



# Accelerate ML Innovation with SageMaker Jumpstart

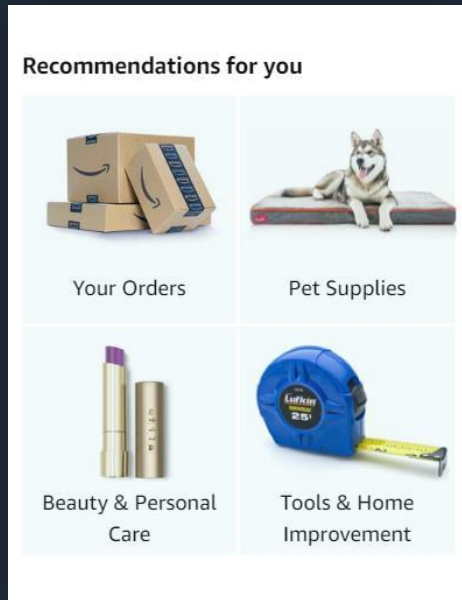


# Machine learning (ML) is at an inflection point

**Key drivers:** Compute capacity increase | Data growth | Model sophistication



# ML innovation is in the Amazon DNA



---

**4,000 products**  
**per minute** sold  
on Amazon.com



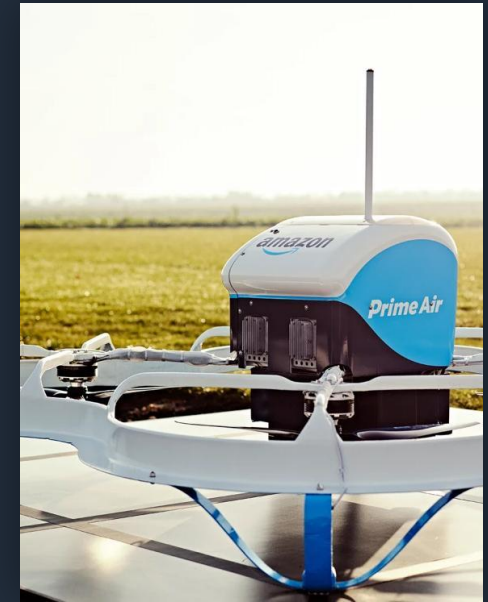
---

**1.6M packages**  
every day



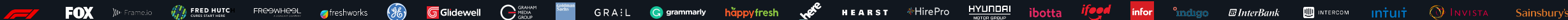
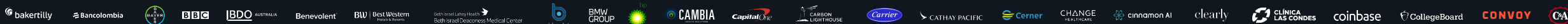
---

**Billions** of Alexa  
interactions each week

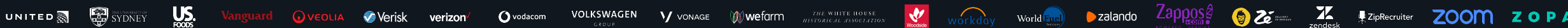


---

First Prime Air delivery  
on **December 7, 2016**



More than **100,000 customers** use AWS for ML





# Our focus has been on democratizing ML



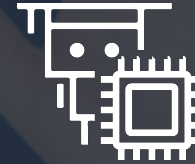
The broadest and  
deepest AI/ML  
capabilities



AWS DeepRacer  
League



Training and  
certification



AWS AI/ML  
Scholarship



Machine Learning  
University

# Amazon SageMaker JumpStart

---

ML hub with foundation models, built-in algorithms, and prebuilt ML solutions that you can deploy with just a few clicks



## Machine learning hub

Browse through 400+ built-in algorithms with pretrained models, pretrained foundation models, solutions, and example notebooks



## Pre-built training and inference scripts

Compatible with SageMaker and configurable with custom dataset



## UI as well as API-based

Use the user interface for single click model deployment or API for the Python SDK-based workflow



## Notebooks with examples

Jump into notebooks to use selected model with examples to guide you through the entire ML workflow



## Share and collaborate within your 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

# Build on top of existing foundation models using Amazon SageMaker JumpStart

stability.ai LightOn

AI21labs



co:here



Products / Machine Learning / Amazon SageMaker JumpStart

## Getting started with Amazon SageMaker JumpStart

Amazon SageMaker JumpStart is a machine learning (ML) hub that can help you accelerate your ML journey. Explore how you can get started with built-in algorithms with pretrained models from model hubs, pretrained foundation models, and prebuilt solutions to solve common use cases. To get started, see documentation or example notebooks that you can quickly execute.

