AI² | Citi event

Dell Chief Technology Office
AI adoption and essential architectures ”

Building An Open & Trusted Capability and community for Ireland aligned to thrive in the New and coming World’s Autonomous Ecosystem

Marc O’Regan
CTO - Dell Technologies EMEA
@reddogmarc

Who Owns AI?

AI Control & Sovereign Implementation Is critical to democratisation

AI Adoption Is Not As Expected

Machines were not built to natively interact and integrate with Human environments

- Underwhelming agentic AI systems
- Polarization and teenage depression caused by social media algorithms
- Slowing performance improvements & scaling risk from current foundation models
- Frustrating search engine inadequacy

These are just some of the failures that stem from a single fundamental weakness:

Today's AIs are limited to language and narrow data - They can't generalize to reality or to humanity

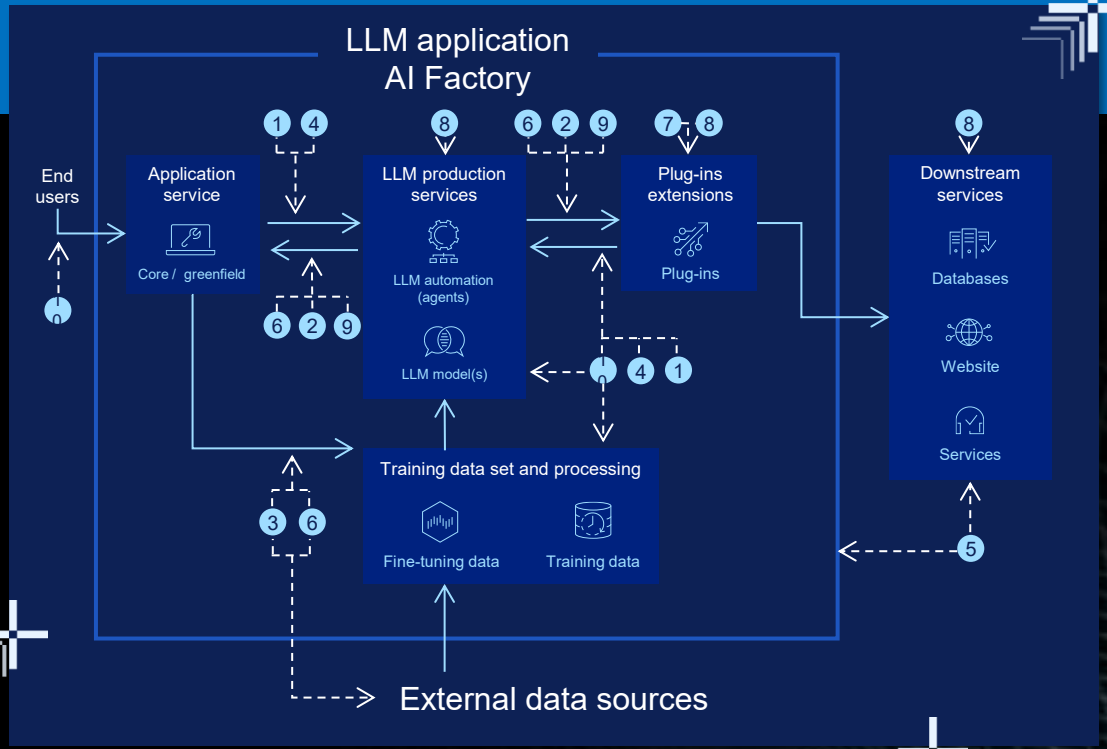
DELL Technologies

Concerns for AI and LLM

Protecting & Securing Models

Threats

- Prompt injection 1
- Insecure output handling 2
- Training data poisoning 3
- Model DoS 4
- Supply chain vulnerabilities 5
- Sensitive information disclosure 6
- Insecure plugin design 7
- Excessive agency 8
- Overreliance 9
- Model theft 0



Methods of Protection

- Model Firewalling & Privatization
- Sanitation Engine implementation
- Traditional Security Controls
- IAM / MDR / SDLC / .RBAC
- Modern Trust Insertion - ZT
- Data Resilience
- Secure Supply Chain
- AI Guardrails
- AI Network visibility & Firewall
- API security
- Logging and monitoring
- Penetration Testing
- Rate limiting

Adapted from OWASP, https://llmtop10.com/

Control Your Language Model

Run Your AI Where You Want It & where you "need" it

Public LLM - Token-based pricing

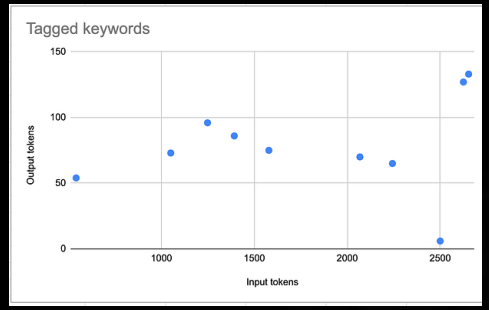
Region: Europe (Frankfurt)

aws Bedrock

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens
Claude Instant	$0.00248	$0.00838
Claude	$0.00800	$0.02400

Azure OpenAI Service

Models	Context	Prompt (Per 1,000 tokens)	Completion (Per 1,000 tokens)
GPT-3.5-Turbo	4K	€0.0014	€0.0019
GPT-3.5-Turbo	16K	€0.003	€0.004
GPT-3.5-Turbo-1106	16K	€0.001	€0.002
GPT-4-Turbo	128K	€0.010	€0.028
GPT-4-Turbo-Vision	128K	€0.010	€0.028
GPT-4	8K	€0.028	€0.055
GPT-4	32K	€0.055	€0.109

