Google Cloud HPC & AI Capabilities

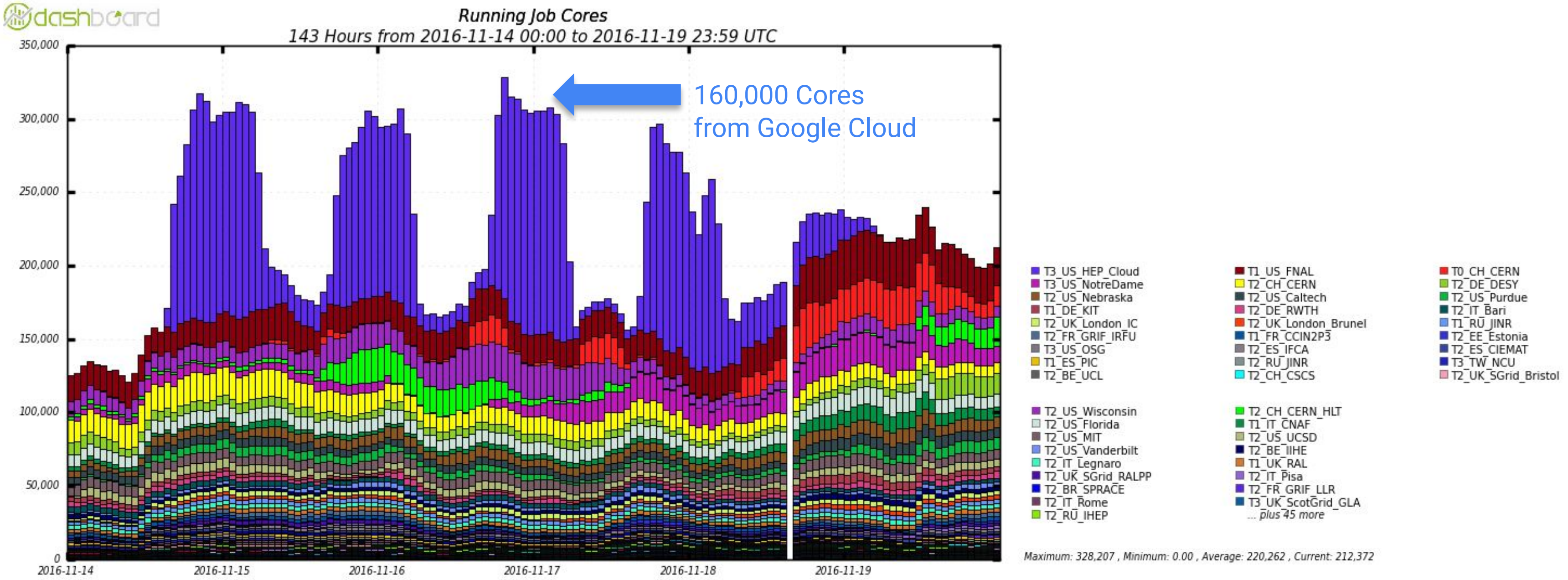
Overview

Leonid Kuligin

Karan Bhatia

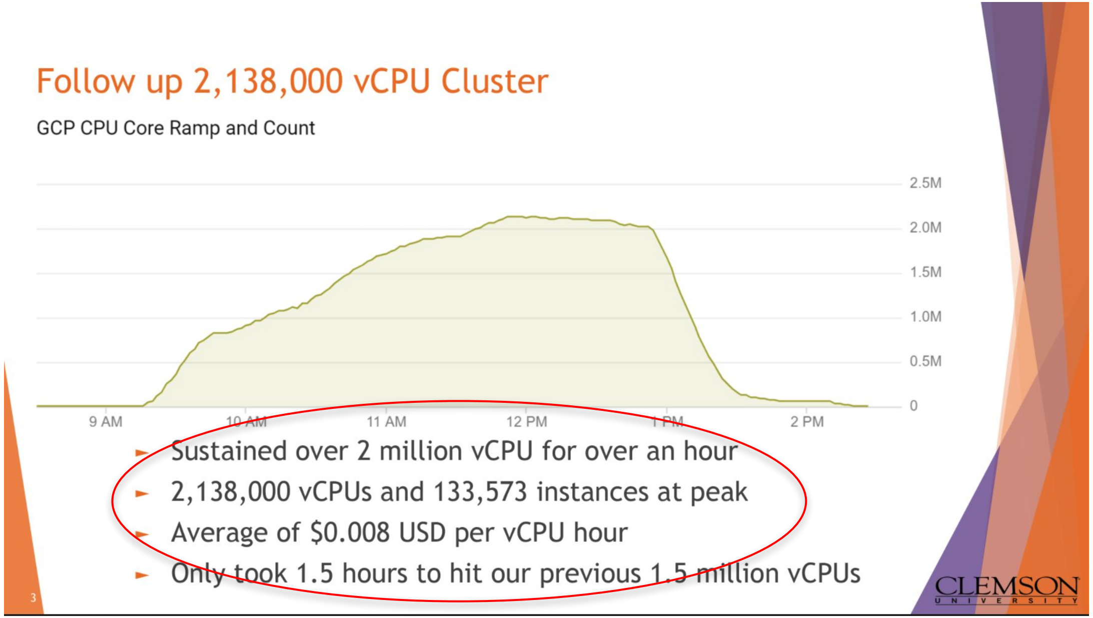
Dec 6 2019

Doubling LHC CMS compute capacity.



