Enhancing Content Using Amazon AI

Julien Simon, AI Evangelist, EMEA
@julsimon
http://medium.com/@julsimon

November 20, 2017

Dog or Muffin?

Confidence	Labels
99.2%	Animal Dog Chihuahua
98.6%	Food Dessert Muffin
97.9%	Collage

Word or Logo?

Algorithm	Viability
OCR	Are you feeling lucky?
Perceptual Hash	Not a chance
Deep Logo Analysis	Bingo

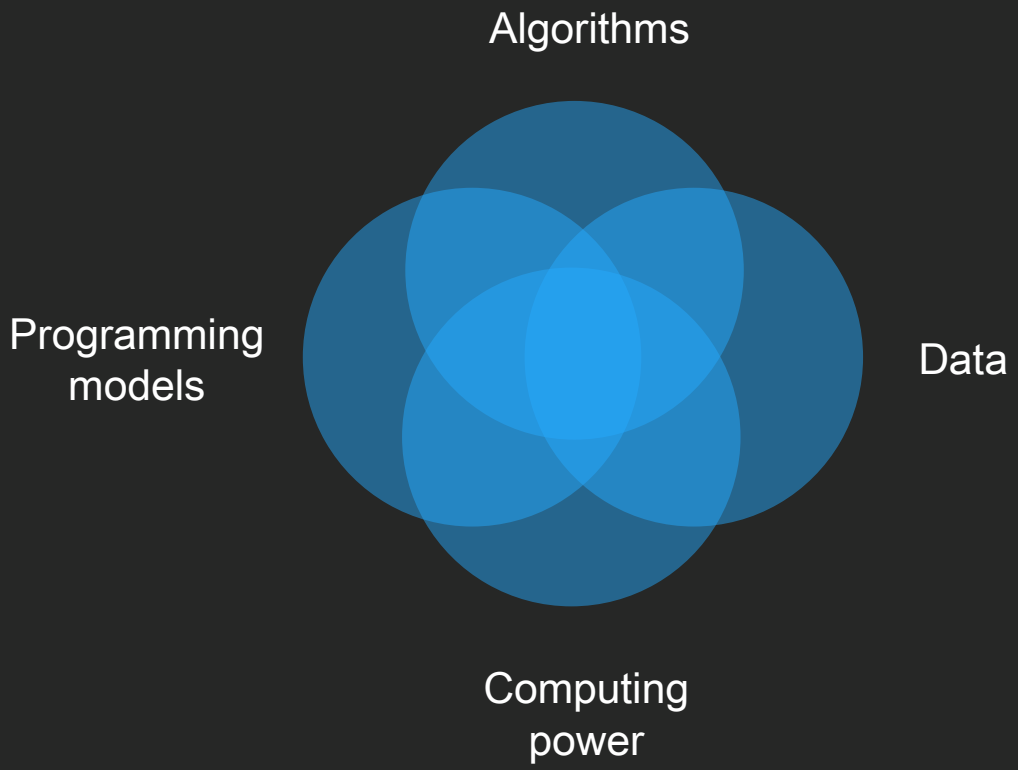
aws

Artificial Intelligence: design software applications which exhibit human-like behavior, e.g. speech, natural language processing, reasoning or intuition

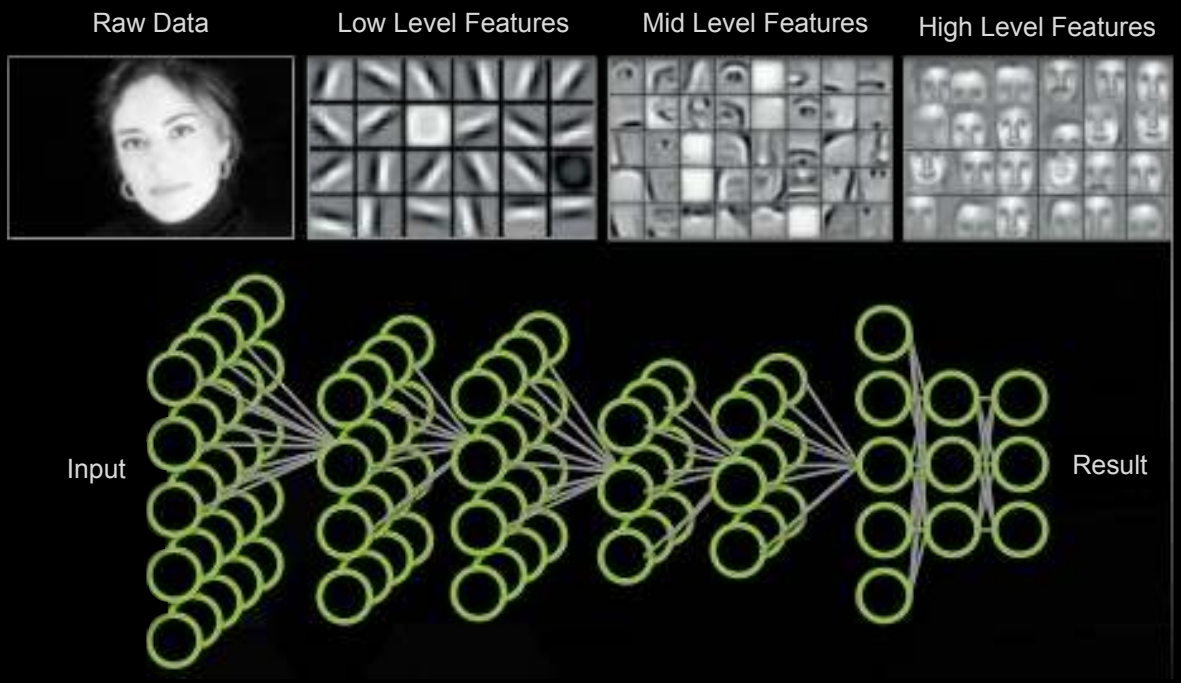
Machine Learning: teach machines to learn without being explicitly programmed