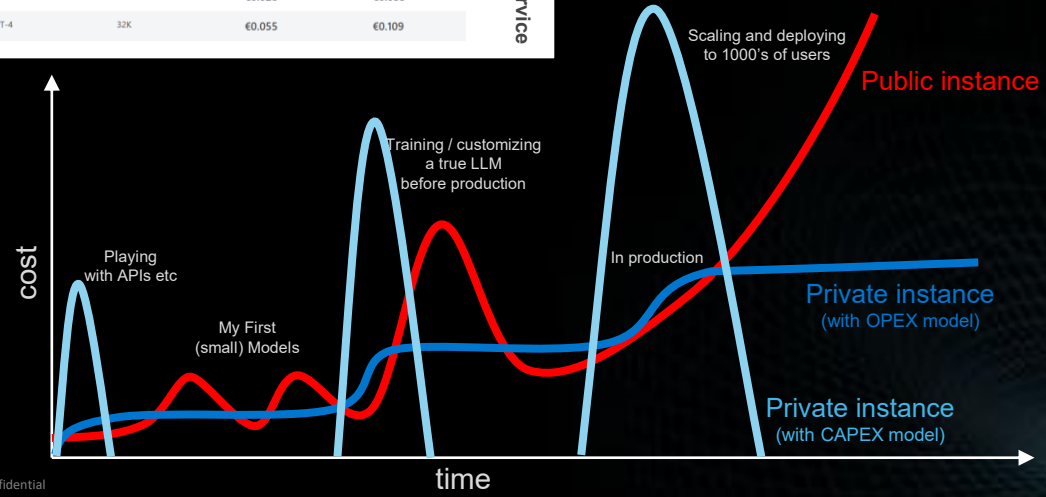


Unpredictable output length

Source :
https://neoteric.eu/blog/how-much-does-it-cost-to-use-gpt-models-gpt-3-pricing-explained/

Public LLM cost model:

- Unpredictable input length (esp. with prompt engineering methods)
- Unpredictable output length (depending on temperature etc)
- Unpredictable use
- Unpredictable pricing policies



Public instance cost model =
No cost predictability
No cost control

VS

Private instance = Buying IT
(as boring and predictable as usual)

DELL Technologies

The Rise Of The Modern Model

The Expansion & Control Of Small Language Models and Others…

OUR CORE BELIEFS

It is more efficient to bring AI models **closer to where the data lives**

The scale of **LLMs continues to grow**

SLMs, general-purpose models, domain-specific and open-source **models differ in their applications, for use-case requirements and outcomes**

We believe **open-source models** create equal opportunity across the tech ecosystem and support the creation of AI breakthroughs

AI models continue to expand and evolve – there is still **"no one-size-fits-all"**

Not every GenAI use case will require a **large infrastructure investment with GPUs**

On-prem Private Models will provide more control and be more cost effective

The unique advantage of running smaller models on prem or on PCs allows orgs to **drive greater specificity in fine tuning**

GenAI is now extending to join "Statistical" AI at the PC experience

INFORM

OUR POINT OF VIEW

While we're still in the early innings, our customers are making their AI investments now. Customers across the board **want to keep critical data and IP on-prem** to minimize risk and maintains control over their Data, their Information and Paterns.

Dell is well positioned for the deployment of AI models of all sizes with our end-to-end portfolio of AI infrastructure and solutions - through our services, servers, storage, networking, data protection and PCs capabilities.

We continue to invest heavily in Research in AI and our focus is on democratized and equitable AI for Business, Orgaisations and institutions, as well as Government and Societal communities in Ireland and around the World

