



Welcome to Cloud OnBoard

Big Data and Machine Learning

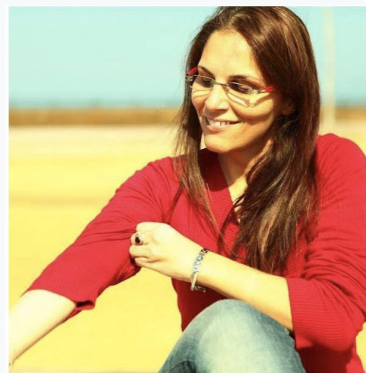
Google Cloud



Welcome



Yoram Ben-Yaacov
Strategic Cloud Engineer



Anat Perly
Strategic Cloud Engineer

An explosion of data



By 2020, some 50 billion smart devices will be connected, along with additional billions of smart sensors, ensuring that the **global supply of data will continue to more than double every two years**”

An explosion of data



... and only about 1% of the data generated
today is actually analyzed”

There is a great demand for data skills

Data Analyst

Analyst

Data Engineer

Data Engineer

Applied ML Engineer

Analyst

Ethicist

Statistician

Social Scientist

Applied ML
Engineer

Researcher

Tech Lead

Analytics
Manager

Decision Maker

Big Data Challenges

Migrating existing
data workloads
(ex: Hadoop, Spark jobs)

Analyzing large
datasets at scale

Building streaming
data pipelines

Applying machine
learning to your data



Agenda

- Intro to Google Cloud Platform infrastructure
- Big data products:
 - Pub/Sub
 - Dataflow
 - BigQuery
- ML products:
 - ML APIs
 - AutoML
 - BigQuery ML

Module 1

Intro to GCP



Agenda

- Intro to Google Cloud Platform infrastructure
- Big data products:
 - Pub/Sub
 - Dataflow
 - BigQuery
- ML products:
 - ML APIs
 - AutoML
 - BigQuery ML

Built on Google infrastructure



This is what makes Google Google: its physical network, its thousands of fiber miles, and those many thousands of servers that, in aggregate, add up to the **mother of all clouds.**"

Wired



Google Cloud

Big Data and ML Products

Compute Power

Storage

Networking

Security



Google Cloud

Big Data and ML Products

Compute Power

Storage

Networking

Security

Machine Learning Models require significant compute resources

Shown: Automatic **Video**

Stabilization for Google Photos

Data sources:

- Image frames (stills from video)
- Phone gyroscope
- Lens motion



A single high-res image represents millions of data points to learn

8 Megapixel resolution

3264 (w)x2448 (h)x3(RGB) = **23,970,816**
data points per image*

* More data = longer model training times + more storage needed



3 “Layers” in depth for Red Blue Green



Google Photos

How many photos are uploaded daily to Google Photos?



Youtube

How many hours of video are uploaded every minute to YouTube?



Google Cloud

Big Data and ML Products

Compute

Storage

Networking

Security



Google Photos

28 billion photos and videos are uploaded to Google Photos **every day**.



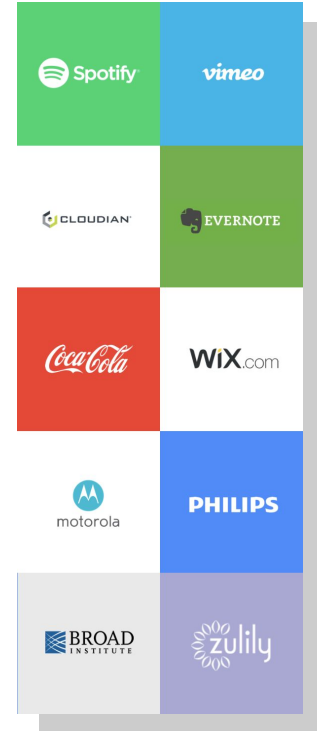
Youtube

Over 1.3PB or **500 hours** of video uploaded **every minute**

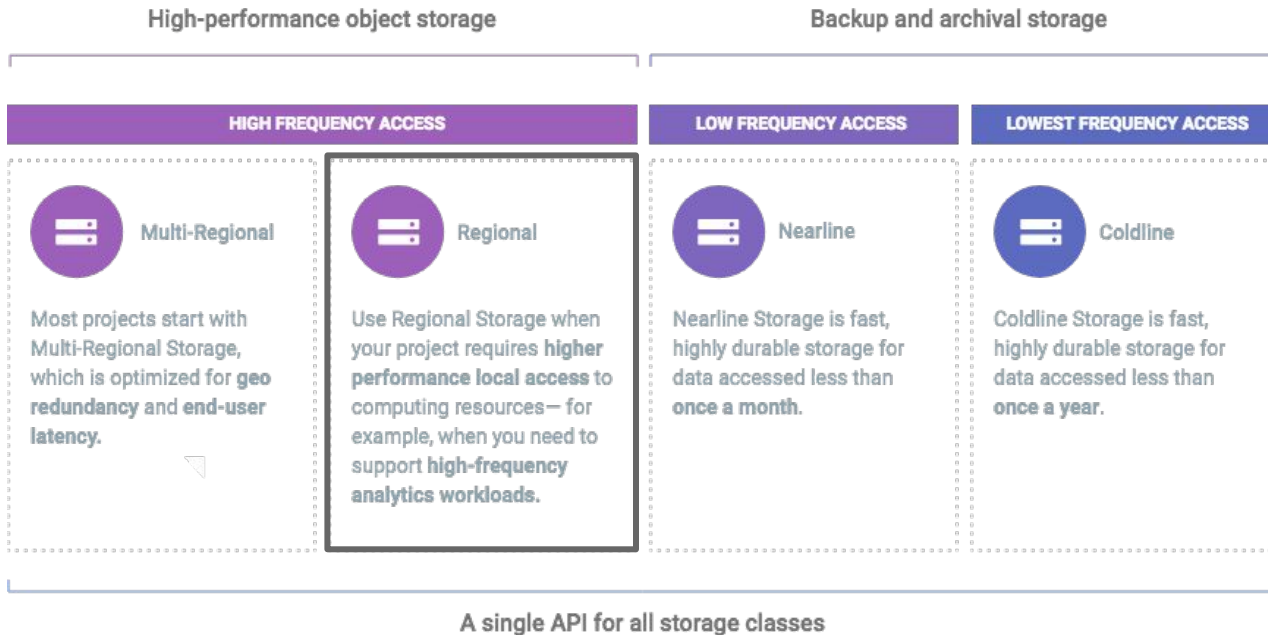
Leverage Google's 99.999999999% durability storage



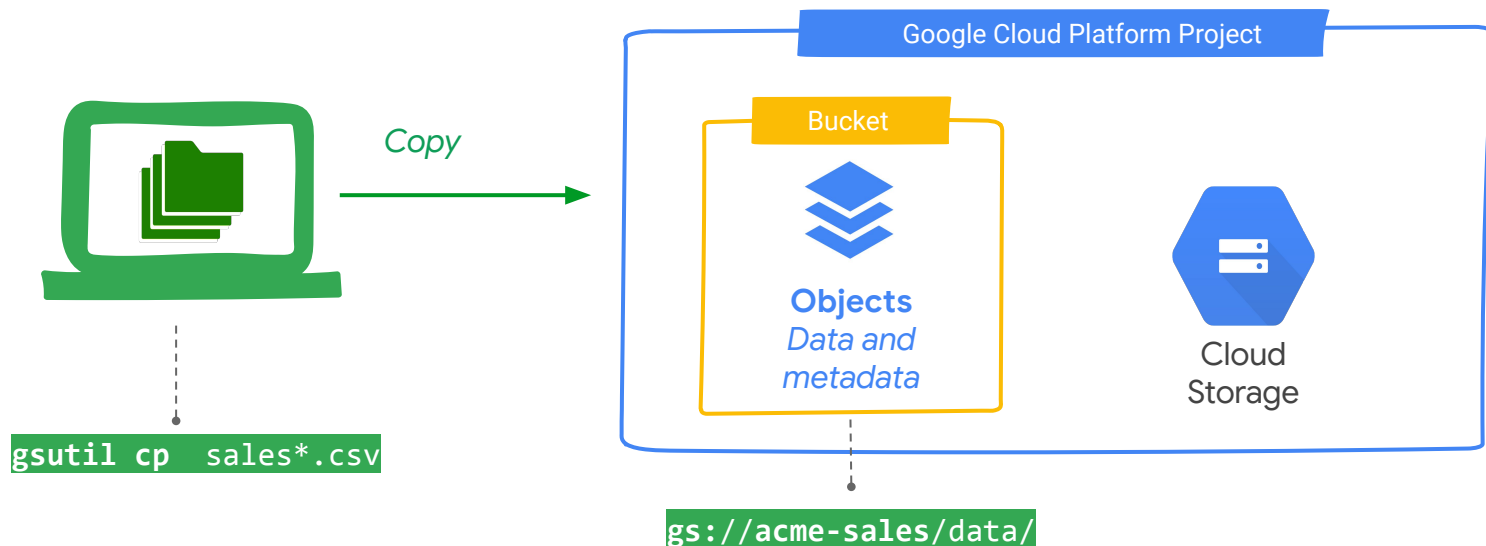
Cloud Storage



Typical big data analytics workloads run in Regional Storage



Got data? Quickly migrate your data to the cloud using gsutil tool





Google Cloud

Big Data and ML Products

Compute

Storage

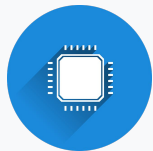
Networking

Security

Google's private network carries as much as 40% of the world's internet traffic every day



Google's data center network speed enables the separation of compute and storage



Servers doing compute tasks don't need to have the data on their disks

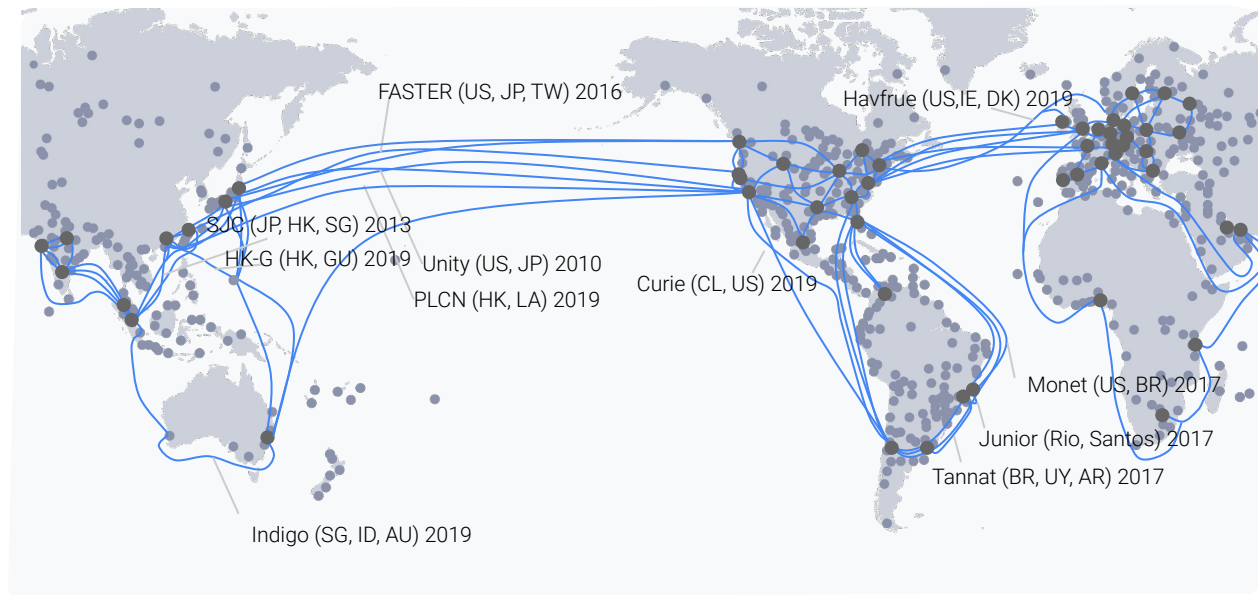
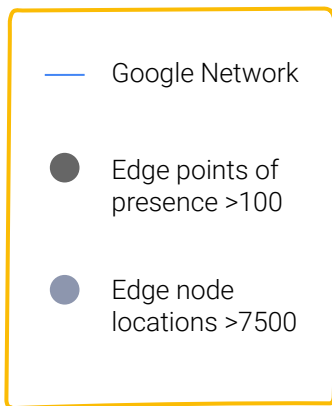


Data can be "shuffled" between compute workers at over 10GBs



1 Petabit/sec of total bisection bandwidth

Google's cable network spans the globe





Google Cloud

Big Data and ML Products

Compute Power

Storage

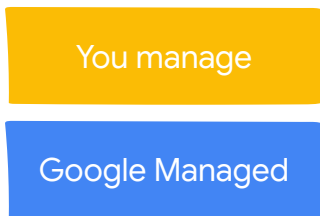
Networking

Security

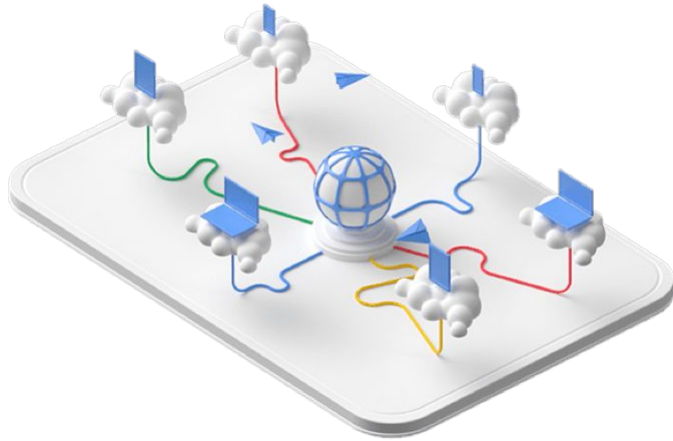
On-premise → you manage all security layers

Responsibility	On-premises
Content	
Access policies	
Usage	
Deployment	
Web application security	
Identity	
Operations	
Access and authentication	
Network security	
OS, data, and content	
Audit logging	
Network	
Storage and encryption	
Hardware	

On-premise → you manage all security layers

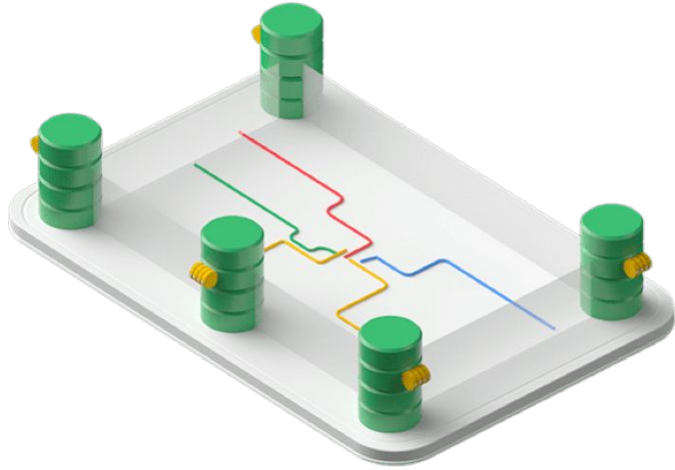


Communications to Google Cloud are encrypted in transit



- In-transit encryption
- Multiple layers of security
- Backed by Google security teams 24/7

Stored data is encrypted at rest and distributed



- Data automatically encrypted at rest
- Distributed for availability and reliability

Module 2

Big data products



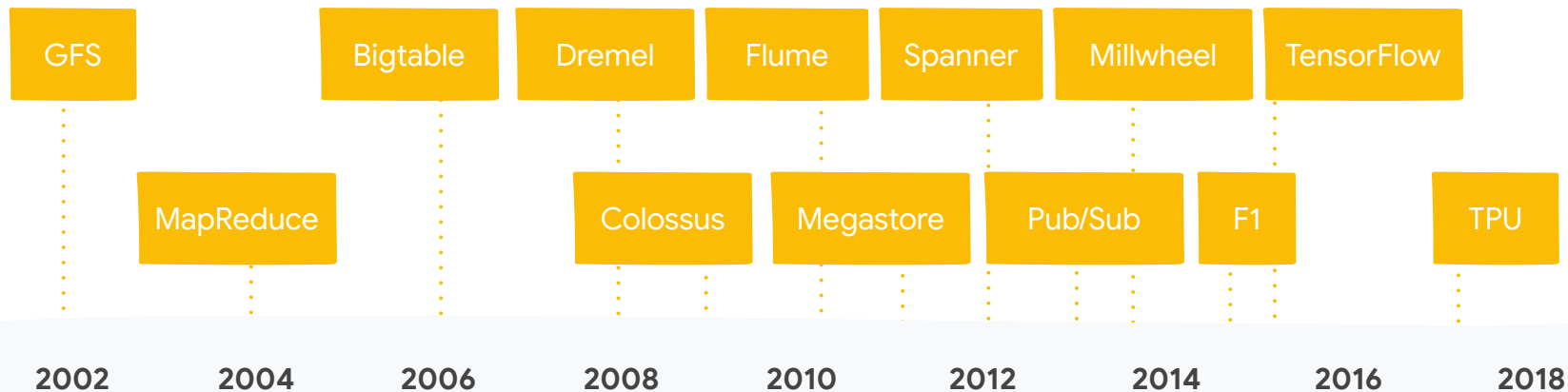
Agenda

- Intro to Google Cloud Platform infrastructure

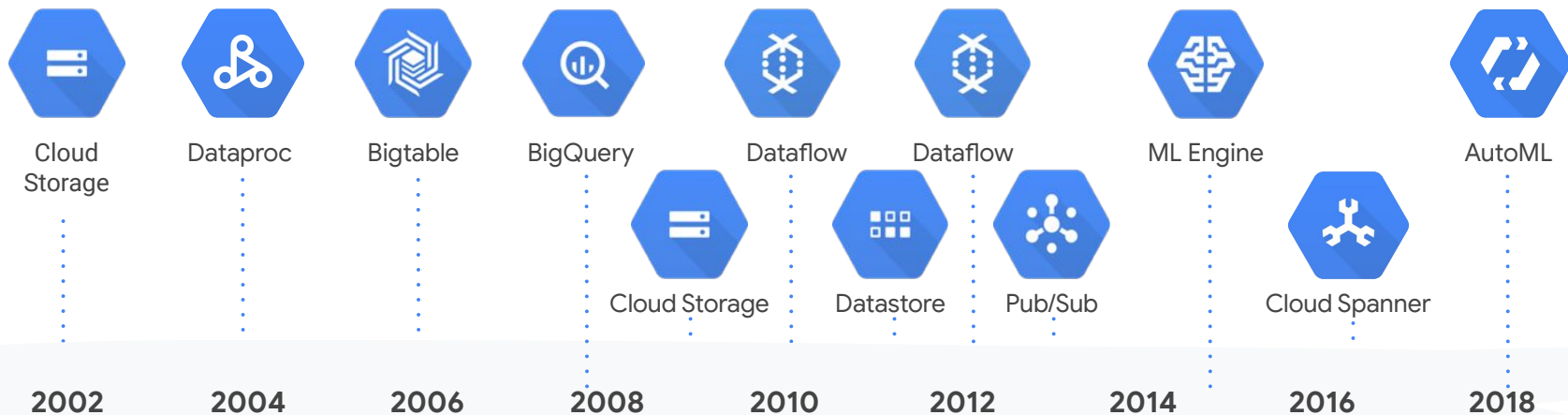
- Big data products:
 - Pub/Sub
 - Dataflow
 - BigQuery

