Vectoring Into The Future: AWS Empowered RAG Systems for LLMs

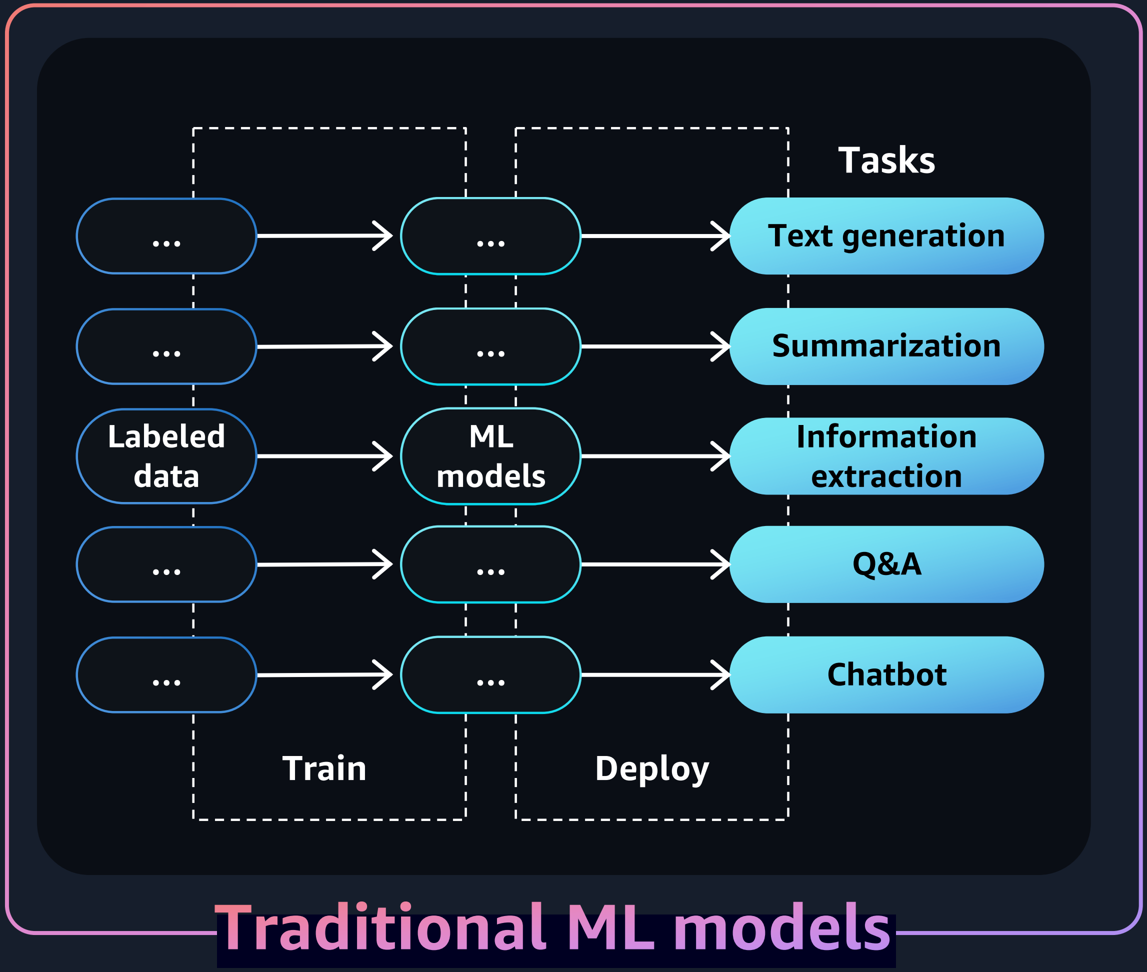
Samuel Baruffi

Principal Solutions Architect @ aws

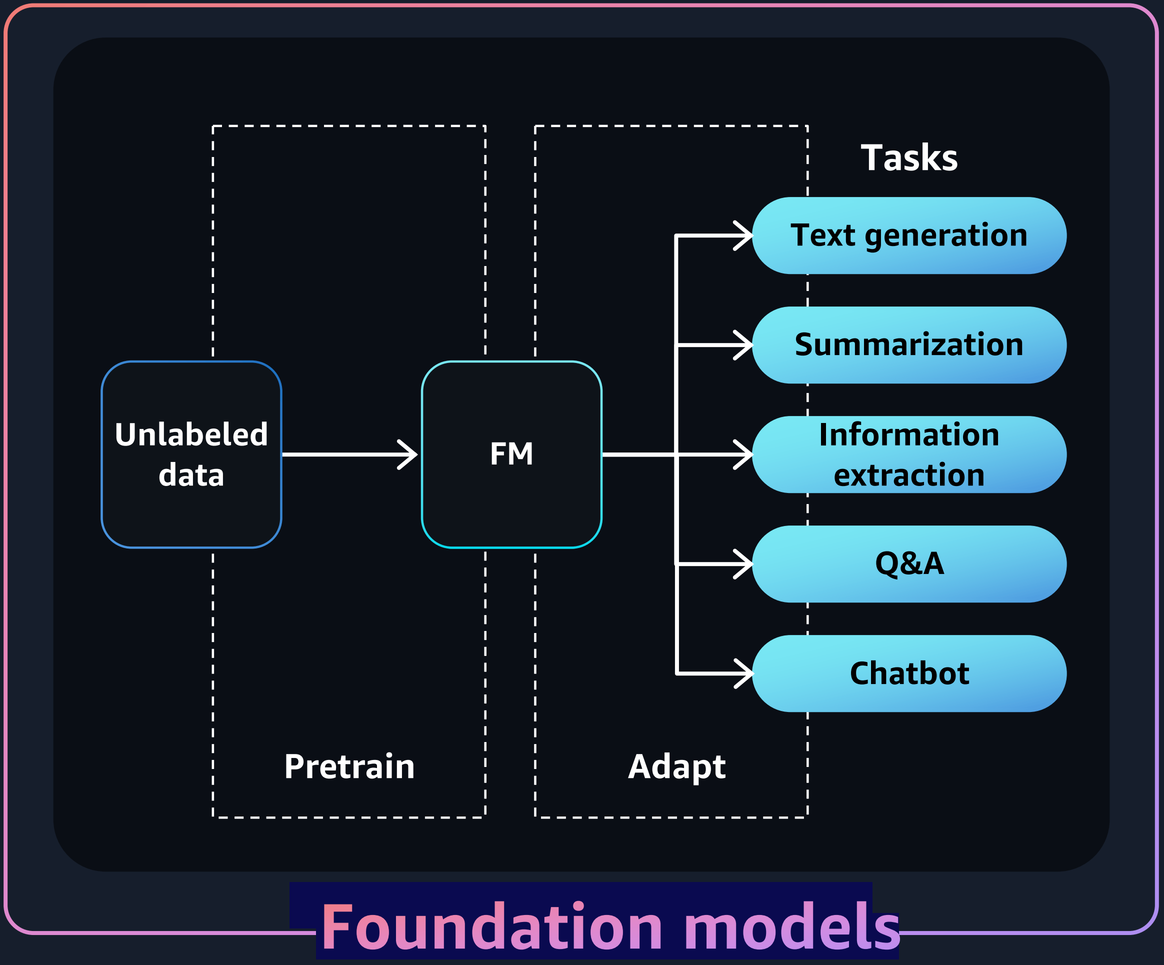
Agenda

- Foundational Models (LLMs)
- AWS GenAI Capabilities
- Limitations of LLMs
- AWS Vector Databases Offering
- Amazon Bedrock
- Amazon Bedrock Knowledge Base
- Demo

Why foundation models?