Research and industry innovation Centre in Ireland

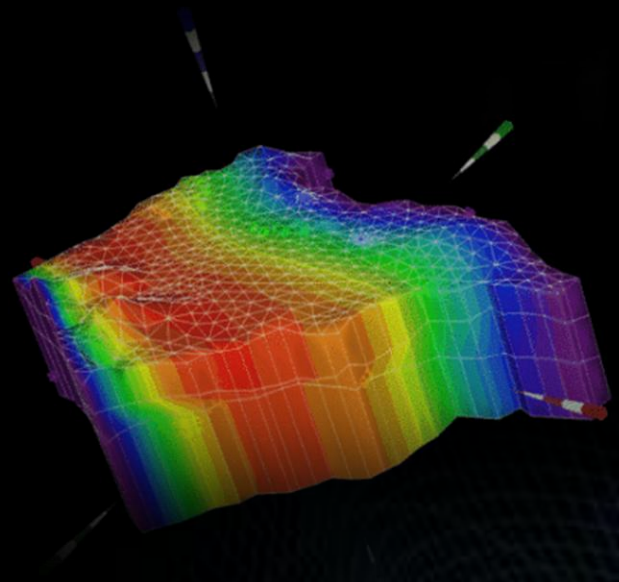
DELL Technologies

The Rise Of SLM

The Expansion & Control Of Small Language Models

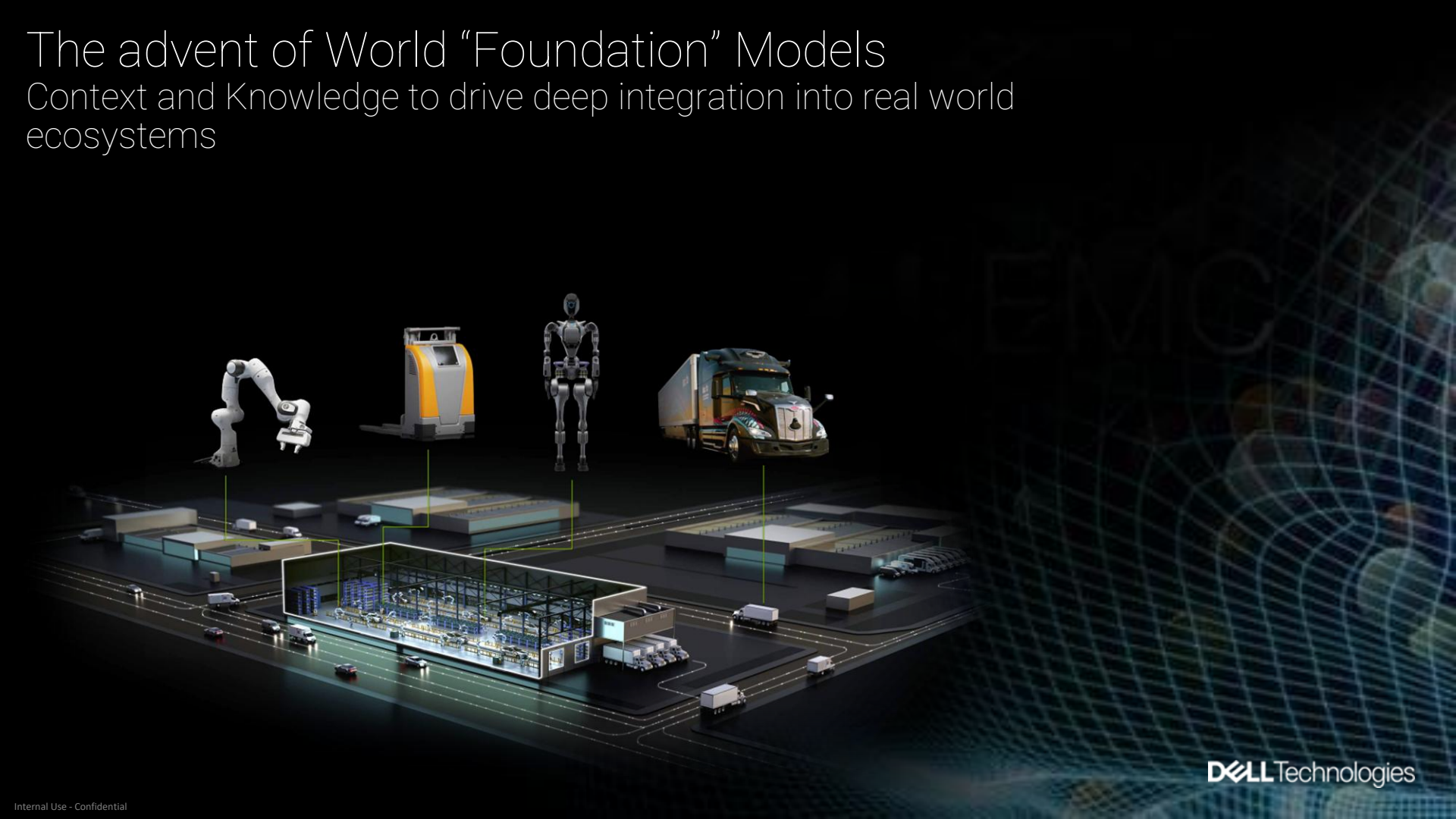
Numerical Models