Reset Filters

Q foundation models

Product Type

Sort By

Popularity

Text Tasks

- ☐ End-to-end Solution
- ☐ Text Classification
- ☐ Text Embedding
- ☐ Text Generation
- ☐ Text Summarization
- ☐ Named Entity Recognition
- ☐ Question Answering
- ☐ Zero-Shot Classification

Vision Tasks

- ☐ End-to-end Solution
- ☒ Image Classification
- ☐ Image Embedding
- ☐ Instance Segmentation

FOUNDATION MODEL PREVIEW

Text Generation

Proprietary Models

Various Providers

Models from AI21 Labs, Cohere, and LightOn in preview. Sign-up for preview with JumpStart in us-east-1 or eu-west-1 SageMaker Console.

Deploy Only

FOUNDATION MODEL FEATURED

Text to Image



Stable Diffusion 2


Stability AI

Model ID: model-txt2img-stabilityai-stable-diffusion-v2. This is a text-to-image model from Stability AI and downloaded from HuggingFace. It takes a textual description as

Fine-tunable

FOUNDATION MODEL FEATURED

Text Generation



AlexaTM (20b)


Pytorch

Model ID: pytorch-textgeneration1-alexa20b. AlexaTM 20B is a multitask, multilingual, large-scale sequence-to-sequence (seq2seq) model, trained on a mixture of Common Crawl

Deploy Only

FOUNDATION MODEL FEATURED

Text Generation



Bloom 1b7

Huggingface

Model ID: huggingface-textgeneration-bloom-1b7. This is a Text Generation model built upon a Transformer model from Hugging Face. It takes a text string as input and predicts next words in the sequence. This model has BigScience Responsible AI License v1.0. Please read the [terms] (https://huggingface.co/spaces/b

Deploy Only



# Why use foundation models on SageMaker JumpStart

1

Choose foundation models offered by model providers

AI21labs

Lightn  
We bring Light to AI

stability.ai

co:here



alexa

2

Try out model and/or deploy



Try out models via  
AWS Console



Deploy the model for  
inference using SageMaker  
hosting options includes  
single node

3

Fine tune model and  
automate ML workflow



Only selected models  
can be fine-tuned



Automate ML  
workflow

**Data stays in your account** including  
model, instances,  
logs, model inputs,  
model outputs

**Fully integrated**  
with Amazon  
SageMaker  
features

# SageMaker JumpStart models and features

## 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



### 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

Light<sup>star</sup>

### Models

Lyra-Fr  
10B

### Tasks

Text Generation  
  
Keyword  
extraction  
  
Information  
extraction  
  
Question  
answering  
  
Summarization  
  
Sentiment  
analysis  
  
Classification

AI21labs

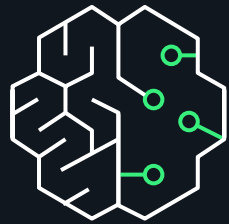
### Models

Jurassic-1  
Grande 17B

### Tasks

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

# Foundation models: how it works



## Amazon SageMaker JumpStart

Access and try out public and proprietary foundation models and easily customize and integrate them into your generative AI applications



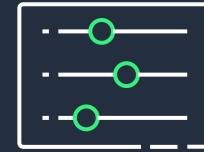
### Browse

Browse public and proprietary foundation models



### Experiment

Experiment with foundation models before choosing a model for deployment



### Customize

Easily customize selected foundation model with your own dataset without training from scratch



### Deploy

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



# Try-out experience

cohere **Cohere Generate Model - Medium**  
By Cohere [🔗](#)

Try a product demo of the capabilities of this model from Cohere. Do not upload any confidential or sensitive information. Use of this feature is for demonstration purposes only. This demo may not accurately represent the actual response times of the product.

