Better Together In the World of Generative AI

Trends, Concepts and Use Cases

PK Gupta
Global CTO & APJ Presales lead
Global Alliances Presales

DELL Technologies

Star Trek Theme

Space: the final frontier. These are the voyages of the starship Enterprise*. Its five-year mission:* ***to explore strange new worlds****. To seek out new life and new civilizations.* ***To boldly go where no man has gone before****!*

IT = INDUSTRY IN TRANSFORMATION

C O M B S

D A N C E

COMBS – Cloud, Open Source, Mobile, Big Data, Social Networks
DANCE – Data, AI/Algorithm, Networks, Cloud, Extreme Compute

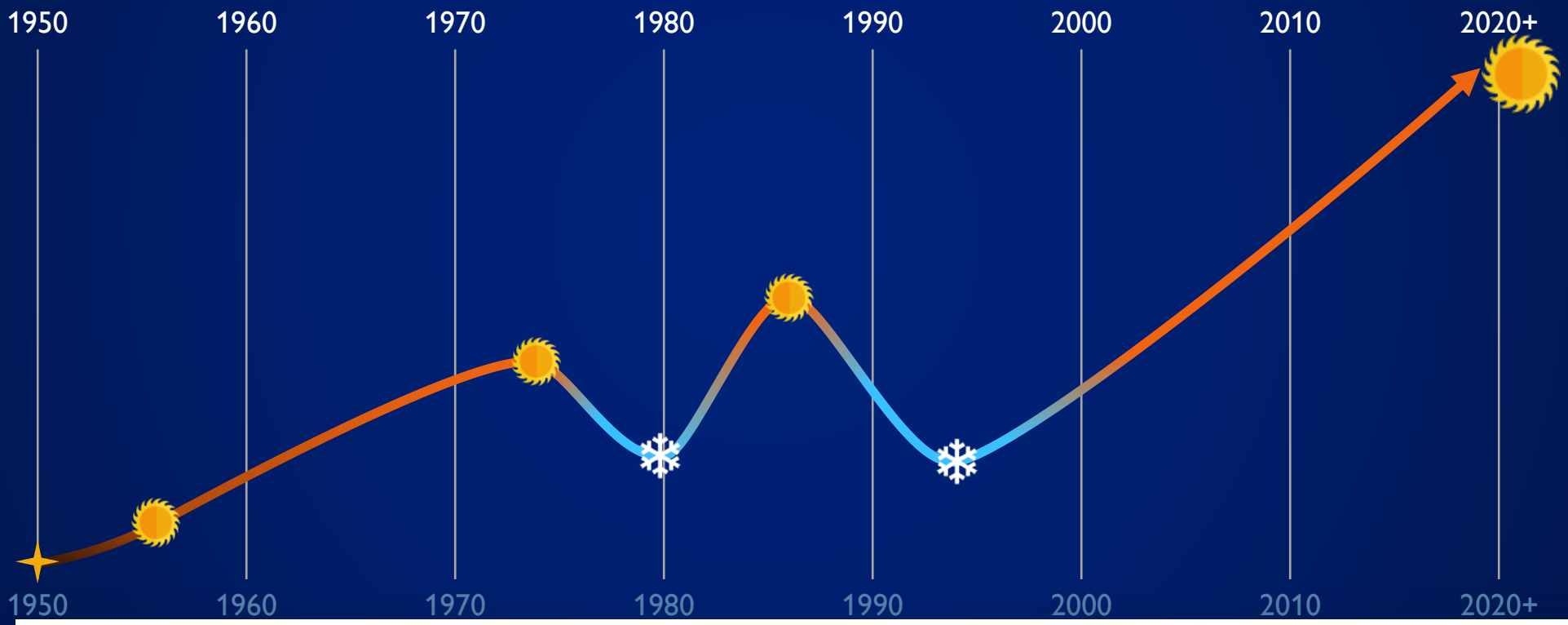
"AI is the defining technology of our industry & our sector for the rest of the decade.

For us, this is all about data… which is really the fuel for AI.

— Michael Dell, Chairman & CEO
Sep 2023

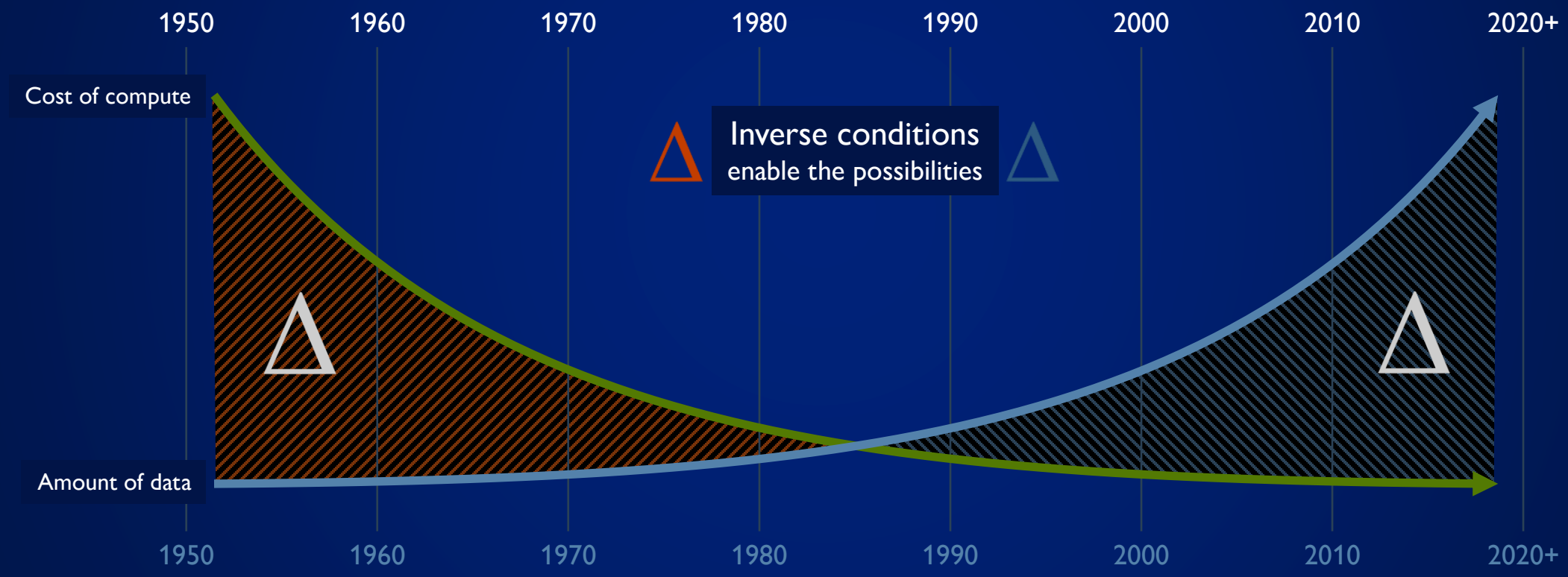
Brief history of Artificial Intelligence

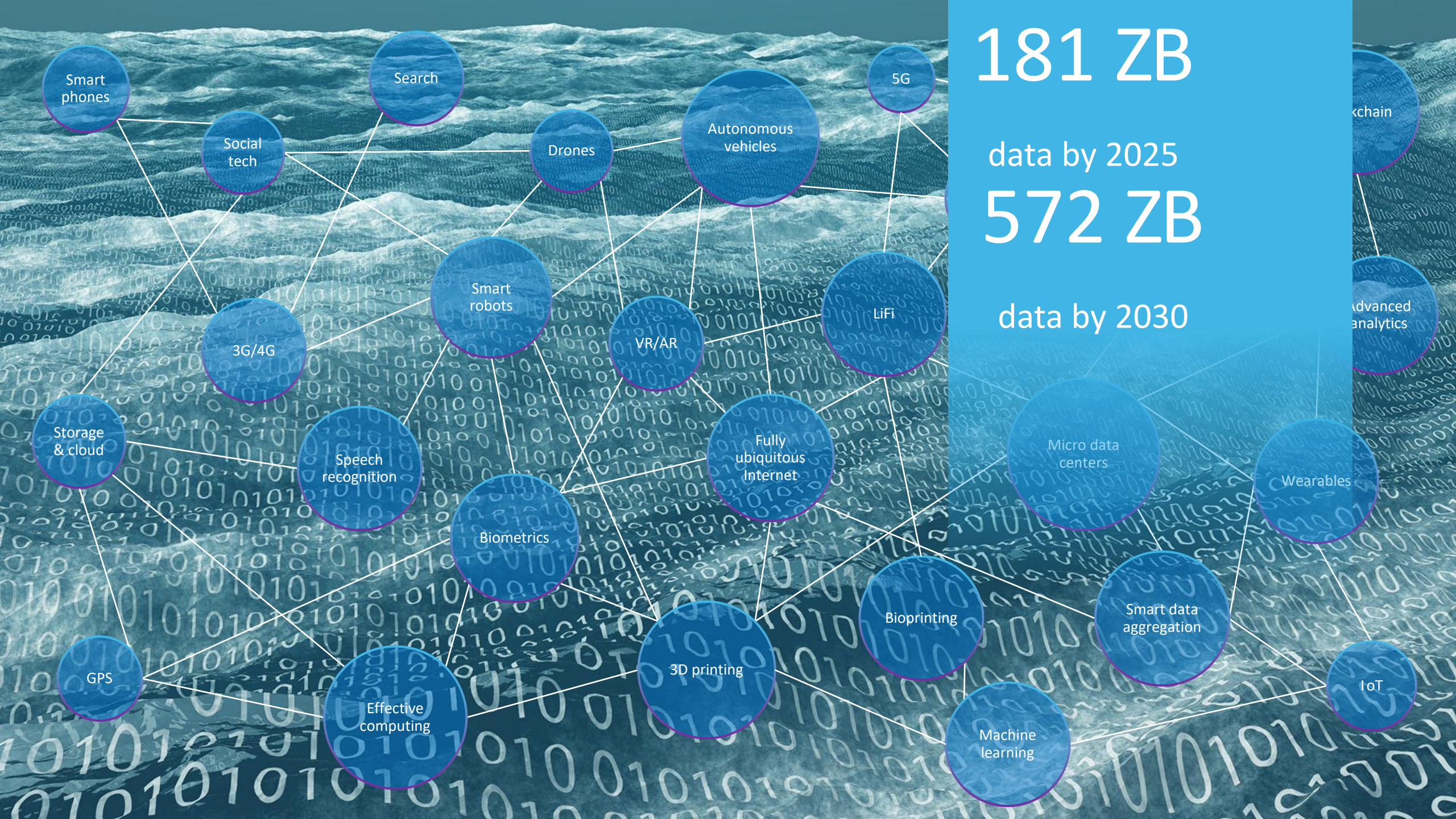
What makes this time any different?



