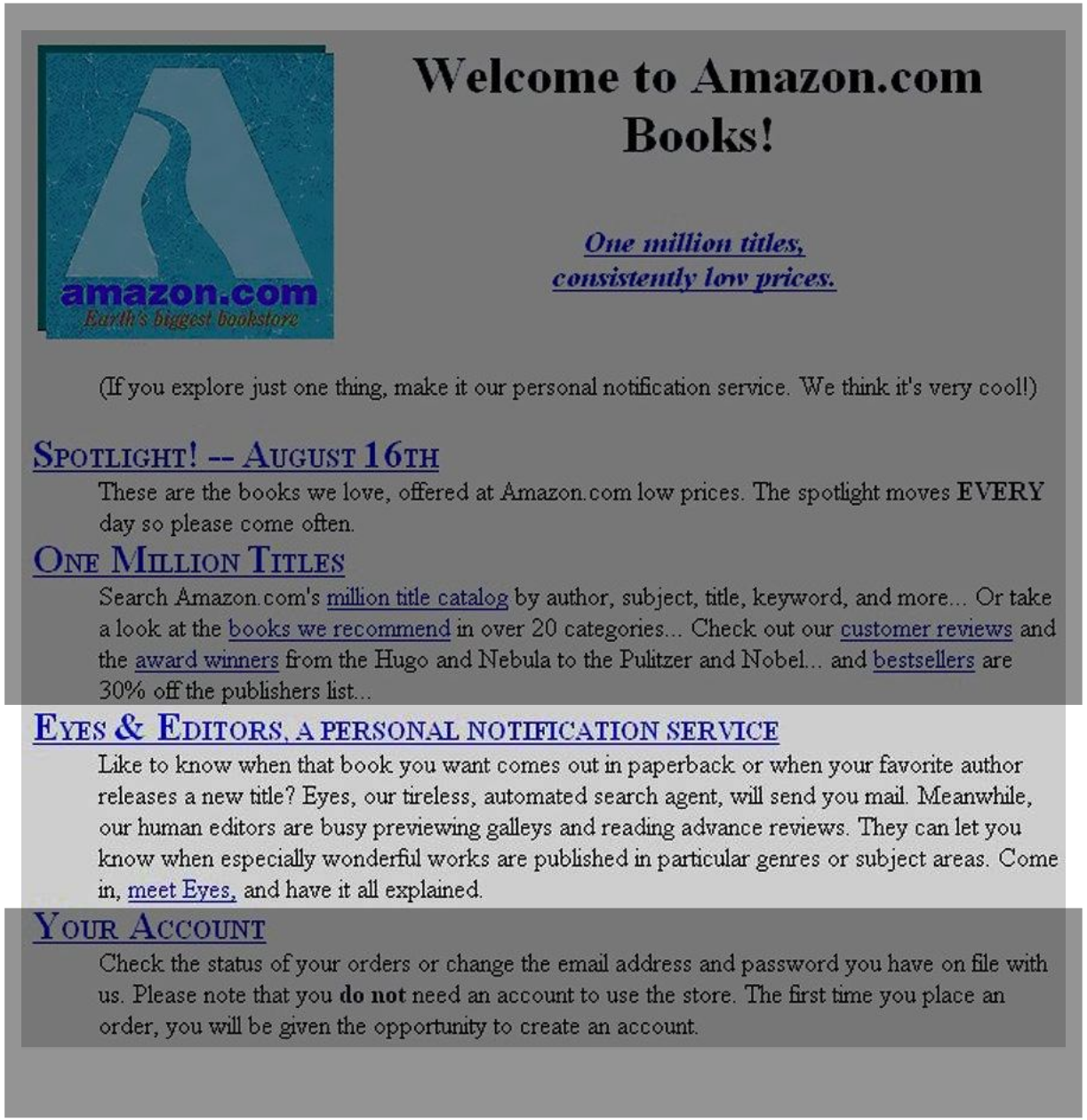
AI and Machine Learning on AWS

Julien Simon, Principal Evangelist, AI/ML
@julsimon



Amazon.com,1995

« *Two Decades of Recommender Systems at Amazon.com* » (2017)

The Test of Time

Two Decades of Recommender Systems at Amazon.com

Amazon is well-known for personalization and recommendations, which help customers discover items they might otherwise not have found. In this update to our original article, we discuss some of the changes as Amazon has grown.

Brent Smith
Amazon.com

Greg Linden
Microsoft

For two decades now,[1] Amazon.com has been building a store for every customer. Each person who comes to Amazon.com sees it differently, because it's individually personalized based on their interests. It's as if you walked into a store and the shelves started rearranging themselves, with what you might want moving to the front, and what you're unlikely to be interested in shuffling further away.

From a catalog of hundreds of millions of items, Amazon.com's recommendations pick a small number of items you might enjoy based on your current context and your past behavior. The algorithms aren't magic; they simply share with you what other people have already discovered. The algorithm does all the work. It's computers helping people help other people, implicitly and anonymously.

Amazon.com launched item-based collaborative filtering in 1998, enabling recommendations at a previously unseen scale for millions of customers and a catalog of millions of items. Since we wrote about the algorithm in *IEEE Internet Computing* in 2003,[2] it has seen widespread use across the Web, including YouTube, Netflix, and many others. The algorithm's success has been from its simplicity, scalability, and often surprising and useful recommendations, as well as desirable properties such as updating immediately based on new information about a customer and being able to explain why it recommended something in a way that's easily understandable.

What was described in our 2003 *IEEE Internet Computing* article has faced many challenges and seen much development over the years. Here, we describe some of the updates, improvements, and adaptations for item-based collaborative filtering, and offer our view on what the future holds for collaborative filtering, recommender systems, and personalization.

The Algorithm

As we described it in 2003, the item-based collaborative filtering algorithm is straightforward. In the mid-1990s, collaborative filtering was generally user-based, meaning the first step of the algorithm was to search across other users to find people with similar interests (such as similar purchase patterns), then look at what items those similar users found that you haven't found yet. Instead, our algorithm begins by finding related items for each item in the catalog. The term "related" could have several meanings here, but at this point,

Published by the IEEE Computer Society 1089-7801/17/$33.00 © 2017 IEEE IEEE INTERNET COMPUTING

Standing the Test of Time

As part of recognizing *IEEE Internet Computing* for its 20 years in publication, I recommended to the editorial board that we pick one of our magazine articles that, over the past 20 years, has withstood the "test of time." In selecting an article, we evaluated the ideas in more than 20 candidate articles that reported on "evergreen" research areas over the past two decades and then assessed these articles based on downloads from IEEE Xplore, citations, and mentions of the work in popular press. This information was presented to a committee consisting of previous Editors in Chief for the magazine. I would like to thank the selection committee from the editorial board — led by Arun Iyengar, and including Fred Douglis, Robert Filman, Michael Huhns, Charles Petrie, Michael Rabinovich, and Munindar Singh. This committee deliberated on the top three articles by evaluating each work's previous importance within the context of its sustained importance in the future.

It's my pleasure to recognize the committee's official "Test of Time" winner: an industry article titled "Amazon.com Recommendations: Item-to-Item Collaborative Filtering" by Greg Linden, Brent Smith, and Jeremy York, from the January/February 2003 issue of *IC* (see doi:10.1109/MIC.2003.1167344). Fourteen years after the publication of this article, it shows 125 downloads from IEEE Xplore in one month, with more than 12,754 downloads since January 2011. The article currently shows 4,258 citations in Google Scholar. I'm delighted that the selection committee recommended an industry article, as it aligns with the magazine's focus of accessibility to academic, research, and industrial populations.

In addition to recognizing the article, we asked the authors to create this retrospective piece discussing research and insights that have transpired since publishing their winning "Test of Time" article, while projecting into the future.

Going forward, the magazine hopes to celebrate a "Test of Time" article every 2–3 years. I hope that you enjoy this retrospective article, and please take a moment to congratulate Greg Linden, Brent Smith, and Jeremy York.

— M. Brian Blake
Editor-in-Chief, *IEEE Internet Computing*
Provost and Distinguished Professor, Drexel University

let's loosely define it as "people who buy one item are unusually likely to buy the other." So, for every item i_0, we want every item i_1 that was purchased with unusually high frequency by people who bought i_0.

Once this related items table is built, we can generate recommendations quickly as a series of lookups. For each item that's part of this customer's current context and previous interests, we look up the related items, combine them to yield the most likely items of interest, filter out items already seen or purchased, and then we are left with the items to recommend.

This algorithm has many advantages over the older user-based collaborative filtering. Most importantly, the majority of the computation is done offline — a batch build of the related items — and the computation of the recommendations can be done in real time as a series of lookups. The recommendations are high quality and useful, especially given enough data, and remain competitive in perceived quality even with the newer algorithms created over the last two decades. The algorithm scales to hundreds of millions of users and tens of millions of items without sampling or other techniques that can reduce the quality of the recommendations. The algorithm updates immediately on new information about a person's interests. Finally, the recommendations can be explained in an intuitive way as arising from a list of items the customer remembers purchasing.

In 2003: Amazon.com, Netflix, YouTube, and More

By the time we published in IEEE in 2003, item-based collaborative filtering was widely deployed across Amazon.com. The homepage prominently featured recommendations based on your past purchases and items browsed in the store. Search result pages recommended items related to your search. The shopping cart recommended other items to add to your cart, perhaps impulse buys to bundle in at the last minute, or perhaps complements to what you were already considering. At the end of your order, more recommendations appeared, suggesting items to order later. Using e-mails, browse pages, product detail pages, and more, many pages on Amazon.com had at least some recommended content, starting to approach a store for every customer.

Others have reported using the algorithm, too. In 2010, YouTube reported using it for recommending videos.[3] Many open source and third-party vendors included the algorithm, and it showed up widely in online retail, travel, news, advertising, and more. In the years following, the recommendations were used so extensively by Amazon.com that a Microsoft Research report estimated 30 percent of Amazon.com's page

G.D. Linden, J.A. Jacobi, and E.A. Benson, Collaborative Recommendations Using Item-to-Item Similarity Mappings,

US Patent 6,266,649, Amazon.com, Patent and Trademark Office, 2001 (**filed 1998**).

https://www.computer.org/csdl/mags/ic/2017/03/mic2017030012.html


amazon
fulfillment

amazon echo


fire TV
TED
amazon
amazon


JUST
WALK
OUT
SHOPPING

Jeff Bezos' letter to Amazon shareholders

"We are solving problems with machine learning and artificial intelligence that were in the realm of science fiction for the last several decades. Natural language understanding, machine vision problems, it really is an amazing renaissance."

https://www.geekwire.com/2017/jeff-bezos-explains-amazons-artificial-intelligence-machine-learning-strategy/

