A quick overview of Google Cloud Platform for Data Science & ML

Amy Unruh
Google Cloud Platform
Developer Relations Engineer
@amygdala (@amy on Astro Hack slack team)


Compute
Compute Engine
App Engine
Container Engine
Container Registry
Cloud Functions
Networking
Cloud Virtual Network
Cloud Load Balancing
Cloud CDN
Cloud Interconnect
Cloud DNS
Storage and Databases
Cloud Storage
Cloud Bigtable
Cloud Datastore
Cloud SQL
Persistent Disk
Big Data
BigQuery
Cloud Dataflow
Cloud Dataproc
Cloud Datalab
Cloud Pub/Sub
Genomics
Machine Learning
Cloud Machine Learning
Vision API
Speech API
Natural Language API
Translation API
Jobs API
Developer Tools
Cloud SDK
Deployment Manager
Cloud Source Repositories
Cloud Tools for Android Studio
Cloud Tools for IntelliJ
Cloud Tools for PowerShell
Cloud Tools for Visual Studio
Google Plug-in for Eclipse
Cloud Test Lab
Management Tools
Stackdriver
Monitoring
Logging
Error Reporting
Trace
Debugger
Deployment Manager
Cloud Endpoints
Cloud Console
Cloud Shell
Cloud Mobile App
Billing App
Cloud APIs
Cloud IAM
Cloud Resource Manager
Cloud Security Scanner
Cloud Platform Security


Compute
Compute Engine
App Engine
Container Engine
Container Registry
Cloud Functions
Networking
Cloud Virtual Network
Cloud Load Balancing
Cloud CDN
Cloud Interconnect
Cloud DNS
Storage and Databases
Cloud Storage
Cloud Bigtable
Cloud Datastore
Cloud SQL
Persistent Disk
Big Data
BigQuery
Cloud Dataflow
Cloud Dataproc
Cloud Datalab
Cloud Pub/Sub
Genomics
Machine Learning
Cloud Machine Learning
Vision API
Speech API
Natural Language API
Translation API
Jobs API
Developer Tools
Cloud SDK
Deployment Manager
Cloud Source Repositories
Cloud Tools for Android Studio
Cloud Tools for IntelliJ
Cloud Tools for PowerShell
Cloud Tools for Visual Studio
Google Plug-in for Eclipse
Cloud Test Lab
Management Tools
Stackdriver
Monitoring
Logging
Error Reporting
Trace
Debugger
Deployment Manager
Cloud Endpoints
Cloud Console
Cloud Shell
Cloud Mobile App
Billing App
Cloud APIs
Cloud IAM
Cloud Resource Manager
Cloud Security Scanner
Cloud Platform Security

A tour and some demos

(Ask me about GCP credits)

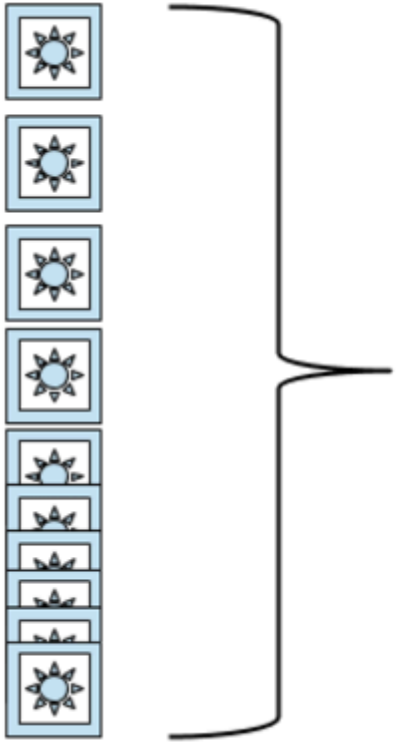
Compute Engine and Container Engine

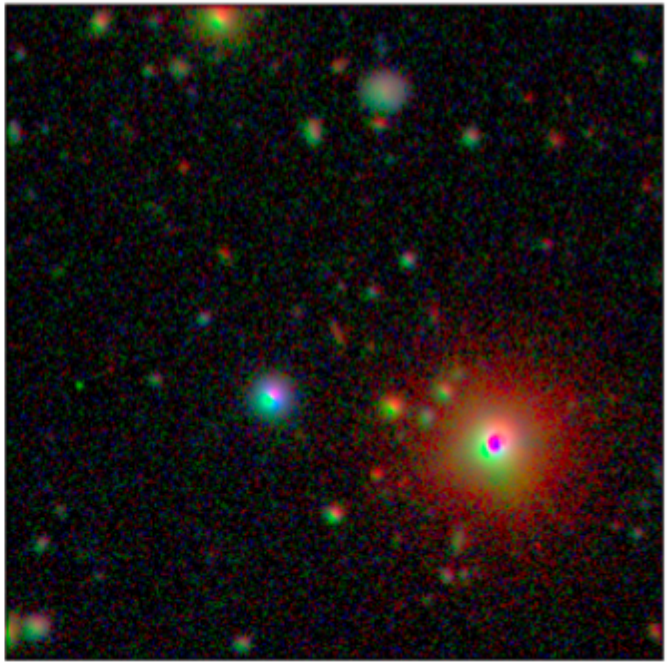
Compute Engine (VMs)

- GPUs: NVIDIA® K80s
  - AMD FirePro S9300 x2 and NVIDIA® Tesla® P100 coming soon
- 1 - 64 processors
- .6 - 416 GB (.9 - 6.5 GB per proc)
- Disk limits
  - 65 TB Persistent, 3 TB SSD, 208 RAM
- Preconfigured images, + can build your own
- Spin up in 10s of seconds

“Compute the Cosmos with GCE”

`https://goo.gl/UmAALC`

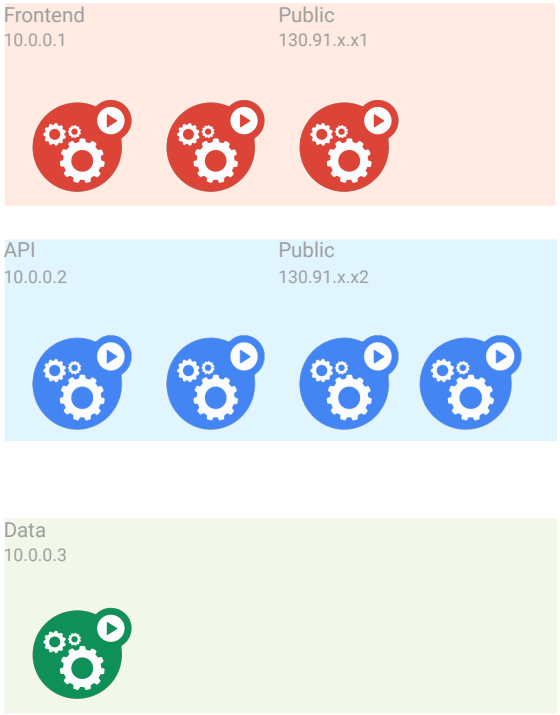




(Process a tile from the Stripe 82 database collected by the SDSS, using the LSST Software Stack)

Kubernetes and Container Engine

- https://kubernetes.io/
- Container Orchestration System
- Allows for running a large number of containers in production
- Organizing them into workloads
    - Microservices
    - Scheduled jobs
    - Batch jobs
- Supports GPU-enabled clusters
- **Google Container Engine: hosted Kubernetes**



Cloud Dataproc

- **Managed YARN Cluster for Hadoop, Spark, Pig, Hive**
- Deploy clusters in 90 seconds and pay per-minute
- Add or remove workers with zero downtime
- Can customize OSS Apache Bigtop image w/initialization actions

BigQuery and GCS

What is BigQuery?

- Fully managed, **no-ops** SQL data warehouse
- Processes **terabytes** of data
  in **tens of seconds**
- The **economy** of cloud
- **ANSI SQL 2011** Compliant, + **User-Defined functions**,
  + rich set of supported functions and operators (incl. statistical aggregate functions)

Structured Data → BigQuery

BigQuery improves access to structured datasets

- Structured data warehouse
- Public data is globally available - `` `allAuthenticatedUsers` ``
- Serverless access - queries via console or API
- Hosters - 10GB **free** storage
- Users - 1TB/mo of **free** queries, additional queries $5/TB
- JOIN multiple datasets (e.g., weather + health + census)

Demo:
BigQuery (& public data sets)

Unstructured Data → GCS

Cloud Storage improves hosting of objects/files

- Object storage - images/files/blobs
- Google pays for hosting
- Globally available - allUsers
- No egress charged to data provider

BigQuery + GCS

Example Structure: The Metropolitan Museum

- 400k+ open art images in GCS
- Meta-data / annotations in BigQuery Table
- Vision API results in BigQuery Table

Ask BigQuery (remember the free tier?)

- I want all the objects available from the "roman" period
- I want all the objects that include the color "blue"
- I want all the objects of "armor"

No need to download Terabytes of images,
get the ones you need.

Public Datasets Program

Democratizing Access to Planetary-scale Data

Google's Core Mission - organize the world's information and make it universally accessible and useful.

Extended to the cloud ...

Public Datasets Program - organize the world's data and make it more universally **accessible and useful** (in the cloud).

- > 2PB in size
- > 5B records/objects
- > 20PB queried (in BigQuery)
- Live tables
- Multi-region
- Not a data-mirror, but focused on improving access and utility

And growing ...

Apache Beam and Cloud Dataflow

Apache Beam and Cloud Dataflow

- ETL, analytics, data pre-processing
- **Unified model** for both **stream** and **batch** processing
- **Event-time processing**
- Cloud Dataflow: **Managed service** with **autoscaling** capabilities

Apache Beam & Cloud Dataflow

Apache Beam is a collection of SDKs for **building** streaming data processing pipelines.

Cloud Dataflow is a fully managed (no-ops) integrated service for **executing** optimized parallelized data processing pipelines.

