Cisco Webex App

Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click "Join the Discussion"
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

**Webex spaces will be moderated
by the speaker until June 7, 2024.**

Agenda

- Looking Back: the Journey to the current AI Era

- Cross-Architecture Considerations
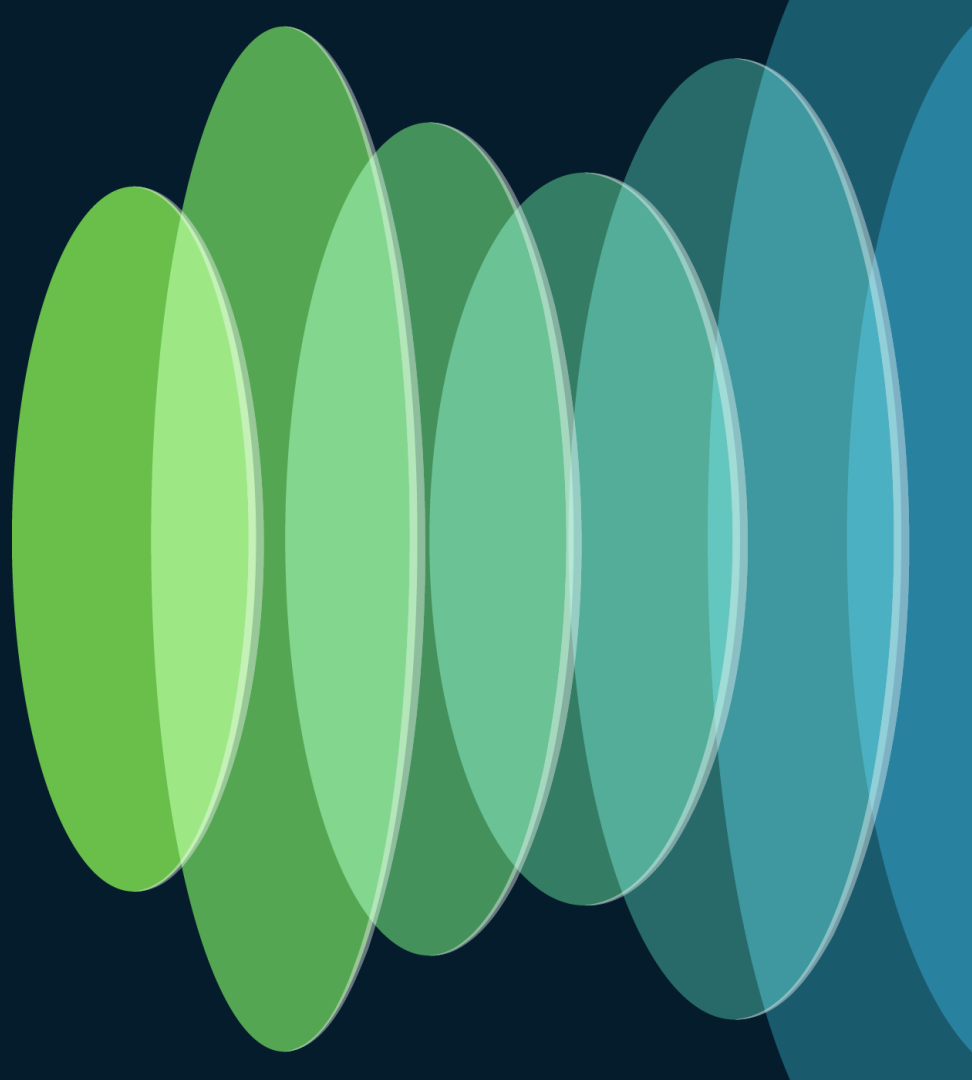
- AI Readiness

- Cisco for AI

Learning Objectives

By the end of this session, you'll be able to:

- Approach your AI strategy with key considerations in mind

- Describe cross architecture AI Readiness

- Understand Cisco's strategy for AI

The AI Era – How did we get here?

1940s	1950	1965	1997	2017	2022+
Math, neurophysiology and computing research	Turing Test	Moore's Law	IBM Deep Blue defeats Kasparov in chess	"Attention is all you need" Transformer architecture, Cambrian Explosion of LLMs	Consumer-facing LLMs

AI

IBM 701 computer plays checkers

Computational models

Machine Learning

Growth in availability of data

Probability models

Statistical analysis techniques

Deep Learning

Artificial Neural Networks

Natural Language Processing

Image recognition

Google Translate

Gen AI

BloombergGPT

ChatGPT

Claude

Foundational Models

Domain-specific Models

Code Generation

Image Generation

What *really* changed?