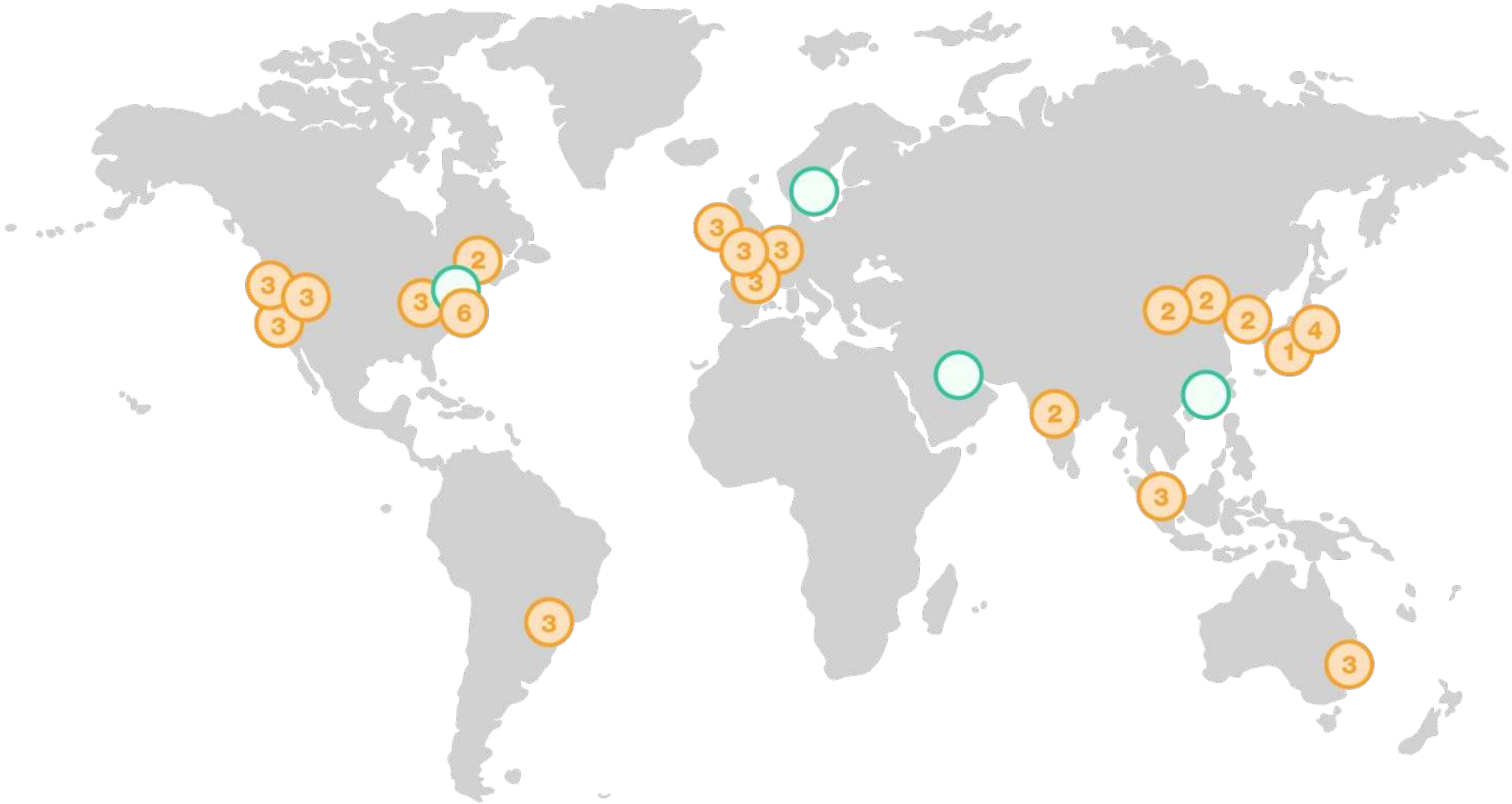
AWS: our mission

Put AI and Machine Learning
in the hands of every developer
and data scientist

Machine Learning requires
a solid infrastructure foundation

AWS Global Infrastructure

18 Regions – **54** Availability Zones – **103** Edge Locations



Gartner: Cloud Infrastructure as a Service, June 2017

Figure 1. Magic Quadrant for Cloud Infrastructure as a Service, Worldwide

Source: Gartner (June 2017)

Machine Learning requires
secure infrastructure

AWS compliance standards

Certifications & Attestations

- Cloud Computing Compliance Controls Catalogue (C5) — DE
- Cyber Essentials Plus — UK
- DoD SRG — US
- FedRAMP — US
- FIPS — US
- IRAP — AU
- ISO 9001
- ISO 27001
- ISO 27017
- ISO 27018
- MLPS Level 3 — CN
- MTCS — SG
- PCI DSS Level 1
- SEC Rule 17-a-4(f) — US

Laws, Regulations and Privacy

- CISPE — EU
- EU Model Clauses — EU
- FERPA — US
- GLBA — US
- HIPAA — US
- HITECH
- IRS 1075 — US
- ITAR — US
- My Number Act — JP
- Data Protection Act – 1988 — UK
- VPAT / Section 508 — US
- Data Protection Directive — EU
- Privacy Act [Australia] — AU
- Privacy Act [New Zealand] — NZ

Alignments & Frameworks

- CIS (Center for Internet Security)
- CJIS (US FBI) — US
- CSA (Cloud Security Alliance)
- Esquema Nacional de Seguridad — ES
- EU-US Privacy Shield — EU
- FISC — JP
- FISMA — US
- G-Cloud — UK
- GxP (US FDA CFR 21 Part 11) — US
- ICREA
- IT Grundschutz — DE
- MITA 3.0 (US Medicaid) — US
- MPAA — US
- NIST — US

AWS is GDPR ready

AWS Security Blog

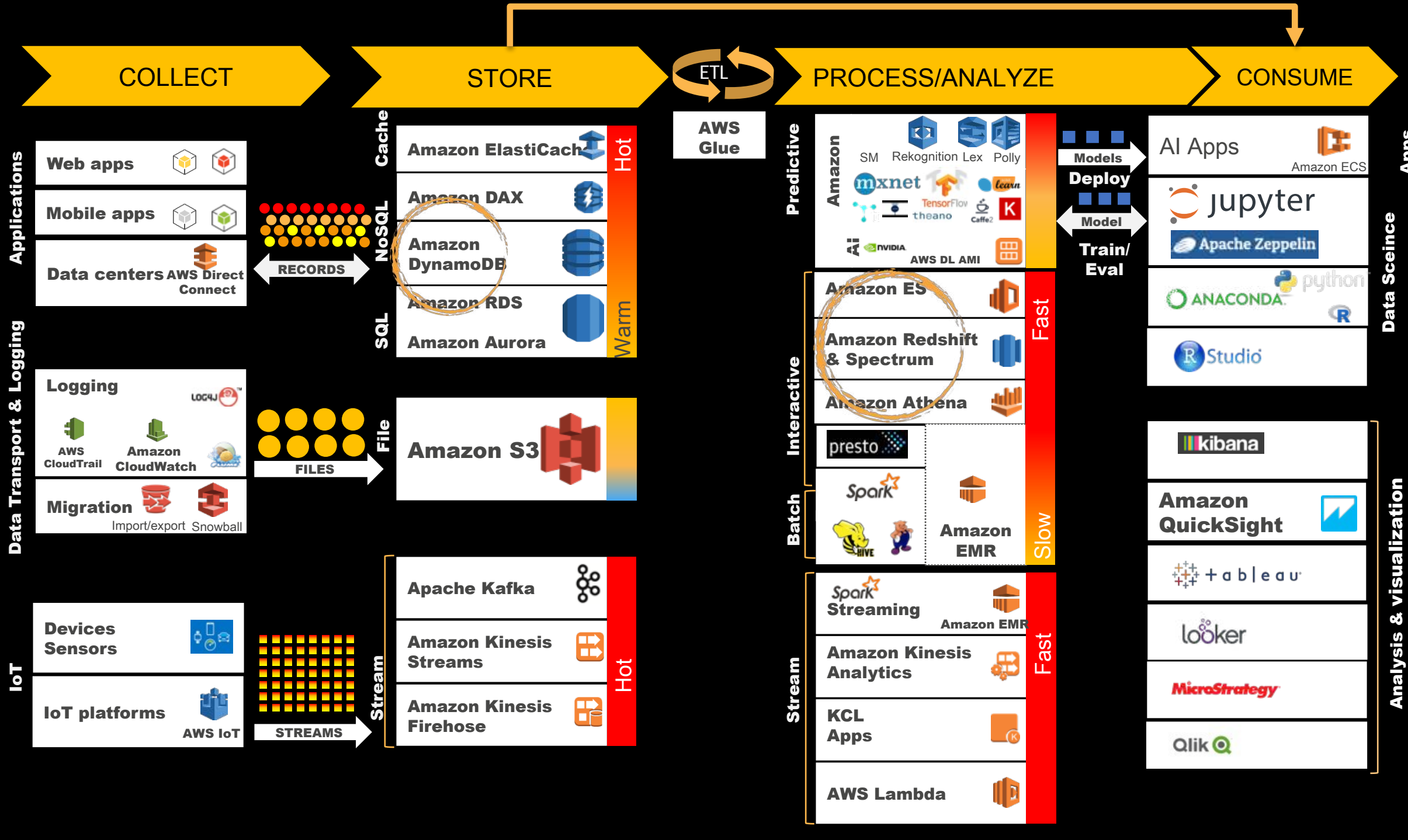
All AWS Services GDPR ready

by Chad Woolf | on 26 MAR 2018 | in Amazon GuardDuty, Amazon Inspector*, Amazon Macie*, AWS Config*, Compliance*, Security, Identity, & Compliance*, Webinars* | Permalink | Comments | Share

Today, I'm very pleased to announce that AWS services comply with the General Data Protection Regulation (GDPR). This means that, in addition to benefiting from all of the measures that AWS already takes to maintain services security, customers can deploy AWS services as a key part of their GDPR compliance plans.

This announcement confirms we have completed the entirety of our GDPR service readiness audit, validating that all generally available services and features adhere to the high privacy bar and data protection standards required of data processors by the GDPR. We completed this work two months ahead of the May 25, 2018 enforcement deadline in order to give customers and APN partners an environment in which they can confidently build their own GDPR-compliant products, services, and solutions.

Machine Learning requires extensive (big) data services


COLLECT
STORE
ETL
PROCESS/ANALYZE
CONSUME
AWS Glue
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
RECORDS
Data Transport & Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Migration
Import/export
Snowball
FILES
IoT
Devices Sensors
IoT platforms
AWS IoT
STREAMS
Cache
Amazon ElastiCache
NoSQL
Amazon DAX
Amazon DynamoDB
SQL
Amazon RDS
Amazon Aurora
Hot
Warm
File
Amazon S3
Stream
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Hot
Predictive
Amazon
SM
Rekognition
Lex
Polly
mxnet
TensorFlow
theano
Caffe2
AWS DL AMI
Models
Deploy
Model
Train/ Eval
Interactive
Amazon ES
Amazon Redshift & Spectrum
Amazon Athena
presto
Batch
Spark
Amazon EMR
Fast
Slow
Stream
Spark Streaming
Amazon EMR
Amazon Kinesis Analytics
KCL Apps
AWS Lambda
Fast
AI Apps
Amazon ECS
jupyter
Apache Zeppelin
ANACONDA
python
R
RStudio
Apps
Data Sceince
kibana
Amazon QuickSight
tableau
looker
MicroStrategy
Qlik
Analysis & visualization

Dynamo (2007)

