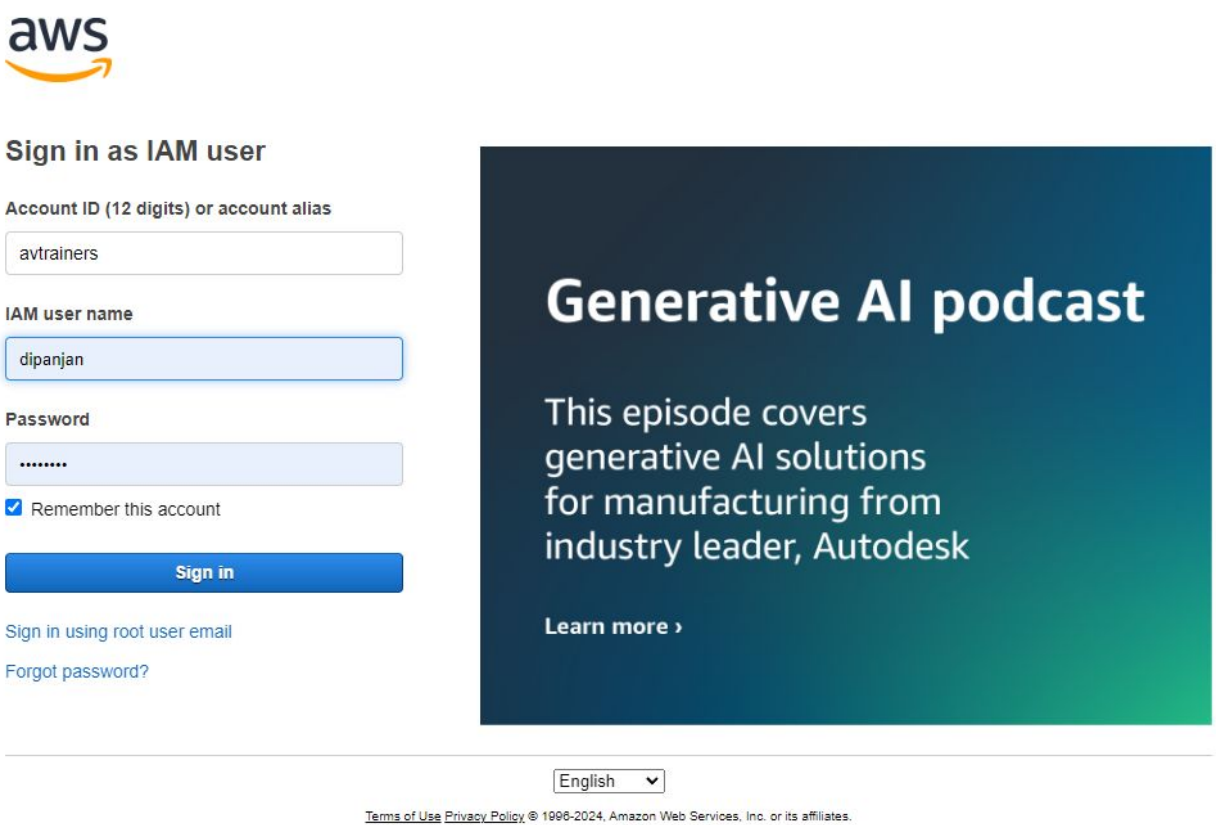
Generative AI on AWS

Prashant Sahu

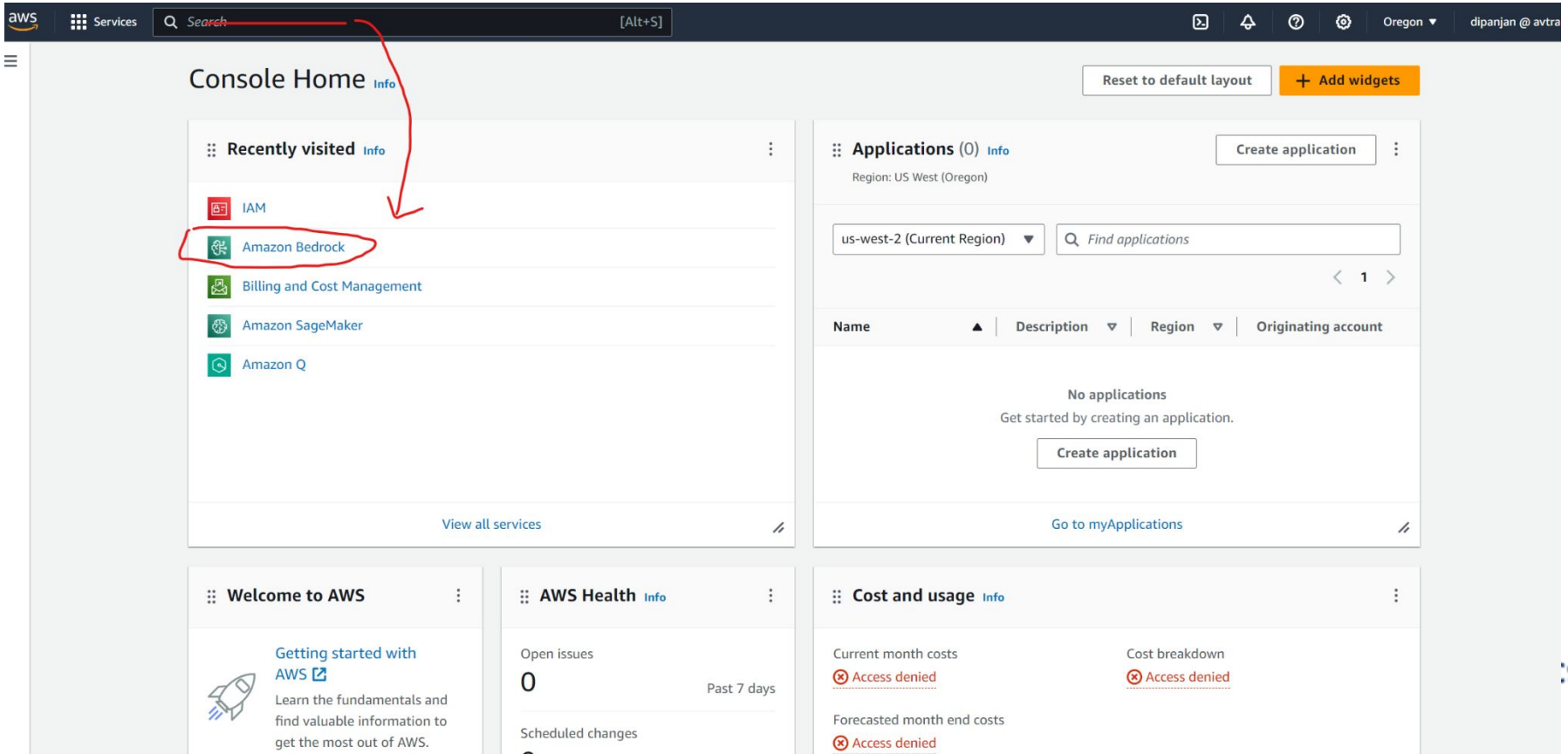
Analytics Vidhya

Create an account and sign-in

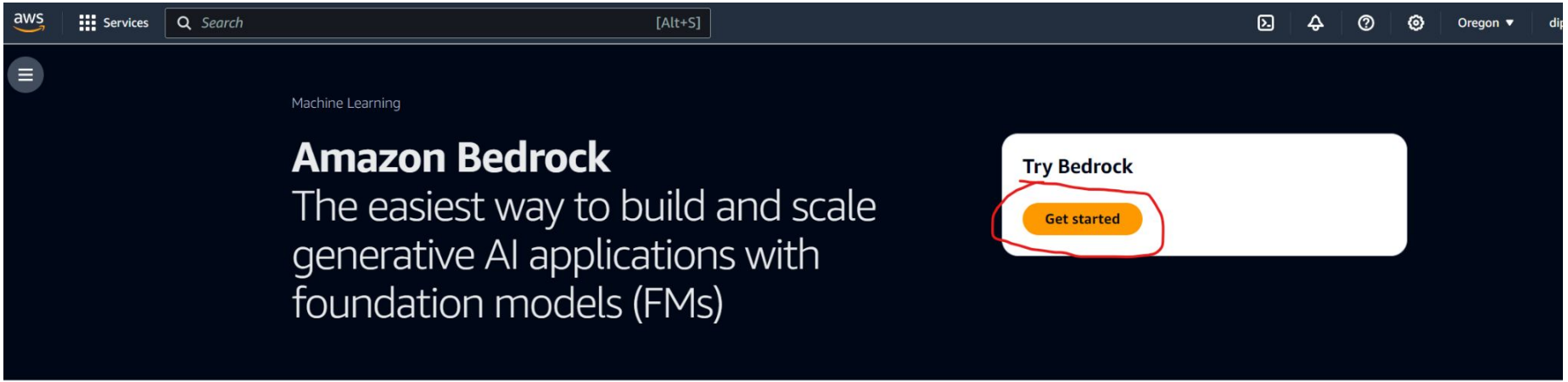


Analytics Vidhya

Search in Services and go to the Amazon Bedrock Page



Click on Get Started



Overview

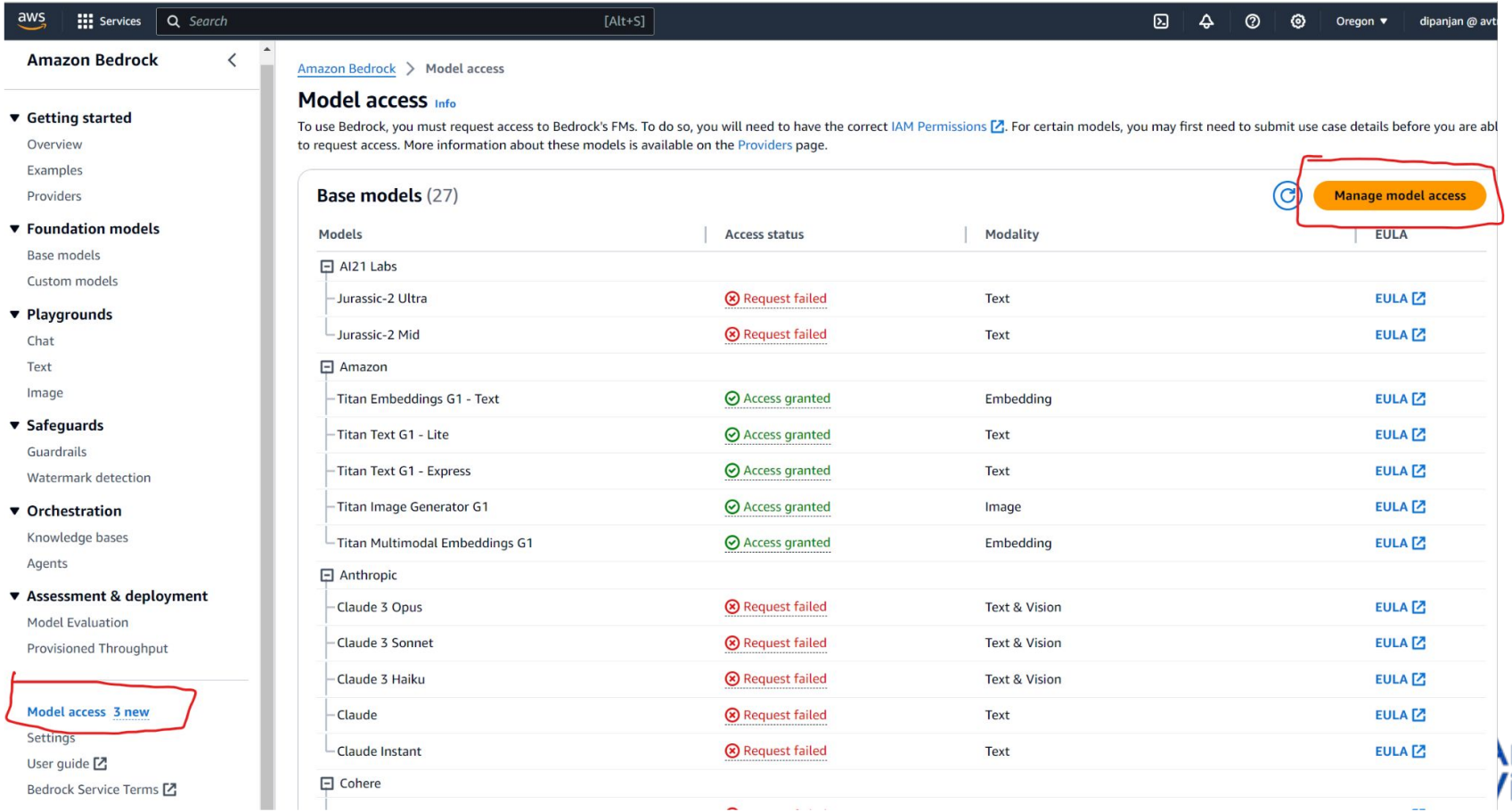
Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your use case. With Bedrock's serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools without having to manage any infrastructure.