160,000 Cores
from Google Cloud

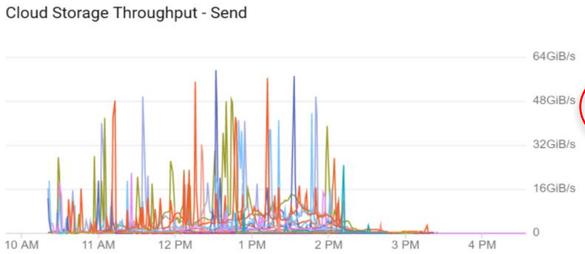
Combating Traffic Congestion using Massive HPC Analytics in the Google Cloud Platform



Follow up 2,138,000 vCPU Cluster

GCP CPU Core Ramp and Count

- Sustained over 2 million vCPU for over an hour
- 2,138,000 vCPUs and 133,573 instances at peak
- Average of $0.008 USD per vCPU hour
- Only took 1.5 hours to hit our previous 1.5 million vCPUs



Cloud Storage Throughput - Send

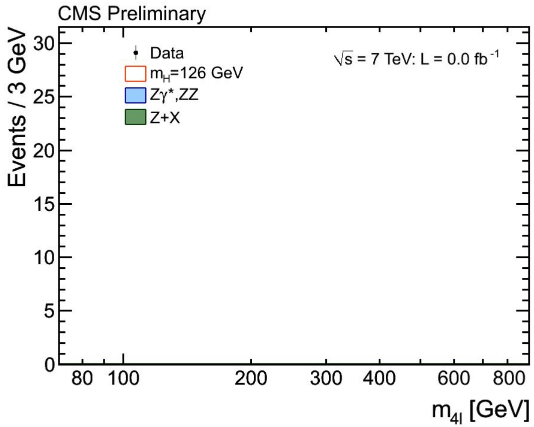
> 1Tbps throughput

- 128 GiB/s peak in Cloud Storage
- 59 GiB/s peak in one bucket

Data Intensive Computing Ecosystems (DICE)
School of Computing, Clemson University

Rediscovering Higgs boson on GCP

- https://github.com/mmm/higgs-tutorial
- Kubecon Barcelona 2019 keynote https://www.youtube.com/watch?v=CTfp2woVEkA



Google Cloud Background

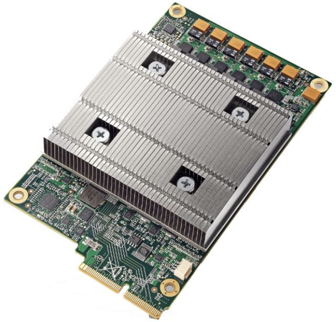
HPC Infrastructure

Compute

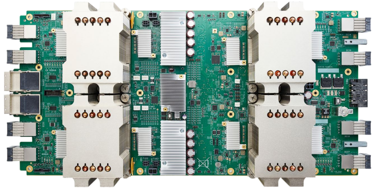
Storage

Networking

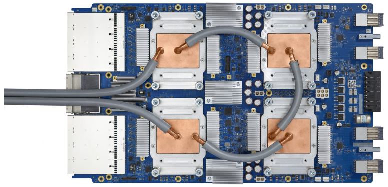
TPUs are ASICs focused on Machine Learning



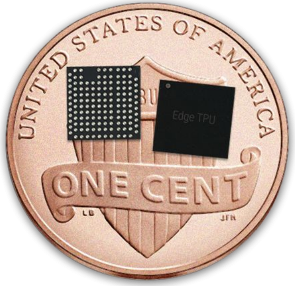
TPU v1
(2015)
92 teraops
First Generation



