Deploying Machine Learning Models at Scale


Google
Google Search
I'm Feeling Lucky

AI is the new ground for gaining competitive edge & creating business value

Competitive advantage ranked as top goal of machine-learning projects for 46% of IT leaders & 50% of adopters can quantify ROI

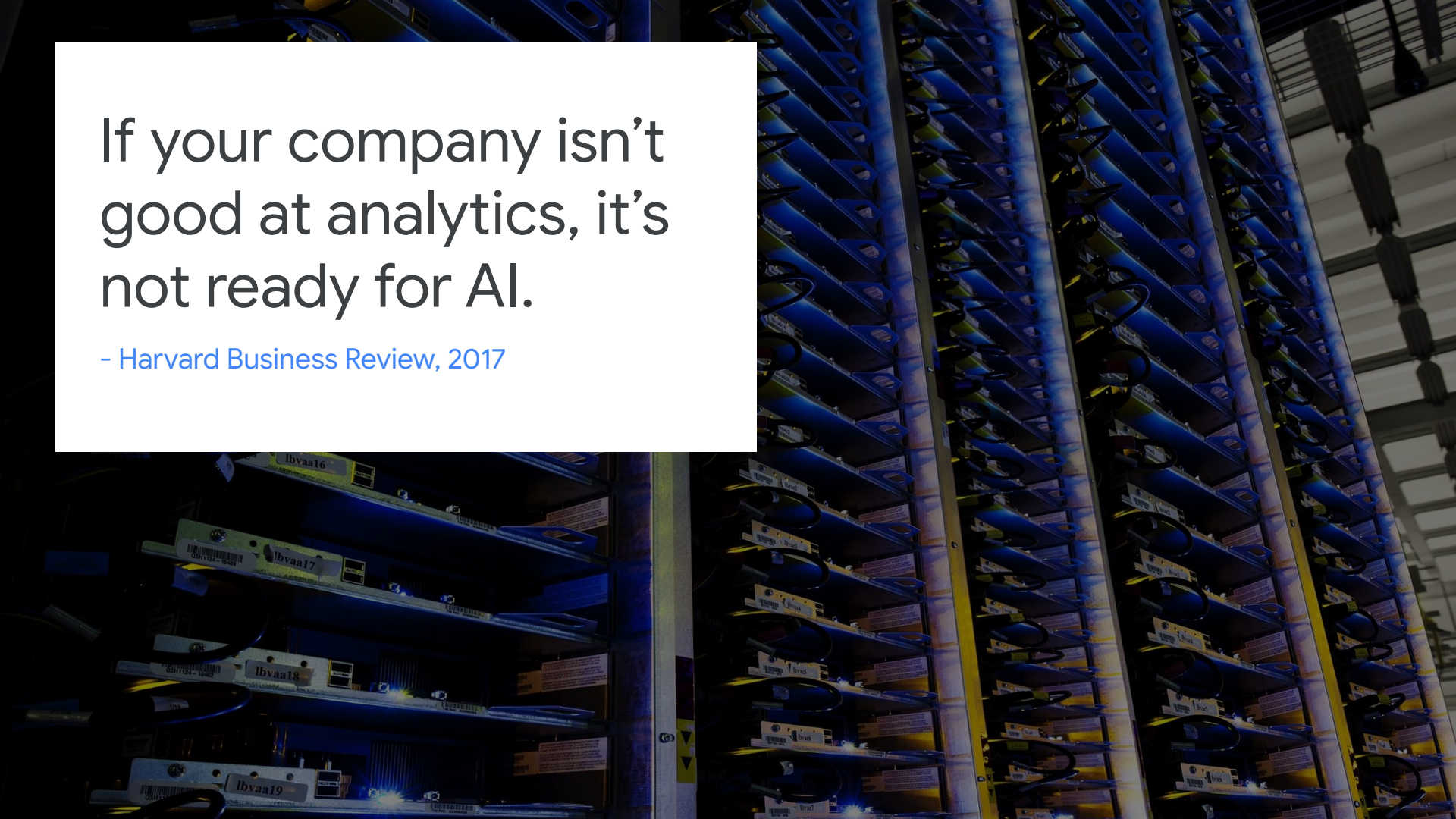
2X more data-driven decisions

5X faster decisions than others

3X faster execution

*Source: MIT Survey 2017; n=375
Bain Consulting Study

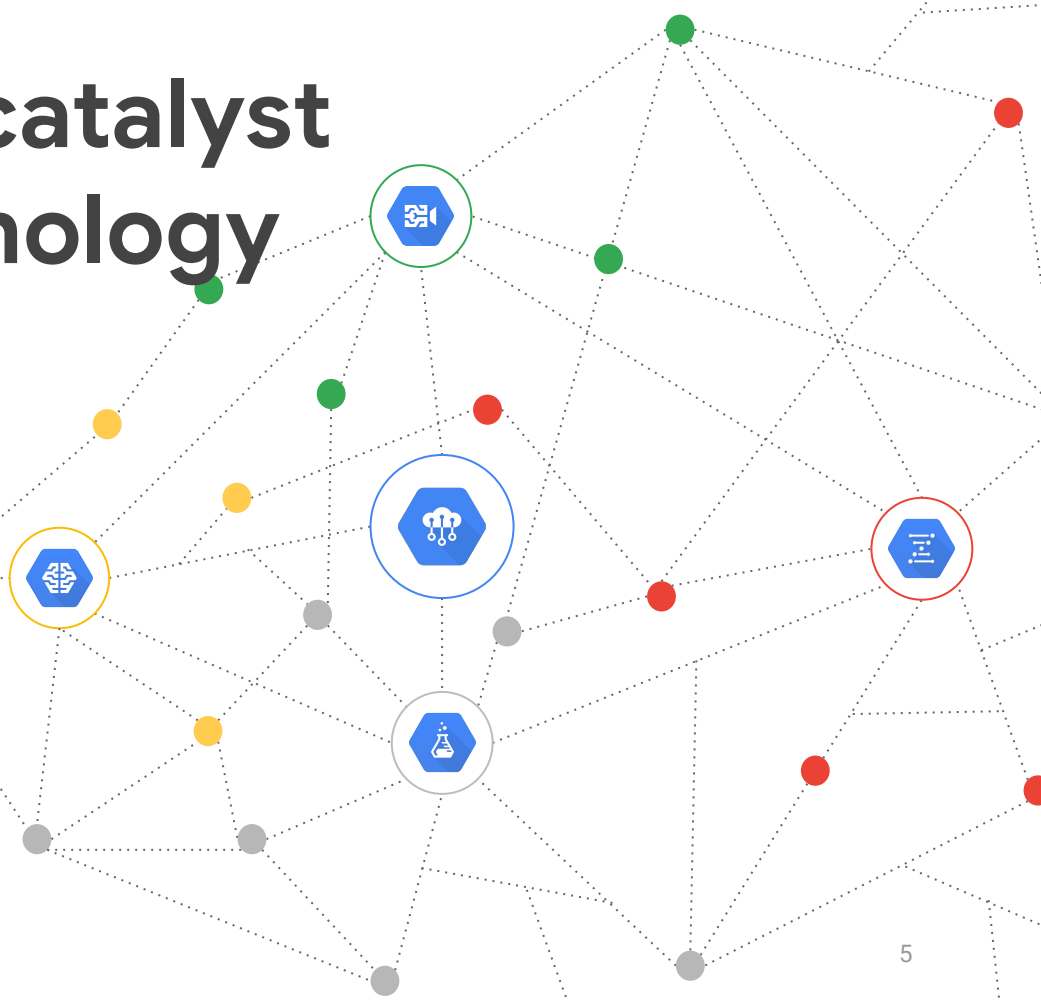
Google Cloud



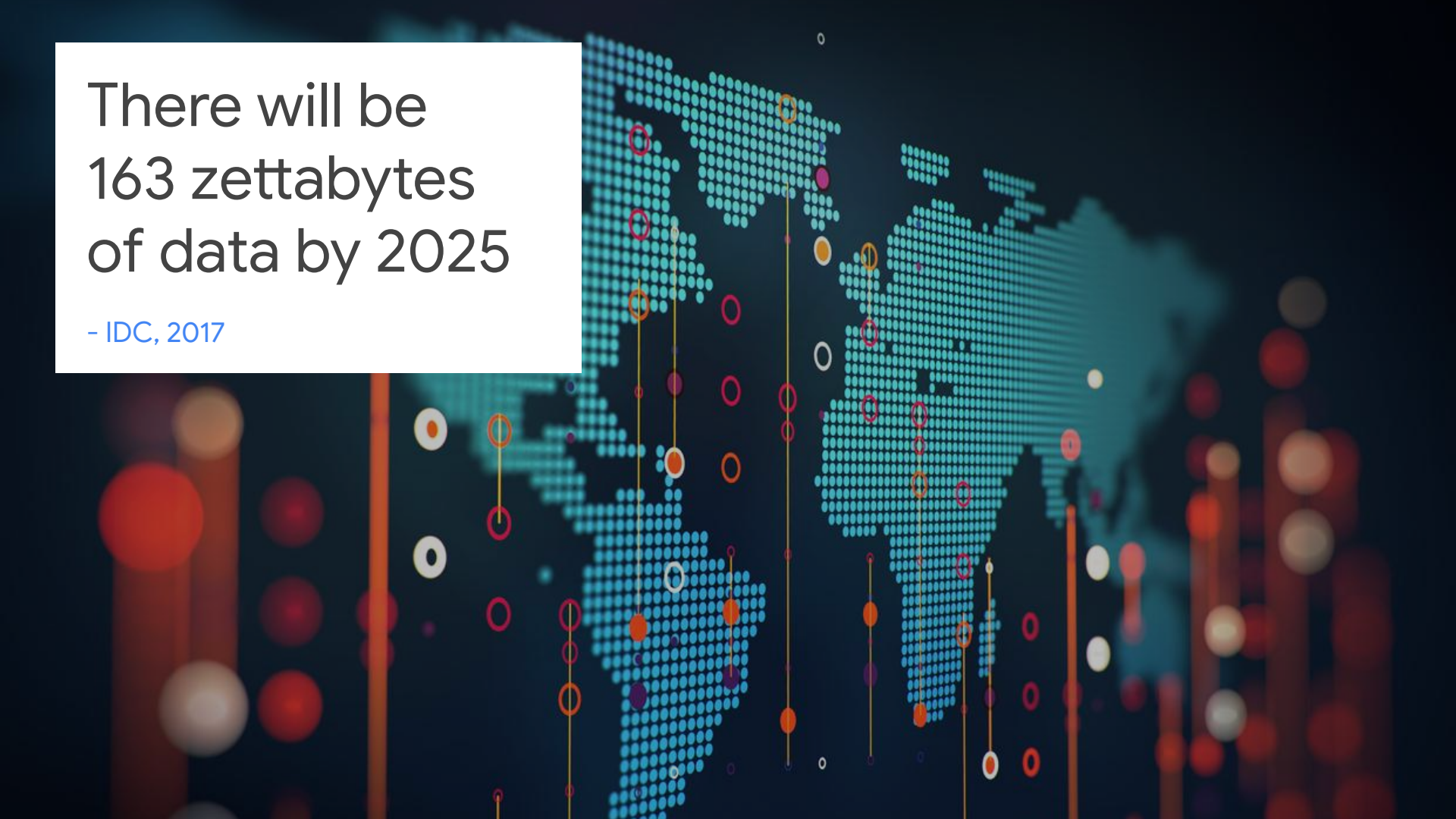
If your company isn't good at analytics, it's not ready for AI.