Benefits

Accelerate development of generative AI applications using FMs through an API, without managing infrastructure.

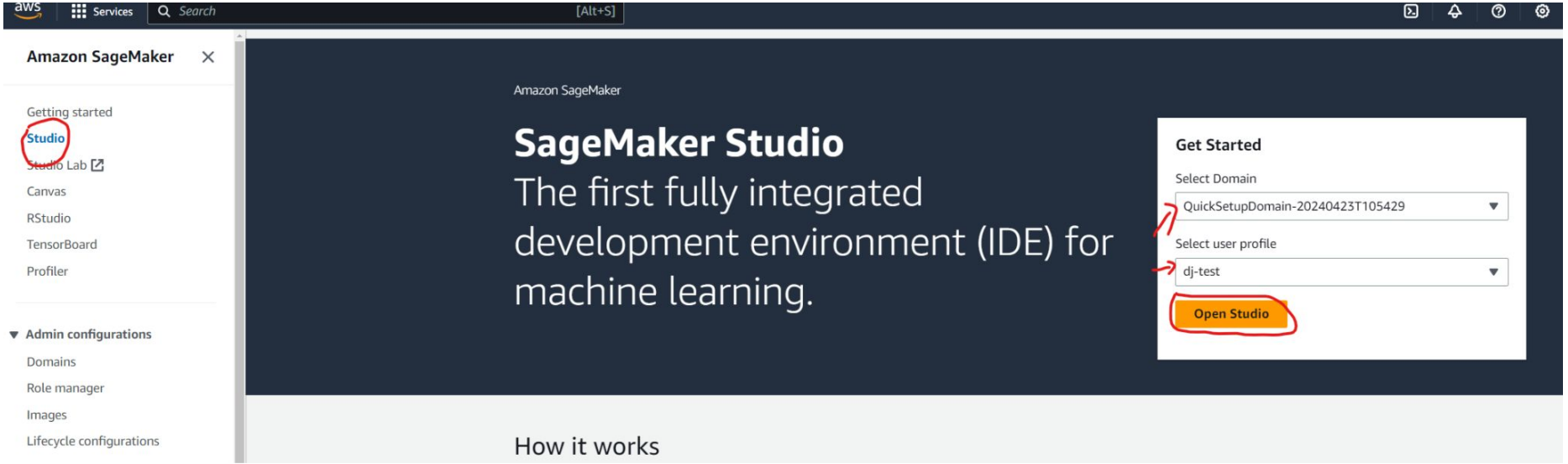
Choose FMs from AI21 Labs, Anthropic, Stability AI, and Amazon to find the right FM for your use case.