Deep Learning: using neural networks, teach machines to learn from data where features cannot be explicitly expressed

The Rise of Deep Learning



The 10,000ft Intro to Deep Learning



Application Components

Task
Identify a Face

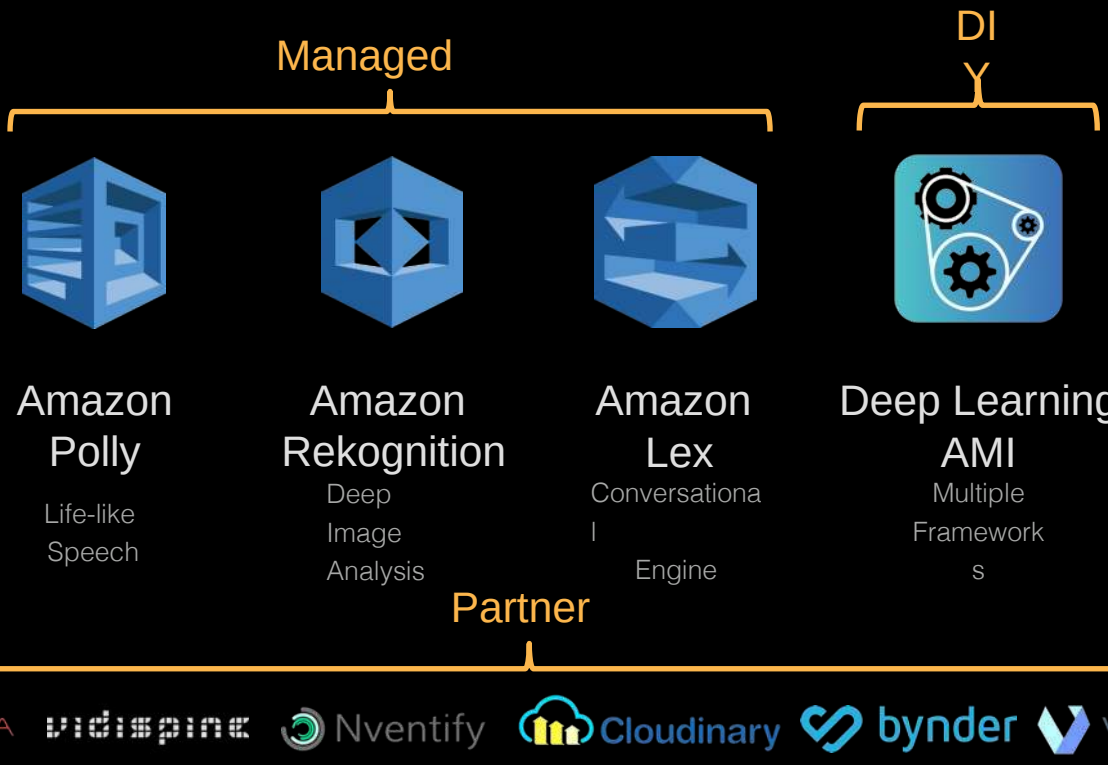
Training
10-100M images

Network
~ 10 layers
1B parameters

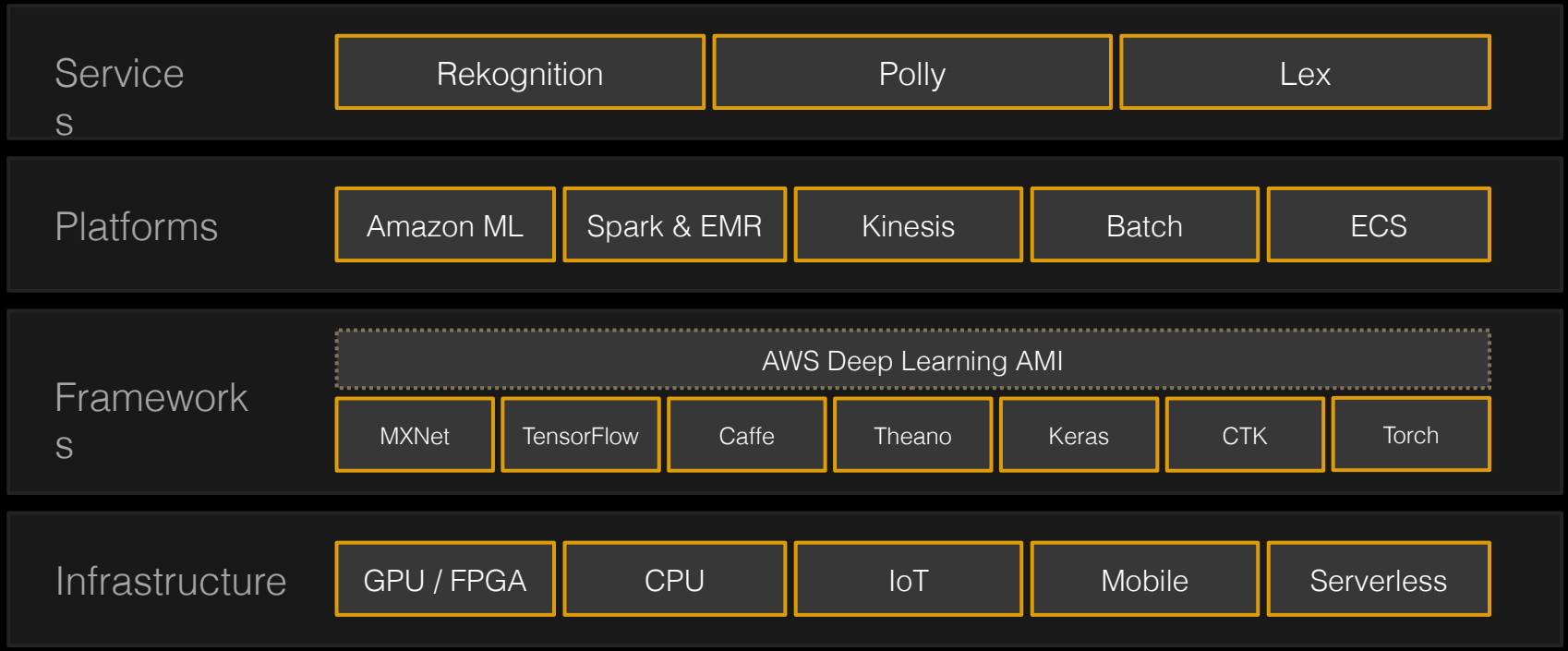
Learning
~ 30 Exaflops
~ 30 GPU days

aws

AWS Services & Partners



The Amazon AI Stack



Choosing the right Instance Type for AI

P3: Distributed Training

NVIDIA V100 GPUs

C5: Inference

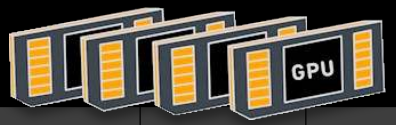
Intel Skylake CPUs

X1: AI/ML/DL at scale

128 vCPUs, 3,904 GiB RAM

F1: FPGA acceleration

Xilinx Ultrascale Plus, 6,800 engines

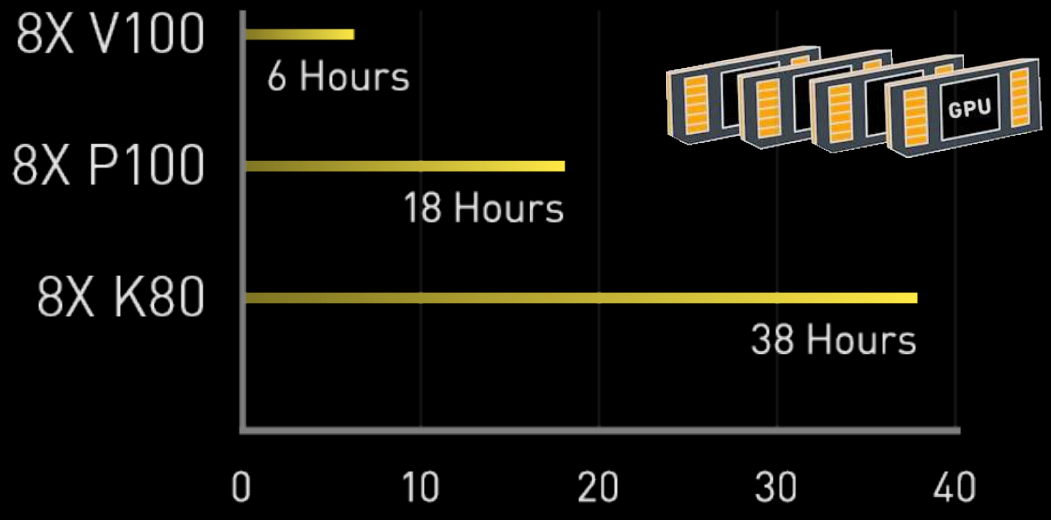


Instance Name	GPU Count	vCPU Count	Memory	Network	EBS
p3.xlarge	1	8	61 GiB	~10Gbps	1.5 Gbps
p3.8xlarge	4	32	244 GiB	10Gbps	7Gbps
p3.16xlarge	8	64	488 GiB	25Gbps	14Gbps