Emergence and availability of consumer and enterprise grade models

Scientific research realm

Consumer and business realm
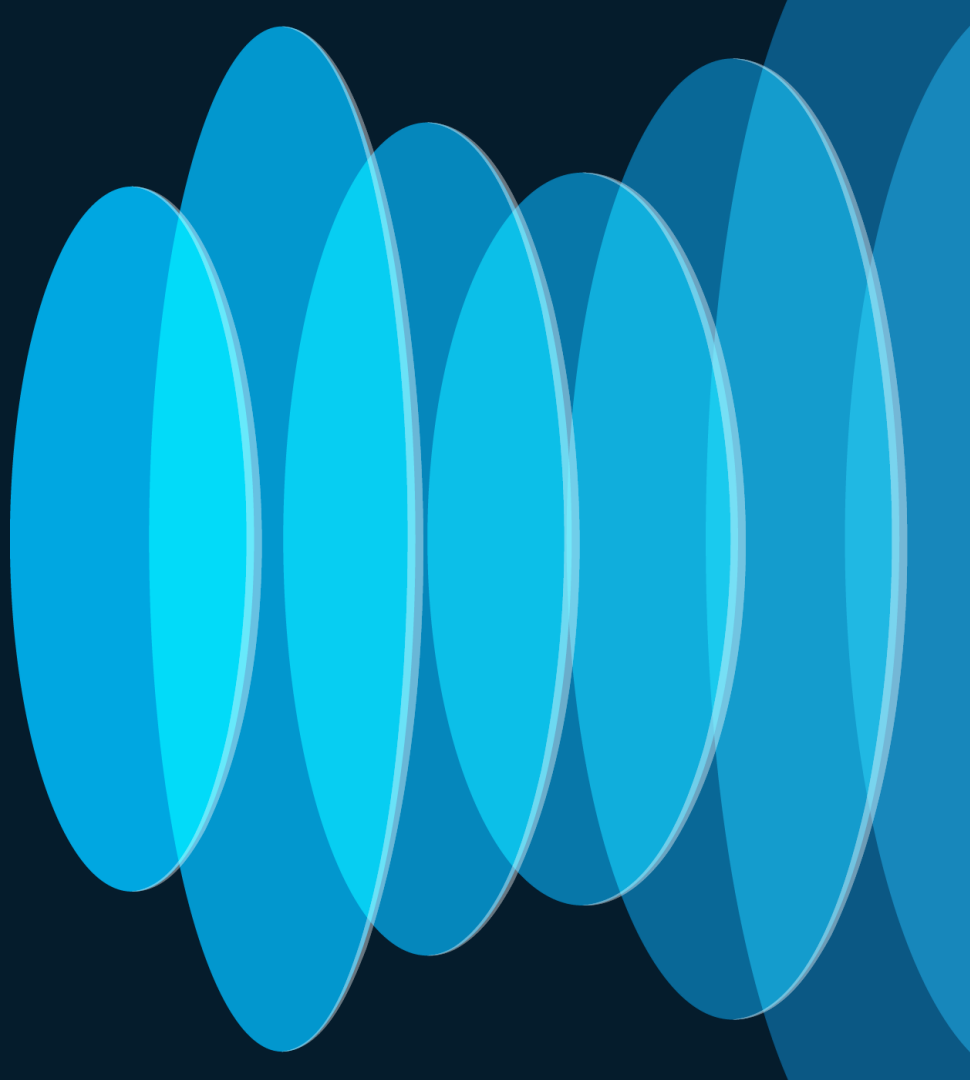
Challenges for Organizations