**Prompt**

Context:  
The United Nations is an intergovernmental organization founded in 1945 with the mission of maintaining international peace and security, promoting human rights, and fostering social and economic development. It is composed of 193 member states and has its headquarters in New York City.

Question:  
What is the mission of the United Nations?

Answer:

**Generate text**

**Output**

The mission of the United Nations is to maintain international peace and security, promote human rights, and foster social and economic development.

**General Info**

Temperature: 0.9

Number of tokens: 100

Top k: 0

Top p: 0.7

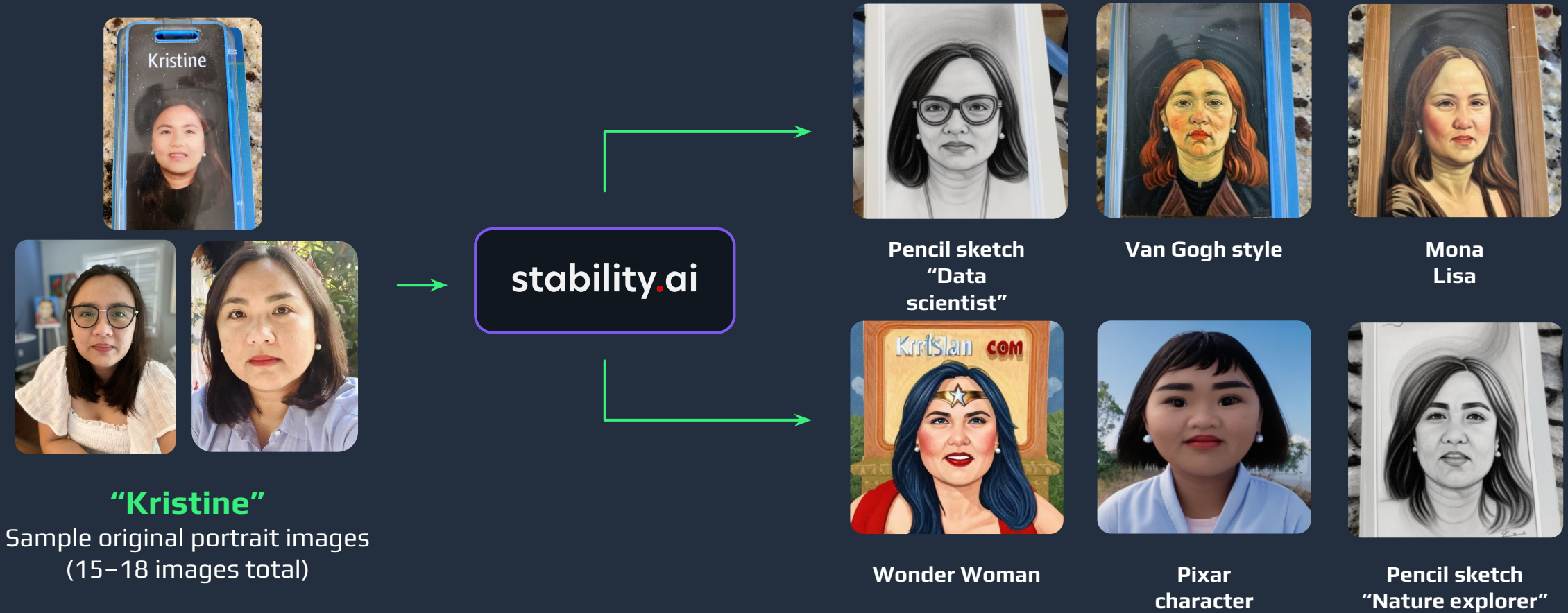
Presence Penalty: 0

Frequency Penalty: 0

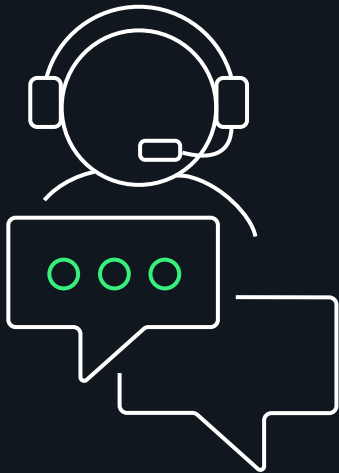
[Copy output](#)