- ML products:
 - ML APIs
 - AutoML
 - BigQuery ML

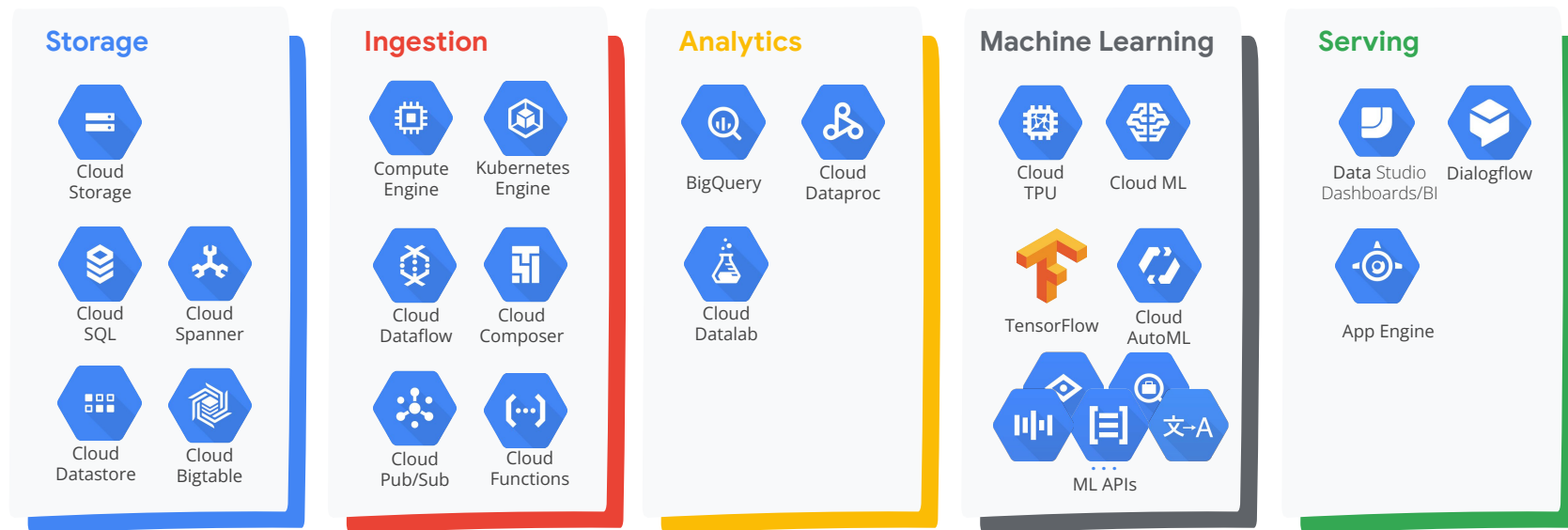
Google invented new data processing methods as it grew



Google Cloud opens up that innovation and infrastructure to you

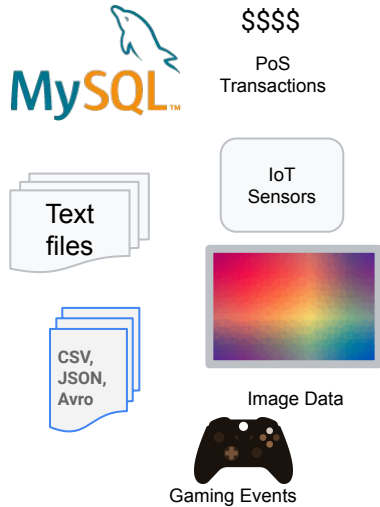


The suite of big data products on Google Cloud Platform



Modern big data pipelines face many challenges

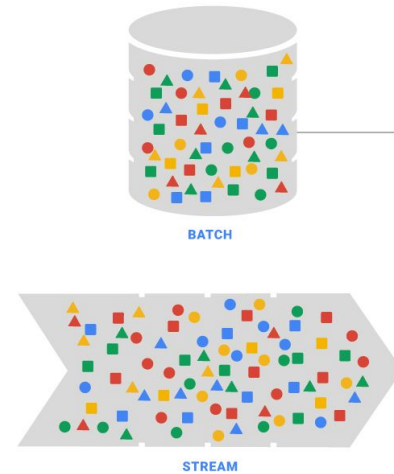
Variety



Volume



Velocity



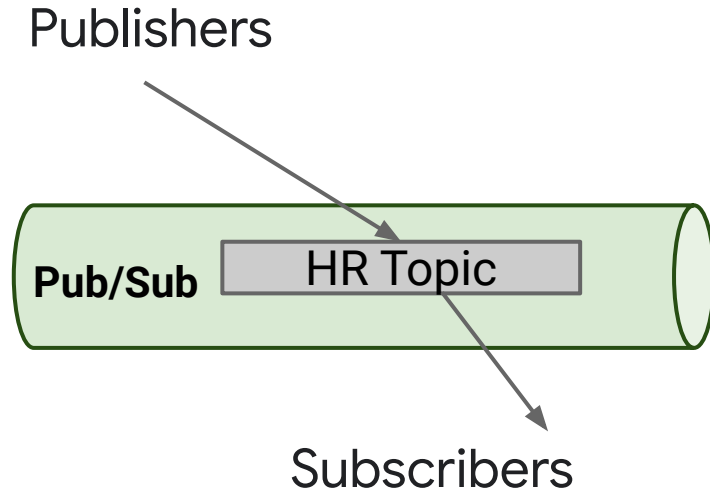
Cloud Pub/Sub offers reliable, real-time messaging

Distributed Messaging with Cloud Pub/Sub



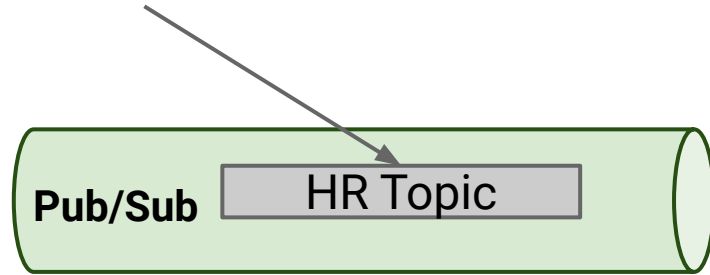
- At-least-once delivery
- Exactly-once processing
- No provisioning, auto-everything
- Open APIs
- Global by default

Pub/Sub topics are like radio antennas

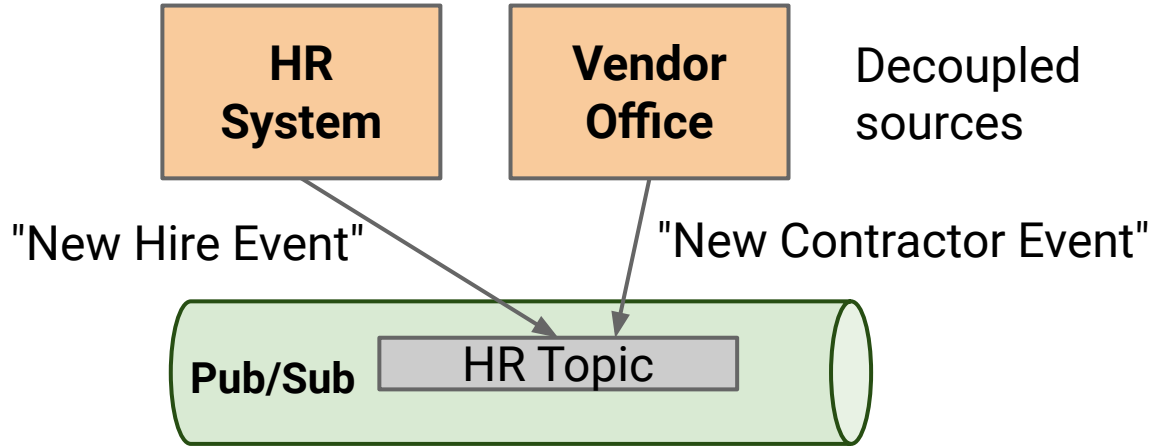


Scenario: HR messaging system

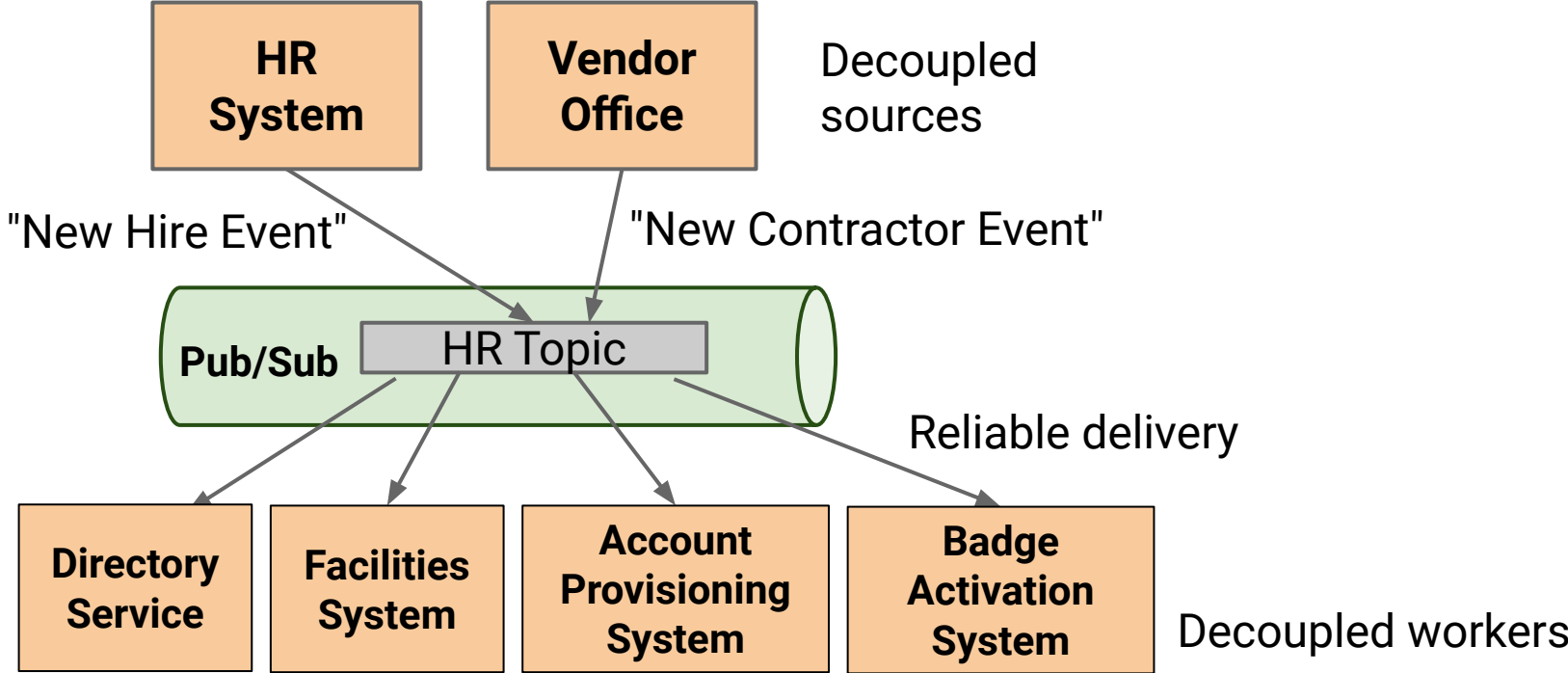
"New Hire Event"



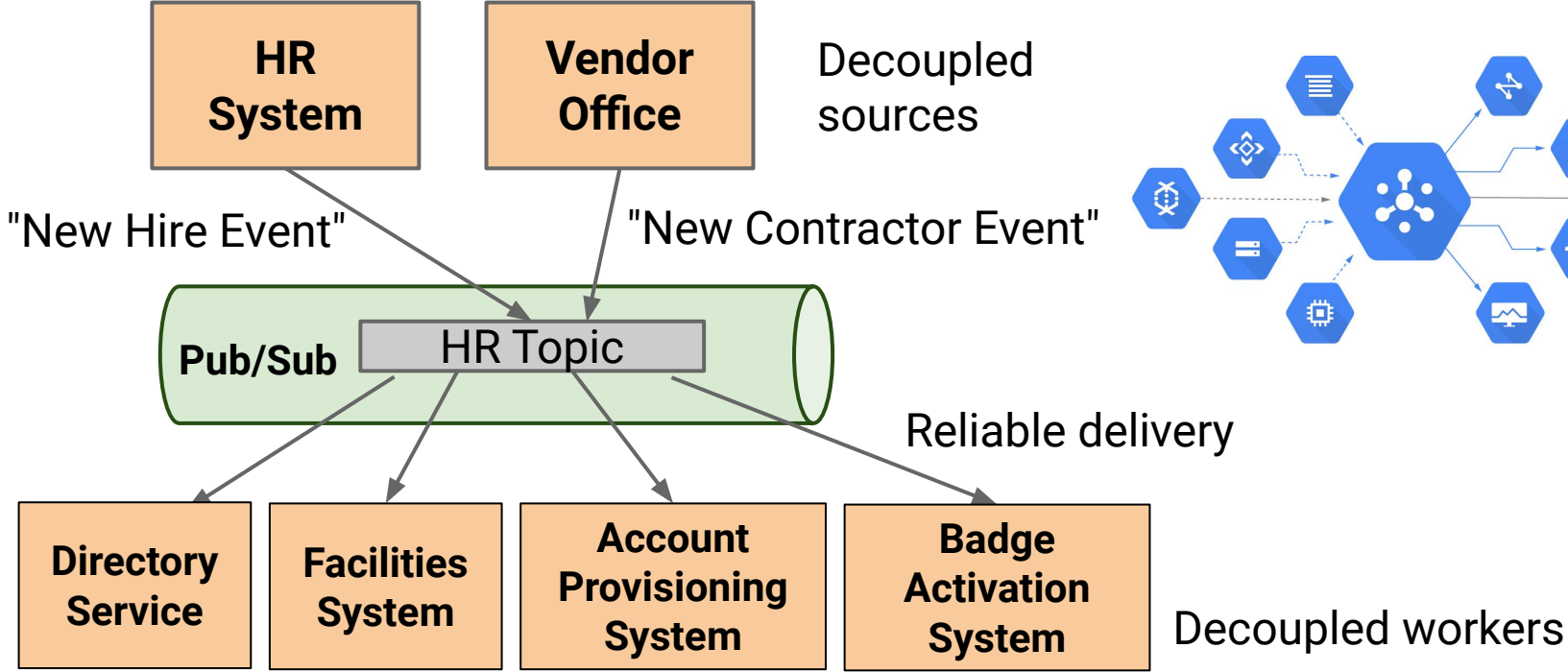
Scenario: HR messaging system



Scenario: HR messaging system




Scenario: HR messaging system



Cloud Dataflow



- Serverless, fully managed data processing
- Unified batch and streaming processing + autoscale
- Open source programming model using  beam
- Intelligently scales to millions of QPS

Data Engineers need to solve two distinct problems

Pipeline design



Implementation



Data Engineers need to solve two distinct problems

Pipeline design with Apache Beam



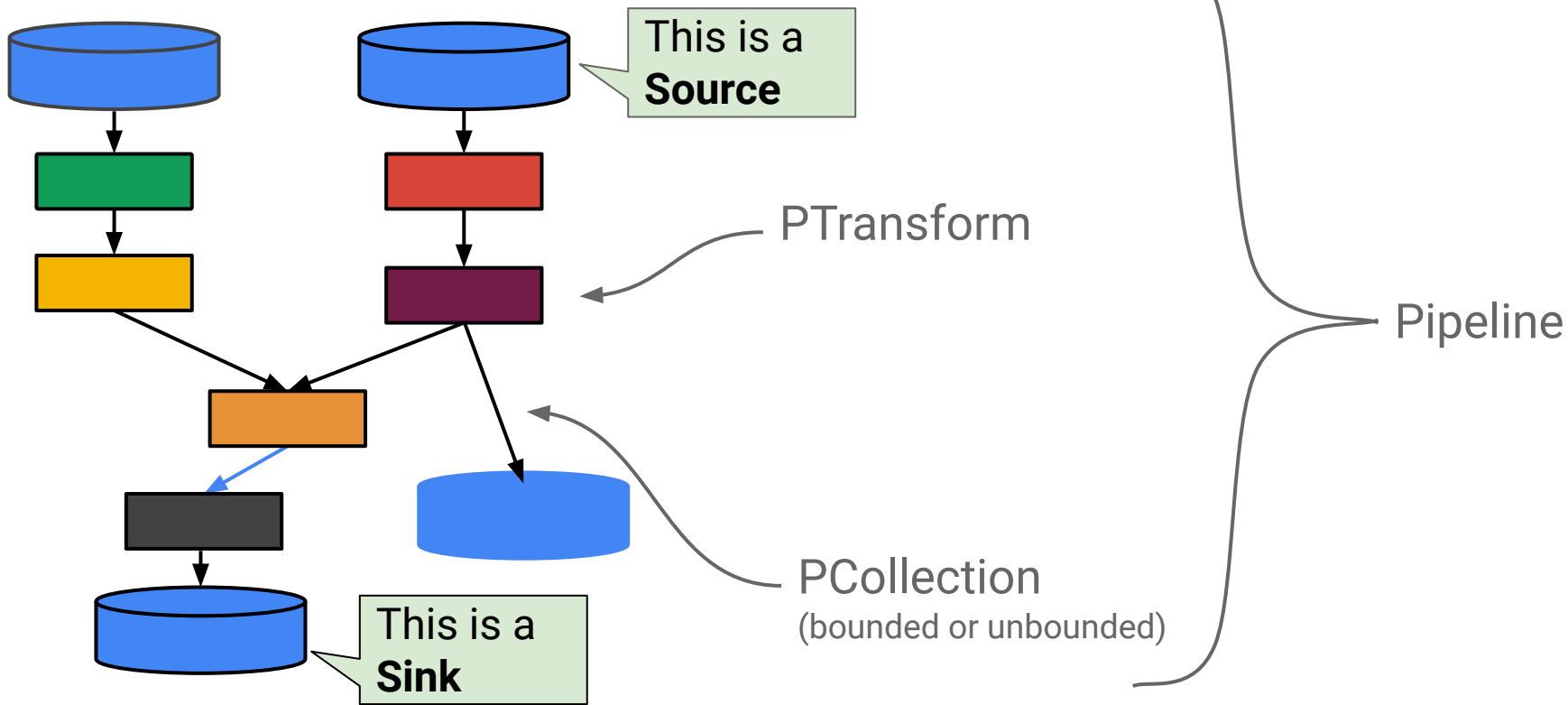
- Will my code work with both batch and streaming data? Yes
- Does the SDK support the transformations I need to do? Likely
- Are there existing solutions? Choose from templates

Start with provided templates and build from there:

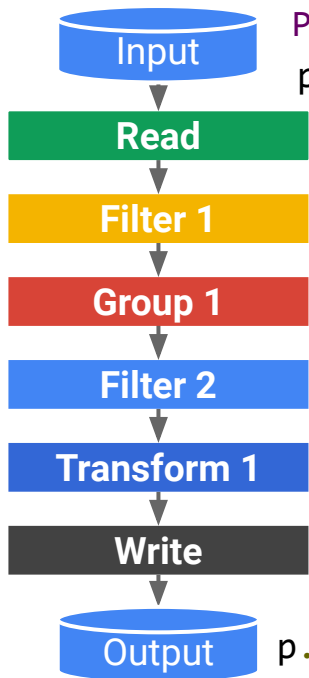
github.com/GoogleCloudPlatform/DataflowTemplates

- BigQuery to Datastore
- Bigtable to GCS Avro
- Bulk Compressor
- Bulk Decompressor
- Datastore Bulk Delete *
- Datastore to BigQuery
- Datastore to GCS Text *
- Datastore to Pub/Sub *
- Datastore Unique Schema Count
- GCS Avro to Bigtable
- GCS Avro to Spanner
- GCS Text to BigQuery *
- GCS Text to Datastore
- GCS Text to Pub/Sub (Batch)
- GCS Text to Pub/Sub (Streaming)
- Jdbc to BigQuery
- Pub/Sub to BigQuery *
- Pub/Sub to Datastore *
- Pub/Sub to GCS Avro
- Pub/Sub to GCS Text
- Pub/Sub to Pub/Sub
- Spanner to GCS Avro
- Spanner to GCS Text
- Word Count

What is a pipeline?