- Harvard Business Review, 2017

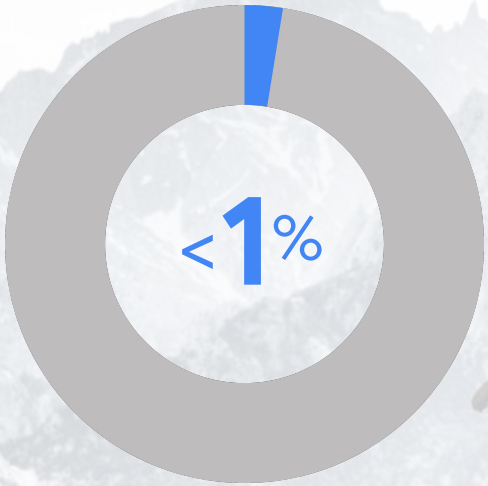
People are the catalyst that drive technology



Google Cloud

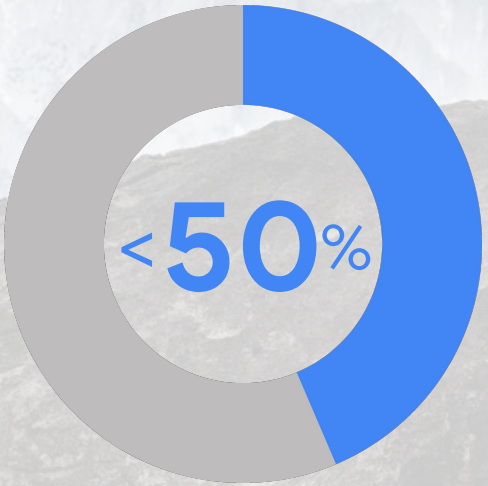

There will be
163 zettabytes
of data by 2025
- IDC, 2017

Data analytics is still too hard



Less than 1% of unstructured data is analyzed or used at all*

Less than 50% of structured data is used to make decisions*



Google Cloud

* Harvard Business Review magazine; May-June 2017

Managing data volume and speed on traditional platforms results in

60-80% → **Higher** up-front, operational and maintenance costs

60%* → **Higher** risk of failure

Google Cloud

*Gartner

Our approach to data analytics

Focus on analytics not infrastructure

Leave scaling, performance, availability and security to Google Cloud's **serverless data analytics** platform

Develop comprehensive solutions

Modern **data warehouse**, streaming data **real-time analytics**, advanced **data visualization** and **AI**

End-to-end ML lifecycle

Operationalize predictive analytics as a logical next step in customers' analytics journey

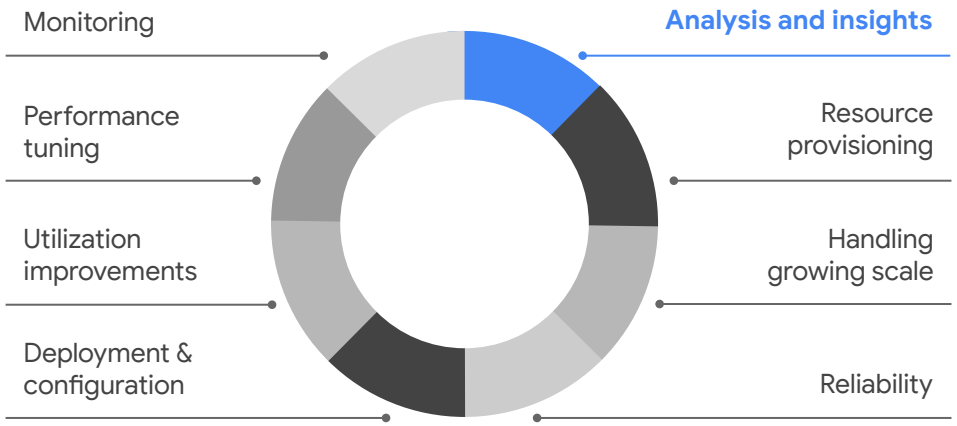
Innovation and proven results

Proven track record of data analytics innovations. Leading enterprises rely on Google Cloud data analytics solutions

Serverless data analytics

From infrastructure to platform for insights

The traditional data analytics platform



The serverless data analytics model

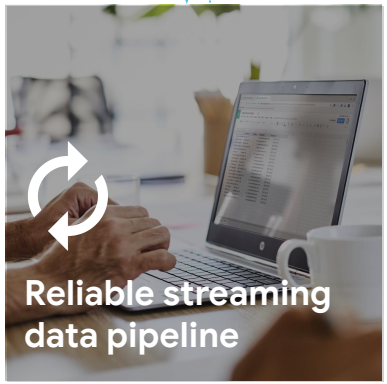
Google Cloud

Complete foundation for data lifecycle

Cloud Pub/Sub

Data Transfer Service

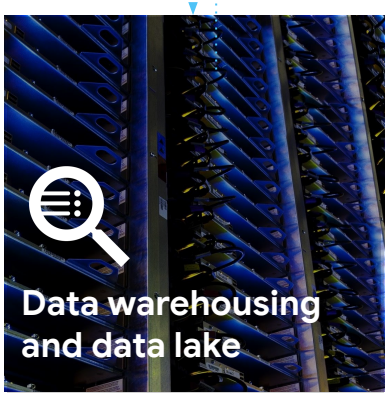
Cloud IoT Core



Cloud Dataflow

Cloud Dataproc

Apache Beam



BigQuery

Cloud Storage

Cloud AI

Google Data Studio

Tensorflow

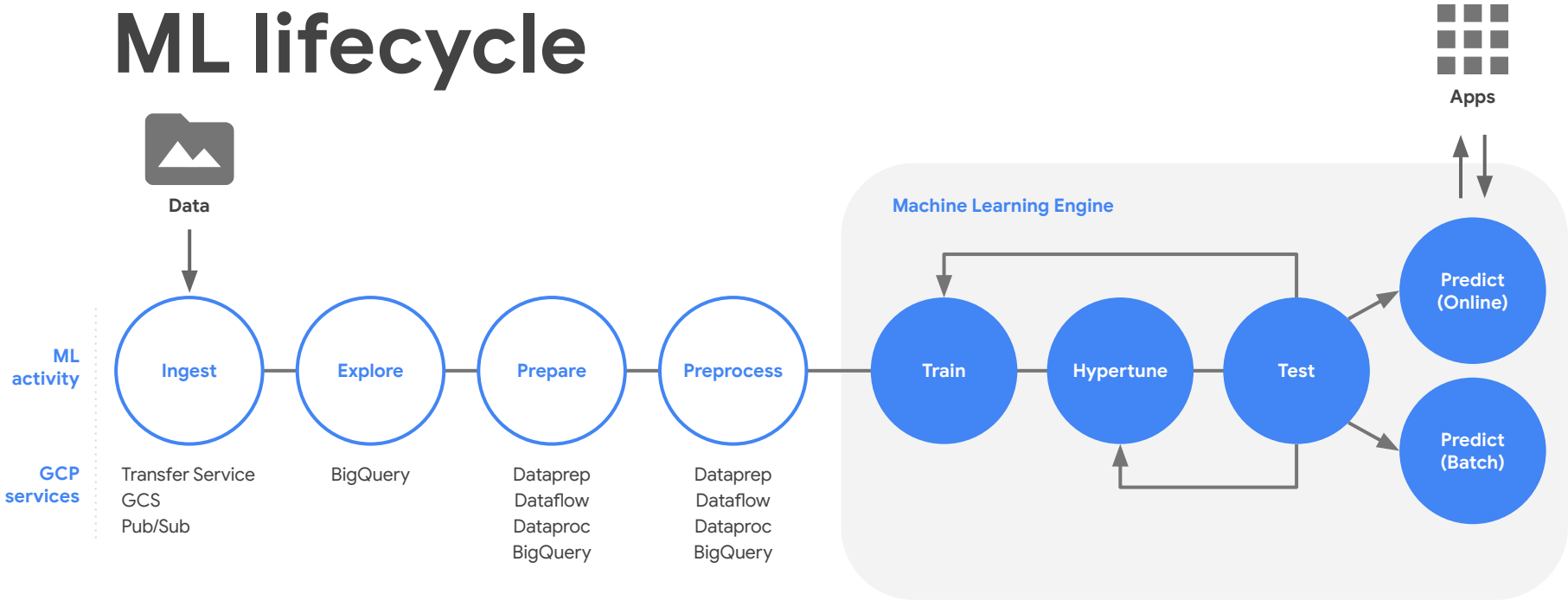
Google Sheets

Cloud Dataprep

Cloud Composer

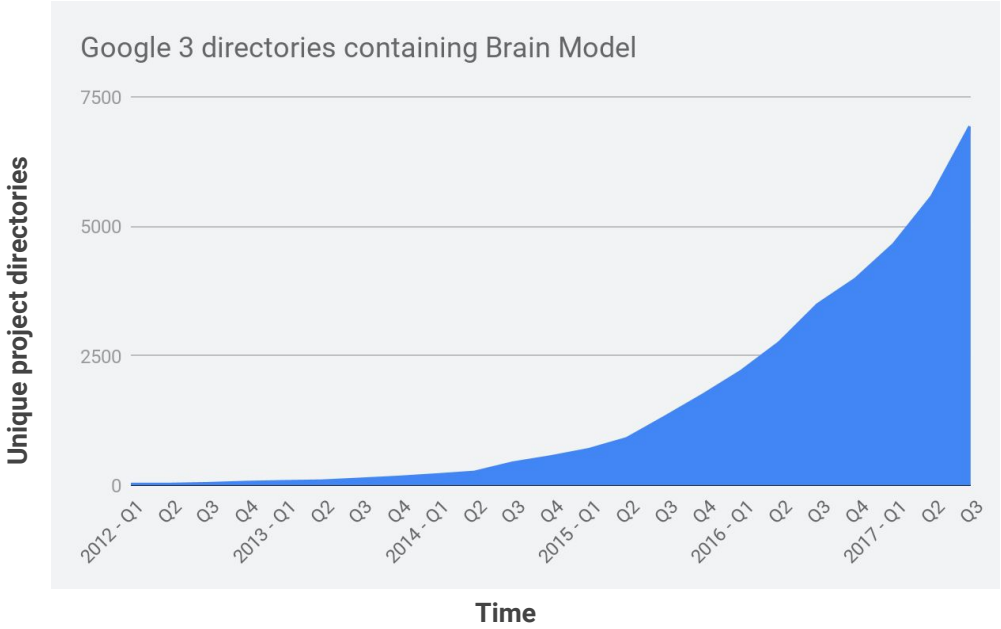
Google Cloud

Serverless analytics for complete ML lifecycle

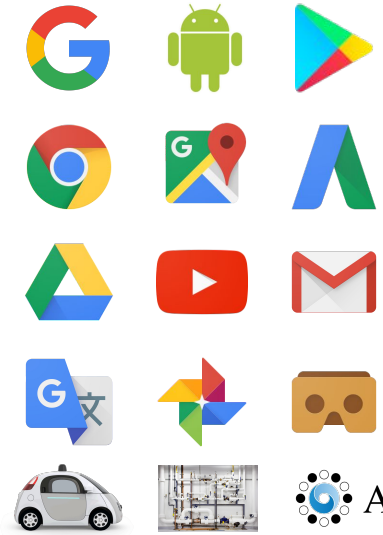


Google Cloud

Google is an AI company



Used across products:

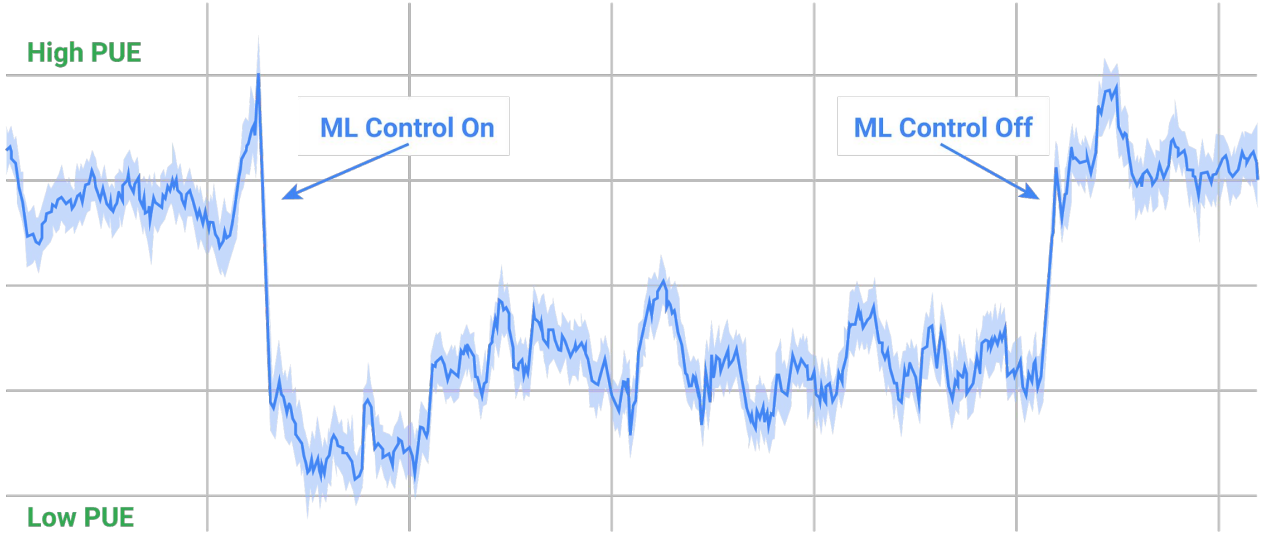


Google Cloud

Google Translate

Google Cloud

Saved data center cooling energy for 40%
Improved power usage effectiveness (PUE) for 15%