Apache Beam's Programming Model

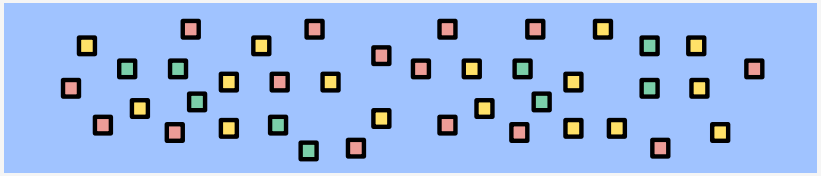
What results are you computing?

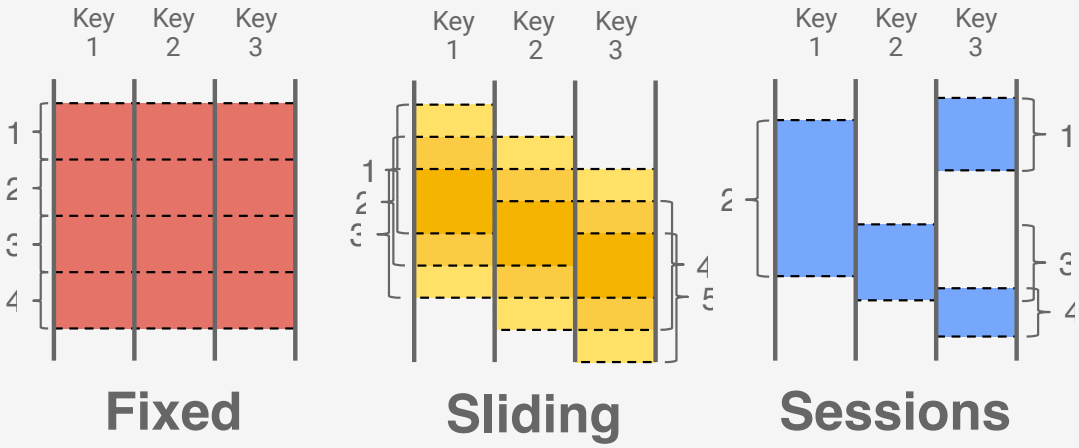
Where in event time are results calculated?

When in processing time are results materialized?

How do refinements of results relate?

What is Windowing in Apache Beam?





Windowing partitions data based on the **timestamps** associated with **events**

Demo: Dataflow twitter analytics pipeline + BigQuery

https://github.com/amygdala/gae-dataflow

Tracking emergent Twitter bigrams

```
SELECT min(ts) as min_ts, w1, w2, FROM [aju-vtests3:gae_dataflow.word_cooccur]
group by  w1, w2 order by min_ts desc, w1, w2 LIMIT 1000
```

No Cached Results | RUN QUERY | Save Query | Save View | Format Query | Show Options | Query complete (2.1s elapse

Results | Explanation | Job Information | Download as CSV | Download as JSON

Row	min_ts	w1	w2
1	2017-08-27 22:02:00 UTC	cross	red
2	2017-08-27 22:02:00 UTC	disaster	natural
3	2017-08-27 22:02:00 UTC	flooding	texas
4	2017-08-27 22:02:00 UTC	from	houston
5	2017-08-27 22:02:00 UTC	from	texas
6	2017-08-27 22:02:00 UTC	harvey	help
7	2017-08-27 22:02:00 UTC	harvey	rain
8	2017-08-27 22:02:00 UTC	harvey	trump
9	2017-08-27 22:02:00 UTC	harvey	weather
10	2017-08-27 22:02:00 UTC	help	please
11	2017-08-27 22:02:00 UTC	inches	rain
12	2017-08-27 22:02:00 UTC	profiling	racial
13	2017-08-27 22:02:00 UTC	safe	stay
14	2017-08-27 17:01:59 UTC	catastrophic	flooding
15	2017-08-27 17:01:59 UTC	flooding	houston
16	2017-08-27 17:01:59 UTC	harvey	houston
17	2017-08-27 14:28:00 UTC	about	arpaio
18	2017-08-27 14:28:00 UTC	flooding	harvey
19	2017-08-27 14:28:00 UTC	identity	politics
20	2017-08-27 14:28:00 UTC	pardoned	trump
21	2017-08-27 09:28:01 UTC	cameron	james
22	2017-08-27 09:28:01 UTC	conor	floyd
23	2017-08-27 09:28:01 UTC	conor	mayweather
24	2017-08-27 09:28:01 UTC	floyd	mayweather
25	2017-08-27 09:28:01 UTC	floyd	mcgregor
26	2017-08-27 09:28:01 UTC	harvey	will
27	2017-08-27 09:28:01 UTC	hurricane	trump
28	2017-08-27 09:28:01 UTC	i'm	like
29	2017-08-27 09:28:01 UTC	letter	resignation
30	2017-08-27 04:28:52 UTC	from	read
31	2017-08-27 04:28:52 UTC	gorka	white
32	2017-08-27 03:13:46 UTC	hurricane	major

Jupyter notebooks, Cloud Datalab

https://github.com/googledatalab/datalab

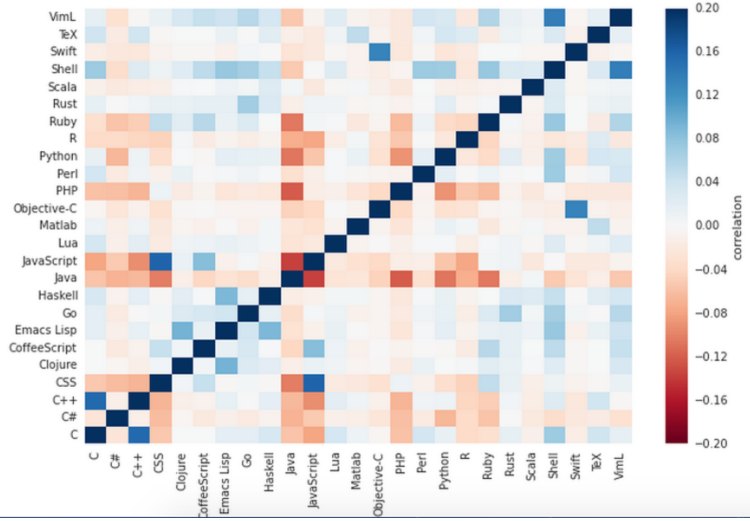
Google Cloud Datalab Programming Language Correlation (autosaved)

Notebook Add Code Add Markdown Delete Move Up Move Down Run Clear Reset Session

BigQuery, SQL + Python, Analysis + Visualization

```
%%sql --module commits
SELECT user, language, pushes
FROM (SELECT timeline.actor AS user, timeline.repository_language AS language,
             COUNT(timeline.repository_language) AS pushes
FROM [githubarchive:github.timeline] AS timeline
JOIN (SELECT repository_language AS language, COUNT(repository_language) as pushes
      FROM [githubarchive:github.timeline]
      WHERE type = 'PushEvent' AND repository_language != '' AND
            PARSE_UTC_USEC(created_at) >= DATE_ADD(CURRENT_TIMESTAMP(), -1, 'YEAR')
      GROUP BY language
      ORDER BY pushes DESC
      LIMIT 25) AS languages
ON timeline.repository_language = languages.language
WHERE type = 'PushEvent' AND PARSE_UTC_USEC(created_at) >= DATE_ADD(CURRENT_TIMESTAMP(), -1, 'YEAR')
GROUP BY user, language)
WHERE ABS(HASH(user)) % 100 < 5
```

```
commits.to_dataframe().pivot(index='user', columns='language', values='pushes').fillna(0)
       .corr(method = 'spearman').plot()
```



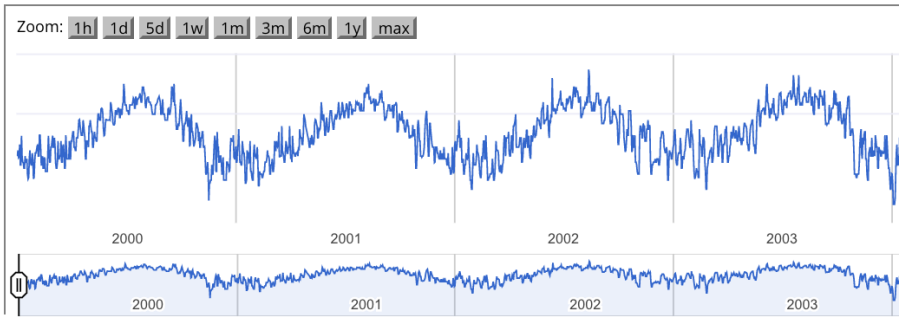
Cloud Datalab: Jupyter Notebooks with GCP integrations

Integrated Google Charts

TimeSeries Chart

```
%%bq query -n weather
SELECT max_temperature AS temperature,
       SAFE_CAST(CONCAT(SAFE_CAST(year AS STRING), '-', SAFE_CAST(mo
FROM `publicdata.samples.gsod`
WHERE station_number = 727930 AND year >= 2000
ORDER BY year DESC, month DESC, day DESC
```

```
%%chart annotation --fields timestamp,temperature --data weather
```



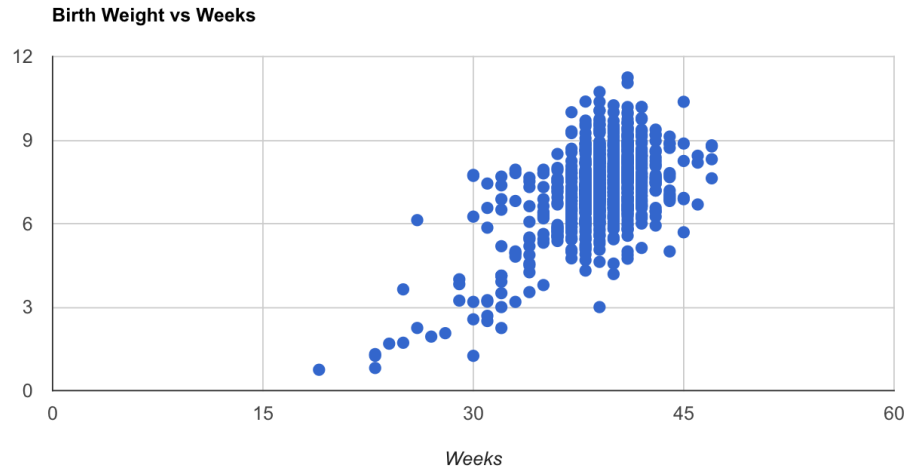
Google Cloud Datalab — Interactive Charts with Google Charting APIs