- Try out the models and model prompts without running code or incurring costs
- Available for proprietary models in Top 10 in HELM benchmarks and public models for comparison purposes
- This is a shared environment in a SageMaker escrow account

# Demo 1: image generation using Stable Diffusion fine-tuning



# Demo 2: text generation using Co:here Medium



## Input

Customer call transcriptions

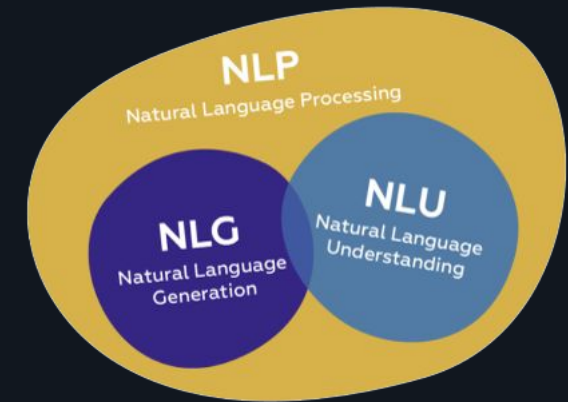


co:here



Amazon  
SageMaker

Foundation model  
in **SageMaker**  
**JumpStart**



## Output

Text summarization  
Abstractive question answering  
Sentiment analysis  
Key phrase extraction



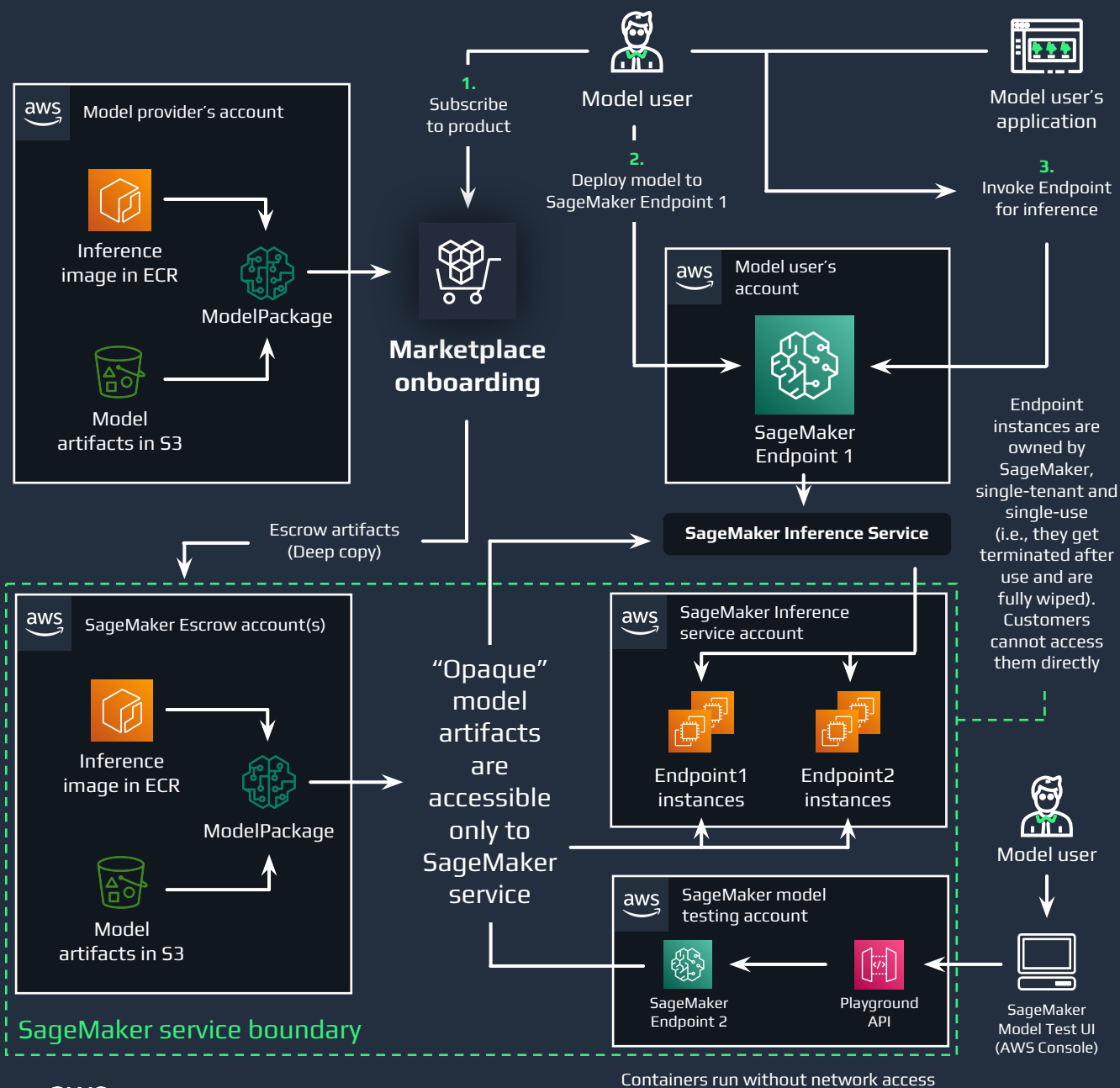
# Choosing the right instance for hosting