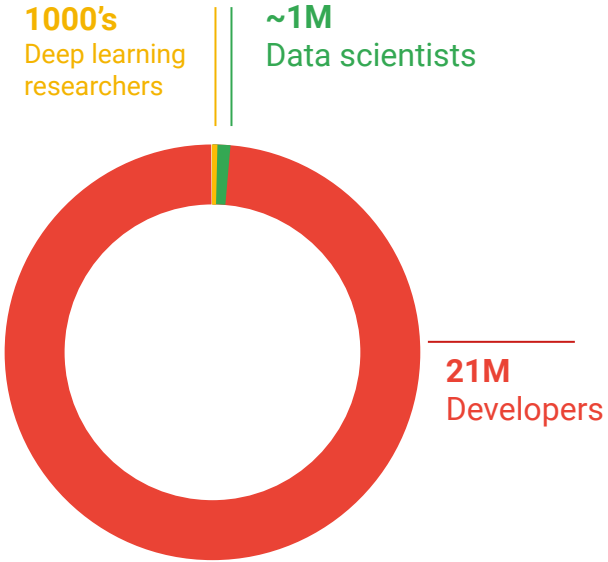


Google Cloud

Who can actually use AI today?

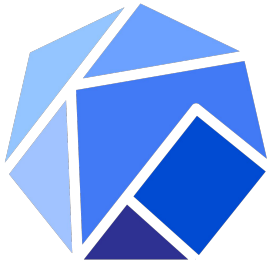
Very few users today can create a custom ML model

We need to make AI accessible to millions more



Google Cloud

Free and Open source tools

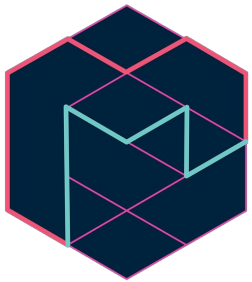


Kubeflow

TensorFlow.js

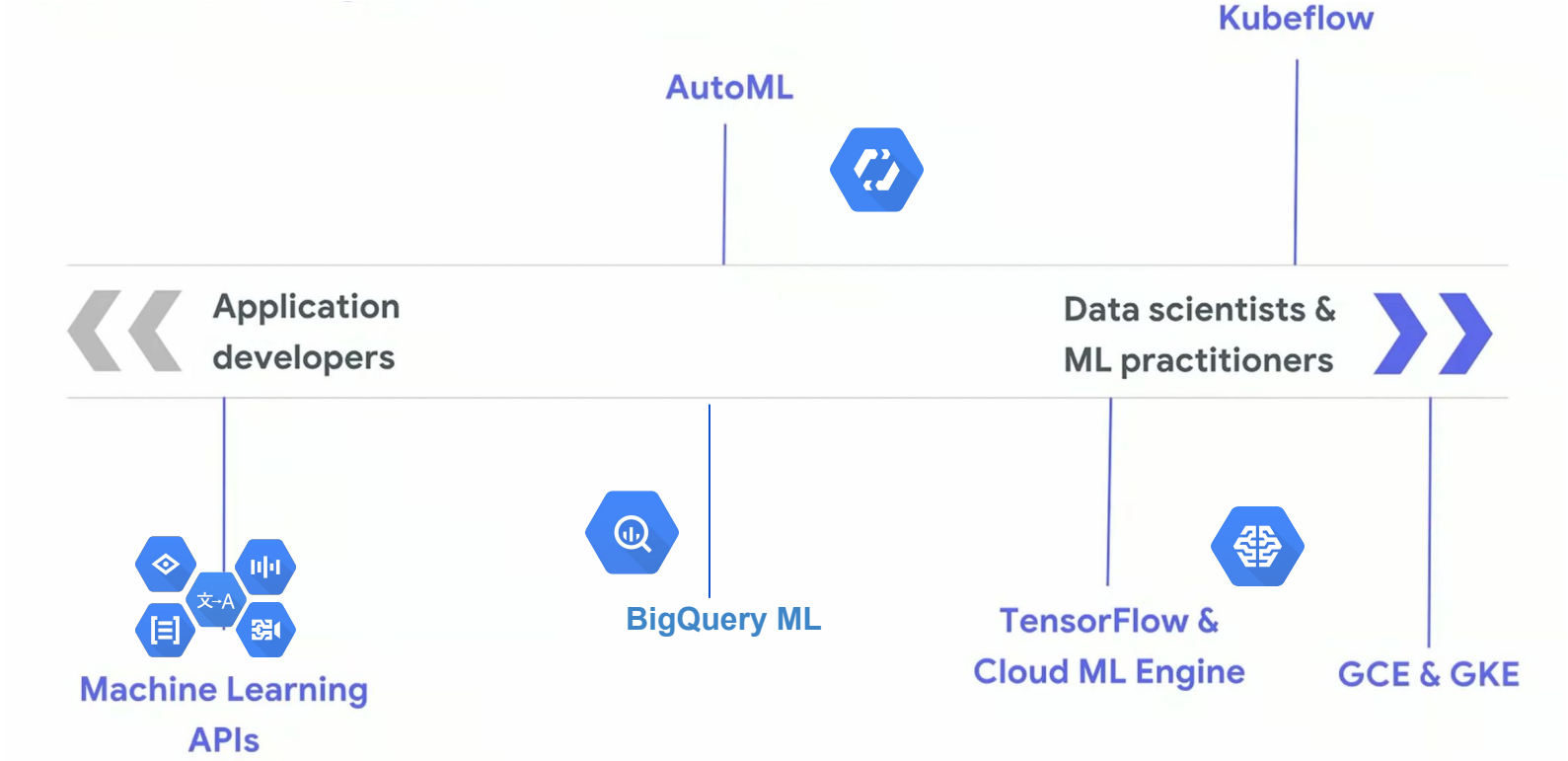
Google Dataset Search Beta

colab



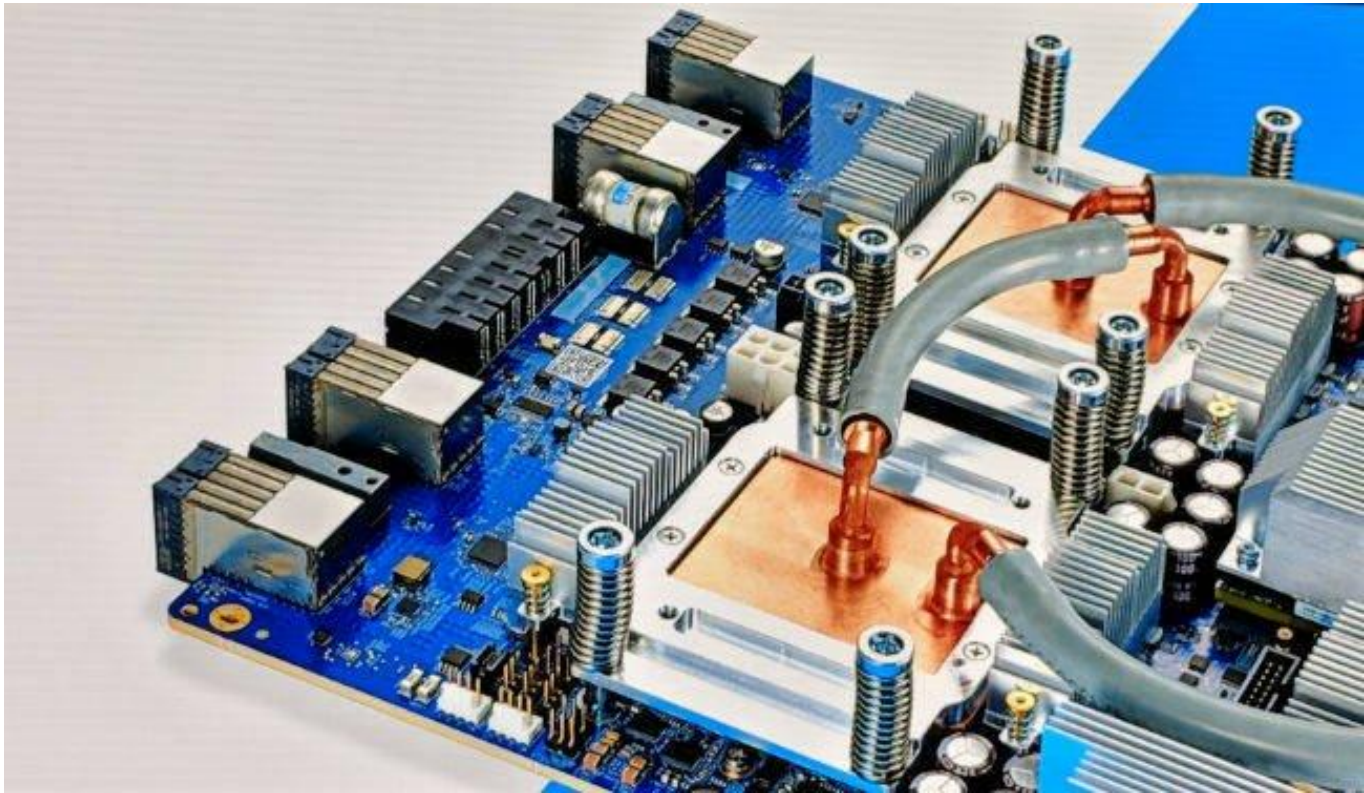
Google Cloud

Meeting users at their skill level



Google Cloud

Hardware for Machine Learning



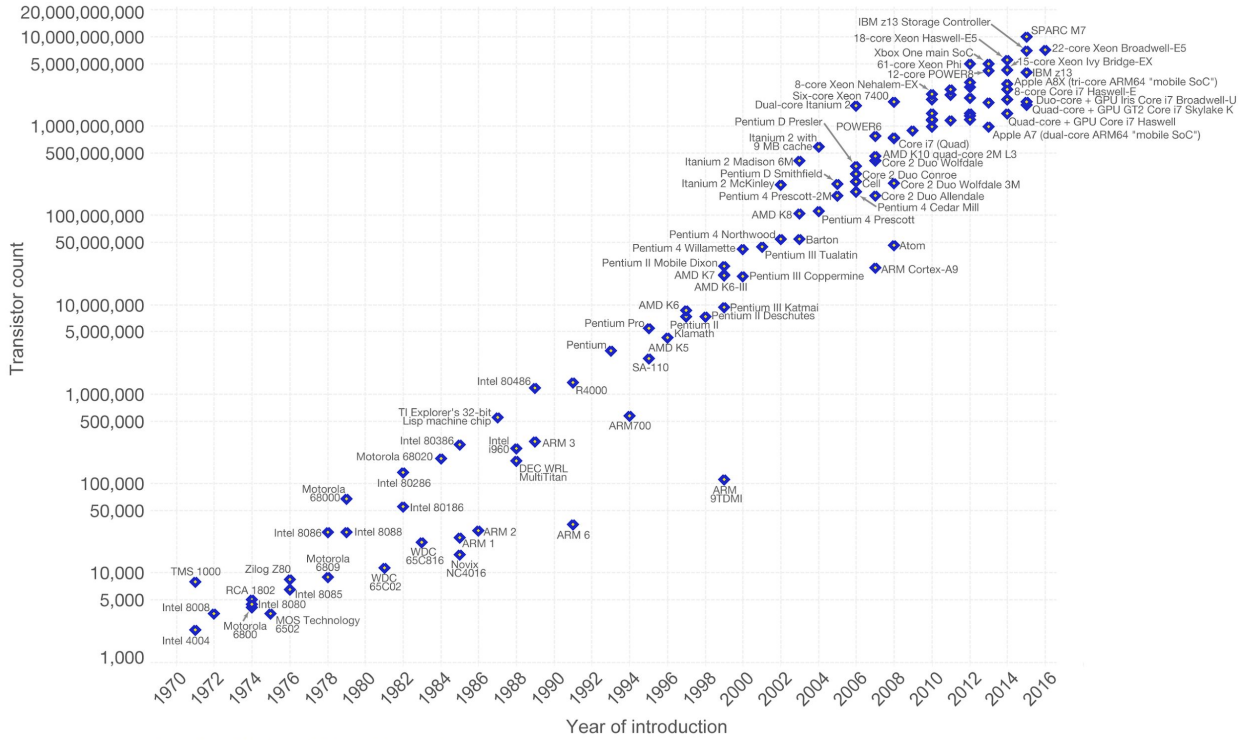
Google Cloud

Google Cloud

Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

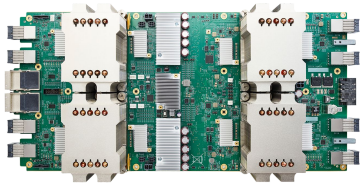
Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore's law.