*P3 Instances Provide up to **1 Petaflop** of mixed precision performance, and 125 Teraflops of single precision floating point*

aws

Why is this Important?



Amazon EC2 Compute & EBS block storage supports second-level billing.
Combined with EC2 SPOT Fleet, this provides a up to ~90% cost savings over on-demand.

Deep Learning Compute

- One-click launch
- Single node or distributed
- CPU, GPU, FPGA
- NVIDIA & Intel libraries
- Anaconda Data Science Platform
- Python w/ AI/ML/DL libraries

aws

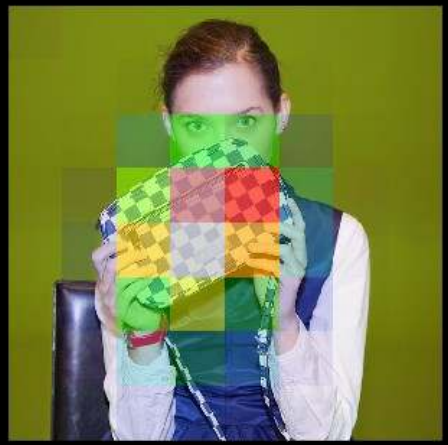
Expedia®

- Expedia have over 10M images from 300,000 hotels
- Using great images boosts conversion
- Using Keras and EC2 GPU instances, they fine-tuned a pre-trained Convolutional Neural Network using 100,000 images
- Hotel descriptions now automatically feature the best available images

https://news.developer.nvidia.com/expedia-ranking-hotel-images-with-deep-learning/

