Sign up for a Google Cloud Platform free trial (https://cloud.google.com) to follow along or try at home.

https://goo.gl/nqsCo5 (this deck)

End-to-end machine learning on structured data

Lak Lakshmanan, Tech Lead, Data & ML Professional Services, Google Cloud

www.vlakshman.com lak_gcp@

Google Cloud

2017-10-12
World Summit AI,
Amsterdam

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Agenda

Machine Learning on Google Cloud

Explore the data

Create the dataset

Build the model

Operationalize model

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

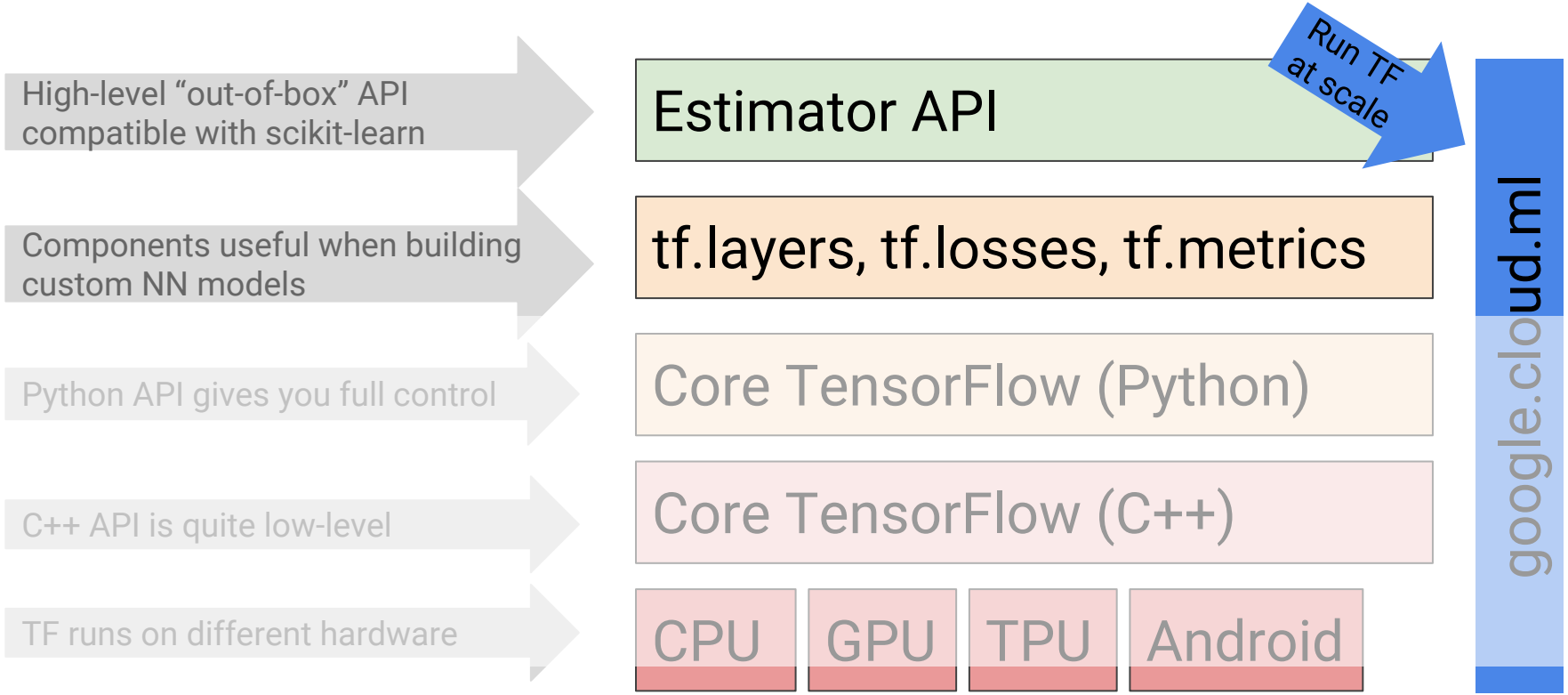
The most common ML models at Google are models that operate on structured data

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	

https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

We will use distributed TensorFlow on Cloud ML Engine



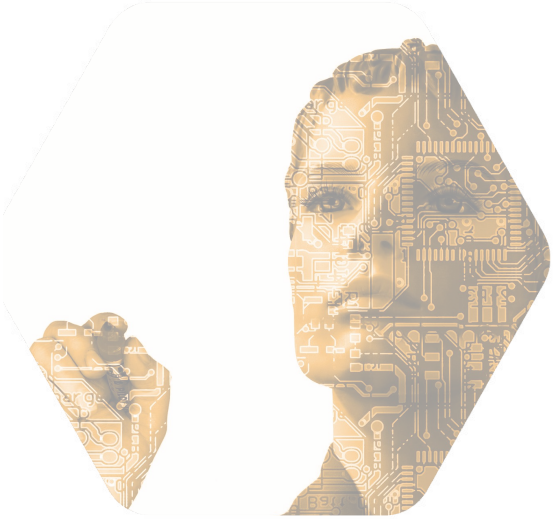
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Many Machine Learning frameworks can handle toy problems

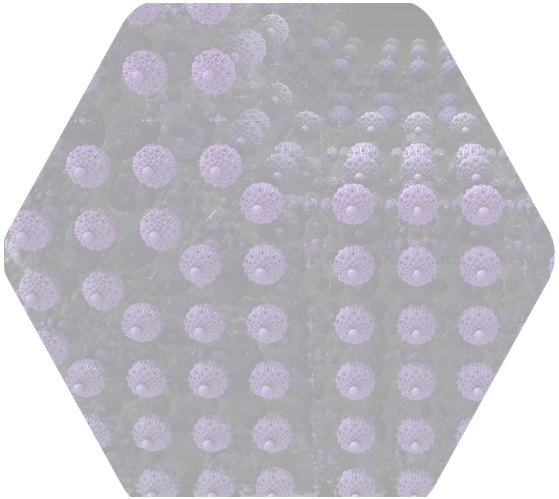
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

To build effective ML, you need:

Big Data



Feature Engineering



Model Architectures

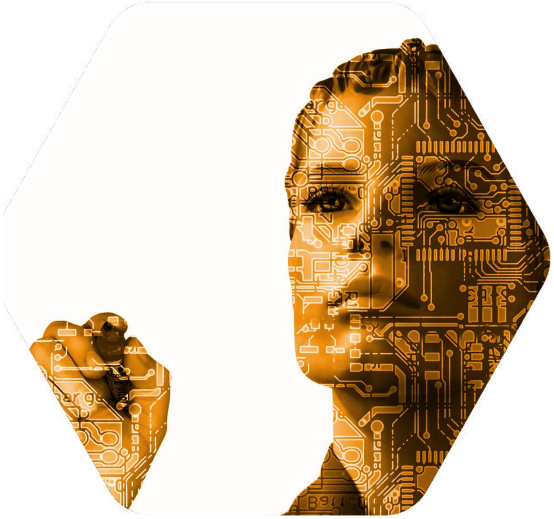
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

As your data size increases, batching and distribution become important

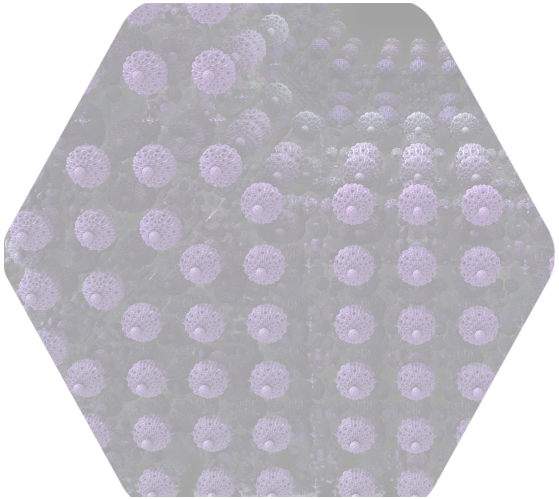
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

To build effective ML, you need:

Big Data



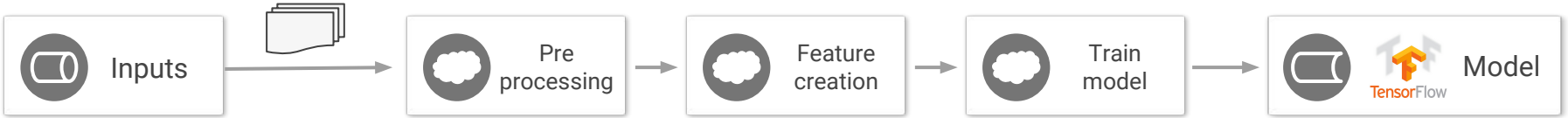
Feature Engineering



Model Architectures

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

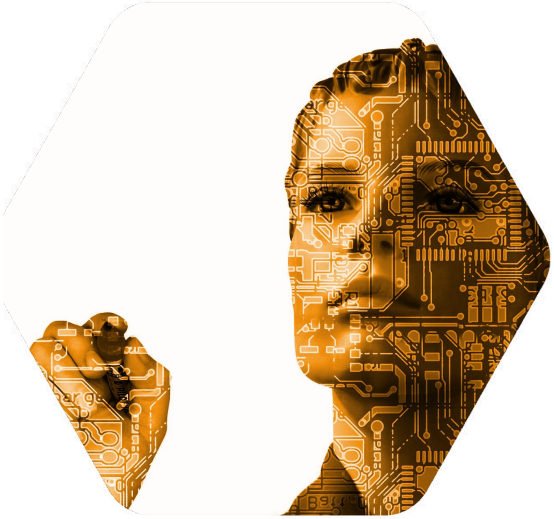
Input necessary transformations



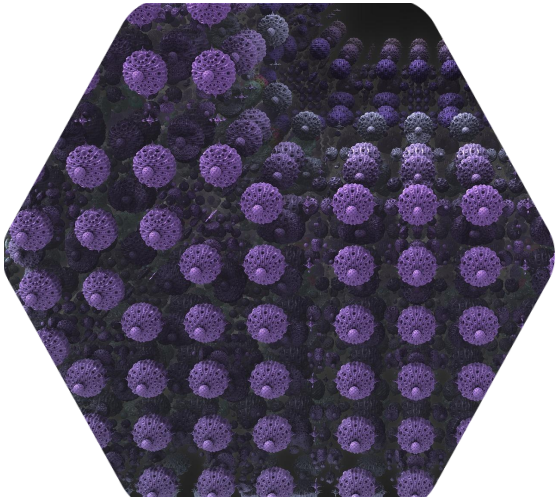
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