Dataflow offers NoOps data pipelines



```
Pipeline p = Pipeline.create();  
p  
  .apply(TextIO.Read.from("gs://...  
  ..."))  
  .apply(ParDo.of(new Filter1()))  
  .apply(new Group1())  
  .apply(ParDo.of(new Filter2()))  
  .apply(new Transform1())  
  .apply(TextIO.Write.to("gs://..."  
  ...));  
p.run();
```

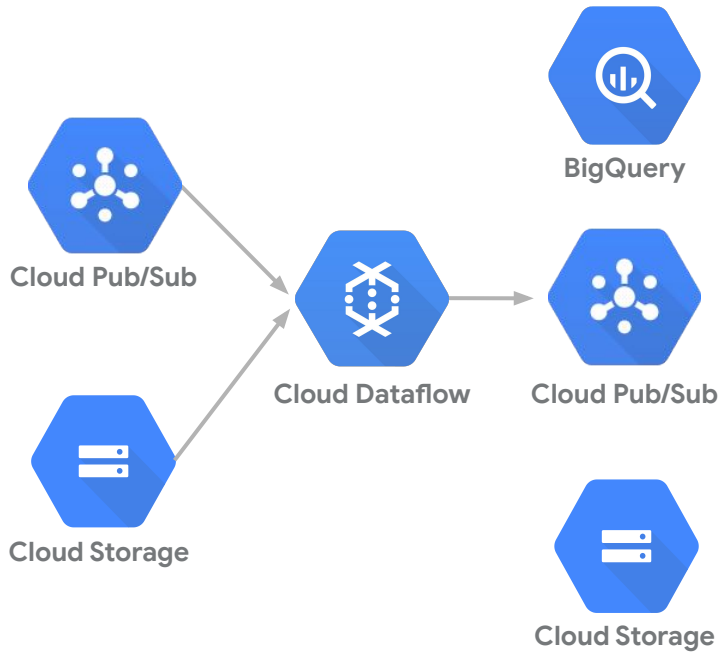
Open-source API (Apache Beam) can be executed on Flink, Spark, etc. also

Parallel task (autoscaled by execution framework)

```
class Filter1 extends DoFn<...> {  
  public void  
  processElement(ProcessContext c) {  
    ... = c.element();  
    ...  
    c.output(...);  
  }  
}
```



Same code does real-time and batch



```
Pipeline p = Pipeline.create();
p.begin()
  .apply(PubsubIO.Read.from("input_topic"))
  .apply(SlidingWindows.of(60, MINUTES))
  .apply(ParDo.of(new Filter1()))
  .apply(new Group1())
  .apply(ParDo.of(new Filter2()))
  .apply(new Transform1())
  .apply(PubsubIO.Write.to("output_topic"));
p.run();
```

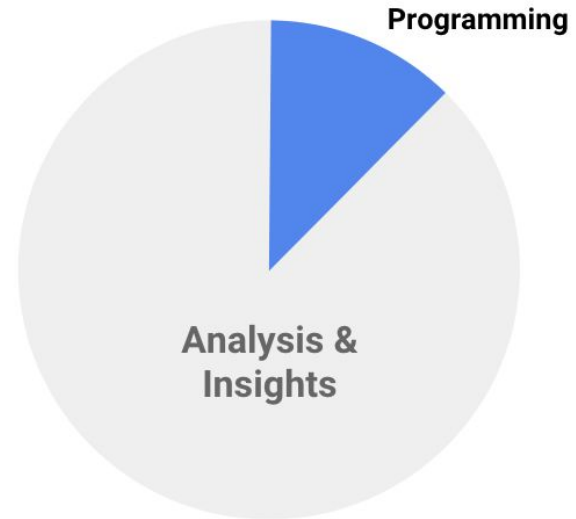
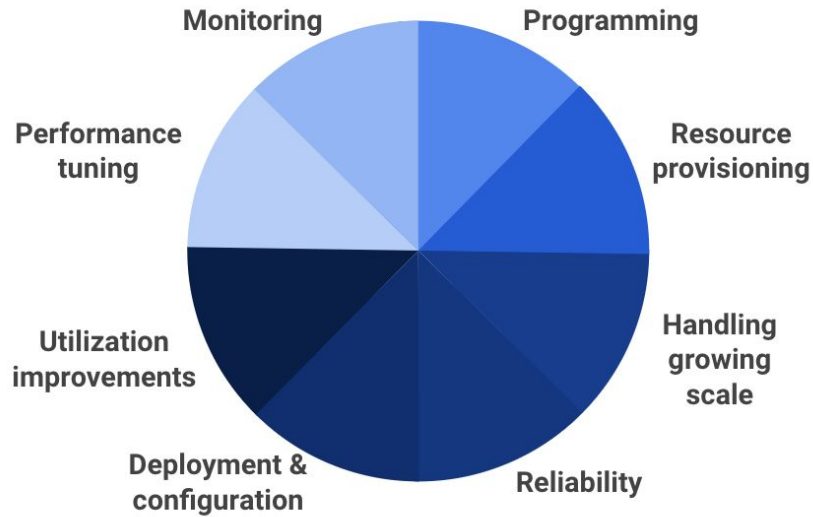
Data Engineers need to solve two distinct problems

Implementation with Google Cloud Dataflow

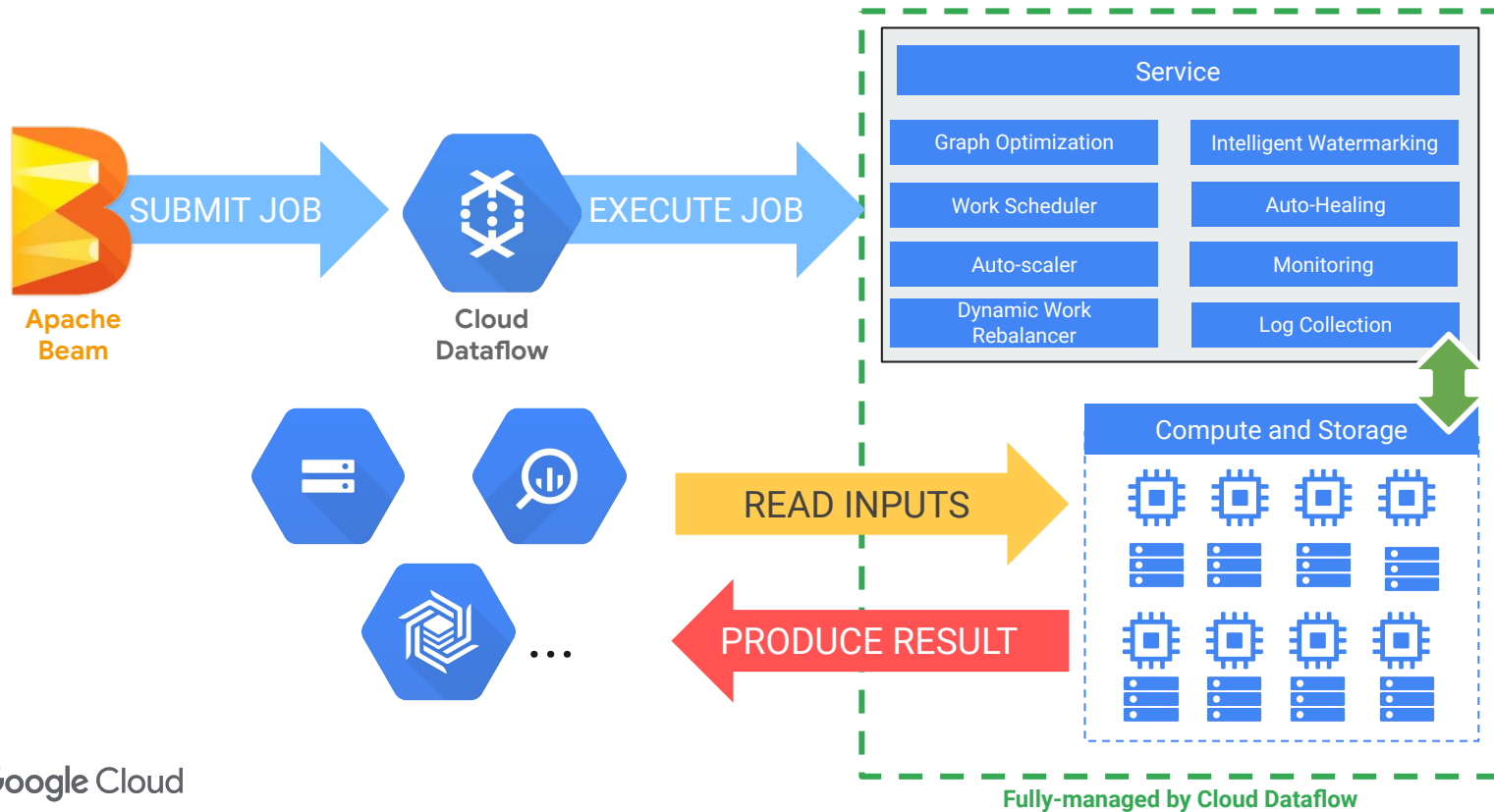


How much maintenance overhead is involved?	Little
Is the infrastructure reliable?	Built on Google infrastructure
How is scaling handled?	Autoscale workers
How can I monitor and alert?	Integrated with Stackdriver
Am I locked in to a vendor?	Run Apache Beam elsewhere

Why Serverless?



Example Dataflow fully-managed workflow



BigQuery is a petabyte-scale fully-managed data warehouse



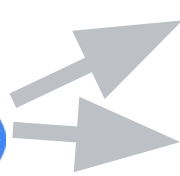
Google
BigQuery

1. It's serverless
2. Flexible pricing model
3. Data encryption and security
4. Geospatial data types & functions
5. Foundation for BI and AI

BigQuery is two services in one



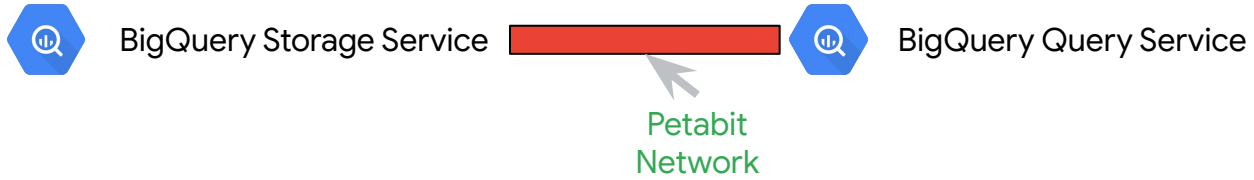
Google
BigQuery



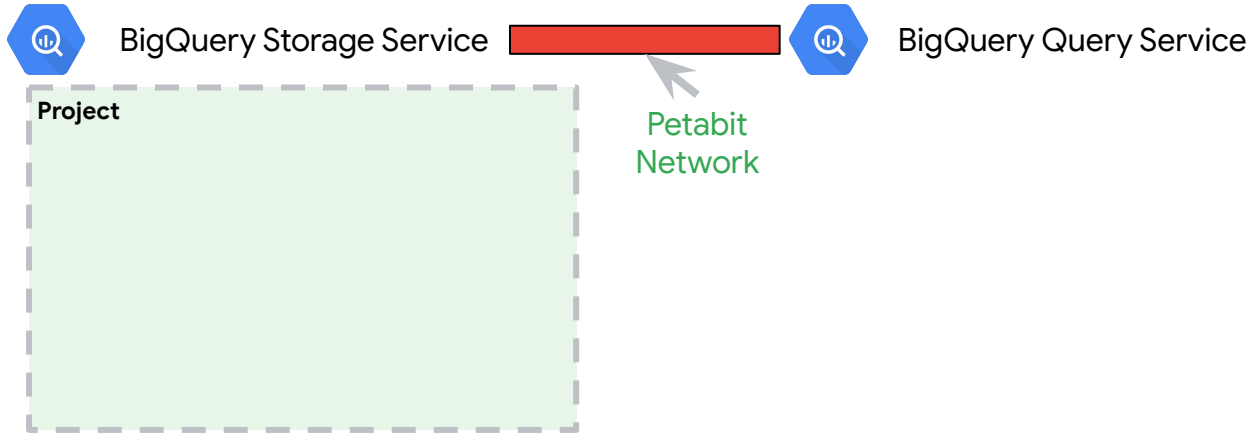
1. Fast SQL Query Engine

2. Managed storage for datasets

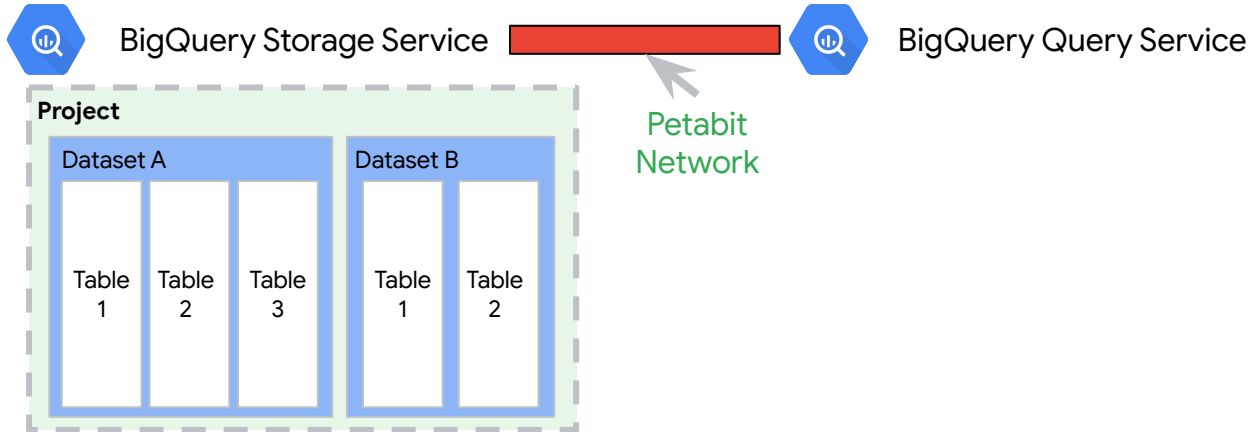
How does BigQuery work?



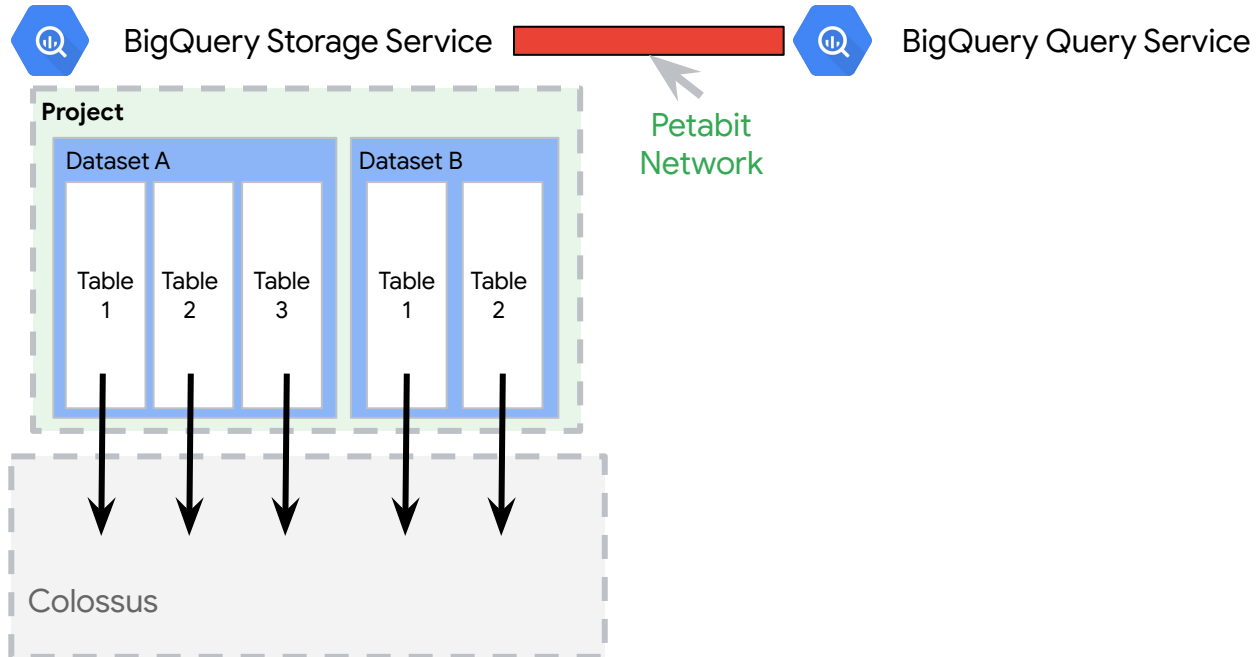
How does BigQuery work?



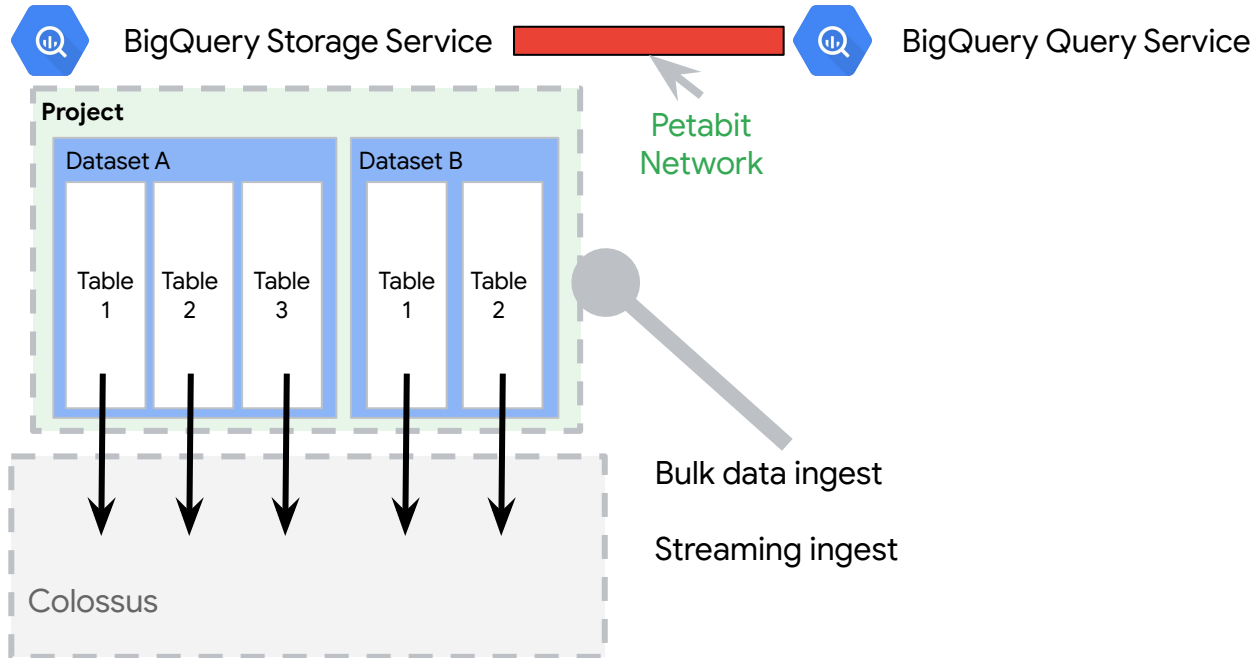
How does BigQuery work?