TPU v2
(2017)
180 teraflops
Available via Google Cloud

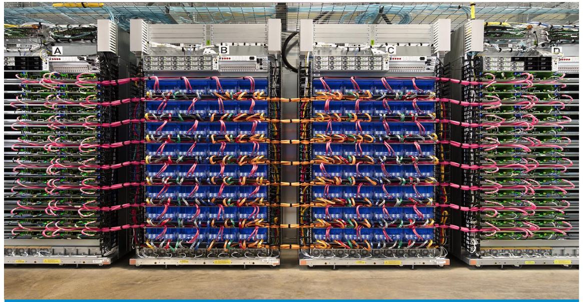


TPU v3
(2018)
420 teraflops
Available via Google Cloud
~2.3x the power of v2



Edge TPU
(2018 EAP)
Inference Accelerator

Cloud TPU Pods - Product Offerings



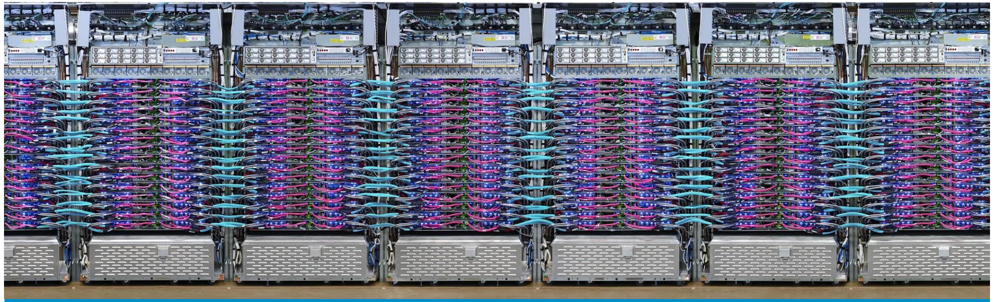
Cloud TPU v2 Pod$^{\text{Beta}}$

11,500 teraflops

Up to 256 chips

4,000 GB HBM

2-D toroidal mesh network



Cloud TPU v3 Pod$^{\text{Beta}}$

100,000+ teraflops

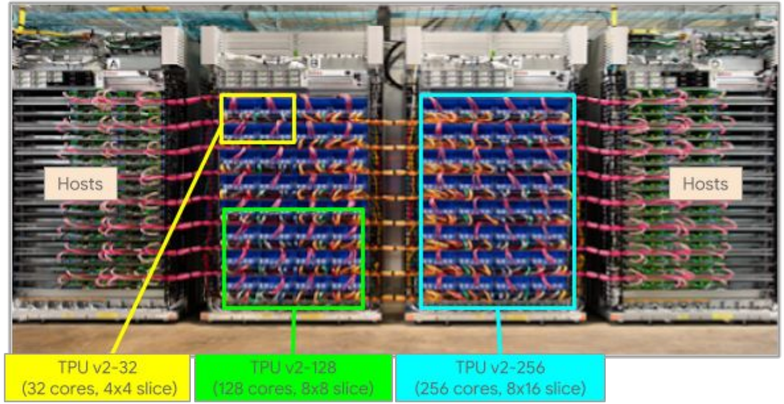
Up to 1,024 chips

32,000 GB HBM

2-D toroidal mesh network

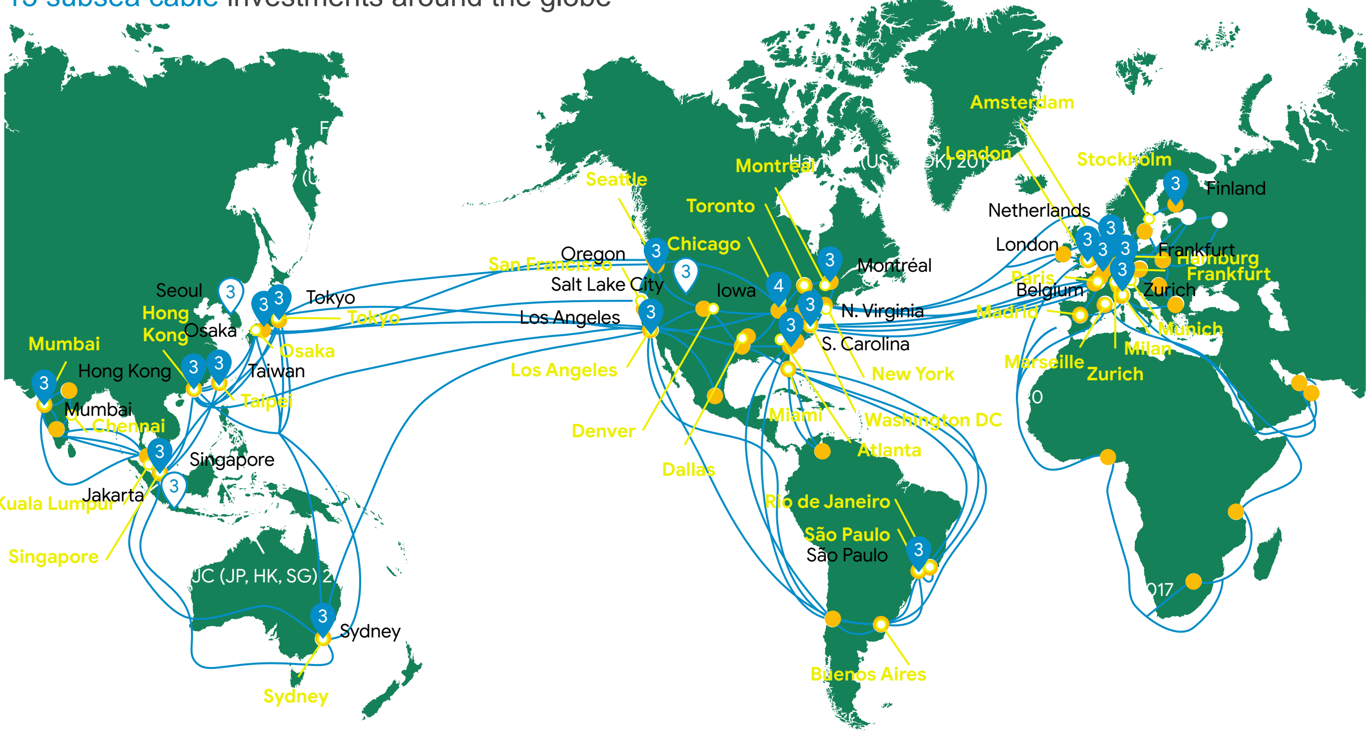
Cloud TPU Configurations

- TPU: The Tensor Processing Unit (TPU) is a custom-design chip, built from the ground up by Google for machine learning workloads.
- Cloud TPU: a device containing four TPU chips along with a fraction of a CPU host.
- Cloud TPU pods: Cloud TPUs are connected via a high-speed 2D toroidal mesh network to form Cloud TPU Pods.
- Cloud TPU slices: Slices, or smaller sections of pods, are scalable to address as much performance is needed for the workload. Slices are internal allocations consisting of different numbers of TPU cores. Pod slices come in 32, 128, 256, 512, 1024, and 2048 core-count configurations.