Dynamo: Amazon's Highly Available Key-value Store

Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels

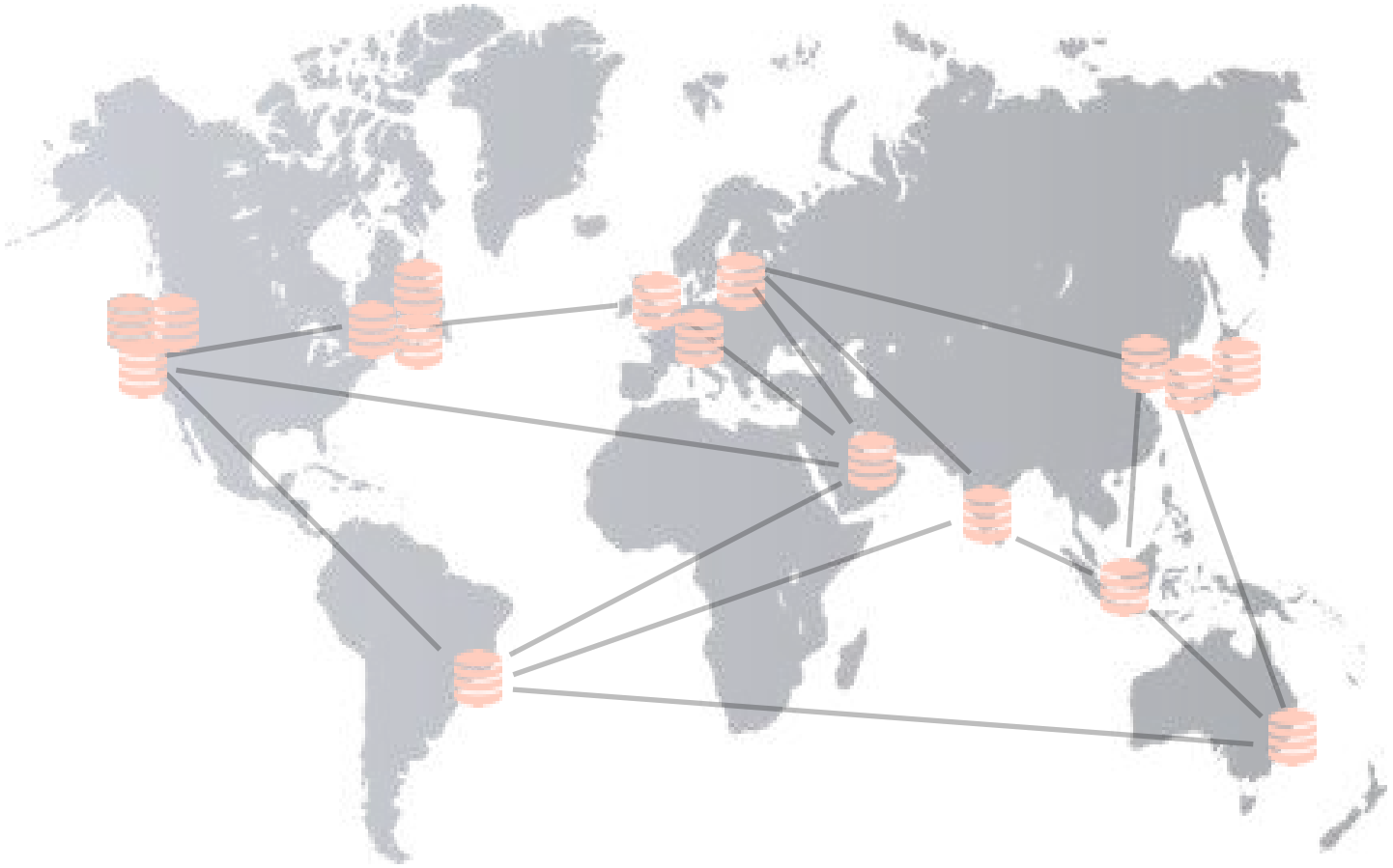
Amazon.com

Launched in 2012, Amazon DynamoDB now powers Alexa, the Amazon.com sites and the Amazon fulfillment centers.

During Prime Day 2017, it handled 3.34 trillion requests, peaking at 12.9 million requests per second.

DynamoDB Global Tables

First fully managed, multi-master, multi-region database



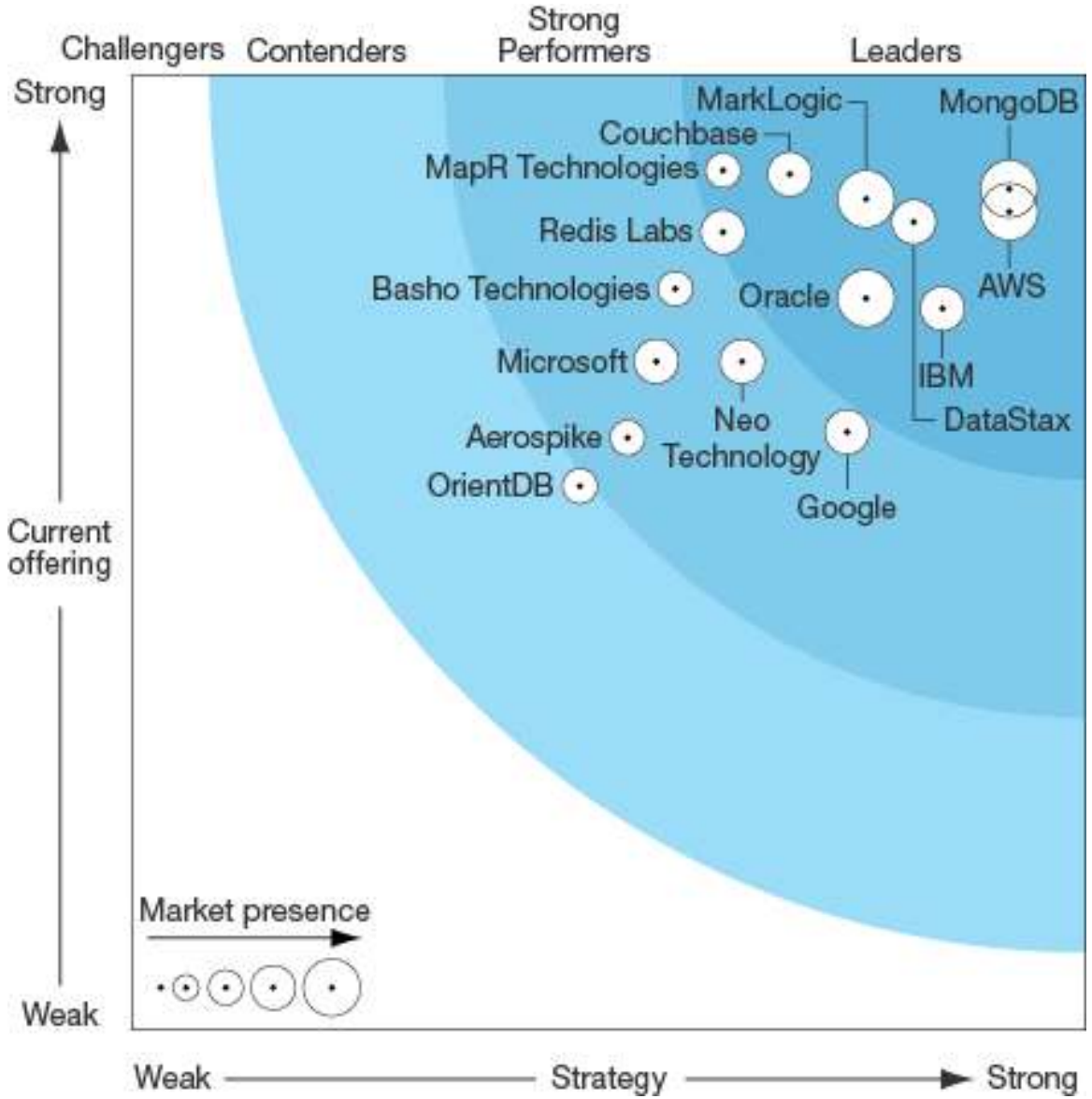
Build high performance, globally distributed applications

Low latency reads and writes to locally available tables

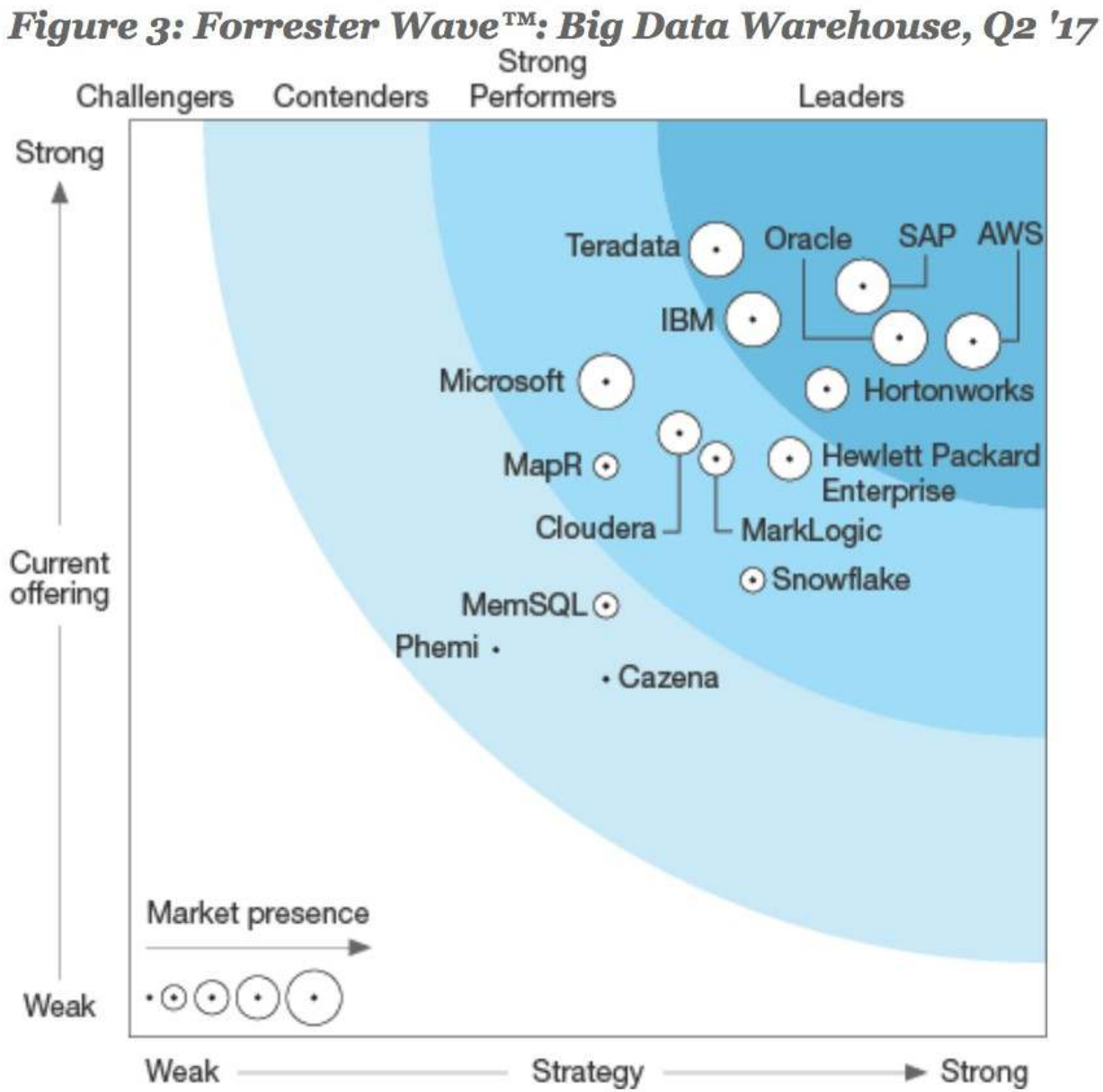
Disaster proof with multi-region redundancy