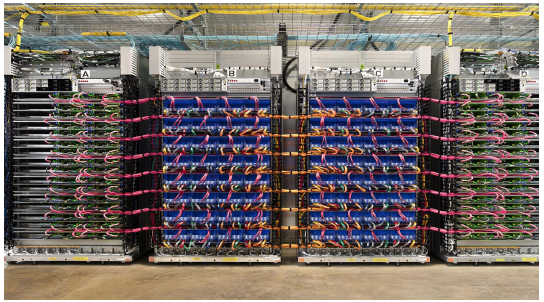


Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

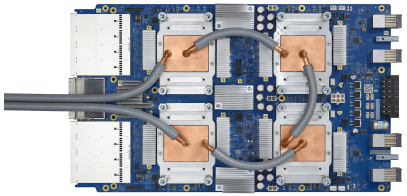
Expanding the AI frontier of performance



Cloud TPU v2
180 teraflops
64 GB HBM
Training and inference



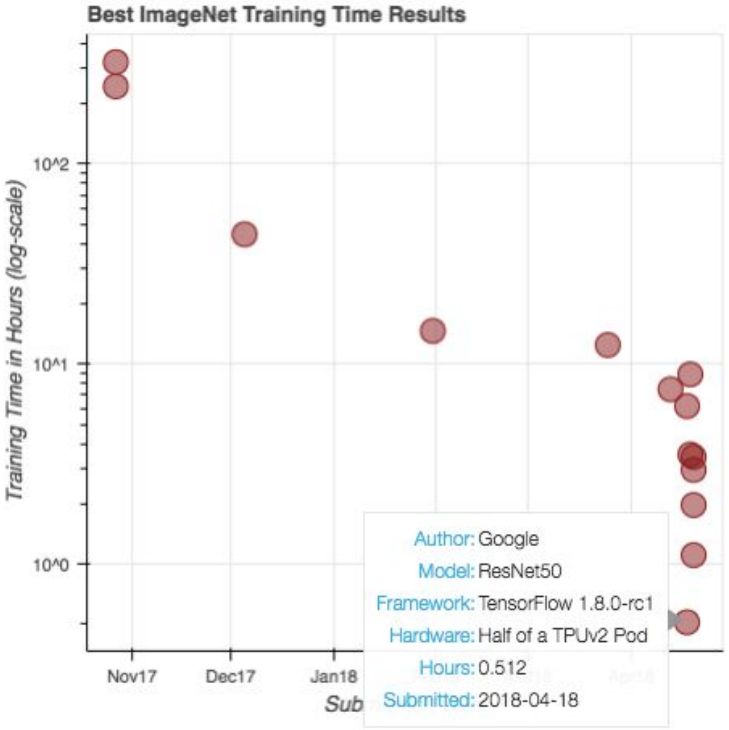
Cloud TPU v2 Pod[ALPHA]
11.5 petaflops
4 TB HBM
2-D toroidal mesh network
Training and inference



Cloud TPU v3[ALPHA]
420 teraflops
128 GB HBM
Training and inference

Google Cloud

Google Cloud



STANFORD DAWN

Training Cost in USD (log-scale)

	TPU Entry	Best non TPU Entry
Cost	25$	72$
Time	24 Minutes	3 Hours

** FAST.AI uses progressive scaling of training images and aggressive learning rate schedule

** This is a V2 POD!

Cloud TPU Performance

#1 DAWNBench Entry, ImageNet Training Time:

ResNet-50 on TPUv2 half-pod

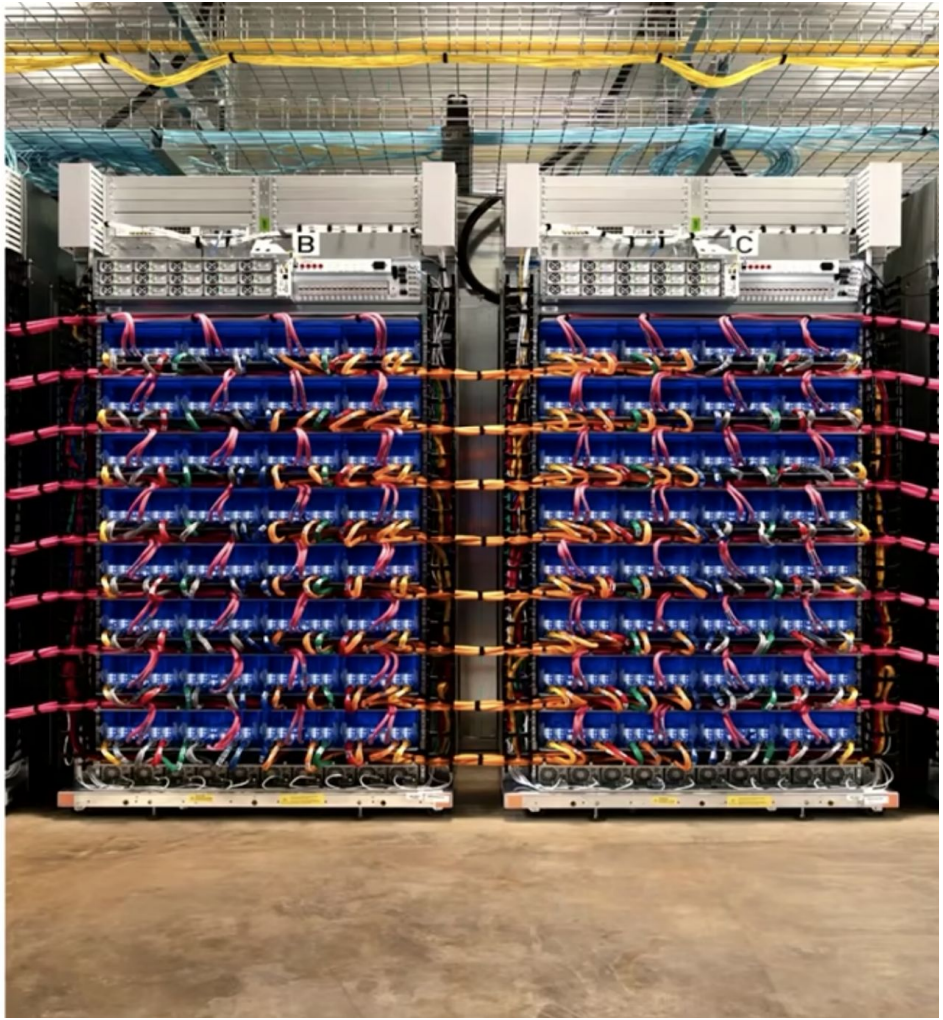
Real data: **77,392** images/sec

Final accuracy: **93%**

Training time: **30 min**

(23.9 min without checkpoints!)

g.co/cloudtpu



Harvard University

Stanford University

University of Arkansas, Littlerock

University of California, Berkeley

University of Illinois, Urbana Champaign

University of Minnesota

University of Texas, Austin

University of Toronto

Alibaba

AMD

Arm

Baidu

Cadence

Cerebras

Cisco

Cray

Dividiti

Enflame Tech

Esperanto

Facebook

Google

Groq

Habana

Huawei

Iluvatar

Intel

In-Q-Tel

MediaTek

Mentor Graphics

Microsoft

Myrtle

Mythic

NetApp

NVIDIA

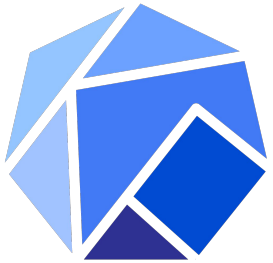
One Convergence

Qualcomm

MLPerf

A broad ML benchmark suite for measuring performance of ML software frameworks, ML hardware accelerators, and ML cloud platforms.

Free and Open source tools

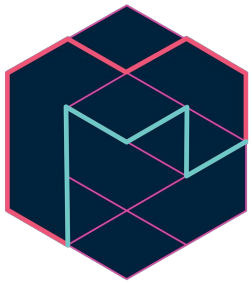


Kubeflow

TensorFlow.js

Google Dataset Search Beta

colab



Google Cloud

Get there faster with Cloud AI

1

Pre-trained Google models

Pre-trained models
Machine learning APIs

Natural chatbot interactions
DialogFlow & Contact Center AI

2

The middle road

Train custom machine learning models
AutoML

ML with SQL
BQML

3

Custom model

Support for custom ML Models
Cloud AI Platform

Hardware optimised for your problem
Cloud TPU, GPU, CPU

Any ML framework
Cloud ML Engine

Google Cloud

Comprehensive set of AI building blocks

Sight

- Cloud Vision
- Cloud Video Intelligence
- New AutoML Vision

Language

- Cloud Translation
- Cloud Natural Language
- New AutoML Translation
- New AutoML Natural Language

