aws

ML Fridays: Using AI to automate content moderation and compliance

Create safe online environments, protect your brand, and reduce moderation costs

Jonathan Hedley

Principal AI/ML
Solutions Architect
AWS

Adhikaar Marwaha

Backend Lead
Rooter Sports Technologies Ltd.

The User Generated Content (UGC) era

80%
of all web content is UGC

Gorilla Group Consultants, 2020

Moderation costs, risk and liability from harmful exposure on the rise

79%
of consumer decisions influenced by UGC

Stackla Customer Survey, 2019

26%
UGC platform industry growth ($10 billion by 2028)

Grand View Research, 2021

Volume and complexity outpacing human capabilities

Inclusion and safety through moderation is expected

50% people create content at least once, daily
Stackla Customer Survey, 2015

40% consumers disengage after one exposure to toxic UGC
BusinessWire, 2021

70% consumers agree that brands are responsible to moderate UGC
BusinessWire, 2021
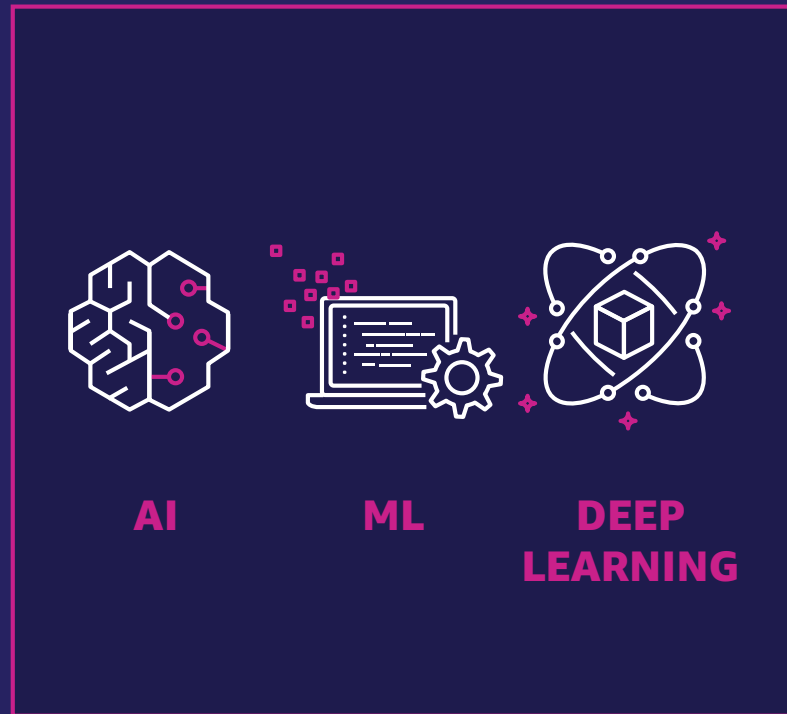
Customers are turning to AI for help

Moderate high volumes of UGC

Keep online communities
safe and inclusive

Better allocate human skill

Meet compliance and guidelines

Reduce moderation costs

AI   **ML**   **DEEP LEARNING**

Improve brand, information, and user safety

Scale coverage

Reduce human moderators' workload

Save time and costs

NEEDS   **BENEFITS**

aws

Content moderation with AWS

Automate **image, video, text, and audio** moderation workflows with fully automated AI services to moderate unwanted content and protect sensitive information.

