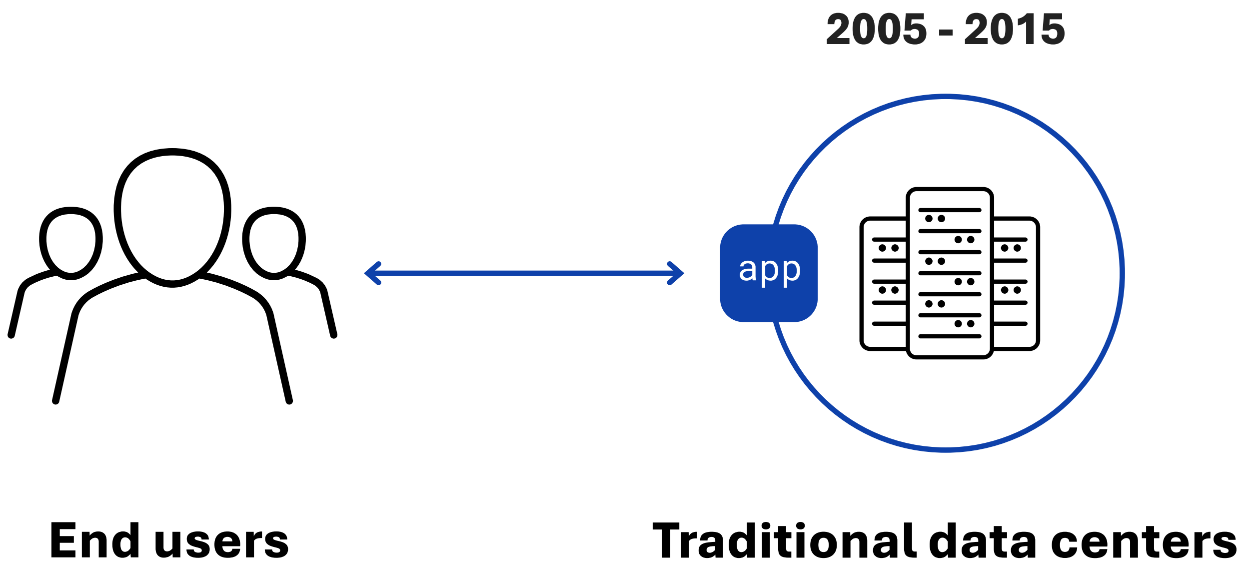
Securing AI in an Evolving Landscape: The Case for an AI Gateway

Michael Christie
Sr Dir, Solutions Engineering
March 2024

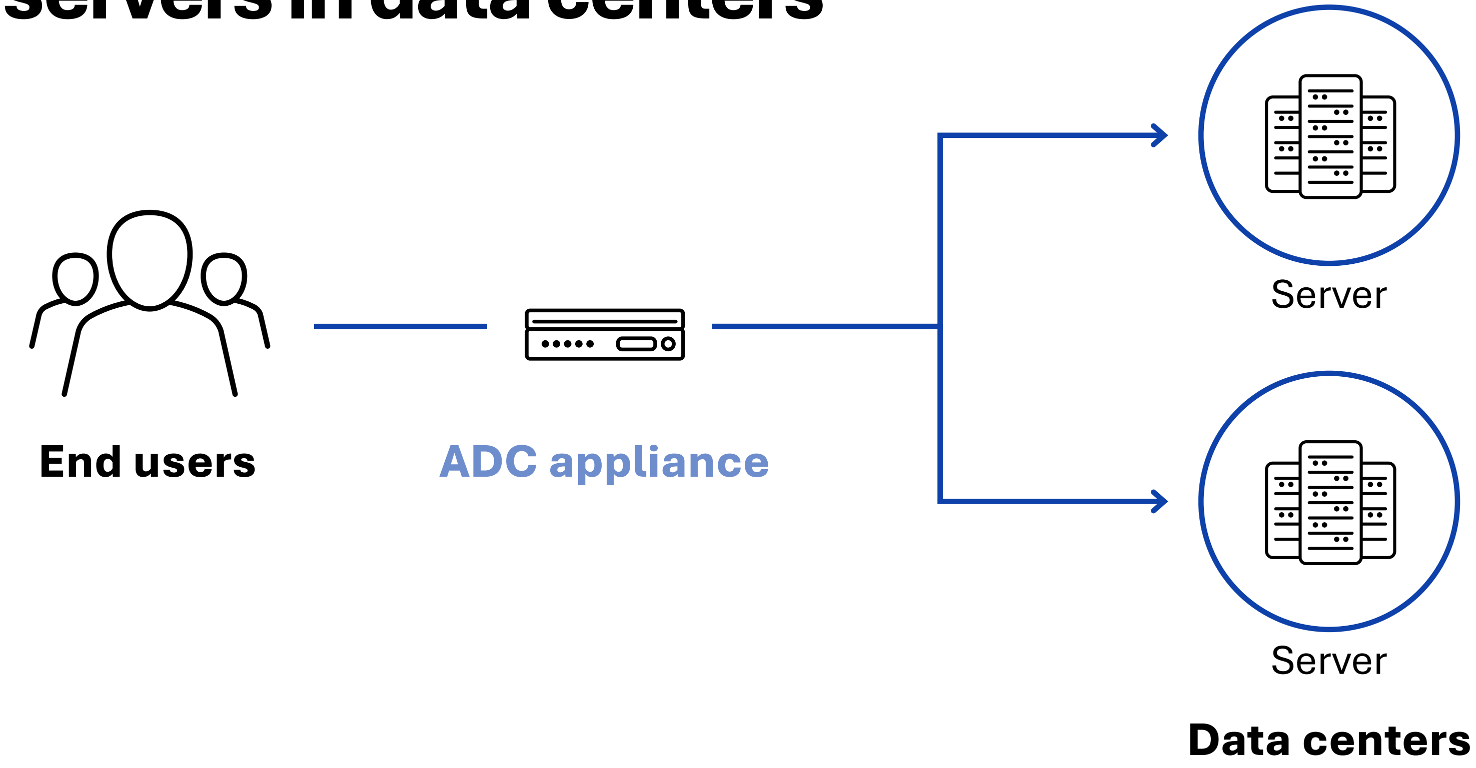
The deployment of applications used to be straight forward



ADC appliance

2005 - 2015
Data center era

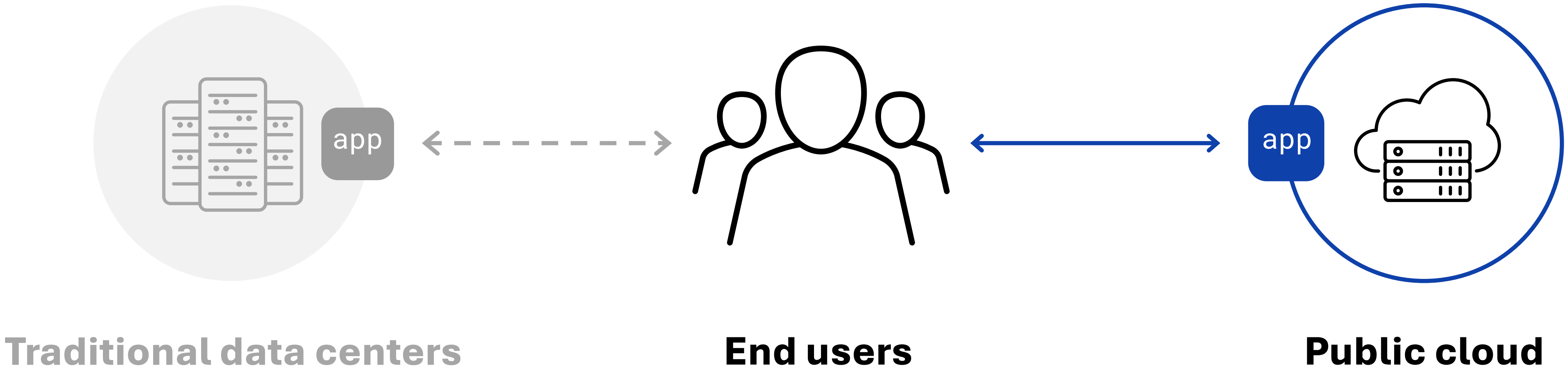
The Application Delivery Controller (ADC) became the primary means of routing application traffic between users and servers in data centers