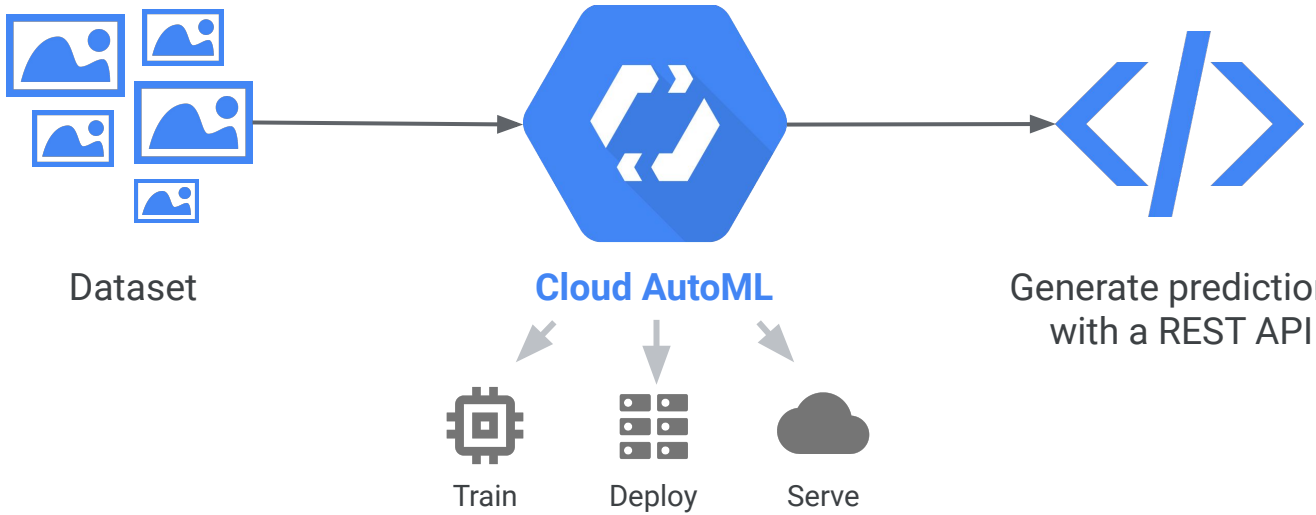
Conversation

- Dialogflow Enterprise Edition
- Cloud Text-to-Speech
- Cloud Speech-to-Text

Google Cloud

Cloud AutoML

ML that creates ML for your problem



Google Cloud

DEMO - AutoML Tables

Machine Learning on structured data
at speed and scale

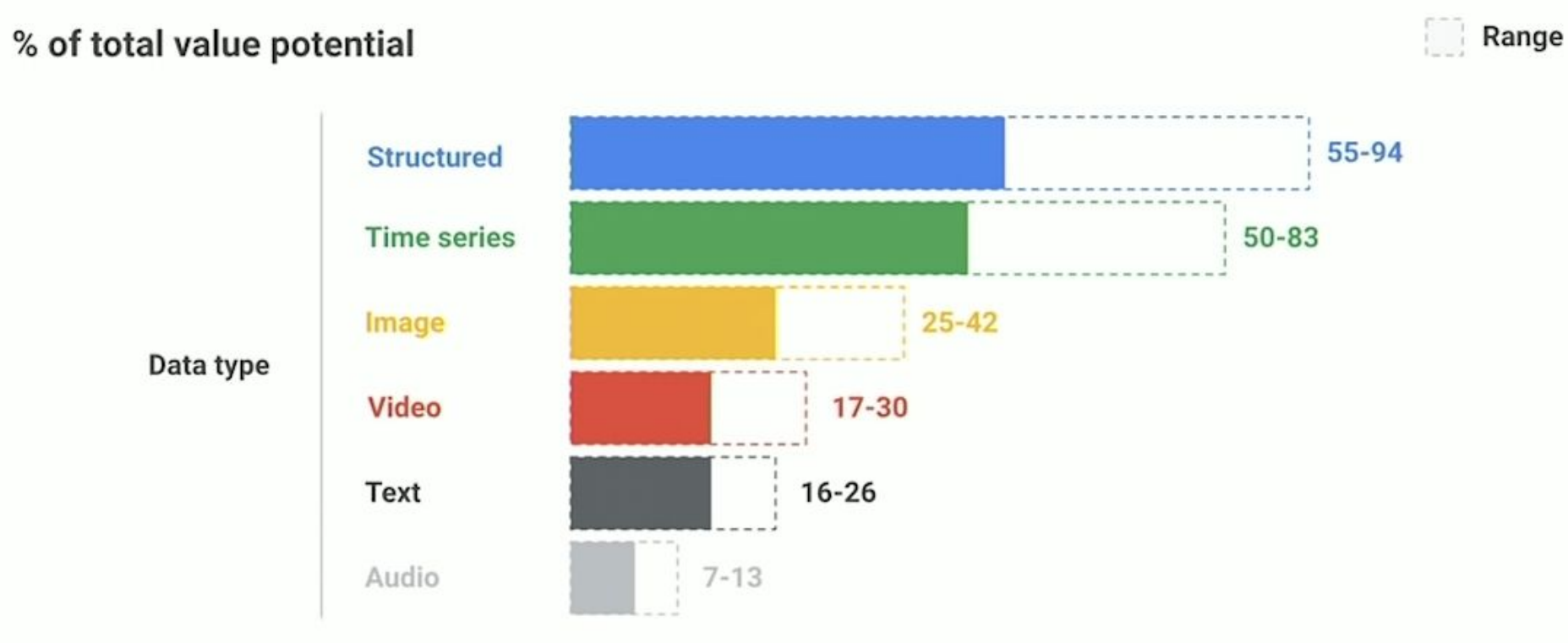
Google Cloud

ML on structured data: 101

Historic offers from marketplace.xyz									
ID	**Geo**	**Domain**	**Posted on:**	**Title**	**Description**	**Category**	**Brand**	**...**	**Price sold:**
104	US	marketA	Feb 1, 2018	"Dark red..."	"Try this soft..."	["A, B, ..."]	Nike	...	$92
204	US	marketB	Jan 20, 2018	"Women's..."	"Medium-size..."	["A, B, ..."]	Adidas	...	$58
302	US	marketA	Jan 12, 2018	"Running..."	"All-terrain..."	["A, B, ..."]	Asics	...	$85
352	EU	marketB	Feb 13, 2018	"Running..."	"All-terrain..."	["A, B, ..."]	Puma	...	?

Target column

This is what matters



Google Cloud

Mission critical use cases across industries

Optimize your product inventory

- Likelihood of stockouts
- Price elasticity

Manage your risk

- Risk of large claims or defaults
- Likelihood of fraud

Understand your customer

- Purchase frequency
- Likelihood to churn
- Lead conversion
- Lifetime value
- Campaign attribution

Maximize your resources

- Incidence of breakdowns
- Drivery supply & demand

Improve patient outcomes

- Hospital asset utilization
- Likelihood of patient recovery
- Liability claims

Google Cloud

A combinatorial explosion in things to worry about

Data preparation

Properly handling:

- Imbalanced data
- Outliers
- Missing values
- High cardinality features
- Highly correlated features
- Target leakage
- Inconsistent feature definition
- Data that doesn't fit local memory
- ...

Feature engineering

Selecting right preprocessing for:

- Numbers
- Classes
- Strings
- Dates
- Lists
- Nested fields
- ...

Multiple options per column, 100s of columns in table

Architecture selection

Selecting the best model architecture from dozens available

- Linear
- Feed forward
- Random forest
- Decision tree
- Residual nets
- ...

Keeping up with the onslaught of newest state of the art

Parameter selection

For each architecture, selecting the right values for each hyperparameter

- Learning rate
- Regularization
- Layers
- Hidden nodes
- Activation fxn
- ...

Potentially more than a dozen values to set

Tuning strategy

Choose efficiently from O(1000s) of combinations.

Selecting right strategy for ensembling

- Simple average
- Boosting
- Bagging
- ...

Model evaluation

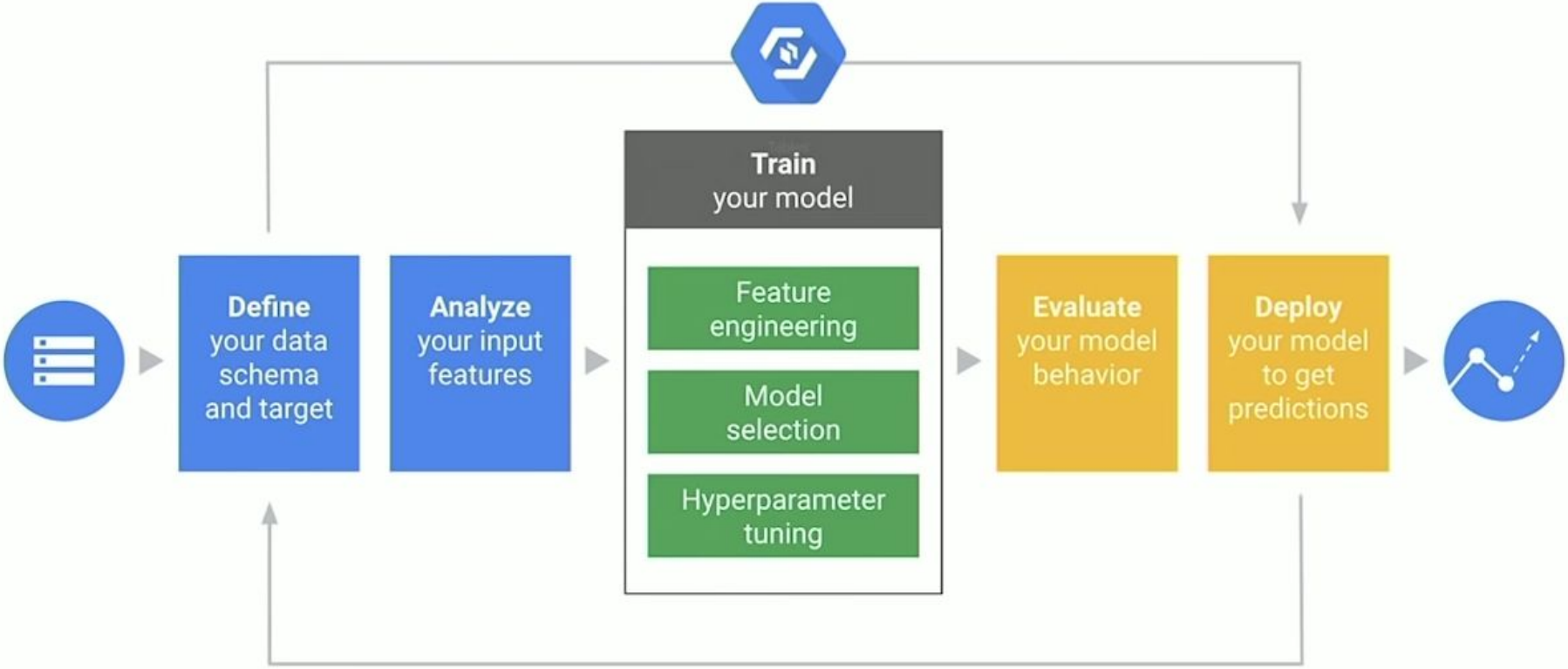
Evaluating model at

- Dataset-level
- Feature-level
- Prediction-level

Ensuring behavior is fully understood before deployment

Rinse & repeat up to 10s of times per use case

Guide users through the end-to-end ML lifecycle



Handle data as found in the wild

Automated feature engineering for:

- Numbers
- Timestamps
- Classes
- Lists
- Strings
- Nested fields

Resilient to + guardrails for:

- Imbalanced data
- Highly correlated features
- Missing values
- High cardinality features (like IDs)
- Outliers

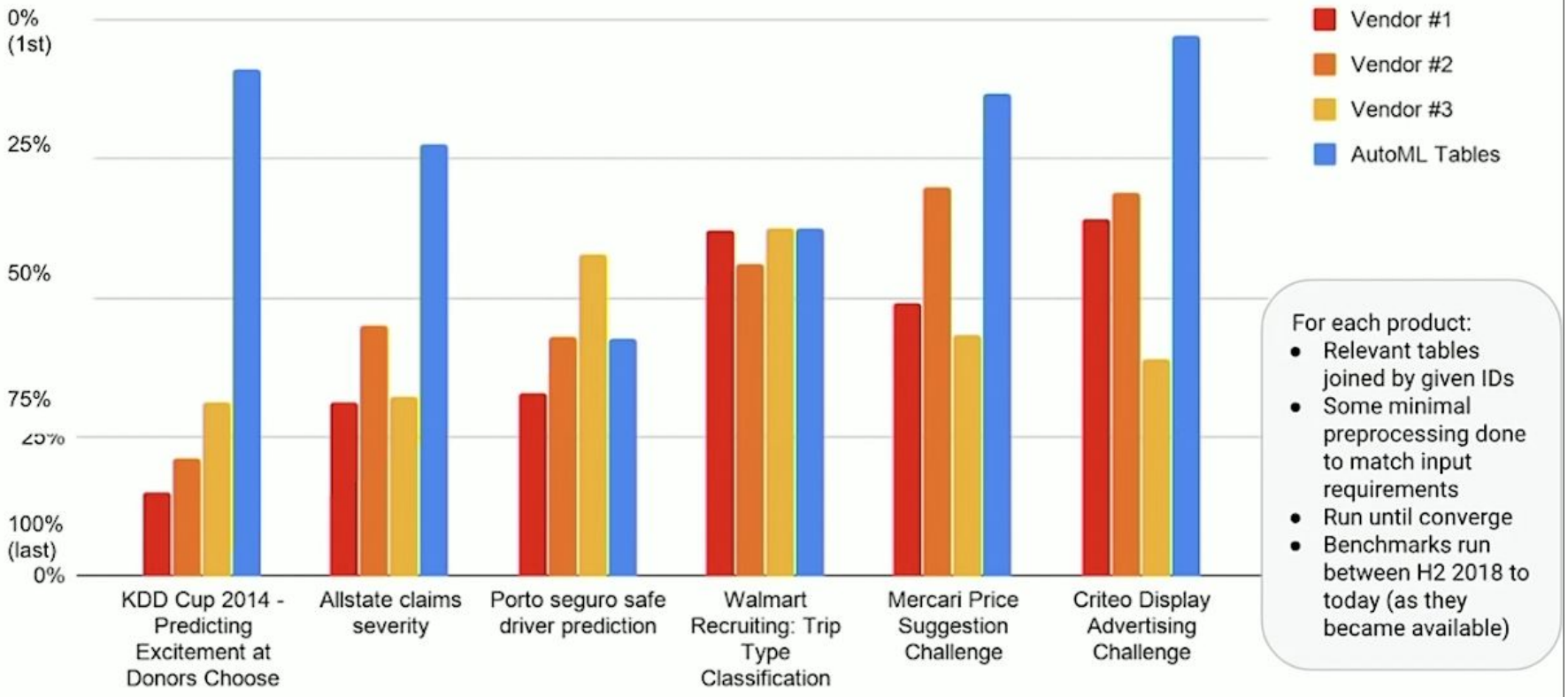
Search the Model Zoo

- Linear or Logistic regression
- Feedforward DNN
- Wide and Deep NN
- Gradient Boosted RF
- DNN + GBRF Hybrid
- Adanet Ensemble
- Neural + Tree Architecture
- … Many more even before publishing