Lower cost and speed to market with AI across industries

IMAGES

VIDEOS

TEXT

AUDIO

+

AI INSIGHTS AND CONNECTIONS

HUMAN REVIEWS

SOCIAL MEDIA

GAMING

E-COMMERCE

ADVERTISING

HIGHLY REGULATED INDUSTRIES

FINSERV

HEALTH

EDUCATION

aws

Multi modal moderation capabilities

IMAGE

Identify objects, text, and scenes

Automatically identify logos and celebrities

Detect and filter unwanted content or associations

VIDEO

Automatically search and index large volumes

Flag unwanted content by frame

Label content with hierarchical categories

TEXT

Automatically identify profanity

Detect and extract text within images and video

Automatically filter and redact personal information

AUDIO

Convert real-time or asynchronous speech to text

Filter profanity and other unwanted vocabulary

Tune models regardless of source and quality

SOCIAL MEDIA

CONTENT TYPE

Personal profiles

Images

Long and short videos

Live feeds Infographics

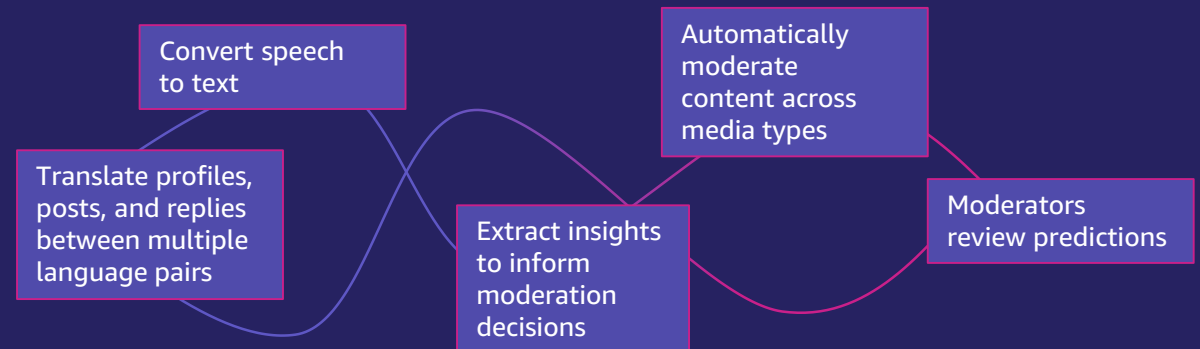Comments and reviews

Media captions and transcripts

WHY

Preventing exposure to harmful content and unwanted brand associations

Increase engagement

Reach responsible social media objectives

Protect network metrics -growth, session length, conversion

Convert speech to text

Automatically moderate content across media types

Translate profiles, posts, and replies between multiple language pairs

Extract insights to inform moderation decisions

Moderators review predictions

GAMING

CONTENT TYPE

User comments

In-game voice and text interactions

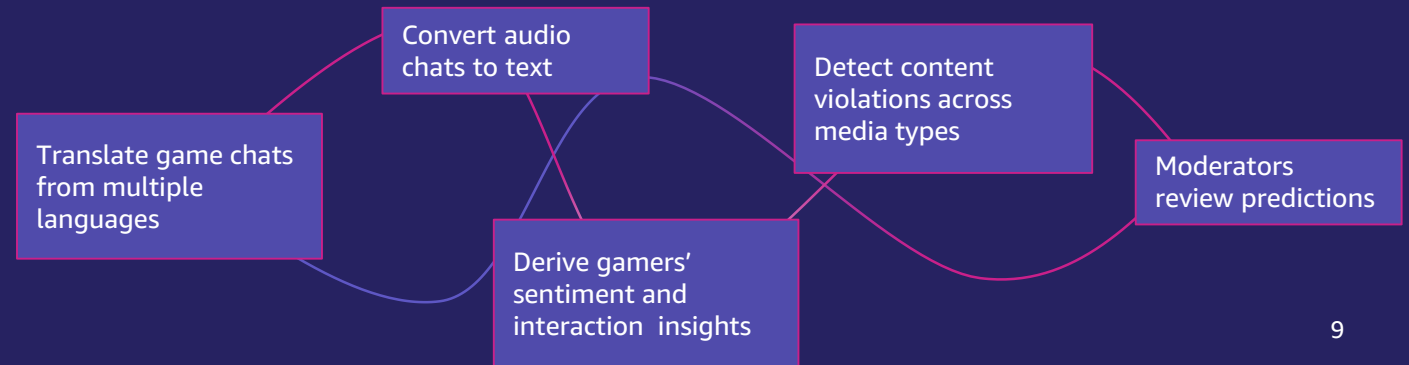Shared images and videos in gaming lounges and forums

Gamers' live streams and recordings

WHY

Increase members' inclusiveness and engagement while protecting brands, users, and sensitive information

Prevent exposure to harmful content (hate, profanity, bullying)

Protect key gaming metrics (positive play guidelines, K-Factor, daily active users, revenue per user)

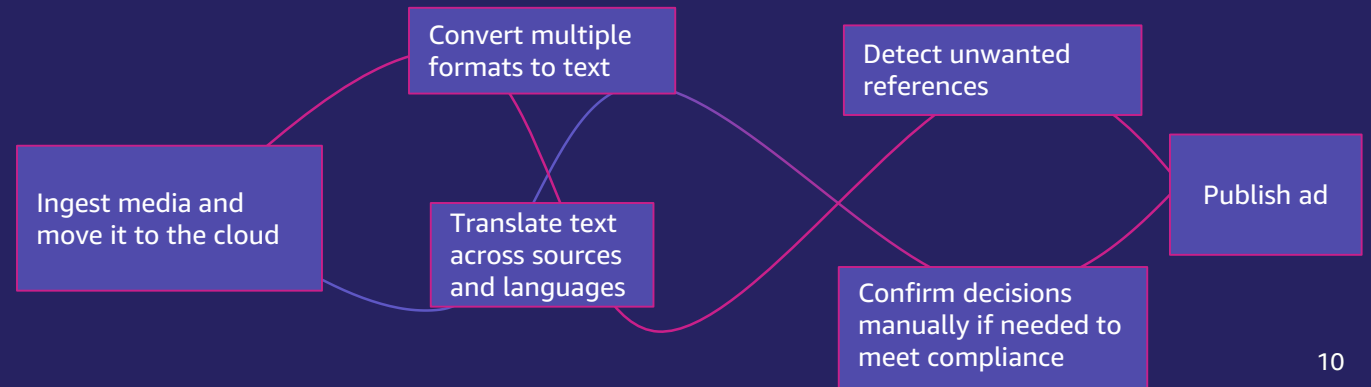Convert audio chats to text

Detect content violations across media types

Translate game chats from multiple languages

Moderators review predictions

Derive gamers' sentiment and interaction insights

CONTENT TYPE

Image	Video	Low and high quality media
Text	Speech	

WHY

Brand safety against unwanted associations

Minimize brand risk

Meet internal and external standards, practices, and regulation

Achieve brand elevation and likeability objectives that lead to revenue growth

Ingest media and move it to the cloud

Convert multiple formats to text

Translate text across sources and languages

Detect unwanted references

Confirm decisions manually if needed to meet compliance

Publish ad

10

HIGHLY REGULATED INDUSTRIES

FINSERV

HEALTH

EDUCATION

PROTECTION OF SENSITIVE INFORMATION

FINSERV	HEALTH	EDUCATION
Financial records	Medical records	Media contributions
Applications	Transcriptions	Chats
Tax documents	Personal Health Information (PHI)	Assignments
Contact center transcripts	Research information	Class transcriptions

WHY

International, federal, and state compliance

FINSERV	HEALTH	EDUCATION
RFPA, GLBA, FCRA, GDPR	HIPAA	FERPA, COPPA, CIPA
Reduce workforce exposure	Secure flow of health information	Foster and increase inclusiveness
	Preserve patients and subjects rights	Eliminate toxic environments

Why AWS AI/ML for content moderation?

CONTENT MODERATION

AWS offers more computer vision solutions and services, including content moderation, than any other cloud provider
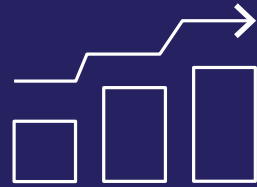
ACCELERATE YOUR ADOPTION OF ML

Self-managed, customizable services, and solutions available to deliver value from day one

BROADEST AND DEEPEST SET OF AI AND ML SERVICES

250 new features and AWS services launched this last year for everyone from ML scientists to application developers

BUILT ON THE MOST COMPREHENSIVE CLOUD PLATFORM

The strongest set of compute, storage, security, database, and analytics capabilities to build upon