Easy to setup and no application re-writes required

The Forrester Wave: Big Data NoSQL, Q3 2016

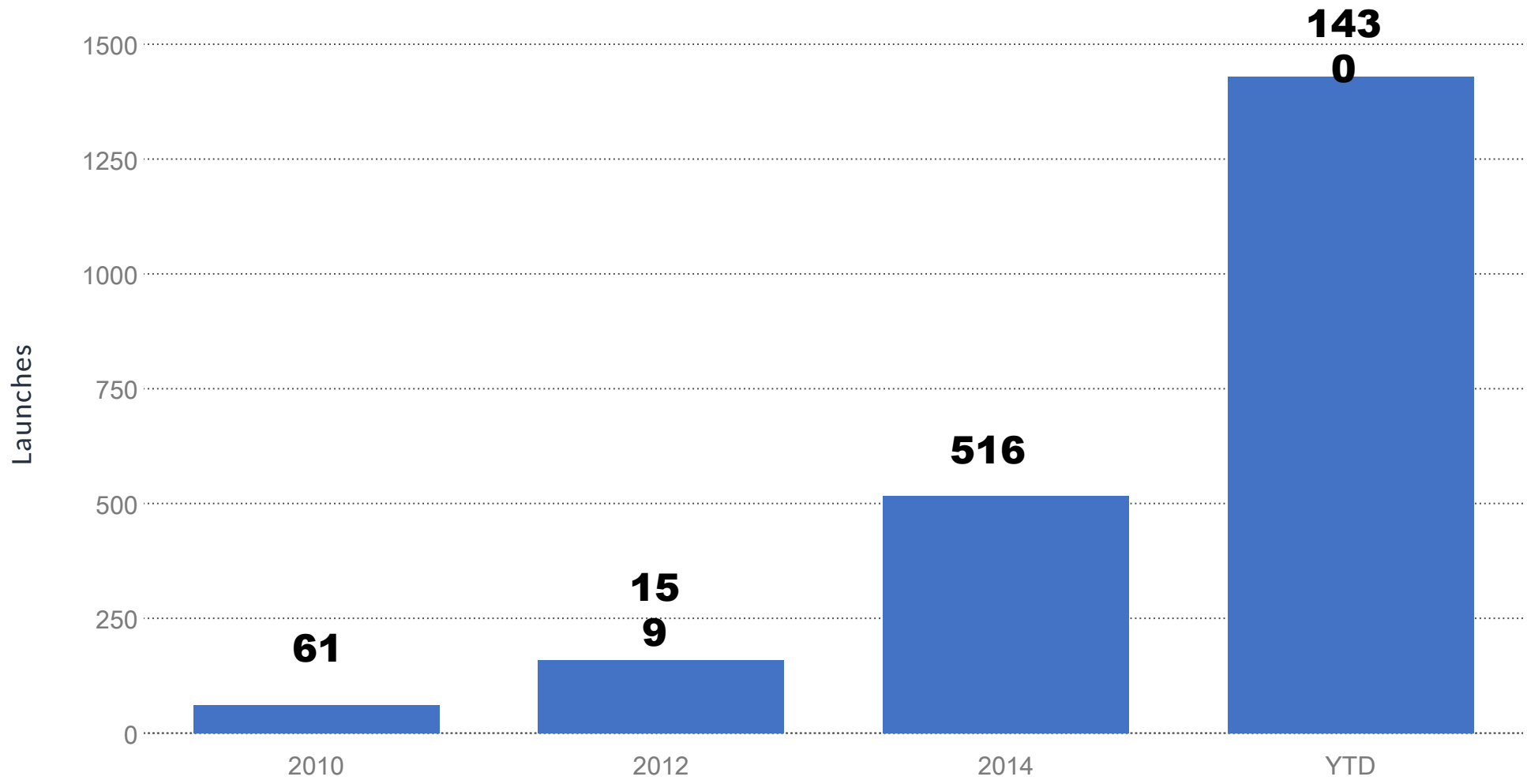


The Forrester Wave: Big Data Warehouse, Q2 2017



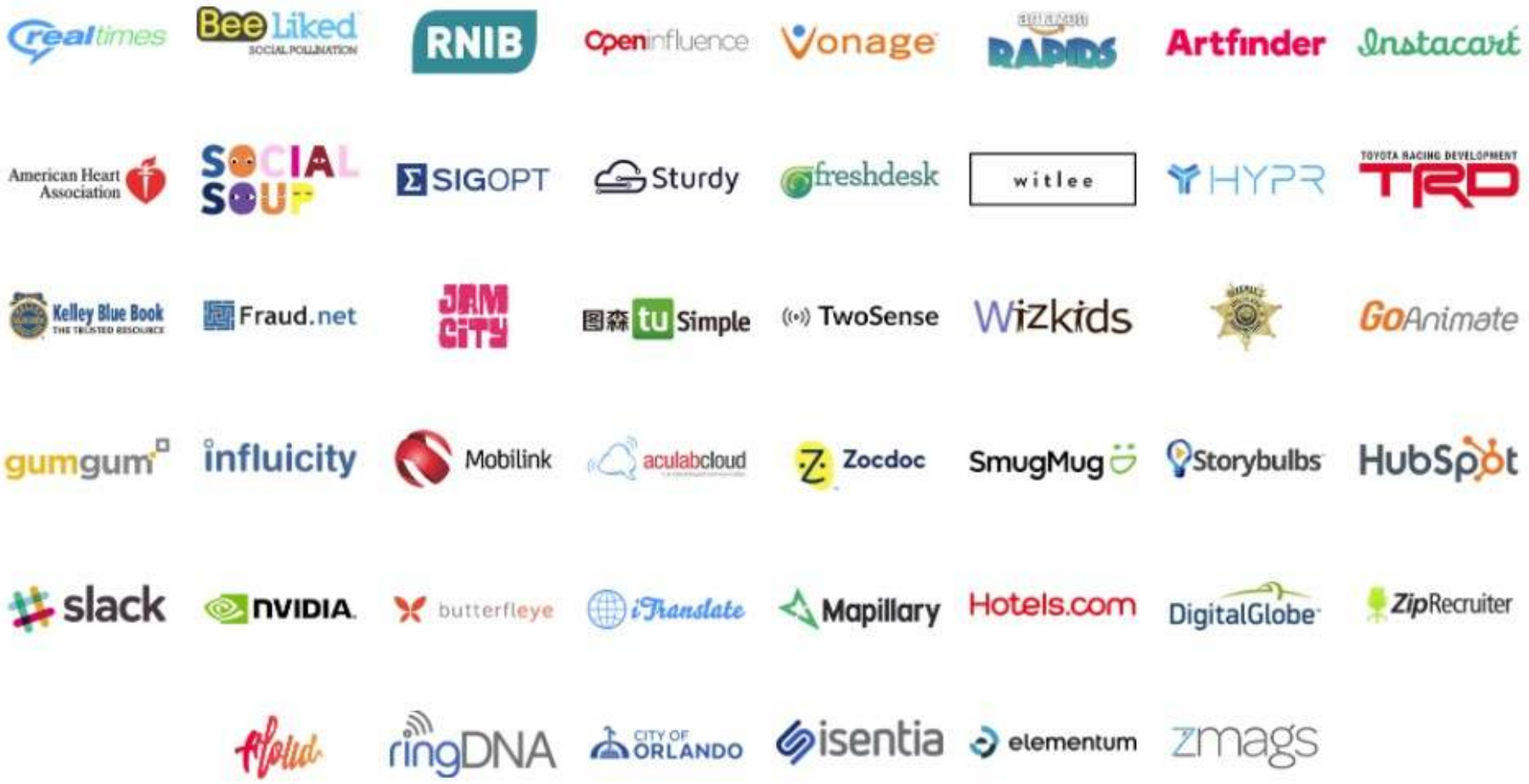
Machine Learning requires
constant innovation

AWS released 1430 features in 2017



More AI/ML is built on AWS than anywhere else

More AI/ML is built on AWS than anywhere else



3 times more customer references
than anyone else

6 times more Enterprise references
than anyone else

AWS AI/ML Stack

Application Services	API-driven services: Vision & Language Services, Conversational Chatbots
Platform Services	Deploy machine learning models with high-performance machine learning algorithms, broad framework support, and one-click training, tuning, and inference.
Frameworks & Infrastructure	Develop sophisticated models with any framework, create managed, auto-scaling clusters of GPUs for large scale training, or run inference on trained models.

Application Services

Vision Services

Amazon Rekognition Image
Deep learning-based image analysis

Learn more »

Amazon Rekognition Video
Deep learning-based video analysis

Learn more »

Conversational chatbots

Amazon Lex
Build chatbots to engage customers

Learn more »

Language Services

Amazon Comprehend
Discover insights and relationships in text

Learn more »

Amazon Translate
Fluent translation of text

Learn more »

Amazon Transcribe
Automatic speech recognition

Learn more »

Amazon Polly
Natural sounding text to speech

Learn more »

Amazon SageMaker

Pre-built notebook instances

Build

Scalable built-in algorithms (11)