Google Cloud

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)

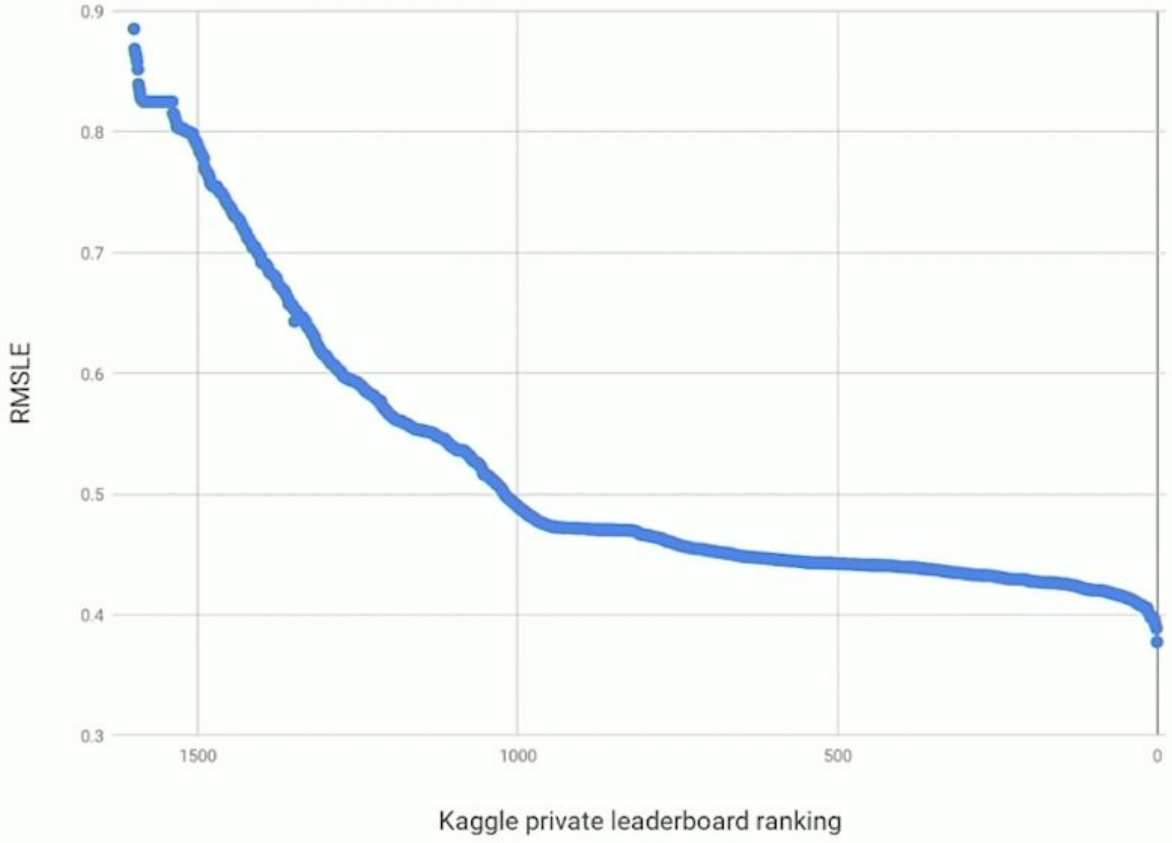
Example: Mercari Price Suggestion Challenge

Goal: Automatically suggest product prices to online sellers

Training data							
ID	**Name**	**Item Condition**	**Categories**	**Brand name**	**Shipping**	**Item description**	**Price**
0	MLB Cincinnati Reds T Shirt Size XL	3	Men, Tops, T-shirts		1	No description yet	**$10**
1	Razer BlackWidow Chroma Keyboard	3	Electronics, Computers & Tablets, Components & Parts	Razer	0	This keyboard is in great condition and works like it came out of the box. All of the ports are tested and work perfectly. The lights are customizable via the Razer Synapse app on your PC.	**$52**
2	AVA-VIV Blouse	1	Women, Tops & Blouses, Blouse	Target	1	Adorable top with a hint of lace and a key hole in the back! The pale pink is a 1X, and I also have a 3X available in white!	**$10**
3	Leather Horse Statues	1	Home, Home Décor, Home Décor Accents		1	New with tags. Leather horses. Retail for [rm] each. Stand about a foot high. They are being sold as a pair. Any questions please ask. Free shipping. Just got out of storage	**$35**

The overall performance curve...

Mercari Price Suggestion Challenge



Google Cloud

Save money

Training
Priced based on compute hours used

Prediction
Priced based on compute hours used

Model deployment
Priced based on gb-hours deployed

AutoML Tables is the best fit when:

1. You know how to create good training data
2. You are willing to wait at least an hour
3. You have a larger, more complicated dataset (for now)

Demo: AutoML Tables

Cloud Machine Learning Engine

TensorFlow K scikit-learn dmlc XGBoost

Fully managed service with integrated notebook experience

Supports multiple frameworks

Training

Design and evaluate model architectures
Train any model at large scale on a managed cluster

Prediction

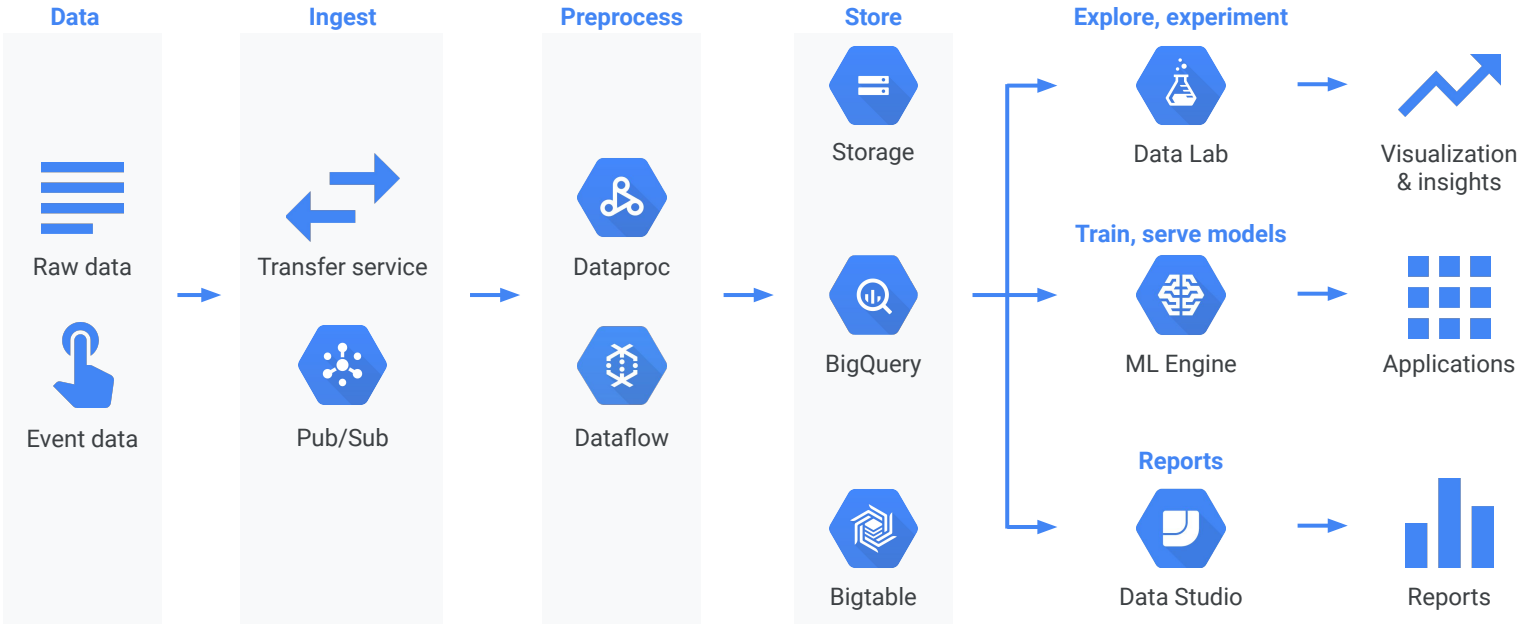
Online *(low latency)* and batch *(high throughput)*

Deploy models into production, no Docker container required

Integrated with RStudio

Google Cloud

Putting it all together



Google Cloud

Cloud AI

Cloud AI solutions

- Cloud Job Discovery
- Contact Center (New)
- Document understanding

ML professional services & partners

- ASL
- Professional services organization

Cloud AI building blocks

Sight

- Cloud Video Intelligence
- AutoML Vision
- Cloud Vision

Language

- Cloud Natural Language
- AutoML NL (New)
- Cloud Translation
- AutoML Translation (New)

Conversation

- Cloud Speech-to-Text
- Dialogflow Enterprise
- Cloud Text-to-Speech

Cloud AI Platform

Machine and deep learning

- Cloud ML Engine
- Cloud Dataflow
- Cloud Dataproc

ML accelerators

- Cloud GPU
- Cloud TPU

ML libraries

- Tensorflow
- Kubeflow
- Spark
- beam
- R
- TORCH
- K
- Spark MLlib
- scikit-learn