Model Access -> Manage Model Access



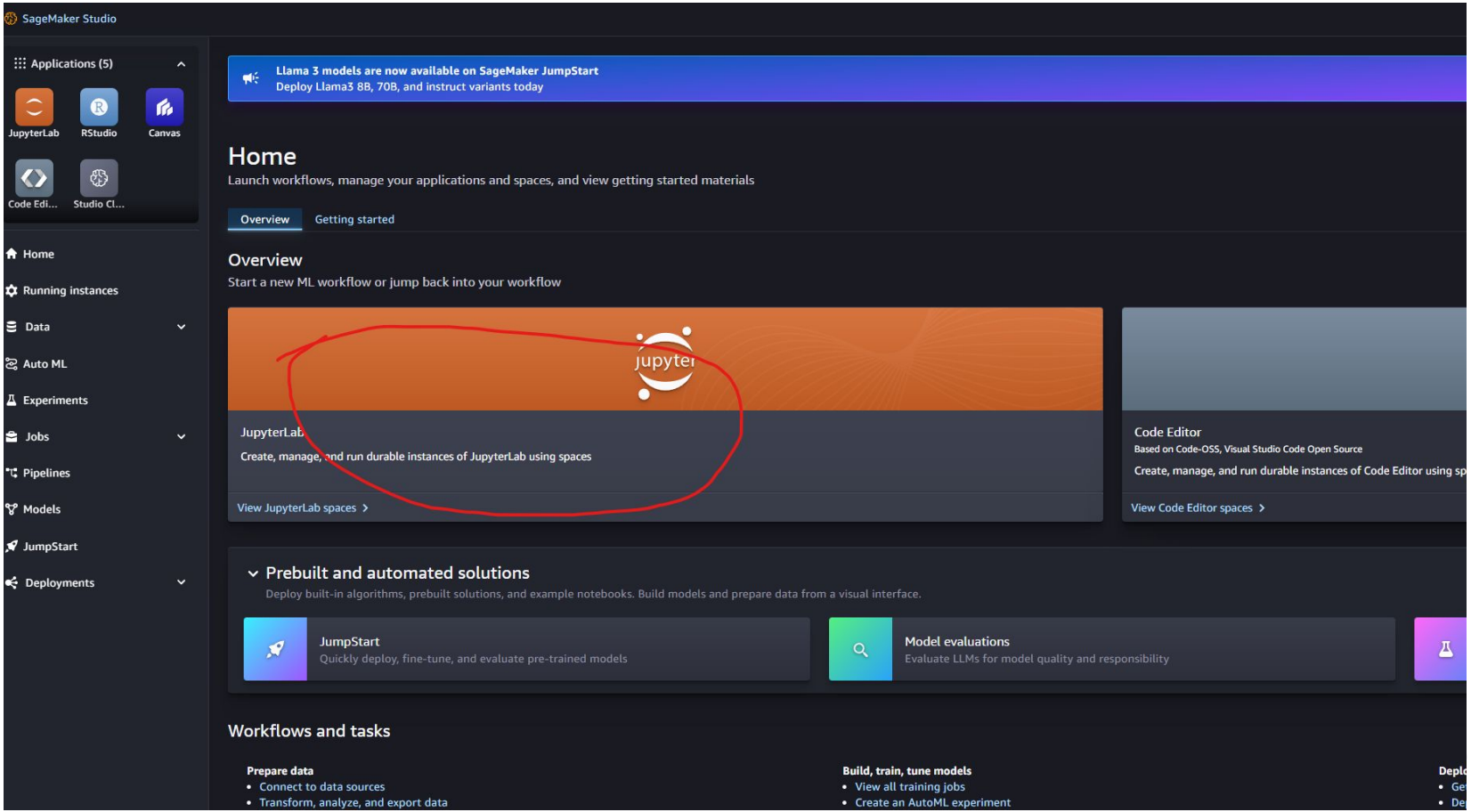
Analytics Vidhya

Open Sagemaker Studio based on your previously created domain and user



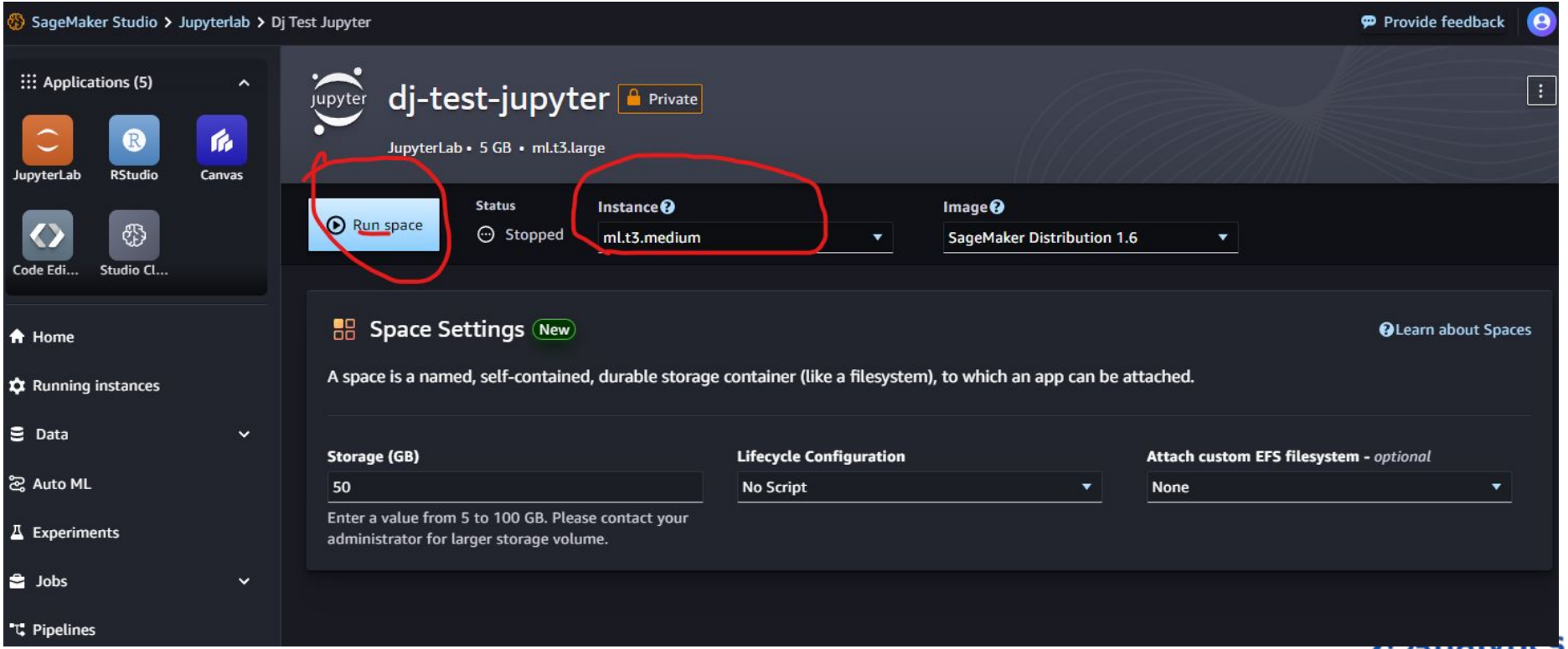
Analytics Vidhya

Go to Jupyterlab

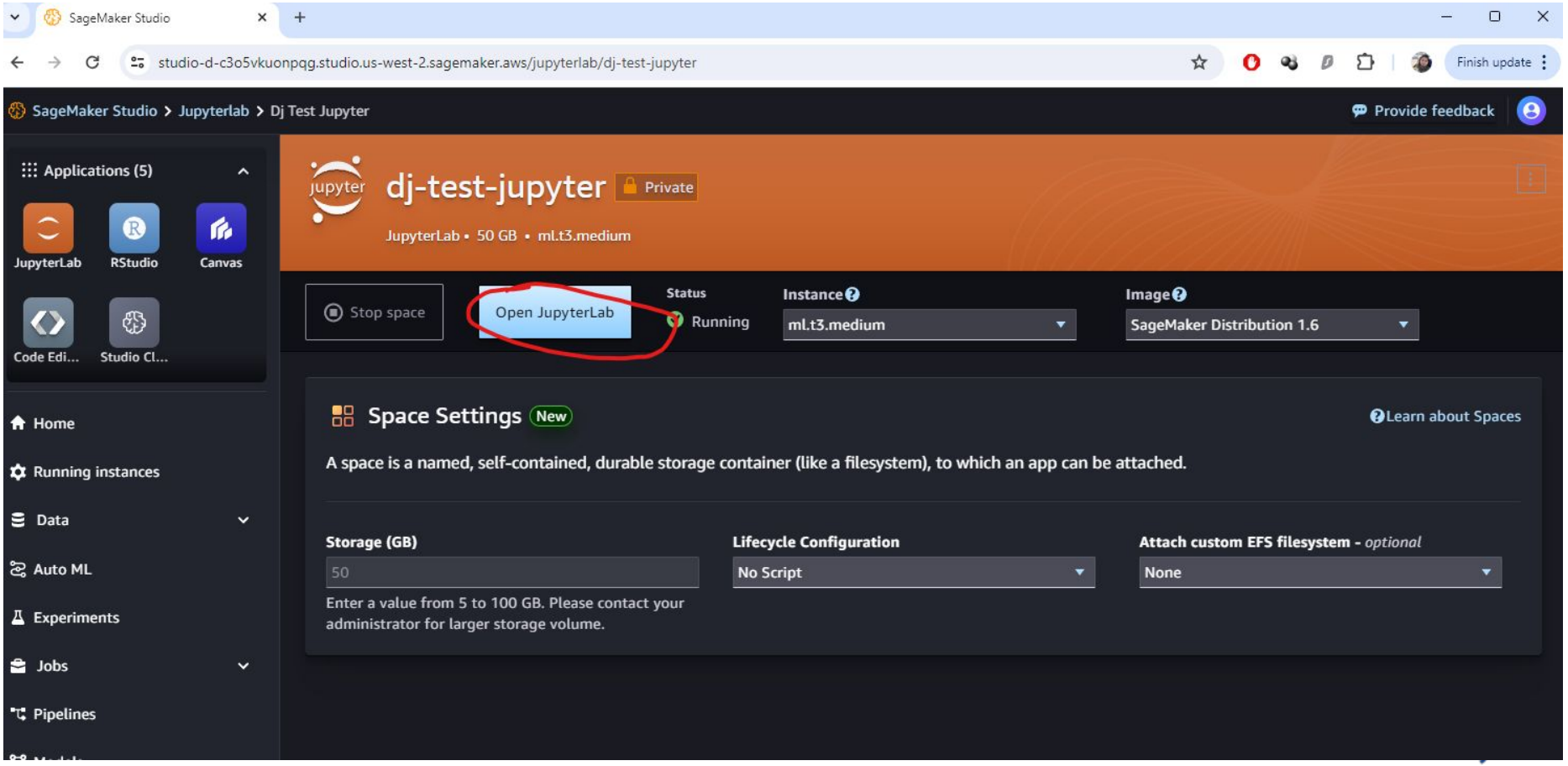


Analytics Vidhya

Select ml.t3.medium, >20GB storage and Run Space



Start Jupyterlab



Generative AI on AWS

What is generative AI?

AI that can produce original content close enough to human generated content for real-world tasks

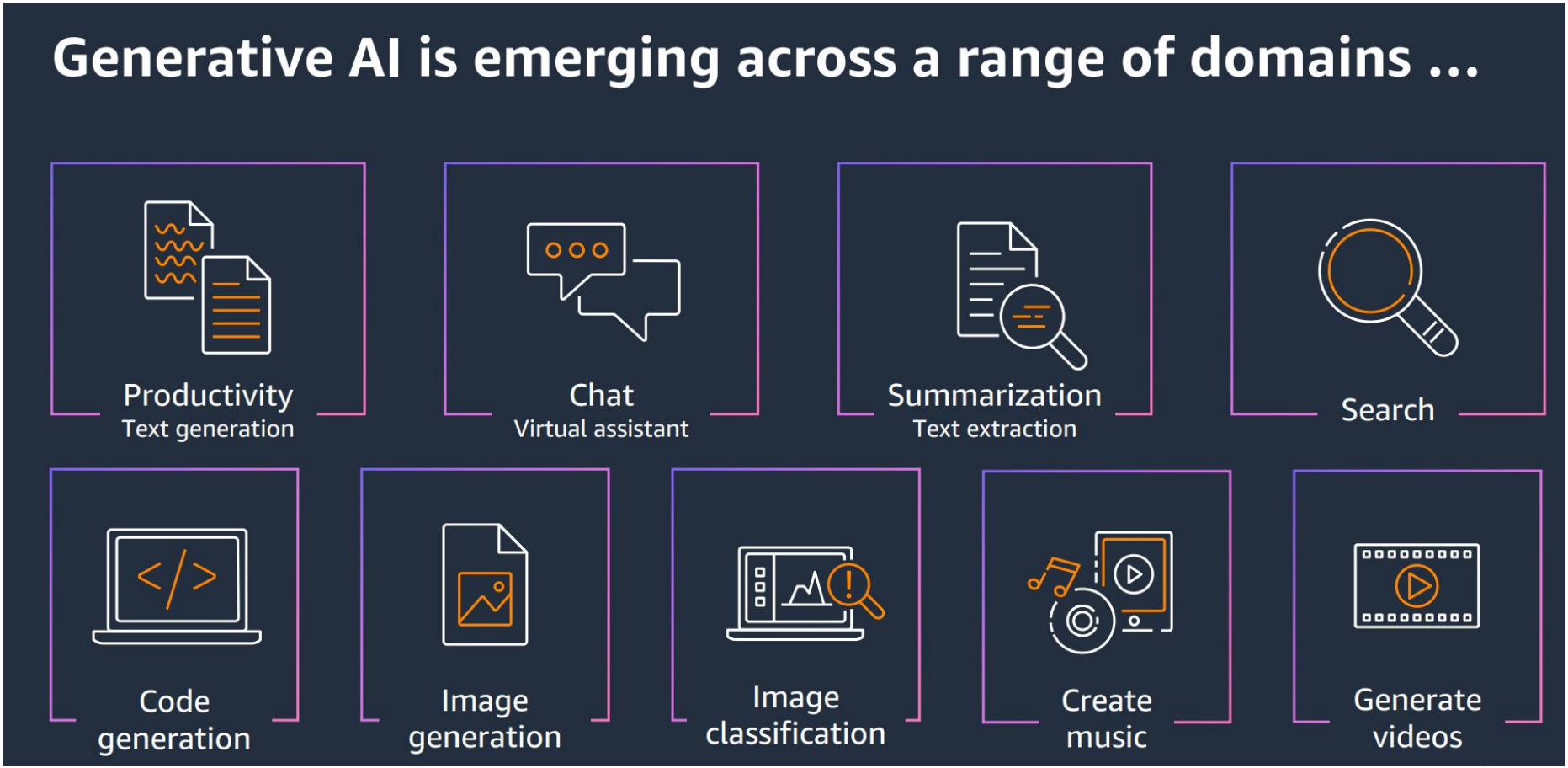
Powered by foundation models pre-trained on large sets of data with several hundred billion parameters

Tasks can be customized for specific domains with minimal fine-tuning

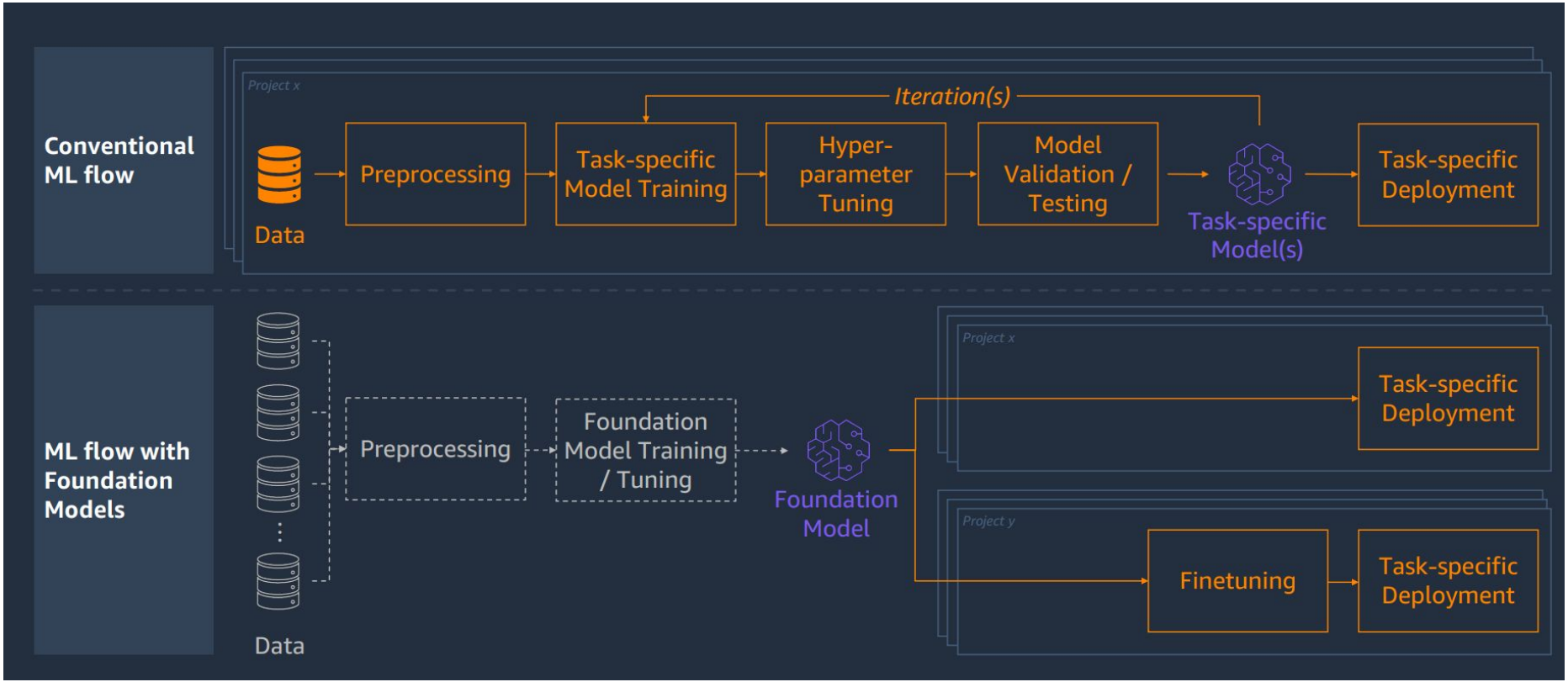
Applicable to many use cases like text summarization, question answering, digital art creation, code generation, etc.