Traditional ML models



Foundation models

Generative AI can be used for a wide range of use cases

- Chatbots & virtual assistants
- Agent assist
- Contact center analytics
- Personalization

Enhance customer experience

- Conversational search
- Content localization
- Text, image, video generation
- Text summarization
- Code generation

Boost employee productivity

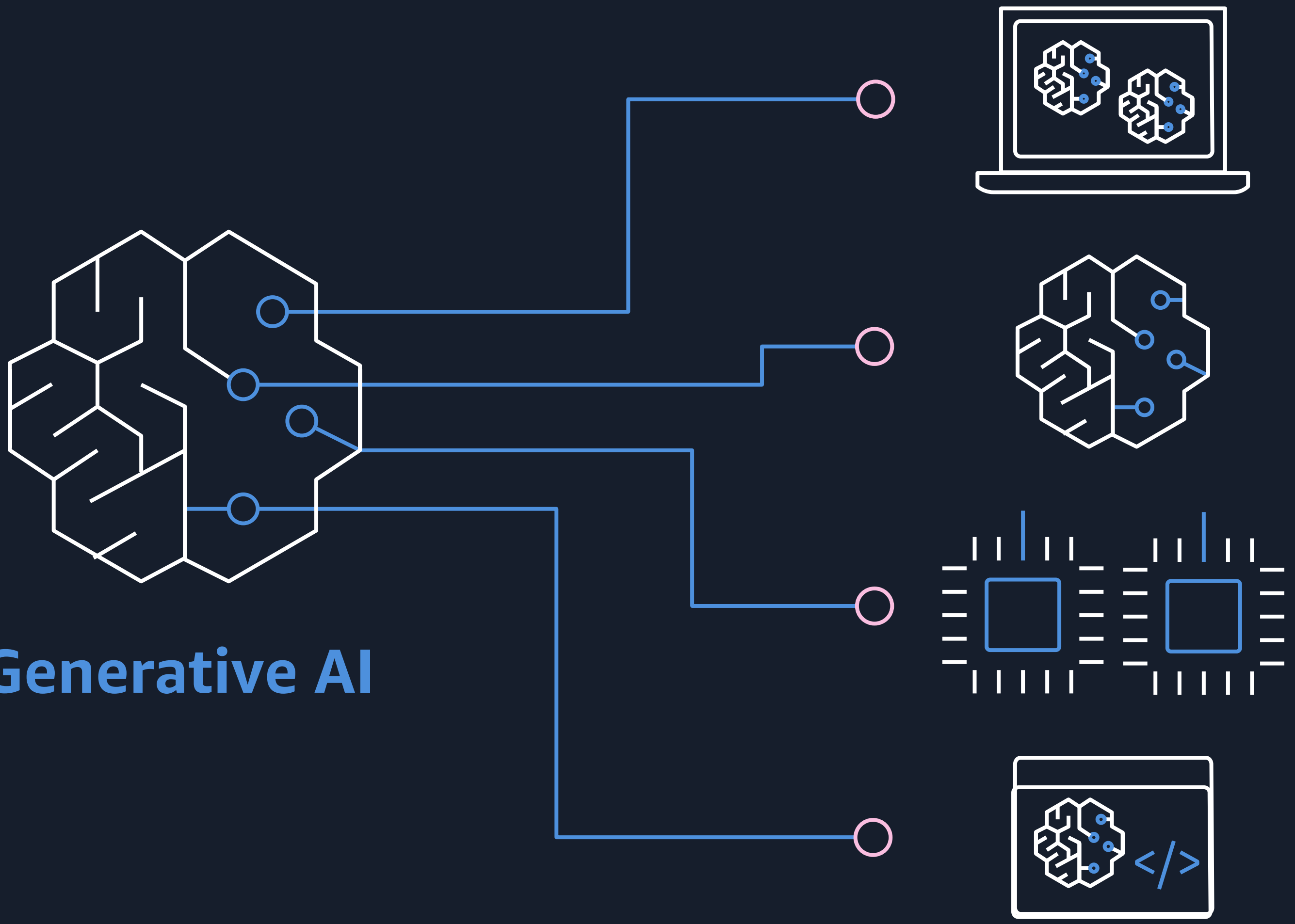
- Document processing
- Content moderation
- Synthetic data creation
- Maintenance assistance
- Anomaly detection

Improve business operations

- Image generation for web pages
- Video enhancement
- Music creation
- Image enhancement
- Creating animations