The resurgence of well defined models



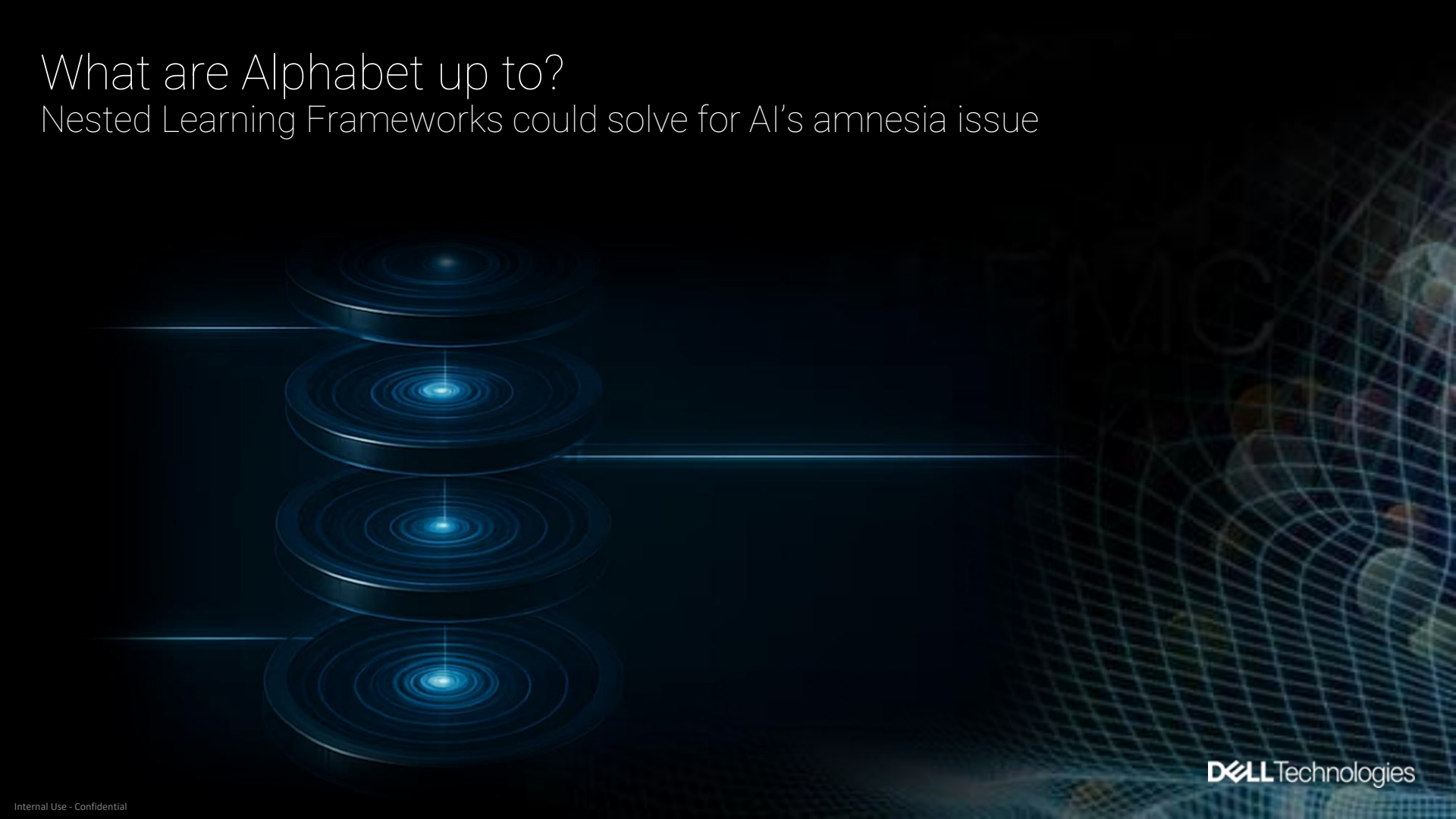
The advent of World “Foundation” Models

Context and Knowledge to drive deep integration into real world ecosystems



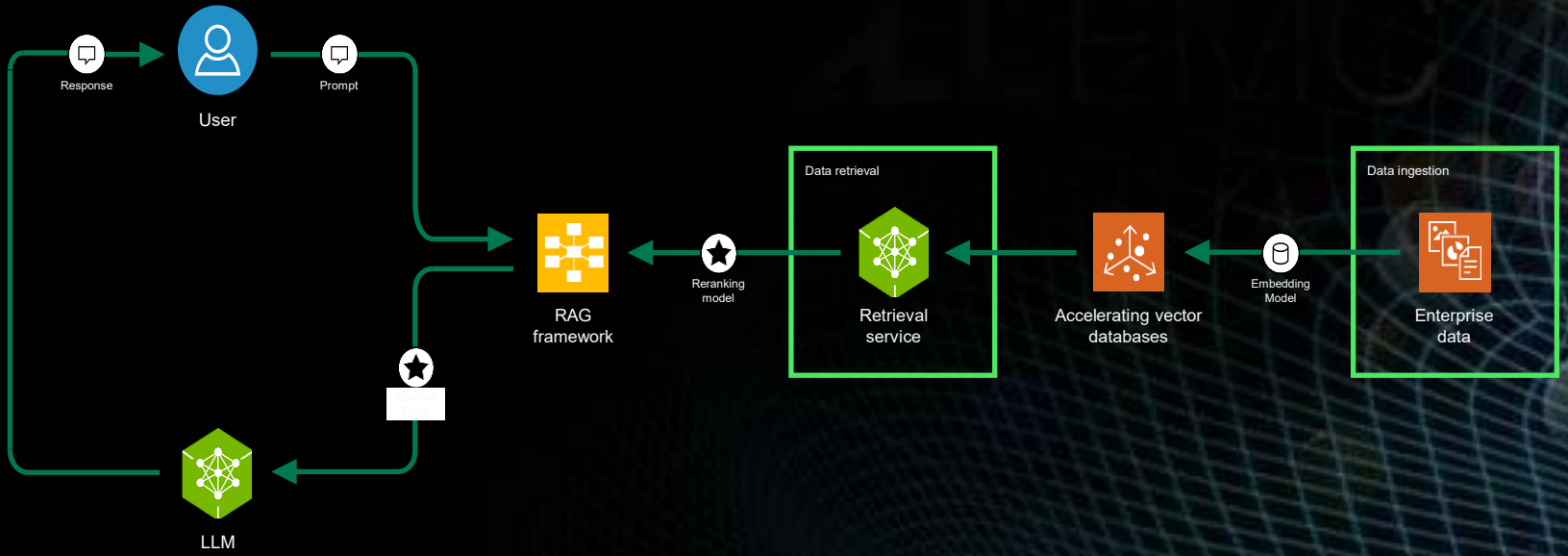
What are Alphabet up to?

