Generative AI on AWS

How to Leverage AWS ML Services to Drive Value for your Customers

Francisco Amaya

Data Partner SA Lead, EMEA

Innovation can **transform industries**

The tipping point for **Generative AI**

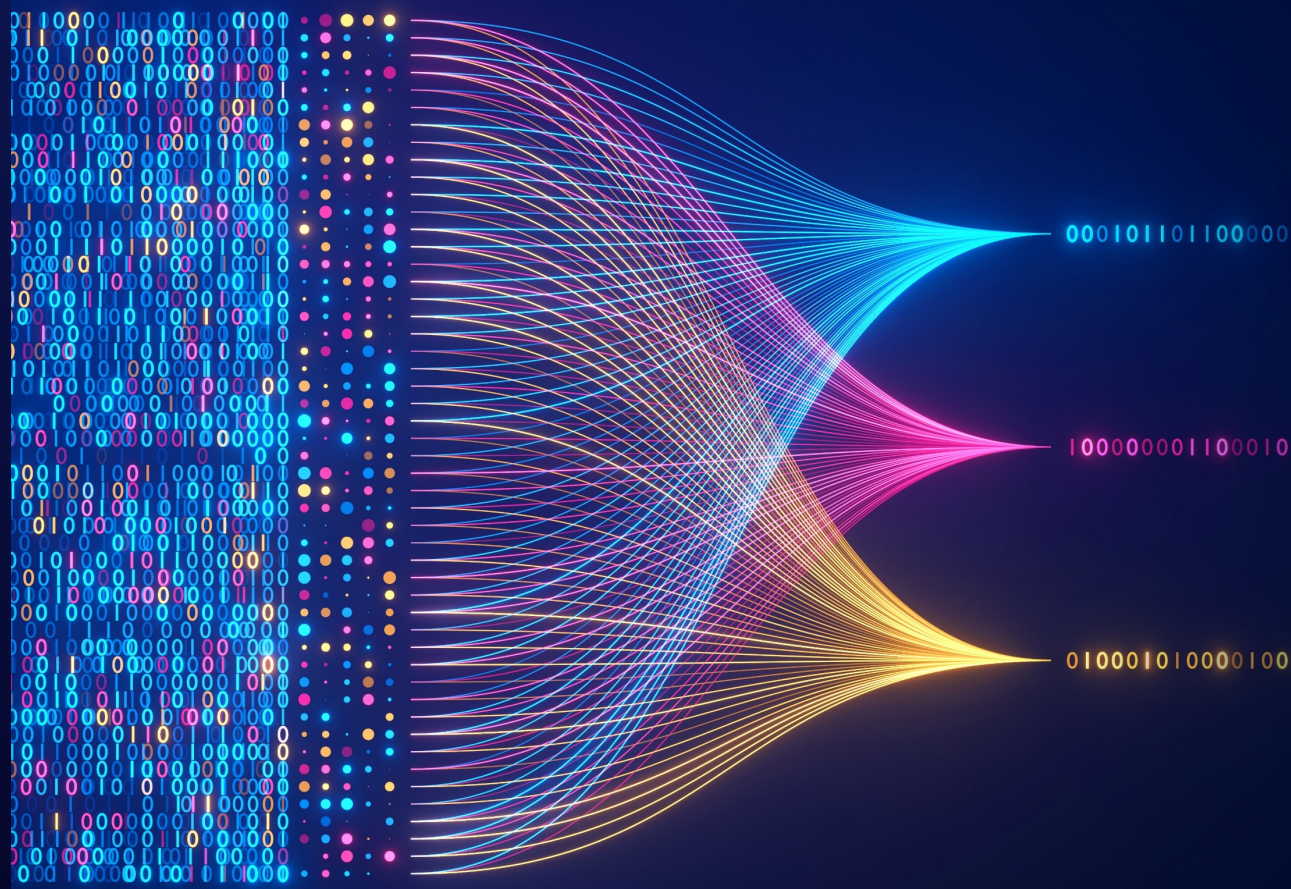
Generative AI is powered by foundation models

Pretrained on vast amounts of unstructured data

Contain large number of parameters that make them capable of learning complex concepts