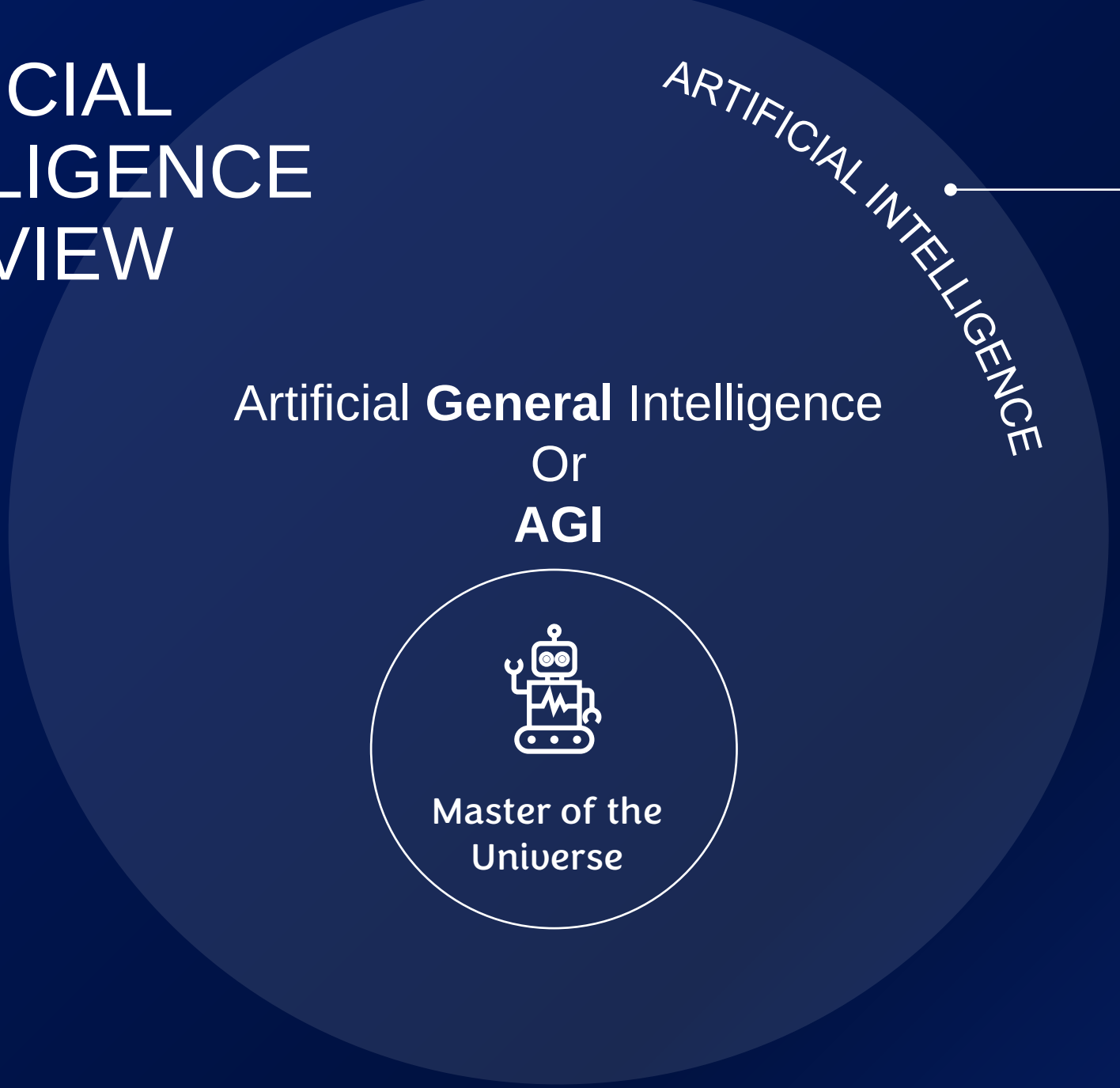
Timing is everything …

What makes this time any different?




181 ZB
data by 2025
572 ZB
data by 2030
Smart phones
Search
5G
Social tech
Drones
Autonomous vehicles
xchain
Smart robots
LiFi
dvanced analytics
3G/4G
VR/AR
Storage & cloud
Speech recognition
Fully ubiquitous Internet
Micro data centers
Wearables
Biometrics
Bioprinting
Smart data aggregation
GPS
3D printing
IoT
Effective computing
Machine learning

ARTIFICIAL INTELLIGENCE OVERVIEW



ARTIFICIAL INTELLIGENCE (AI)
Any technique that enables computers to mimic human intelligence.

What can AGI do?

Everything.

Better, and faster than humans

ARTIFICIAL INTELLIGENCE OVERVIEW

ARTIFICIAL INTELLIGENCE

ARTIFICIAL INTELLIGENCE (AI)
Any technique that enables computers to mimic human intelligence.

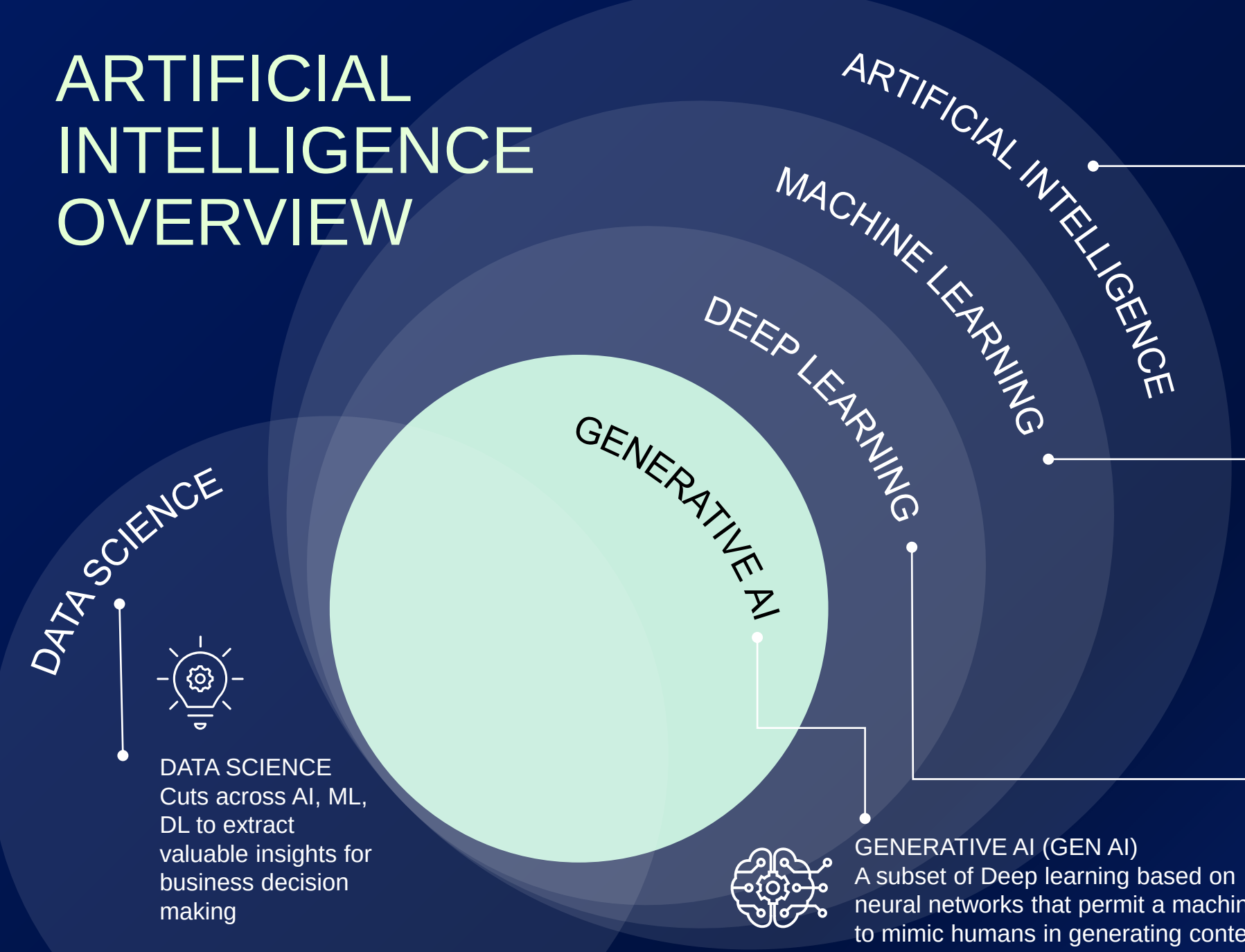
Artificial **Narrow** Intelligence
Or
Narrow AI

What can Narrow AI do?

One task at a time:

Play chess
Speech Recognition
Image Recognition
Generate response
Generate music

ARTIFICIAL INTELLIGENCE OVERVIEW



DATA SCIENCE
Cuts across AI, ML, DL to extract valuable insights for business decision making

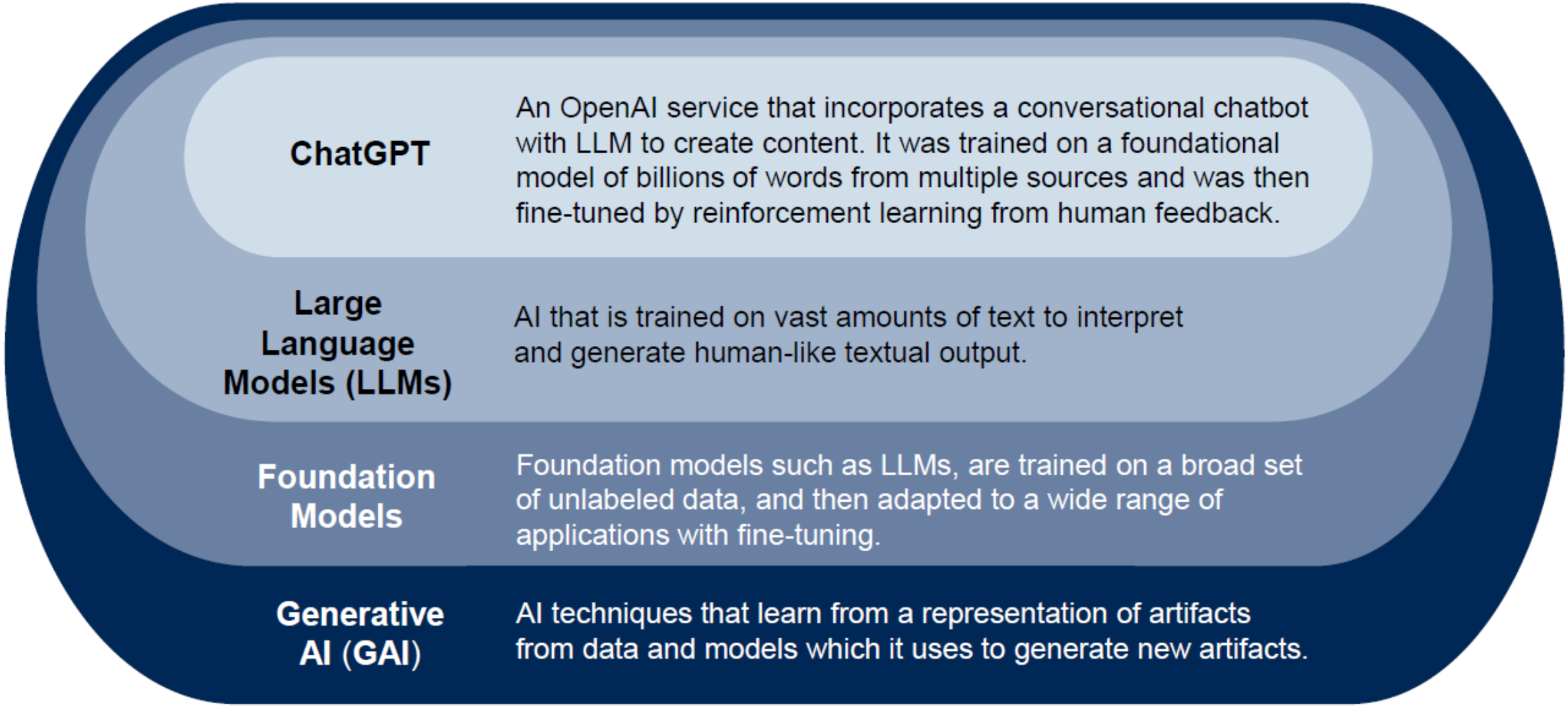
GENERATIVE AI (GEN AI)
A subset of Deep learning based on neural networks that permit a machine to mimic humans in generating content

DEEP LEARNING (DL)
A subset of machine learning based on neural networks that permit a machine to train itself to perform a task.