CONDÉ NAST

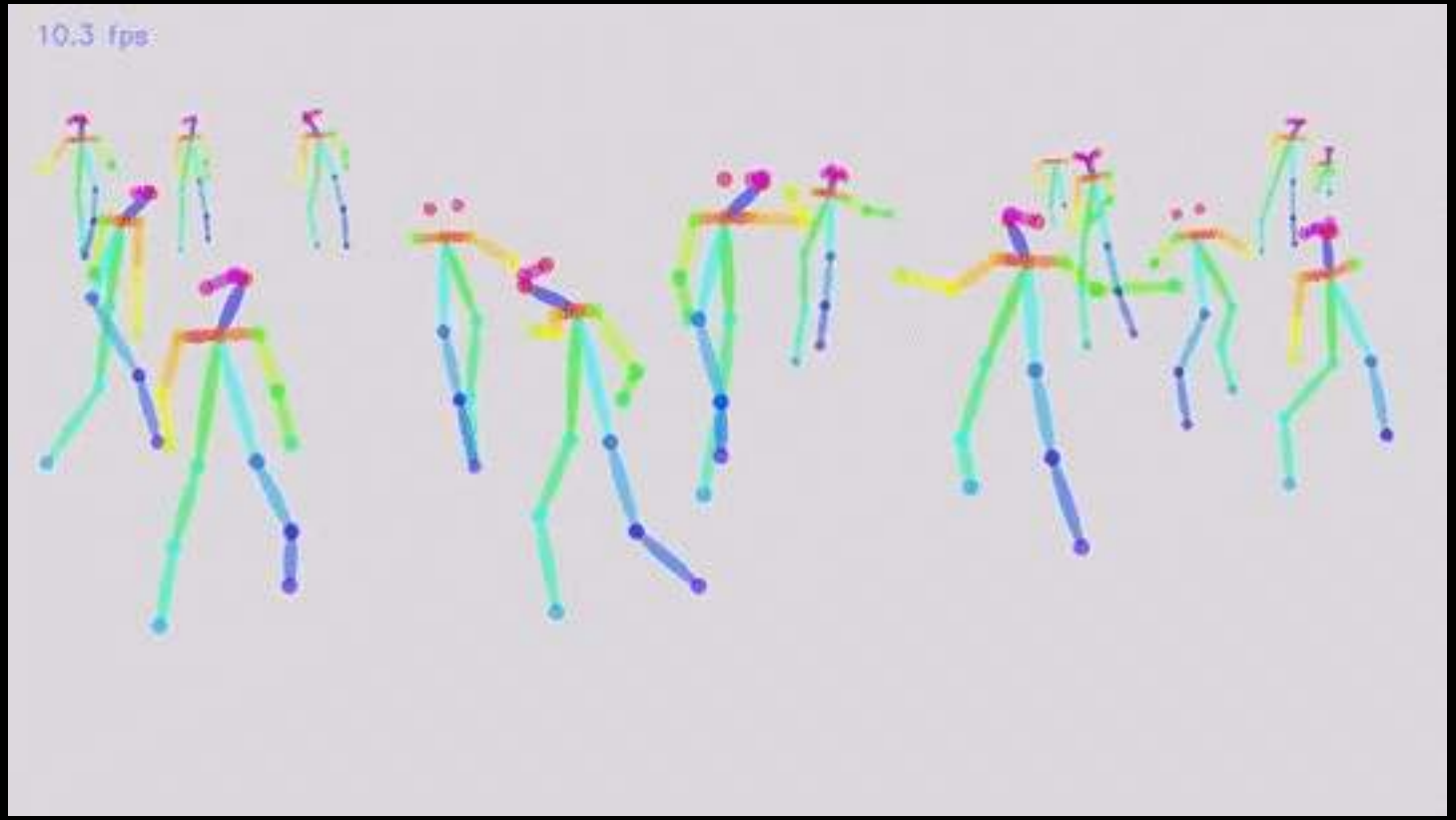
- 17,000 images from Instagram
- 7 brands
- Deep Learning model pre-trained on ImageNet
- Fine-tuning with TensorFlow and EC2 GPU instances
- Additional work on color extraction



	Chanel	Coach	Gucci	Marc Jacobs	Kate Spade	No Handbag	Prada	Vuitton
Chanel	0.83	0.00	0.01	0.02	0.00	0.00	0.00	0.01
Coach	0.01	0.85	0.00	0.05	0.05	0.01	0.04	0.03
Gucci	0.01	0.00	0.85	0.02	0.00	0.01	0.01	0.02
Marc Jacobs	0.00	0.03	0.01	0.78	0.00	0.01	0.03	0.00
Kate Spade	0.00	0.01	0.01	0.01	0.87	0.00	0.00	0.00
No Handbag	0.09	0.06	0.08	0.09	0.04	0.97	0.04	0.09
Prada	0.03	0.03	0.02	0.03	0.01	0.00	0.85	0.01
Vuitton	0.01	0.00	0.00	0.02	0.00	0.01	0.01	0.81

https://technology.condenast.com/story/handbag-brand-and-color-detection

Real-Time Pose Estimation



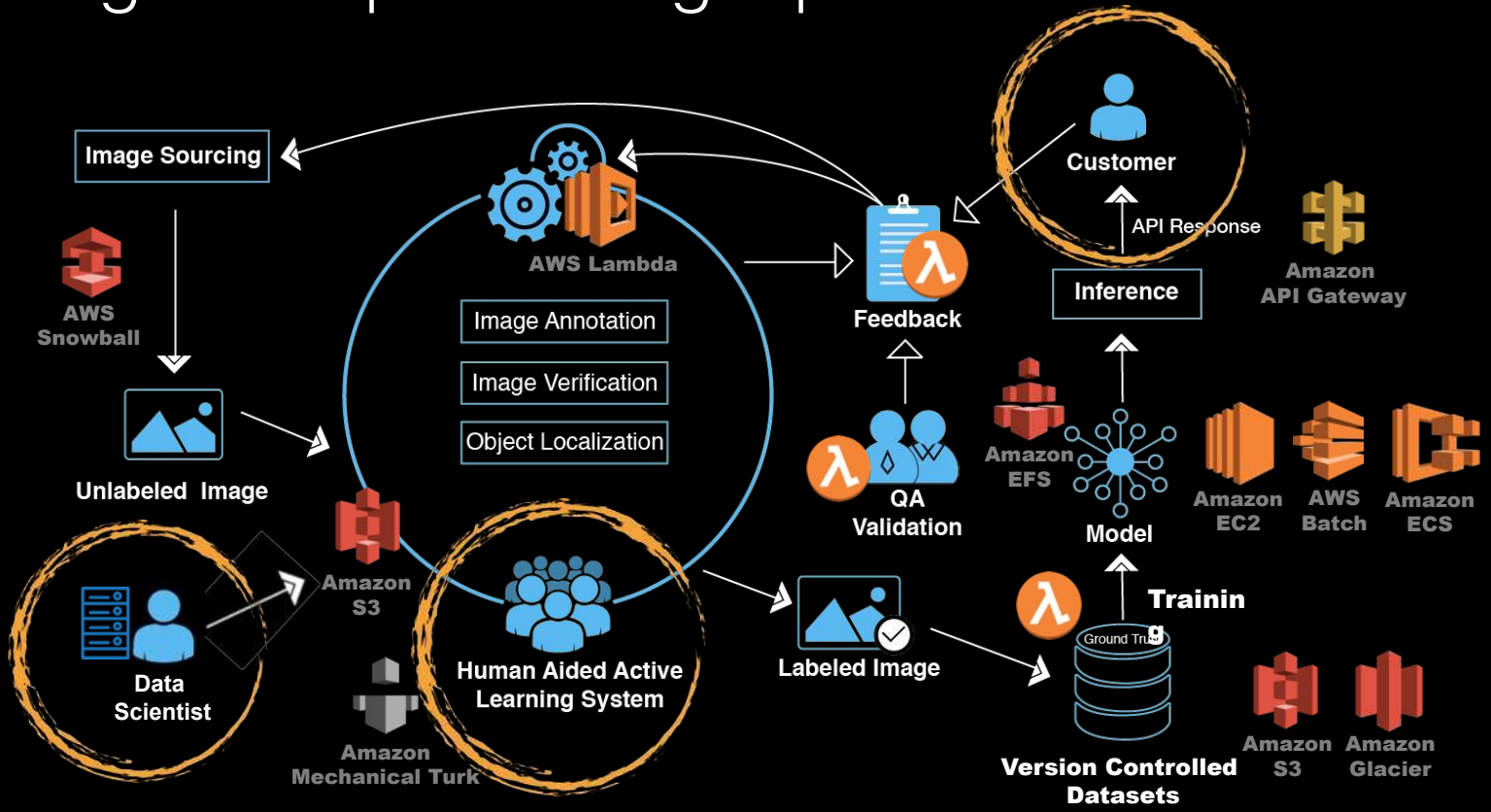
https://github.com/dragonfly90/mxnet_Realtime_Multi-Person_Pose_Estimation