Can be applied in a wide range of contexts

Customize FMs using your data for domain specific tasks



AI

MACHINE LEARNING

DEEP LEARNING

FOUNDATION MODELS

SIMPLE

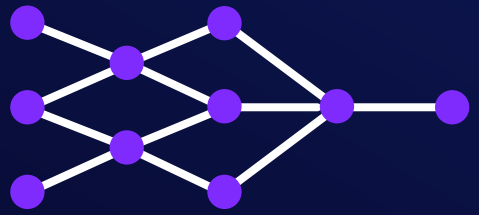
INPUTS

SIMPLE

OUTPUTS

COMPLEX

INPUTS

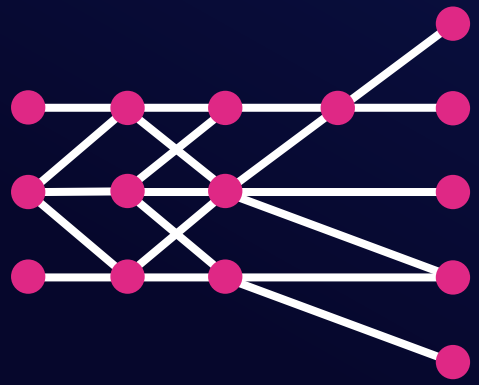


SIMPLE

OUTPUTS

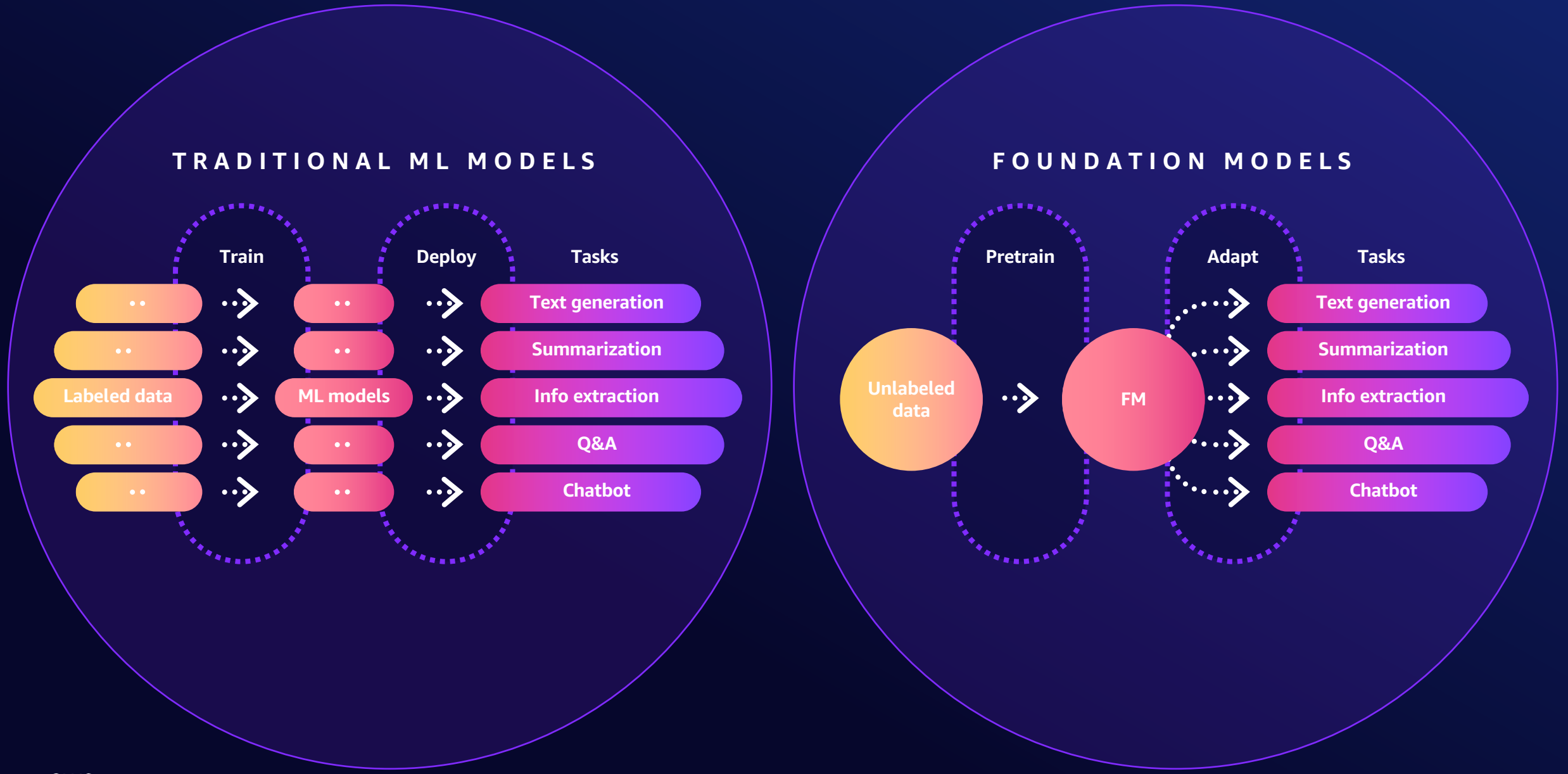
COMPLEX

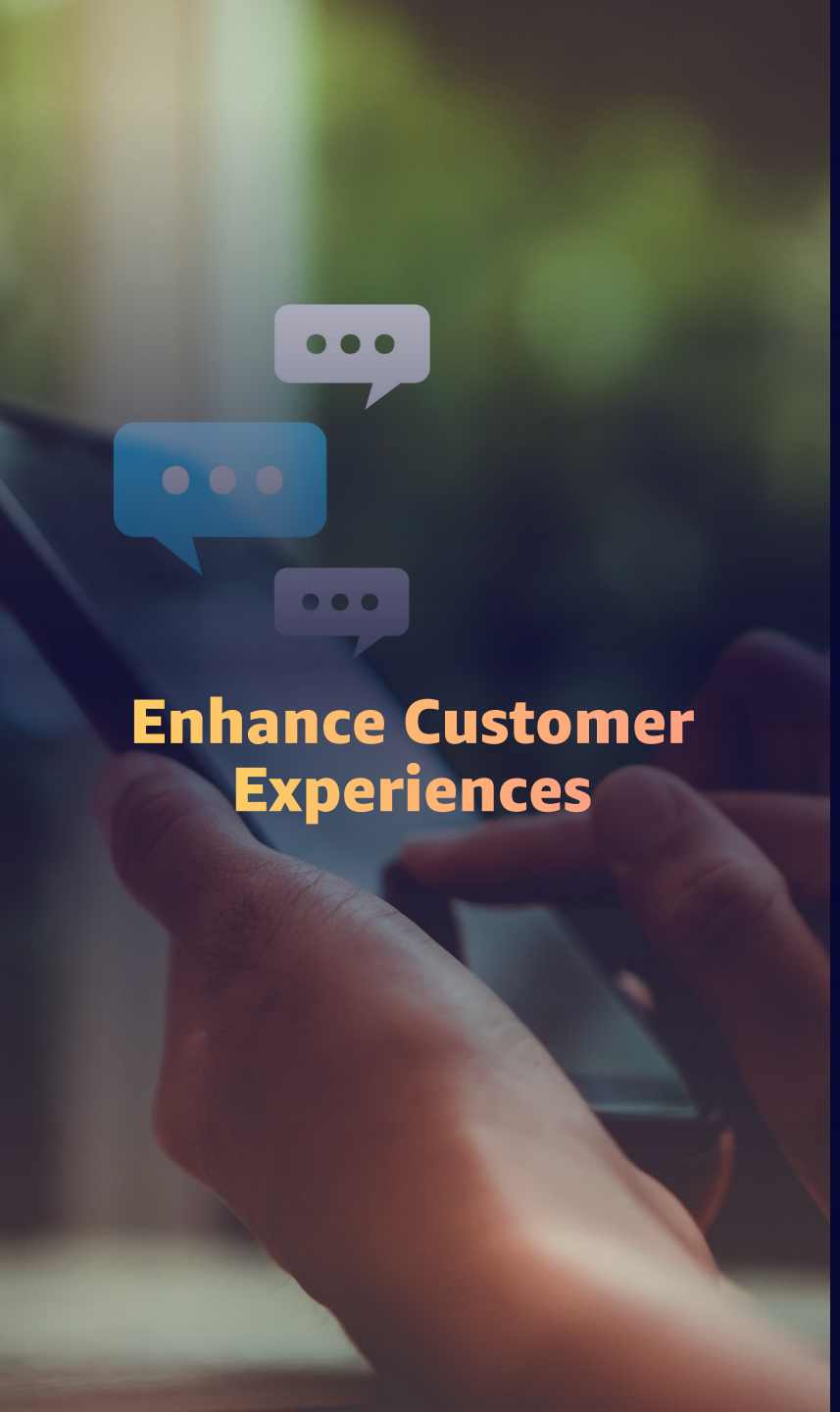
INPUTS



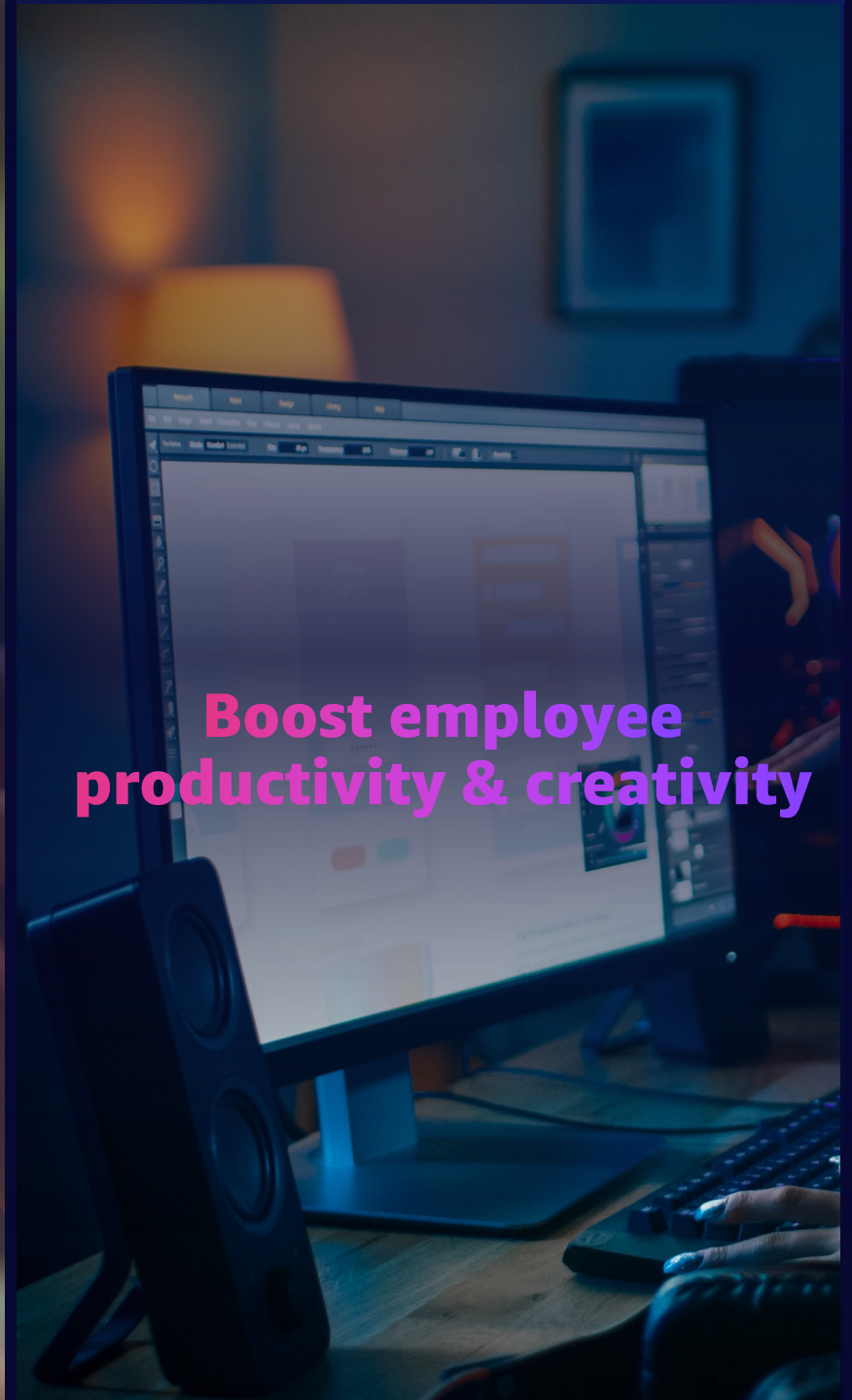
COMPLEX

OUTPUTS



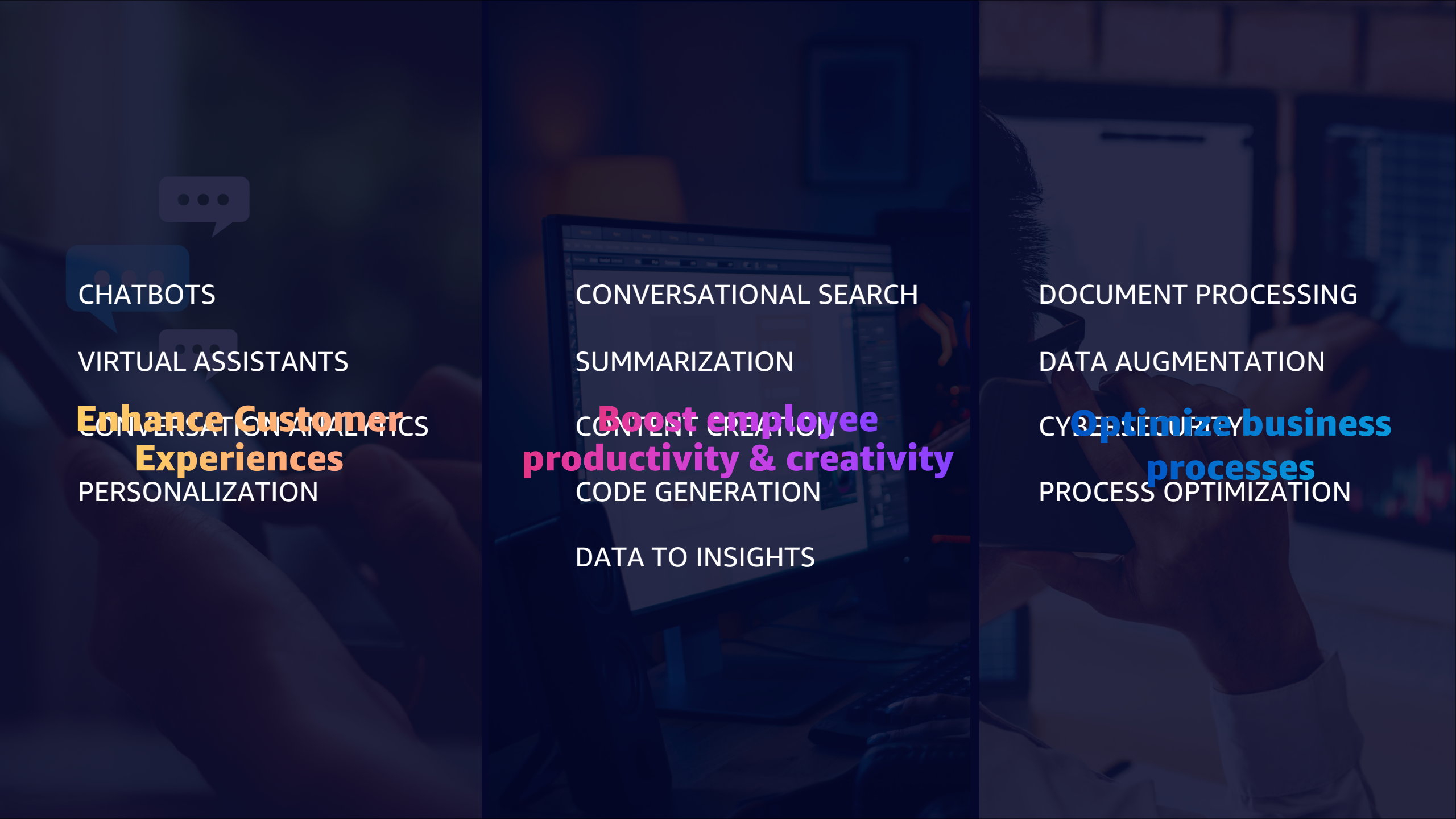


Enhance Customer Experiences



Boost employee productivity & creativity

Optimize business processes


CHATBOTS
VIRTUAL ASSISTANTS
CONVERSATION ANALYTICS
PERSONALIZATION
Enhance Customer Experiences
CONVERSATIONAL SEARCH
SUMMARIZATION