Nested Learning Frameworks could solve for AI’s amnesia issue



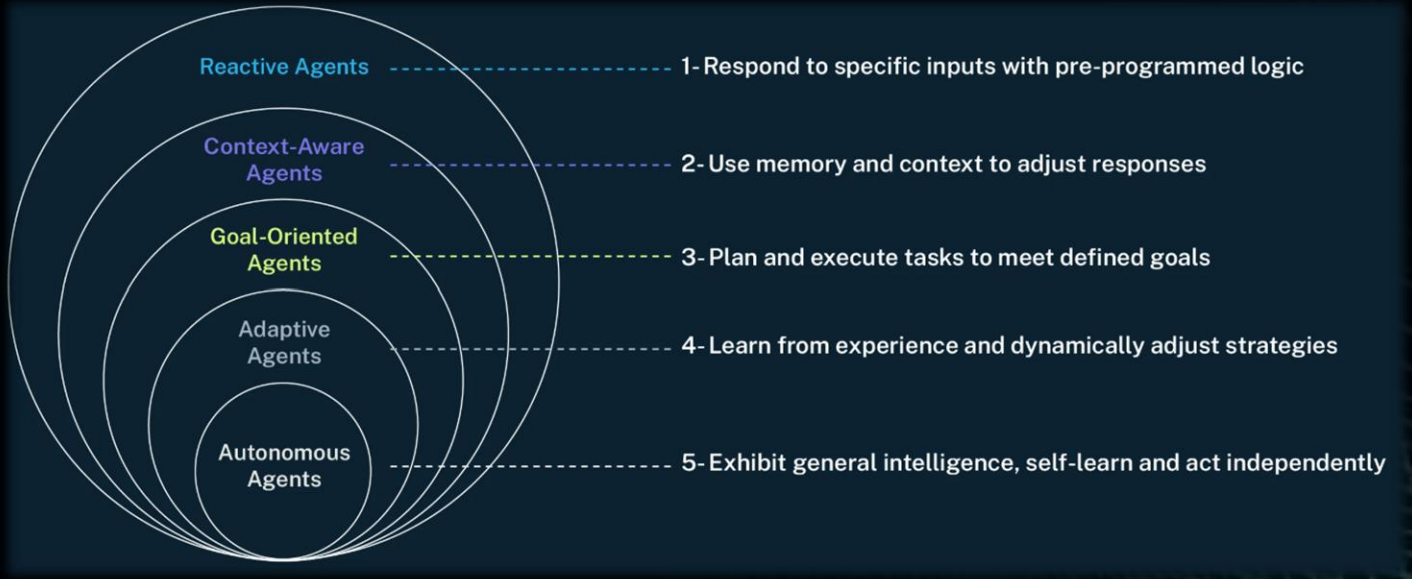
Moving Beyond The RAG Process Approach

Enriching The Domain Knowledge Of Your Model's and Function's



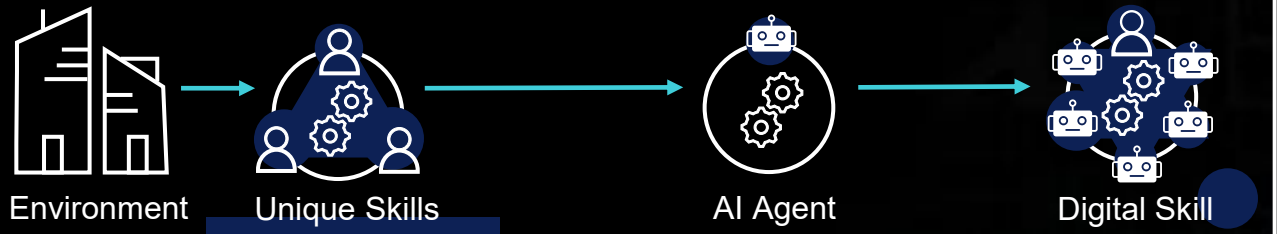
What are AI Agents

Machine entities that act on and interact with their environment autonomously



Introducing Digital Skills From Office of the CTO

Bringing Enterprise Advanced Skills into the AI Cycle



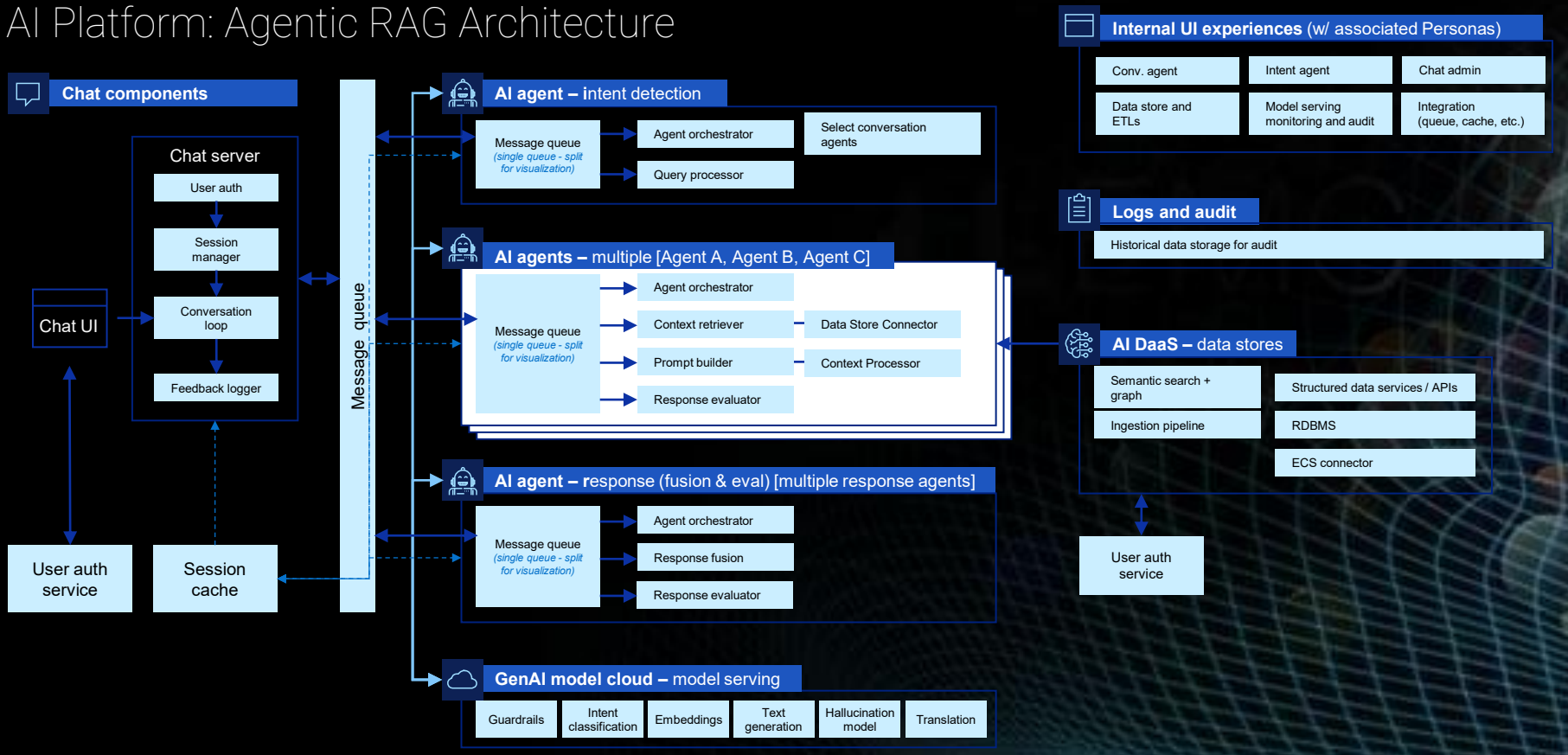
Digital Skills excel in three key areas:

1.Seamless execution,
2. collaborative problem-solving,
3. continuous adaptation through learning

