cisco Live!

Let’s go

The bridge to possible

Harnessing the Power of APIs in Artificial Intelligence

DEVNET-2220

Shannon McFarland – CCIE#5245
VP, Engineering
Cisco DevNet
@eyepv6

Other Sessions

- DEVNET-2703: Securing APIs from Left to Right, and Everywhere in Between
- DEVWKS-1704: AI Code Warrior – Wielding Artificial Intelligence Tools as a Developer
- DEVNET-2708: Empowering Business with Security, Private and Sovereign AI: A Guide to Deploying Large Language Models
- DEVNET-2714: Explore Generative AI Capabilities
- DEVNET-3707: Network Telemetry and AI for Network Incident Response
- DEVNET-2850: Build an LLM-based Application in 45mins!

Agenda

- Level-set: Why APIs matter
- Artificial Intelligence - Uses
- State of the Union on AI APIs
- The Age of Agents
- Your AI Strategy

Why APIs Matter

© Damian Sobczyk / stock.adobe.com

Applications & Developers Drive Business

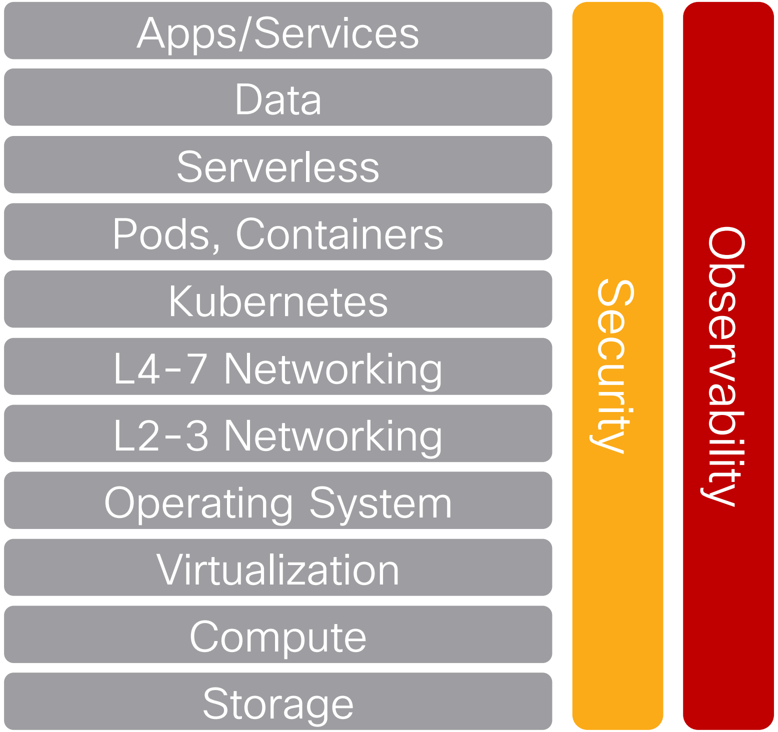
Without applications, all our networks do is send control plane traffic

© Mongta Studio / stock.adobe.com

Developers Create APIs – We Use Them

We use APIs everywhere

- APIs are the face of infra, apps and services
- There are different APIs for different uses



APIs are the face of new tech

Artificial Intelligence

- AI for Software
- Software for AI

Sustainability

- Footprint Reduction
- Energy Observation

Low-Code/No-Code

- Rapid Prototyping
- Workflow Applications

Artificial Intelligence - Uses

Artificial Intelligence isn’t new

Infrastructure Use Cases

- Network Traffic
- Resource Allocation in Storage & Compute
- Application Performance Prediction
- Load Balancing
- Data Breach

Business Use Cases

- Sales Forecasting
- Manufacturing Maintenance
- Fraud Detection
- Customer Behavior
- Healthcare

Generative AI – the new(er) kid on the block

Infrastructure Use Cases

- Network Optimization
- Resource Allocation
- Security Threat Detection
- Disaster Recovery Planning

Business Use Cases

- Content Creation
- Data Augmentation
- Automated Programming
- Personalized Marketing