Reduces time and cost to develop ML models and innovate faster

Generative AI on AWS



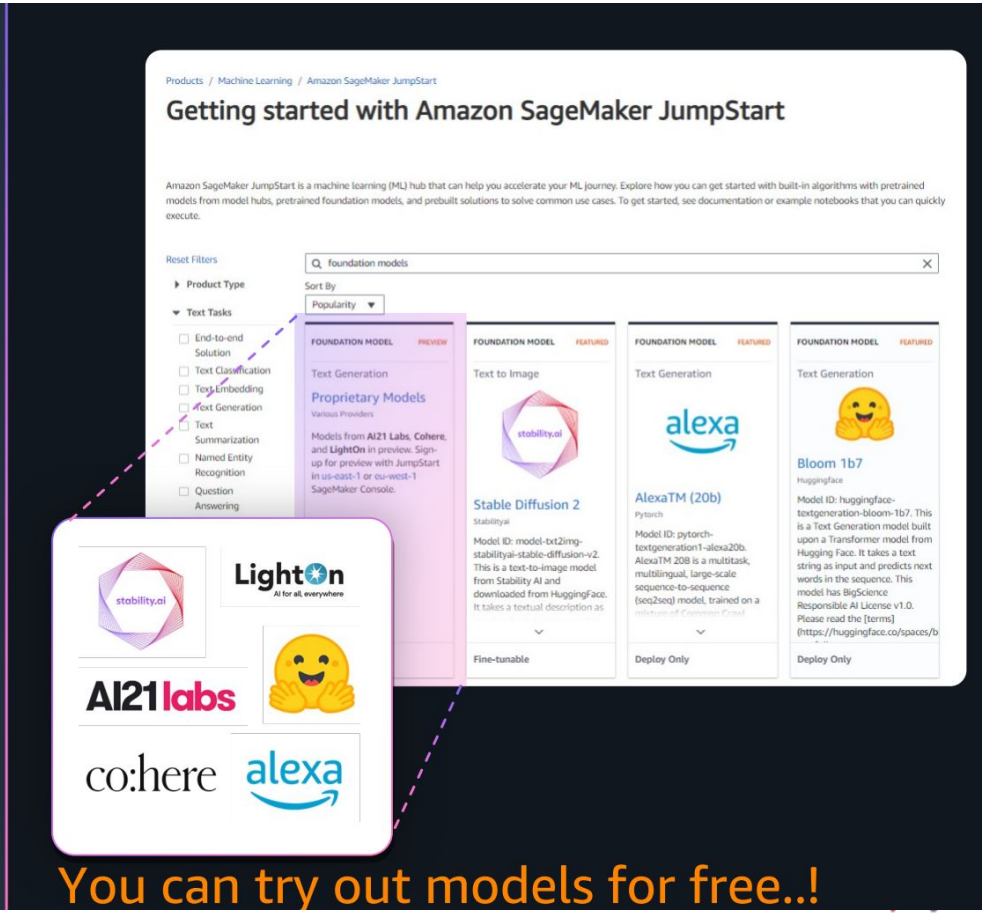
AWS Foundation Models



Analytics Vidhya

Gen AI on AWS

Option 1: Foundation models available through SageMaker JumpStart



You can try out models for free..!

aws

Analytics Vidhya

Gen AI on AWS

SageMaker JumpStart

Easily access ML assets and quickly bring ML applications to market

Machine Learning Hub for SageMaker
Browse through ~400 contents including, built-in algorithms with pre-trained models, (New) Foundation Models, solution templates, and example notebooks

Pre-built training and inference scripts
Compatible with SageMaker and configurable with custom dataset

UI as well as API based machine learning
Use UI for single click model deployment or API for Python SDK based workflow

Notebook with examples
JumpStart lets you jump into notebook to use selected model with examples to guide customers through entire ML workflow

(New) Share and collaborate within an organization
Share models and notebooks with others within your organization, and allow them to train with their own data or deploy as-is for inferencing

aws

Analytics Vidhya

Gen AI on AWS

Foundation models with SageMaker JumpStart: How it works

Amazon SageMaker JumpStart

Access and try out public and proprietary FMs, and customize and integrate them into your generative AI applications

Browse

Browse public and proprietary FMs

Experiment

Experiment with FMs before choosing a model for deployment

Customize

Customize selected FM with your own dataset without training from scratch

Deploy

Deploy the model and run inference for your generative AI use case

Gen AI on AWS

Foundation models available on SageMaker JumpStart for self-managed access

Publicly available

stability.ai

Models
Text2Image
Upscaling

Tasks
Generate photo-realistic images from text input
Improve quality of generated images

Features
Fine-tuning on SD 2.1 model

alexa

Models
AlexaTM 20B

Tasks
Machine translation
Question answering
Summarization
Annotation
Data generation

Models
Flan T-5 models (8 variants)
DistilGPT2, GPT2
Bloom models (3 variants)

Tasks
Machine translation
Question answering
Summarization
Annotation
Data generation

Proprietary models

co:here

Models
Cohere generate-med

Tasks
Text generation
Information extraction
Question answering
Summarization

Lighton

Models
Lyra-Fr 10B

Tasks
Text generation
Keyword extraction
Information extraction
Question answering
Summarization
Sentiment analysis
Classification

AI21labs

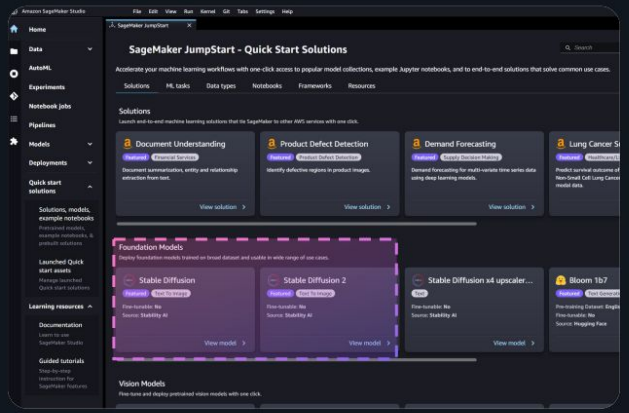
Models
Jurassic-1 Grande 17B

Tasks
Text generation
Long-form generation
Summarization
Paraphrasing
Chat
Information extraction
Question answering
Classification

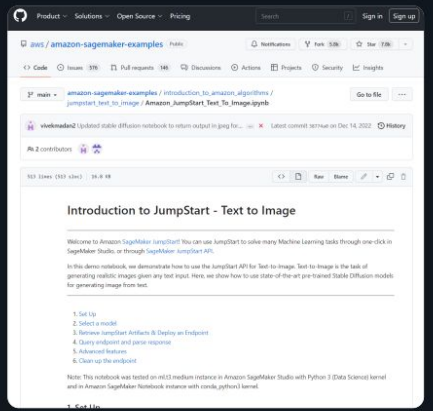
Gen AI on AWS

3 ways to use foundation models with SageMaker JumpStart

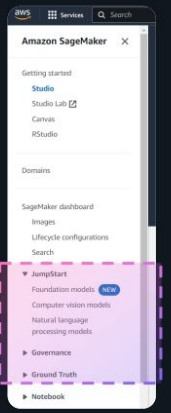
SageMaker Studio
One-step deploy



SageMaker Notebooks



AWS Management Console Preview



Gen AI on AWS

Key factors in decision making

Cost

Optimize for cost with a variety of models, size, and instance for your needs with AWS pay-as-you go pricing

Accuracy

Use highly accurate models per HELM benchmarks

Speed (latency)

Optimize for performance with different model sizes and instance types

Ease of use

Instantly try models in playground; deploy with SageMaker using managed inference scripts

Data security

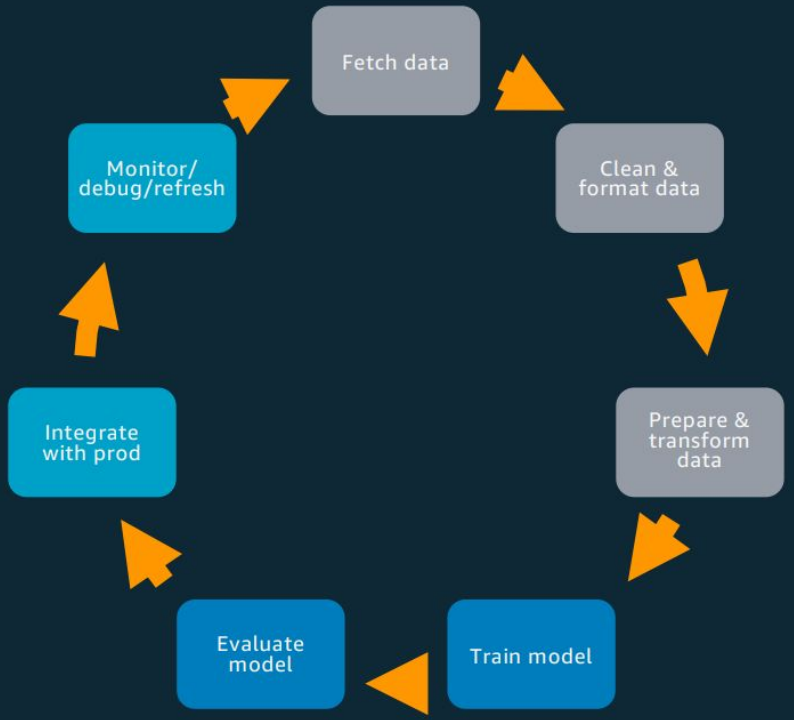
Host models on customer-dedicated endpoints inside your VPC

Gen AI on AWS

Analytics Vidhya

Gen AI on AWS

Machine learning process is hard…



1. Data wrangling

- Setup and manage Notebook environments
- Get data to notebooks securely

2. Experimentation

- Setup and manage clusters
- Scale/distribute ML algorithms

3. Deployment

- Setup and manage inference clusters
- Manage and auto scale inference APIs
- Testing, versioning, and monitoring