Size of model (# of parameters)	Large 3B–10B	Mega 11B–20B	Massive 100B+*
Task Type	Image generation Simple text classification (Short form)	Natural language understanding (NLU)	Natural language generation (NLG) (long form)
Minimum instance required	p3.2xlarge g5.2xlarge	p3.8xlarge g5.12xlarge	p4de.24xlarge p4d.24xlarge
Pricing	\$4/hr \$2/hr	\$15/hr \$9/hr	\$47/hr \$38/hr

Scale vertically (larger instances) to improve latency

Scale horizontally (more instances) to support higher traffic

\*P4d instances will have limited availability, escalate to S-Team for support

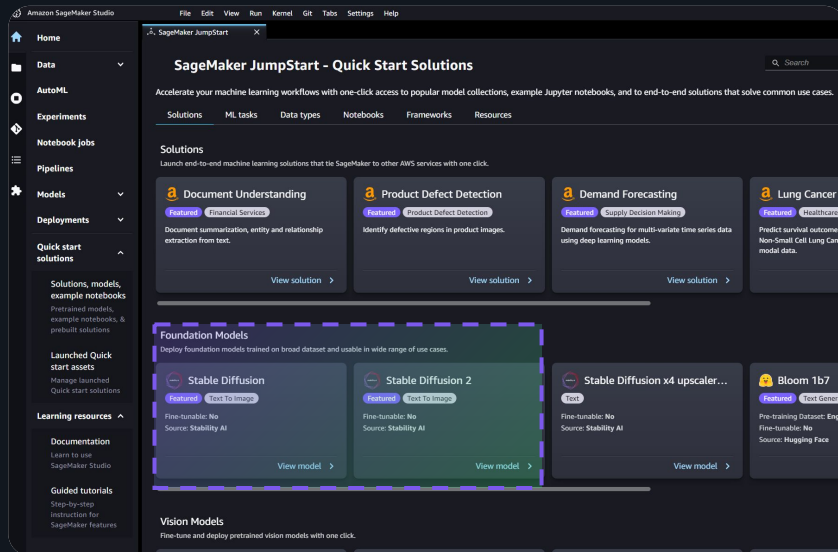


# SageMaker JumpStart protects your data and model provider IP

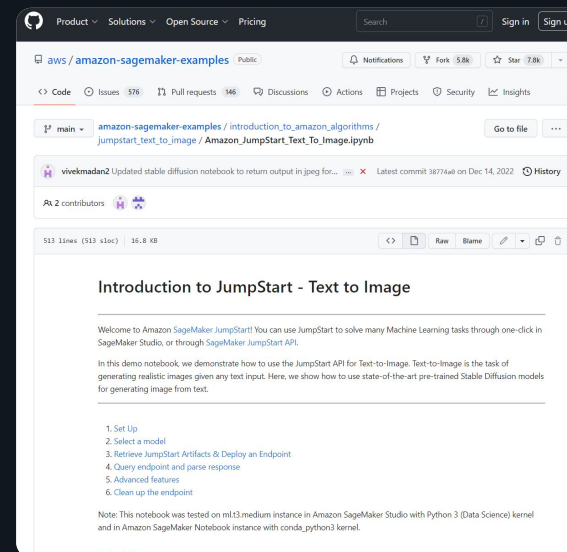
- Proprietary model package and endpoint is hosted in SageMaker owned escrow account
- Containers have no outbound network access; user data and model provider IP is protected the same time
- No data is used to update/train the base model that JumpStart provides to customers

# 3 ways to use foundation models with SageMaker JumpStart

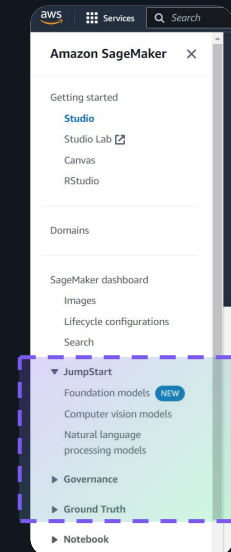
## SageMaker Studio One-click deploy



## SageMaker Notebooks







## AWS console Preview





# Built-in algorithms and pretrained models

400+ ALGORITHMS AND PRE-TRAINED, STATE-OF-THE-ART, PUBLICLY AVAILABLE MODELS FROM PYTORCH HUB, TENSORFLOW HUB, AND HUGGING FACE HUB THAT ARE SECURELY STORED IN AN AWS-OWNED ENVIRONMENT

	Tasks		Algorithms/models