Notebook ▾ | Run ▾ | Clear ▾ | Reset ▾ | python2 ▾

Scatter Chart

```
%%bq query -n births
SELECT gestation_weeks AS weeks, weight_pounds AS weight
FROM `publicdata.samples.natality`
WHERE gestation_weeks < 99
LIMIT 1000
```

```
%%chart scatter --data births
title: Birth Weight vs Weeks
height: 400
width: 900
hAxis:
  title: Weeks
vAxis:
  title: Weight
legend: none
```

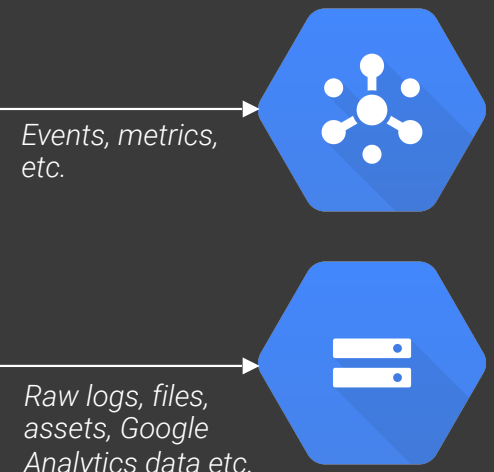


demo: Datalab with BigQuery and Facets

(Inspecting the ‘NHTSA Fatalities’ public data set)

https://pair-code.github.io/facets/ ,
https://github.com/amygdala/code-snippets/tree/master/datalab/facets

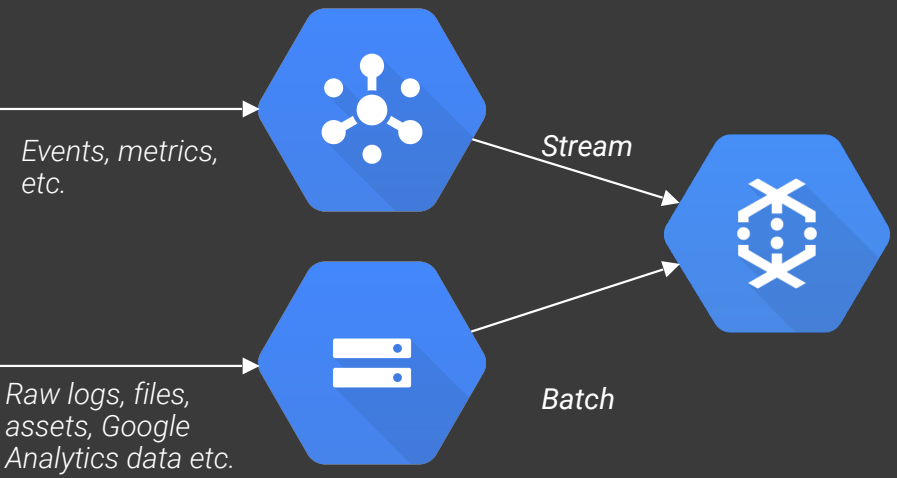
A common configuration: *ingest and store*



Cloud Pub/Sub
Serverless, global and fully managed async messaging platform

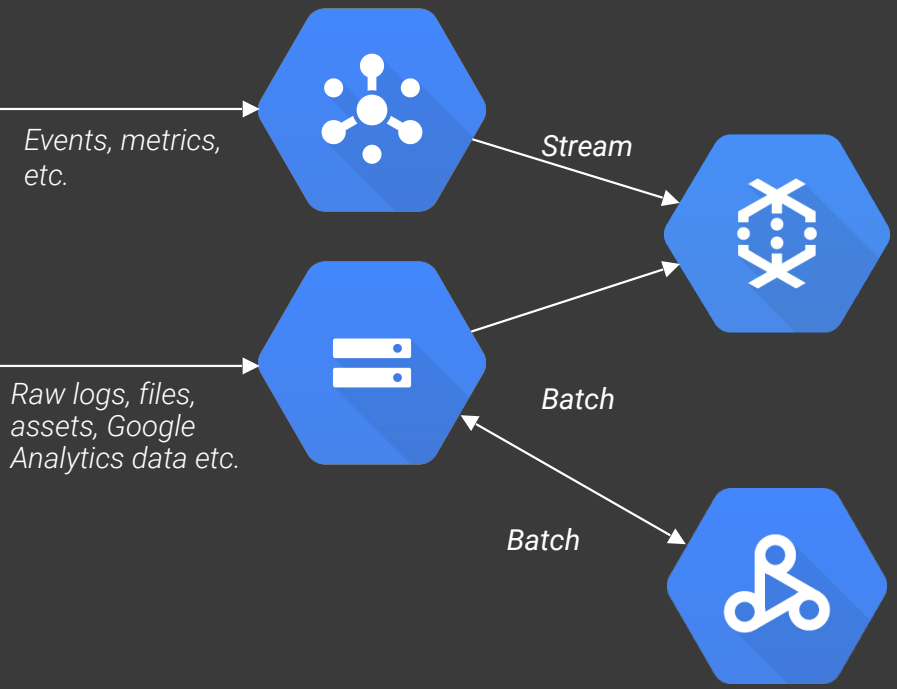
Cloud Storage
Performant, unified and cost-effective Object Storage for binary and blob data

A common configuration: *process and transform*



Cloud Dataflow
Data processing framework for batch and stream processing

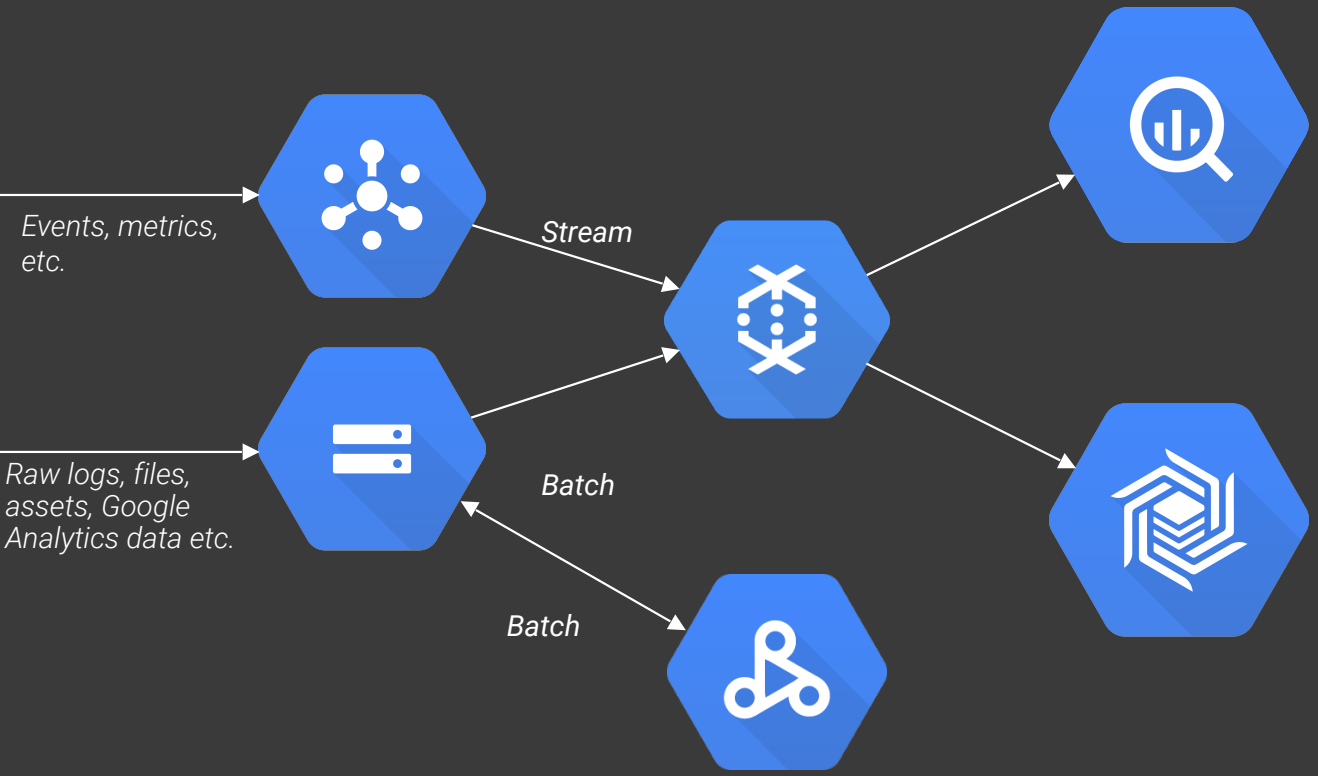
A common configuration: *process and transform*