Creativity

AWS offers a broad choice of generative AI capabilities



Generative AI

Amazon SageMaker foundation model hub

Deployment and fine-tuning of open source and third-party FMs

Amazon Bedrock

API access to fully managed first-party and third-party FMs

Amazon EC2 Trn1n and Amazon EC2 Inf2

Specialized chips built for best cost/performance model training and serving

Amazon CodeWhisperer

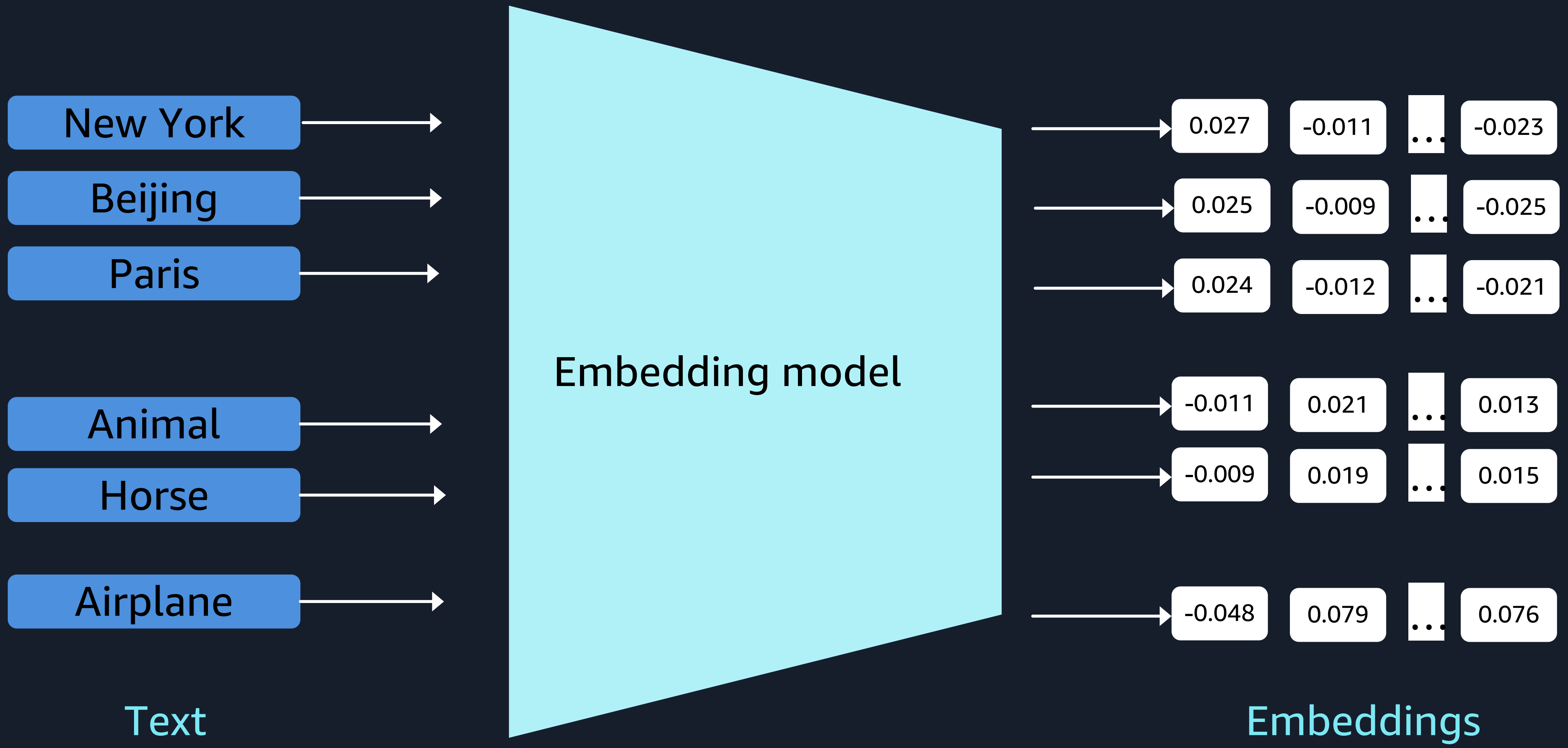
Generative AI-powered coding companion for automatic code completion

Limitations of LLMs

- Limited contextual understanding
- Lack of domain-specific knowledge
- Lack of explainability and interpretability
- Inaccurate information

Vector embeddings

What are vector embeddings?