To build effective ML, you need:

Big Data



Feature Engineering



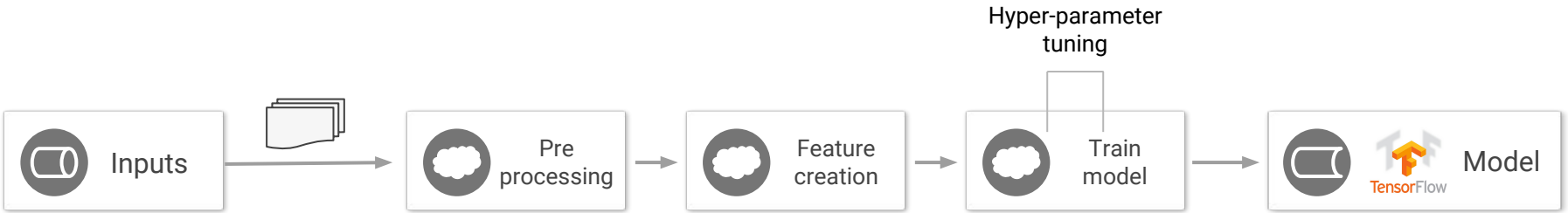
Model Architectures

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

What else does a ML framework need to provide?

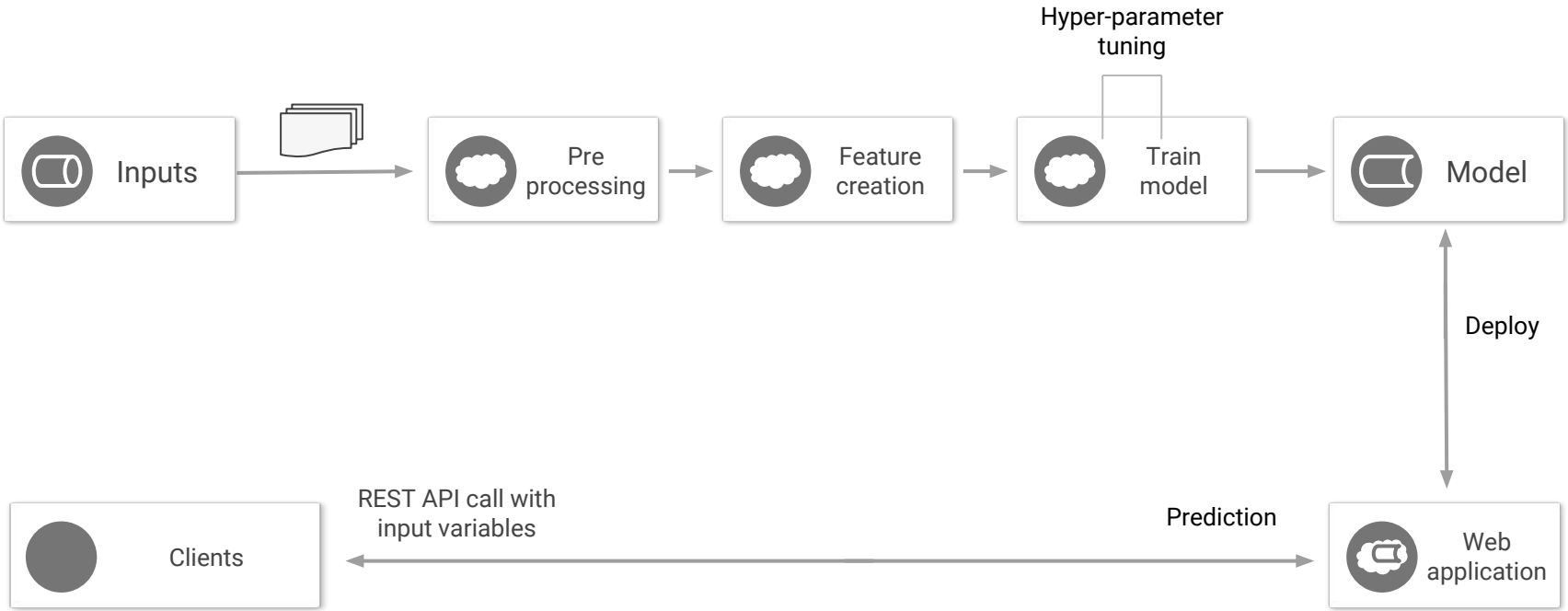
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Hyperparameter tuning might be nice



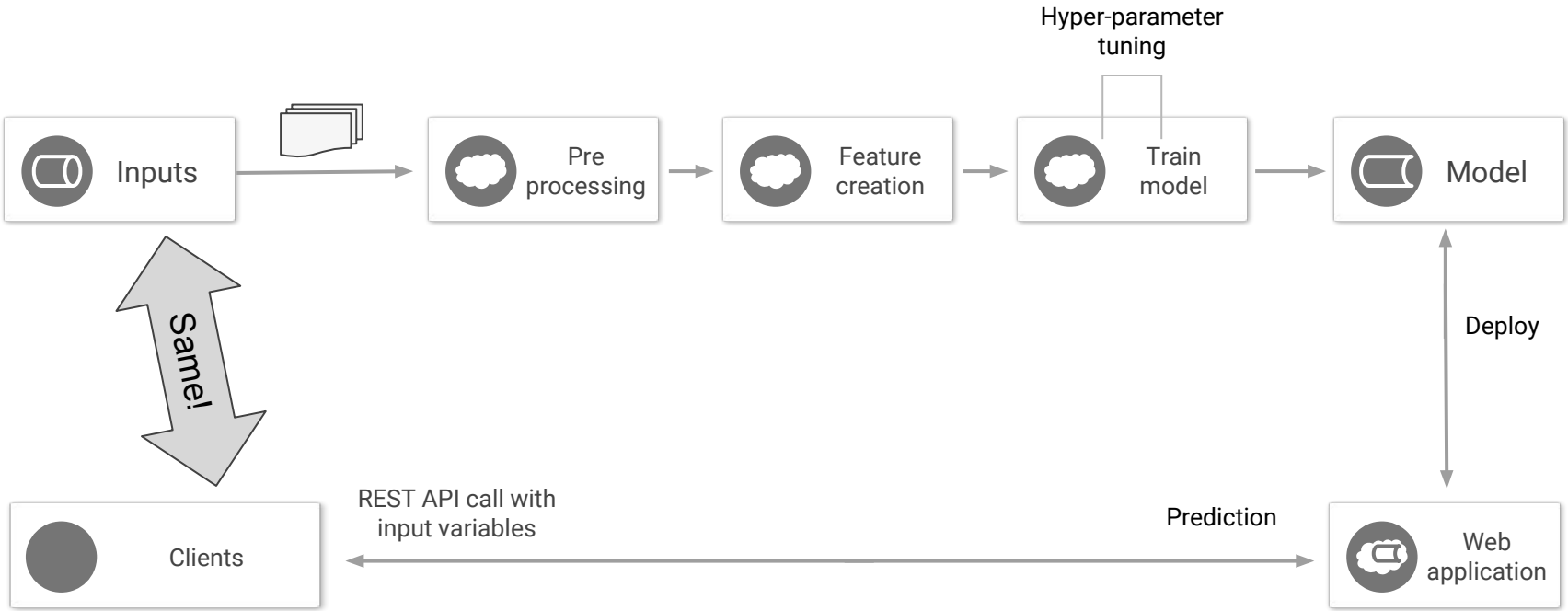
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Need to autoscale prediction code



