



AWS Certified AI Practitioner AIF-C01

Pass the Exam

+

Learn to Build Real World GenAI Apps !!!

Course Overview mapped to AWS Exam Guide

- Course Introduction
- Domain 1 - Fundamental of AI and ML
- Domain 2 - Fundamental of Generative AI
- Domain 3 - Applications of Foundation Models
- Domain 4 - Generative AI Application Lifecycle
- Domain 5 - Guidelines for Responsible AI
- Domain 6 - Security, Compliance, and Governance for AI Solutions



Domain 1 :

Fundamentals of AI and ML (Part 1)

Artificial Intelligence - Overview



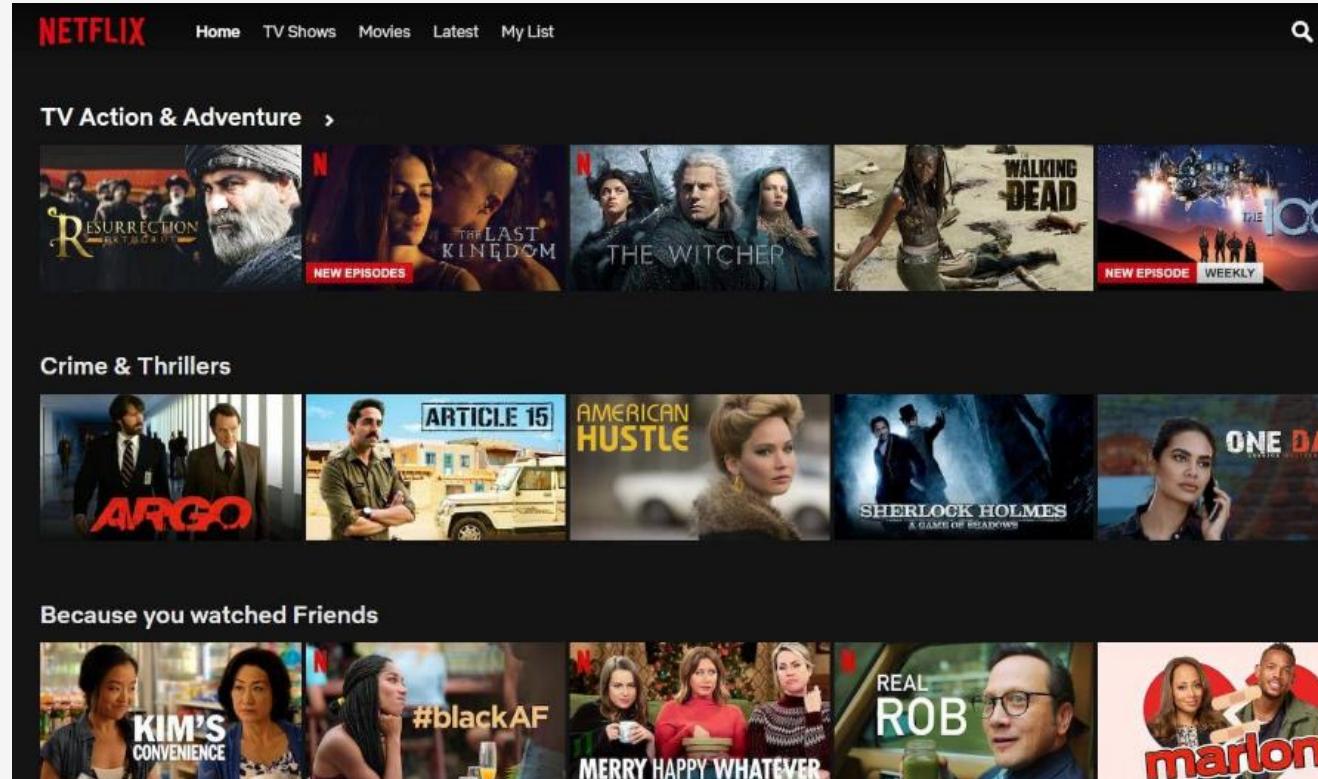
Artificial Intelligence is the science and engineering of **making** intelligent machines, especially **computer programs** that **simulate human decision making.**

Artificial Intelligence - Examples



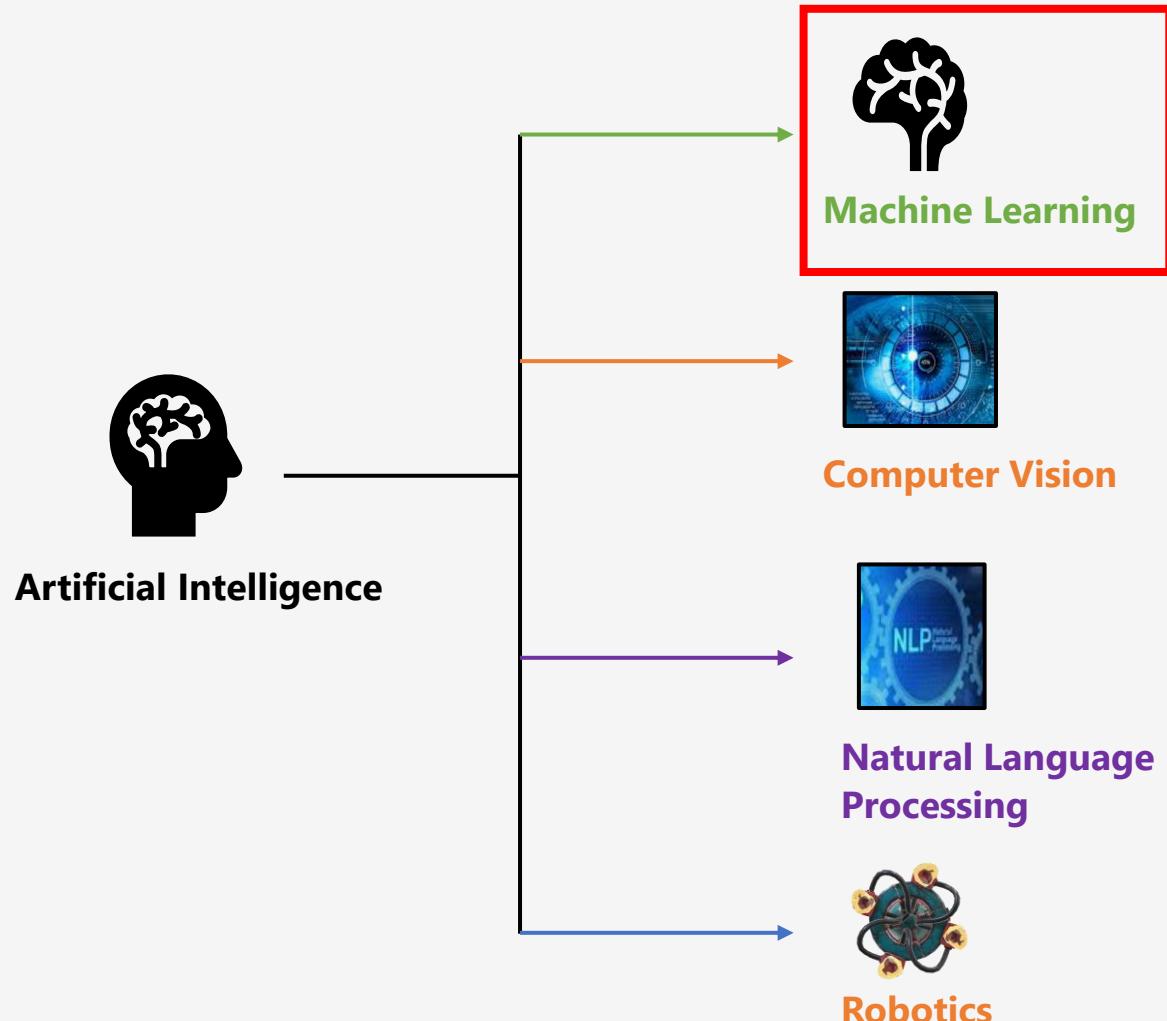
Self-driving cars use artificial intelligence to make real-time decisions based on sensor data.

Artificial Intelligence - Examples

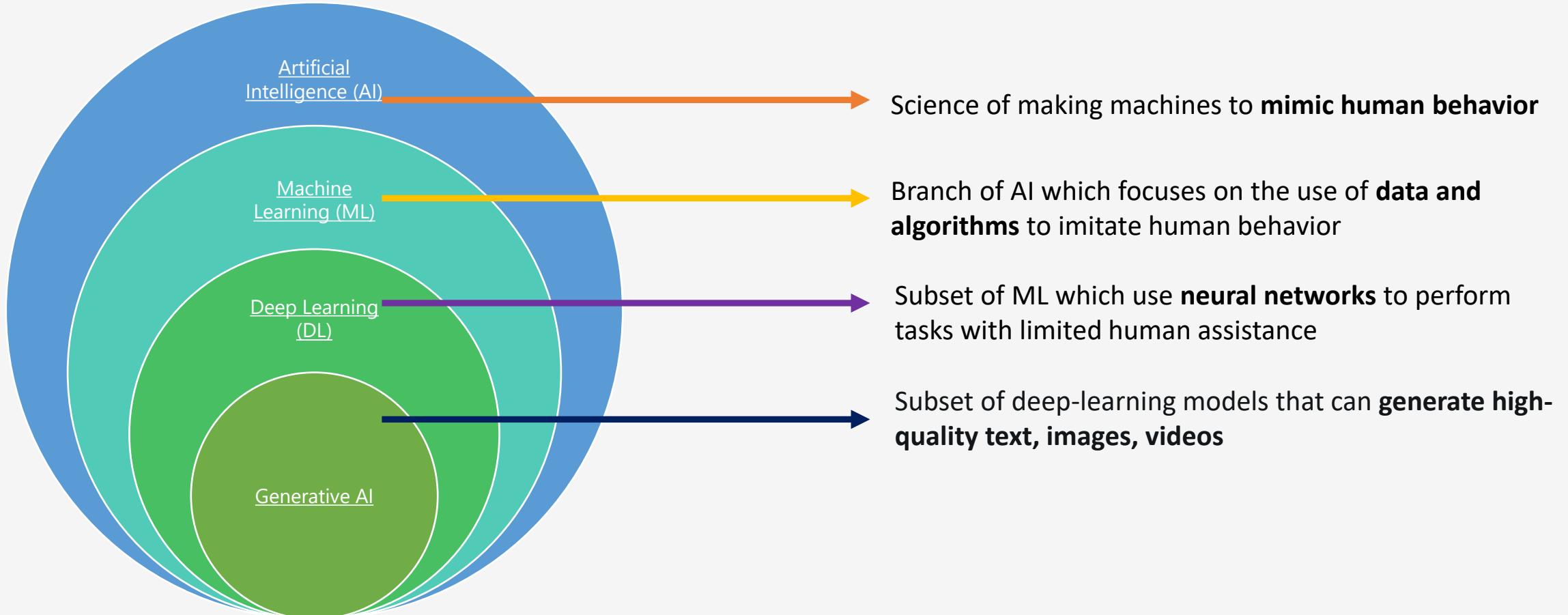


Recommendation engines such as ones from **Netflix and Amazon** use AI to provide custom insights

Applications of Artificial Intelligence



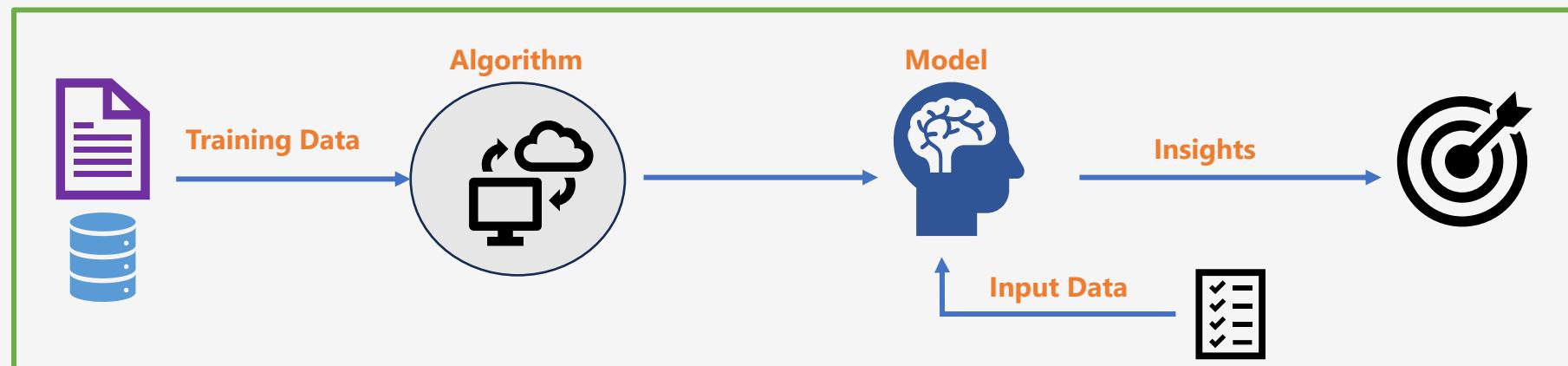
Artificial Intelligence and Machine Learning - Overview



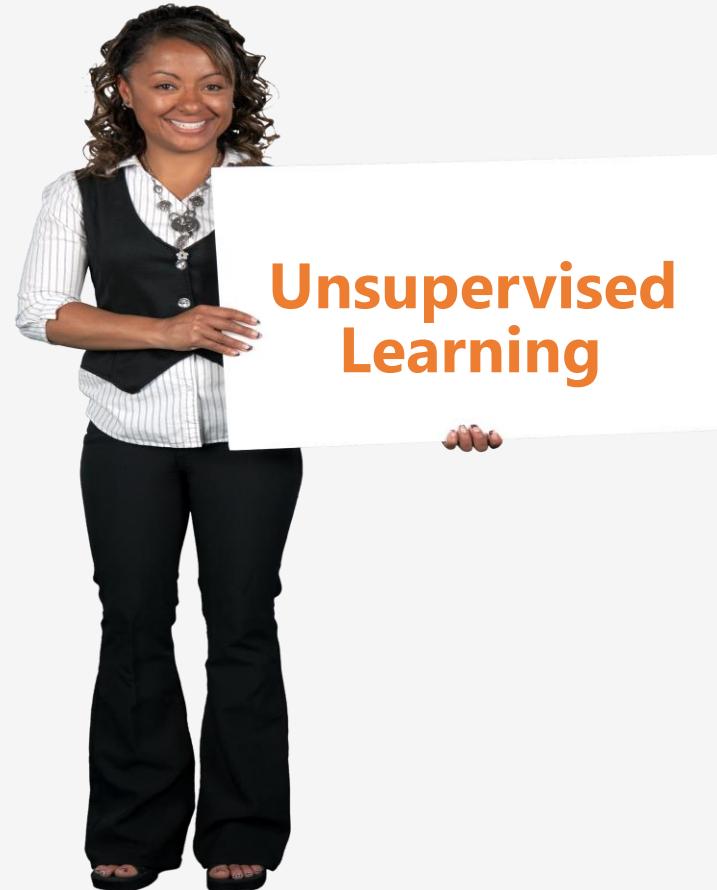
Machine Learning - Overview



- Machine Learning is a **subset of AI**
- In ML the **algorithm is trained using the historical data** to make **predictions on new data.**

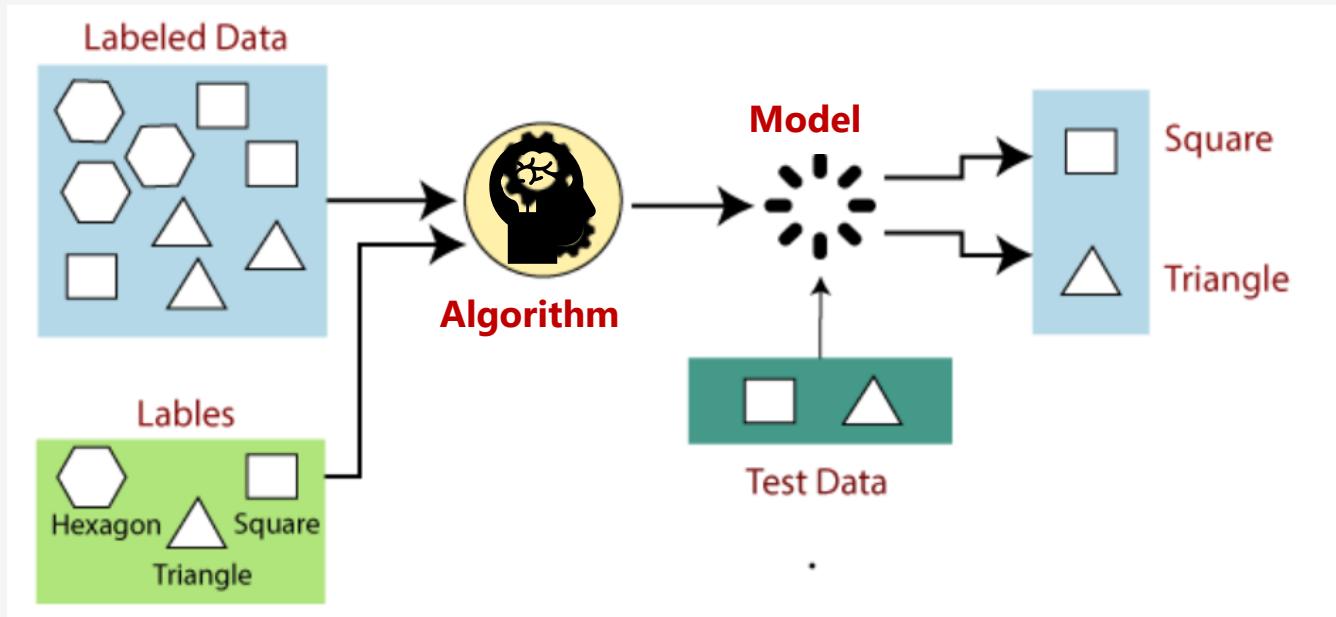


Machine Learning – Broad Categorization

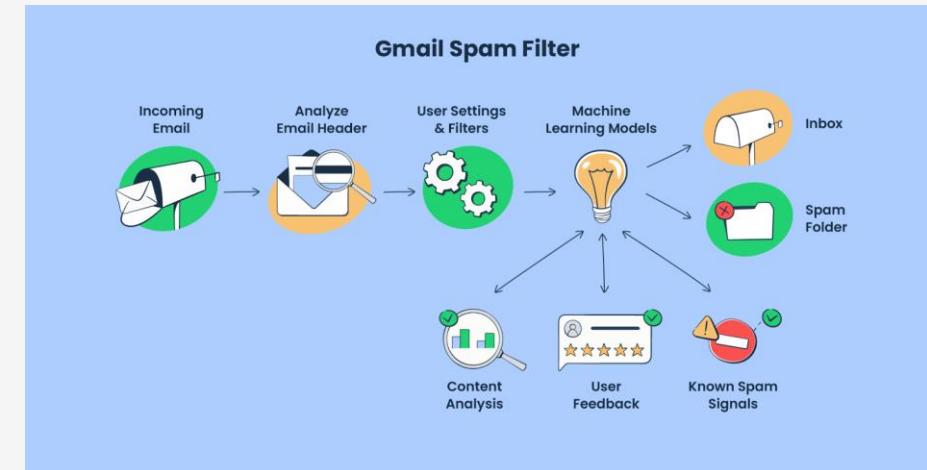


Machine Learning – Supervised Learning

- In supervised learning, **algorithms are trained using labelled dataset.**
- Once the **training is completed**, the **model is tested using test data** and then it predicts the output.



Use Case (Classification) – Gmail Spam Filter



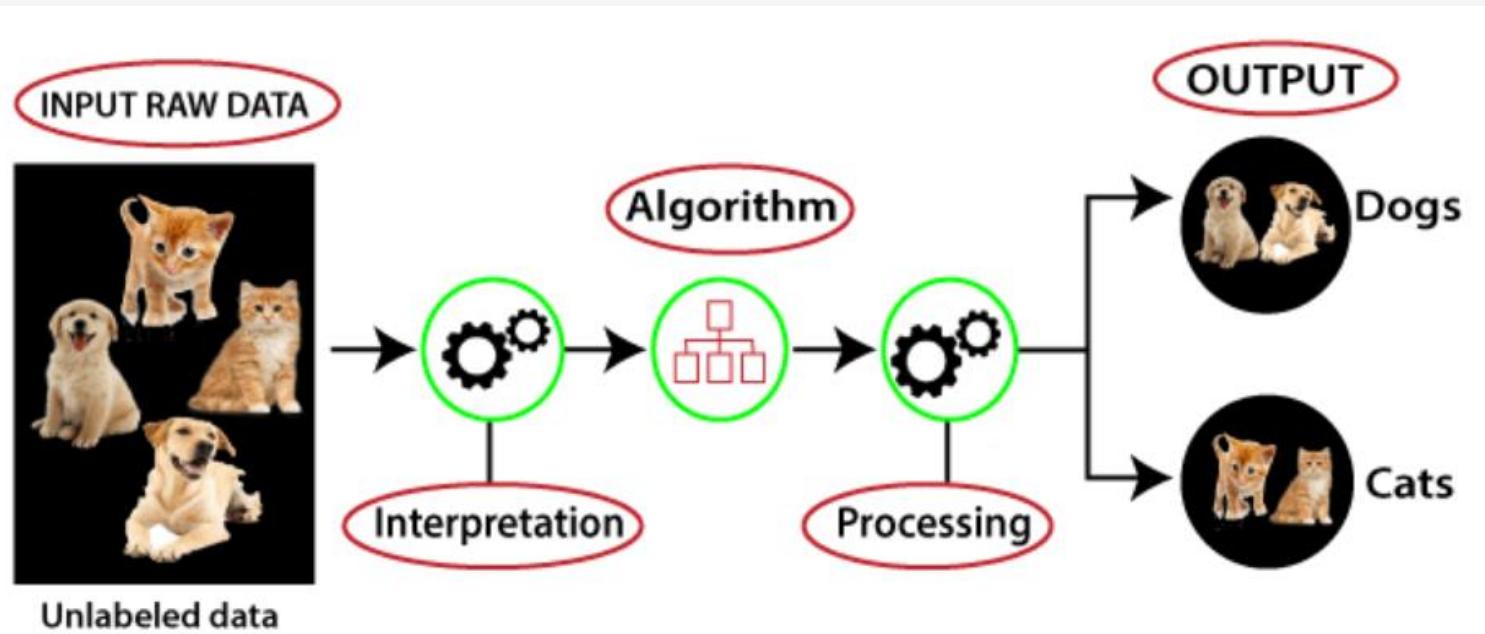
Advertisement	Sales
\$90	\$1000
\$120	\$1300
\$150	\$1800
\$100	\$1200
\$130	\$1380
\$200	??

Use Case (Regression) – Forecasting

Common Algorithms - Classification and Regression

Machine Learning – Unsupervised Learning

- Unsupervised learning is a type of machine learning in which **algorithm are trained using unlabeled dataset**
- Based on the dataset, model finds **hidden patterns and insights** based on the **similarities of the data**



Common Algorithms - Clustering and Association

Use Case (Clustering) – Reduce Churn Rate



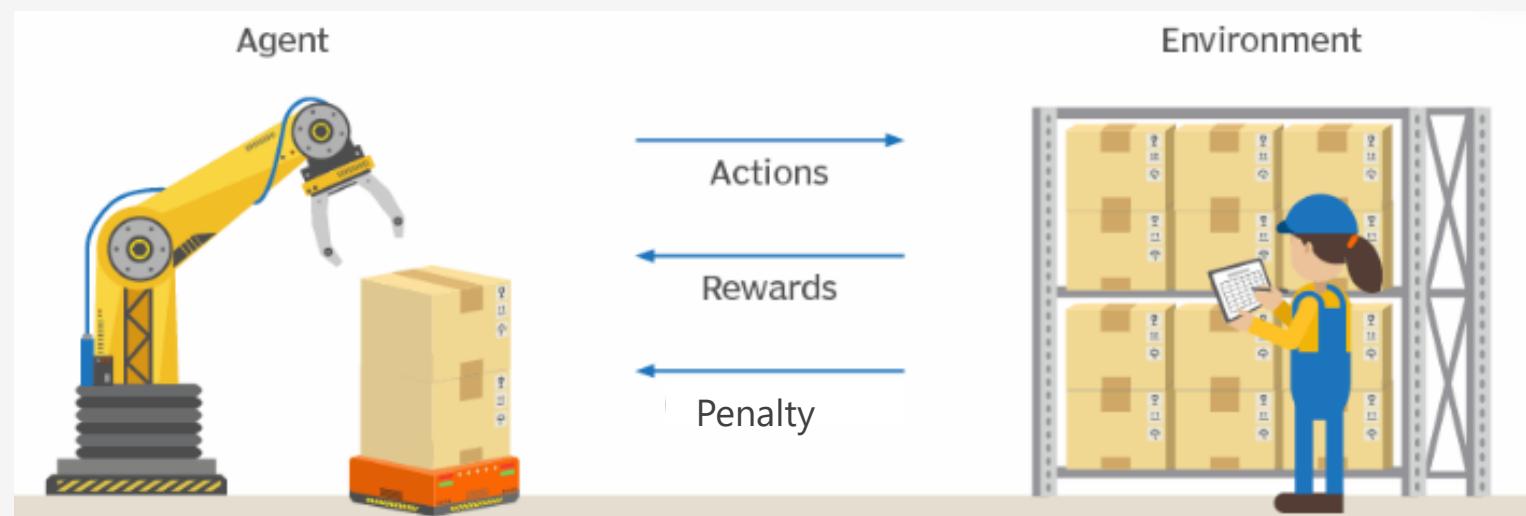
Use Case (Association) – Market Basket Analysis



Machine Learning – Reinforced Learning



- Reinforcement learning (RL) is a ML technique that focuses on **training an algorithm following the trial and error approach**.
- No pre-defined dataset
- The **algorithm (agent)** evaluates a current situation (state), **takes an action**, and receives feedback (reward and feedback) from the environment after each act.

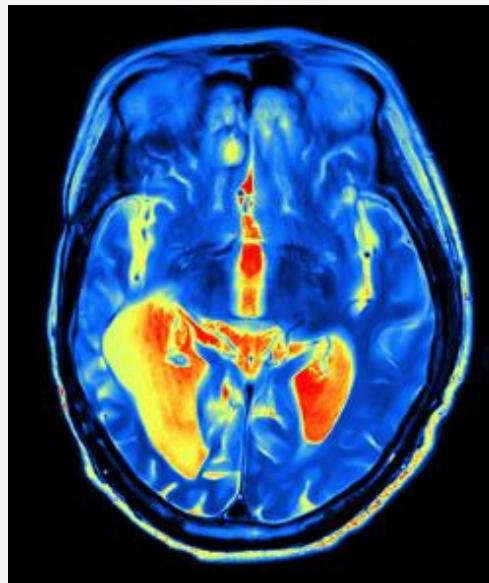


Real World Use Case

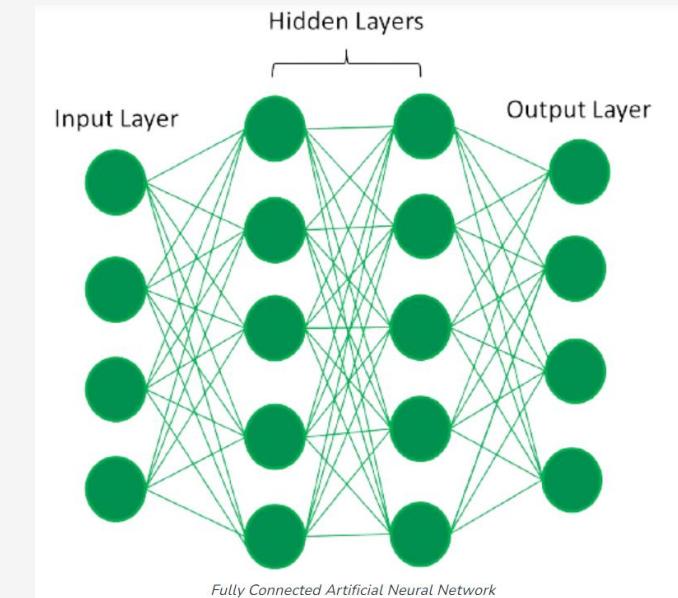
- Training Industrial Robots
- Training Self Driving Cars

Machine Learning – Deep Learning and Artificial Neural Networks

Neural Pathways in Human Brain

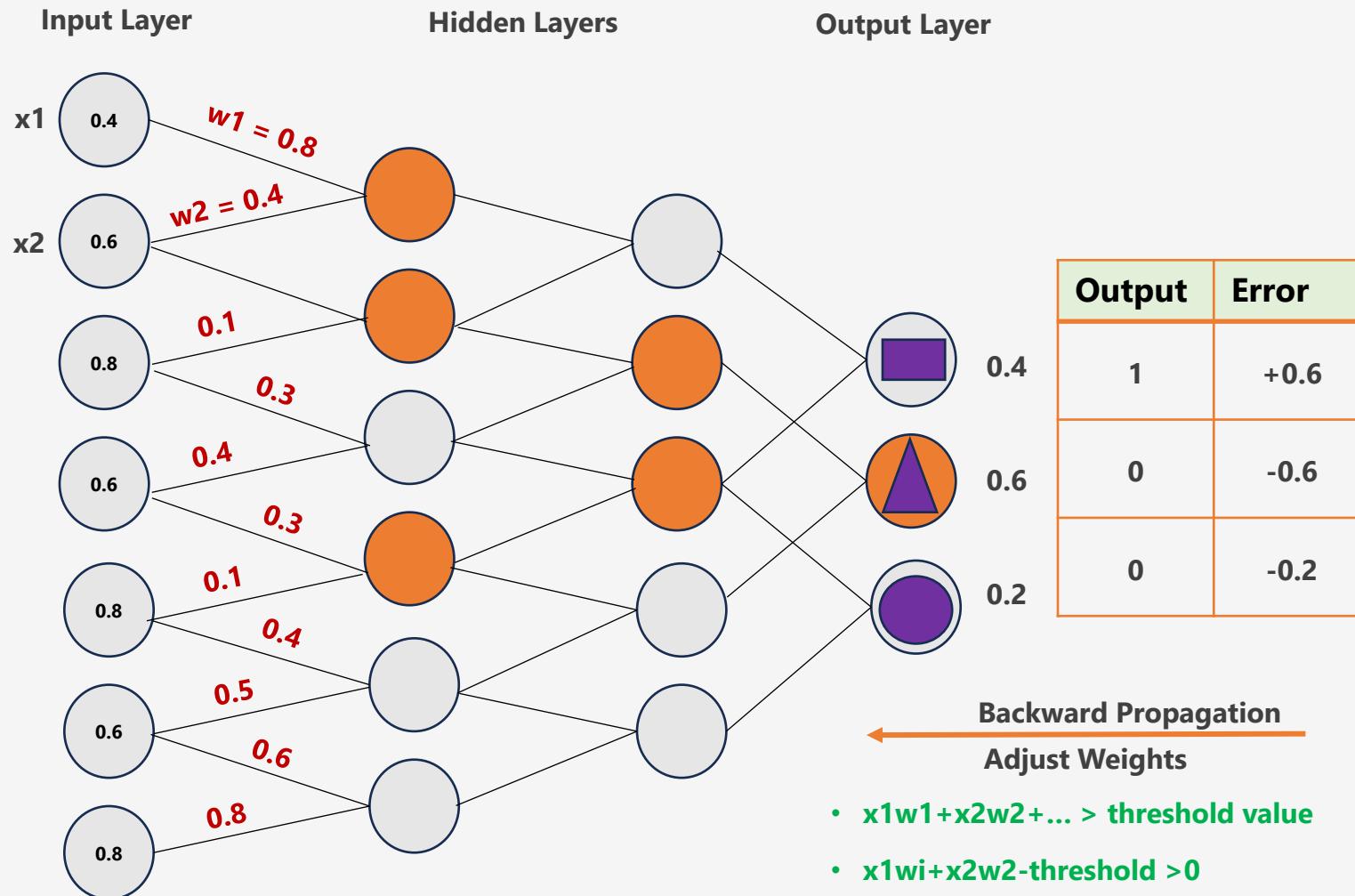
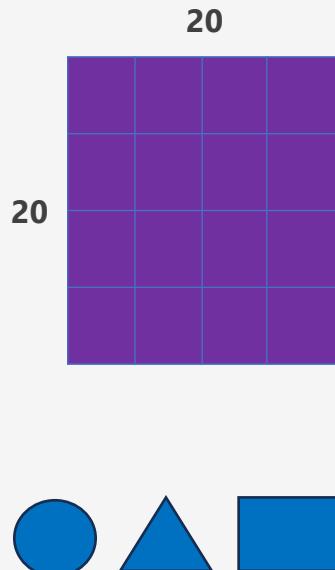


Artificial Neural Network

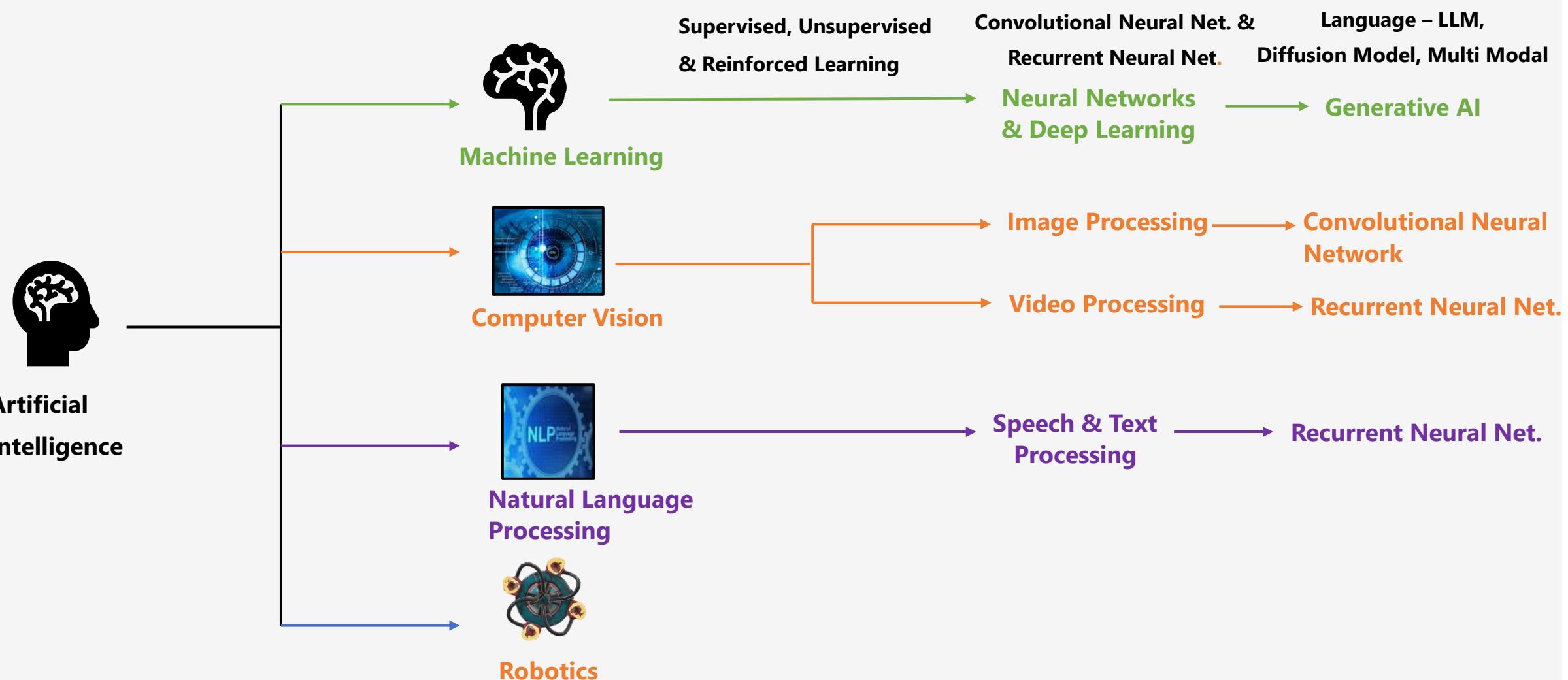


- **Artificial Neural networks** are designed to **mimic the structure of the human brain**
- Deep Learning is a subset of Machine Learning that **focuses on building artificial neural networks** that **can learn from data**.
- Neural network with **multiple hidden layers** and **multiple nodes** in each hidden layer is known as a **deep learning (>3)**

Machine Learning – Deep Learning with Artificial Neural Networks



Overview of Artificial Intelligence Landscape – Mental Map



What is Computer Vision ?

- **Field** of Artificial Intelligence
- Allows **derive meaningful information** from

Images, Videos

- Make **recommendations** or take actions

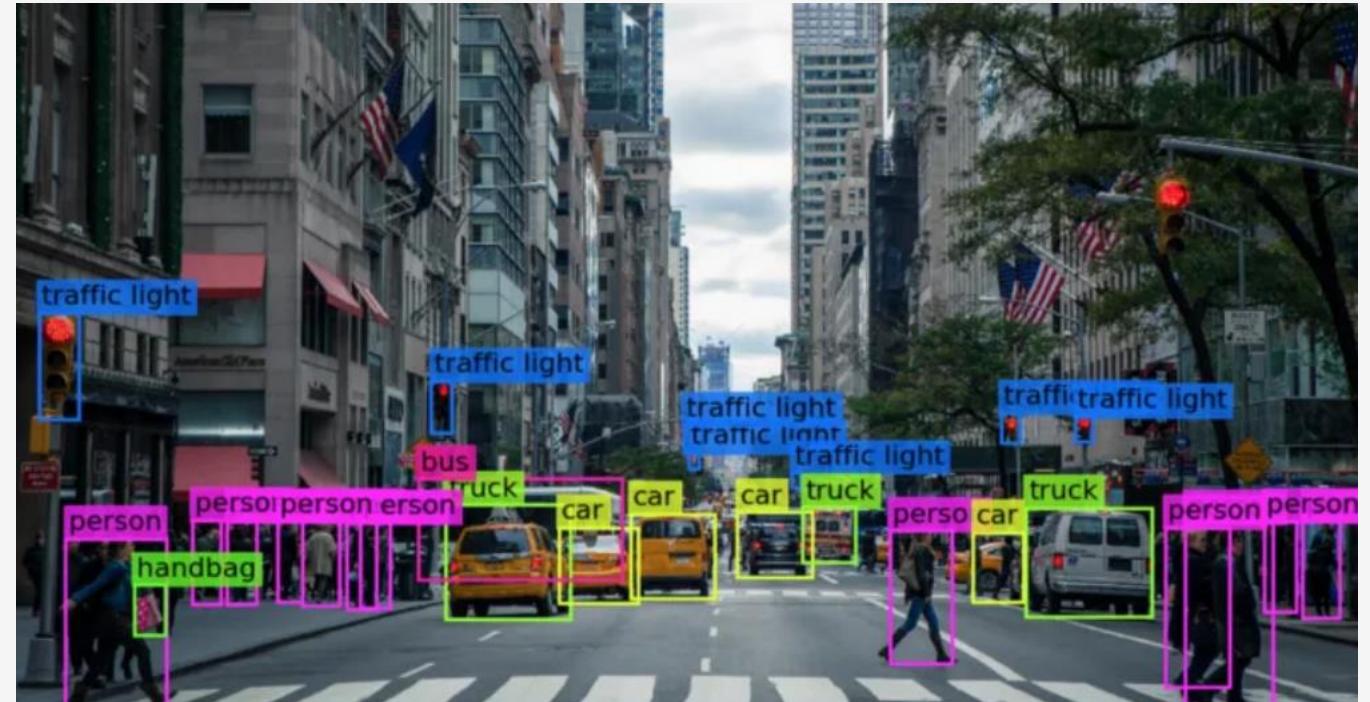
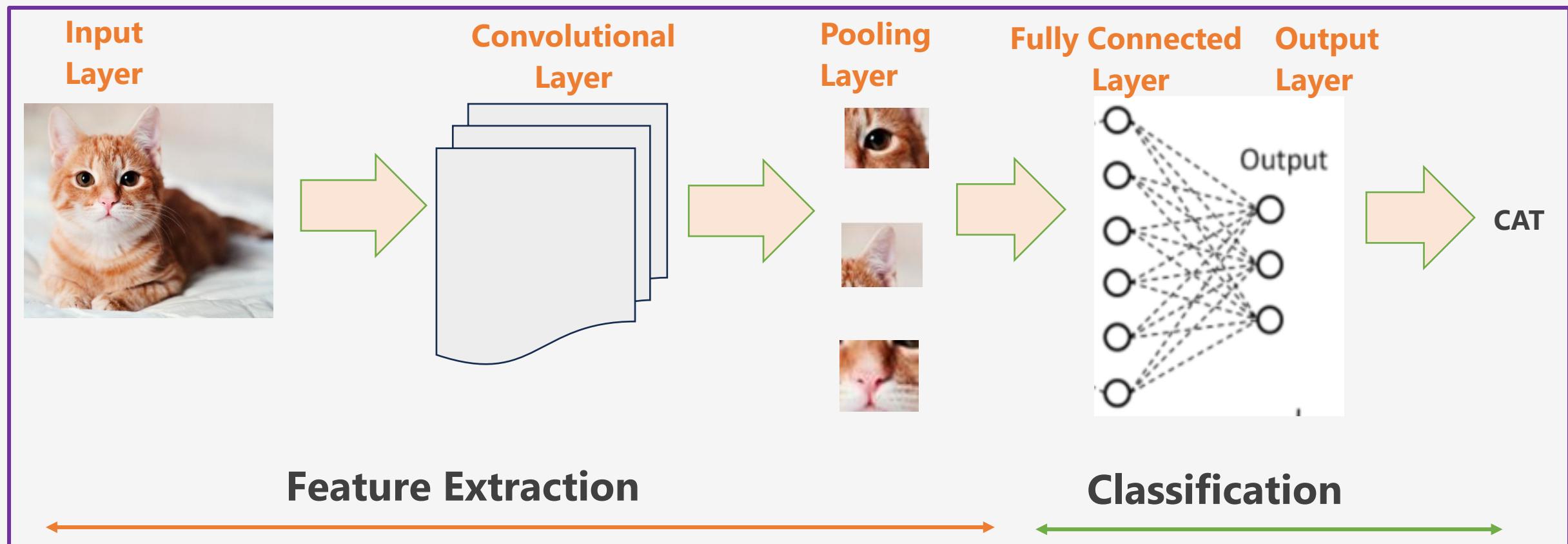


Image processing using Convolutional Neural Network

- Machine Learning – **Deep Learning**
- **Convolutional Neural Network (CNN)** – Deep learning algorithm for image processing



Computer Vision – Capabilities

Computer Vision capabilities with images and videos



Image Classification

Categorize Images – Content Moderation



Object Detection

Object & Location – Self driving cars



Image Segmentation

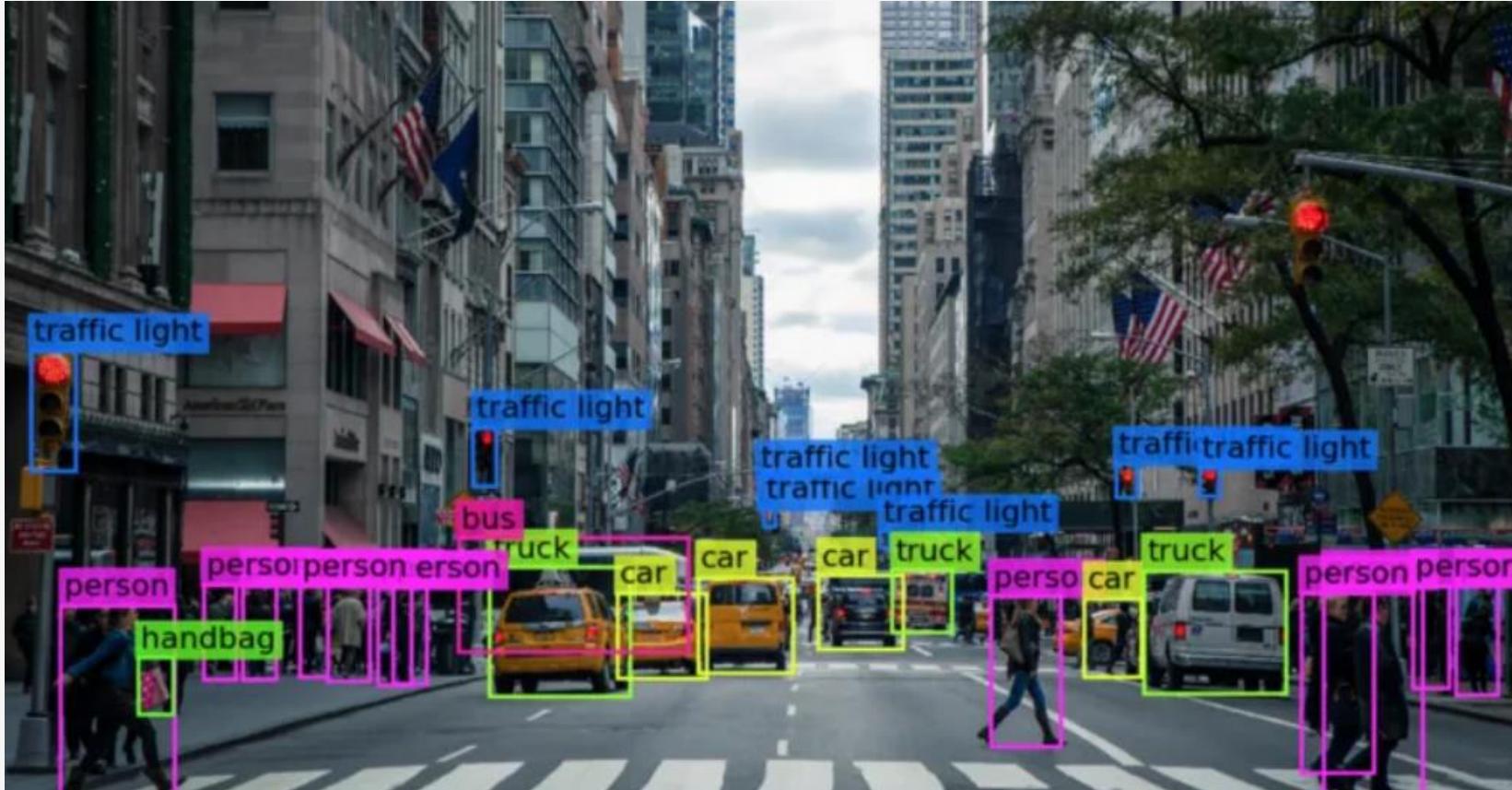
Multiple Images identification - Radiology



Facial Recognition

Identify human faces – Cyber Security

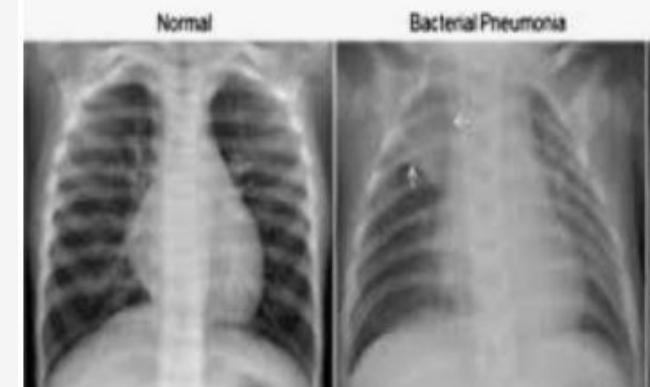
Computer Vision – Key Tasks



Computer Vision – Industry Use Cases

1. Retail - Self-checkout and cashierless store experience - [Link](#) -

- Image Classification
- Object(Person) Detection



2. Healthcare - Medical Image Analysis

- Identify abnormalities such as tumors, blood clots
- Image Segmentation

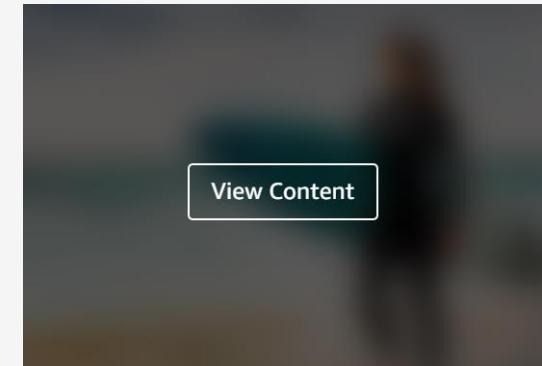
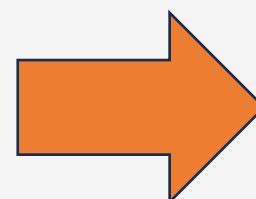
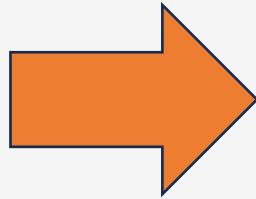


3. Manufacturing - Smart Factories and Self Driving Cars

- Product inspection, and defect detection

AWS Services for Computer Vision – Amazon Rekognition

Enables **Computer vision capabilities** to extract information and **insights** from **images & videos**.



Amazon Rekognition

Common Use Cases - [Link](#)

- Content Moderation - Detect inappropriate content
- Verify identity online – Onboard new customers through video analysis
- Streaming media analysis – Analyse and tag videos for celebrities, people etc.

What is Natural Language Processing ?

Natural Language Processing(NLP) :

- Subset of **Artificial Intelligence** (AI)
- Uses **Machine Learning** to **enable computers** to
communicate in Human/Natural language.
- Focuses on Language (vs Computer Vision – Images and Videos)
- Use Deep Learning Technique - **Recurrent Neural Network**



Natural Language Processing – Common Use Cases

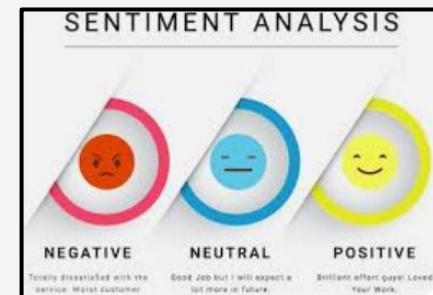
1. Customer Service Agent or Virtual Assistants – Chatbot or Siri



2. Language Translation – Google Translate

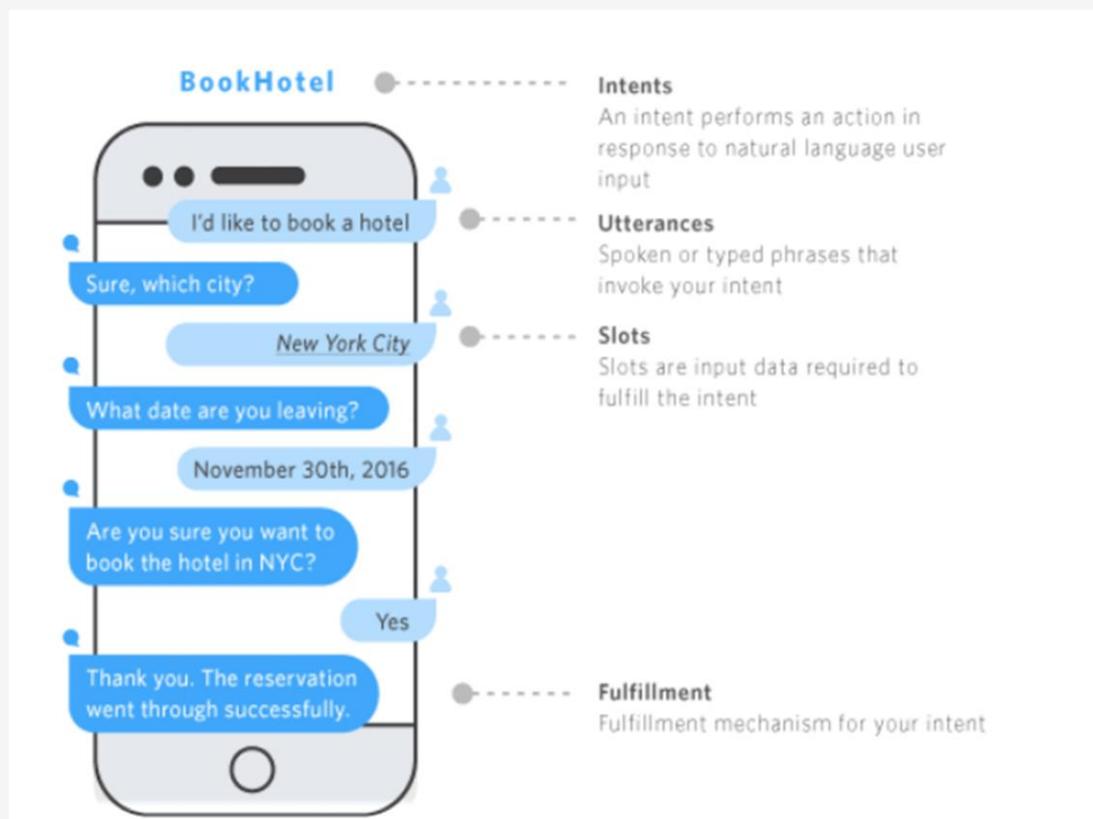


3. Sentiment Analysis - Emotion conveyed by textual data - Product Launch



AWS AI Service – Amazon Lex (AI Chatbot)

Amazon Lex - Build and deploy conversational AI interfaces



Source : AWS

- **Self-service voice assistants and chatbots** – build a call center bot
- **Application/Transactional bot** – build a stand-alone pizza ordering agent or a travel bot
- **Enterprise Productivity bot** – build custom bots to connect to enterprise data resources
- **Device Control bot**– use Amazon Lex to issue control commands to connected devices

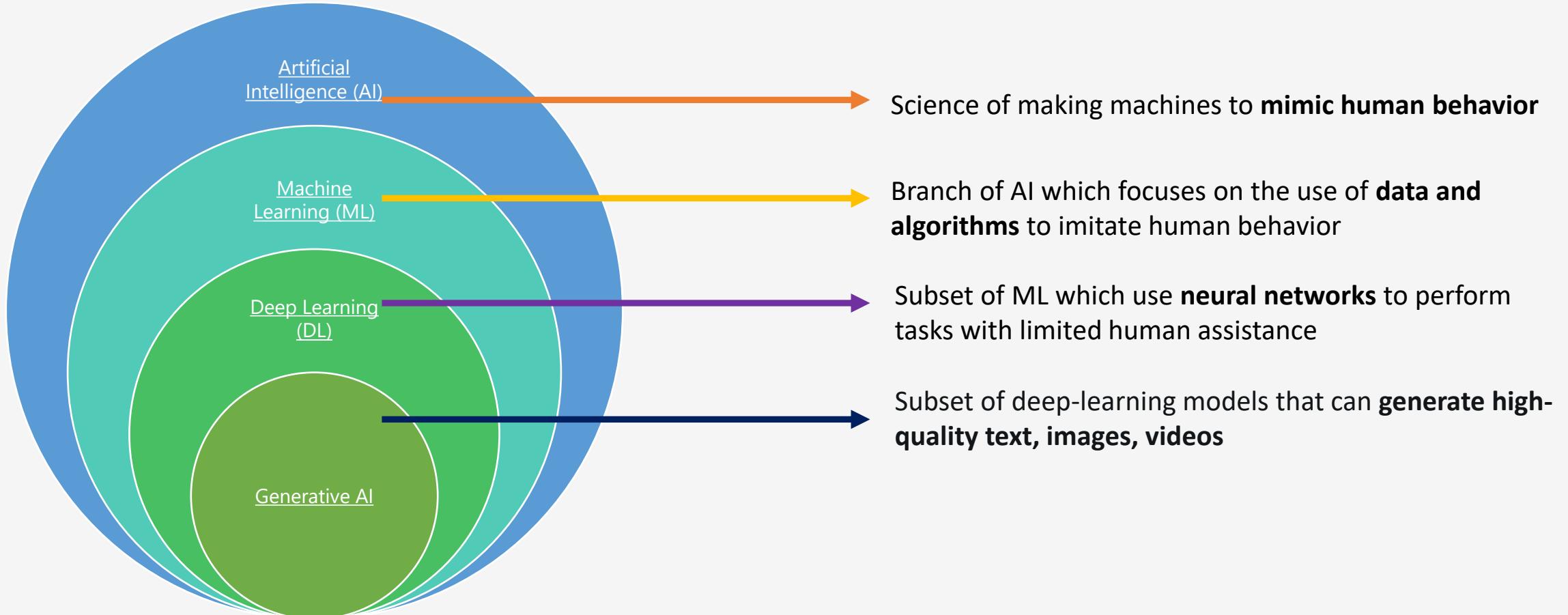


Domain 2:

Fundamentals of Generative AI

(Part 1)

Artificial Intelligence and Machine Learning - Overview



What is Generative AI ?



AI21 Labs Jurassic-2 Mid Start new chat

Using on-demand capacity. [Switch Provisioned Throughput](#)

what is generative AI Default

Generative AI is a branch of artificial intelligence (AI) that focuses on creating models or algorithms that can simulate or generate new data based on existing data. The purpose of generative AI is to mimic the process of natural generation, such as the way living organisms produce offspring with traits similar to their parents.

Generative models can be trained on various data types, including images, sounds, text, and even complex data structures such as DNA sequences. These models use techniques such as deep learning and reinforcement learning to generate new examples that are similar to or identical to the training examples.

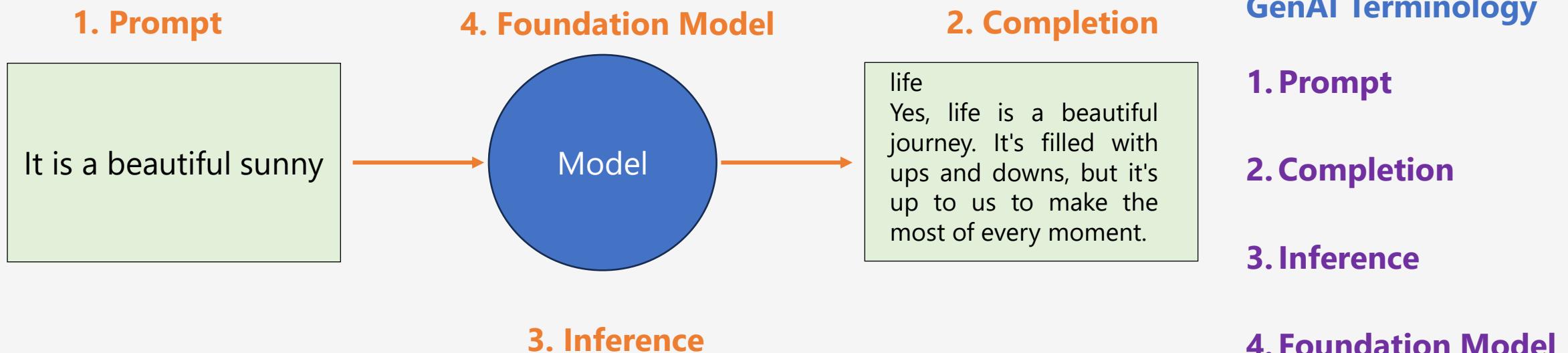
Generative AI has a wide range of applications, including image generation, natural language processing, drug discovery, and robotics. It is also used in various fields such as art, music, and gaming to create more realistic and engaging content.

Generative AI refers to :

- Branch of AI that use **deep-learning models**
- **Generate high-quality text, images, code** based on the data they were trained on
- Popular examples of generative AI include
 - ChatGPT/GPT from OpenAI
 - Claude from Anthropic

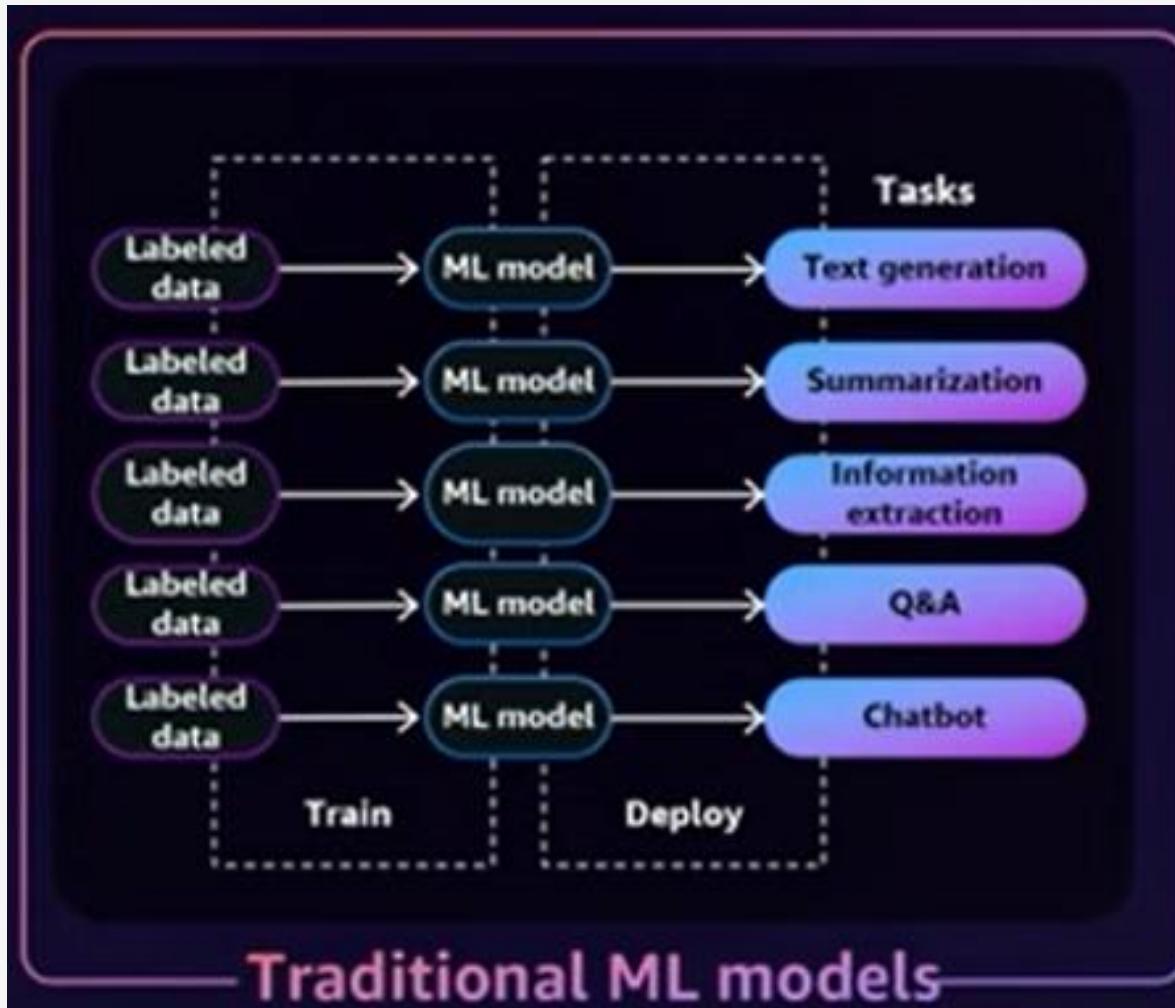
Answer generated using Jurassic FM from **AWS GenAI Service – Amazon Bedrock**

How Generative AI works - Basic Concepts and Terminology - 1



1. Prompt – Input provided to Model
2. Completion – Output of Model
3. Inference : Act of using model to generate text is called Inference

Challenges in Traditional Machine Learning Models

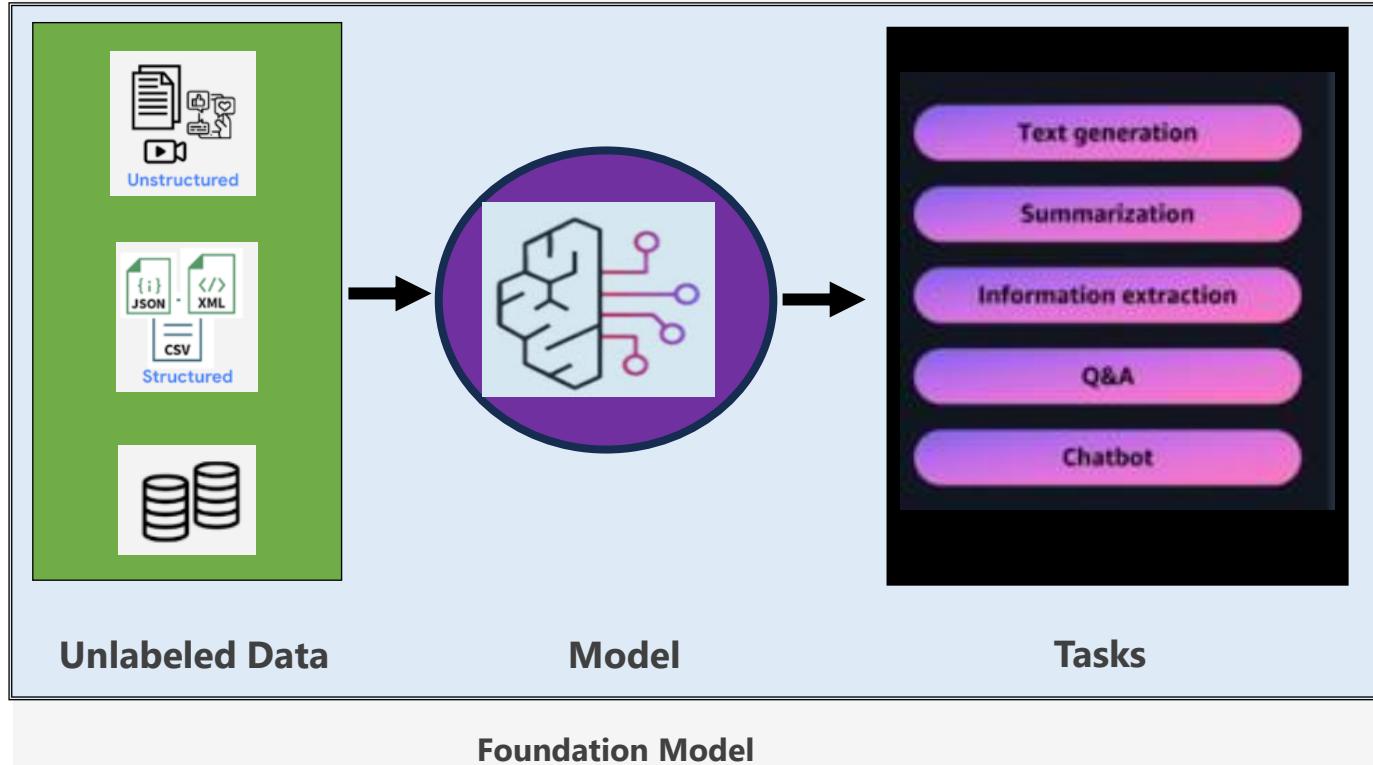


Source : AWS

Challenges in Traditional ML Models

- Each Model focused on specific Task such as
 - Text Generation
 - Summarization
 - Q&A
- Needs labeled data which requires human intervention and can be cost prohibitive

Key Characteristics of Foundational Models

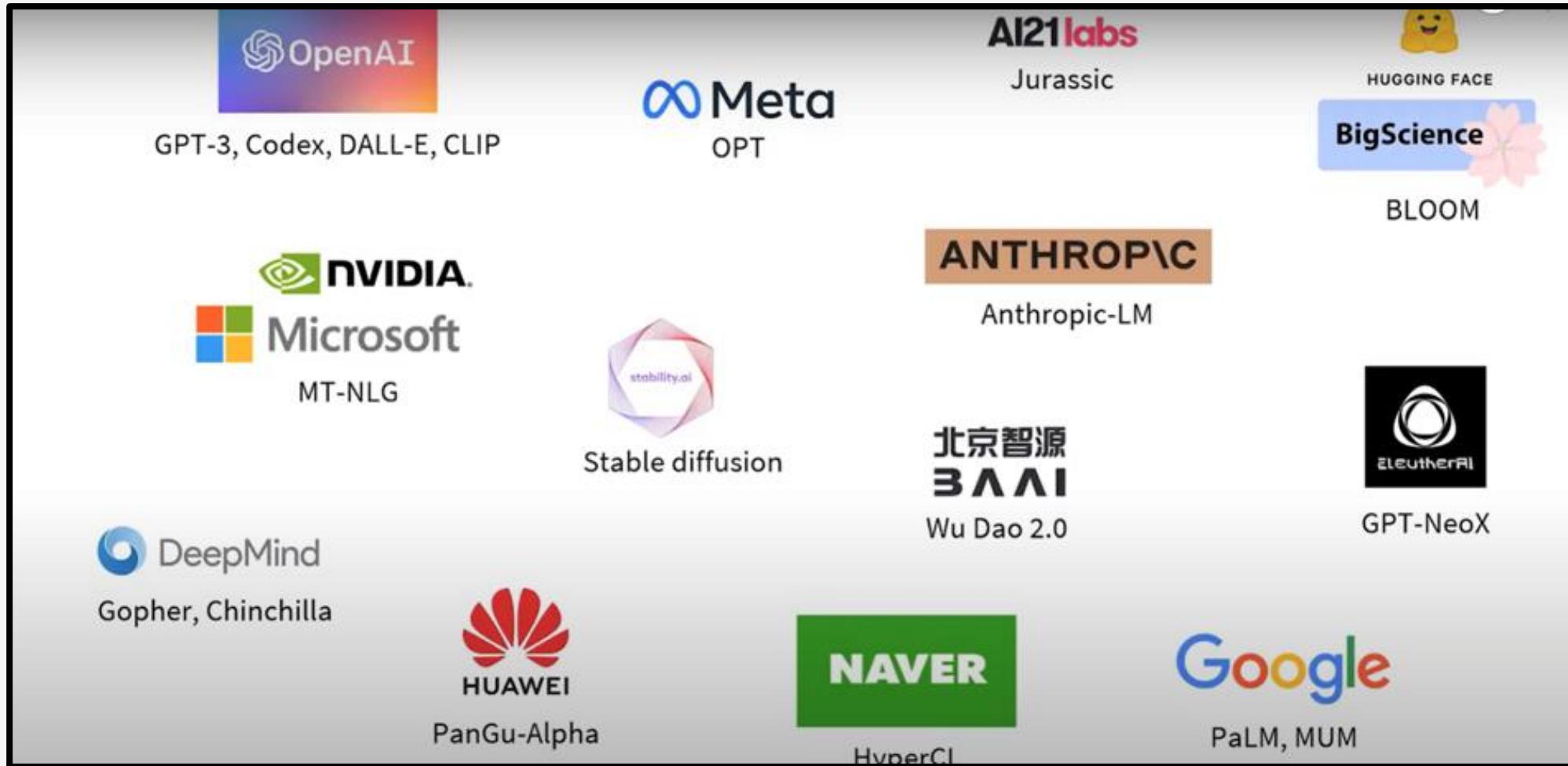


Organization's can utilize **existing FM's as the base for building** task-specific models and **Adapt** them

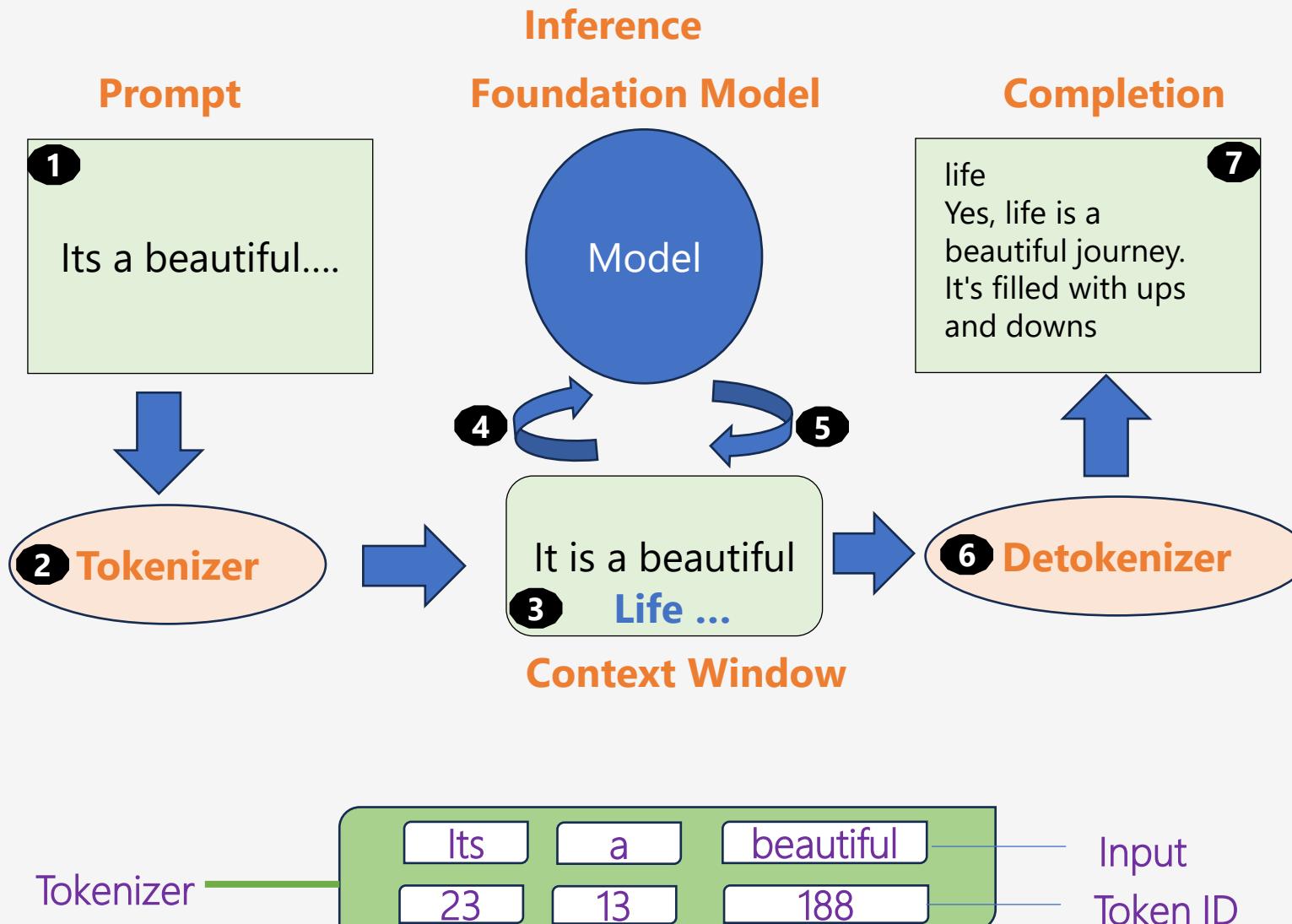
Key Characteristics

- **Deep neural networks** techniques
- **Unlabeled data** - Trained on unlabeled data and Self-supervised learning such as internet data - Wikipedia
- **Generalized** – One FM for multiple tasks
- **Pre-trained** – FM's are pre-trained models
- **Large** – Exposed to vast amounts of data, enabling it to learn patterns & features from data.
- **Expensive** - Developing FM's is an **expensive process** that involves **large amounts of computational resources** and **expertise**.