Gen AI on AWS

Amazon SageMaker

Build, train, and deploy machine learning models at scale

End-to-End Machine Learning Platform

Zero setup

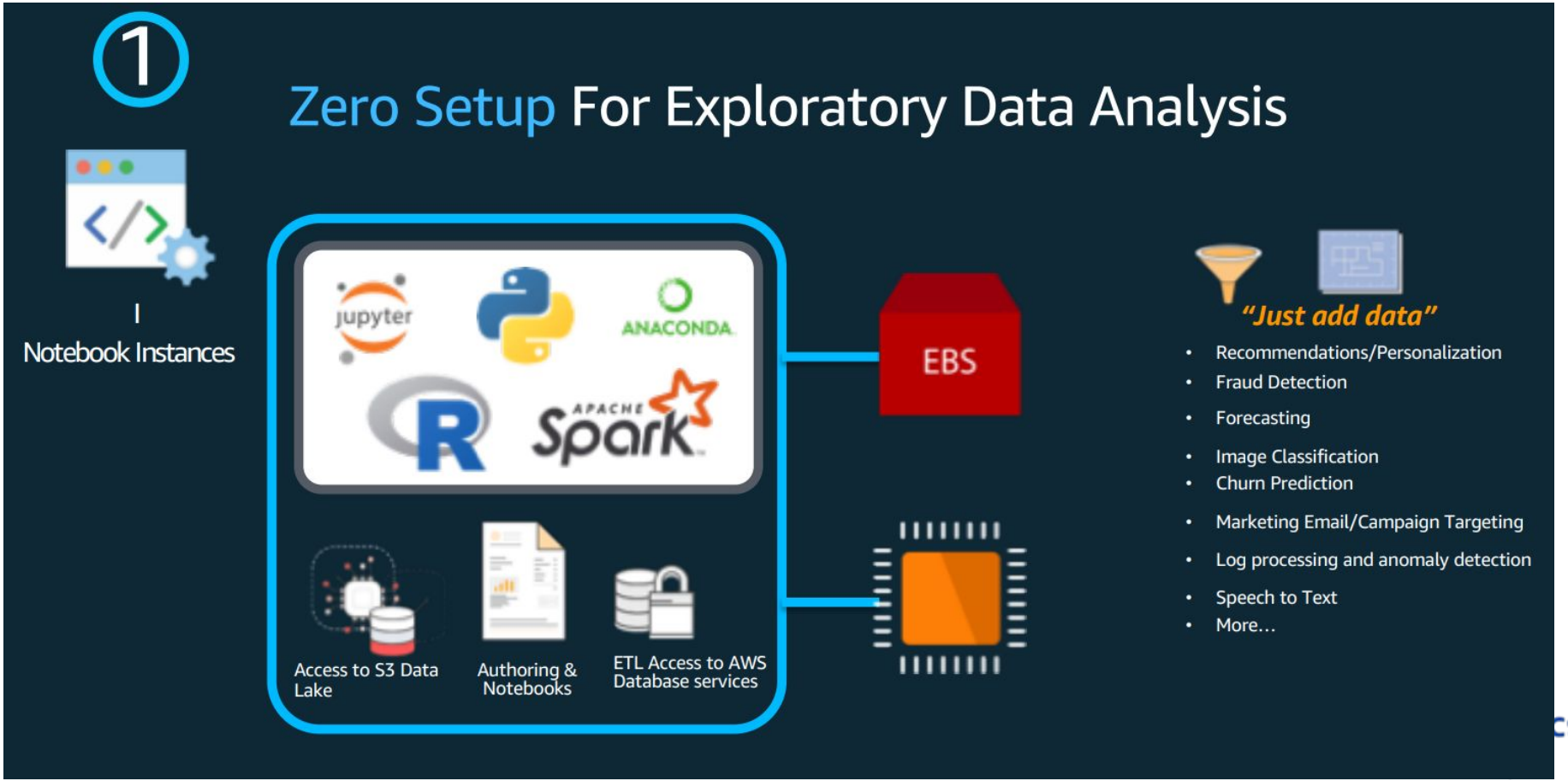
Flexible Model Training

Pay by the second

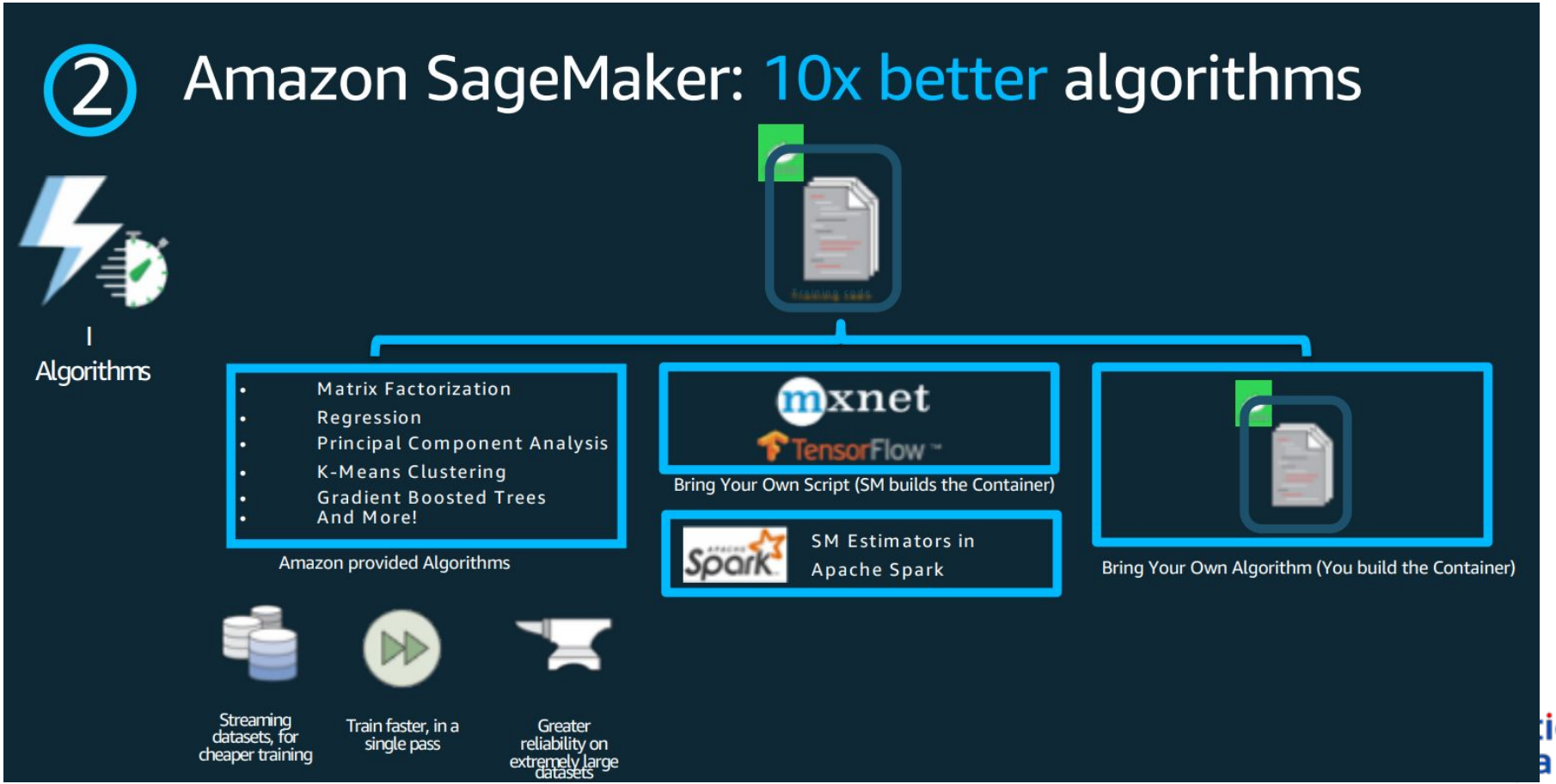
Gen AI on AWS


Amazon SageMaker
1
2
3
4
Notebook Instances
Algorithms
ML Training Service
ML Hosting Service

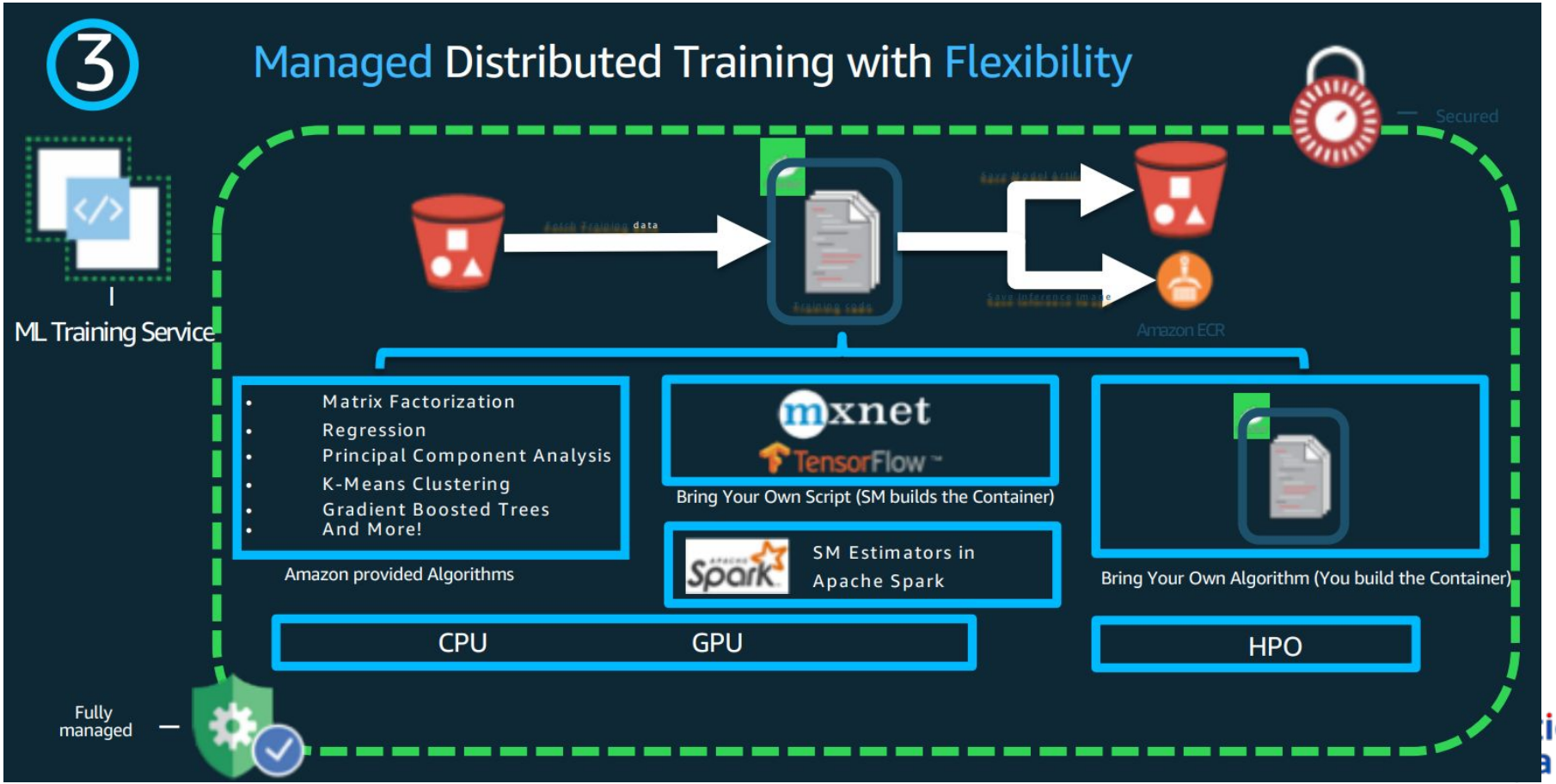
Gen AI on AWS


1
Zero Setup For Exploratory Data Analysis
Notebook Instances
jupyter
ANACONDA
R
APACHE Spark
EBS
Access to S3 Data Lake
Authoring & Notebooks
ETL Access to AWS Database services
"Just add data"
Recommendations/Personalization
Fraud Detection
Forecasting
Image Classification
Churn Prediction
Marketing Email/Campaign Targeting
Log processing and anomaly detection
Speech to Text
More…