Core AWS AI services for content moderation

Amazon Rekognition

Image and video analysis

Amazon Transcribe

Speech to text

Amazon Comprehend

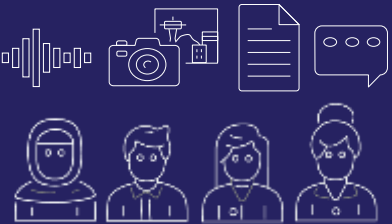Natural language processing

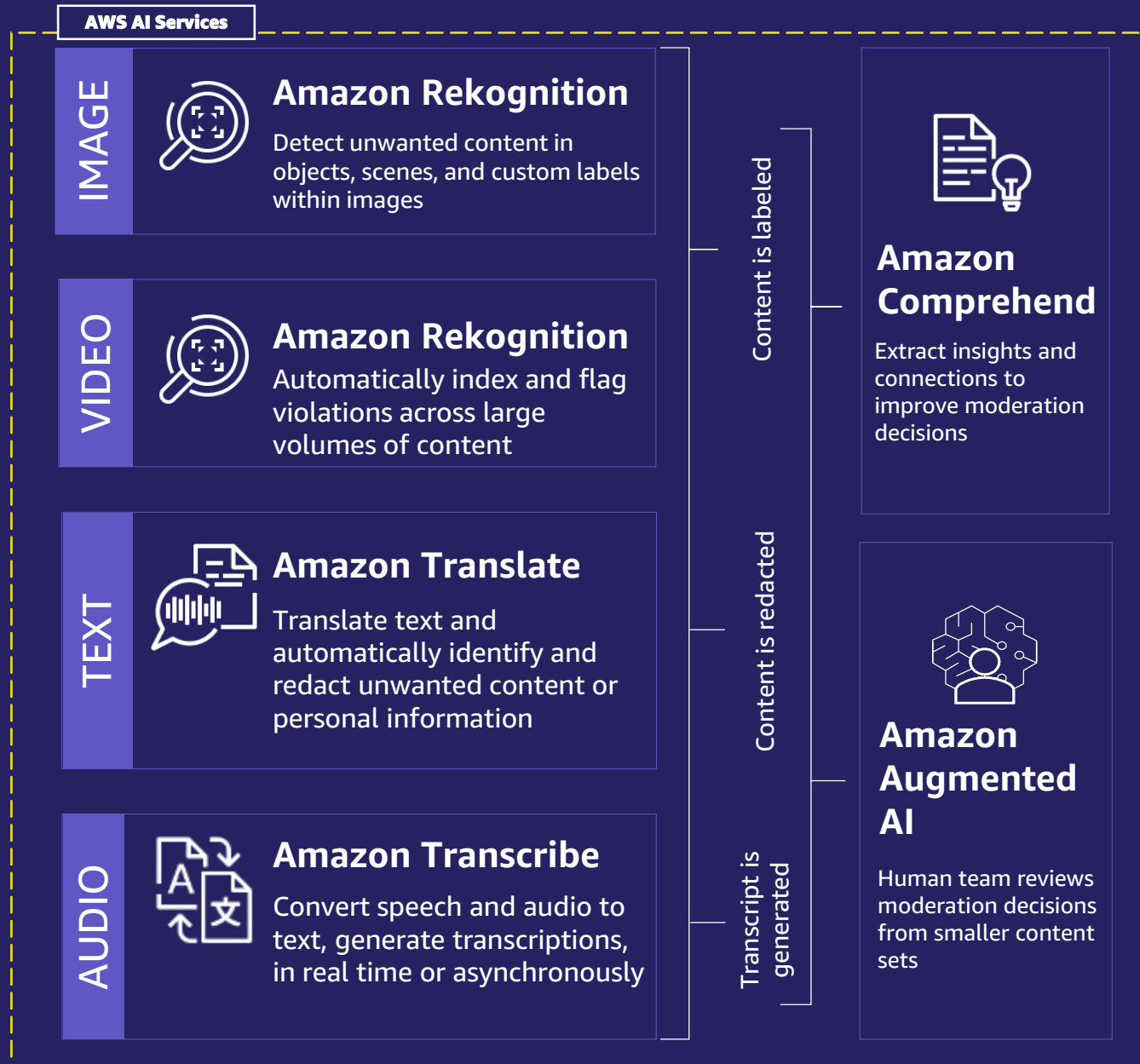**Amazon Translate**

Translation

Amazon Augmented AI

Human reviews

How it works

Users generate and upload content in one or more formats to share in an online community

AWS AI Services

IMAGE

Amazon Rekognition
Detect unwanted content in objects, scenes, and custom labels within images

VIDEO

Amazon Rekognition
Automatically index and flag violations across large volumes of content

TEXT

Amazon Translate
Translate text and automatically identify and redact unwanted content or personal information

AUDIO

Amazon Transcribe
Convert speech and audio to text, generate transcriptions, in real time or asynchronously

Content is labeled

Content is redacted

Transcript is generated

Amazon Comprehend
Extract insights and connections to improve moderation decisions

Amazon Augmented AI
Human team reviews moderation decisions from smaller content sets

Moderated content flows back into the community

aws

Amazon Rekognition

A fully managed AI service that automatically extracts information from images and video files with machine learning to build computer vision apps and models **faster, less costly, and customizable** to your business needs.