The network matters

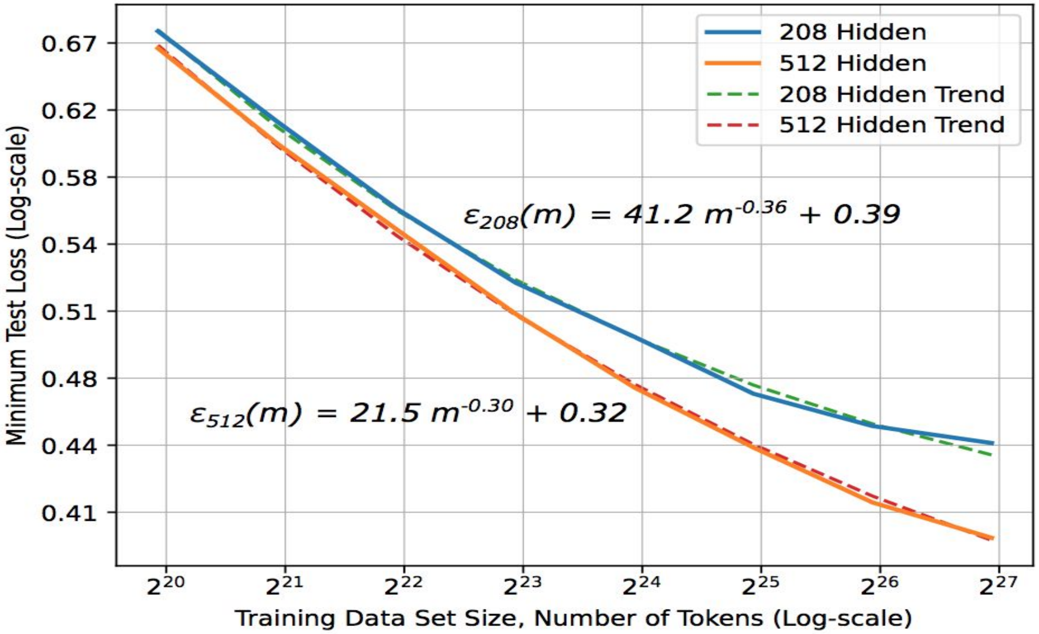
134 points of presence and 13 subsea cable investments around the globe



- ⬤ Current regions and number of zones
- ○ Future regions and number of zones
- Edge points of presence
- ● CDN nodes
- — Network
- ○ Dedicated Interconnect

HPC → Machine Learning

ML improves with data size

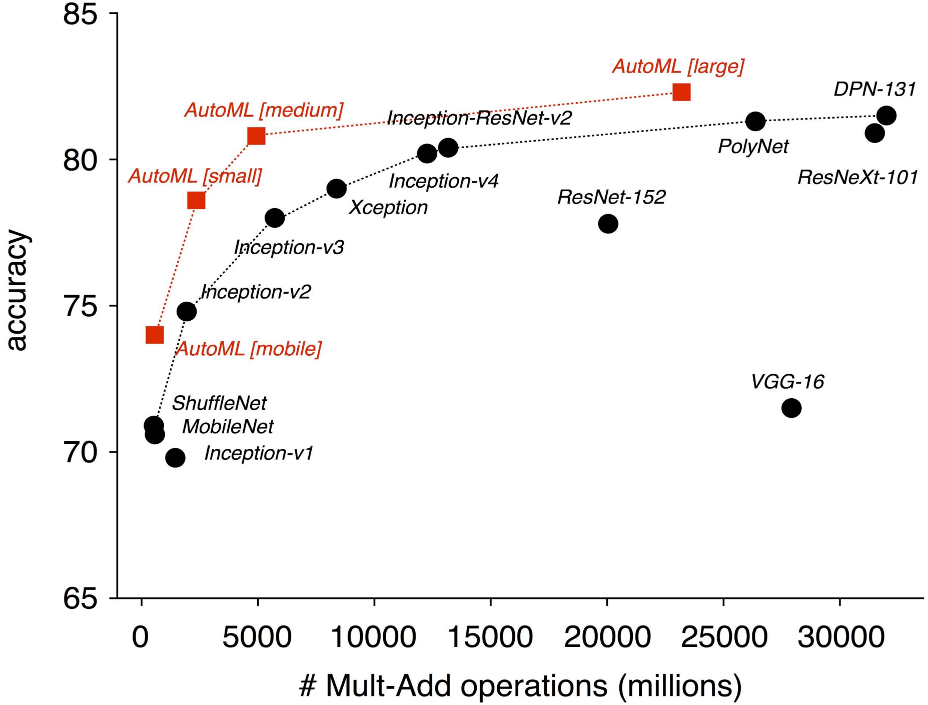


The unreasonable effectiveness of data
https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf

Deep Learning scaling is predictable, empirically
https://arxiv.org/abs/1712.00409

Google Cloud

Increases in accuracy require much more compute.



Learning Transferable Architectures for Scalable Image Recognition
Barret Zoph, Vijay Vasudevan, Jonathon Shlens, Quoc V. Le
https://arxiv.org/abs/1707.07012