State of the Union on AI APIs

Use Cases inside of Use Cases

AI API Use Cases

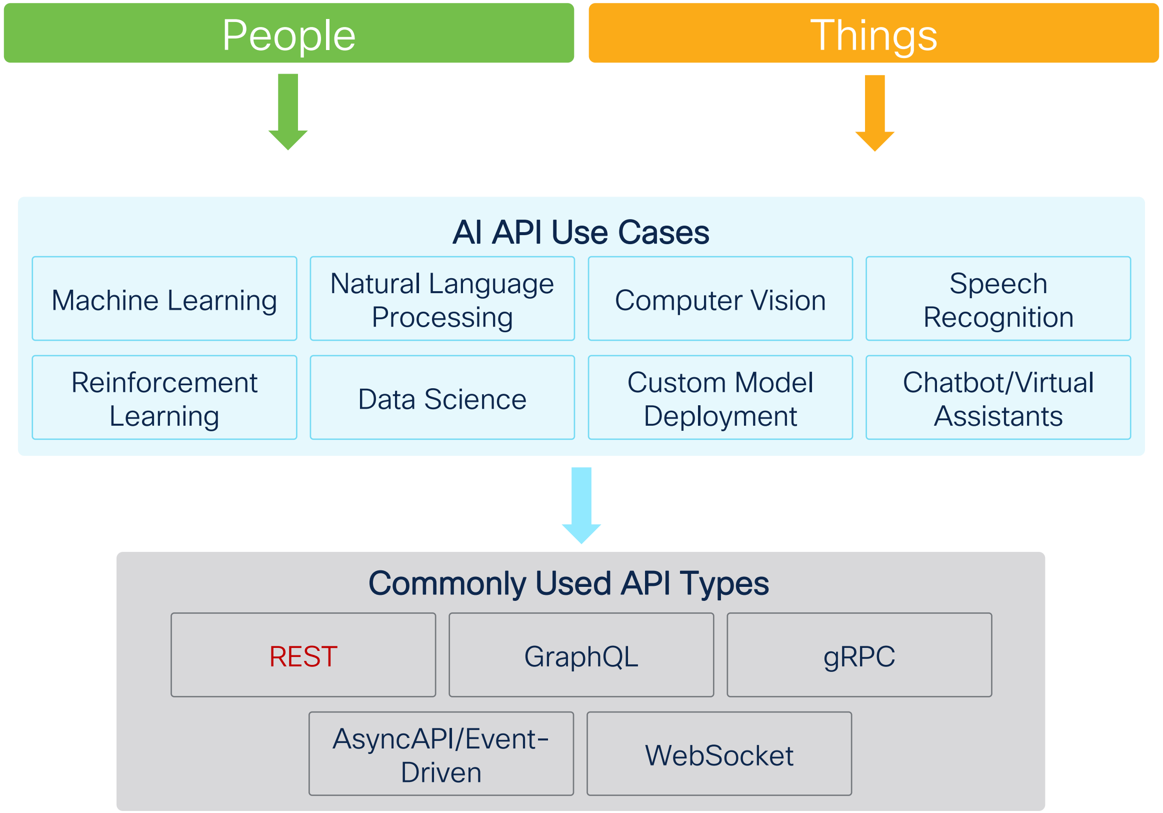
Machine Learning	Natural Language Processing	Computer Vision	Speech Recognition
Reinforcement Learning	Data Science	Custom Model Deployment	Chatbot/Virtual Assistants

- Speech, text, language
  - Speech-to-text
  - Text-to-speech
  - Translation

- Image & video
  - Text-to-image
  - Video intelligence

- Document & data
  - Optical Character Recognition (OCR)
  - Parsers
  - Warehouse

AI API Use Cases & types



People

Things

AI API Use Cases

- Machine Learning
- Natural Language Processing
- Computer Vision
- Speech Recognition
- Reinforcement Learning
- Data Science
- Custom Model Deployment
- Chatbot/Virtual Assistants

Commonly Used API Types

- REST
- GraphQL
- gRPC
- AsyncAPI/Event-Driven
- WebSocket

OpenAPI + OpenAI = “Wait, what?!”

RESTful APIs – OpenAPI Specification (OAS)

OpenAPI

- https://www.openapis.org/
- Originally based on the Swagger Specification (Donated by SmartBear Software)
- OAS provides a standard, language-agnostic interface to RESTful APIs
- Design, build, document and consume RESTful APIs

OpenAI

https://openai.com/
Artificial General Intelligence
ChatGPT hotness
Uses OpenAPI:

- https://platform.openai.com/docs/api-reference/introduction
- https://github.com/openai/openai-openapi/blob/master/openapi.yaml