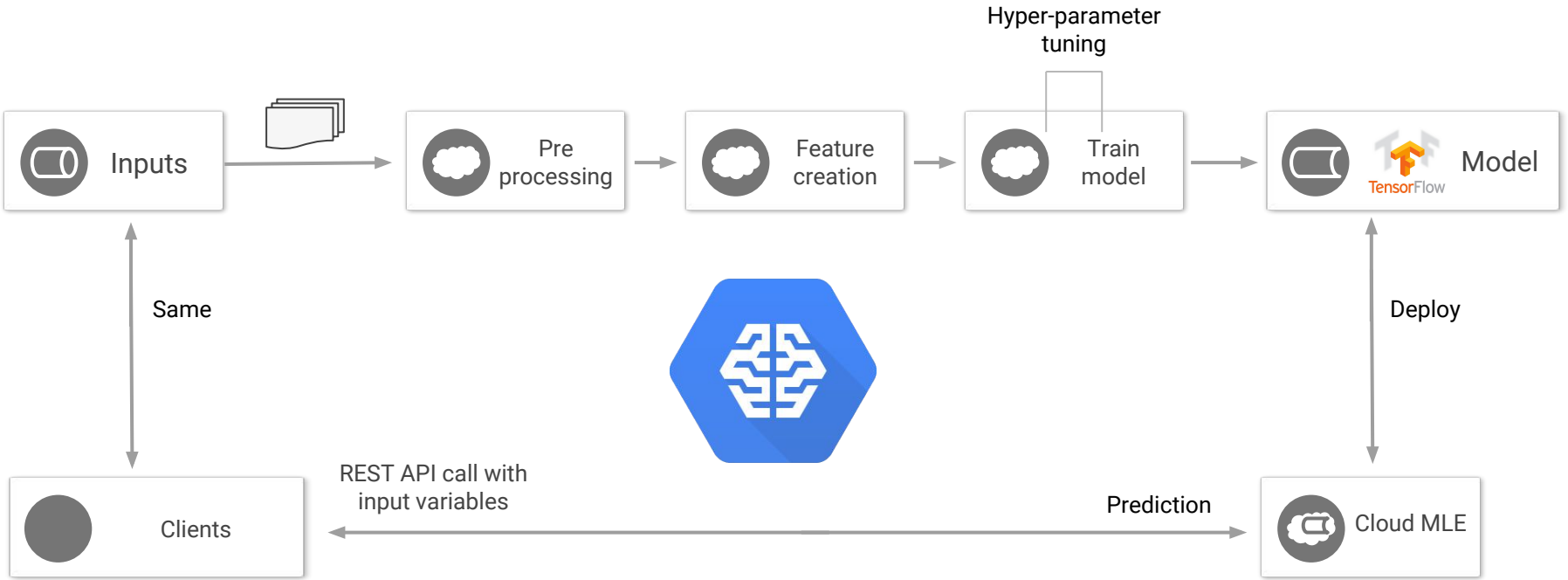
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Who does the preprocessing?



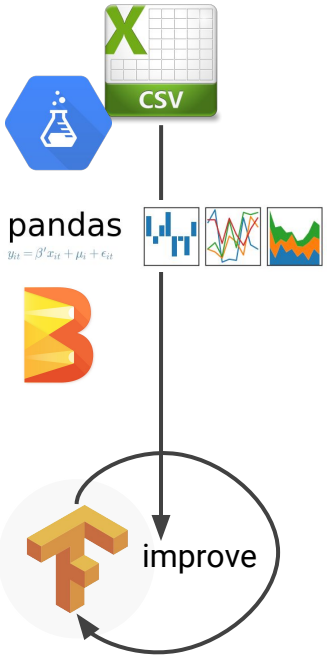
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Cloud Machine Learning—repeatable, scalable, tuned



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

In Datalab, start locally on sampled dataset



Google Cloud Datalab tfclassic (unsaved changes)

Notebook · Add Code · Add Markdown · Delete · Move Up · Move Down · Run · Clear · Reset Session · Widgets · Navigation · Help

2b. Working with low-level TensorFlow

This notebook is Lab2b of CPB 102, Google's course on Machine Learning using Cloud ML.

In this notebook, we will work with relatively low-level TensorFlow functions to implement a linear regression model. We will use this notebook to demonstrate early stopping -- a technique whereby training is stopped once the error on the validation dataset starts to increase.

```
import datalab.bigquery as bq
import tensorflow as tf
import pandas as pd
import numpy as np
import shutil
```

Code to read data and compute error is the same as Lab2a.

```
def read_dataset(filename):
  return pd.read_csv(filename, header=None, names=['pickuplon','pickuplat','dropofflon','dropofflat','passengers','fare_amount'])

df_train = read_dataset('../lab1a/taxi-train.csv')
df_valid = read_dataset('../lab1a/taxi-valid.csv')
df_test = read_dataset('../lab1a/taxi-test.csv')
df_train[:5]
```

```
FEATURE_COLS = np.arange(0,5)
TARGET_COL   = 'fare_amount'
```

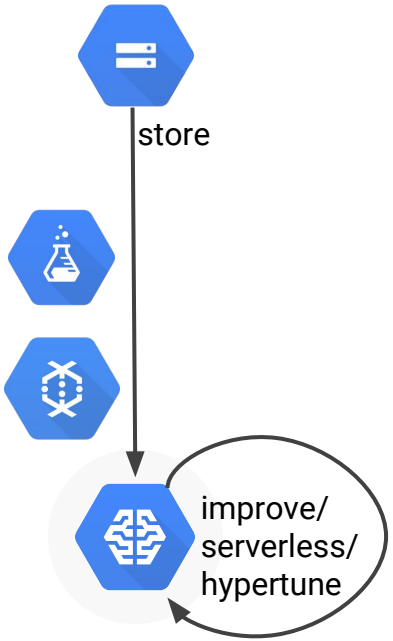
```
def compute_rmse(actual, predicted):
  return np.sqrt(np.mean((actual-predicted)**2))

def print_rmse(model):
  print "Train RMSE = {0}".format(compute_rmse(df_train[TARGET_COL], model.predict(df_train.iloc[:,FEATURE_COLS].values)))
  print "Valid RMSE = {0}".format(compute_rmse(df_valid[TARGET_COL], model.predict(df_valid.iloc[:,FEATURE_COLS].values)))
```

Linear Regression

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Then, scale it out to GCP using serverless technology



Google Cloud Datalab cloudml (autosaved)

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

Training on cloud

In order to train on the cloud, we have to copy the model and data to our bucket on Google Cloud Storage (GCS).

```
%bash
rm -rf taxifare.tar.gz taxi_trained
tar cvfz taxifare.tar.gz taxifare
gsutil cp taxifare.tar.gz gs://$BUCKET/taxifare/source/taxifare.tar.gz
gsutil cp ../labla/*.csv  gs://$BUCKET/taxifare/input/
gsutil -m rm -r -f gs://$BUCKET/taxifare/taxi_preproc
gsutil -m rm -r -f gs://$BUCKET/taxifare/taxi_trained
```

Running...

When you run your preprocessor, you have to change the input and output to be on GCS.