Kaggle / datasets

- kaggle
- Datasets

Google Cloud

Demo: ML Engine Code

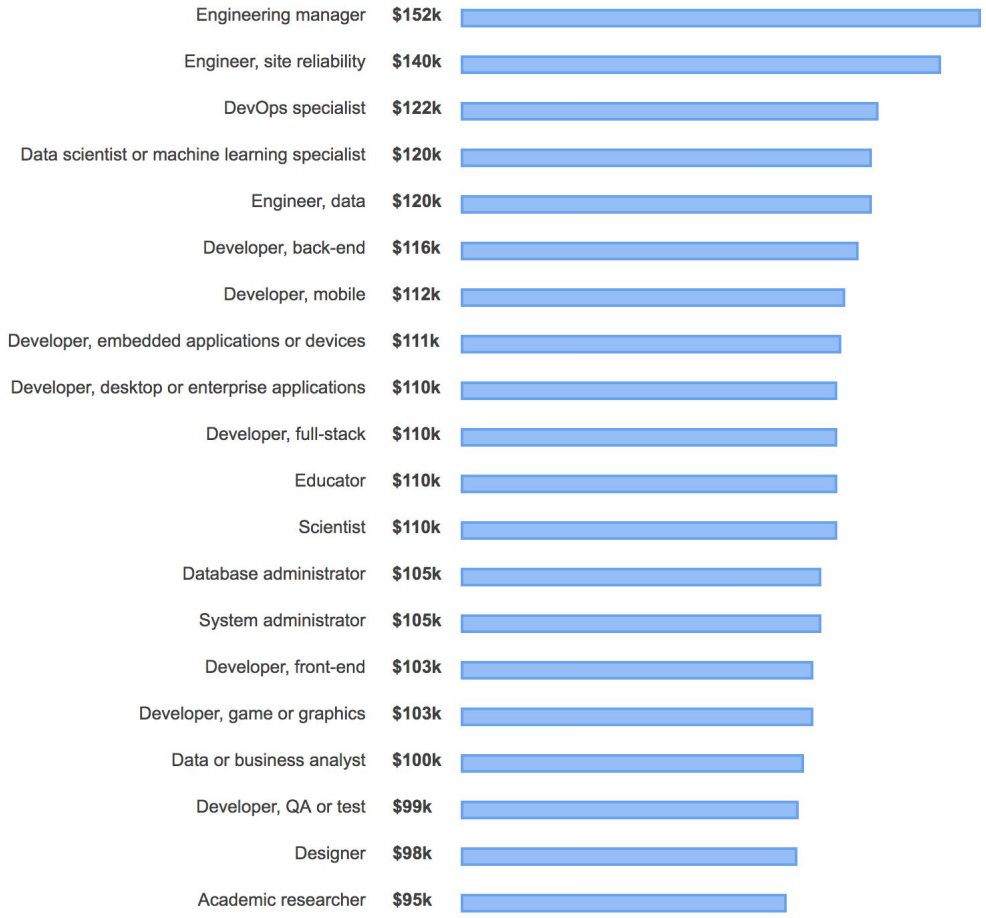
Examples
Codealong

Google Cloud

Stackoverflow Developer Insights Survey

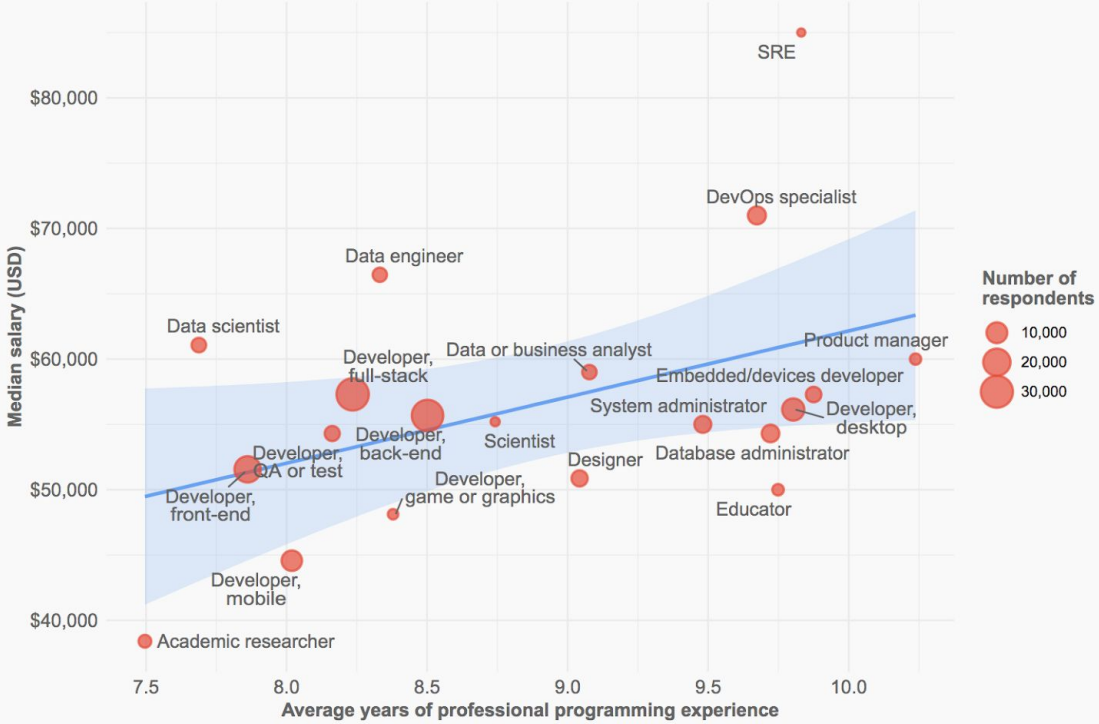
Google Cloud

Global | United States



Median of 14,661 responses; USD

Google Cloud



WORK

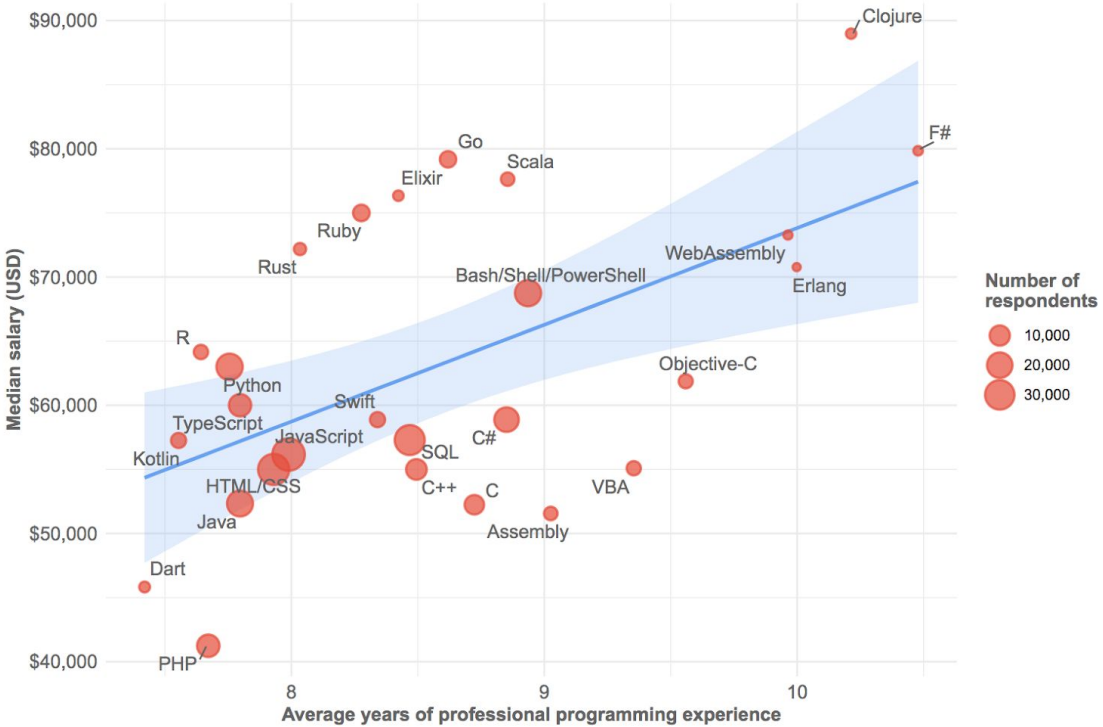
Salary and Experience by Developer Type

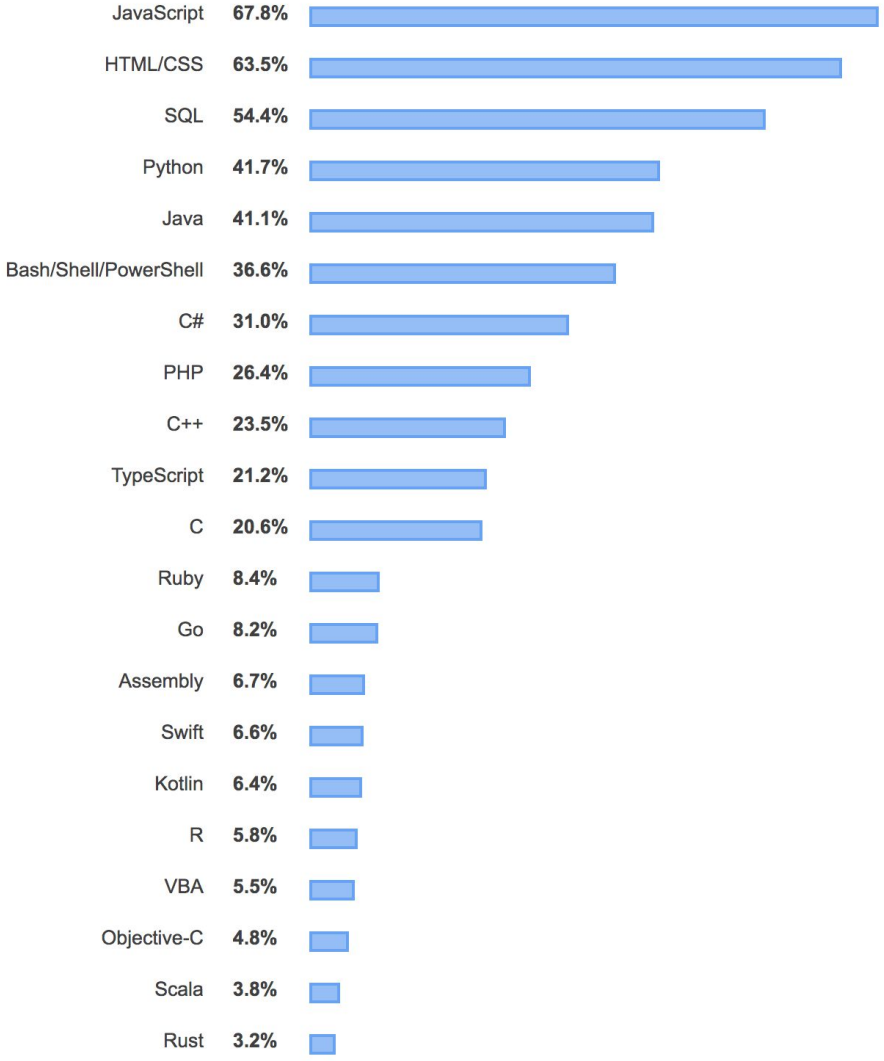
This plot shows global median salary and years of programming experience. Naturally, developers with more years of experience are paid more. However, **some types of coding work are paid more highly at the same level of experience.** Developers who work with data (data scientists and engineers) and those who work in DevOps and site reliability are high earners for their level of experience, while academic researchers and educators are paid less at their experience levels.

WORK

Salary and Experience by Language

Developers using languages that appear above the line in this chart, such as Clojure, Scala, Go, Rust, and R, are being paid more even given how much experience they have. Developers using languages below the line, like PHP, Assembly, and VBA, however, are paid less even given years of experience. The size of the circles in this chart represents how many developers are using that language compared to the others.