Cloud Dataflow
Data processing framework for batch and stream processing

Cloud Dataproc
Managed Spark and Hadoop

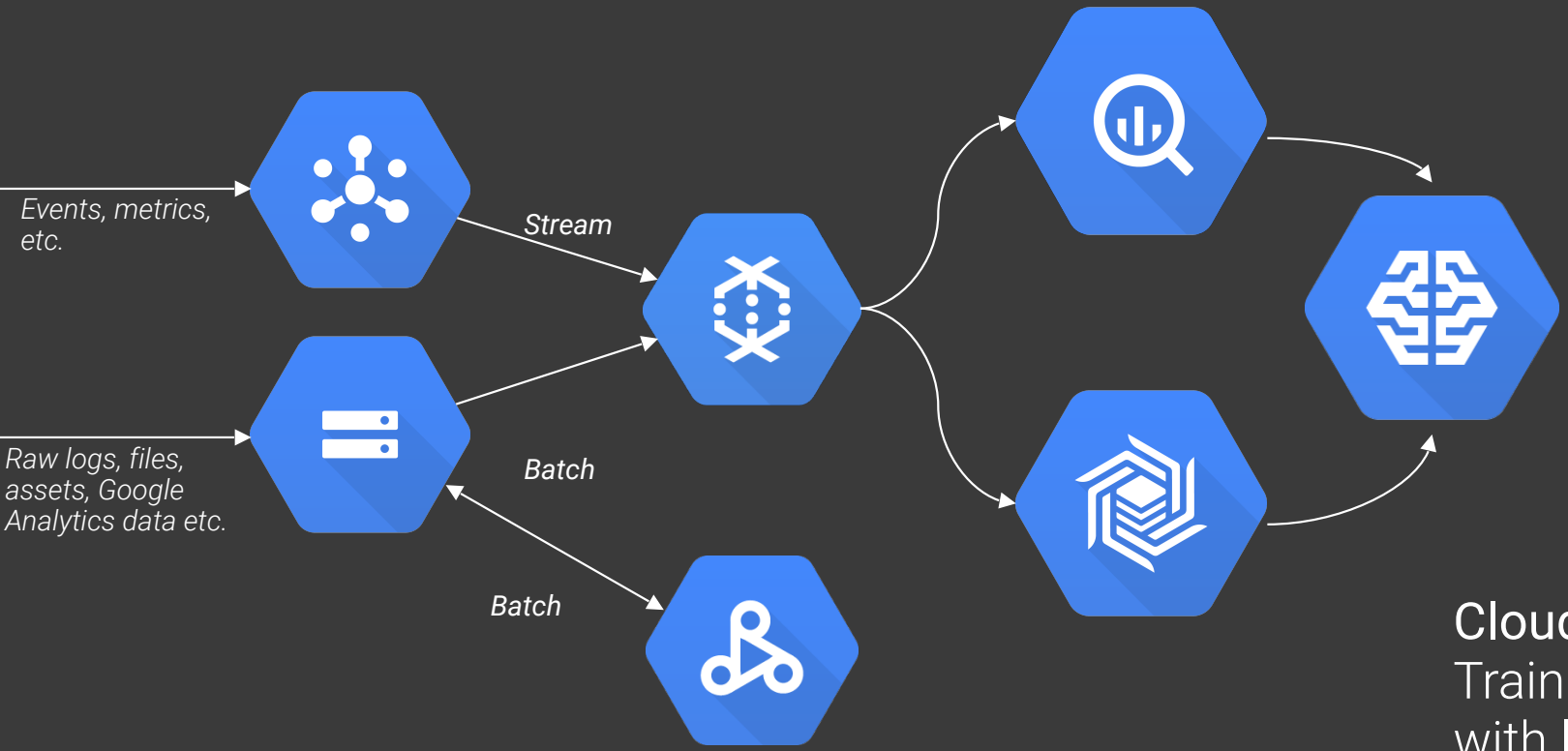
A common workflow: *analyze and store*



BigQuery
Extremely fast and cheap on-demand analytics engine

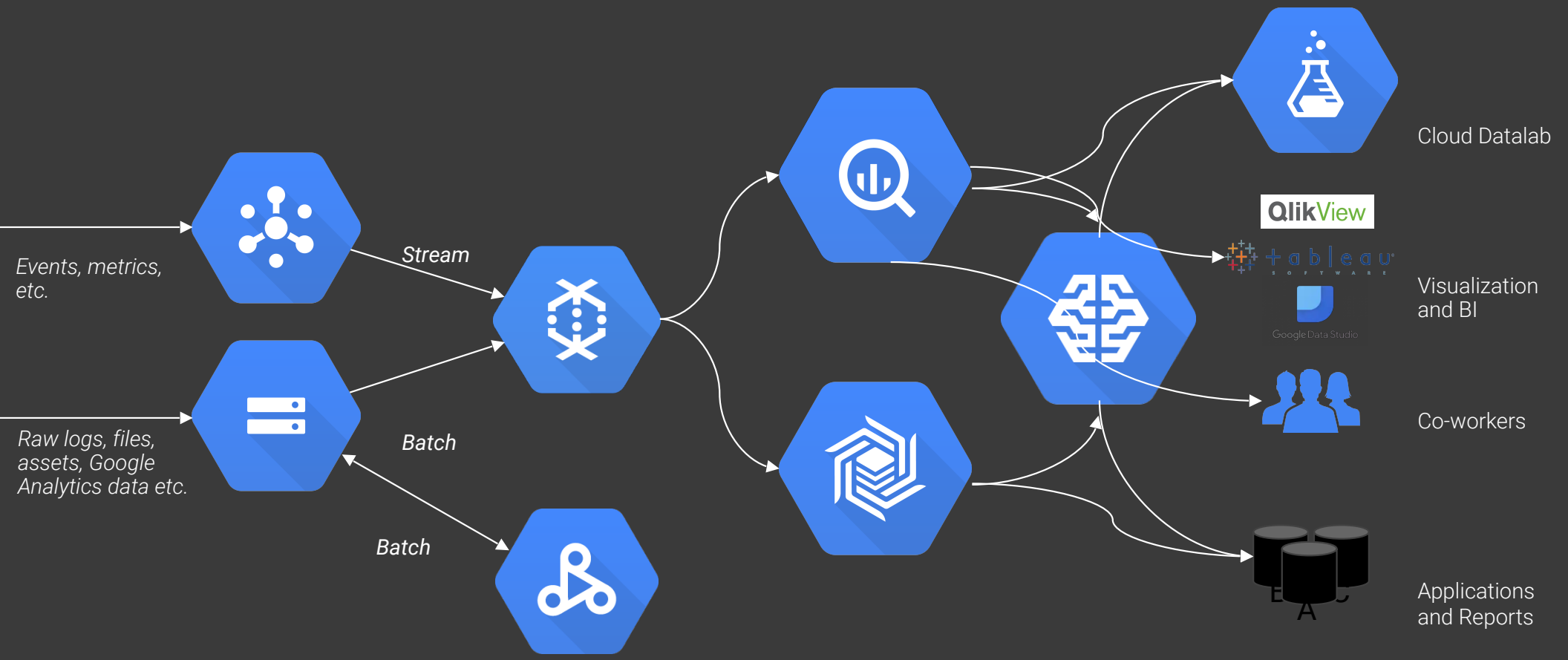
Bigtable
High performance NoSQL database for large workloads