Using DirectPipelineRunner runs Dataflow locally, but the inputs & outputs are on the cloud. Using BlockingDataflowPipelineRunner will use Cloud Dataflow (and take much longer because of the overhead involved for such a small dataset). To see the status of your BlockingDataflowPipelineRunner job, visit https://console.cloud.google.com/dataflow

```
# imports
import apache_beam as beam
import google.cloud.ml as ml
import google.cloud.ml.dataflow.io.tfrecordio as tfrecordio
import google.cloud.ml.io as io
import os

# Change as needed
#RUNNER = 'DirectPipelineRunner'  #
RUNNER = 'BlockingDataflowPipelineRunner'

# defines
feature_set = TaxifareFeatures()
OUTPUT_DIR = 'gs://{0}/taxifare/taxi_preproc'.format(BUCKET)
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Agenda

Machine Learning on Google Cloud

Explore the data

Create the dataset

Build the model

Operationalize model

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Goal: predict weight of a newborn

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

An open dataset of births is available in BigQuery

Births recorded in the 50 states of the USA from 1969 to 2008

Table ID	bigquery-public-data:samples.natality
Table Size	21.9 GB
Long Term Storage Size	21.9 GB
Number of Rows	137,826,763

https://bigquery.cloud.google.com/table/bigquery-public-data:samples.natality

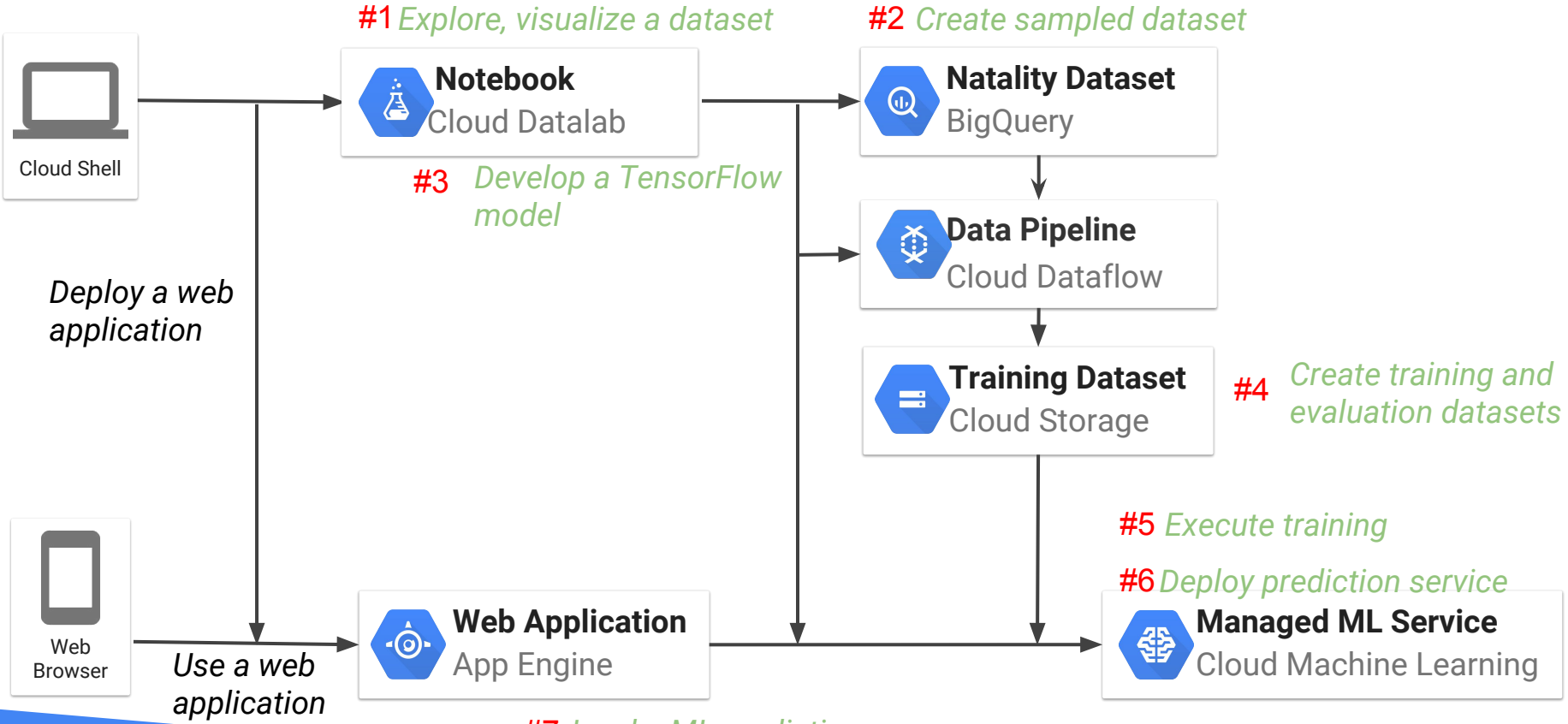
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

The data set includes details about the pregnancy

Date of birth	**year**	INTEGER	NULLABLE	Four-digit year of the birth. Example: 1975.
	month	INTEGER	NULLABLE	Month index of the date of birth, where 1=January.
	day	INTEGER	NULLABLE	Day of birth, starting from 1.
	wday	INTEGER	NULLABLE	Day of the week, where 1 is Sunday and 7 is Saturday.
Location of birth (US state)	**state**	STRING	NULLABLE	The two character postal code for the state. Entries after 2004 do not include this value.
Baby's birth weight (lbs)	**weight_pounds**	FLOAT	NULLABLE	Weight of the child, in pounds.