=

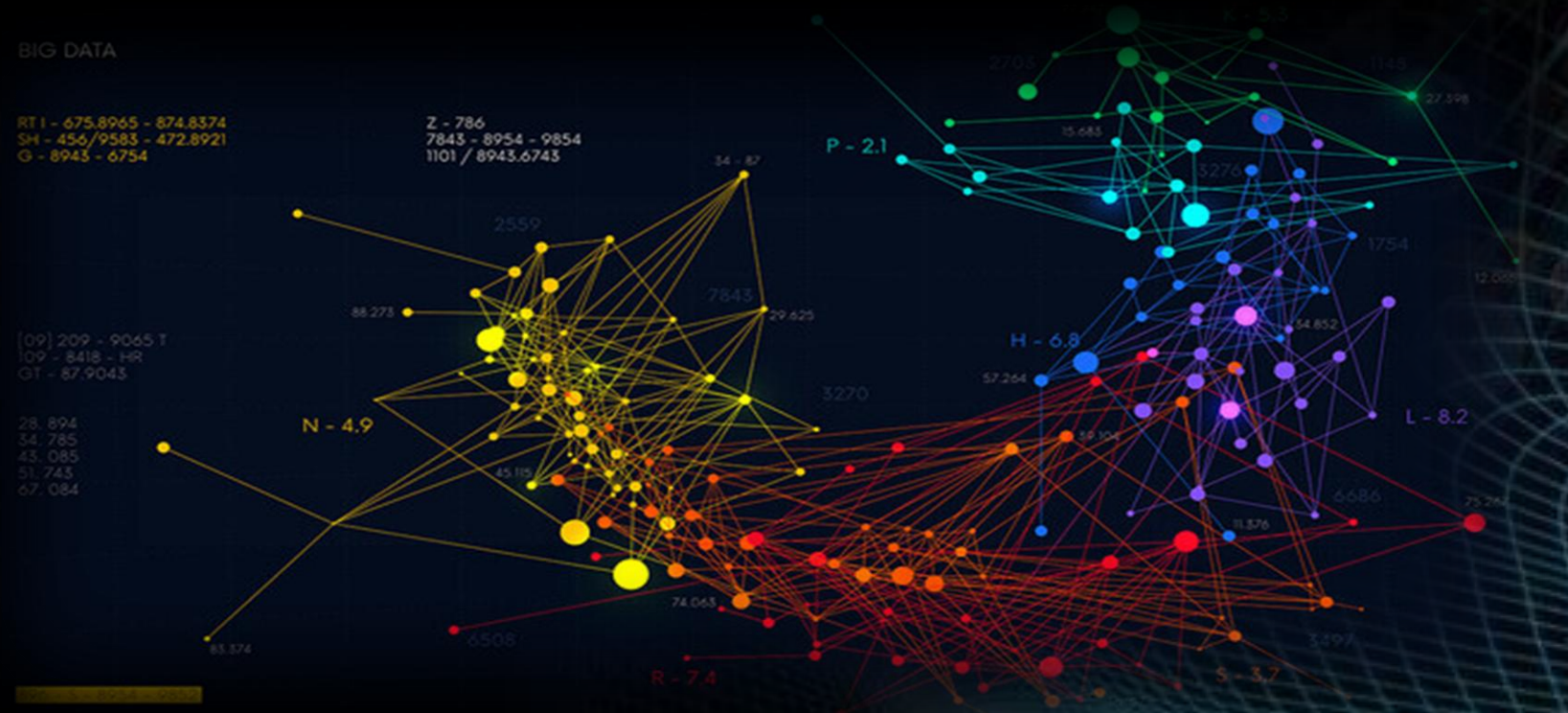
Architecture details

AI Platform: Agentic RAG Architecture



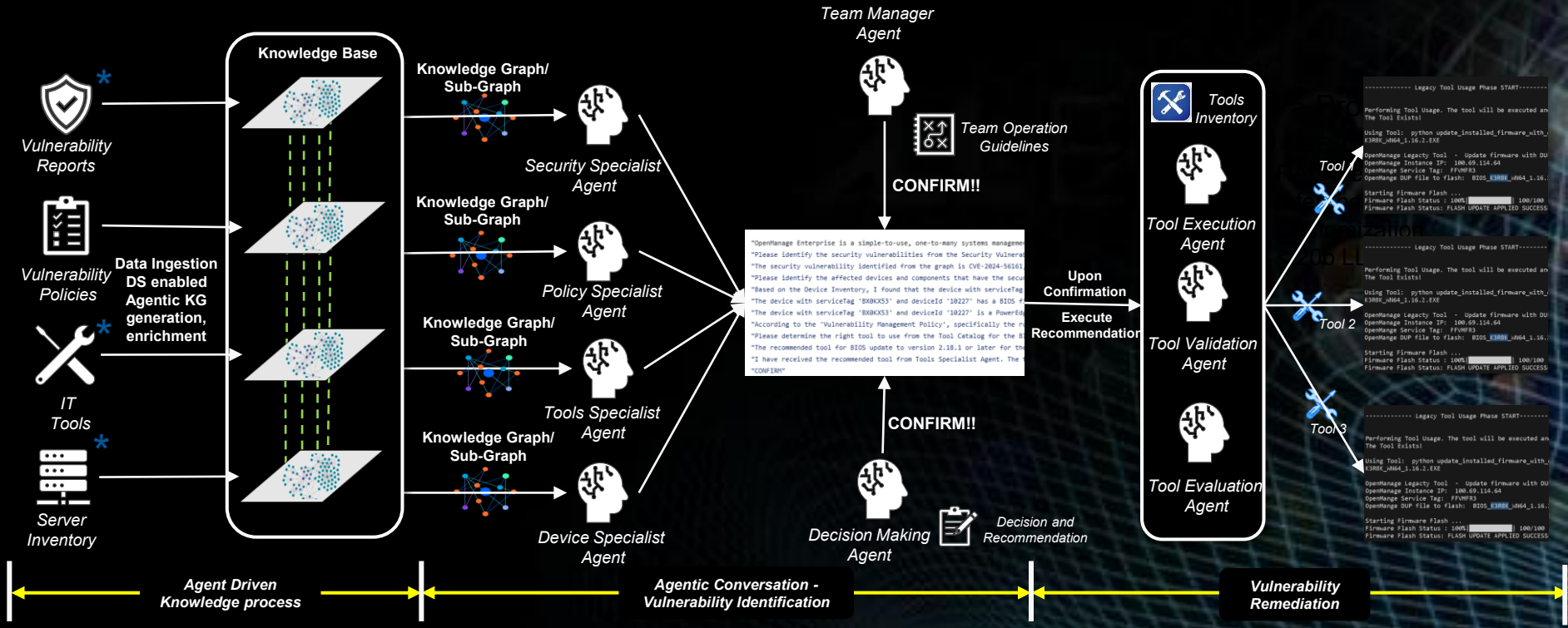
knowledge Graphs

Architecture Is Based on Relationship V's "Similarities"