 <b>Tabular</b>	Classification, regression, time-series		LightGBM, CatBoost, AutoGluon, TabTransformer, XGBoost, DeepAR
 <b>Vision</b>	Image classification Image embedding	Object detection Semantic segmentation	ResNet, Inception, MobileNet, SSD, Faster RCNN, YOLO, and more
 <b>Text</b>	Sentence classification Text classification Question answering	Summarization Text generation, translation, Named-entity recognition	AlexaTM, Bloom, Stable Diffusion 2.0, BERT, RoBERTa, DistilBERT, Distillbart xsum, GPT2, ELECTRA, and more
 <b>Audio</b>	Audio embedding		TRILL, TRILLsson, TRILL-Distilled, FRILL

# Solutions with SageMaker JumpStart



## Extract & analyze data from documents

Document understanding  
Handwriting recognition  
Intelligently fill in missing form data  
Privacy-based NLP



## Forecasting & optimization

Deep demand forecasting  
Price optimization  
Purchase modelling  
Filling missing value  
Lung cancer survival prediction



## Classification

Detect malicious users and transactions  
Fraud detection in financial transactions  
Financial payment classification  
Privacy sentiment classification



## Credit risk prediction

Corporate credit rating prediction  
Graph-based credit scoring  
Explain credit decisions



## Predictive maintenance

Detecting potential equipment failure for manufacturing  
Predictive maintenance for vehicle fleets



## Computer vision

Product defect detection  
Bird species object detection



## Autonomous driving

Visual perception with active learning  
Reinforcement learning for battlesnake AI