CONTENT CREATION
CODE GENERATION
DATA TO INSIGHTS
Boost employee productivity & creativity
DOCUMENT PROCESSING
DATA AUGMENTATION
CYBERSECURITY
PROCESS OPTIMIZATION
Optimize business processes

ENHANCE CUSTOMER EXPERIENCE

USE CASE EXAMPLES

Chatbots & Virtual Assistants

Automate responses for customer service queries through generative AI-powered chatbots, voice bots, and virtual assistants.

Conversation analytics

Provide insights from recorded conversations, forms, or surveys to better understand customer needs, identify call drivers, and detect emerging trends.

Agent assist

Enhance agent performance to improve first contact resolution, augment tasks such as knowledge search, call summarization, and problem-solving.

Personalization

Deliver better personalized experiences and increase customer engagement with individually curated offerings and communications.

BOOSTING YOUR WORKFORCE'S PRODUCTIVITY & CREATIVITY

USE CASE EXAMPLES

Content creation

Empower employees to create content faster and smarter across departments—from marketing to sales to engineering

Report generation

Unlock powerful insights from your data and enable every employee to make faster and better, data-driven decisions

Search, summarization and analysis

Summarize data from various sources into actionable insights and perform comparative analysis on your enterprise content

Code generation

Accelerate application development with automatic code recommendations based on the code and comments in your IDE