Data Governance | Scalable Systems | Risk Management

Cross Architecture Considerations

Differentiate Features from AI vs Systems for AI

Features from AI

Background noise removal

Self-service tools

Predictive analytics in t-shoot
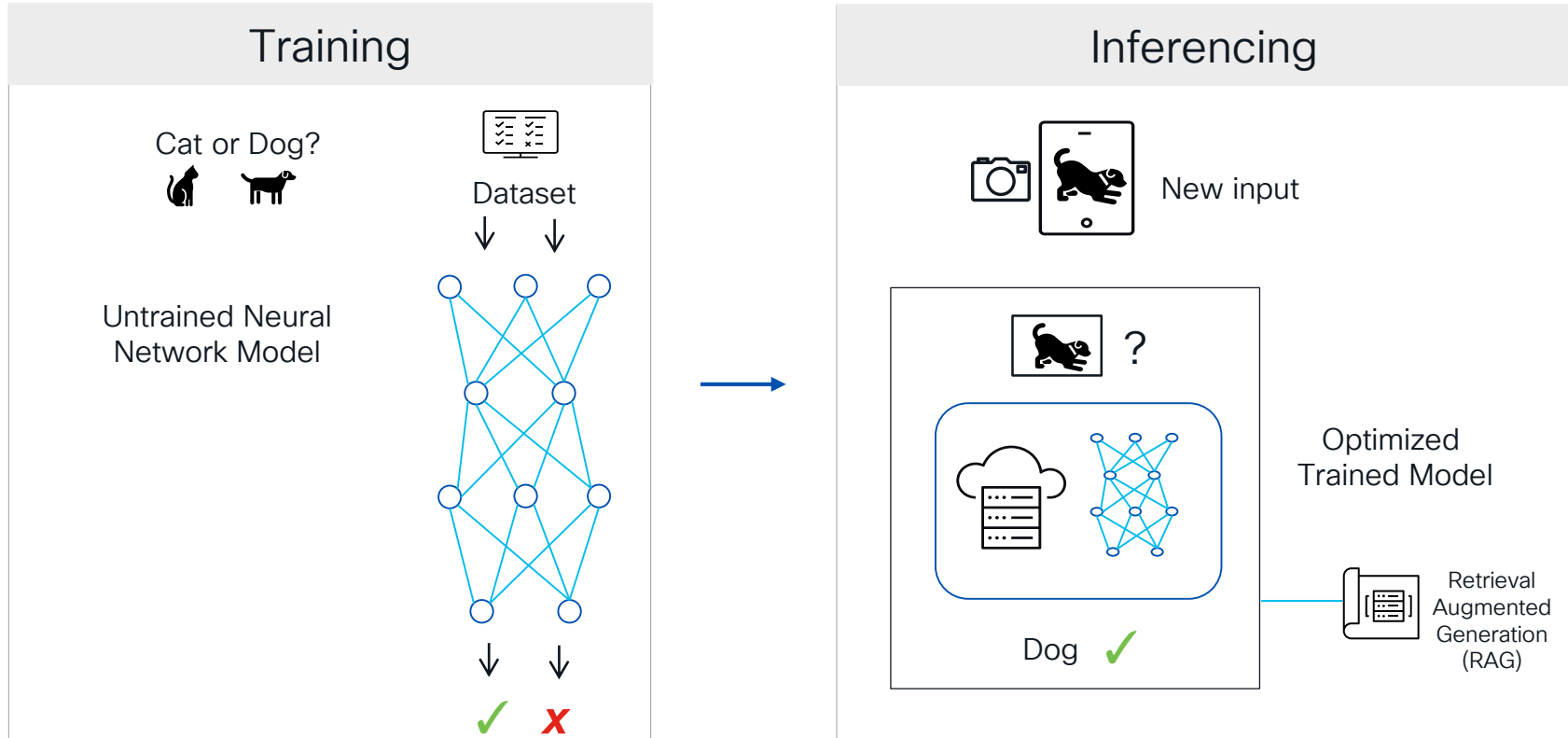
AI assistants