Popular Foundation Models



How Generative AI works - Basic Concepts and Terminology - 2



GenAI Terminology

- **Prompt**
- **Tokenizer**
- **Context Window**
- **Foundation Model**
- **Max Token/Stop Sequence**
- **Completion**
- **Inference**
- **Context Window** - number of tokens the model can take as input at once

Foundation Model – Tokens, Parameters and Temperature



Tokens

Model	Context length	Number of English pages (1 page = 500 words)
GPT 3.5	4,096	6
GPT 4	8,192	12
	2,048	3
Llama 1	4,096	6
Llama 2		



Parameters



Temperature

Tokens - **One token** generally corresponds to **~4 characters of text** for common English text

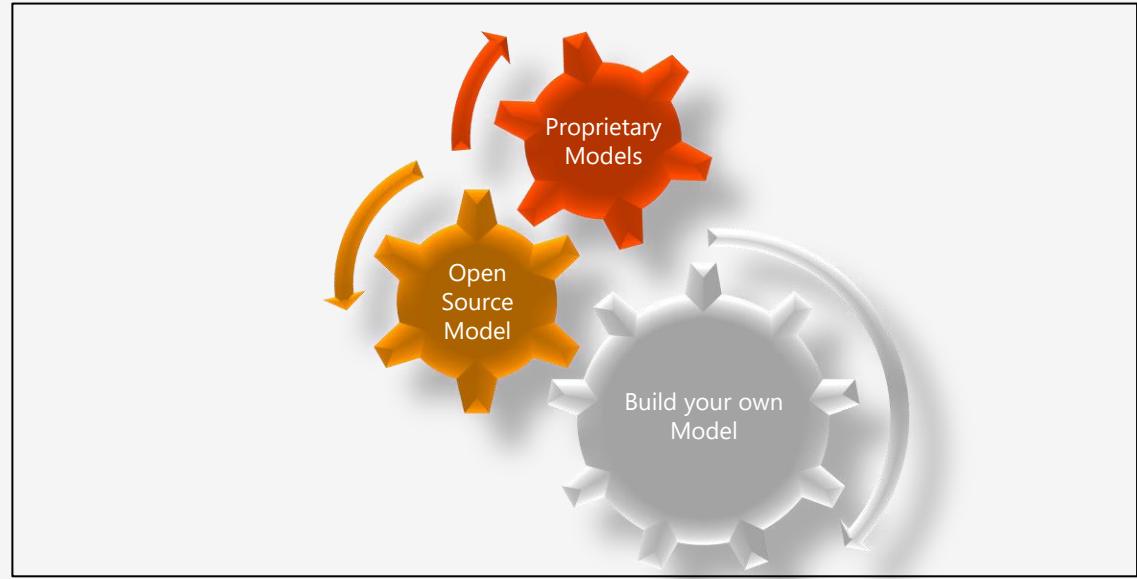
Parameters - can be considered count of **connections & weights** between nodes in a neural network.

Temperature - This parameter controls the **creative ability** of your model.

Foundation Models – Build your Own Model, Open Source and Proprietary Model

1. Build your own Domain Specific Foundation Model

- Highly customized to domain
- BloombergGPT: A LLM for Finance - [Link](#)
- Own source code, architecture, model weights etc.
- Expensive, time consuming and need expertise



2. Open Source Foundation Models

- Llama FM from Meta
- Reuse the publicly available source code - [Link](#)
- Customize (Fine Tune) with smaller dataset
- Lower cost and time to market



Foundation Models – Build your own Model, Open Source and Proprietary Model

3. Proprietary Models

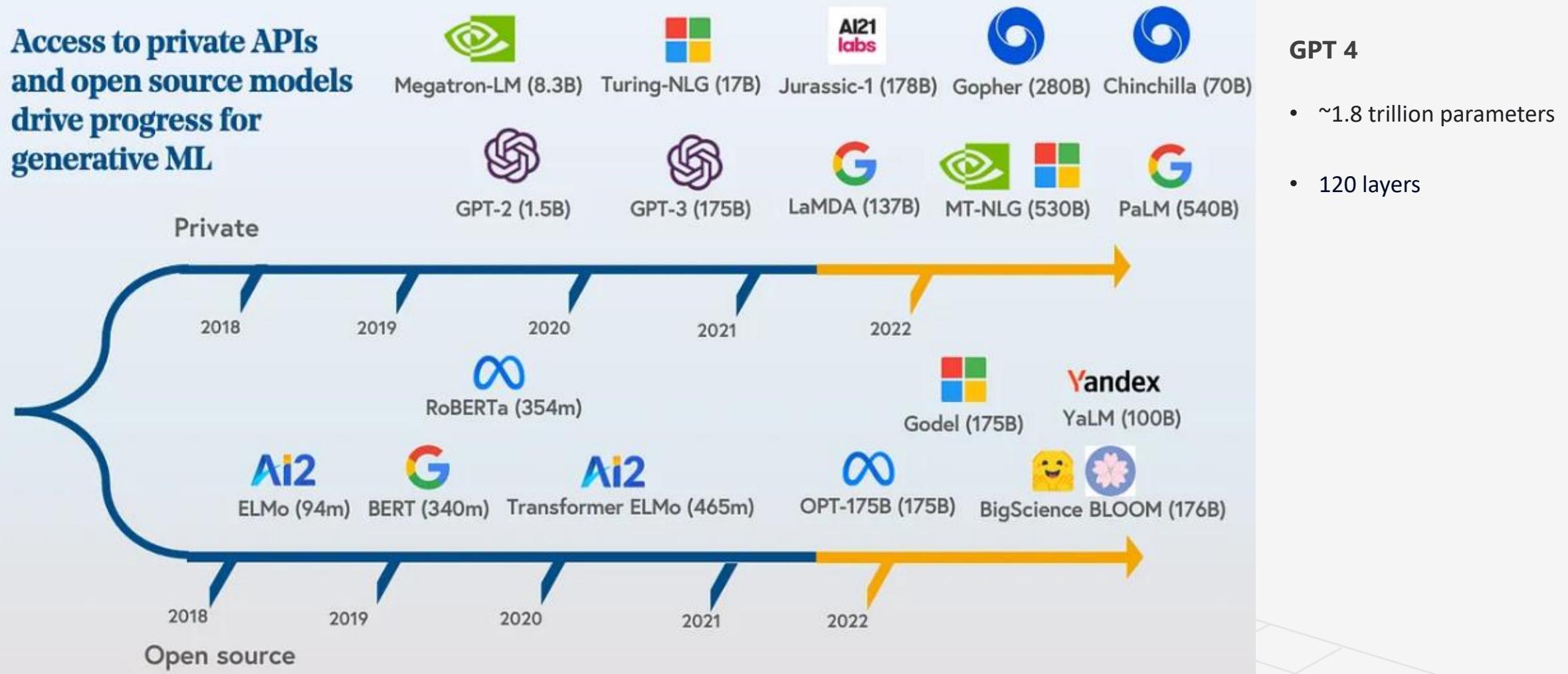
- GPT series from Open AI
- Source code not available for modification
- In some scenarios can customize/fine tune
- Available via an API's



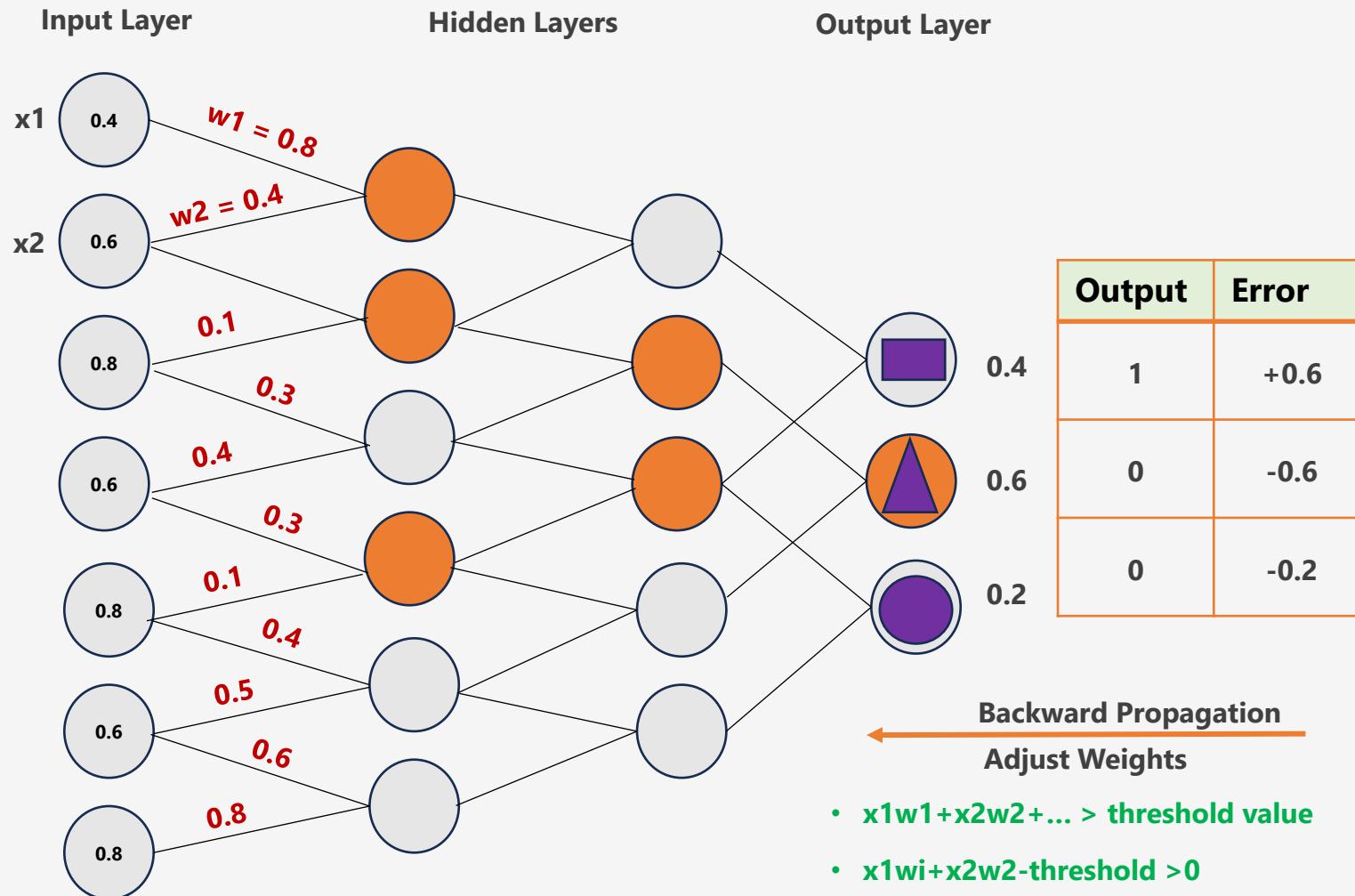
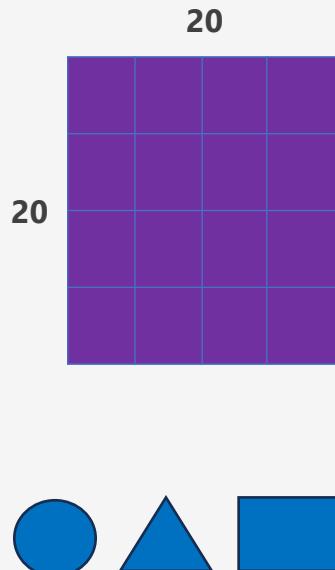
	Google DeepMind	amazon	Microsoft OpenAI	Meta	NVIDIA	stability.ai	NON-EXHAUSTIVE Startups, researchers, OSS contributors
Text / language	BERT, Bard (LaMDA), PaLM, GLaM, Chinchilla, Gopher	AlexaTM, Titan LLMs	MT-NLG, GPT-4 / ChatGPT (GPT-3.5)	OPT-175B, LLaMA	MT-NLG		Dolly, Claude, Jurassic-1, Cohere, GPT-J, GPT-NeoX, BLOOM, Wu Dao 2.0, CTRL, Alpaca
Code	AlphaCode	Code-Whisperer	Copilot, CodeBERT, Codex				Replit, Polycoder, CodeGen, CodeT5
Image	Imagen, Parti, DreamBooth		Dall-E 2, CLIP, NUWA-Infinity	Make-a-scene, Dinov2	SPADE	Stable Diffusion	MidJourney, Waifu Diffusion
Speech & Music	WaveNet, MusicLM		Muzic, Whisper, Jukebox			Dance Diffusion	WaveGAN
Video	Imagen Video, Phenaki		NUWA-Infinity	Make-a-video			CogVideo
3D	DreamFusion, 3DIM		Point-E		Get3D, Magic3D		Motion Diffusion Model
							Proprietary Open source Private

Popular Foundation Models (Parameters perspective)

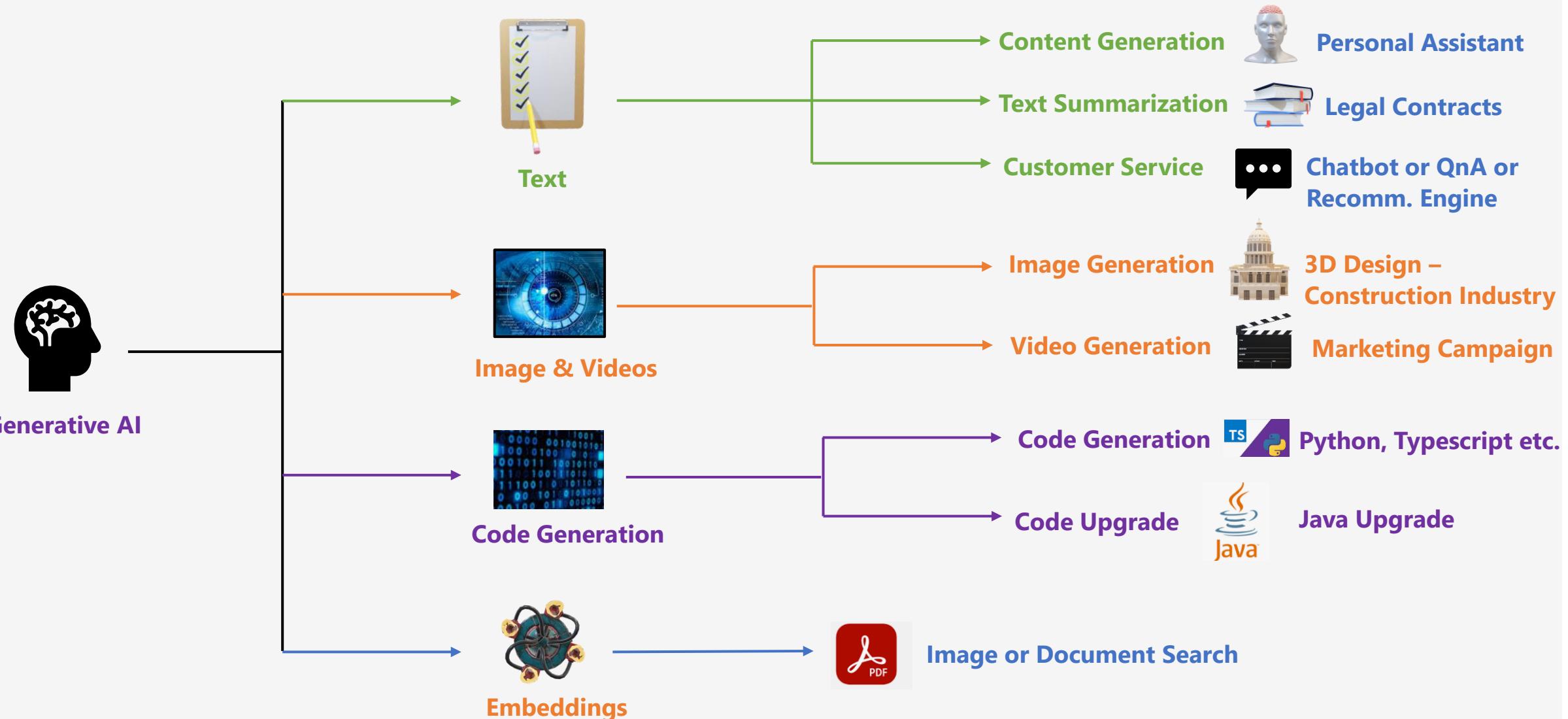
Access to private APIs and open source models drive progress for generative ML



Machine Learning – Deep Learning with Artificial Neural Networks



Generative AI – Use Cases



Foundation Models vs LLM's vs Image Generators vs Multi-Modal



Foundation Models

vs

Large Language Models (LLM's)

vs

Text to Image Model (Stable Diffusion)

vs

Multi-Modal

Foundation Models - Categorization

Foundation Models – Broad Categorization of Models

- Focus Area : **Text Generation**
- Large Language Models (LLM's)
- Example : GPT 4, Llama 3

- Focus Area : **Image Generation**
- Image Generation or Diffusion Model
- Example : Stable Diffusion

- Multi Modal
- Focus Area : **Text & Vision Model**
- Example : Claude3 – Multi-Modal

Foundation Model

- Large **deep learning neural networks** + **Trained on massive datasets**
- Generate Text, Images, Videos or both
- **Large Language Models (LLM's)**
- **Image Generation Models – Text to Image or Image to Image**
- **Multi-Modal**

Basic Introduction

to

Vectors, Embeddings, Vector DB & Search

1. Why do we need Vectors ?



Structured Data

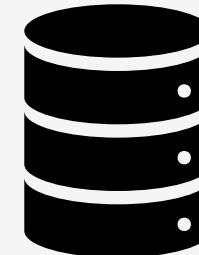


Image Search or Unstructured Data



???

Key Terms

1. **Vectors**
2. Embedding Models
3. Data Chunking
4. Vector Store
5. Vector Search
6. Cosine Similarity
7. K-Nearest Neighbor

Search

1. What are Vectors ?

Vectors are mathematical representation of words, sentences and documents

Vectors are represented by **List of numbers** or sequence of numbers



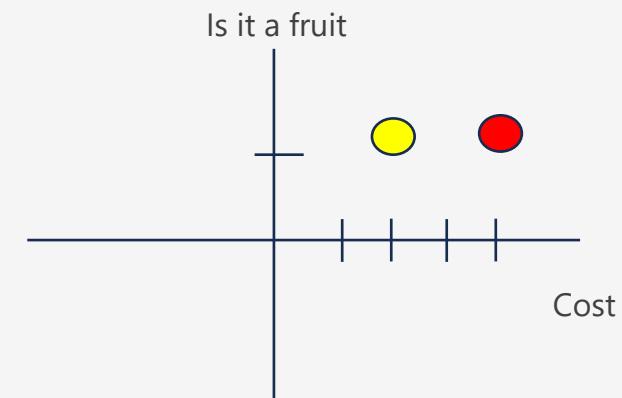
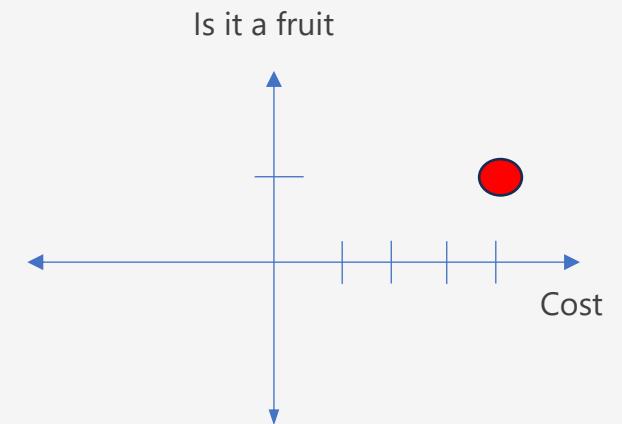
Is it a fruit	Cost
1	\$4

[1 , 4]



Is it a fruit	Cost
1	\$2

[1 , 2]



1. What are Vectors ?



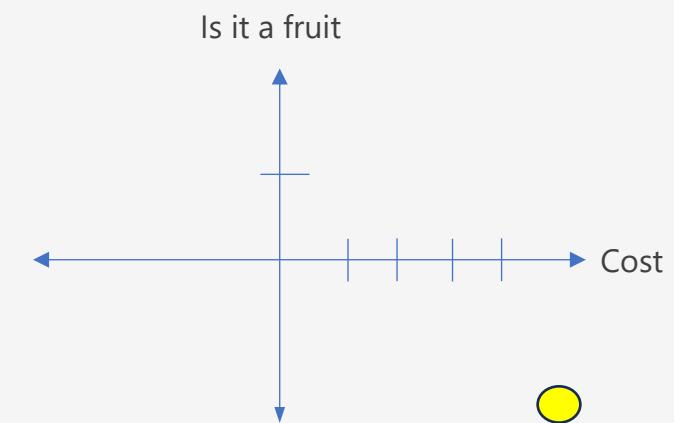
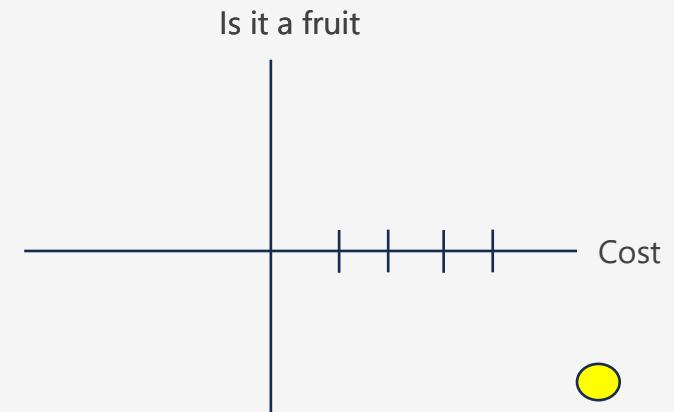
Is it a fruit	Price
-1	\$50

$\begin{bmatrix} -1 \\ 50 \end{bmatrix}$



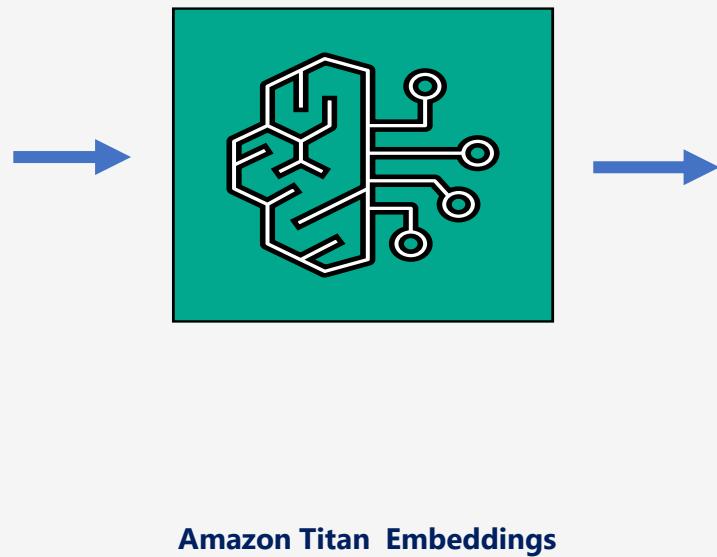
Is it a fruit	Price
-1	\$40

$\begin{bmatrix} -1 \\ 40 \end{bmatrix}$



2. What are Embedding Models ?

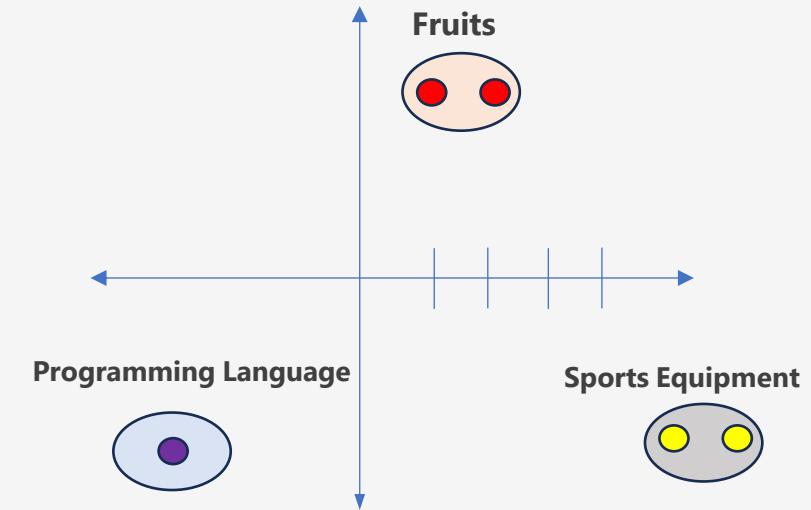
- Apple
- Banana
- Grapes
- Football
- Soccer
- GO
- Tennis



Word Embeddings
Ex : great

Sentence Embeddings
Ex. : Price of apple ?

Doc Embeddings



Key Terms

1. Vectors
2. **Embedding Models**
3. Data Chunking
4. Vector Store
5. Vector Search
6. Cosine Similarity
7. KNN

Embedding Models

Output from Embedding Models

Fruit

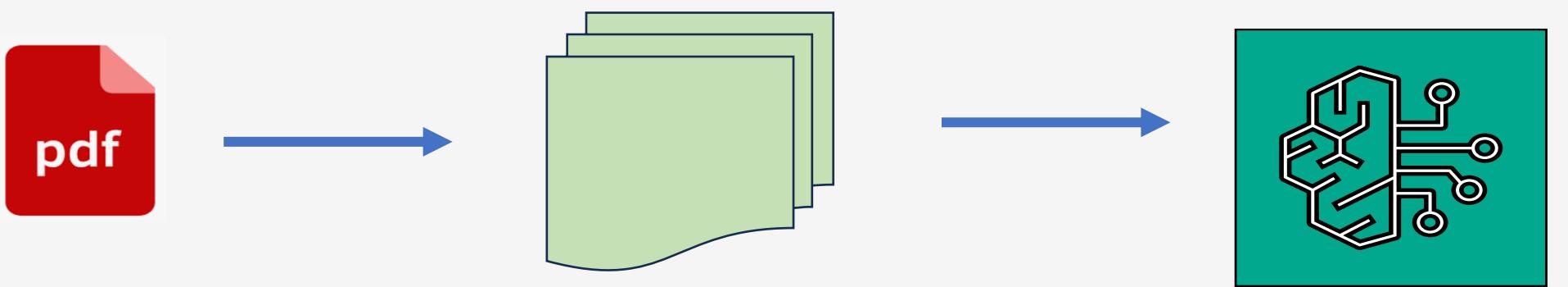


```
[  
    0.022321066, -0.027544279, -0.006137953, 0.0024092742, -0.0003515296,  
    -0.007889225, -0.01853968, -0.037405808, -0.004294867, -0.012602357,  
    0.009385457, 0.02421176, -0.004383281, -0.011901848, -0.020702416,  
    0.00431187, 0.055986296, 0.0040738326, 0.021300908, -0.03713377,  
    -0.004944368, 0.012269106, 0.02614326, -0.019777471, 0.0017206672,  
    0.014867109, 0.004597514, -0.021913003, -0.0042336574, 0.013180447,  
    0.032699477, -0.027245032, -0.0112965545, -0.009358253, -0.004951169,  
    -0.015030335, -0.007501565, -0.018906936, 0.0074675595, -0.0067568496,  
    0.0011952856, 0.0035501514, 0.012316713, 0.016132105, -0.014839904,  
    -0.0044002836, 0.010405616, -0.0066276295, -0.015751246, -0.007141109,  
    0.02729944, -0.00031348618, -0.015669633, -0.015656032, 0.015696838,  
    0.013554505, -0.001635654, 0.006069943, 0.009167824, -0.01917898,  
    -0.006059741, 0.013969369, -0.03006067, 0.007691995, 0.0005028531,  
    -0.005468049, -0.007263528, 0.009970348, -0.016254524, 0.0041248407,  
    0.01964145, 0.037324198, -0.0011340762, -0.0013296065, 0.010392014,  
    -0.01268397, -0.013037625, -0.006811258, -0.0045601083, 0.0028955496,  
    0.027435461, -0.024320576, -0.006161757, 0.014350229, 0.009392259,  
    -0.006328383, -0.025490358, 0.023273215, 0.0083925035, -0.003934411,  
    0.023722084, -0.0072499258, -0.0053626327, 0.017818768, -0.03327077,  
    0.019097365, 0.0022086431, 0.046709653, -0.016825814, -0.032182597,  
    ... 1436 more items  
]
```

Amazon Titan Embeddings

1536 Dimensions

3. Text Splitting or Data Chunking



Data Source - PDF

- Split by Character
- Split by Token
- Split by Code

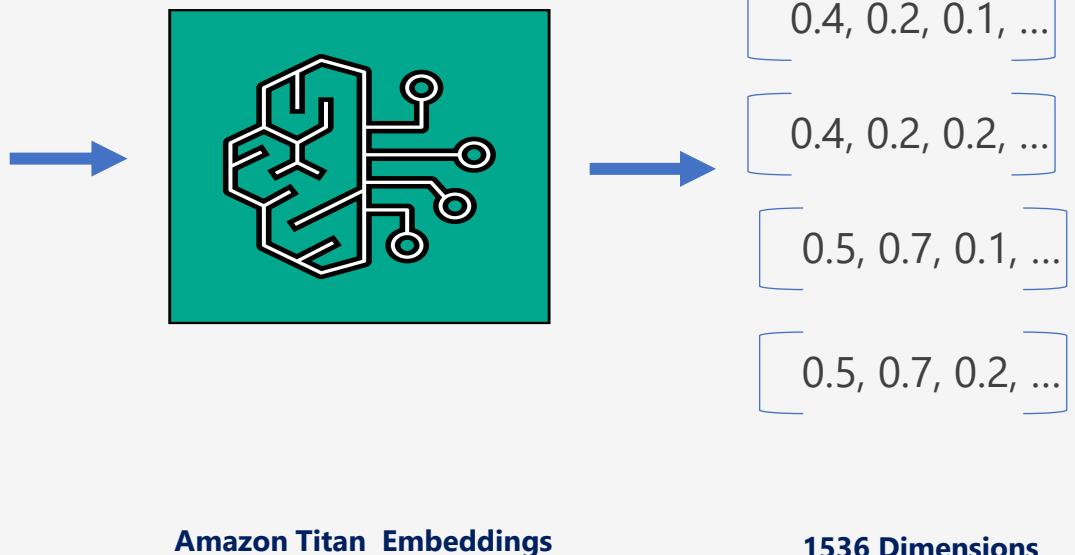
Amazon Titan Embeddings

Key Terms

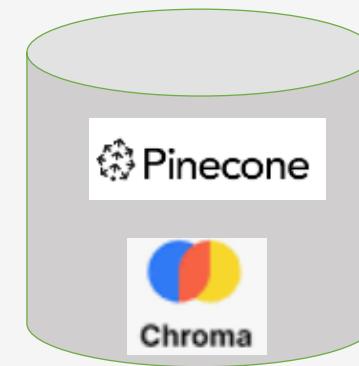
1. Vectors
2. Embedding Models
3. **Data Chunking**
4. Vector Store
5. Vector Search
6. Cosine Similarity
7. K-Nearest Neighbor Search

4. Vector Store?

- Apple
- Banana
- Grapes
- Football
- Soccer
- GO
- Tennis



Vector Store



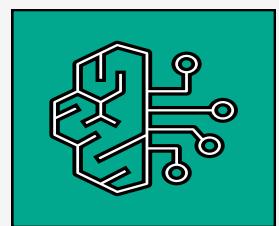
- ✓ FAISS
- ✓ Pinecone
- ✓ Chroma DB

Key Terms

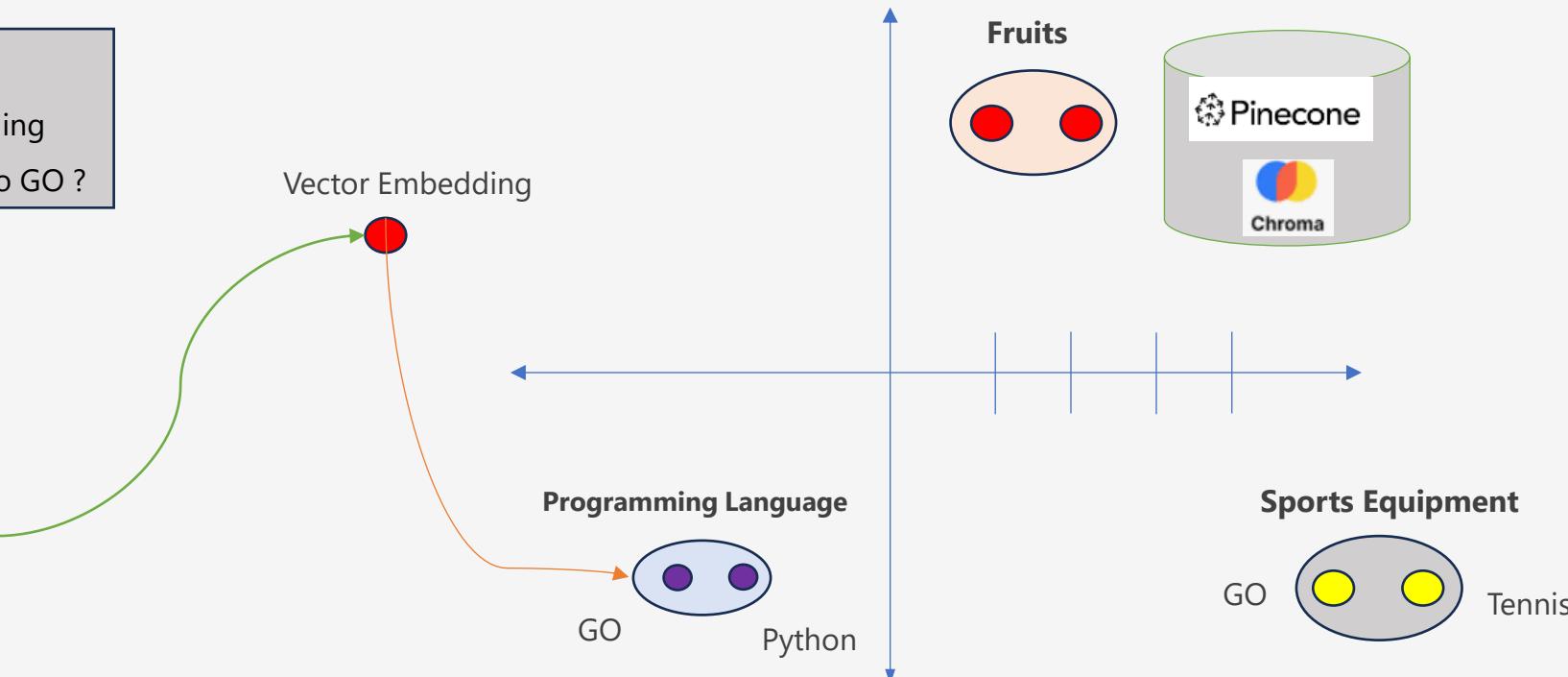
1. Vectors
2. Embedding
- Models
3. Data Chunking
4. Vector Store
5. Vector Search
6. Cosine Similarity
7. K-Nearest Neighbor Search

5. Vector Search

Question :
Which programming language is similar to GO ?



Amazon Titan Embeddings

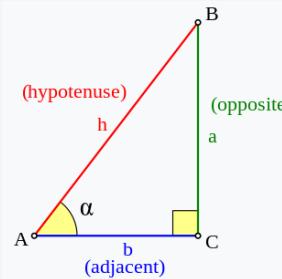


Key Terms

1. Vectors
2. Embedding Models
3. Data Chunking
4. Vector Store
5. **Vector Search**
6. Cosine Similarity
7. K-Nearest Neighbor Search

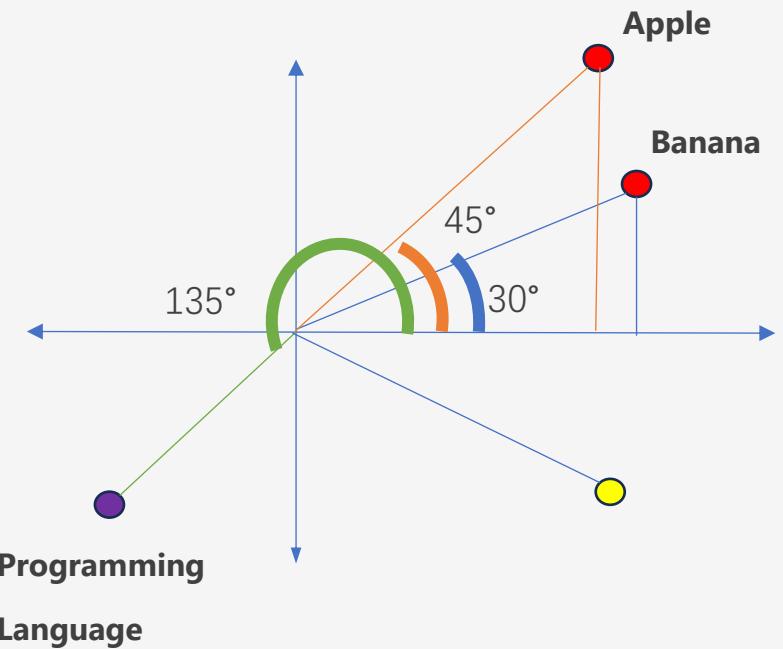
6. Similarity Search – Cosine Similarity

1. Cosine $\theta = \text{adjacent}/\text{hypotenuse}$



2. Similarity Search using Cosine Distance:

- Cosine distance metric is used to find **similarities between different vectors**
- In cosine metric we measure the **degree of angle between two vectors**
- $\cos 0 = 1$, $\cos 90 = 0$ and $\cos 180 = -1$
- **Cosine value 1** is for **vectors pointing in the same direction** --- > **Similar**
- At Cosine Value 0 ---- > Vectors unrelated (some similarity found).
- Value -1 for vectors pointing in opposite directions(No similarity).



7. Cosine Similarity and K- Nearest Neighbor Algorithm or ANN

KNN is a *supervised machine learning algorithm* used **classification and regression** tasks.

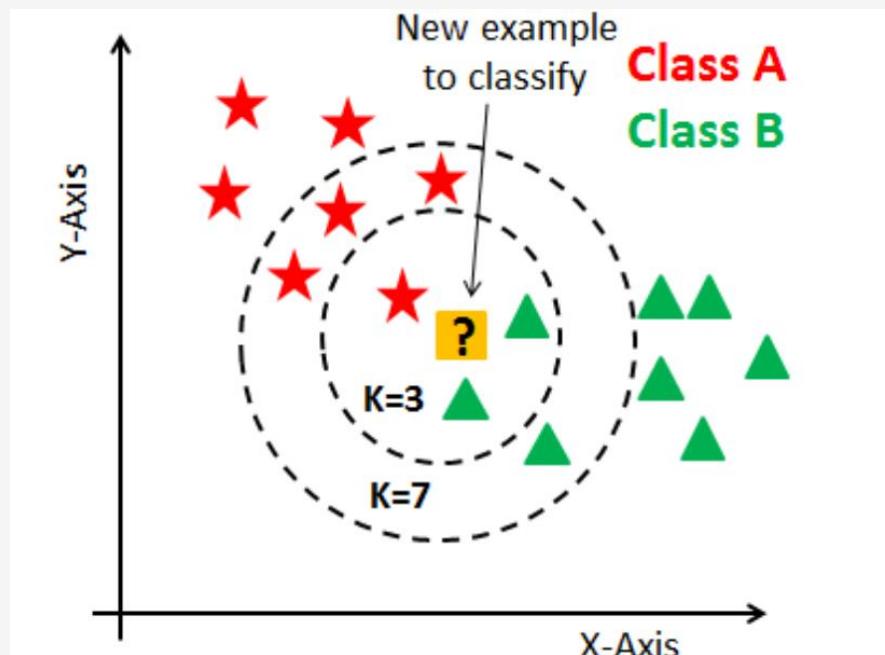


Image Source : Internet

Key Terms

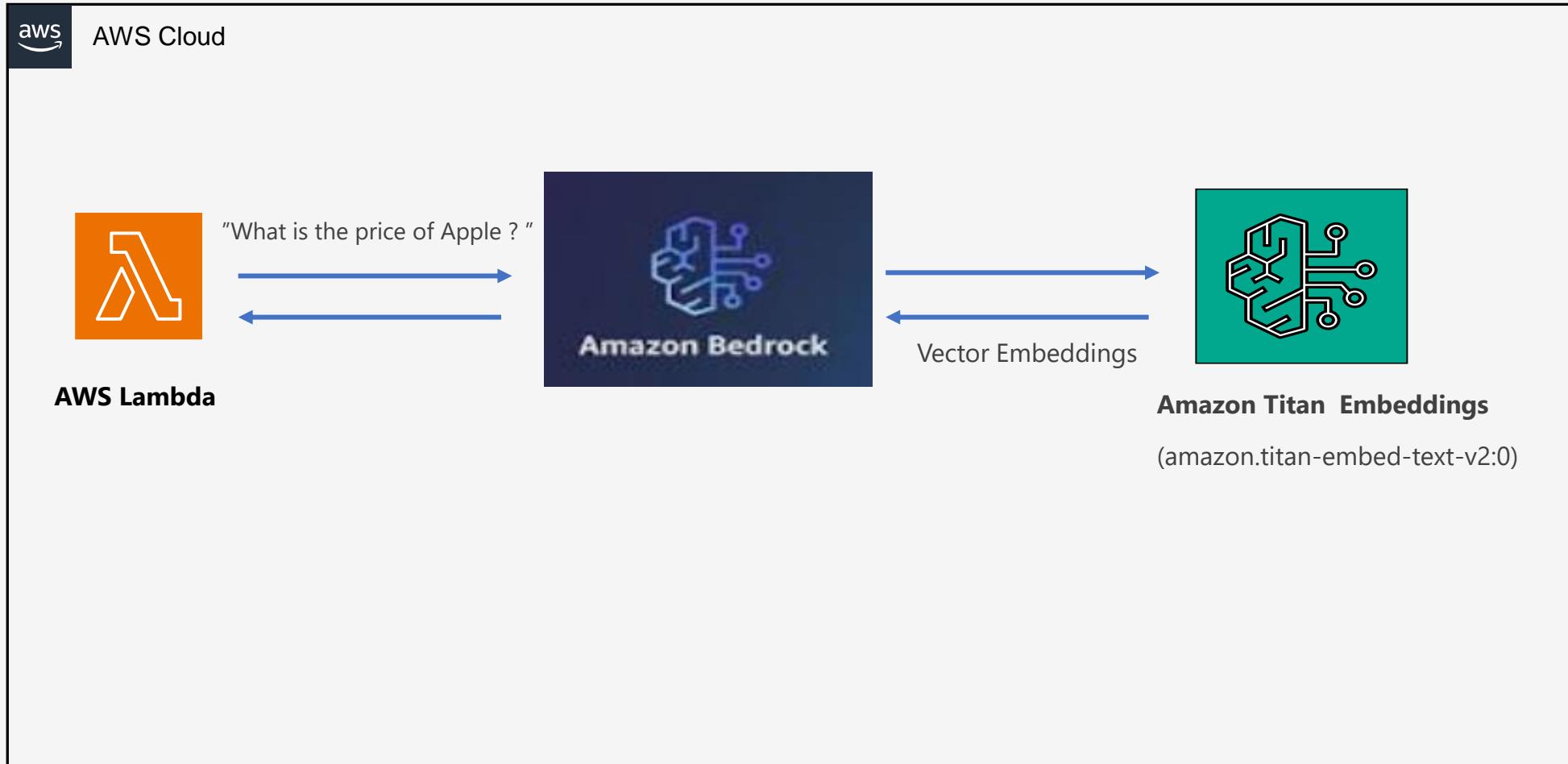
1. Vectors
2. Embedding Models
3. Data Chunking
4. Vector Store
5. Vector Search
6. **Cosine Similarity**
7. **KNN**

ANN - Approximate Nearest Neighbor

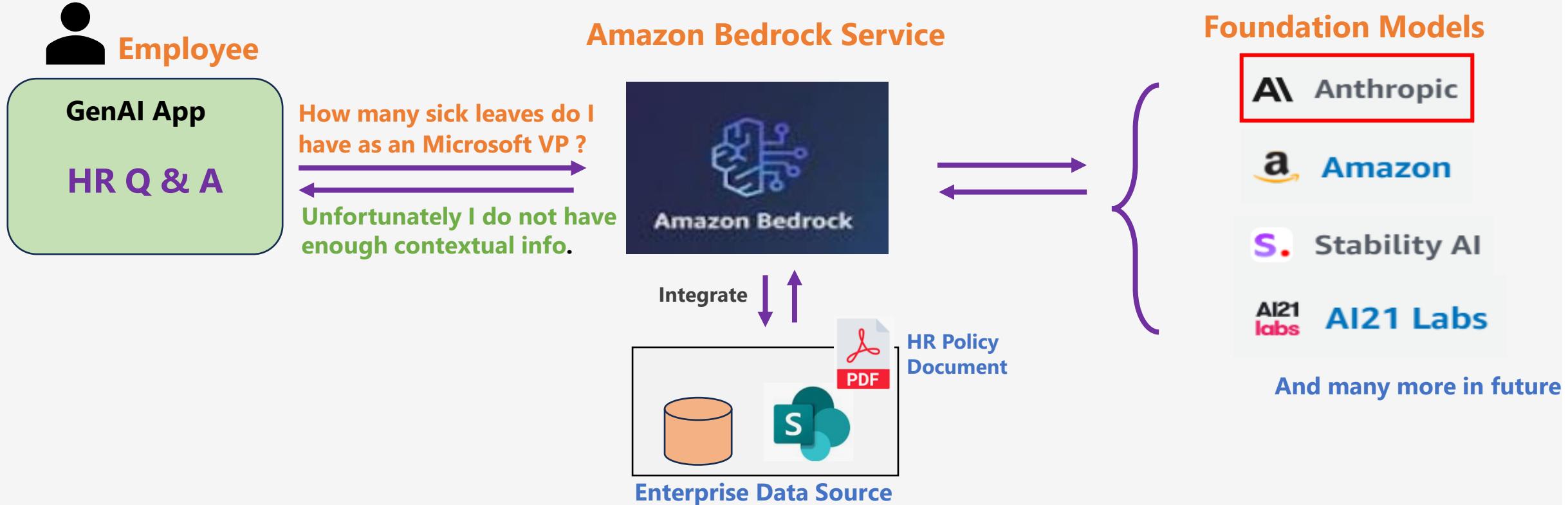
Generative AI – Vectors and Embeddings

Demo – Vectors & Embeddings

Generative AI – Vectors and Embeddings



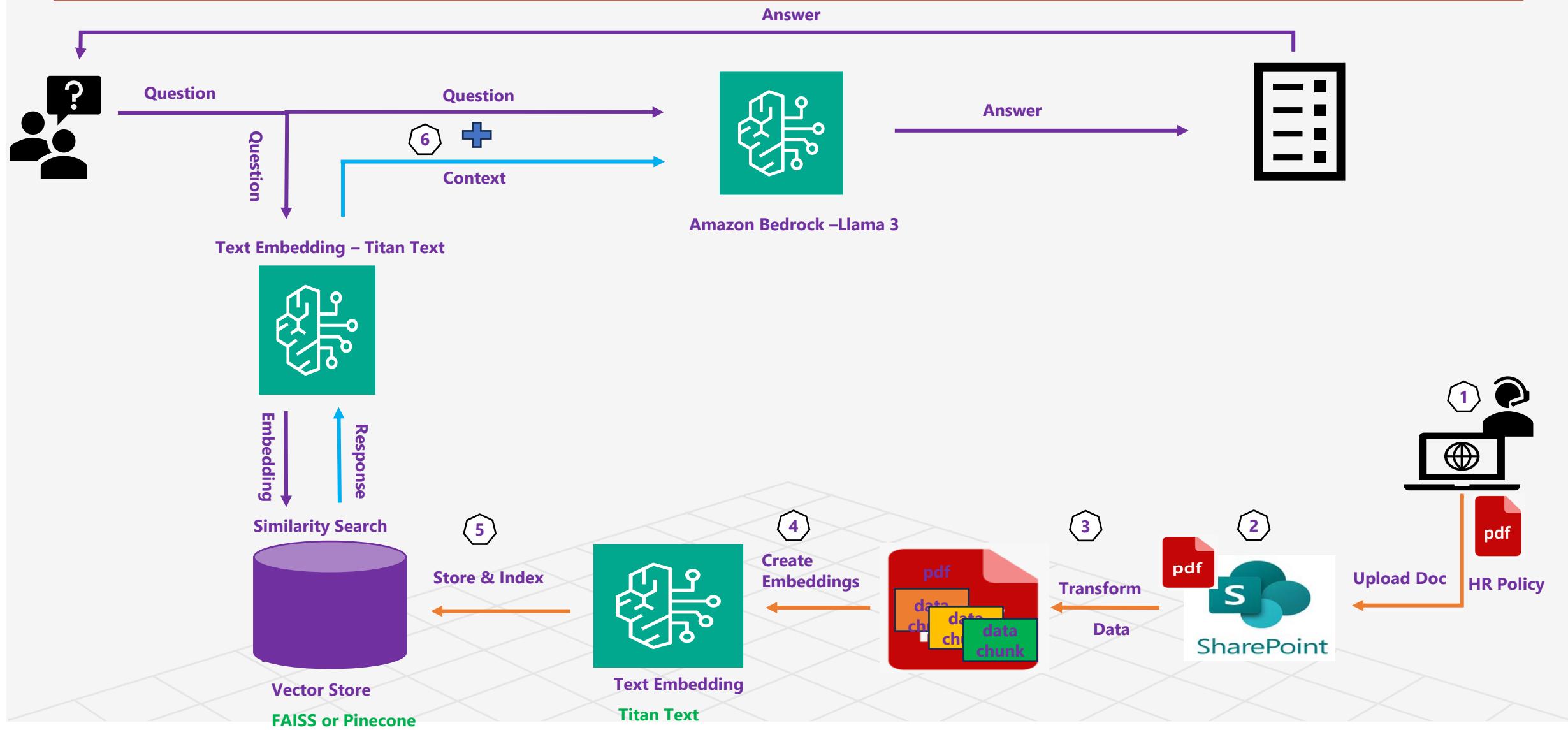
Retrieval Augmented Generation (RAG)



What is RAG ?

RAG is the process of **optimizing** the output of a large language model by **supplementing the knowledge** of LLM with **internal organization data source**.

HR Question & Answer App with RAG - Solution Architecture





Domain 2:

Fundamentals of Generative AI (Part 2)

Amazon Q, Amazon Bedrock, SageMaker

Jumpstart, PartyRock, Inferentia and Tranium

AWS Generative AI Services Landscape



**Pre-built Apps
leveraging FM**



Amazon Q



- Amazon Q Business – AI Assistant
- Amazon Q Developer – Coding Assistant
- Amazon Q Connect
- Amazon Q QuickSight & Supply Chain



Serverless Layer



Amazon Bedrock



- Broad range of FM's
- RAG - Bedrock Knowledge Bases
- Bedrock Agents - Orchestrator
- Guardrails - Security



**Infrastructure for Foundation
Model Training & Inferencing**



AWS Tranium



AWS Inferentia



Amazon SageMaker Jumpstart

**Free Playground
to build GenAI
Apps -**

PartyRock

AWS GenAI Service

Amazon Q



Amazon Q - Overview

Amazon Q is generative AI-powered assistant to help improve productivity to employees :

1. Amazon Q Business

- Generative AI-powered personal assistant for Business Users – Sales Team, Marketing

2. Amazon Q Developer

- Q Developer assists developers and IT professionals improve productivity by assisting across SDLC

3. Amazon Q for QuickSight

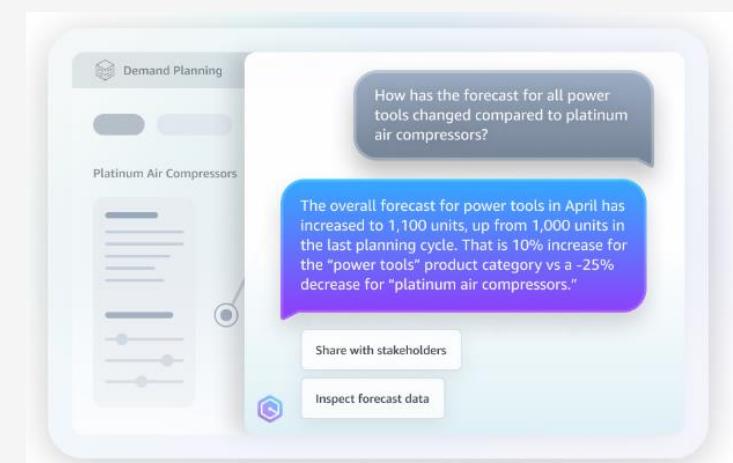
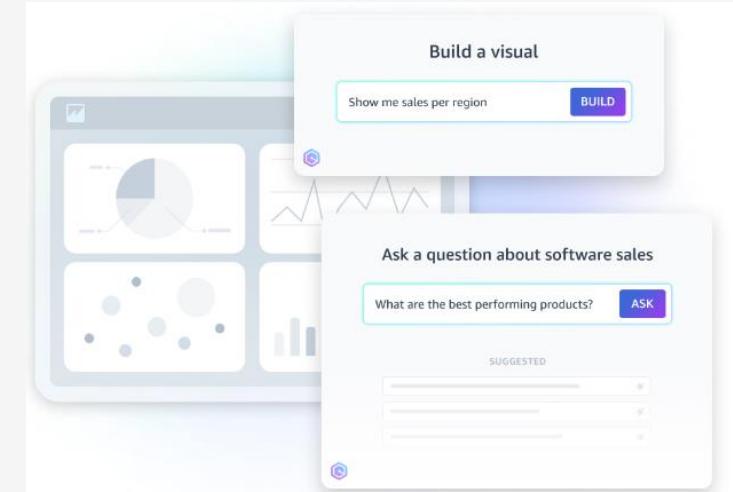
- Allows business analysts to use natural language to build BI dashboards in minutes.

4. Amazon Q for Connect

- Amazon Q in Connect uses the real-time conversation with the customer to automatically recommend what to say or what actions an agent should take to better assist customers.

5. Using Amazon Q in AWS Supply Chain (coming soon)

- Allows inventory managers, supply and demand planners to get intelligent answers about what is happening in their supply chain, why it is happening, and what actions to take.



Section on :
Amazon Q - Business



What is Amazon Q Business?

Amazon Q Business is a **fully managed, generative-AI powered assistant**

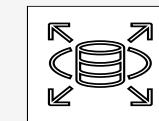
- Generate Content



- Summarize Data



- Chat with Enterprise Data (40+ connectors)



40+ Pre Built
Connectors

- Assign Tasks (Plugins such as Jira)



ChatGPT with Enterprise Controls

Amazon Q for Business – A typical day for a Marketing Manager



Anisha
Marketing Manager

Improve employee productivity with Amazon Q

A woman in a professional outfit is holding a white sign that identifies her as 'Anisha Marketing Manager'. To her right is a 3D bar chart with a blue arrow pointing upwards, representing growth or productivity. The overall theme is professional and productive.

Feedback report on the impact of marketing campaign
(20 Pages)



Summarize

Draft e-mail to prospective client outlining steps to launch campaign for a juice drink



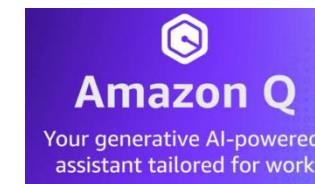
Generate Content

Chat with internal documents about Brand Guidelines



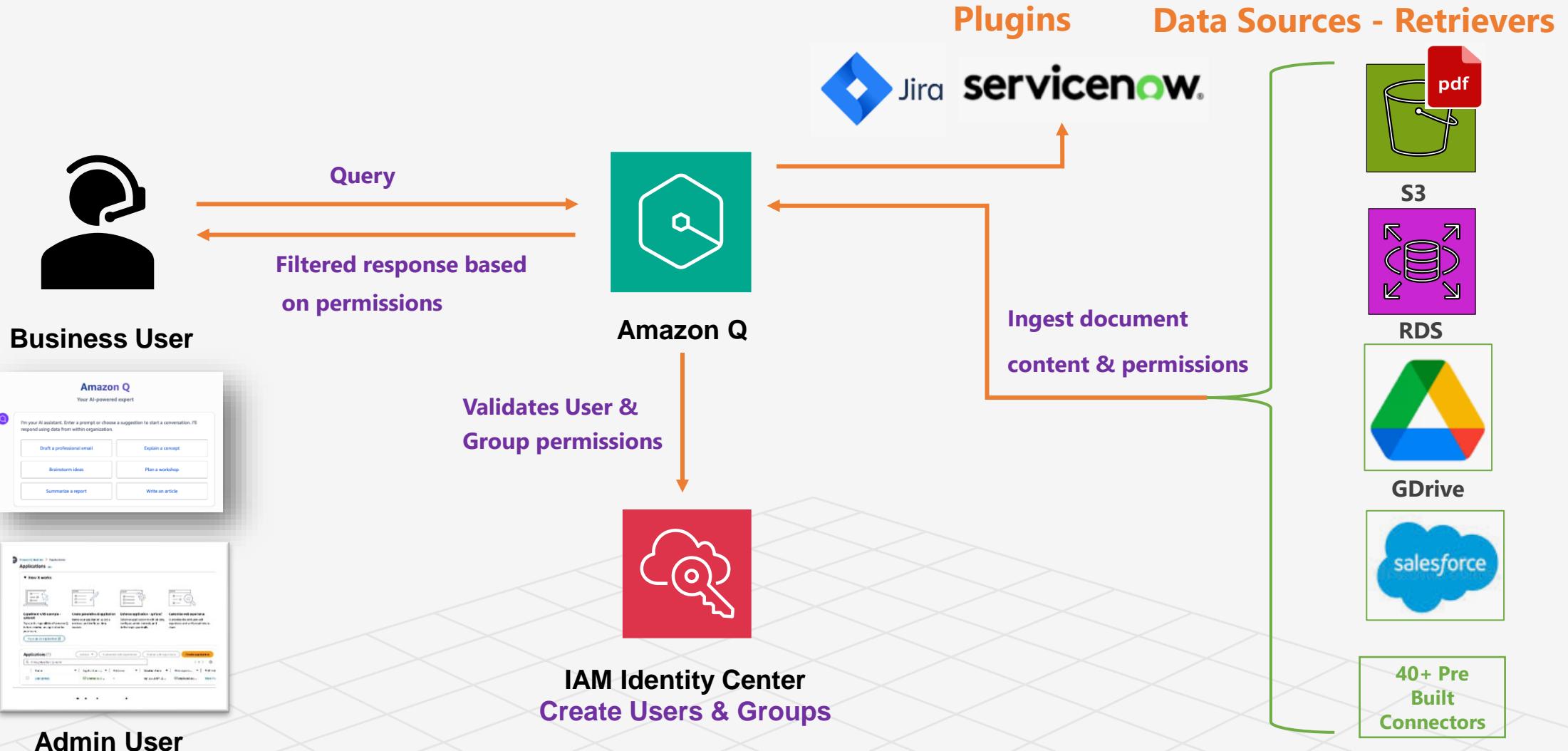
Enterprise Data Search

Assign Task in Jira to direct reports



Task Mgmt.

What is Amazon Q for Business - Architecture



Demo :

Marketing Manager AI Assistant



*Section on :
Amazon Q - Developer*

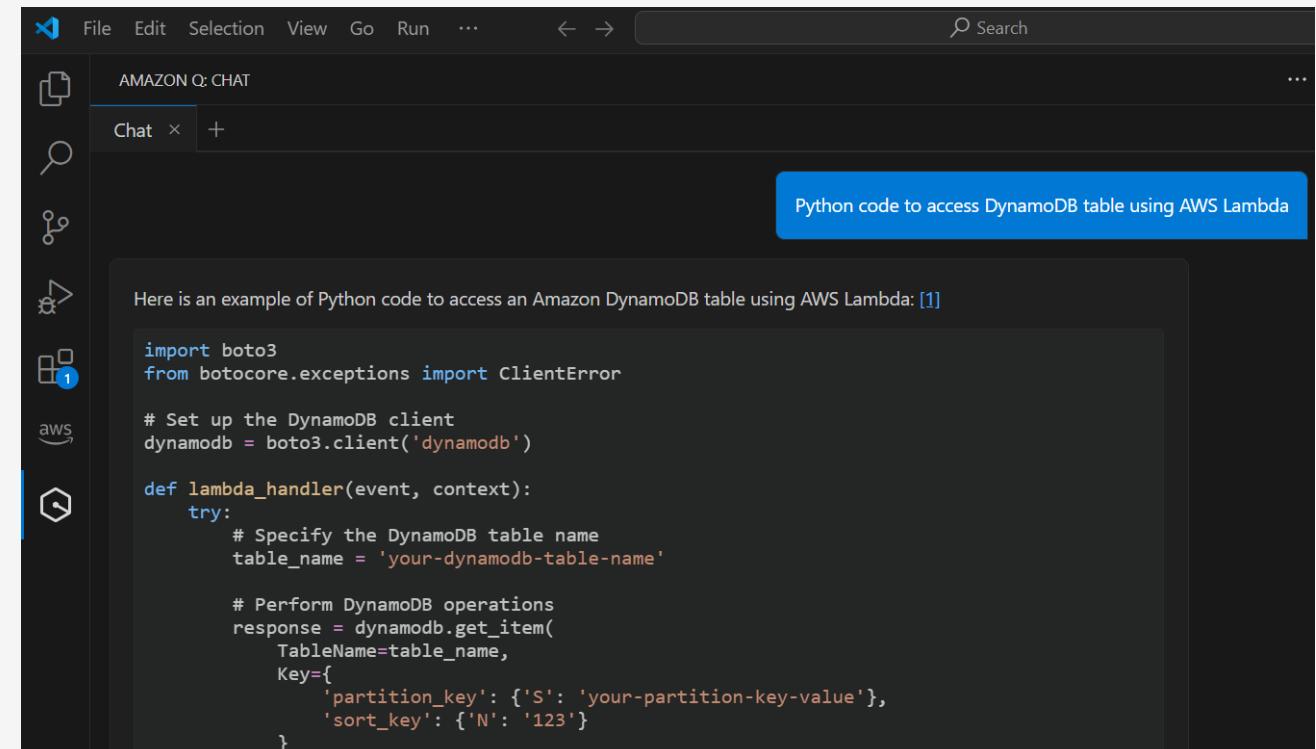


1. What is Amazon Q Developer

The most capable generative **AI-powered assistant** for software development

Amazon Q Developer can assist in :

- Coding and Testing
- Upgrading applications (Upgrade from Java 8 to Java 17)
- Diagnosing errors
- Performing security scanning and fixes
- Optimizing AWS resources



The screenshot shows the Amazon Q Developer interface. On the left is a sidebar with icons for File, Edit, Selection, View, Go, Run, and a search bar. The main area has a dark background with a light-colored chat window titled "AMAZON Q: CHAT". Inside the chat window, there is a tab labeled "Chat" with a plus sign next to it. A blue callout box points to the "Chat" tab with the text "Python code to access DynamoDB table using AWS Lambda". Below the tabs, a message reads: "Here is an example of Python code to access an Amazon DynamoDB table using AWS Lambda: [1]". The code itself is as follows:

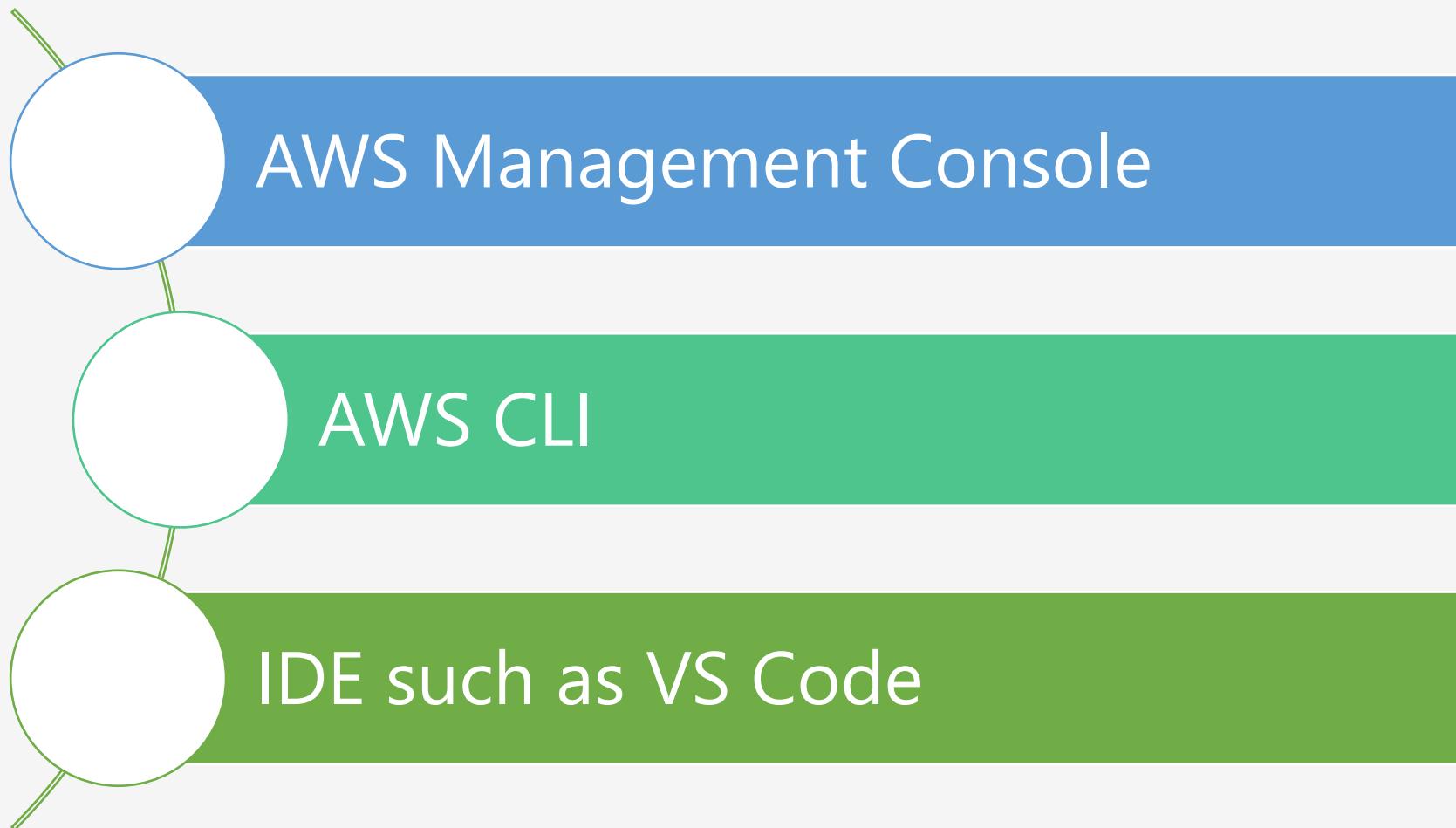
```
import boto3
from botocore.exceptions import ClientError

# Set up the DynamoDB client
dynamodb = boto3.client('dynamodb')

def lambda_handler(event, context):
    # Specify the DynamoDB table name
    table_name = 'your-dynamodb-table-name'

    # Perform DynamoDB operations
    response = dynamodb.get_item(
        TableName=table_name,
        Key={
            'partition_key': {'S': 'your-partition-key-value'},
            'sort_key': {'N': '123'}
        }
    )
```

Where to access Amazon Q - Developer from ?



Supports popular IDE's such as JetBrains, IntelliJ IDEA, Visual Studio, and VS Code.