Generative adversarial networks

these faces are not real, they have been generated!

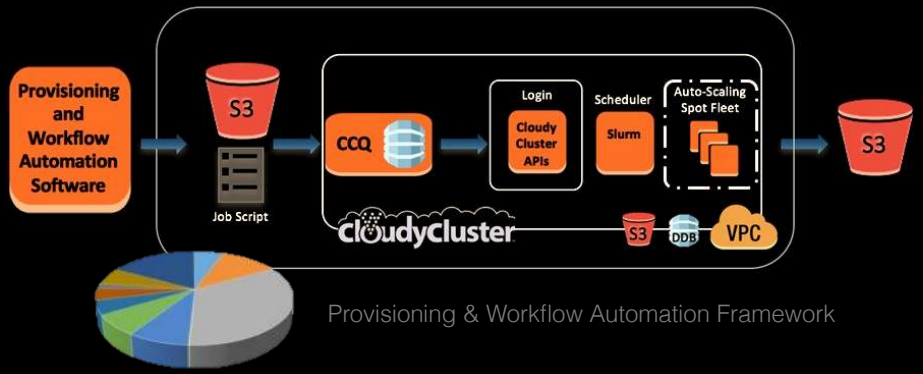
https://github.com/tkarras/progressive_growing_of_gans

Building A Deep Learning Pipeline in the Cloud



Large Scale Document Analysis

- NLP Topic Modeling @ Clemson University
- 533,560 Documents, 32,551,540 Words
- 1.1 million vCPUs over ~3hrs
- EC2 Spot, Single AWS Region
- SLURM scheduler - overlay virtual workflow automation
- Per second billing for EBS & EC2