Typically delivered via SaaS

Systems for AI

Deep Neural Network model design

Back-end and front-end networks

Data storage

Deep Learning frameworks and SDKs
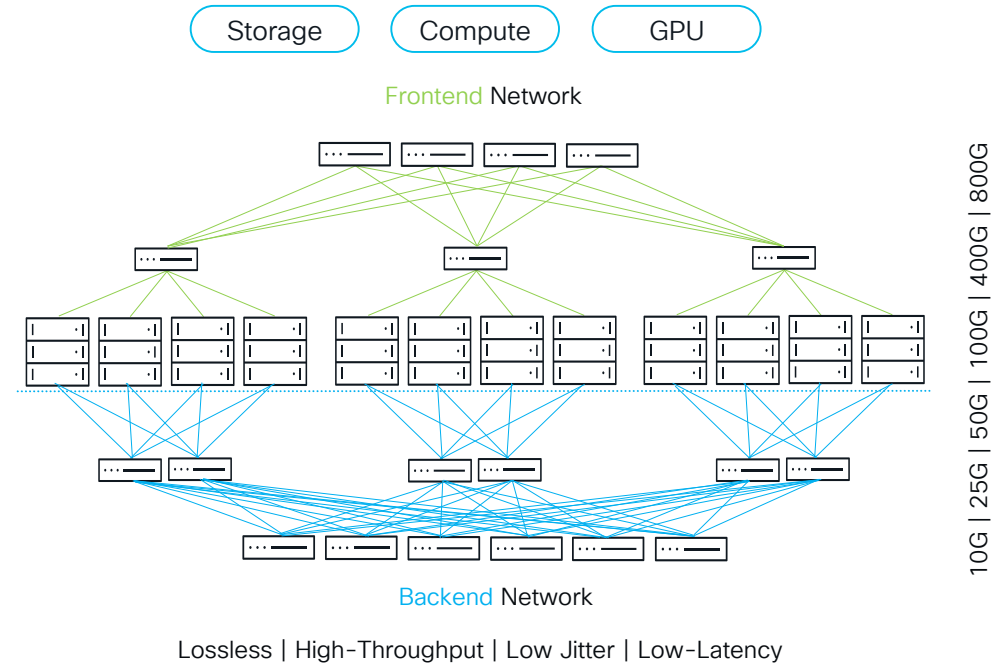
On-prem, Cloud

What does a system for AI entail?

Training — Cat or Dog? Dataset → Untrained Neural Network Model → ✓ ✗

Inferencing — New input → ? Optimized Trained Model → Dog ✓ — Retrieval Augmented Generation (RAG)

What does infrastructure for AI entail?

- High throughput, lossless fabric for training

- Can require extensive power and cooling

- Training and Inference can differ in compute density and utilization

Storage Compute GPU

Frontend Network

10G | 25G | 50G | 100G | 400G | 800G

Backend Network

Lossless | High-Throughput | Low Jitter | Low-Latency

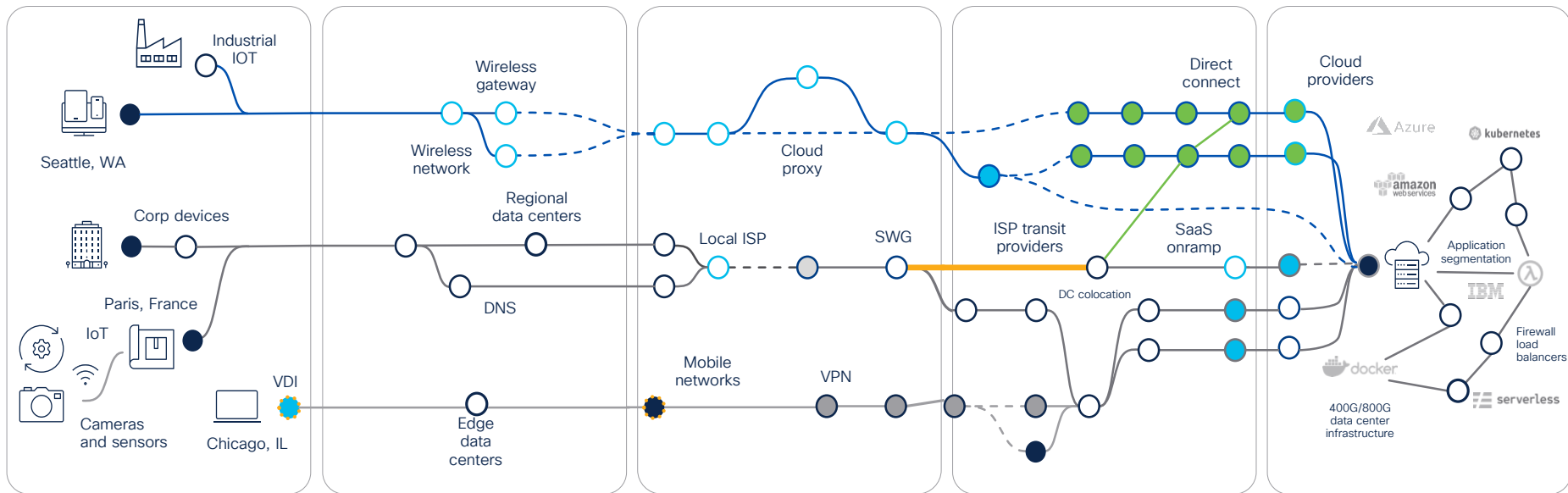It's all about the **data**

It's a cross-architecture endeavor...