A common workflow: analyze, *learn and recommend*

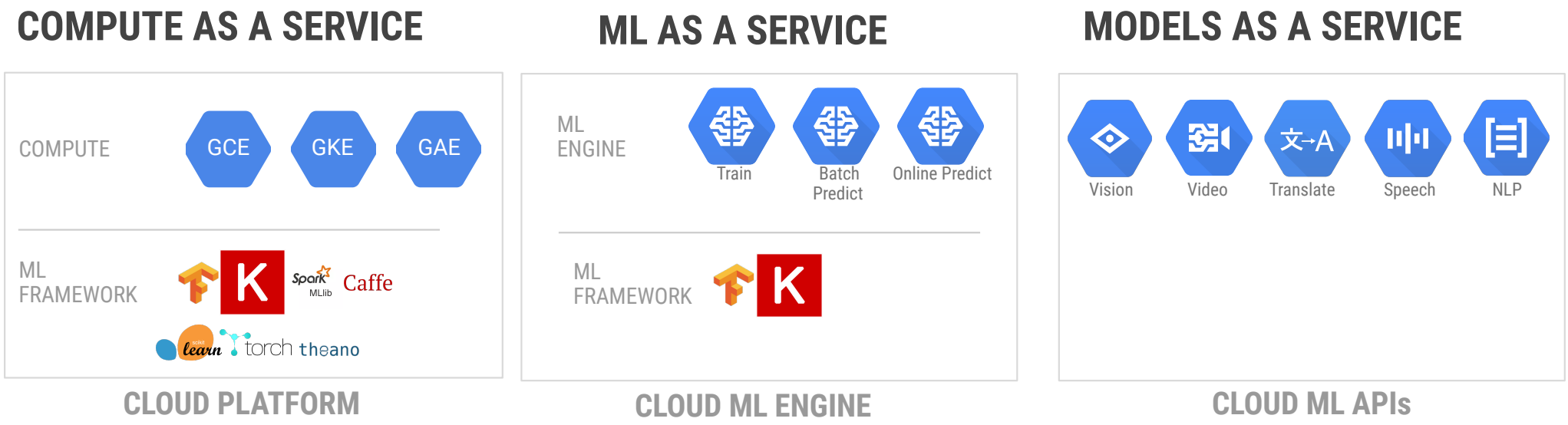


Cloud Machine Learning
Train your own models with large datasets

A common workflow: *share and visualize*

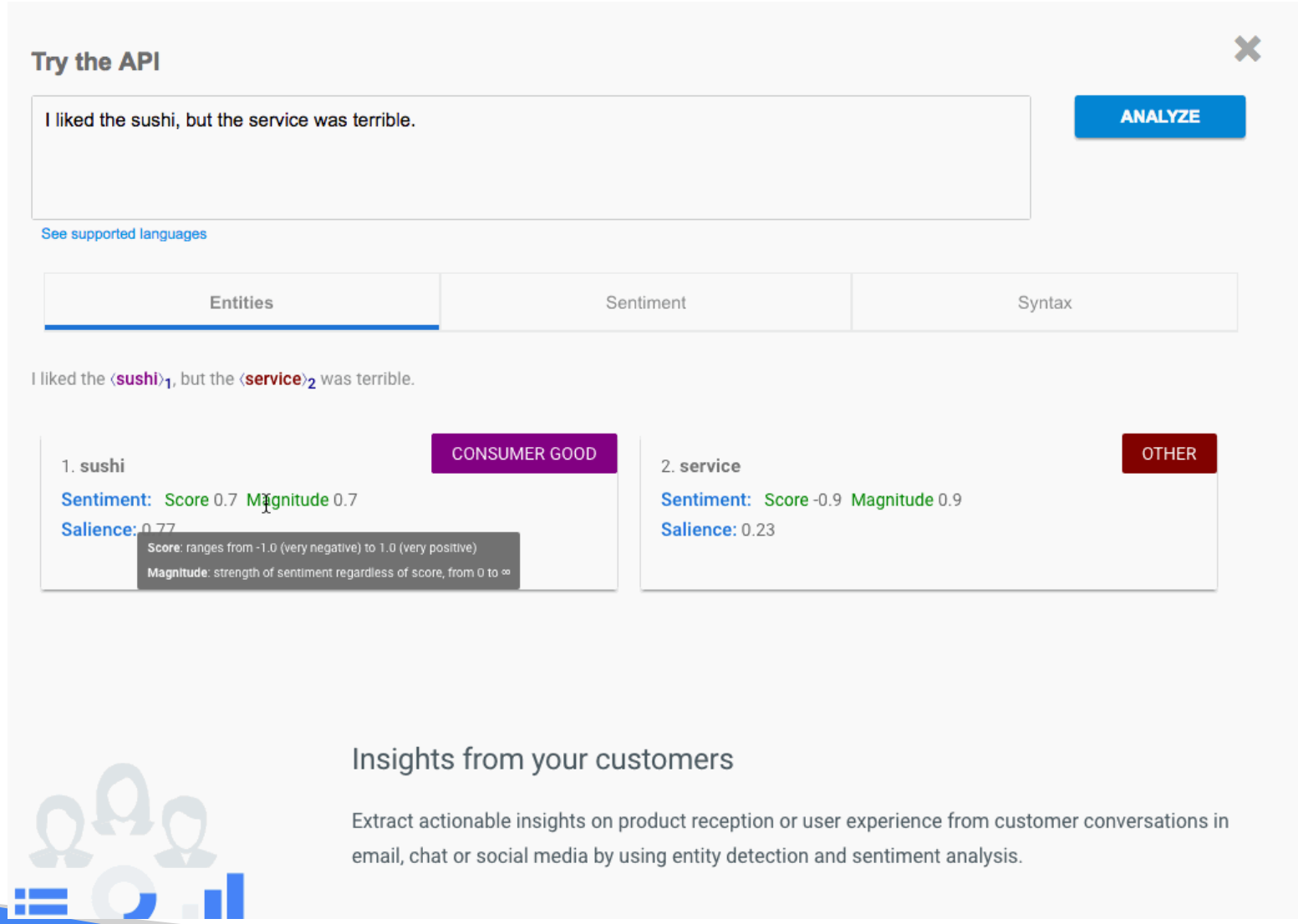


Machine Learning on Google Cloud Platform: A spectrum of offerings



Natural Language API

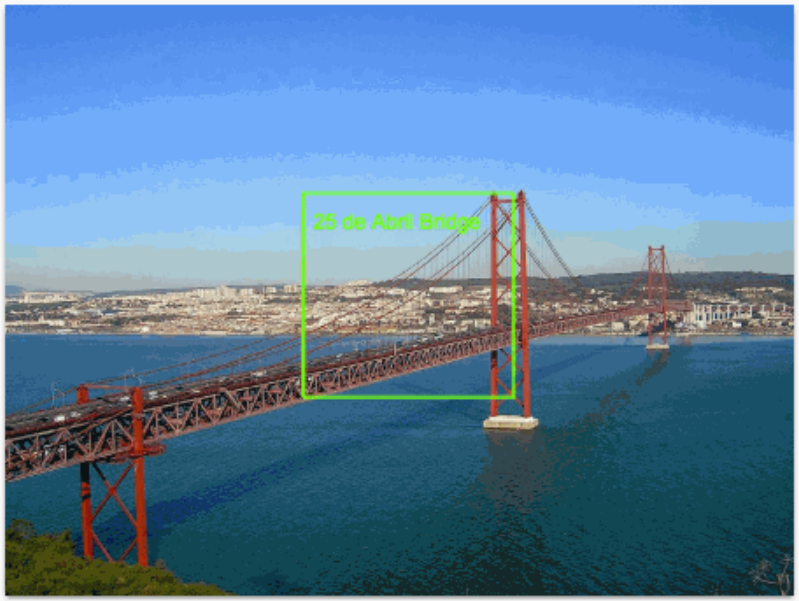
Extract entities, sentiment, and syntax from text


Try the API
I liked the sushi, but the service was terrible.
ANALYZE
See supported languages
Entities
Sentiment
Syntax
I liked the ⟨sushi⟩1, but the ⟨service⟩2 was terrible.
1. sushi
CONSUMER GOOD
Sentiment: Score 0.7 Magnitude 0.7
Salience: 0.77
Score: ranges from -1.0 (very negative) to 1.0 (very positive)
Magnitude: strength of sentiment regardless of score, from 0 to ∞
2. service
OTHER
Sentiment: Score -0.9 Magnitude 0.9
Salience: 0.23
Insights from your customers
Extract actionable insights on product reception or user experience from customer conversations in email, chat or social media by using entity detection and sentiment analysis.

Vision API

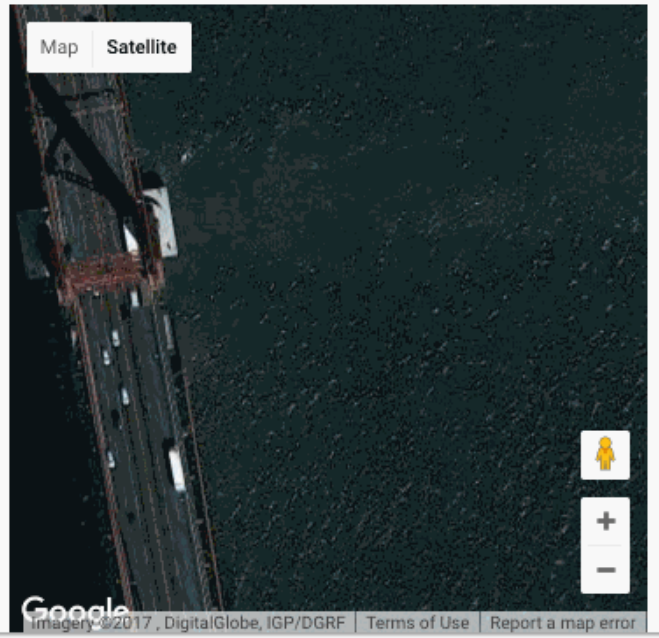
Complex image detection
with a simple REST request

Landmarks | Labels | Web | Text | Document | Properties | Safe Search | JSON



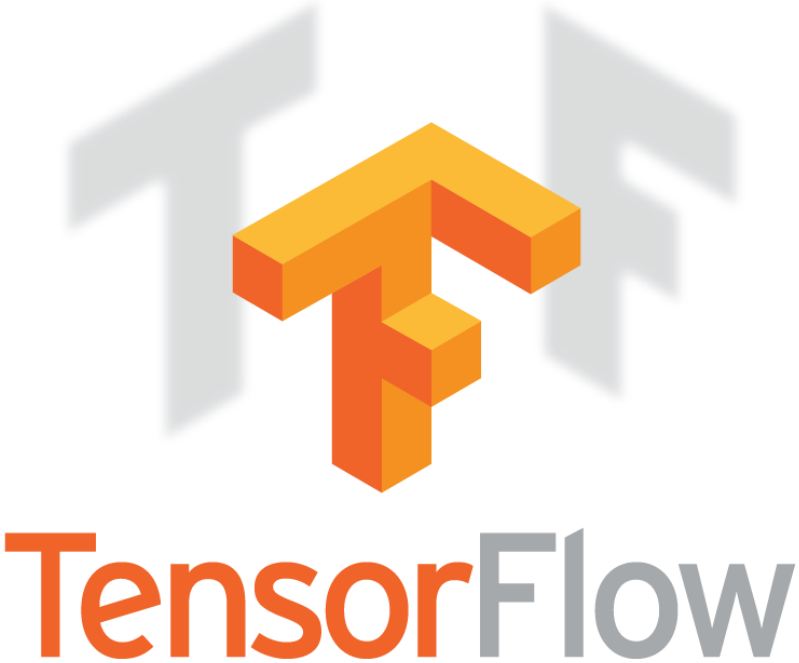
1200px-Ponte25Abril1.jpg

25 de Abril Bridge 52%



Cloud Machine Learning Engine (ML Engine)

- Fully managed **distributed training** and **prediction**
- Scales to **tens of CPUs and GPUs**
- Supports custom **TensorFlow** graphs
- **HyperTune** for hyper-parameter tuning automation
- **General Availability** - cloud.google.com/**ml**



tensorflow.org

- Fast, flexible, and scalable open-source machine learning library
- For research and production
- Apache 2.0 license

https://research.googleblog.com/2017/02/announcing-tensorflow-10.html

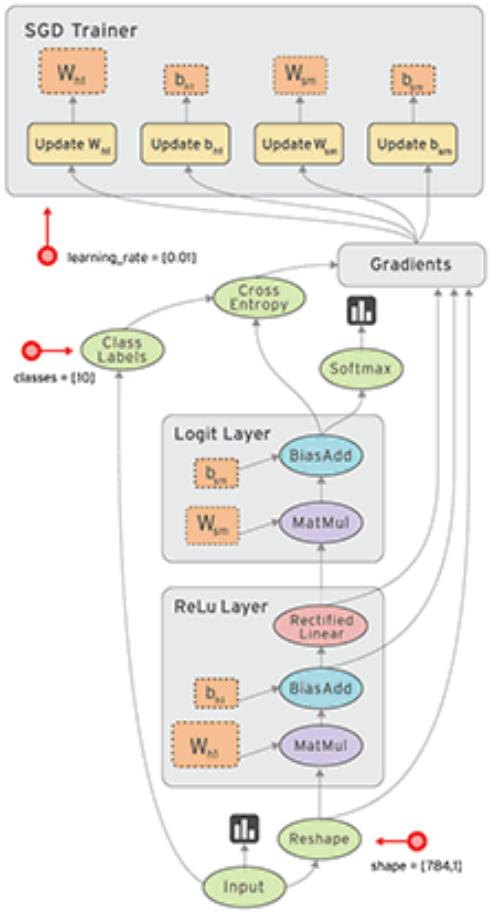
TensorFlow

A multidimensional array.