scikit learn

mxnet

GLUON

Fully-managed training for ML, DL, and custom algorithms

TensorFlow

PYTORCH

Train

Easier training with hyperparameter optimization

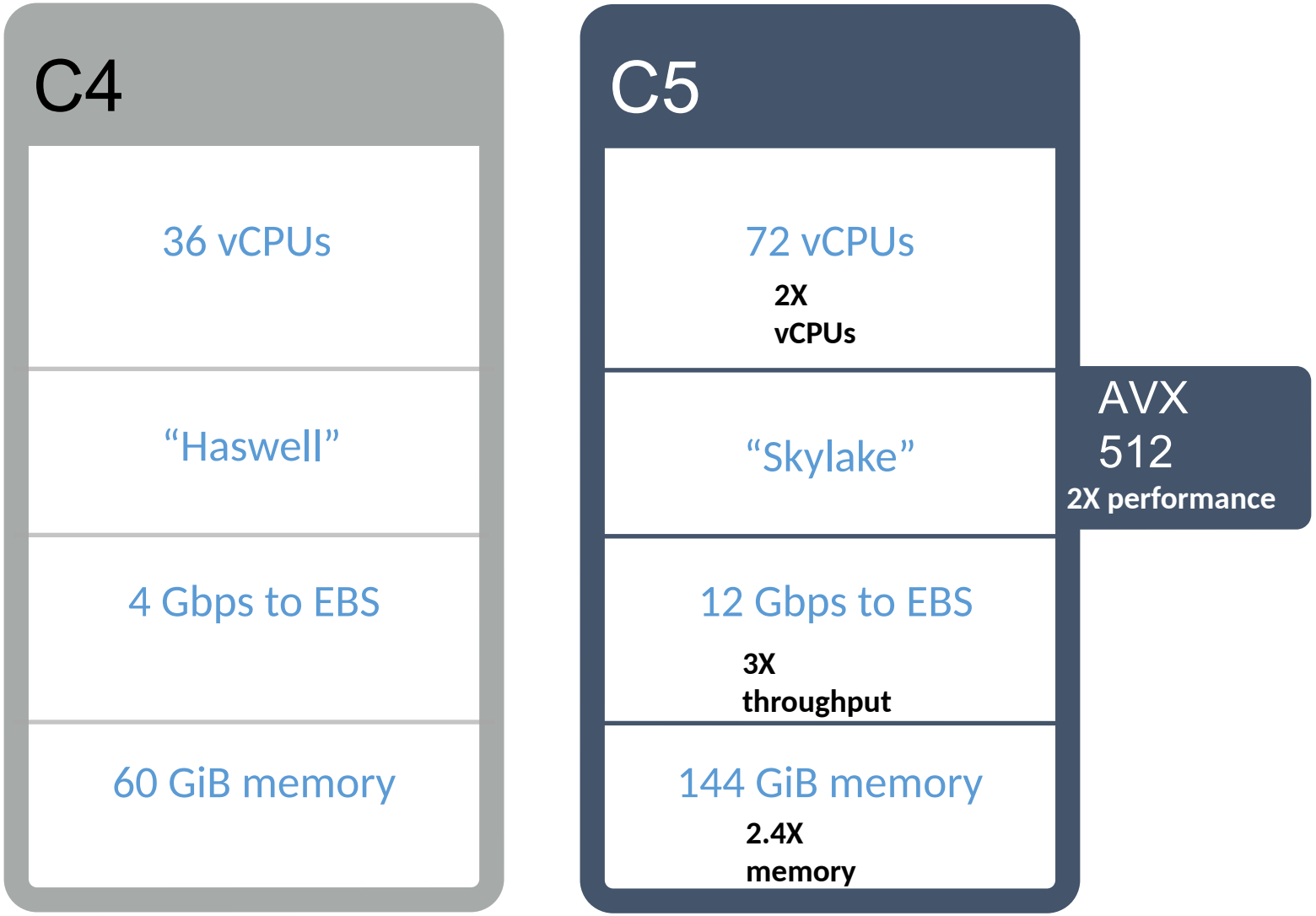
Deploy

Deployment without engineering effort

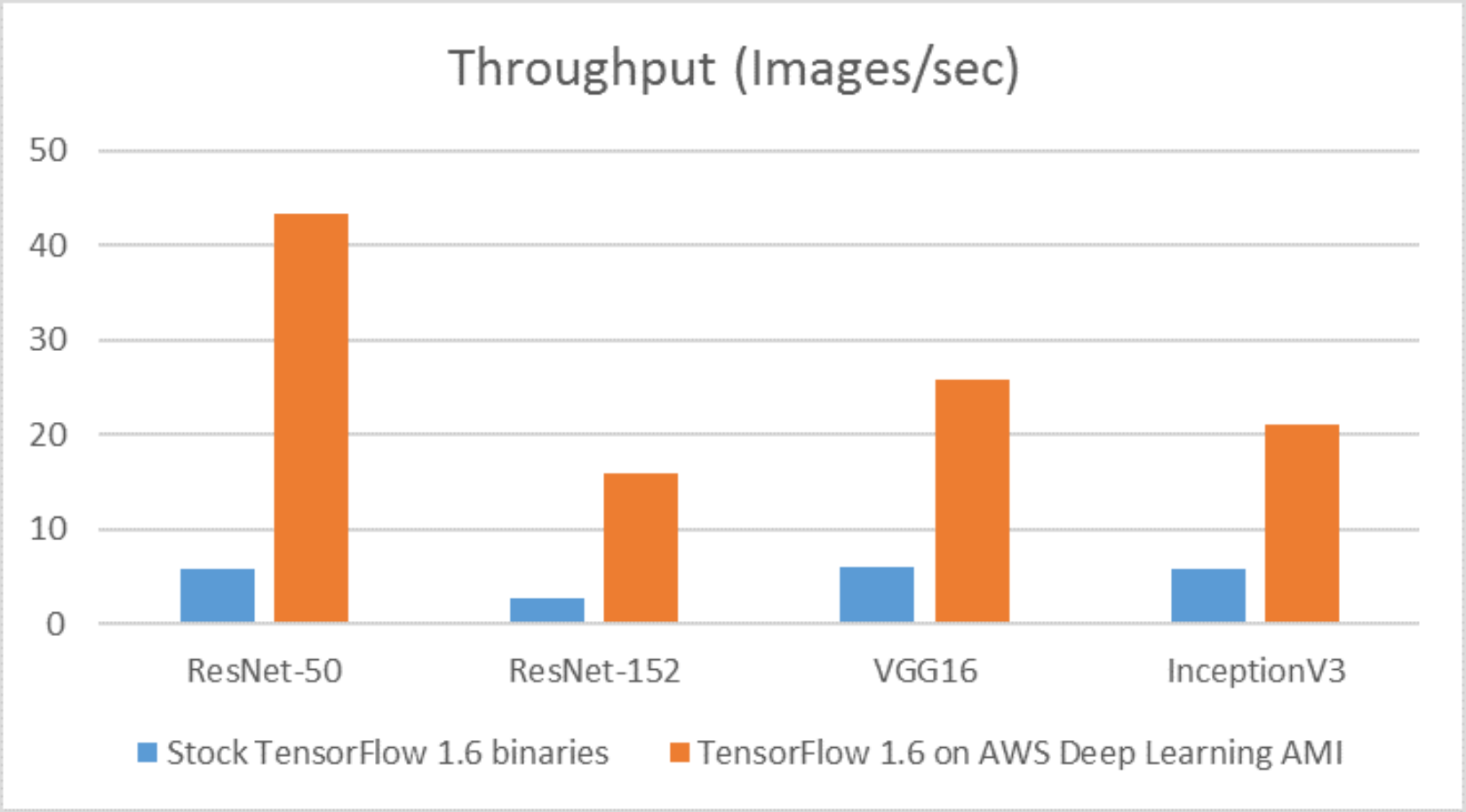
Fully-managed hosting at scale

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Infrastructure – Amazon EC2 C5



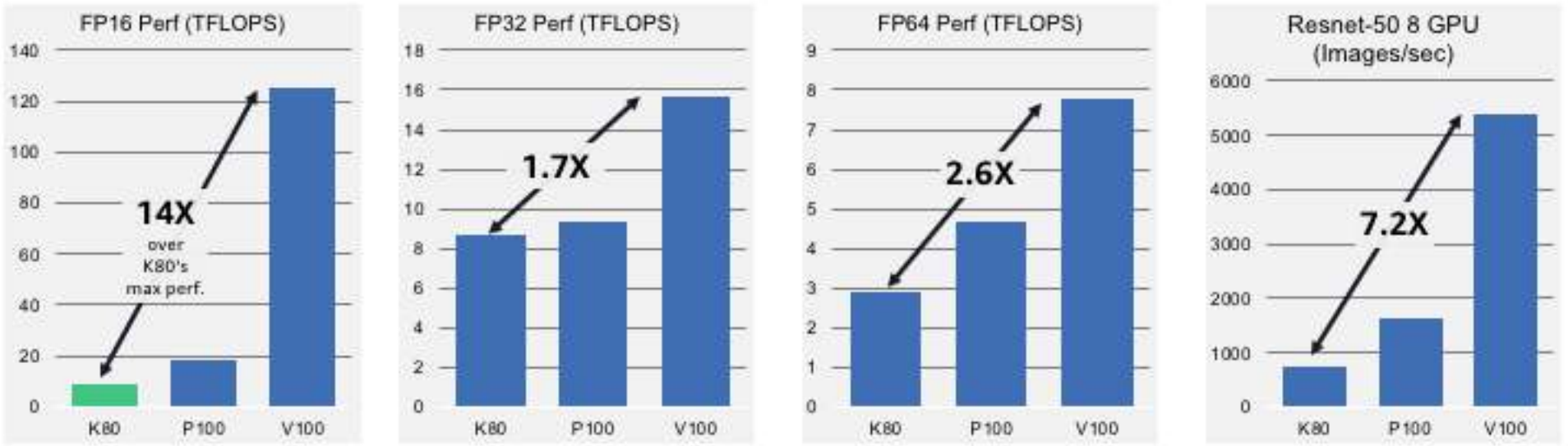
Train TensorFlow 7x faster on AWS



https://aws.amazon.com/blogs/machine-learning/faster-training-with-optimized-tensorflow-1-6-on-amazon-ec2-c5-and-p3-instances/

Infrastructure - GPU instances

- NVIDIA GPU Architectures: Kepler > Maxwell > Pascal > Volta
- P2 Instances use K80 Accelerator (Kepler Architecture)
- P3 Instances use V100 Accelerator (Volta Architecture)



P3 family launched 25/10/2017: up to 8 GPUs in the same server (46,000+ cores)
Now available in 7 regions (US, EU, APAC)

NVIDIA V100 availability

	K80	P100	V100
GCE	Available	Available	-
Azure	Available	Available	Available 6/3, 1 US region, 4 GPUs
Oracle	N/A	Available	Available 27/3, 1 US region
IBM	Available	Available	Available 31/1, bare metal, 2 GPUs
NVIDIA 	N/A	Available	Available

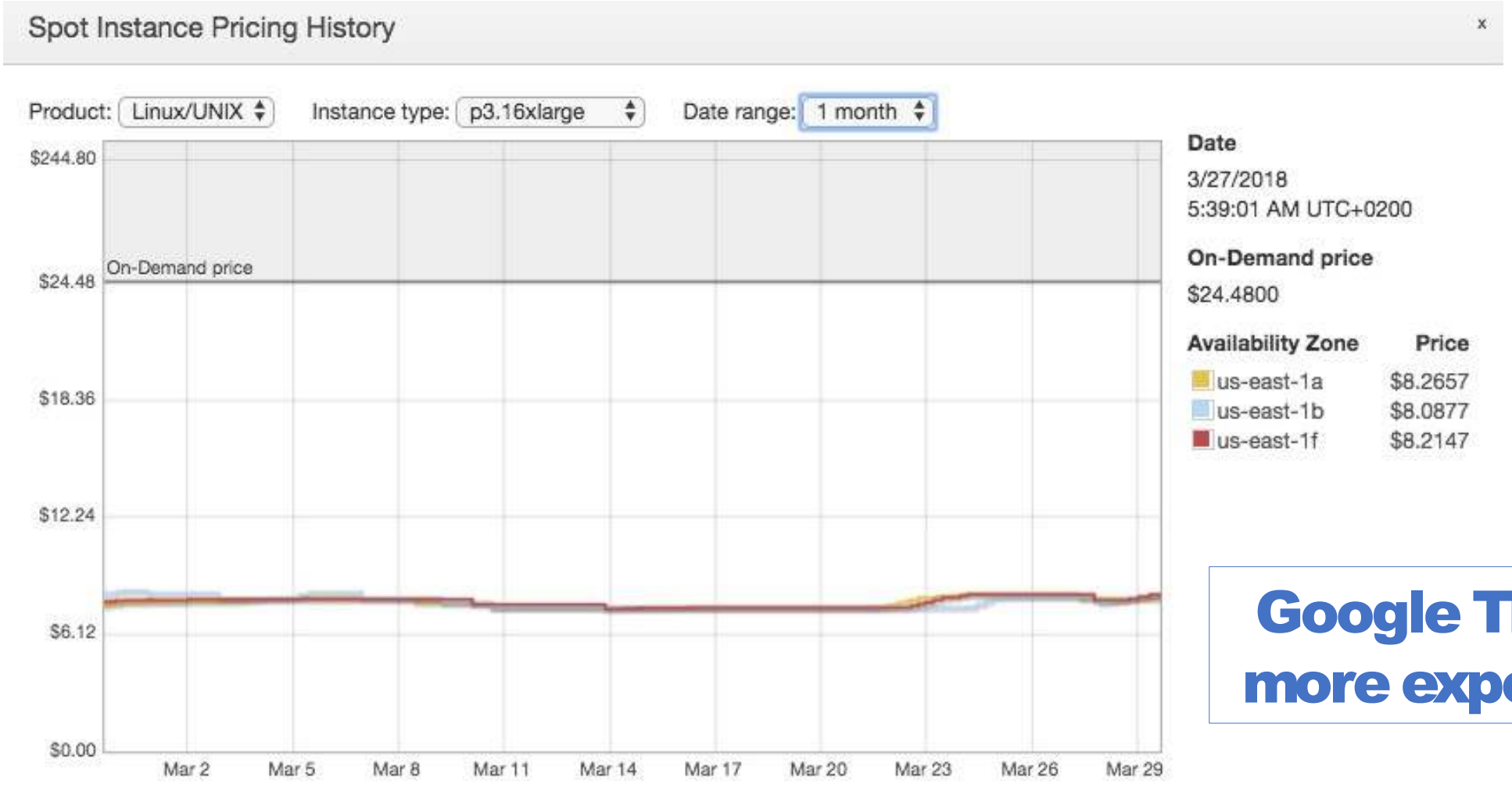
« Google Announces Expensive Cloud TPU Availability »