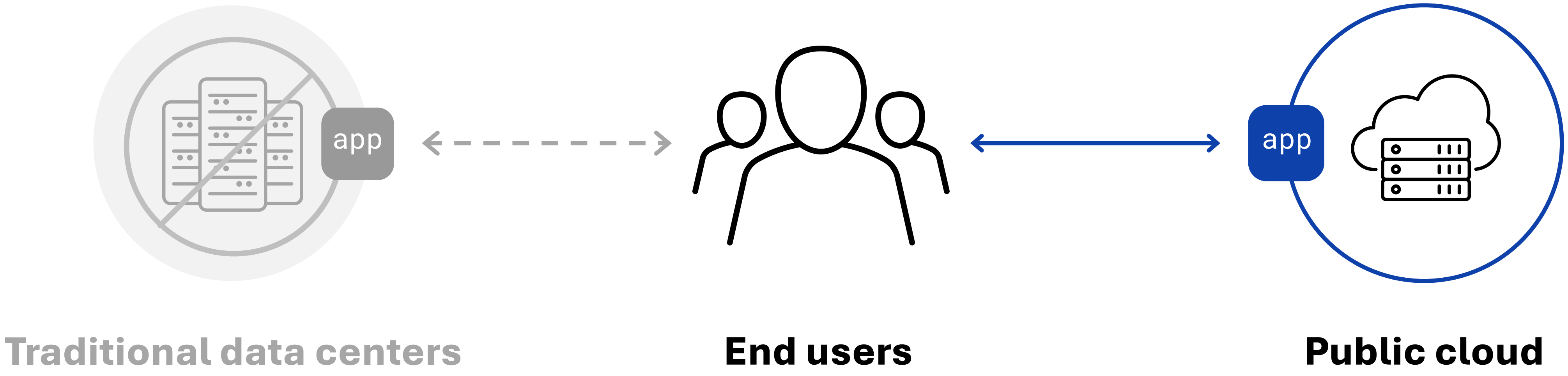
Capabilities

- Load balancing and traffic acceleration
- Web application security
- Access management

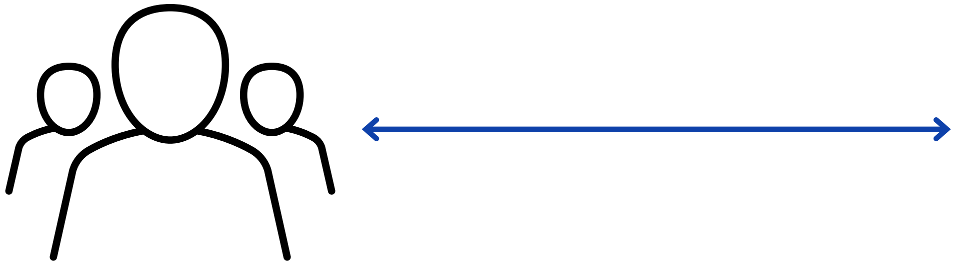
However, enterprises started moving applications to the cloud