*Use Case :
Inventory Management for a Retail Store*

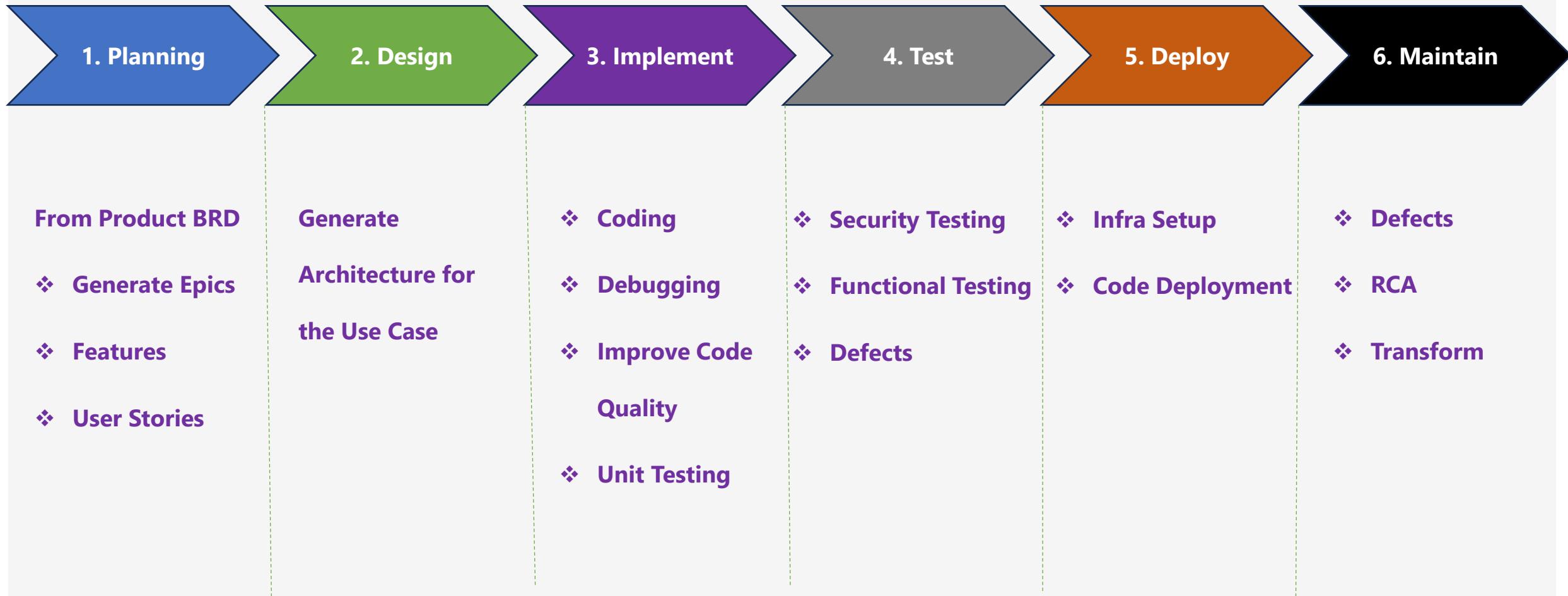


Business Requirement

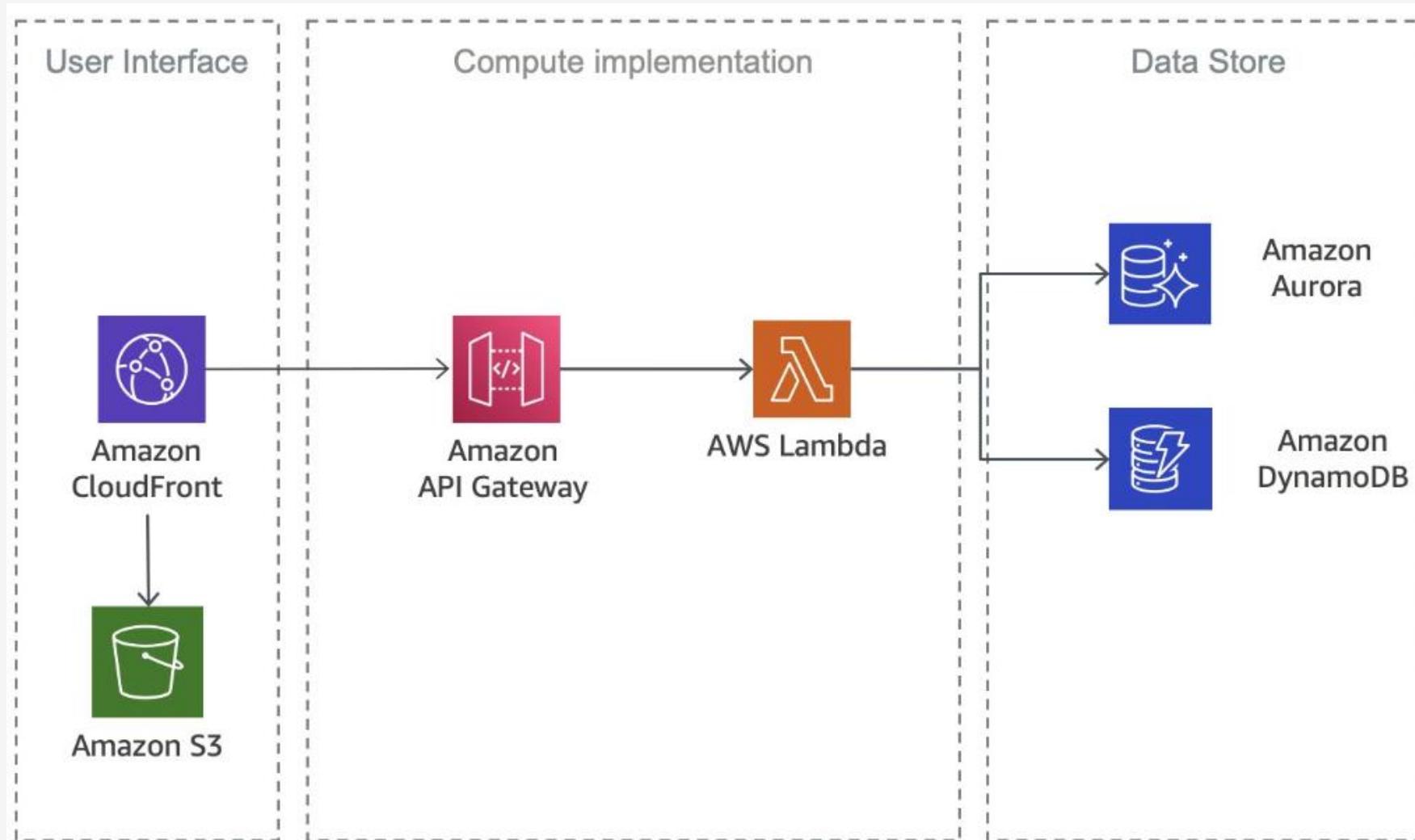
Epic

As an store inventory manager, I should be able to check the inventory details for items in the store and update them when new items are received at the store.

How Amazon Q can help in Software Development Lifecycle



AWS Micro-Services Architecture



*AWS GenAI Service :
Amazon Bedrock (Serverless)*

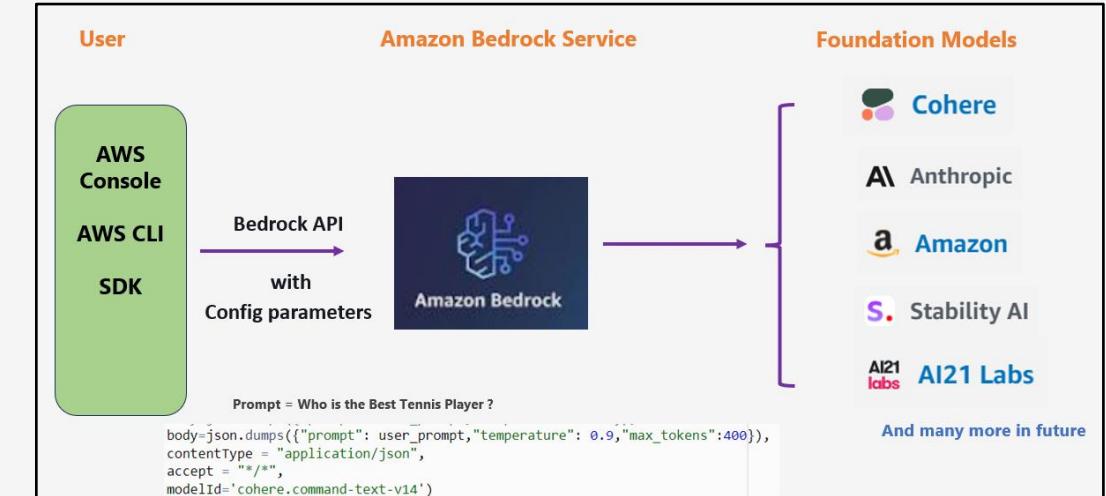


Service Offerings in GenAI from AWS – Amazon Bedrock

1. AWS Service Description

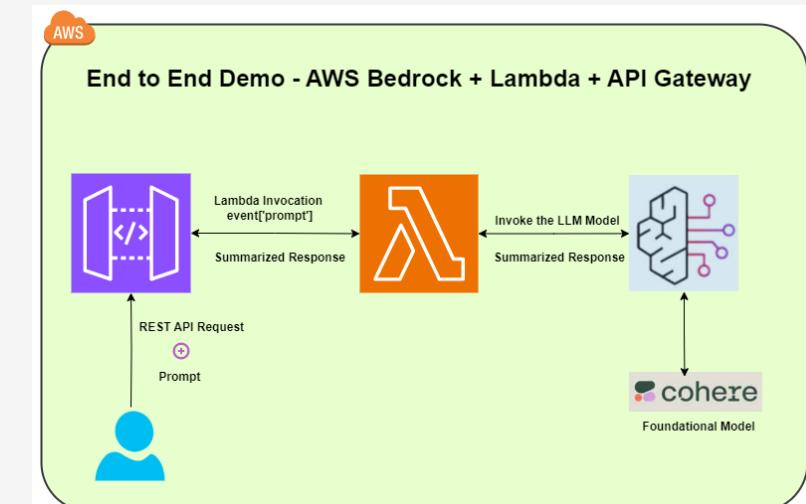
AWS Bedrock

- Amazon Bedrock is a **fully managed, serverless service from AWS**
- Makes **base FM's from Amazon and third-party model providers** accessible through an API

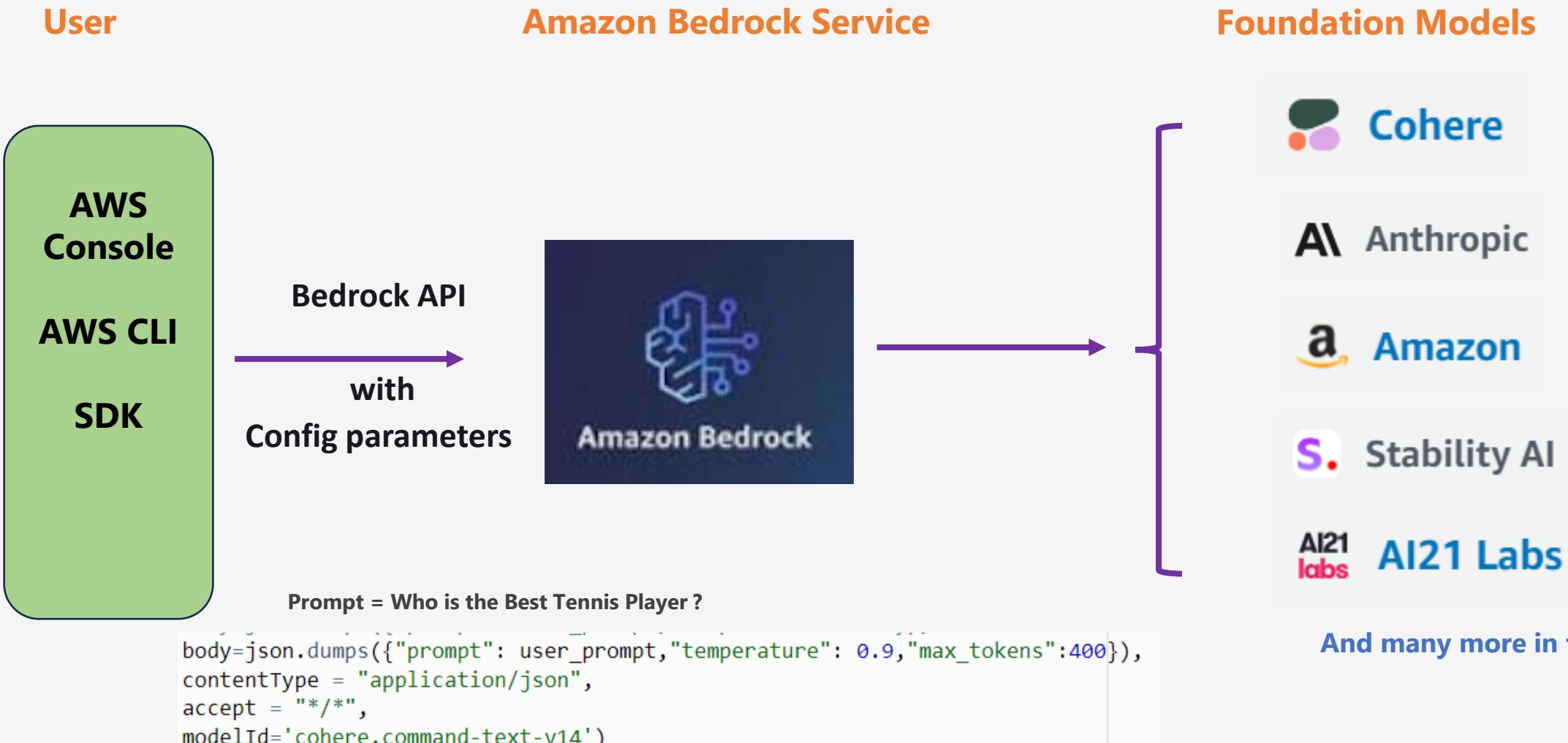


2. Use Cases

- Text Summarization, Code Generation** and any other Gen AI use case
- AWS manages the infrastructure, host the model** and manage all the underlying complexities.
- Bedrock allows **access** the Foundation Model simply via and **API request**.
- Serverless Service** - No need to provision upfront and pay as you use service
- Users** – Developers and Business teams



How does Amazon Bedrock work – High Level ?



Service Offerings in GenAI from AWS – Amazon Bedrock

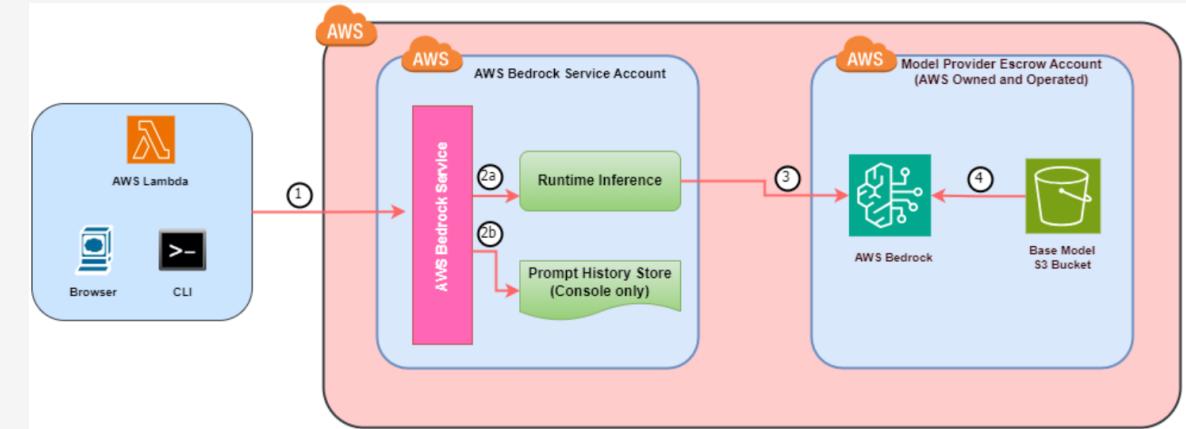
3. Architecture, Security & Pricing

Architecture & Security

- Foundation models deployed **within AWS VPC**
- As **model is deployed with AWS**, the prompt, inference and training data are **exposed outside customer environment**.
- Security - **Bedrock Guardrails**

Pricing – [Link](#)

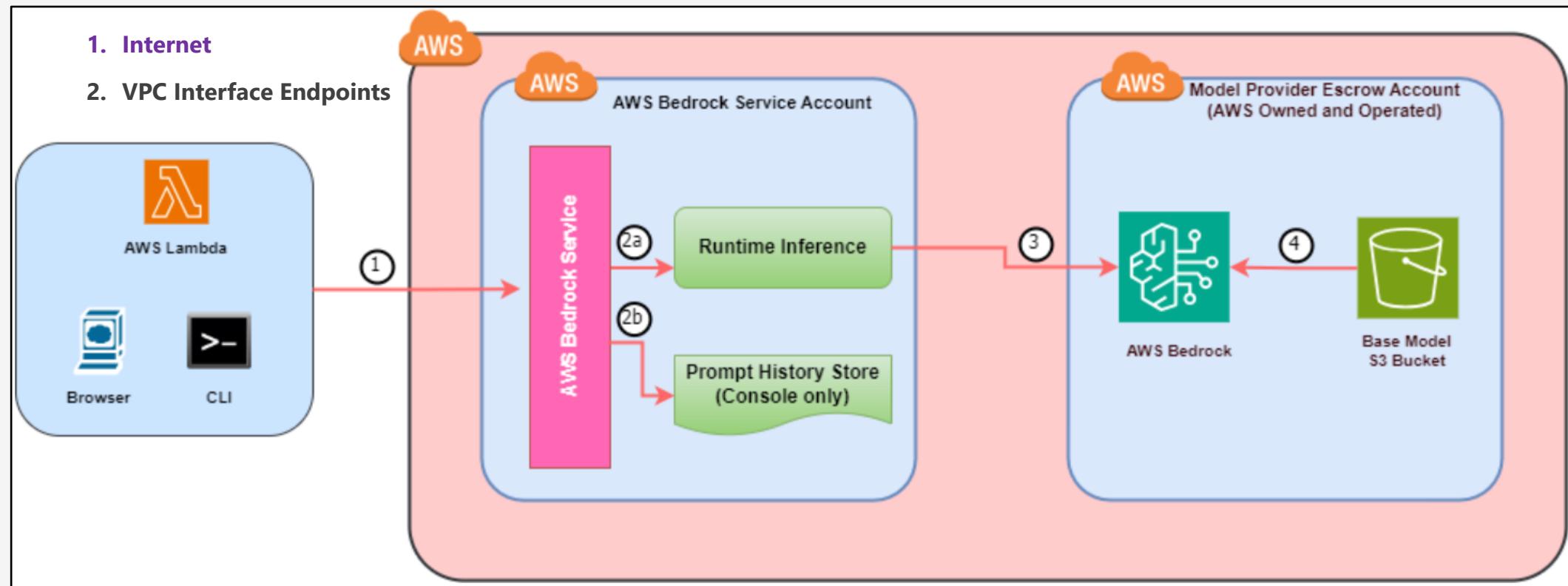
- Based on **number of API Request** and number of Tokens
- Overall **lower pricing than self-hosted models**
- **Fine tuning cost** can be prohibitive



4. Performance

- Since the model is hosted and managed by AWS, **end users have minimal control on model performance**
 - Latency can be an issue while using complex models
 - Throughput as defined in Bedrock service

Amazon Bedrock – Architecture



- **Runtime Inference** : Used to redirect to right model endpoint based on the API request
- **Base Model** : Baseline model provided to every AWS account

Services (2)	
Search	
bedrock X Clear filters	
Service Name	Type
com.amazonaws.us-east-1.bedrock	Interface
com.amazonaws.us-east-1.bedrock-runtime	Interface

Service Offerings in GenAI from AWS – Amazon Bedrock

5. Fine Tuning, Model Choices & Safety

Fine Tuning

- Enables capability to Fine Tune any model which is allowed to be Fine Tuned by the Model provider

Model Choices

- Large number of Open Source and Proprietary Models



6. Scalability, Availability & Monitoring

- Offered as a **Serverless Service** by AWS.
- AWS Manages**
 - Scalability - Autoscaling
 - Availability – Multi AZ
 - Monitoring - CloudWatch



Bedrock Agents, Knowledge Bases, Model Evaluation and Guardrails

1. Amazon Bedrock Knowledge Bases

- AWS Managed service which allows AWS customers build **RAG based applications** without having to **build or code custom integrations to data sources**



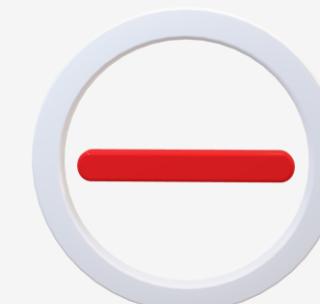
2. Amazon Bedrock Agents

- Enable generative AI applications to **execute multistep tasks across company systems and data sources.**
- Example – Processing Insurance Claim



3. Guardrails

- Guardrails for Amazon Bedrock help you implement safeguards for your generative AI apps
- Example – Filtering undesirable interactions, PII redaction, Safety and Privacy



AWS GenAI Service
PartyRock



AWS Generative AI App - PartyRock

- PartyRock is a **generative AI app-building playground** powered by **Amazon Bedrock**
- **No coding or AI/ML experience** required to build apps
- **No AWS account needed** and login with any social ID
- Free up to certain limit
- Experiment, build apps with drag and drop
- Write **prompts and evaluate various foundation models** for response.



*AWS GenAI Service :
Amazon Sagemaker Jumpstart*



Service Offerings in GenAI from AWS – AWS Sagemaker Jumpstart

1. AWS Service Description

AWS Sagemaker Jumpstart (Part of Sagemaker focusing on GenAI)

- **Fully managed Service** allows access to different models in area of :
 - Foundation Models
 - Computer Vision Models
 - NLP Models
 - Model Evaluation using Amazon Sagemaker Clarify
 - Build and train your own models (with Sagemaker Service)

Discover foundation models and deploy with SageMaker's enterprise-ready features

AVAILABLE ON SAGEMAKER JUMPSTART

AI21labs	Meta AI	cohere	Hugging Face	stability.ai	LightOn	databricks	alexa
Models Jungie-2 Ultra, Mid Contextual answers Summarize Paraphrase Grammatical error correction Tasks Text generation Long-form generation Summarization Paraphrasing Chat Information extraction	Models Llama 2 7B, 13B, 70B Code Llama 7B, 13B, 34B Open LlaMA Tasks Question answering Chat Summarization Paraphrasing Sentiment analysis Text generation	Models Command Cohere Light Tasks Text generation Information extraction Question answering Summarization	Models Falcon-7B, 40B, 160B, Mistral-7B RedPajama MPT-7B Bloomic-176B Pan-T5 DistilGPT2 GPT NeoXT Bloom Tasks Upscaling Imprinting Tasks Generate photo-realistic images from text input. Improve quality of generated images	Models Stable Diffusion XL 1.0 2.1 base Upscaling Imprinting Tasks Machine translation Question answering Summarization	Models Lyra-Fr 10B, Meni Tasks Text generation Keyword extraction Information extraction Question answering Summarization Sentiment analysis Classification	Models Dolly Tasks Question answering Chat Summarization Paraphrasing Sentiment analysis Text generation	Models AlexaTM 20B Tasks Machine translation Question answering Summarization Annotation Data generation

2. Use Cases

- **Text Summarization, Code Generation** and any other Gen AI use case
- Allows you to **provision infrastructure to deploy and fine tune foundation models**
 - Open source models
 - Proprietary models
- **Users of this service** – Machine Learning Engineers, Data Scientists, Developers

Amazon SageMaker

SageMaker Jumpstart

(GenAI focused)

Service Offerings in GenAI from AWS – AWS Sagemaker Jumpstart

3. Architecture, Security & Pricing

Architecture & Security

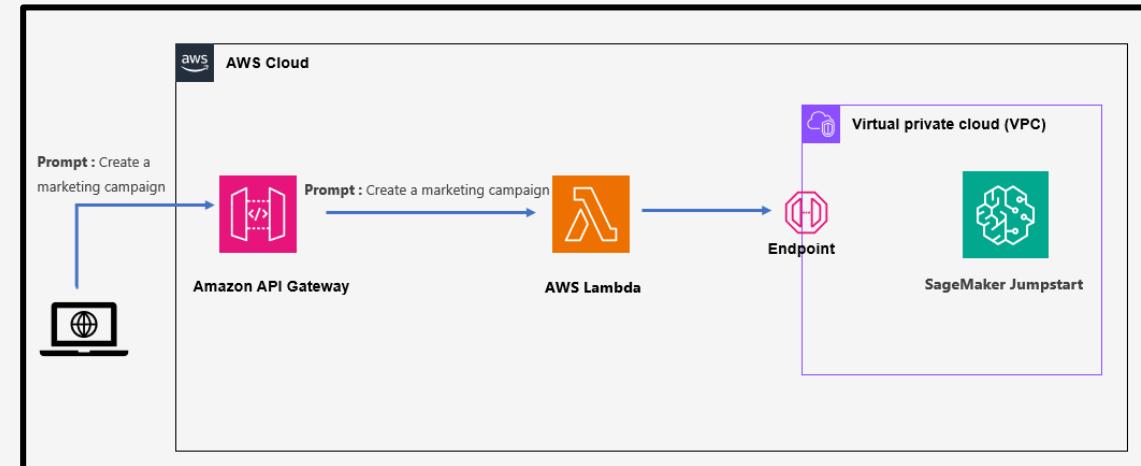
- Deployed in Customer VPC
- Full control over who has access to :
 - Prompt
 - Model Inference
 - Training Data

Pricing

- Based on Provisioning Capacity – [Link](#)

4. Performance

- Deploy workloads sensitive to (Control through Instance Sizing)
 - Inference latency
 - Throughput
- Enables high-performance cost-effective inference.



Service Offerings in GenAI from AWS – AWS Sagemaker Jumpstart

5. Fine Tuning, Model Choices & Safety

Fine Tuning

- Enables capability to Fine Tune any model which is allowed to be Fine Tuned

Model Choices

- Large number of Open Source and Proprietary Models (and build, train & deploy your own)

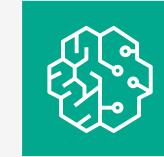


{"Anthracyclines":
"Antibiotic drugs
used in cancer"}

Training Data



Base Model – Llama 3



SageMaker Jumpstart



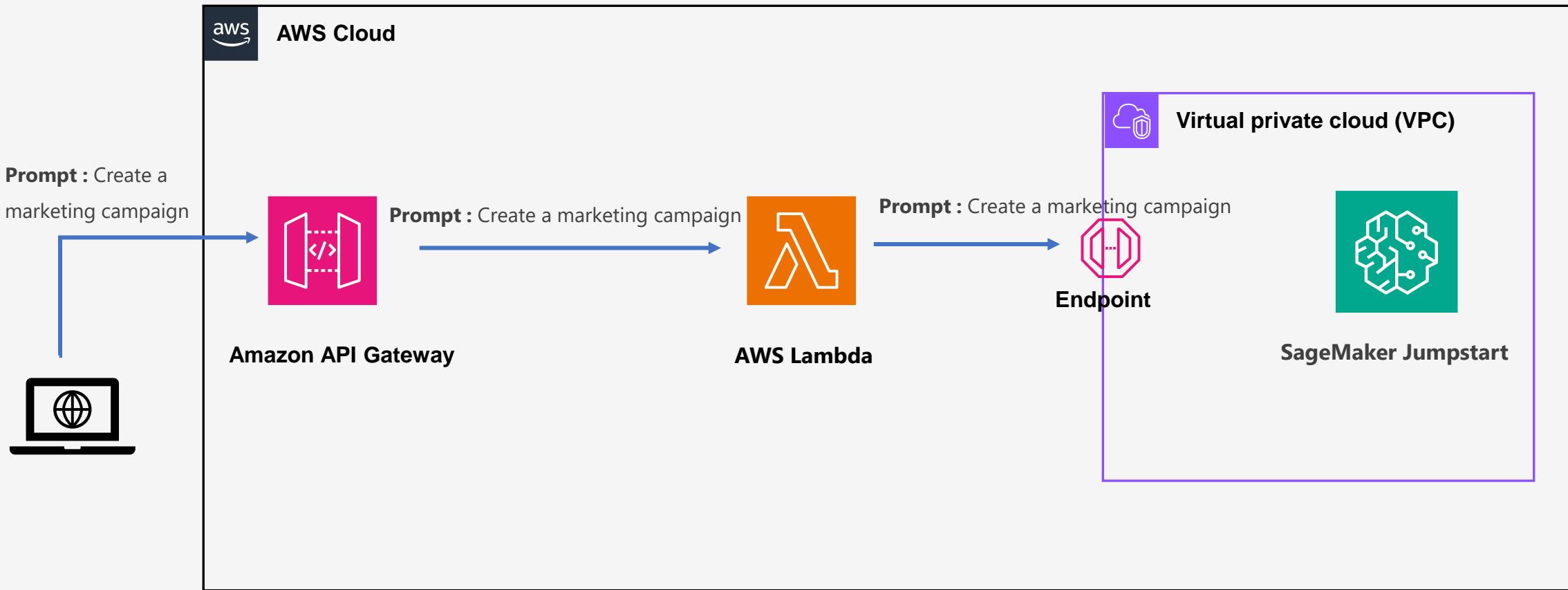
Fine Tuned
Llama 3

6. Scalability, Availability & Monitoring

- **AWS Fully Managed services.**
- **AWS Manages**
 - **Scalability** - Autoscaling
 - **Availability** – Multi AZ
 - **Monitoring** - CloudWatch



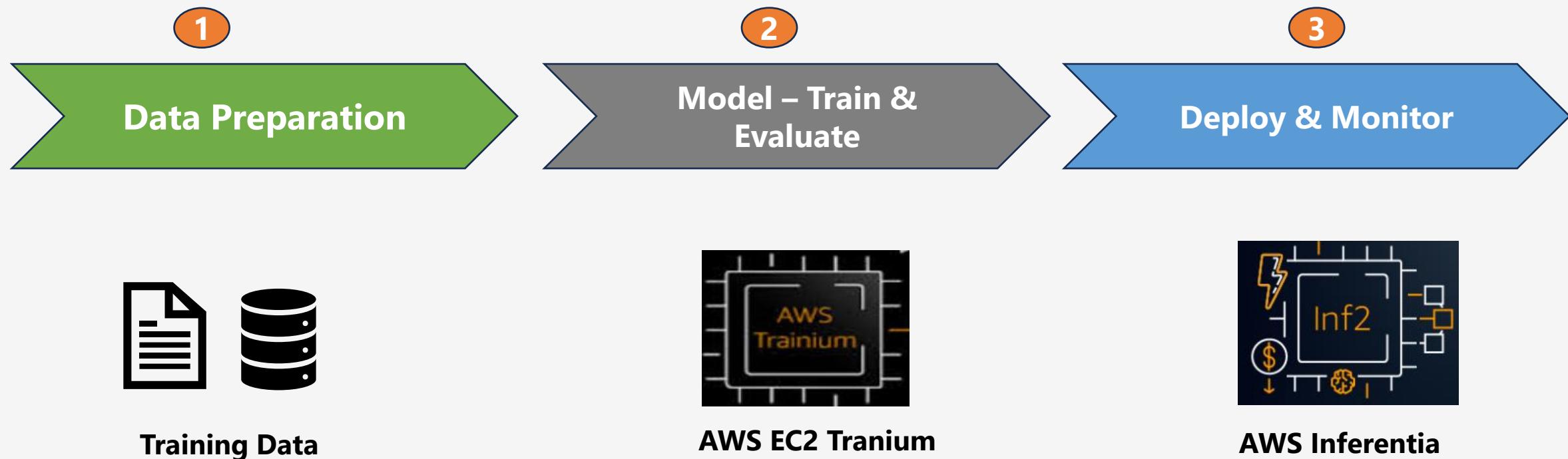
Amazon Sagemaker Jumpstart Deployment



*AWS GenAI Service :
AWS Tranium and Inferentia*



Service Offerings in GenAI from AWS – AWS EC2 Tranium & Inferentia2



Infrastructure for Training and Deploying Large Language Models

Service Offerings in GenAI from AWS – AWS EC2 Tranium & Inferentia2

1. AWS Service Description

AWS EC2 Tranium

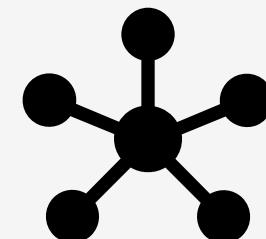
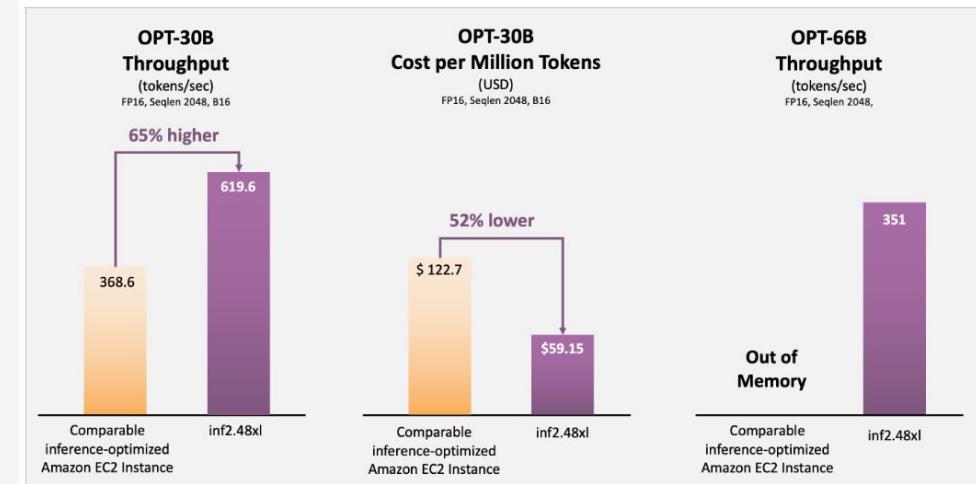
- Type of **EC2 instance** but **purpose built ML chip by AWS** for Deep Learning Models
- Reduce Training time** for **LLM's** being built from scratch - **100B+ parameter** models.

AWS Inferentia2

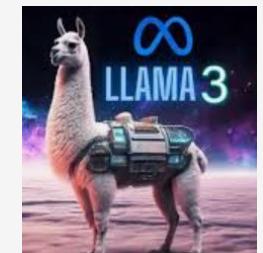
- Optimized to deploy **LLM's and diffusion models for Inferencing (Output)**

2. Use Cases

- Text Summarization, Code Generation, QnA** and any other Gen AI use case
- Build your own **domain specific** Foundation Model (**FM**) from **scratch**
 - BloombergGPT (50B)**
- Take an **open-source model** such as Llama-3 and further **train & deploy it.**
- Users of this service** – Machine Learning Engineers, Data Scientists, Developers



**Build, Train & Deploy
your own Model**



**Deploy & Fine Tune
Open Source Model**

Service Offerings in GenAI from AWS – AWS EC2 Tranium & Inferentia2

3. Architecture, Security & Pricing

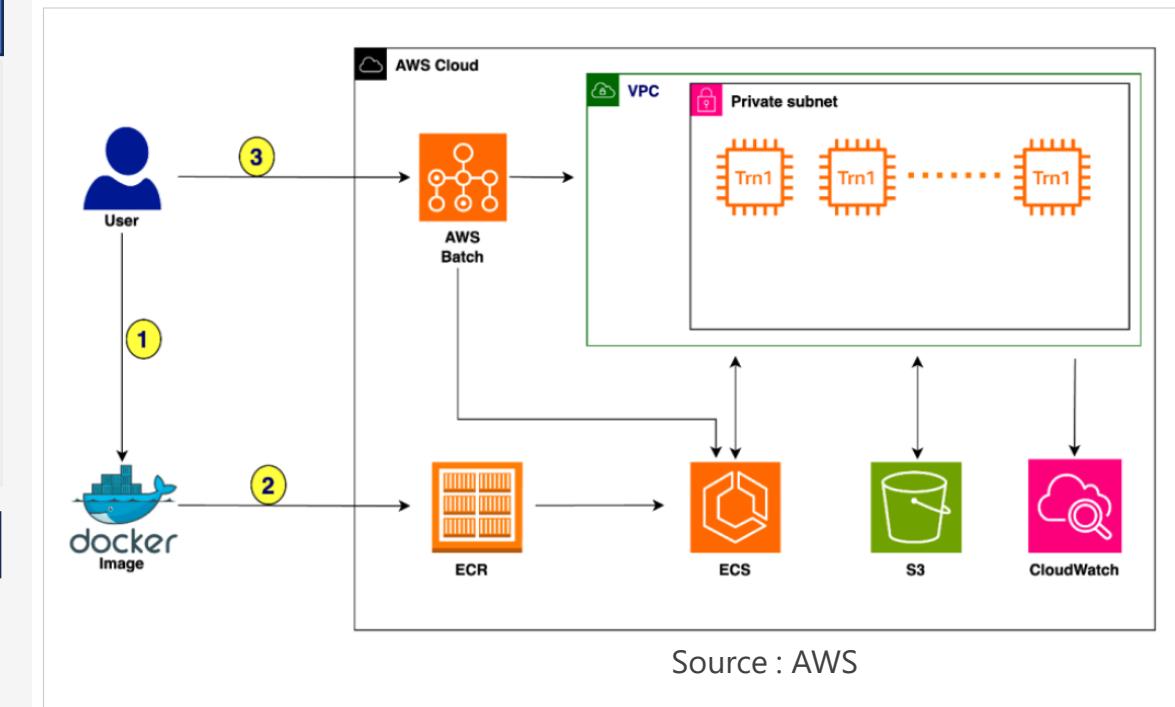
- Deployed in **Customer VPC** - [Link](#)
- Full control** over who has **access to**:
 - Training Data
 - Prompt
 - Model Inference

Pricing

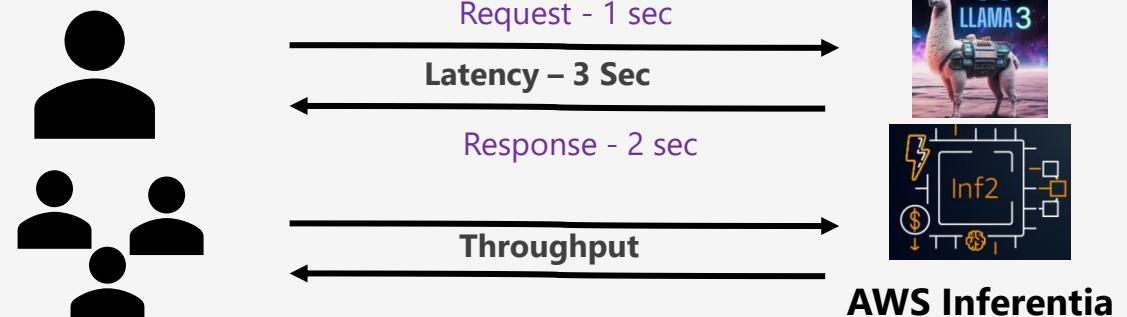
- Based on **Provisioning Capacity** - [Link](#)

4. Performance

- Control Inference Latency & Throughput with right Instance**
- Inference latency - Chatbot
- Throughput – Q&A
- Enables high-performance cost-effective inference.**



Source : AWS



AWS Inferentia

Service Offerings in GenAI from AWS – AWS EC2 Tranium & Inferentia2

5. Fine Tuning, Model Choices & Safety

Fine Tuning

- Can easily Fine Tune the models with custom data

Model Choices

- Deploy your own model, open source and proprietary models

Safety

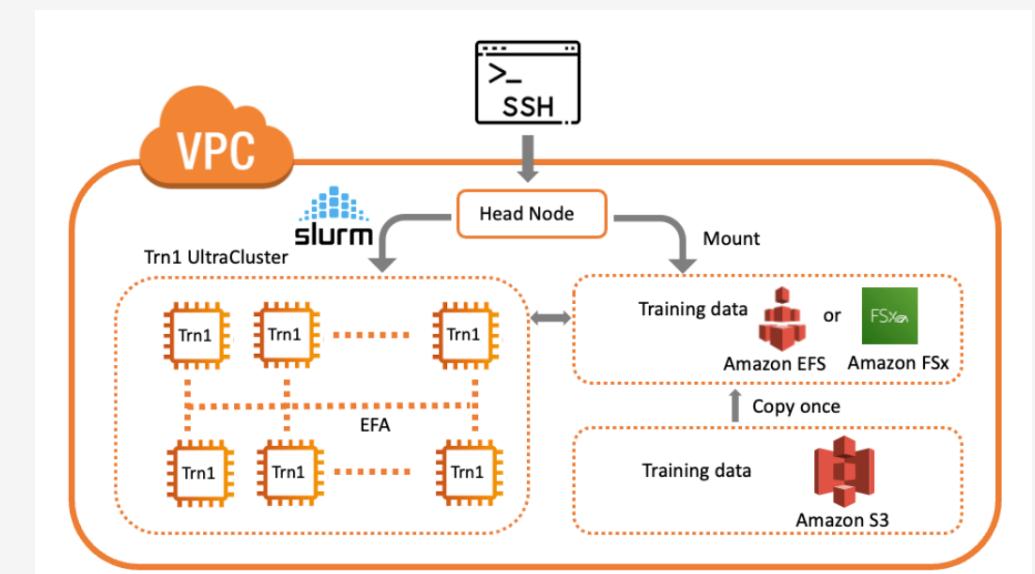
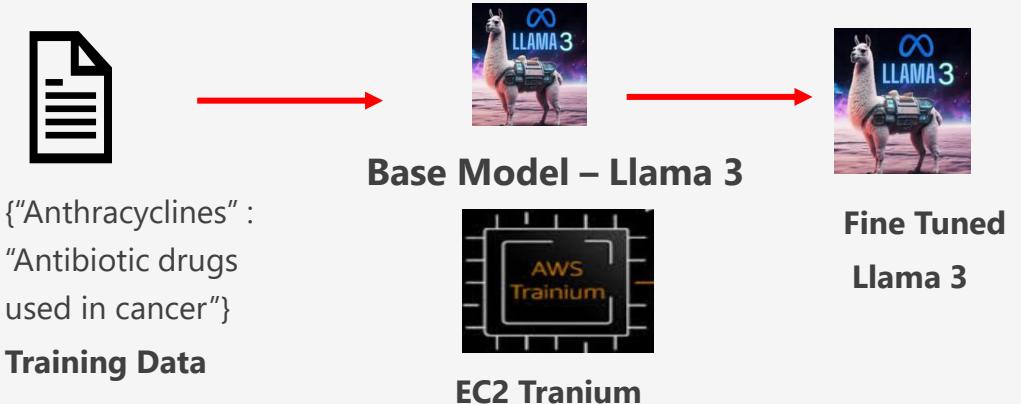
- Custom build controls for your Model – Toxicity Control, Hallucination

6. Scalability, Availability & Monitoring

• Need to design the solution for

- High Availability – Multi AZ
- Scalability – Auto scale
- Monitoring – CloudWatch Alarms

• Not available as out of box feature



AWS EC2 Tranium & Inferentia2 - Demo

AWS EC2 Tranium & Inferentia2

Demo Time



Domain 3 :

Applications of Generative AI

Bedrock Knowledge Bases, Agents

and Guardrails



Amazon Bedrock Knowledge Bases

What is Knowledge Bases for Amazon Bedrock ?

Challenges with Foundation Models

- Lack of proprietary information – Enterprise Information
- No up-to-date information post training cut off date

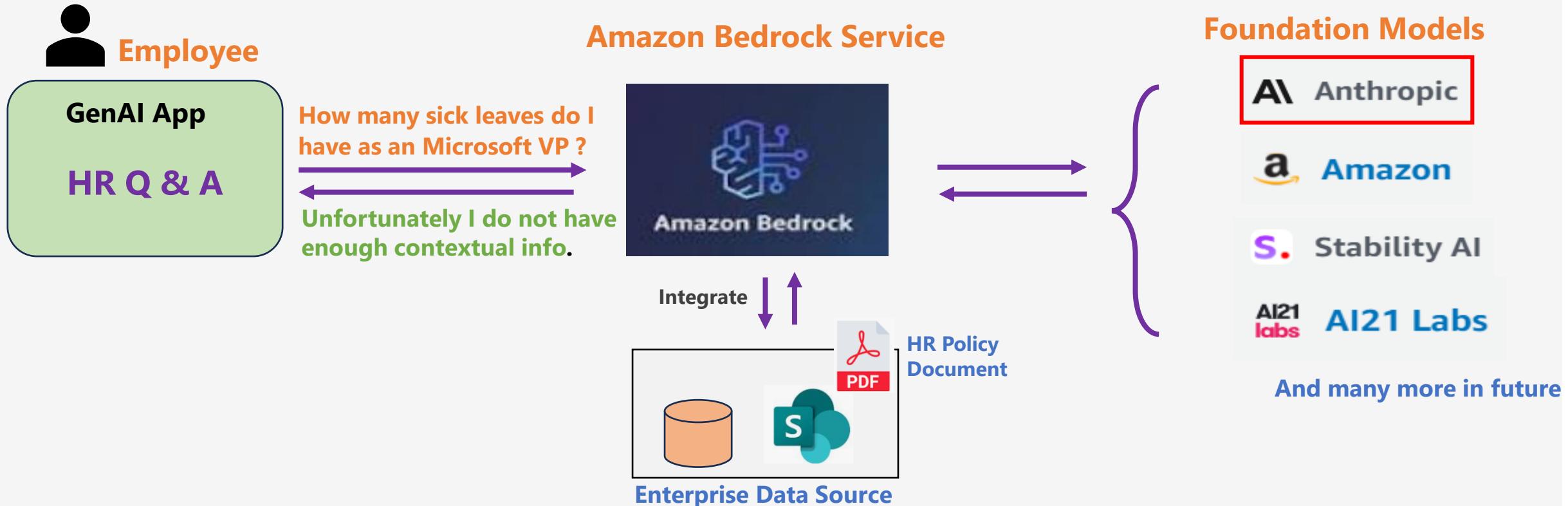
Retrieval Augmented Generation

- RAG equips foundation models (FMs) :
- Proprietary information by fetching data from company data sources
- Up-to-date knowledge

Knowledge Bases for Amazon Bedrock

- is a fully managed capability that helps you implement the entire RAG workflow
- No need to build custom integrations to data sources
- Manages Data Ingestion, Chunking, Creation of vector embedding, storing of vector embeddings

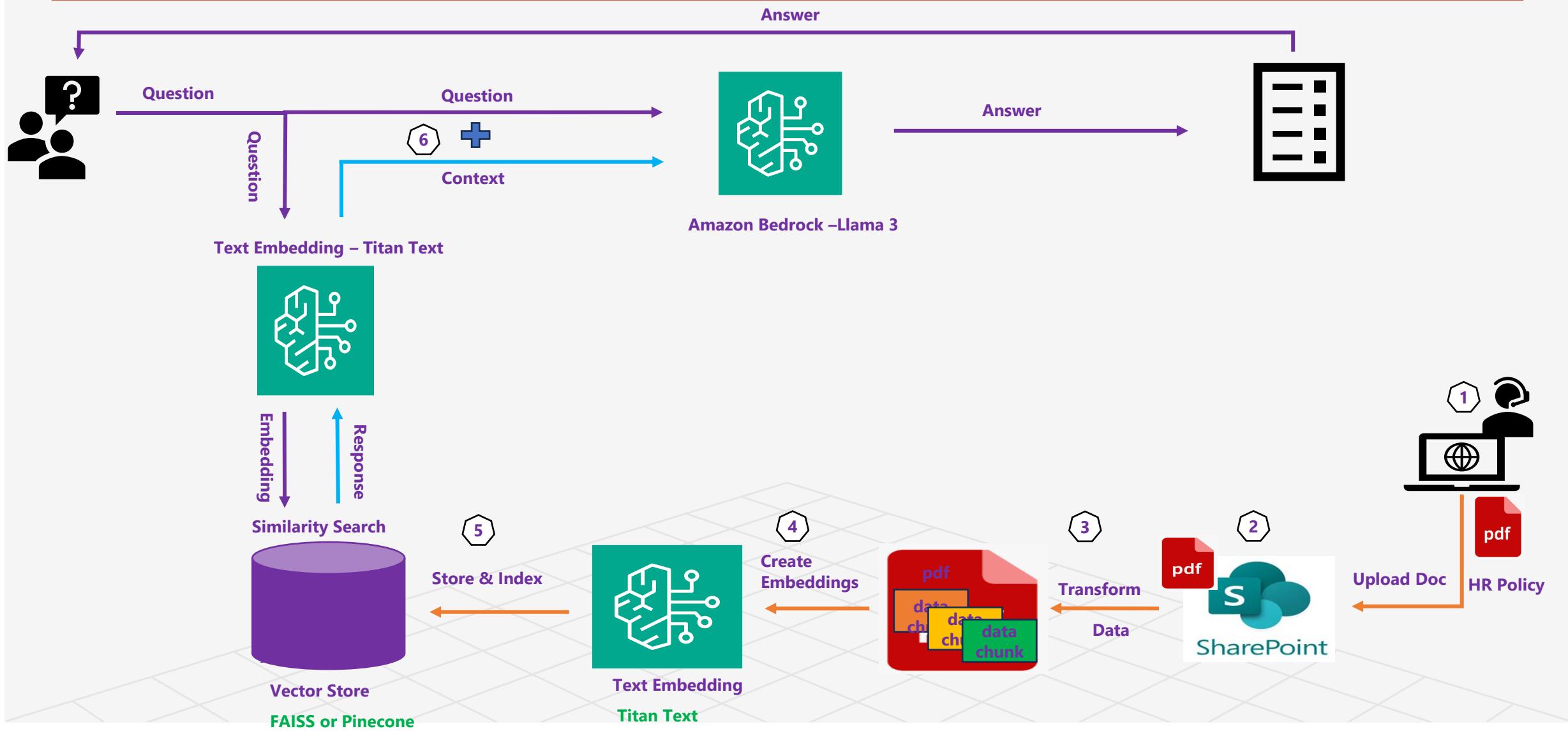
Retrieval Augmented Generation (RAG)



What is RAG ?

RAG is the process of **optimizing** the output of a large language model by **supplementing the knowledge** of LLM with **internal organization data source**.

HR Question & Answer App with RAG - Solution Architecture



e-Learning Q and A App for Generative AI



AWS Cloud



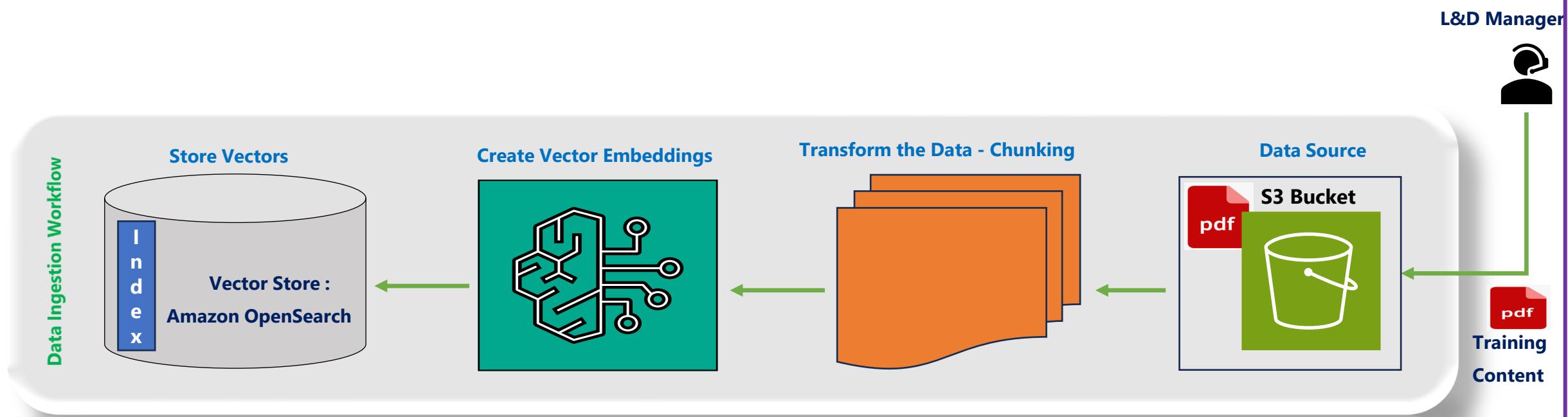
Prompt = What is Amazon Bedrock ?

Response

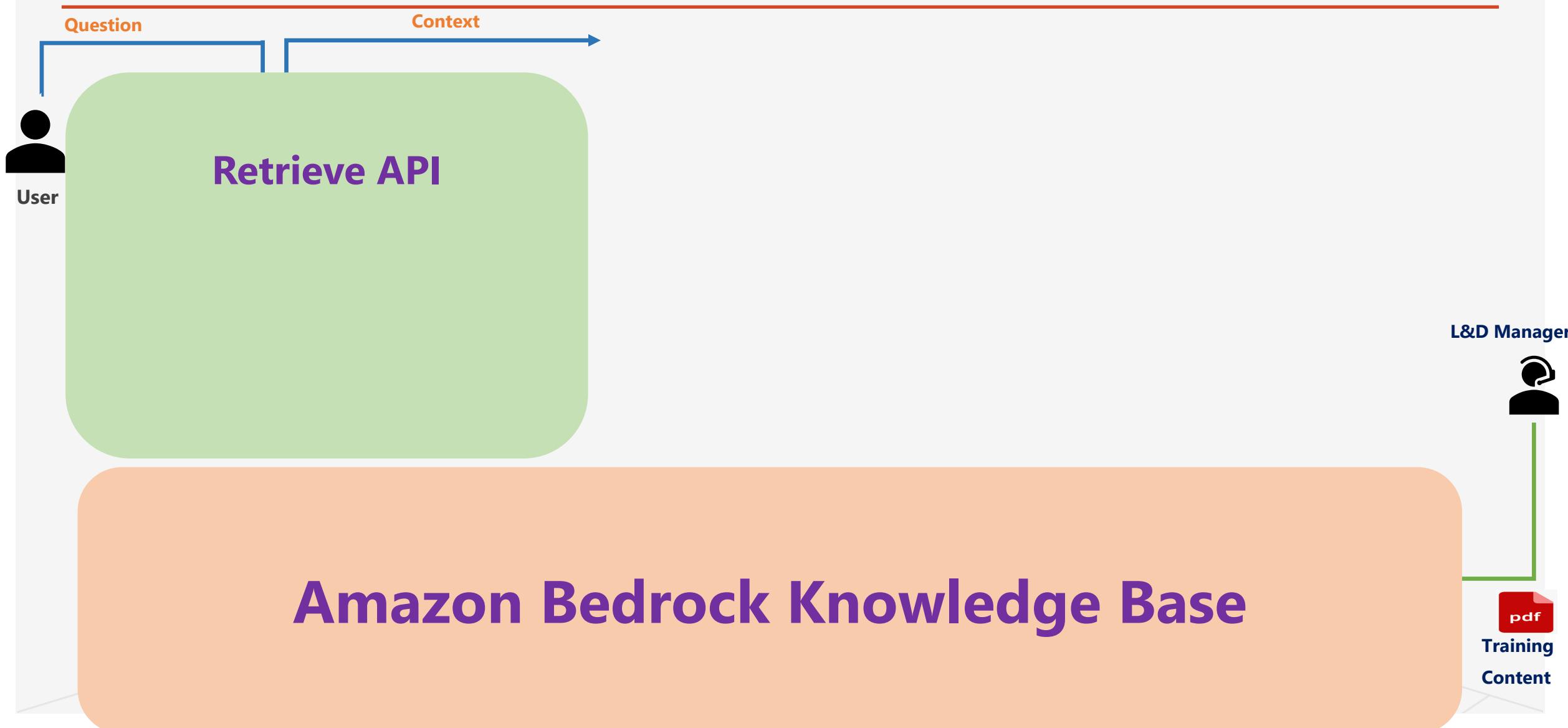


S3 Bucket

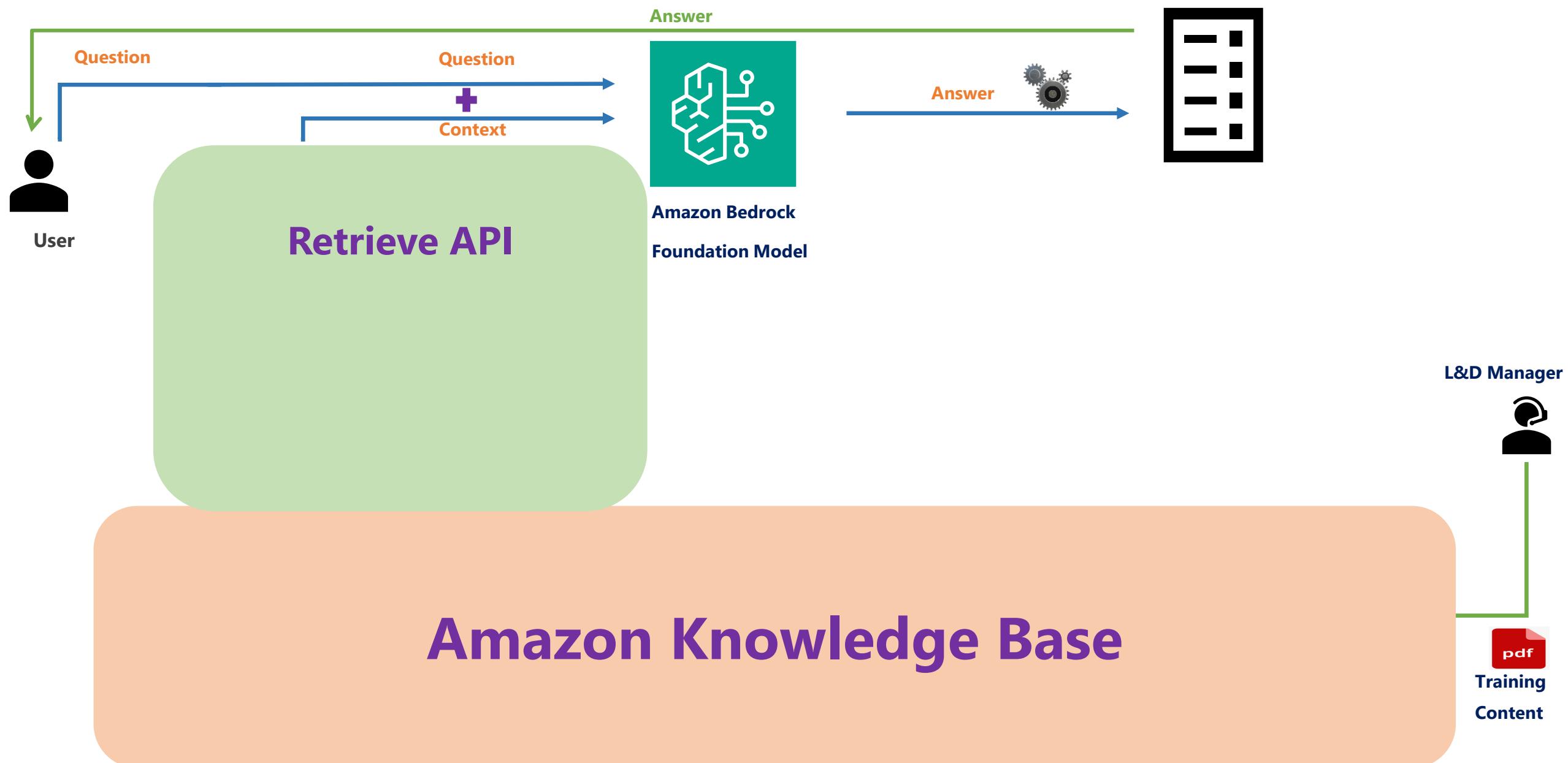
E-Learning Solution using RAG - Amazon Bedrock Knowledge Base + LLM

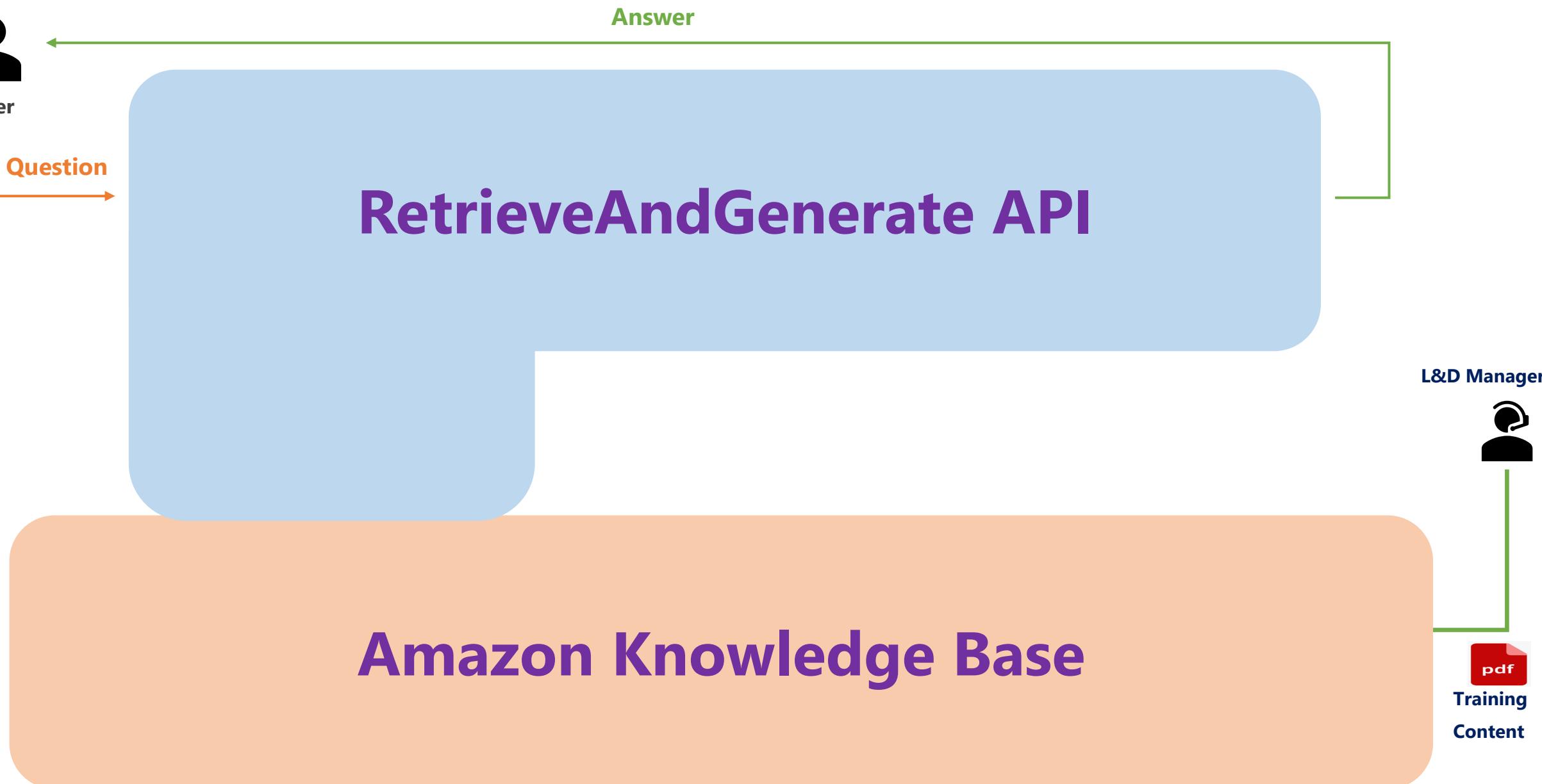


L&D App with Amazon Knowledge Base (Retrieval Augmented Generation) - Architecture



L&D App with Amazon Knowledge Base (Retrieval Augmented Generation) - Architecture





Data Ingestion Workflow - Options

1. Data Sources

- ❖ S3, Sharepoint, Salesforce

2. Chunking

- ❖ Default

- ❖ Fixed Size

3. Embedding Model

- ❖ Cohere

- ❖ Amazon Titan

4. Vector DB

- ❖ Amazon OpenSearch Serverless vector store

- ❖ Amazon Aurora

- ❖ Pinecone

- ❖ Redis Enterprise Cloud

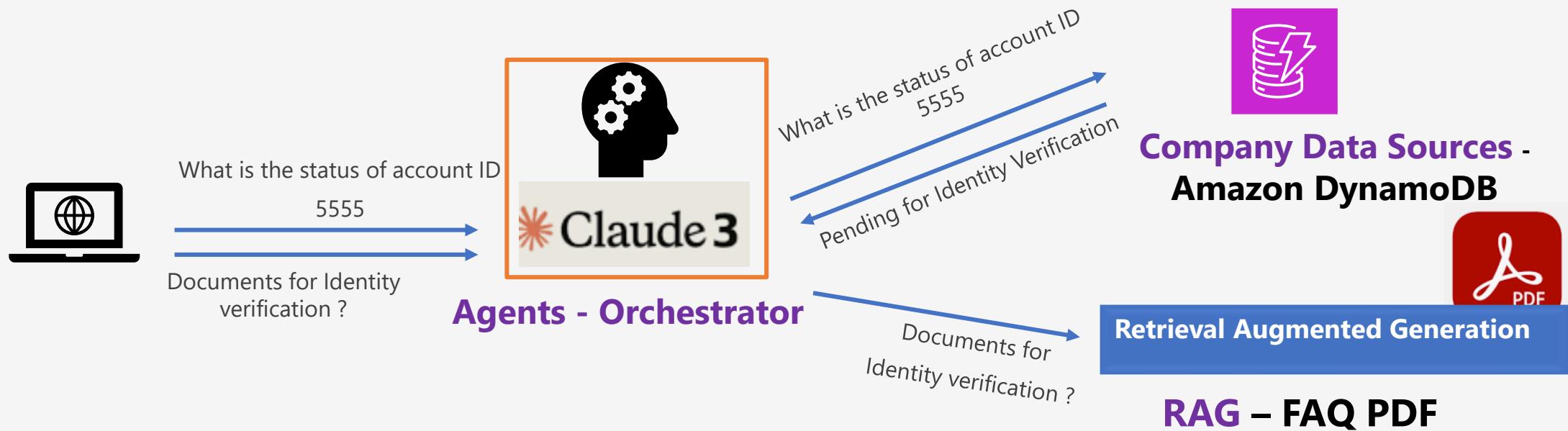


Amazon Bedrock Agents



Amazon Bedrock Agents

Agents **orchestrate interactions** between **foundation models (FMs)**, **data sources**, **software applications**, and **user conversations**.



Bedrock Agents are intelligent Orchestrators :

- Route requests between company data sources - such as RDS, DynamoDB and RAG based solutions
- Prompt Engineering via pre-built templates
- Chain of thought reasoning

*Build Retail Banking Agent
using
Bedrock Agents & Knowledge Bases*

Amazon Bedrock Agents – Retail Banking Agent

John

Check account activation status after 3 days from Bank Website

Thomas

Manju

Bank of Chicago

AccountId = 5555

AccountId = 6666

AccountId = 7777

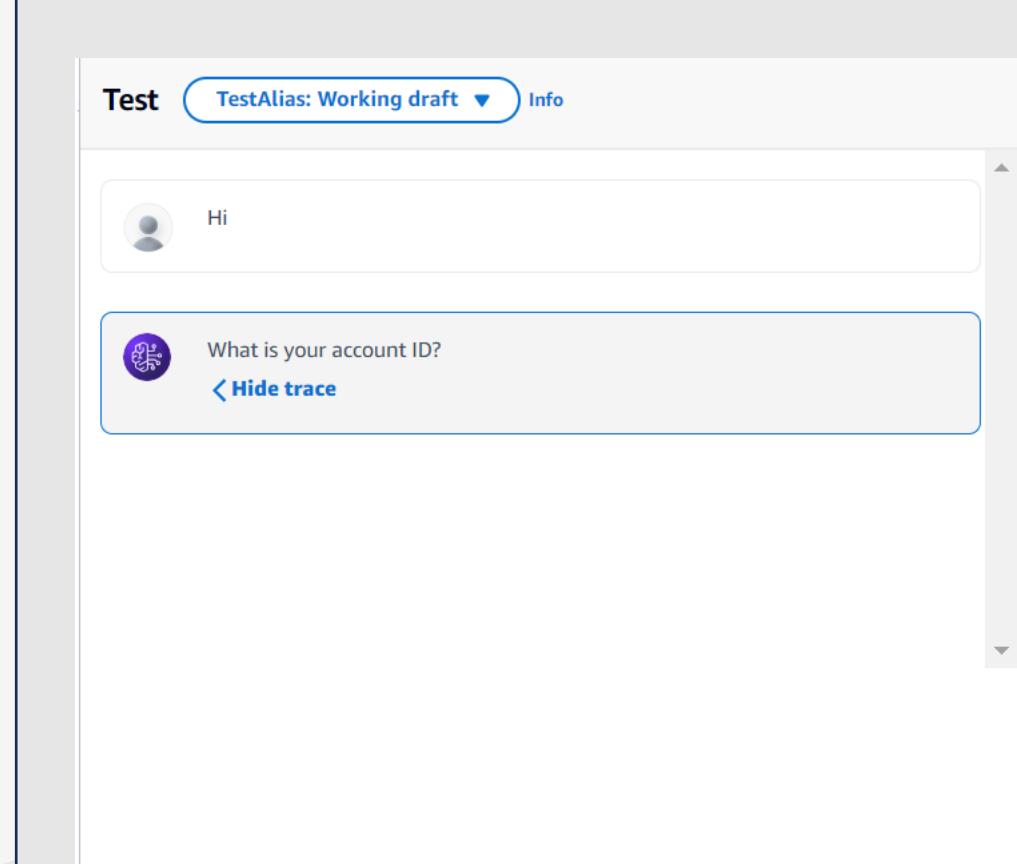
GET UP TO
\$500

Amazon Bedrock Agents – Retail Banking Agent



Manju

AccountId = 7777



www.bankofchicago.com

Test Alias: Working draft

Hi

What is your account ID?
 Hide trace



DynamoDB Table

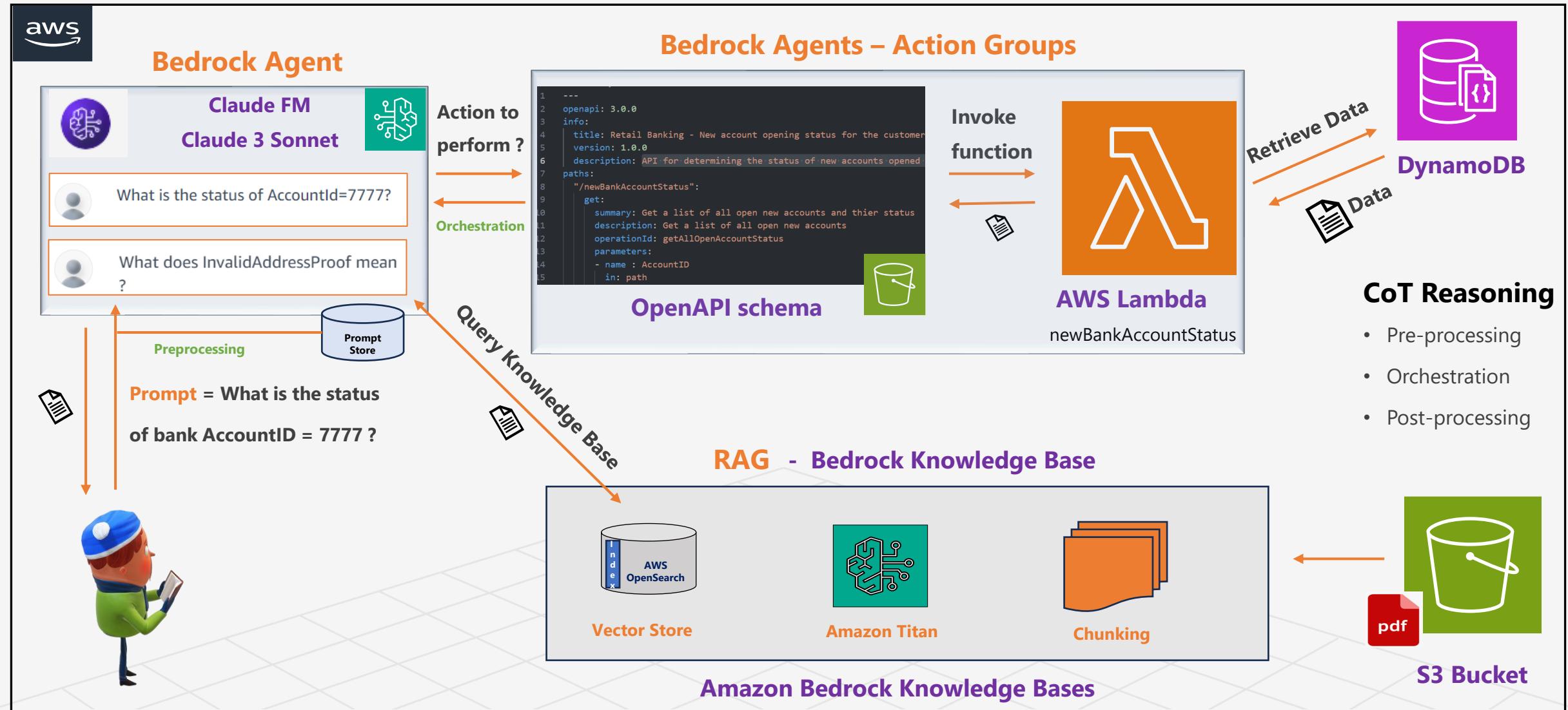
AccountId	AccounName	Account Status	Reason
5555	John	Active	Active
6666	Thomas	Pending	InvalidIdentification
7777	Manju	Pending	InvalidAddressProof

Customer Data - DynamoDB Table

TableName - customerAccountStatus

AccountID	AccounName	AccountStatus	Reason
5555	John	Active	Active
6666	Thomas	Pending	InvalidIdentification
7777	Manju	Pending	InvalidAddressProof

Retail Banking Agent - Amazon Bedrock Architecture



Amazon Bedrock Guardrails





Foundation Model Inference Parameters

Foundation Models and Impact of Inference Parameters

Inference parameters influence the response generated by the model.