	Cloud TPU (4x TPU2)	AWS P3.2xlarge (1xV100)	AWS P3.8xlarge (4xV100)	AWS P3.16xlarge (8xV100)
RN50 Image/Second	1369	819	3242	6309
RN50 Time to Train (hr)	23.4	39.1	9.9	5.1
$/Hr	$6.88	$3.06	$12.24	$24.48
$ to Train ImageNet, 90 epochs	$161	$120	$121	$124
Availability	Beta	Now	Now	Now

*« Net it out, and the Google part costs ~**33% more** to do the same work. »*

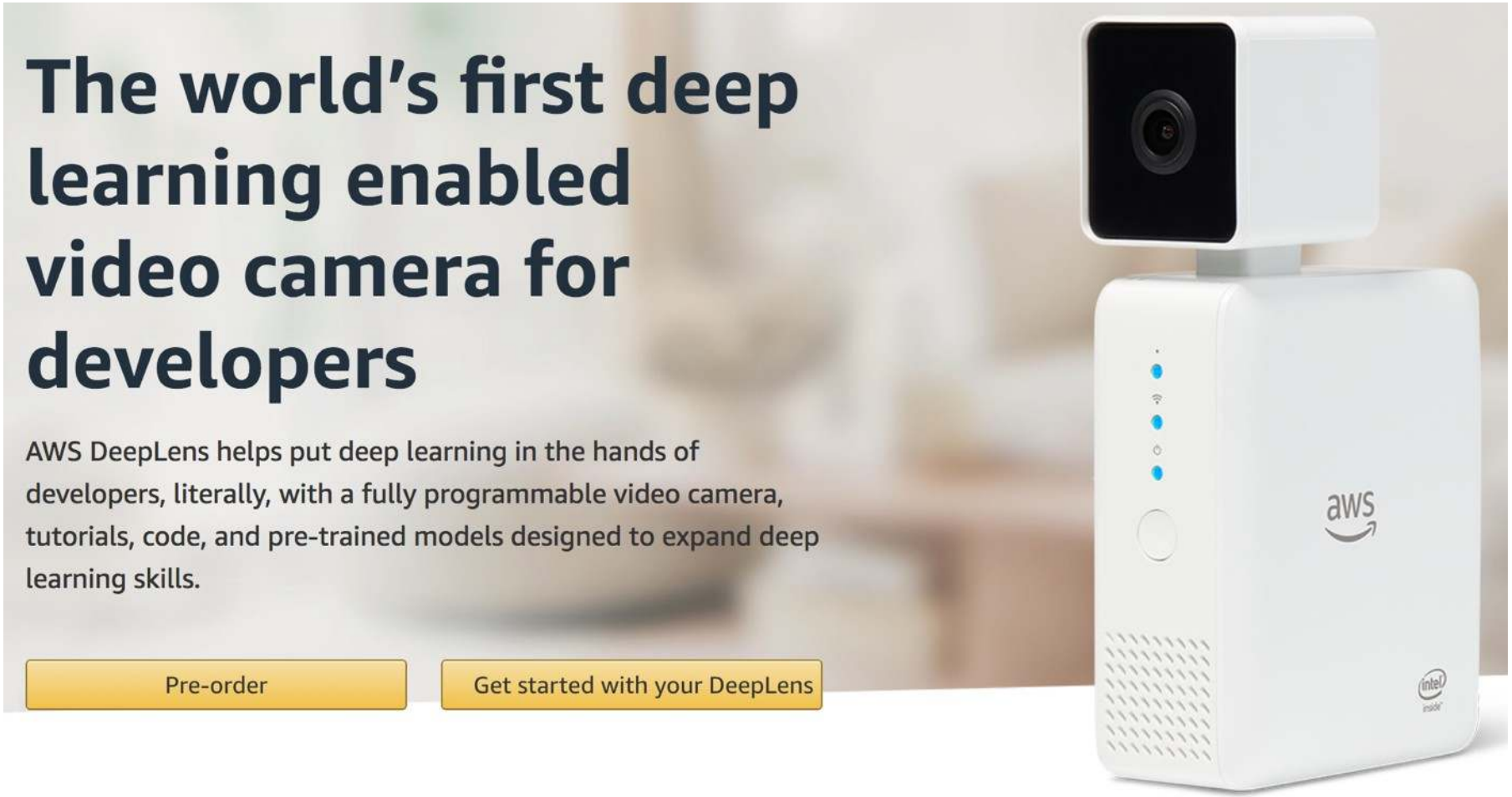
https://www.forbes.com/sites/moorinsights/2018/02/13/google-announces-expensive-cloud-tpu-availability/

EC2 Spot Instances: 67% discount on p3

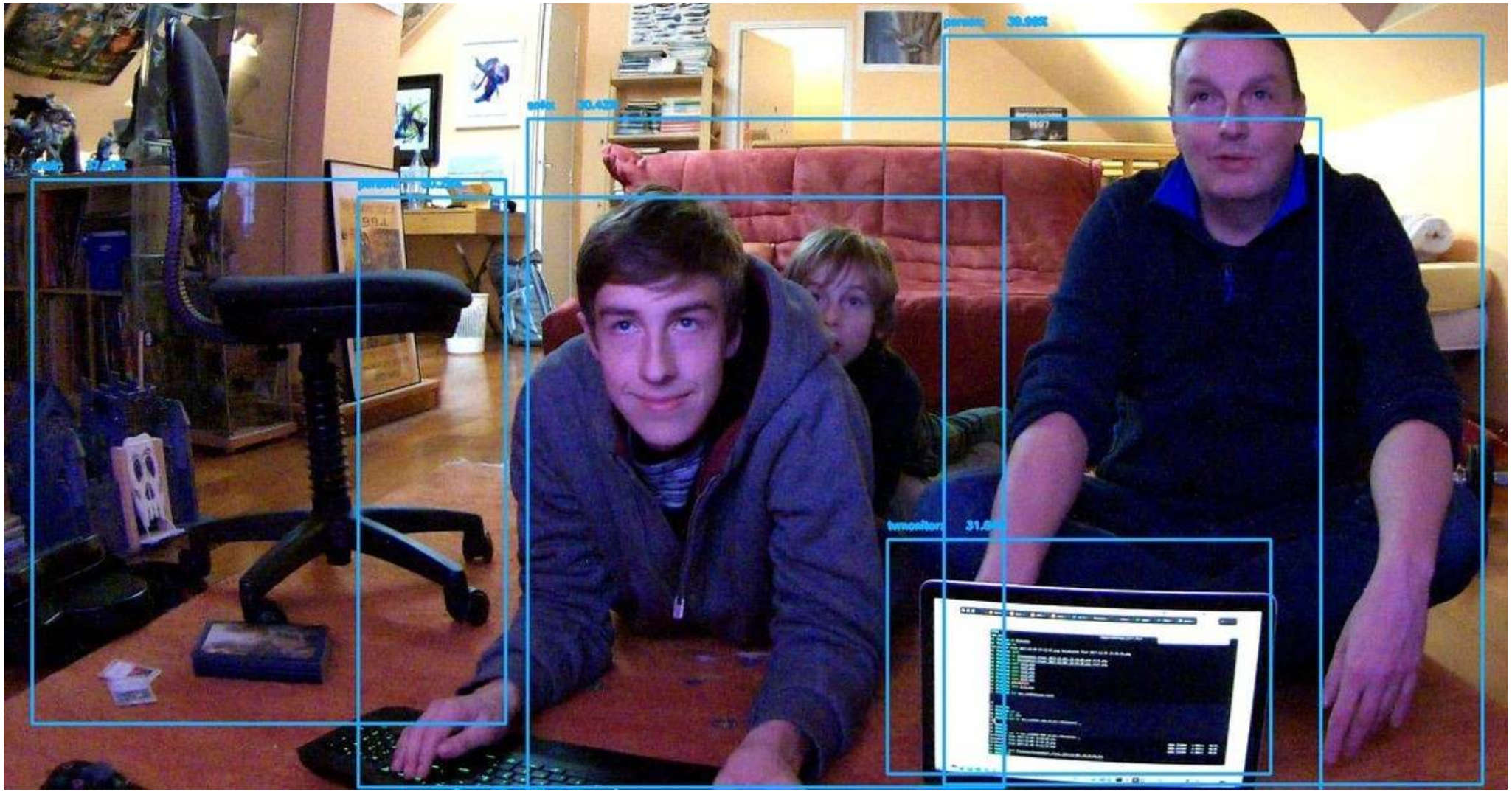


Google TPU 4x more expensive

AWS DeepLens



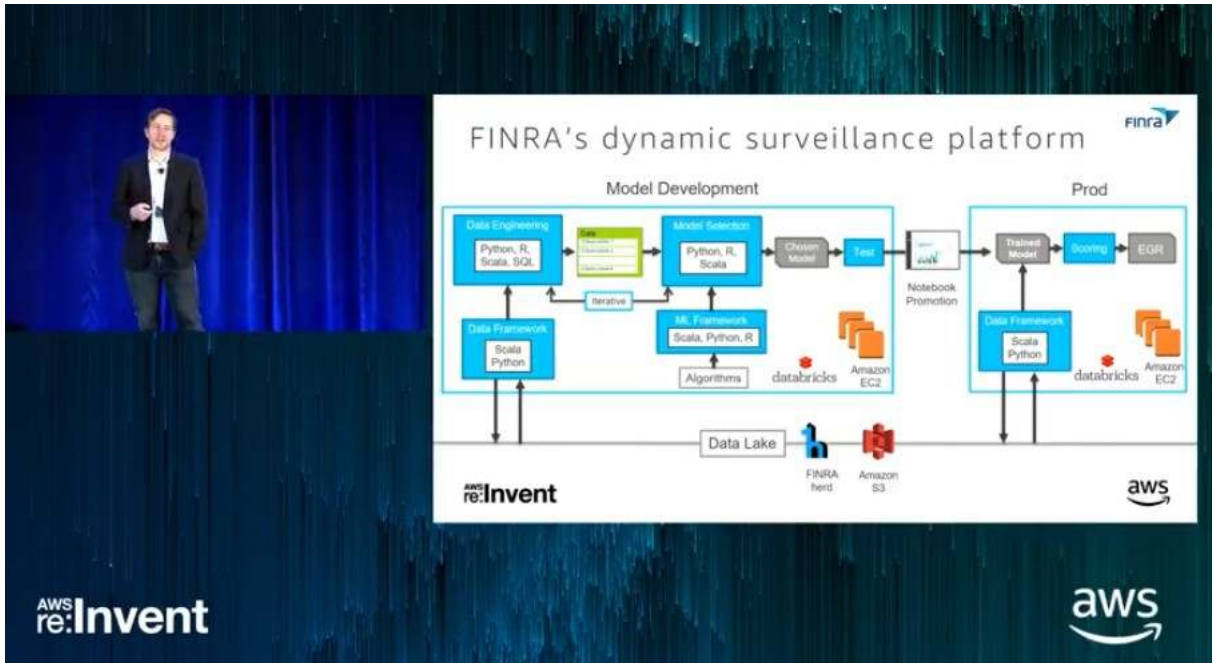
Object detection with AWS DeepLens



Machine Learning at scale

FINRA

- FINRA is the primary regulatory agency for stock brokers in the US.
- They handle about 75 billion market events every day and run hundreds of surveillance algorithms.
- A typical complex query that was taking an hour now takes five to ten seconds.
- Cost savings amount up to $20 million annually.