MACHINE LEARNING (ML)
A subset of AI that includes techniques that enable machines to improve at tasks with experience. It includes *deep learning*.

ARTIFICIAL INTELLIGENCE (AI)
Any technique that enables computers to mimic human intelligence. It includes *machine learning*.

What's Generative AI?



Gartner - Trillions $ spent on AI, especially Generative AI

Between 2023 and 2027,
$3 Trillion dollars
will be spent on AI

By 2027

- AI spending will be 16% of total IT spending
- GenAI represents 36% of total AI Spending

By 2030

- Every dollar of GDP, created anywhere on the planet, will be influenced by AI.
- 100% of global IT spending will be directly on AI or indirectly supporting AI
- 100% of IT development will incorporate AI in the design, development, testing or supporting production
- Every person in a developed country will have an interaction with at least one AI instance every single day.

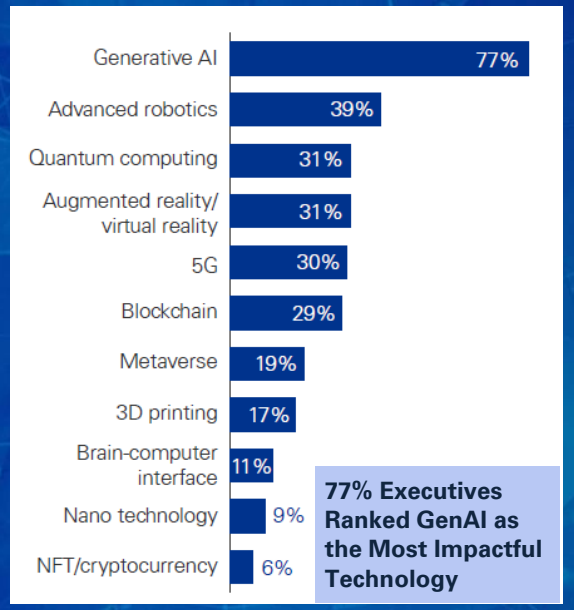
Source: Gartner 3Q23 Update

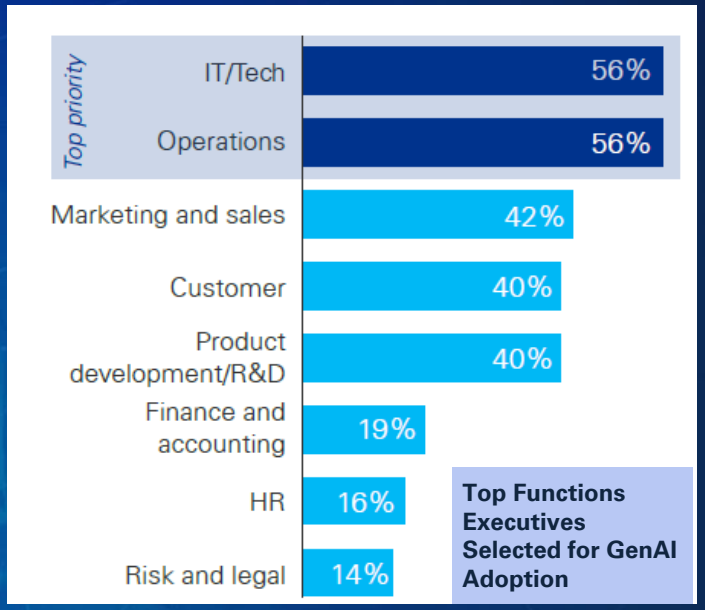
Gartner

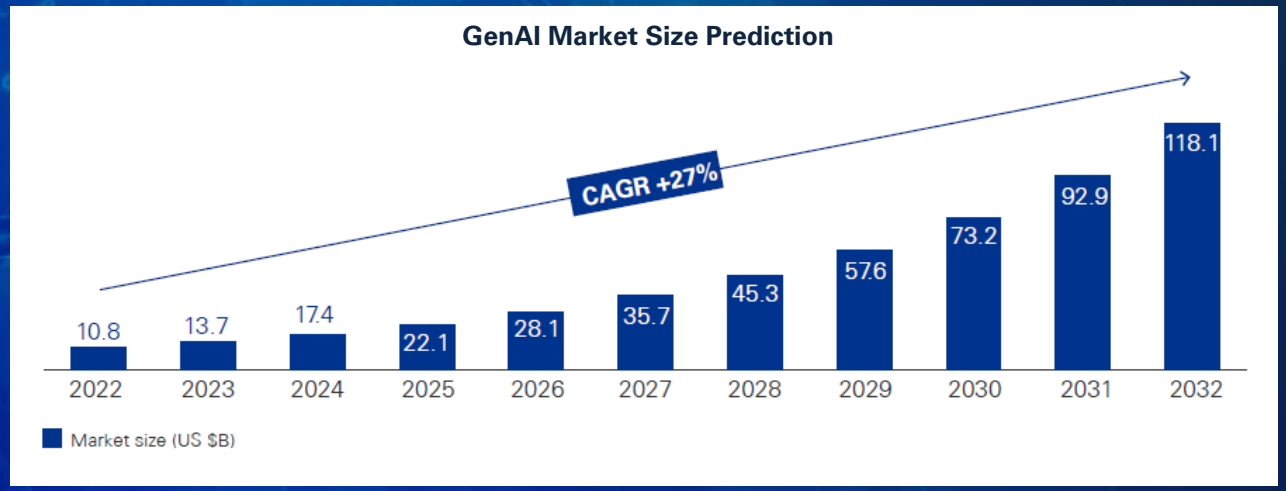
>70% CxO believe GenAI bring largest impact, plan to implement GenAI solutions within 2 years

Survey of business executives' opinions on Generative AI for 2024 and beyond:

- **77%** expect GenAI to have the **largest impact** on their business among emerging technologies
- **73%** believe GenAI will **increase workforce productivity**
- **71%** will implement their **first GenAI solution** within the next two years
- **64%** believe GenAI will help their business gain a **competitive advantage** over competitors







https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2023/generative-ai-survey.pdf

Generative AI is the catalyst for multi-decade productivity

Generative AI and automation everywhere era will drive

$1 trillion in productivity gains

by 2026*

Organizations will spend

346 billion on products and services

to implement Generative AI from 2024-2027*

AI will drive

$17 trillion global economic growth impact

over a 10-year period*

*Based on the IDC InfoBrief commissioned by Dell Technologies: Workforce Upskilling for the AI Era, January 2024
https://www.delltechnologies.com/asset/en-us/solutions/business-solutions/industry-market/idc-infobrief-workforce-upskilling-for-the-ai-era.pdf.external

GenAI is the multi-decade productivity driver

Represented by an upwards of **20%** jump in productivity gains across main use cases

Productivity Gains across four main use cases[4]	
Customer Operations	↑ ~40%
Content Creation and Management	↑ ~30%
Software Development	↑ ~50%
Sales	↑ ~5%

Gen AI Expected Impacted

1% OF US GDP - EXPECTED US INVESTMENT IN AI BY 2030[1]

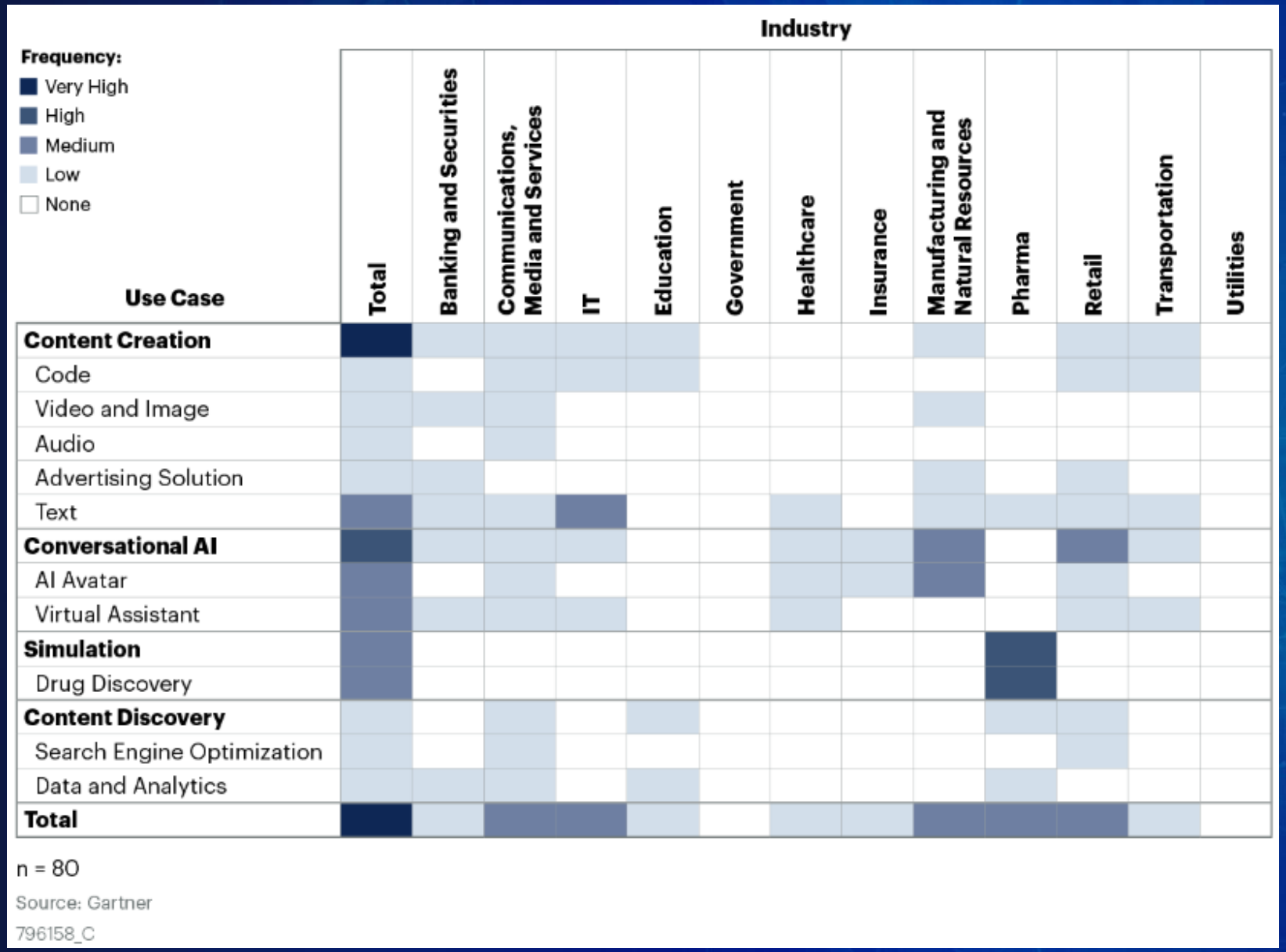
$17T IN GLOBAL ECONOMIC GROWTH DRIVEN BY AI OVER A 10-YEAR PERIOD[2]