Amazon Rekognition

Builders innovate faster with fully managed moderation, image, and video APIs

Content Moderation

Amazon Rekognition

OBJECT, SCENE, AND ACTIVITY

TEXT DETECTION

VIDEO SEGMENTS & SHOTS

FACE DETECTION AND ANALYSIS

CELEBRITY RECOGNITION

LIVE STREAM VIDEO & PATHING

FACE COMPARE & SEARCH

CONTENT MODERATION

CUSTOM LABELS

Why Amazon Rekognition

Build faster
Ready to integrate SDKs & APIs

High quality results

No ML Expertise required

Simplify operations
No ML models to build or manage

Reduce manual work

Continuous improvements

Customize easily
Amazon Rekognition Custom Labels

Detect business-specific labels

Run trained models in your account

KEY FEATURES

Detect
unwanted content in images and videos

Obtain granular output across
10 main and 35 sub categories

Apply your own policies and define
what is safe for your customers

Broadest and
deepest taxonomy

Detect objects,
scenes, and actions

Detect text in 7
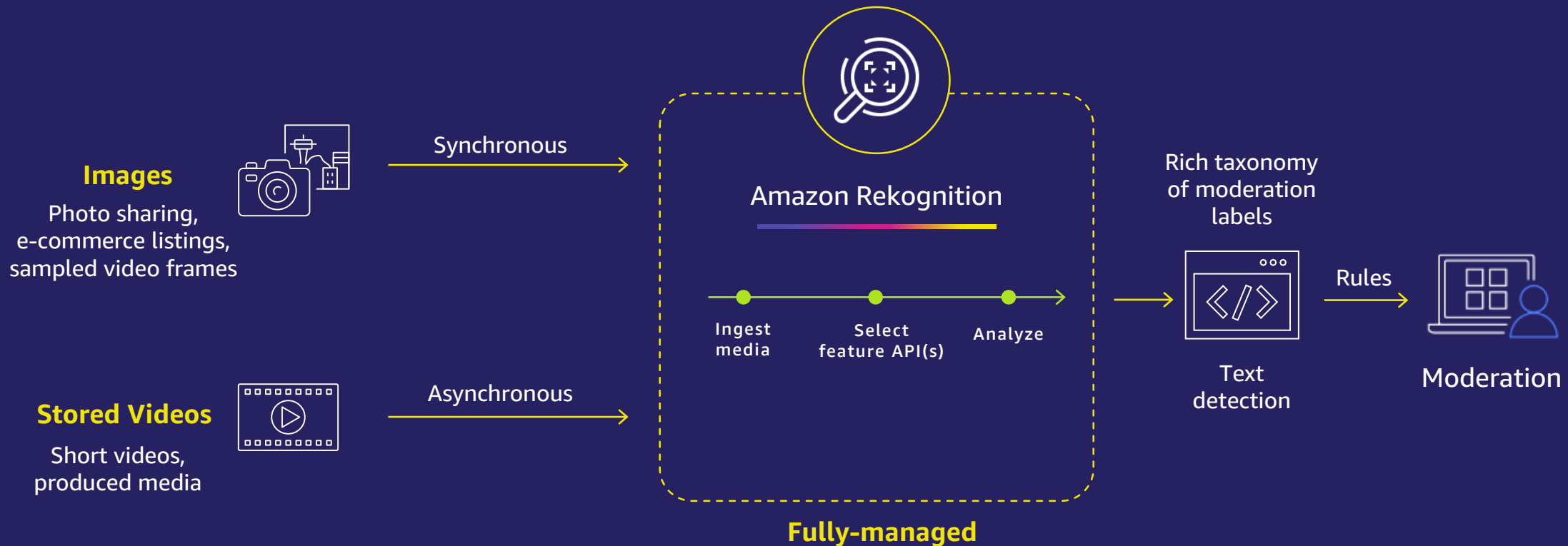languages

Confidence scores
for each detection

Pre-trained and
trainable models

Fully managed
APIs

aws

How Amazon Rekognition works

Images
Photo sharing, e-commerce listings, sampled video frames

Synchronous →

Stored Videos
Short videos, produced media

Asynchronous →

Amazon Rekognition

Ingest media — Select feature API(s) — Analyze

Fully-managed

Rich taxonomy of moderation labels →

Text detection

Rules →

Moderation

Content Moderation

Top-level category	Second-level category
Explicit Nudity	Nudity Graphic Male Nudity Graphic Female Nudity Sexual Activity Illustrated Nudity Adult Toys
Suggestive	Female Swimwear Or Underwear Male Swimwear Or Underwear Partial Nudity Revealing Clothes
Violence	Graphic Violence Or Gore Physical Violence Weapon Violence Weapons Self Injury
Visually Disturbing	Emaciated Bodies Corpses Hanging

Top-level category	Second-level category
Rude Gestures	Middle Finger
Drugs	Drug Products Drug Use Pills Drug Paraphernalia
Tobacco	Tobacco Products Smoking
Alcohol	Alcoholic Beverages Drinking
Gambling	Gambling
Hate Symbols	Nazi Party White Supremacy Extremist

Amazon Rekognition does not flag content as "safe" or "unsafe"

Instead, you get granular output about the nature of the content. You can apply your own policies that define what is safe for your customers.

aws

Rekognition Content Moderation API

Request

```
{
    "Image": {
        "S3Object": {
            "Bucket": "bucket",
            "Name": "input.jpg"
        }
    },
    "MinConfidence": 60
}
```

Response

```
{
"ModerationLabels": [
    {
        "Confidence": 99.24723052978516,
        "ParentName": "",
        "Name": "Explicit Nudity"
    },
    {
        "Confidence": 99.24723052978516,
        "ParentName": "Explicit Nudity",
        "Name": "Graphic Male Nudity"
    },
    {
        "Confidence": 88.25341796875,
        "ParentName": "Explicit Nudity",
        "Name": "Sexual Activity"
    }
]
}
```

Amazon Rekognition Custom Labels

Detect objects and scenes unique to your business

Amazon Rekognition
Custom Labels

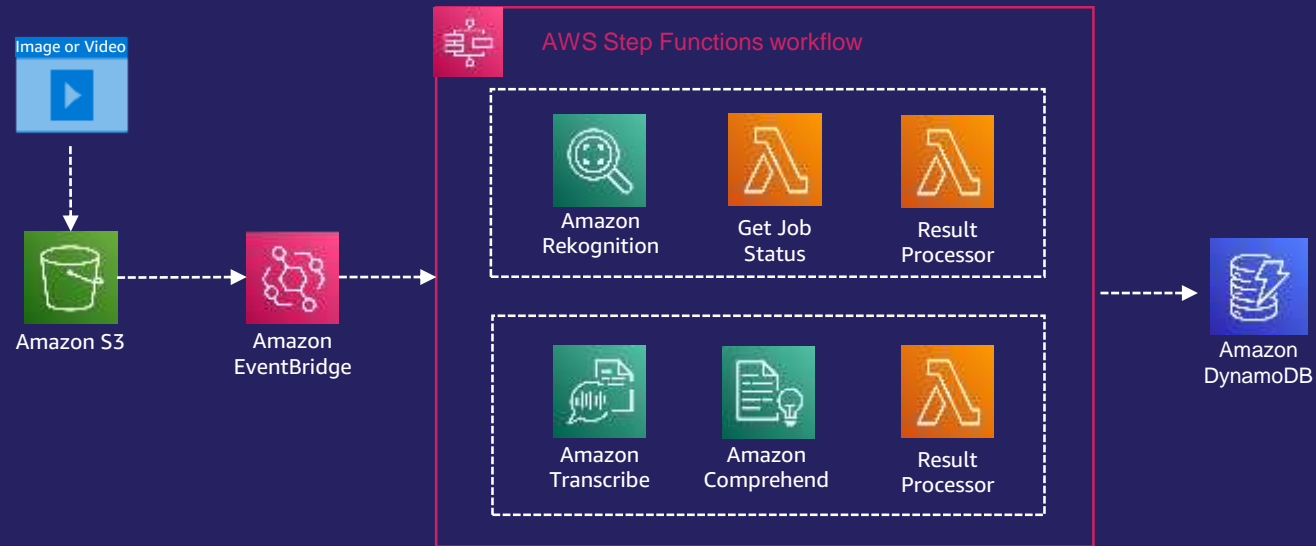**Guided step-by-step** experience or programmatic APIs to create and manage datasets

Train with high accuracy ML models with one click/API call, and with as few as 10 images

Analyze images at scale using endpoint

Image and Video Processing

Demo
Amazon Rekognition Content Moderation

aws

AWS Computer Vision: an IDC leader in Asia

AWS offers more computer vision solutions and services than any other cloud provider, relevant to key use cases

- Productivity

- End user experience

- Decision recommendation

IDC MarketScape: Asia/Pacific (Excluding Japan) Vision Artificial Intelligence Software Platform 2021

Leaders

Amazon Web Services (AWS)

Google

Microsoft

NEC

Alibaba Group

Tencent

Major Players

SAS

Samsung SDS

Capabilities

Contenders

Participants

Strategies

Source: IDC, 2021

Amazon Transcribe

A fully managed and continuously trained automatic speech recognition (ASR) service that **converts speech in audio and videos** to text, while checking for use of profanities, PII, or other sensitive information in the process.