Google Cloud

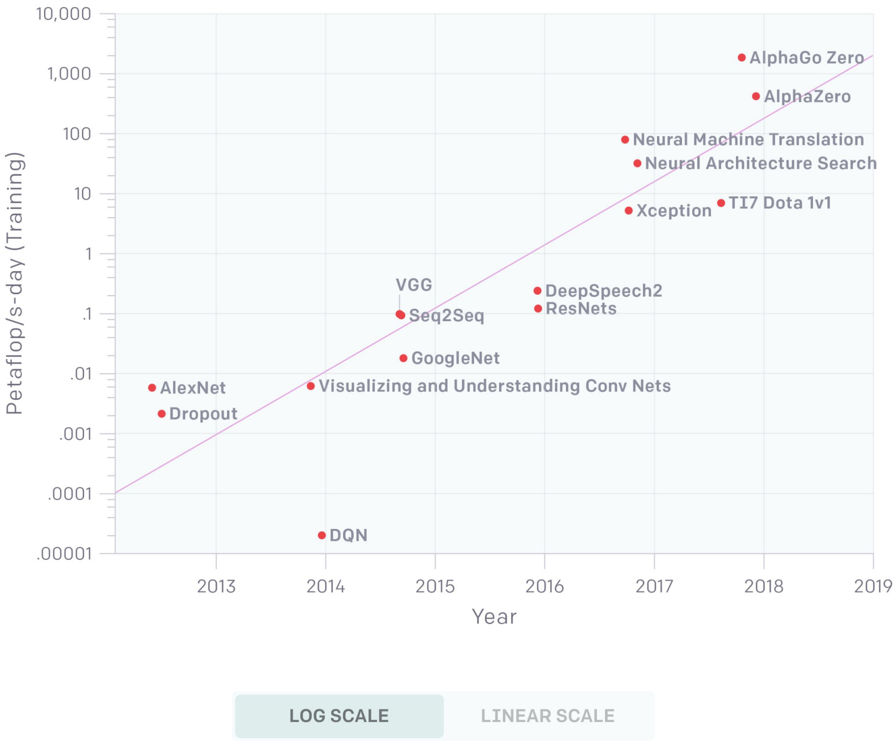
AI and Compute

MAY 16, 2018

OpenAI

https://blog.openai.com/ai-and-compute/

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



Google Cloud

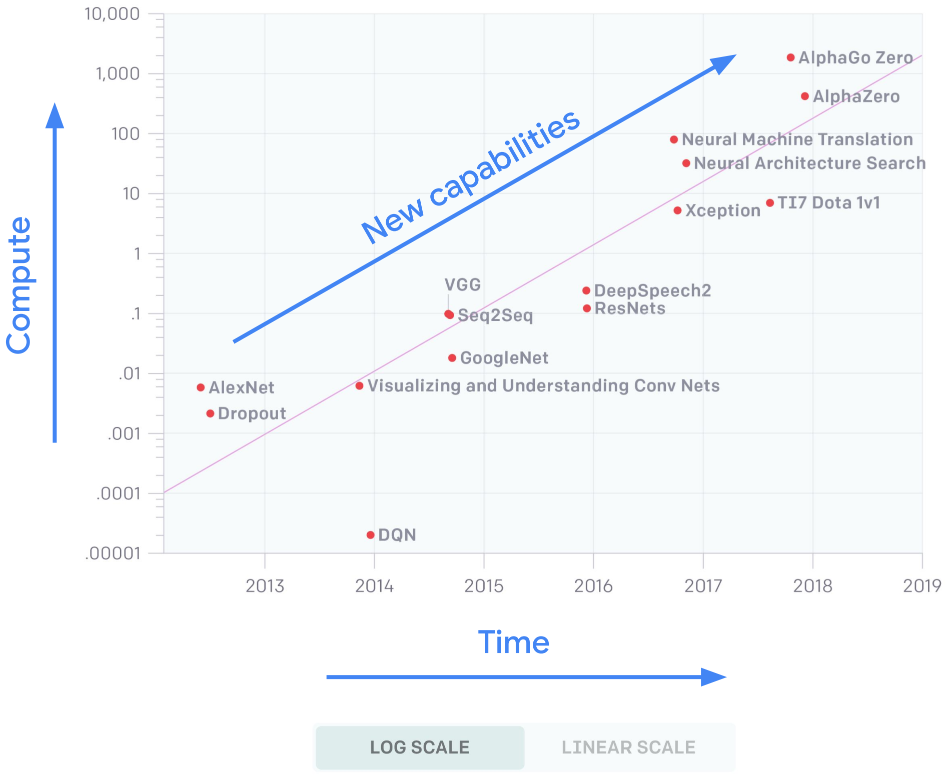
AI and Compute

MAY 16, 2018

OpenAI

https://blog.openai.com/ai-and-compute/

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



AI and Compute

MAY 16, 2018

... since 2012, the amount of compute used in the largest AI training runs **has been increasing exponentially with a 3.5-month doubling time** (by comparison, Moore's Law had an 18-month doubling period).

Since 2012, **this metric has grown by more than 300,000x** (an 18-month doubling period would yield only a 12x increase).