https://developer.cisco.com/learning/tracks/Coding-APIs-v0/

OpenAI's OpenAPI Spec

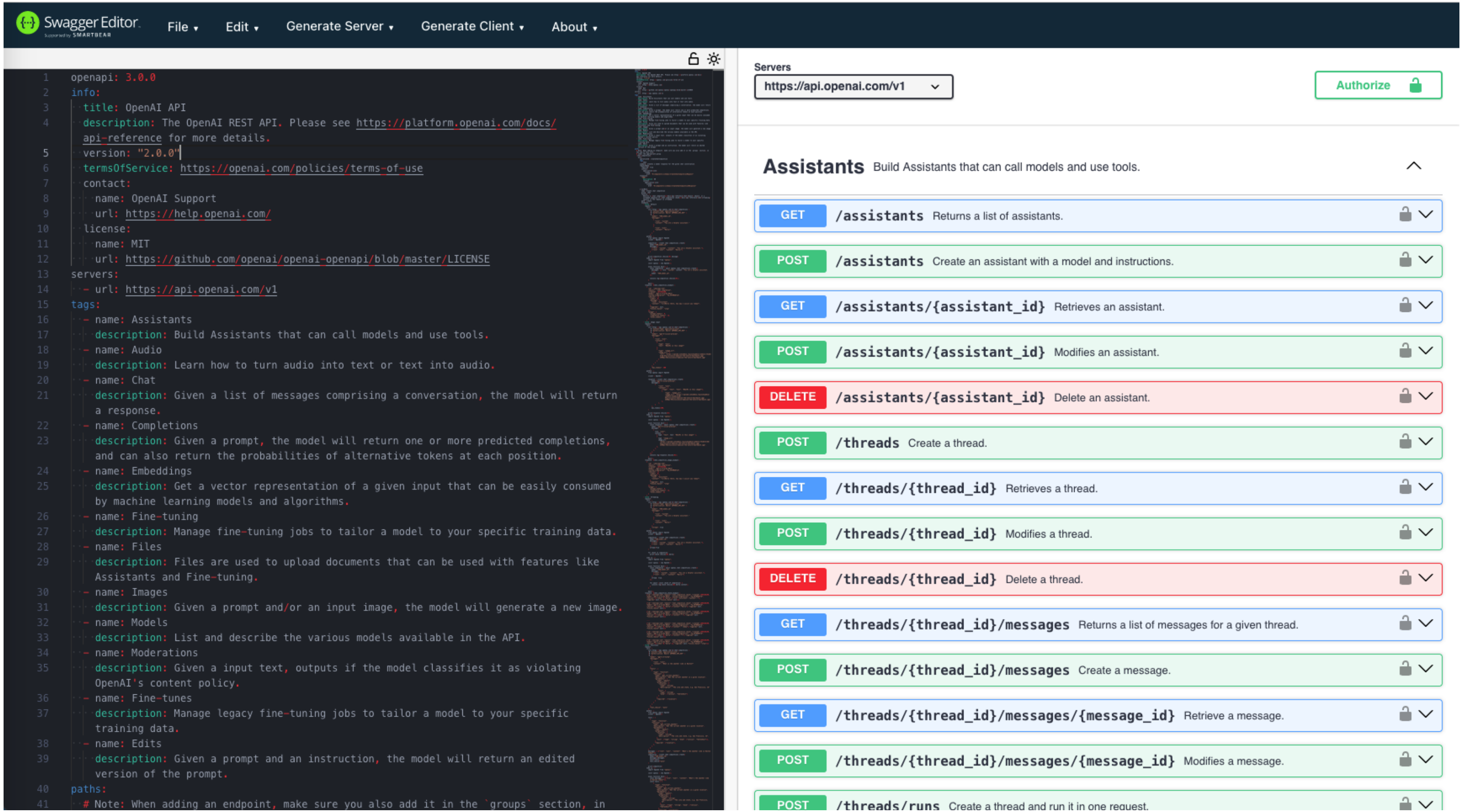
- OpenAI Platform Overview: https://platform.openai.com/docs/overview
- OpenAI API Docs: https://platform.openai.com/docs/api-reference/introduction
- OpenAPI Document for OpenAI: https://github.com/openai/openai-openapi/blob/master/openapi.yaml

API Server: `https://api.openai.com/v1`

API Paths:

```
/assistants/
/audio/
/chat/
/completions/
/embeddings/
/fine_tuning/
/files/
/images/
/models/
/moderations/
```

Operations: `GET, POST, DELETE`



```yaml
openapi: 3.0.0
info:
  title: OpenAI API
  description: The OpenAI REST API. Please see https://platform.openai.com/docs/api-reference for more details.
  version: "2.0.0"
  termsOfService: https://openai.com/policies/terms-of-use
  contact:
    name: OpenAI Support
    url: https://help.openai.com/
  license:
    name: MIT
    url: https://github.com/openai/openai-openapi/blob/master/LICENSE
servers:
  - url: https://api.openai.com/v1
tags:
  - name: Assistants
    description: Build Assistants that can call models and use tools.
  - name: Audio
    description: Learn how to turn audio into text or text into audio.
  - name: Chat
    description: Given a list of messages comprising a conversation, the model will return a response.
  - name: Completions
    description: Given a prompt, the model will return one or more predicted completions, and can also return the probabilities of alternative tokens at each position.
  - name: Embeddings
    description: Get a vector representation of a given input that can be easily consumed by machine learning models and algorithms.
  - name: Fine-tuning
    description: Manage fine-tuning jobs to tailor a model to your specific training data.
  - name: Files
    description: Files are used to upload documents that can be used with features like Assistants and Fine-tuning.
  - name: Images
    description: Given a prompt and/or an input image, the model will generate a new image.
  - name: Models
    description: List and describe the various models available in the API.
  - name: Moderations
    description: Given a input text, outputs if the model classifies it as violating OpenAI's content policy.
  - name: Fine-tunes
    description: Manage legacy fine-tuning jobs to tailor a model to your specific training data.
  - name: Edits
    description: Given a prompt and an instruction, the model will return an edited version of the prompt.
paths:
  # Note: When adding an endpoint, make sure you also add it in the `groups` section, in
```

Servers

https://api.openai.com/v1 ⌄

Authorize 🔓

Assistants Build Assistants that can call models and use tools.

- GET `/assistants` Returns a list of assistants.
- POST `/assistants` Create an assistant with a model and instructions.
- GET `/assistants/{assistant_id}` Retrieves an assistant.
- POST `/assistants/{assistant_id}` Modifies an assistant.
- DELETE `/assistants/{assistant_id}` Delete an assistant.
- POST `/threads` Create a thread.
- GET `/threads/{thread_id}` Retrieves a thread.
- POST `/threads/{thread_id}` Modifies a thread.
- DELETE `/threads/{thread_id}` Delete a thread.
- GET `/threads/{thread_id}/messages` Returns a list of messages for a given thread.
- POST `/threads/{thread_id}/messages` Create a message.
- GET `/threads/{thread_id}/messages/{message_id}` Retrieve a message.
- POST `/threads/{thread_id}/messages/{message_id}` Modifies a message.
- POST `/threads/runs` Create a thread and run it in one request.