25% OF CURRENT WORK TASKS COULD BE AUTOMATED BY AI IN THE US AND EUROPE[3]

$4.4T POTENTIAL ADDITIONAL TO THE GLOBAL GDP[4]

1. Goldman Sachs - The potentially large effects of AI on economic growth (Briggs/Kodnani), March 2023
2. Goldman Sachs - Generative AI - Part I: Laying Out the Investment Framework, March 2023
3. Goldman Sachs - The potentially large effects of AI on economic growth (Briggs/Kodnani), March 2023
4. McKinsey Report

Every industry has different focus and uses for Generative AI



n = 80

Source: Gartner

796158_C

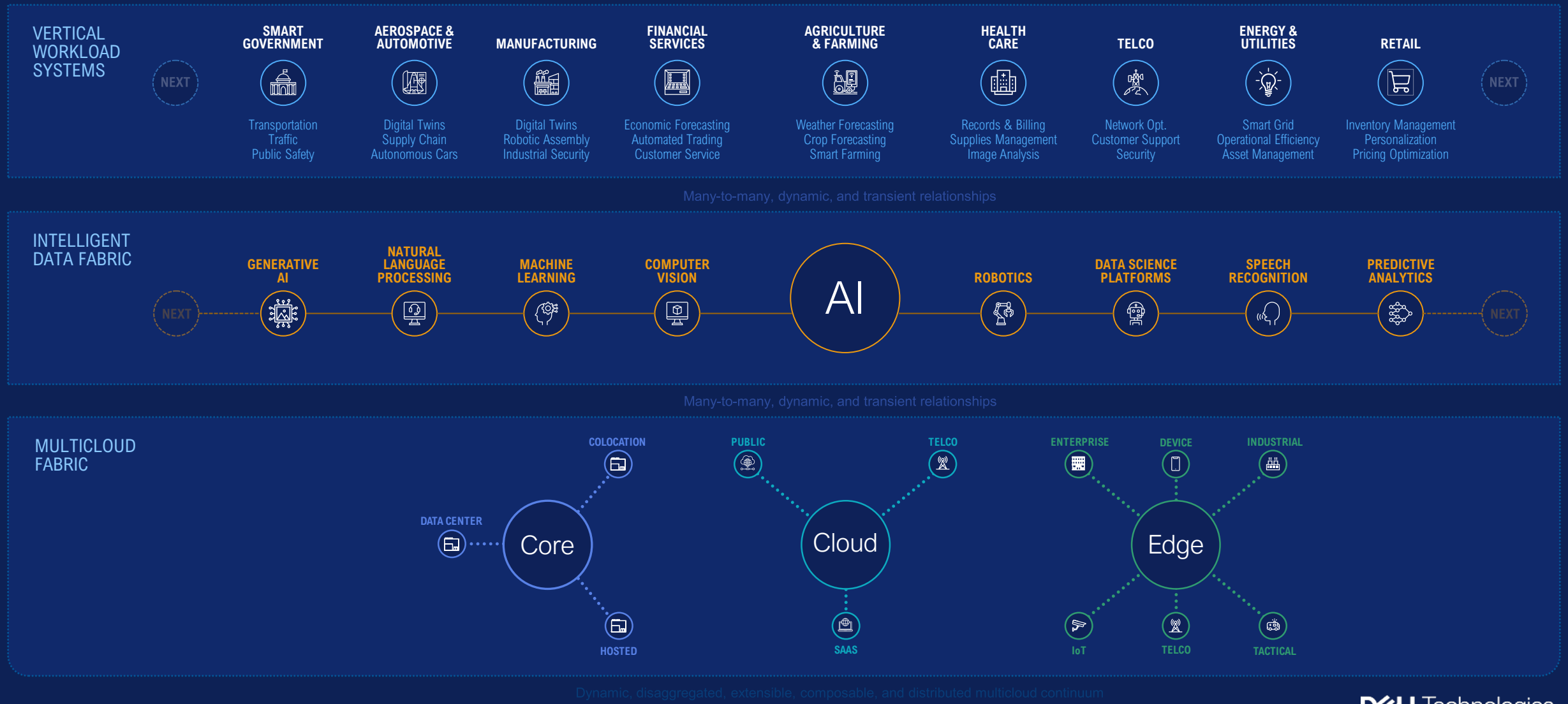
Primarily creation of text-based content

Image, video and voice generation, as well as AI avatars, used in marketing and human resources departments.

Virtual assistants and content discovery tools highly leveraging GenAI

Drug discovery, is an early application of GenAI that has gained notable traction within the pharmaceutical industry.

Multicloud: AI Landscape



Multicloud: AI Landscape

Multicloud
by design.

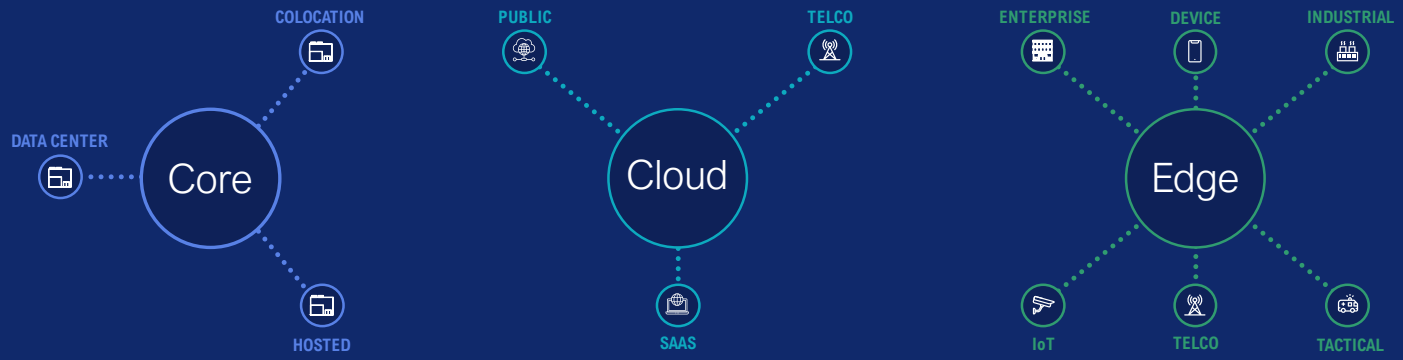
Total Addressable Markets (TAM) by Type of AI

Sources in notes

INTELLIGENT DATA FABRIC

Many-to-many, dynamic, and transient relationships

MULTICLOUD FABRIC



Dynamic, disaggregated, extensible, composable, and distributed multicloud continuum

DELL Technologies

Bring AI to Data

Move quickly

Drive business transformation

Increase productivity

Traditional AI

Data set inputs

Focused on detecting patterns

Classify data, make decisions

1000s to millions of data points

Data

Applications

Generative AI

Drive insights from new modalities

Produce text, graphics, video

Transform data into intelligence, faster and at scale

Synthesize billions of data points

Applications

AI Developer experience

AI Framework, Models

AI Infrastructure

DELL Technologies PARTNER PROGRAM

DELL Technologies

AI Headwinds Slowing Down AI Adoption

76% of IT and business leaders believe GenAI will deliver transformative value for their organization[1]

Top challenge for enterprise AI adoption is an **AI skills and talent strategy**[2]

The #1 factor considered for choosing GenAI models is **security**[3]

Top challenge organizations face when implementing AI is **cost**

[1] Dell Technologies Innovation Catalyst Study, February 2024
[2] IDC Report: From Breakthrough Innovation to Impact: Monetizing the AI Moment Philip Carter, Directions 2024
[3] Dell Technologies Generative AI Pulse Survey, August and September 2023, www.dell.com/GenAIPulse
[4] IDC, Global AI Buyer Sentiment, Adoption and Business Value Survey, October 2023.

Dell's Strategy for Accelerating AI Adoption

SIMPLE

Comprehensive AI services and validated, optimized solutions make AI easier, augmenting skills gaps and addressing data readiness

SECURE

Bringing AI to customers' most valuable data on-premises and at the edge, maintains control and protects intellectual property