https://www.youtube.com/watch?v=gdKmMO5PVOY

Autodesk

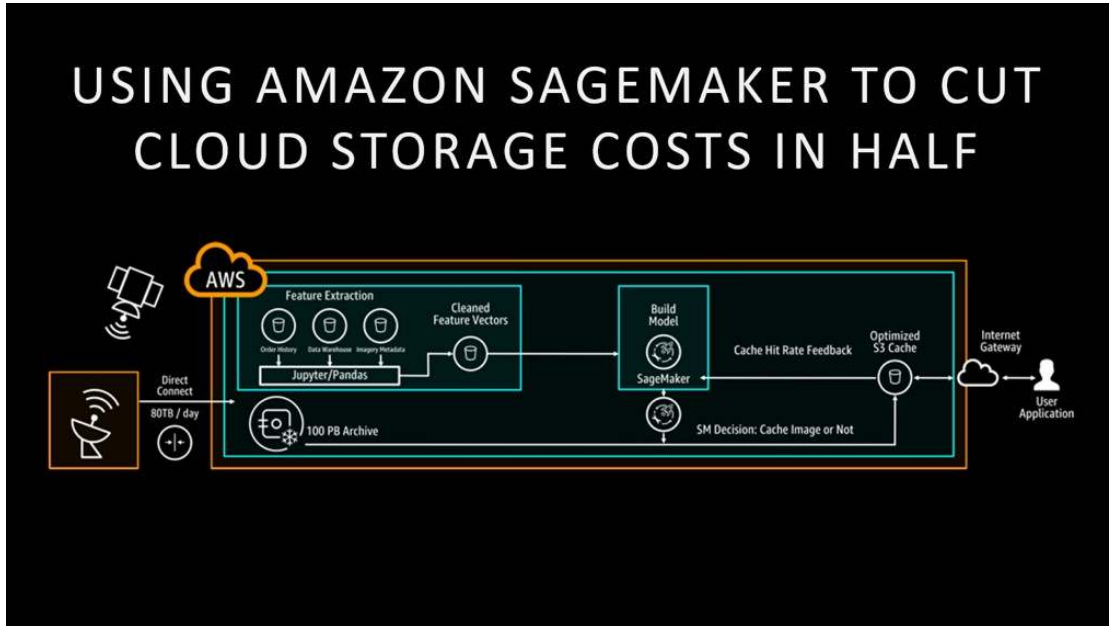
- Autodesk demonstrates the power of generative design to explore all the possible permutations of a solution quickly.
- Autodesk relies on high performance computing on AWS to scale its solutions.

https://www.youtube.com/watch?v=A31A8KDC9S4

Digital Globe

- In the last 18 years DigitalGlobe has been operating Earth imaging satellites, they have collected over 100 PB of imagery.
- They make extensive use of Machine Learning on Amazon SageMaker to extract information.
- Working with the AWS ML Lab, Digital Globe also built a predictive model that will reduce cloud storage costs by 50%.

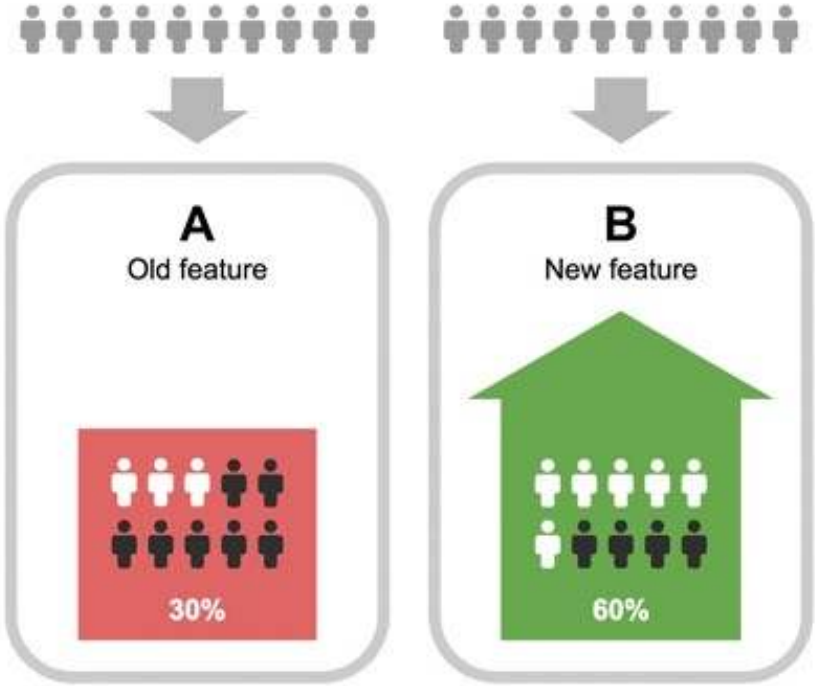
https://aws.amazon.com/solutions/case-studies/digitalglobe-machine-learning/



Speech

Duolingo

- Duolingo is the most popular language-learning platform and the most downloaded education app in the world, with more than 170 million users.
- They have run six A/B tests, testing an Amazon Polly voice against a voice from other TTS providers.
- For all of these experiments, the winning condition was the Amazon Polly voice



https://aws.amazon.com/blogs/machine-learning/powering-language-learning-on-duolingo-with-amazon-polly/

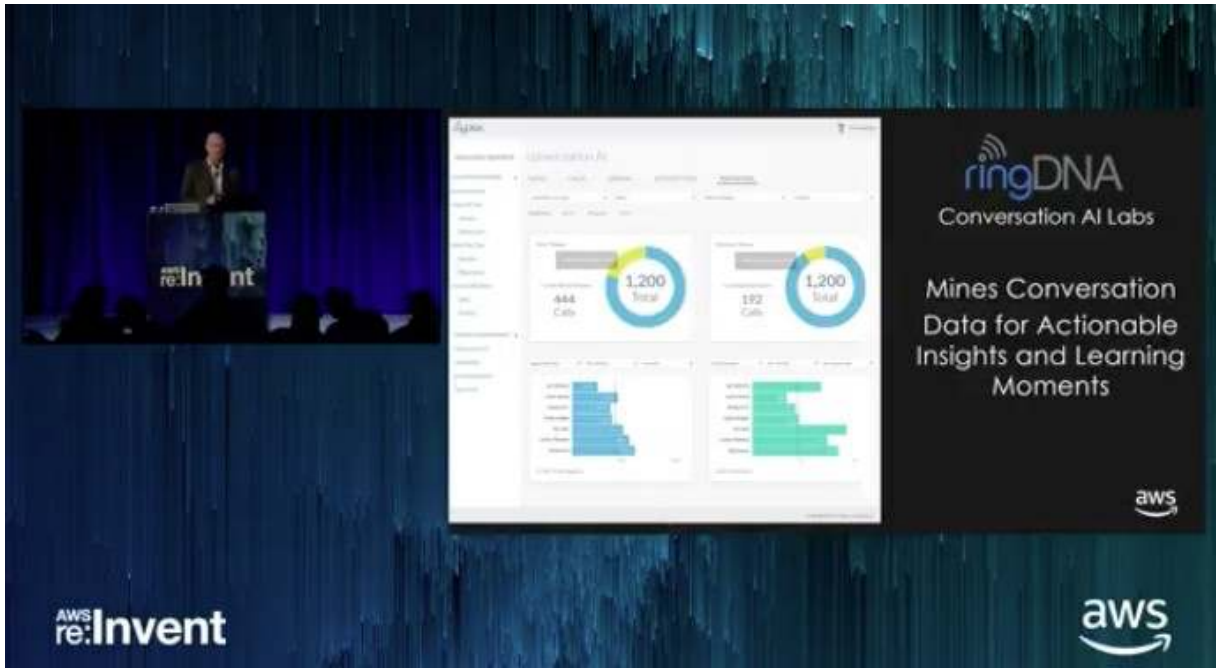
ringDNA

- RingDNA is an end-to-end communications platform for sales teams.
- Hundreds of enterprise organizations use RingDNA to dramatically increase productivity, engage in smarter sales conversations, gain predictive sales insights and improve their win rate.

Speech to Text

"A critical component of RingDNA's Conversation AI requires best of breed speech-to-text to deliver transcriptions of every phone call. RingDNA is excited about **Amazon Transcribe** since it provides high-quality speech recognition at scale, helping us to better transcribe every call to text "

Howard Brown, CEO & Founder, RingDNA



https://www.youtube.com/watch?v=1ZJ_f1bDdog

Natural Language Processing

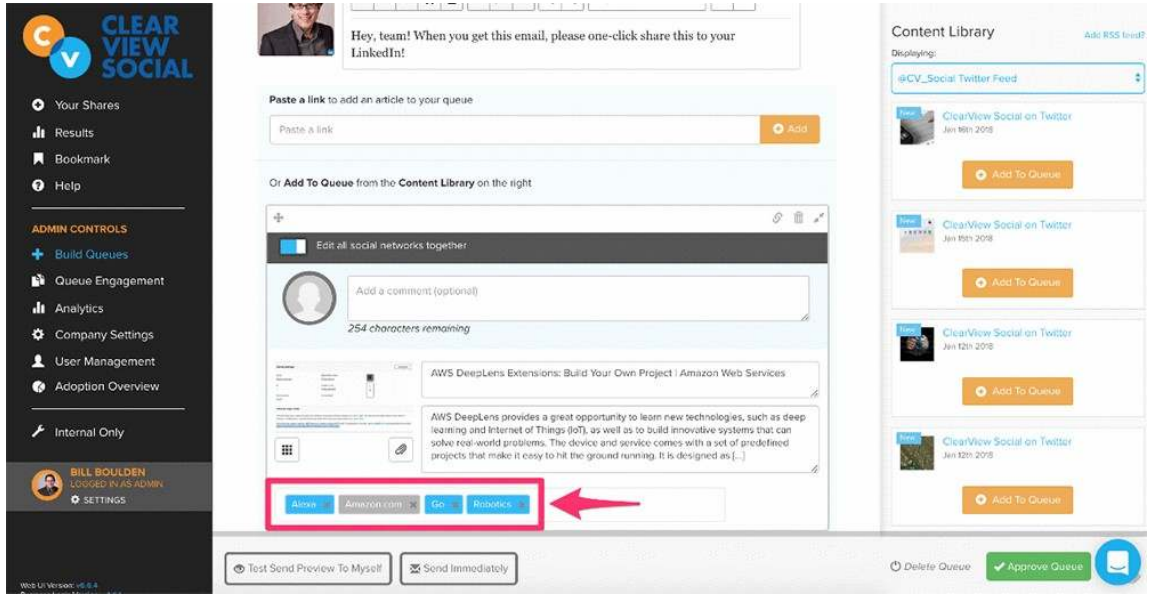
ClearView Social

- ClearView Social enables a company's employees to share approved content on LinkedIn, Twitter, and other social networks.
- To eliminate the reliance on manual tagging, ClearView Social turned to Amazon Comprehend to find insights and relationships in text.