AI Agents and Knowledge Platform

Agentitic Architecture Identifying & Remediating Autonomously

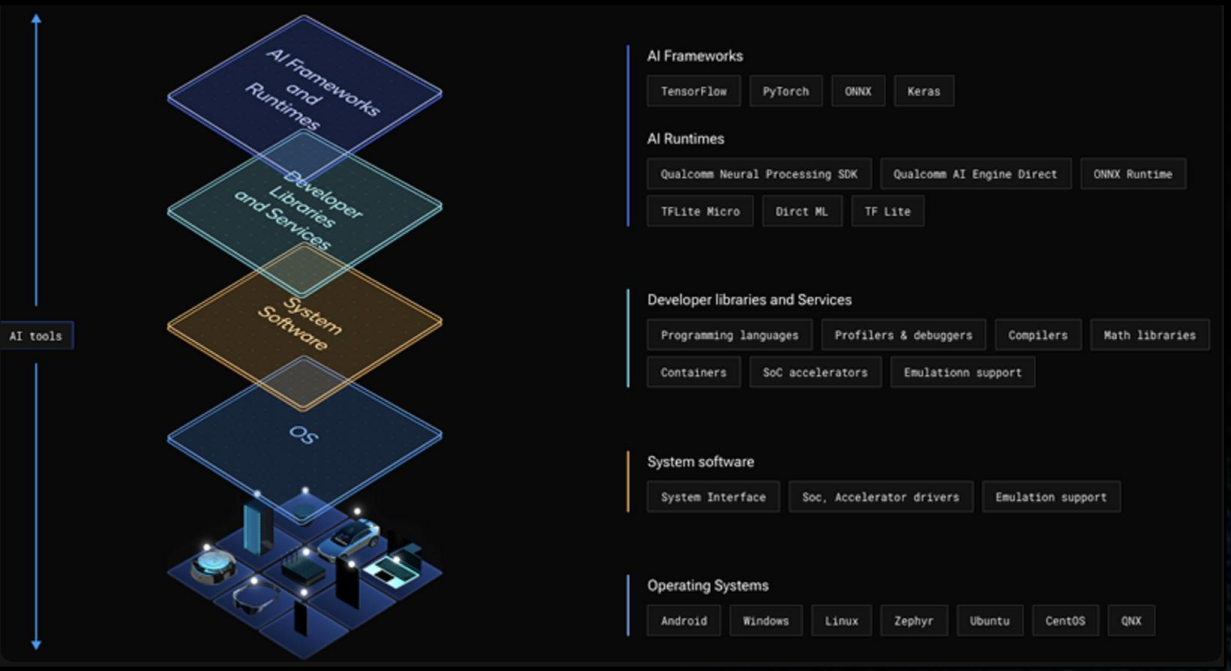


* Vulnerability Reports to "https://www.dell.com/support/security/en-us/"
Legacy IT Tools to "https://github.com/dell/OpenManage-Enterprise/tree/main/Python"
Server Inventory to "https://100.69.114.64/core/console/console.html#/"