GraphQL

- https://graphql.org/
- An open-source query and mutation language for APIs
- Allows clients to specify exactly what data they need: https://www.howtographql.com/basics/1-graphql-is-the-better-rest/
- Reduces the amount of data that is transferred between the client and the API
- AI use case: Identify text & language, translate text, convert to speech
- OpenAI and many other services do not yet support GraphQL natively

Reference

REST & GraphQL – Example

REST

1

```
curl -X 'GET' \ 'https://api.example.com/v1/models’ \
-H 'accept: application/json'
```

2

```
curl -X 'GET' \ 'https://api.example.com/v1/models/gpt-3.5-turbo’ \
-H 'accept: application/json’
```

GraphQL

1

```
query {
  model(name: "gpt-3.5-turbo") {
    id
    name
    created
    /* other fields related to the model */
  }
}
```

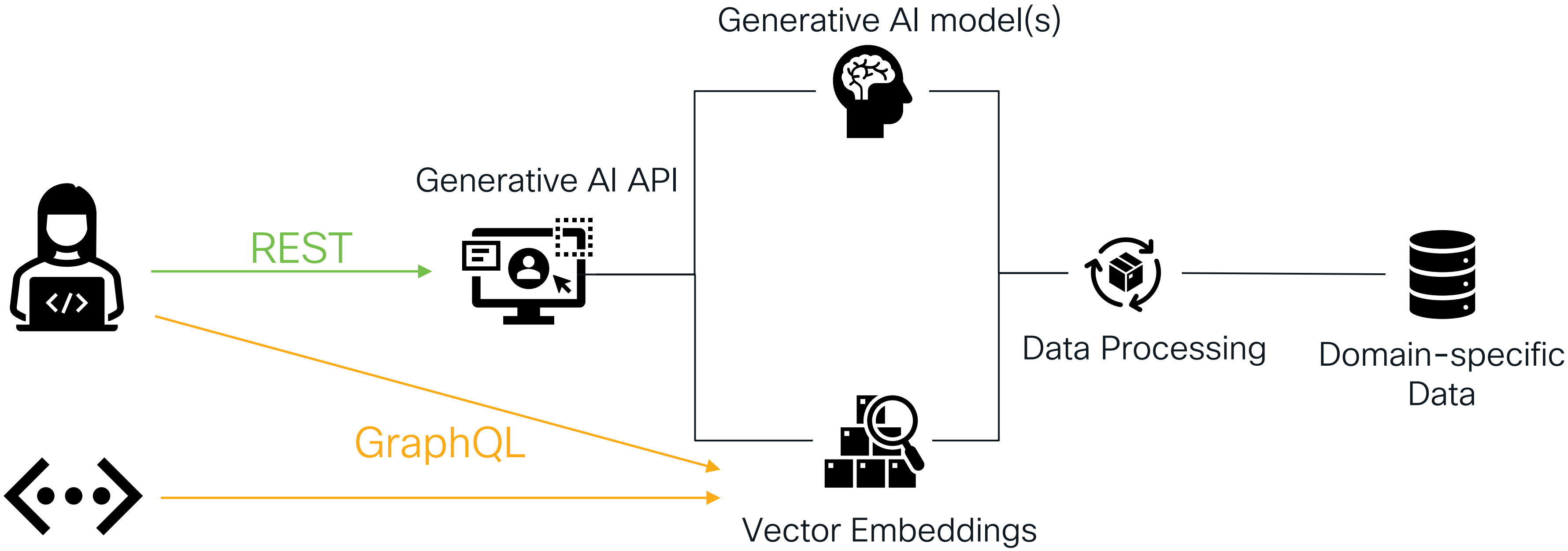
or

```
query {
  models(first: 4) {
    edges {
      node {
        id
        name
        created
      }
    }
  }
}
```