Google Cloud

OpenAI

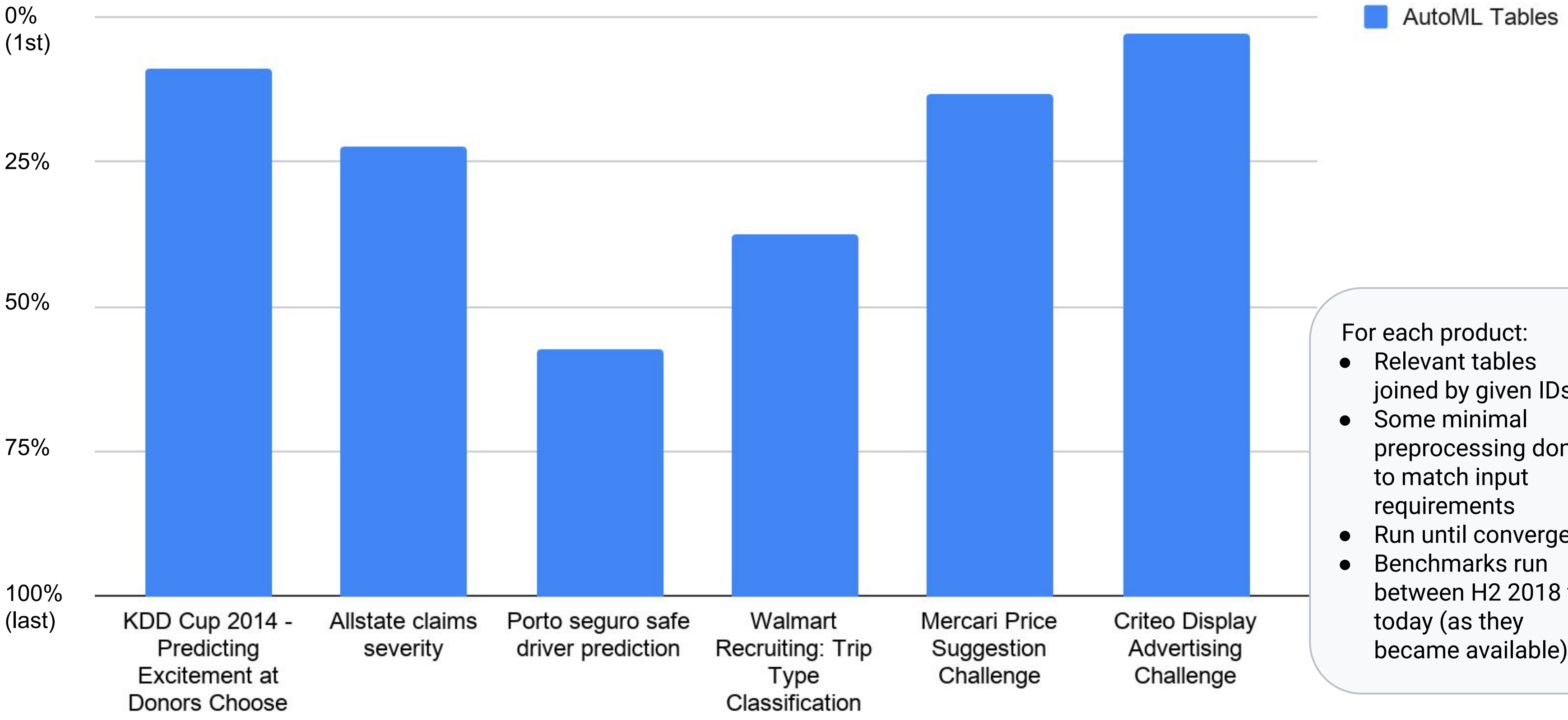
... within many current domains, **more compute seems to lead predictably to better performance**, and is often complementary to algorithmic advances.

... we believe **the relevant number** is not the speed of a single GPU, nor the capacity of the biggest datacenter, but **the amount of compute that is used to train a single model** – this is the number most likely to correlate to how powerful our best models are.