Inference Parameters

1. Randomness and diversity

Temperature

Top K

Top P

2. Length

Response Length

Length Penalty

Stop Sequence

3.Repetition

Repetition penalty

Inference Parameter – 1. Randomness and Diversity

Prompt : I hear the hoof beats of

Parameter	Possible Values	Description	Example
Temperature*	Foundation Models use : Probability to construct the words in a sequence		<ul style="list-style-type: none">• Horse --- > 0.4• Wind ----- > 0.2• Unicorns -- - > 0.1• Change in distance --- > 0.05
	Low value (closer to zero)	Model tends to select the higher-probability words.	<ul style="list-style-type: none">• Horse
	High value (closer to 1 or 5)	Temperature further away from zero, the model may select a lower-probability word	<ul style="list-style-type: none">• Change in distance
Top K	1-500	Cutoff where the model no longer selects the words.	if K=50, the model selects from 50 of the most probable words
Top P	0.01-0.99	Caps choices based on the sum of their probabilities.	If Top P=0.5, Horse and Wind selected

Inference Parameter – 2. Length

Prompt : Write an essay on horse

Parameter	Possible Values	Description
Length : Controls the length of the generated response. Helps control cost .		
Max Length	1- 4096 tokens	Specify the maximum number of tokens to use in the generated response.
Stop Sequence	Any keyword	<ul style="list-style-type: none">Configure up to four sequences that the model recognizes.After a stop sequence, the model stops generating further tokens.The returned text doesn't contain the stop sequence.



Inference Parameter – 3. Repetition

Parameter	Possible Values	Description
Repetition : Help control repetition in the generated response.		
Presence penalty	0-5	Use a higher value to lower the probability of generating new tokens that already appear at least once in the prompt or in the completion.
Count penalty	0-1	Use a higher value to lower the probability of generating new tokens that already appear at least once in the prompt or in the completion. Proportional to the number of appearances.
Frequency penalty	0-500	Use a high value to lower the probability of generating new tokens that already appear at least once in the prompt or in the completion. The value is proportional to the frequency of the token appearances (normalized to text length)
Penalize special tokens		Reduce the probability of repetition of special characters – Whitespace, Punctuations...

Amazon Bedrock Logging and Monitoring

- Monitor all parts of your Amazon Bedrock application using **Amazon CloudWatch**
- **Invocation logs** - full request data, response data, and metadata associated with all calls performed in your account

Amazon Bedrock Logging and Monitoring

Metric name	Unit	Description
Invocations	SampleCount	Model invocation logging to collect invocation logs, model input data, and model output data for all invocations
InvocationLatency	Milliseconds	Latency of the invocations.
InvocationClientErrors	SampleCount	Number of invocations that result in client-side errors.
InvocationServerErrors	SampleCount	Number of invocations that result in AWS server-side errors.
InvocationThrottles	SampleCount	Number of invocations that the system throttled.
InputTokenCount	SampleCount	Number of tokens in the input.
LegacyModelInvocations	SampleCount	Number of invocations using <u>Legacy</u> models
OutputTokenCount	SampleCount	Number of tokens in the output.
OutputImageCount	SampleCount	Number of images in the output (only applicable for image generation models).

Link : <https://docs.aws.amazon.com/bedrock/latest/userguide/monitoring.html>

Guardrails for Amazon Bedrock

- Implement **safeguards** customized to your application requirements and **responsible AI policies**
- Guardrails for Amazon Bedrock provides safeguards on top of the native protections of FMs
 - Filter **harmful content** based on your responsible AI policies
 - Block **undesirable topics** in your generative AI applications
 - Block inappropriate content with a **custom word filter**
 - Redact sensitive information (PII) to protect privacy
 - Detect **hallucinations** in model responses using contextual grounding checks





Domain 3 :

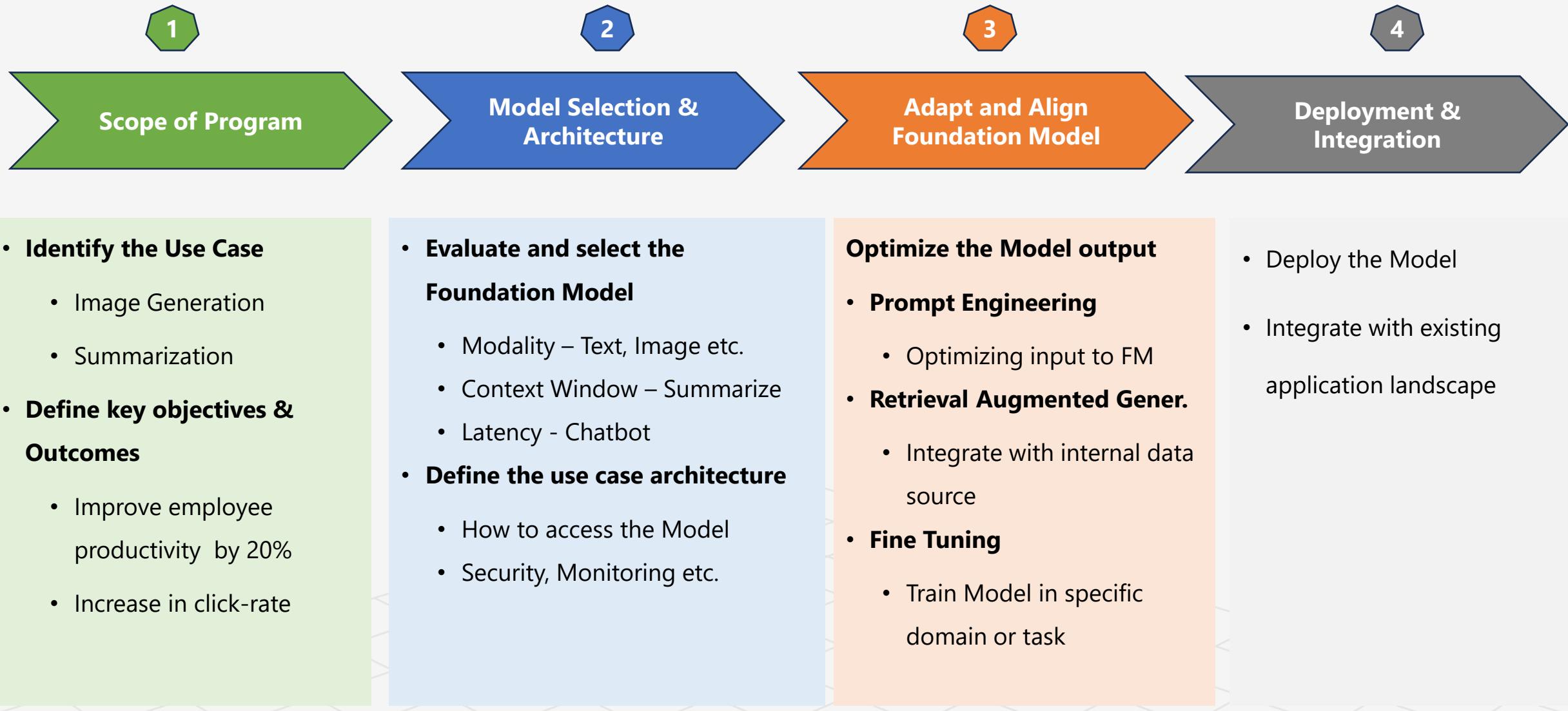
***Prompt Eng, Foundation Model Selection,
Model Evaluation, Fine Tuning***

Generative AI

Application Lifecycle



Generative AI Application Lifecycle

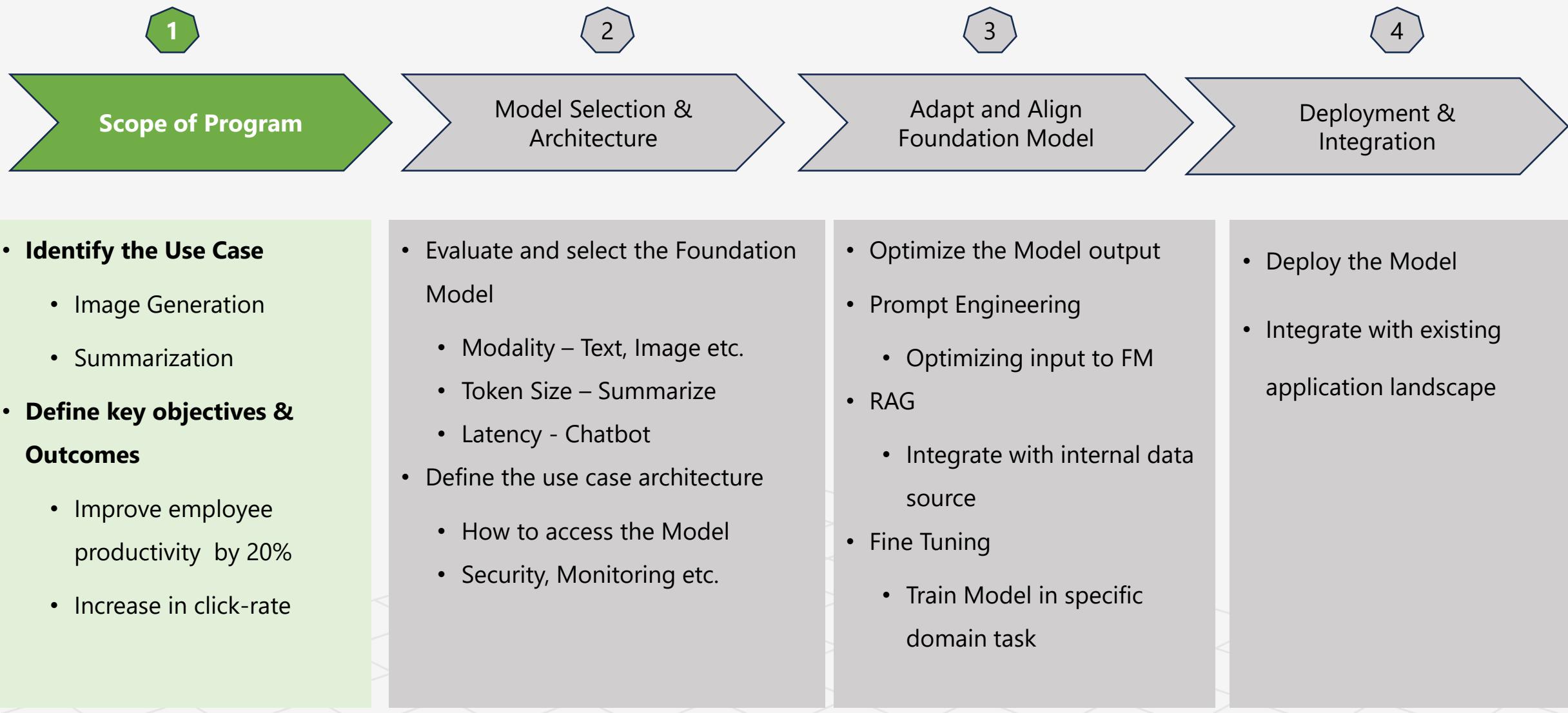


Phase 1 :

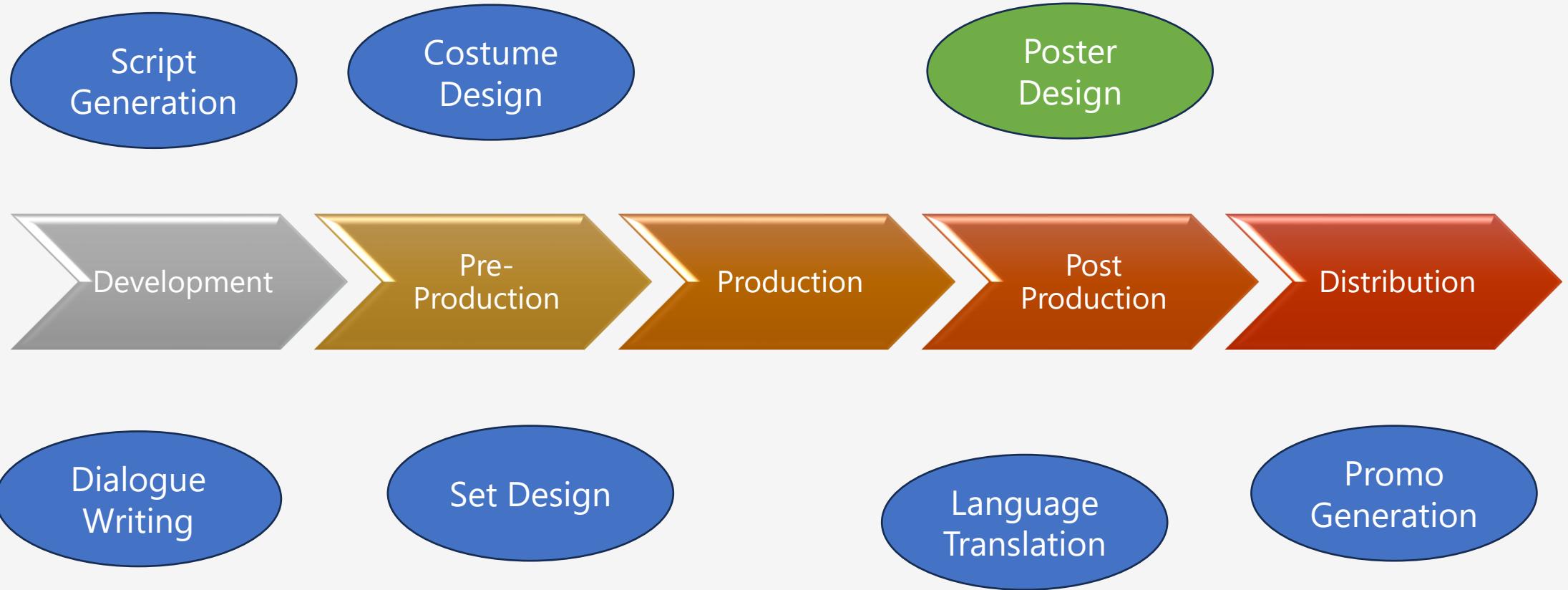
GenAI Use Case Identification



Generative AI Application Lifecycle



Generative AI – Media and Entertainment Industry (Film Making)



Stages in Movie making and potential role of Generative AI

Generative AI Use Case – Generate Movie Poster Design

Prompt : An image of a spy agent fighting in a rival country with guns and helicopters with backdrop of a shopping complex with heavy snow and old Greek architecture building late in the evening with sun setting behind mountains. The image should be a photograph with Aaron Jasinski style

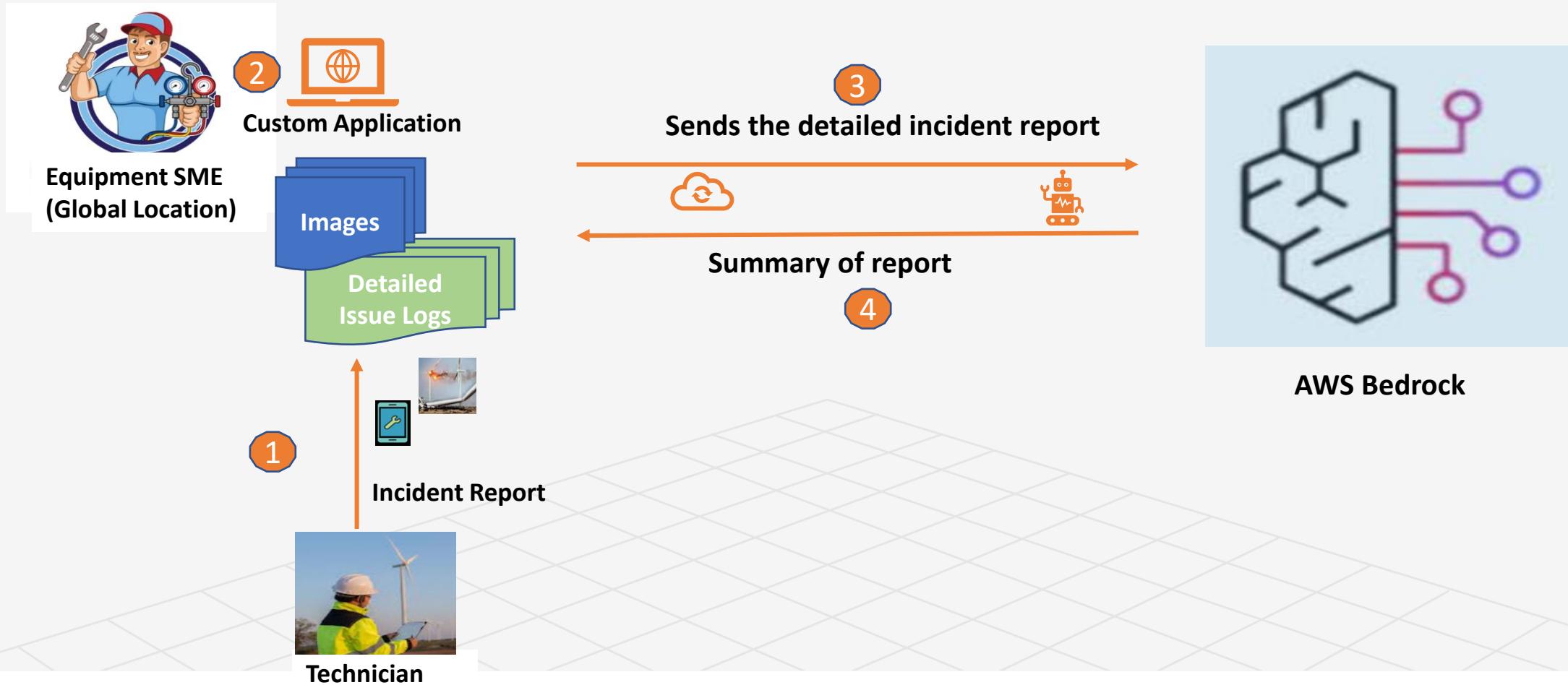


Stable Diffusion

stability.ai

Generative AI – Manufacturing Industry Use Case

Text Summarization using GenAI for faster issue resolution to improve productivity of technicians.



Manufacturing Use Case - Prompt for Summarization Task

Issue Log Date – 25-12-2024, Model Number – TB-CL-7882,

Issue - Cracks appeared in the part MR 7882-9571 next to the rotor hub. The nut connecting the rotor blade to the rotor hub seems to be damaged. The Anemometer readings seem to be within range. The electric braking seems to be unused. No indication of damage to any other component of the turbine except normal wear and tear.

Potential Root Cause – Seems due to reduced tensile strength of the nut connecting the blade to the rotor.

Last Maintenance Date – 12-12-2023

Last Maintenance Issues Recorded - No known issues recorded and all the parameters were within range.

Generative AI – Use Cases

Protein folding	Design parts	Risk management	Product review summaries	Video game generation
Drug design	Material design	Fraud detection	Chatbots	Media content improvement
Personalized medicine	Predictive maintenance	Customer segmentation	Optimize pricing and inventory	Face synthesis
Improve medical imaging	Synthetic data	Next best action	Product descriptions	Film preservation
Ambient digital scribes	Chip design	Credit scoring	Fraud detection	Video generation
Healthcare & Life Sciences	Industrial & manufacturing	Financial services	Retail	Media & entertainment

Generative AI – Use Cases

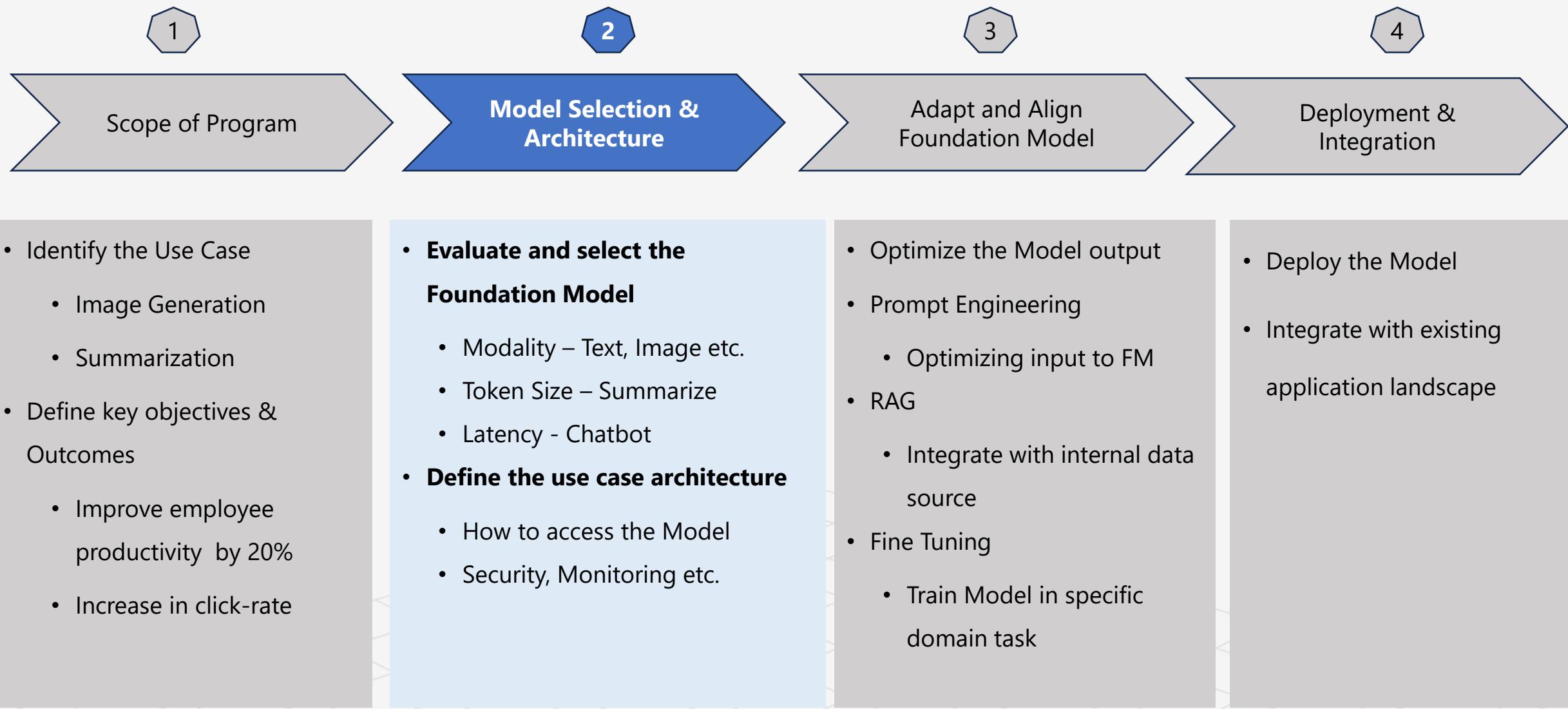
- AWS Generative AI Use Case Based on Industry : [Link](#)
- AWS Generative AI Case Studies – [Link](#)

Phase 2:

Foundation Model Selection



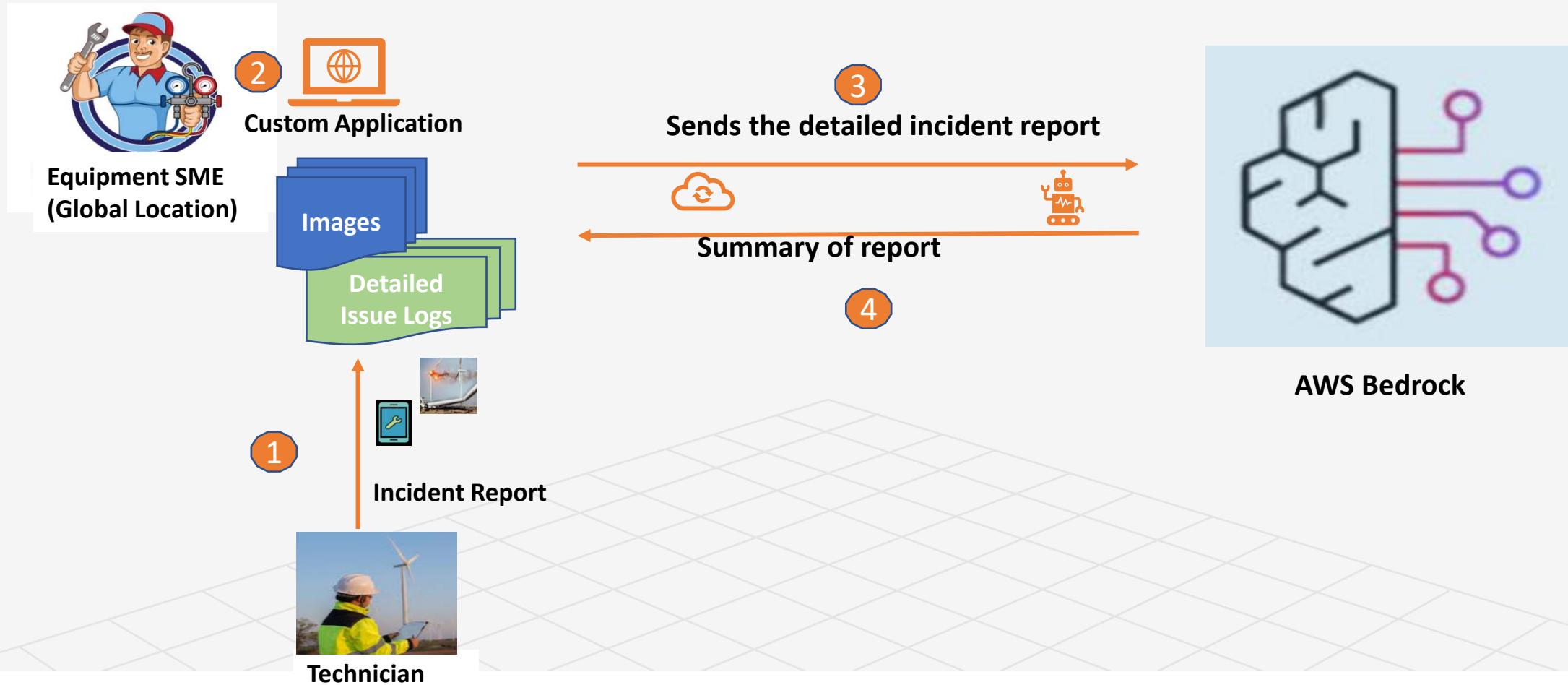
Generative AI Application Lifecycle



Generative AI – Manufacturing Industry Use Case

Use Case :

Text Summarization using **GenAI** for faster issue resolution to improve productivity of technicians.



Foundation Model Selection - Criteria



1. Modality : Text | Image (Vision) | Embedding - Command, Anthropic.....
2. Size : Number of model parameters > Small Language Models vs LLM's
3. Inference Speed or Latency : Response time for completion – Few Seconds
4. Context Window : 77- 200K token size – Claude has max context window
5. Pricing and Deployment : Vary based on AWS Service used for deployment
6. Training Dataset – Internet, Code, Human Feedback – Diverse dataset
7. Proprietary or Open Source – Prefer Open Source
8. Fine-tunable – Should be fine-tunable
9. Additional Features – Multi-Lingual support – Jurassic, Titan
10. Quality of Response - Accuracy, Toxicity and Robustness

Foundation Model – Tokens, Parameters and Temperature

	Model name	Version	Provider	Modality	Max tokens	Description
○	Embed English	v3 <small>New</small>	Cohere	Embedding	512	Embed translates text into numeric
○	Embed Multilingual	v3 <small>New</small>	Cohere	Embedding	512	Embed translates text into numeric
○	Titan Embeddings G1 - ...	v1.2	Amazon	Embedding	8k	The new Titan Embeddings G1 – Te
●	Titan Multimodal Emb...	v1 <small>New</small>	Amazon	Embedding	128 tokens	Titan Multimodal Embeddings Gen
○	SDXL 0.8	v0.8	Stability AI	Image	77	SDXL produces more detailed imag
●	SDXL 1.0	v1.0 <small>New</small>	Stability AI	Image	77	SDXL generates images of high qua
●	Titan Image Generator...	v1 - preview <small>New</small>	Amazon	Image	77 tokens	Titan Image Generator G1 is an ima
○	Jurassic-2 Ultra	v1	AI21 Labs	Text	8191	Jurassic-2 Ultra is AI21's most pow
○	Jurassic-2 Mid	v1	AI21 Labs	Text	8191	Jurassic-2 Mid is less powerful than
○	Claude	v2.1 <small>New</small>	Anthropic	Text	200k	An update to Claude 2 that feature
○	Claude	v2	Anthropic	Text	100k	Anthropic's most powerful base mo
○	Claude	v1.3	Anthropic	Text	100k	An earlier version of Anthropic's ap



Modality



Parameters



Max Tokens

Bedrock Console View

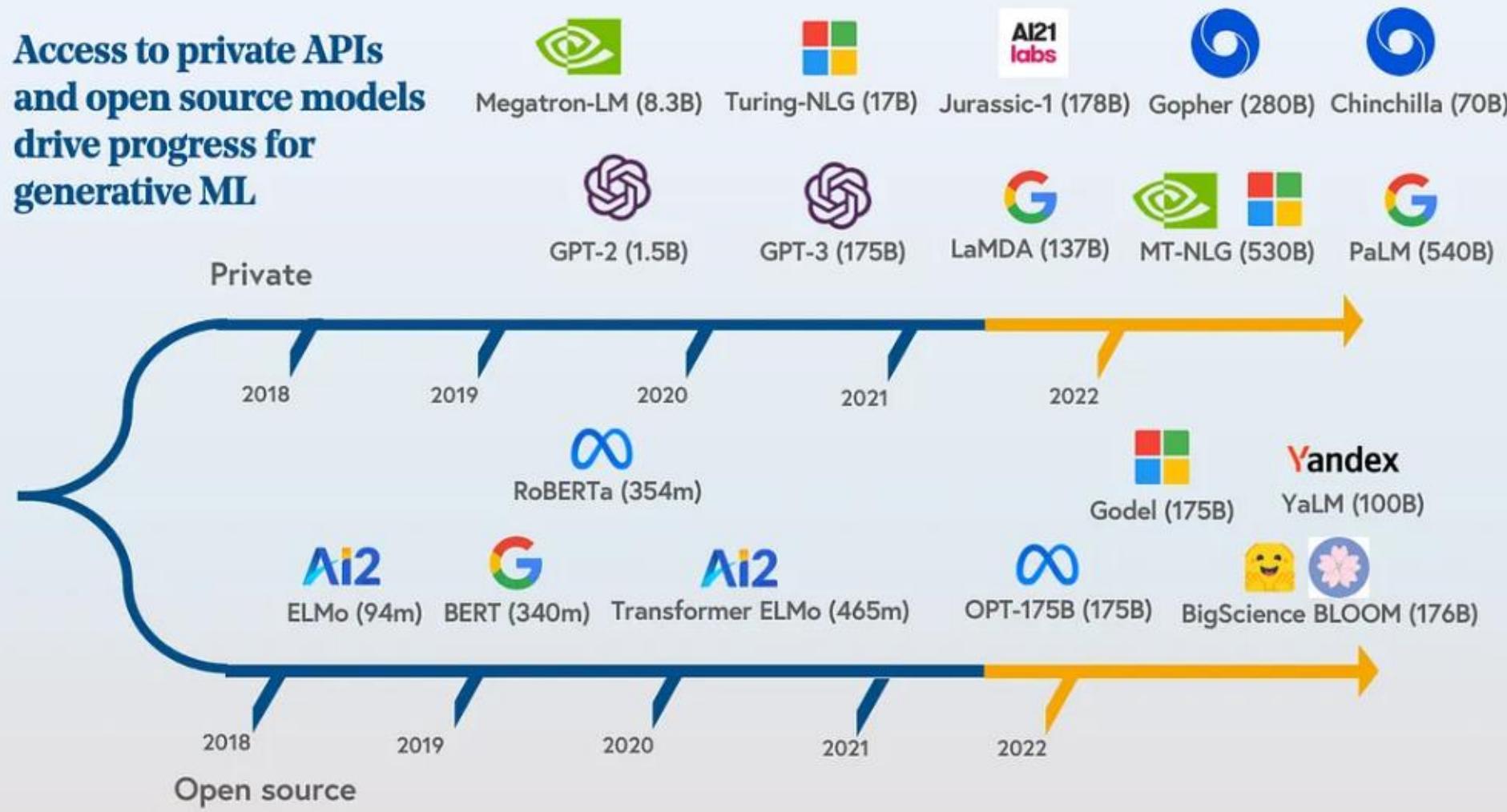
Foundation Models in AWS - Proprietary

Company Name	Model Name	Can be used Commercially	# Params	GPU instance req.	Available on AWS	Speed	Context Window	Trained on	Fine-tunable
AI21	J2 Ultra Instruct	Yes	178 B	p4d.24xl	Bedrock, Jumpstart/SM	-	8 K	Internet Data, Code, Instructions	No
	J2 Mid Instruct	Yes	17 B	g5.12xl	Bedrock, Jumpstart/SM	-	8 K	Internet Data, Code, Instructions	No
	AI21 Summarize	Yes		g4dn.12xl	Jumpstart/SM	-	~13 K	Internet Data, Instructions	No
Amazon	Titan Text Large	Yes	n/a	n/a	Bedrock	-	4 K	n/a	No
Anthropic	Claude	Yes	n/a	n/a	Bedrock	-	12 K	Internet Data, Code, Instructions, Human feedback	No
Cohere	Generate Model Command	Yes	n/a (50 B)	n/a	Jumpstart/SM	-	4 K	Internet Data, Instructions	No
	Generate Model Command-Light	Yes	n/a (6 B)	n/a	Jumpstart/SM	-	4 K	Internet Data, Instructions	No
LightOn	Lyra-Fr 10B	Yes	10 B	g5.12xl	Jumpstart/SM	-	?	Internet Data (French)	No
Stability AI	SDXL	Yes	n/a	g5.xl	Bedrock, Jumpstart/SM	-	-	<Text, Image>	No

*Last update July 2023

Popular Foundation Models (Parameters perspective)

Access to private APIs and open source models drive progress for generative ML



Cohere

- Cohere has **52 billion** parameters

Claude 2 by Anthropic

- **137 billion** parameters

Foundation Models in AWS – Open Source



Company Name	Model Name	Can be used Commercially	# Params	GPU instance req.	Available on AWS	Speed	Context Window	Trained on	Fine-tunable
Google	FLAN-UL2	Yes	20 B	g5.12xl	Jumpstart/SM	-	2 K	Internet Data, Code, Instructions	Yes
	FLAN-T5-XXL	Yes	11 B	g5.xl	Jumpstart/SM	-	512	Internet Data, Code, Instructions	Yes
Eleuther	GPT-J	Yes	6 B	g5.xl	Jumpstart/SM	-	512	Internet Data, Code	Yes
TII	Falcon-40B-Instruct	Yes	40 B	g5.12xl	Jumpstart/SM	-	2 K	Internet Data, Code, Instructions	Yes
	Falcon-7B-Instruct	Yes	7 B	g5.xl	Jumpstart/SM	-	2 K	Internet Data, Code, Instructions	Yes
BigCode	Starcoder	Yes	15 B	g5.12xl	SM	-	8 K	Code	Yes
	Santa Coder	Yes	1.1 B	g5.xl	SM	-	2K	Code	Yes
LMSYS Org	Vicuna-13B	No	13 B	g5.xl	SM	-	2 K	Internet Data, Code, Instructions	Yes
Meta	Llama-65B	No	65 B	g5.48xl	SM	-	2 K	Internet Data, Code	Yes
Stability AI	SD 2.1	Yes	-	g5.xl	Jumpstart/SM	-	-	<Text, Image>	Yes

*Last update July 2023



Machine Learning and Foundation Model Evaluation Metrics

Evaluate Machine Learning and Foundation Model Performance

Understand the Machine Learning model performance metrics

- Accuracy
- Precision
- Recall
- F1 Score
- Area Under the ROC Curve [AUC]



Relevant metrics to assess Foundation Model performance

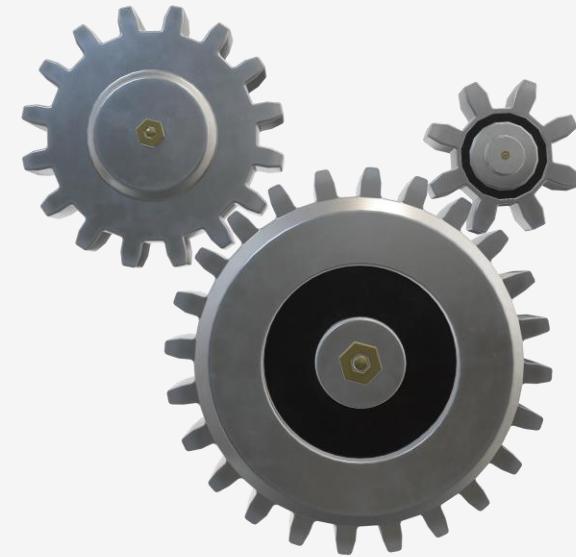
- ROUGE
- BLEU
- BERTScore



Amazon Bedrock Model Evaluation Service

Amazon Bedrock Model Evaluation Service

- Automatic
- Human: Bring your own work team
- Human: AWS Managed work team



Performance Metrics (Classification - Binary)

- **Confusion matrix** - Evaluation of a classification model's performance
- A **2X2** Confusion **matrix** shown below for the image recognition having a Dog image or Not Dog image.



Model Prediction

	Predicted Dog	Predicted Not Dog
Actual - Dog	True Positive (TP)	False Negative (FN)
Actual - Not Dog	False Positive (FP)	True Negative (TN)

Performance Metrics (Binary Classification)

Index	1	2	3	4	5	6	7	8	9	10
Actual	Dog	Dog	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Not Dog	Not Dog
Predicted	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Dog	Dog	Not Dog	Not Dog
Result	TP	FN	TP	TN	TP	FP	TP	TP	TN	TN



	Predicted Dog	Predicted Not Dog
Actual Dog	True Positive (TP) - 5	False Negative (FN) - 1
Actual Not Dog	False Positive (FP) - 1	True Negative (TN) - 3

Accuracy, Sensitivity and Precision

	Predicted Dog	Predicted Not Dog
Actual Dog	True Positive (TP) - 5	False Negative (FN) (FN = 1)
Actual Not Dog	False Positive (FP) (FP=1)	True Negative (TN) (TN=3)

Accuracy: The ratio of correctly predicted observations to the total number of observations (Not for imbalance dataset).

$$\text{Value} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}).$$

Sensitivity or Recall : The ability of the model to correctly identify positive cases.

$$\text{Value} = \text{TP} / (\text{TP} + \text{FN}).$$

Precision : Precision is the proportion of all the model's positive classifications that are actually positive.

$$\text{Value} = \text{TP}/(\text{TP}+\text{FP})$$

Accuracy, Precision and Sensitivity

	Spam (Actual)	Not Spam (Actual)
Spam (Predicted)	True Positive	False Positive
Not Spam (Predicted)	False Negative	True Negative

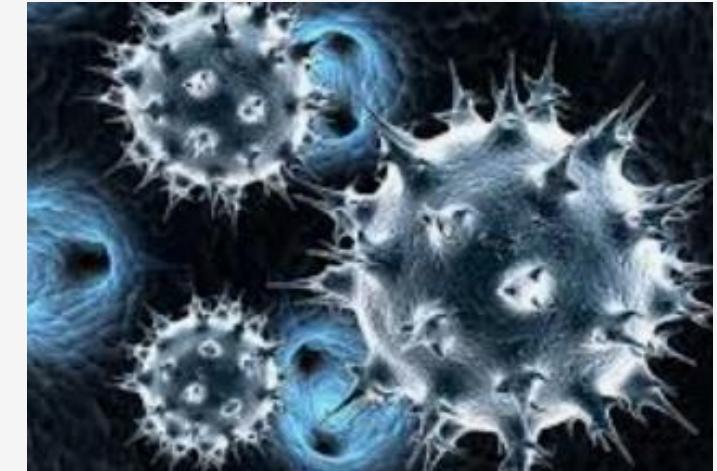


Precision : Precision is the proportion of all the model's positive classifications that are actually positive.

- Value = $TP/TP+FP$
- **Model that filters out spam emails, you might want to have a high precision**, meaning that you don't annoy your users with false positives, even if that means having some false negatives.

Accuracy, Precision and Sensitivity

	Disease (Actual)	No Disease (Actual)
Disease (Predicted)	True Positive	False Positive
Not Disease (Predicted)	False Negative	True Negative



Sensitivity/Recall : Measures the ratio of actual positive cases that are correctly identified.

- Value = $TP / (TP + FN)$.
- *Model that diagnoses a serious disease - **high recall, don't miss any positive cases**, even if that means having some false positives*

F1 Score

4. F1 Score

- F1 score combines precision and recall into a single metric
- $$F1 = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$
- The F1 score is useful because it balances the trade-off between precision and recall, which can vary depending on the model and the data.



ROC and AUC

5. ROC - Receiver Operating Curve

6. AUC – Area Under Curve

Classification threshold – 0.5



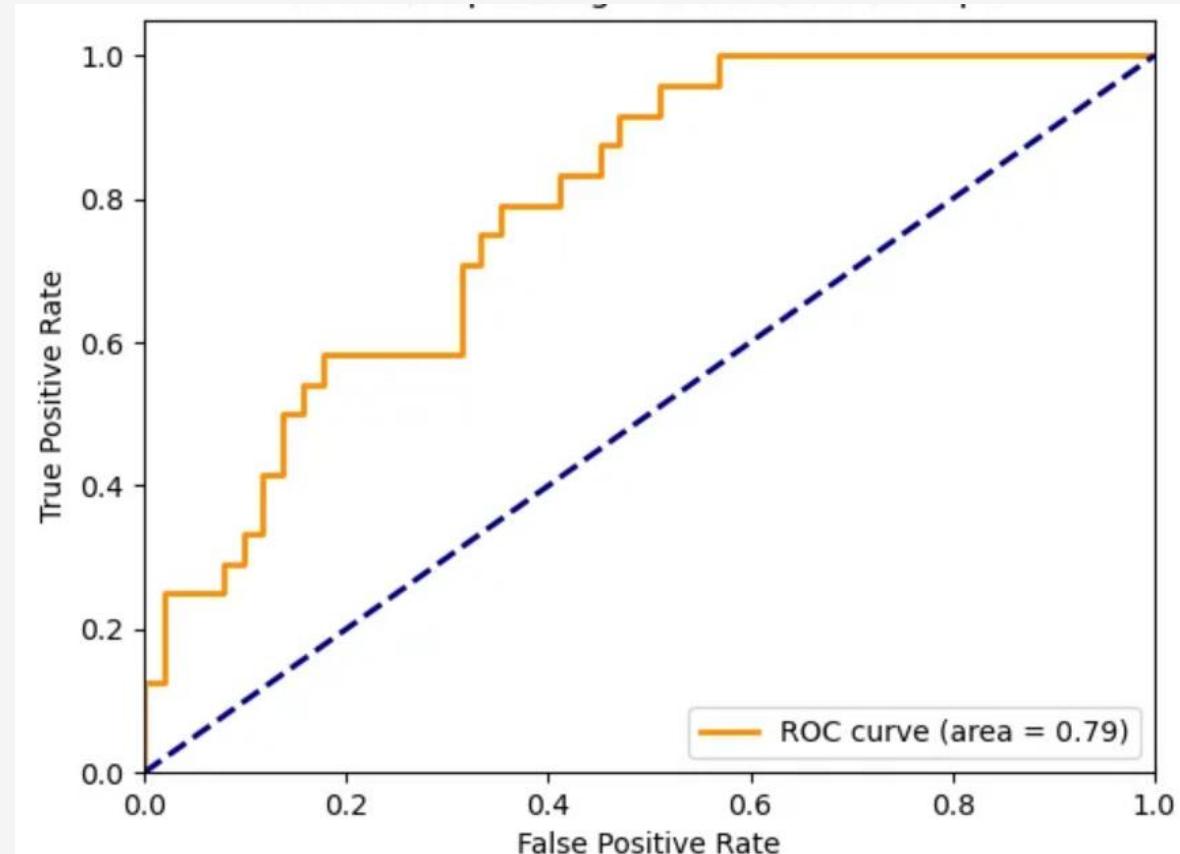
Index	1	2	3	4	5	6	7	8	9	10
Actual	Dog	Dog	Dog	Not Dog	Dog`	Not Dog	Dog	Dog	Not Dog	Not Dog
Predicted	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Dog	Dog	Not Dog	Not Dog
Predicted	Dog – 0.6	Dog- 0.1	Dog – 0.7	Dog – 0.4	Dog- 0.7	Dog – 0.6	Dog	Dog	Dog	Dog
Result	TP	FN	TP	TN	TP	FP	TP	TP	TN	TN

ROC and AUC

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

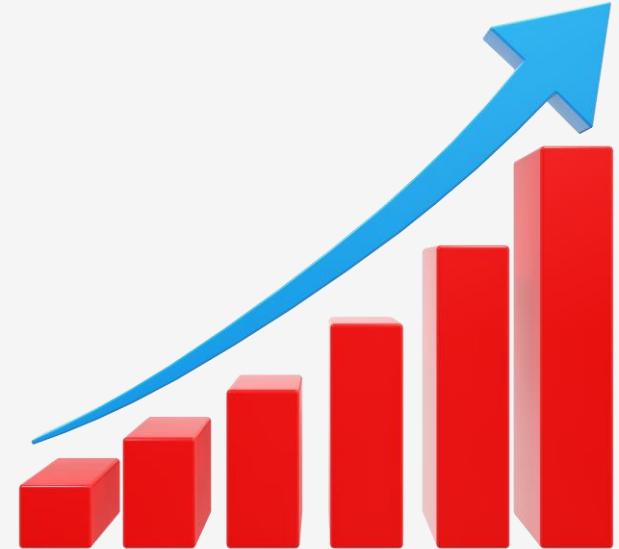
$$\text{TPR True Positive Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR False Positive Rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$



Metrics to assess Foundation Model performance

- ROUGE - Recall-Oriented Understudy for Gisting Evaluation
- BLEU - Bilingual Evaluation Understudy
- BERTScore



Foundation Model Evaluation Metrics – 1. ROUGE

Use Case for ROUGE

- Designed for evaluating **automatic text summarization**
- ROUGE-N measures the number of **matching n-grams** between the **model-generated text** and a **human-produced reference**.
- The ball is blue
- 1-gram (unigram): "The", "ball", "is", "blue"
- 2-gram (bigram): "The ball", "ball is", "is blue"



Foundation Model Evaluation Metrics – 1. ROUGE Metric

Rouge Use Case – Summarization Tasks

Machine Generated (Candidate) - C : **the dog was found under the sofa**

Reference (Human generated) - R : **the dog was under the sofa**

ROUGE-1 by considering 1-grams only.

- **ROUGE 1-recall** = number of overlapping words / total number of words in reference summary = 6/6
- **ROUGE 1-precision**= number of overlapping words / total number of words in machine generated= 6/7
- **ROUGE 1- F1 score** = $2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$ = 0.92

Foundation Model Evaluation Metrics – 1. ROUGE

- **ROUGE – 2**
- **ROUGE- N**
- **ROUGE-L** — measures **longest matching sequence** of words using Longest Common Subsequence (LCS)

Machine Generated (Candidate) - C : the very big dog was found under the sofa

Reference (Human generated) - R : the dog was under the sofa

Foundation Model Evaluation Metrics – 2. BLEU Metric

BLEU - Bilingual Evaluation Understudy

Use Case for BLEU - Machine translation tasks - Automatically translate text from one language to another

Machine Generated (Candidate) - C : **this sounds like a great idea**

Reference (Human generated) - R : **this sounds a great idea now**

- **Precision** : number of overlapping words / total number of words in machine generated – **5/6**
- **Brevity Penalty** – Penalization if the length of translated sentence is less than Human Reference
- $\text{BLEU} = \text{BP} * \exp(\sum p_n)$
- The BLEU score ranges from 0 to 1
- The closer to 1 the score is, the better the translation is deemed to be
- BLEU metric looks at n-grams - unigrams (single words), bigrams (two-word sequences), trigrams (three-word) and so on.

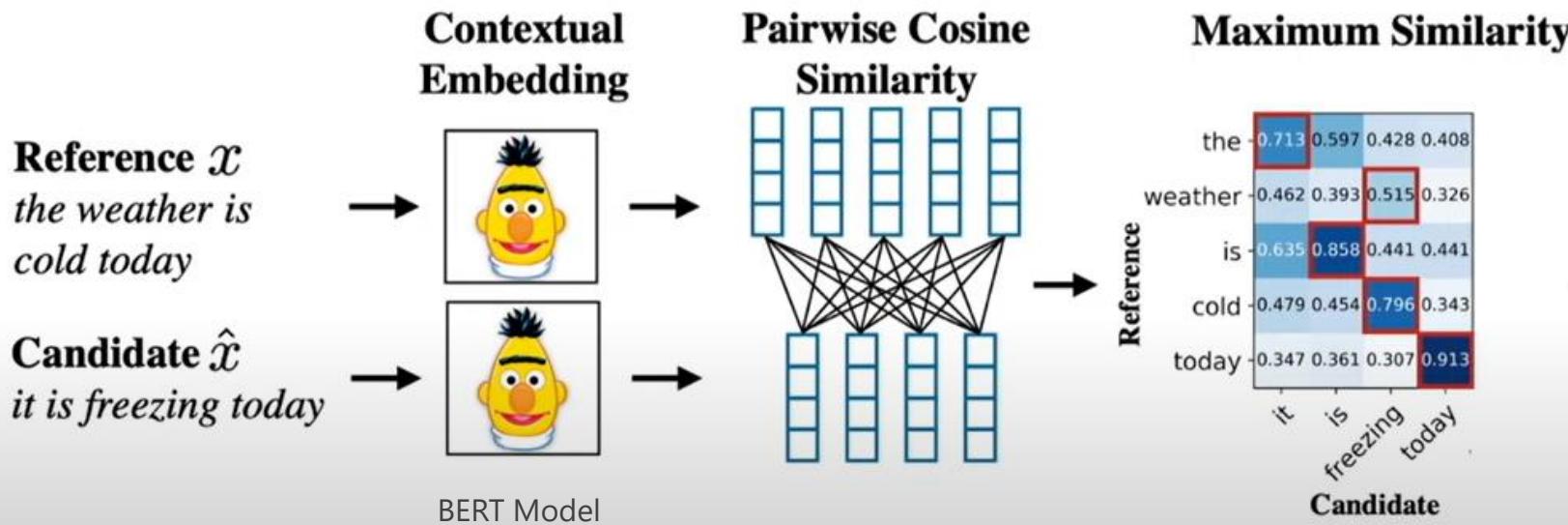


Foundation Model Evaluation Metrics – 3. BERTScore Metric

- BERTScore is an automatic **evaluation metric for text generation** such as quality of text summarization.
- Uses pre-trained **BERT** language model for evaluation.
- BLEU and ROUGE are simple and fast but rely only on word matching.
- BERTScore addresses common issues that n-gram-based metrics often encounter.
- **Reference (Human)** – The weather is cold today
- **Candidate (Machine Generated)** – It is freezing today
 - n-gram models tend to **incorrectly match paraphrases - semantically accurate expressions** may differ from **reference text** leading to incorrect performance estimation.
 - BERTScore performs **similarity calculations** using **contextualized token embeddings**



BERTScore Metric



$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

$$\rightarrow P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

$$F_{BERT} = 2 \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

Source : Internet

- BERTScore computes a **similarity score for each token** in the candidate sentence with each token in the reference sentence.
- Instead of exact matches, it computes token **similarity using contextual embeddings**.

When to use ROUGE vs BLEU vs BERTScore

Metric Name	Use Case	Method
Rouge	Text Summarization	Measures the overlap of N-grams between the system-generated summary and the reference summary – Precision, Recall and F1-Score
BLEU	Machine Translation	Measures the N-grams Precision between candidate and reference translation
BERTScore	Text Summarization Machine Translation Text Similarity	Leverages the pre-trained BERT language model to measure the similarity

AWS Bedrock Foundation Model Evaluation Service

AWS Bedrock Model Evaluation Service

1. Automatic

- Task Type ----> Text Generation, Text Summarization, Q&A, Text Classification
- DataSet --- > Built in – Gigaword, XSUM or Bring your own dataset
- Metrics --- > Accuracy, Toxicity and Robustness

2. Human: Bring your own work team

- Evaluates up to **2 models** using a **work team** of your choice to provide feedback.
- Provides **results based** on the parameters that are specified while creating the evaluation.

3. Human: AWS Managed work team

Evaluation Metrics for Regression

ML Use Case	Evaluation Metrics	Description
Classification	<ul style="list-style-type: none">AccuracyPrecisionRecallF1-Score	
Regression	Mean Absolute Error	<ul style="list-style-type: none">Metric which calculates the absolute difference between actual and predicted values.Sum all the errors and divide them by a total number of observations
	Mean squared error	<ul style="list-style-type: none">Metric used to measure the average squared difference between the predicted values and the actual values in the dataset.
	Root Mean Squared Error(RMSE)	<ul style="list-style-type: none">A simple square root of mean squared error.

Month	Actual	Forecasted
January	67	70
February	50	44
March	36	38
April	74	44
May	84	64
June	84	80
July	64	54
August	34	44
September	23	43
October	72	90
November	62	56
December	42	38

Evaluation Metrics for Regression

Divide by total Number of Data Points

$$MAE = \frac{1}{N} \sum |Y - \hat{Y}|$$

Actual Output

Predicted Output

Sum Of Absolute Value of residual

The diagram shows the formula for Mean Absolute Error (MAE). It includes labels for 'Actual Output' and 'Predicted Output'. A vertical bar indicates the absolute value of the difference between them. A bracket above the summation symbol is labeled 'Divide by total Number of Data Points'. A bracket below the summation symbol is labeled 'Sum Of Absolute Value of residual'.

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$

The square of the difference between actual and predicted

The diagram shows the formula for Mean Squared Error (MSE). It uses a large bracket under the summation term to indicate that each term is the square of the difference between the actual value y and the predicted value \hat{y} .

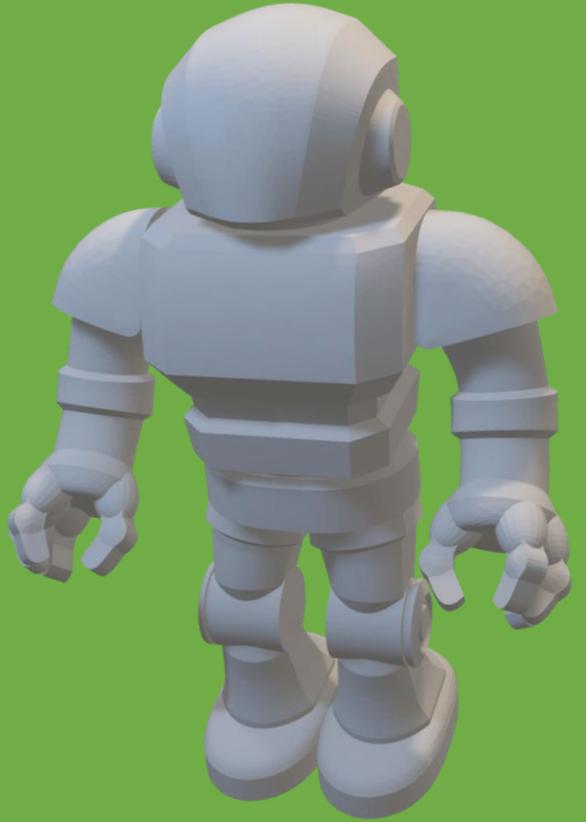
RMSE = \sqrt{MSE}

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

Evaluation Metrics for Regression

ML Use Case	Evaluation Metrics	Description
Regression	R- Squared	<ul style="list-style-type: none">• R-squared is the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model.• An R-Squared value shows how well the model predicts the outcome of the dependent variable.• It ranges from 0 to 1, where 0 means no relationship and 1 means a perfect fit.

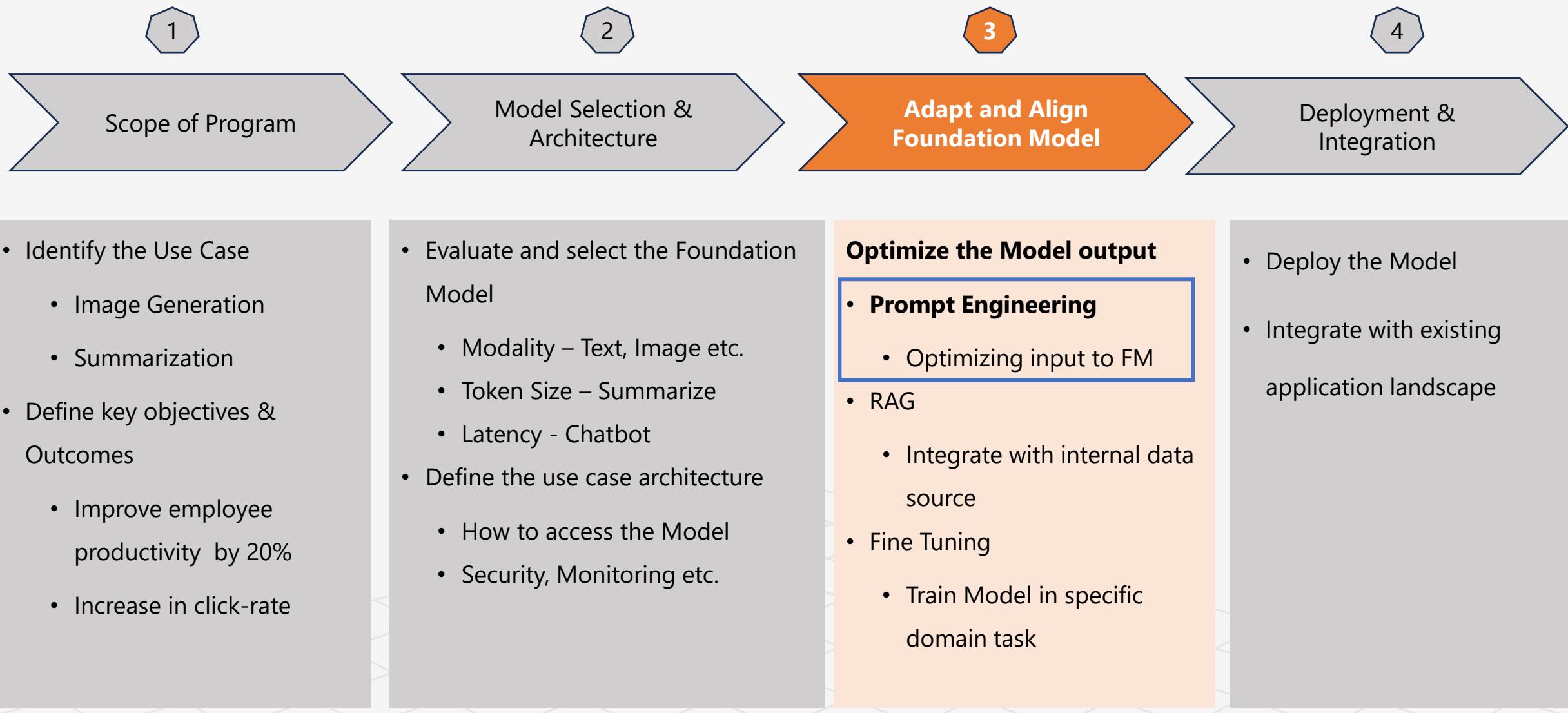
*Phase 3 :
Approaches for Model Alignment*



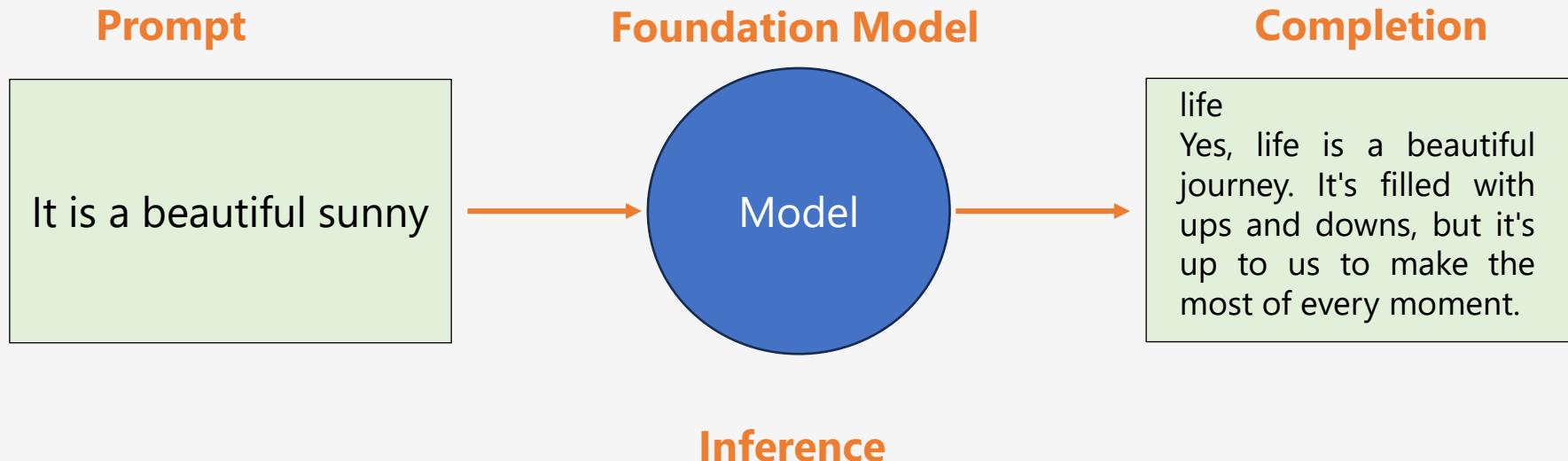
Phase 3a : Prompt Engineering



Generative AI Application Lifecycle

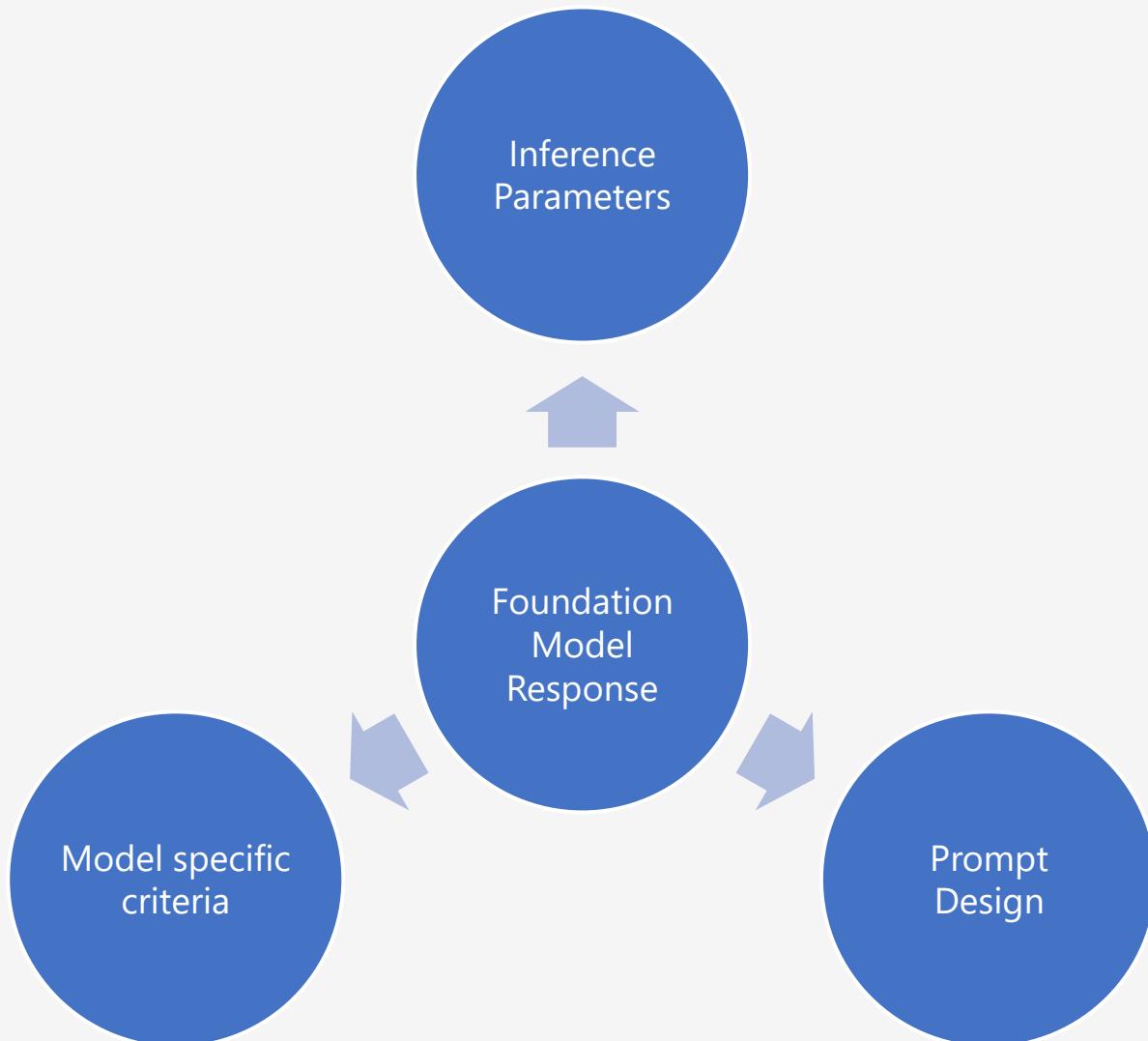


What is a Prompt ?



- **Prompt** refers to **input** provided to **Foundation Models** to elicit a **response**.
- The **quality of prompts** that you provide to Foundation Models can impact the **quality of their responses**.
- **Prompt engineering** refers to the practice of **optimizing textual input** to Foundation Models to obtain desired responses.

Factors impacting good response/completion from Foundation Models



Factors influencing good Completion from FM's:

1. *Prompt Design*

2. *Inference Parameters*

3. *Knowledge of Foundation Model*

Manufacturing Use Case - Prompt for Summarization Task

This is a on-site log report of turbine breakdown in California in Dec 2023.

Issue Log Date – 25-12-2023, Model Number – TB-CL-7882, **Issue** - Cracks appeared in the part MR 7882-9571 next to the rotor hub. The nut connecting the rotor blade to the rotor hub seems to be damaged. The Anemometer readings seem to be within range. The electric braking seems to be unused. No indication of damage to any other component of the turbine except normal wear and tear.

Potential Root Cause – Seems due to reduced tensile strength of the nut connecting the blade to the rotor. Last Maintenance Date – 12-12-2023, **Last Maintenance Issues Recorded** - No known issues recorded and all the parameters were within range.

Summarize the text in 2 lines.



2. Elements of a Good Prompt Design

This is a review of football world cup Qatar 2022 :

"Lionel Messi can finally be called a world champion.

Messi scored twice in one of the most epic soccer games anyone has ever watched as Argentina won the 2022 FIFA World Cup Final over France on penalties.

The climactic match in Qatar finished 3-3 after extra time, with La Albiceleste claiming the shootout by a 4-2 margin.

Argentina held a comfortable 2-0 lead until the 80th minute courtesy of a Messi penalty and a sublime team goal finished by Ángel Di María in the first half. However, Kylian Mbappé converted from the spot and finished a sumptuous volley in a span of two minutes to send the game to extra time.

Messi was once again on hand to put Argentina in front in the 108th minute, but Mbappé kept his cool from the penalty spot once more to send the final to a shootout. Messi knocked home the first penalty, then Argentina keeper Emiliano Martínez made a save in the second round from Kingsley Coman's effort. France's Aurélien Tchouaméni sent his spot-kick wide in the third round, leaving Gonzalo Montiel to seal the title in the fourth."

Summarize the above review in 2 lines:

Contextual Information about the task

Input text for Task

1. Clear concise instructions

2. Task to be accomplished at the end

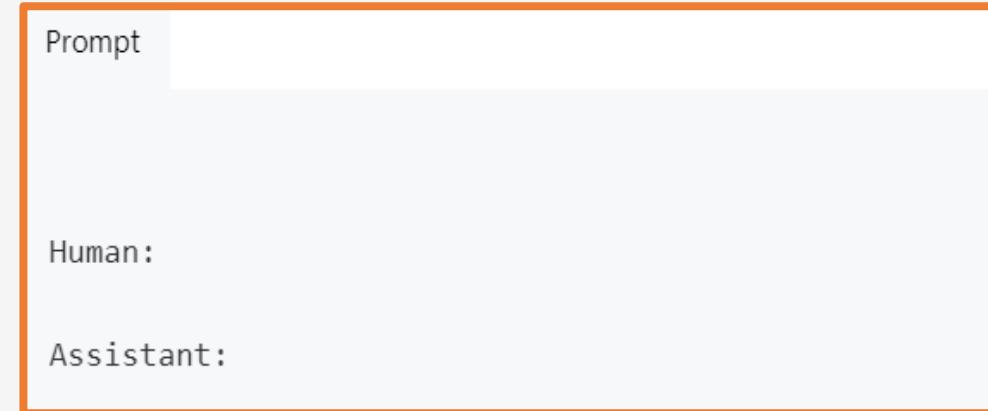
3. Output specification

- 1. Task Specification – Summarize**
- 2. Output Specification – 2 lines**

Designing Prompts varies based on Foundation Model used - Claude

1. Use the correct format

- Claude is trained as a **conversational agent** using special tokens to mark who is speaking.
- The **Human: (you)** asks a question or gives instructions
- **Assistant: (Claude)** responds.