Amazon Transcribe

Automatically convert speech to text

CONTENT MODERATION

Amazon Transcribe

AUDIO INPUTS

READABLE TRANSCRIPTS

OUTPUT CUSTOMIZATION

SAFETY AND PRIVACY

CALL ANALYTICS

TRANSCRIBE MEDICAL

Why Amazon Transcribe

Fully-managed and continuously trained

Highly accurate, efficient, and scalable

Low latency for live streaming

KEY FEATURES

Enable content filtering and privacy with vocabulary filtering and content redaction

Customize transcription output with alternative transcriptions, custom vocabulary and custom language models

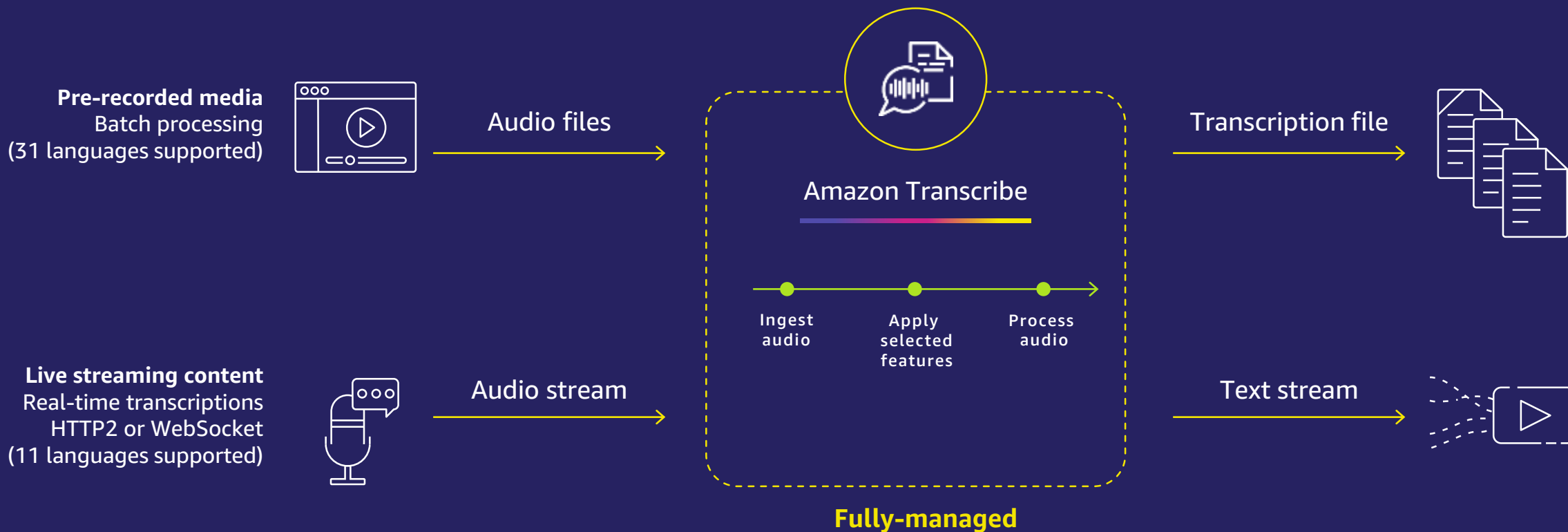| Language Identification	Word-level confidence scores	Word-level time stamps	Channel Identification	Speaker labeling	Punctuation + Capitalization

aws

How Amazon Transcribe works

Pre-recorded media
Batch processing
(31 languages supported)

Audio files →

Amazon Transcribe

Ingest audio → Apply selected features → Process audio

Transcription file →

Live streaming content
Real-time transcriptions
HTTP2 or WebSocket
(11 languages supported)

Audio stream →

Text stream →

Fully-managed

Amazon Comprehend

A fully managed natural-language processing (NLP) service that uses AI to **understand valuable insights and connections in text.**

Amazon Comprehend

Differentiate your business with valuable insights from text within documents

CONTENT MODERATION

Amazon Comprehend

PRE-TRAINED AND CUSTOM ENTITY RECOGNITION

CUSTOM CLASSIFICATION

EVENTS DETECTION

PII IDENTIFICATION AND REDACTION

KEYPHRASE EXTRACTION

MULTI-LINGUAL SUPPORT

LANGUAGE DETECTION

SYNTAX ANALYSIS

SENTIMENT ANALYSIS

TOPIC MODELING

How Amazon Comprehend works

Social media
Gaming/chat forums, reviews

text →

Media content
Broadcast media, talk radio, video content, and audio files

audio →

Amazon Transcribe

text →

Amazon Comprehend

Entities
+ Custom Entities

Sentiment

Key Phrases

Syntax

PII
(Personally Identifiable Information)

Language

Events

Topics

Document Classification
Custom Classifiers

Moderate

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

PII Detection

Hi, my name is John Doe. For verification, the last 4 digits of my social are 6789 and my dob is 01/01.

I paid for my credit card 1111-0000-1111-0000 last week from my bank account XXXXXX1111 with the routing number XXXXX0000. The check was mailed from 100 Main Street, Anytown, WA 98121.

Please confirm receipt by calling me at 206-555-0199 or emailing at john.doe@anycompany.com.

Value	Entity
John Doe	Entity: NAME
6789	Entity: SSN
01/01	Entity: DATE_TIME
1111-0000-1111-0000	Entity: CREDIT_DEBIT_NUMBER
XXXXXX1111	Entity: BANK_ACCOUNT_NUMBER
XXXXX0000	Entity: BANK_ROUTING
100 Main Street, Anytown, WA 98121	Entity: ADDRESS
206-555-0199	Entity: PHONE
john.doe@anycompany.com	Entity: EMAIL

PII handing options: Detect, Redact, Mask

Detect

Hi, my name is John Doe. For verification, the last 4 digits of my social are 6789 and my dob is 01/01.

I paid for my credit card 1111-0000-1111-0000 last week from my bank account XXXXXX1111 with the routing number XXXXX0000. The check was mailed from 100 Main Street, Anytown, WA 98121.

Please confirm receipt by calling me at 206-555-0199 or emailing at john.doe@anycompany.com.

Redact

Hi, my name is [NAME]. For verification, the last 4 digits of my social are [SSN] and my dob is [DATE_TIME].

I paid for my credit card [CREDIT_DEBIT_NUMBER] last week from my bank account [BANK_ACCOUNT_NUMBER] with the routing number [BANK_ROUTING]. The check was mailed from [ADDRESS].