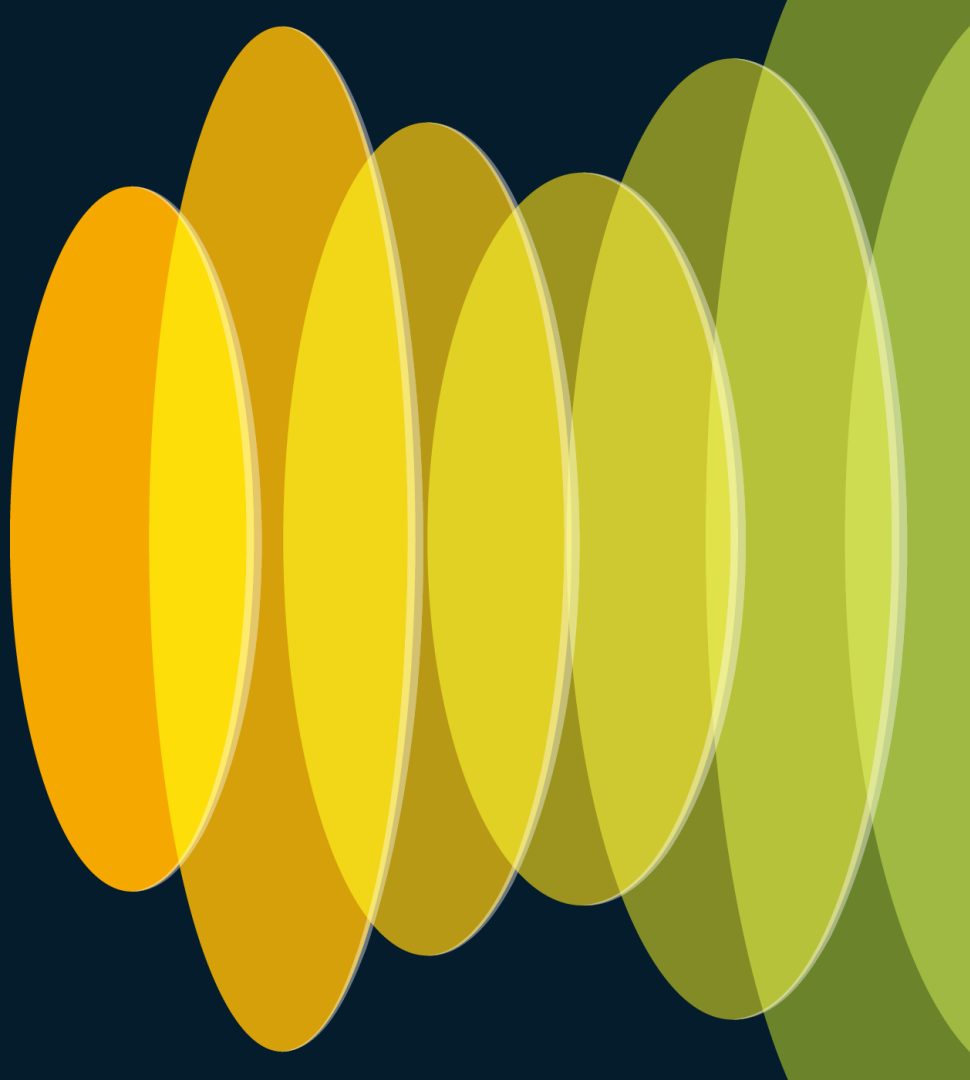People, places and things · Access networks · Network services · Cloud connectivity infrastructure · Apps (data center, cloud)

Use Cases Across Industries

Finance	Healthcare	MFG	Automotive	Public sector	Retail

Content Creation | AI-Enabled Customer Service
Drug Discovery | Virtual Try-Ons | Automated Robotics | Fraud Detection

Large →

→ **Small**

Cisco Deployment Templates

"Journey to AI" Guide	Cisco Validated Designs	Industry Partnerships

Training and fine-tuning

Inferencing and Retrieval Augmented Generation (RAG)