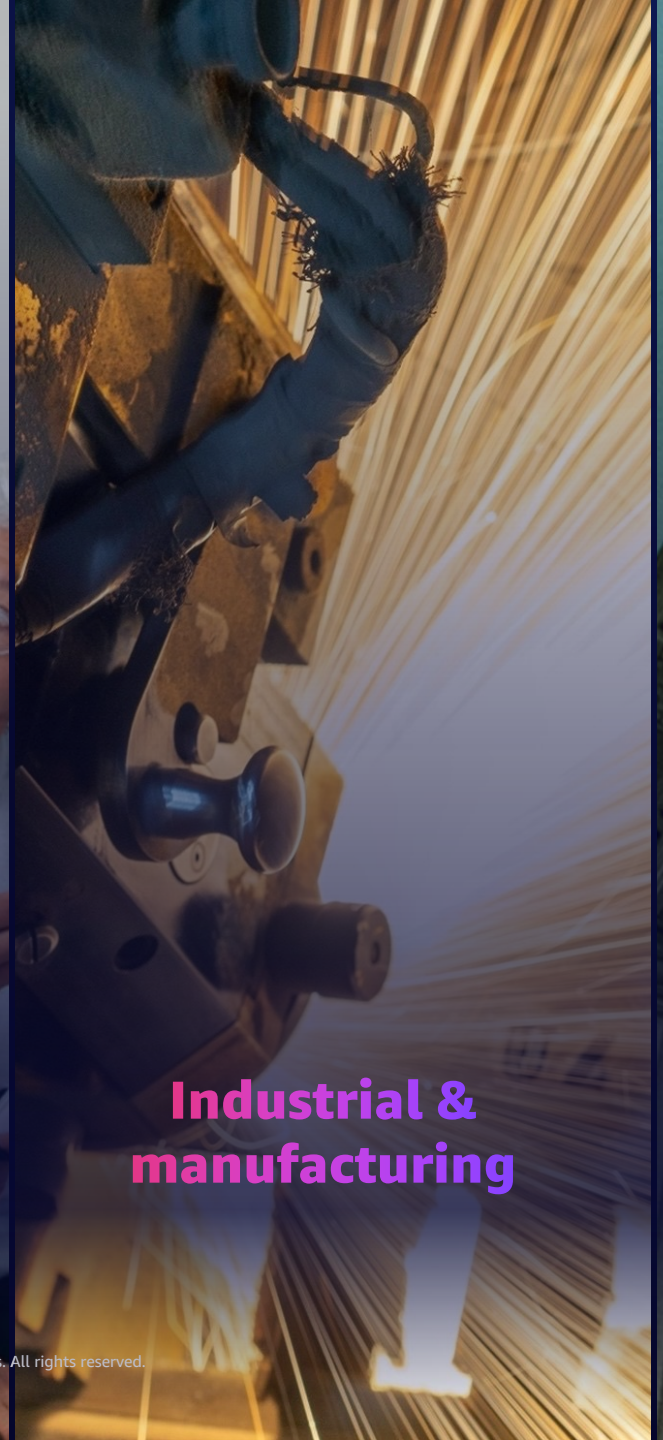
IMPROVE BUSINESS PROCESSES

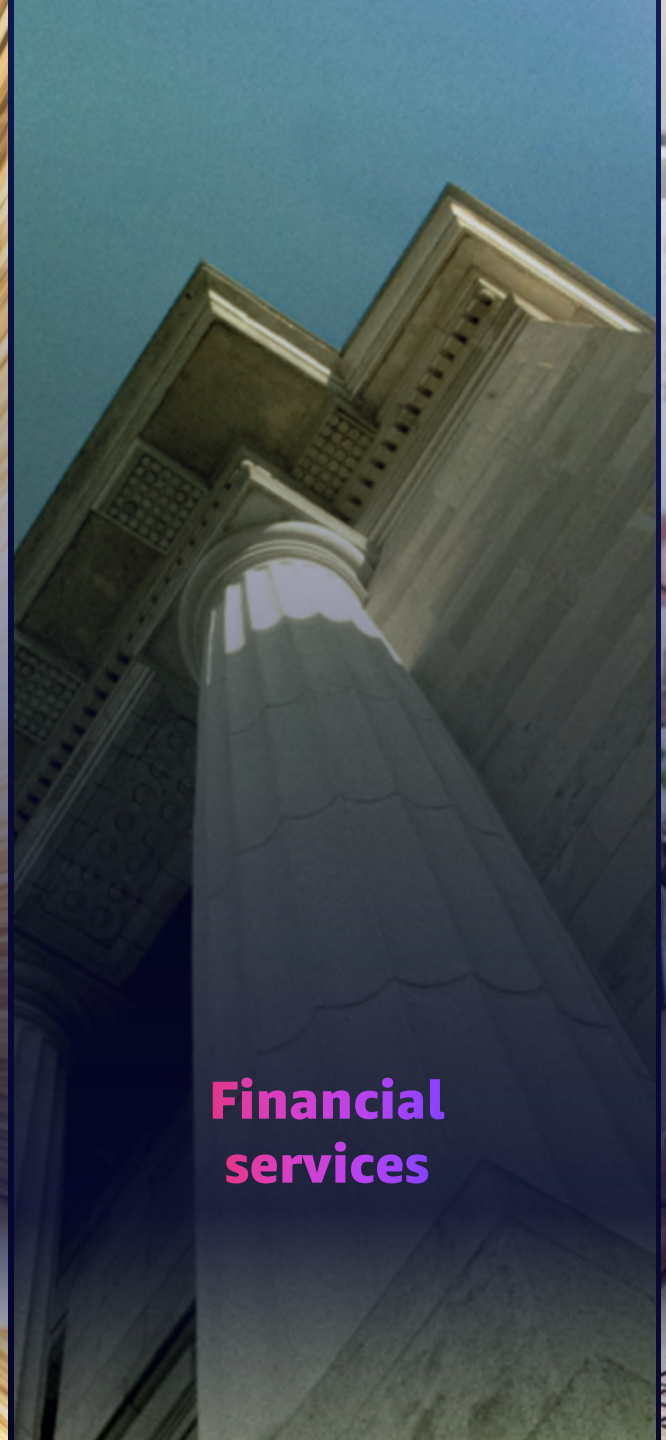
Document processing

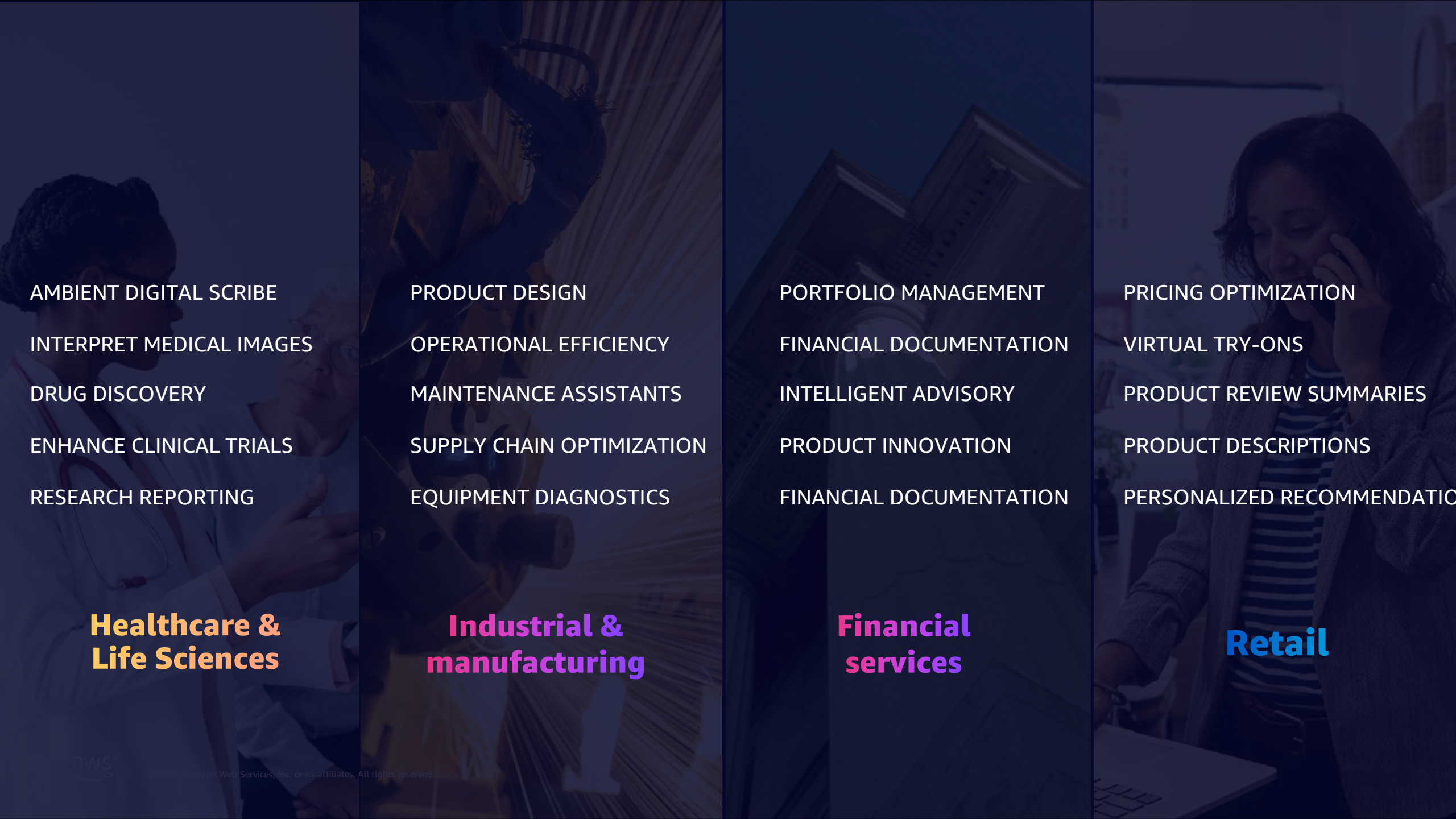
Improve document processing by automatically extracting and summarizing data and insights from documents.

Data augmentation

Generate synthetic data to train ML models, when the original dataset is small, imbalanced or sensitive.







AMBIENT DIGITAL SCRIBE

INTERPRET MEDICAL IMAGES

DRUG DISCOVERY

ENHANCE CLINICAL TRIALS

RESEARCH REPORTING

Healthcare & Life Sciences

PRODUCT DESIGN

OPERATIONAL EFFICIENCY

MAINTENANCE ASSISTANTS

SUPPLY CHAIN OPTIMIZATION

EQUIPMENT DIAGNOSTICS

Industrial & manufacturing

PORTFOLIO MANAGEMENT

FINANCIAL DOCUMENTATION

INTELLIGENT ADVISORY

PRODUCT INNOVATION

FINANCIAL DOCUMENTATION

Financial services

PRICING OPTIMIZATION

VIRTUAL TRY-ONS

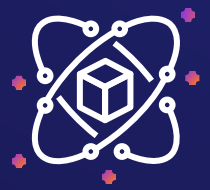
PRODUCT REVIEW SUMMARIES

PRODUCT DESCRIPTIONS

PERSONALIZED RECOMMENDATIO

Retail

Everything you need to accelerate your generative AI journey



Easiest way to build

with leading foundation models

Differentiate with your data