Embedding model

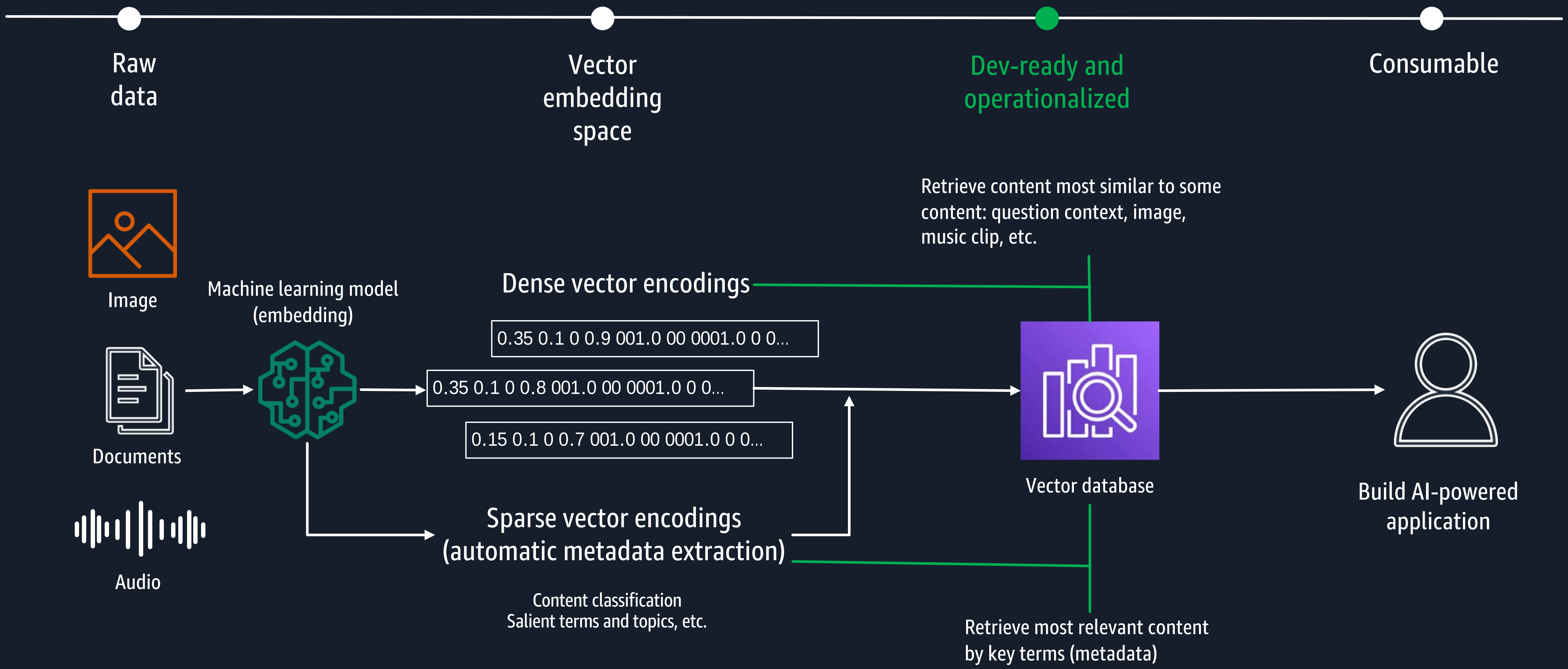
New York	0.027	-0.011	…	-0.023
Beijing	0.025	-0.009	…	-0.025
Paris	0.024	-0.012	…	-0.021
Animal	-0.011	0.021	…	0.013
Horse	-0.009	0.019	…	0.015
Airplane	-0.048	0.079	…	0.076

Text

Embeddings

Vector databases

What is a vector database?