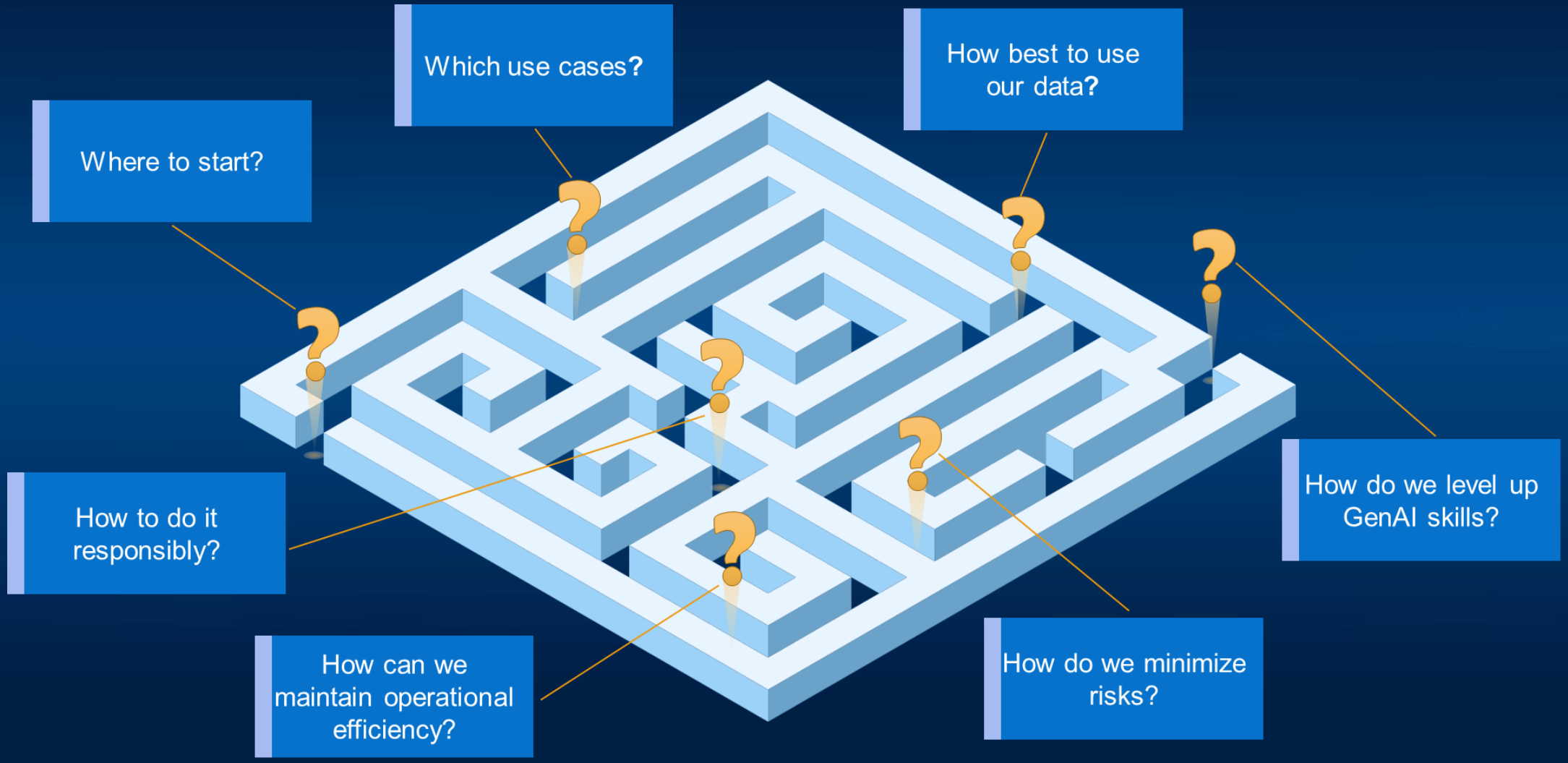
ECONOMICAL

Right sizing AI investments with the industry's broadest AI portfolio and leveraging on premises implementations to lower TCO by up to 75%

DELL Technologies

How can you benefit from AI *today* ?

Customer needs guidance



AI Journey: Two Common Questions

How do I start?

Where do I start?

VALUE PROPOSITION

- High Value/Impact use cases
- Demonstrate ROI
- Data Driven Transformation

PEOPLE

- Skillsets and Expertise
- Knowledge retention

PROCESS

- AI/ML Pipeline
- Feature Store
- Streamlined processes
- Analytics Assets

PLATFORM

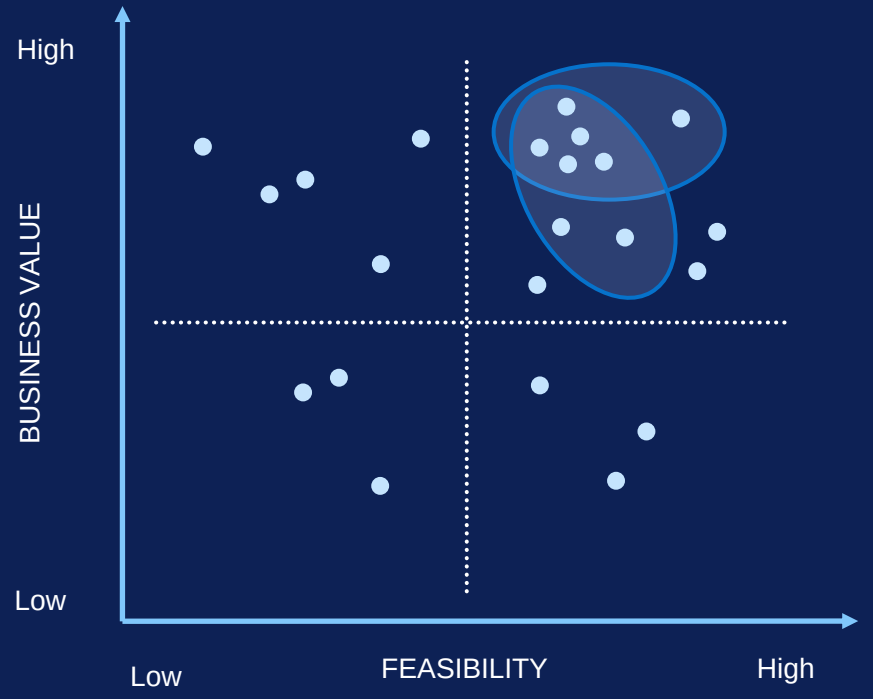
- Secure environment
- Right tool for the job
- Stop shadow IT
- Focus on core skills
- Significant time/effort savings

Prioritizing Generative AI use cases

Criteria for consistently evaluating use cases

Business value criteria

- More
- Better
- Faster
- Cheaper



Feasibility criteria

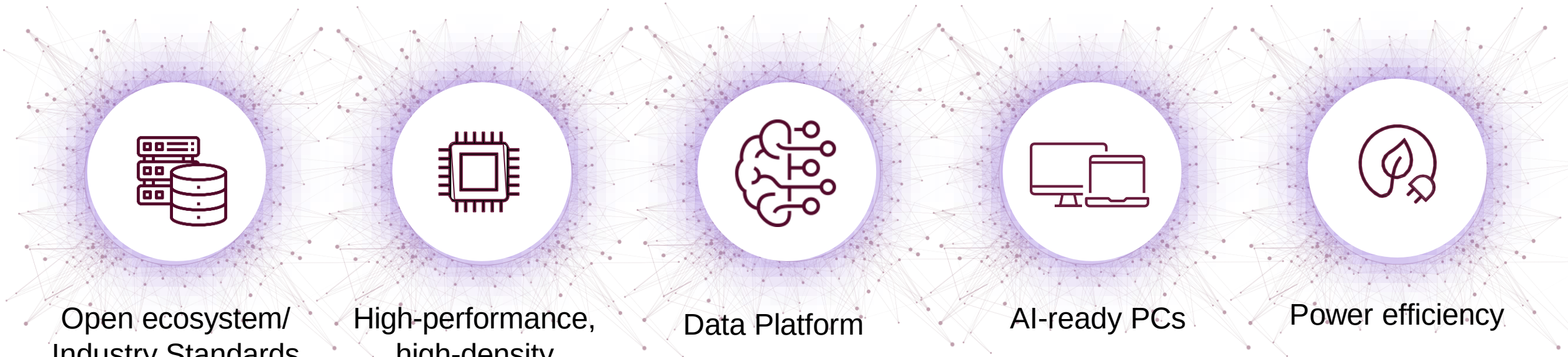
- Data
- AI model
- Process
- People
- Platform

Content creation | **Digital assistant** | **Natural language search** | **Design and data creation** | **Code generation** | **Document automation**

DELL Technologies PARTNER PROGRAM

Use cases

The Dell approach for AI technology foundation



Open ecosystem/ Industry Standards

High-performance, high-density compute

Data Platform

AI-ready PCs

Power efficiency

Dell durable advantages

Technology leadership, winning portfolio, operational capability, go-to-market, ecosystem

OPEN, DIVERSE,SCALABLE

Creating a consistent experience – from workstation to edge to core to cloud – for IT Ops, Data scientists, and Developers

High-performance, High-density Compute

Market Trends

- **Large AI Training + Inference Adoption**
- **Silicon Diversity**
- **Rise of Large Language Models**
- **Increasing CPU Capabilities**

Silicon Diversity

QCOM

Inspiron (TBA – 8CX Gen 4)

Latitude (TBA – 8CX Gen 4)

XPS (TBA – 8CX Gen 4)

NVIDIA

XE9680 H100

XE8640 H100

XE9640 H100

Precision 7960 Tower

Dual GB200

B100

"Archer"

GB200

Falcon Shores

MI400

INTEL

R760xa Xeon Flex Series

XE9640 Max Series

Precision 7780

Gaudi 3

AMD

R7625 EPYC MI Instinct 210

Precision 7875 Tower

MI300X

MI350X

DELL Technologies

Right-sizing your AI investment

Business

Business strategy and operations management	Product innovation and R&D	Manufacturing and supply chain	Marketing, sales, and customer service	IT, HR ,and Finance	Users / Dataset	Location / •••

Example Outcome

Reduce cost of goods	Increase speed of innovation	Logistics Optimization	Content creation Customer intimacy	Employee satisfaction Automate issue resolution	Users / Dataset	Location / •••

Use Case

- **Content creation**
- **Digital assistant**
- **Natural language search**
- **Design and data creation**
- **Code generation**
- **Document automation**

Patterns

- **Pre-trained model/ inferencing**
- **Model augmentation**
- **Fine-tuning**
- **Model training**

NVIDIA AI Enterprise

Includes AI workflows, frameworks, pre-trained models, AI and data science development and deployment tools as well as cloud-native management /orchestration, and infrastructure optimization