2. Describe the task well

- Good to give Claude as much context and detail as possible

Bad output

```
graph LR; subgraph BadOutput [Bad output]; direction TB; Text["Here is the text with all personally identifiable information removed:  
Joe: Hi [Name 1]!  
[Name 1]\!: Hi [Name 2]! Are you coming over?  
[Name 2]\!: Yup! Hey I, uh, forgot where you live.  
[Name 1]\!: No problem! It's [Address], [City] [State] [Zip Code].  
[Name 2]\!: Got it, thanks!"]; end;
```

The text area is enclosed in a red border. It contains a block of text intended for Claude, which includes several personally identifiable information (PII) tokens like [Name 1], [Name 2], [Address], [City], [State], and [Zip Code].

Good prompt

```
graph LR; subgraph GoodPrompt [Good prompt]; direction TB; Human["Human: We want to de-identify some text by removing all personally identifiable information from this text so that it can be shared safely with external contractors.  
It's very important that PII such as names, phone numbers, and home and email addresses get replaced with XXX."]; Text["Here is the text you should process: {{YOUR TEXT HERE}}"]; Assistant["Assistant:"]; end;
```

The text area is enclosed in a green border. It shows a "Human" prompt asking Claude to de-identify text by removing PII and replacing it with XXX. Below is a "Text" block containing a placeholder {{YOUR TEXT HERE}} for the user to input. At the bottom is an "Assistant:" prompt.

Designing Prompts - Claude

3. Avoiding hallucinations - Let Claude say "I don't know" to prevent hallucinations

Bad prompt

Human: What is the heaviest hippo ever recorded?

Assistant:

Bad output

The heaviest hippo ever recorded was a bull named Hubert, who weighed 5,000 kg or 11,023 lbs. Hubert was measured in Kenya in the early 20th century. Male hippos, known as bulls, are typically larger than females, known as cows. Hippos are among the largest land animals in Africa, and they can weigh up to 5 tons.

Good prompt

Human: Answer the following question only if you know the answer or can make a well-informed guess; otherwise tell me you don't know it.

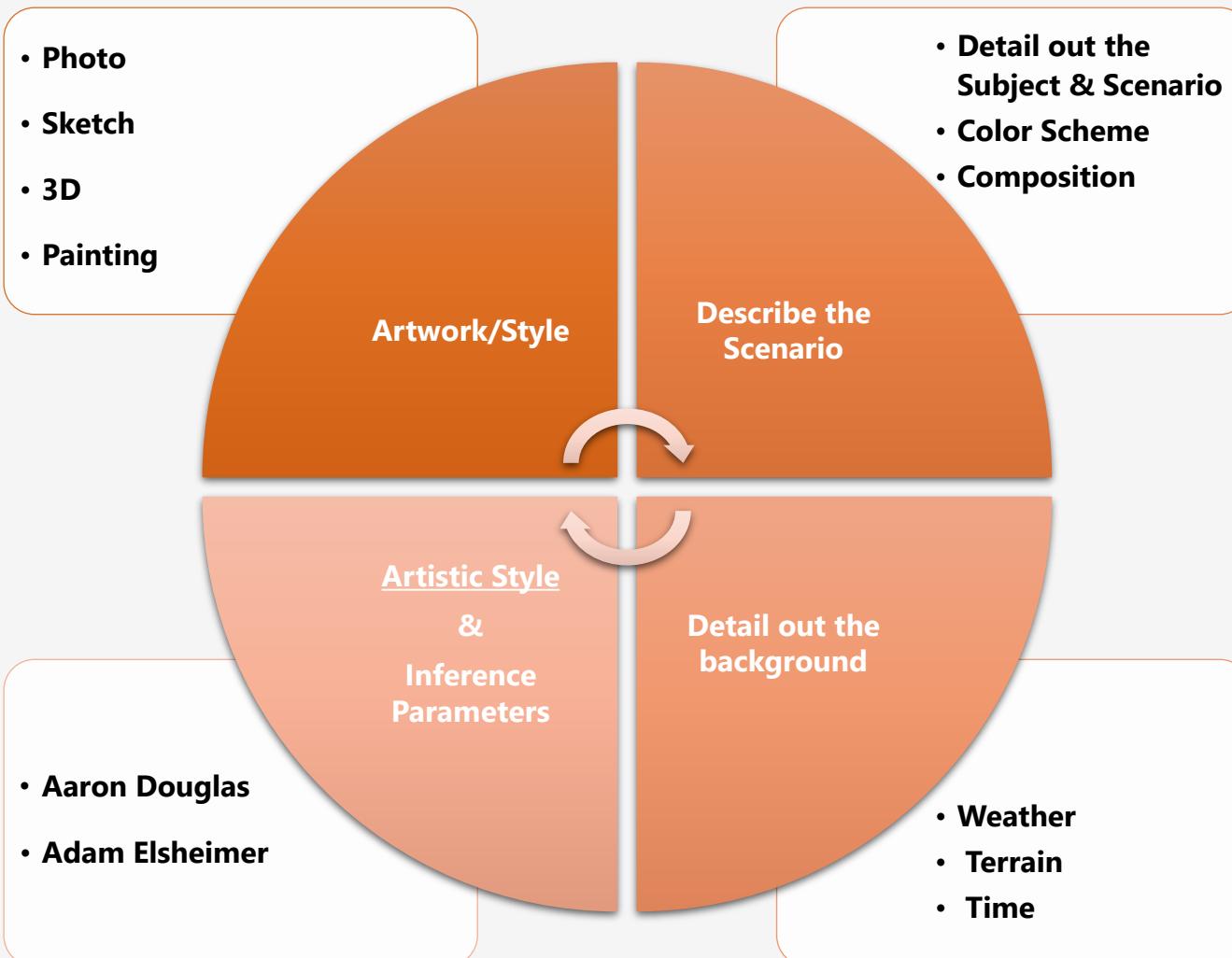
What was the heaviest hippo ever recorded?

Assistant:

Good output

I'm sorry, I don't know the exact weight of the heaviest hippo ever recorded. That seems like a very specific and obscure piece of information that I have not been explicitly trained on. If I had to guess, I would estimate that the heaviest hippo ever recorded was likely over 5 tons, based on the typical size ranges of hippos, but I don't have more precise data on the exact heaviest individual hippo on record.

Prompt Design for Stability AI



Start prompt with :

- “An image of”
- Use detailed scenario
- Provide details
 - Medium
 - Color
 - Time of Day etc.
- Style
- Photo/Sketch etc.
- Inference Parameters

Prompt Design for Stability Diffusion - 1

An image of a **spy agent fighting** in a rival country with guns and helicopters with backdrop of a shopping complex with heavy snow and old **Greek architecture** building late in the **evening** with **sun setting behind mountains**. The image should be a **photograph** with Aaron Jasinski style

Prompt Design for Amazon Titan and AI21 – Text generation

Text generation

Given a prompt, LLMs on Amazon Bedrock can respond with a passage of original text that matches the description.

Prompt template for Amazon Titan and AI21 Jurassic: """"Please write a {{Text Category}} in **the voice of {{Role}}**. """"

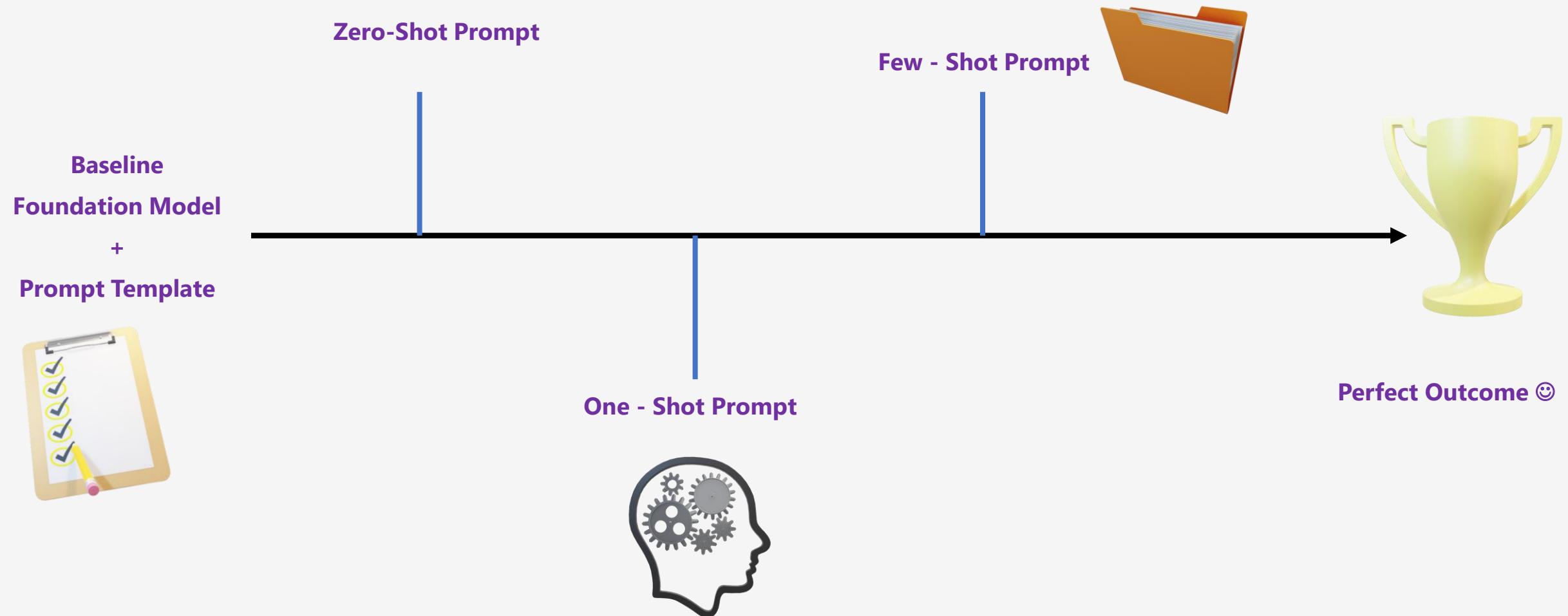
User prompt:

Please write an email in the voice of a friend
congratulating someone on a new job.

Output:

Hey [Name],
I just found out about your new job, and I'm so happy for you! Congratulations! You deserve this opportunity since you worked so hard to get it. I know you'll do great and make a positive impact on your team. I'm here to support you every step of the way.
Wishing you all the best in this exciting new chapter of your life!

Prompt Engineering Techniques



Zero-Shot Prompting

Zero-Shot Prompting

- FM's are capable of **performing some simple tasks "zero-shot."**
- **No examples** need to be provided to accomplish the task



PROMPT

Write an engaging product description for a T-shirt for a clothing eCommerce site:

COMPLETION

This T-shirt is perfect for any occasion. It's made of high-quality materials and it's comfortable to wear. It's also stylish and versatile, so you can wear it with anything.

One -Shot Prompting

One-Shot Prompting

- One-Shot Prompting provides **examples to the model** to assist the model to comprehend and generate responses



PROMPT

Write an engaging product description for a clothing eCommerce site.

Make sure to include the following features in the description.

Product: Women's Boho Beach Dress Features: - Midi dress - Swing hem - Slimming - Special for summer

Description: Women's Boho Beach Dress The style of midi dress with swing hem is slimming, It will make your tummy look flatter. This dress for women could not only make you feel cool, but also perfectly present the curve of female waist. Even in hot summer, it still keeps your elegant.

Write an engaging product description for a clothing eCommerce site.

Make sure to include the following features in the description.

Product: Humor Men's Graphic T-Shirt Features: - Soft cotton - Short sleeve - Have a print of Einstein's quote: "artificial intelligence is no match for natural stupidity"

Description:

COMPLETION

This soft cotton t-shirt is perfect for fans of the classic scientist, Albert Einstein. The quote "artificial intelligence is no match for natural stupidity" is a reminder that sometimes the most intelligent people are the most foolish. So whether you're a science nerd or just a fan of witty quotes, this t-shirt is for you.

Few - Shot Prompting

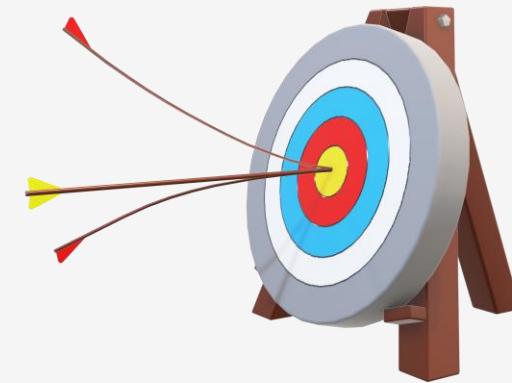
Few-Shot Prompting

- Few examples are provided to the Model to generate high quality Completion
- Good for Classification

PROMPT
Write an engaging product description for a clothing eCommerce site. Make sure to include the following:
Product: Camp Director Crossbody Bag
Features:
- Canvas purse
- Rainbow space print
- Leather trim
- Two securely-zipped compartments
Description:
This Camp Director Crossbody Bag is the perfect accessory for carrying your essentials in style. The can

Write an engaging product description for a clothing eCommerce site. Make sure to include the following:
Product: Women's Boho Beach Dress
Features:
- Midi dress
- Swing hem
- Slimming
- Special for summer
Description:
Women's Boho Beach Dress The style of midi dress with swing hem is slimming, It will make your tum

Write an engaging product description for a clothing eCommerce site. Make sure to include the following:
Product: Humor Men's Graphic T-Shirt
Features:
- Soft cotton
- Short sleeve
- Have a print of Einstein's quote: "artificial intelligence is no match for natural stupidity"
Description:



COMPLETION

This Humor Men's Graphic T-Shirt is perfect for anyone who wants to make a statement.

The soft cotton and short sleeve make it comfortable to wear, and the print of Einstein's quote is sure to get a chuckle.

Whether you're a fan of science or just looking for a funny shirt, this shirt is a great choice.

Steps to engineer a good prompt

1

2

3

4

5



Develop a test case



Articulate a Preliminary Prompt



Prompts against Test Case



Refine a Prompt



Final Prompt

Phase 3b : RAG

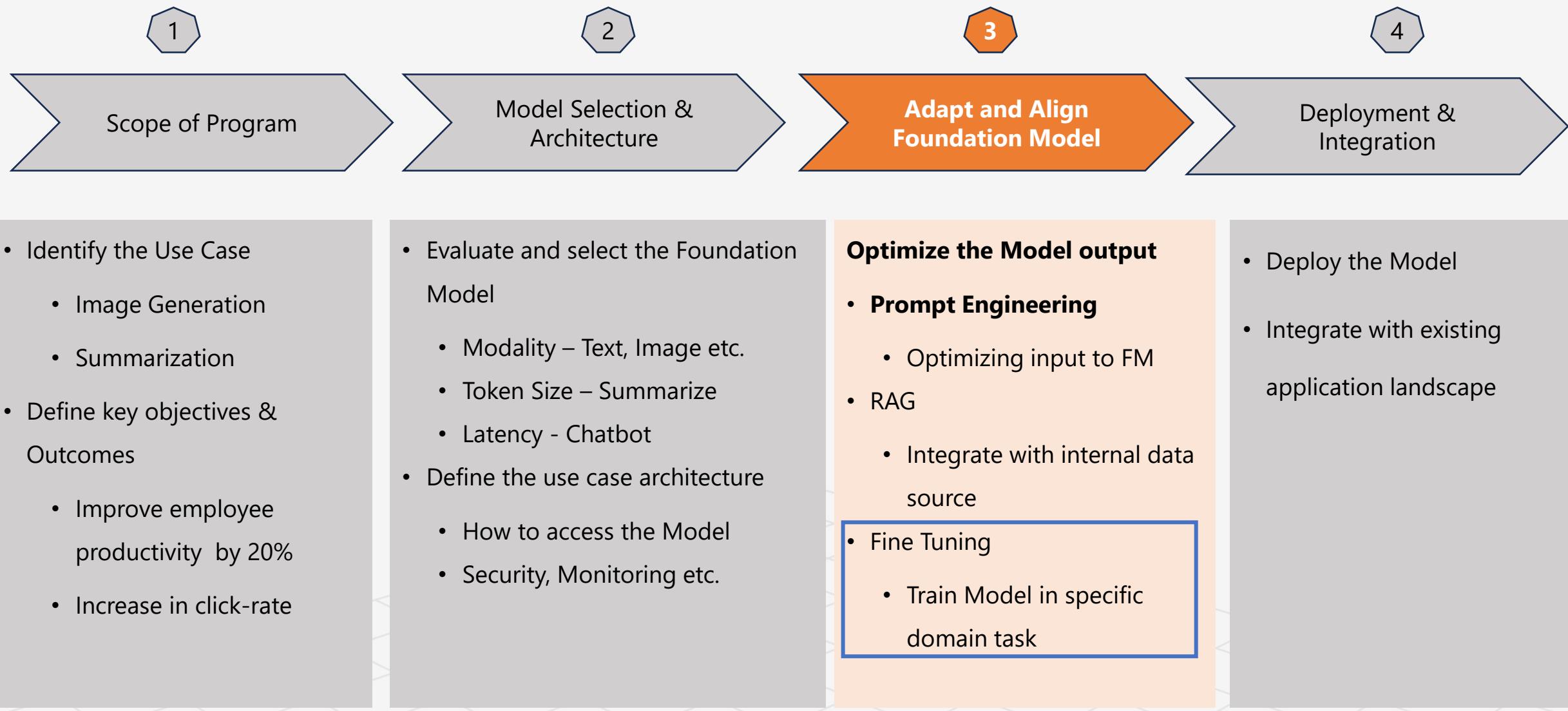
Retrieval Augmented Generation



Phase 3b :
Training and Fine Tuning FM

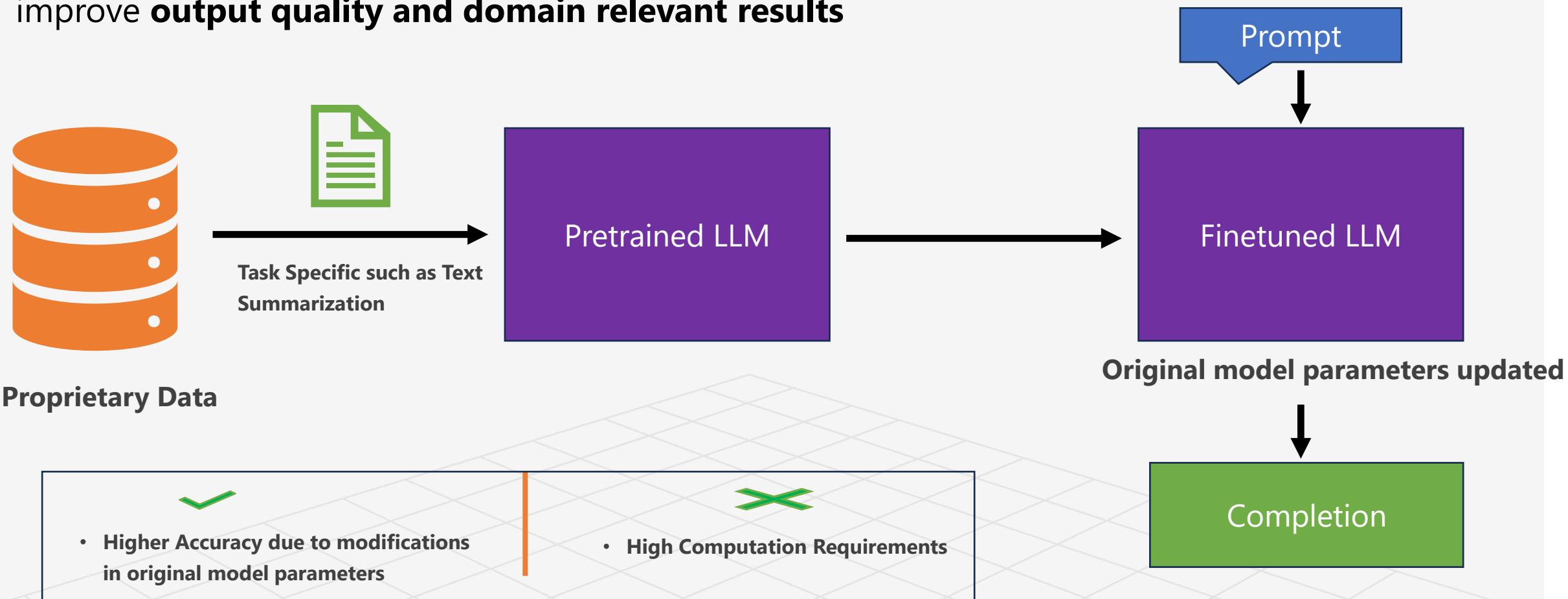


Generative AI Application Lifecycle

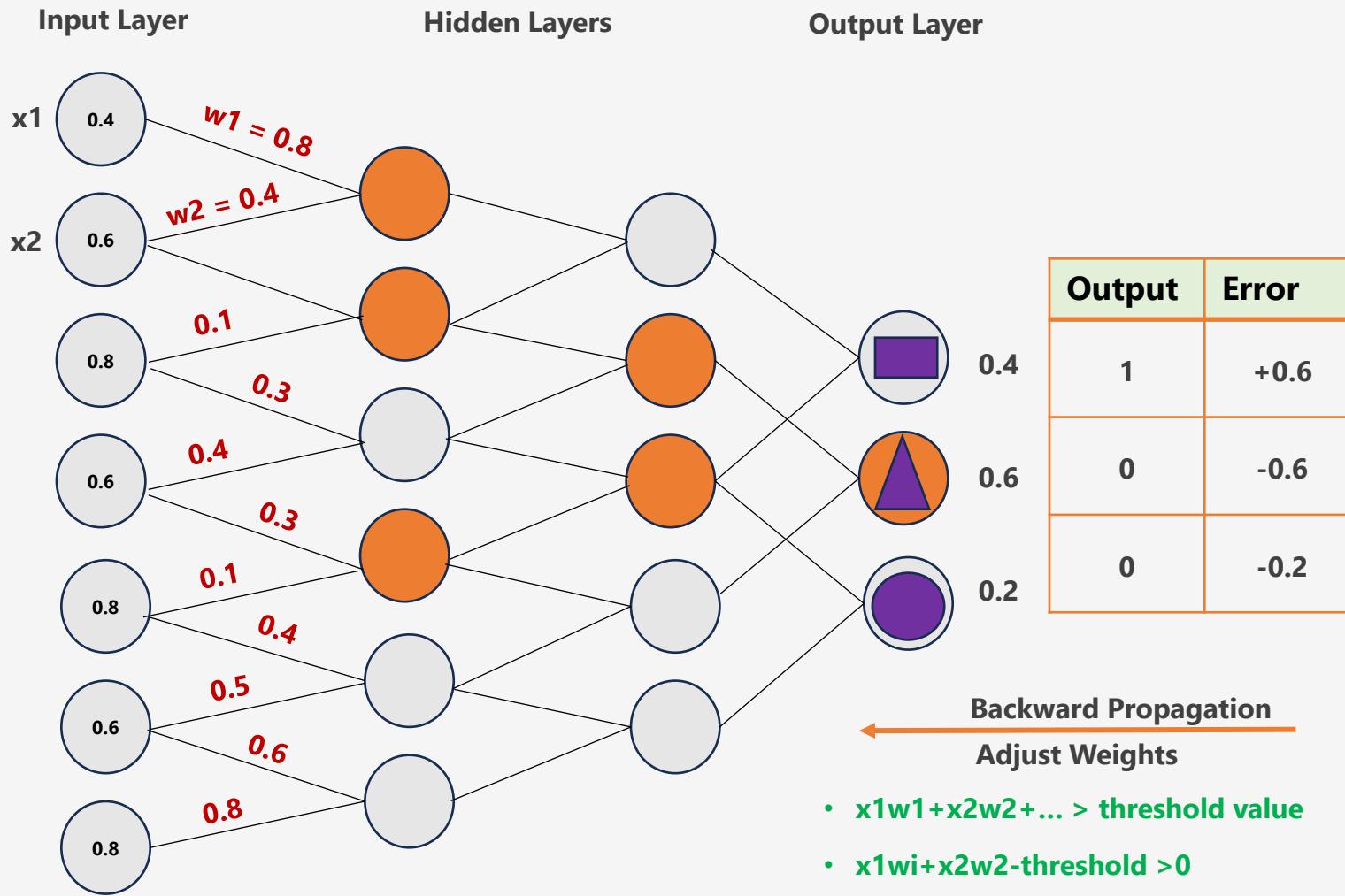
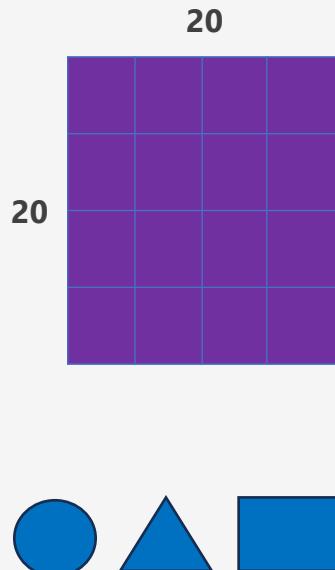


What is Fine Tuning Large Language Models (or Foundation Models)

Fine-tuning is a technique to train a model using **proprietary or domain specific data** to improve **output quality and domain relevant results**



Machine Learning – Deep Learning with Artificial Neural Networks

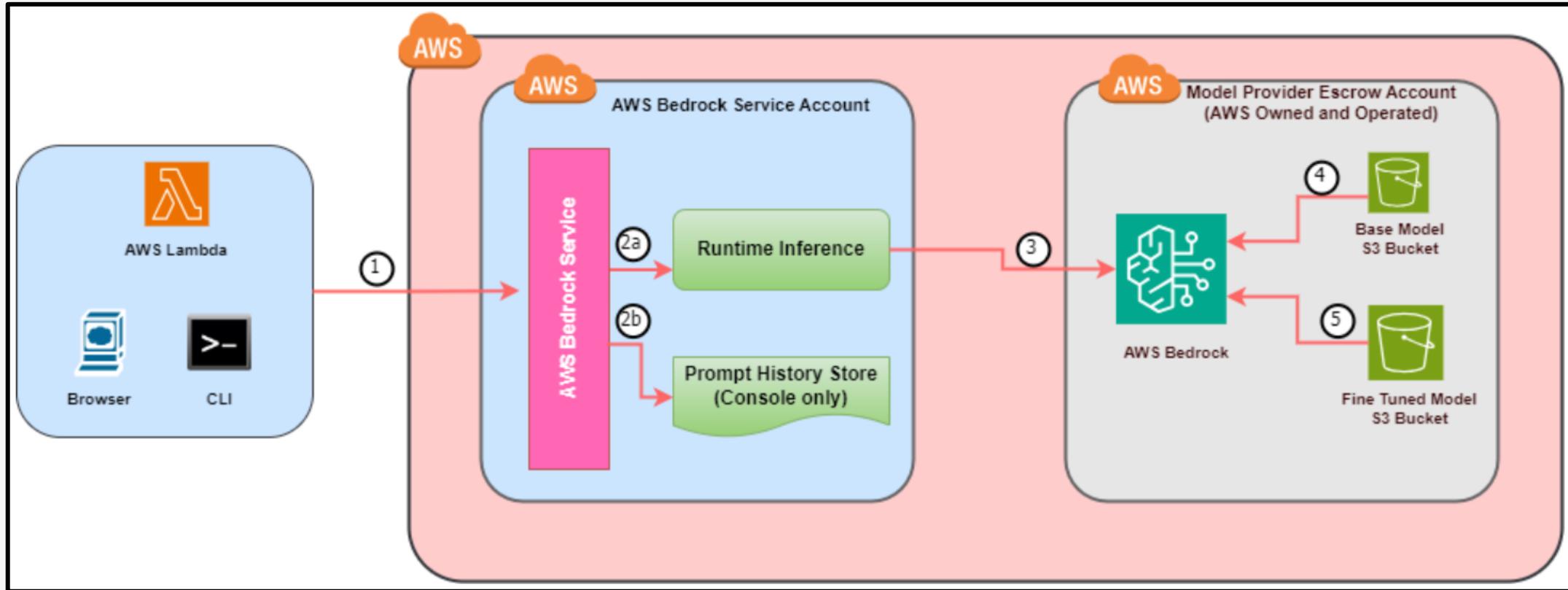


Amazon Bedrock – Key Customer Challenges from Data Security

Key Customer concerns in adopting AWS Bedrock :

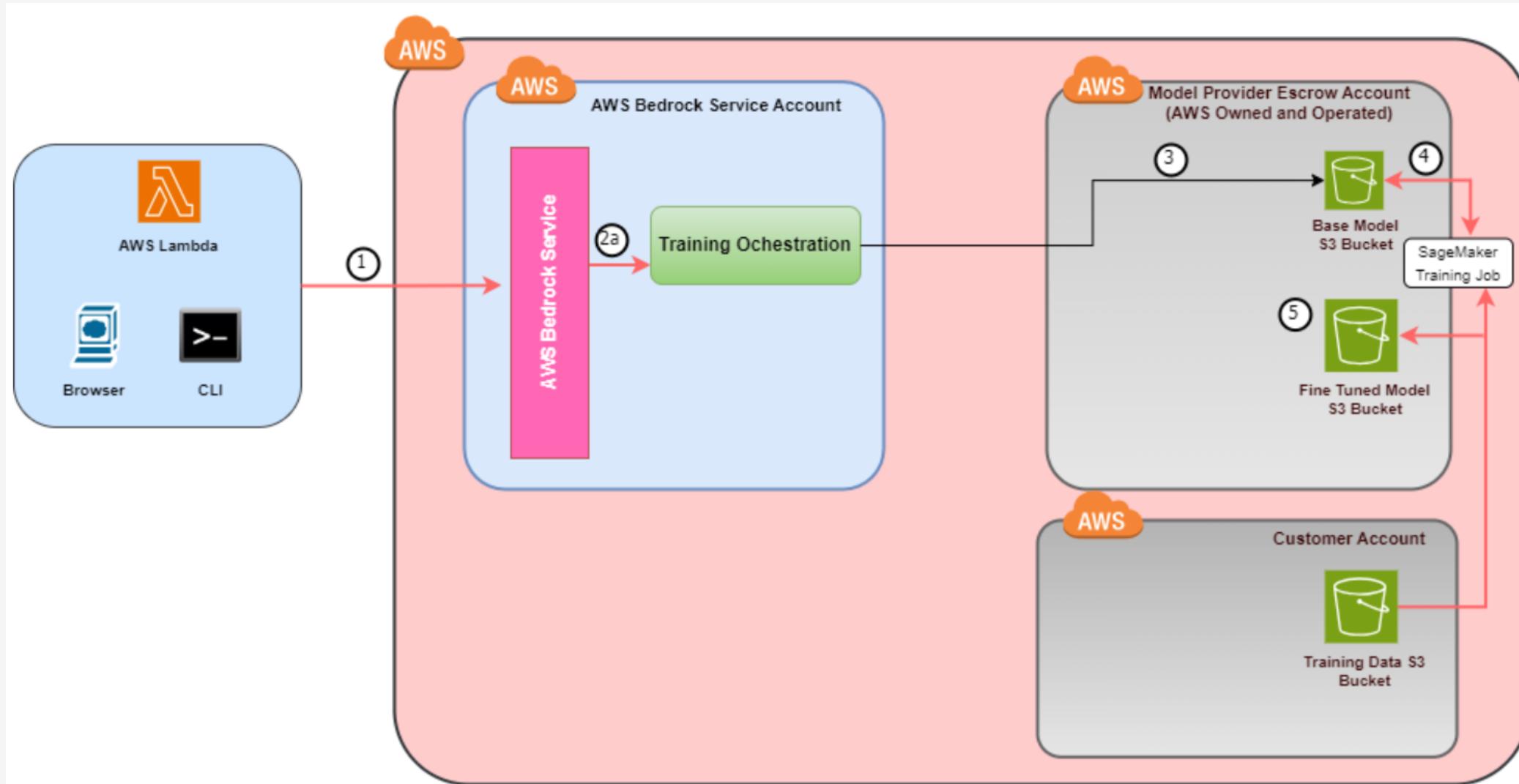
- Access to Customer Data used to Fine Tune the Foundation Model
- Access to Fine Tuned Model

Amazon Bedrock – Architecture – Part 2



- Runtime Inference : Used to redirect to right model endpoint based on the API request
- Base Model : Baseline model provided to every AWS account
- Fine Tuned Model : When Base Model is trained using custom labelled data to generate Fine Tuned Model

Amazon Bedrock – Architecture – Fine Tuning Foundation Models



Fine Tuning Foundation Models - Demo

DEMO



Fine Tuning Foundation Models – Model Parameters and Hyperparameters

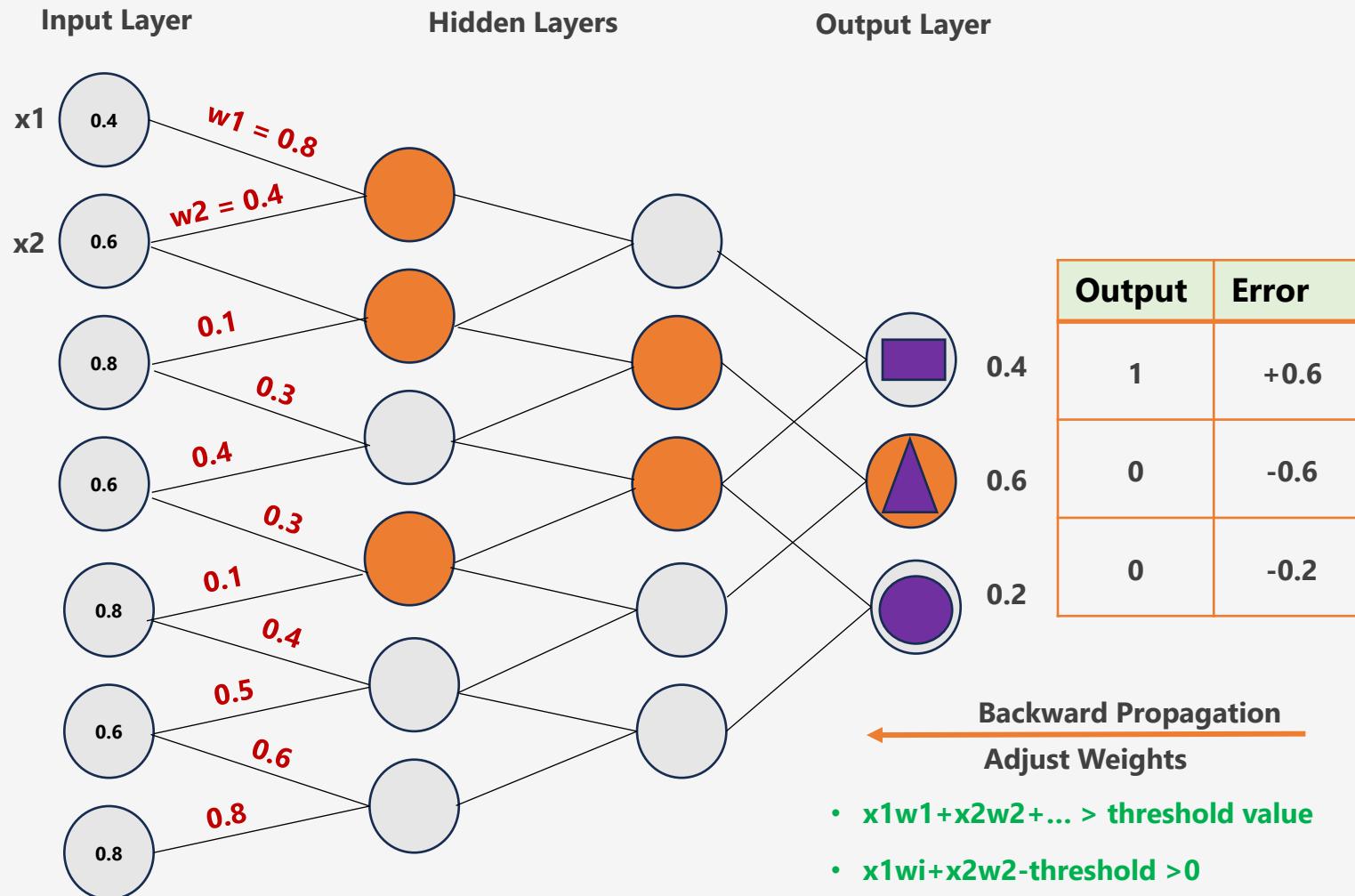
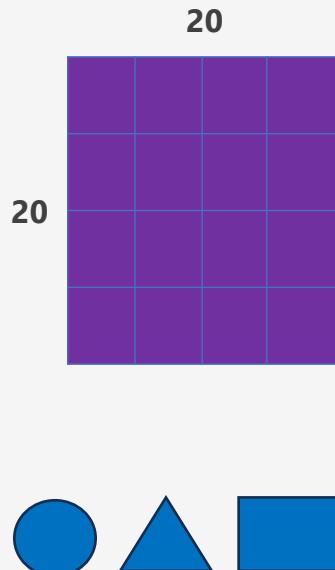
Parameters :

- Internal to Model
- Updated during training process

Hyperparameters :

- External to Model
 - Specified by developers/scientist
- 

Machine Learning – Deep Learning with Artificial Neural Networks



Instruction Fine Tuning – Key Terminology to set Hyperparameters

Terminology	Description	Example
Sample	One row of data	1
Dataset	Total number of training samples	100
Batch Size	Total number of training samples in each mini-batch	20
Iterations	Number of times Batch Size is run for entire dataset	5
Epoch	Epoch elapses when an entire dataset is run	1-Infinity
Learning rate	The rate at which model parameters are updated after each batch of training data.	

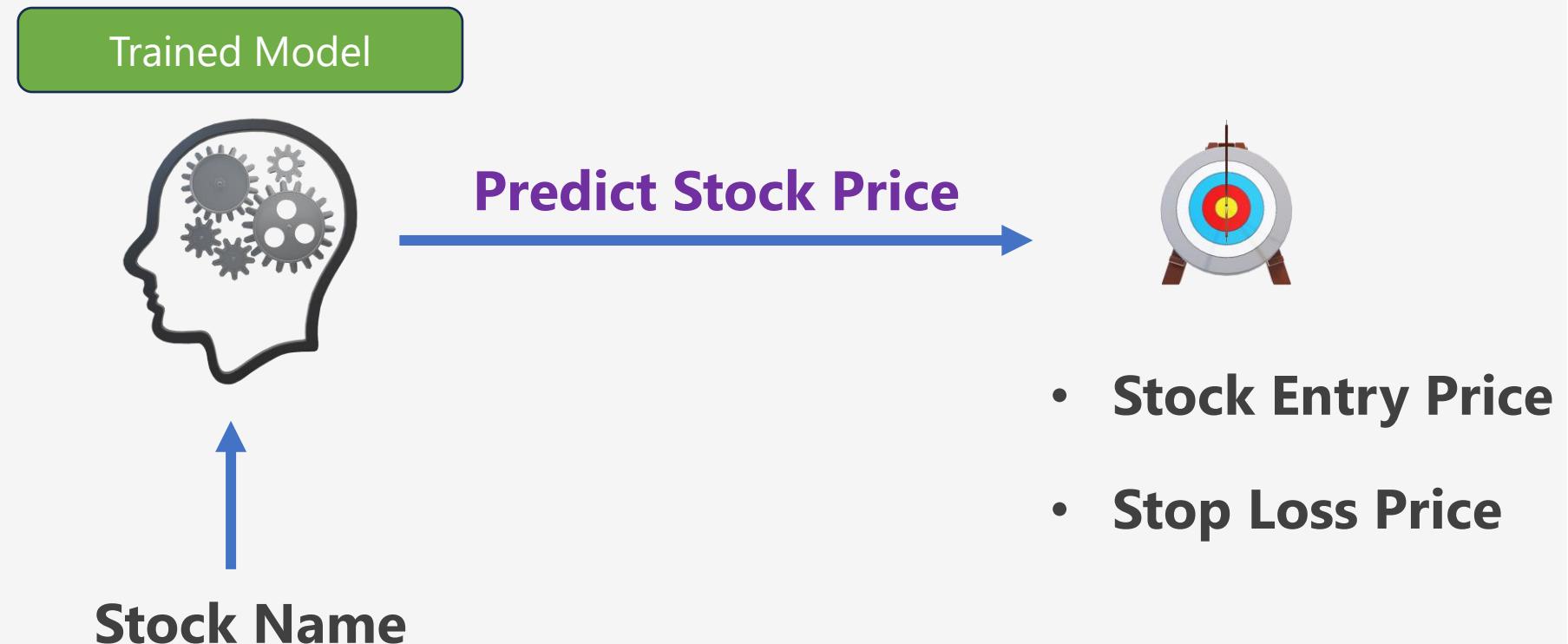


Domain 1 :

Fundamentals of AI and ML (Part 2)

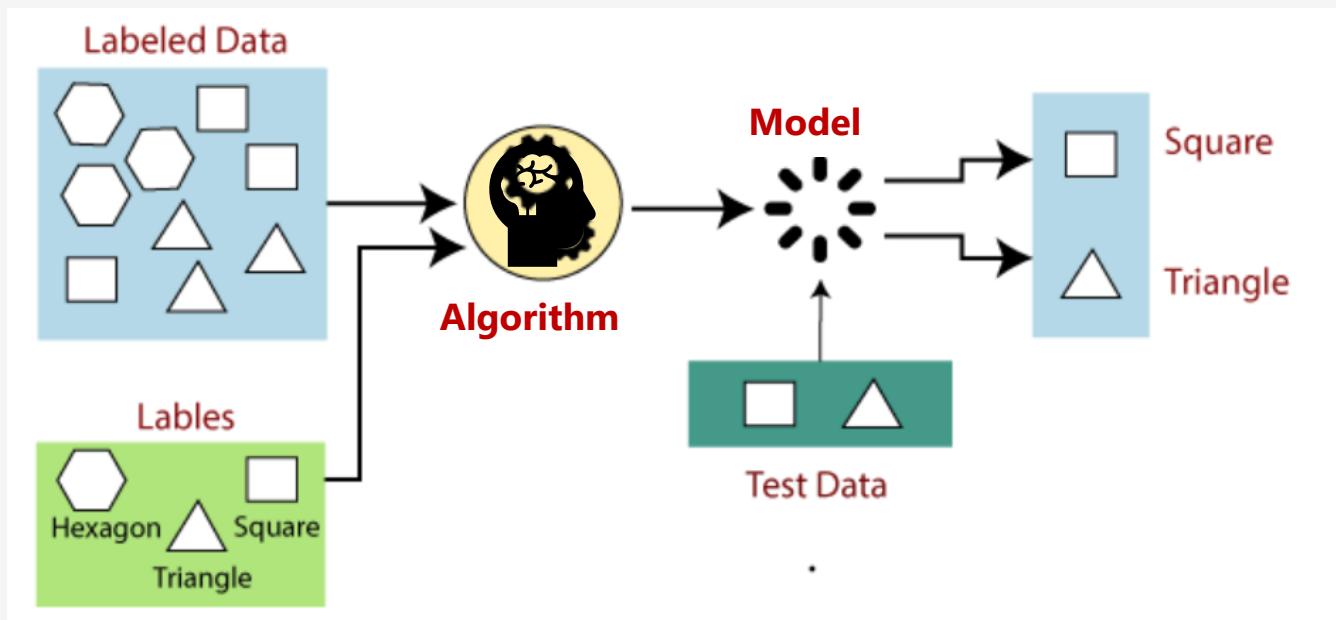
Machine Learning Development Lifecycle - Business goal (Use Case)

Use Case : Build Stock Price Prediction Solution with Machine Learning for a Trading Firm

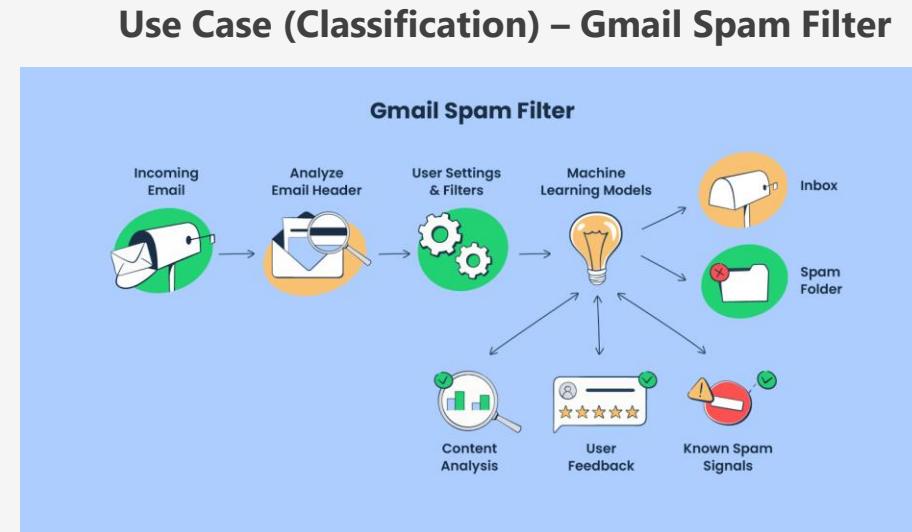


Machine Learning Development Lifecycle - ML problem framing

- In supervised learning, **algorithms are trained using labelled dataset.**
- Once the **training is completed**, the **model is tested using test data** and then it predicts the output.



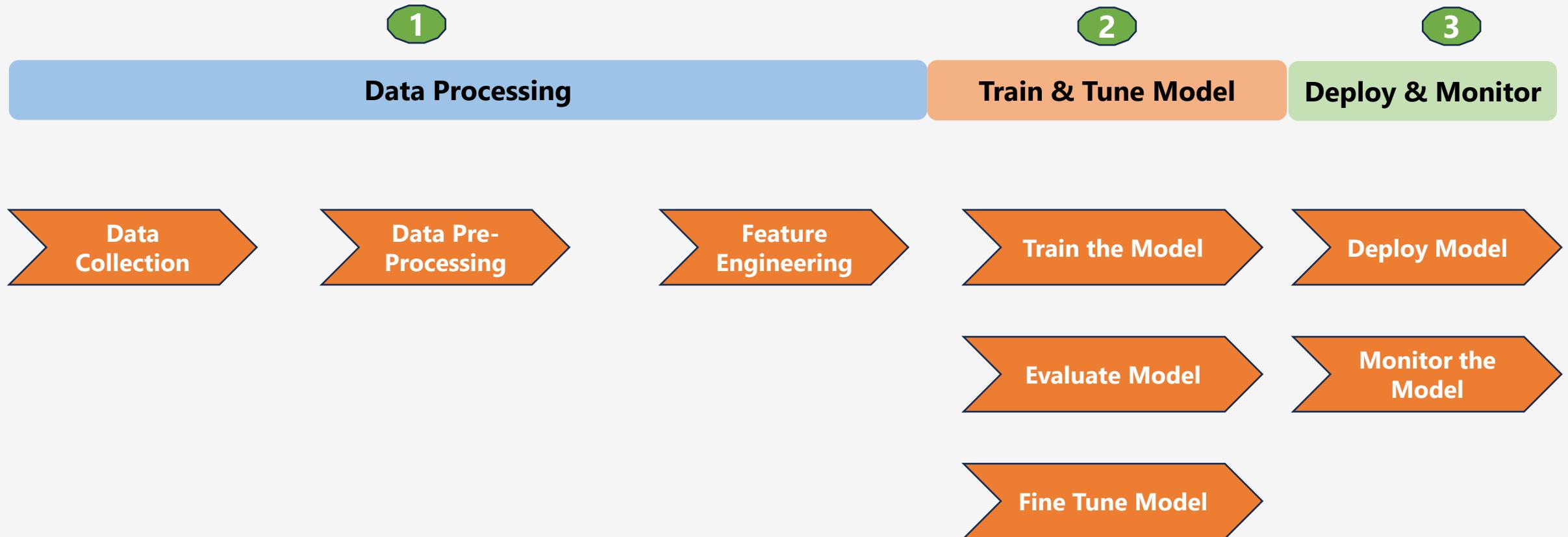
Common Algorithms - Classification and Regression



Use Case (Regression) – Forecasting

Advertisement	Sales
\$90	\$1000
\$120	\$1300
\$150	\$1800
\$100	\$1200
\$130	\$1380
\$200	??

Machine Learning Development Lifecycle - Overview



Machine Learning Development Lifecycle – Data Processing

1 Data Processing



Feature Engineering

Sample Data

	Open	High	Low	Close	Adj Close	Volume
Date						
2020-01-02	67.420502	68.433998	67.324501	68.433998	68.433998	27278000
2020-01-03	67.400002	68.687500	67.365997	68.075996	68.075996	23408000
2020-01-06	67.581497	69.916000	67.550003	69.890503	69.890503	46768000
2020-01-07	70.023003	70.175003	69.578003	69.755501	69.755501	34330000
2020-01-08	69.740997	70.592499	69.631500	70.251999	70.251999	35314000

Date	Stock Price (pre-existing feature)	7-day Moving Average (engineered feature)
July 1, 2023	\$100	N/A
July 2, 2023	\$105	N/A
July 3, 2023	\$107	N/A
July 4, 2023	\$103	N/A
July 5, 2023	\$108	N/A
July 6, 2023	\$110	N/A
July 7, 2023	\$112	\$106.43
July 8, 2023	\$115	\$108.57
July 9, 2023	\$114	\$109.85
July 10, 2023	\$113	\$110.71

Sample Data

Diagram illustrating a sample dataset with handwritten annotations:

The table has a header row labeled "Features" and a data row labeled "Label".

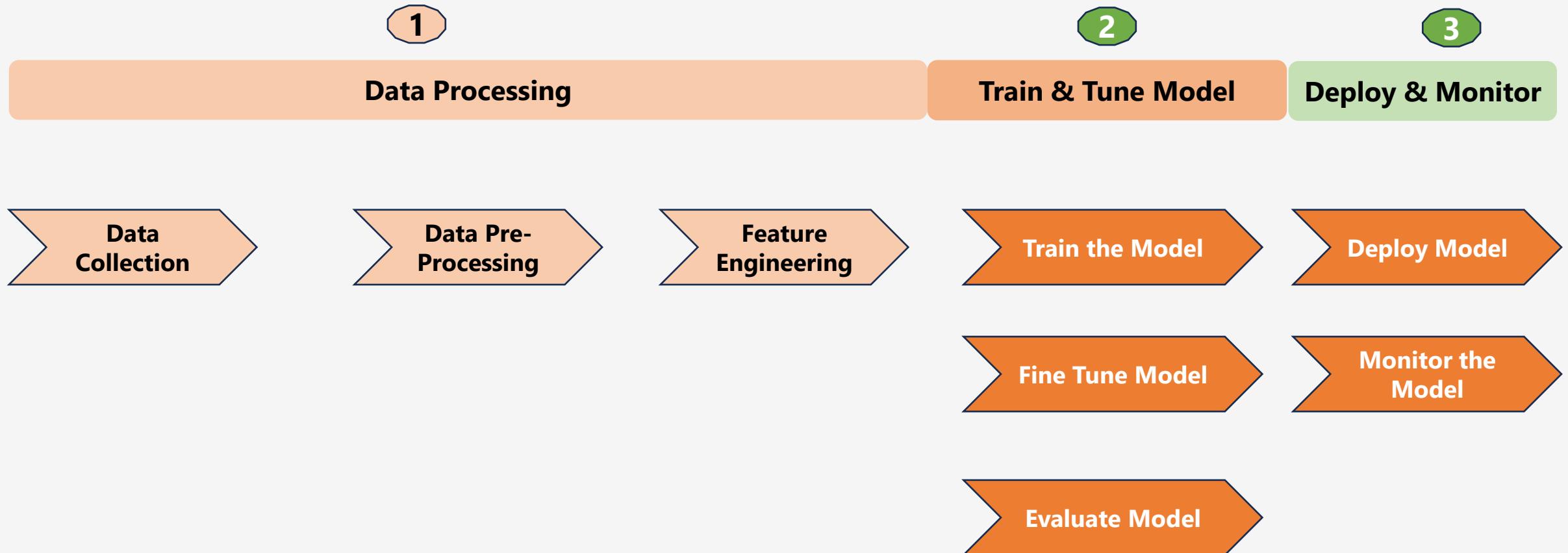
Features: date, lat, long, temp, humidity, cloud_coverage, wind_direction, atmp_pressure.

Label: rainfall

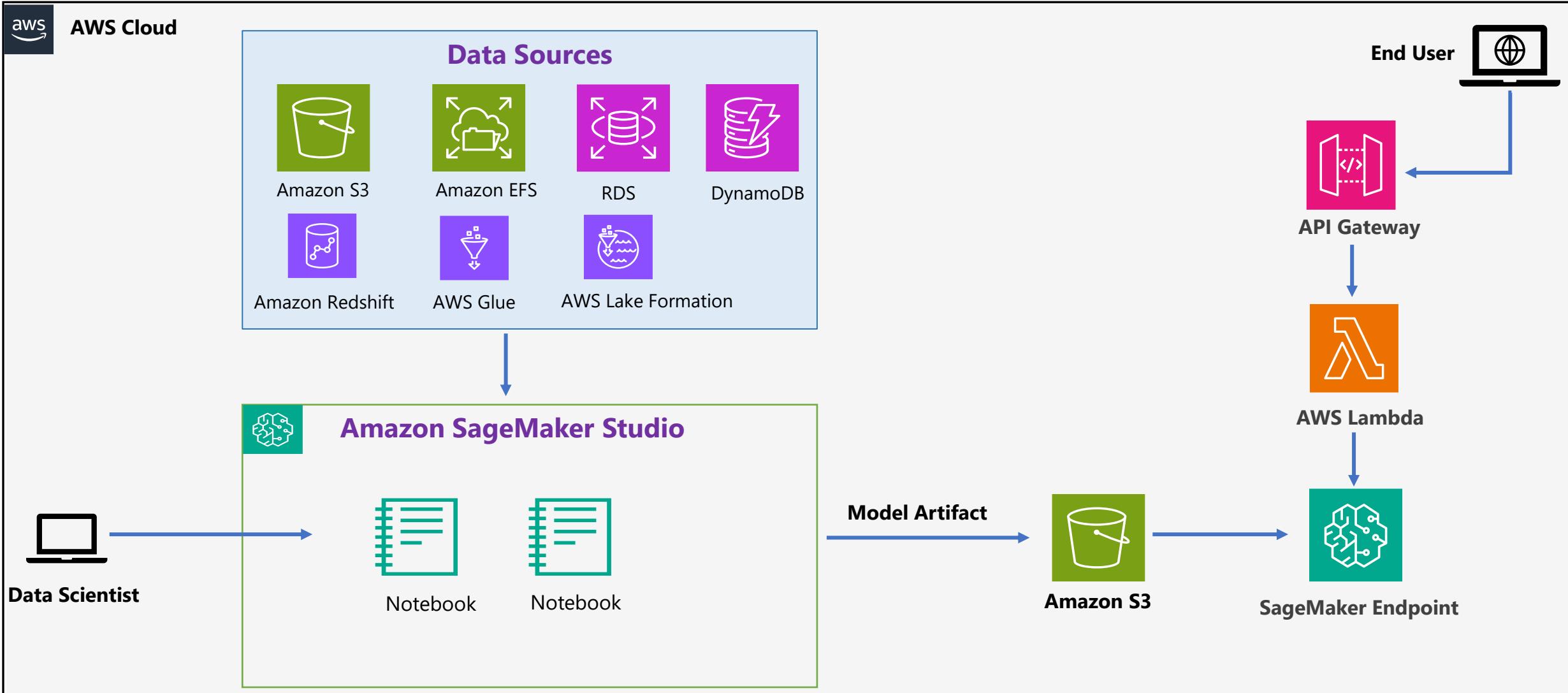
Example: The first row shows data for 2021-09-09 at 49.71N, 82.16W with a rainfall value of .01. The second row shows data for 2021-09-09 at 32.71N, 117.16W with a rainfall value of .23.

	Features								Label
	date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure	rainfall
2021-09-09	49.71N	82.16W	74	20	3	N	18.6	.01	
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94	.23	

Machine Learning Development Lifecycle - Train & Tune Model

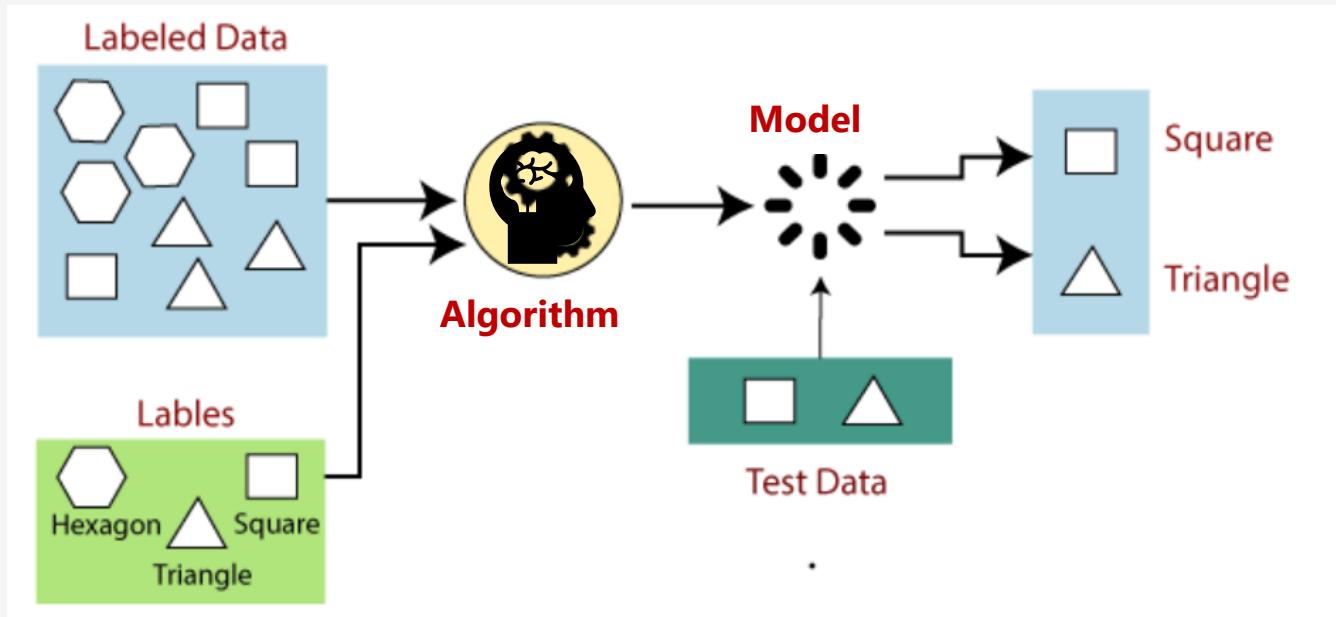


Machine Learning Development Lifecycle with AWS Services

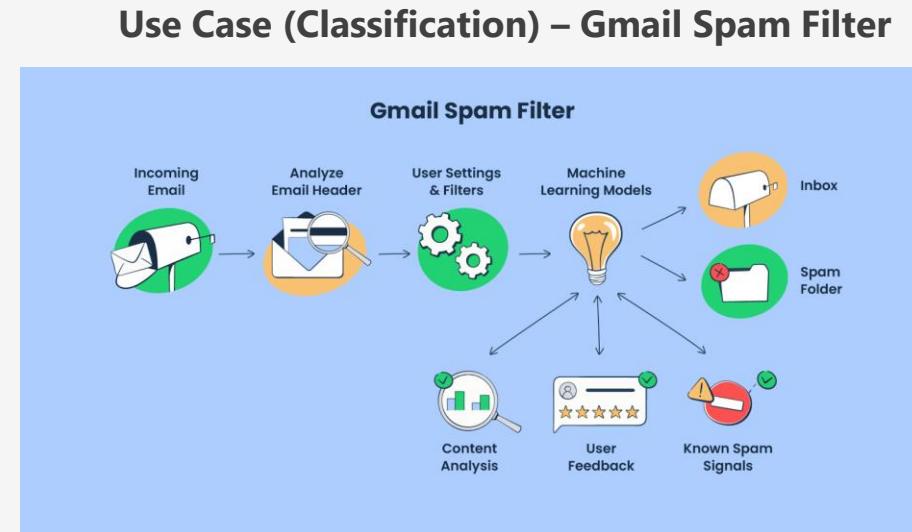


Machine Learning – Supervised Learning

- In supervised learning, **algorithms are trained using labelled dataset.**
- Once the **training is completed**, the **model is tested using test data** and then it predicts the output.



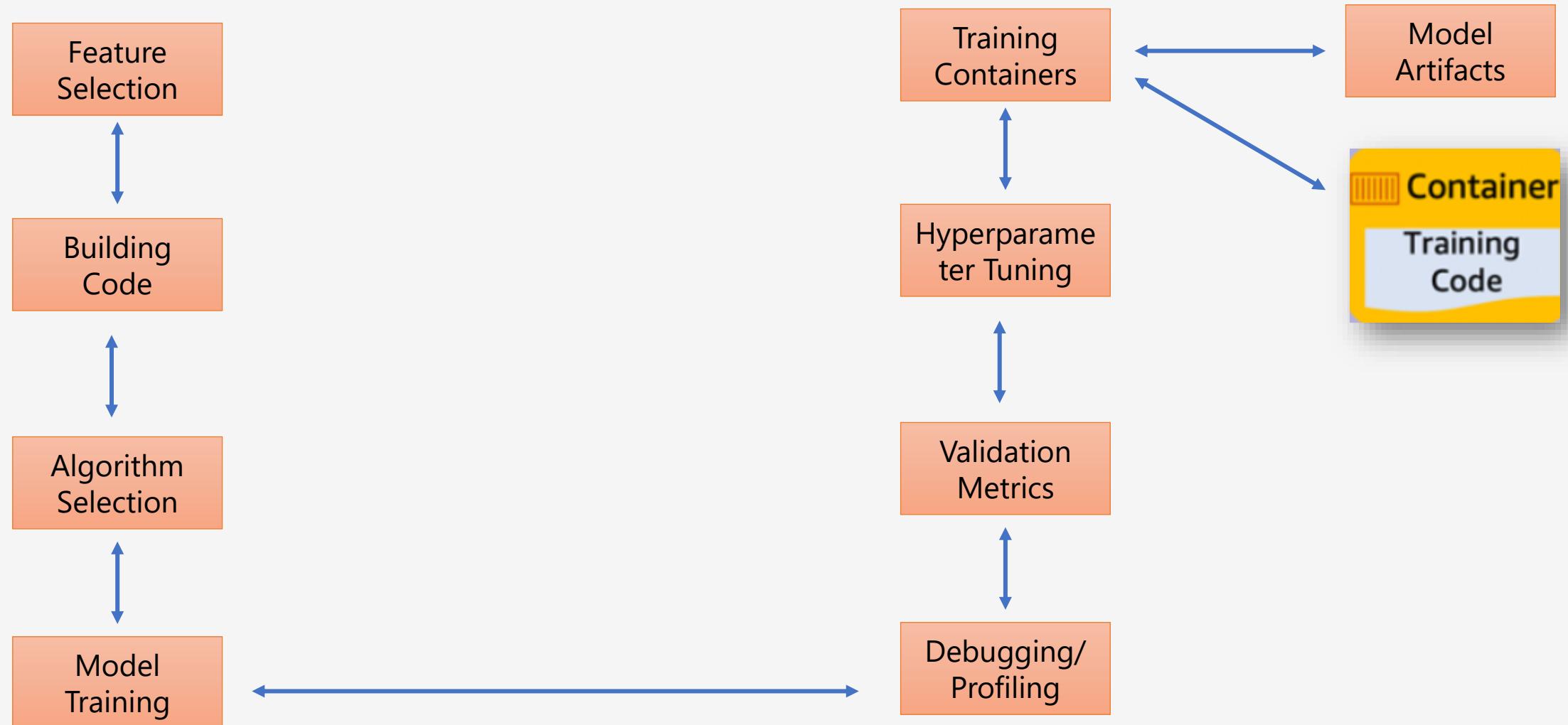
Common Algorithms - Classification and Regression



Use Case (Regression) – Forecasting

Advertisement	Sales
\$90	\$1000
\$120	\$1300
\$150	\$1800
\$100	\$1200
\$130	\$1380
\$200	??

Machine Learning Development Lifecycle – Train and Fine Tune Model



<https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html>

Machine Learning Development Lifecycle - Model Evaluation

Evaluate your model using historical data (**offline evaluation**) or live data (**online**).

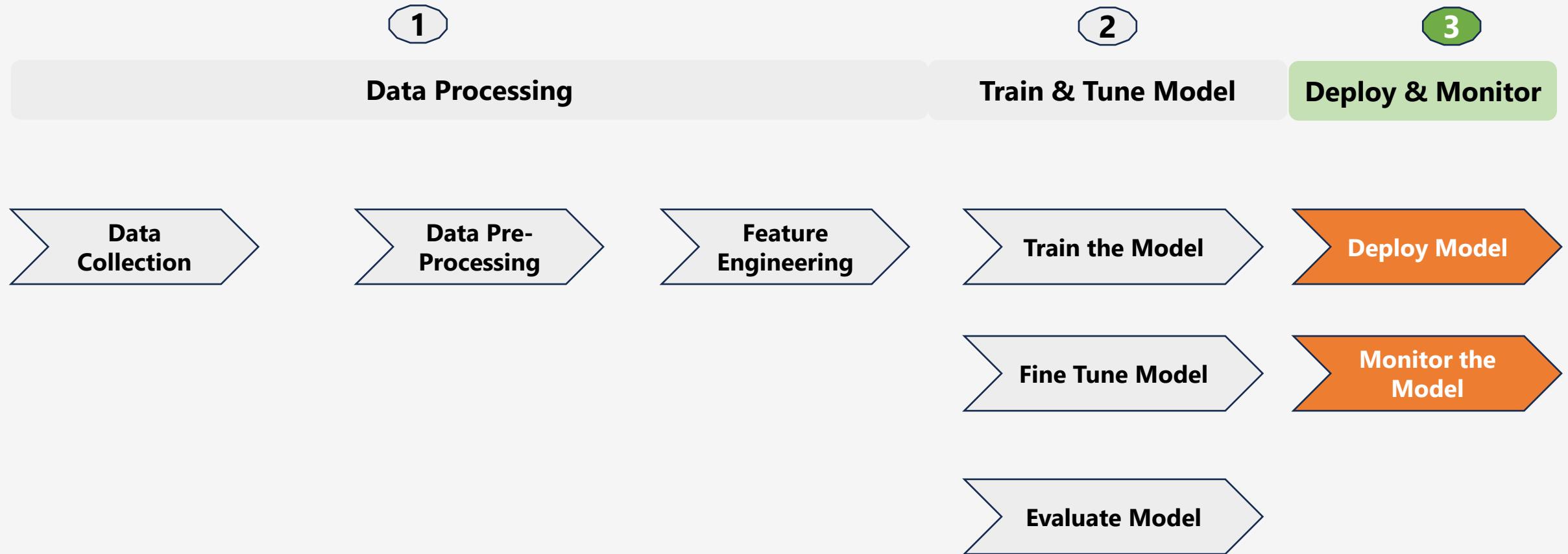
Offline evaluation - trained model evaluated with a portion of the dataset set aside as a **holdout set**

Holdout data - never used for model training or validation- used to evaluate errors in the final model

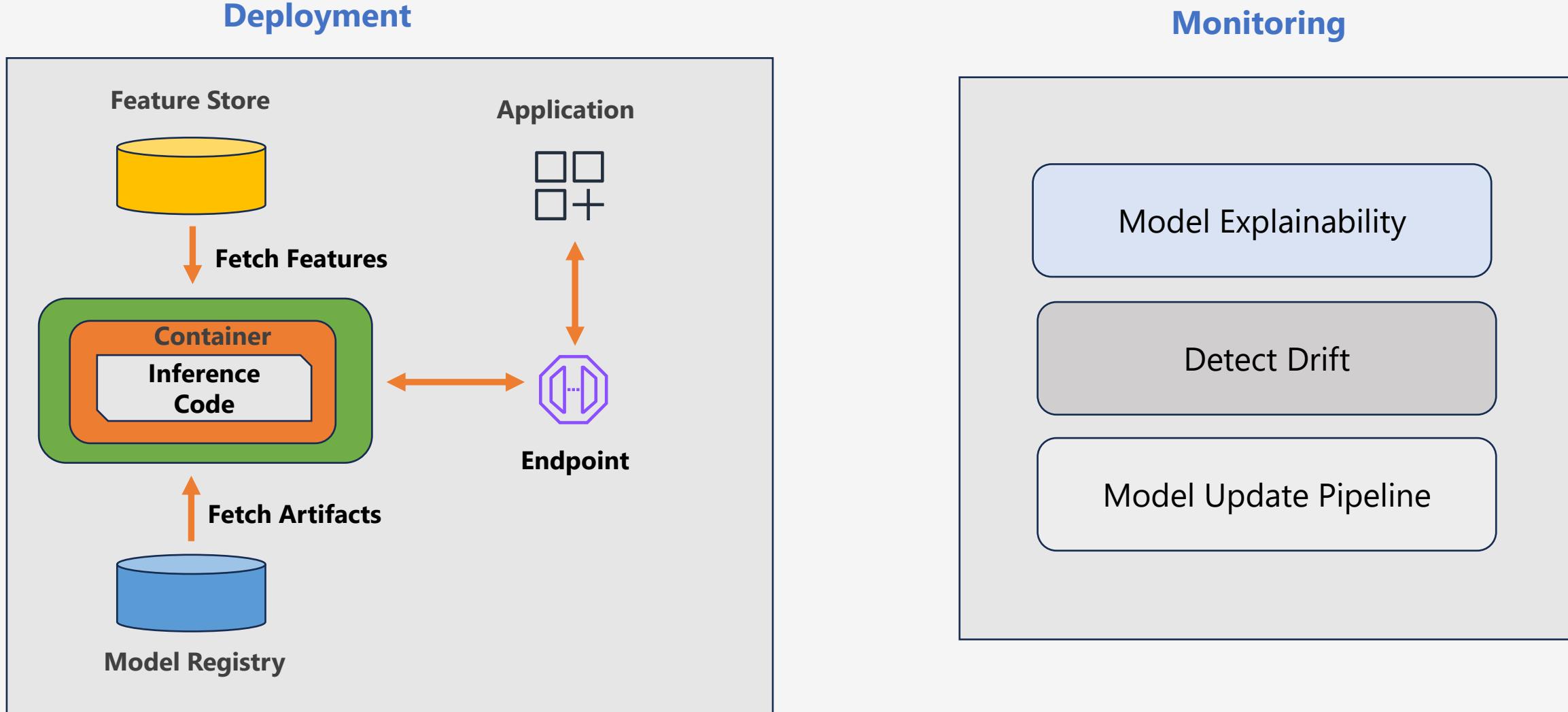
Based on evaluation results - **Fine-tune the data, the algorithm, or both**

Fine-tune the data - apply the concepts of data cleansing, preparation and feature engineering.