DELL Technologies

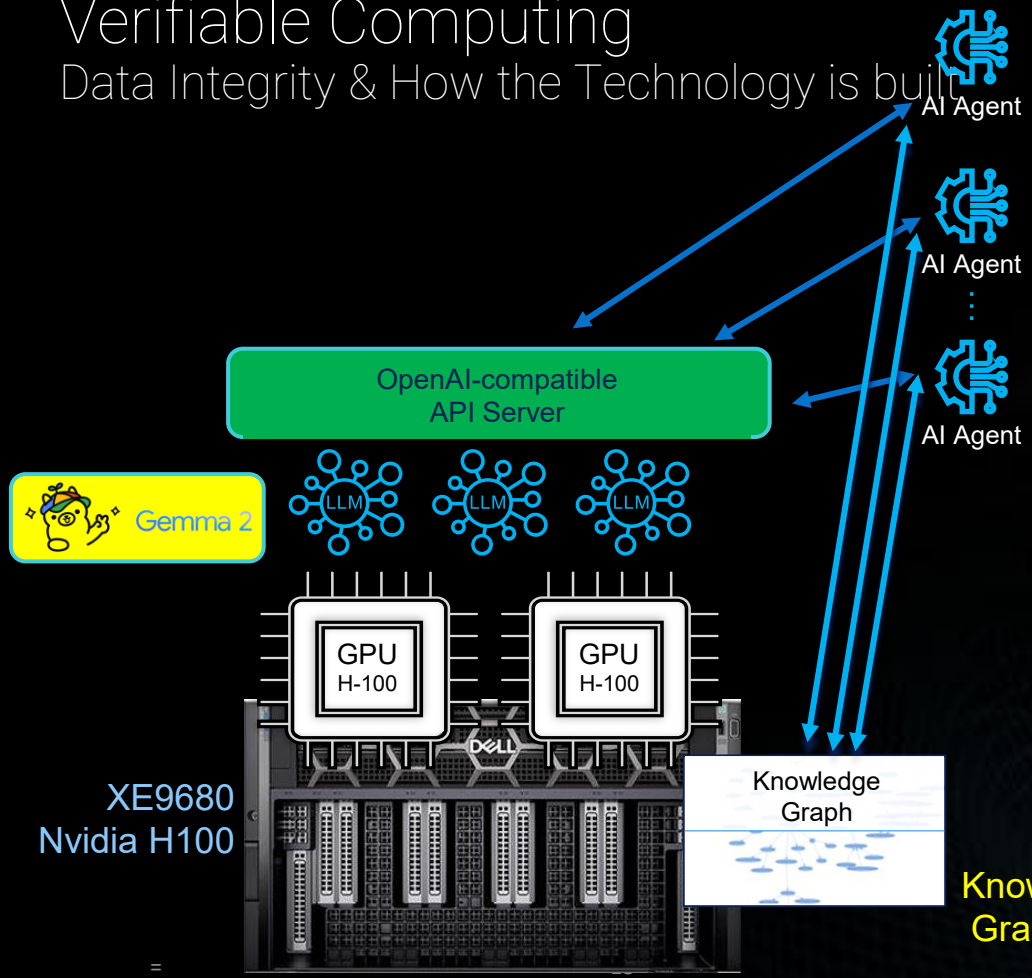
Platform For Your Models

AI Factory Full Stack Platform



Verifiable Computing

Data Integrity & How the Technology is built



Precision 5580 Laptop

LangGraph

- Agent Host leveraging a modified version of LangGraph
- Reasoning Engine, Gemma 2 70B LLM on a H100/XE9680
- Shared Memory framework form LangGraph + additional Knowledge Graph on central Dell Server to enhance reasoning through contextual relationships
- Zero Trust architecture capability
- Verifiable Compute for Trusted and confidential compute capability

=

EQTYLAB

intel. NVIDIA.

Introduction

Whitepaper
November 2024

Verifiable Compute

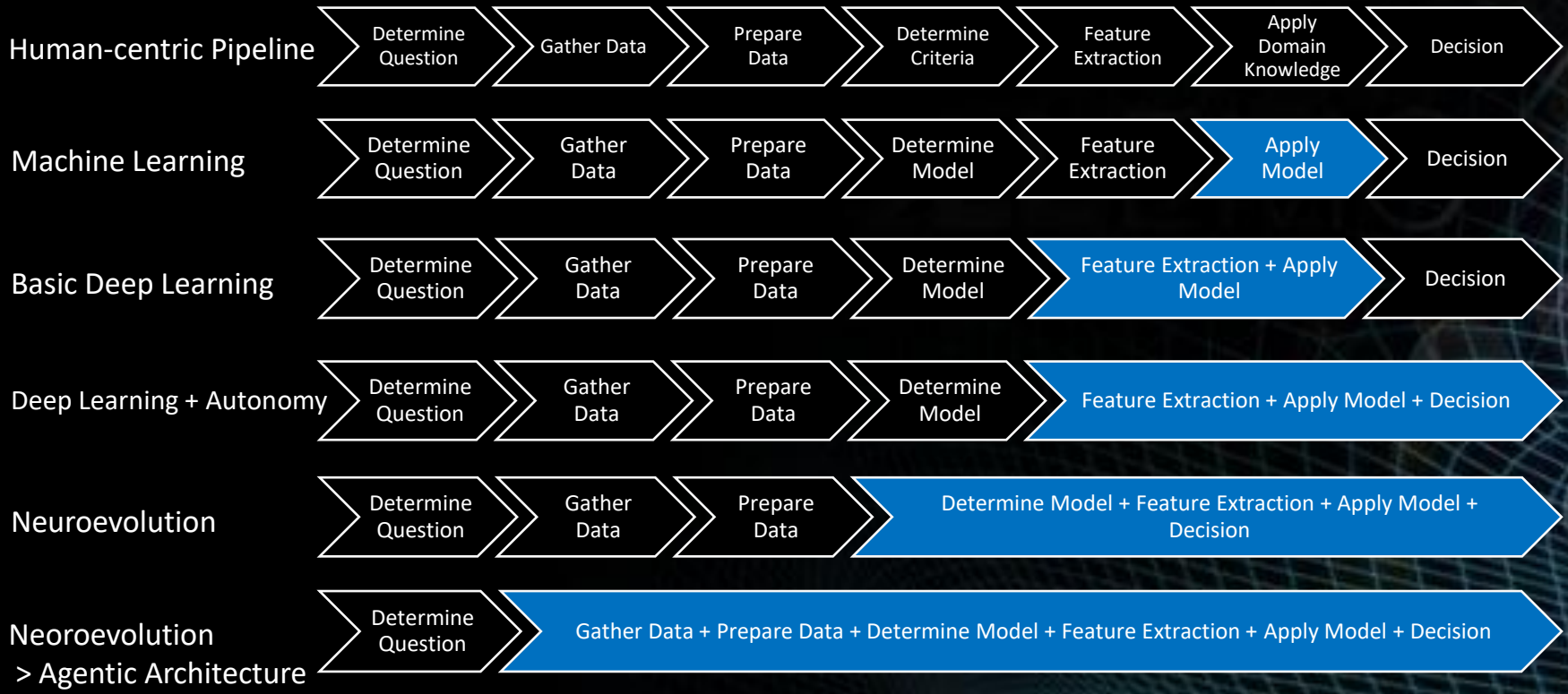
VERIFY TO TRUST, AI

Computing Ready for the Agentic AI Era

Demo Time
Agentic Confidential Computing Solution

https://d2t401rg7d25b1.cloudfront.net

The ever-expanding role of the computer



Looking forward

- Many artificial intelligent techniques (some of which are based on nature) will join neural networks in the limelight
- The joining of these approaches (as with neuroevolution) will allow computers to take over greater fractions of the human work
  - @Edge – AI – Tasks>Thinking Tasks
  - Robotics & Humanoids
  - Full Autonomy with Human On The Loop
- The computational power required to execute these algorithms in reasonable time will continue to increase
  - Scaling will be more important than ever
- AI and Quantum Technologies will Harmonize in the years to come

Thank you!

@reddogmarc

@DellTech-CTO Office EMEA