Please confirm receipt by calling me at [PHONE] or emailing at [EMAIL].

Mask

Hi, my name is ********. For verification, the last 4 digits of my social are **** and my dob is *****.

I paid for my credit card ******************* last week from my bank account ********** with the routing number **********. The check was mailed from ************************.

Please confirm receipt by calling me at ************ or emailing at **********************.

Bringing all AWS AI services together

Automated video content analysis with **AWS Content Analysis**

MEDIA INSIGHTS ENGINE

control plane

9 AWS Step Functions workflow

media analysis and transformation operators

Amazon Transcribe · Amazon Translate · Amazon Comprehend

Amazon Rekognition · AWS Elemental MediaConvert · custom operators

6 Amazon DynamoDB

5 AWS Lambda · **7** Amazon Simple Queue Service · **8** AWS Lambda

4 Amazon API Gateway

3 Amazon Cognito user pool · Amazon Cognito identity pool · IAM

1 Amazon CloudFront · **2** Amazon S3 web source bucket · Users

data plane

10 Amazon API Gateway

11 AWS Lambda

data pipeline

media and metadata storage

12 Amazon DynamoDB data plane table · Amazon S3 data plane bucket

13 Amazon Kinesis Data Streams

data pipeline consumers

OpenSearch consumer

AWS Lambda · Amazon OpenSearch Service (successor to Amazon Elasticsearch Service)

Moderate and innovate with our partners

AWS PARTNER NETWORK

Let AWS certified partners with moderation expertise help you with a POC and build integrations with your current systems to productionize

CONSULTANCY PARTNER

INDEPENDENT SOFTWARE VENDOR (ISV)

Next steps

Try and learn

Try all our moderation [AI services free](), and access developer guides, tutorials, Github repositories, use case videos, webinars, getting started guides, pricing, and more for each service.

[Amazon Rekognition]()   [Amazon Translate]()
[Amazon Transcribe]()   [Amazon Comprehend]()
[Amazon Augmented AI]()

Build with us

AWS experts and Solution Architects can help you with additional information or resources needed to get you started.

[Contact us]()

Find a partner

Contact us to help you find the right partner, or explore the AWS partner network to get a POC running.

[AWS Partners]()

Voice of the customer AWS Rekognition @ Rooter

Adhikaar Marwaha
Backend Lead
Rooter Sports Technologies Ltd.

India is the **5th** largest mobile gaming market

7 hrs & 20 mins
Avg time spent per week by 81% of Gen Z gamers

40%
mobile gamers spend INR 230/month

675Mn
2025

510Mn
2022

$7 Billion
Gaming Market[1]
in India by 2026

rooter

END TO END IMMERSIVE PLATFORM

PLAY

Let's CREATE
Gaming Content

Let's WATCH
Streaming / Videos

Click-of-a-Button Creation

Fan Engagement

Cross-platform Streaming

Live Chat

Top trending Overlays & Thumbnails

Discover Live Content

Personalized Feed

Chat with Creators

Earn Rewards/Points

Video on demand

An intuitive & personalized UI with a feature rich and easy to use UX

rooter

6

FAVOURITE AMONGST THE NEW-AGE GAMING AUDIENCE

#1

App in 'Sports' Category
On the Google Play store (Since Oct 2020)
Consistently amongst top 100 apps on google play store

App Installs

35M

Avg Time Spent

32 Mins/day

MAU/DAU

9.3M/850K

Funding Raised

$30M till date

1:28

52%

Rooter: Watch
Gaming & Esports
Rooter Sports
Contains ads

4.4★   ⬇   12+   10M+
377K reviews   27 MB   Rated for 12+ ⓘ   Download

Install

About this app   →

India's Biggest Gaming & Esports Streaming App.
Win Giveaways upto 5 Lakhs.

#1 top free in sports

rooter

LEADING ORGANISER OF PROFESSIONAL & AMATEUR ESPORTS TOURNAMENTS

Reaching out to a large gaming audience

150K+	30K+	200	19	100K
Total ESPORTS players	Amateur teams including schools and colleges	Professional Esports Teams	Total Gaming IPs Owned by Rooter	Mins of Content Produced

100M+	7M+	250+
Avg. Total Viewership	Avg. Per Event Engagement	Avg. Content Pieces

rooter

10

ALREADY SERVING THE TOP BRANDS IN THE COUNTRY

Social

Josh

Chingari

moj

Education

BYJU'S
The Learning App

unacademy

Vedantu

Gaming

DREAM11

MPL

WINZO

Entertainment

voot

ZEE5

prime video

Shopping

Flipkart shopsy

yaari

amazon

Finance

CoinDCX

PhonePe

ICICI Bank

Others

Godrej

TISSOT

vodafone

WINGS

betway

PARI MATCH

~700K/Month
Conversions Facilitated

~75 Brands/Month

rooter

14

WELL FUNDED TO EMERGE AS A LEADER IN GAME STREAMING IN INDIA

Existing Investors

Lightbox

Helping in strategy & industry connects

$400M | AUM

MARCH GAMING

Access to US investors & gaming companies

$1B+ | AUM

FUNDING · THE ECONOMIC TIMES

Gaming startup Rooter raises $25 million from Lightbox, others

Jan 13, 2022

FUNDING · YOURSTORY

[Funding alert] Rooter raises $1.7M led by Paytm, leAD Sports, Rockstud Capital, others

By Sujata Sangwan | May 4, 2020   May 4, 2020

ADvantage

Duane Park

GOAL VENTURES

9UNICORNS India's First Accelerator VC

Paytm

Anthill

leAD SPORTS & HEALTH TECH

VentureCatalysts India's 1st Integrated Incubator

FBC FounderBank Capital

User Content Upload Flow
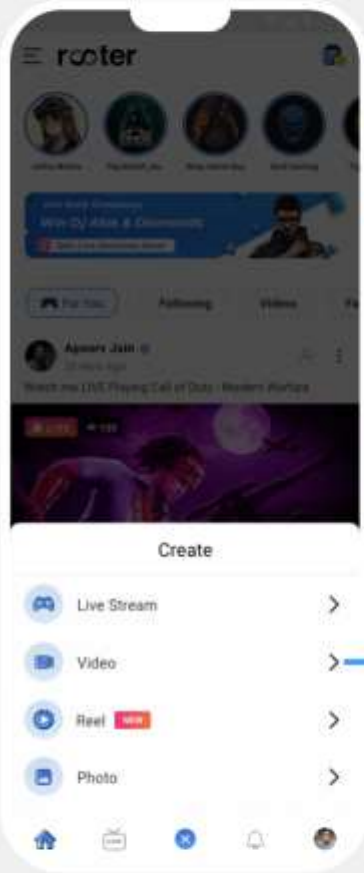
User Selects Creation Option

