. In addition, the solutions offer a cost-effective and scalable approach for customers who wish to accelerate their AI journey.

AI Solutions @ Scale

www https://highfens.com

@HighFens

@ info@highfens.com

in https://www.linkedin.com/company/highfens