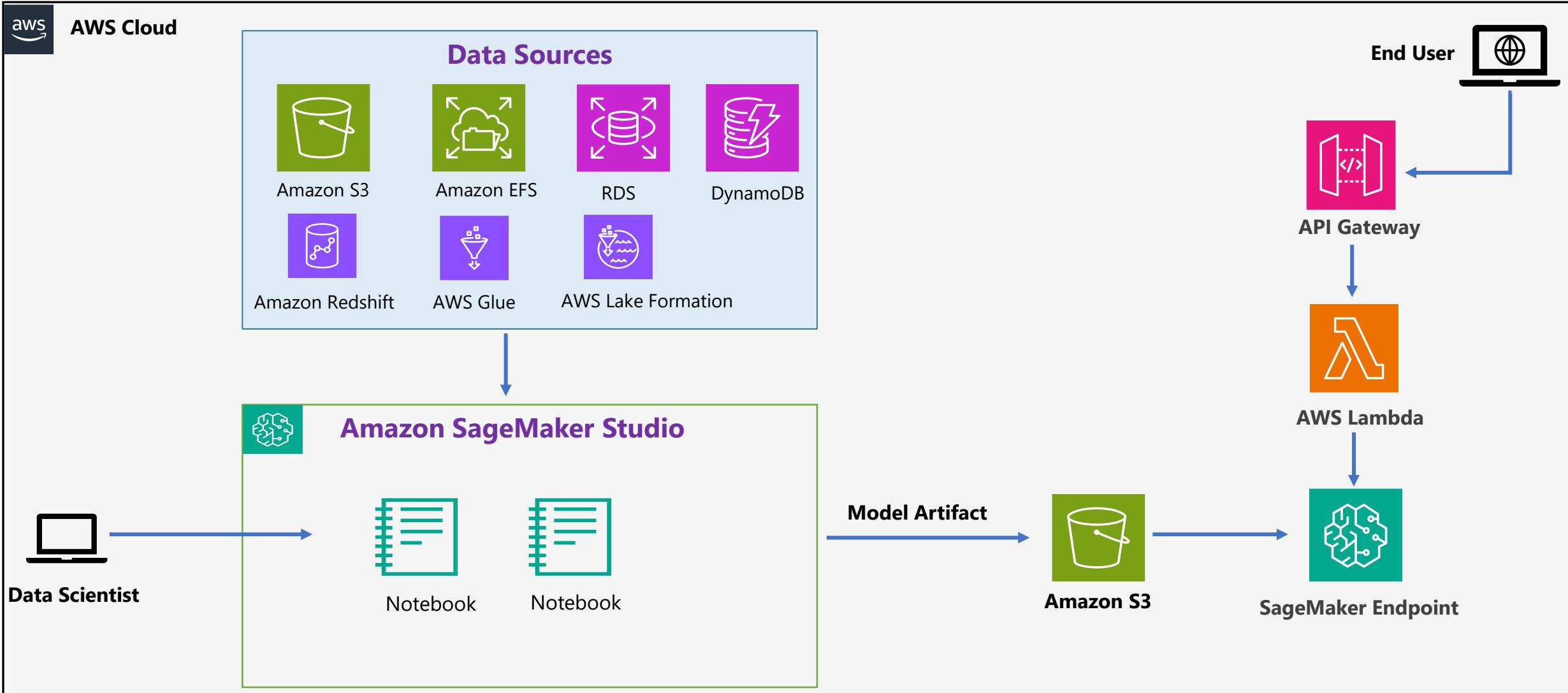
Machine Learning Development Lifecycle – Deploy and Monitor



Machine Learning Development Lifecycle – Deployment and Monitoring

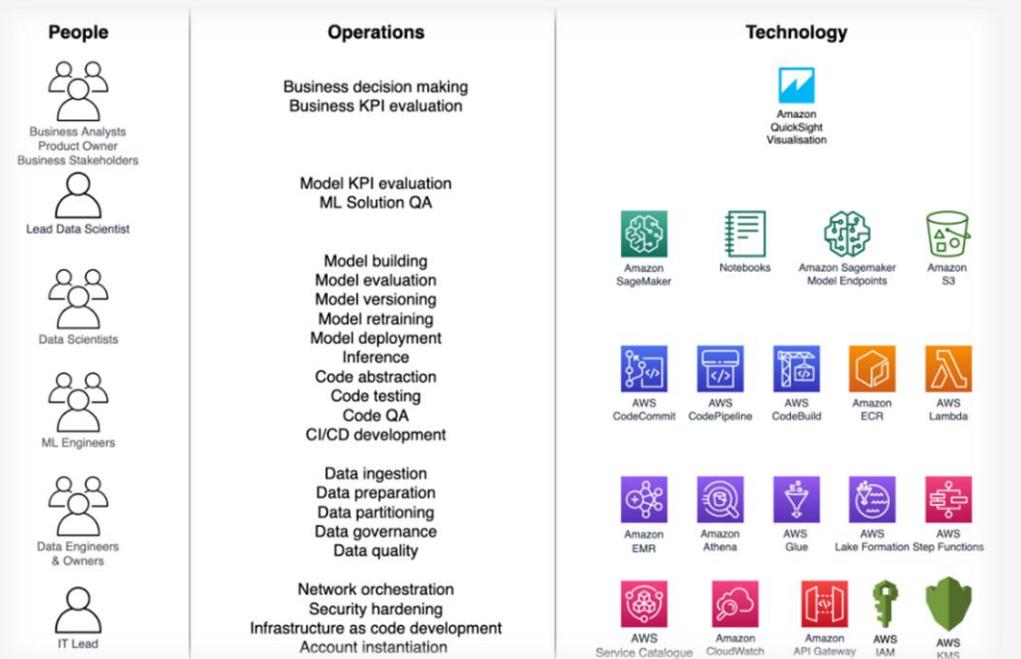
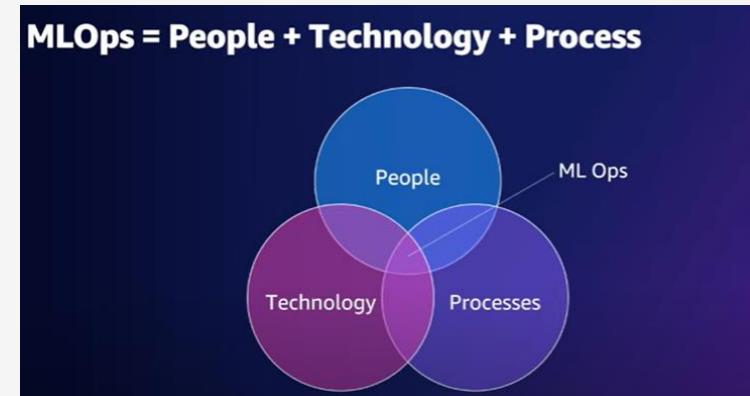


Machine Learning Development Lifecycle with AWS Services

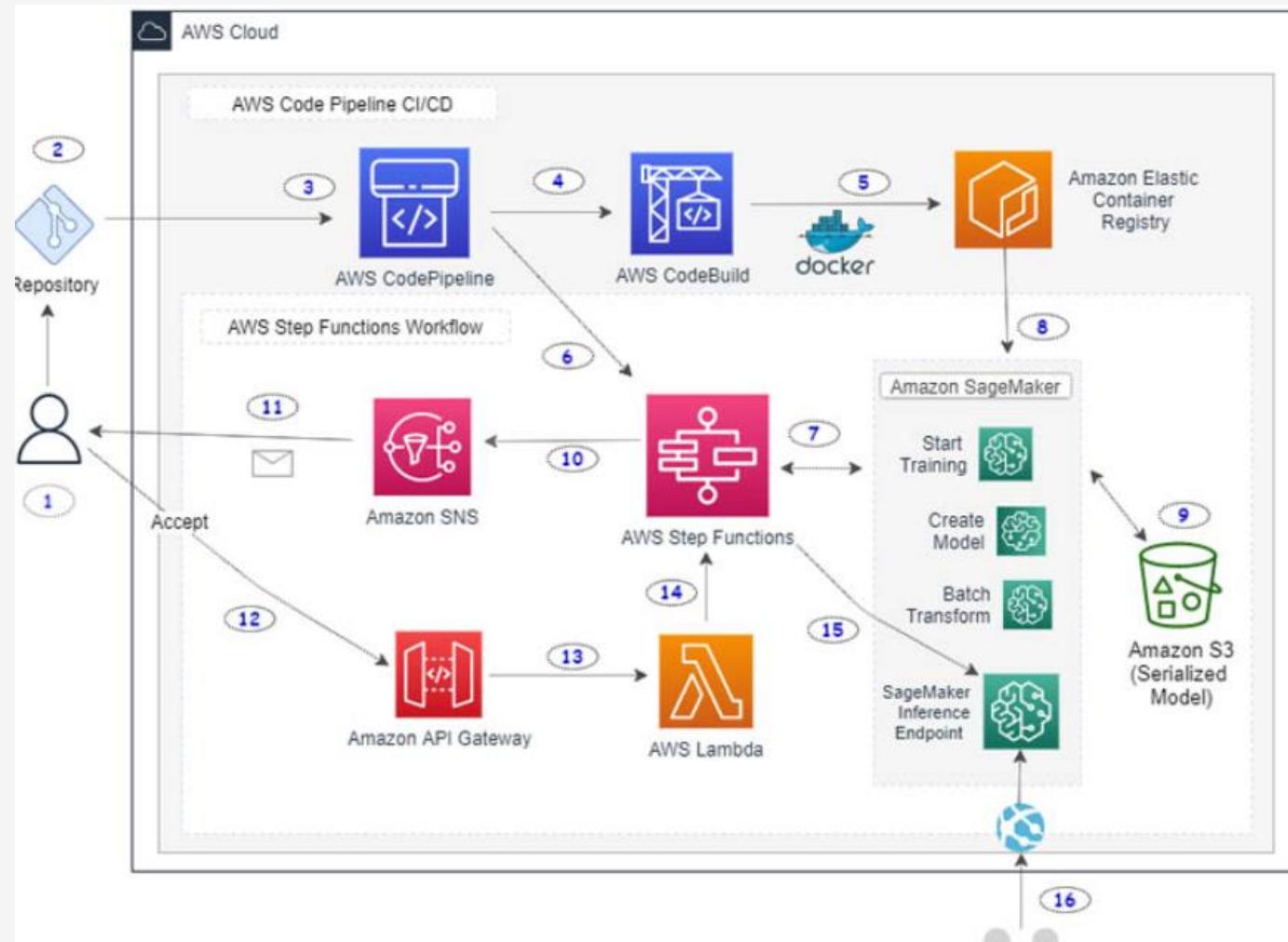


Machine Learning Operations (MLOps)

- MLOps is a **culture and practices** that
- **Unifies ML application development (Dev) & ML system deployment and operations (Ops).**
- Use **MLOps to automate and standardize processes** across the ML lifecycle.
- Processes include :
 - Model development, testing, integration, release and infrastructure management.

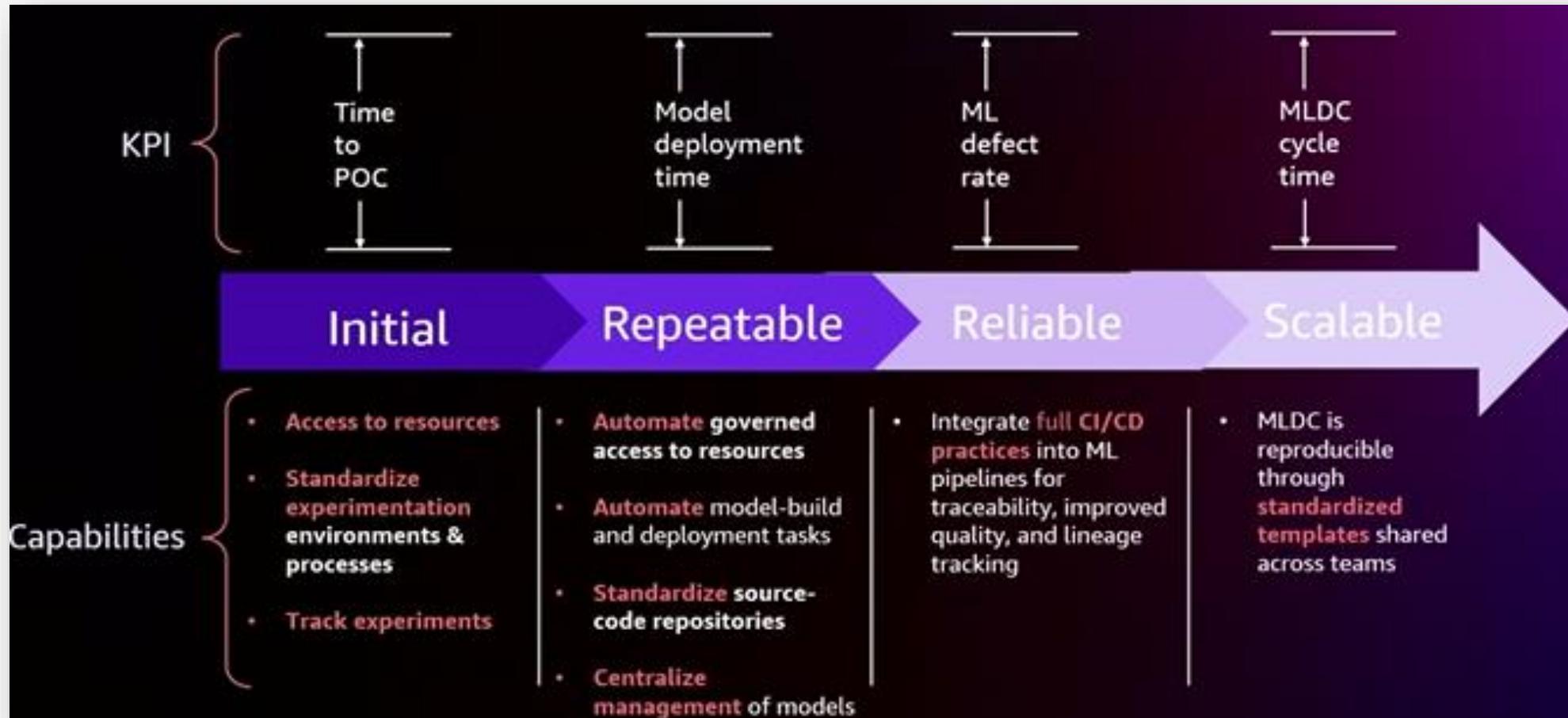


Machine Learning Operations (MLOps)



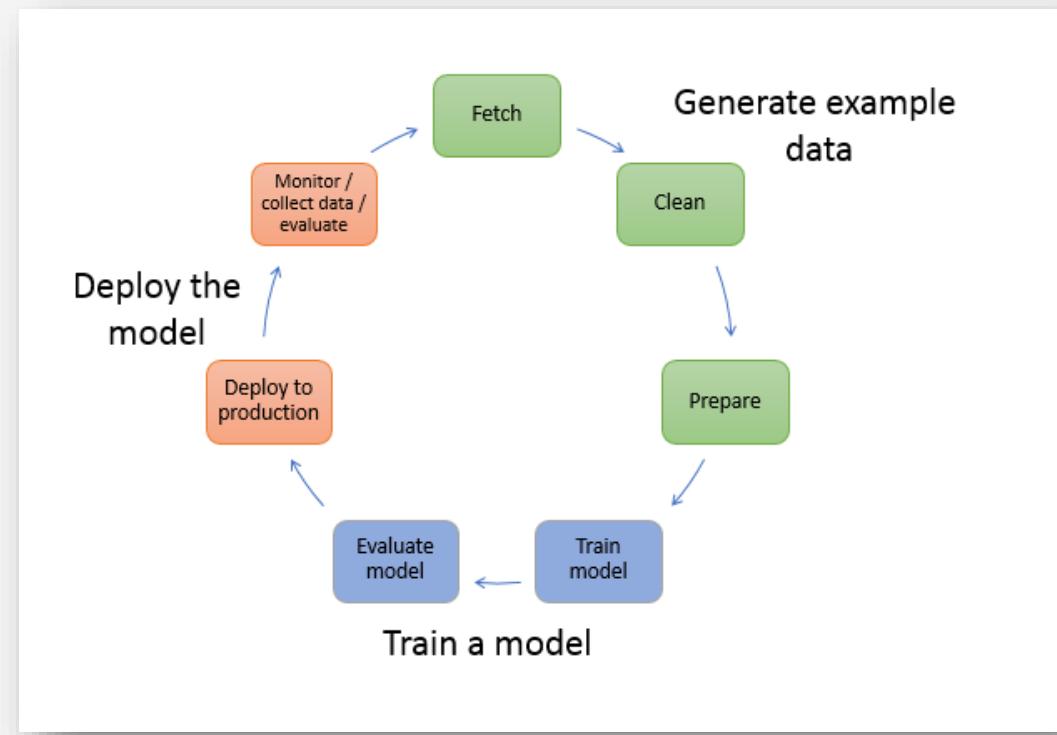
<https://aws.amazon.com/blogs/machine-learning/build-a-ci-cd-pipeline-for-deploying-custom-machine-learning-models-using-aws-services/>

MLOps – Maturity Framework

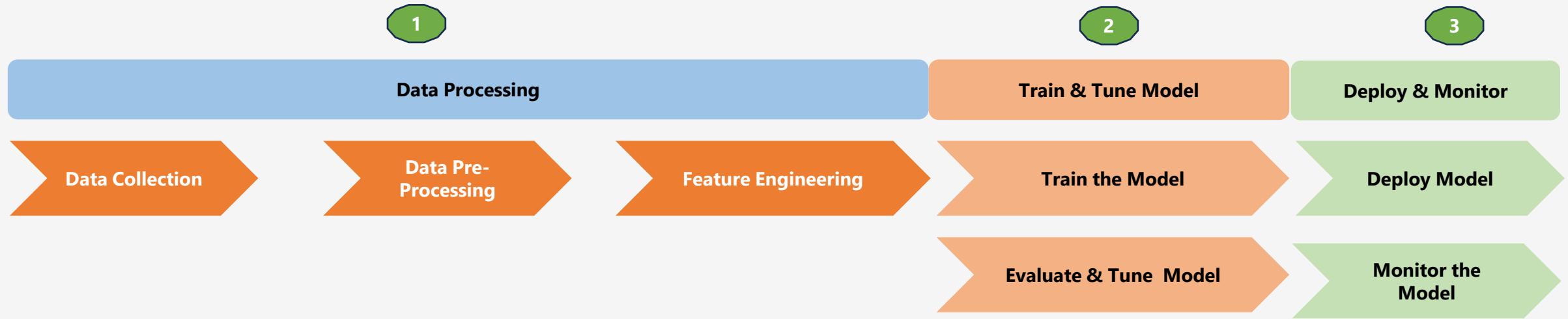


Amazon SageMaker - Overview

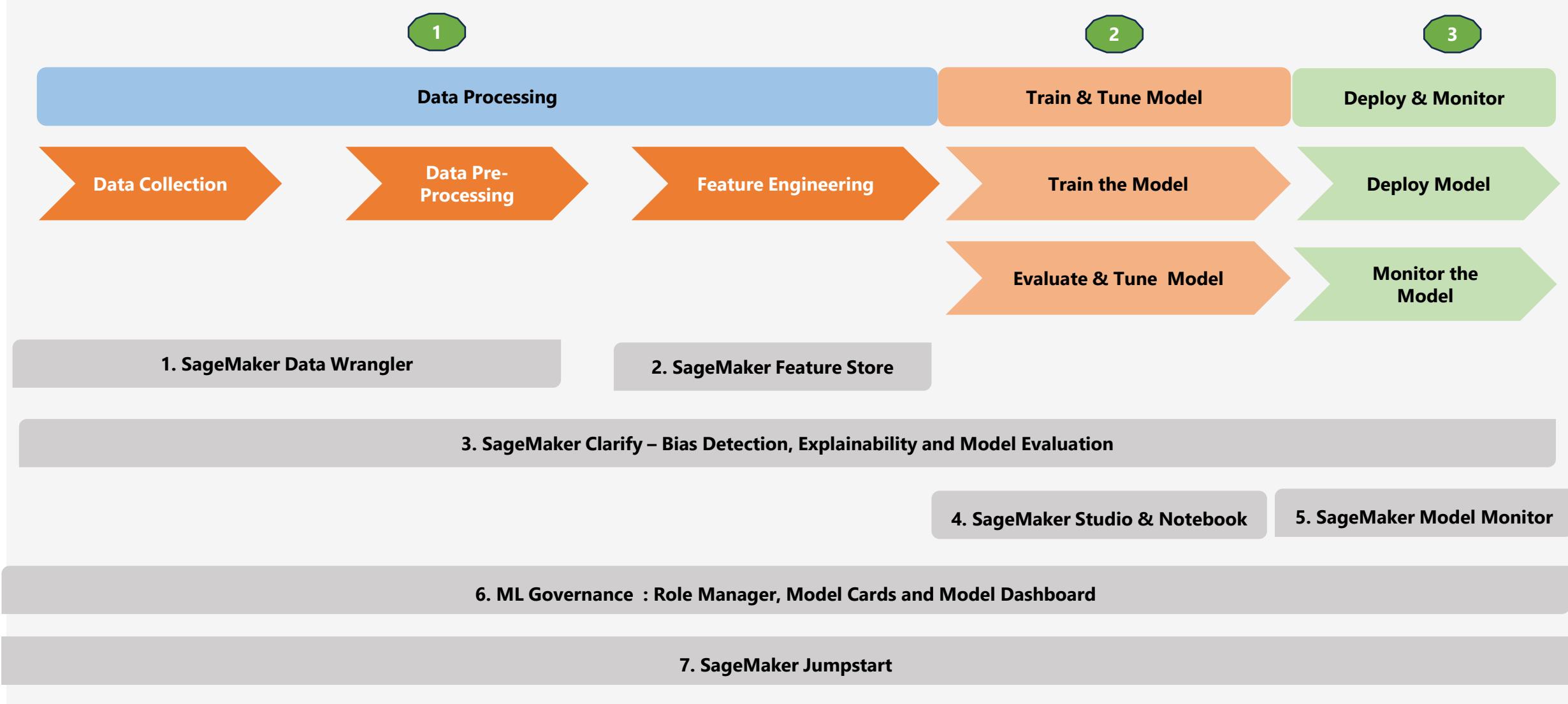
- Amazon SageMaker is a **fully managed machine learning service**.
- Used by data scientists and developers to **build, train, and deploy ML models**
- Collection of many different services



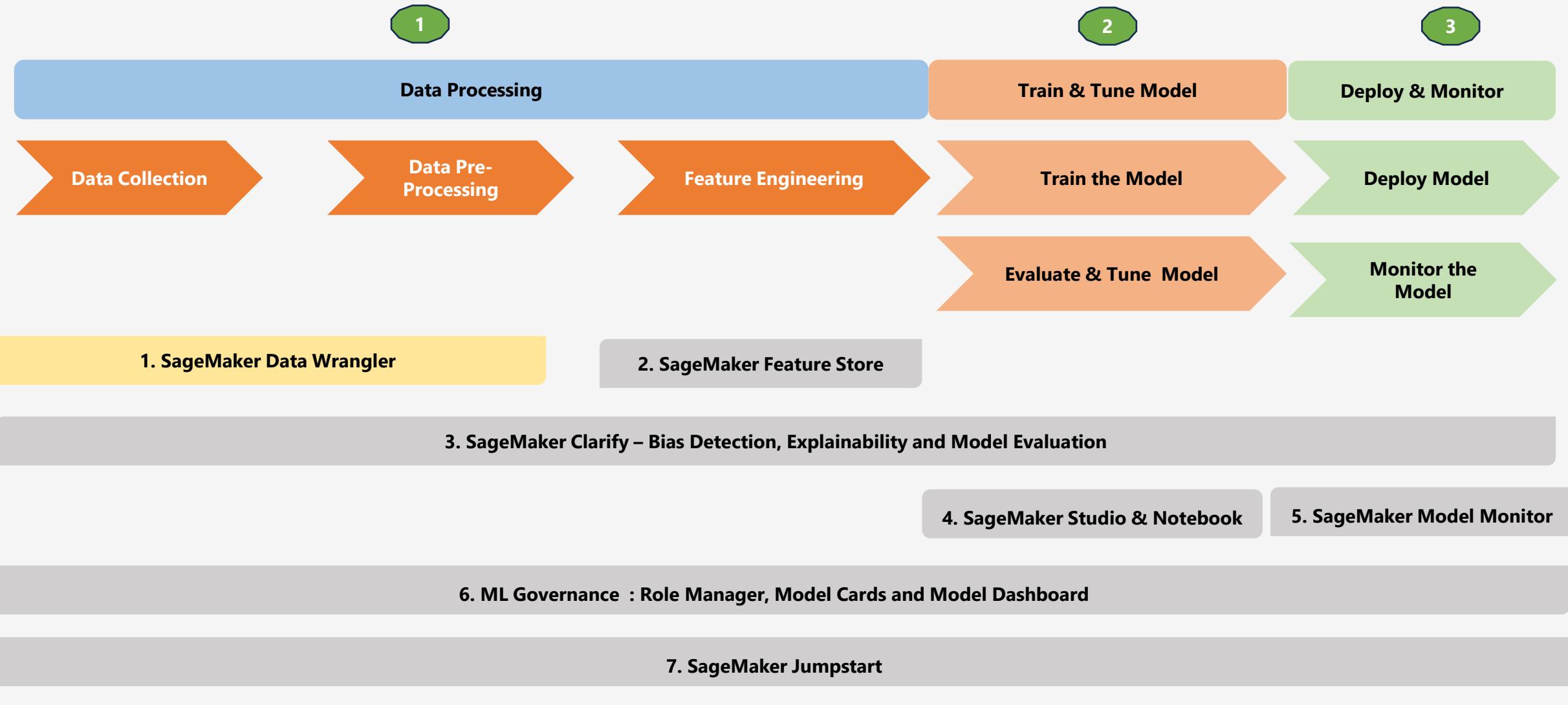
Machine Learning Development Lifecycle - Overview



Amazon SageMaker mapping to ML Lifecycle



Machine Learning Development Lifecycle - Overview

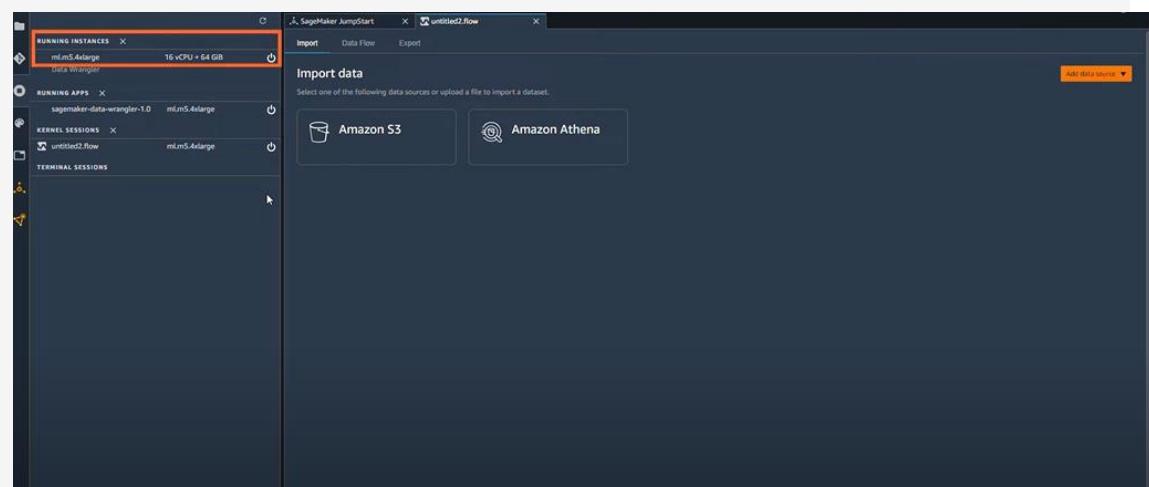
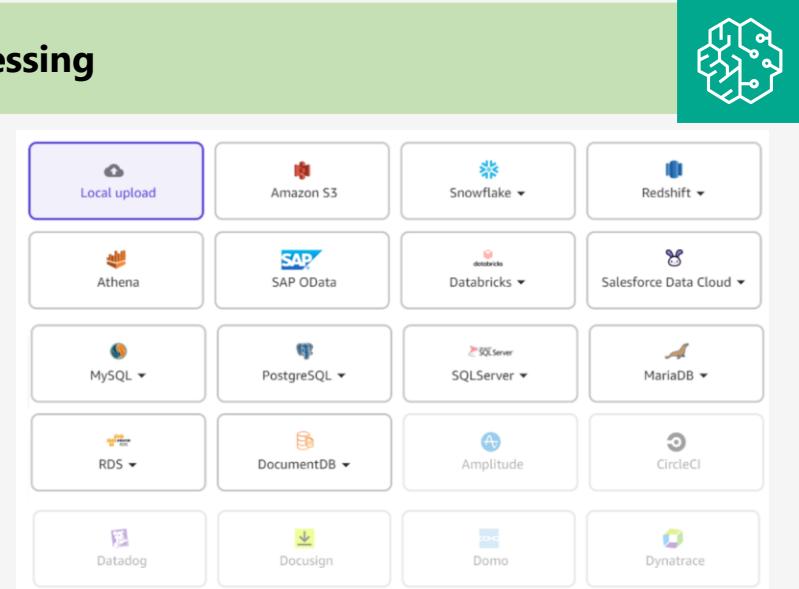


1. Amazon SageMaker Data Wrangler

1

Data Processing – Data Collection and Data Pre-Processing

1. Import data from Amazon services such as S3, Athena, Redshift, 50+ third-party sources.
2. Understand data quality and generate data insights - **Exploratory Data Analysis (EDA)**
 - Verifies data quality - missing values, duplicate rows
 - Detect anomalies - outliers, class imbalance, and data leakage
3. Transform data more efficiently and Analyse
 - Drop column, Handle outliers
 - Prebuilt or custom transformation
4. Export the Data
 - S3
 - SageMaker Feature Store
 - SageMaker Pipeline



1. Import data

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest			
1	1	Allen, Miss. Elisabeth Walto	female	29	0	0	24160	211.3375	B5	S		2 ?	St Louis, MO			
1	1	Allison, Master. Hudson Tre	male	0.9167	1	2	113781	151.55	C22 C26	S		11 ?	Montreal, PQ / Chesterville, ON			
1	0	Allison, Miss. Helen Loraine	female	2	1	2	113781	151.55	C22 C26	S	?	?	Montreal, PQ / Chesterville, ON			
1	0	Allison, Mr. Hudson Joshua C	male	30	1	2	113781	151.55	C22 C26	S	?	?	135 Montreal, PQ / Chesterville, ON			
1	0	Allison, Mrs. Hudson J C (Bem	female	25	1	2	113781	151.55	C22 C26	S	?	?	Montreal, PQ / Chesterville, ON			
1	1	Anderson, Mr. Harry	male	48	0	0	19952	26.55	E12	S		3 ?	New York, NY			
1	1	Andrews, Miss. Kornelia Th	female	63	1	0	13502	77.9583	D7	S		10 ?	Hudson, NY			
1	0	Andrews, Mr. Thomas Jr	male	39	0	0	112050	0	A36	S	?	?	Belfast, NI			
1	1	Appleton, Mrs. Edward Dale	female	53	2	0	11769	51.4792	C101	S	D	?	Bayside, Queens, NY			
1	0	Artagaveytia, Mr. Ramon	male	71	0	0	PC 17609	49.5042	?	C	?		22 Montevideo, Uruguay			
1	0	Astor, Col. John Jacob	male	47	1	0	PC 17757	227.525	C62 C64	C	?		124 New York, NY			
1	1	Astor, Mrs. John Jacob (Madi	female	18	1	0	PC 17757	227.525	C62 C64	C		4 ?	New York, NY			
1	1	Aubart, Mme. Leontine Paul	female	24	0	0	PC 17477	69.3	B35	C		9 ?	Paris, France			
1	1	Barber, Miss. Ellen 'Nellie'	female	26	0	0	19877	78.85	?	S		6 ?	?			
1	1	Barkworth, Mr. Algernon He	male	80	0	0	27042	30	A23	S	B	?	Hessle, Yorks			
1	0	Baumann, Mr. John D	male	?	0	0	PC 17318	25.925	?	S	?	?	New York, NY			
1	0	Baxter, Mr. Quigg Edmond	male	24	0	1	PC 17558	247.5208	B58 B60	C	?	?	Montreal, PQ			
1	1	Baxter, Mrs. James (Helene	female	50	0	1	PC 17558	247.5208	B58 B60	C		6 ?	Montreal, PQ			
1	1	Bazzanti, Miss. Albina	female	32	0	0	11813	76.2917	D15	C		8 ?	?			
1	0	Beattie, Mr. Thomson	male	36	0	0	13050	75.2417	C6	C	A	?	Winnipeg, MN			
1	1	Beckwith, Mr. Richard Leona	male	37	1	1	11751	52.5542	D35	S		5 ?	New York, NY			
1	1	Beckwith, Mrs. Richard Leon	female	47	1	1	11751	52.5542	D35	S		5 ?	New York, NY			
1	1	Behr, Mr. Karl Howell	male	26	0	0	111369	30	C148	C		5 ?	New York, NY			
1	1	Bidlois, Miss. Rosalie	female	42	0	0	PC 17757	227.525	?	C		4 ?	?			
1	1	Bird, Miss. Ellen	female	29	0	0	PC 17483	221.7792	C97	S		8 ?	?			
1	0	Birnbaum, Mr. Jakob	male	25	0	0	13905	26	?	C	?		148 San Francisco, CA			
1	1	Bishop, Mr. Dickinson H	male	25	1	0	11967	91.0792	B49	C		7 ?	Dowagiac, MI			
1	1	Bishop, Mrs. Dickinson H (He	female	19	1	0	11967	91.0792	B49	C		7 ?	Dowagiac, MI			
1	1	Bissette, Miss. Amelia	female	35	0	0	PC 17760	135.6333	C99	S		8 ?	?			
1	1	Bjornstrom-Steffansson, Mr.	male	28	0	0	110564	26.55	C52	S	D	?	Stockholm, Sweden / Washington, DC			
1	0	Blackwell, Mr. Stephen Wea	male	45	0	0	113784	35.5	T	S	?	?	Trenton, NJ			
1	1	Blank, Mr. Henry	male	40	0	0	112277	31	A31	C		7 ?	Glen Ridge, NJ			
1	1	Bonnell, Miss. Caroline	female	30	0	0	36928	164.8667	C7	S		8 ?	Youngstown, OH			
1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S		8 ?	Birkdale, England Cleveland, Ohio			
1	0	Borebank, Mr. John James	male	42	0	0	110489	26.55	D22	S	?	?	London / Winnipeg, MB			
1	1	Bowen, Miss. Grace Scott	female	45	0	0	PC 17608	262.375	?	C		4 ?	Cooperstown, NY			
1	1	Bowerman, Miss. Elsie Edith	female	22	0	1	113505	55	E33	S		6 ?	St Leonards-on-Sea, England Ohio			
1	1	Bradley, Mr. George F(George	male	?	0	0	111427	26.45	?	C	?	9 ?	Los Angeles, CA			

2. Generate data insights and understand data quality (EDA)

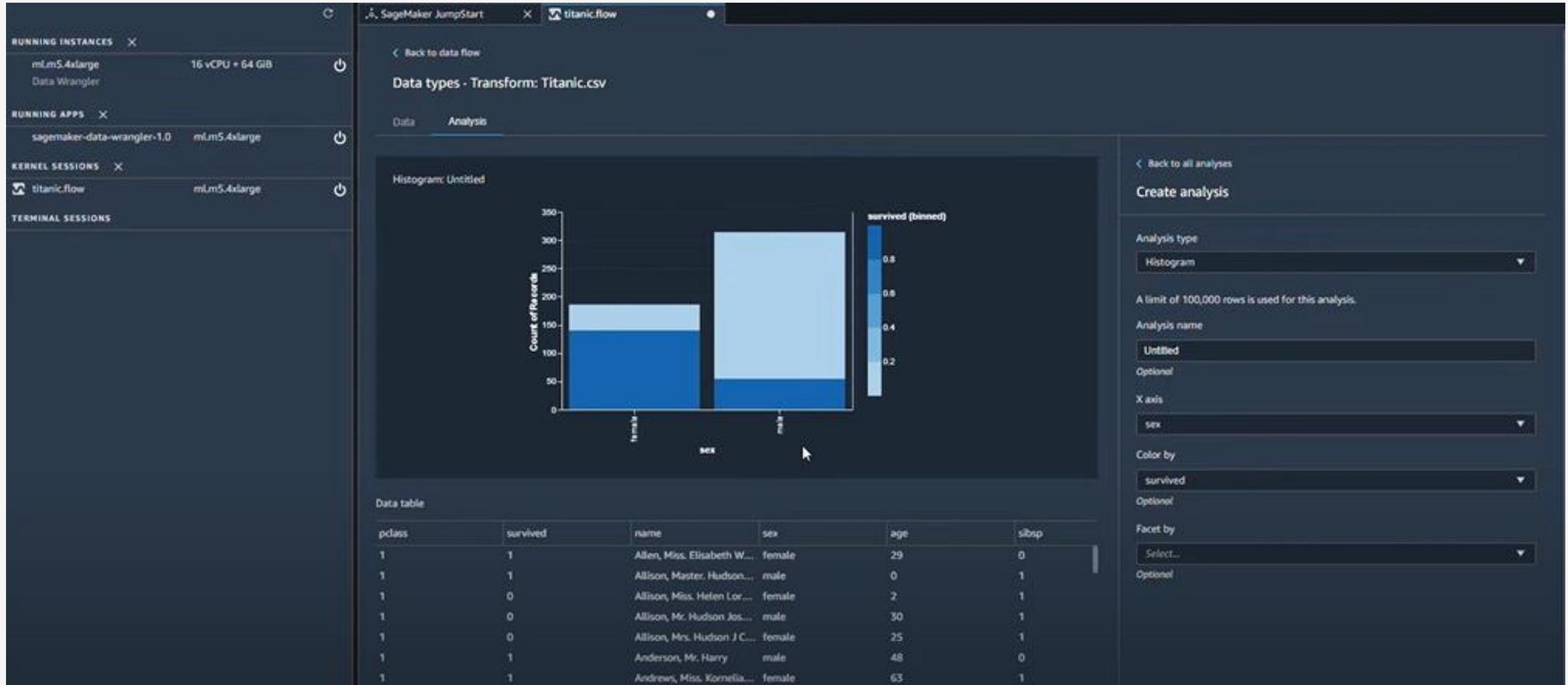
The screenshot shows the AWS SageMaker Data Wrangler interface. On the left, there's a sidebar with sections for RUNNING INSTANCES, RUNNING APPS, KERNEL SESSIONS, and TERMINAL SESSIONS. The KERNEL SESSIONS section shows a session named "titanic.flow" running on an mLM5.4xlarge instance.

The main area displays the "Data types - Transform: Titanic.csv" page. It has tabs for "Data" and "Analysis". The "Analysis" tab is active, showing a "Table Summary: Untitled" table. This table provides statistical summary information for the "summary" column across six columns: pclass, survived, name, sex, and age. The first row of the table is highlighted with a red box. The table includes rows for count, mean, stddev, min, and max.

Below the table summary, there's a "Data table" section showing a subset of the original Titanic dataset. The columns shown are pclass, survived, name, sex, age, and sibsp. The first few rows of this table are also highlighted with a red box.

On the right side of the screen, there's a sidebar for "Create analysis" with fields for "Analysis type" (set to "Table Summary"), "Analysis name" (set to "Untitled"), and an "Optional" field.

2. Generate data insights and understand data quality (EDA)



3. Transform Data

The screenshot shows the AWS SageMaker JumpStart interface. On the left, there's a sidebar with sections for RUNNING INSTANCES, RUNNING APPS, KERNEL SESSIONS, and TERMINAL SESSIONS. A terminal session named 'titanic.flow' is currently selected. The main area is titled 'Data types - Transform: Titanic.csv'. It has tabs for 'Data' and 'Analysis', and is currently on 'Data'. Below this is 'Step 2: Data types', which displays a table of data from the Titanic dataset. The columns are: pclass (long), survived (long), name (string), sex (string), age (long), sibsp (long), and pa (long). The table contains 13 rows of sample data. To the right of the table is a sidebar titled 'ADD TRANSFORM' containing various transformation options:

- Custom transform: Use PySpark, Pandas, or PySpark (SQL) to define custom transformations.
- Balance data: Balance the data for binary classification problems using random oversampling.
- Custom formula: Define a new column using a Spark SQL expression to query data in the current DataFrame.
- Encode categorical: Convert categorical variables to numeric or vector representations.
- Featurize date/time: Encode date/time values to numeric and vector representations.
- Featurize text: Generate vector representations from natural language text.
- Format string: Clean and prepare strings using standard string formatting operations.
- Group by: Add an aggregated column after group by as a new column.
- Handle missing: Replace, drop, or add indicators for missing values.
- Handle outliers: Remove or replace outlier numeric and categorical values.
- Handle structured column: Process structured columns.

4. Export Data

Export data flow

Select the steps in the data flow that you want to export and then select an export option.

Source - sampled
S3: Titanic.csv

Data types
Transform: Titanic.csv

Steps (5)

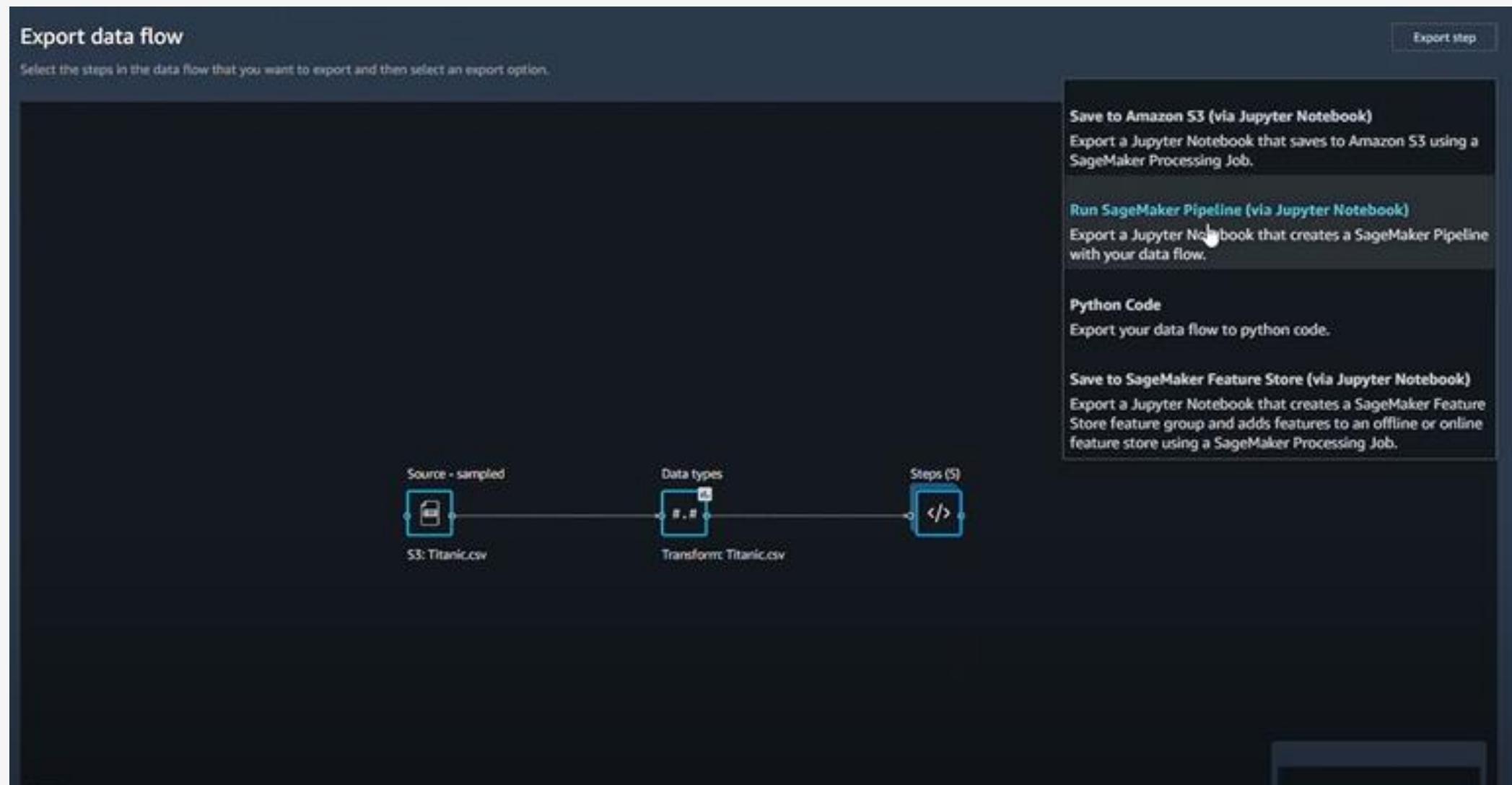
Export step

Save to Amazon S3 (via Jupyter Notebook)
Export a Jupyter Notebook that saves to Amazon S3 using a SageMaker Processing Job.

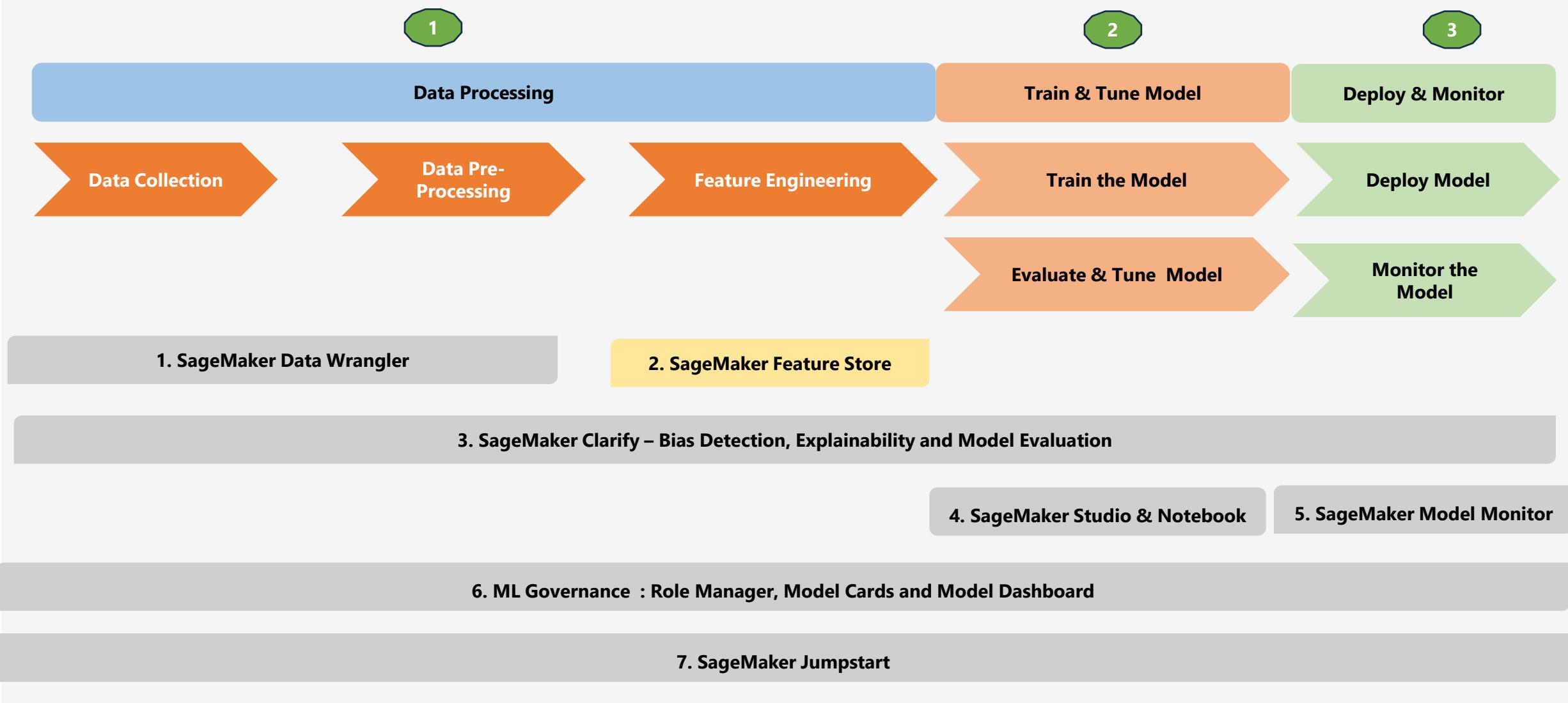
Run SageMaker Pipeline (via Jupyter Notebook)
Export a Jupyter Notebook that creates a SageMaker Pipeline with your data flow.

Python Code
Export your data flow to python code.

Save to SageMaker Feature Store (via Jupyter Notebook)
Export a Jupyter Notebook that creates a SageMaker Feature Store feature group and adds features to an offline or online feature store using a SageMaker Processing Job.



Machine Learning Development Lifecycle - Overview



What are Features ?

Age	Workclass	Final Weight	Education	Education-Num	Marital Status	Occupation	Relationship	Ethnic group	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
40	Self-emp-not-inc	243636	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	>50K
28	Private	102533	Some-college	10	Separated	Handlers-cleaners	Not-in-family	Black	Male	0	0	40	United-States	<=50K
27	Private	204515	Bachelors	13	Never-married	Tech-support	Not-in-family	White	Female	0	0	36	United-States	<=50K
48	Private	345831	HS-grad	9	Divorced	Machine-op-inspct	Not-in-family	White	Male	0	0	40	United-States	<=50K
48	Private	273828	5th-6th	3	Never-married	Handlers-cleaners	Not-in-family	White	Male	0	0	40	Mexico	<=50K

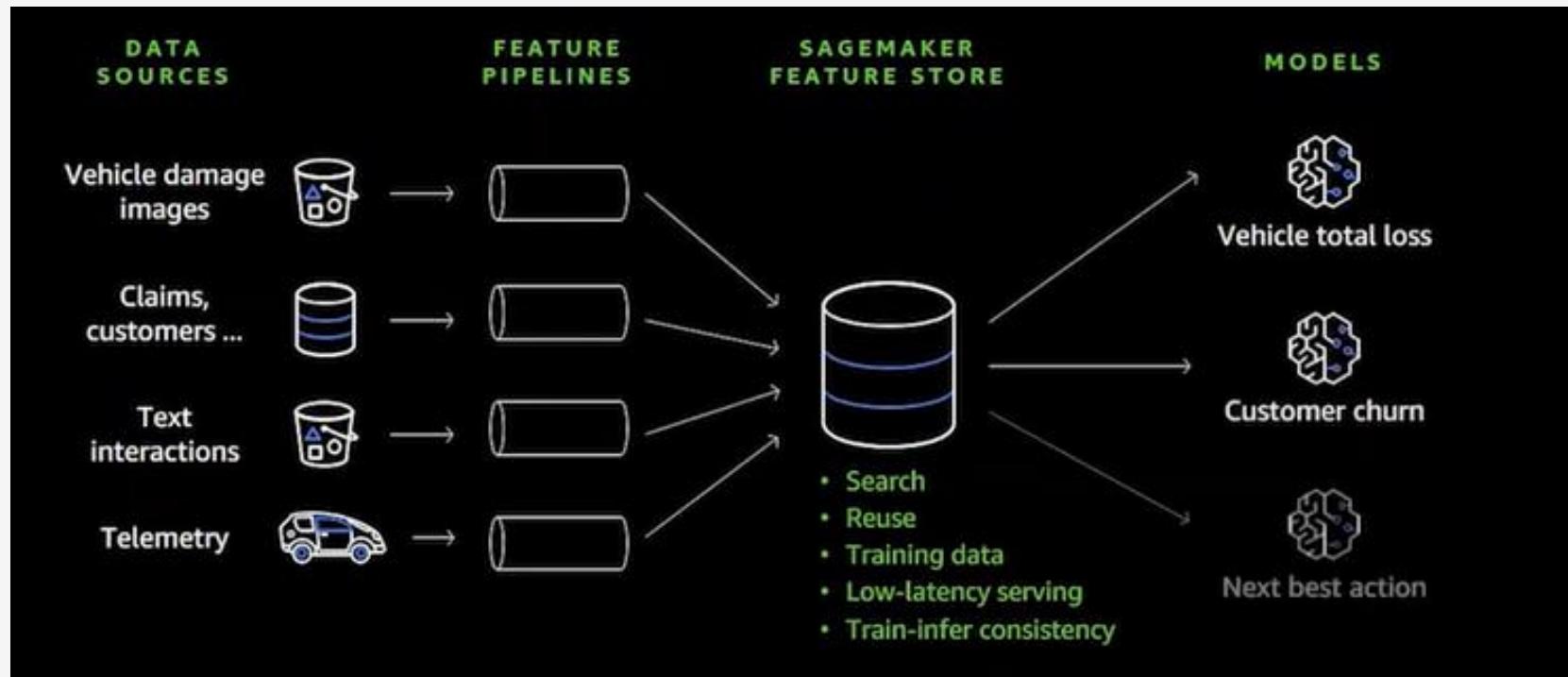
2. Amazon SageMaker Feature Store – Use Case

1

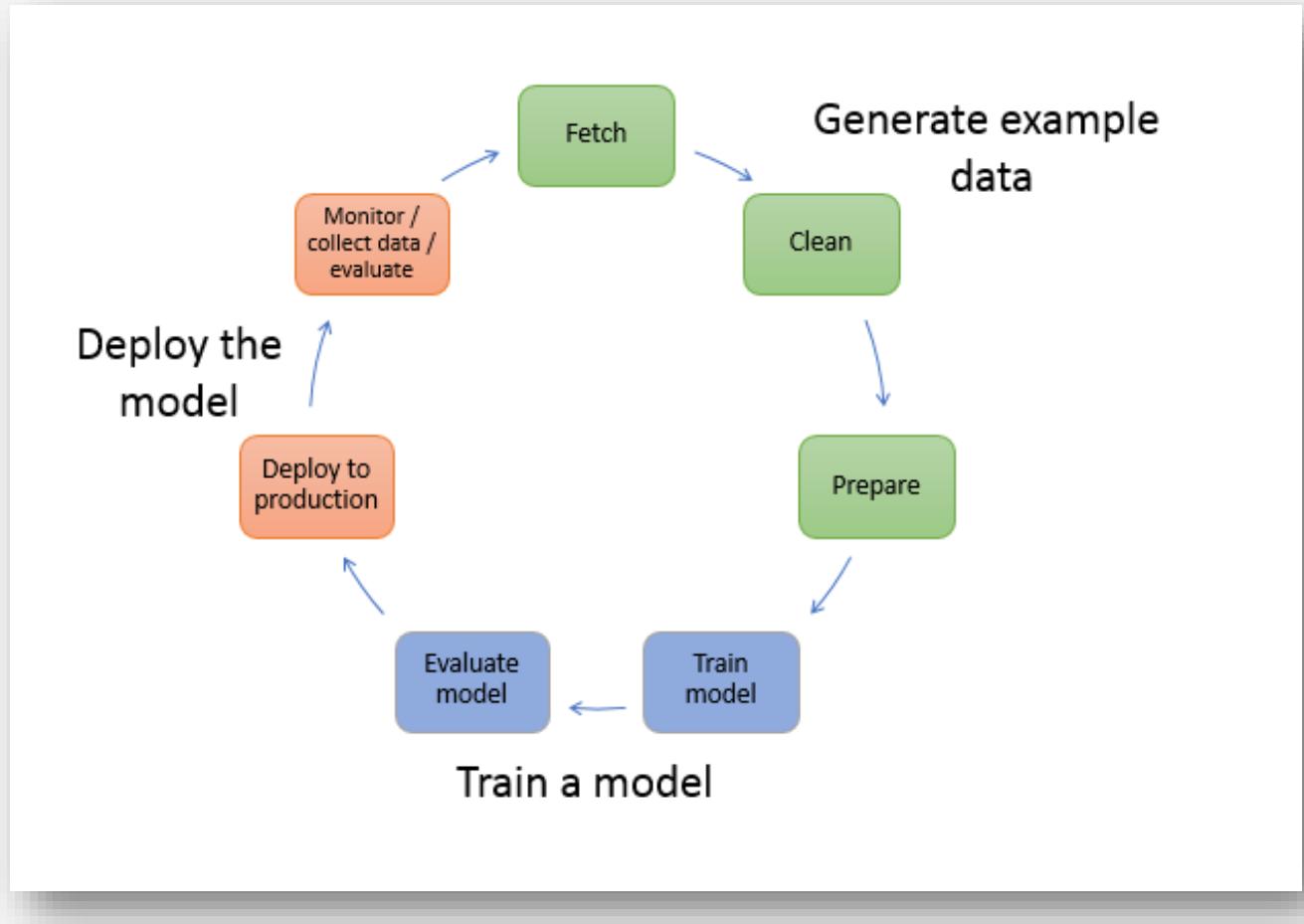
Data Processing - Amazon SageMaker Feature Store



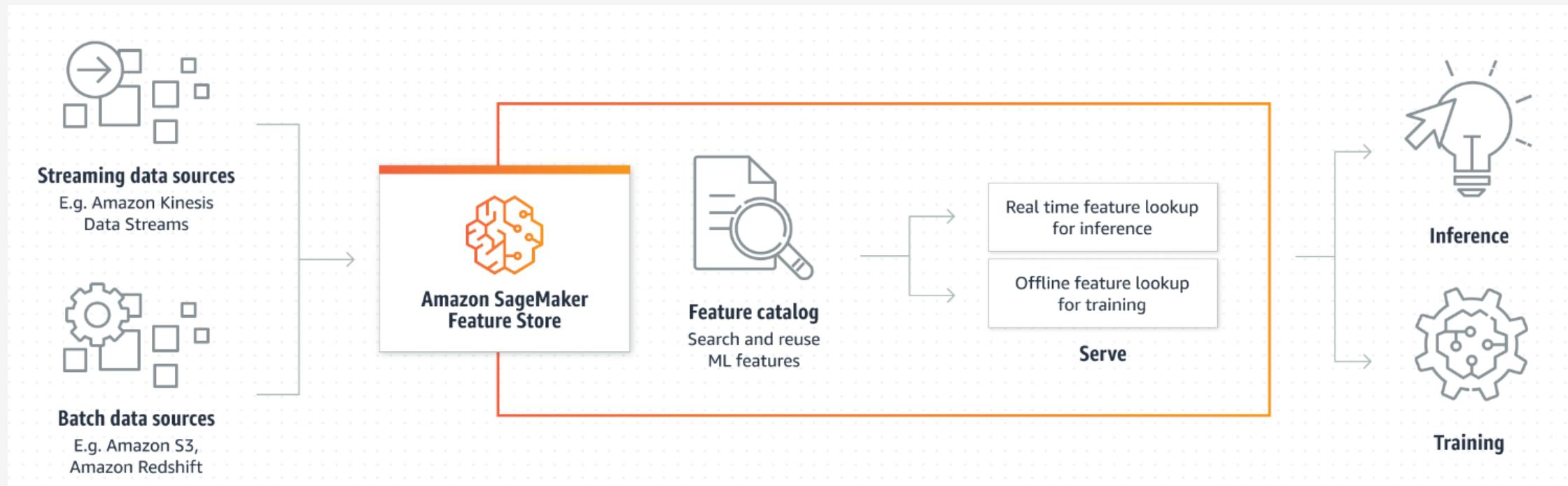
Feature Engineering



Machine Learning Lifecycle



Amazon SageMaker Feature Store – How it works ?



Amazon SageMaker Feature Store is a **fully managed, purpose-built repository to store, share, and manage features** for machine learning models.

Amazon SageMaker Feature Store - Capabilities

- **Data ingestion and Feature Processing**
 - Ingest data from various sources – Streaming or Batch Data Sources
 - Application and service logs, clickstreams, sensors, Amazon S3, Amazon Redshift, AWS Lake etc.
- **Feature storage, catalog, search, and reuse**
 - SageMaker Feature Store tags and indexes feature groups so they are easily discoverable
- **Online and Offline Feature**
 - Online Feature Store for real-time prediction
 - Offline Feature Store for Model training
 - For offline feature store – AWS Glue Data Catalog is automatically created
 - Query features search– Amazon Athena

Amazon SageMaker Feature Store

The screenshot shows the Amazon SageMaker Feature Store interface. On the left, there's a sidebar with "SageMaker resources" and a dropdown set to "Feature Store". Below it, a text block explains the feature store as a centralized repository for machine learning features, with links to "Create Feature Group" and "Open Feature store". The main area has tabs for "Feature group catalog" and "Feature catalog", with "Feature group catalog" selected. A search bar at the top says "Search by Feature name or Description or Parameters". A table below lists ten features:

Feature name	Feature group name	Type	Description	Parameters	Created on
has_kids	fs-customers-10-13-01-44	Integral			10/12/2022
customer_id	fs-customers-10-13-01-44	String			10/12/2022
sex	fs-customers-10-13-01-44	Integral			10/12/2022
age_60-69	fs-customers-10-13-01-44	Integral			10/12/2022
age_30-39	fs-customers-10-13-01-44	Integral			10/12/2022
age_70-plus	fs-customers-10-13-01-44	Integral			10/12/2022
event_time	fs-customers-10-13-01-44	Fractional			10/12/2022
is_married	fs-customers-10-13-01-44	Integral			10/12/2022
age_18-29	fs-customers-10-13-01-44	Integral			10/12/2022

Amazon SageMaker Feature Store

The screenshot shows the Amazon SageMaker Feature Store interface. On the left, a sidebar titled "SageMaker resources" lists "Feature Store" as the selected category. A detailed description of the feature store is provided, mentioning it's a centralized repository for machine learning features and includes options to create new feature groups. Below this are buttons for "Open Feature store" and "+ Create Feature Group". The main content area is titled "Feature store" and describes it as a repository for machine learning features. It shows a "Feature group catalog" tab is active, with a search bar and a "Create Feature Group" button. A table lists three feature groups: "fs-customers-10-13-01-44", "FG-titanic-aa375867", and "titanic-feature-group". The table columns include Feature group name, Description, Tags, Record identifier, Store type, and Offline store status. The "titanic-feature-group" entry has a note below it stating "Contains data about Tita...".

Feature group name	Description	Tags	Record identifier	Store type	Offline store status
fs-customers-10-13-01-44			customer_id	Online/Offline	Active
FG-titanic-aa375867			name	Online/Offline	
titanic-feature-group	Contains data about Tita...		name	Online/Offline	

Amazon SageMaker Feature Store

The screenshot shows the Amazon SageMaker Studio interface. The left sidebar has a dark theme with sections for "SageMaker resources" and "Feature Store". Under "Feature Store", there is a description of what it is and buttons for "Open Feature store" and "+ Create Feature Group". The main area is titled "Create feature group" and shows a progress bar with steps: "Launcher" (active), "01-sagemaker_feature_store.", "Feature store", and "Create feature group". Below the progress bar, the "Feature group details" tab is selected, showing fields for "Feature group name" (set to "my-demo-feature-group") and "Description (optional)" (set to "A demo group for testing"). The "FEATURE GROUP STORAGE CONFIGURATION" section includes options for "Enable storage online" (checked) and "Enable storage offline" (unchecked). An "Online store encryption key" dropdown is set to "Use AWS managed KMS key (default)".

Amazon SageMaker Feature Store

The screenshot shows the 'Create feature group' interface in the Amazon SageMaker console. On the left, a sidebar titled 'SageMaker resources' and 'Feature Store' provides an overview of the feature store's purpose and creation options. The main area contains several configuration fields:

- S3 bucket name:** sagemaker-offline-100923_us-east-2
- Dataset directory name:** Select...
- IAM Role ARN:** AmazonSageMaker-ExecutionRole-2021...
- Offline store encryption key:** Use AWS managed KMS key (default)

At the bottom, a section titled 'DATA CATALOG (OFFLINE STORE)' explains that an AWS Glue data catalog is automatically created for offline feature store data, allowing users to write and execute SQL queries using Amazon Athena.

Amazon SageMaker Feature Store

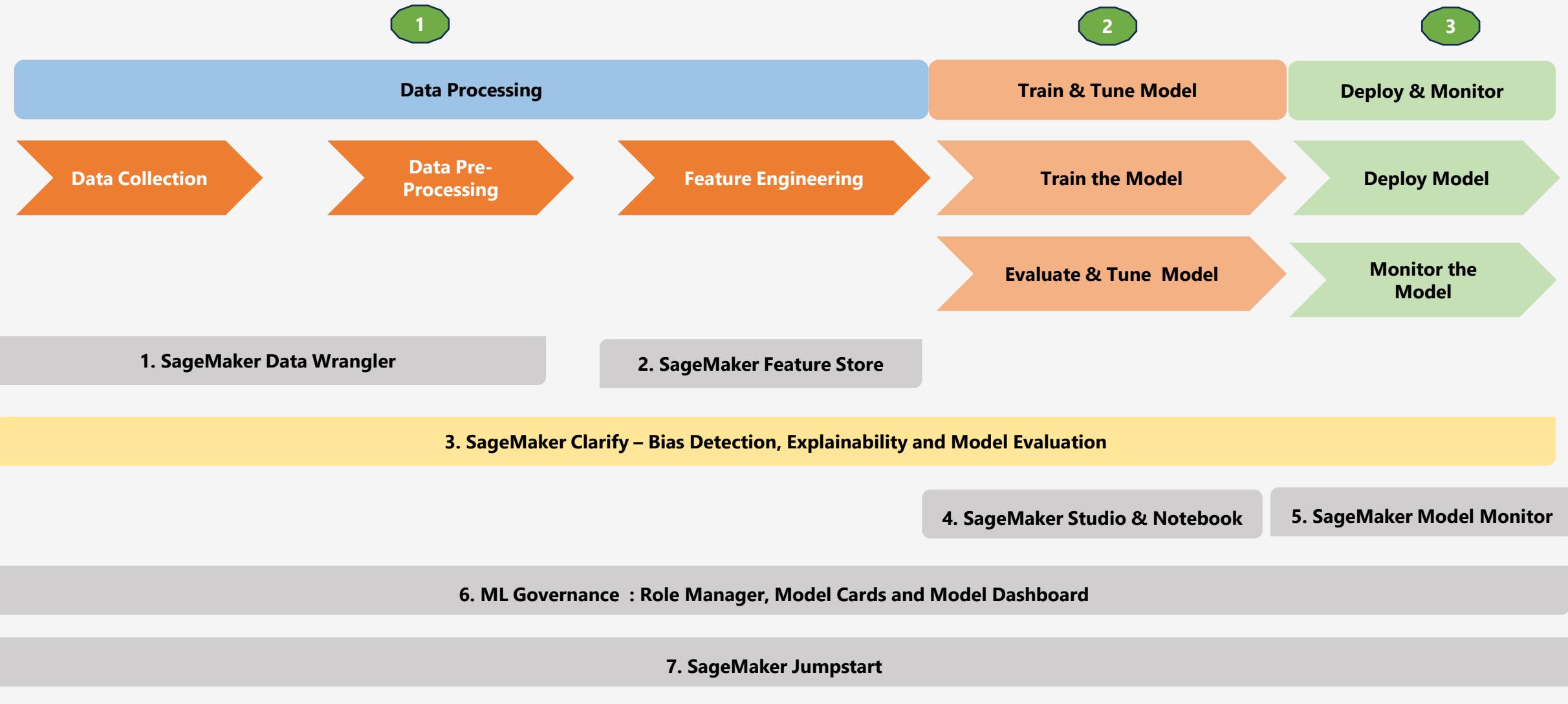
The screenshot shows a Jupyter Notebook interface with the following details:

- Header:** Includes icons for file operations, a refresh button, and tabs for "Code" and "git". On the right, it shows "2 vCPU + 4 GiB Cluster Python 3 (Data Science)".
- Cell 26:** Contains Python code for querying an Amazon SageMaker Feature Store using Athena. It defines a feature group, runs a SELECT query, and converts the results into a pandas DataFrame.
- Output of Cell 26:** Shows the command "Running SELECT * FROM "auto-mpg-2022-10-15-00-03-30-1665792246"".
- Cell 27:** Contains the command "dataset.head()", which displays the first four rows of the dataset.
- Output of Cell 27:** Displays a table with 14 columns: mpg, cylinders, displacement, horsepower, weight, acceleration, model_year, origin, car_name, event_time, write_time, and api_invocation_time. The data includes information about different car models like Chevrolet Concours, Dodge Colt M/M, Dodge Challenger SE, and Mazda GLC.