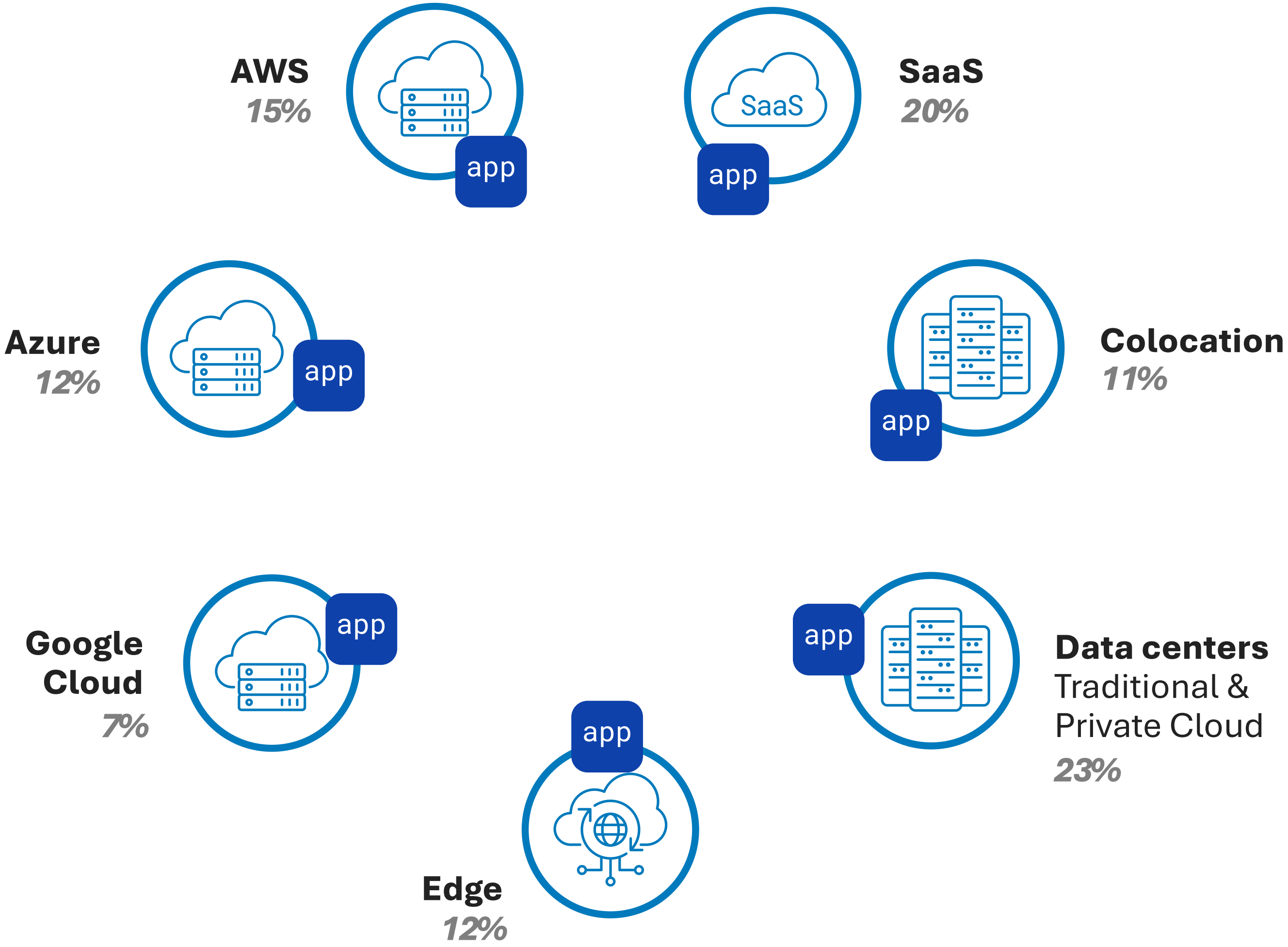


It was commonly assumed that apps would move to a single public cloud



In reality, enterprises deploy their apps in multiple environments



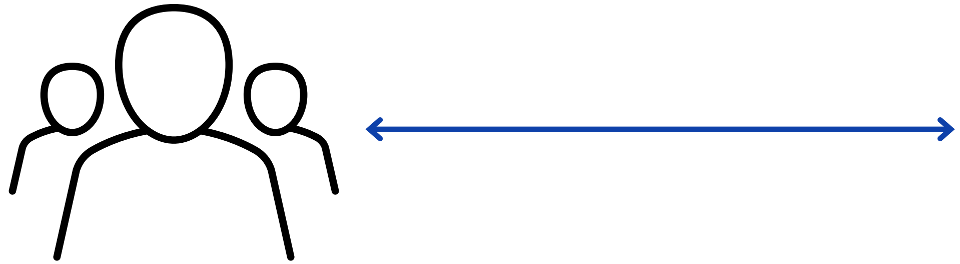


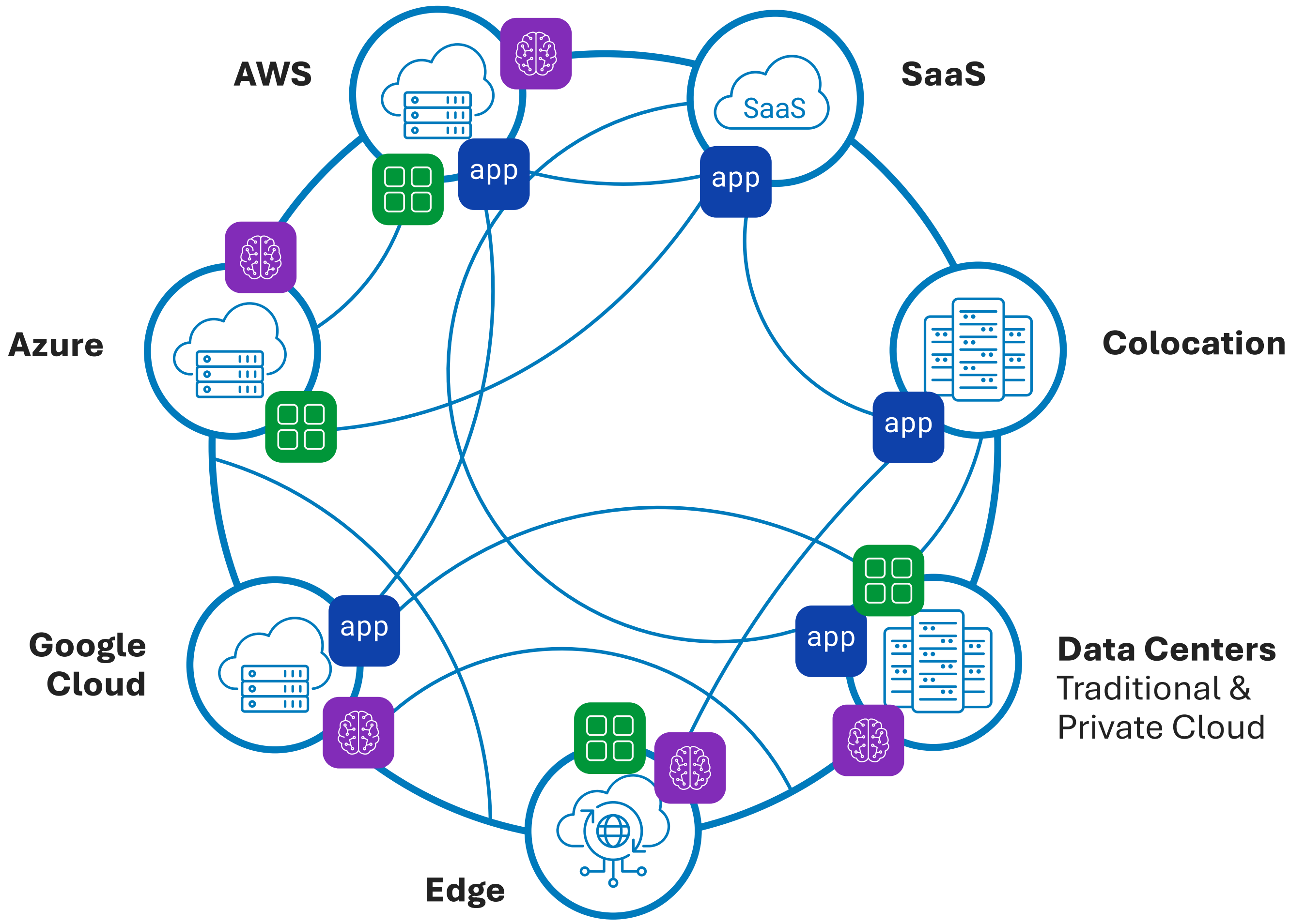
9/10

large organizations have
accepted hybrid multicloud

SOURCE: 2024 F5 State of Application Strategy

“Modern” apps, that are complex and distributed, have multiplied





51%

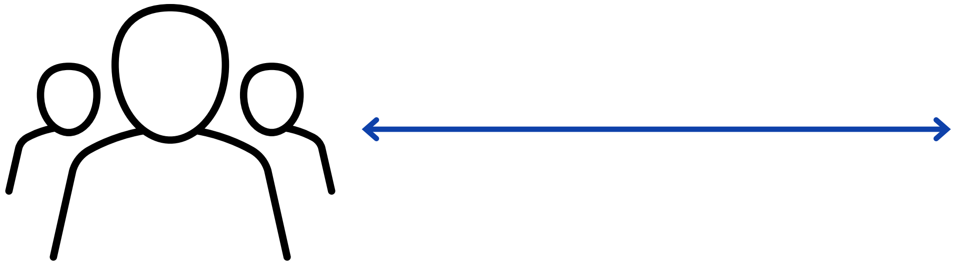
apps are modern apps

18%

apps are AI apps

SOURCE: 2024 F5 State of Application Strategy, F5 Corporate Strategy

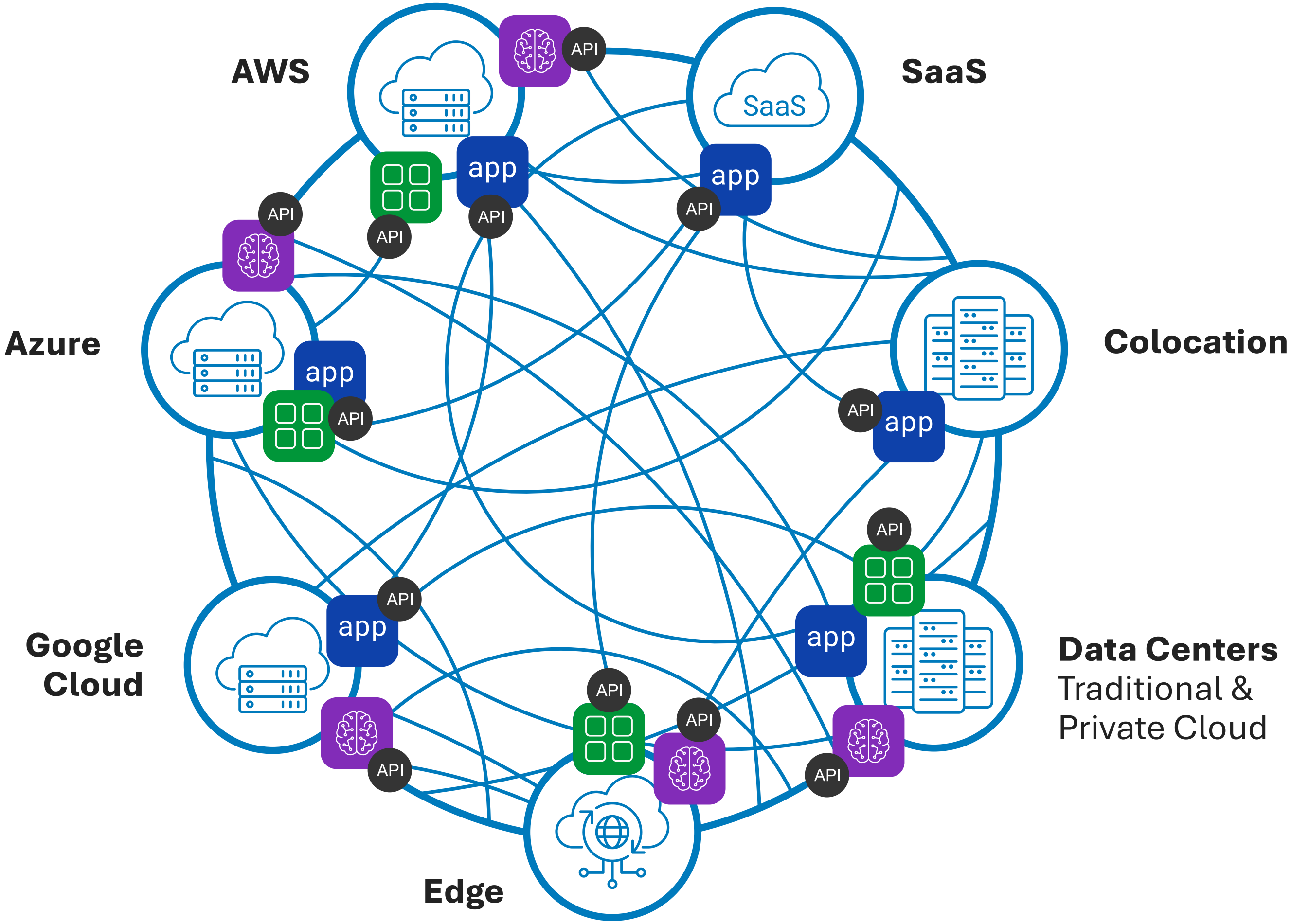
APIs and app-to-app connections have increased complexity