*REST allows for the use of query parameters

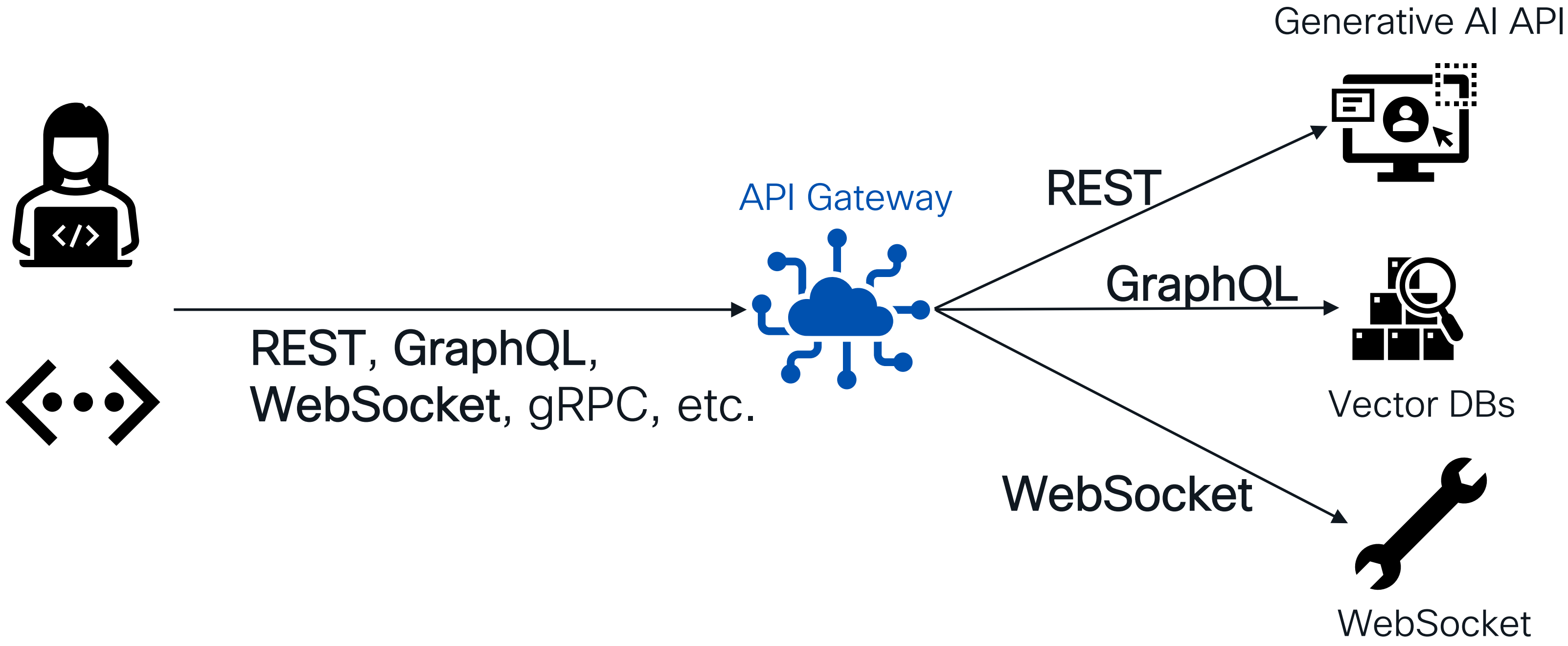
Multi-API Systems – Using Vector DB with GraphQL

More Accuracy for GenAI



The Role of API Gateways

Multi-API Support



API Security

- The APIs used in AI are subject to the same (or more) attack vectors as anything else
- Open Worldwide Application Security Project (OWASP) Top Ten certainly still apply: https://owasp.org/API-Security/editions/2023/en/0x11-t10/
- OWASP Top Ten for LLMs: https://owasp.org/www-project-top-10-for-large-language-model-applications/
- MITRE ATLAS: https://atlas.mitre.org/
- Threats to pay particular attention to:
  - Weak 3rd Party Authentication (OWASP API10:2023) – Multi-service AI APIs
  - Data Injection – Pass malicious data, configurations or programs into AI apps
  - Code Injection – IDE plugins and AI-authored code can be used to inject unknown or misunderstood code
  - Shadow, Zombie & Rogue APIs – Unknown/Undocumented or deprecated APIs, especially those built by AI
  - Prompt requests/responses format is free-form text, which is easy to manipulate

Panoptica—Comprehensive Code to Cloud Security – www.panoptica.app

Code + Build Security

Prioritize and Remediate:

- Full Development Lifecycle
- Governance Policies
- Infrastructure as Code

Cloud Security Posture Management

Automate and Simplify:

- Compliance Monitoring
- Resource Management

Cloud Workload Protection

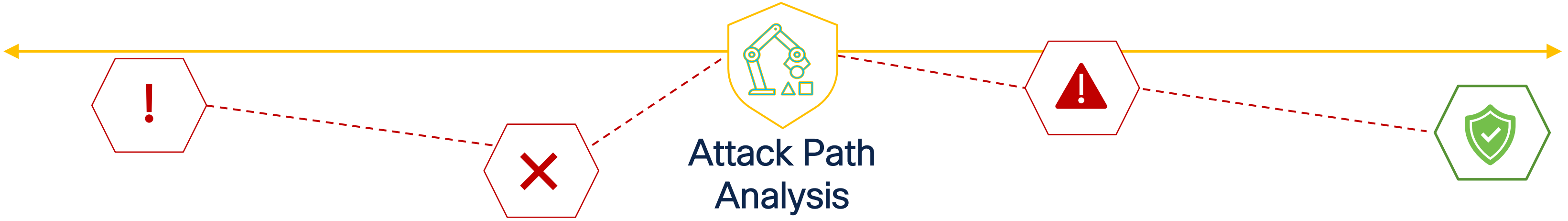
Continuous Risk Management:

- Virtual Machines
- Containers
- Serverless

API Security

Assess and Monitor:

- Internal & External APIs
- API Tokens



Attack Path Analysis

Prioritize with precision.
Remediate the risks that matter—first.

The Age of Agents

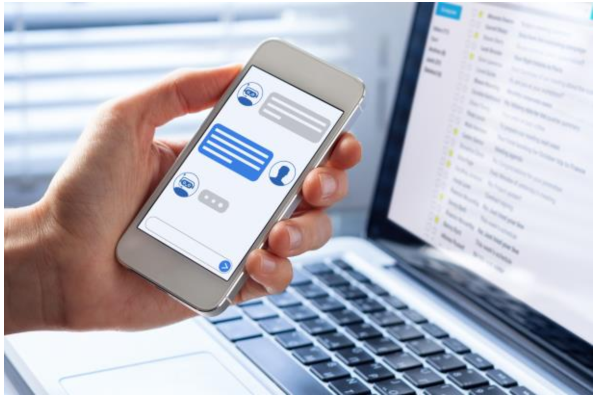
AI Agents – another tool in the toolbox

- An intelligent system that is designed for Natural Language Processing and to more efficiently interact with complex data
- Rule-based to complex ML algorithms
- Semi-to-Fully autonomous learning & decision making
- Conversational interactions
- Real-time data processing