Raw data

Vector embedding space

Dev-ready and operationalized

Consumable

Image

Documents

Audio

Machine learning model (embedding)

Dense vector encodings

0.35 0.1 0 0.9 001.0 00 0001.0 0 0...

0.35 0.1 0 0.8 001.0 00 0001.0 0 0...

0.15 0.1 0 0.7 001.0 00 0001.0 0 0...

Sparse vector encodings (automatic metadata extraction)

Content classification
Salient terms and topics, etc.

Retrieve content most similar to some content: question context, image, music clip, etc.

Vector database

Retrieve most relevant content by key terms (metadata)

Build AI-powered application

Enabling vector search across AWS services

Amazon OpenSearch Service

Amazon OpenSearch Serverless

Amazon Aurora PostgreSQL

Amazon RDS for PostgreSQL

Amazon DocumentDB

Amazon DynamoDB via zero-ETL

Amazon MemoryDB for Redis

Amazon Neptune Analytics

Amazon Aurora with PostgreSQL compatibility

Vector data stores in AWS

High performance, cloud-native RDBMS

Provisioned and serverless deployment options

Vector capabilities provided by **pgvector** extension

Supports k-NN and ANN with HNSW and IVFFlat

For PostgreSQL apps, no driver changes needed

Ideal for existing PostgreSQL users, or any users who prefer relational DBs

Using pgvector in AWS

Vector data stores in AWS

Available in both Amazon **Aurora PostgreSQL** compatible and Amazon **RDS for PostgreSQL**

Aurora is **integrated** with Amazon **Bedrock knowledge base**, Amazon SageMaker and Amazon Comprehend via **Aurora ML**

Configurable recall rate via HNSW ef_search, IVFFlat probes

Scalable to support **over 1 billion vectors** & **16,000 dimensions** (2,000 indexed)

Amazon OpenSearch Service

Vector data stores in AWS

Search and analytics engine

Managed service or serverless deployment options

Vector capabilities provided by the **k-nn** plugin

Supports k-NN and ANN with HNSW and IVFFlat

Vectorize Amazon DynamoDB data using Zero-ETL

Ideal for OpenSearch users, users who prefer NoSQL, or hybrid search uses

Using OpenSearch in AWS

Vector data stores in AWS

Available as **Amazon OpenSearch Service** (provisioned domains with k-nn plugin) and **Vector engine for Amazon OpenSearch Serverless**

OpenSearch is **integrated** with Amazon **Bedrock knowledge base**, the **quick create** feature uses OpenSearch Serverless vector search collections. OpenSearch's **Neural Search plugin** provides **seamless text to vector** transformation via external LLM

Configurable recall rate via segments and NMSLIB ef_search

Scalable to support **over 1 billion vectors** & **16,000 dimensions** (max. 1,024 for Lucene engine)

Amazon DocumentDB

Vector data stores in AWS

- Fast cloud-native document database
- MongoDB compatible
- Provisioned deployment option
- Supports k-NN and ANN with IVFFlat

Ideal for existing DocumentDB and MongoDB users

Amazon MemoryDB

⚠ **Vector search feature is released in preview**

Vector data stores in AWS

Fully durable, in-memory cloud-native database

Redis compatible

Provisioned deployment option

Supports k-NN and ANN with HNSW

Up to 32,768 dimensions

Ideal for Redis users, workloads requiring in-memory latencies & throughput

Amazon Neptune Analytics

Vector data stores in AWS

- Analytical, memory-optimized graph DB engine
- Discrete capacity deployments*
- HNSW similar algorithm
- Up to 65,536 dimension vectors
- Complements Amazon Neptune Database

Ideal for graph neural network use cases, vector search in graph traversals

*Amazon Neptune Database is also available for serverless deployments; Amazon Neptune Analytics supports only discrete capacity levels