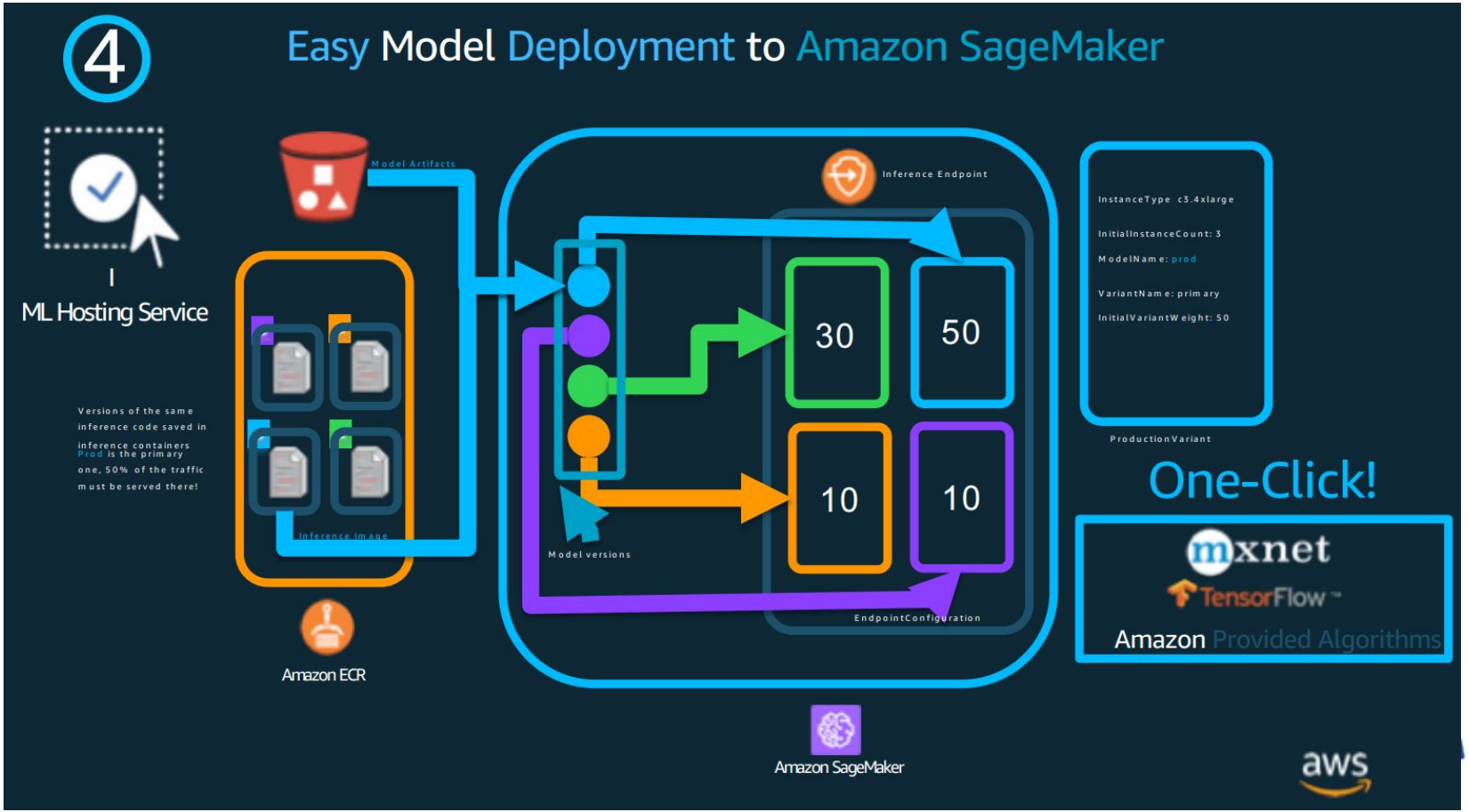
Gen AI on AWS


2
Amazon SageMaker: 10x better algorithms
Algorithms
Matrix Factorization
Regression
Principal Component Analysis
K-Means Clustering
Gradient Boosted Trees
And More!
Amazon provided Algorithms
mxnet
TensorFlow
Bring Your Own Script (SM builds the Container)
Spark
SM Estimators in Apache Spark
Bring Your Own Algorithm (You build the Container)
Streaming datasets, for cheaper training
Train faster, in a single pass
Greater reliability on extremely large datasets

Gen AI on AWS


3
Managed Distributed Training with Flexibility
Secured
ML Training Service
data
Amazon ECR
Matrix Factorization
Regression
Principal Component Analysis
K-Means Clustering
Gradient Boosted Trees
And More!
Amazon provided Algorithms
mxnet
TensorFlow
Bring Your Own Script (SM builds the Container)
Spark
SM Estimators in
Apache Spark
Bring Your Own Algorithm (You build the Container)
CPU
GPU
HPO
Fully
managed

Gen AI on AWS



Analytics Vidhya

Gen AI on AWS

NEW

Option 2 - Amazon Bedrock

The easiest way to build and scale generative AI applications with FMs

Analytics Vidhya

Gen AI on AWS

Amazon Bedrock key benefits

Accelerate development of generative AI applications using FMs through an API, without managing infrastructure

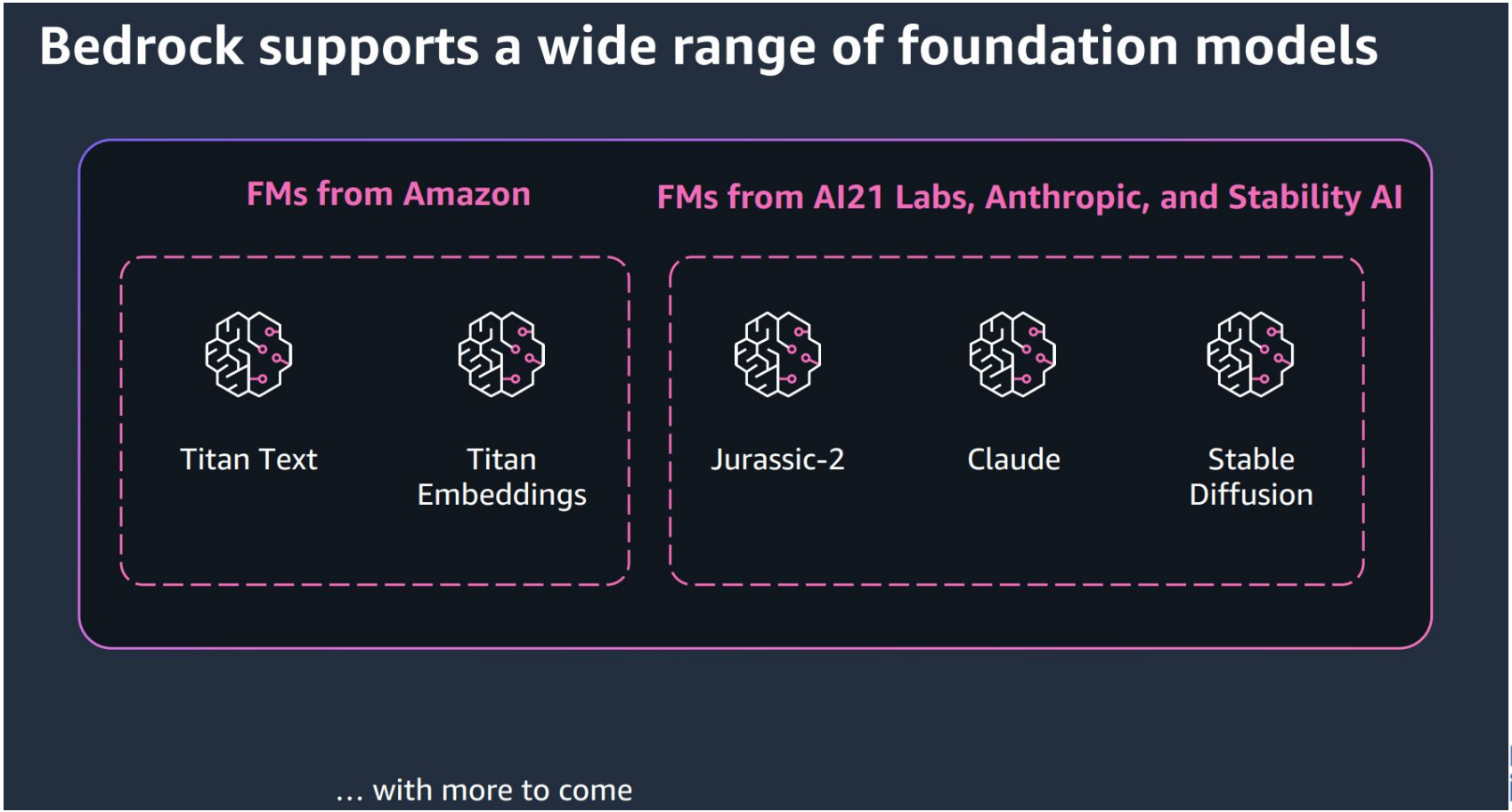
Choose FMs from AI21 Labs, Anthropic, Stability AI, and Amazon to find the right FM for your use case

Privately customize FMs using your organization's data

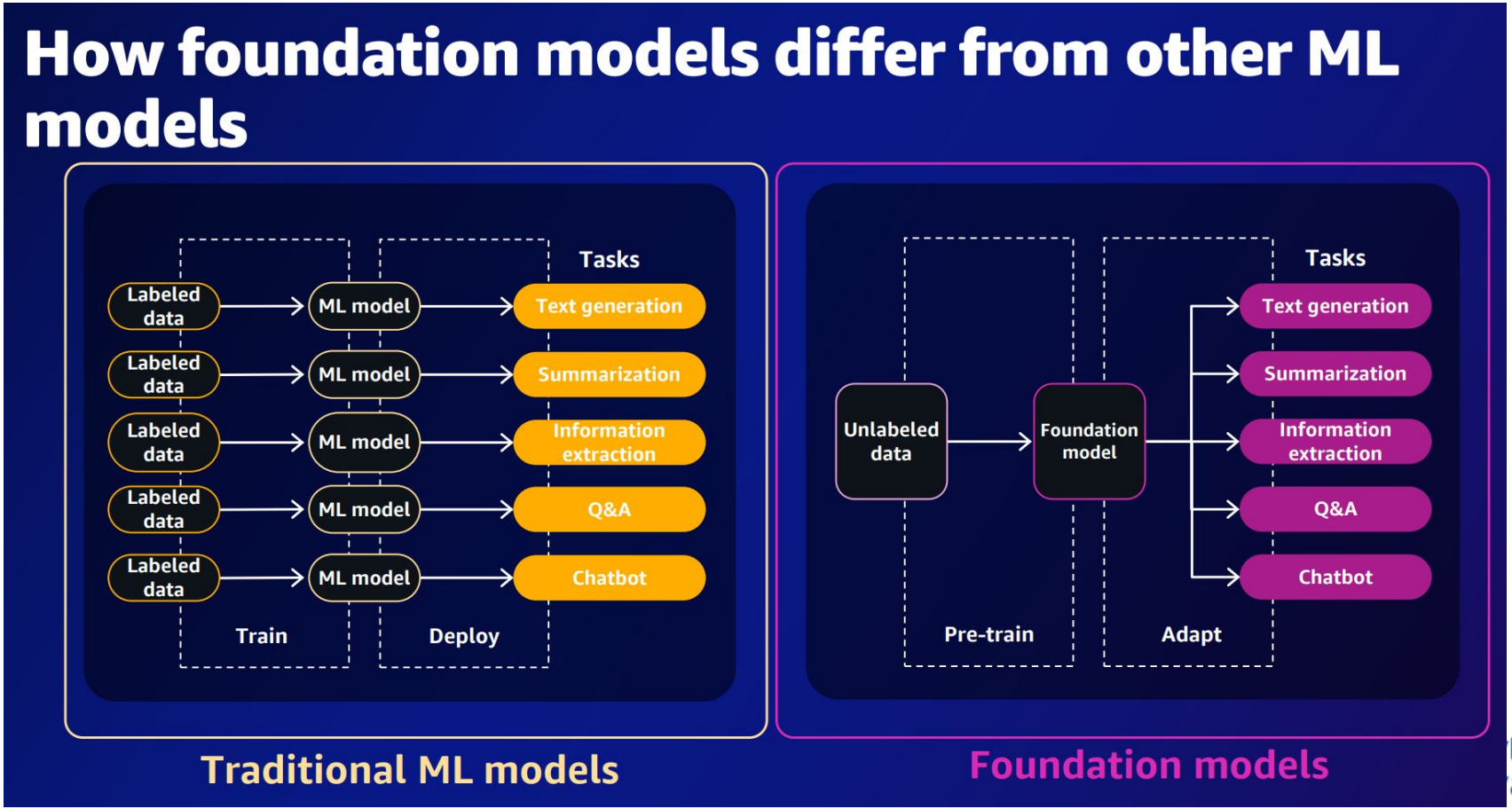
Enhance your data protection using comprehensive AWS security capabilities