API

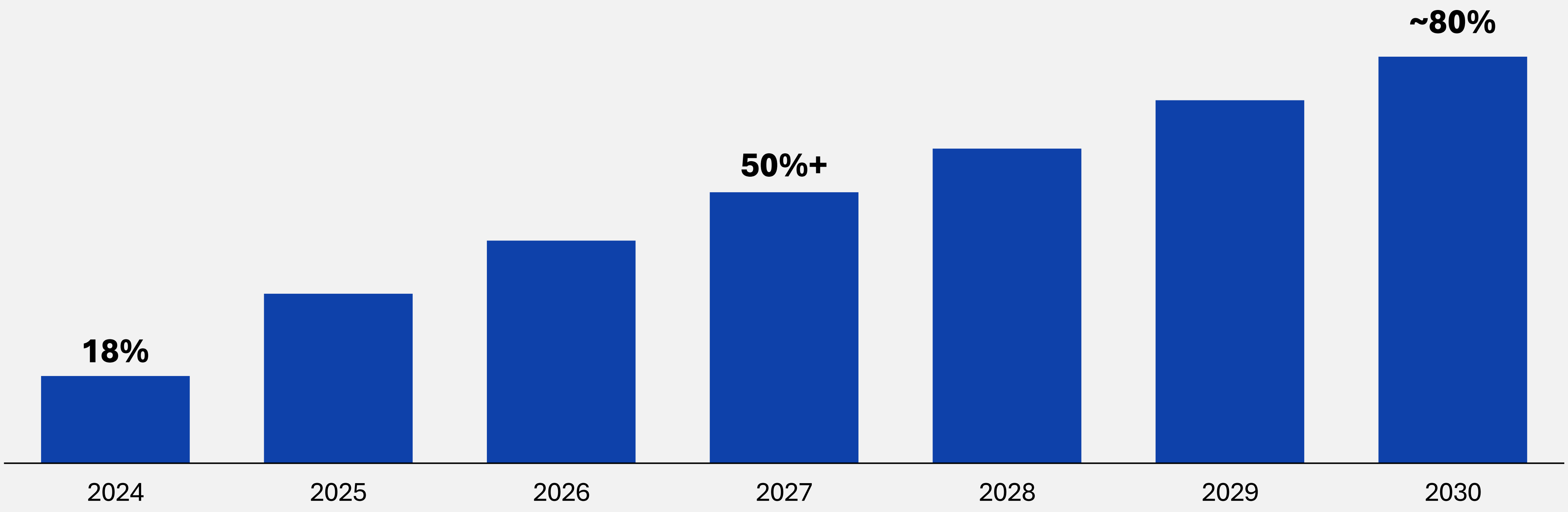
41%

manage as many APIs as apps



But there's more:
Nearly every app will become an AI app

Proportion of AI applications



Source: IDC, Gartner, and F5 Corporate Strategy

AI is not new…

1950 Turing Test

1955 The Term “AI”

1956 Perceptron, K-Means Clustering

1960 GANs

1966 Eliza

1970s – 1980s AI Winter, Genetic Programming

1986 Backpropagation

1995 SVMs, Neural Networks

1997 Deep Blue, LSTMs

2005 Self-Driving Vehicles

2006 Deep Neural Networks

2011 Watson

2014 Alexa

2017 AlphaGo

2017 Automated recommenders surpass humans

2018 OpenAI - GPT

2020 OpenAI launch

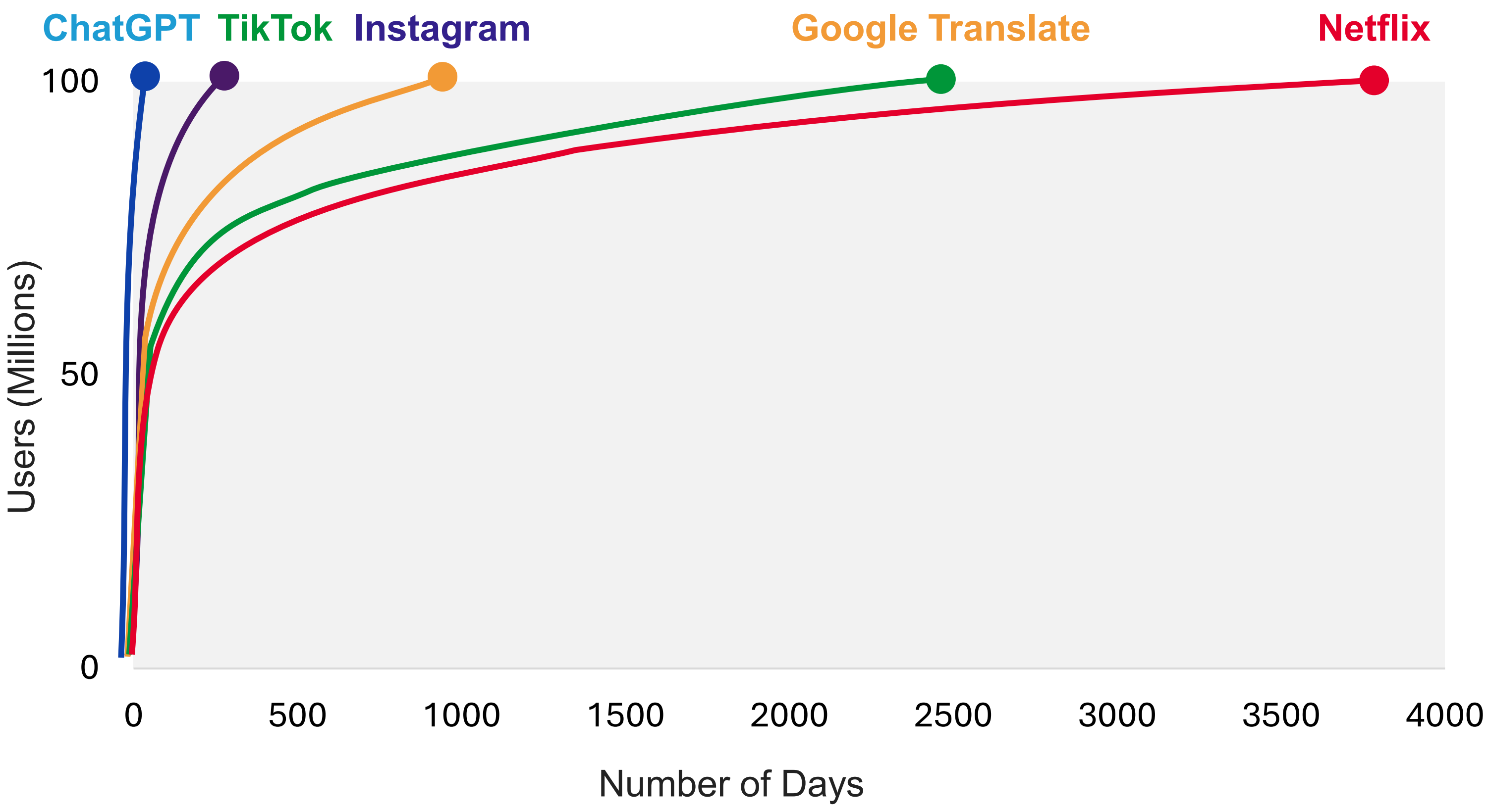
2021 DALL-E

2021 Cancer Identification

2022 ChatGPT (GPT-3)

2023 - 2025 Open Source LLMs, LangChain, Vector Databases, GPT-4, DeepSeek

...but Generative AI has changed the game

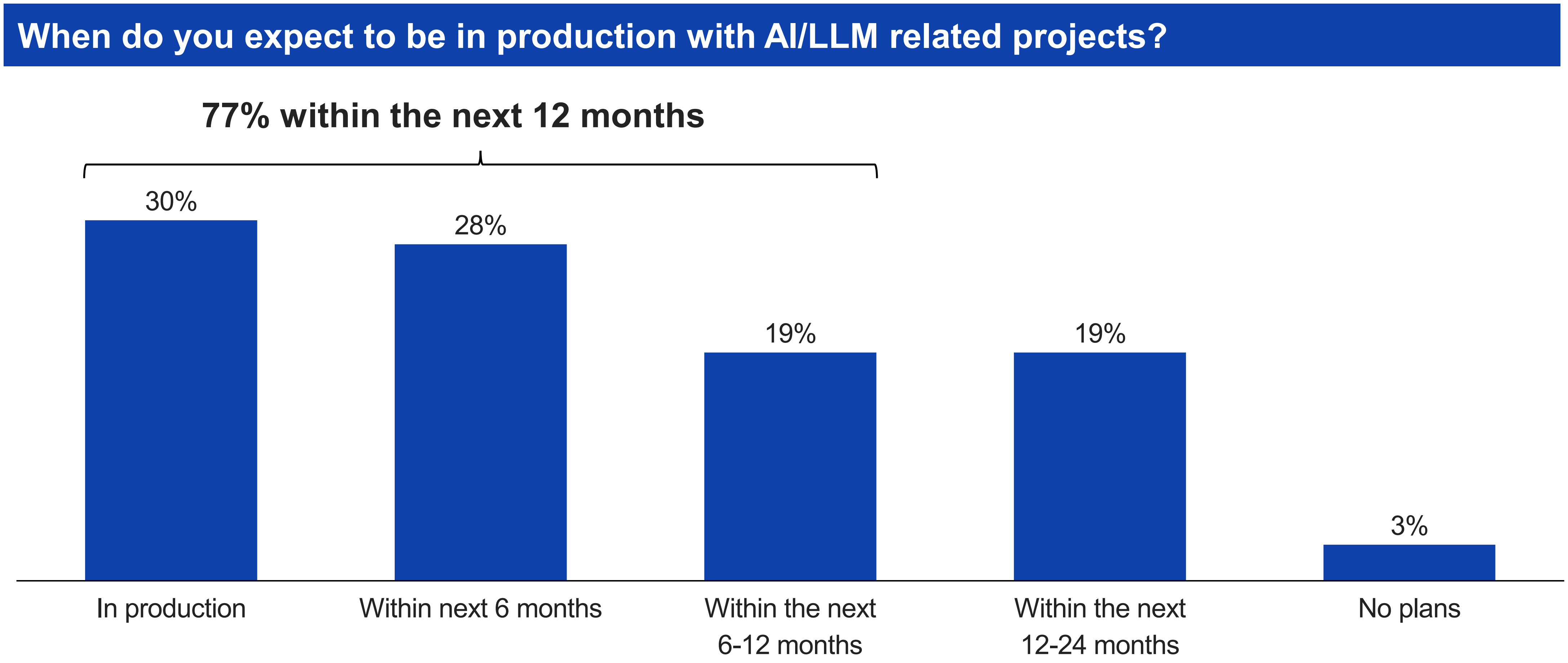


Sources: Global X ETFs with info derived from: BBC News. (2018, Jan 23). Netflix's history: From DVD rentals to streaming success; Cerullo, M. (2023, Feb 1). ChatGPT user base is growing faster than TikTok. CBS News.