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models

Choice of industry-leading FMs available via a single API

Customize your models using your organization’s data

Enterprise-grade security and privacy

Amazon Bedrock

Broad choice of models

AI21 labs	amazon	ANTHROPIC	cohere	Meta	MISTRAL AI_	stability.ai
Jurassic-2 Ultra Jurassic-2 Mid	Titan Text Embeddings Titan Multimodal Embeddings Titan Text Lite Titan Text Express Titan Image Generator	Claude 3 Claude 2.1 Claude 2 Claude Instant	Command + Embed Cohere Command Light Cohere Embed English Cohere Embed Multilingual	Llama 2 Llama 2 13B Llama 2 70B	Mistral 7B Mixtral 8x7B	Stable Diffusion XL1.0
Contextual answers, summarization, paraphrasing	Text summarization, generation, Q&A, search, Image generation	Summarization, complex reasoning, writing, coding	Text generation, search, classification	Q&A and reading comprehension	Text summarization, Q&A, Text classification, Text completion, code generation	High-quality images and art

Knowledge bases for Amazon Bedrock

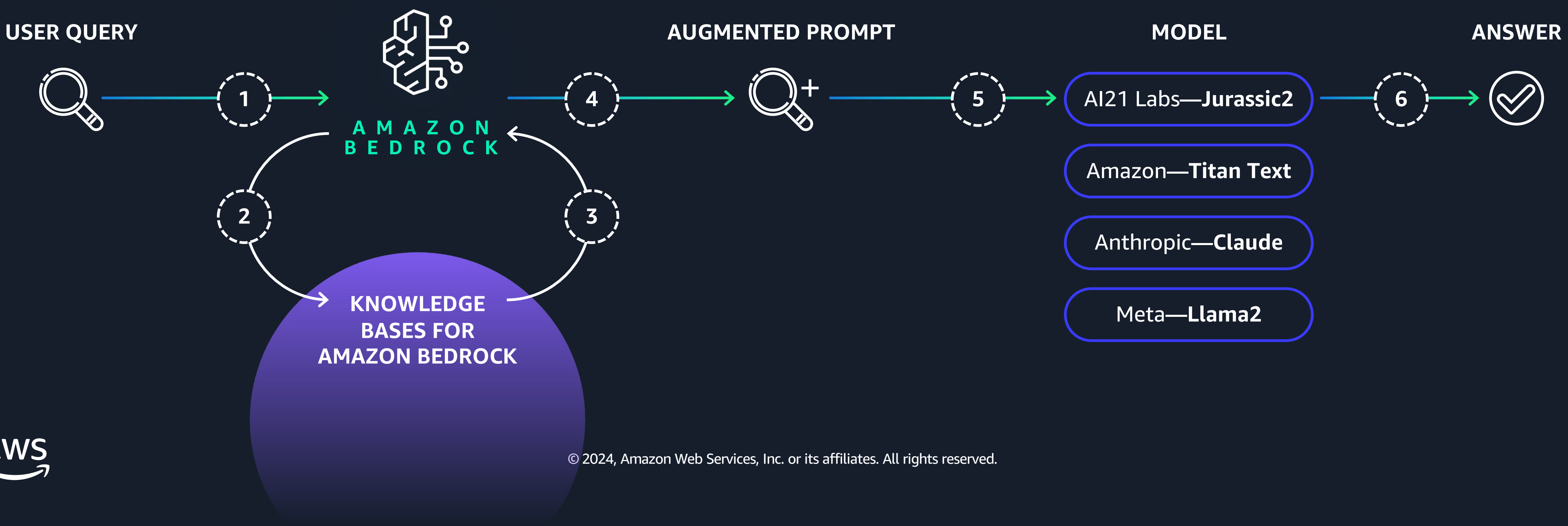
Native support for Retrieval Augmented Generation (RAG)