A graph of operations.

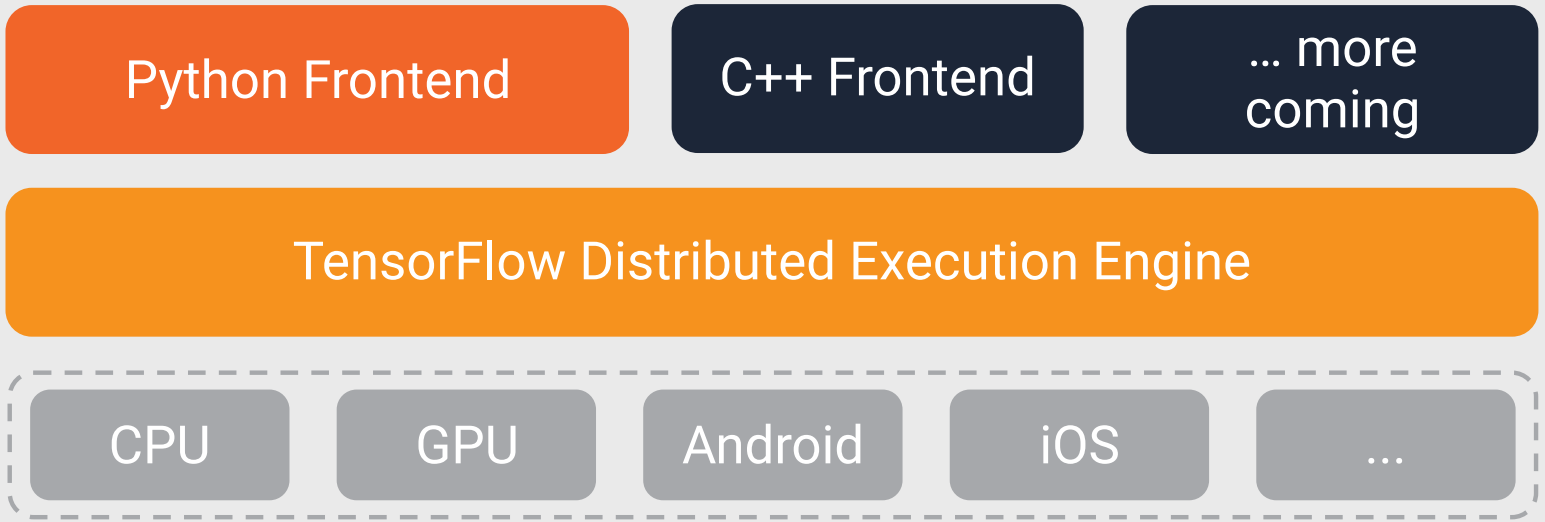
Operates over **tensors**: *n-dimensional arrays*
Using a **flow graph**: *data flow computation framework*

- Flexible, intuitive construction
- Support for threads, queues, and asynchronous computation; **distributed runtime**
- **automatic differentiation**
- Train on CPUs, GPUs, ...and coming soon, **TPUS**...
- Run wherever you like