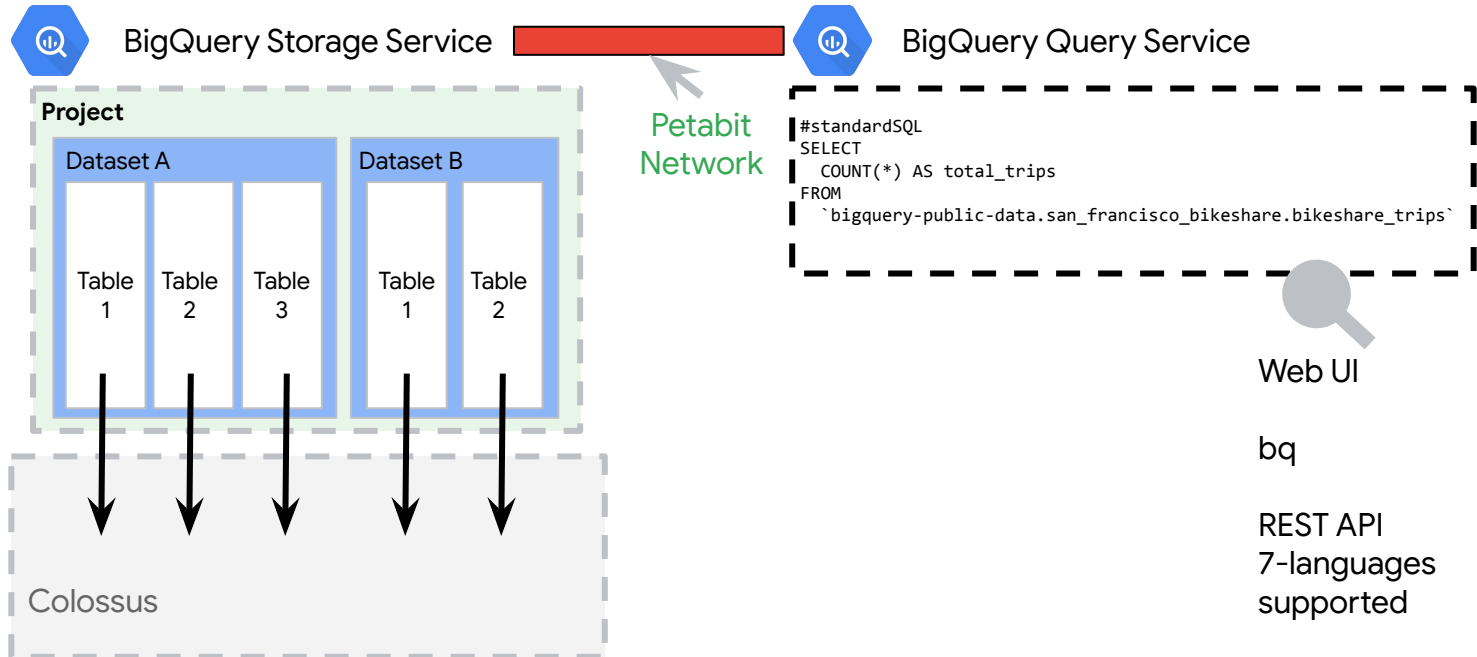
How does BigQuery work?



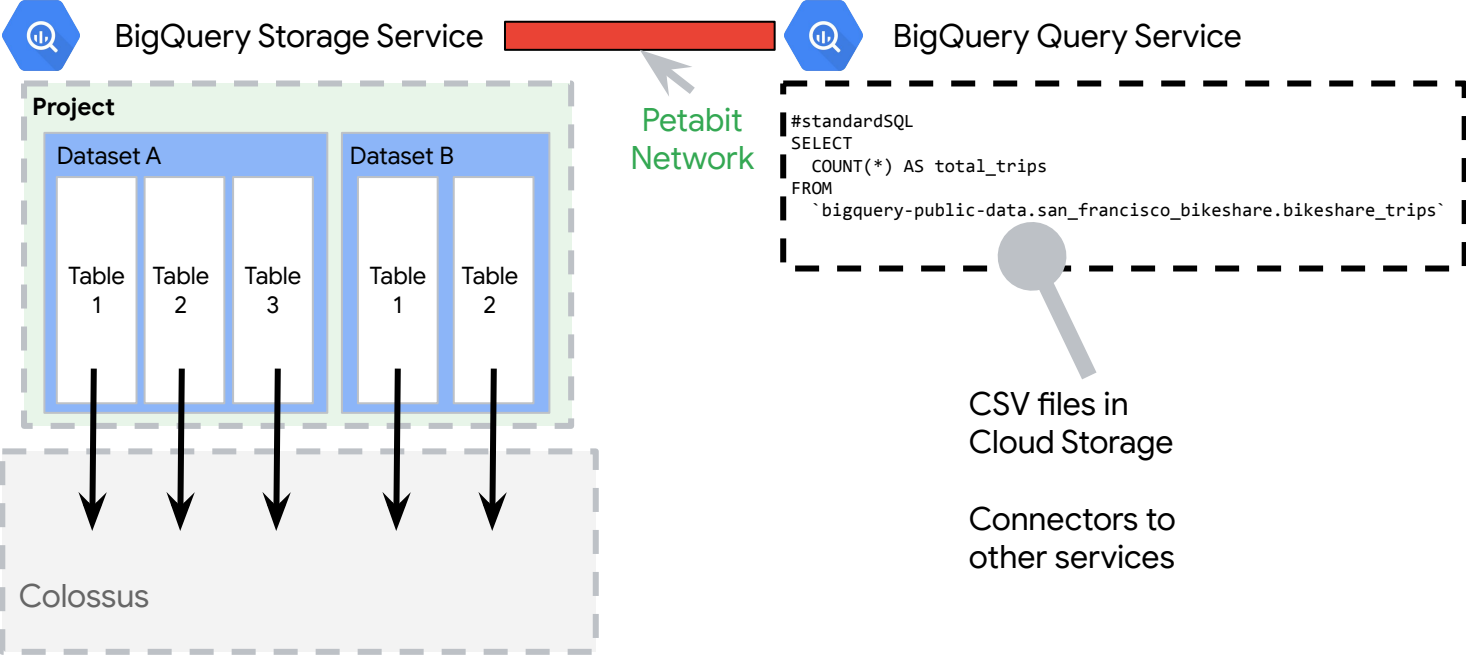
How does BigQuery work?



How does BigQuery work?



How does BigQuery work?



BigQuery supports standard SQL queries for analysis

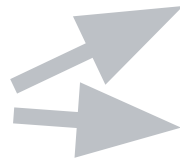
```
#standardSQL
SELECT
  COUNT(*) AS total_trips
FROM
  `bigquery-public-data.san_francisco_bikeshare.bikeshare_trips`
```

Row	total_trips
1	1947419

BigQuery is two services in one



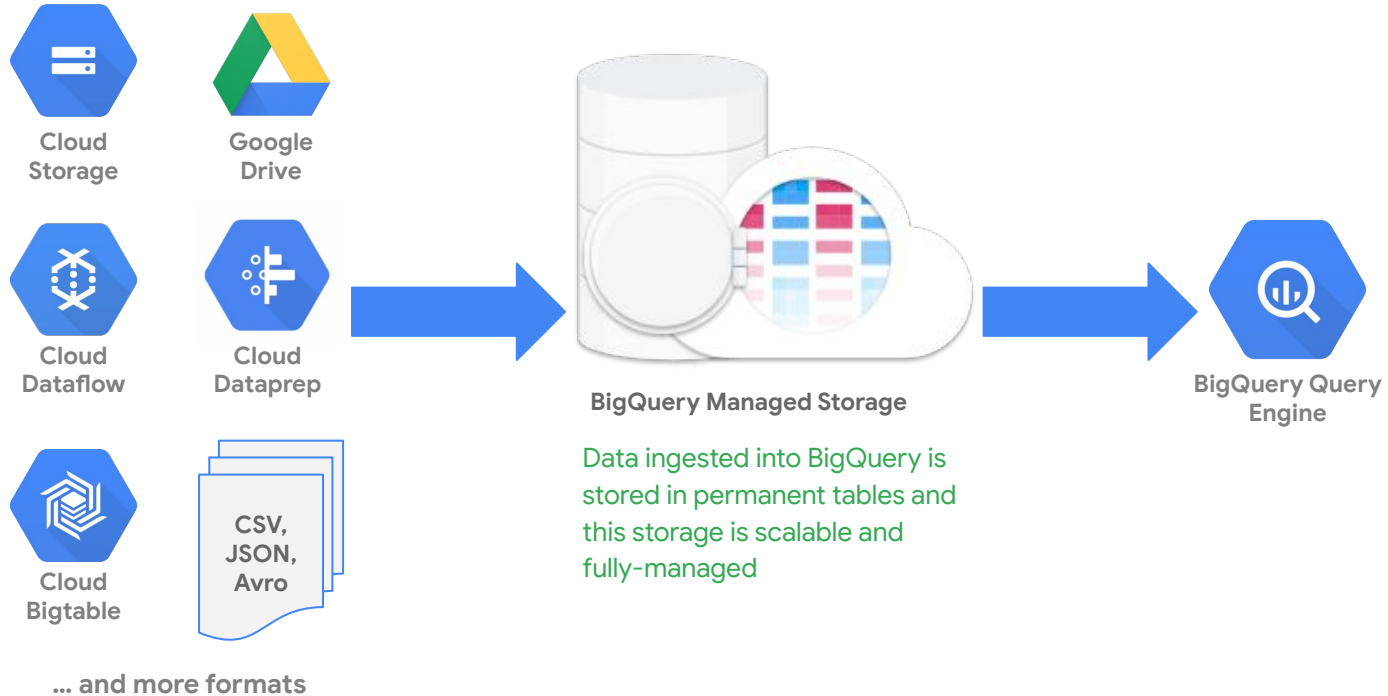
Google
BigQuery



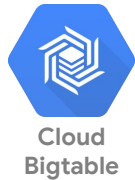
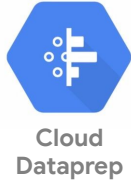
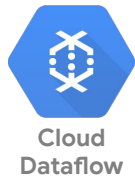
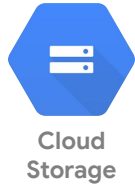
1. Fast SQL Query Engine

2. Managed storage for
datasets

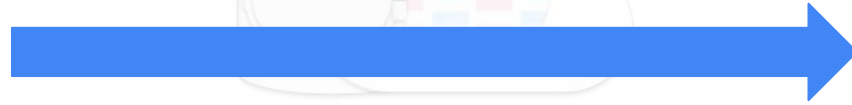
Use native BigQuery storage for the highest performance



BigQuery can query external (aka federated) data sources in GCS and Drive directly

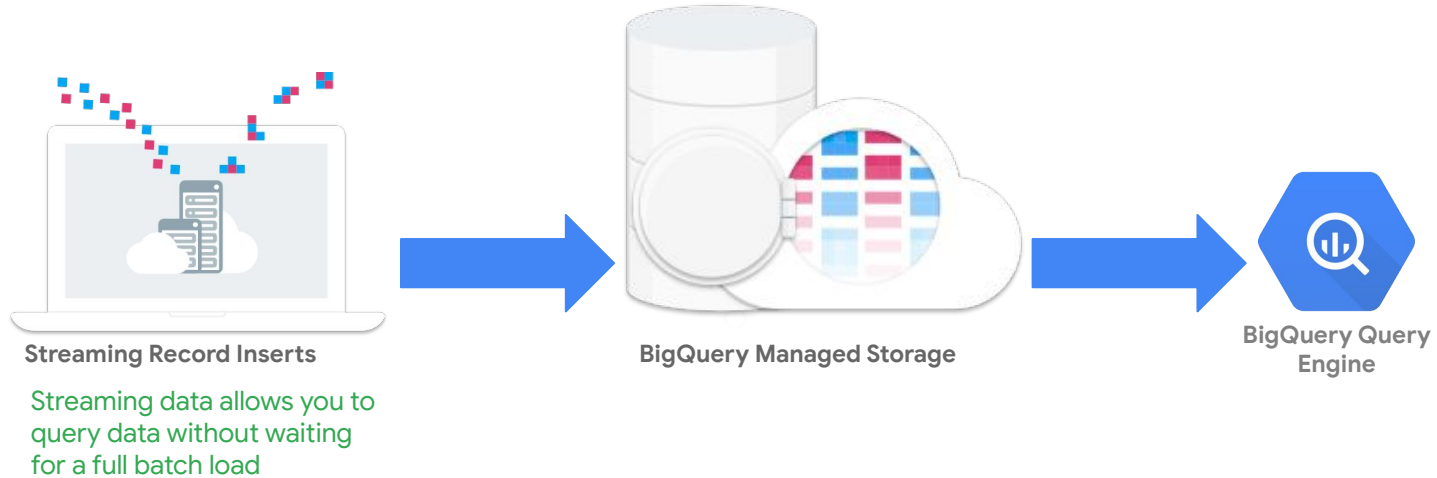


... and more formats



You can query external data sources directly from BigQuery which bypasses managed storage

Streaming records into BigQuery through the API



Explore Data Studio insights right from within BigQuery

Query editor

```
1 # which days did it rain in SF?
2 WITH rainy_sf AS (
3 SELECT
4 wban,
5 stn,
6 rain_drizzle,
7 fog,
8 PARSE_DATE("%F", CONCAT(year, '-', mo, '-', da)) AS date
9 FROM bigquery-public-data.noaa_gsod.gsod2018
10 WHERE wban = '93816'
11 ORDER BY rain_drizzle DESC, date
12 )
13
```

[Run](#) [Save query](#) [Save view](#) [Schedule query](#) [More](#)

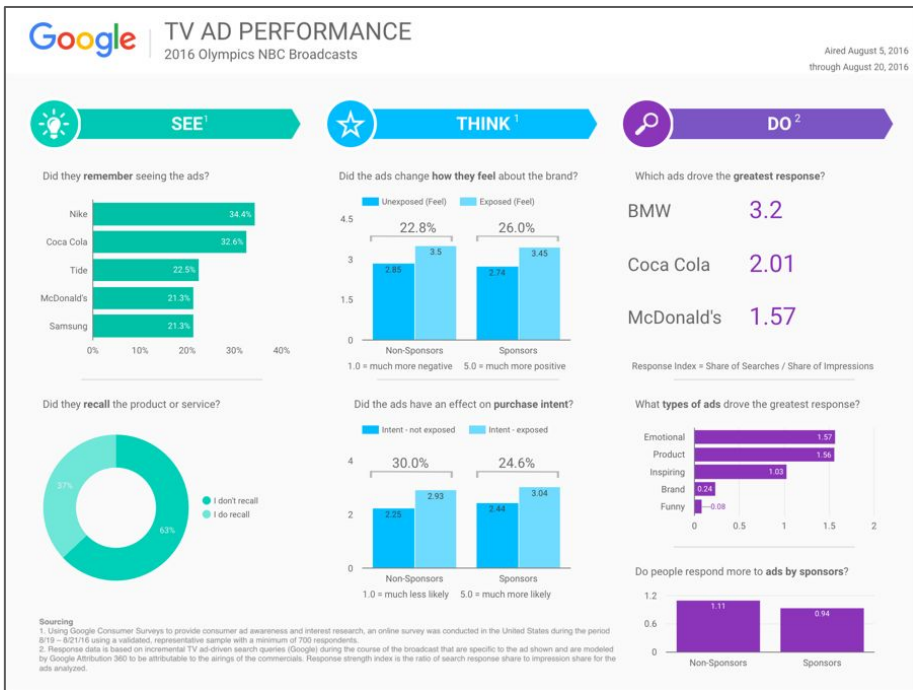
Query results [SAVE RESULTS](#) [EXPLORE IN DATA STUDIO](#)

Query complete (2.5 sec elapsed, 118.1 MB processed)

Job information [Results](#) JSON Execution details

Row	date	total_trips	rain_drizzle	fog
1	2018-01-07	1382	1	0
2	2018-01-08	805	1	0
3	2018-01-10	3459	1	1

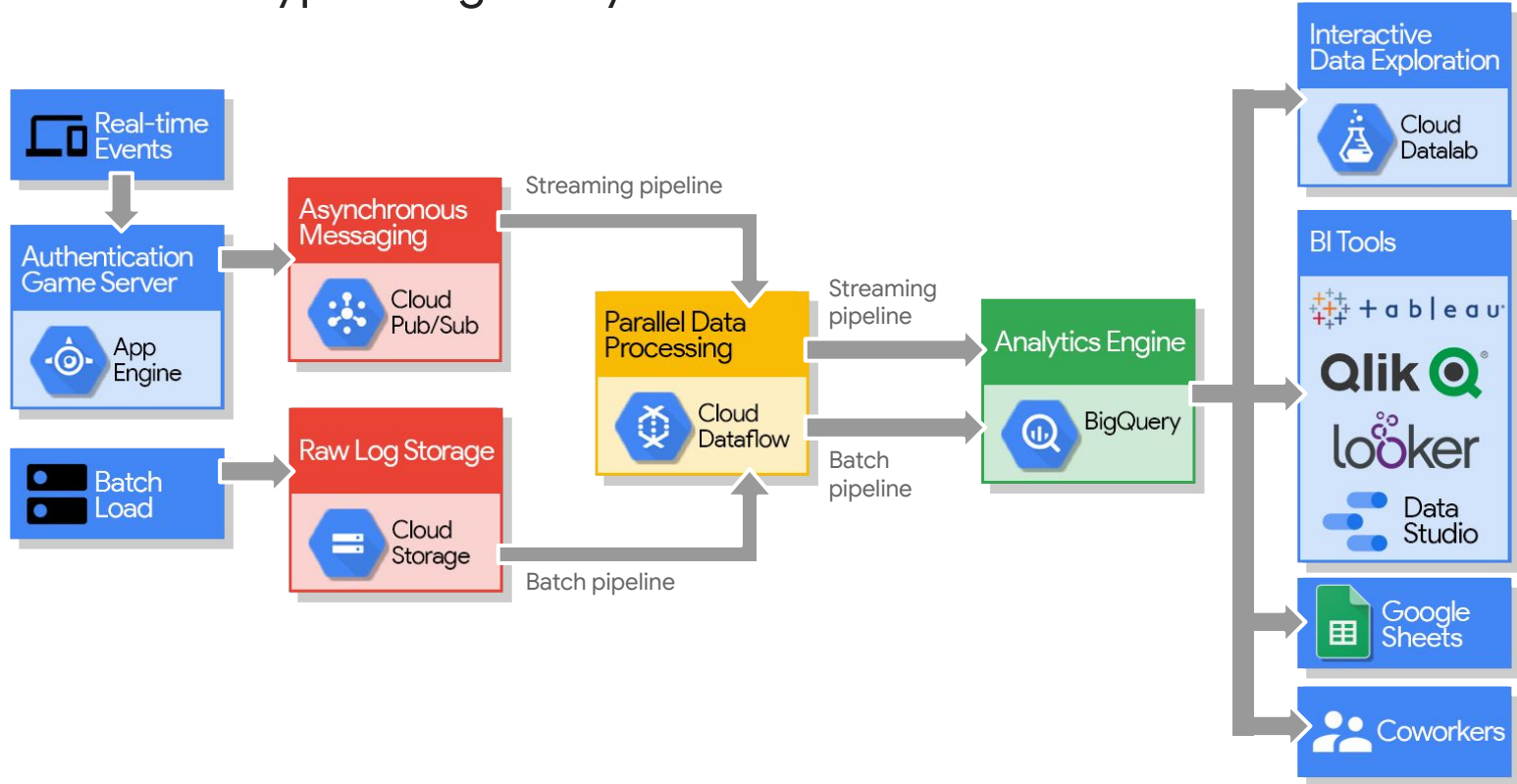
Build, collaborate, and share your dashboards



Tell a clear story with your data

Share and collaborate on reports with others

Typical BigQuery data warehouse architecture



Demo

Break - 15 min

Module 3

Deriving Insights using ML



Agenda

- Intro to Google Cloud Platform infrastructure
- Big data products:
 - Pub/Sub
 - Dataflow
 - BigQuery

- ML products:
 - ML APIs
 - AutoML
 - BigQuery ML

The popular imagination of what ML is



Lots of data



Complex mathematics in
multidimensional spaces



Magical results

In reality, ML is...