Use AWS tools and capabilities that you are familiar with to deploy scalable, reliable, and secure generative AI applications

Gen AI on AWS



Gen AI on AWS


How foundation models differ from other ML models
Tasks
Labeled data
ML model
Text generation
Labeled data
ML model
Summarization
Labeled data
ML model
Information extraction
Labeled data
ML model
Q&A
Labeled data
ML model
Chatbot
Train
Deploy
Traditional ML models
Unlabeled data
Foundation model
Tasks
Text generation
Summarization
Information extraction
Q&A
Chatbot
Pre-train
Adapt
Foundation models

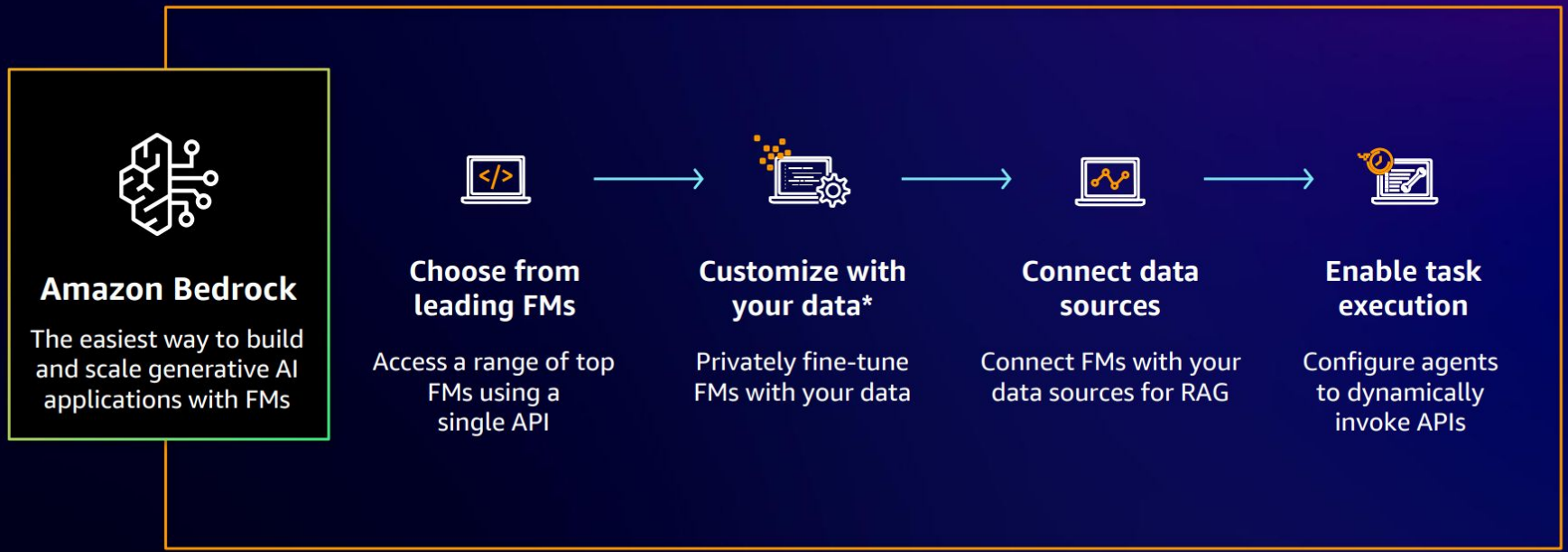
Gen AI on AWS

Types of foundation models

Input	FM	Output
"Summarize articles on the impact of walking on heart health"	**Text-to-text** Generate text from natural language prompts	"Ten thousand steps per day is optimum for maintaining a healthy heart"
"hand soap"	**Text-to-embeddings** Generate numerical representation of text	Numerical representation of "Hand soap refills Hand soap dispenser Hand soap antibacterial"
"a photo of an astronaut riding a horse on mars"	**Multimodal** Create and edit images using natural language prompts	

Gen AI on AWS

How does Amazon Bedrock work?

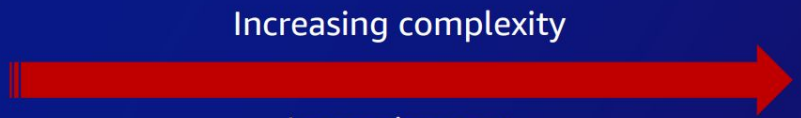


*your content is not used to improve the base models and is not shared with third-party model providers.

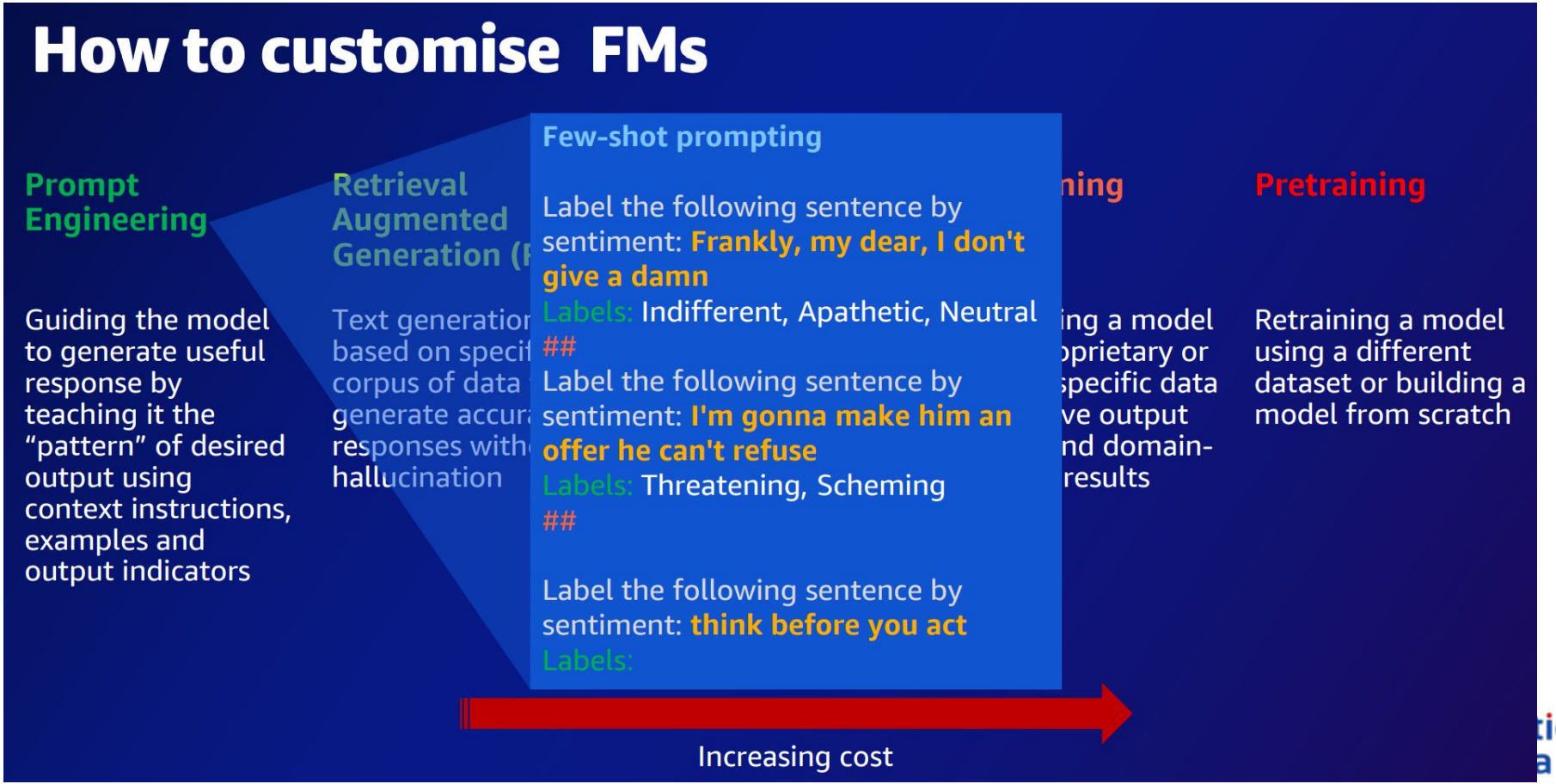
Gen AI on AWS

How to customise FMs

Prompt Engineering	Retrieval Augmented Generation (RAG)	Instruction fine-tuning	Fine-tuning	Pretraining
Guiding the model to generate useful response by teaching it the “pattern” of desired output using context instructions, examples and output indicators	Text generation based on specific corpus of data to generate accurate responses without hallucination	Fine-tuning a language model on a collection of tasks improves the zero-shot performance of language models on unseen tasks	Fine-tuning a model using proprietary or domain specific data to improve output quality and domain-relevant results	Retraining a model using a different dataset or building a model from scratch



Gen AI on AWS


How to customise FMs
Prompt Engineering
Guiding the model to generate useful response by teaching it the "pattern" of desired output using context instructions, examples and output indicators
Retrieval Augmented Generation (R
Text generation based on specif corpus of data generate accura responses with hallucination
Few-shot prompting
Label the following sentence by sentiment: Frankly, my dear, I don't give a damn
Labels: Indifferent, Apathetic, Neutral
##
Label the following sentence by sentiment: I'm gonna make him an offer he can't refuse
Labels: Threatening, Scheming
##
Label the following sentence by sentiment: think before you act
Labels:
ning
ing a model oprietary or specific data ve output nd domain- results
Pretraining
Retraining a model using a different dataset or building a model from scratch
Increasing cost

Gen AI on AWS

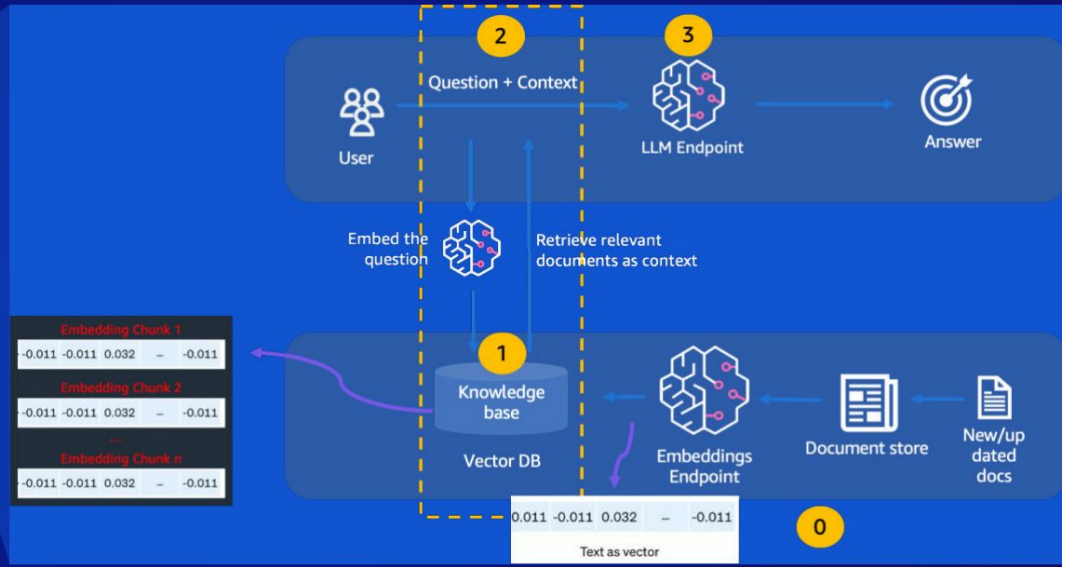
How to customise FMs

Prompt Engineering

Guiding the model to generate useful response by teaching it the “pattern” of desired output using context instructions, examples and output indicators