https://blog.openai.com/ai-and-compute/

Leading to increased model quality

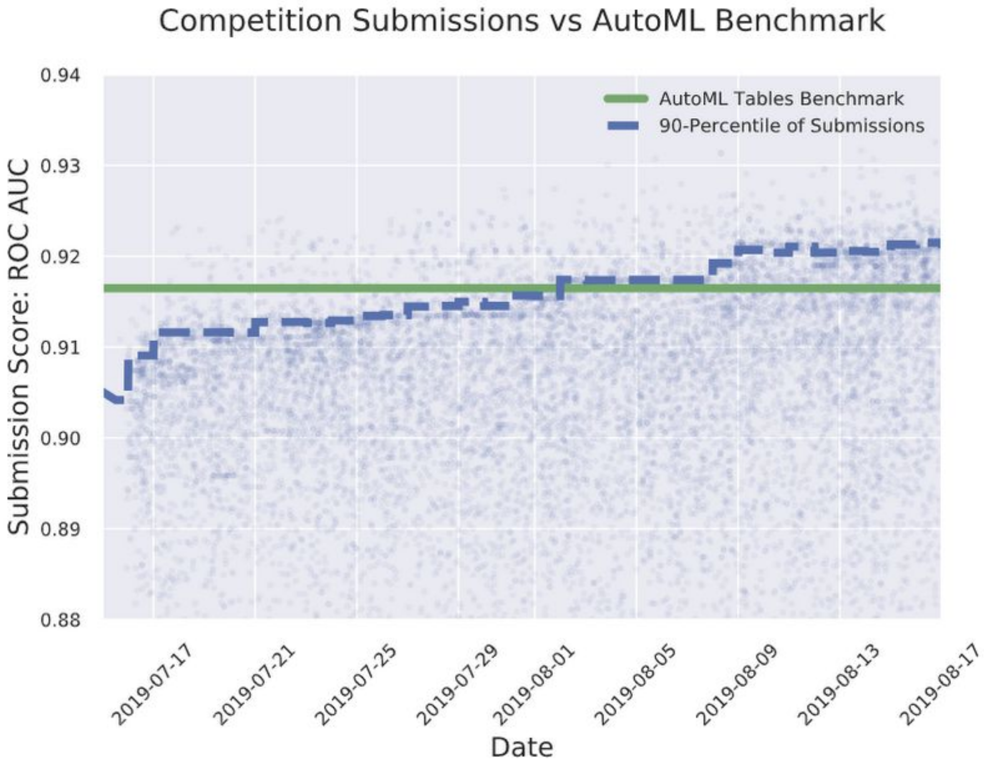
% ranking on Kaggle private leaderboard

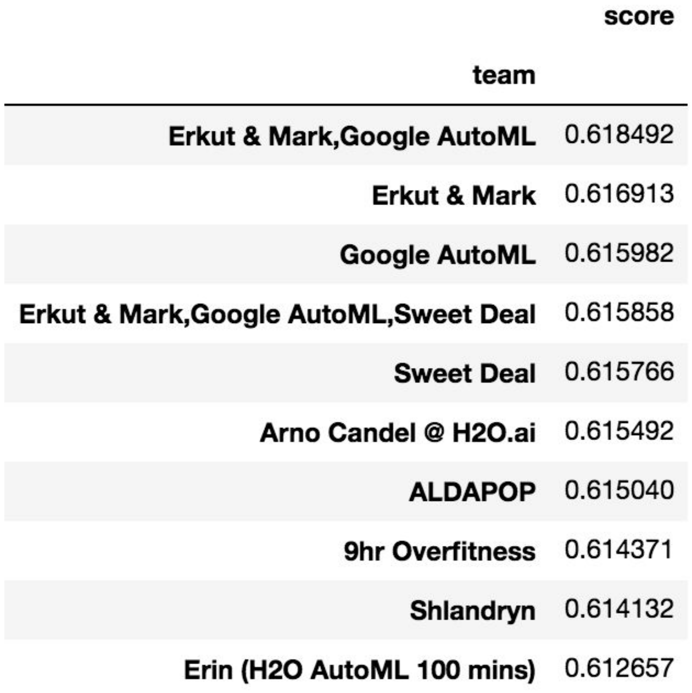


For each product:

- Relevant tables joined by given IDs
- Some minimal preprocessing done to match input requirements
- Run until converge
- Benchmarks run between H2 2018 to today (as they became available)

AutoML + experts = even better!



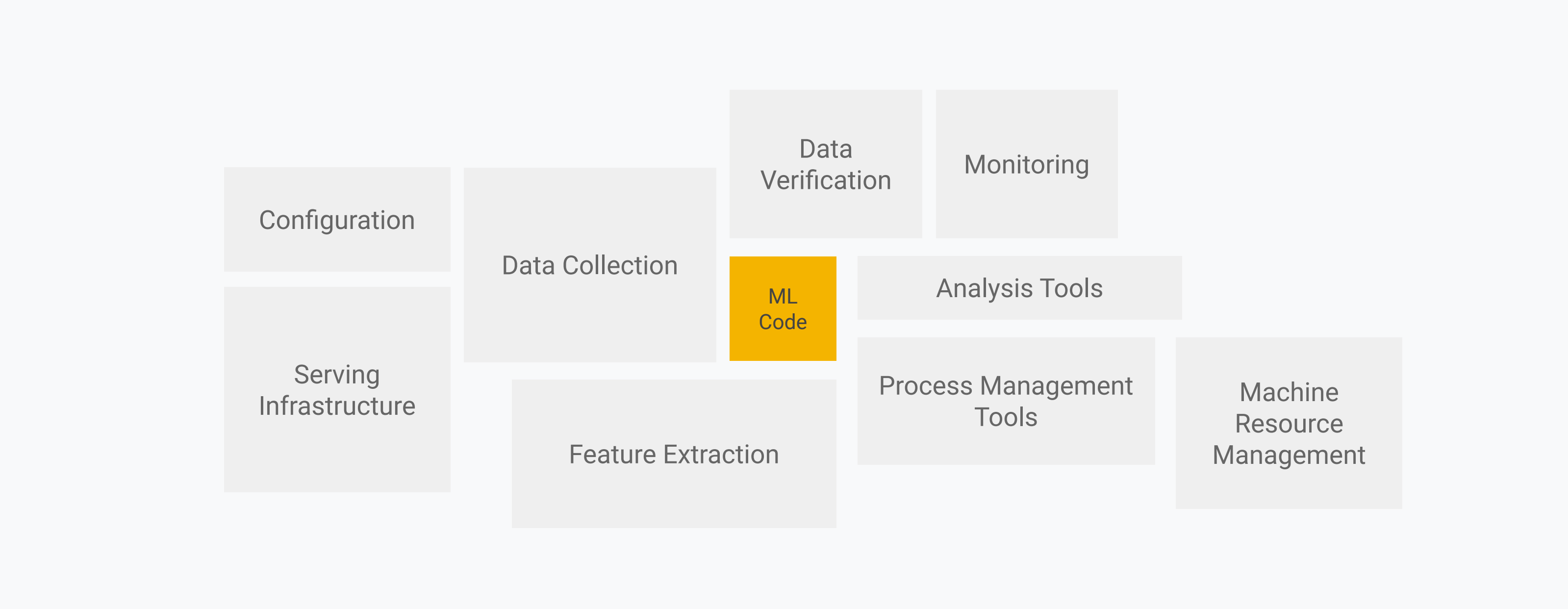


	score
team	
Erkut & Mark,Google AutoML	0.618492
Erkut & Mark	0.616913
Google AutoML	0.615982
Erkut & Mark,Google AutoML,Sweet Deal	0.615858
Sweet Deal	0.615766
Arno Candel @ H2O.ai	0.615492
ALDAPOP	0.615040
9hr Overfitness	0.614371
Shlandryn	0.614132
Erin (H2O AutoML 100 mins)	0.612657

To do ML in production, in addition to the actual ML...

ML Code

...you have to worry about so much more.