User selects media to upload
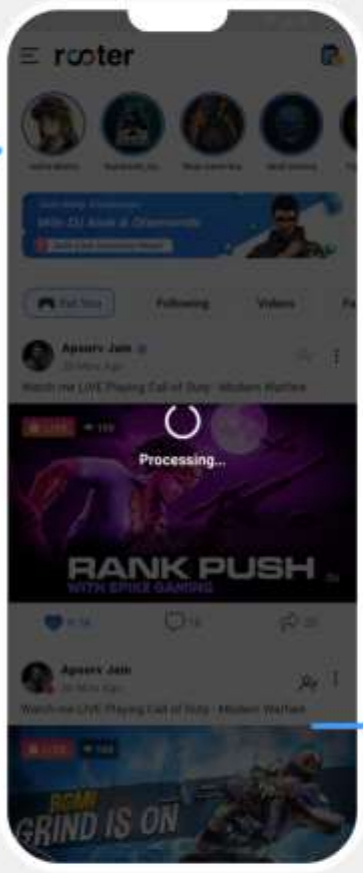
Video Processing.

(Amazon Rekognition detects Moderation Labels)
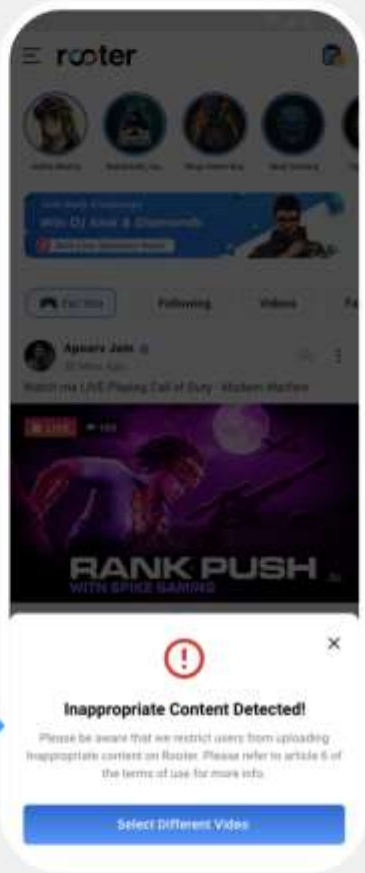
User gets feedback.

(If Inappropriate labels are detected)

Typical Gaming Scene

Challenges with Gaming Content

- Avoiding flagging certain moderation labels

- Violent Content – Weapons, Explosions are commonplace

- Semi Nude Characters are typical in games

Challenges with Gaming Content

Top-level Category	Second-level Category
Explicit Nudity ✗	Nudity
	Graphic Male Nudity
	Graphic Female Nudity
	Sexual Activity
	Illustrated Explicit Nudity
	Adult Toys
Suggestive ✓	Female Swimwear Or Underwear
	Male Swimwear Or Underwear
	Partial Nudity
	Barechested Male
	Revealing Clothes
	Sexual Situations
Violence ✓	Graphic Violence Or Gore
	Physical Violence
	Weapon Violence
	Weapons
	Self Injury
Visually Disturbing ✓	Emaciated Bodies
	Corpses
	Hanging
	Air Crash
	Explosions and Blasts

Confidence(Min is 0 and Max is 100)

Drop or Click here to upload image!

10

Open Rekognition Guide

CONFIDENCE	NAME	PARENT NAME
31.592300415039062	Suggestive	
31.592300415039062	Revealing Clothes	Suggestive
13.796099662780762	Weapon Violence	Violence
13.796099662780762	Violence	

Image Moderation

- Calling the Rekognition DetectModerationLabels API Operation

- Block content upload if Labels are detected above a threshold minimum

  Confidence Level

```
const params = {
  Image: {
    Bytes: data,
  },
  MinConfidence: config.aws_rekognition_min_confidence,
};

const detectModerationLabels = util.promisify(rekognition.detectModerationLabels.bind(rekognition));
try {
  const res = await detectModerationLabels(params);
  console.log('res:', res);
  const detections = [];
  res.ModerationLabels.forEach((currentLabel) => {
    if (Object.keys(config.aws_rekognition_labels).includes(currentLabel.Name)) {
      if (currentLabel.Confidence > config.aws_rekognition_labels[currentLabel.Name]) {
        detections.push({ label: currentLabel.Name, confidence: currentLabel.Confidence });
      }
    }
  });
  if (videoFrame) {
    fs.unlinkSync(image);
  }
  return detections;
```

Video Moderation

- Extracting Equidistant Frames

- Calling the Rekognition detect ModerationLabels API on the frames

- Block content upload if Labels are detected with a threshold minimum Confidence Level

```
if (options.lengthInSec <= 60) {
  const noOfFrames = 1;

  const frames = await extractEquidistantFrames(videoPath, extension, options.lengthInSec, noOfFrames);

  if (frames && frames.length) {
    let ext = frames[0].split('.');
    ext = ext[ext.length - 1];
```

Refinement Steps in Pipeline

- Directly using Rekognition Video Moderation APIs operations StartContentModeration and GetContentModeration

- Using Rekognition Custom Labels by creating Label Training Datasets

- Moderate sized training datasets would work since Rekognition internally uses Transfer Learning techniques

- Human Moderators using Amazon Augmented AI to review flagged content

Business Outcomes

- Around 100 Hours / month of Manual work saved by using Automation

- 15% increase in Feed Interaction due to better feed experience

- Content Reports have come down by 75%

- Less burden on the Legal Team regarding explicit content reports

- Identify Spam Creators and Adding Trust Factors in Good Creators, helping in improving the feed algorithms

Personalisation Roadmap

Personalisation

General content doesn't work anymore. Customers want their interactions with technology to feel more fulfilling, they want it to feel unique. Personalization is the way forward.