Infrastructure

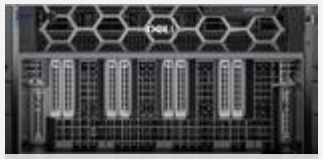
Powered by NVIDIA GPUs

Laptops

Workstations

Servers

Dell Validated Designs



Compute

Network

Storage

Professional Services

- Consultancy
- Advisory
- Data science

DELL Technologies

The Dell AI Factory

The world's broadest AI solutions portfolio from desktop to data center to cloud

DATA: FUELING THE AI FACTORY

Data is the raw material that powers the AI Factory

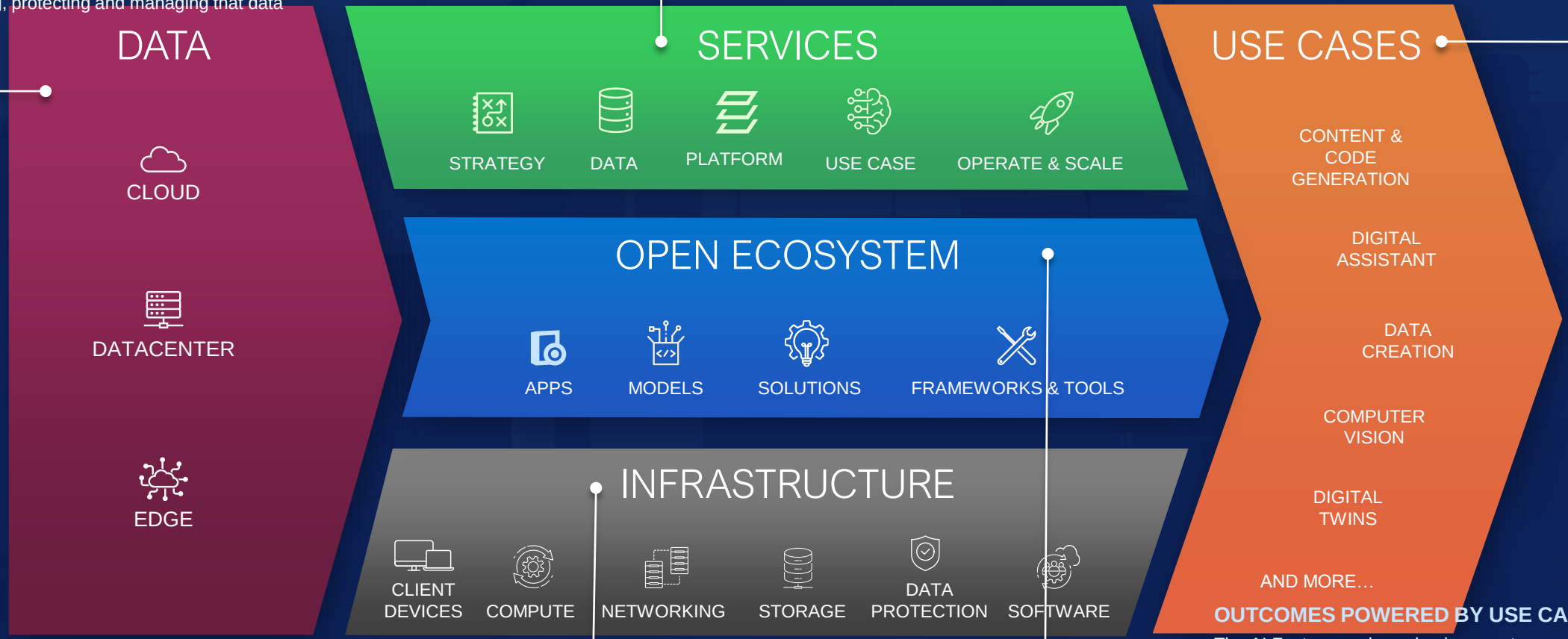
Most valuable data is on-premises and at the edge

Dell helps you bring AI to your most valuable data and is a leader in storing, protecting and managing that data

EXPERT AI SERVICES

An effective AI Factory needs a skilled team to succeed, but AI-ready skills are in short supply and the ecosystem is diverse

Dell has extensive experience guiding customers through their AI journeys, accelerating AI outcomes aligned to business objectives while utilizing the right technical solutions at scale



AI OPTIMIZED INFRASTRUCTURE

Infrastructure is the foundation of the AI Factory

AI use cases are diverse and AI technologies are rapidly evolving

Dell has the broadest AI portfolio from client to cloud allowing you to right size your AI investment and giving you the flexibility to run AI anywhere

AI OPEN ECOSYSTEM

No single vendor can meet all your AI needs

Dell has a proven history of nurturing open ecosystems and is committed to building a broad AI ecosystem giving customers greater access to innovation and flexibility

OUTCOMES POWERED BY USE CASES

The AI Factory produces business outcomes powered by your highest priority use cases

Dell simplifies the deployment of your most important AI uses cases with validated solutions and tailored services

DELL Technologies

Rapidly deliver digital assistants with a proven solution

GenAI Development Frameworks & Partner SW

- Interface
- Dell Middleware
- Information Store
- NVIDIA Audio2Face

GenAI Deployment Tools

- NVIDIA Triton Inference Server
- NVIDIA TensorRT-LLM
- NVIDIA GPU Operator

Infrastructure Management

- Container Storage Interface Drivers
- Kubernetes
- Enterprise Linux

Infrastructure from Device to Data Center

Dell Compute with NVIDIA GPUs

Dell Storage

Dell, NVIDIA Networking

Smart City Example

How long will my building permit take to process?

The initial review of simple permit applications is typically 2-3 weeks. Complex permits take around 8 weeks

Example life-size holographic display

Multiple Display Options

Professional Services

Strategize → Implement → Adopt → Scale

Strong partner ecosystem to support your GenAI initiatives

Models

Hugging Face · NVIDIA · Meta

ML Frameworks

PyTorch · TensorFlow · RAPIDS · ONNX

Software Models

NVIDIA · OMNIA · slurm workload manager · kubernetes · Kubeflow · mlflow

Operating Environment

Windows · Red Hat · vmware · SUSE · Ubuntu

Global GTM Partners*

World Wide Technology, Inc. · Cognizant · HCLTech · Deloitte. · TCS TATA CONSULTANCY SERVICES · accenture · Infosys · kyndryl · EY · DXC TECHNOLOGY

Silicon Vendors

intel · AMD · NVIDIA · Qualcomm

COLOCATION PARTNERS

EQUINIX · DIGITAL REALTY · FLEXENTIAL · CyrusOne · DATABANK · switch

HOSTERS

rackspace technology

Data Management

Data Discovery & Access

Starburst · teradata. · CLOUDERA

Data Ingestion

Starburst · elastic · CONFLUENT · DELL Technologies STREAMING DATA PLATFORM

Data Transformation & Data Engineering

Starburst · DELL Technologies STREAMING DATA PLATFORM · databricks · CLOUDERA · snowflake · teradata.