Mother's age at birth	**mother_age**	INTEGER	NULLABLE	Reported age of the mother when giving birth.
Duration of pregnancy	**gestation_weeks**	INTEGER	NULLABLE	The number of weeks of the pregnancy.
Mother's use of cigarette or alcohol	**cigarette_use**	BOOLEAN	NULLABLE	True if the mother smoked cigarettes. Available starting 2003.
	alcohol_use	BOOLEAN	NULLABLE	True if the mother used alcohol. Available starting 1989.
Mother's weight gain (lbs)	**weight_gain_pounds**	INTEGER	NULLABLE	Number of pounds gained by the mother during pregnancy.

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

The end-to-end machine learning set of labs



= Lab Exercises

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

https://goo.gl/nqsCo5 (this deck)

codelabs.developers.google.com/cloud-quest-scientific-data

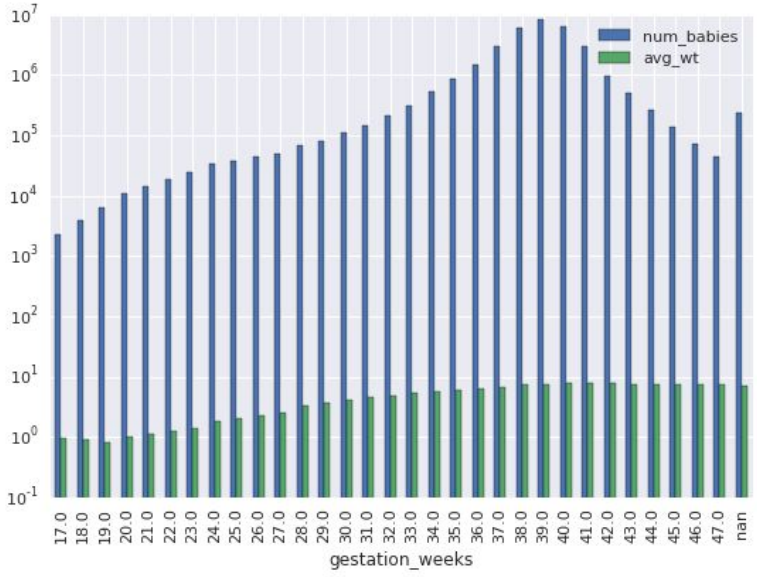
github.com/GoogleCloudPlatform/training-data-analyst/blob/master/blogs/babyweight/

Lab #1: Exploring a BigQuery dataset to find features to use in model

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Lab #1: Exploring a BigQuery dataset to find features to use in model

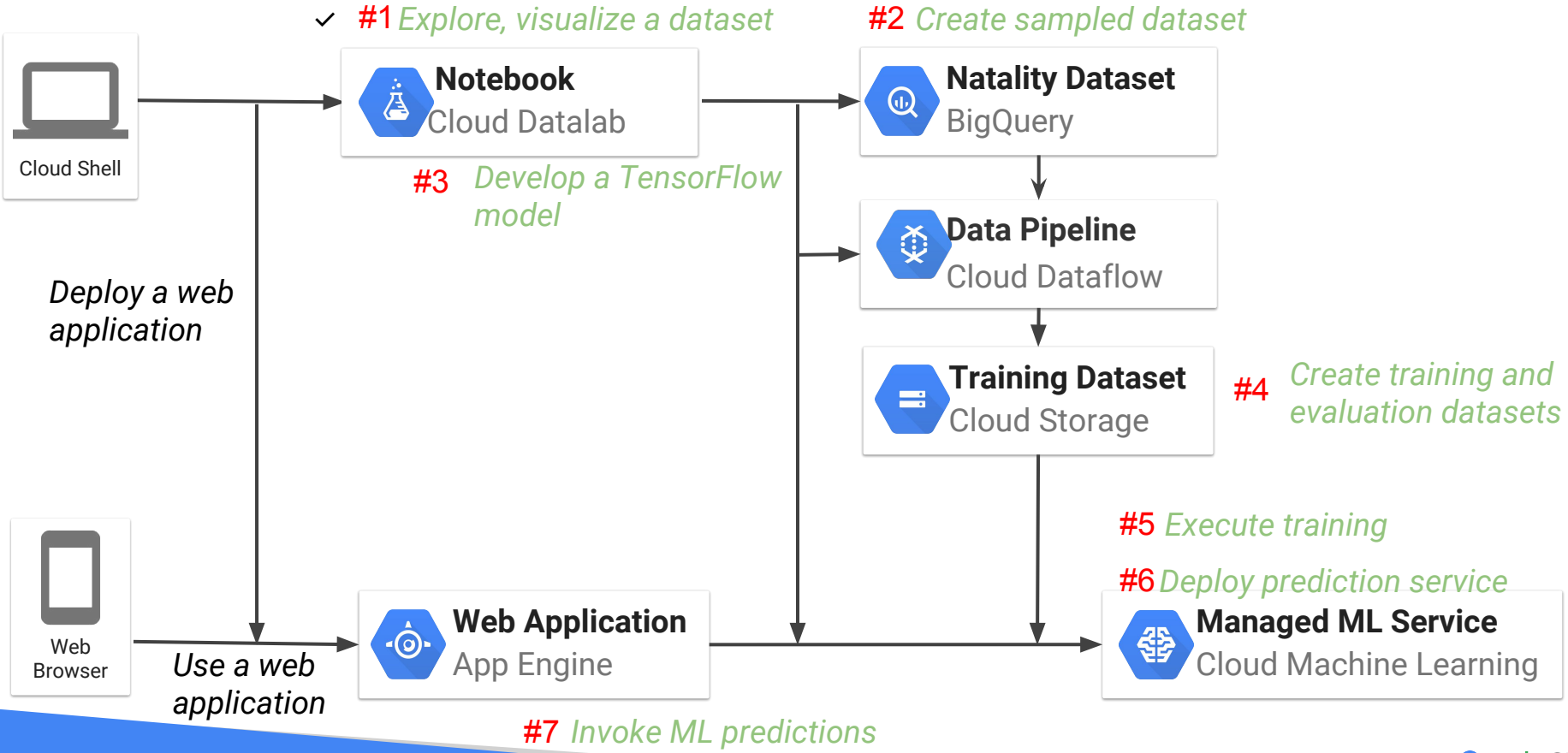
BigQuery: `publicdata.samples.natality`



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

The End-to-End process

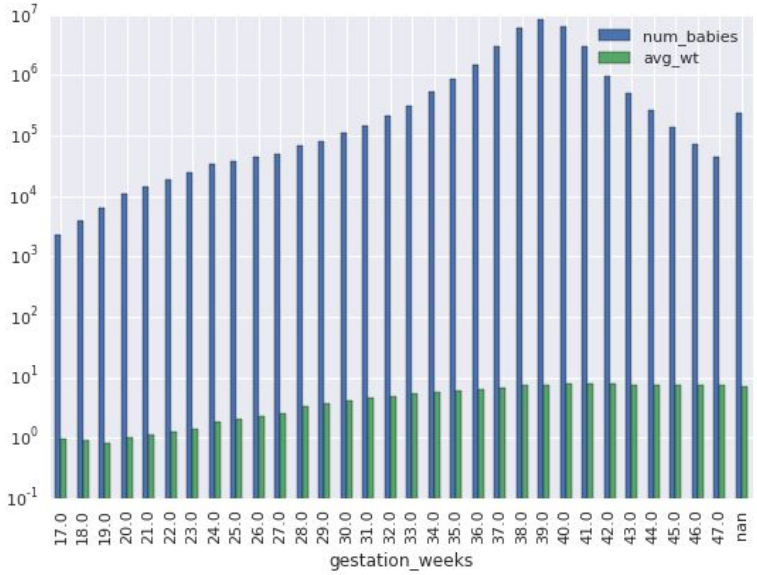


= Lab Exercises

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Debrief

What features did you use? Why?



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Is it fair practice to use mother's race as a feature?

How would we check whether the use/non-use of mother's race amplifies injustice?

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Agenda

- Machine Learning on Google Cloud
- Explore the data
- Create the dataset
- Build the model
- Operationalize model

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Building a ML model involves:

Create the dataset

Build the model

Operationalize the model

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

https://goo.gl/nqsCo5 (this deck)