Collect
data



Organize
data



Create
model



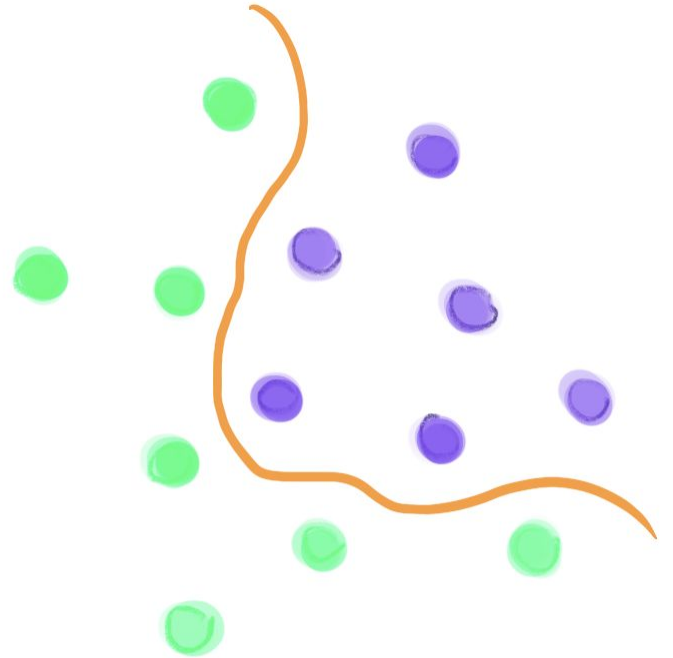
Experiment
a lot



Magical results

What is machine learning?

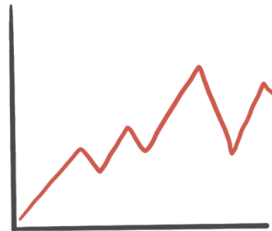
Finding patterns in data



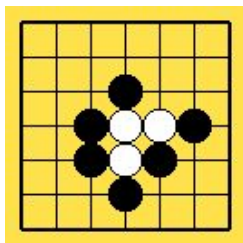
Recommendation



Forecasting



Game Strategy



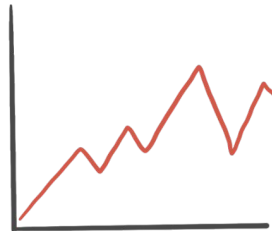
Classification



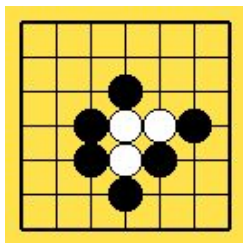
Recommendation



Forecasting



Game Strategy

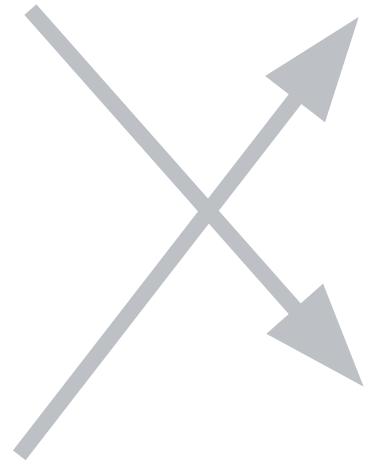


Classification

"CAT"



Classification



Dog

Cat

Traditional Programming:

```
if (animal) :  
    has a long tail  
    and  
    likes fish  
    and  
    hates people  
then:  
    return "cat"
```

Traditional Programming:

```
if (animal) :  
    has a long tail  
    and  
    likes fish  
    and  
    hates people  
then:  
    return "cat"
```



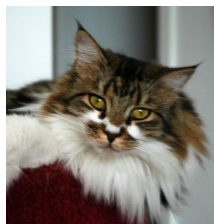
rules

Machine Learning:

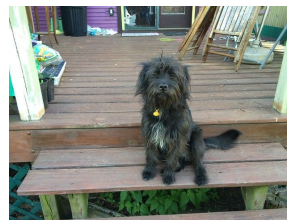
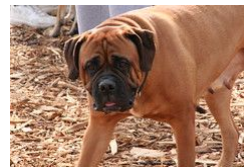
Learn by **examples**, not **rules**

Labeled Training Dataset

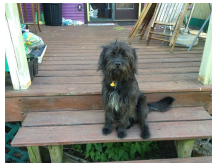
Examples of cats



Examples of dogs



Training a Model

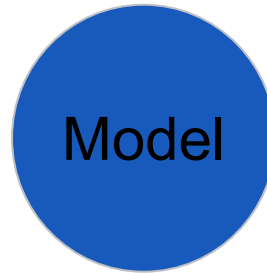


Machine
Learning
Algorithm



Model

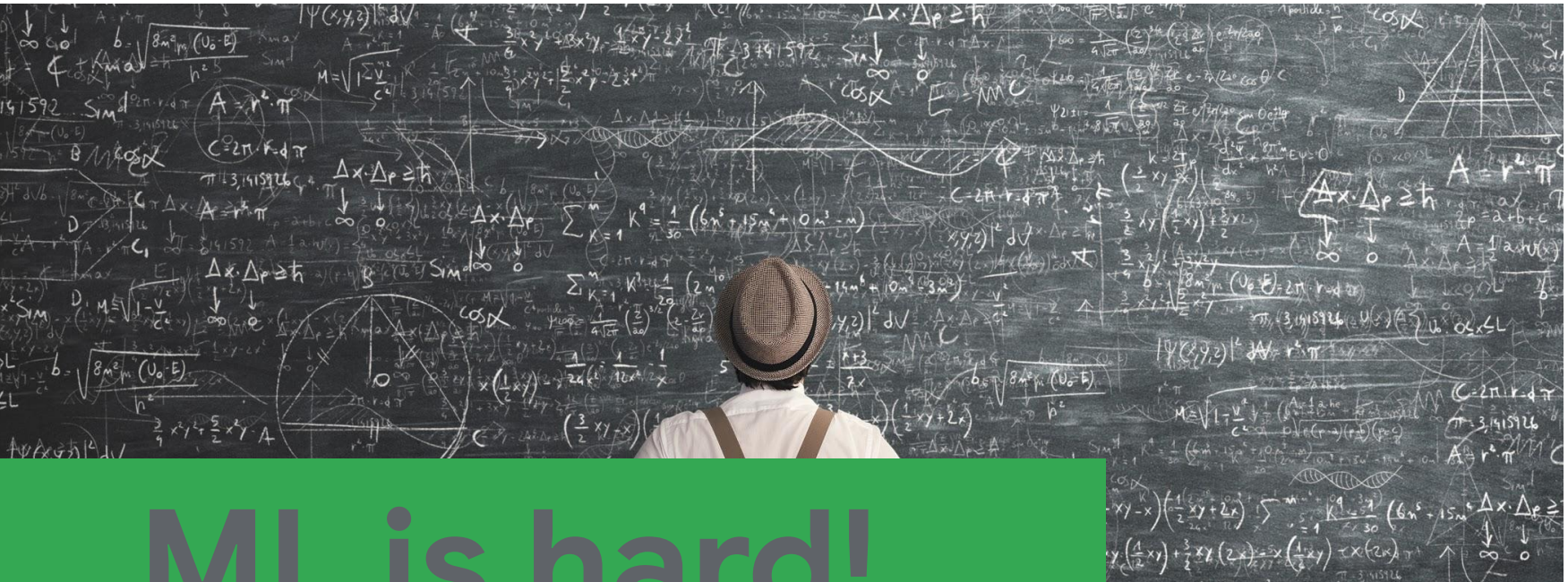
Making Predictions



“Dog”

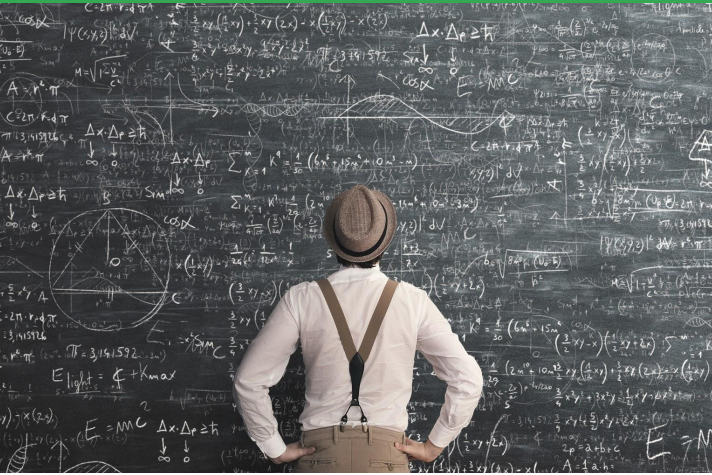


ML is great!



ML is hard!

Most folks



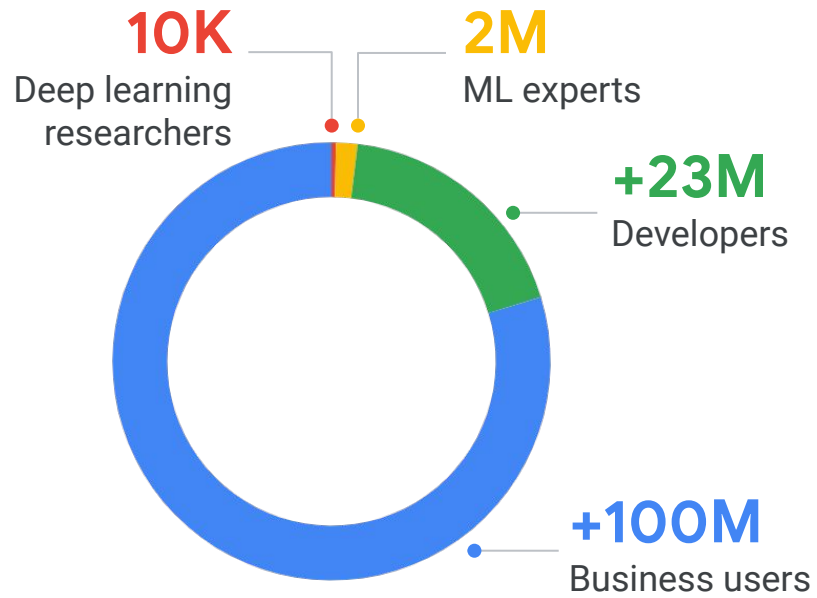
Lots of pain



Magical AI goodness

If ML is a rocket engine, data is the fuel







Democratising Machine Learning

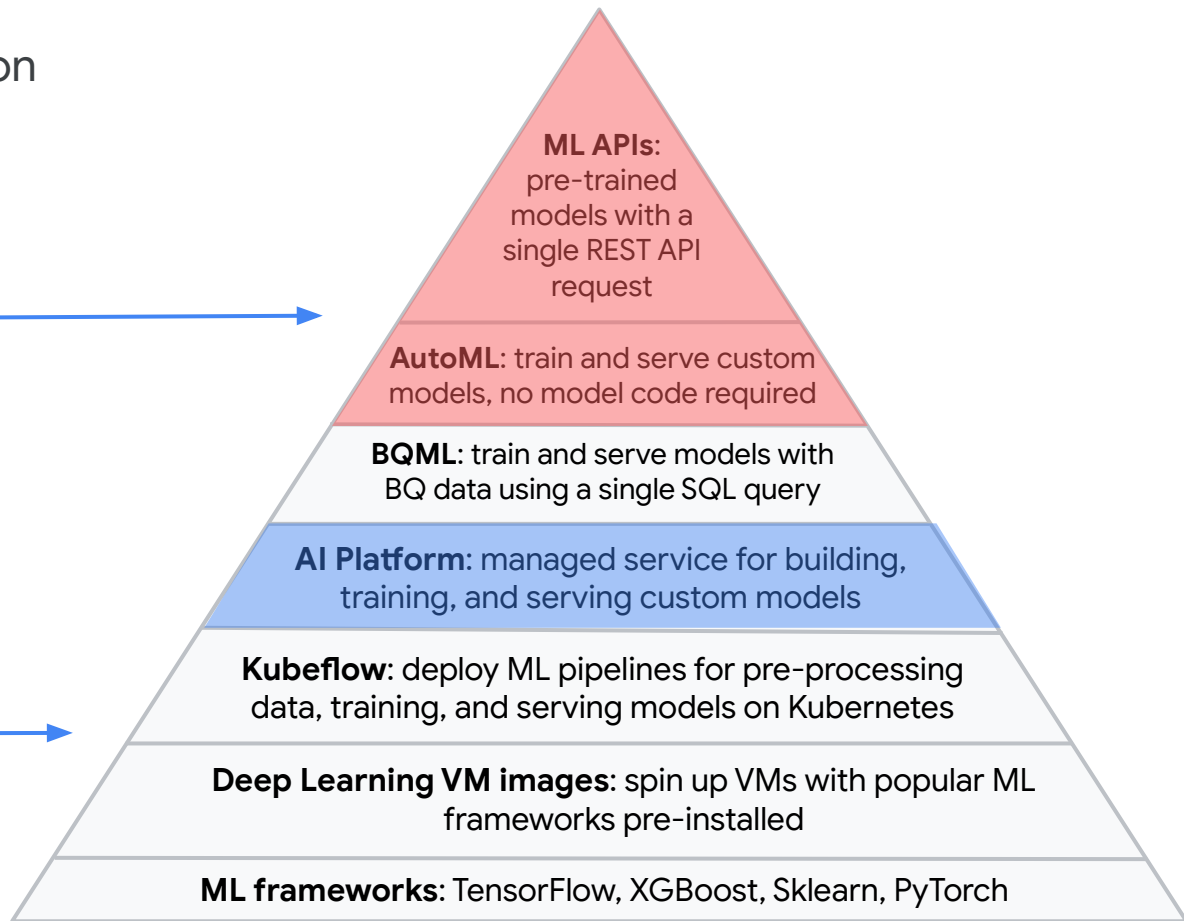
ML @ Google Cloud:

Choose your level of abstraction

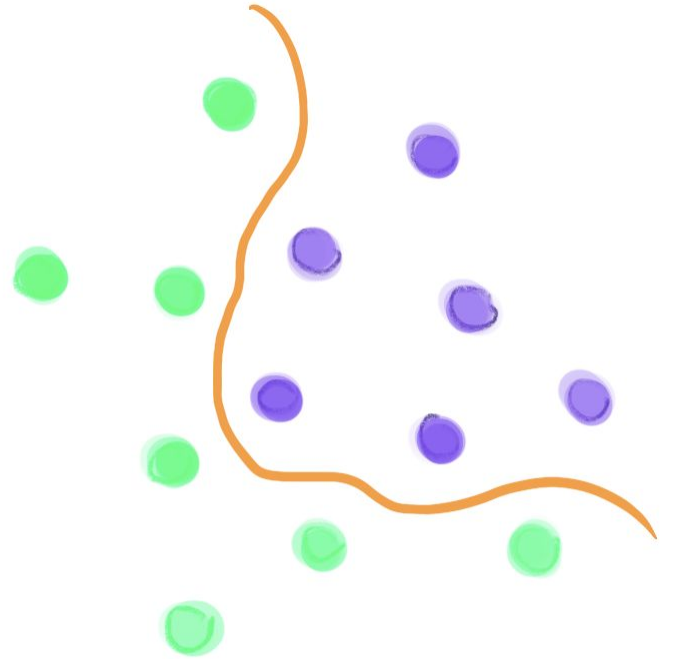
Application
developers



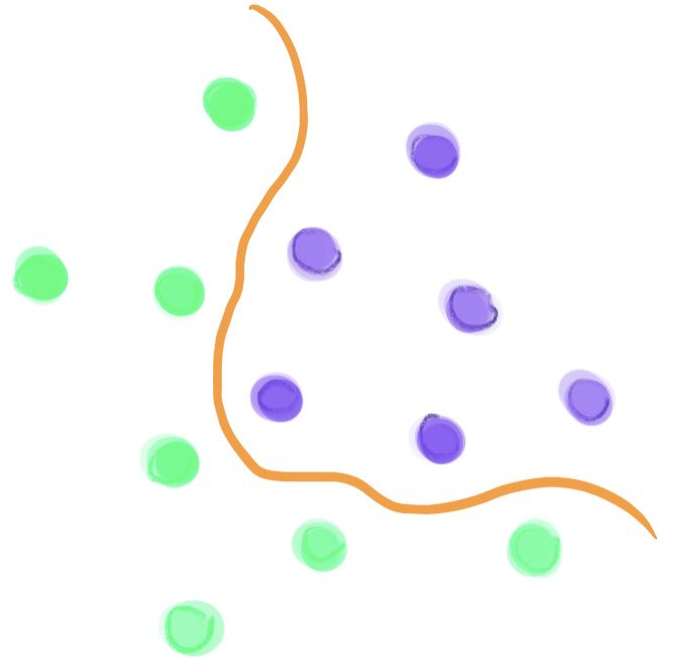
Data scientists &
ML engineers



Finding patterns in data



Finding patterns in data
*requires data



No data?

No model :(

Use Google's models

Use a **pre-trained model** to accomplish common ML tasks



Cloud Vision



Cloud Video Intelligence



Cloud Speech-to-Text &
Text-to-Speech



Cloud Natural Language



Cloud Translation



Cloud Vision API

Faces

Faces, facial landmarks, emotions



Label

Detect entities from furniture to transportation



OCR

Support for > 50 languages, images, PDF, TIFF



Landmarks

Detect landmarks using Google index



Logos

Identify product logos



Safe Search

Detect explicit content - adult, violent, medical and spoof



Web Detection

Leverage power of Google Search



Product Search

Identify products from your catalog



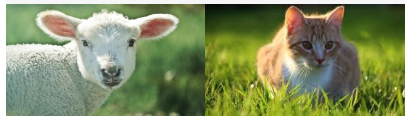
Image Properties

Dominant colors



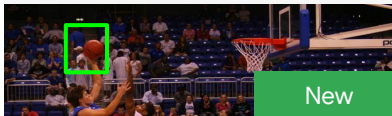
Crop hints

Detect salient image patches



Object Localizer

Retrieve object coordinates



Handwriting OCR

Extract handwritten text from your documents





labels

Landmark	96%
Place Of Worship	83%
Architecture	79%
Temple	74%
Hindu Temple	74%
Temple	72%
Tourist Attraction	71%
Building	68%
Amusement Park	64%



emotion

Joy	■	Very Unlikely
Sorrow	■	Very Unlikely
Anger	■	Very Unlikely
Surprise	■	Very Unlikely
Exposed	■	Very Unlikely
Blurred	■	Very Unlikely
Headwear	■ ■ ■ ■ ■	Very Likely

Roll: -2° Tilt: 4° Pan: 1°

Confidence 53%

Celebrity Recognition **NEW**



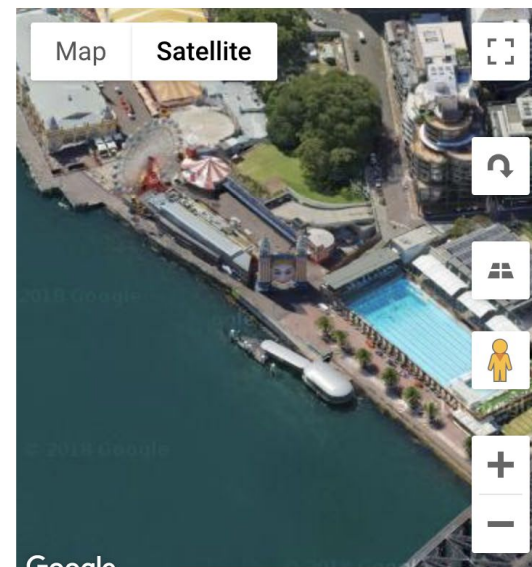
"BRAD PITT"



landmark detection

Luna Park Sydney

42%





ocr

"LUNA PARK"



Video Intelligence API



Track Objects

Speech Transcription

Explicit Content Detection

... and more

Demo



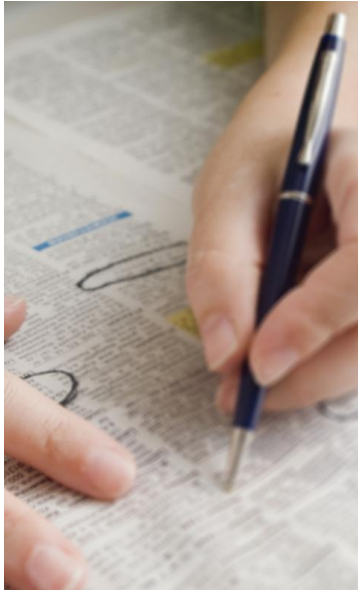
Speech-to-Text and Text-to-Speech

Demo

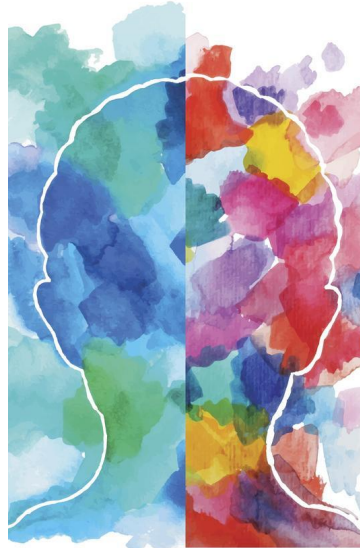




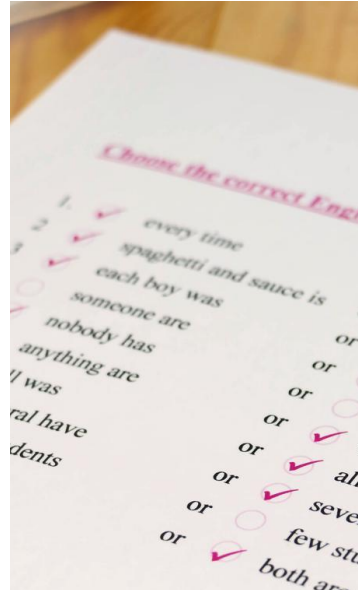
Cloud Natural language api



Extract
entities



Detect
sentiment



Analyze
syntax



Classify
content



Cloud Natural Language

☰ 


PLAY THE CROSSWORD

TECH FIX

Protecting Your Internet Accounts Keeps Getting Easier. Here's How to Do It.