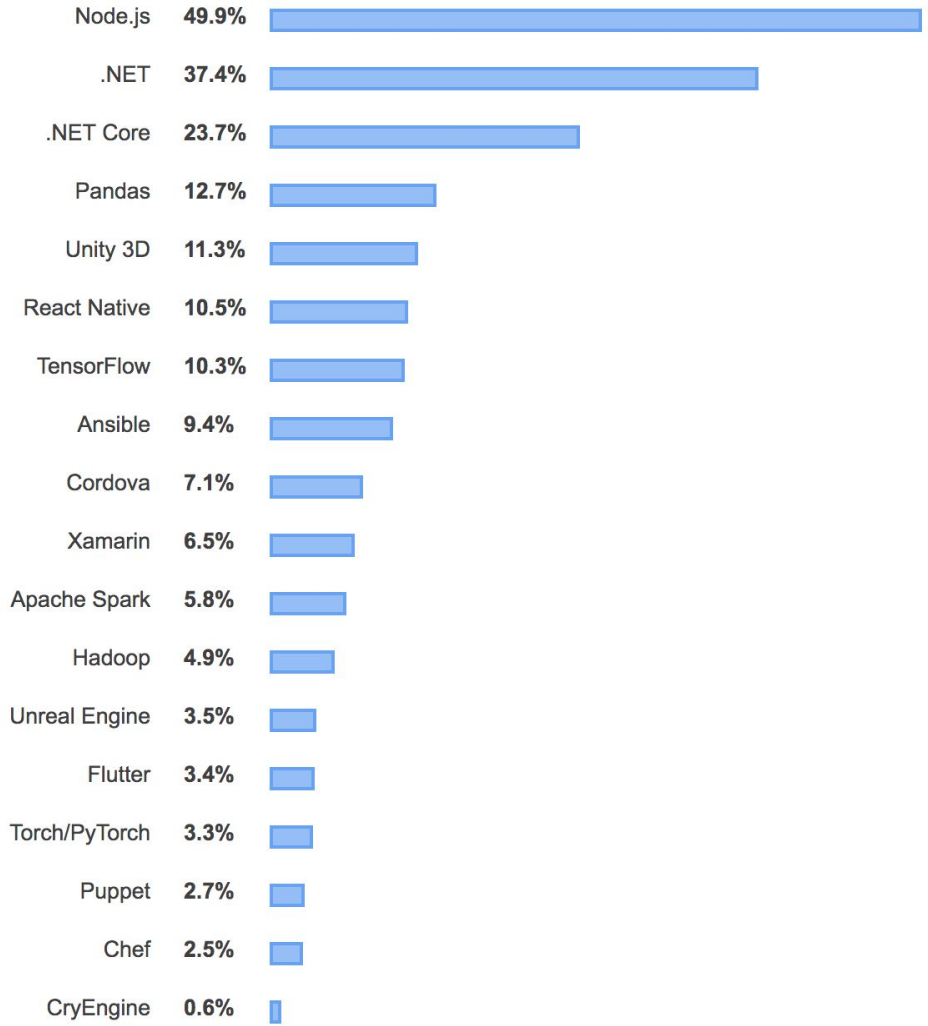


Google Cloud

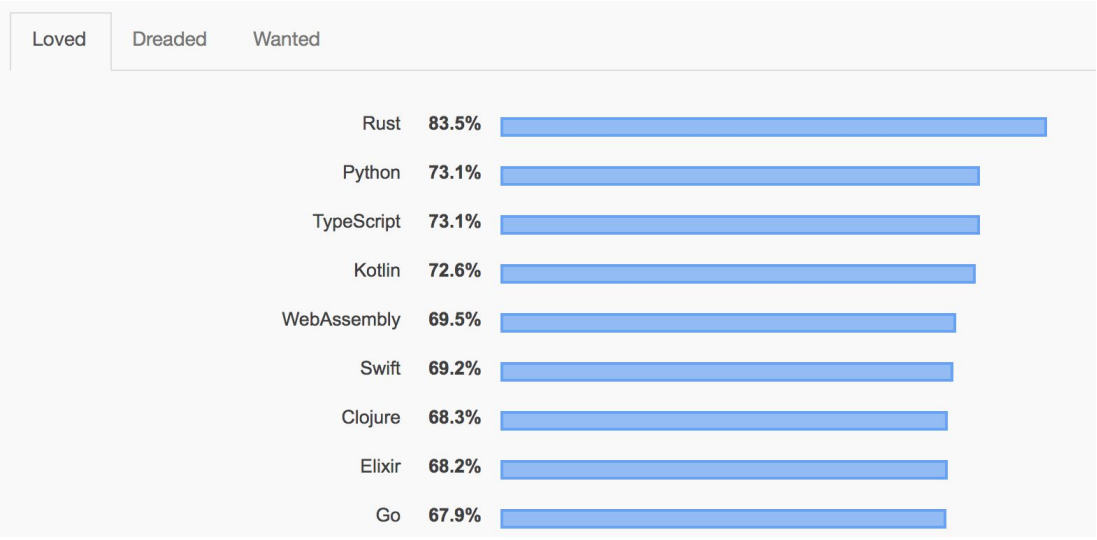
TECHNOLOGY

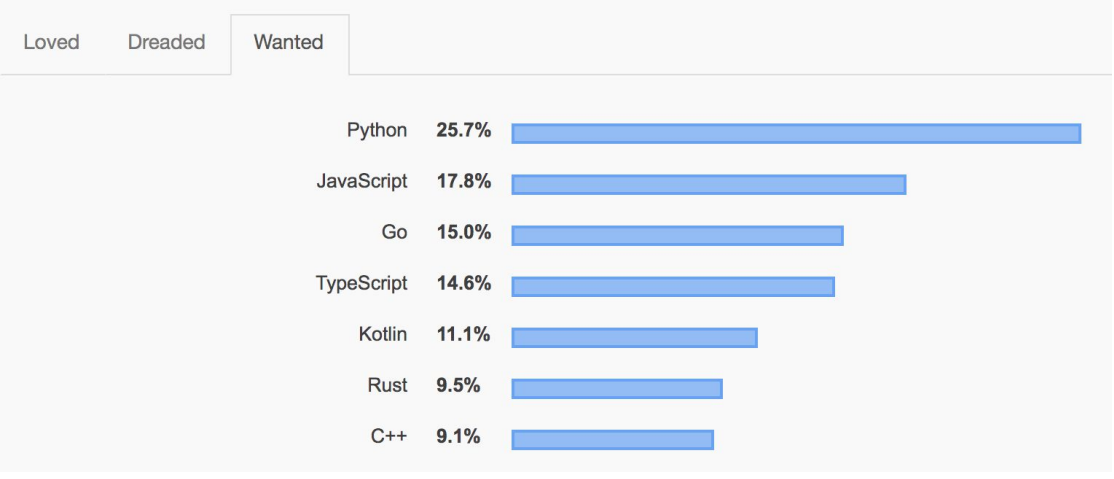
Most Popular Other Frameworks, Libraries, and Tools

This is the first year we've asked about many of the technologies in this category, and Node.js is the most commonly used. More developers say they use .NET than .NET Core, and **the deep learning framework TensorFlow is many times more popular than the deep learning framework Torch/PyTorch.**



Google Cloud





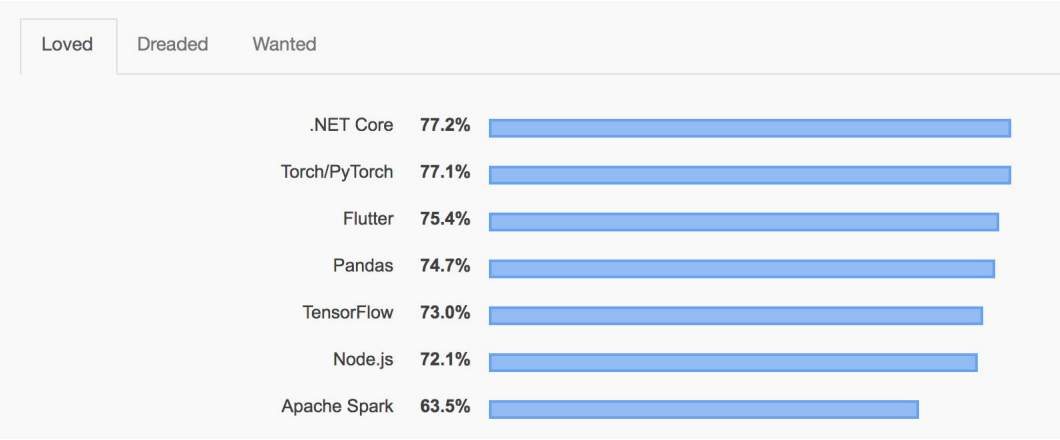
TECHNOLOGY

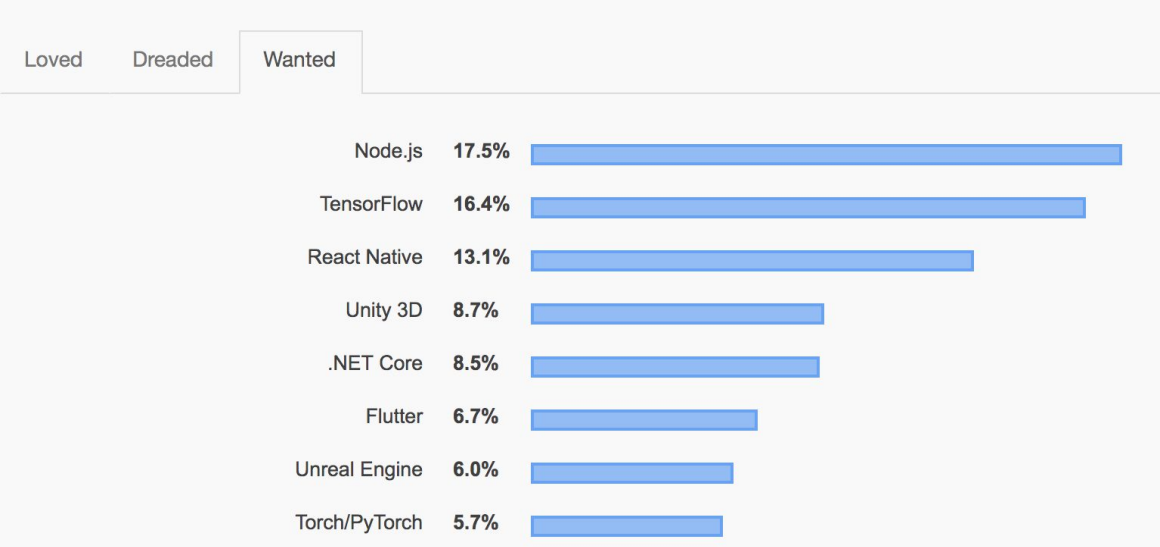
Most Loved, Dreaded, and Wanted Languages

For the fourth year in a row, **Rust is the most loved programming language among our respondents**, followed close behind by Python, the fastest-growing major language today. This means that proportionally, more developers want to continue working with these than other languages.

VBA and Objective-C rank as the most dreaded languages this year. Most dreaded means that a high percentage of developers who are currently using these technologies express no interest in continuing to do so.

Python is the most wanted language for the third year in a row, meaning that developers who do not yet use it say they want to learn it.

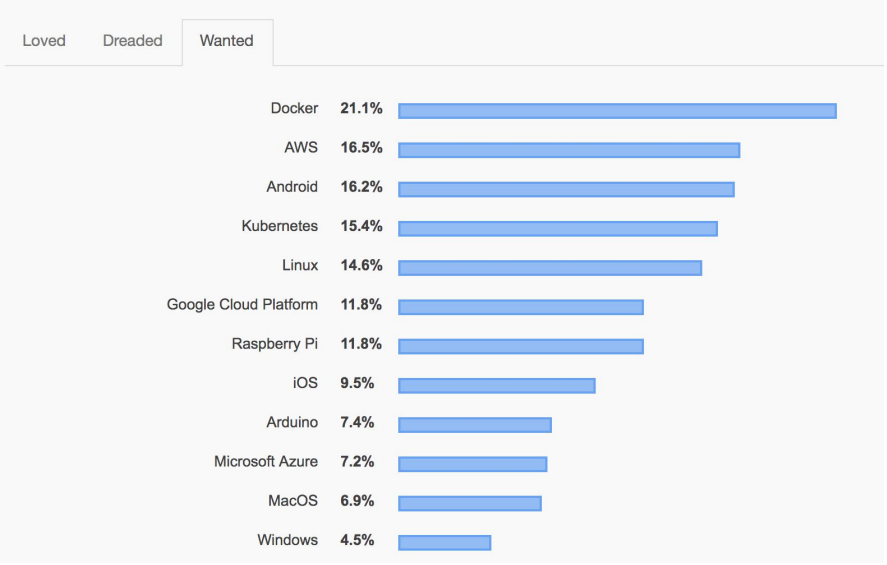


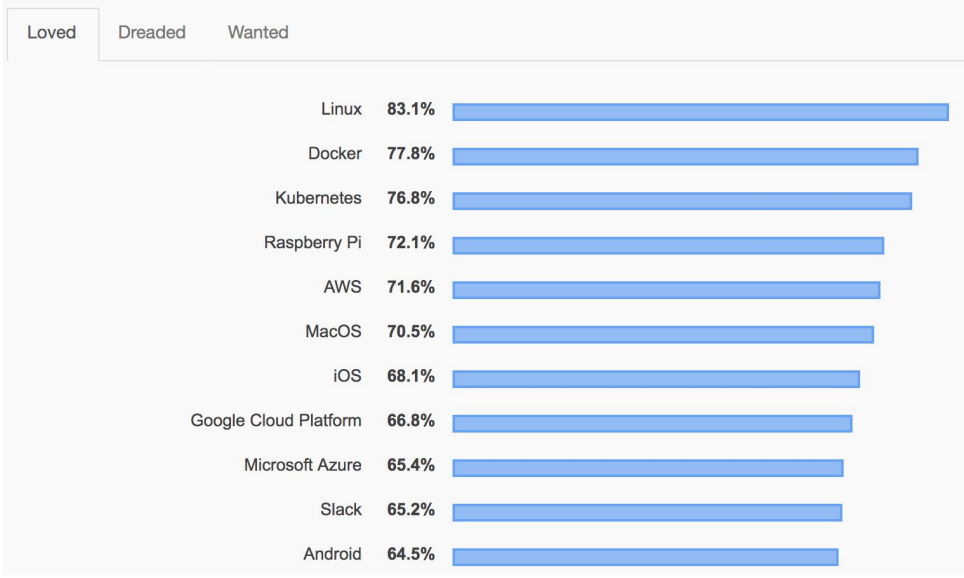


TECHNOLOGY

Most Loved, Dreaded, and Wanted Other Frameworks, Libraries, and Tools

.NET Core and Torch/PyTorch are both used less than other counterparts in their respective ecosystems (.NET and Tensorflow, respectively) but are loved by developers more. Chef and Cordova rank as the most dreaded in this category of frameworks, libraries, and tools.





TECHNOLOGY

Most Loved, Dreaded, and Wanted Platforms

Linux is once again the most loved platform for development, with both Docker and Kubernetes also highly loved this year. WordPress is the most dreaded development platform, and many developers say they want to start developing using Docker and AWS.