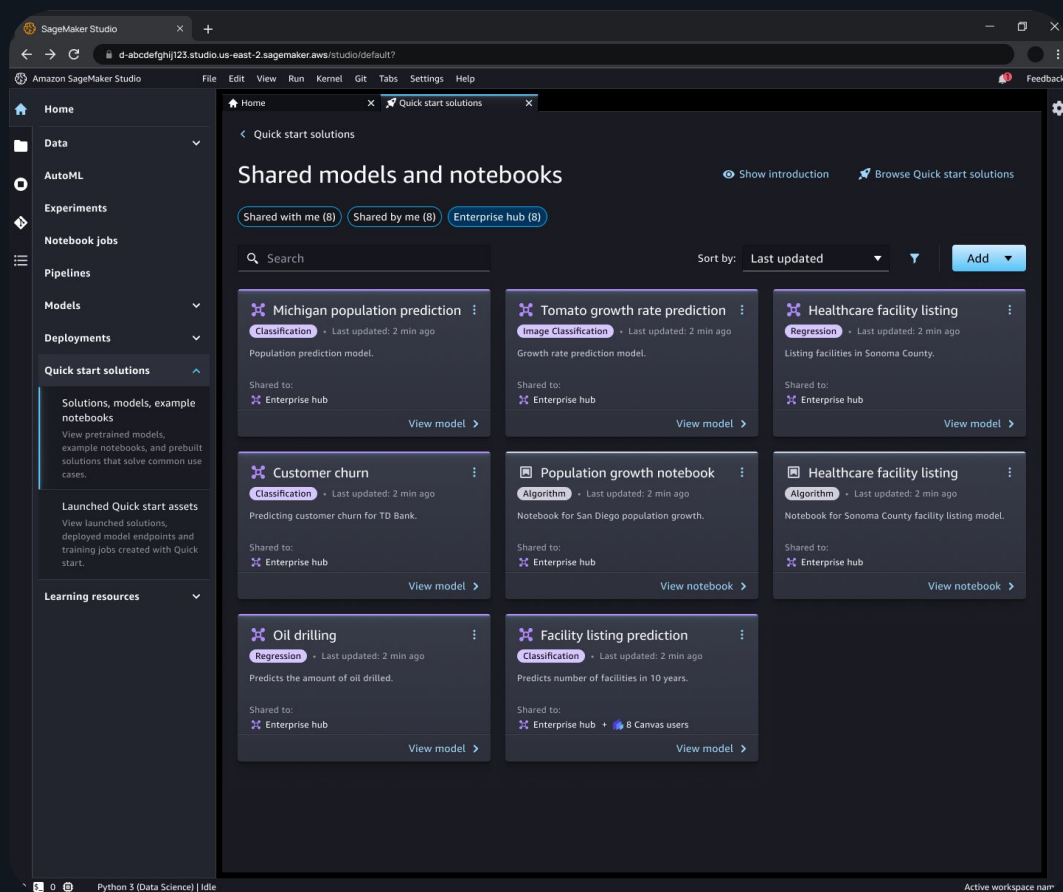
## Personalized recommendations

Entity resolution for smart advertising  
Purchase modeling



## Churn prediction

Churn prediction  
Churn prediction for mobile phone customers



# ML artifact sharing in SageMaker JumpStart

Enables data scientists to share ML artifacts securely within the enterprise and reuse alongside SageMaker built-in content

- Share with other users in your organization
- Discover shared contents easily and start fine-tuning with your own data
- Monitor and control what contents are shared