> 75% of organizations are implementing AI assistance in 2024+, up from just 17% in 2020

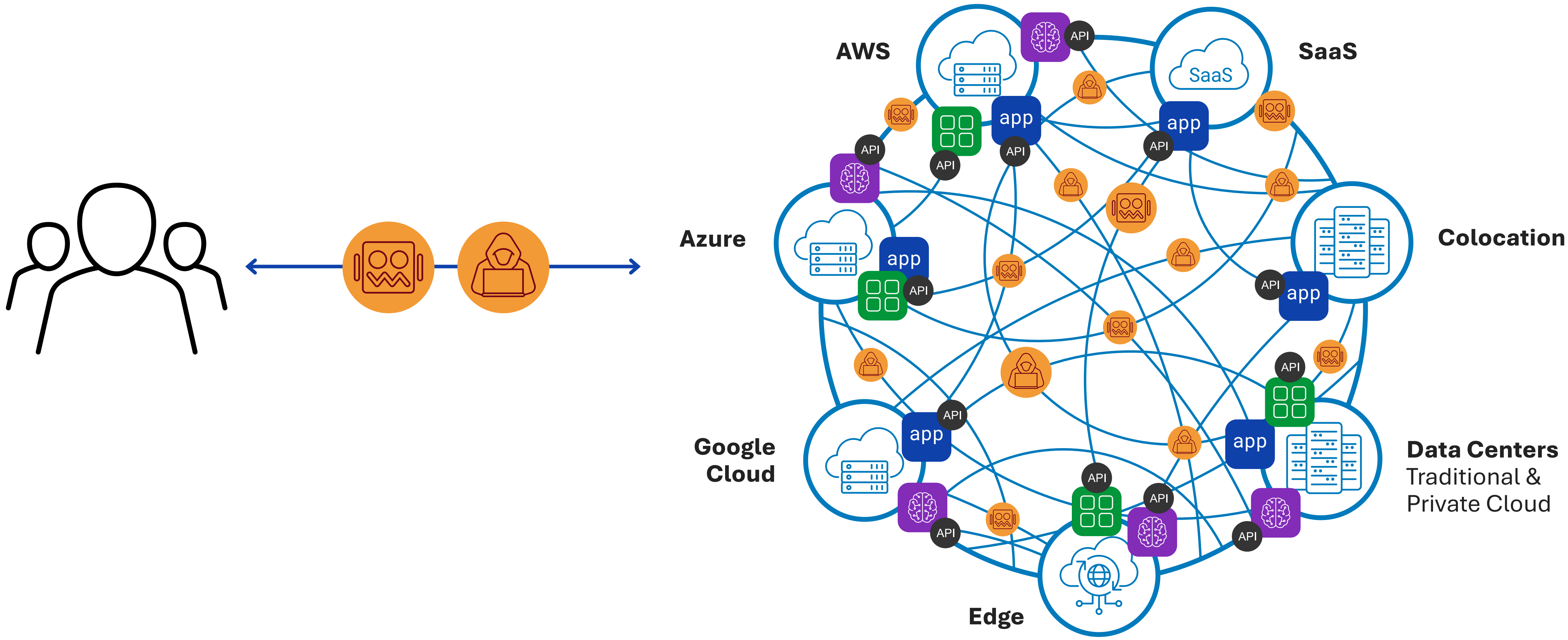
F5 2024 State of AI Application Strategy Report.