codelabs.developers.google.com/cloud-quest-scientific-data

github.com/GoogleCloudPlatform/training-data-analyst/blob/master/blogs/babyweight/

Lab #2: Creating a sampled dataset

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

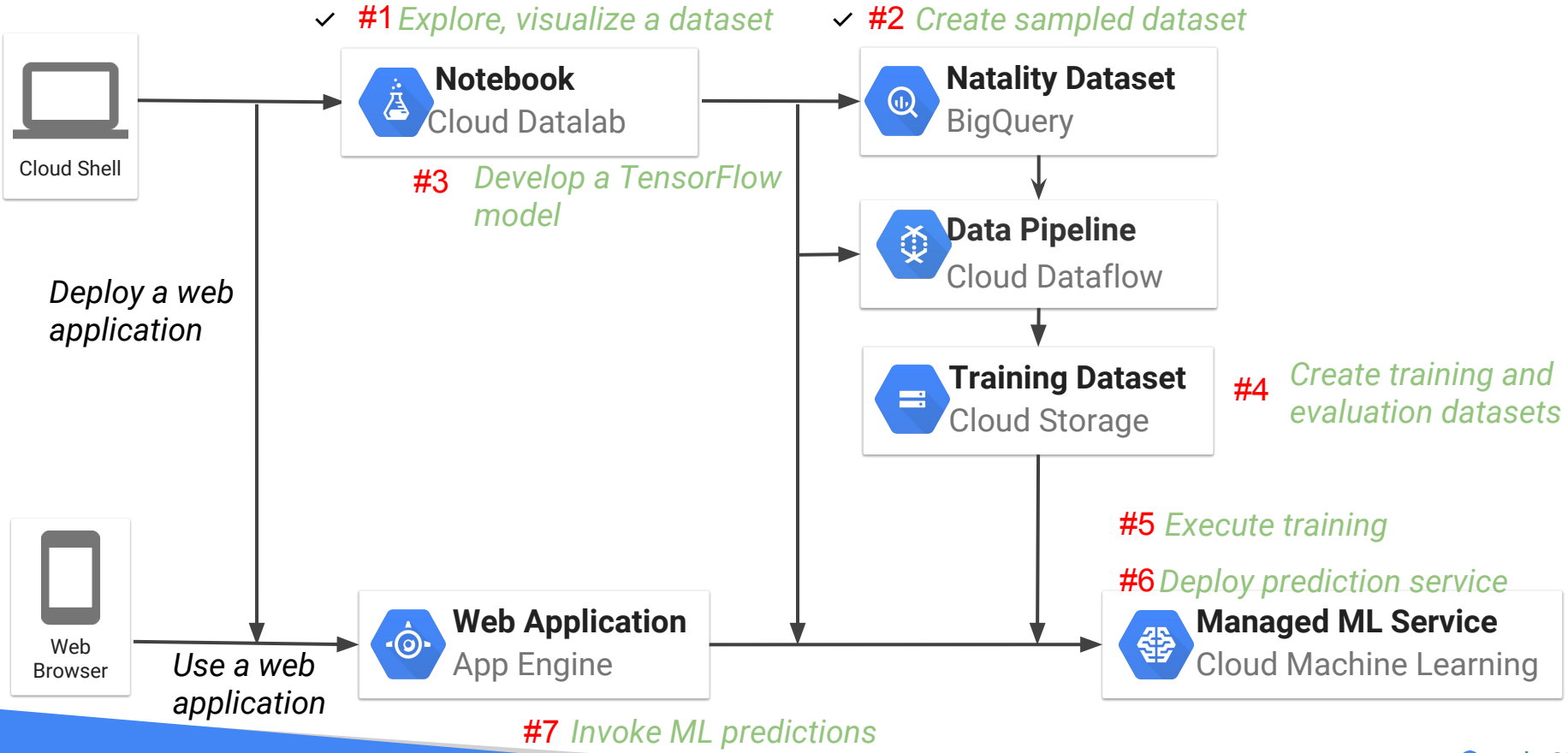
Lab #2: Creating a sampled dataset

1. Sample the full BigQuery, creating a smaller dataset so that you can use it for model development and local training

2. Make sure that this smaller dataset is representative of the full dataset.

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

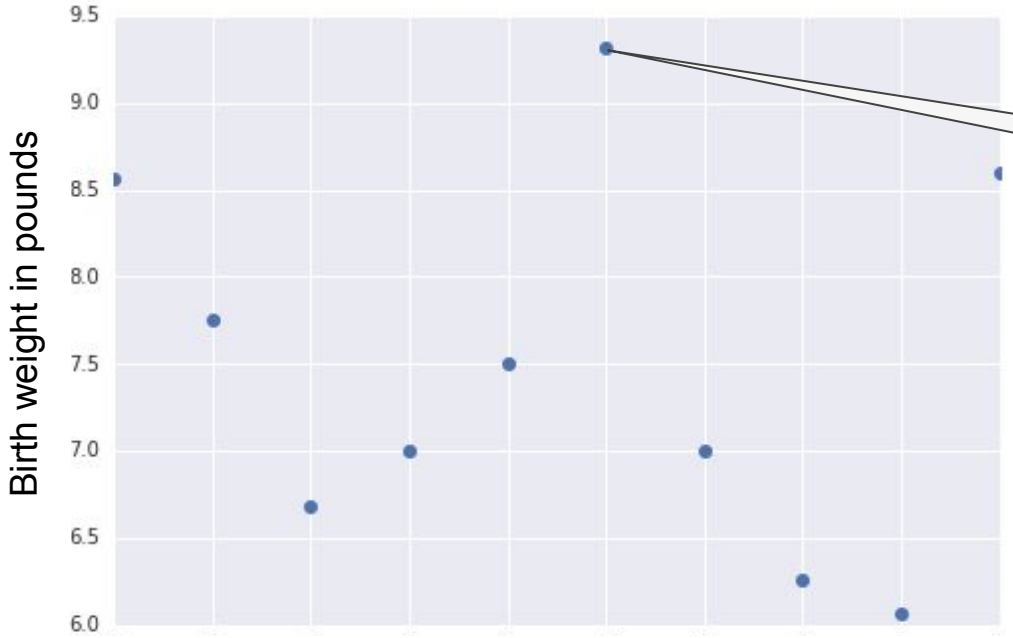
The End-to-End process



= Lab Exercises

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Debrief: The simplest option is to sample rows randomly



- Each data point is a birth record from the natality dataset

weight	year	mother_age	gestation_weeks	cigarette_use	alcohol_use
6.03	2004	29	39	false	false

- Random sampling eliminates potential biases due to order of the training examples but ...

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

But what about triplets?

3 rows with essentially the same data!

How can we make this data unique?

How can we solve this?

Use FARM_FINGERPRINT so that you get a repeatable, nonoverlapping sample.

Export train.csv and eval.csv (2 files)

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Agenda

Machine Learning on Google Cloud

Explore the data

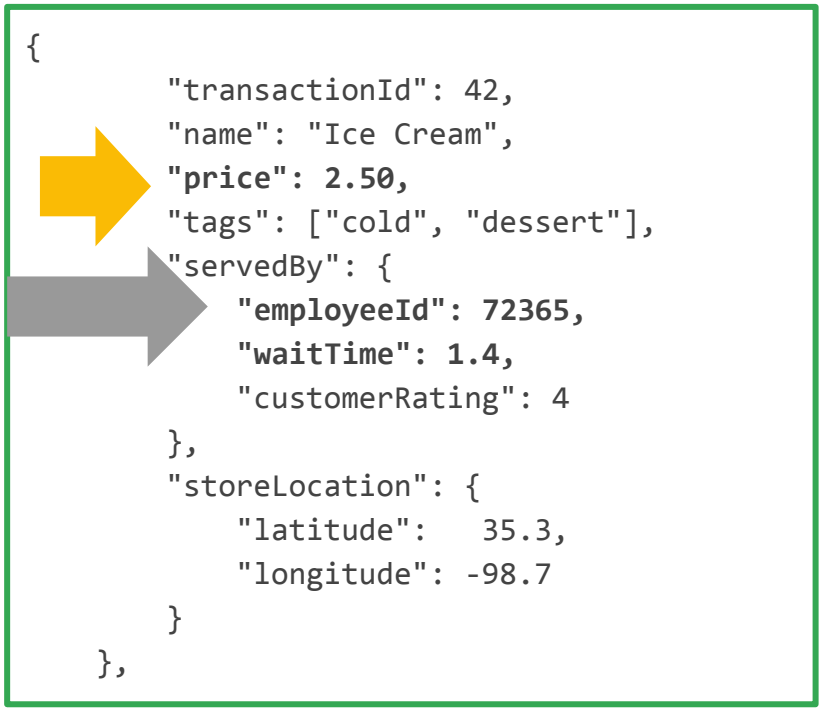
Create the dataset

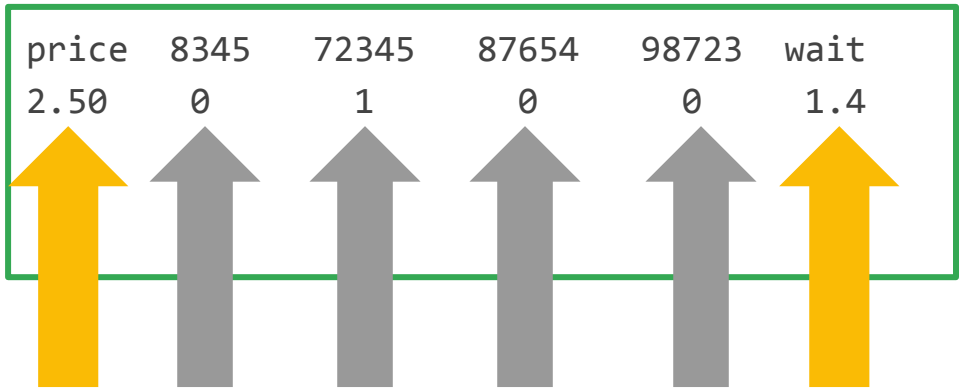
Build the model

Operationalize model

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Two types of features: Dense & sparse

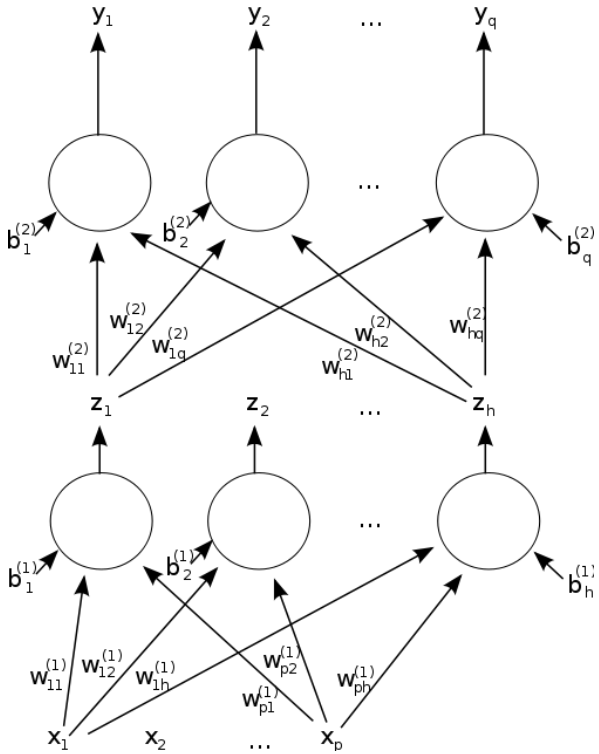




© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

DNNs good for dense, highly correlated

pixel_values ()



1024^2 input nodes

10 hidden nodes -> 10 image features

https://commons.wikimedia.org/wiki/File:Two_layer_ann.svg

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

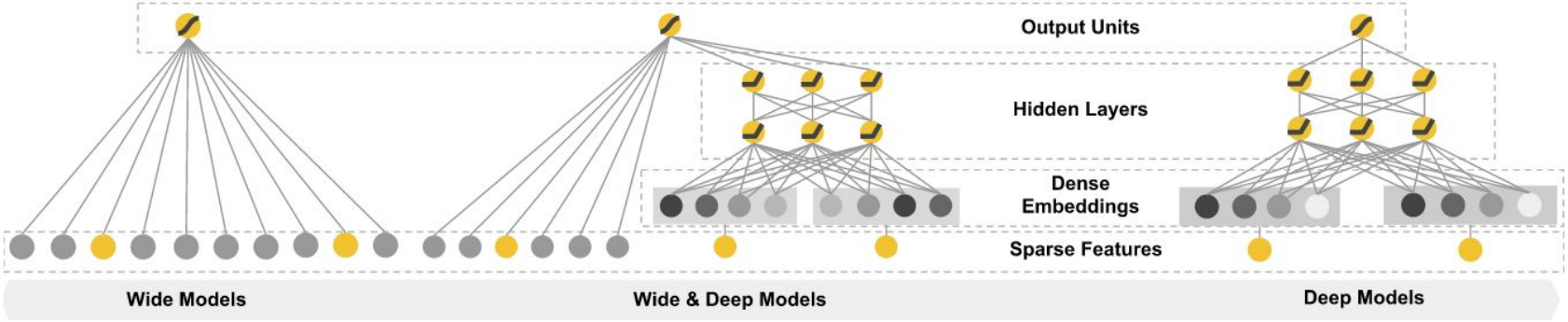
Linear for sparse, independent features

One hot encoding