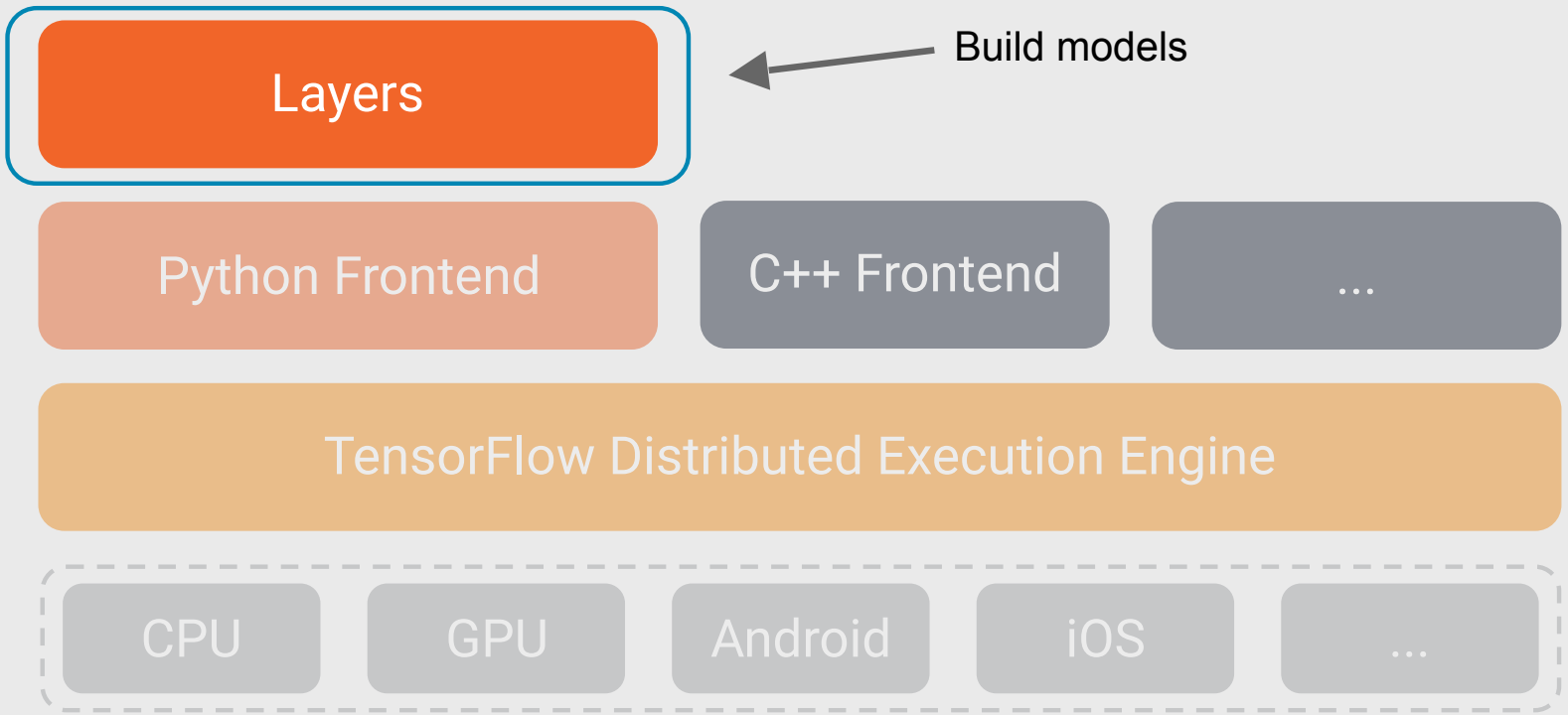


A Brief Tour of TensorFlow Abstraction Levels

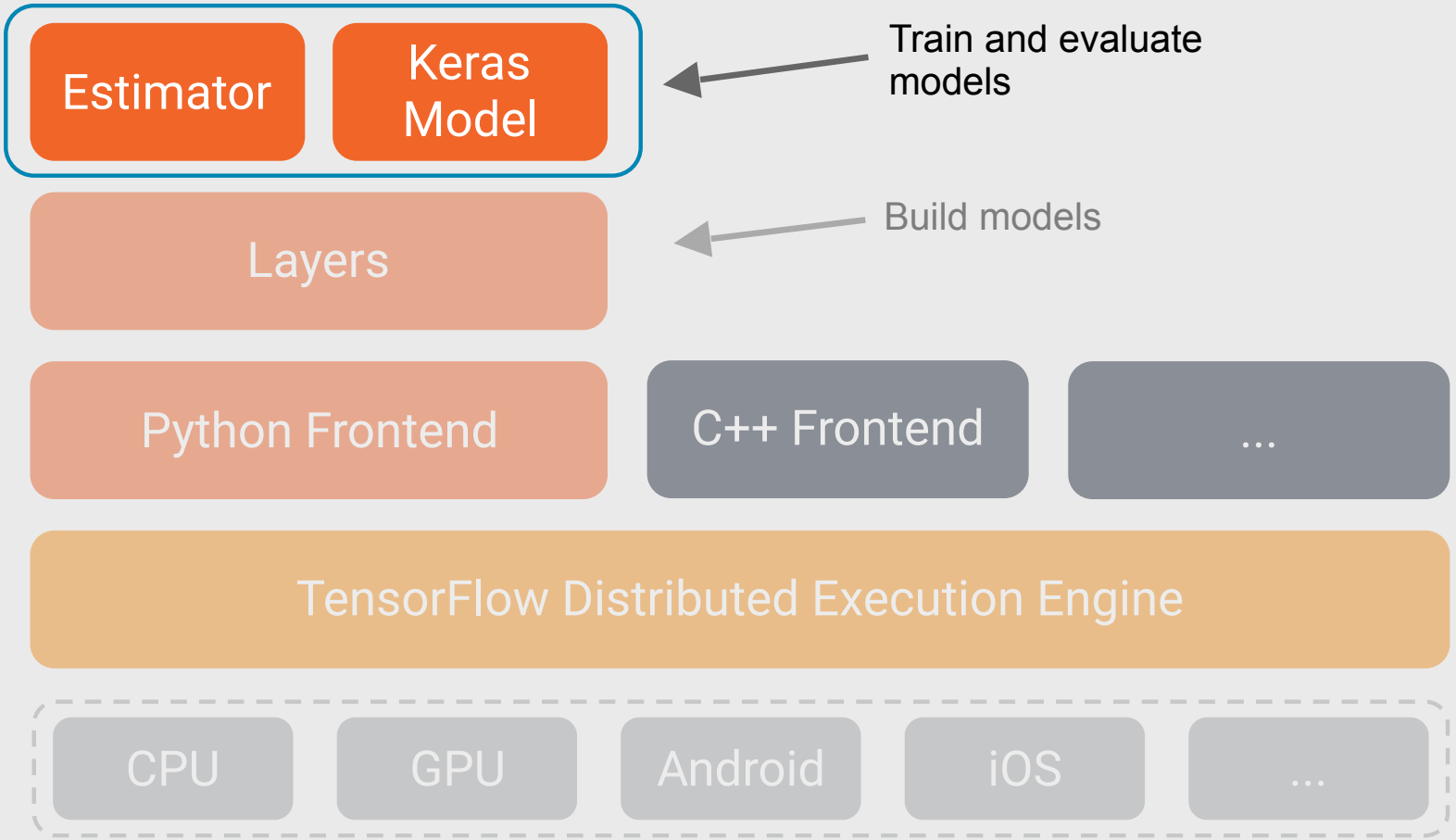
Low-Level Frontends

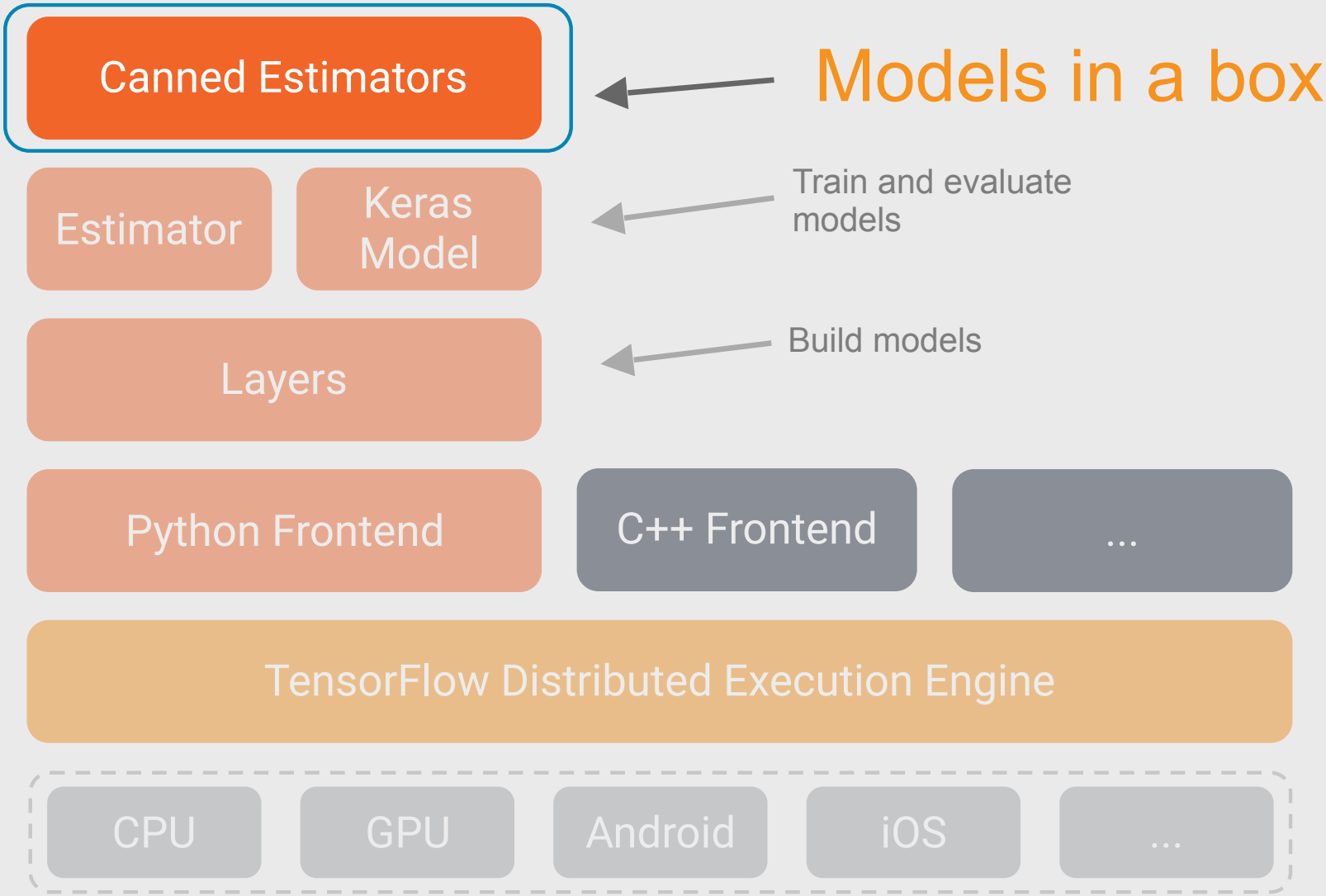


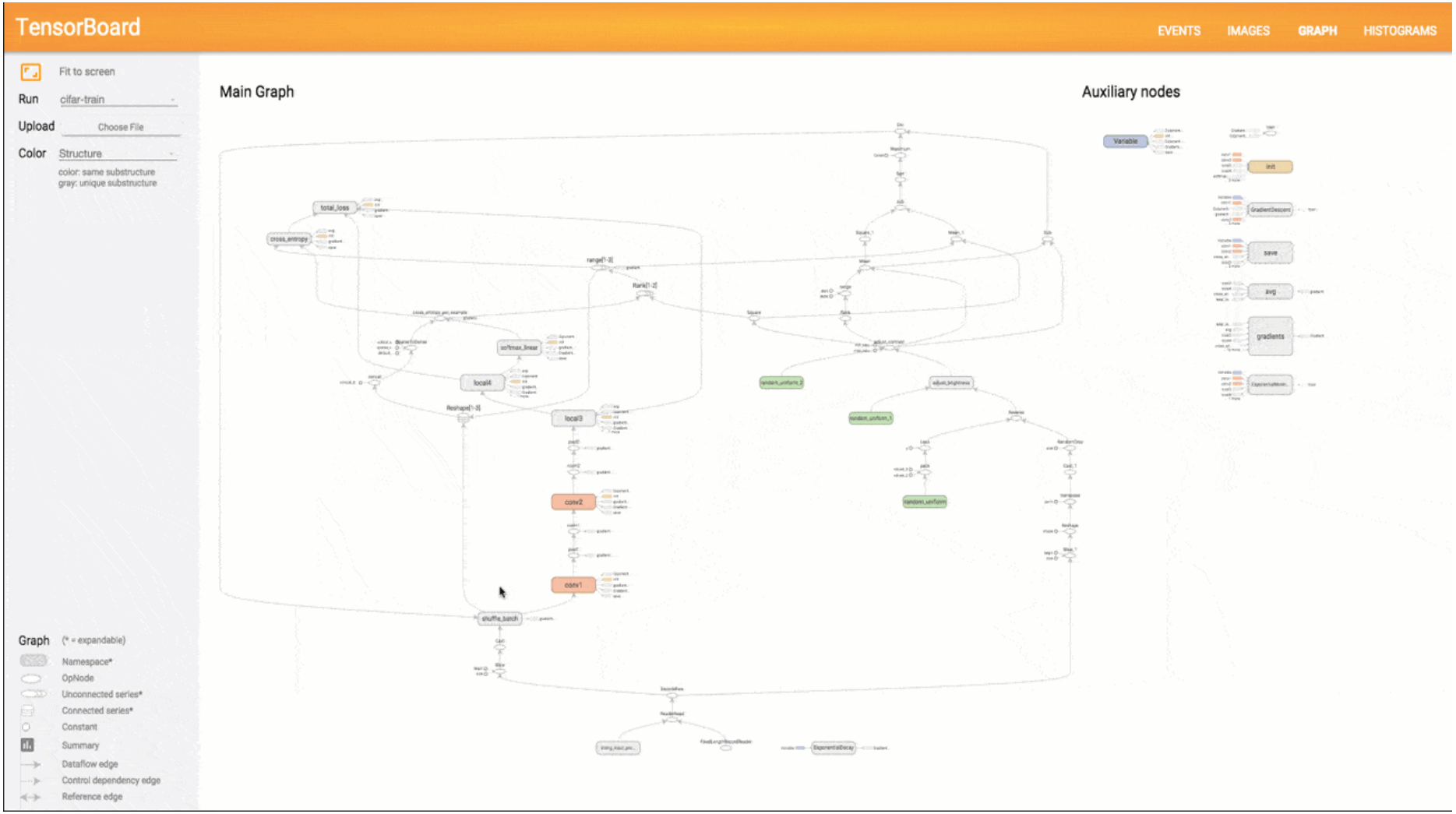
High-Level APIs - Build



High-Level APIs - Train/Eval/Predict

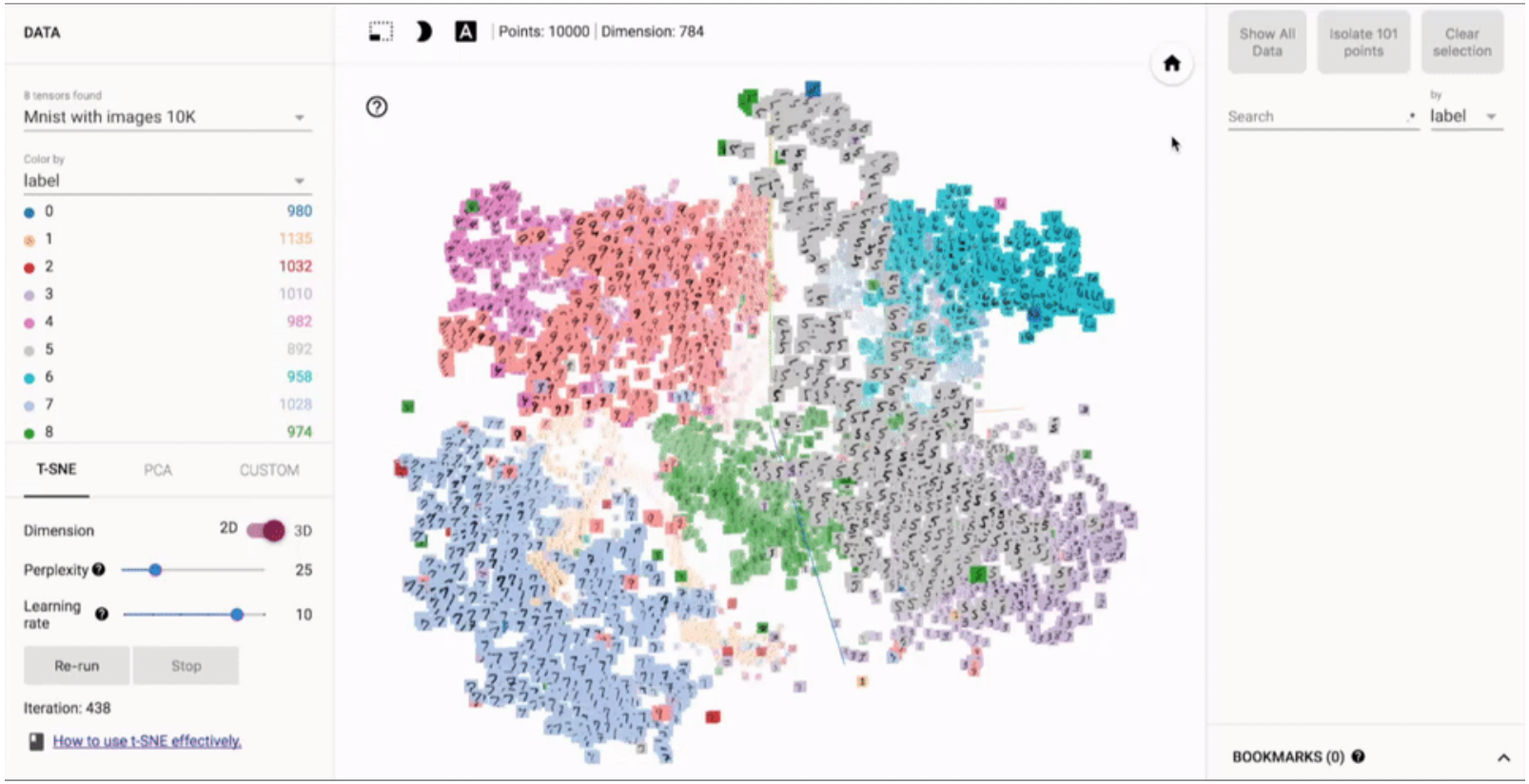



Canned Estimators
Models in a box
Estimator
Keras Model
Train and evaluate models
Layers
Build models
Python Frontend
C++ Frontend
...
TensorFlow Distributed Execution Engine
CPU
GPU
Android
iOS
...



http://bit.ly/tf-aiconf