AI & Machine Learning

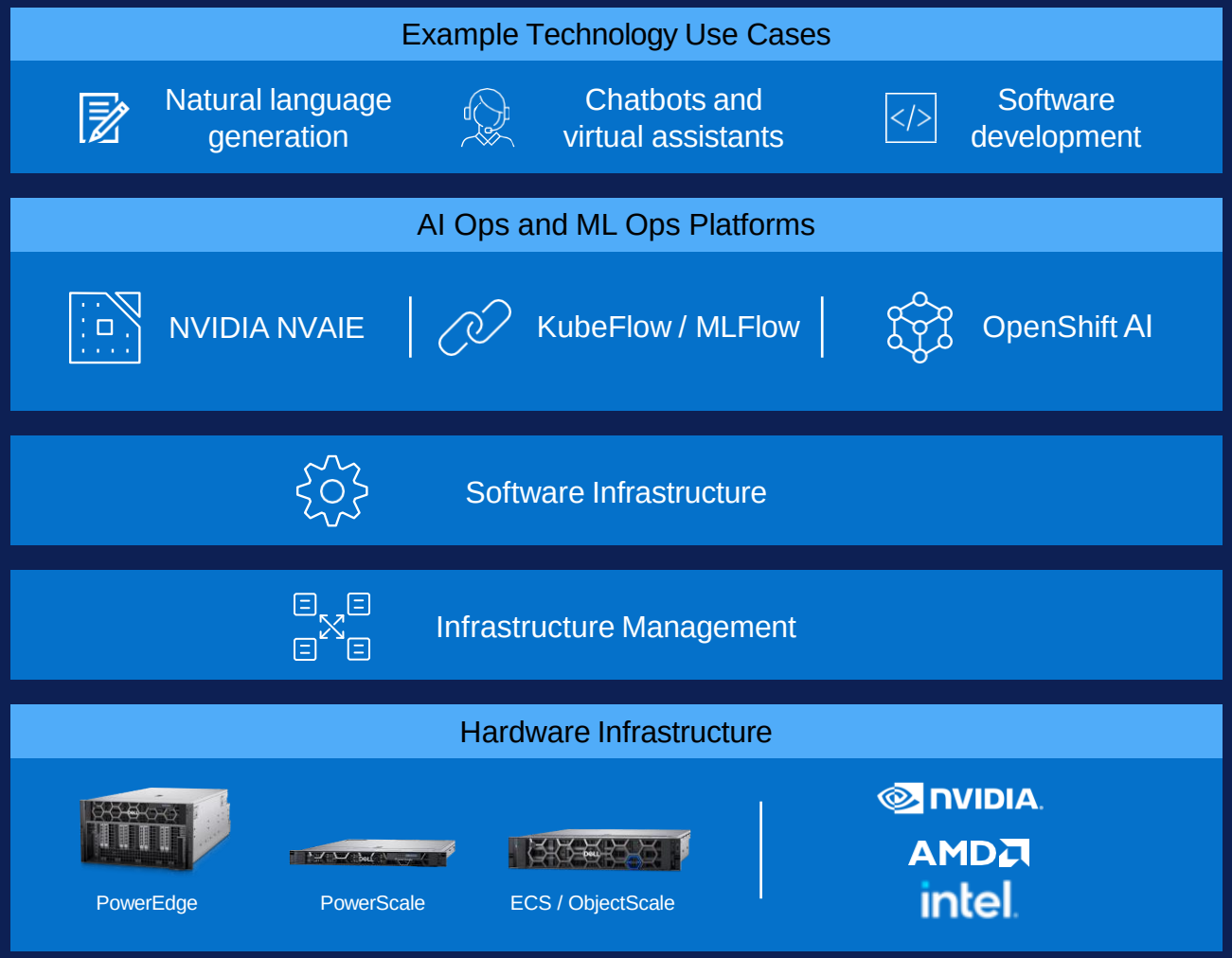
NVIDIA · cnvrg.io · DOMINO · H_2O.ai

*representative not comprehensive

DELL Technologies

Dell Validated & Reference Designs

Silicon diversity, use-case driven



Dell Reference Designs

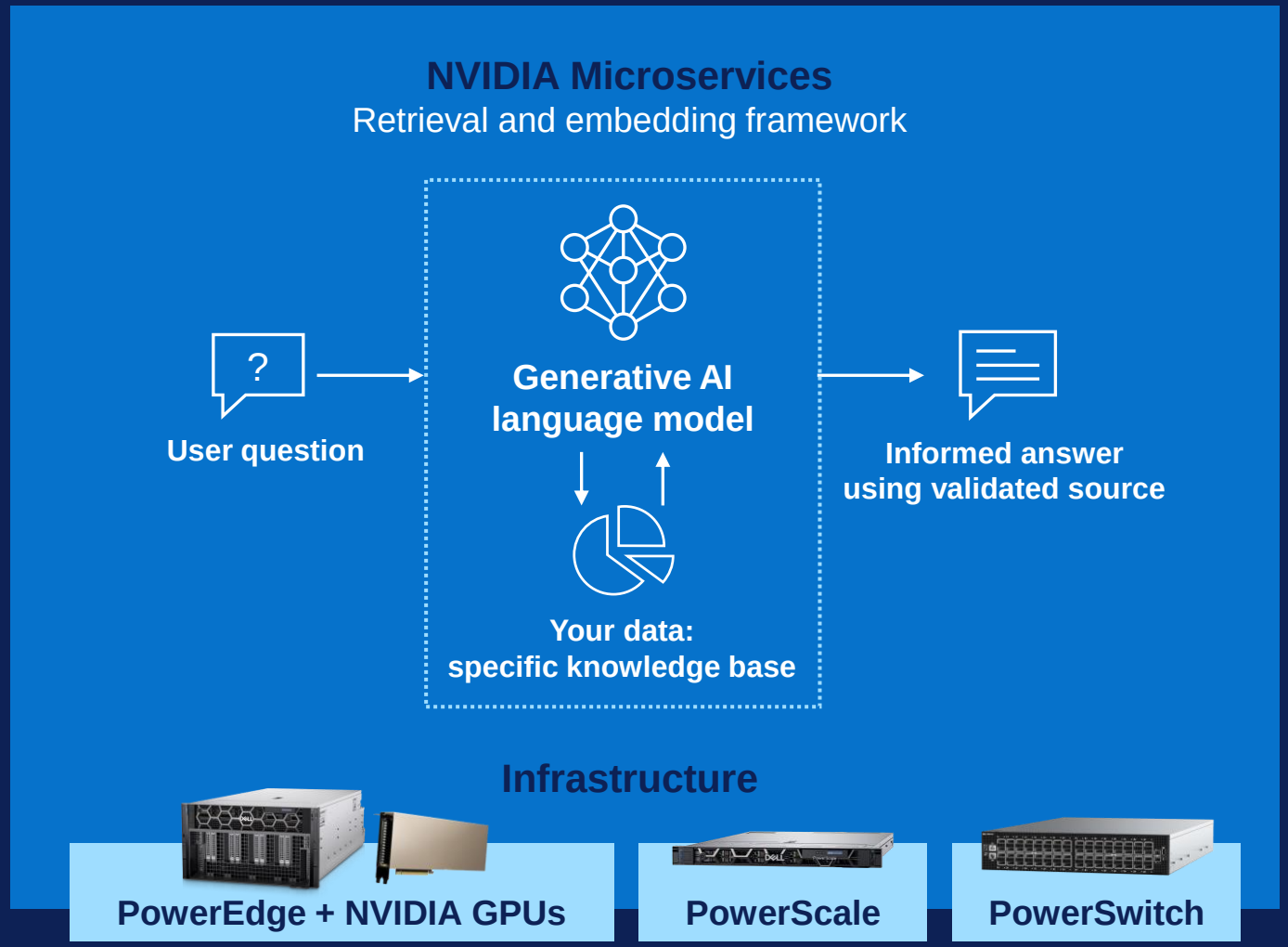
Primarily developer-forward use case validations, sometimes including code artifacts and example playbooks

Dell Validated Designs

Primarily infrastructure-forward engineering validations, based predominantly on compute, fabric, and storage, basic use case validation up to MLOps layer, possible MLPerf benchmarking

Quickly delivering a more accurate and reliable GenAI application

Retrieval Augmented Generation (RAG) design with NVIDIA Microservices



- Simplify deployment of scalable information retrieval frameworks for Generative AI models
- Enable more precise and trustworthy answers for their Generative AI models
- Get better model performance via a retrieval-based approach
- Reduce dependence on pre-trained, open-source LLMs with limited relevant data

DELL Technologies PARTNER PROGRAM

Ecosystem

AI and Network

AI – IN-THE-NW

Making telco networks smarter by embedding AI directly into them

Results in making networks more intelligent and more automated

Ex: Neural Rx, Neural Scheduler etc.

AI – ON-THE-NW

Designing networks to deliver the outcomes for customer AI workloads

Establishes a new class of workload made specifically for AI

Ex: Sovereign AI Factory, Edge Inferencing etc.

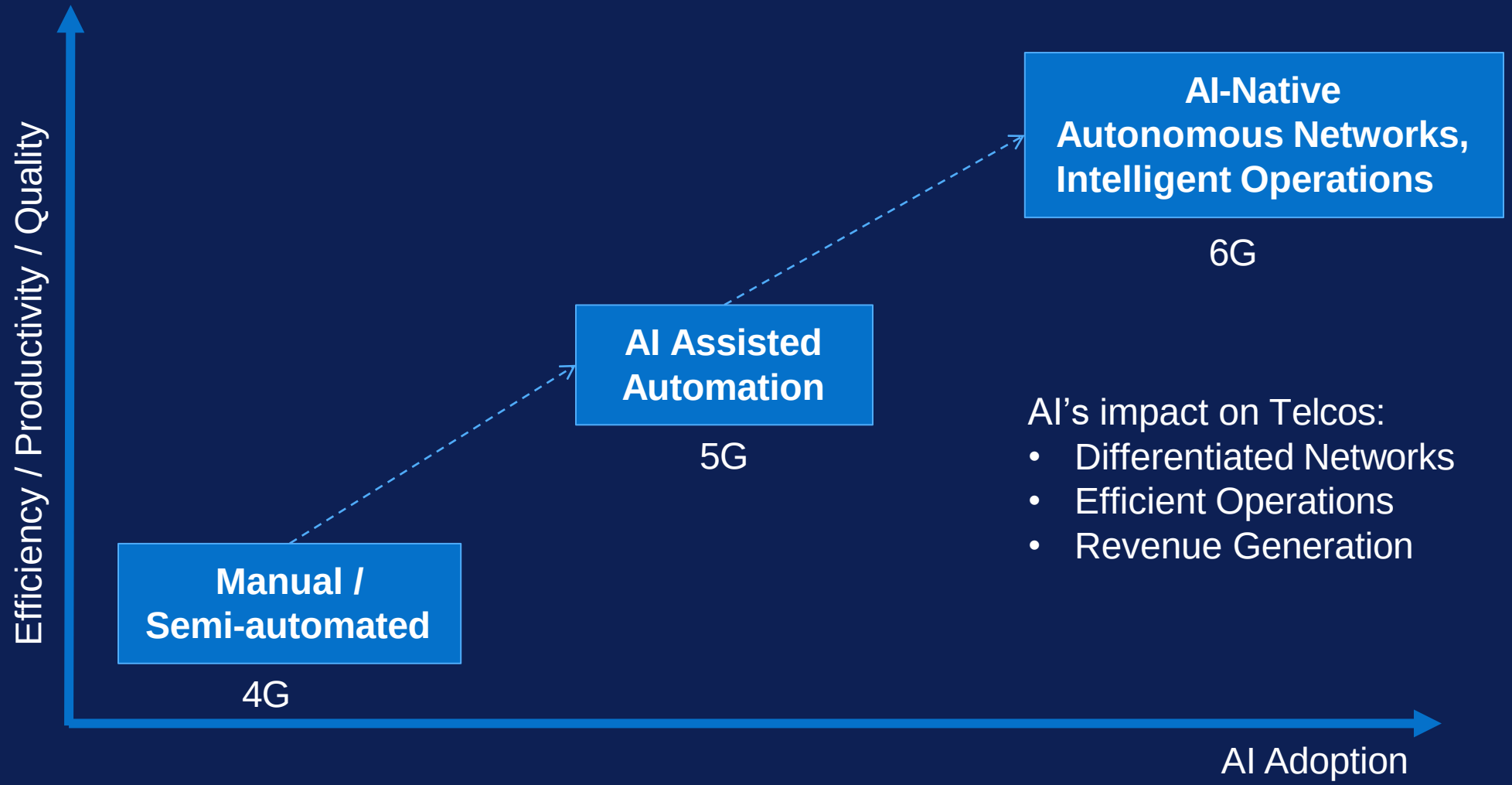
AI – AND-THE-NW

AI enable infrastructure shared by NW Functions and AI workloads

Enables Telcos to run their own NFs and monetize 3rd party AI workload, multi-tenant

Ex: Multi-tenant Edge and RAN workloads

From AI-Assisted to AI-Native Telco Journey



AI's impact on Telcos:

- Differentiated Networks
- Efficient Operations
- Revenue Generation

All aspects of Telco business will be enhanced by AI
Networks, Operations, Sustainability, Marketing, Customer Care etc.

AI Use Cases for Telcos

Autonomous Networks

- Planning, Forecasting
- Configuration Recommender
- RF Optimization, Cell Shaping
- Digital Twins
- Performance
- Green, Energy Efficiency
- Service Assurance
- Closed-Loop-Automation, Control Loops, Fault Management

Differentiation

Traditional AI/ML/DL

Efficient Operations

- Customer Care
- Digital Marketing, text-to-video
- Call Center Assistant
- Chatbots, Q&A, Summarization
- Churn Prediction, Sentiment Analysis
- NOC-of-the-Future, Dynamic Dashboarding
- Code Development Copilot
- Field Support Co-pilot

Productivity

GenAI / LLM

Revenue Generation

- Training – GPUaaS
- Inferencing – Latency, Data Gravity
- Data Security
- Data Sovereignty
- Regulations – EU AI Act

Growth

AIaaS

Finance / Audit

- Fraud Prevention
- Payout prediction
- Internal Audit
- Investment Optimization
- Money Laundering Detection
- Email Surveillance
- Sales Forecasting
- Revenue Prediction
- Spend Analysis

Operations

- 5G Utilization
- Network Optimization
- Predictive Maintenance
- Load Balancing
- Capacity Planning
- Anomaly Detection

Telco Customer Service

- Automated Responses to FAQs
- Support Calls prediction
- Personalized packages
- Customer Support Chatbots
- Technical Troubleshooting
- Service Outage Notifications
- Network Coverage Information

Sales and Marketing

- Content Generation
- Sales Forecasting
- Surveys and Feedback Forms
- Market Sentiment Analysis
- 5G Propensity Modeling
- Lead Prioritization
- Cumulative Payment Prediction
- Cross-Selling and Upselling
- Dynamic Pricing Strategies
- Personalized Email Campaigns

Research and Development

- Suspicious Activity Detection
- Alert Optimization
- Predictive Traffic Management
- Real-time Translation Services
- Document Processing
- Data Extraction and Validation
- Customer Feedback Analysis
- Predictive Maintenance
- Real-time Production Monitoring
- Scenario Modeling

Legal / Regulations

- Contract Generation and Review
- Regulatory Filings
- Language Translation and Interpretation
- Legal Research and Analysis
- Privacy Policy Generation
- Regulatory Communication Templates
- Compliance Audits
- Anonymity Assurance
- Automated Regulatory Reporting