Orchestration, Observability, Policy Enforcement Tools

Models and Frameworks | Software for AI

Pre-Trained Foundational Models | Domain-specific Models | Model Repositories

Accelerators | AI Infrastructure | Scalable Systems

AI-ready infrastructure from the data center to the edge

Cisco® UCS	Nexus® switches

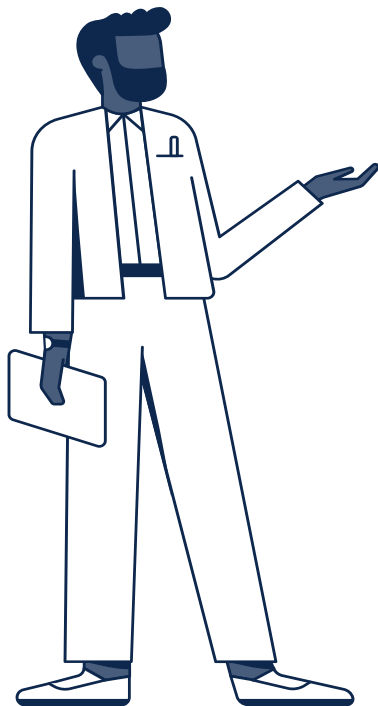**AI specific** →

→ **General purpose**

How can I start?

Initial Questions to Define Cross Architecture Requirements

- What is the use case?

- Will I need to Train? Inference? Augment?

- What is my model size?

- How will I manage my compute resource utilization?

- What security policies will I need to implement?
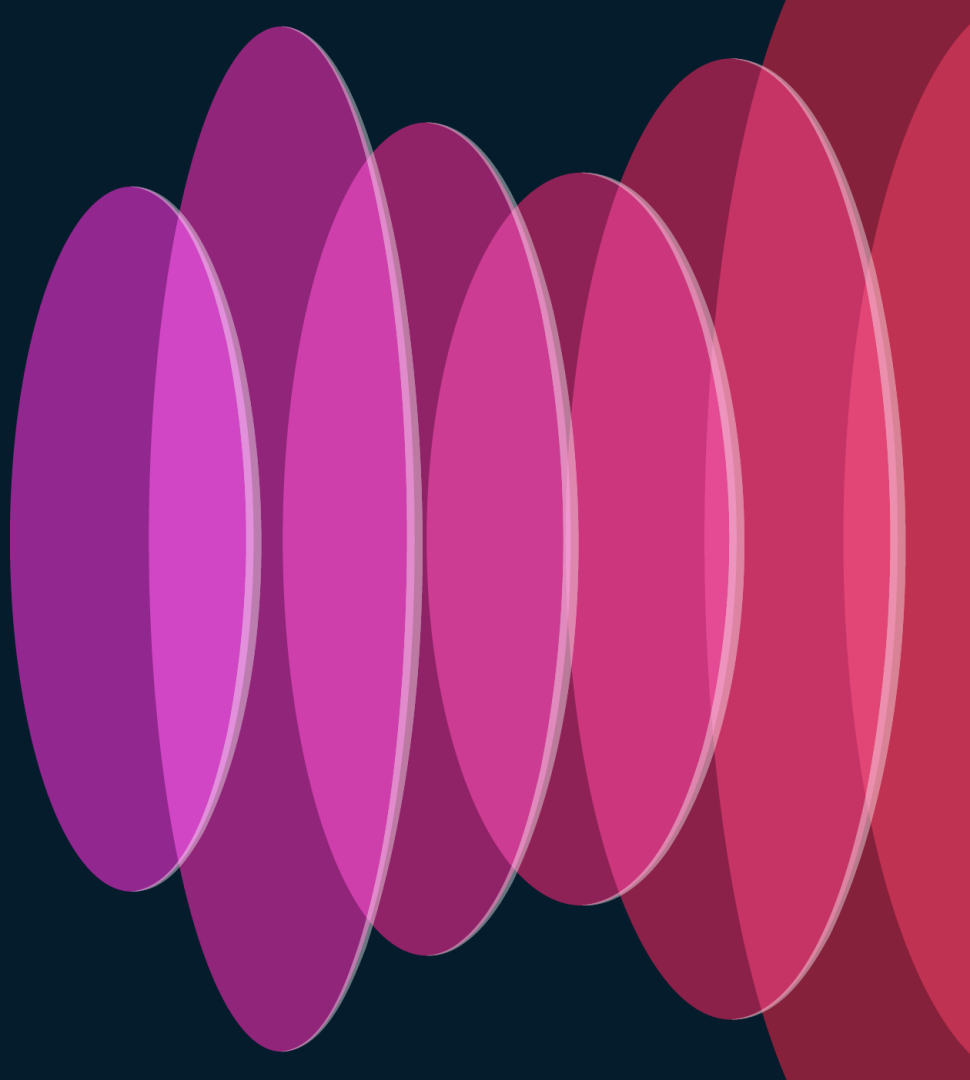
- Will I need to train my staff?