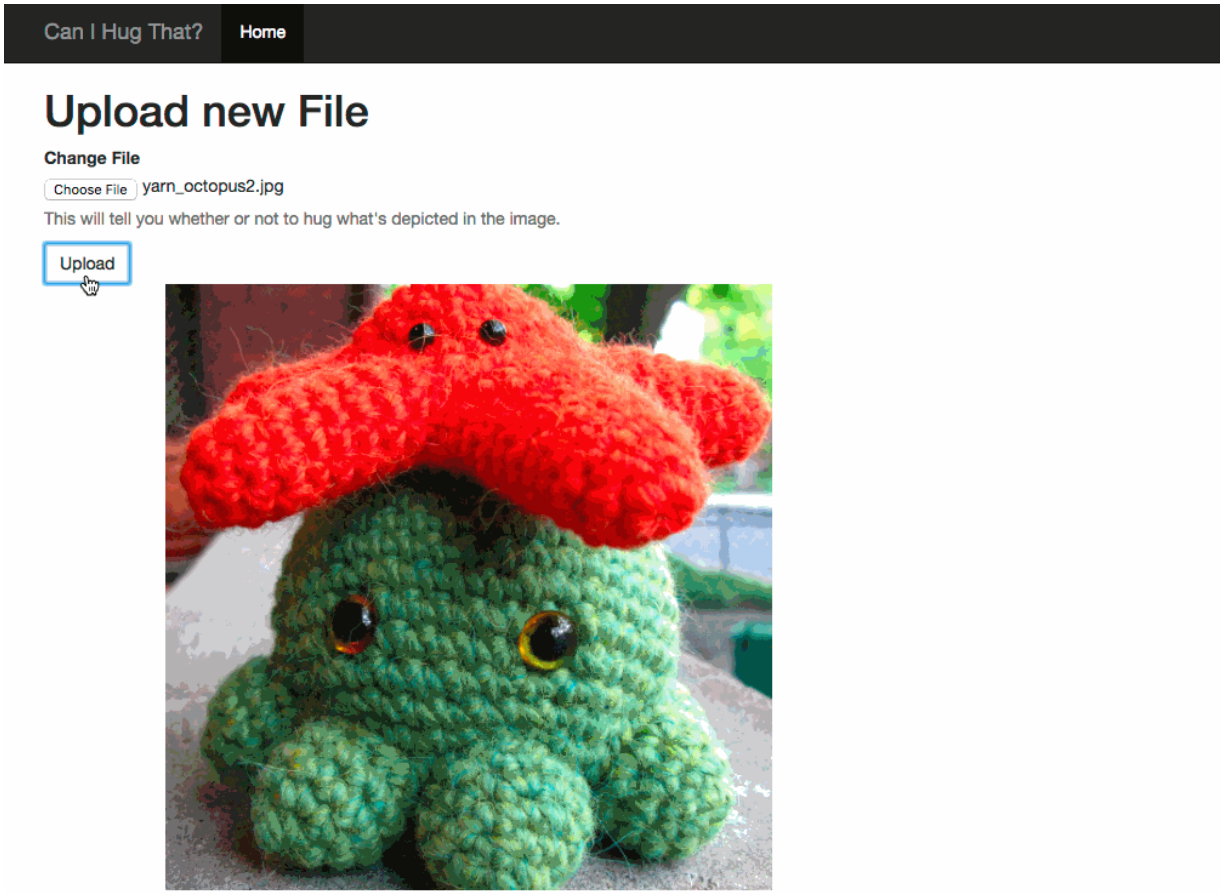
TensorBoard



https://www.tensorflow.org/versions/r0.12/how_tos/embedding_viz/index.html

Open-Source TensorFlow Models

github.com/tensorflow/models



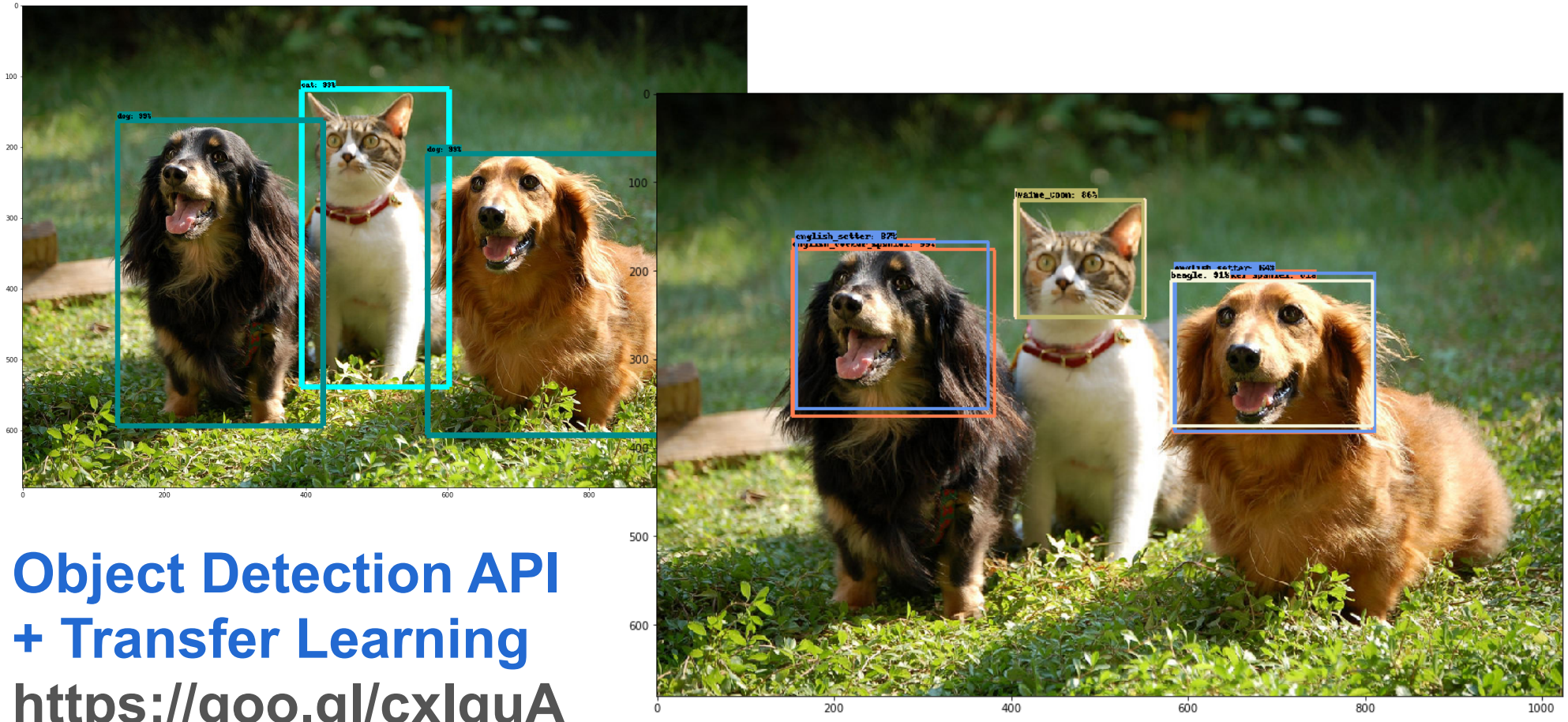
Inception model with **Transfer learning**: "Can I hug that?"

(Thanks to Julia Ferraioli)

goo.gl/nr6Qgh

TensorFlow **Object Detection API**

research.googleblog.com/2017/06/supercharge-your-computer-vision-models.html



Object Detection API + Transfer Learning

https://goo.gl/cxlquA

https://goo.gl/images/xf45oH, by https://www.flickr.com/photos/raneko/

thank you