Enterprises are already deploying Generative AI to production

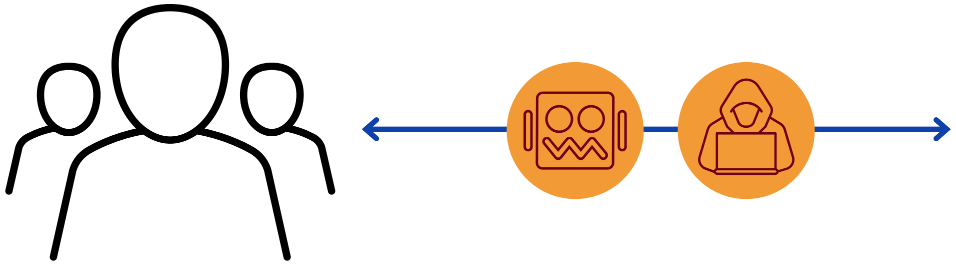


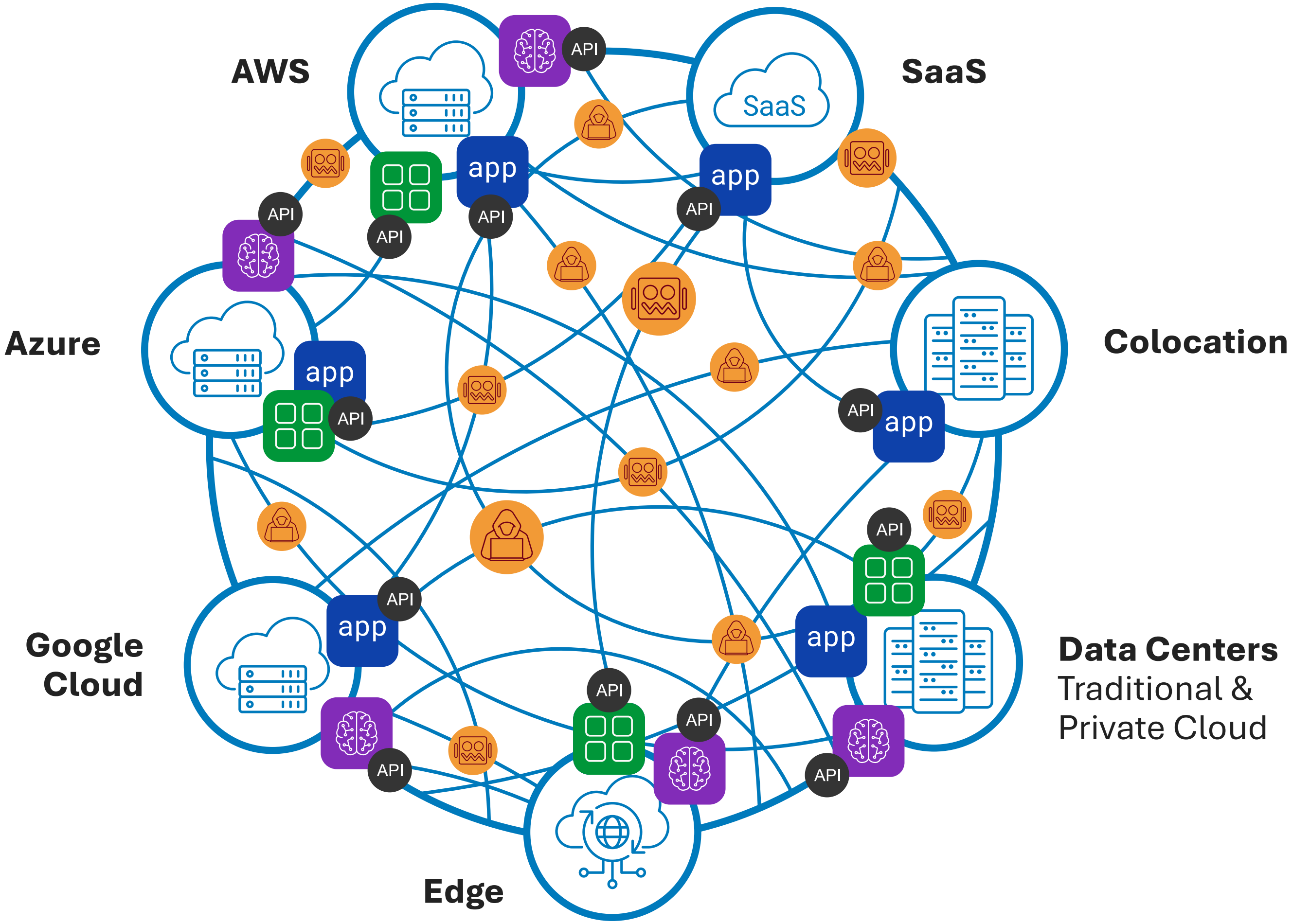
Source: RBC CIO Survey 2024

AI apps further increase complexity and the attack surface

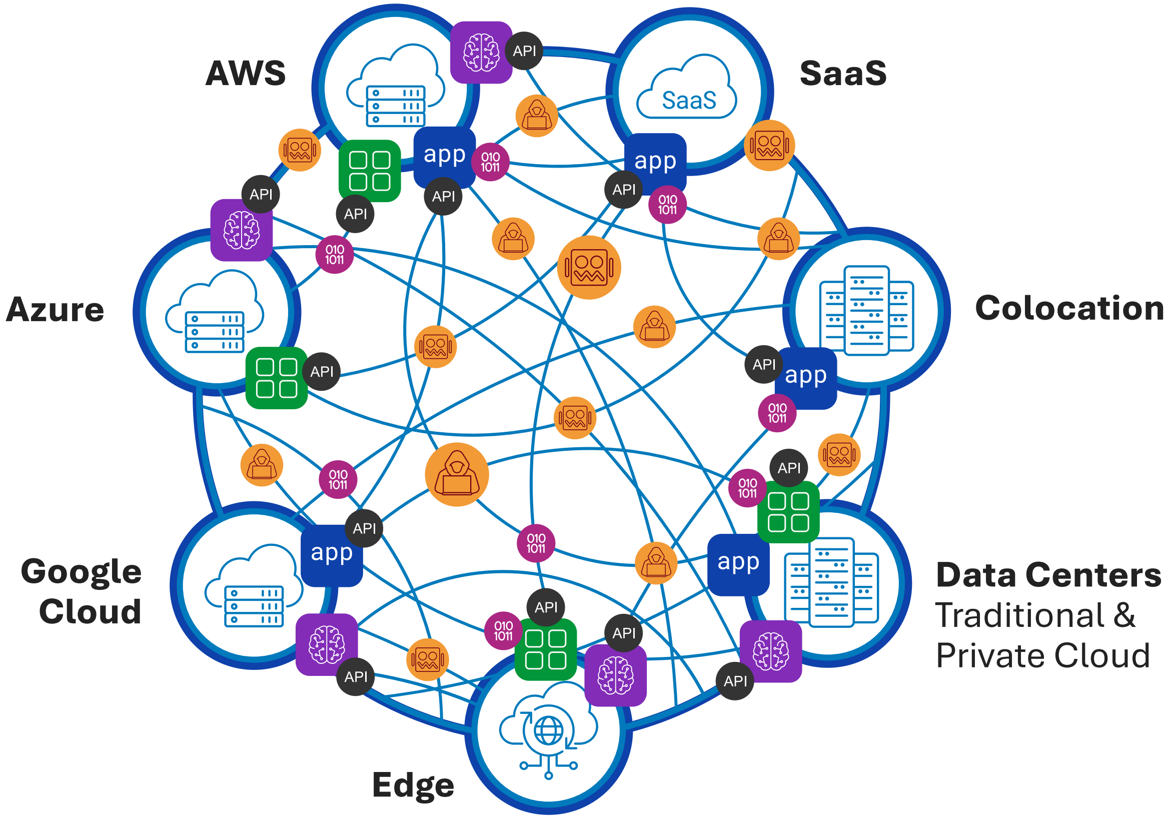


Leaving us to deal with a “Ball of Fire”