Considerations for AI-Readiness

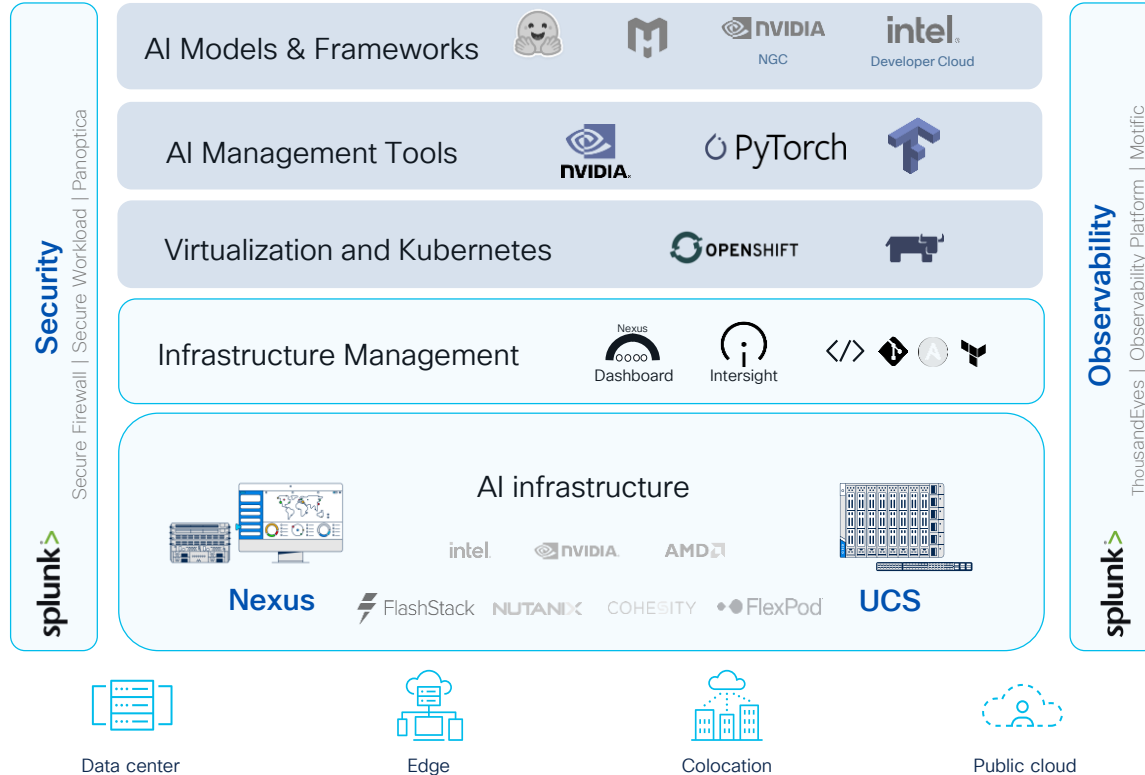*From the Cisco "[Journey to AI](#)" guide*