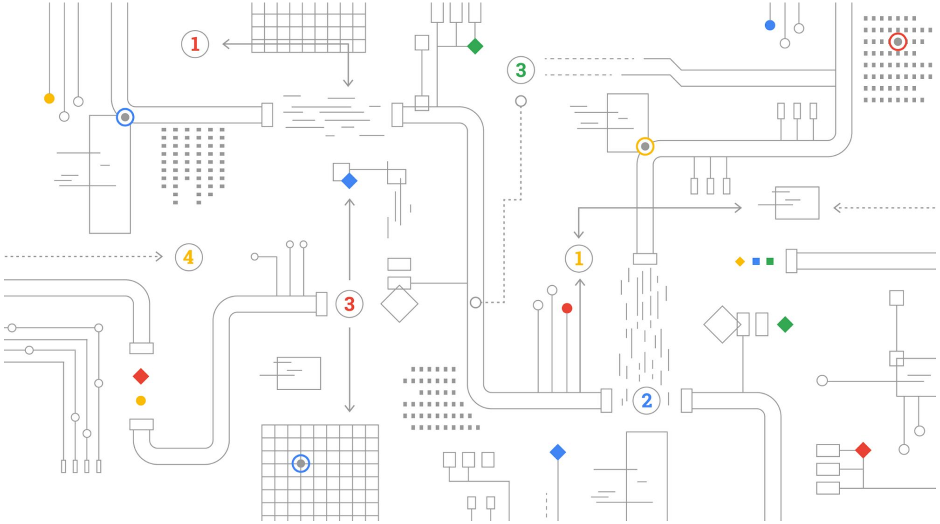


Source: Sculley et al.: Hidden Technical Debt in Machine Learning Systems

Google Cloud

Operatize ML pipelines, not ML models

- Input validation
- Model retraining
- Reusable and shareable components
- ML microservices
- Serverless



Google Cloud

AI Hub

Public Content

By Google

Unique AI
assets by Google

AutoML, TPUs, kaggle
Cloud AI Platform, etc.

Research at Google

DeepMind

By Partners

Created, shared
& monetized
by anyone.

+ Private Content

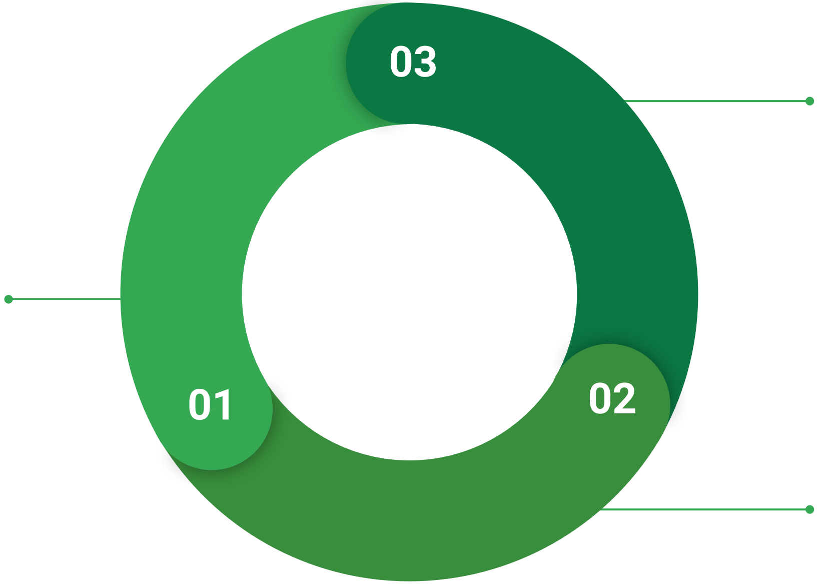
By Customers

Content shared Securely within
and with other organizations.

AI Platform Overview

<table>
<tr><th></th><th>Data Readiness</th><th>Feature Engineering</th><th>Training/ HP-Tuning</th><th>Model Management</th><th>Prediction</th><th>Understanding / Tuning</th><th>Productionize</th></tr>
<tr><td rowspan="2">Fully Managed</td><td colspan="6">AI Platform</td><td></td></tr>
<tr><td>Labeling</td><td></td><td>Training</td><td>Hub Metadata</td><td>Prediction</td><td>Explainability</td><td></td></tr>
<tr><td rowspan="2">Semi- Managed</td><td></td><td colspan="4">DL Environment (DL VM + DL Container)</td><td></td><td></td></tr>
<tr><td></td><td colspan="4">Notebooks</td><td></td><td></td></tr>
<tr><td rowspan="3">On Prem &amp; DIY</td><td></td><td></td><td colspan="5">Kubeflow</td></tr>
<tr><td></td><td></td><td>Training</td><td>Hub / Metadata</td><td>Prediction</td><td>Catib</td><td>Serving</td></tr>
<tr><td></td><td></td><td colspan="5">Pipelines</td></tr>
</table>

Choose operational complexity based on level of control needed



Cloud AI Platform

Turnkey serverless training and batch/online predictions

Kubeflow pipelines on Google Kubernetes Engine

Fully customizable

Deep Learning VM
Cloud Functions
Cloud Dataflow

Customized training, online/batch predictions

Takeaways

GCP for HTC and AI

GCP well suited for high throughput computing with many partners, schedulers and cost effective solutions.

AI capabilities both for quick prototyping as well as scaled training

Cloud is ideal both for quick prototyping, sharing and reusing code and ML models, as well as for reproducible workloads (both large-scale training and high-throughput inference)

Scale and hybrid approach

You can use GCP capabilities to achieve enormous scale to add them to your existing on-premise resources

Thank you

Google

Karan Bhatia
karanbhatia@google.com
Leonid Kuligin
kuligin@google.com