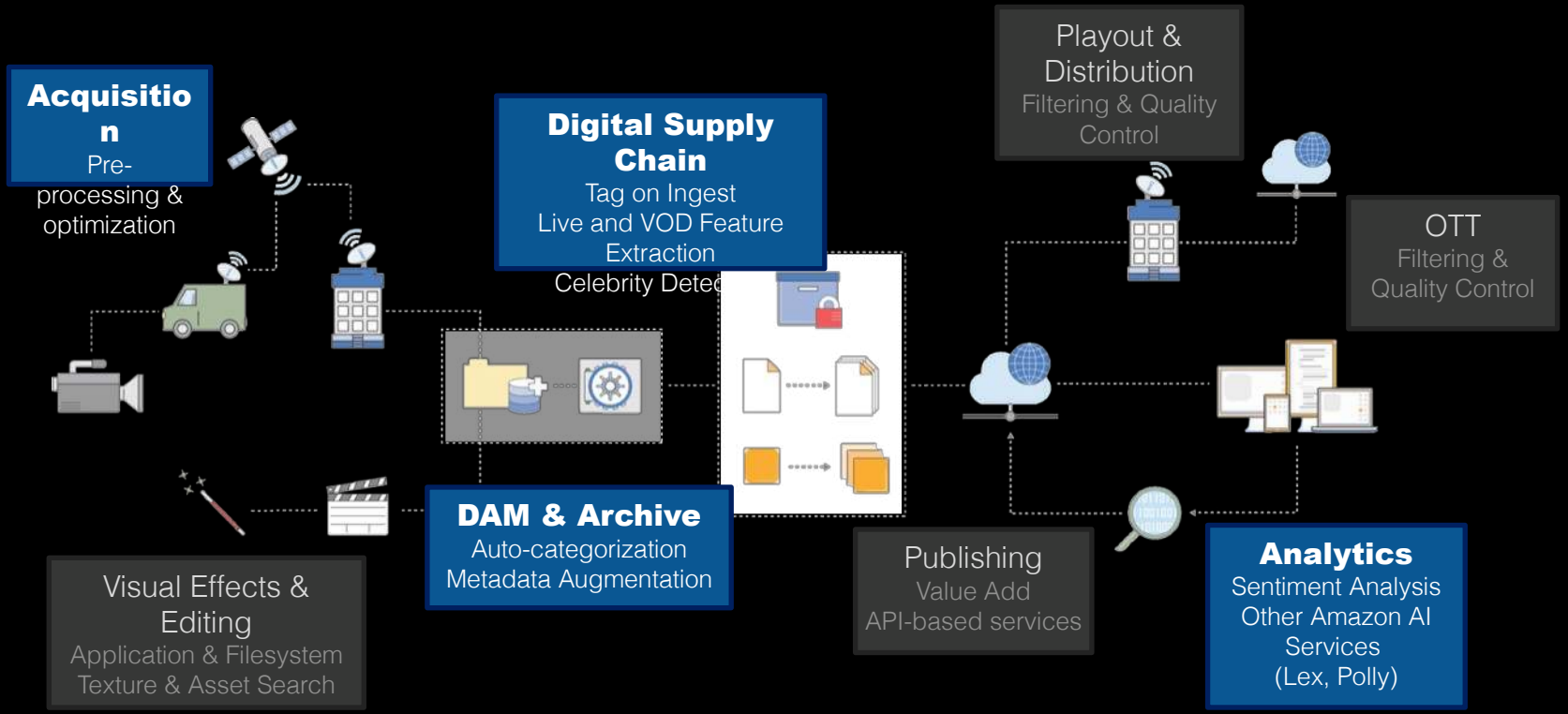


Provisioning & Workflow Automation Framework

17 years of computer science journal abstracts and full text papers
from the NIPS (Neural Information Processing Systems) Conference (2,484 documents and 3,280,697 words)