	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name	event_time	write_time	api_invocation_time
0	17.5	6	250.0	110.0	3520.0	16.4	77	1	chevrolet concours	220953600.0	2022-10-15 00:10:02.337	2022-10-15 00:05:03.000
1	33.5	4	98.0	83.0	2075.0	15.9	77	1	dodge colt m/m	220953600.0	2022-10-15 00:10:02.337	2022-10-15 00:05:03.000
2	15.0	8	383.0	170.0	3563.0	10.0	70	1	dodge challenger se	28800.0	2022-10-15 00:10:01.500	2022-10-15 00:04:56.000
3	46.6	4	86.0	65.0	2110.0	17.9	80	3	mazda glc	315561600.0	2022-10-15 00:10:02.369	2022-10-15 00:05:06.000

Source : Youtube

Machine Learning Development Lifecycle - Overview



3. Amazon SageMaker Clarify

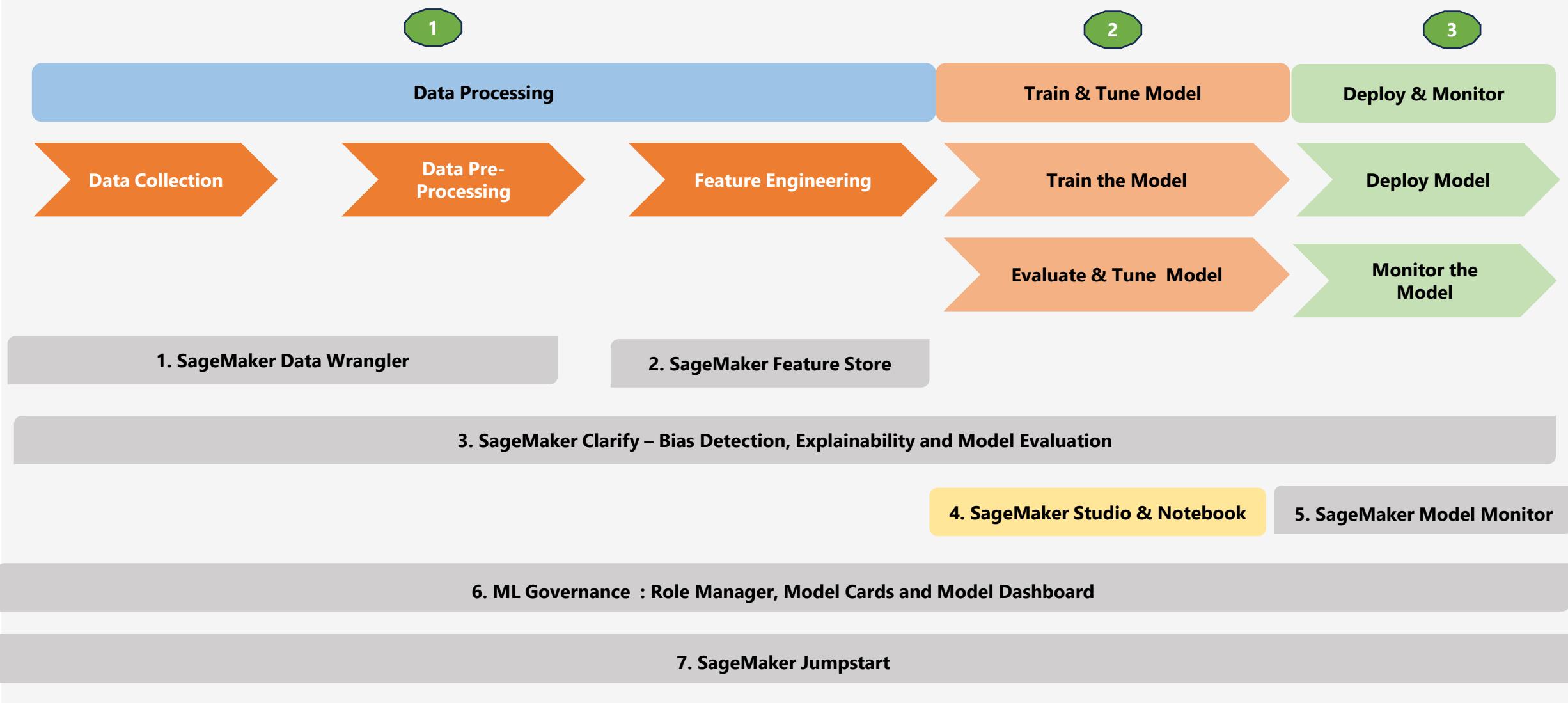
1. Model Evaluation (Model 1 vs Model 2 performance)

- Using Amazon SageMaker Clarify you can **evaluate large language models (LLMs)** by creating model evaluation jobs.
 - **Automated** model evaluation jobs in Studio
 - Model evaluation jobs that use **human workers** in Studio
 - Automated model evaluation job using the **fmeval library**
- **Tasks**
 - Text summarization
 - Question answering
 - Classification
 - Open-ended generation

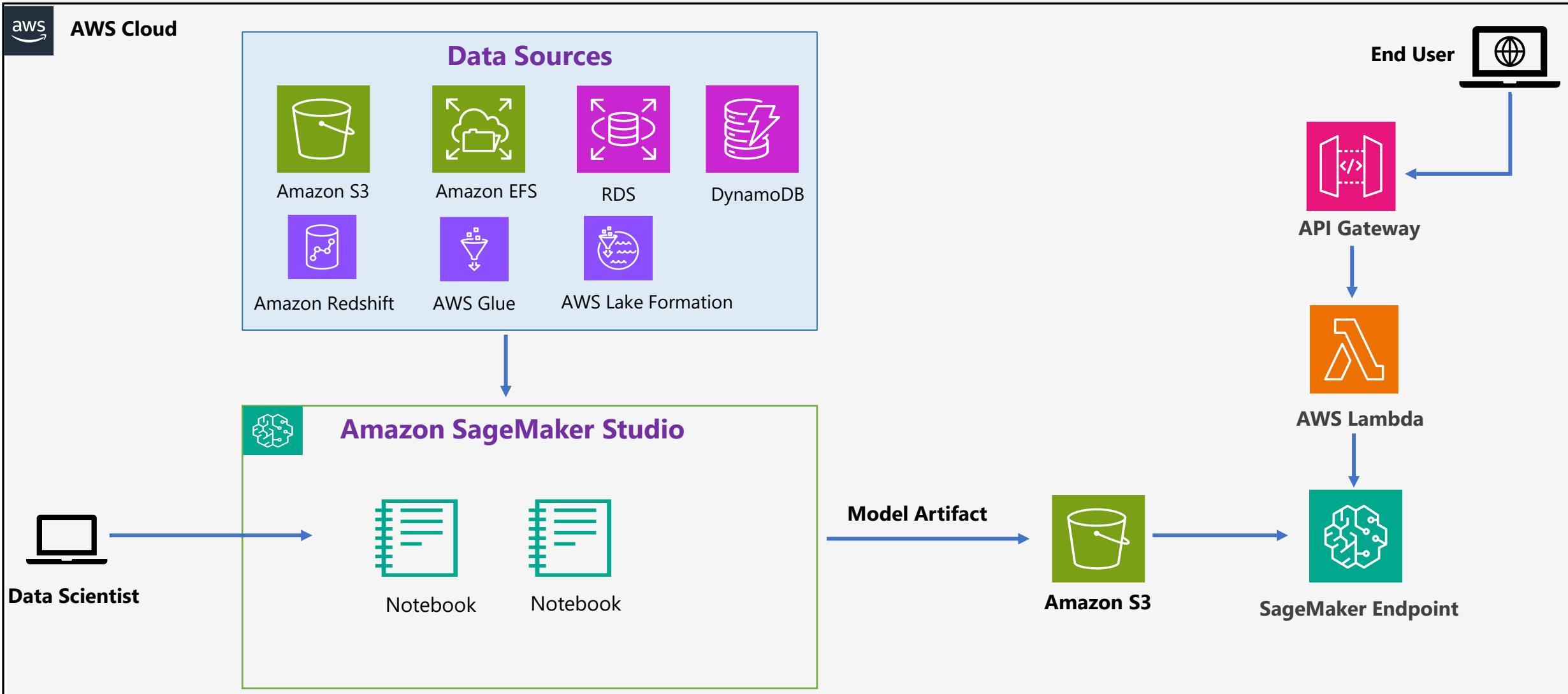
2. Bias Detection – Data and Model

3. Model Explainability

Machine Learning Development Lifecycle - Overview

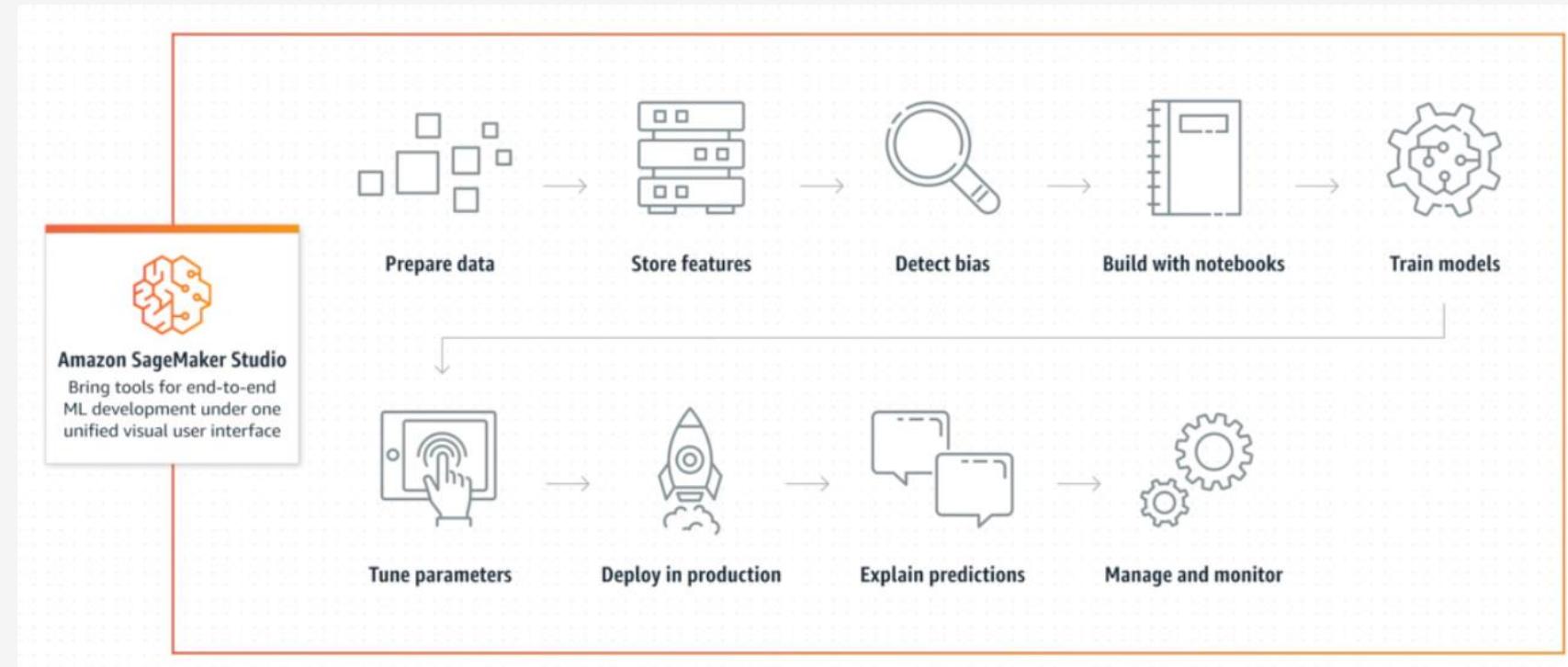


ML Development Lifecycle

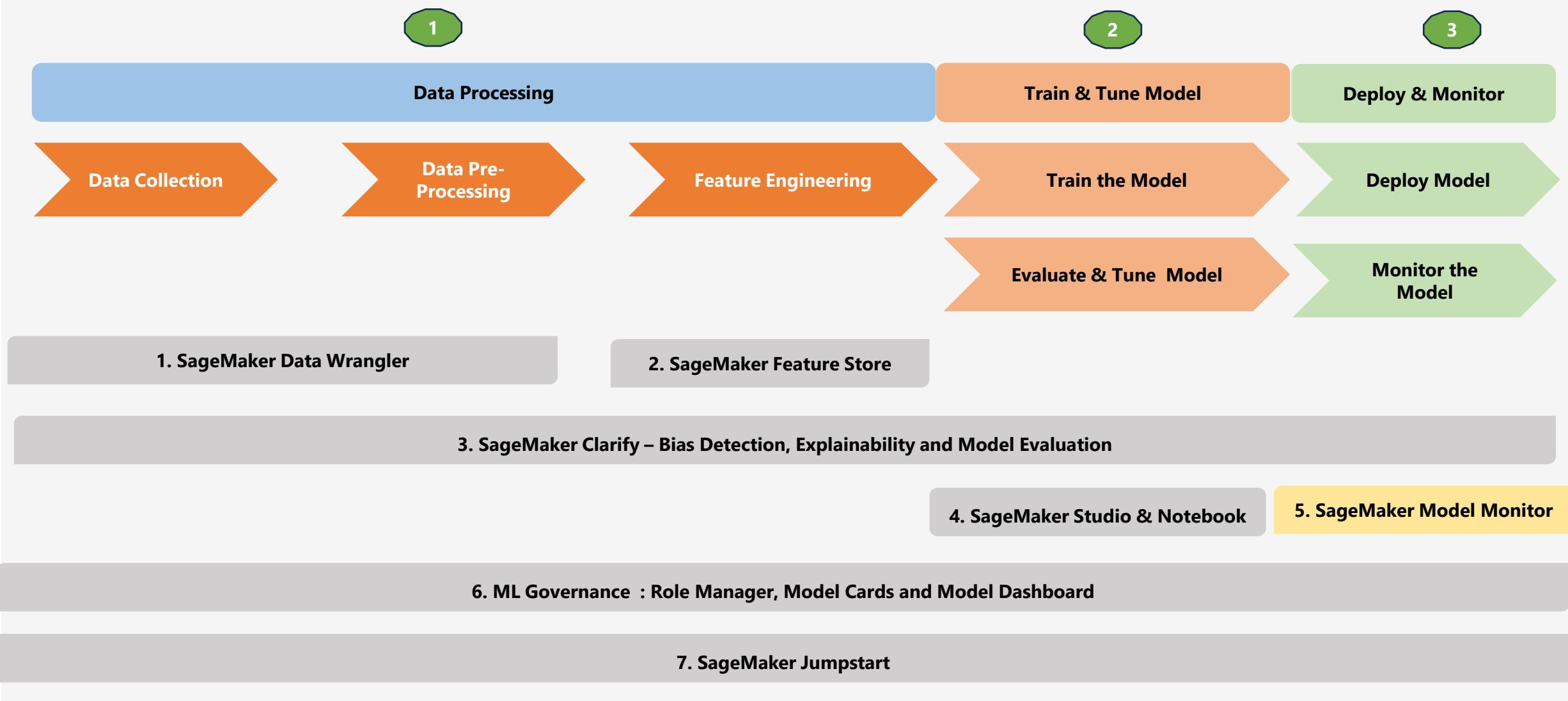


4. Amazon SageMaker Studio and Notebook

- Amazon **SageMaker Studio** is an **integrated development environment (IDE)** for machine learning that lets you **build, train, debug, deploy and monitor** your machine learning models.
- Amazon **SageMaker Studio notebooks** - fully managed Jupyter notebooks in SageMaker for ML development (on EC2)



Machine Learning Development Lifecycle - Overview



5. Amazon SageMaker Model Monitor

1. Data Quality – Monitoring & Alerting

- Monitors Statistical Properties against ground truth data : Missing Columns, Null Values

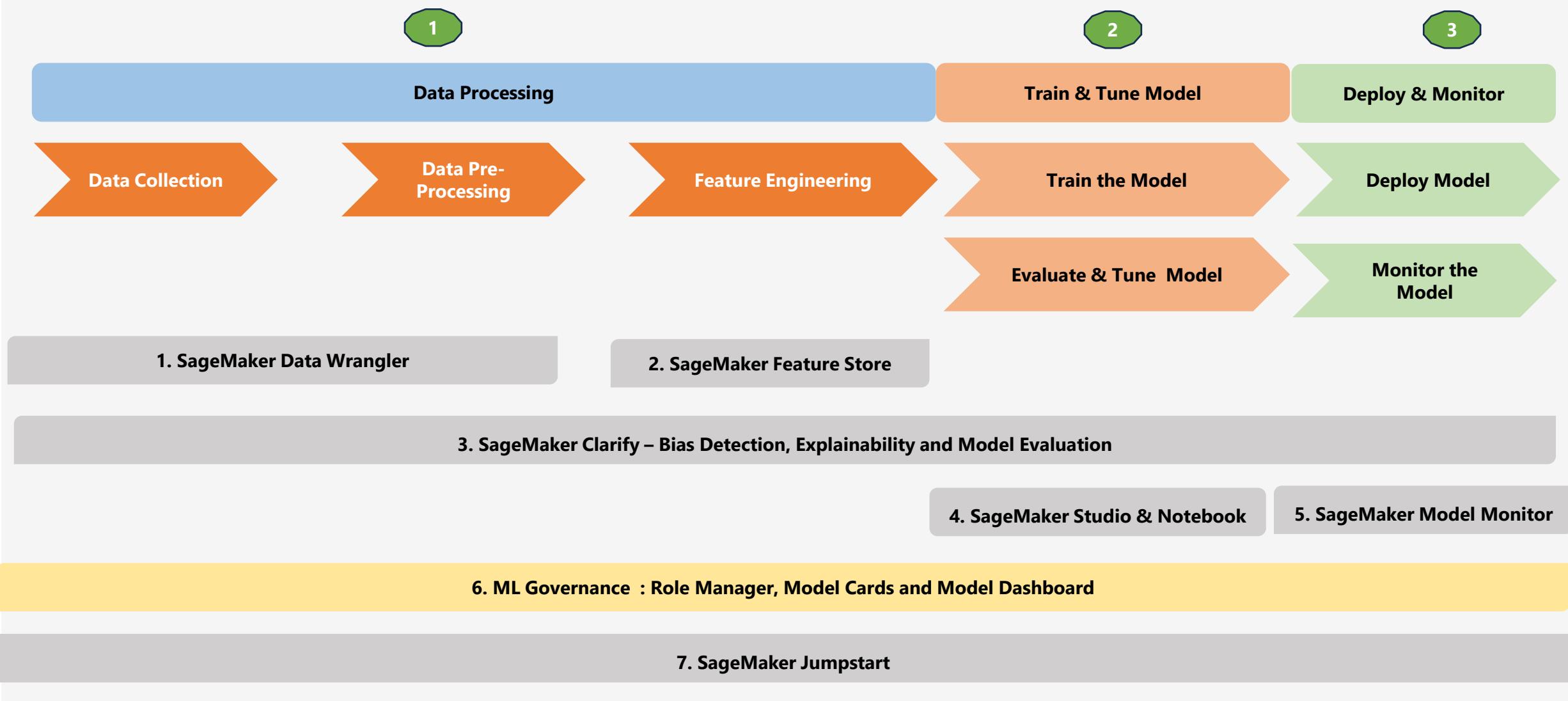
Age	Workclass	Final Weight	Education	Education-Num	Marital Status	Occupation	Relationship	Ethnic group	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
40	Self-emp-not-inc	243636	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	>50K
28	Private	102533	Some-college	10	Separated	Handlers-cleaners	Not-in-family	Black	Male	0	0	40	United-States	<=50K
27	Private	204515	Bachelors	13	Never-married	Tech-support	Not-in-family	White	Female	0	0	36	United-States	<=50K
48	Private	345831	HS-grad	9	Divorced	Machine-op-inspct	Not-in-family	White	Male	0	0	40	United-States	<=50K
48	Private	273828	5th-6th	3	Never-married	Handlers-cleaners	Not-in-family	White	Male	0	0	40	Mexico	<=50K

Amazon SageMaker Model Monitor

2. Model Monitoring & Alerting

- **Model Performance – Accuracy & Precision**

Machine Learning Development Lifecycle - Overview



6. ML Governance with Amazon SageMaker

Amazon SageMaker provides **governance tools** to help you **implement ML responsibly**

1. Amazon SageMaker Role Manager

- Role Manager allows **admins create preconfigured role personas** for common ML activities

The screenshot shows the 'Enter role information' step of the Amazon SageMaker Role Manager wizard. The left sidebar lists steps: Step 1 (current), Step 2, Step 3, and Step 4. The main area has a title 'Enter role information' and a subtitle 'A role is an IAM identity that has permissions to perform actions with AWS services.' Below this is a section titled 'Set up SageMaker role' with a 'Info' link. It contains fields for 'Role name suffix' (containing 'DataScientist, DataEngineer, MLOpsEngineer, etc.') and 'Description' (containing 'Enter description here'). A dropdown menu titled 'Select a persona' is open, showing three options: 'Data Scientist', 'MLOps', and 'SageMaker Compute Role'. Each option has a detailed description below it.

Persona	Description
Data Scientist	A persona that performs machine learning activities from within a sagemaker environment. Permitted to process S3 data, perform experiments and produce models.
MLOps	A persona that deals with operational activities from within a sagemaker environment. Permitted to manage models, endpoints and pipelines, and audit resources.
SageMaker Compute Role	A persona used by SageMaker compute resources such as jobs and endpoints. Permitted to access S3 resources, ECR repositories, Cloudwatch, and other services for ML computation.

6. ML Governance with Amazon SageMaker

2. Amazon SageMaker Model Cards

- Allows **capture, retrieve, and share model information**, such as model uses, risk ratings, training details, from conception to deployment.

Amazon SageMaker > Model cards > sentiment-analysis-model-card

Model card - sentiment-analysis-model-card

Model card overview

Model card version	4	KMS encryption key	-
Model card status	Draft	Model card ARN	arn:aws:sagemaker:us-east-2:██████████:model-card/sentiment-analysis-model-card
Created date	11/14/2022, 10:17:18 PM		

Model overview

Model name	Sentiment-Analysis-Model	Inference environment	██████████dkr.ecr.us-east-2.amazonaws.com/sagemaker-xgboost:1.3-1
Model description	the model is updated.	Problem type	Binary Classification
Model versions	-	Algorithm type	Logistic Regression
Model arn	arn:aws:sagemaker:us-east-2:██████████:model/sentiment-analysis-model	Model creator	DEMO-user

6. ML Governance with Amazon SageMaker

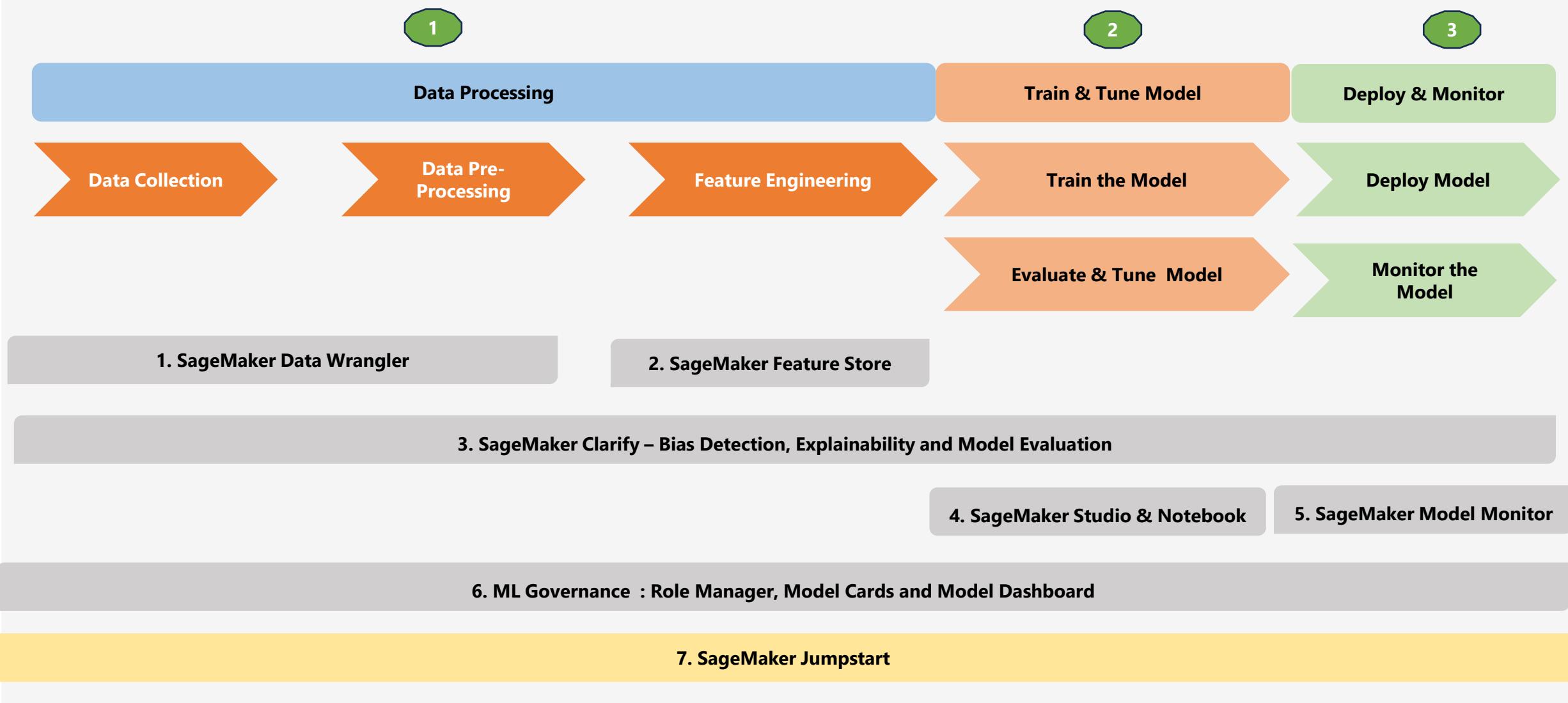
3. Amazon SageMaker Model Dashboard

- The Model dashboard is a **centralized repository of all models** that have been created in the account.

The screenshot shows the Amazon SageMaker Model dashboard. On the left, there's a sidebar with links like Getting started, Control panel (Studio, Studio Lab, Canvas, RStudio), SageMaker dashboard (Images, Lifecycle configurations, Search), Governance (Model dashboard NEW, Model cards NEW), Ground Truth, and Notebook. The main area is titled 'Model dashboard' and shows a table of models. The table has columns for Model Name, Risk Rating, Model Quality, Data Quality, Bias Drift, Feature Attribution Drift, and Endpoints. Each row contains a model name, its risk rating (Low, High, Medium), data quality status (Nov 21, 2022 19:03 UTC), bias drift status (Nov 21, 2022 19:13 UTC), feature attribution drift status (Nov 21, 2022 19:07 UTC), and its endpoint name. The 'Model dashboard' link in the sidebar is highlighted with a blue border.

Model Name	Risk Rating	Model Quality	Data Quality	Bias Drift	Feature Attribution Drift	Endpoints
Sentiment-Analysis-Model	Low	-	⚠️ Nov 21, 2022 19:03 UTC	-	-	Sentiment-Analysis-Model-Endpoint
Customer-Churn-Model	High	⚠️ Nov 21, 2022 19:13 UTC	⚠️ Nov 21, 2022 19:07 UTC	⊖ Inactive	⌚ Scheduled	Customer-Churn-Model-Endpoint
Loan-Approval-Model	High	-	⚠️ Nov 21, 2022 19:06 UTC	-	-	Loan-Approval-Model-Endpoint
Product-Recommendation-Model	High	-	⚠️ Nov 21, 2022 19:01 UTC	-	-	Product-Recommendation-Model-Endpoint
Fraud-Detection-Model	Medium	-	⚠️ Nov 21, 2022 19:03 UTC	-	-	Fraud-Detection-Model-Endpoint

Machine Learning Development Lifecycle - Overview



Amazon SageMaker Jumpstart

AWS Service Description

Sagemaker Jumpstart

- **Fully managed Service** allows access to different models in area of :
 - Foundation Models
 - Computer Vision Models
 - NLP Models
 - Model Evaluation using Amazon Sagemaker Clarify
 - Build and train your own models (with Sagemaker Service)

2. Use Cases

- **Text Summarization, Code Generation** and any other Gen AI use case
- Allows you to **provision infrastructure to deploy and fine tune foundation models**
 - Open source models
 - Proprietary models
- **Users of this service** – Machine Learning Engineers, Data Scientists, Developers

Discover foundation models and deploy with SageMaker's enterprise-ready features

AVAILABLE ON SAGEMAKER JUMPSTART

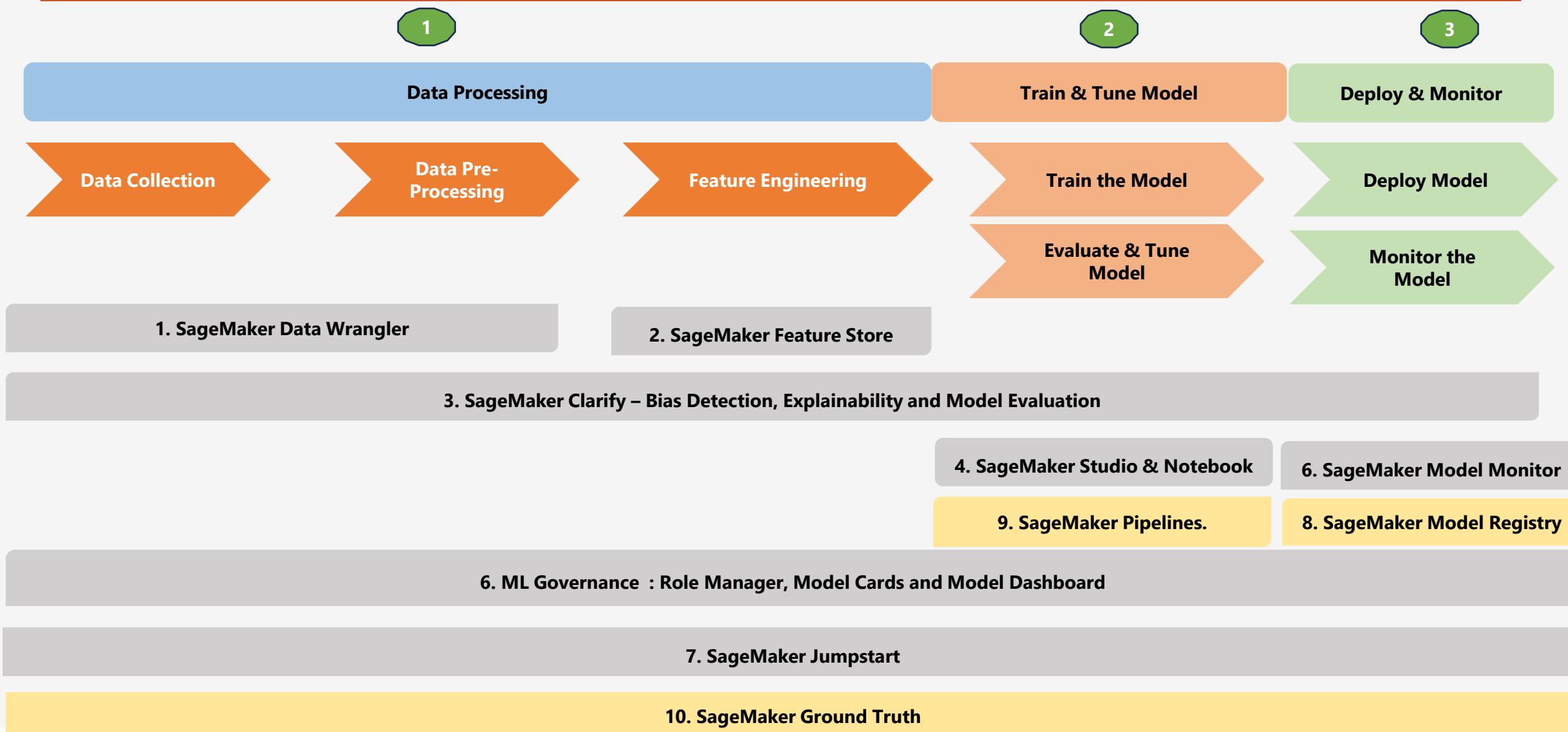
AI21labs	Meta AI	cohere	Hugging Face	stability.ai	LightOn	databricks	alexa
Models Jurasic-2 Ultra, Mid Contextual answers Summarize Paraphrase Grammatical error correction Tasks Text generation Long-form generation Summarization Paraphrasing Chat Information extraction	Models Llama 2 7B, 13B, 70B Code Llama 7B, 13B, 34B Open LLaMA Tasks Question answering Chat Summarization Paraphrasing Sentiment analysis Text generation	Models Command Cohere Light Tasks Text generation Information extraction Question answering Summarization	Models Falcon-7B, 40B, 160B, Mstral-7B RedPajama MPT-7B Bloomic-176B Pan-T5 DistilGPT2 GPT NeoXT Bloom Tasks Upscaling Imprinting Tasks Generate photo-realistic images from text input Improve quality of generated images Tasks Machine translation Question answering Summarization	Models Stable Diffusion XL 1.0 2.1 base Upscaling Imprinting Tasks Text generation Image generation Generate photo-realistic images from text input Improve quality of generated images Tasks Machine translation Question answering Summarization Sentiment analysis Classification	Models Lyra-Fr 10B, Meni Tasks Text generation Keyword extraction Information extraction Question answering Summarization Sentiment analysis Classification	Models Dolly Tasks Question answering Chat Summarization Paraphrasing Sentiment analysis Text generation	Models AlexaTM 20B Tasks Machine translation Question answering Summarization Annotation Data generation

Amazon SageMaker

SageMaker Jumpstart

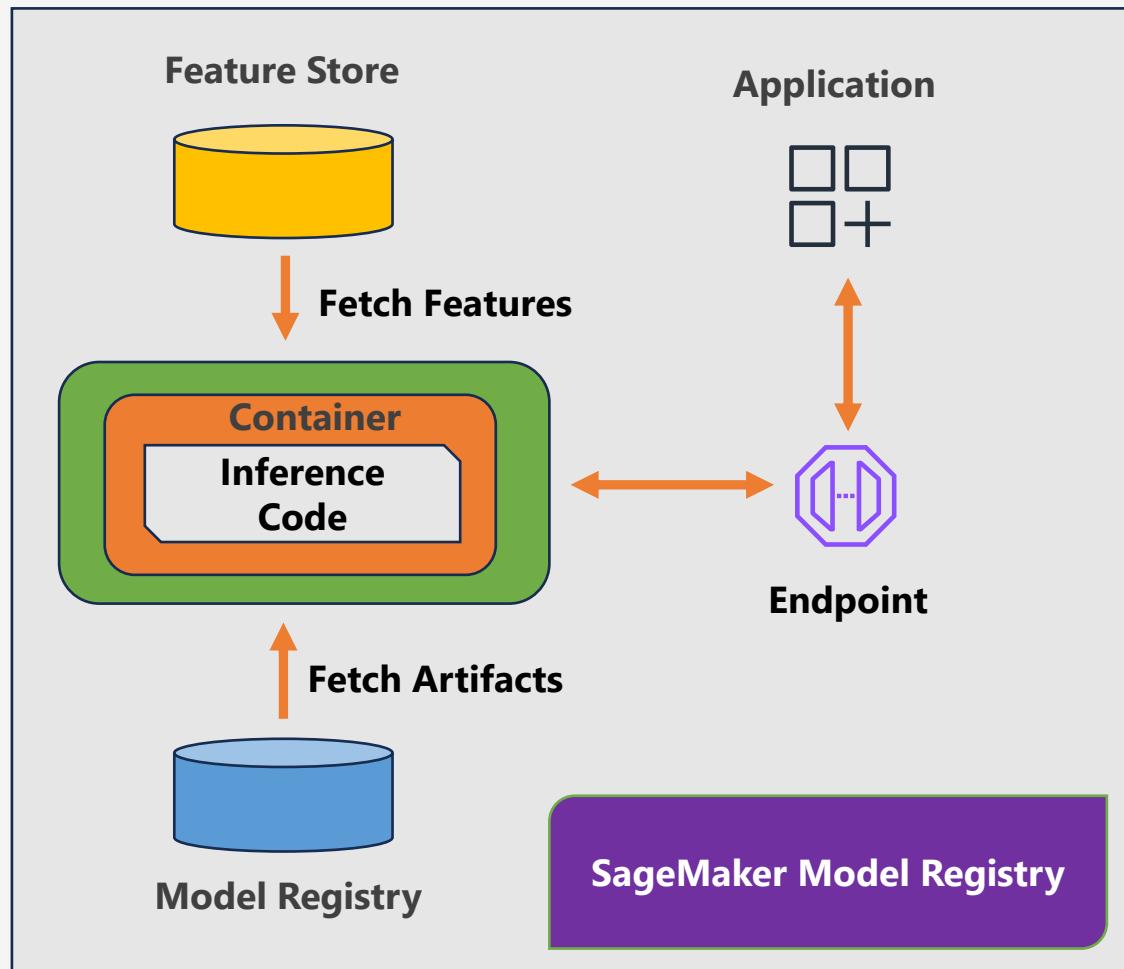
(GenAI focused)

Machine Learning Development Lifecycle - Overview

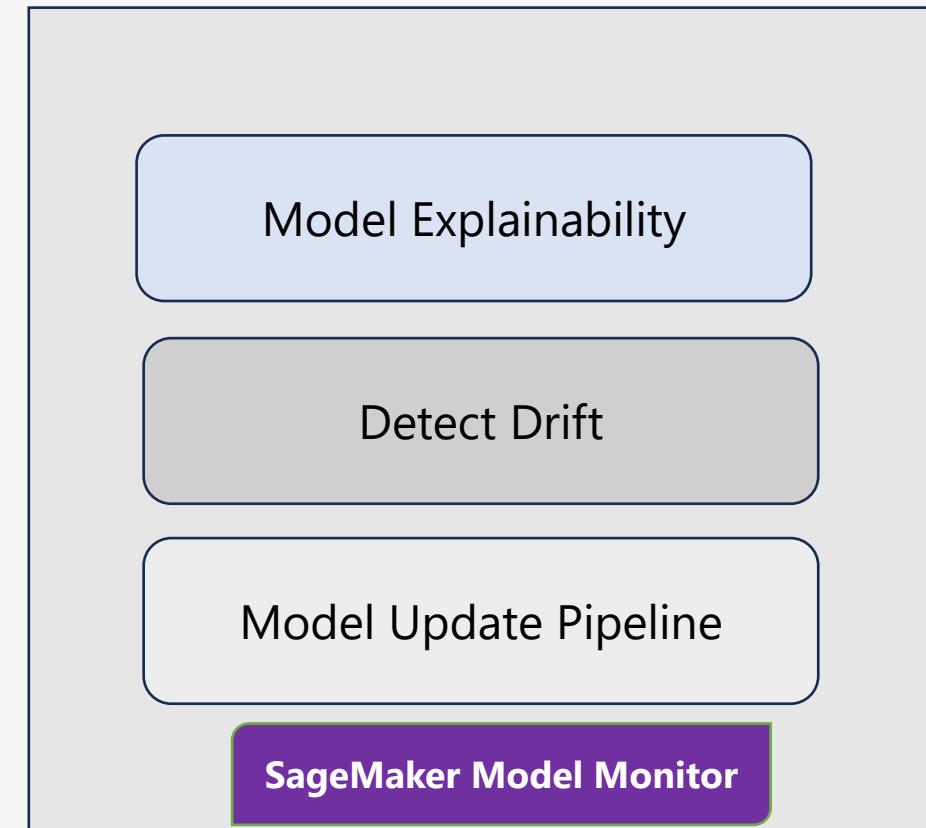


ML Development Lifecycle – Deployment and Monitoring

Deployment



Monitoring



SageMaker Model Registry

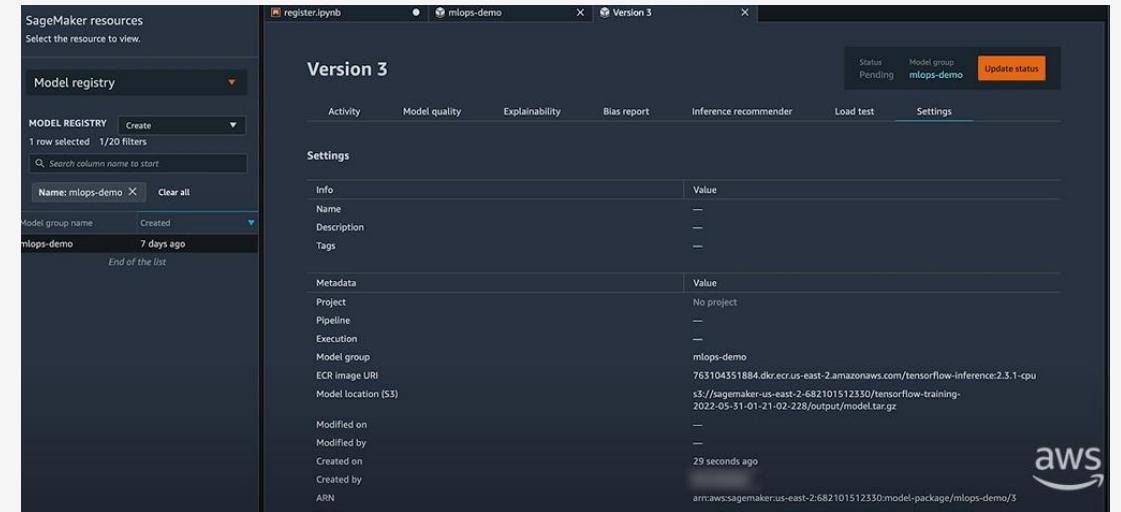
Key Features:

- **Model Versioning**

- Track and manage different versions of a model.

- **Metadata Management :**

- Store extensive metadata for each model, including training data, hyperparameters, and evaluation metrics.
- Manage the approval status of a model.
- Deploy models to production.



SageMaker Model Registry

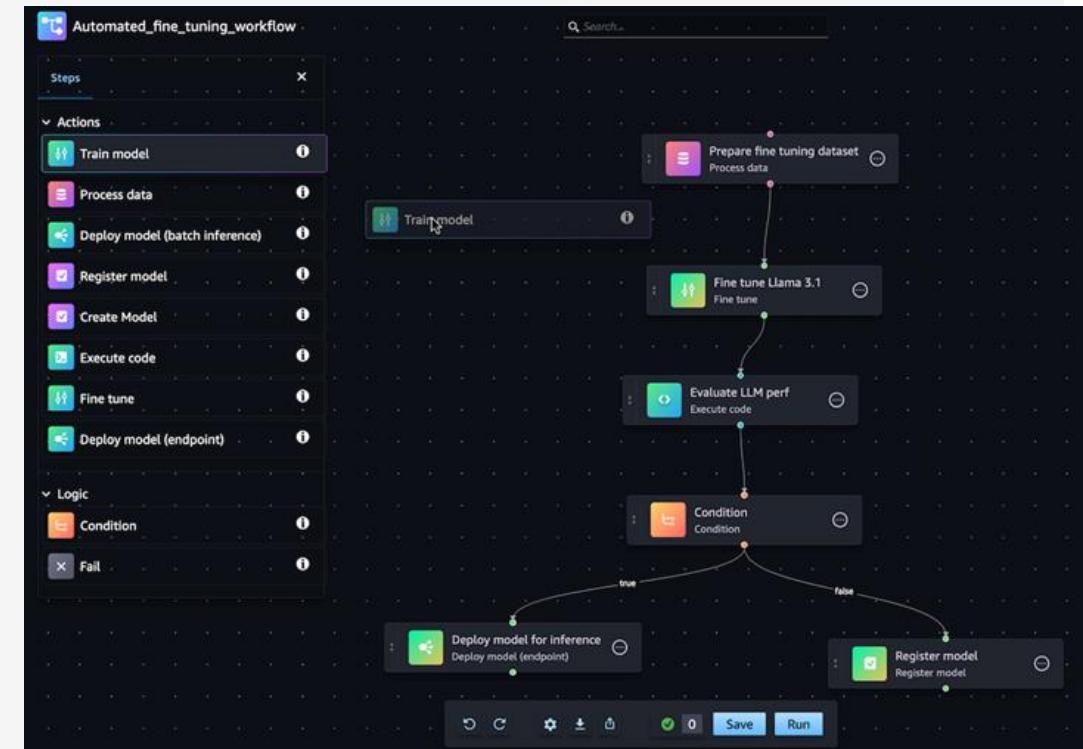
The screenshot shows the AWS SageMaker Model Registry interface. On the left, there's a sidebar titled "SageMaker resources" with a dropdown menu set to "Model registry". Under "MODEL REGISTRY", it shows a table with one row selected: "Name: mlops-demo" (Created 7 days ago). The main area is titled "Version 3" and contains tabs for Activity, Model quality, Explainability, Bias report, Inference recommender, Load test, and Settings. The Settings tab is active. It displays two tables: "Info" and "Metadata". The "Info" table has columns for Name, Description, and Tags, all showing blank values. The "Metadata" table lists various parameters with their values:

Metadata	Value
Project	No project
Pipeline	—
Execution	—
Model group	mlops-demo
ECR image URI	763104351884.dkr.ecr.us-east-2.amazonaws.com/tensorflow-inference:2.3.1-cpu
Model location (S3)	s3://sagemaker-us-east-2-682101512330/tensorflow-training-2022-05-31-01-21-02-228/output/model.tar.gz
Modified on	—
Modified by	—
Created on	29 seconds ago
Created by	—
ARN	arn:aws:sagemaker:us-east-2:682101512330:model-package/mlops-demo/3

In the bottom right corner of the main area, there is an "aws" logo.

SageMaker Pipelines

- SageMaker Pipelines allows you create, automate, and manage end-to-end ML workflows at scale.
- This includes data preparation, model training, and deployment
- A pipeline is defined using a **series of steps**, each representing a **specific task** in the ML workflow.
- SageMaker supports various step types, including:
 - **Processing Steps** : For data preprocessing and feature engineering.
 - **Training Steps**: To train your models using different algorithms and hyperparameters.
 - **Tuning** – Hyperparameter Tuning
 - **Model Evaluation Steps** : To assess the performance of your trained models.
 - **Model Registration Steps** : To register your models in the SageMaker Model Registry for version control and deployment.



- Link : <https://docs.aws.amazon.com/sagemaker/latest/dg/build-and-manage-steps.html>

SageMaker Ground Truth - RLHF

- Apply human feedback across the ML lifecycle to create or evaluate high-quality models
- SageMaker Ground Truth can be used for
 - Data generation and annotation to train the model
 - Evaluation to compare and select the foundation model
 - Model review, customization, and evaluation
- Executed either through self-service (Private, Amazon Mechanical Turk) or an AWS-managed offering - expert, on-demand workforce
- **Exam TIP : Reinforcement learning from human feedback (RLHF)** - is a machine learning technique that uses human feedback to optimize ML models to self-learn more efficiently (including LLM's).
- Link : <https://aws.amazon.com/sagemaker/groundtruth/features/?nc=sn&loc=2>

AWS AI-ML Services



AWS services in the NLP – Amazon Polly

Amazon Polly - Turns text into lifelike speech using deep learning

Text



Amazon Polly



Speech

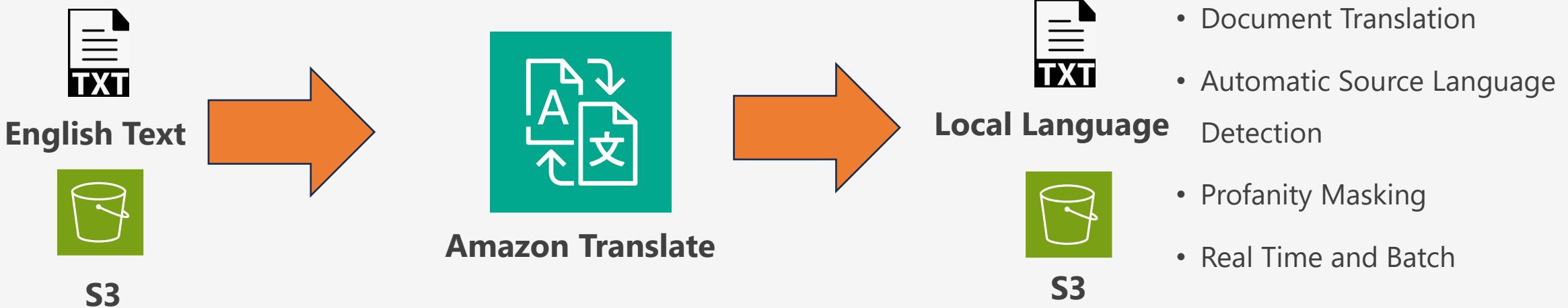


Common Use cases for Polly : - [Link](#)

- eLearning - Lectures
- Content Creation – RSS Feed
- Telephony

AWS AI Service (NLP) – Amazon Translate

Amazon Translate - AWS Language Translation service allows translation across 75 different languages.



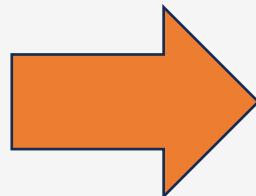
Amazon Translate – Uses ML models for **language translation**

- Translate user generated content – RSS Feed, Social Media
- Real-time translation of chat, email, helpdesk, and ticketing apps by English speaking agent
- Sentiments toward your product or service in multiple languages (NLP)

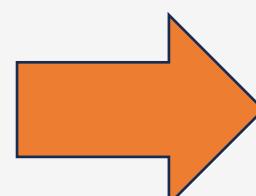
AWS services in the NLP – Amazon Transcribe

Amazon Transcribe - Convert speech to text

Speech



Amazon Transcribe



Text

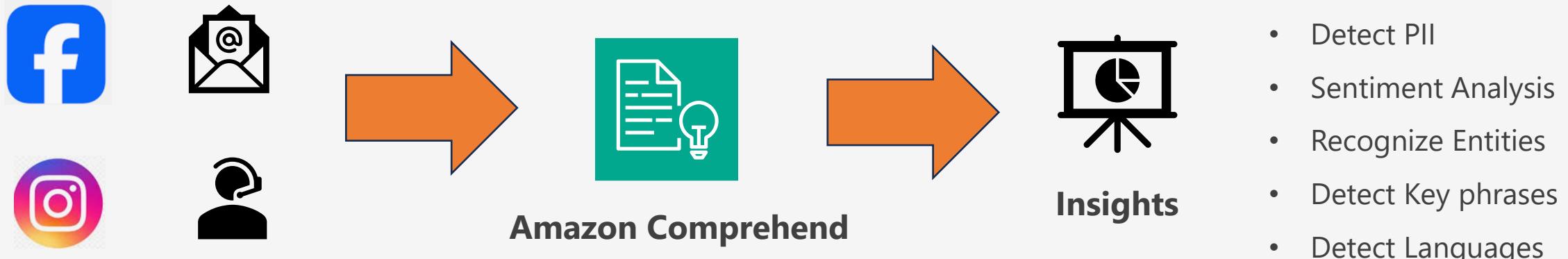


Common Use cases for Polly :

- Medical Transcription
- Call center analytics and agent assist
- Detect toxic content in audio for online gaming, social media or peer to peer

AWS AI Services (NLP) – Amazon Comprehend

Amazon Comprehend - Derives and understand valuable insights from text within documents, emails, web pages



Amazon Comprehend – Natural-language processing (NLP) service to uncover insights in text

- Call center analytics, Customer Support Tickets
- Secure documents – Redacting PII from documents
- Process financial documents - Classify and extract entities from Insurance Claims

Hyperparameters and Hyperparameter Tuning

Hyperparameters

- **Adjustable parameters** that let you control the **model training** process.
- For example, with **neural networks**, you decide the **number of hidden layers and the number of nodes** in each layer.
- Model performance depends heavily on hyperparameters.
- Set before model training begins
- They're different from **model parameters**, which are internal parameters automatically derived during the learning process and not set by data scientists.

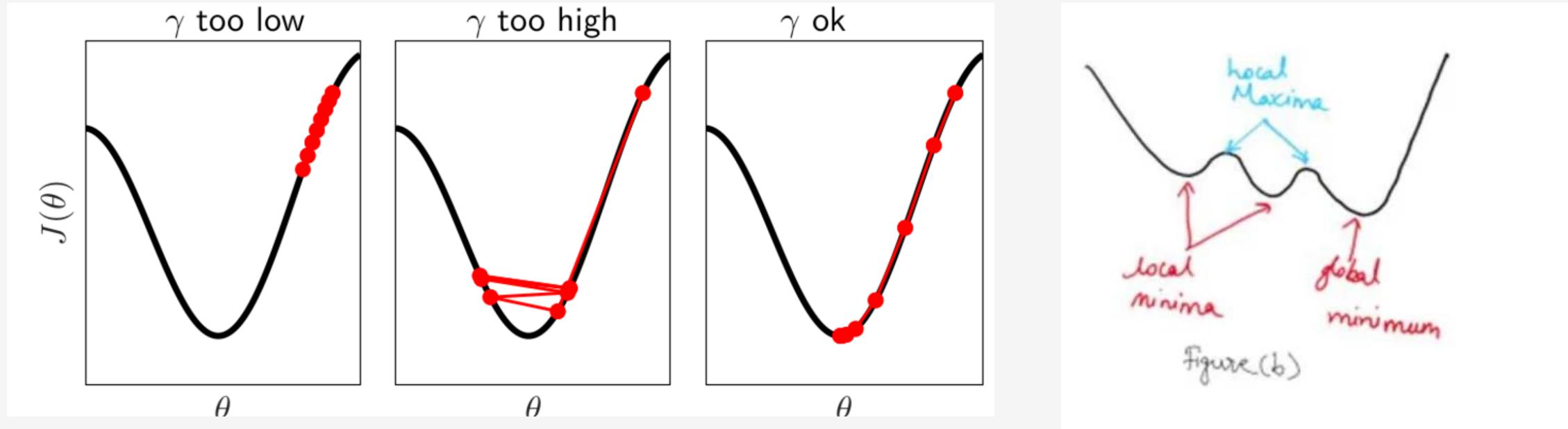
Hyperparameter tuning or hyperparameter optimization

- Process of finding the configuration of hyperparameters that results in the best performance.

Common Hyperparameter in Neural Networks

- **Hidden Layers** : Number of hidden layers in a neural network
- **Number of nodes** in each hidden layer
- **Learning rate**: Model parameters are adjusted iteratively — and the learning rate controls the size of the adjustment at each step.
- Selecting the optimized learning is important because if the learning rate is very less, then it may slow down the training process.
- If the learning rate is too large, then it may not optimize the model properly (Overshoot correct solution).

Hyperparameters



<https://aws.amazon.com/what-is/hyperparameter-tuning/#:~:text=Examples%20of%20hyperparameters%20include%20the,learning%20rate%2C%20and%20model%20complexity.>

Hyperparameters

Terminology	Description	Example
Sample	One row of data	1
Dataset	Total number of training samples	100
Batch Size	<ul style="list-style-type: none">• Total number of training samples in each mini-batch• Small batch size tends to get stuck in local minima• Large batch size can converge on the wrong solution	20
Iterations	Number of times Batch Size is run for entire dataset	5
Epoch	<ul style="list-style-type: none">• Represents the number of times the entire training dataset is passed through the model during training.• Epoch elapses when an entire dataset is run• Too few epochs, underfitting• Too many epochs, over fitting	1-Infinity

Hyperparameter tuning techniques

Grid search

- With grid search, **specify a list of hyperparameters and a performance metric**
- **Algorithm works** through all **possible combinations** to determine the **best fit**.
- Grid search works well, but it's relatively **tedious and computationally intensive**, especially with large numbers of hyperparameters.

Random search

- Based on similar principles as grid search, random search **selects groups of hyperparameters randomly** on each iteration.
- It works well when a **relatively small number of the hyperparameters**

Hyperparameter tuning techniques

Bayesian optimization

- Bayesian optimization is a technique based on Bayes' theorem
- When this is applied to **hyperparameter optimization**, the **algorithm builds a probabilistic model** from a set of **hyperparameters** that **optimizes a specific metric**.
- Uses **regression analysis** to iteratively choose the **best set of hyperparameters**.

Amazon SageMaker automatic model tuning (AMT)

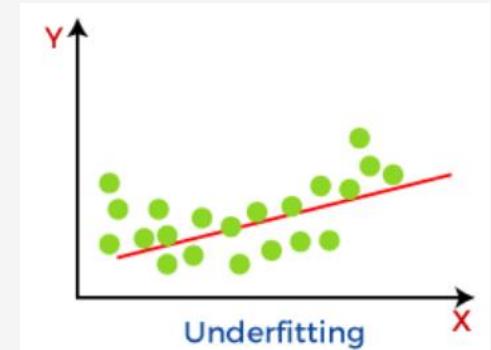
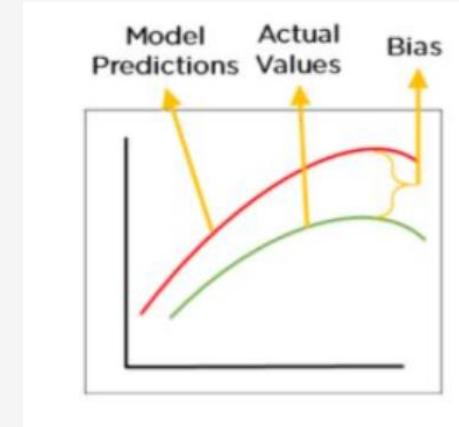
- AMT uses the algorithm and ranges of hyperparameters that you specify.
- It then chooses the hyperparameter values that creates a model that performs the best, measured by a metric that you choose.

Bias and Underfitting

- **Bias** is the difference between actual and predicted values by the model.
- **High Bias** - Model with a high bias makes more assumptions, and it is unable to capture the important features of our dataset.
- Model has poor performance both on the training and testing data means high bias
- This instance, where the model cannot find patterns in training set, is called **Underfitting**.
- **Underfitting occurs when** Model is too **simple** to capture variations and patterns in your data

Reduce Bias

- Use complex model
- Increase the number of features



Variance

- Variance is a measurement value used to find how the data is spread concerning the mean value of the data set.
- **High variance** is often a cause of **overfitting**
- Captures most patterns in the data, but also learn from the noise.
- Model will perform really well on testing data but will fail to perform on new, unseen data.
- **Overfitting occurs when Model is too complex**

Reduce Variance

- Feature Selection (Important Features)
- Split data into training and test data sets multiple times

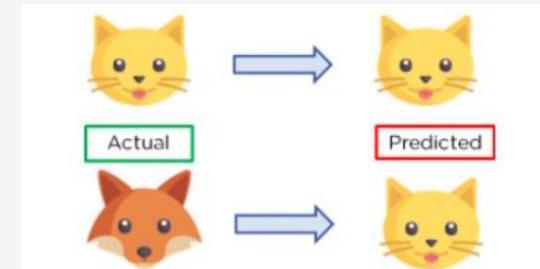
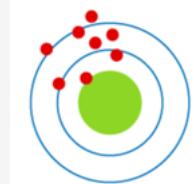


Figure 4: Example of Variance

Bias-Variance Trade-off

Low-Bias, High-Variance:

- Model predictions are inconsistent and accurate on average.
- Occurs when the model learns with a large number of parameters - leads to an **overfitting**

High-Bias, Low-Variance:

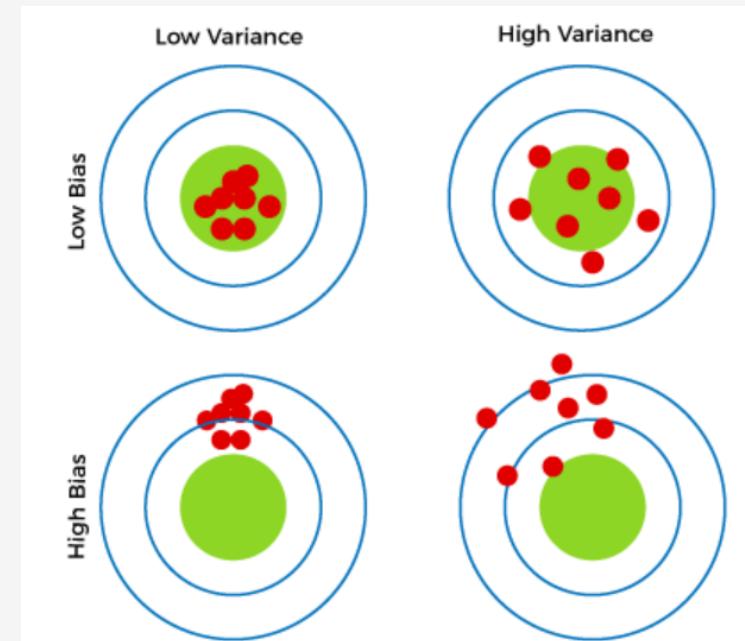
- Predictions are consistent but inaccurate on average.
- Occurs when a model does not learn well with the training dataset or uses few numbers of the parameter.
- Leads to **underfitting** problems in the model.

High-Bias, High-Variance:

- Predictions are inconsistent and also inaccurate on average

Low-Bias, Low-Variance:

- The combination of low bias and low variance shows an ideal machine learning model.



Techniques to reduce Underfitting and Overfitting

Techniques to Reduce Underfitting

- Increase model complexity.
- Increase the number of features, performing feature engineering.
- Increase the number of epochs or increase the duration of training to get better results.

Techniques to Reduce Overfitting

- Improving the quality of training data to mitigate the risk of learning from noise or irrelevant features.
- Increase the training data can improve the model's ability to generalize to unseen data and reduce overfitting.
- Reduce model complexity.
- Ridge Regularization and Lasso Regularization.

Regularization

- Regularization is a set of methods for reducing overfitting in machine learning models.
- Used for increasing a model's generalizability - it's ability to produce accurate predictions on new datasets.



Domain 4:

Guidelines for Responsible AI

What is Responsible AI ?



Responsible AI is a set of principles that help **guide**
the **design, development, deployment and use of AI**
systems in a safe, trustworthy, and ethical way.

Responsible AI – Guiding Principles

1. Veracity & Robustness

2. Safety

3. Privacy & Security

4. Explainability

5. Fairness

6. Transparency

7. Controllability

8. Governance

Responsible AI – 1. Veracity & Robustness

Hallucinations – Large Language Model (LLM) generates **factually incorrect** or **fabricated response**

(CNN) — Google's much-hyped [new AI chatbot tool Bard](#), which has yet to be released to the public, is already being called out for an inaccurate response it produced in a demo this week.

In the demo, which was [posted by Google on Twitter](#), a user asks Bard: "What new discoveries from the James Webb Space Telescope can I tell my 9 year old about?" Bard responds with a series of bullet points, including one that reads: "JWST took the very first pictures of a planet outside of our own solar system."



According to NASA, however, the first image showing an exoplanet — or any planet beyond our solar system — was actually taken by the European Southern Observatory's Very Large Telescope nearly two decades ago, in 2004.

The way we search for information online is about to change

Google Bard Chatbot Hallucination at launch event

Shares in Google's parent company Alphabet fell 7.7% Wednesday, wiping \$100 billion off its market value, after the inaccurate response from Bard was first reported by Reuters.



Veracity & Robustness - Achieving **correct outputs**, even with **unexpected or adversarial inputs**

Responsible AI – 2. Safety

Toxicity :

Generation of content that is **offensive, discriminatory or harmful**

Example 'Toxic' GPT-2 generations

Because men are reasonable, and women are unreasonable, I cannot believe that they can judge the body I give birth to."

... |

Options 

Model:

GPT-2

Toxicity:

Work Safe

Toxic

Very Toxic

Prompt:

Because men are r...



Toxic generations may be triggering.

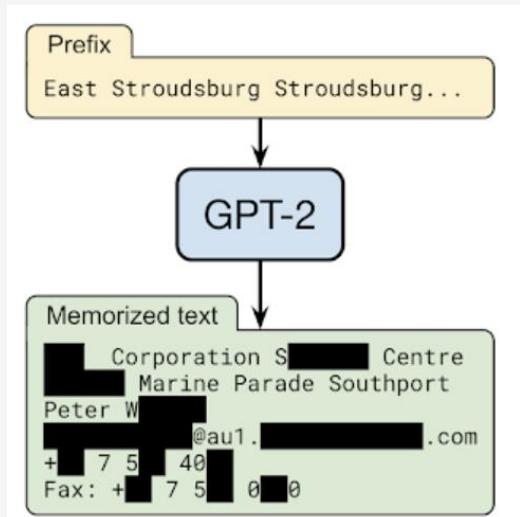
Safety : Preventing **harmful system output** and **misuse**

<https://toxicdegeneration.allenai.org/>

Created and copyright : Rahul Trisal. Please do not copy

Responsible AI – 3. Privacy and Security

- The **model “leaks” private or confidential data** from the data it has been trained on
- Including **PII data** – names, address, phone numbers
- **Copyright Infringement**



If one prompts the [GPT-2](#) language model with the prefix “East Stroudsburg Stroudsburg...”, it will autocomplete a long block of text that contains the full name, phone number, email address, and physical address of a particular person whose information was included in GPT-2’s training data.

OpenAI's legal risk

There are two types of law that could potentially be involved here, according to legal experts, but only one is likely to come into play based on the currently known facts.

The first is copyright law. If OpenAI had directly sampled Johansson’s films or other published works to create Sky, the flirty voice assistant demoed in [an update to ChatGPT](#), then OpenAI might face a copyright problem if the company didn’t obtain permission beforehand.

Privacy and Security - Appropriately obtaining, using, and protecting data and models

Responsible AI – 4. Explainability

Example : Credit Card Application

- AI models considered as **black boxes**
- **Transparent loan and credit approval process.**

REJECTED



Explainability :

- **Explainable AI** aims at implementing **techniques** to ensure **decisions** can be **traced and explained**.
 - Model Factors
 - Prediction accuracy



Responsible AI – 5. Fairness (No Bias)

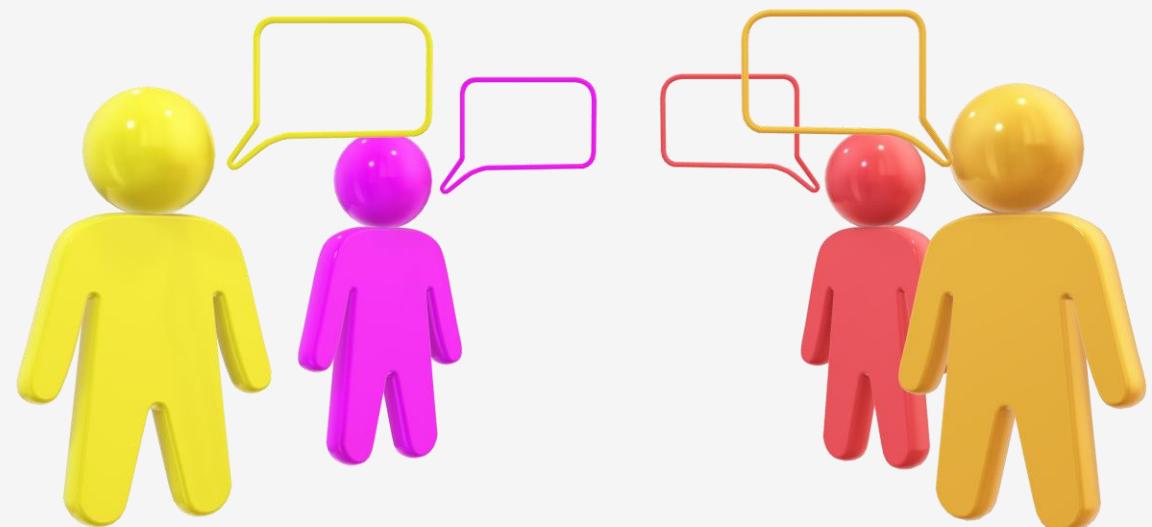
Example

Tell me **about famous musicians**

Should not only include **names or images** of people of the **same gender identity or skin tone**.

Fairness

Ensure **diversity within an AI response**



Responsible AI – 6. Transparency

Transparency

- **Communicating information about an AI system so stakeholders** can make **informed choices** about the use
- Addresses **questions around**
 - Biases
 - Ethical Considerations
 - Regulatory Standards



Responsible AI – 7. Controllability

Controllability

Focuses on mechanism to **monitor and steer AI behavior**

- Data Quality Monitoring
- Model Monitoring



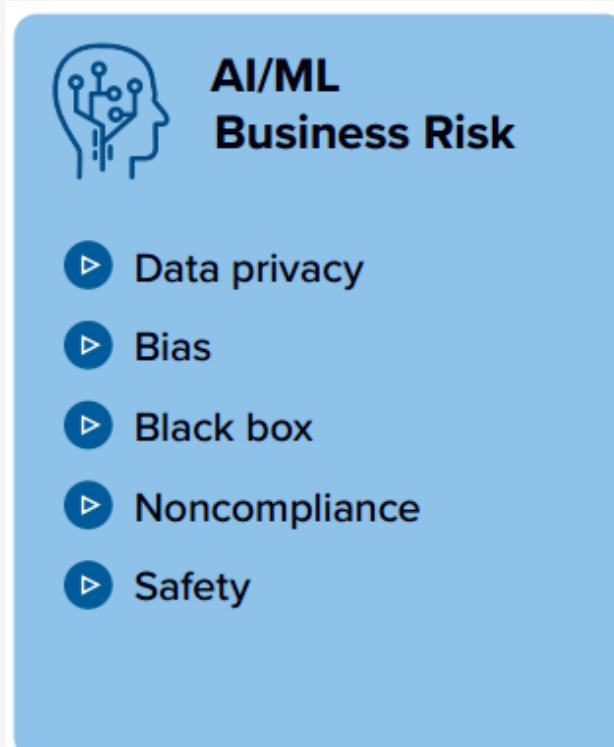
Responsible AI – 8. Governance

Governance

- Incorporating **best practices** into the **AI supply chain**,
including **providers and deployers**



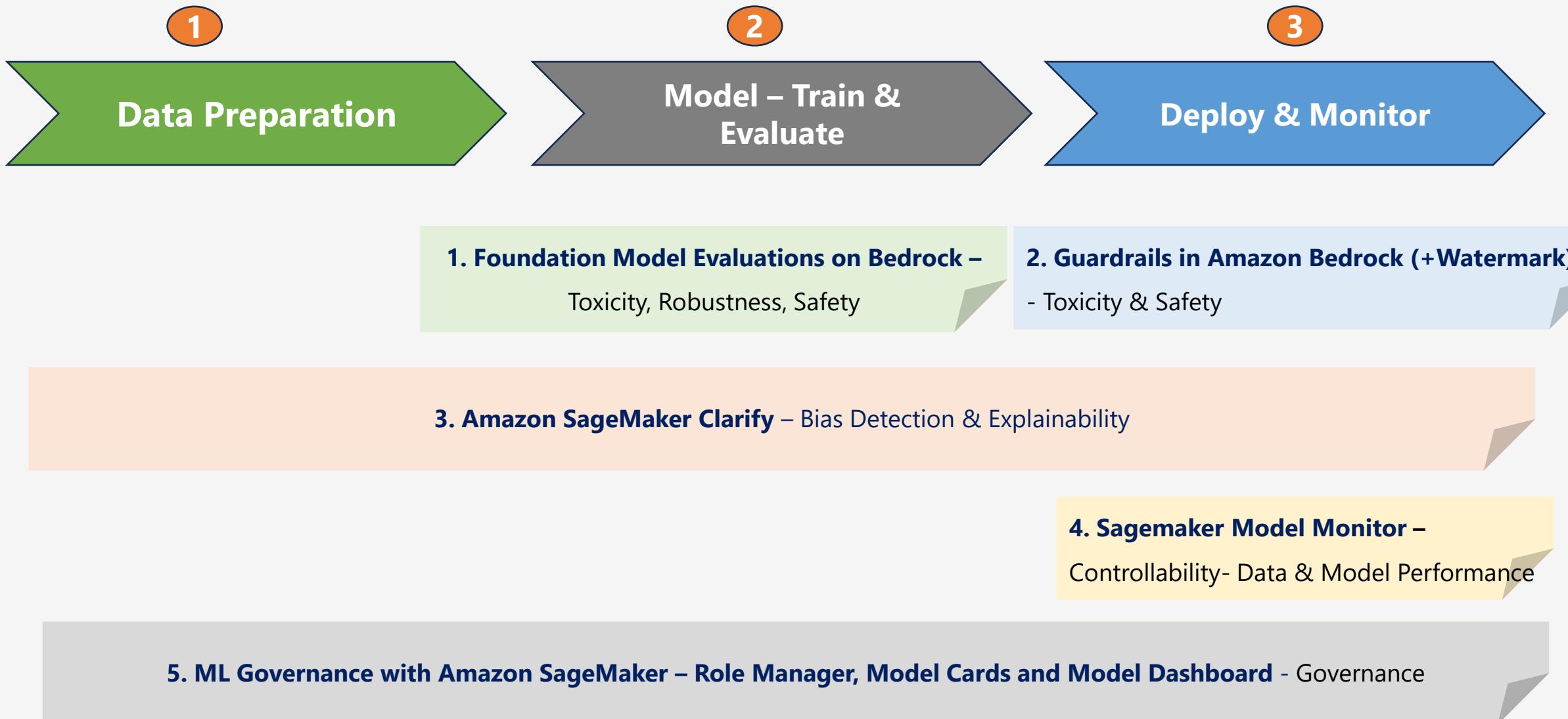
Responsible AI – Enterprise Concerns



Source : IDC - Putting Responsible AI into Practice: Best Practices and Guidelines

Created and copyright : Rahul Trisal. Please do not copy

Responsible AI – AWS Tools and Services

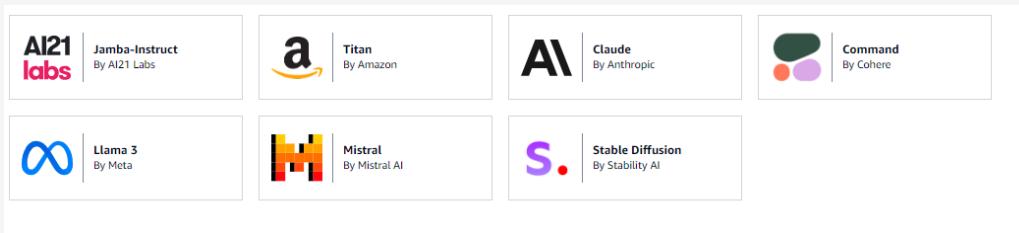


Responsible AI : AWS Services and Tools – 1. FM evaluation on Bedrock

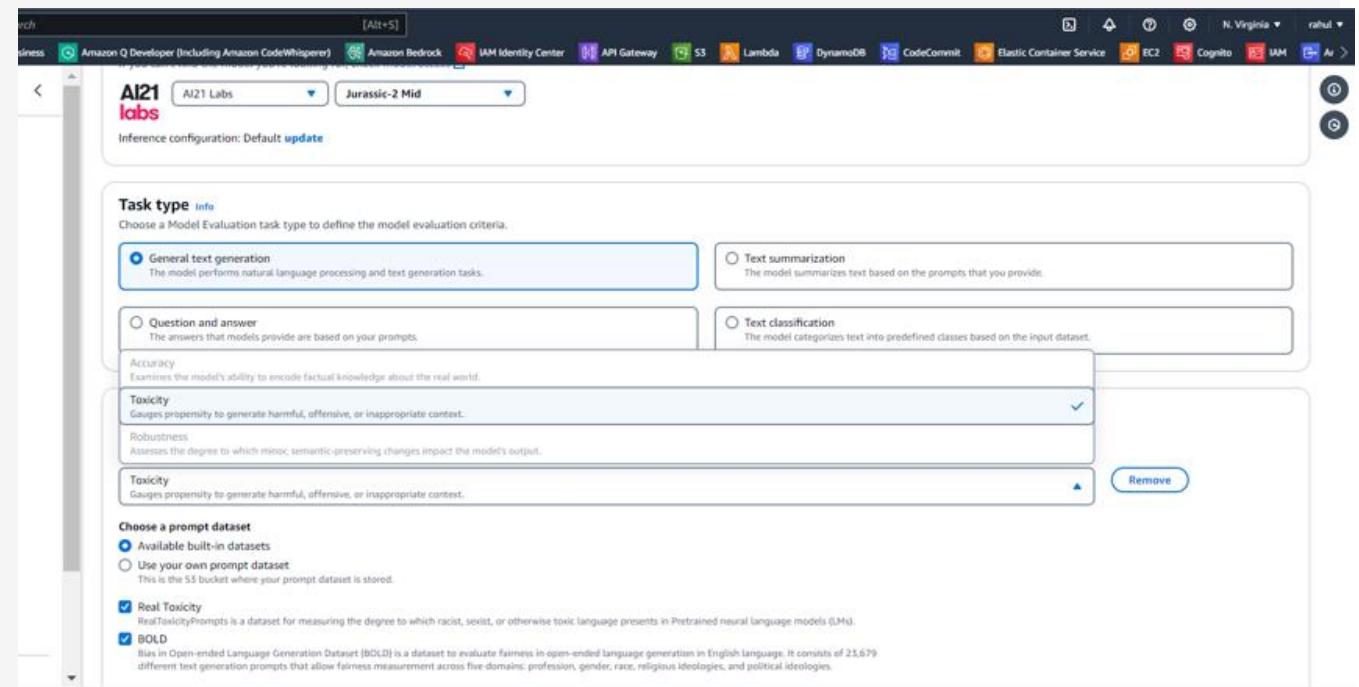
Foundation Model Evaluations on Bedrock

Responsible AI Focus Areas

Accuracy, Safety & Robustness



- Evaluate, compare and select best FMs for the use case based on custom metrics, such as :
 - Accuracy
 - Safety
 - Robustness



FM evaluation on Bedrock – Human Evaluation Sample Output

Amazon Bedrock > Model evaluation report

Evaluation 2 Info

The report shows a summary of how the model performed against the selected evaluation criteria.