in a secure and private environment

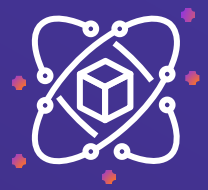
Increase productivity

with generative AI applications and services

Most performant, low cost

infrastructure to scale generative AI

Everything you need to accelerate your generative AI journey



Easiest way to build

with leading foundation models

Differentiate with your data

in a secure and private environment

Increase productivity

with generative AI applications and services

Most performant, low cost

infrastructure to scale generative AI

NOW GENERALLY AVAILABLE

Amazon **Bedrock**

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

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

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

Privately customize FMs using your organization's data

Amazon Bedrock

Choice of foundation models

AI21labs

JURASSIC-2

Multilingual LLMs for text generation in Spanish, French, German, Portuguese, Italian, and Dutch

ANTHROPIC

CLAUDE 2

LLM for conversations, question answering, and workflow automation based on research into training honest and responsible AI systems

co:here

COMMAND

Text generation model for business applications like summarization, copywriting, dialog, extraction, and question answering

Meta AI

LLAMA 2

Pre-trained and fine-tuned LLMs for natural language tasks like question answering and reading comprehension

stability.ai

SDXL 1.0

Generation of unique, realistic, high-quality images, art, logos, and designs

amazon

AMAZON TITAN

Text summarization, generation, classification, open-ended Q&A, information extraction, embeddings and search

aws

Build with publicly available foundation models

AVAILABLE ON SAGEMAKER JUMPSTART

AI21labs

Models

Jurassic-2 Ultra, Mid

Contextual answers

Summarize

Paraphrase

Grammatical error correction

Tasks

Text generation

Long-form generation

Summarization

Paraphrasing

Chat

Information extraction

Meta AI

Models

Llama 2 7B, 13B, 70B

Tasks

Question answering

Chat

Summarization

Paraphrasing

Sentiment analysis

Text generation

cohere

Models

Cohere

Command XL

Tasks

Text generation

Information extraction

Question answering

Summarization

Hugging Face

Models

Falcon-7B, 40B

Open LlaMA

RedPajama

MPT-7B

BloomZ 176B

Flan T-5 models (8 variants)