As we have multiple types of content on our platform, personalising them via traditional methods can be challenging especially if we are starting from scratch, and that is where AWS comes in.
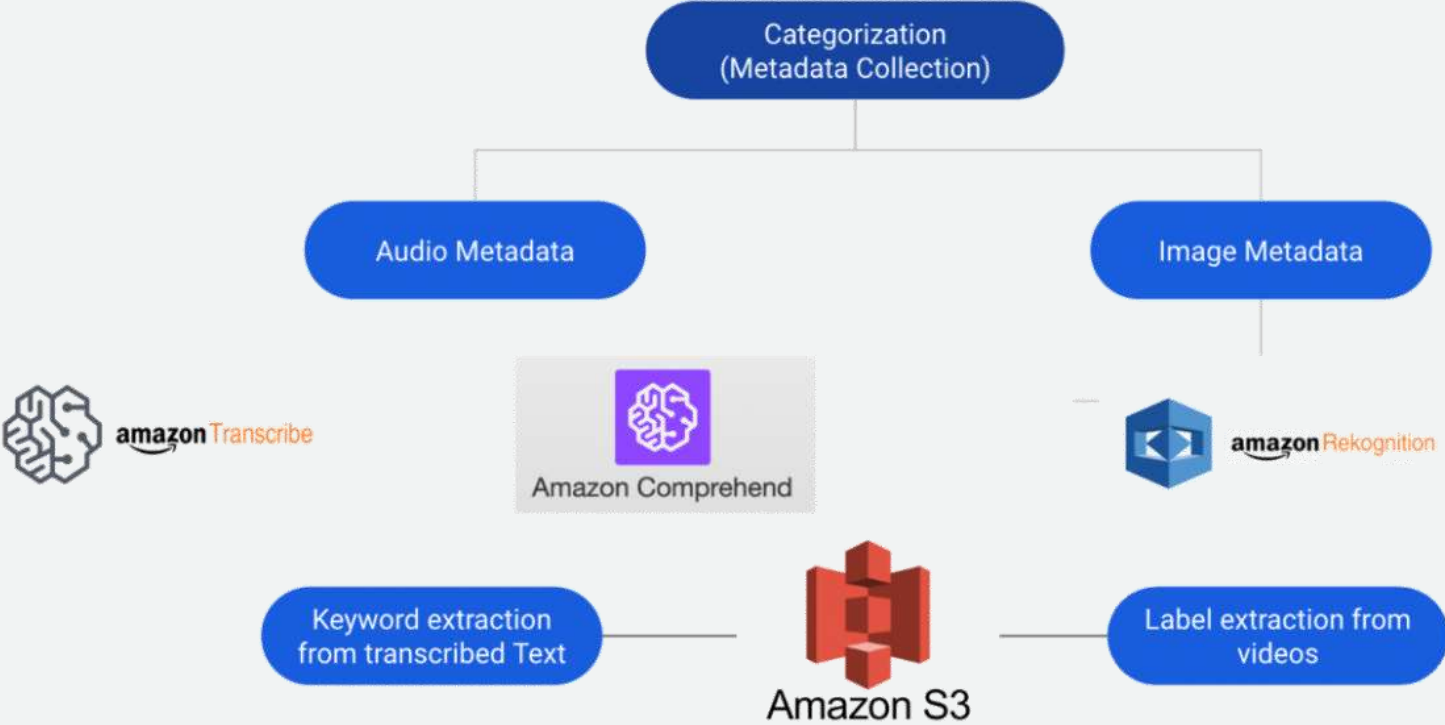
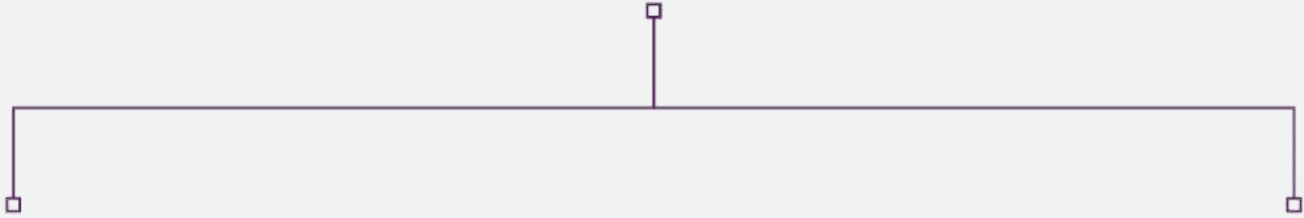The following steps are required in order to create a reasonable personalised recommendation engine.

Step 1	Step 2	Step 3	Step 4
Content Identification	**Metadata Collection**	Feature Selection	Recommendation
Separating video and Live content is necessary as the next steps would depend on it	Identifying the language, game and other aspects of the content is extremely necessary for categorizing them into various buckets	Be it user attributes or content metadata, it is vital to identify the properties that affect personalisation the most, so that we can minimise noise.	Based on the features, interest graph of each user is to be built.

Personalisation Roadmap

Personalisation Roadmap
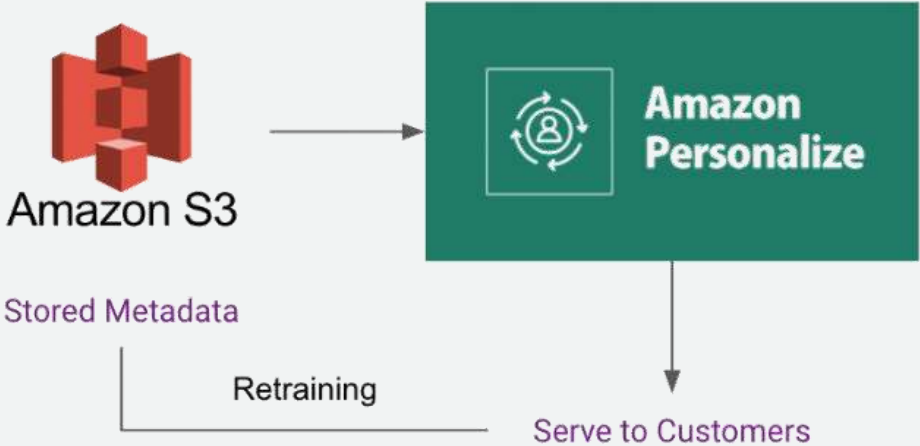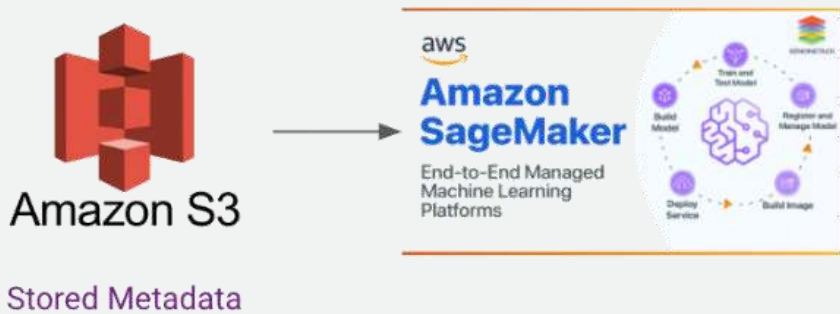
Thank you!

ML Enthusiasts & Professionals

Get In Touch:

adhikaar.marwaha@rooter.io

Q&A