- Form an AI committee and establish clear ownership

- Project the required increase in GPUs

- Ensure hygiene, quality, and reliability of in-house and external data sources

- Implement mechanisms to detect bias and checks on AI algorithms

- Establish proficiency for responsible usage of AI

Cisco for AI

The Cisco AI Stack

AI Models & Frameworks

NGC

intel Developer Cloud

AI Management Tools

NVIDIA

PyTorch

Virtualization and Kubernetes

OPENSHIFT

Infrastructure Management

Nexus Dashboard

Intersight

AI infrastructure

intel NVIDIA AMD

Nexus

FlashStack NUTANIX COHESITY FlexPod

UCS

Security — Secure Firewall | Secure Workload | Panoptica

splunk>

Observability — ThousandEyes | Observability Platform | Motific

splunk>

Data center

Edge

Colocation

Public cloud

Cisco's AI approach

Using AI to drive operational efficiency for **better experiences**

AI-Native IT operations

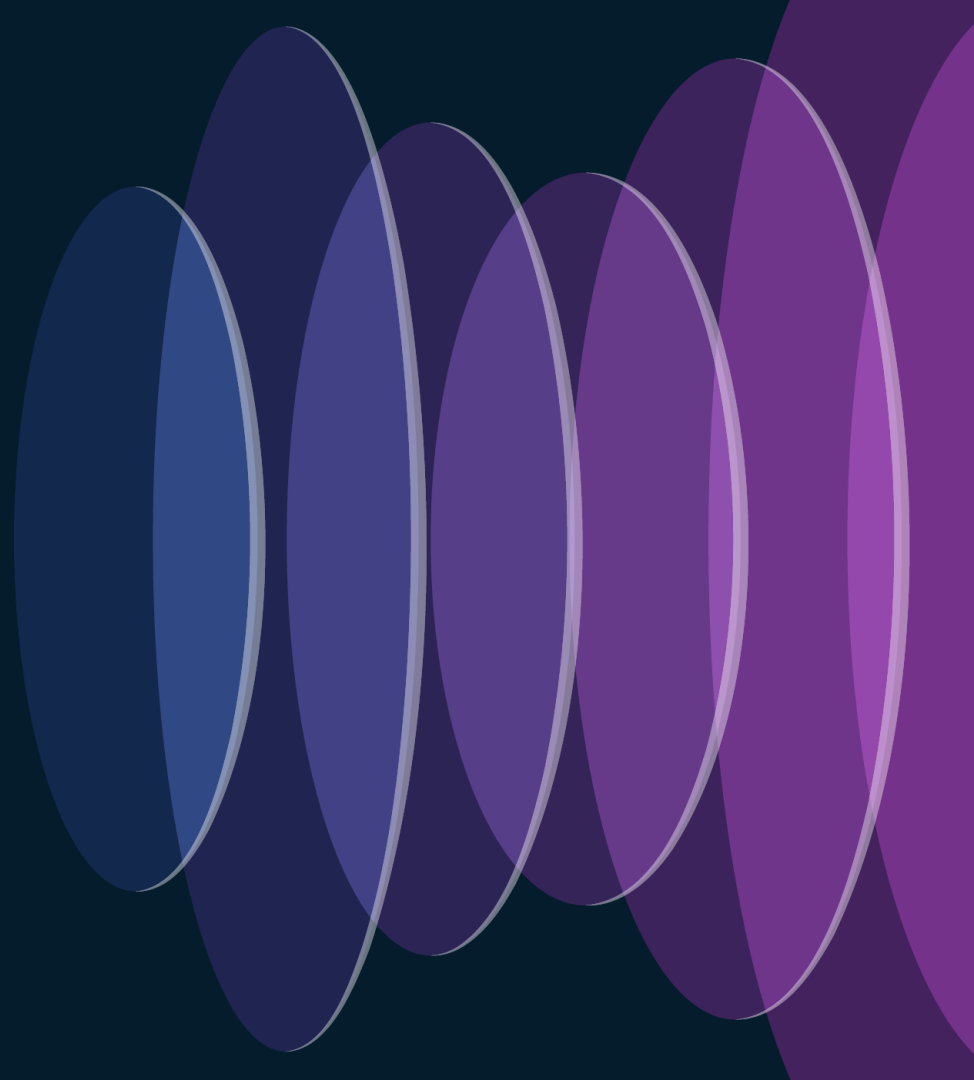Networking | Assurance | Collaboration Security | Responsible AI

Enabling **Your infrastructure** to support adoption of AI applications
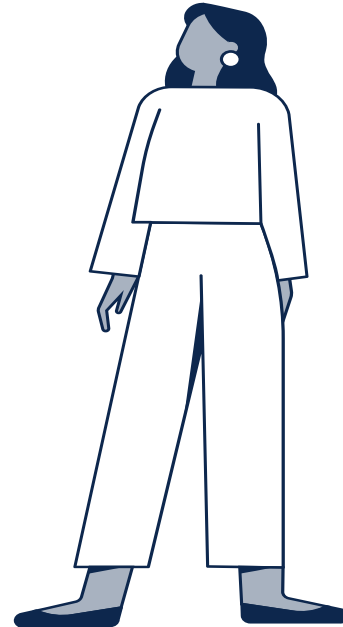
AI-Native infrastructure

Data Center Networking | Compute | Storage | Security | Trusted AI

What should I do next?

CISCO *Live!*

Next Steps

- Determine your AI readiness across your environment

- Continue building your AI literacy

- Develop a cross architecture roadmap for AI

AIHUB-1009

Continue your education

- Visit the Cisco Showcase for related demos

- Book your one-on-one Meet the Engineer meeting

- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs

- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Contact me at: dquiroga@cisco.com