The solution to the Ball of Fire will need numerous capabilities



1 Security and delivery for *every* app

2 Any form factor

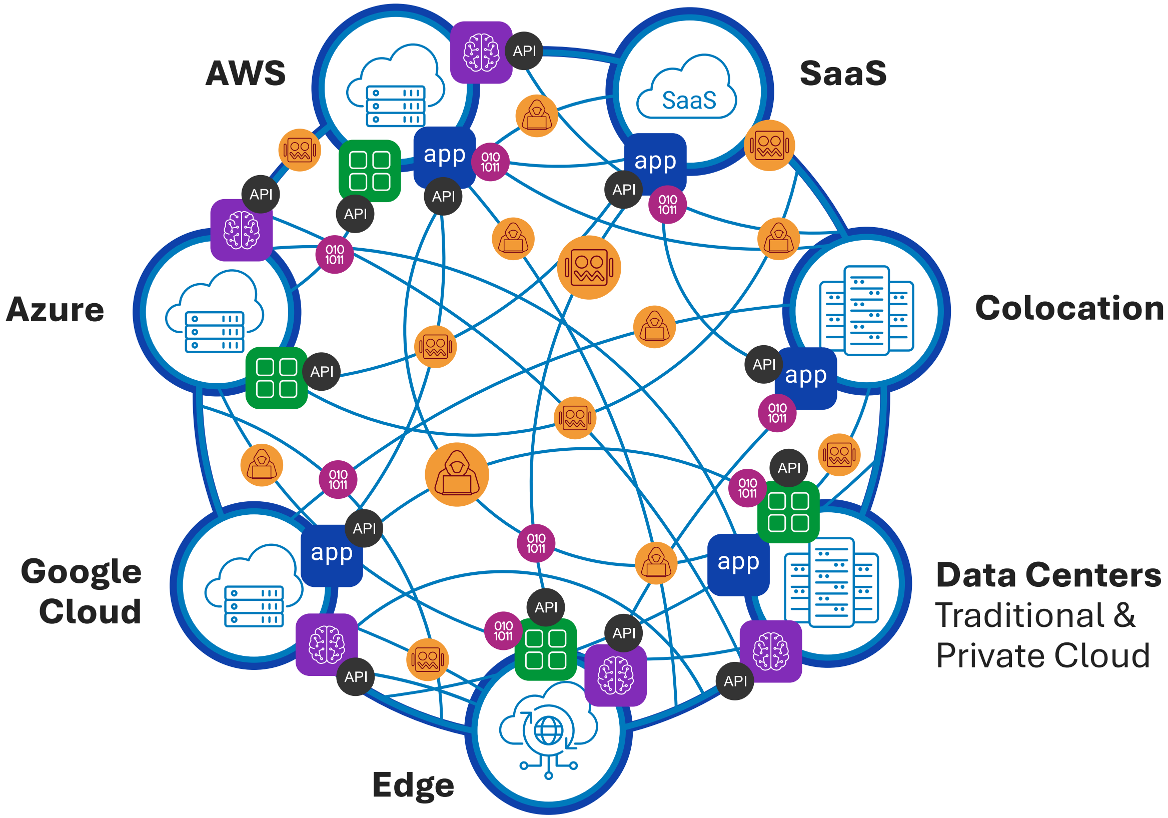
3 Ubiquitous policies

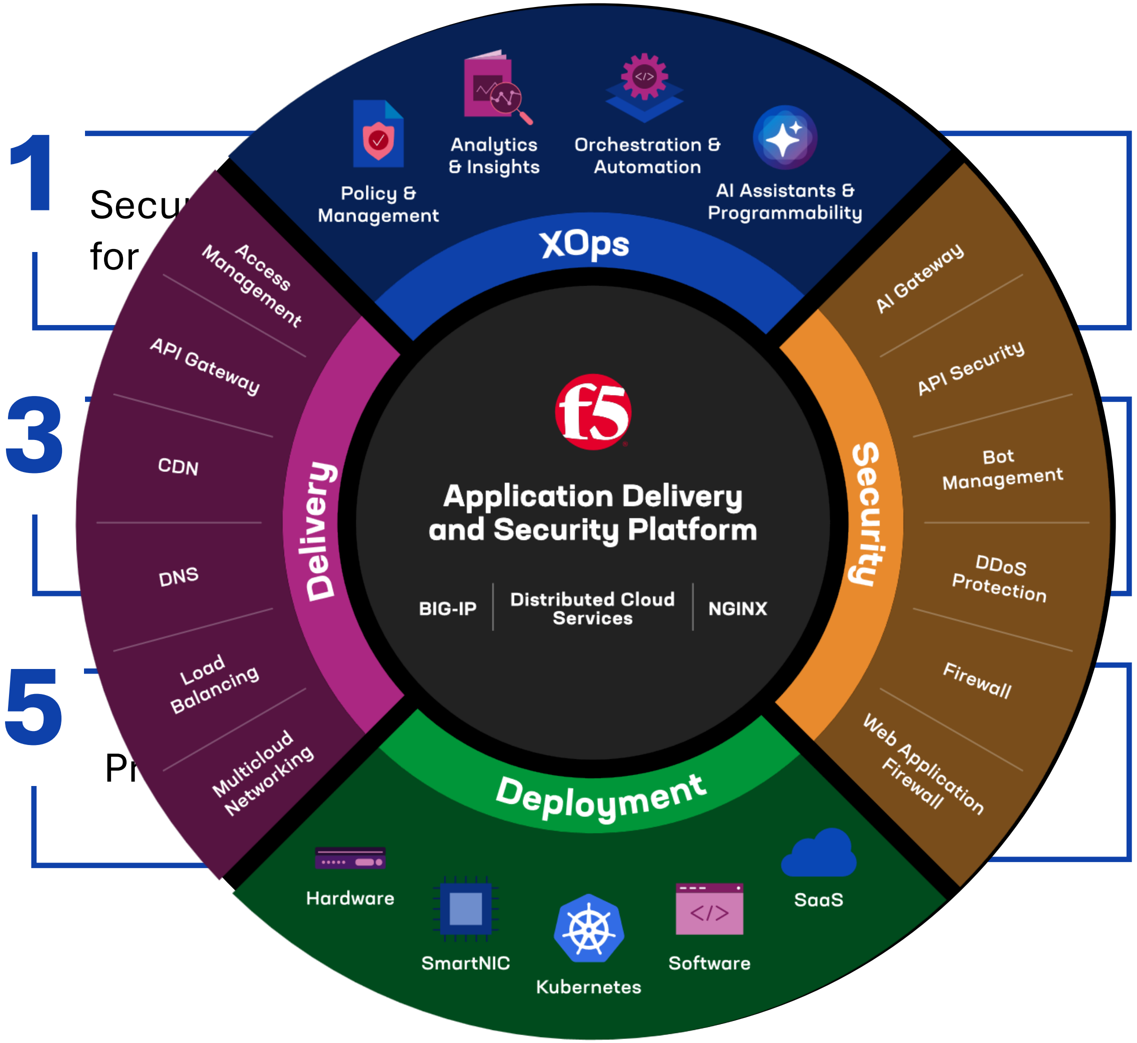
4 Rich analytics and insights

5 Programmable

6 Automated

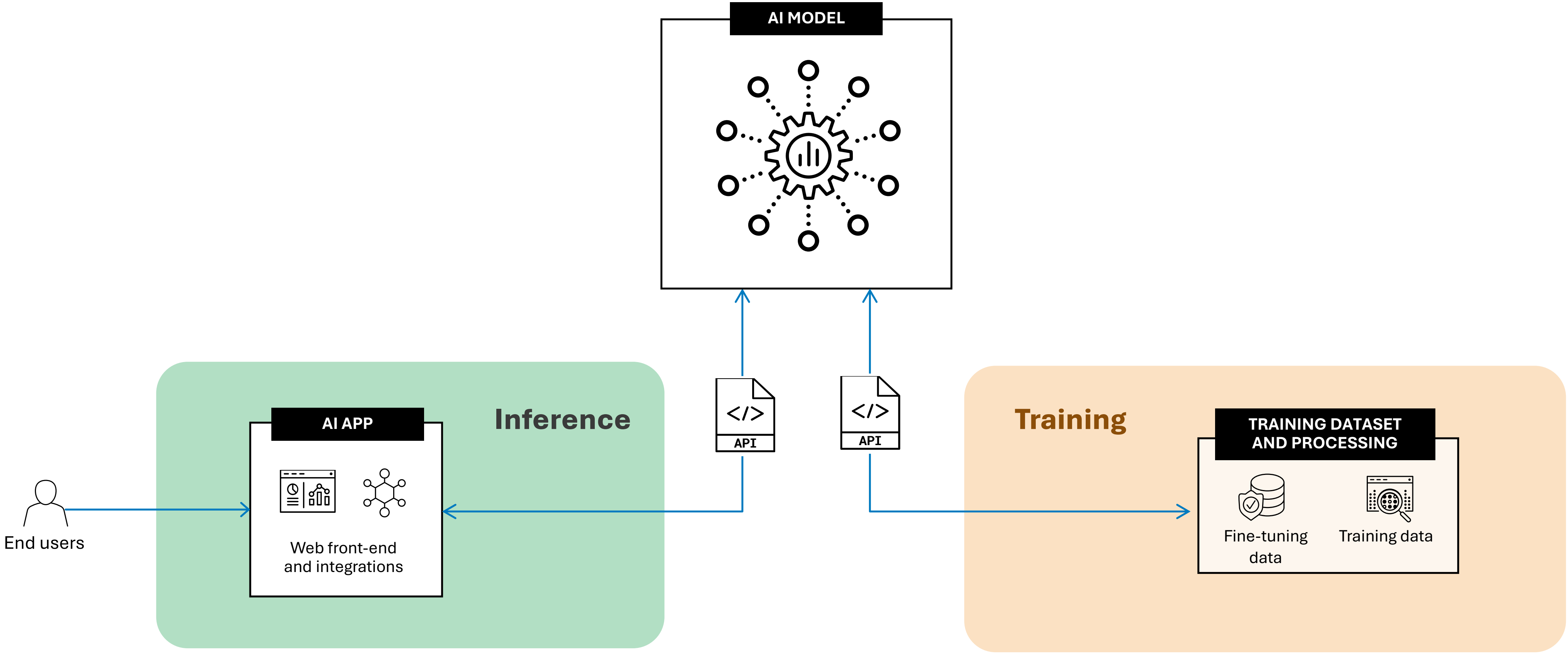
ADC 3.0: F5 Application Delivery and Security Platform





Artificial Intelligence Primer

AI uses a model of a dataset that can be queried



Basic AI (machine learning) uses supervised training

TRAINING:
These are dogs

AI Model

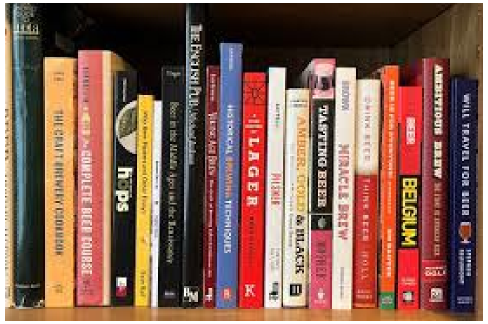
(Computer Program designed to recognize patterns and make decisions)

INFERENCE:
Is this a dog?

85% confident this is a dog

>85% = Yes

Advanced AI uses unsupervised training



Here are 100 million books

Here is wikipedia

Large Language AI Model

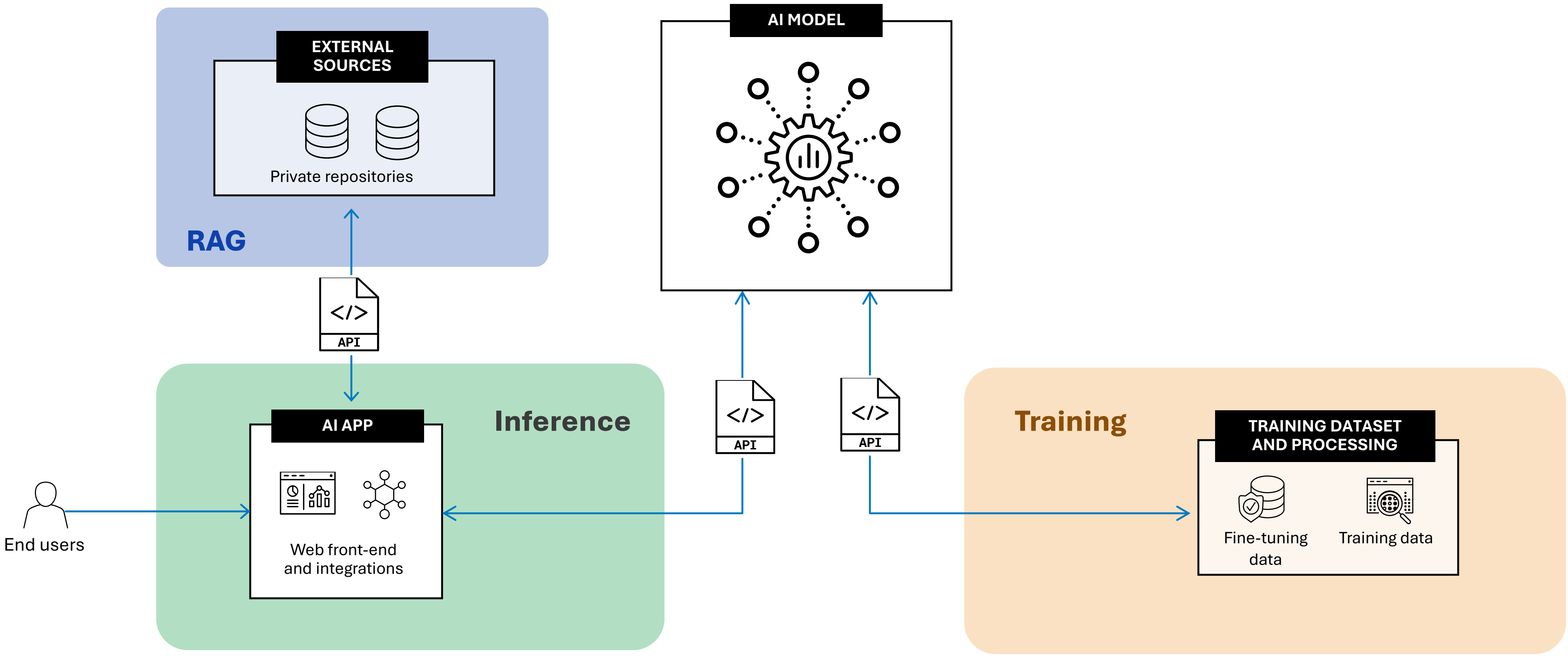
(Computer Program designed to recognize patterns and make decisions)

"Write me a 1 minute story about ancient Greece where the main character is James the lovable hippo."