© Mustafa / stock.adobe.com

AI Agents and AI Assistants

© Mustafa / stock.adobe.com



© NicoElNino / stock.adobe.com

AI Agent: Autonomously perform actions to achieve specific goals

AI Assistant: Perform pre-determined or human-assisted tasks

Common Use Cases – Agents and Assistants

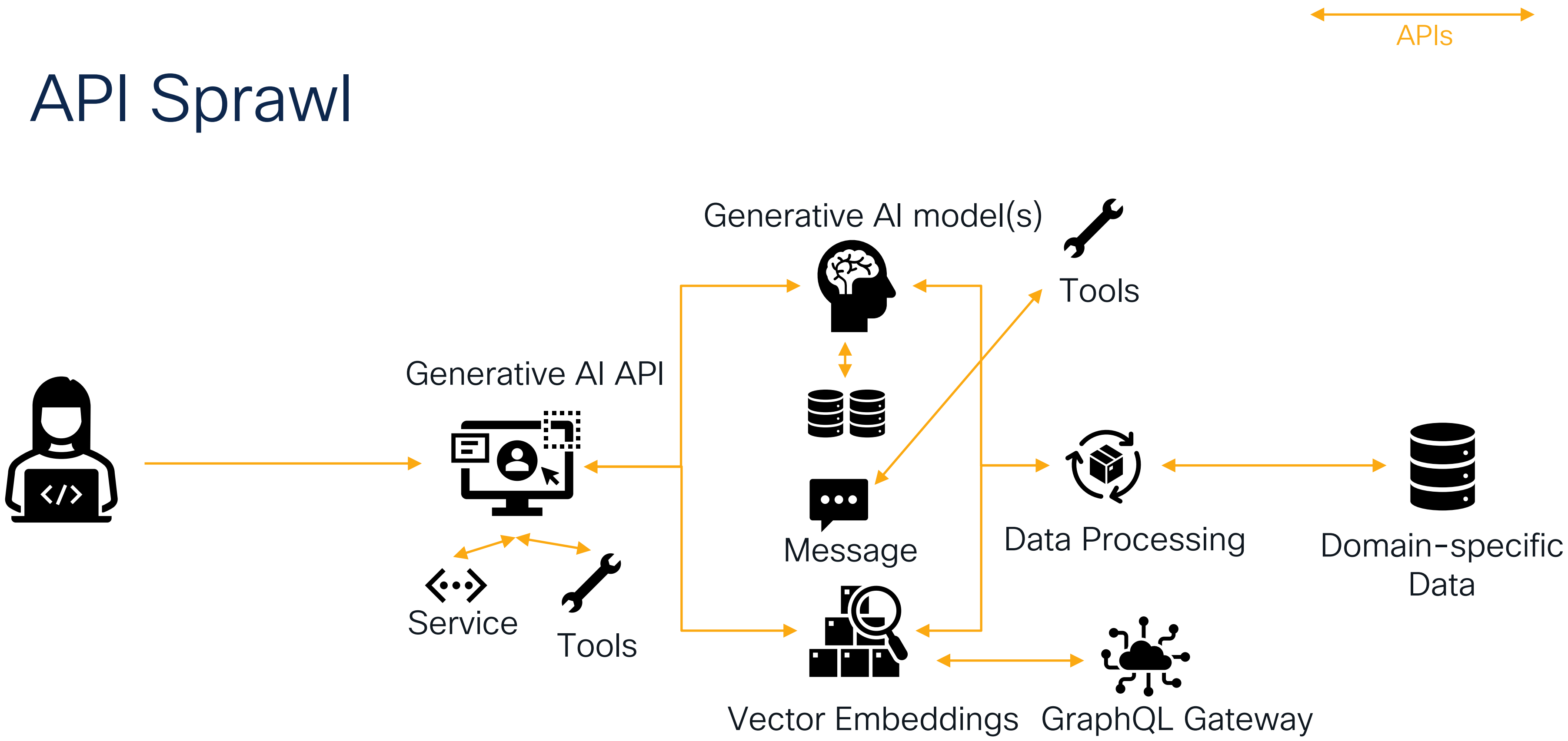
Agents

- Self-driving cars
- Recommendation systems
- Robotics
- Healthcare

Assistants

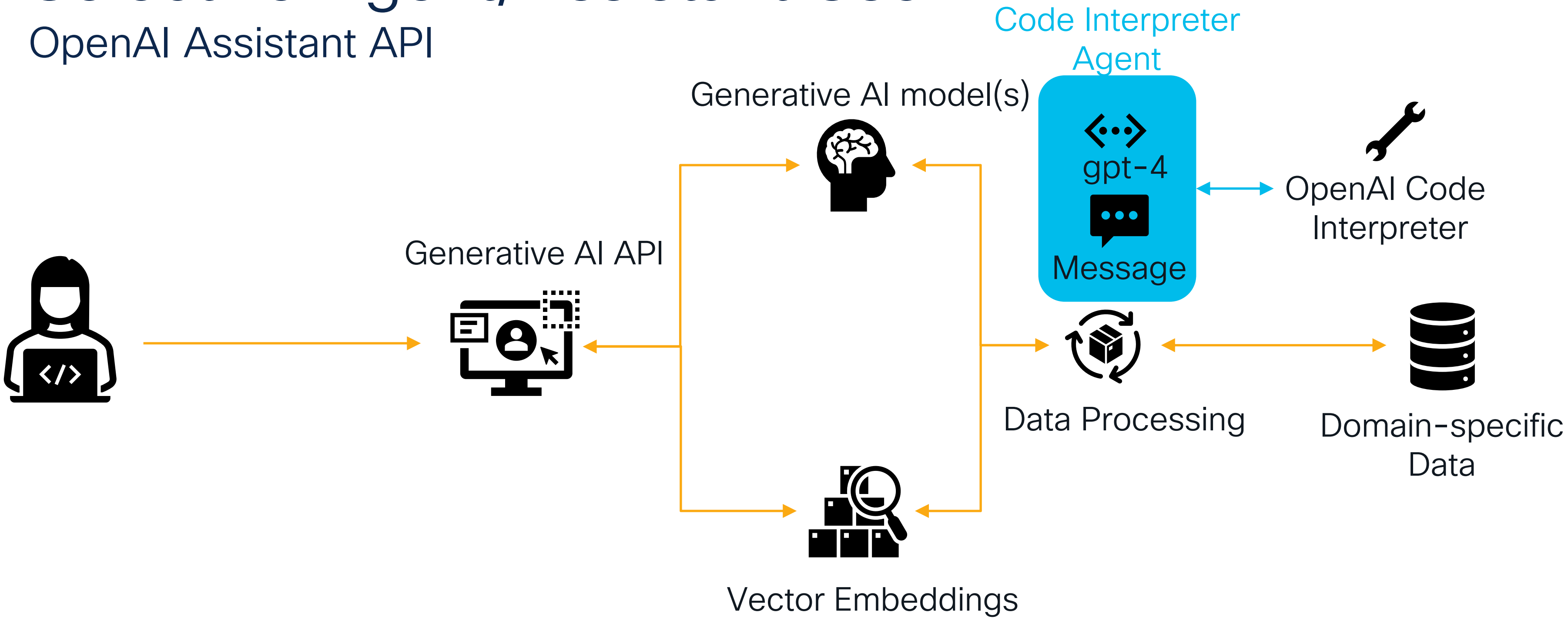
- Webex AI Assistant:
  - https://blog.webex.com/innovation/advanced-ai-powered-hybrid-work-platform/
  - https://www.webex.com/ai-assistant.html
- Chatbots
- Amazon Alexa
- Apple Siri
- Google Assistant
- Microsoft Cortana

API Sprawl


APIs
Generative AI model(s)
Tools
Generative AI API
Message
Data Processing
Domain-specific Data
Service
Tools
Vector Embeddings
GraphQL Gateway