Retrieval Augmented Generation (RAG)

Text generation based on specific corpus of data to generate accurate responses without hallucination



Increasing complexity

Increasing cost

Gen AI on AWS

How to customise FMs

Prompt Engineering

Guiding the model to generate useful response by teaching it the "pattern" of desired output using context instructions, examples and output indicators

Retrieval Augmented Generation (RAG)

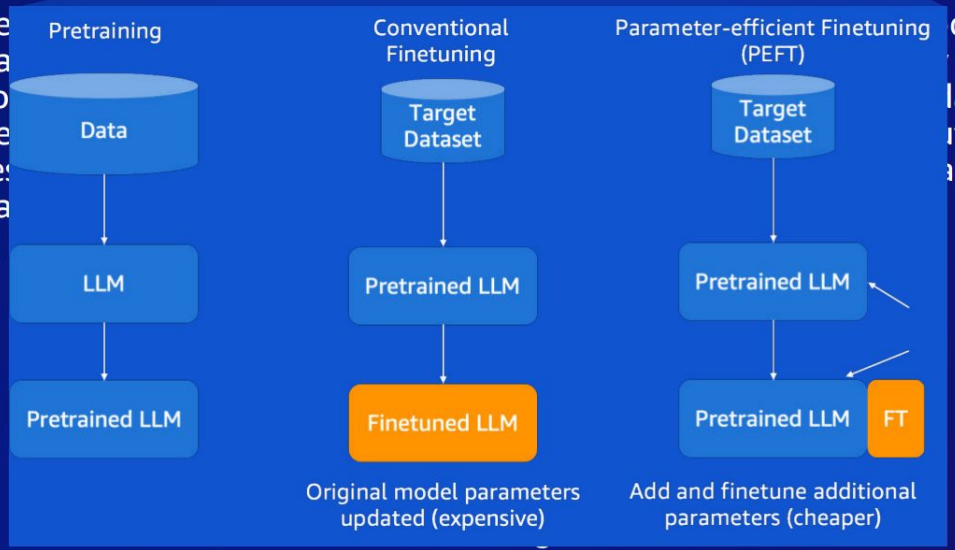
Te
ba
co
ge
re
ha

Instruction fine-tuning

Fine-tuning

Pretraining

Retraining a model using a different dataset or building a model from scratch



Gen AI on AWS

Amazon Bedrock

THE EASIEST WAY TO BUILD AND SCALE GENERATIVE AI APPLICATIONS WITH FMS

- Accelerate development of generative AI applications using FMs through an API
- No need to manage infrastructure
- Privately customize FMs using your organization's data
- Comprehensive AWS security capabilities
- Enable generative AI apps to complete tasks in just a few clicks using **agents** for Bedrock

amazon	AI21labs	ANTHROPIC	cohere	stability.ai
Amazon Titan	**Jurassic-2**	*new* **Claude 2**	*new* **Command + Embed**	*new* **Stable Diffusion XL 1.0**
Text summarization, generation, classification, open-ended Q&A, information extraction, embeddings, and search	Multilingual LLMs for text generation in Spanish, French, German, Portuguese, Italian, and Dutch	LLM for thoughtful dialogue, content creation, complex reasoning, creativity, and coding, based on Constitutional AI and harmlessness training	Text generation model for business applications and embeddings model for search, clustering, or classification in 100+ languages	Generation of unique, realistic, high-quality images, art, logos, and designs

Gen AI on AWS

Enabling foundation models to complete tasks

1

Define instructions and orchestration

2

Configure FM to access data sources

3

Complete actions with API calls

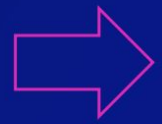
4

Manage cloud hosting and security

Gen AI on AWS

Bedrock core API: InvokeModel

```
bedrock.invoke_model(
        modelId = model_id,
        contentType = "…",
        accept = "…",
        body = body)
```



Access foundation models

- Amazon Titan models
- Third-party models
- Fine-tuned models

```
prompt_data = """Write me a blog about making strong business decisions as a leader"""

config = {"maxTokenCount":512,"stopSequences":[],"temperature":0.5,"topP":0.9}
body = json.dumps({"inputText":prompt_data,"textGenerationConfig": config})
modelId = "amazon.titan-tg1-large"
accept = "*/*"
contentType = "application/json"
response = bedrock.invoke_model(
    body=body, modelId=modelId, accept=accept, contentType=contentType
)
response_body = json.loads(response.get("body").read())
print(response_body.get("results")[0].get("outputText"))
```

Titan Text

```
prompt_data = """Write me a blog about making strong business decisions as a leader"""

body = json.dumps({"prompt": prompt_data,
                   "max_tokens_to_sample": 300,
                   "temperature": 0.5,
                   "top_k": 250,
                   "top_p": 1,
                   "stop_sequences": ["\n\nHuman:"]})
modelId = "anthropic.claude-instant-v1"
accept = "*/*"
contentType = "application/json"
response = bedrock.invoke_model(
    body=body, modelId=modelId, accept=accept, contentType=contentType
)
response_body = json.loads(response.get("body").read())
print(response_body.get("completion"))
```

Anthropic Claude

Gen AI on AWS

Data protection, privacy, and security

Data protection and privacy

- Your data used with Bedrock is not used for service improvement and not shared with third party model providers
- Private connectivity between Amazon Bedrock service and your virtual private cloud (VPC)
- Your data is encrypted in transit and at rest
- Customize FMs privately, retaining control over how your data is used and encrypted

Secure Generative AI application

- Use AWS security services to form your defense-in-depth security strategy
- Your customized FMs are encrypted using AWS Key Management Service (AWS KMS) keys and stored encrypted
- Control access to your customized FMs using AWS Identity and Access Management Service (IAM)

Gen AI on AWS

Model tenancy

Single-tenant endpoint

1. Deployment available to a single customer
2. Holds a single version of a baseline 1P/3P model that has been fine-tuned by a customer

Multi-tenant endpoint

1. Deployment available to all customers
2. Holds a baseline version of each supported 1P/3P model

3. No inference request's input or output text is used to train any model(s) in the deployment
4. Model deployments are inside an AWS account owned and operated by the Bedrock service team
5. Model vendors have no access to any customer data

Amazon Bedrock private fine-tuning

Organization-specific information

- Deliver more contextual and personalized responses by incorporating your organization's data
- Create secure integration with your organization data sources without retraining the FM
- Agents identify the data sources, retrieve the relevant information, incorporate the information into user query, and provide a response

Customized search capabilities

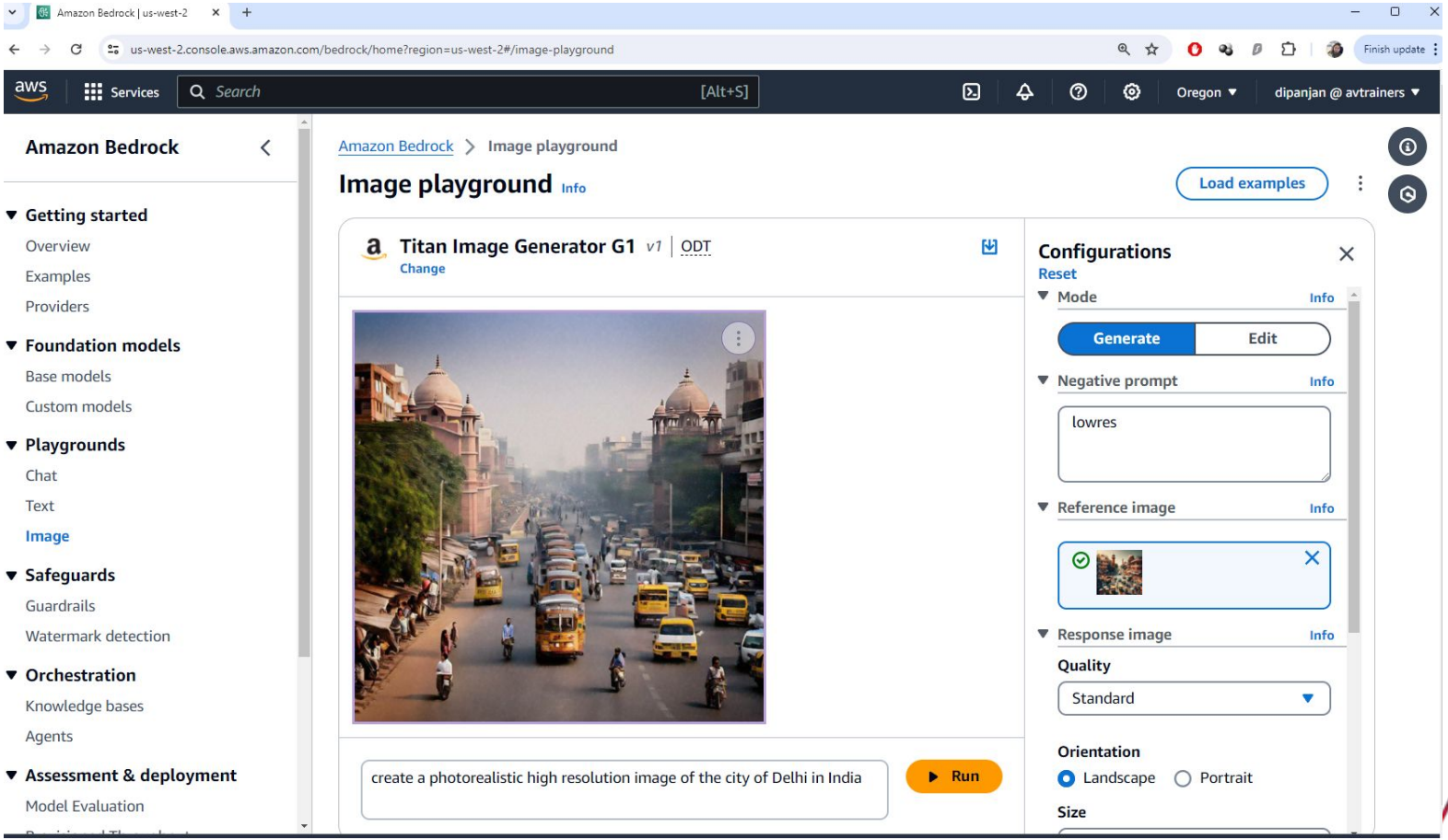
- Use Amazon Titan Embeddings or Cohere Embed to create vector of company data to enable semantic search
- The embedding can be stored in a database to use for quick and accurate searches

Gen AI on AWS

Base components and design patterns

- Models
- Prompts
- Memory
- Chains
- Tools
- Agents

Try Foundation Models in AWS Playground



Analytics Vidhya