Once upon a time, in ancient Greece, a lovable hippo named James roamed the vibrant hills of Athens. James wasn't just any hippo, he was curious and fascinated by the grand architecture and the tales of mighty gods. One day, as he waddled through the Agora, he came across a golden olive branch glowing under the sun. It was said thatthis branch had been a gift from Athena herself, granting wisdom to anyone who touched it.

James, with his gentle heart, decded to carry it to the Parthenon, believing it belonged there. Along the way, people laughed and marveled at the sight of a hippo strolling through the streets, but James remained focused on his quest. When he finally reached the temple, Athena herself appeared in a vision, smiling at the kind-hearted creature. "You, James, with your pure spirit, have brought my gift home."

Adding context with Retrieval-Augmented Generation (RAG)



F5 for AI

The new AI stack brings new risks, vulnerabilities and costs

SECURITY VULNERABILITIES

"Cybersecurity researchers have disclosed six security flaws in the Ollama artificial intelligence (AI) framework that could be exploited by a malicious actor to perform various actions, including denial-of-service, model poisoning, and model theft.

The Hacker News

BOOMING COSTS

OpenAI

"LLMs have also grown way more expensive, the HAI authors also observe. "For example, OpenAI's GPT-4 used an estimated $78 million worth of compute to train, while Google's Gemini Ultra costs $191 million for compute," they estimate.

Forbes

CORPORATE LIABILITY

AIR CANADA

"in this case, the CRT found that Air Canada owed a duty of care to its chatbot users, and the airline "did not take reasonable care to ensure its chatbot was accurate" and that these inaccuracies amounted to a negligent misrepresentation.

JDSUPRA

GOVERNANCE COMPLIANCE

THE CITY

"Unveiled in October 2024, MyCity was intended to help provide New Yorkers with information on starting and operating businesses in the city, as well as housing policy and worker rights. ...MyCity falsely claimed that business owners could take a cut of their workers' tips, fire workers who complain of sexual harassment, and serve food that had been nibbled by rodents....

CIO

2025 OWASP Top 10 List for LLM and Gen AI

LLM01:25

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02:25

Sensitive Information Disclosure

Sensitive info in LLMs includes PII, financial, health, business, security, and legal data. Proprietary models face risks with unique training methods and source code, critical in closed or foundation models.

LLM03:25

Supply Chain

LLM supply chains face risks in training data, models, and platforms, causing bias, breaches, or failures. Unlike traditional software, ML risks include third-party pre-trained models and data vulnerabilities.

LLM04:25

Data and Model Poisoning

Data poisoning manipulates pre-training, fine-tuning, or embedding data, causing vulnerabilities, biases, or backdoors. Risks include degraded performance, harmful outputs, toxic content, and compromised downstream systems.

LLM05:25

Improper Output Handling

Improper Output Handling involves inadequate validation of LLM outputs before downstream use. Exploits include XSS, CSRF, SSRF, privilege escalation, or remote code execution, which differs from Overreliance.

LLM06:25

Excessive Agency

LLM systems gain agency via extensions, tools, or plugins to act on prompts. Agents dynamically choose extensions and make repeated LLM calls, using prior outputs to guide subsequent actions for dynamic task execution.

LLM07:25

System Prompt Leakage

System prompt leakage occurs when sensitive info in LLM prompts is unintentionally exposed, enabling attackers to exploit secrets. These prompts guide model behavior but can unintentionally reveal critical data.

LLM08:25

Vector and Embedding Weaknesses

Vectors and embeddings vulnerabilities in RAG with LLMs allow exploits via weak generation, storage, or retrieval. These can inject harmful content, manipulate outputs, or expose sensitive data, posing significant security risks.

LLM09:25

Misinformation

LLM misinformation occurs when false but credible outputs mislead users, risking security breaches, reputational harm, and legal liability, making it a critical vulnerability for reliant applications.

LLM10:25

Unbounded Consumption

Unbounded Consumption occurs when LLMs generate outputs from inputs, relying on inference to apply learned patterns and knowledge for relevant responses or predictions, making it a key function of LLMs.

F5 Delivery and Security for AI

1

Traffic Management for Data Ingest

High-performance load balancing to ingest data for multi-billion-parameter AI models

2

API Security

Automatically discover and secure all AI training and inference endpoints

3

Secure Multicloud Networking

High-throughput connectivity across on-prem, clouds, and edge for AI model training & inference

4

Distributed Inference

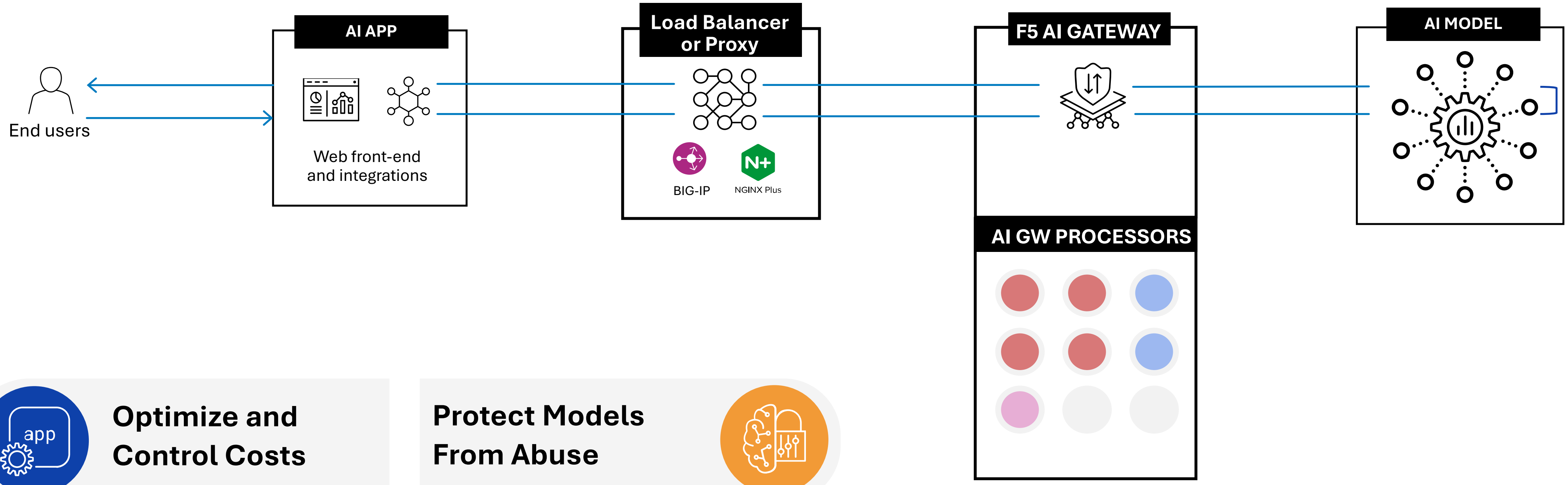
Run AI inference models anywhere cloud to edge with lifecycle management update, patch, and version manage centrally.

5

Secure AI App Delivery in Kubernetes

Simplify and streamline model operations and security at scale.

Introducing F5 AI Gateway



Optimize and Control Costs

Protect Models From Abuse

Prevent Unauthorised Output

SDK for Extensibility

AI Gateway functionality today

	Watermarking		
	LLM Response Rejection	Malicious Command Detection	Processor Resource Exhaustion
	Data Generalization	Sensitive System/Data Info Detection	API Rate Limiting
Malicious Command Detection	Data Redaction	PII Detection	Malicious Command Detection
Prompt Injection Detection	Data Filtering / labeling	Code Sanitization	Code Sanitization
Behavior Reinforcement	Sensitive System / Data info Detection	Code Detection	Prompt Injection Detection
Max Input Limit	PII Detection	Behavior Reinforcement	Max Input Limit
Prompt Injection	**Sensitive Information Disclosure**	**Improper Output Handling**	**Unbounded Consumption**
LLM Attack Vectors			

Threat & IP Protection

- Prompt inspection/blocking
- Large Language Model (LLM) integrity

Application Services & Acceleration

- Content-based model routing
- Authentication, Authorization, and Role-based Access in conjunction with BIG-IP or NGINX

Observability

- OpenTelemetry logging

Software Development Kit (SDK)

- Python

Coming soon…

- Rust
- Go

SDKs enable customers / partners to customize AI Gateway to secure & control their AI Apps to meet business needs