Selective Agent/Assistant Use

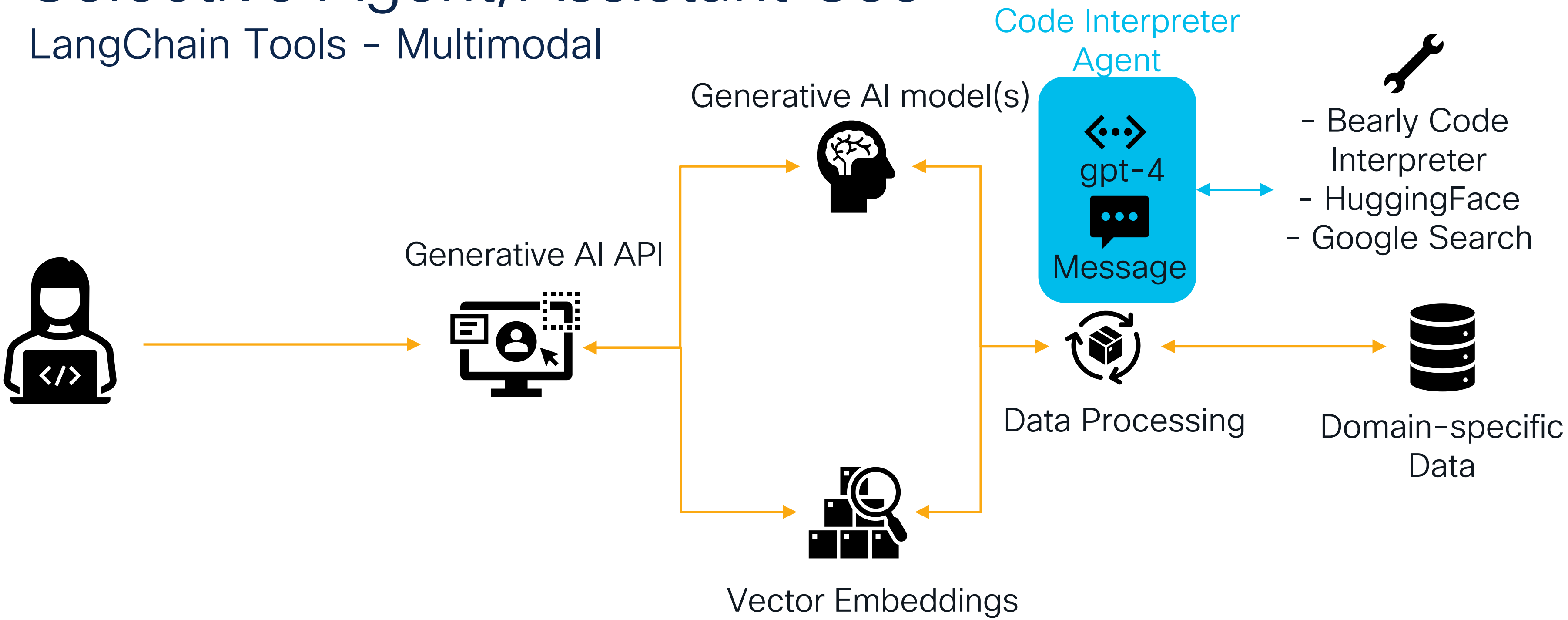
OpenAI Assistant API



OpenAI Assistant API: https://platform.openai.com/docs/assistants/how-it-works

Selective Agent/Assistant Use

LangChain Tools - Multimodal



OpenAI Assistant API: https://platform.openai.com/docs/assistants/how-it-works
LangChain + OpenAI Assistant: https://python.langchain.com/docs/modules/agents/agent_types/openai_assistants
LangChain Integrations: https://python.langchain.com/docs/integrations/tools

Your AI Strategy

Consuming vs. Building – It is likely both

Consuming

- Resources/Cost – Training is computationally heavy
- Data – Models may require large amounts of public data
- Expertise – Special skills & experience
- Time – Results are needed quickly
- Customization - Use the service “as-is”

Building

- Data Privacy – Prevent data leakage or the use of internal-only datasets
- Regulatory Compliance – Data Sovereignty, geo-specific laws
- Customization – Unique AI use for your industry
- Cost – Business-specific AI/ML-optimized infrastructure & staff

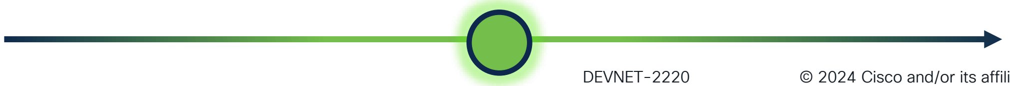
Consuming vs. Building – It is likely both

Consuming

- Resources/Cost – Training is computationally heavy
- **Data – Models may require large amounts of public data**
- Expertise – Special skills & experience
- Time – Results are needed quickly
- Customization - Use the service “as-is”

Building

- **Data Privacy – Prevent data leakage or the use of internal-only datasets**
- Regulatory Compliance – Data Sovereignty, geo-specific laws
- **Customization – Unique AI use for your industry**
- Cost – Business-specific AI/ML-optimized infrastructure & staff

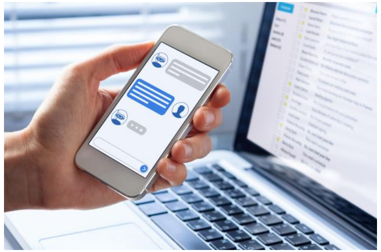


You Are Here

APIs + Agents + Assistants

- For now, you will be doing a lot of API work
- REST, GraphQL, WebSockets, etc. will still be used
- API Gateways are your friend
- Additional API security is needed for AI use cases

- Agents will grow in popularity
- Agents will talk to Agents: https://microsoft.github.io/autogen/docs/getting-started
- Agents will work with APIs and become more and more powerful