Securely connect FMs to data sources for RAG to deliver more relevant responses

Fully managed RAG workflow including ingestion, retrieval, and augmentation

Built-in session context management for multi-turn conversations

Automatic citations with retrievals to improve transparency



Vector databases for Amazon Bedrock

Vector Engine For Amazon OpenSearch Serverless

Redis Enterprise Cloud

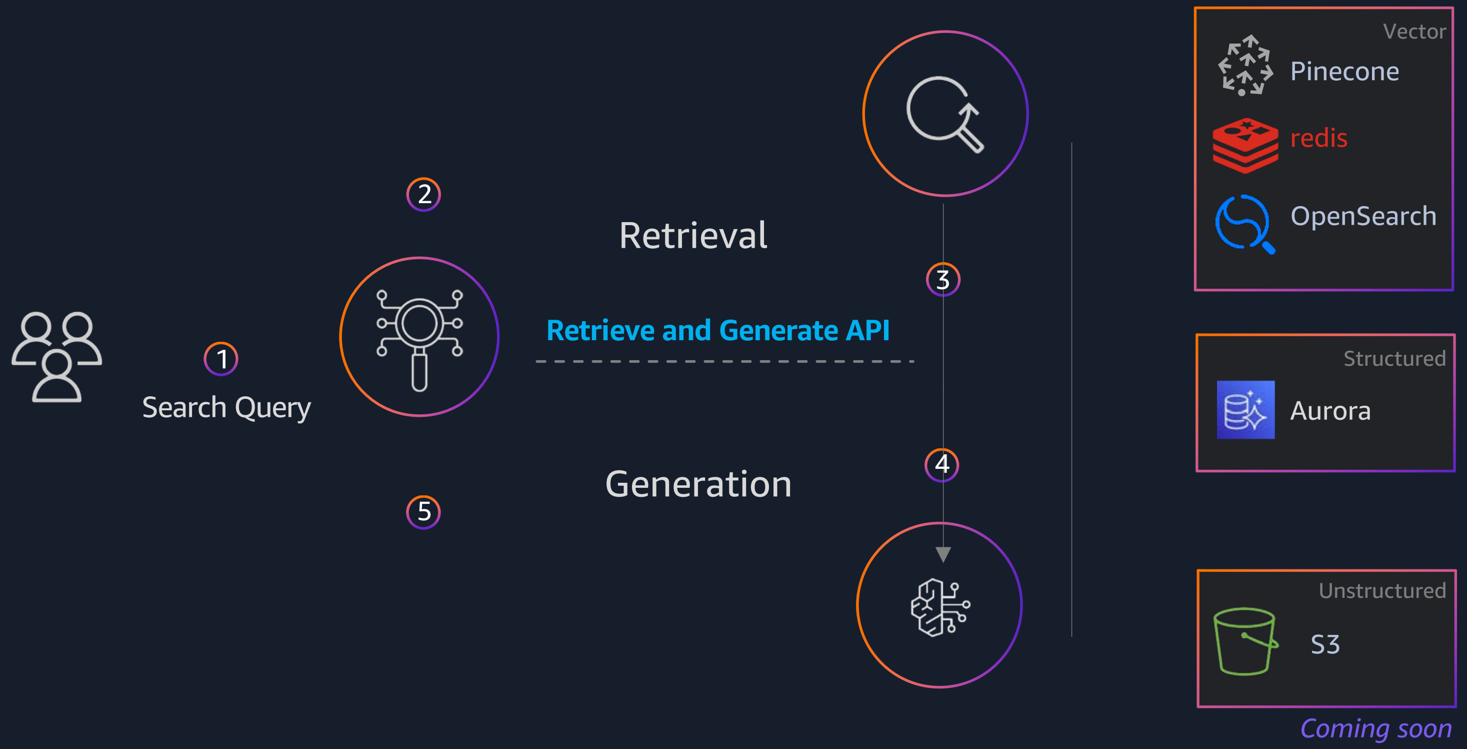
Pinecone

Amazon Aurora

COMING SOON

MongoDB

Retrieve and Generate API



1 Search Query

2

Retrieval

3

Retrieve and Generate API

4

Generation

5

Vector

Pinecone

redis

OpenSearch

Structured

Aurora

Unstructured

S3

Coming soon

✓ **Single API**	One API for RAG
✓ **Comprehensive**	Vector, Structured data
✓ **Session**	Built-in session management

Retrieve and generate API will enable a simplified RAG solution

Demo Time