Security, Ethics & Business Alignment

AI Personality & Alignment

AI Security

Prioritized Use Cases

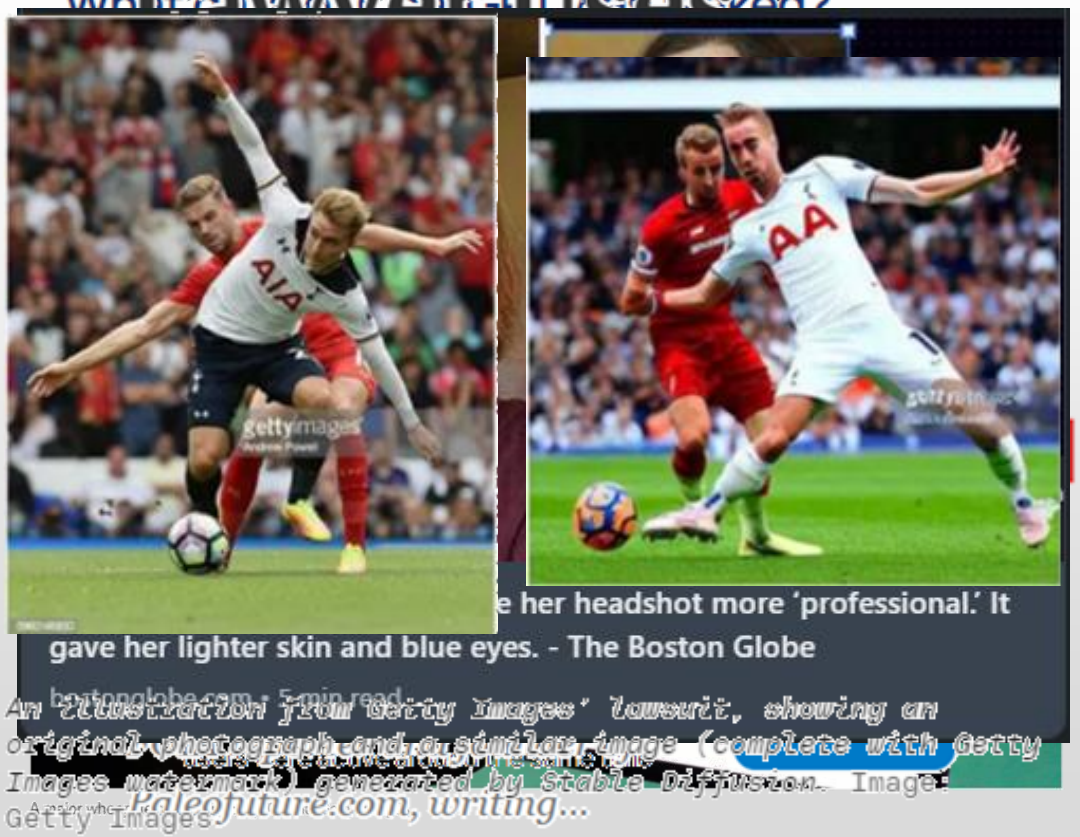
Multi Disciplinary Team Approach

The Rise of the Chief AI Officer

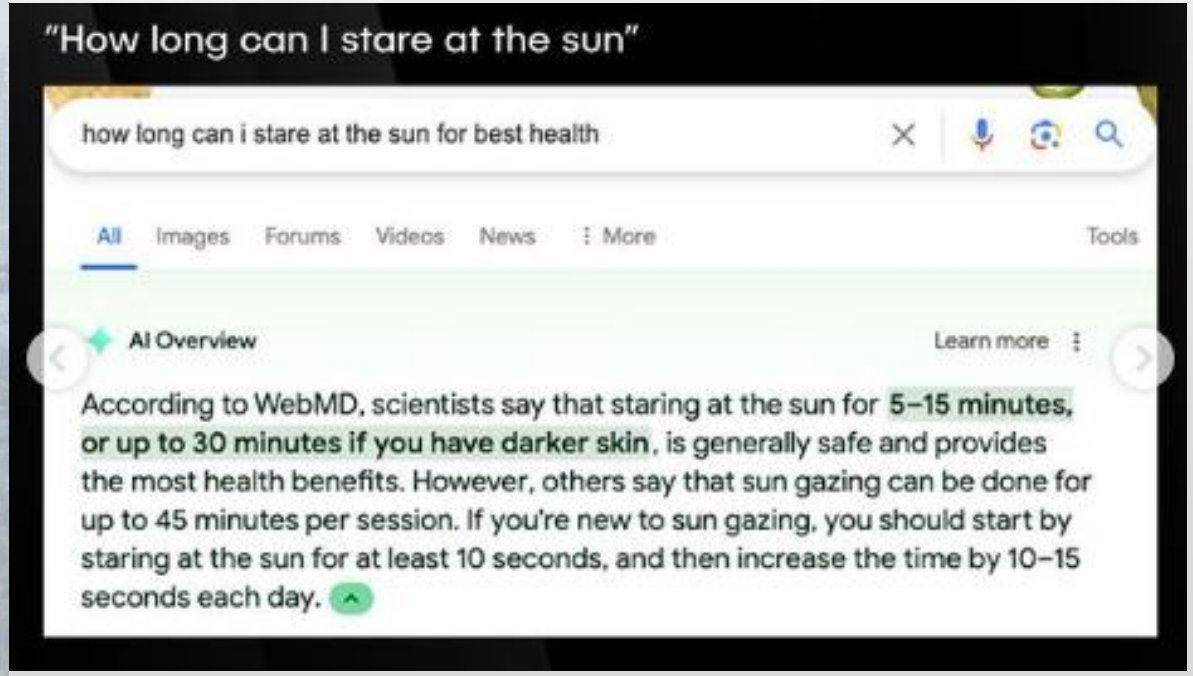
Risks of getting GenAI wrong

Don’t make the headlines

- Intellectual property loss
- Data leakage
- Privacy issues
- Compliance violations
- Credibility & integrity risks
- Bias
- IP infringement



Risks of getting GenAI wrong



Cost is a major consideration for LLM inferencing

Dell commissioned a study with Enterprise Strategy Group (ESG) comparing the expected costs to inference LLMs on-prem with Dell infrastructure vs. **public cloud IaaS** and **API services**[1].

Over a three-year period, Dell can provide inferencing that is up to:

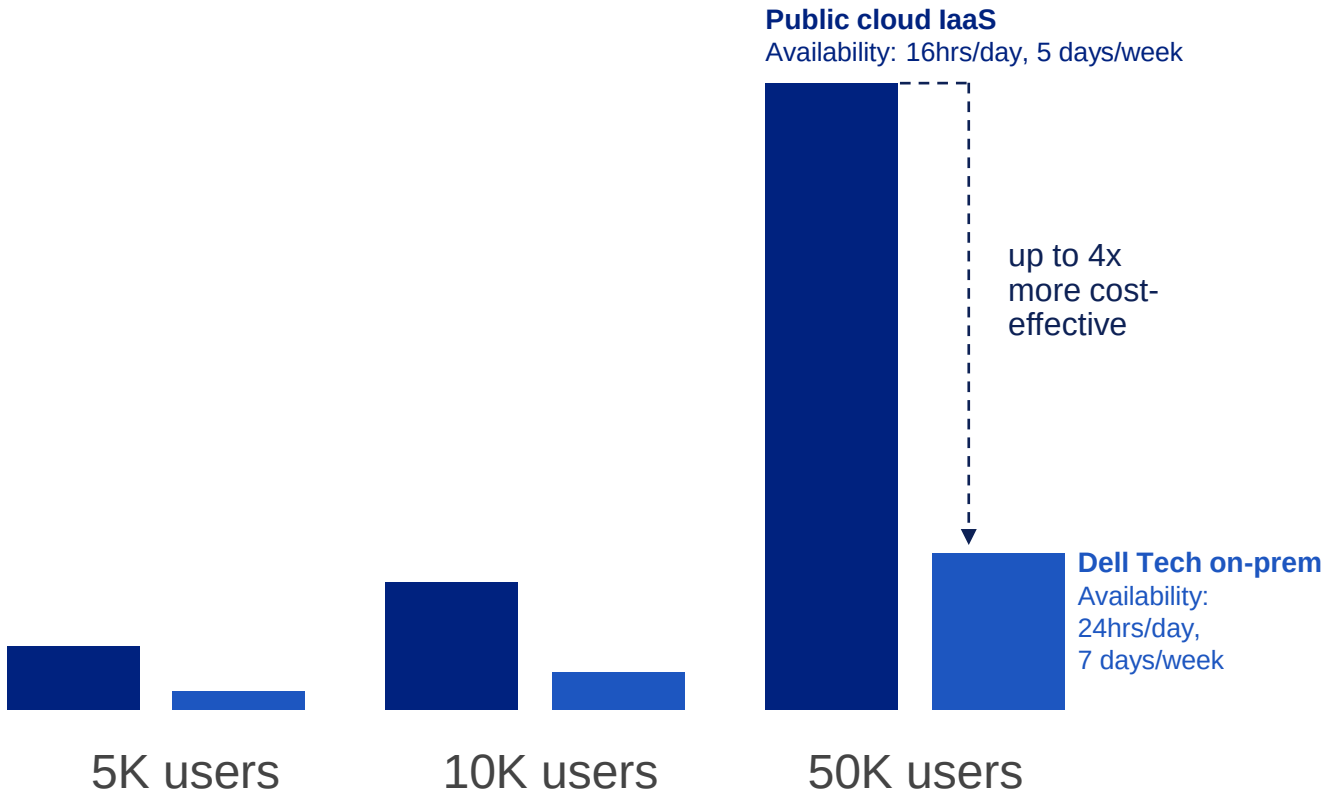
4x

more cost-effective than public cloud IaaS

8x

more cost-effective than token-based API services

Expected Cost to Deliver Inferencing for 70B Parameter Llama 2 LLM Using RAG



1. Based on Enterprise Strategy Group research commissioned by Dell, comparing on-premises Dell infrastructure versus native public cloud infrastructure as a service, April, 2024. Analyzed models show a 7B parameter LLM leveraging RAG for an organization of 5k users being up to 38% more cost effective while a 70B parameter LLM leveraging RAG for an organization of 50k users being up to 75% more cost effective. Actual results may vary.

Forbes | June 17, 2024 | Edge, Data Center/Colo, Public Cloud

- The rapid growth of AI initiatives is driving the need for a new type of infrastructure that delivers accelerated computing, with existing Hyperscalers like AWS, Azure, and GCP revising their legacy infrastructure strategy to accommodate workflows that juggle billions or even trillions of parameters.
- Despite deployment flexibility, over 80% of organizations are expecting to repatriate some compute and storage resources to private cloud or non-cloud environments, with 93% of IT leaders involved in cloud repatriation projects in the past three years, citing rising public cloud costs and performance concerns.
- Organizations deploying extremely latency-sensitive apps (LLM, GenAI, finance, aerospace, autonomous driving, or life sciences) may find on-premises or colocation service levels more predictable and effective, as "GPU-based servers, parallel file systems and fiber optic networks—that can now deliver speeds measured in petabits per second—are available on-premises".

DELL Technologies
PARTNER PROGRAM

2024 NVIDIA GTC

"Nobody is better at building end-to-end systems of very large scale for the enterprise than Dell."

Jensen Huang
Founder, President & CEO
NVIDIA

DELL Technologies

Why Dell is the right GenAI Partner

Broadest GenAI solutions portfolio*

Dell's Partner ecosystem of GTM and Technology Partners

More control and secure access

Right-size AI needs

Faster Time-to-value for your customers

*Based on Dell analysis, August 2023. Dell Technologies offers solutions engineered to support AI workloads from Workstations PCs (mobile and fixed) to Servers for High-performance Computing, Data Storage, Cloud Native Software-Defined Infrastructure, Networking Switches, Data Protection, HCI and Services. CLM-009177
Please do not use this claim in Argentina, Chile, China, Colombia, Costa Rica, Guatemala, Hong Kong, Korea, Malaysia, Russia, South Africa, Taiwan, Ukraine, Venezuela, Vietnam, Thailand, and the Philippines

Dell Technologies

The future of AI is not being written!

It is being reimagined!

DELL Technologies


delltechnologies.com/ai ▪ InfoHub
DELL Technologies