- Assistants will grow in capability, but will likely be relegated to more ‘simplistic’ functions
- They will work in combination with Agents and APIs as part of a full “stack”

Summary

- Some AI/ML APIs are proprietary, but a many are built on solid standards-based API specifications such as OpenAPI
- Things are moving fast – watch for broken APIs and lack of backward compatibility – changelogs are your friend
- As you mature your AI strategy, you will need to understand which APIs work best for which use cases
- Know the various API types such as REST, GraphQL, WebSocket, gRPC, etc..
- You will likely end up with a hybrid AI strategy:
  - Public AI Services
  - Internal AI Services (Commercial + Home Grown)

Reference

More Stuff to Learn

- Cisco AI Solutions: https://www.cisco.com/site/us/en/solutions/artificial-intelligence/index.html
- Cisco AI Security: https://www.cisco.com/c/en/us/products/security/artificial-intelligence-ai.html
- Cisco AI Observability: http://cs.co/9000RcAsy
- LangChain: https://python.langchain.com/docs/get_started/introduction
- Hugging Face: Inference API - https://huggingface.co/docs/huggingface_hub/guides/inference
- AWS AI Services: https://aws.amazon.com/ai/
- Azure AI Services: https://learn.microsoft.com/en-us/azure/ai-services/what-are-ai-services
- GCP AI Services: https://cloud.google.com/products/ai

Other Sessions

- DEVNET-2703: Securing APIs from Left to Right, and Everywhere in Between
- DEVWKS-1704: AI Code Warrior – Wielding Artificial Intelligence Tools as a Developer
- DEVNET-2708: Empowering Business with Security, Private and Sovereign AI: A Guide to Deploying Large Language Models
- DEVNET-2714: Explore Generative AI Capabilities
- DEVNET-3707: Network Telemetry and AI for Network Incident Response
- DEVNET-2850: Build an LLM-based Application in 45mins!

Continue to Learn, Code and Build with Cisco DevNet!

Get access to an exclusive learning module filled with digital learning opportunities on topics including Security and more.

Scan QR Code to get started.

The bridge to possible

Thank you

The bridge to possible

Thank you

cisco Live!

Let's go