Natural Language Processing

"We use Amazon Comprehend to read an article and extract topics, which are automatically tagged using machine learning. This automatic tagging helps customers easily estimate the market value of their engagement according to the current bid prices from the Google AdWords API"

Bill Boulden, CTO, ClearView Social



https://aws.amazon.com/blogs/machine-learning/clearview-social-uses-amazon-comprehend-to-measure-the-impact-of-social-sharing/

Digital Genius

DigitalGenius
Artificial Intelligence for Brands

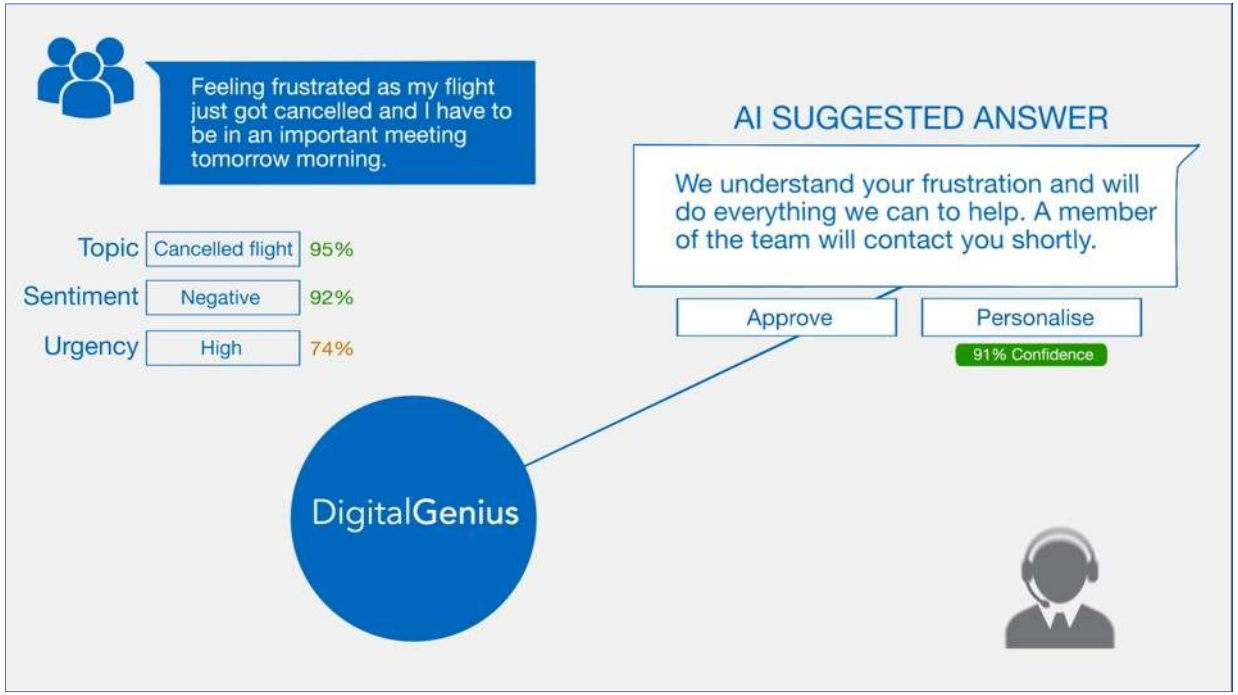
Digital Genius is an AI-powered customer service solution that monitors social media, texts and email for negative customer sentiment, identifies cases that need more attention, and suggests ideal responses to queries.

Scaling on the cloud

« We're unlocking the intelligence value of historical data while helping customer service agents deliver a faster and more accurate experience for their consumers »

« NVIDIA GPUs make it possible for us to train very large neural nets with millions of parameters in a matter of hours rather than days »

Mikhail Naumov, President, DigitalGenius



https://blogs.nvidia.com/blog/2017/01/27/faster-customer-service-with-ai/

Computer Vision

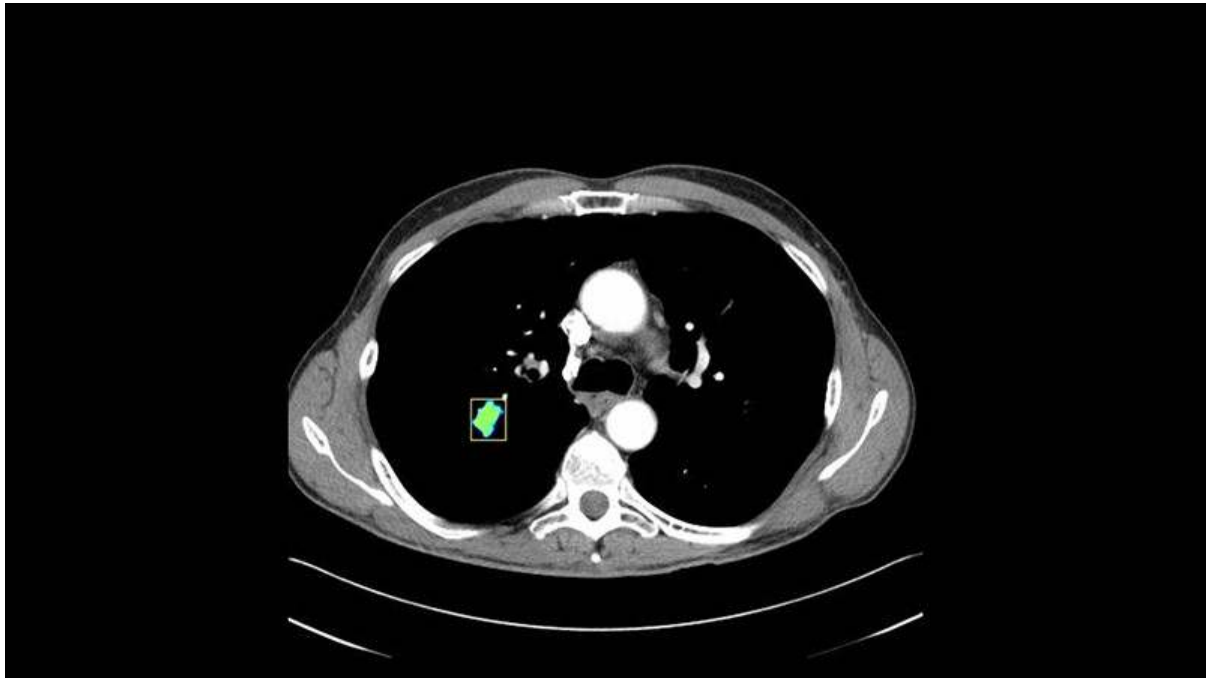
GumGum

- GumGum unlocks the value of images in advertising &professional sports.
- Mantii, GumGum's real-time social media listening tool can analyze the more than 1.8 billion social images posted daily, where 80% of them lack text to help marketers find them.
- Using Deep Learning with Apache MXNet and GPU instances, the company reduced model training time from 21 hours to 6 hours.

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

Matrix Analytics

- Matrix Analytics is helping to save lives. The Colorado-based startup uses Deep Learning on AWS (Deep Learning AMI, TensorFlow, GPU instances) to track disease progression for patients diagnosed with pulmonary nodules in their lungs.
- The Matrix Analytics tools are able to outperform previous methods in their ability to diagnose cancer from a CT image.
- The software automates follow-up care to ensure that each patient follows through with recommendations in order to monitor changes in their condition.



Deep Learning

« Using the convenience of the [Deep Learning] AMI on AWS gives us the opportunity to offer up different business models, which allows us to become excellent technology partners as the market evolves at an ever-increasing pace.

Dr. Aki Alzubaidi, Founder, Matrix Analytics

https://aws.amazon.com/blogs/machine-learning/matrix-analytics-uses-deep-learning-on-aws-to-boost-early-cancer-detection/

PowerScout

PowerScout

- PowerScout is changing the sales model for solar-powered homes by using Deep Learning to identify households likely to embrace solar panels.
- Moving away from the door-to-door sales approach, the company's deep learning models analyze satellite imagery to evaluate solar-worthy factors for each home.
- PowerScout uses GPU instances, the CUDA parallel processing platform and the cuDNN deep neural network library.

https://blogs.nvidia.com/blog/2016/12/27/ai-solar-powered-homes/

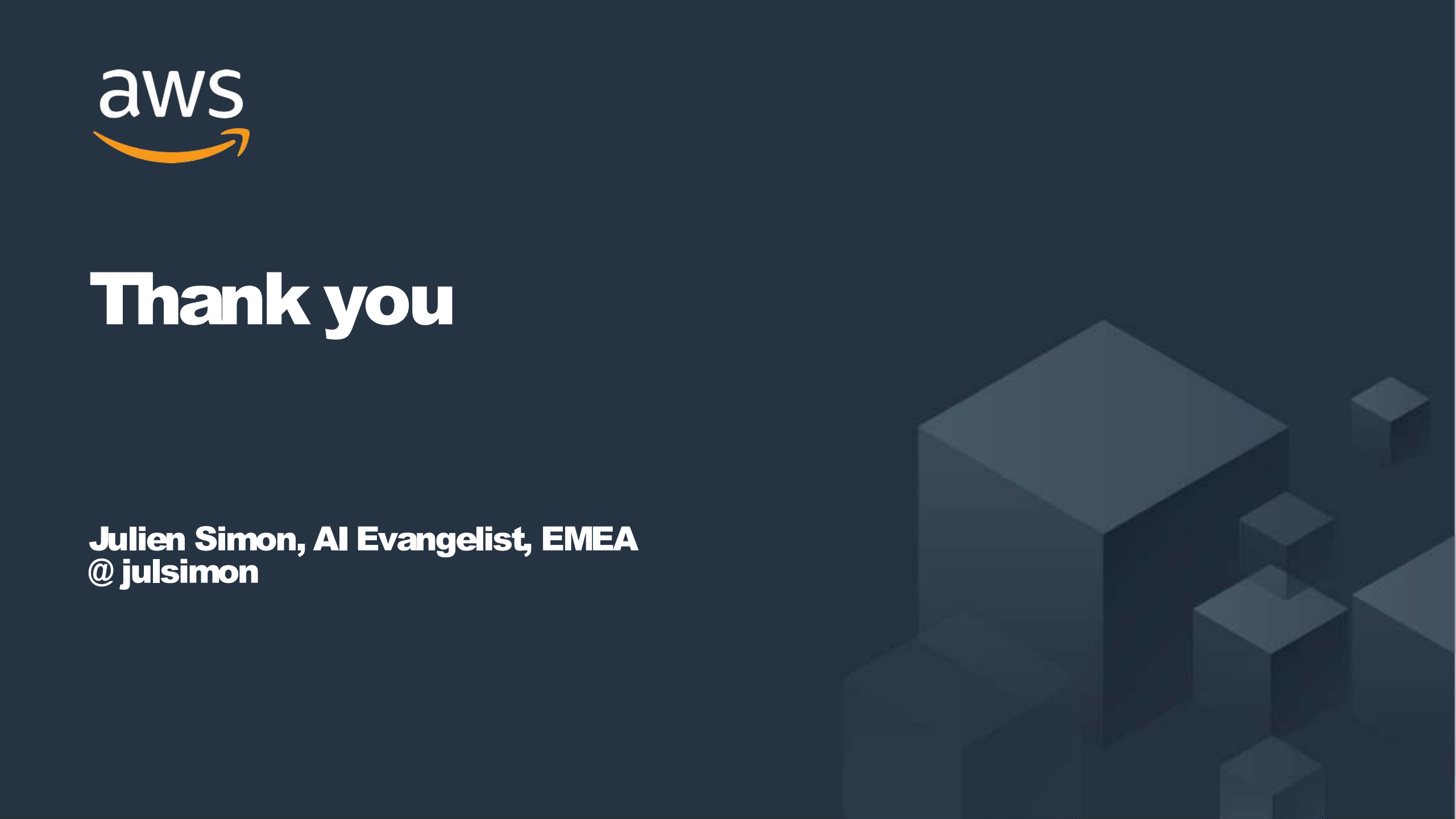
TuSimple

- TuSimple, a leader in self-driving technology, uses Deep Learning to build sophisticated algorithms for computer vision and driving simulation.
- They rely on Apache MXNet to teach computers how to recognize and track objects and to make decisions to avoid collisions and prioritize safety.
- They simulated a billion miles of road driving with a wide range of variables and driving conditions—the largest simulation of its kind in history.

https://www.oreilly.com/ideas/self-driving-trucks-enter-the-fast-lane-using-deep-learning

More AI/ML is built on AWS than anywhere else

https://aws.amazon.com/machine-learning/



Thank you

Julien Simon, AI Evangelist, EMEA
@ julsimon