```
[ 0.  1.  0.  0.  0.  0.  0.  0.  0.  0.]
[ 0.  0.  0.  0.  0.  0.  0.  1.  0.  0.]
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  1.]
[ 0.  1.  0.  0.  0.  0.  0.  0.  0.  0.]
[ 0.  0.  0.  0.  0.  0.  0.  1.  0.  0.]
[ 0.  0.  0.  1.  0.  0.  0.  0.  0.  0.]
[ 0.  0.  0.  0.  0.  0.  1.  0.  0.  0.]
[ 0.  1.  0.  0.  0.  0.  0.  0.  0.  0.]
[ 0.  0.  0.  0.  1.  0.  0.  0.  0.  0.]
[ 0.  1.  0.  0.  0.  0.  0.  0.  0.  0.]
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Wide-and-deep models let you handle both



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Wide & Deep

memorization
relevance

generalization
diversity

Wide-and-deep network in Estimator API

```
model = tf.contrib.learn.DNNLinearCombinedClassifier(
    model_dir=...,
    linear_feature_columns=wide_columns,
    dnn_feature_columns=deep_columns,
    dnn_hidden_units=[100, 50])
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

https://goo.gl/nqsCo5 (this deck)

codelabs.developers.google.com/cloud-quest-scientific-data

github.com/GoogleCloudPlatform/training-data-analyst/blob/master/blogs/babyweight/

Lab #3: Create Tensorflow model

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Lab #3: Create Tensorflow model

In this lab, you use the high-level Estimator API for wide-and-deep model

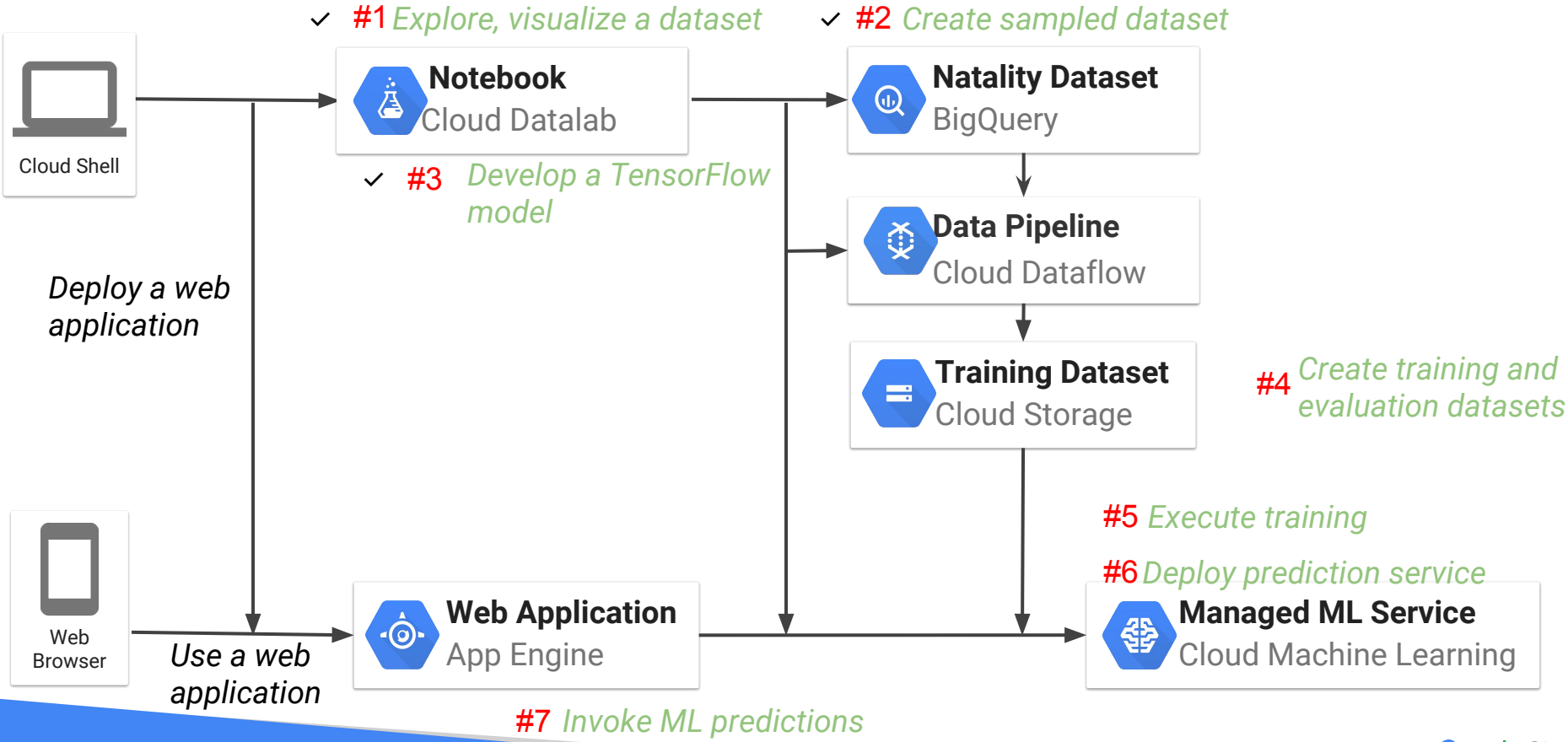
1. Determine your wide columns and deep columns

2. Determine if you need any embeddings

3. Feel free to iterate and add further preprocessing of features

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

The End-to-End process



= Lab Exercises

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Agenda

Machine Learning on Google Cloud

Explore the data

Create the dataset

Build the model

Operationalize model

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

https://goo.gl/nqsCo5 (this deck)

codelabs.developers.google.com/cloud-quest-scientific-data

github.com/GoogleCloudPlatform/training-data-analyst/blob/master/blogs/babyweight/

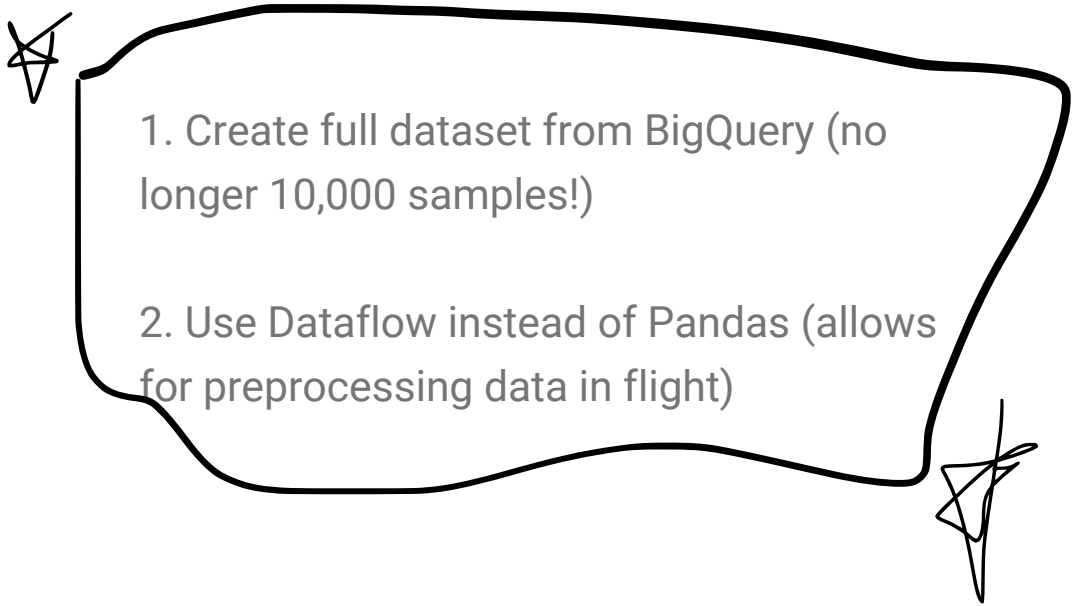
Lab #4: Preprocess using Dataflow