aws

Deep Learning (& AI) for Media



aws

Amazon Rekognition

Deep learning-based image recognition service

Search, verify, and organize millions of images

Object and Scene Detection

Facial Analysis

Face Comparison

Facial Recognition

Celebrity Recognition

Image Moderation

aws

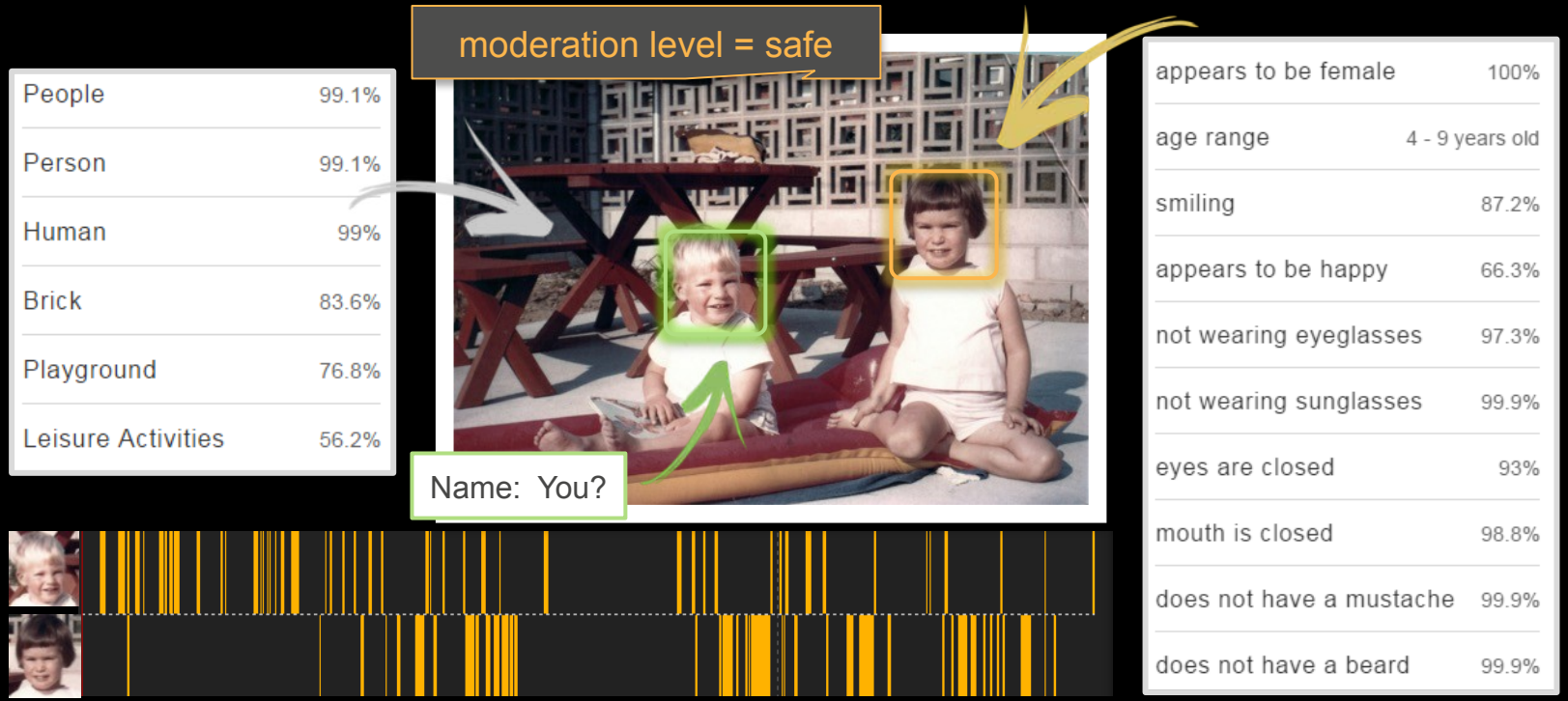
Deterministic Response Time

~500ms Object & Scene Detection

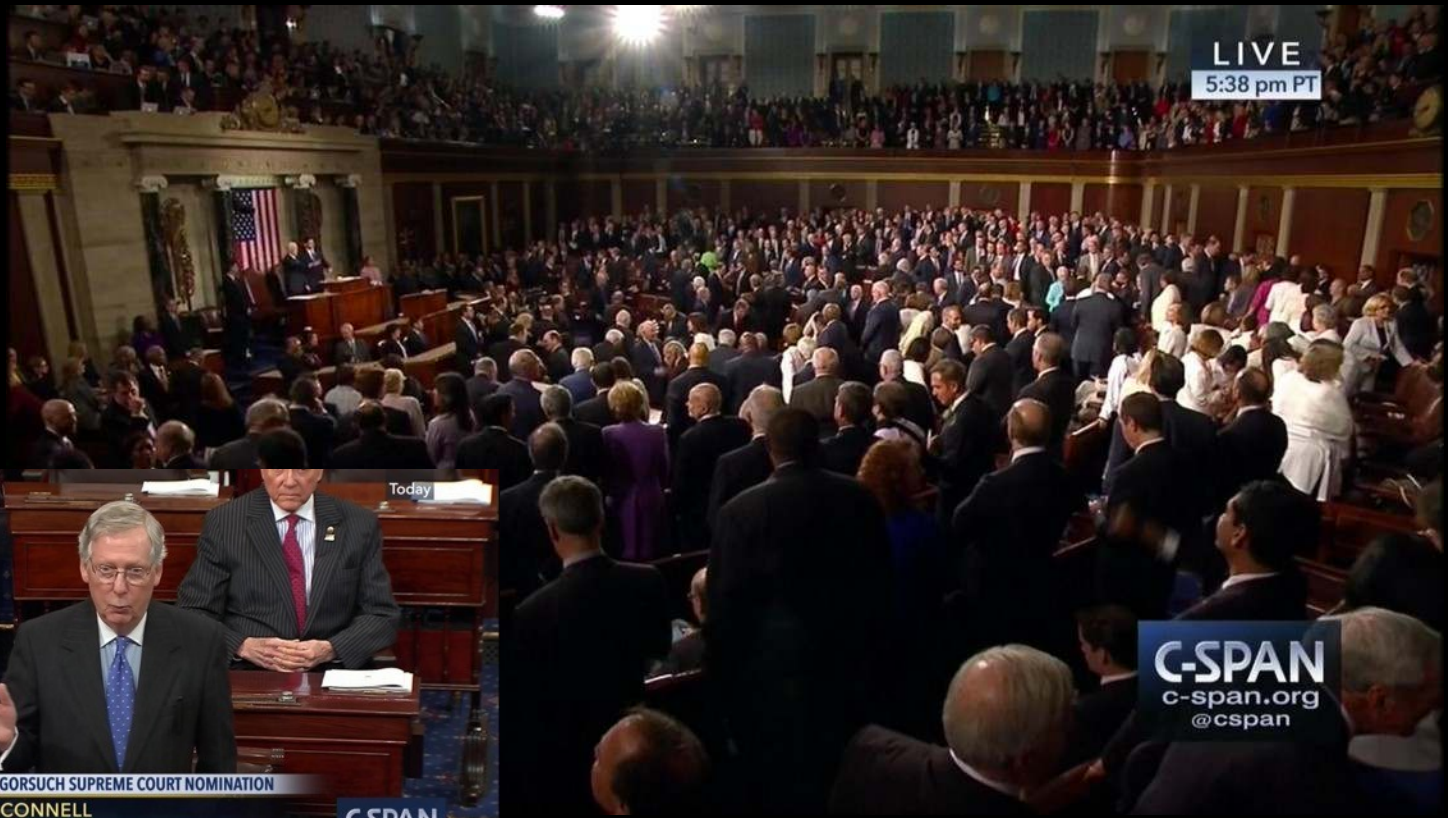
~1.5s Search for 1mil Face Collection

aws

Building Rich Metadata Indexes using Rekognition



C-SPAN

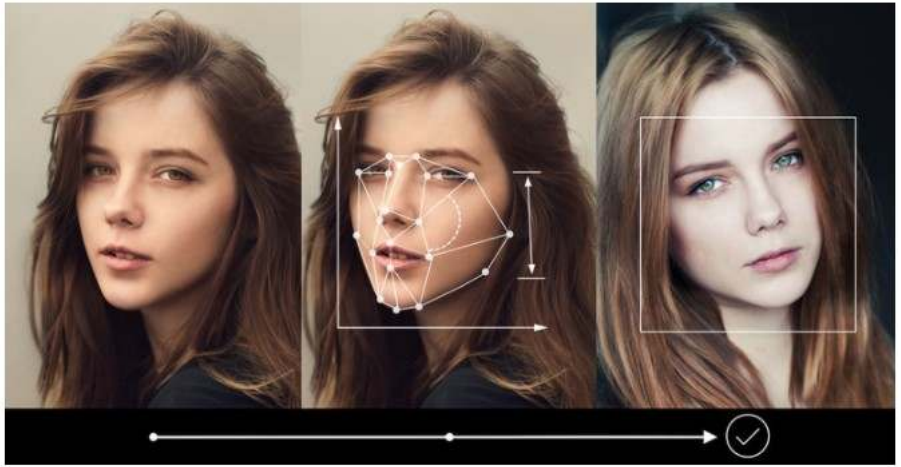


https://aws.amazon.com/solutions/case-studies/cspan/

October 19, 2017

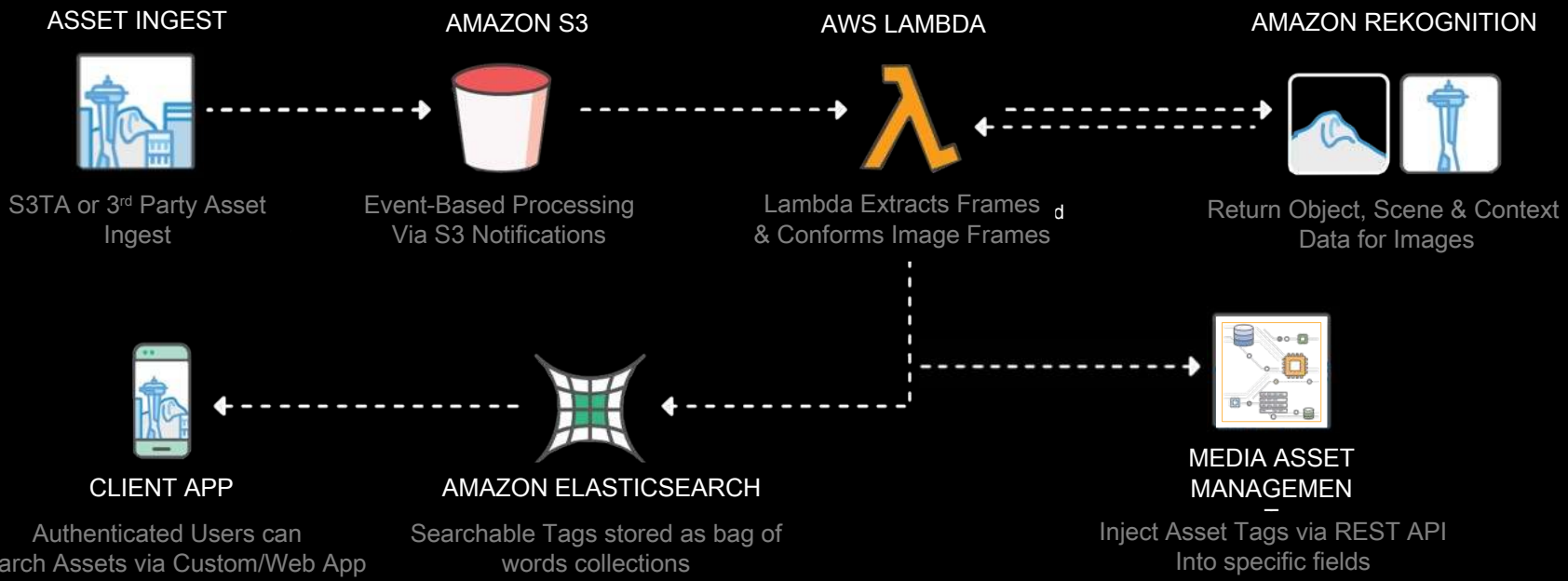
Amazon Rekognition Helps Marinus Analytics Fight Human Trafficking

Marinus Analytics provides law enforcement with tools, founded in artificial intelligence, to turn big data into actionable intelligence. The Marinus flagship software, Traffic Jam, is a suite of tools for use by law enforcement agencies on sex trafficking investigations.