DistilGPT2

GPT NeoXT

Bloom models (3 variants)

Tasks

Machine translation

Question answering

Summarization

stability.ai

Models

Stable Diffusion XL 1.0

2.1 base

Upscaling

Inpainting

Tasks

Generate photo-realistic images from text input

Improve quality of generated images

Features

Fine-tuning on Stable Diffusion 2.1 base model

LightOn

Models

Lyra-Fr

10B, Mini

Tasks

Text generation

Keyword extraction

Information extraction

Question answering

Summarization

Sentiment analysis

Classification

databricks

Models

Dolly

Tasks

Question answering

Chat

Summarization

Paraphrasing

Sentiment analysis

Text generation

alexa

Models

AlexaTM 20B

Tasks

Machine translation

Question answering

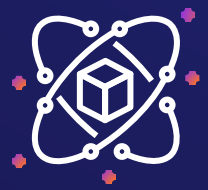
Summarization

Annotation

Data generation

aws

Everything you need to accelerate your generative AI journey



Easiest way to build

with leading foundation models

Differentiate with your data

in a secure and private environment

Increase productivity

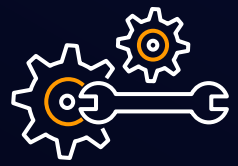
with generative AI applications and services

Most performant, low cost

infrastructure to scale generative AI

Your data is **your differentiator**

Privately customize foundation models using your organization's data



Fine-tune

PURPOSE — Maximizing accuracy for specific tasks

DATA NEED — Small number of labeled examples

Keeping your data private and secure

None of the customer's data is used to train the underlying model

All data is encrypted at rest and PrivateLink support allows access to Bedrock APIs via customer's VPC endpoints

Customized foundation models and the customer-specific data that trains them remain private

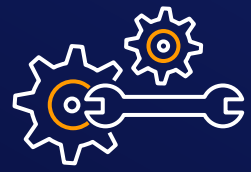
Build a data strategy to fuel your generative AI applications

Comprehensive

Comprehensive set of services for storing and querying structured unstructured and vector data

Integrated

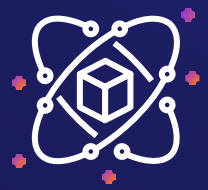
Choices for integrating data including zero-ETL so you can easily connect to all your data



Governed

End-to-end data governance capabilities

Everything you need to accelerate your generative AI journey



Easiest way to build

with leading foundation models

Differentiate with your data

in a secure and private environment

Increase productivity

with generative AI applications and services

Most performant, low cost

infrastructure to scale generative AI

Generative BI capabilities in **Amazon QuickSight**

New FM-powered capabilities for business users to extract insights, collaborate and visualize data

PUBLIC PREVIEW

Easily author, fine-tune and add visuals to dashboards

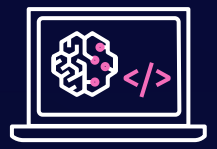
COMING SOON

Automatically generate data stories with natural language

GENERALLY AVAILABLE

Amazon **CodeWhisperer**

Build apps faster and more securely with an AI coding companion



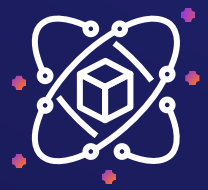
Generate code suggestions in real-time

Scan code for hard-to-find vulnerabilities

Flag code that resembles open-source training data or filter by default

FREE FOR INDIVIDUAL TIER

Everything you need to accelerate your generative AI journey



Easiest way to build

with leading foundation models

Differentiate with your data

in a secure and private environment

Increase productivity

with generative AI applications and services

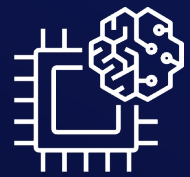
Most performant, low cost

infrastructure to scale generative AI

Deep investments
in global infrastructure

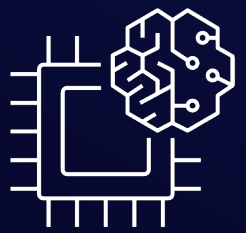
Broad choice of ML accelerators

High performance, low-cost ML infrastructure



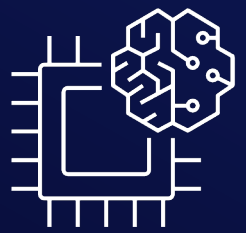
10+ years of silicon innovation

Purpose-built accelerators
for generative AI



AWS Trainium

Up to 50% savings on training costs
over comparable Amazon EC2 instances



AWS Inferentia2

Up to 40% better price performance
than comparable Amazon EC2 instances

Everything you need to accelerate **your generative AI journey**

Easiest and most secure way to build generative AI applications

Data as your differentiator and strategic asset for generative AI

Most performant, low cost infrastructure for generative AI

Generative AI applications to enhance productivity

Thank you!