# 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

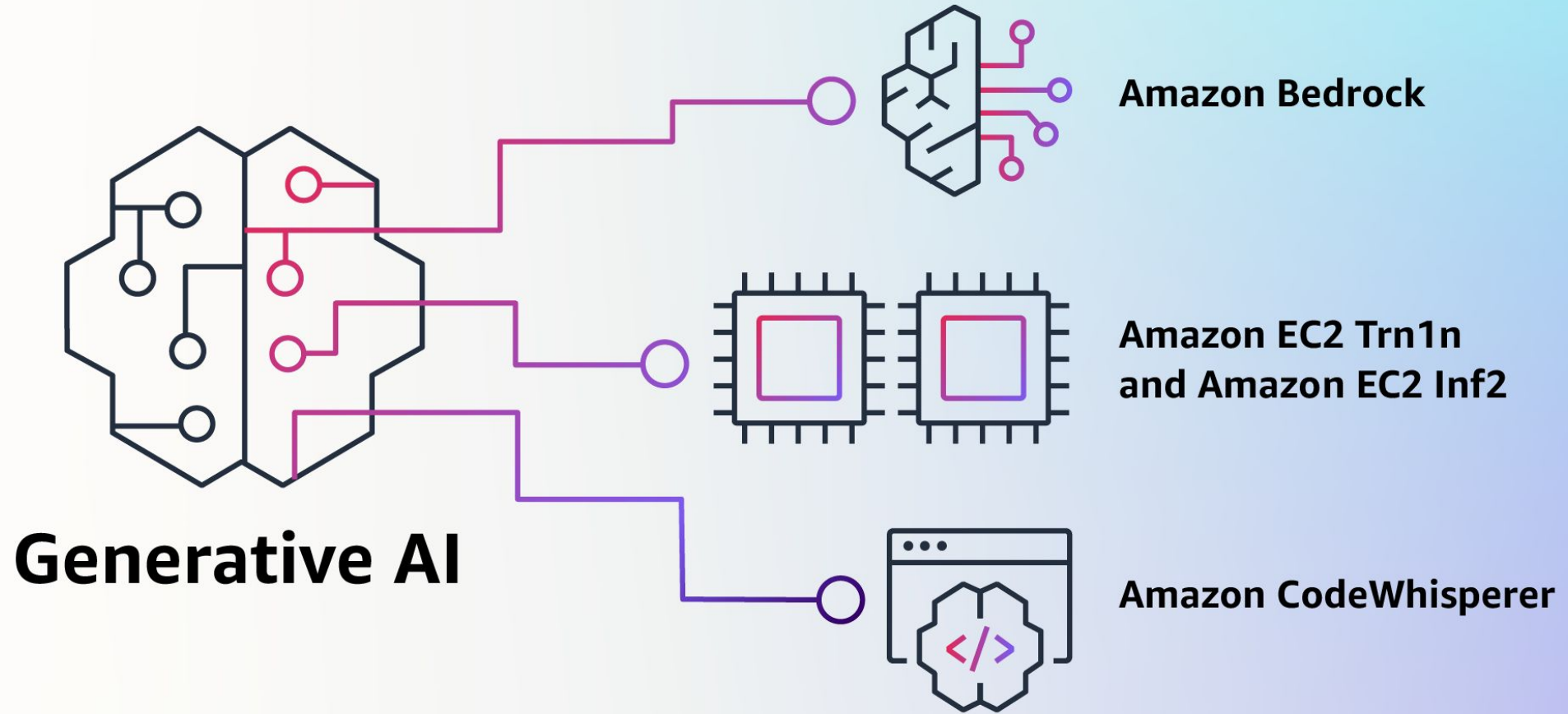
Instantly try models in playground; deploy with SageMaker using managed inference scripts



## Data security

Host models on customer-dedicated endpoints inside your VPC

# Building with generative AI on AWS



# Why AWS for generative AI?



Flexibility



Secure  
customization



The most  
cost-effective  
infrastructure



The easiest way  
to build with FMs



Generative AI-  
powered solutions



# Getting started with SageMaker JumpStart



**Visit PDP and explore**



[JumpStart Product Detail Page](#)



**Engage AWS AI/ML Specialist**

Engage your account team  
or AI/ML Specialists

**1**

JumpStart  
Hands-on  
Workshop

**2**

Design  
Partner/POC  
in a Box

**3**

ML  
Solutions  
Lab



# Thank you!

Kristine Pearce  
Principal BDM  
pearck@amazon.com

Meena Thandavarayan  
Sr. AI/ML SA  
thandavm@amazon.com