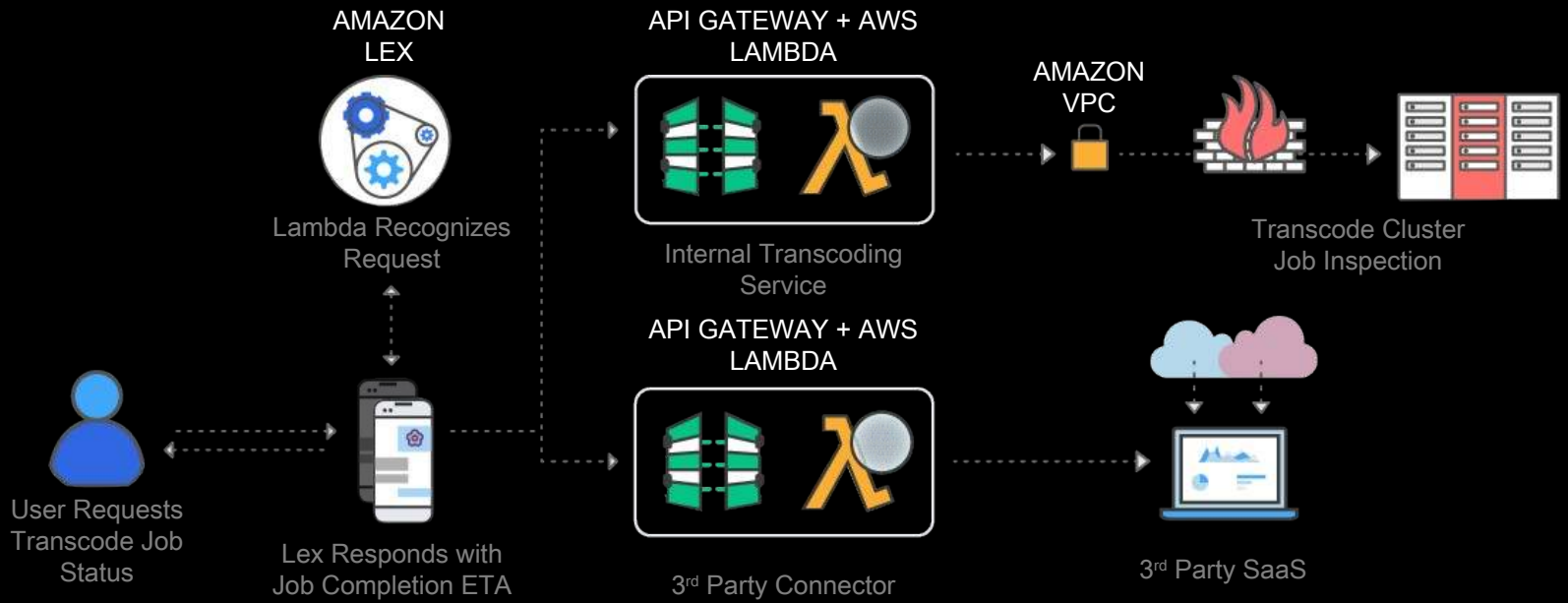


http://www.marinusanalytics.com/articles/2017/10/17/amazon-rekognition-helps-marinus-analytics-fight-human-trafficking

Metadata Enrichment using Amazon Rekognition



Service Enhancement using Amazon Lex + Polly



Key Takeaways

- Building your own Deep Learning infrastructure is hard and costly
- Managed services can be used to eliminate ‘undifferentiated heavy lifting’, allowing for niche AI focus
- AI for media is a cross-functional tech undertaking
- Many traditional ‘in the cloud’ paradigms map to deep learning
- AI technology provides opportunities to enhance existing media services
- Utilize compute diversification across GPU, CPU & FPGA, combined with Object Storage (S3) & Fractional Billing (SPOT)

aws

Thank You!

Julien Simon, AI Evangelist, EMEA
@julsimon
http://medium.com/@julsimon