There are many tools for setting up two-factor authentication, a security mechanism that prevents improper access. These four methods are the most compelling.




☰ 

GET UPDATES

NONFICTION

The Two Artist Couples Who Helped Start American Modernism



☰ 

GET UPDATES

RESTAURANT REVIEW

What Has New York Pizza Been Missing? Little Old Rhode Island



Pizza, Hot Off the Grill

 10 Photos | [View Slide Show >](#)



Cloud Natural Language

☰ 👤

PLAY THE CROSSWORD

TECH FIX

Protecting Your Internet Accounts Keeps Getting Easier. Here's How to Do It.

There are many tools for setting up two-factor authentication, a security mechanism that prevents improper access. These four methods are the most compelling.

Computers & Electronics

☰ 👤

GET UPDATES

NONFICTION

The Two Artist Couples Who Helped Start American Modernism

Arts & Entertainment / Visual Art & Design

☰ 👤

GET UPDATES

RESTAURANT REVIEW

What Has New York Pizza Been Missing? Little Old Rhode Island

Pizza, Hot Off the Grill

📷 10 Photos

Food & Drinks / Restaurants / Pizzerias

entity extraction

Try the API



SINGAPORE (AP) — President Donald Trump and North Korea's Kim Jong Un concluded an extraordinary nuclear summit Tuesday by signing a document in which Trump pledged "security guarantees" to the North and Kim reiterated his commitment to "complete denuclearization of the Korean Peninsula." The leaders also offered lofty promises, with the American president pledged to handle a "very dangerous problem" and Kim forecasting "major change for the world."

ANALYZE

[See supported languages](#)

Entities

Sentiment

Syntax

Categories

⟨SINGAPORE⟩₅ (⟨AP⟩₆) — ⟨President⟩₁ ⟨Donald Trump⟩₁ and ⟨North Korea⟩₃'s ⟨Kim Jong Un⟩₂ concluded an extraordinary nuclear ⟨summit⟩₇ Tuesday by signing a ⟨document⟩₄ in which ⟨Trump⟩₁ pledged "⟨security guarantees⟩₉" to the ⟨North⟩₃ and ⟨Kim⟩₂ reiterated his ⟨commitment⟩₈ to "complete ⟨denuclearization⟩₁₀ of the ⟨Korean Peninsula⟩₁₃." The ⟨leaders⟩₁₁ also offered lofty ⟨promises⟩₁₅, with the ⟨American⟩₂₁ ⟨president⟩₁₄ pledged to handle a "very dangerous ⟨problem⟩₁₇" and ⟨Kim⟩₂ forecasting "major ⟨change⟩₁₆ for the ⟨world⟩₂₅." The broad ⟨agreement⟩₁₂ was ⟨light⟩₁₂ on ⟨specifics⟩₂₆, largely reiterating previous public ⟨statements⟩₁₈ and past ⟨commitments⟩₁₉. It did not include an ⟨agreement⟩₂₀ to take ⟨steps⟩₂₂ toward ending the technical ⟨state⟩₂₃ of ⟨warfare⟩₂₄ between the ⟨U.S.⟩₂₁ and ⟨North Korea⟩₃.

1. Donald Trump

PERSON

Sentiment: Score 0 Magnitude 0.9

[Wikipedia Article](#)

Salience: 0.26

2. Kim Jong Un

PERSON

Sentiment: Score 0 Magnitude 1.3

[Wikipedia Article](#)

Salience: 0.11

Demo



Cloud AutoML

Use your data to extend
Google's pretrained models



Video



Vision



Translation



Tables *new*



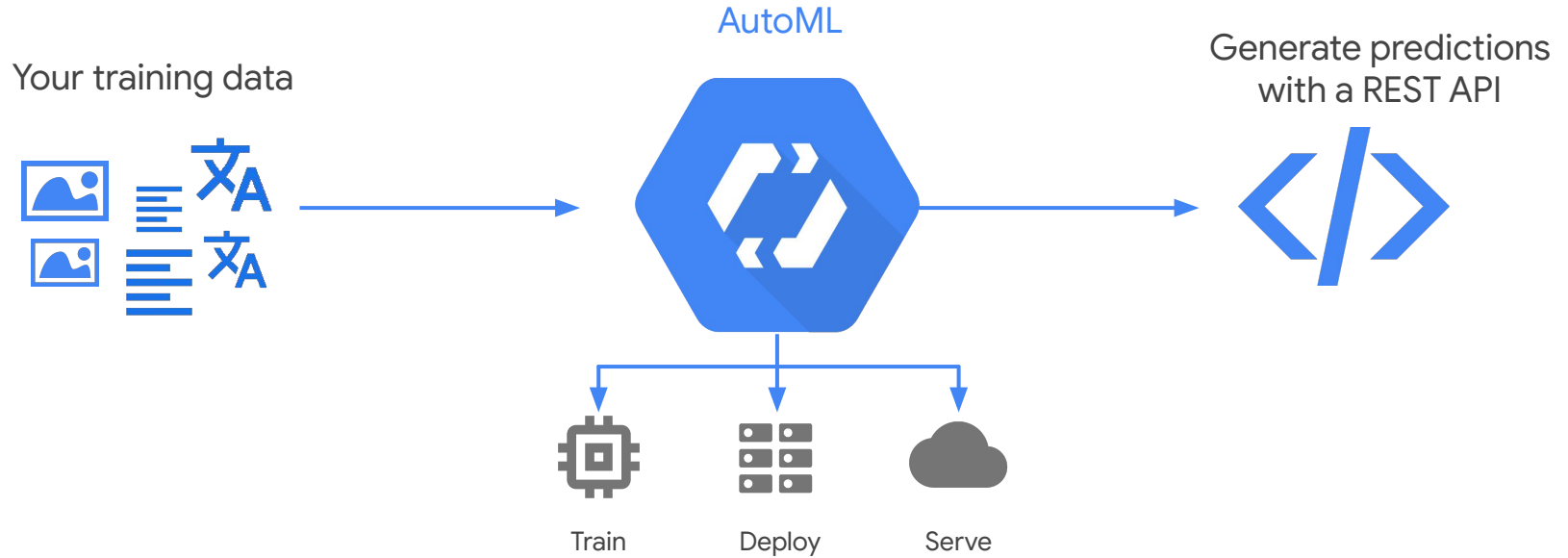
Language



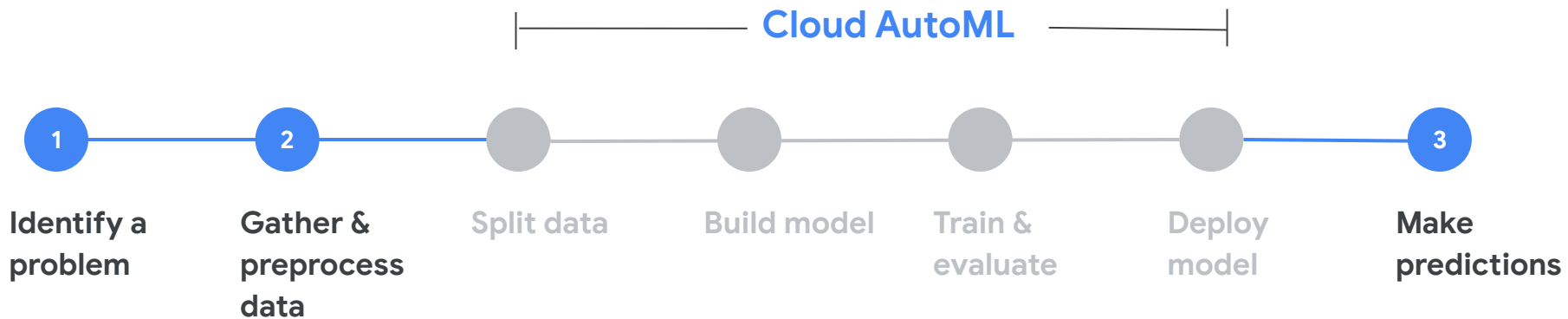
Recommendations *new*



What is Cloud AutoML?



How can Cloud AutoML help?



AutoML Vision

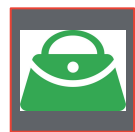
AutoML Vision

How it works?

Upload and label images

Train your model

Evaluate



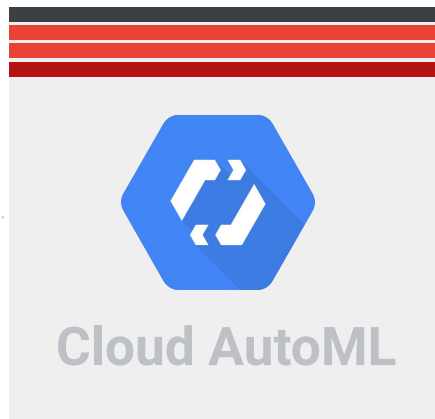
Handbag



Shoe



Hat



Model is now trained and ready to make prediction.
This model can scale as needed to adapt to customer demands.

Select the type of model



Create new dataset

Dataset name *

untitled_1579212004665



Use letters, numbers and underscores up to 32 characters.

Select your model objective



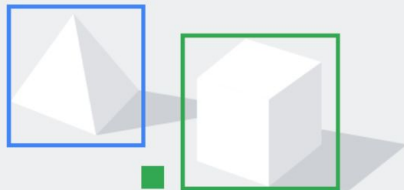
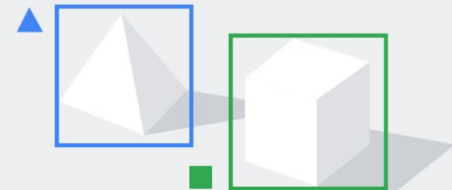
Single-Label Classification

Predict the one correct label that you want assigned to an image.



Multi-Label Classification

Predict all the correct labels that you want assigned to an image.



Object detection

Predict all the locations of objects that you're interested in.

CANCEL

CREATE DATASET

Import Images

← untyped_1579212004665 || LABEL STATS EXPORT DATA

IMPORT IMAGES TRAIN EVALUATE TEST & USE

Select files to import

To build a custom model, you first need to import a set of images to train it. Each image should be categorized with a label. (Labels are essential for telling the model how to identify an image.)

- Each label should have at least 100 images for best results.

- Upload images from your computer
- Select a CSV file on Cloud Storage

Select a CSV file on Cloud Storage

If you haven't already, upload your files to [Cloud storage](#). The CSV file should be a list of GCS paths to your images. Images can be in JPG, PNG, GIF, BMP or ICO formats.

Optionally, you can specify the TRAIN, VALIDATE, or TEST split.

Sample CSV format

```
[set, ]image_path[, label]
TRAIN,gs://My_Bucket/sample1.jpg,cat
TEST,gs://My_Bucket/sample2.jpg,dog
```

gs:// * BROWSE

CONTINUE

Review dataset

Datasets

[+ NEW DATASET](#)

	Name	Type	Total images	Labeled images	Last updated	Status	
	untitled_1579212004665 ICN1037466186620600320	Single-Label Classification	0	0	Jan 16, 2020, 2:02:03 PM	Running: Importing images	
	Vision_dataset1 ICN8336791830913875968	Single-Label Classification	3,667	3,666	Dec 12, 2019, 12:43:59 PM	Success: Training model	
	Vision_dataset2 ICN6465264710764724224	Single-Label Classification	3,667	3,666	Dec 12, 2019, 12:40:49 PM	Success: Training model	
	Vision_dataset3 ICN8451633621411823616	Single-Label Classification	3,667	3,666	Dec 12, 2019, 12:40:39 PM	Success: Training model	

Review labels

Vision

Vision_dataset1 LABEL STATS EXPORT DATA

IMPORT IMAGES TRAIN EVALUATE TEST & USE

All Images 3,667

Labeled 3,666

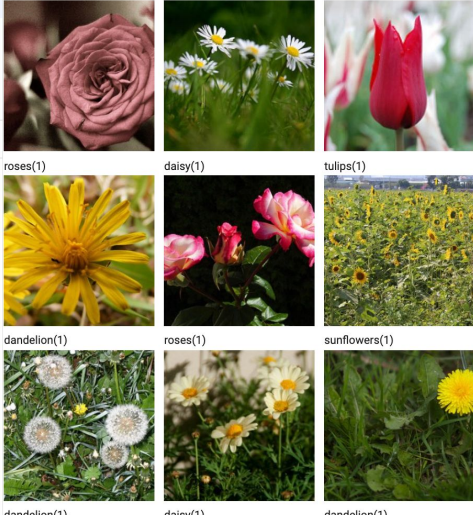
Unlabeled 1

Filter labels

daisy	633
dandelion	898
roses	640
sunflowers	697
tulips	798

ADD NEW LABEL

Filter images



roses(1) daisy(1) tulips(1)

dandelion(1) roses(1) sunflowers(1)


dandelion(1) daisy(1) dandelion(1)

Image 22 of 50

Filter labels

- daisy
- dandelion
- roses
- sunflowers
- tulips

Unlabeled



gs://automl-doc-filtering-vcv/img/flower_photos/roses/2788276815_8f730bd942.jpg

SAVE CANCEL < >

Train model


← Vision_dataset1 [LABEL STATS](#) [EXPORT DATA](#)

IMPORT IMAGES **TRAIN** EVALUATE TEST & USE Single

Models

[TRAIN NEW MODEL](#)

Vision_model5



Average precision [?](#)
0.996

Precision* [?](#) 96.17%
Recall* [?](#) 95.91%

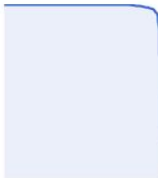
* Using a score threshold of 0.5

Model ID ?	ICN3665591447098228736
Created	Feb 6, 2020, 12:21:32 PM
Base model	None
Data	3,666 images
Model type	Cloud
Train cost	40 node hours
Deployment state	Not deployed

[SEE FULL EVALUATION](#)

[RESUME TRAINING](#)

Vision_model4



Average precision [?](#)
0.995

Precision* [?](#) 96.99%
Recall* [?](#) 96.46%

* Using a score threshold of 0.5

Model ID ?	ICN1790123677275127808
Created	Feb 6, 2020, 9:43:13 AM
Base model	None
Data	3,666 images
Model type	Cloud
Train cost	45 node hours
Deployment state	Not deployed

[SEE FULL EVALUATION](#)

[RESUME TRAINING](#)

Deploy Model

IMPORT

IMAGES

TRAIN

EVALUATE

TEST & USE

Single-Label Classification

Model

Vision_model1



To use online prediction, deploy your model to the cloud. Deployed model charges are per hour and number of machines used. [Pricing guide](#)

DEPLOY MODEL



Notice for beta users: The v1beta1 API endpoint is scheduled for deletion after GA release. If your beta models have not been [redeployed since October 17, 2019](#), please do so now to avoid interruption when the old service is shut down.

More AutoML Vision features: Edge models

Train new model

Model name
leaf_types_v20190416180326

Model type

Cloud-hosted
Host your model on Google Cloud for online predictions.

Edge
Download your model for offline/mobile use. Typically has lower accuracy than Cloud-hosted models.

Format model for Core ML (iOS / macOS)

Optimize model for:

Lowest latency Latency: 2 msec Size: 858 KB Accuracy: Typically lower	Best trade-off Latency: 3 msec Size: 3.7 MB Accuracy: Best trade-off	Higher accuracy Latency: 5 msec Size: 6.8 MB Accuracy: Typically higher
---	--	---

Show latency estimates for

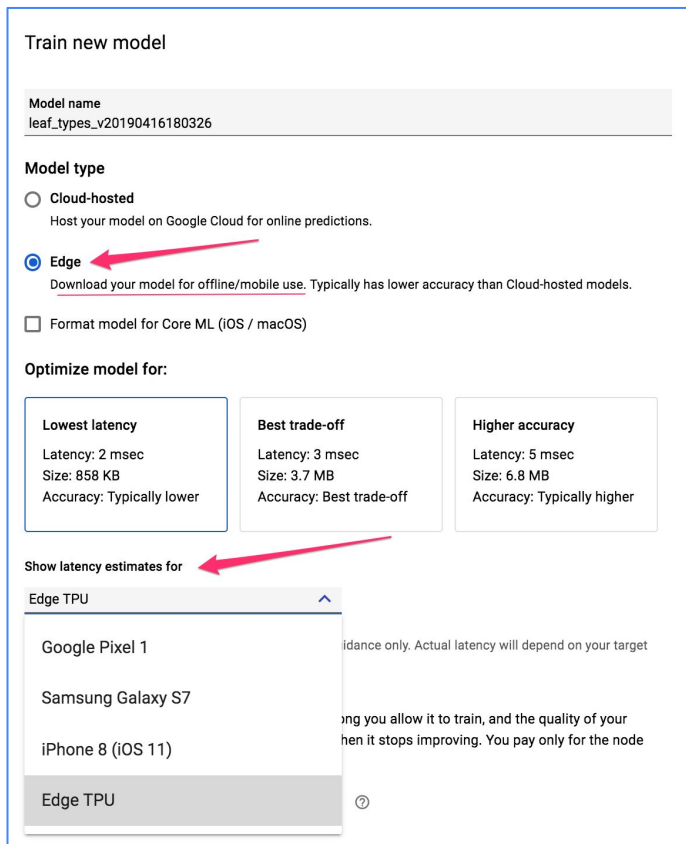
Edge TPU

Google Pixel 1

Samsung Galaxy S7

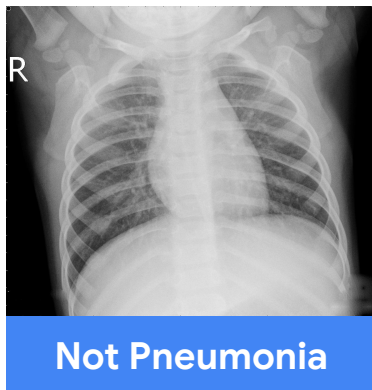
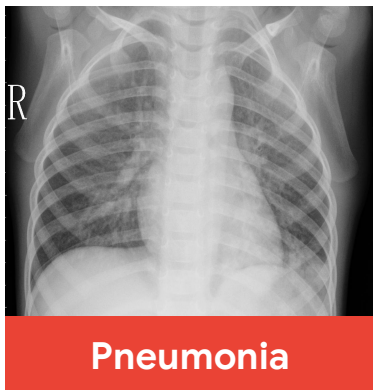
iPhone 8 (iOS 11)

Edge TPU



AutoML for Pneumonia detection

Distinguishing if an x-ray has pneumonia in it is a task that even a human can struggle with. AutoML Vision Classification can train a model that is over 99% accurate for this task.