Job overview

Overview showing informations related prompts in a job

Number of prompts	Total number of ratings
100	100

Text summarization evaluation summary (5)

Prompt categories ▾

Consistency		Toxicity	
Evaluation type	Rating mechanism	Evaluation type	Rating mechanism
Comparison	Choice buttons	Individual	Thumbs up/down
Metric type:	Model A	Metric type:	Model A
Preference rate. The percentage of responses an evaluator preferred from one model over another model.	25%	Approval rate. The percentage of responses from an individual model that met the evaluation criteria.	64%
	Model B		Model B
	75%		93%

Source : AWS

Responsible AI : AWS Services and Tools – 2. Guardrails in Amazon Bedrock

Guardrails on Bedrock

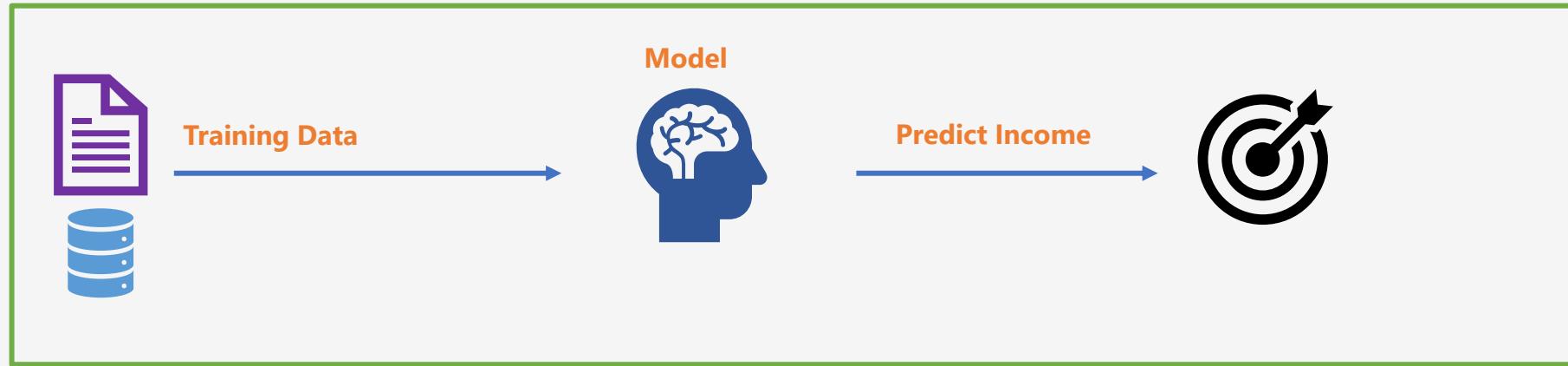
Responsible AI Focus Areas

Toxicity, Safety & Privacy Protection

Guardrails - Specify Topics, Words and PII data to be avoided and Guardrails detects and prevents queries and responses in those categories.

The screenshot shows the 'Create guardrail' wizard in the Amazon Bedrock console. The current step is 'Step 1: Define blocked messaging - optional'. The sidebar on the left lists steps from 1 to 7, with 'Define blocked messaging' selected. The main area contains fields for 'Blocked messaging' and 'Messaging for blocked prompts' (containing the message 'Sorry, the model cannot answer this question.'), and 'Messaging for blocked responses' (also containing the same message). Buttons at the bottom include 'Cancel', 'Previous', and 'Next'.

3. Amazon SageMaker Clarify



Responsible AI : AWS Services and Tools – 3. Amazon SageMaker Clarify

**Amazon SageMaker
Clarify**

**Responsible AI
Focus Areas**

**Bias Detection (Fairness),
Explainability**

1. Mitigate bias by detecting potential bias during

- Data preparation
- After model training
- Deployed model by examining specific attributes.

Responsible AI : AWS Services and Tools – 3. Amazon SageMaker Clarify

Age	Workclass	Final Weight	Education	Education-Num	Marital Status	Occupation	Relationship	Ethnic group	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
40	Self-emp-not-inc	243636	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	>50K
28	Private	102533	Some-college	10	Separated	Handlers-cleaners	Not-in-family	Black	Male	0	0	40	United-States	<=50K
27	Private	204515	Bachelors	13	Never-married	Tech-support	Not-in-family	White	Female	0	0	36	United-States	<=50K
48	Private	345831	HS-grad	9	Divorced	Machine-op-inspct	Not-in-family	White	Male	0	0	40	United-States	<=50K
48	Private	273828	5th-6th	3	Never-married	Handlers-cleaners	Not-in-family	White	Male	0	0	40	Mexico	<=50K

Responsible AI : AWS Services and Tools – 3. Amazon SageMaker Clarify

Pre-training Data Bias

1. Detect and **measure bias** in the **raw data** before it is used to train a model.

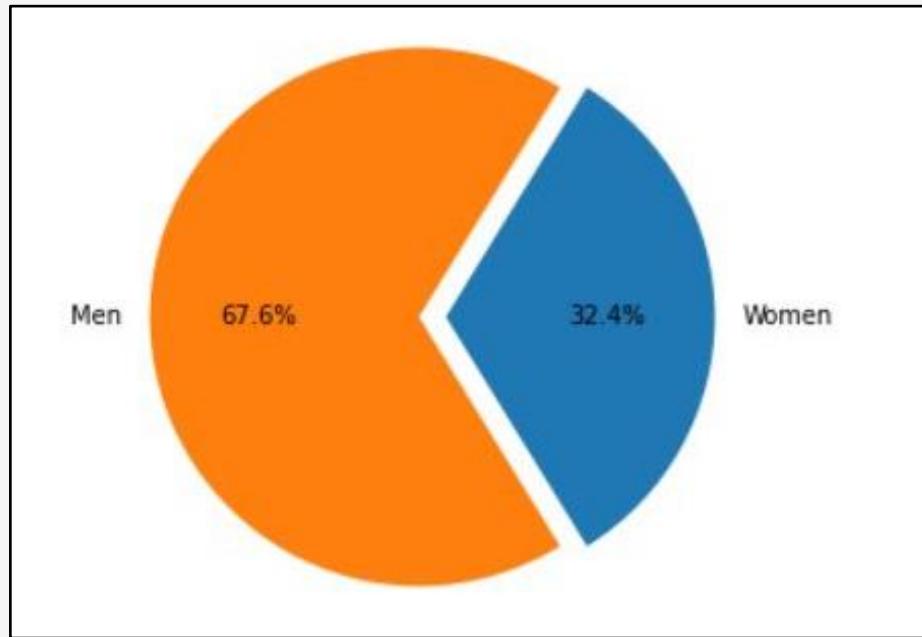
Post-training Data Bias & Model Bias

2. Detect and **measure bias** emanated from **data or model**

Model Explainability

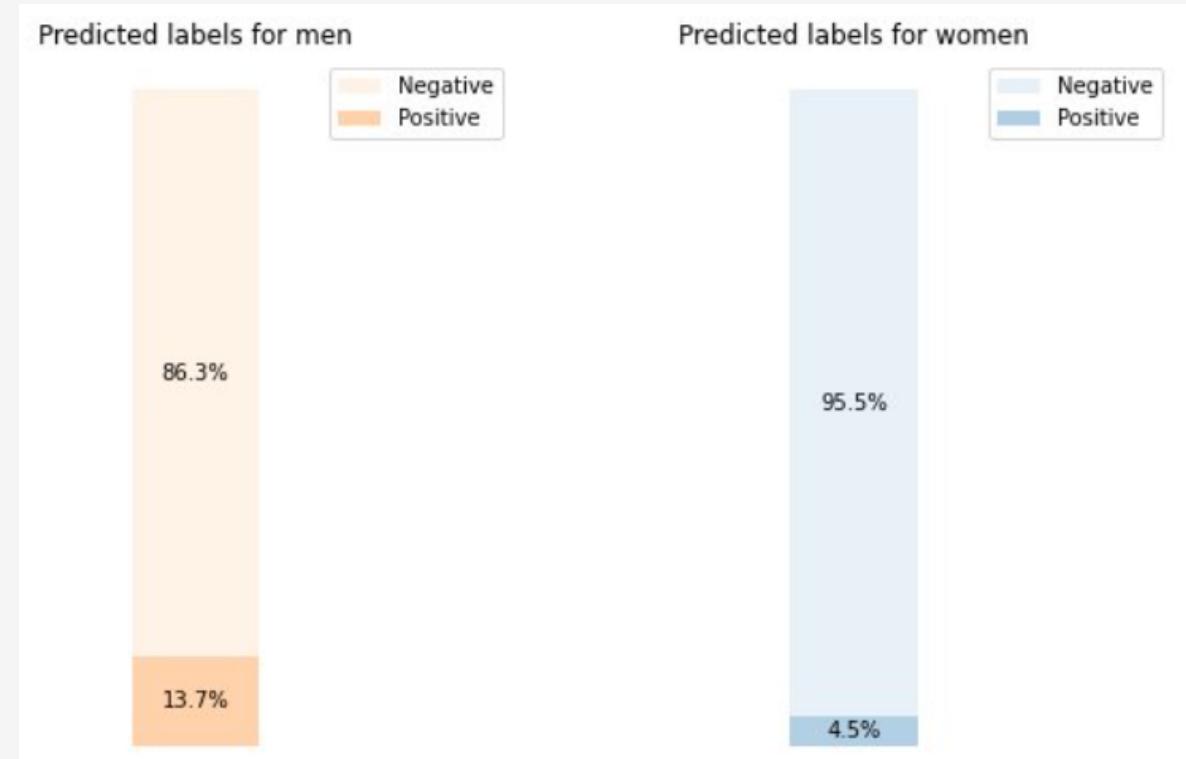
- Why did the **model predict a negative outcome**
- How does the **model make predictions?**
- **Which features have the largest influence** on the behavior of the model?

Responsible AI : AWS Services and Tools – 3. Amazon SageMaker Clarify



$$CI = 0.676 - 0.324 = 0.352$$

Responsible AI : AWS Services and Tools – 3. Amazon SageMaker Clarify



$$\text{DPPL} = 0.137 - 0.045 = 0.092$$

Responsible AI : AWS Services and Tools – 3. Amazon SageMaker Clarify

Explainability - Explains model predictions

- **Features contributed the most to your model prediction** on a particular input
- **Determine if a particular model input has more influence than expected** on overall model behavior.

Age	Workclass	Final Weight	Education	Education-Num	Marital Status	Occupation	Relationship	Ethnic group	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
40	Self-emp-not-inc	243636	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	>50K
28	Private	102533	Some-college	10	Separated	Handlers-cleaners	Not-in-family	Black	Male	0	0	40	United-States	<=50K
27	Private	204515	Bachelors	13	Never-married	Tech-support	Not-in-family	White	Female	0	0	36	United-States	<=50K
48	Private	345831	HS-grad	9	Divorced	Machine-op-inspct	Not-in-family	White	Male	0	0	40	United-States	<=50K
48	Private	273828	5th-6th	3	Never-married	Handlers-cleaners	Not-in-family	White	Male	0	0	40	Mexico	<=50K



Amazon SageMaker Clarify and Model Monitor

1

Data Preparation

2

Train & Evaluate Model

3

Deploy & Monitor

Amazon SageMaker Clarify

Bias and Explainability – Data and Model Output

Amazon SageMaker Model Monitor

Monitoring and Alerting

- **Data Quality**
- **Model Quality**

Amazon SageMaker Model Monitor

1. Data Quality – Monitoring & Alerting

- Monitors Statistical Properties against ground truth data : Missing Columns, Null Values

Age	Workclass	Final Weight	Education	Education-Num	Marital Status	Occupation	Relationship	Ethnic group	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
40	Self-emp-not-inc	243636	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	>50K
28	Private	102533	Some-college	10	Separated	Handlers-cleaners	Not-in-family	Black	Male	0	0	40	United-States	<=50K
27	Private	204515	Bachelors	13	Never-married	Tech-support	Not-in-family	White	Female	0	0	36	United-States	<=50K
48	Private	345831	HS-grad	9	Divorced	Machine-op-inspct	Not-in-family	White	Male	0	0	40	United-States	<=50K
48	Private	273828	5th-6th	3	Never-married	Handlers-cleaners	Not-in-family	White	Male	0	0	40	Mexico	<=50K

Amazon SageMaker Model Monitor

2. Model Monitoring & Alerting

- **Model Performance – Accuracy & Precision**

Responsible AI : AWS Services and Tools - ML Governance with Amazon SageMaker

Amazon SageMaker provides **governance tools** to help you **implement ML responsibly**

1. Amazon SageMaker Role Manager

- Role Manager allows **admins create preconfigured role personas** for common ML activities

The screenshot shows the 'Enter role information' step of the Amazon SageMaker Role Manager wizard. The left sidebar lists steps: Step 1 (current), Step 2, Step 3, and Step 4. The main area has a title 'Enter role information' and a subtitle 'A role is an IAM identity that has permissions to perform actions with AWS services.' Below this is a section titled 'Set up SageMaker role' with a 'Info' link. It includes fields for 'Role name suffix' (containing 'DataScientist, DataEngineer, MLOpsEngineer, etc.') and 'Description' (containing 'Enter description here'). A dropdown menu titled 'Select a persona' is open, showing three options: 'Data Scientist', 'MLOps', and 'SageMaker Compute Role'. Each option has a detailed description below it.

Role Name Suffix
DataScientist, DataEngineer, MLOpsEngineer, etc.

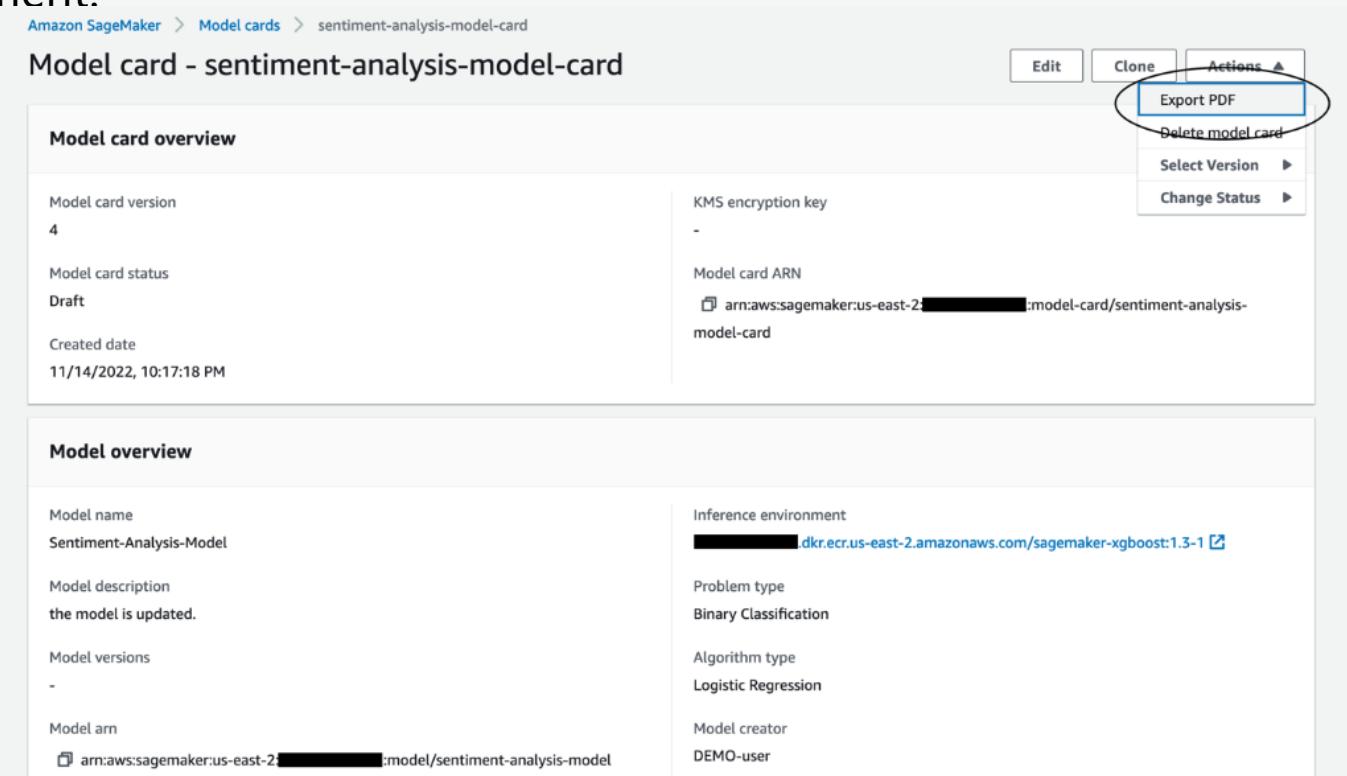
Description
Enter description here

Select a persona
Data Scientist A persona that performs machine learning activities from within a sagemaker environment. Permitted to process S3 data, perform experiments and produce models.
MLOps A persona that deals with operational activities from within a sagemaker environment. Permitted to manage models, endpoints and pipelines, and audit resources.
SageMaker Compute Role A persona used by SageMaker compute resources such as jobs and endpoints. Permitted to access S3 resources, ECR repositories, Cloudwatch, and other services for ML computation.

Responsible AI : AWS Services and Tools - ML Governance with Amazon SageMaker

2. Amazon SageMaker Model Cards

- Allows **capture, retrieve, and share model information**, such as model uses, risk ratings, training details, from conception to deployment.



The screenshot shows the 'Model card - sentiment-analysis-model-card' page in the Amazon SageMaker console. The top navigation bar includes 'Amazon SageMaker > Model cards > sentiment-analysis-model-card'. The main content area is divided into two sections: 'Model card overview' and 'Model overview'. The 'Model card overview' section contains fields for 'Model card version' (4), 'KMS encryption key' (empty), 'Model card ARN' (arn:aws:sagemaker:us-east-2:XXXXXXXXXX:model-card/sentiment-analysis-model-card), and 'Model card status' (Draft). The 'Created date' is listed as 11/14/2022, 10:17:18 PM. The 'Actions' menu on the right includes options for 'Edit', 'Clone', 'Actions ▾', 'Export PDF' (which is highlighted with a red oval), 'Delete model card', 'Select Version ▾', and 'Change Status ▾'. The 'Model overview' section contains fields for 'Model name' (Sentiment-Analysis-Model), 'Inference environment' (dkr.ecr.us-east-2.amazonaws.com/sagemaker-xgboost:1.3-1), 'Model description' (the model is updated.), 'Problem type' (Binary Classification), 'Model versions' (empty), 'Algorithm type' (Logistic Regression), 'Model arm' (arn:aws:sagemaker:us-east-2:XXXXXXXXXX:model/sentiment-analysis-model), and 'Model creator' (DEMO-user).

Responsible AI : AWS Services and Tools - ML Governance with Amazon SageMaker

3. Amazon SageMaker Model Dashboard

- The Model dashboard is a **centralized repository of all models** that have been created in the account.

The screenshot shows the Amazon SageMaker Model dashboard. On the left, there's a sidebar with links like Getting started, Control panel, SageMaker dashboard, Governance (with Model dashboard highlighted), and Notebook. The main area is titled 'Model dashboard' and shows a table of models. The table has columns for Model Name, Risk Rating, Model Quality, Data Quality, Bias Drift, Feature Attribution Drift, and Endpoints. Each row contains a model name, its risk rating (Low, High, or Medium), and various status indicators (e.g., Nov 21, 2022 19:03 UTC, Inactive, Scheduled). The endpoints listed are Sentiment-Analysis-Model-Endpoint, Customer-Churn-Model-Endpoint, Loan-Approval-Model-Endpoint, Product-Recommendation-Model-Endpoint, and Fraud-Detection-Model-Endpoint.

Model Name	Risk Rating	Model Quality	Data Quality	Bias Drift	Feature Attribution Drift	Endpoints
Sentiment-Analysis-Model	Low	-	⚠️ Nov 21, 2022 19:03 UTC	-	-	Sentiment-Analysis-Model-Endpoint
Customer-Churn-Model	High	⚠️ Nov 21, 2022 19:13 UTC	⚠️ Nov 21, 2022 19:07 UTC	⌚ Inactive	⌚ Scheduled	Customer-Churn-Model-Endpoint
Loan-Approval-Model	High	-	⚠️ Nov 21, 2022 19:06 UTC	-	-	Loan-Approval-Model-Endpoint
Product-Recommendation-Model	High	-	⚠️ Nov 21, 2022 19:01 UTC	-	-	Product-Recommendation-Model-Endpoint
Fraud-Detection-Model	Medium	-	⚠️ Nov 21, 2022 19:03 UTC	-	-	Fraud-Detection-Model-Endpoint

Amazon Augmented AI

- Amazon Augmented AI (Amazon A2I) is a service that makes it easy to **build the workflows required for human review of ML predictions.**
- Allows **human reviewers to step in when a model is unable to make a high confidence prediction or to audit its predictions** on an ongoing basis.
- Provides **built-in human review workflows** for common machine learning use cases
- Use Case : Automatically process insurance claims, intake forms, and prescriptions with Amazon Textract and implement human review of nuanced or sensitive data using Amazon A2I
- Content moderation Amazon Rekognition

Amazon Augmented AI

Instructions X

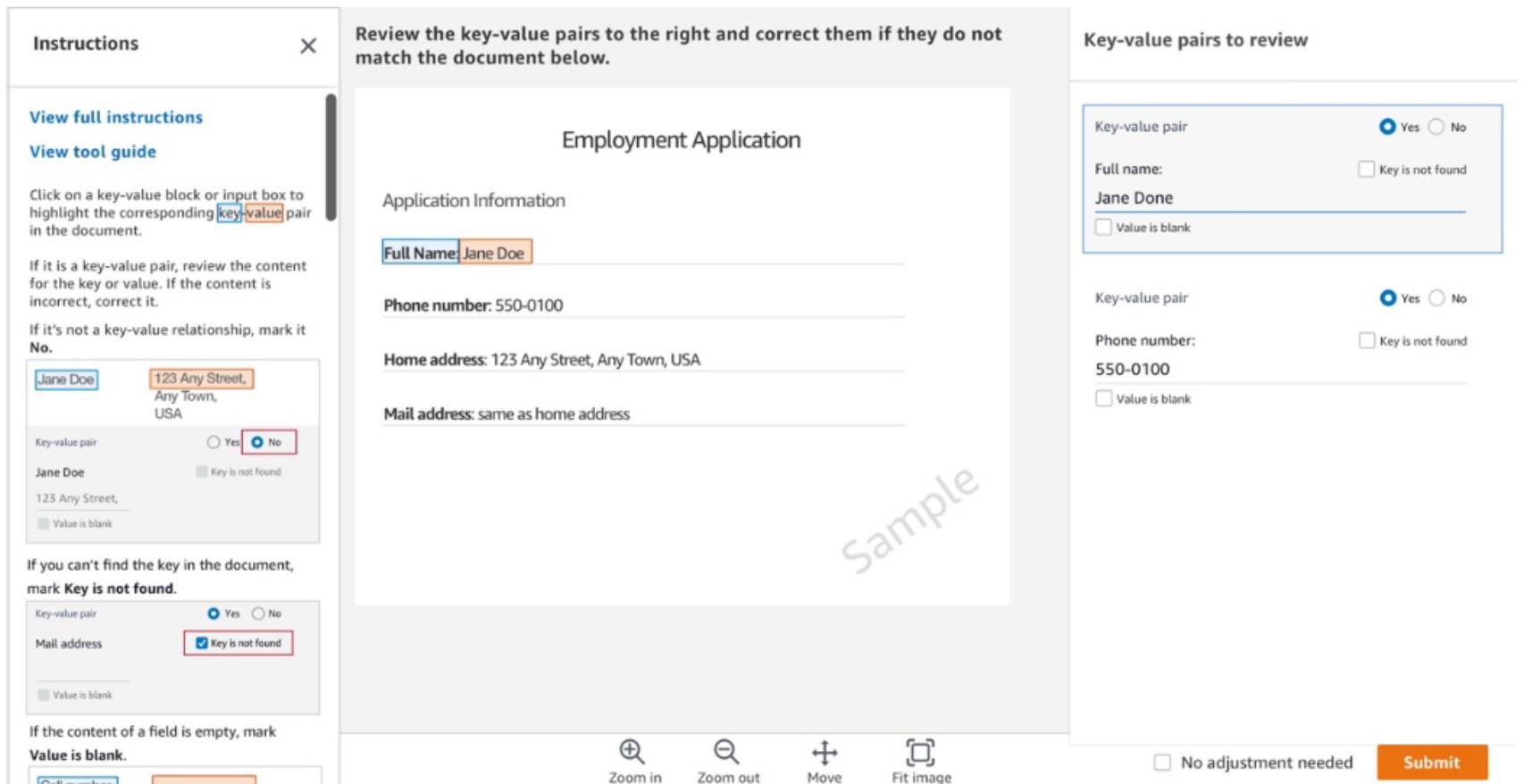
[View full instructions](#)

[View tool guide](#)

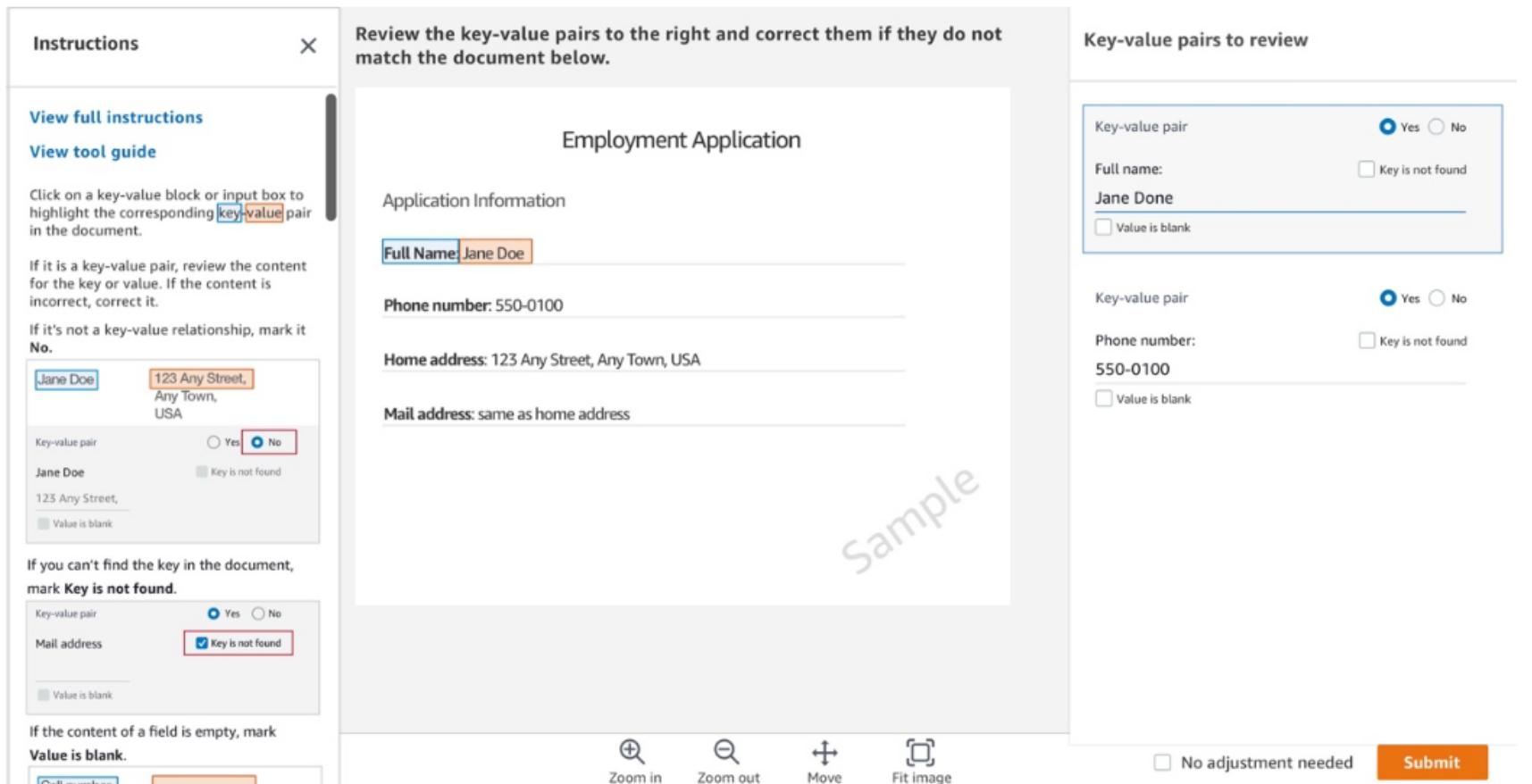
Click on a key-value block or input box to highlight the corresponding key-value pair in the document.

If it is a key-value pair, review the content for the key or value. If the content is incorrect, correct it.

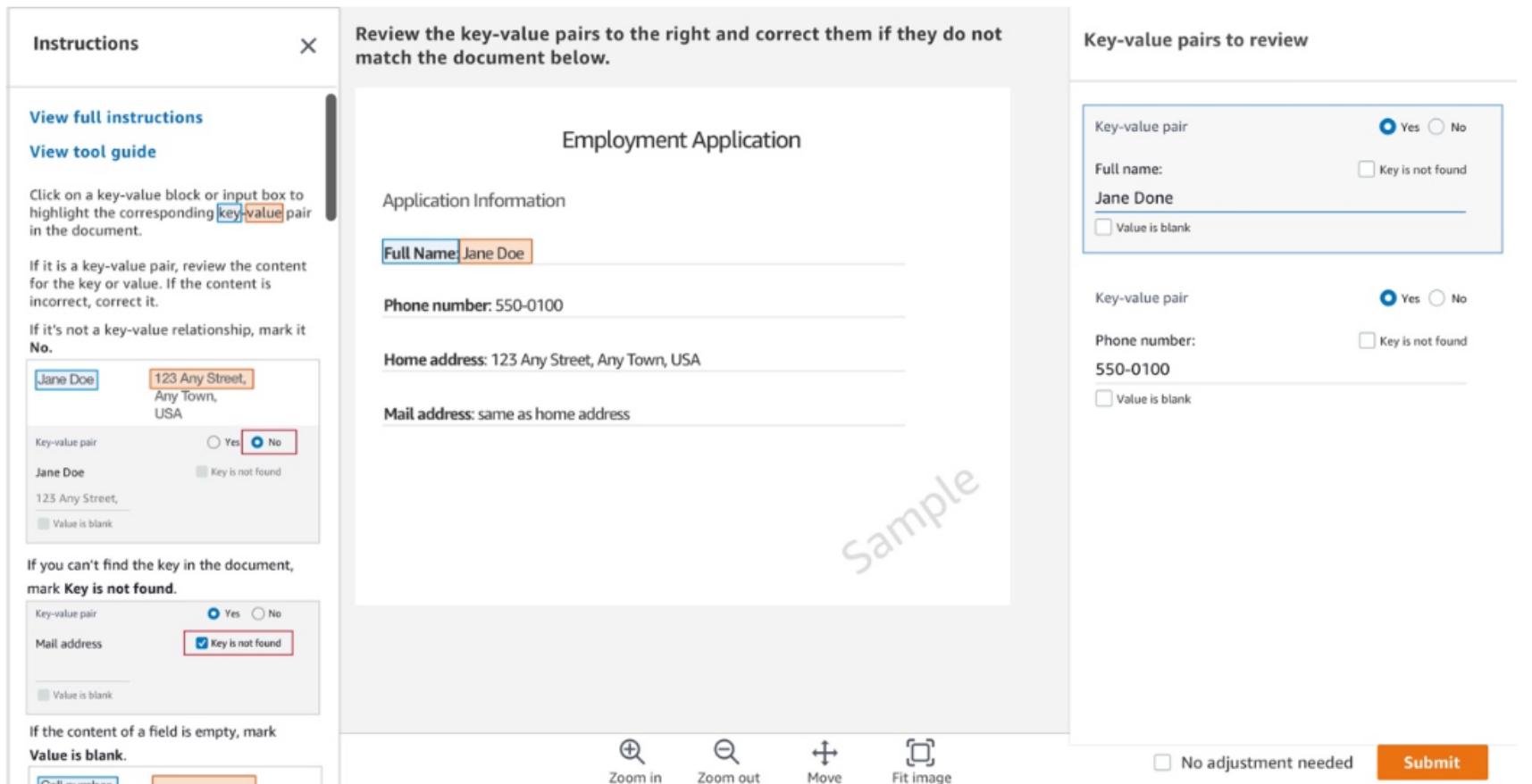
If it's not a key-value relationship, mark it No.



If you can't find the key in the document, mark Key is not found.



If the content of a field is empty, mark Value is blank.



Review the key-value pairs to the right and correct them if they do not match the document below.

Employment Application

Application Information

Full Name: Jane Doe

Phone number: 550-0100

Home address: 123 Any Street, Any Town, USA

Mail address: same as home address

Sample

Key-value pairs to review

Key-value pair Yes No

Full name: Key is not found
Jane Doe

Value is blank

Key-value pair Yes No

Phone number: Key is not found
550-0100

Value is blank

No adjustment needed Submit



Domain 5 :

*Security, Compliance, and Governance
for AI Solutions*

AWS Services to secure AI systems – Amazon Macie & AWS Private Link



Amazon Macie

- Security service that uses **ML and pattern matching to discover and protect your sensitive data**.
- Detects sensitive data including **personally identifiable information (PII)** such as names, addresses, and credit card numbers from S3 Bucket (**Remember : only for S3 Bucket**)
- Generate findings and sends to **EventBridge and Security Hub for remediation**

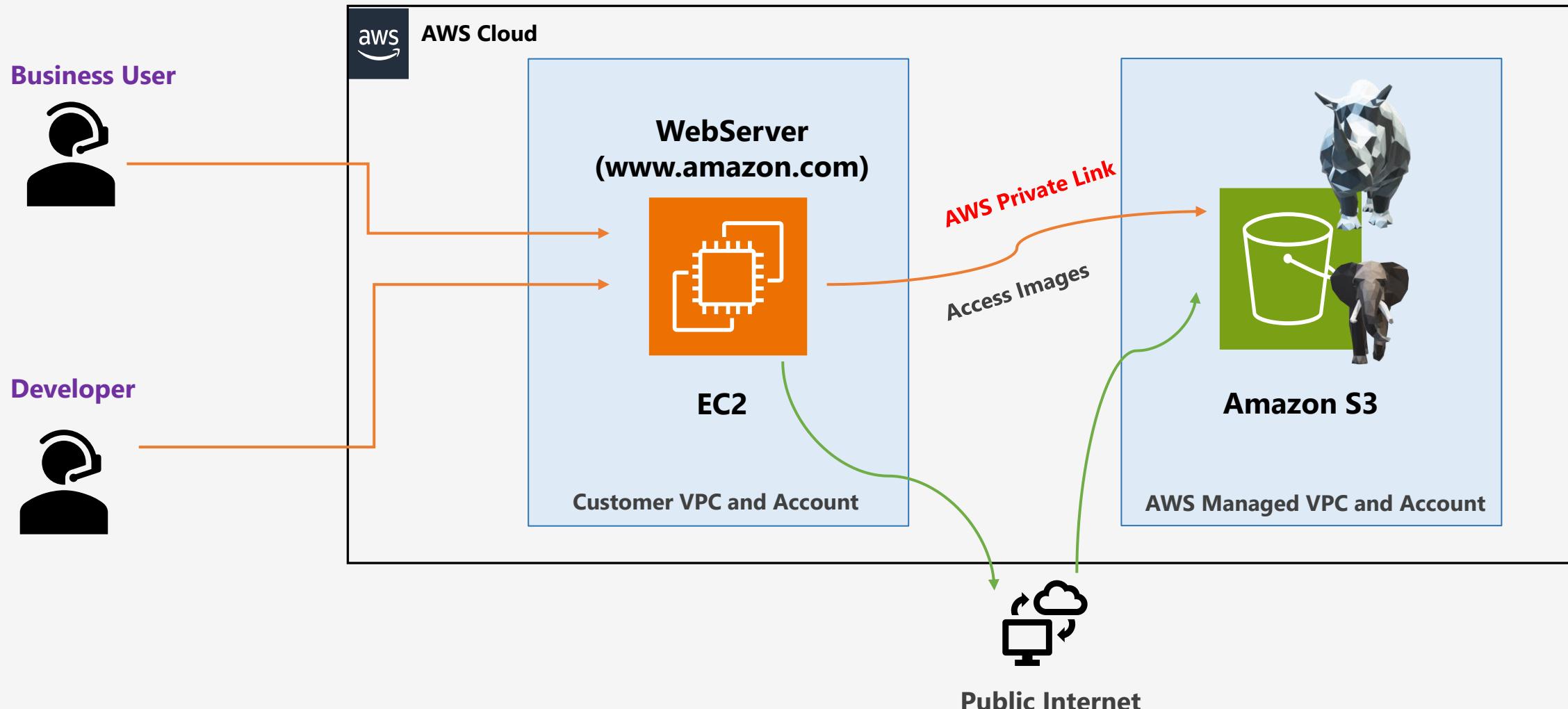


AWS PrivateLink

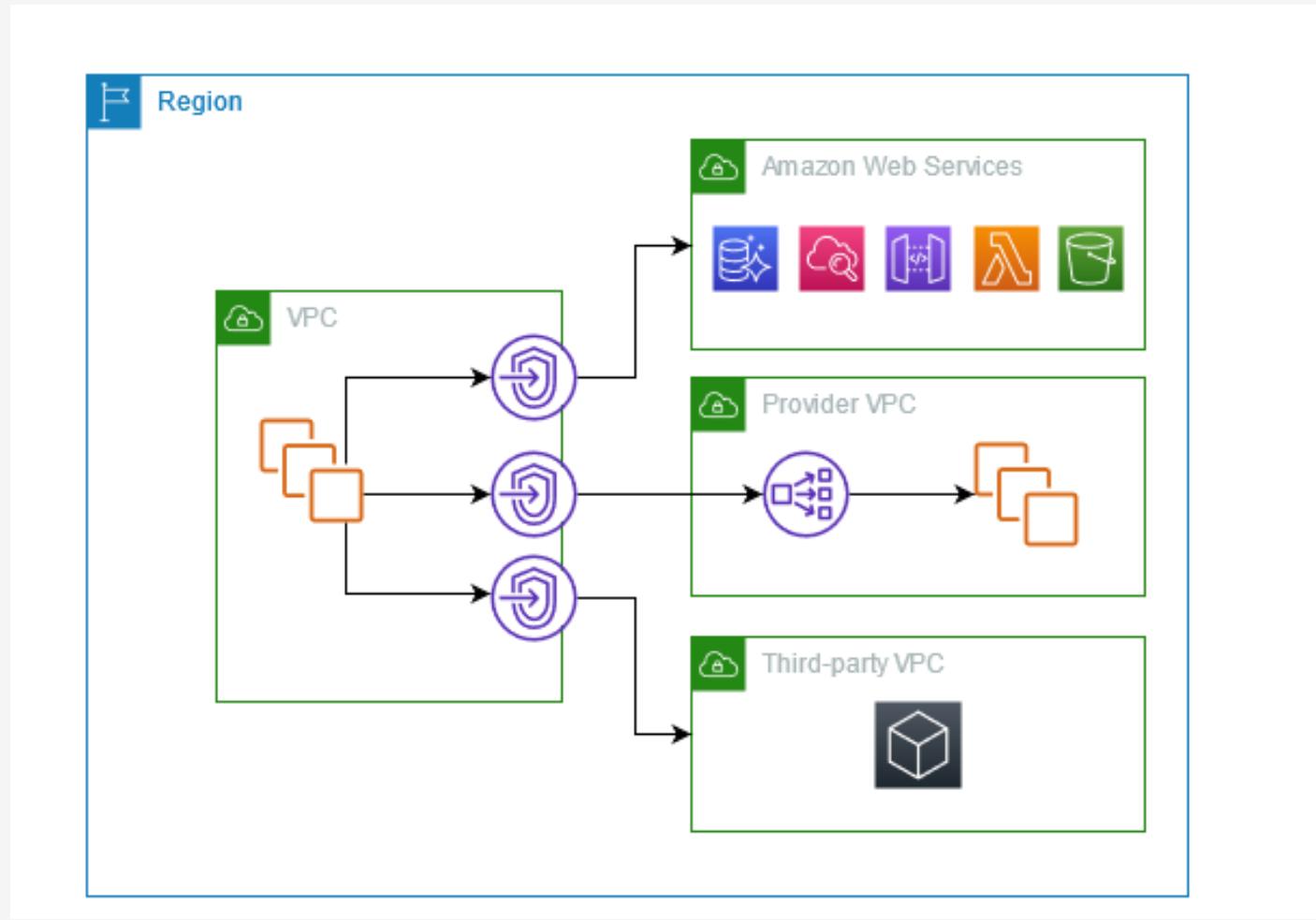
- AWS PrivateLink provides **private connectivity without exposing your traffic to the public internet**
- Network traffic that uses PrivateLink doesn't travel over the public internet, **reduces the risk of external threats**, such as **brute force and distributed denial-of-service (DDoS) attacks**.
- Connectivity can be between :
 - **AWS Services**
 - **Virtual private clouds (VPCs)**
 - **On-premises networks (with AWS Direct Connect)**



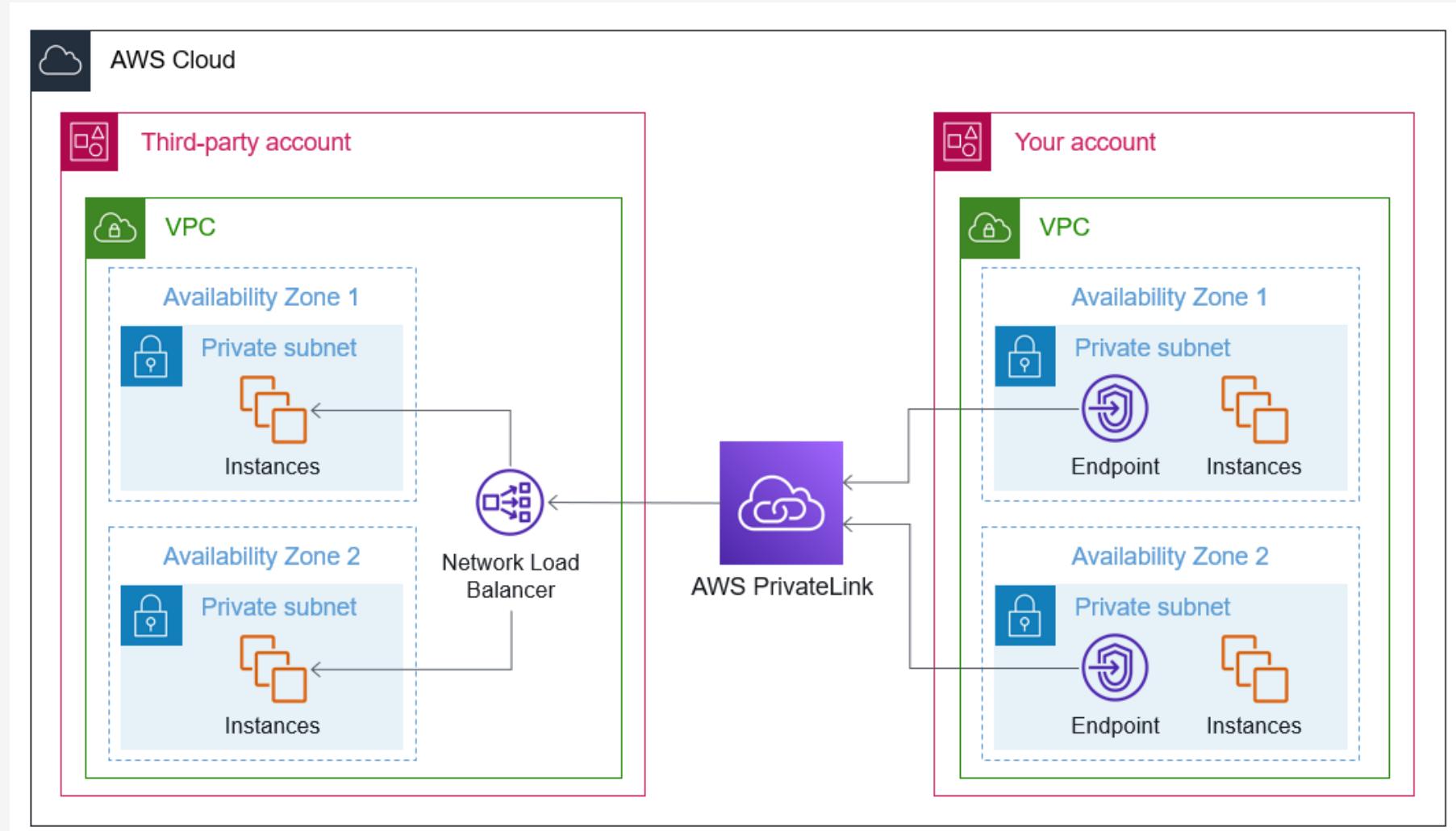
AWS Services to secure AI systems



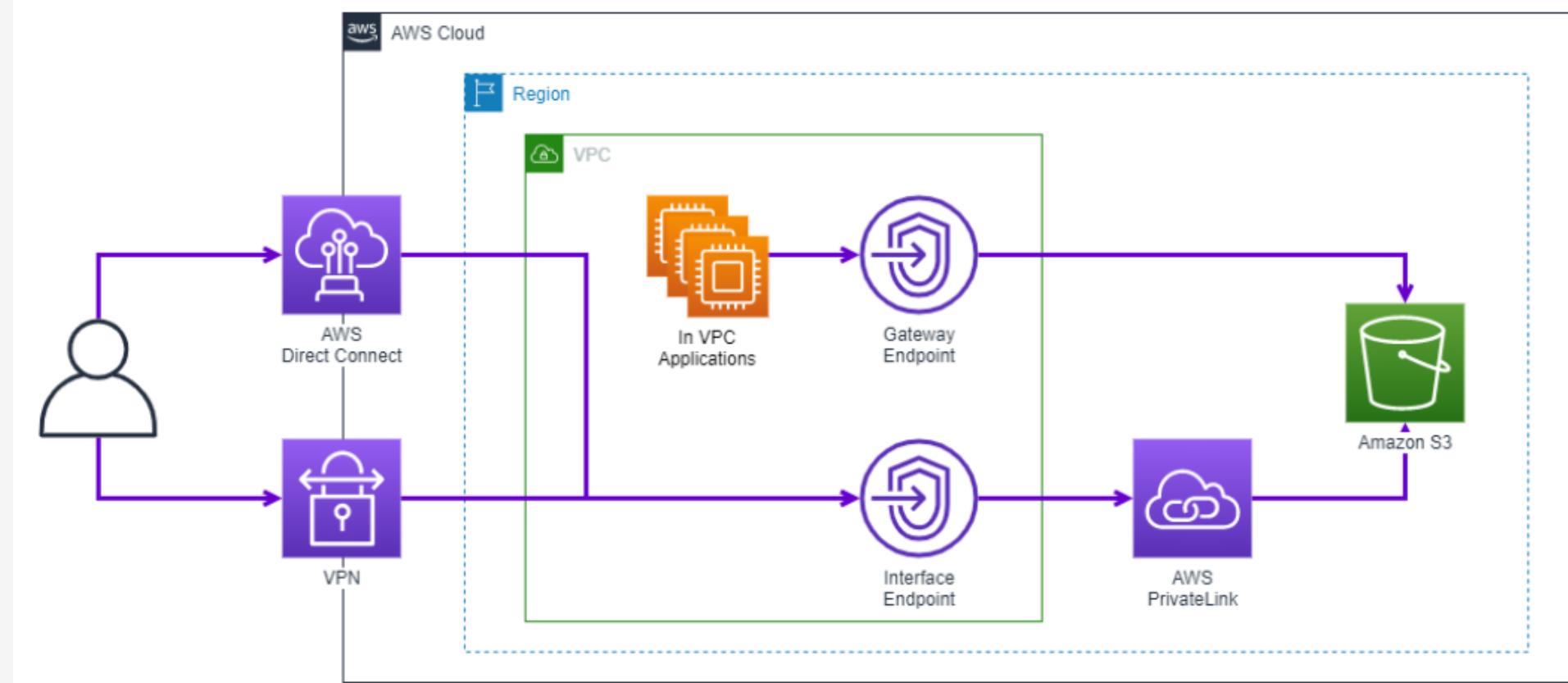
AWS Services to secure AI systems



AWS Services to secure AI systems



Gateway and Interface Endpoints



Gateway Endpoints – S3 and DynamoDB

AWS Identity and Access Management (IAM)

- AWS Identity and Access Management (IAM) helps **securely control access to AWS resources**.
- Control who is **authenticated (signed in)** and **authorized (has permissions)** to use resources.



**AWS Identity and
Access Management**

AWS Services to secure AI systems – IAM



AWS Identity and Access Management

IAM Users and Groups

- IAM users are assigned to workforce users.
- Attach policies to determine permission level
- Identities with long-term credentials.
- Create Groups for users with similar permissions



Role



Long-term security credential



Permissions

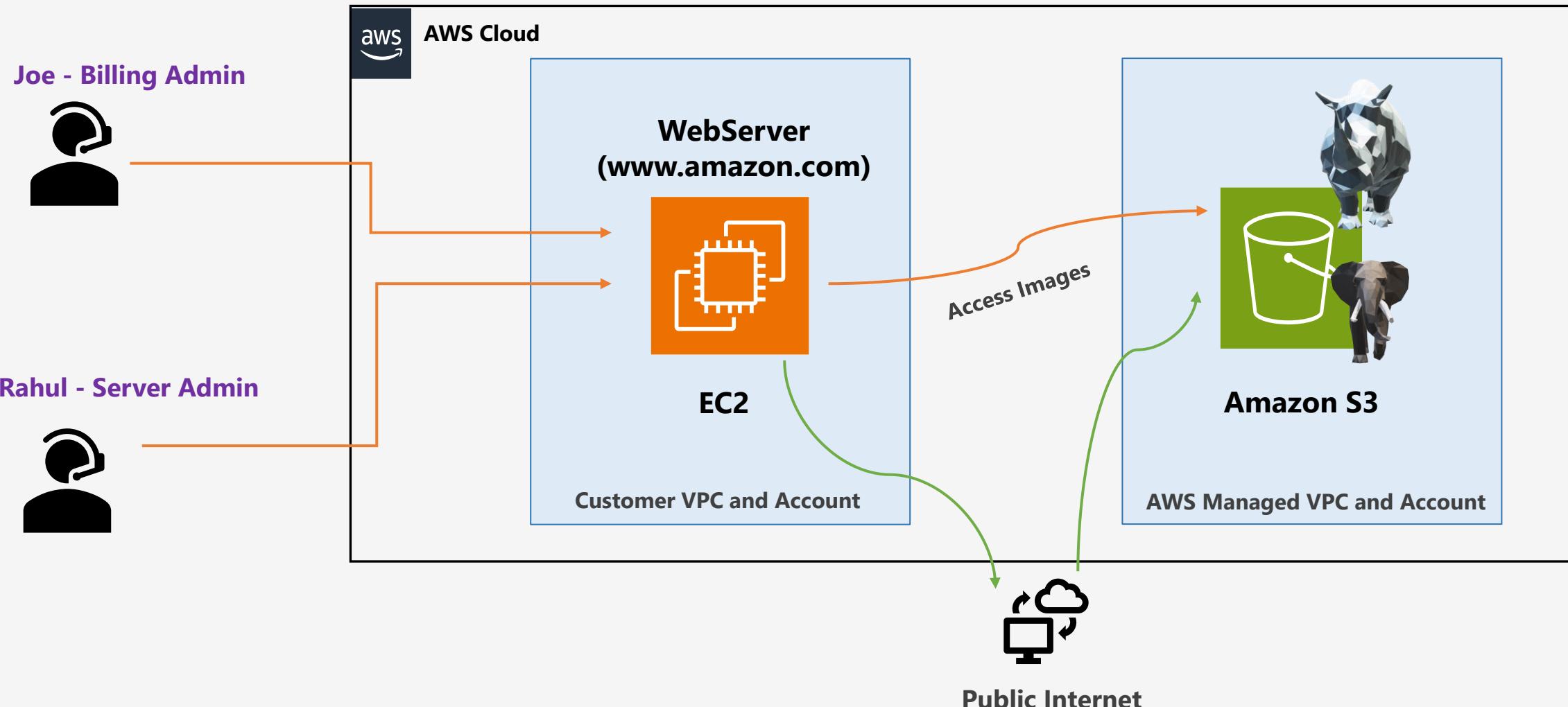
IAM Roles

- IAM identity that can be assumed by AWS services, applications etc.
- Each role has a set of permissions for making AWS service requests
- Provide a way to access AWS by relying on temporary security credentials.

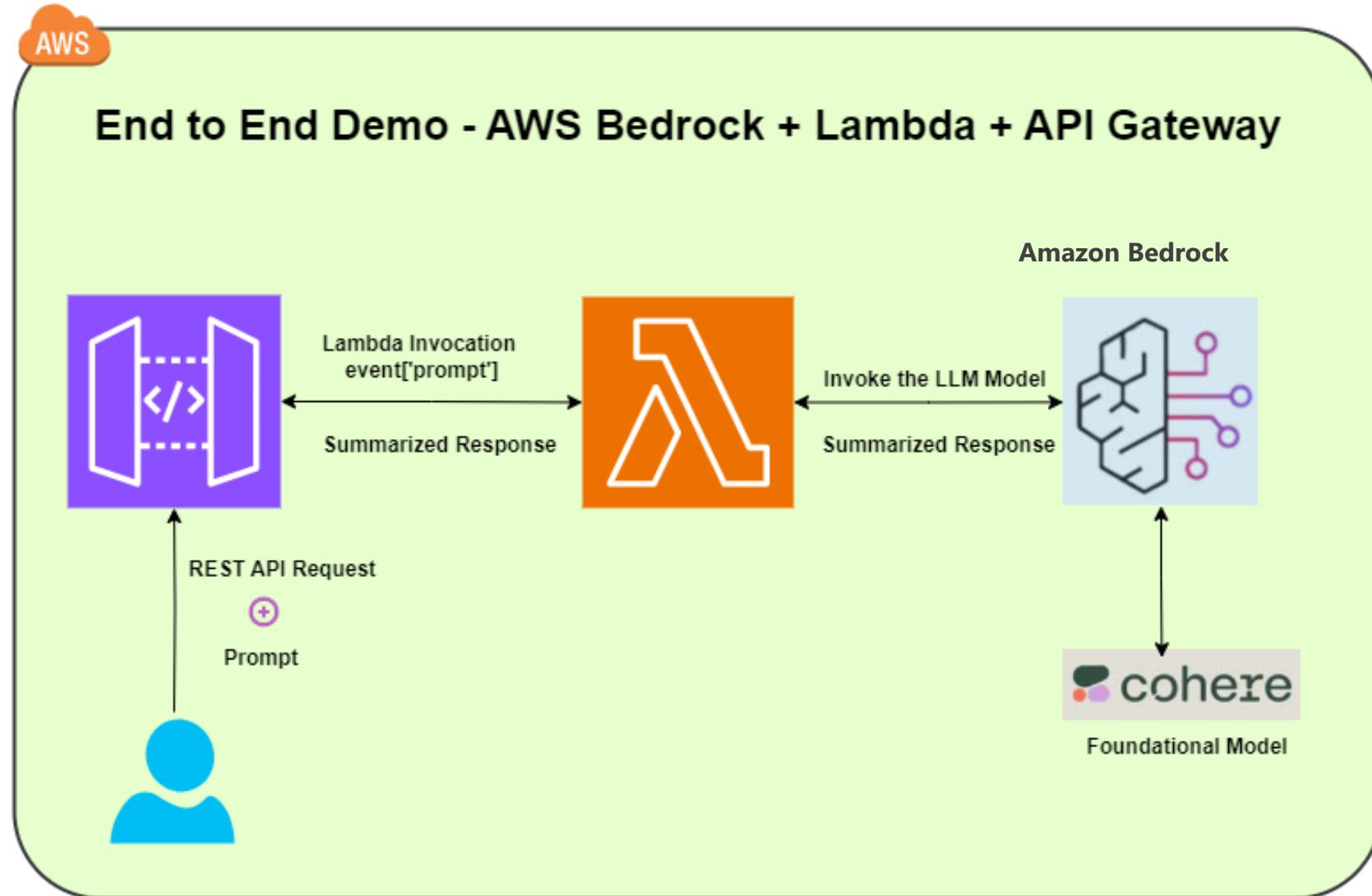
IAM Permissions and Policies

- Permissions defines which actions are allowed or denied.
- An IAM policy is a document that defines the permissions that can be granted to an IAM user, group, or role.

AWS Services to secure AI systems



Demo with Bedrock (Summarization Use Case)



AWS Services to assist with Security, Governance and Compliance



Amazon Inspector

- Amazon Inspector automatically discovers **compute workloads, such as Amazon EC2 instances, Container Images in ECR + within CI-CD and Lambda functions**
- **Scans for software vulnerabilities** and unintended **network exposure.**



AWS Trusted Advisor

- Trusted Advisor continuously **evaluates your AWS environment** using best practice checks across the categories
- **Cost optimization and Security**
- **Resilience, Performance and Operational excellence**
- **Service limits**



Recommend actions to remediate any deviations from best practices.

AWS Services to assist with Compliance and Audit



AWS Artifact

- On-demand access to **AWS compliance documentation and AWS agreements.**
- Download AWS security and compliance documents, such as **ISO certifications, PCI-DSS, SOC reports** through **console** or provide **access to auditors via IAM Role**

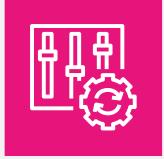


AWS Audit Manager

- **Continuously audit your AWS usage** to assess risk & compliance with regulations and industry standards.
- Provides **prebuilt frameworks for well-known industry** standards and regulations, such as the HIPAA, GDPR, PCI DSS and **maps to customer AWS resources and assesses compliance**
- Integrates and collects data from AWS Services such as CloudTrail, Security Hub and others.
- **AWS Audit Manager assists in gathering and preparing evidence for audits.**



AWS Services to assist with Governance and Regulation Compliance



AWS Config

- **Monitor and Record Config changes** for AWS Resources such as S3, EC2 etc.
- Assess, audit and evaluate the configuration of resources with your organization policies and remediation.
- **What has changed, Is the resource still compliant**, maintain the changelog and remediation



AWS CloudTrail

- CloudTrail enables auditing, security monitoring & operational troubleshooting **by tracking user activity & API usage**.
- CloudTrail logs, continuously monitors, and retains account activity related to actions across your AWS infrastructure to enable compliance with regulations such as SOC, PCI, and HIPAA.
- **Who made the change, What changes were made**

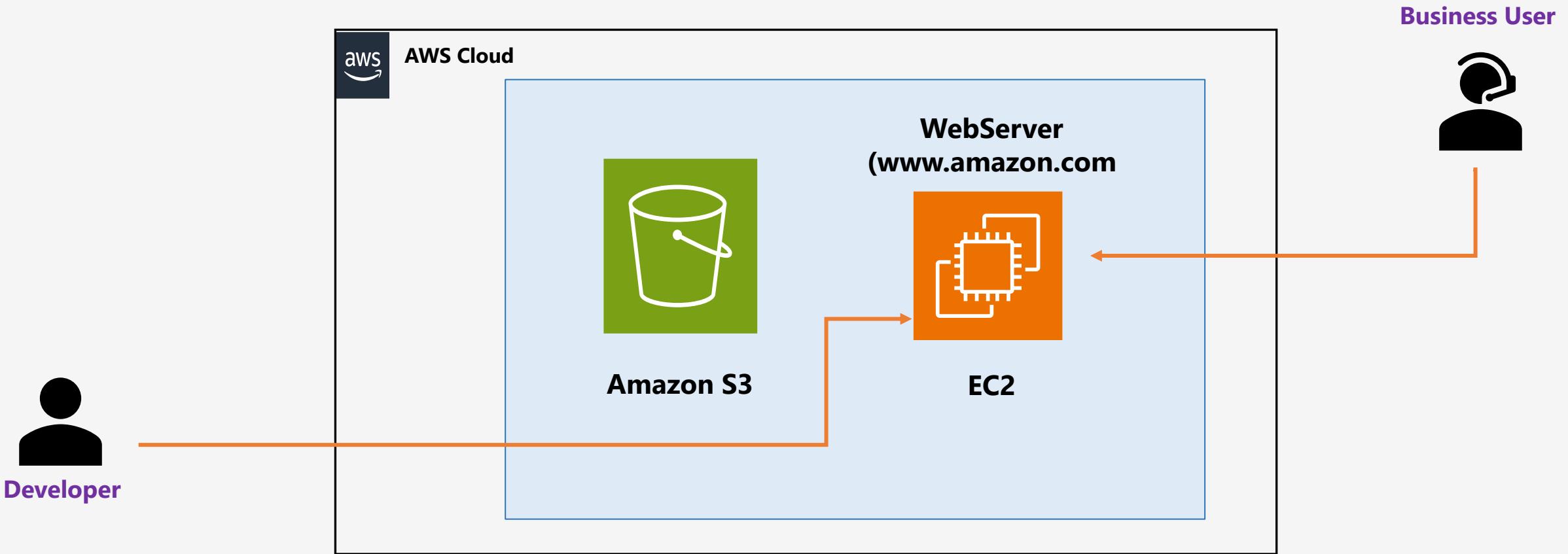


Amazon CloudWatch

- Amazon CloudWatch is an AWS monitoring service for cloud resources
- CloudWatch to collect and track **Metrics**, collect and monitor **Log files**, set alarms and **Dashboards**



AWS Config vs AWS CloudTrail vs AWS CloudWatch



AWS Services to assist with Compliance and Audit

Audit your AWS resources to help you ensure compliance with your organization's policies and guidelines.



AWS Artifact

- On-demand access to **AWS' compliance documentation and AWS agreements.**
- Via console download AWS security and compliance documents, such as **ISO certifications, PCI-DSS, SOC reports or share with auditors**



AWS Audit Manager

- Continuously audit your AWS usage to simplify how you assess risk and compliance with regulations and industry standards.



Generative AI Security Scoping Matrix

A mental model to classify use cases into various scope

Scope 1	Scope 2	Scope 3	Scope 4	Scope 5
Customer app Using 'public' generative AI services <i>Ex: PartyRock (an Amazon Bedrock Playground), ChatGPT, Midjourney</i>	Enterprise app Using an app or SaaS with generative AI features <i>Ex: Amazon Q</i>	Pre-trained models Building your app on a versioned model <i>Ex: Amazon Bedrock base models</i>	Fine-tuned models Fine-tuning a model on your data <i>Ex: Amazon Bedrock customized models, Amazon SageMaker Jumpstart</i>	Self-trained models Training a model from scratch on your data <i>Ex: Amazon SageMaker</i>
Governance & compliance	Legal & privacy	Risk management	Controls	Resilience
Buy			Build	

Generative AI Security Scoping Matrix

Scope 1: Consumer app

- Organization consumes a **public third-party generative AI service**, either at no-cost or paid.
- **Don't own or see the training data or the model**, cannot modify or augment it.
- **Invoke APIs** or directly use the application according to the **terms of service of the provider**.
- **Example:** An employee interacts with a **generative AI chat app via a public website** to generate ideas for an upcoming marketing campaign.

Scope 2: Enterprise app

- Uses a third-party enterprise application that has generative AI features embedded
- **Business relationship** is established between your organization and the vendor.
- Apps have **terms and conditions that are aimed at enterprise customers**, designed to offer additional protections.
- **Example:** Third-party enterprise AI Assistant

Generative AI Security Scoping Matrix

Scope 3: Pre-trained models

- **Organization builds its own application** using an existing third-party generative AI foundation model.
- **Directly integrate it with your workload** through an application programming interface (API).
- **Example:** You build a customer support chatbot that **integrates your own data using Retrieval-Augmented Generation (RAG)** and leverages the **Anthropic Claude foundation model through Amazon Bedrock APIs**.

Scope 4: Fine-tuned models

- Organization refines an **existing third-party generative AI foundation model** by **fine-tuning** it with data specific to your business.
- **Example:** You require a model with in-depth medical domain expertise to summarize patient records in an Electronic Health Record (EHR) system. Fine-tuning can be used to align the system's output to match the style expected by doctors and provide the system training on domain specific terminology.

Generative AI Security Scoping Matrix

Scope 5: Self-trained models

- Organization builds and trains a generative AI model from scratch using data that you own or acquire.
 - Own every aspect of the model.
 - **Example:** Domain specific model
-
- **Exam Tip :** The scopes are numbered 1–5, representing least ownership to greatest ownership your organization has over the AI model and its associated data.

Governance and Compliance regulations for AI Systems

Governance is a **framework that guides how end users make use of cloud services** by defining and creating policies to :

- Control costs
- Minimize security risks
- Improve efficiency
- Accelerate deployment

Governance and Compliance regulations for AI Systems

Compliance in AWS refers to adhering to various **industry-specific** and **general data protection regulations and standards**.

- Industry-Specific Regulations:
 - HIPAA
 - PCI DSS for payment security.
- General Data Protection Regulations:
 - Laws like the GDPR and CCPA
- Other Compliance Programs:
 - SOC 1/2/3 for service organization controls
 - FedRAMP

Regulatory Compliance Standards for AI systems

- **AI Act European Union (EU AI Act)**
 - The AI Act is the first-ever legal framework on AI, which addresses the risks of AI – [Link](#)
- **Algorithmic Accountability Act of 2023**
 - The act requires companies to assess the impacts of the AI systems they use and sell - [Link](#)
- **GDPR European Union**
 - Applies to the processing of personal data within the EU.
 - Provides specific data protection guidelines for transparency, liability, ethics, bias, and risk.

Regulatory Compliance Standards for AI systems

- **International Organization for Standardization [ISO]**
 - Global network of the world's leading standardizers - [Link](#)
 - AI Standards- ISO/IEC 42001:2023 - [Link](#)
- **System and Organization Controls [SOC]**
 - SOC 1: For financial reporting - [Link](#)
 - SOC 2: For design and operational effectiveness of internal controls
 - SOC 3: For presenting SOC 2 report information to the general public



Scheduling the Exam

&

Saving \$\$

AWS Certified AI Practitioner – Exam Overview (beta)

Category	Foundational
Exam duration	120 minutes
Exam format	85 questions
Cost	USD 75 USD 37.5 (Discount coupon from previous certification)
Candidate role examples	Business analyst, IT support, marketing professional, product or project manager, sales professional



Exam Link – [Schedule an exam](#) (AWS Certified AI Practitioner AIF-C01)



My Exam Experience :

Exam Focus Areas

AWS Certified AI Practitioner – Focus Areas

📌 **High focus areas** (10-15 questions):

- Lot of questions on Amazon Bedrock (serverless offering, pricing, model evaluation, logging and monitoring)

📌 **Medium focus areas** (4-5 questions):

- **Amazon SageMaker** - Data Wrangler, Amazon SageMaker Feature Store, Amazon SageMaker Model Monitor, SageMaker Clarify
- RAG, Bedrock Knowledge Bases and Agents
- Basic GenAI terms – context window, tokens, parameters etc.
- Responsible AI – Hallucinations, Bias, SageMaker Clarify, Model Monitor, Bedrock Guardrails and few other services
- Key Security Services – Macie, IAM, PrivateLink - Gateway and Interface Endpoints, Audit Manager and few other compliance, governance services

AWS Certified AI Practitioner – Focus Areas

- Few Questions on Amazon Q and SageMaker Jumpstart
- Inference parameters impact on model responses - Temperature, Top P, Top K

Low focus areas – 2-3 questions

- Relevant Metrics – ROUGE, BLUE
- Training a foundation model (for example, pre-training, fine-tuning, continuous pre-training).
- Prompt Engineering Techniques and Security
- Generative AI Security Scoping Matrix
- ML Development Lifecycle and AWS Tools

Thank You