5,863 high resolution images from [public dataset on Kaggle](#) with a mix of image formats and resolutions.

99+% Accuracy!

Beating almost all of the [of the models on Kaggle](#)



AutoML Natural Language

Classification

U.S. Congress bill topic categorization

A bill to provide additional financial assistance for educational and biological programs pertaining to U.S. fisheries.



Agriculture

A bill to provide for a temporary increase in the public debt limit.



Macroeconomics

Dataset source:

congressionalbills.org/credits.html

Custom Sentiment



Frequent Flyer

@frequentflyer11

@Alta I was stuck waiting on the tarmac for hours!



Frequent Flyer

@frequentflyer11

@JetGreen has so much legroom in coach!

Custom Sentiment



Frequent Flyer

@frequentflyer11

@Alta I was stuck waiting on the tarmac for hours!

Very Negative



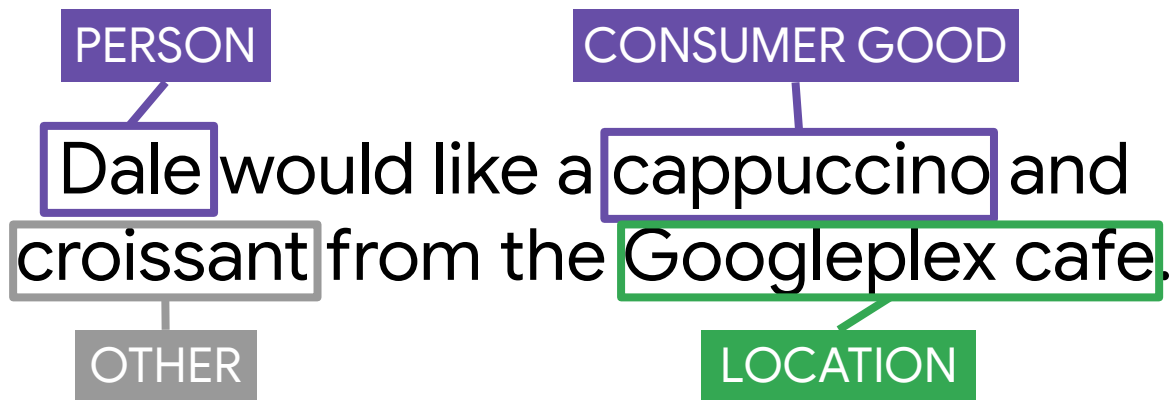
Frequent Flyer

@frequentflyer11

@JetGreen has so much legroom in coach!

Very Positive

Pretrained Entity Extraction



Custom Entity Extraction



Dataset

The Movies Dataset

Metadata on over 45,000 movies. 26 million ratings from over 270,000 users.



Rounak Banik · updated 2 years ago (Version 7)

[Data](#)[Kernels \(84\)](#)[Discussion \(11\)](#)[Activity](#)[Metadata](#)[Download \(228 MB\)](#)[New Kernel](#)

License CC0: Public Domain



Tags popular culture, film

Description

Context

These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.

This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

Movie Genre Classification

An epic drama of
adventure and exploration

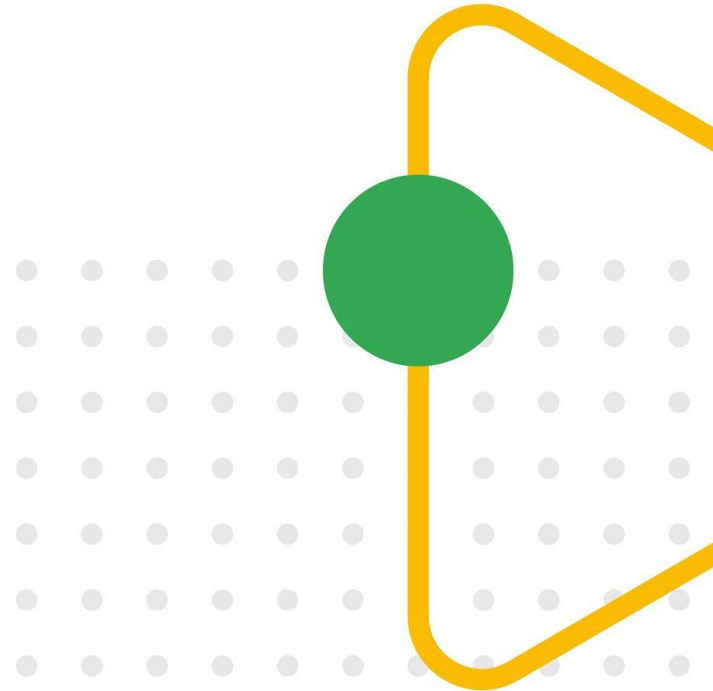


"After discovering a mysterious artifact buried beneath the lunar surface, mankind sets off on a quest to find its origins with help from intelligent supercomputer HAL 9000."



sci-fi

AutoML Tables



AutoML products announced so far



Vision



Video



Natural Language



Translation

Missing structured data!

AutoML Tables

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	...	?

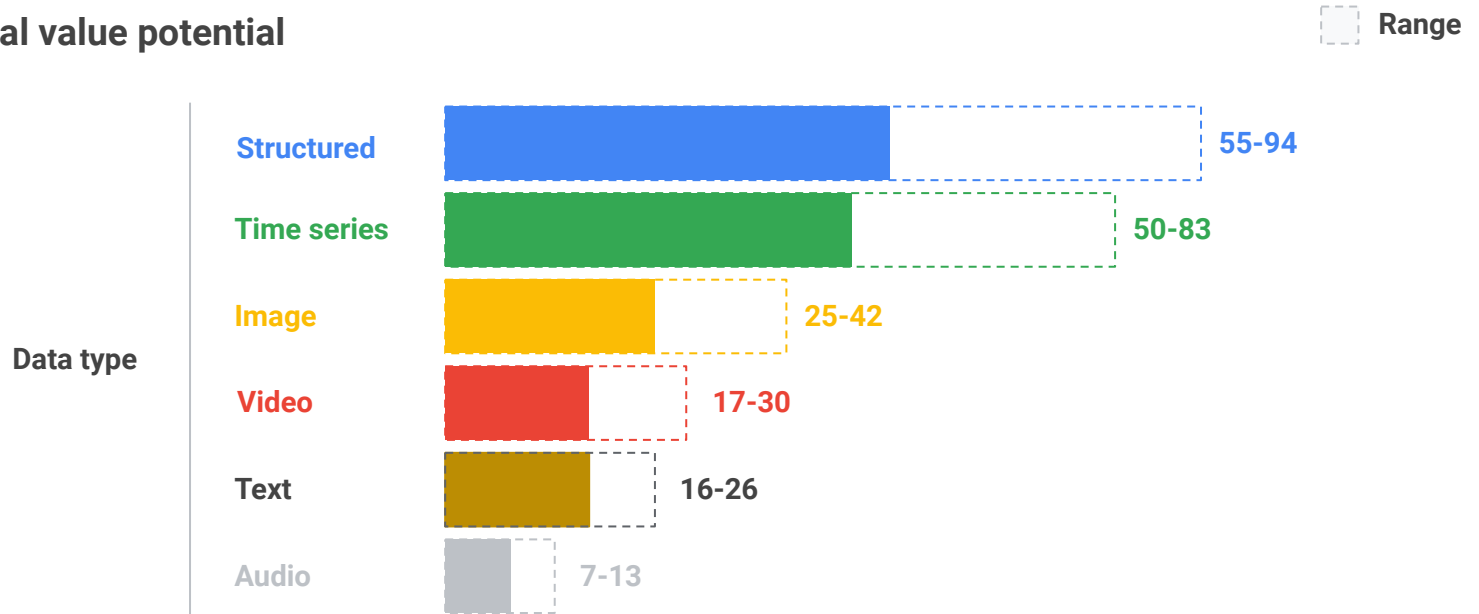
AutoML Tables

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

Structured data is likely to drive most of AI's impact

% of total value potential



Source: McKinsey Global Institute

Introducing AutoML Tables

Enable your entire team to automatically build and deploy state-of-the-art ML models on structured data at massively increased speed and scale.



Automatically search through Google's whole model zoo...

Linear, logistic

Feedforward DNN

Wide and Deep NN

Gradient Boosted Decision Tree (GBDT)

DNN + GBDT Hybrid

Adanet ensemble

Neural + Tree Architecture Search

...and more!



IMPORT

SCHEMA

ANALYZE

TRAIN

EVALUATE

PREDICT

Import your data

AutoML Tables uses tabular data that you import to train a custom machine learning model. Your dataset must contain at least one input feature column and a target column. Optional columns can be added to configure parameters like the data split, weights, etc. [Preparing your training data](#)

Table from BigQuery

The table must be in the US regional location

CSV from Cloud Storage

The bucket containing the CSV must be in the us-central1 region. [CSV formatting](#)

BROWSE

IMPORT



IMPORT

SCHEMA

ANALYZE

TRAIN

EVALUATE

PREDICT

Select a target

Select a column to be the target (what you want your model to predict) and add optional parameters like weight and time columns

Target column 

RESET













Deposit 

The selected column is categorical data. AutoML Tables will build a classification model, which will predict the target from the classes in the selected column. [Learn more](#)

Additional parameters (Optional) 

Before continuing, review your dataset schema to make sure each column has the appropriate data type and nullability setting

CONTINUE

Column name 	Variable type 	Nullability 
Age	Numeric 	<input type="checkbox"/> Nullable
Job	Categorical	<input type="checkbox"/> Nullable
MaritalStatus	Categorical	<input type="checkbox"/> Nullable
Education	Categorical	<input type="checkbox"/> Nullable
Default	Categorical	<input type="checkbox"/> Nullable
Balance	Numeric 	<input type="checkbox"/> Nullable
Housing	Categorical	<input type="checkbox"/> Nullable
Loan	Categorical	<input type="checkbox"/> Nullable
Contact	Categorical	<input type="checkbox"/> Nullable
Day	Categorical 	<input type="checkbox"/> Nullable
Month	Categorical	<input type="checkbox"/> Nullable
Duration	Numeric 	<input type="checkbox"/> Nullable
Campaign	Categorical 	<input type="checkbox"/> Nullable
PDays	Numeric 	<input type="checkbox"/> Nullable
Previous	Numeric 	<input type="checkbox"/> Nullable
POOutcome	Categorical	<input type="checkbox"/> Nullable
 Deposit Target	Categorical 	<input type="checkbox"/> Nullable

IMPORT

SCHEMA

ANALYZE

TRAIN

EVALUATE

PREDICT

⚠ Not up to date. Click the "Continue" button on the Schema tab to regenerate statistics.

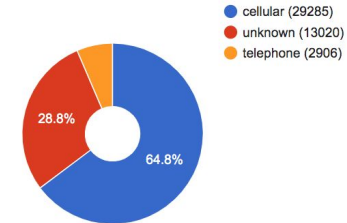
Filter instances

Feature name ↑	Type	Missing ?	Distinct values ?	Correlation with Target ?	Mean ?
All features 17					
Numeric 5					
Age	Numeric	0%	77	0.065	40,936
Balance	Numeric	0%	7,168	0.095	1,362.272
Categorical 12					
Campaign	Categorical	0%	48	0.083	---
Contact	Categorical	0%	3	0.144	---
Day	Categorical	0%	31	0.122	---
Default	Categorical	0%	2	0.028	---
Deposit	Categorical	0%	2	---	---
Duration	Numeric	0%	1,573	0.333	258,163
Education	Categorical	0%	4	0.071	---
Housing	Categorical	0%	2	0.117	---
Job	Categorical	0%	12	0.134	---
Loan	Categorical	0%	2	0.073	---
MaritalStatus	Categorical	0%	3	0.059	---
Month	Categorical	0%	12	0.245	---
PDays	Numeric	0%	559	0.181	40,198
POutcome	Categorical	0%	4	0.313	---
Previous	Numeric	0%	41	0.181	0.58

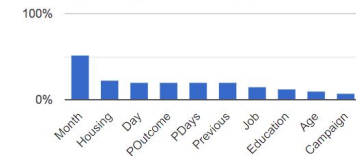
Rows per page: 50 1 - 17 of 17

Details

Distribution



Top correlated features to Contact



Train your model

Model name *
banking_20190410095716

Training budget

Enter a number between 1 and 72 for the maximum number of node hours to spend training your model. If your model stops improving before then, AutoML Tables will stop training and you'll only be charged for the actual node hours used. [Training pricing guide](#)

Budget * maximum node hours ?

Input feature selection

By default, all other columns in your dataset will be used as input features for training (excluding target, weight, and split columns).

16 feature columns *
All columns selected

Summary

Model type: Binary classification model

Data split: Automatic

Target: Deposit

Input features: 16 features

Rows: 45,211 rows

Blue Jeans Meeting

Optimization objective ▾

Depending on the outcome you're trying to achieve, you may want to train your model to optimize for a different objective. [Learn more](#)

TRAIN MODEL CANCEL

IMPORT

SCHEMA

ANALYZE

TRAIN

EVALUATE

PREDICT

Models

TRAIN MODEL

Binary classification model

banking_20190403100832



AUC PR ?

0.628

AUC ROC ?

0.936

Accuracy ?

90.98%

Log loss ?

0.195

Metrics are generated based on the less common label being the positive class.
Accuracy is based on a score threshold of 0.5

Model ID	TBL1263030997058846720
Created on	Apr 3, 2019, 10:08:38 AM
Target	Deposit
Feature columns	15 included
Test rows	4,546
Optimization objective	AUC ROC
Status	Deployed

[SEE FULL EVALUATION](#)

Binary classification model

banking_20190313051647



AUC PR ?

0.596

AUC ROC ?

0.924

Accuracy ?

90.81%

Log loss ?

0.209

Metrics are generated based on the less common label being the positive class.
Accuracy is based on a score threshold of 0.5

Model ID	TBL2539625569557938176
Created on	Mar 14, 2019, 3:06:46 PM
Target	Deposit
Feature columns	16 included
Test rows	4,546
Optimization objective	AUC ROC
Status	Deployed

[SEE FULL EVALUATION](#)

IMPORT

SCHEMA

ANALYZE

TRAIN

EVALUATE

PREDICT

Model

banking_20190403100832

Binary classification mo

Apr 3, 2019, 10:08:38 AM

Target

Deposit

Feature columns

15 included

4,546 test rows

Optimized for

AUC ROC

AUC PR ?

0.628

AUC ROC ?

0.936



Accuracy ?

91.0%

Log loss ?

0.195

Metrics are generated using the least-common class as the positive class. Accuracy based on score threshold of 0.5

→ EXPORT PREDICTIONS ON TEST DATASET TO BIGQUERY

You have up to 30 days to export your test dataset to BigQu

Filter labels



2

Score threshold 0.50

F1 score ? 0.557

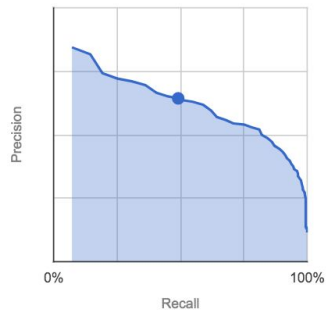
Accuracy ? 91.0% (4,136/4,546)

Precision ? 64.3% (258/401)

True positive rate (Recall) ? 49.1% (258/525)

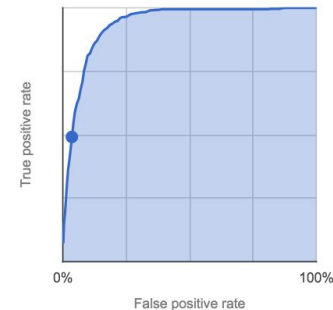
False positive rate ? 0.036 (143/4,021)

The score threshold determines the minimum level of confidence needed to make a prediction positive. [Learn more about model evaluation](#)



AUC: 0.628

PRC ?



AUC: 0.936

ROC ?

IMPORT

SCHEMA

ANALYZE

TRAIN

EVALUATE

PREDICT

BATCH PREDICTION

ONLINE PREDICTION

Model

banking_20190403100832



Your model was deployed and is available for online prediction requests. Your model size is 1,131.127 MB. [Learn more](#)

Test and use your model


Online prediction deploys your model so you can send real-time REST requests to it. Online prediction is useful for time-sensitive predictions (for example, in response to an application request). [Learn more](#)


Online prediction pricing is based on the size of your model and the length of time your model is deployed. [View pricing guide](#)

Predict label

Deposit

Prediction result

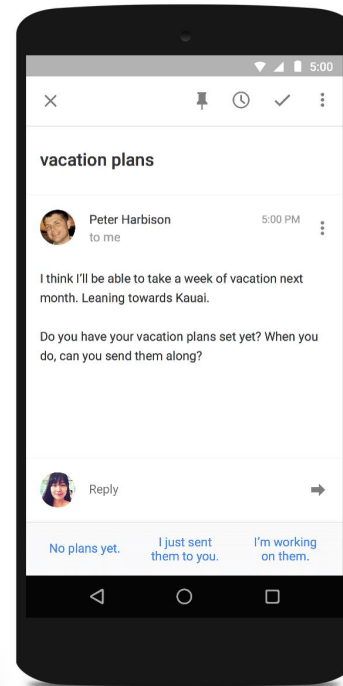
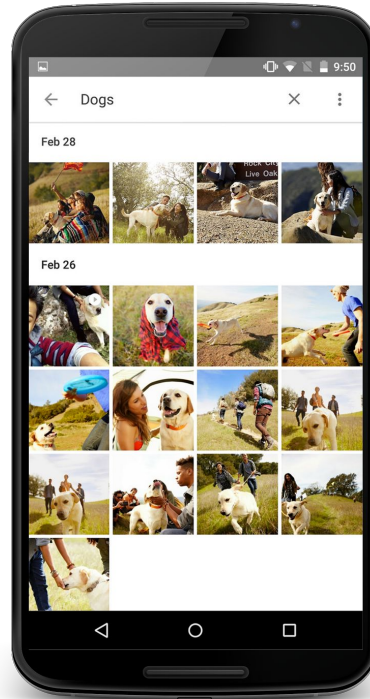
1
 Confidence score: 0.992

2
 Confidence score: 0.008

```
5     "values": [  
6         "technician",  
7         "married",  
8         "secondary",  
9         "no",  
10        "52",  
11        "no",  
12        "no",  
13        "cellular",  
14        "12",  
15        "aug",  
16        "96",  
17        "?"
```

When you hear “AI or ML,” you probably think of:

Image models
Sequence models
Neural Networks



The most common ML models at Google are those that operate on structured data

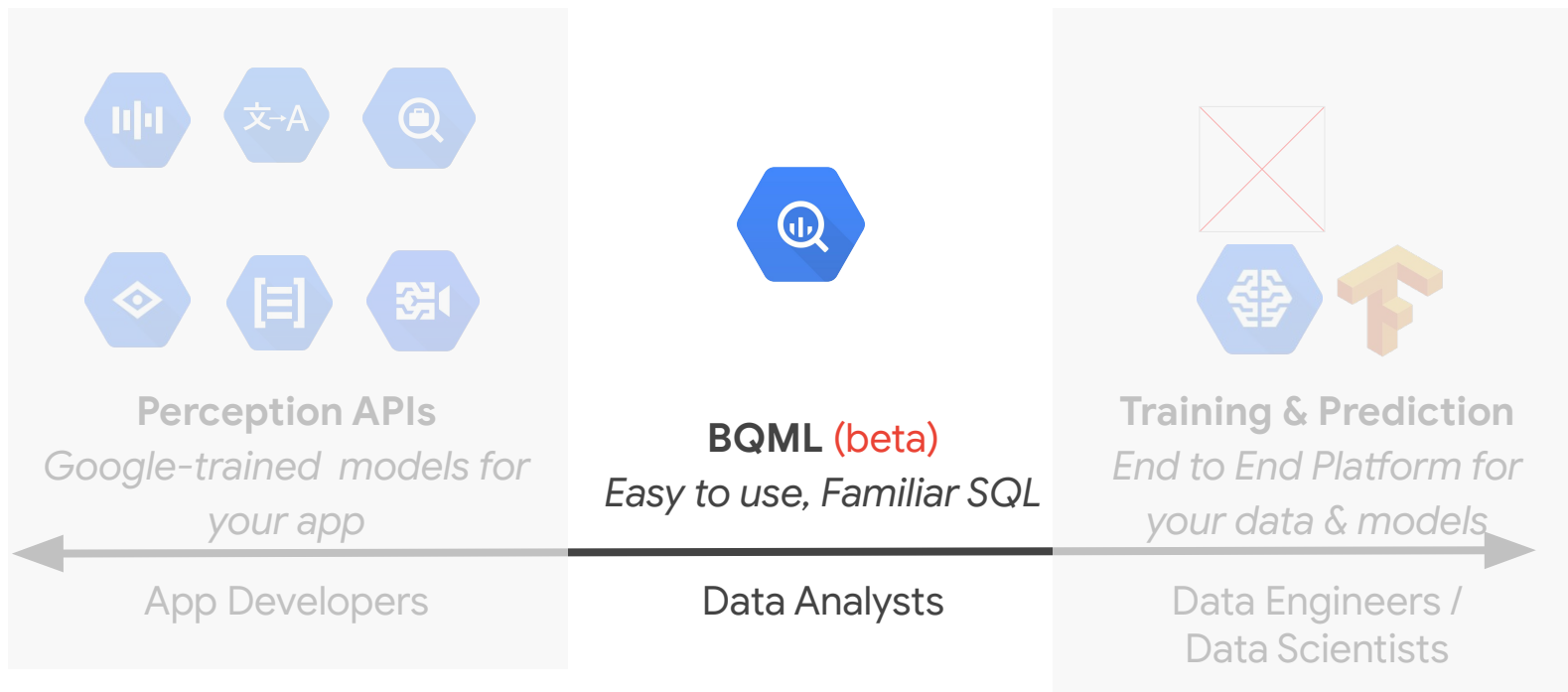
ML on structured data drives value

Type of network	# of network layers	# of weights	% of deployed models
MLP0	5	20M	61%
MLP1	4	5M	
LSTM0	58	52M	29%
LSTM1	56	34M	
CNN0	16	8M	5%
CNN1	89	100M	

It can take days to months to create an ML model



BQML is a way to easily build machine learning models



Cloud ML Engine



Working with BigQuery ML

1

Select your training data using SQL

2

Create a model, specifying model type

3

Evaluate model and verify that it meets requirements

4

Predict using model on new data from BigQuery

Working with BigQuery ML



1 Dataset

2 Create/train

3 Evaluate

4 Predict/classify

```
CREATE MODEL `bqml_tutorial.sample_model`  
OPTIONS(model_type='logistic_reg') AS  
SELECT
```

```
FROM  
ML.EVALUATE (MODEL  
`bqml_tutorial.sample_model`,  
TABLE eval_table)
```

```
FROM  
ML.PREDICT (MODEL  
`bqml_tutorial.sample_model`,  
table game_to_predict) )  
AS predict
```





Table info

Table ID	nyc-tlc:yellow.trips
Table size	129.72 GB
Long-term storage size	129.72 GB
Number of rows	1,108,779,463

pickup_datetime	dropoff_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	rate_code	passenger_count
2010-03-04 00:35:16 UTC	2010-03-04 00:35:47 UTC	-74.035201	40.721548	-74.035201	40.721548	1	1
2010-03-15 17:18:34 UTC	2010-03-15 17:18:35 UTC	0.0	0.0	0.0	0.0	1	1
2015-03-18 01:07:02 UTC	2015-03-18 01:07:07 UTC	0.0	0.0	0.0	0.0	1	5
2015-03-09 18:24:03 UTC	2015-03-09 18:25:37 UTC	-73.93724822998047	40.758201599121094	-73.93726348876953	40.7581901550293	1	1
2010-03-06 06:33:41 UTC	2010-03-06 06:36:06 UTC	-73.785514	40.6454	-73.784564	40.648681	1	2
2013-08-07 00:42:45 UTC	2013-08-07 00:58:43 UTC	-74.025817	40.763044	-74.046752	40.78324	5	1
2015-04-26 02:56:37 UTC	2015-04-26 03:00:01 UTC	-73.98765563964844	40.77165603637695	-73.98755645751953	40.771751403808594	1	1
2015-04-29 18:45:03 UTC	2015-04-29 18:49:01 UTC	0.0	0.0	0.0	0.0	1	1
2010-03-11 21:24:48 UTC	2010-03-11 21:46:51 UTC	-74.571511	40.9108	-74.628928	40.964321	1	1
2013-08-24 01:58:23 UTC	2013-08-24 01:58:23 UTC	-73.972171	40.759439	0.0	0.0	5	4

Select data



[Photo from Unsplash](#)

```
SELECT
    fare_amount,
    pickup_longitude,
    pickup_latitude,
    dropoff_longitude,
    dropoff_latitude,
    passenger_count

FROM
    `nyc-tlc.yellow.trips`
```

Build and train with **CREATE MODEL**



[Photo from Unsplash](#)

```
CREATE OR REPLACE MODEL
  mydataset.model_linreg

OPTIONS(
  input_label_cols=['fare_amount'],
  model_type='linear_reg') AS

SELECT
  fare_amount,
  pickup_longitude,
  pickup_latitude,
  dropoff_longitude,
  dropoff_latitude,
  passenger_count

FROM
  `nyc-tlc.yellow.trips`
```

Evaluate with ML.EVALUATE



[Photo from Unsplash](#)

```
SELECT
  *

FROM
  ML.EVALUATE(
    MODEL mydataset.model_linreg
  )
```

Use the model with ML.PREDICT



[Photo from Unsplash](#)

```
SELECT
  *

FROM
  ML.PREDICT(MODEL mydataset.model_linreg,
  (
    SELECT
      fare_amount,
      pickup_longitude,
      pickup_latitude,
      dropoff_longitude,
      dropoff_latitude,
      passenger_count
    FROM
      `nyc-tlc.yellow.trips`
  ))
```

Supported BigQuery ML models

Classification

- Logistic regression
- DNN classifier (TensorFlow)
- XGBoost
- AutoML Tables

Other Models

- k-means clustering
- Time series forecasting
- Recommendation: Matrix factorization

Regression

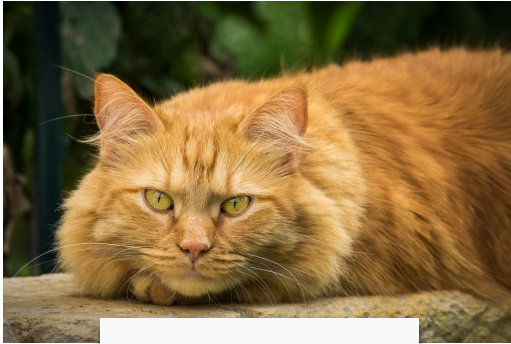
- Linear regression
- DNN regressor (TensorFlow)
- XGBoost
- AutoML Tables

Model Import/Export

- TensorFlow models for batch and online prediction

What about custom tasks?

Generic Task



“cat”

Custom Task



“Maine coon”

Healthy or Pneumonia?

Normal



Bacterial Pneumonia



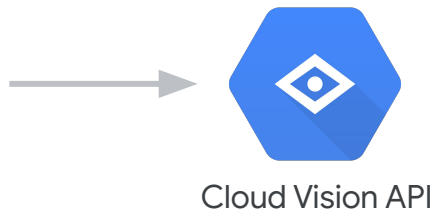
Viral Pneumonia



Scenario 1: cat or not?



Input



Model

```
"labelAnnotations": [  
  {  
    "mid": "/m/01yrx",  
    "description": "cat",  
    "score": 0.9925717  
  },  
  ...  
]
```

Prediction

Scenario 2: what breed is this cat?



Input



?

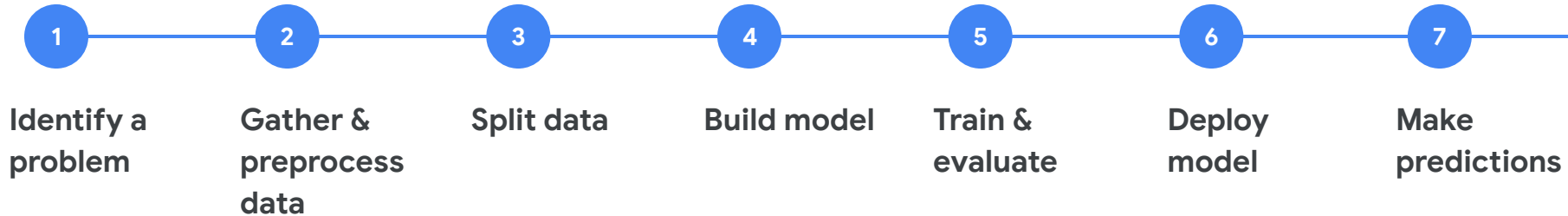
Model



```
"predictions": [  
  {  
    "maine_coon": 0.97  
  },  
  ...  
]
```

Prediction

Scenario 2: building a cat breed prediction model



Step 2: Gather & preprocess data

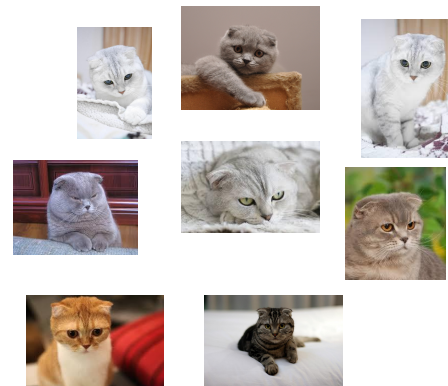
Maine Coon



Abyssinian



Scottish Fold



1

Identify a problem

2

Gather & preprocess data

3

Split data

4

Build model

5

Train & evaluate

6

Deploy model

7

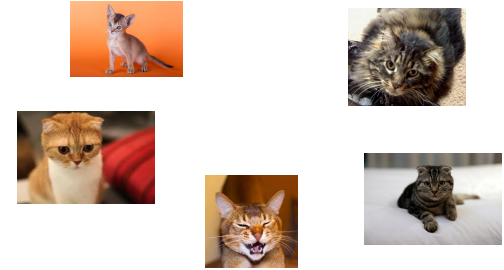
Make predictions

Step 3: Split data

Train



Test



1

Identify a problem

2

Gather & preprocess data

3

Split data

4

Build model

5

Train & evaluate

6

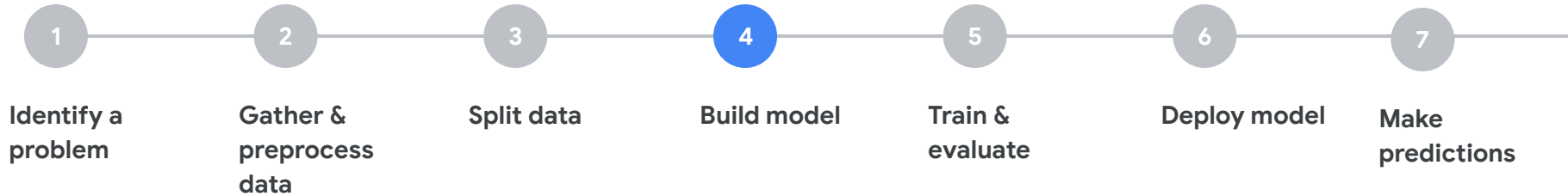
Deploy model

7

Make predictions

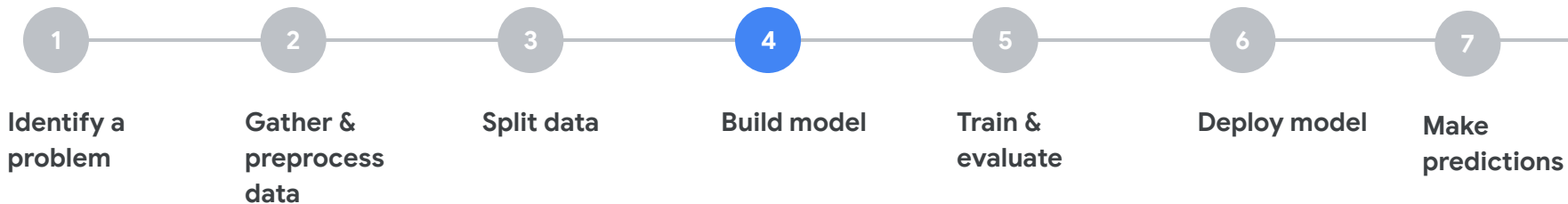
Step 4: Build model

```
from tensorflow import keras
from keras.models import Sequential
from keras.layers import Dense, Dropout, Flatten
from keras.layers import Conv2D, MaxPooling2D
```



Step 4: Build model

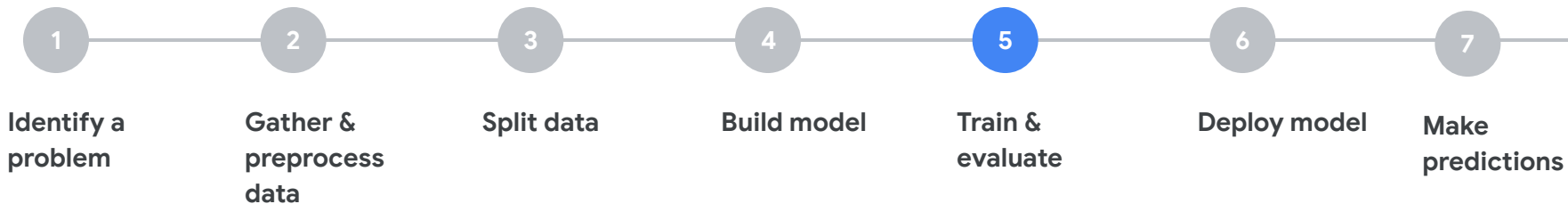
```
model = Sequential()  
model.add(Conv2D(32, kernel_size=(3, 3),  
                activation='relu',  
                input_shape=input_shape))  
model.add(Conv2D(64, (3, 3), activation='relu'))  
model.add(MaxPooling2D(pool_size=(2, 2)))  
model.add(Flatten())  
model.add(Dense(128, activation='relu'))  
model.add(Dense(num_classes, activation='softmax'))
```



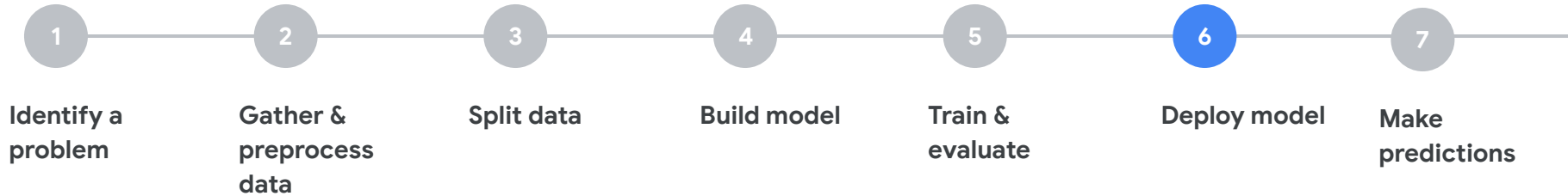
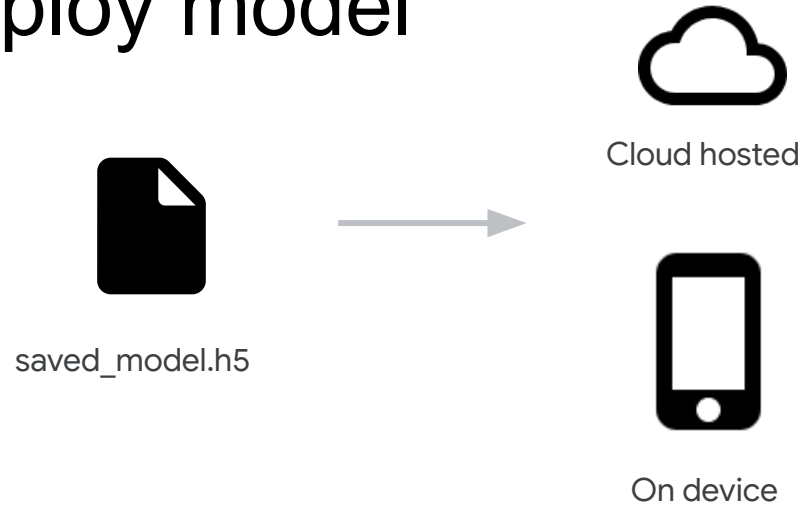
Step 5: Train & evaluate

```
# Train
model.fit(x_train, y_train,
          batch_size=batch_size,
          epochs=epochs,
          verbose=1,
          validation_data=(x_test, y_test))

# Evaluate
score = model.evaluate(x_test, y_test)
```

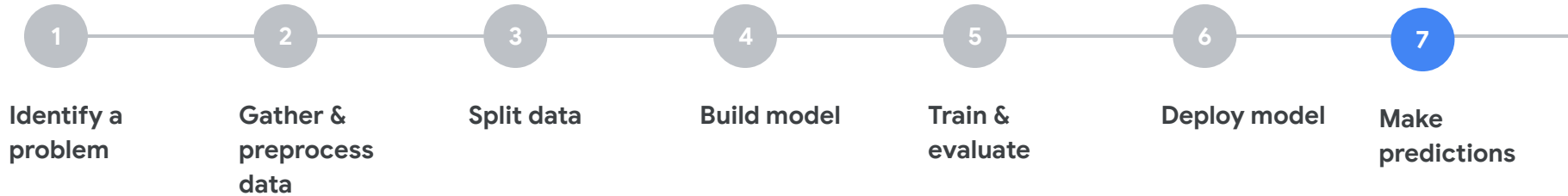


Step 6: Deploy model

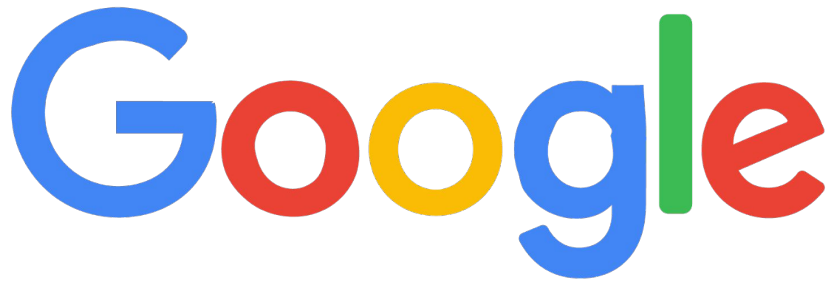


Step 7: Make predictions

```
model.predict(image_array)
```



New Google Training Portal for Nimbus Bookmark!



Register for Google Webinars in the Nimbus Bootcamp



Google Cloud
Fundamentals: Core
Infrastructure

Oct 14 | 10:00-13:00



Google Cloud
Fundamentals: Big Data &
Machine Learning

Oct 28 | 10:00-13:00



Fundamentals of
Security in
Google Cloud

Nov 4 | 10:00-13:00



Google Cloud
Digital Leader
(tech and non-tech)

Nov 18 | 10:00-13:00

googlecloud.folloze.com/nimbus



Thank you

Google Cloud