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Lab #4: Preprocess using Dataflow

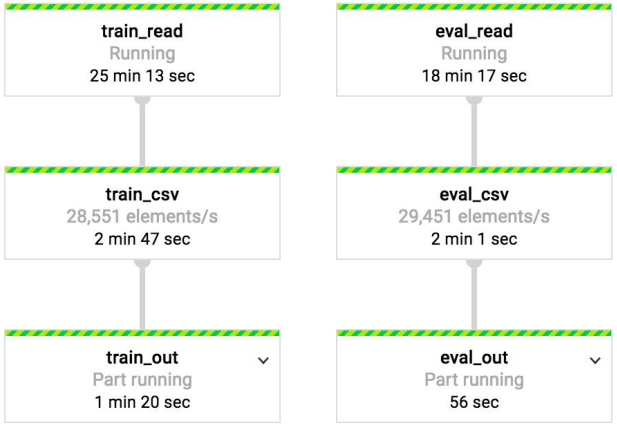
While Pandas is fine for experimenting, for operationalization of your workflow, it is better to do preprocessing in Apache Beam (Dataflow)

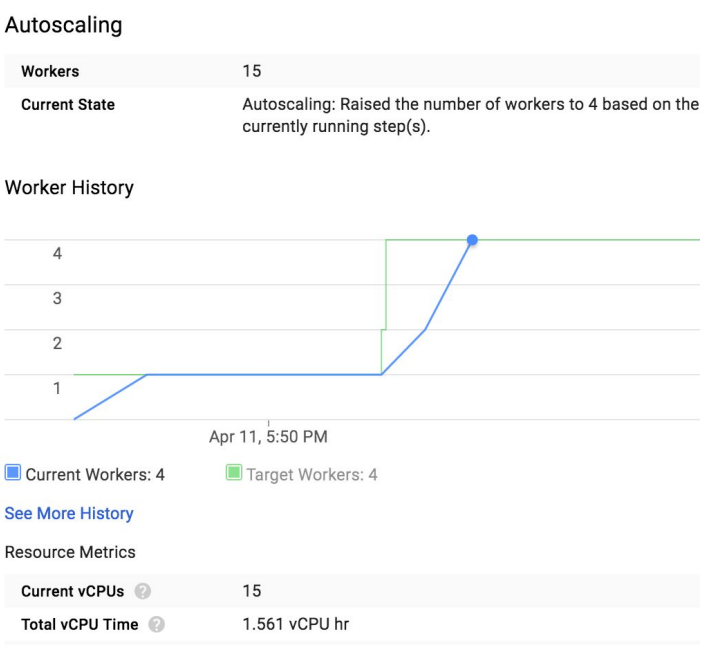


© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Split full dataset into train/eval and do preprocessing

BigQuery -> Dataflow -> CSV

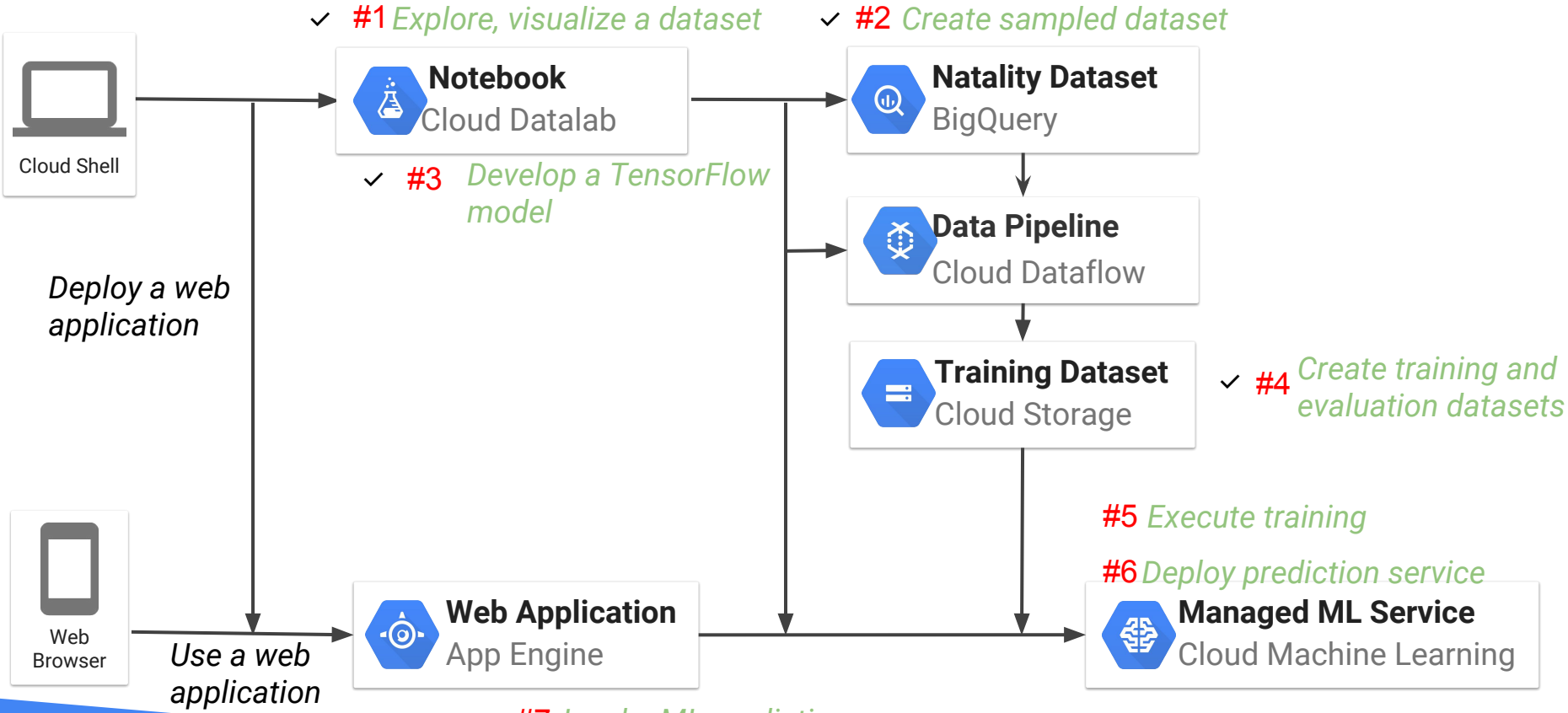




© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

The End-to-End process



= Lab Exercises

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

https://goo.gl/nqsCo5 (this deck)

codelabs.developers.google.com/cloud-quest-scientific-data

github.com/GoogleCloudPlatform/training-data-analyst/blob/master/blogs/babyweight/

Lab #5: Train on Cloud ML Engine

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Lab #5: Train on Cloud ML Engine

This lab illustrates distributed training and hyperparameter tuning on Cloud ML Engine

1. change batchsize if necessary
2. use train_steps instead of num_epochs

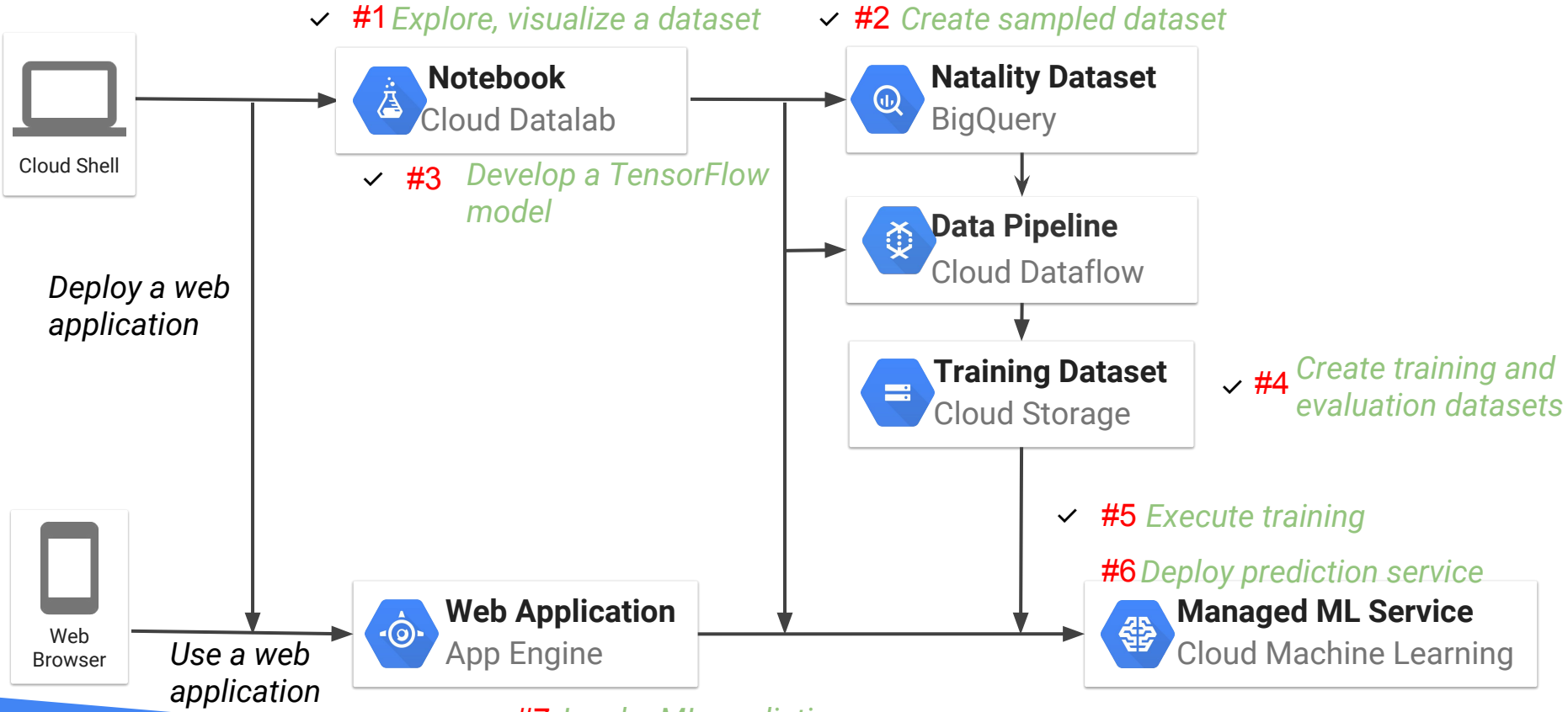
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Train on Cloud ML Engine

Submit training job on full dataset, monitor using TensorBoard

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

The End-to-End process



= Lab Exercises

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

https://goo.gl/nqsCo5 (this deck)

codelabs.developers.google.com/cloud-quest-scientific-data

github.com/GoogleCloudPlatform/training-data-analyst/blob/master/blogs/babyweight/

Lab #6: Deploy and predict

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Lab #6: Deploy and predict

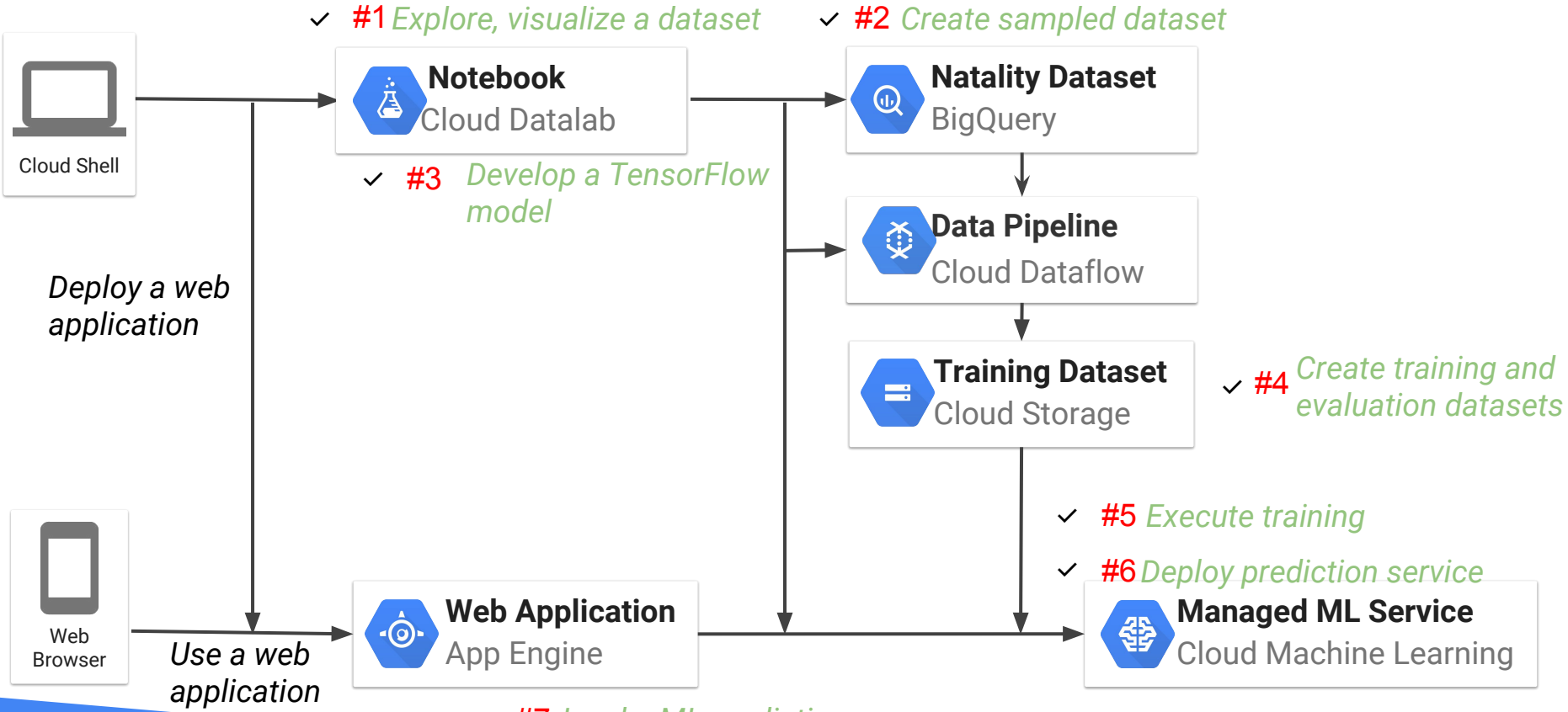
In this lab, you will:

1. Deploy trained model to Cloud ML Engine

2. Send JSON request to model to get predictions

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

The End-to-End process



= Lab Exercises

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

https://goo.gl/nqsCo5 (this deck)

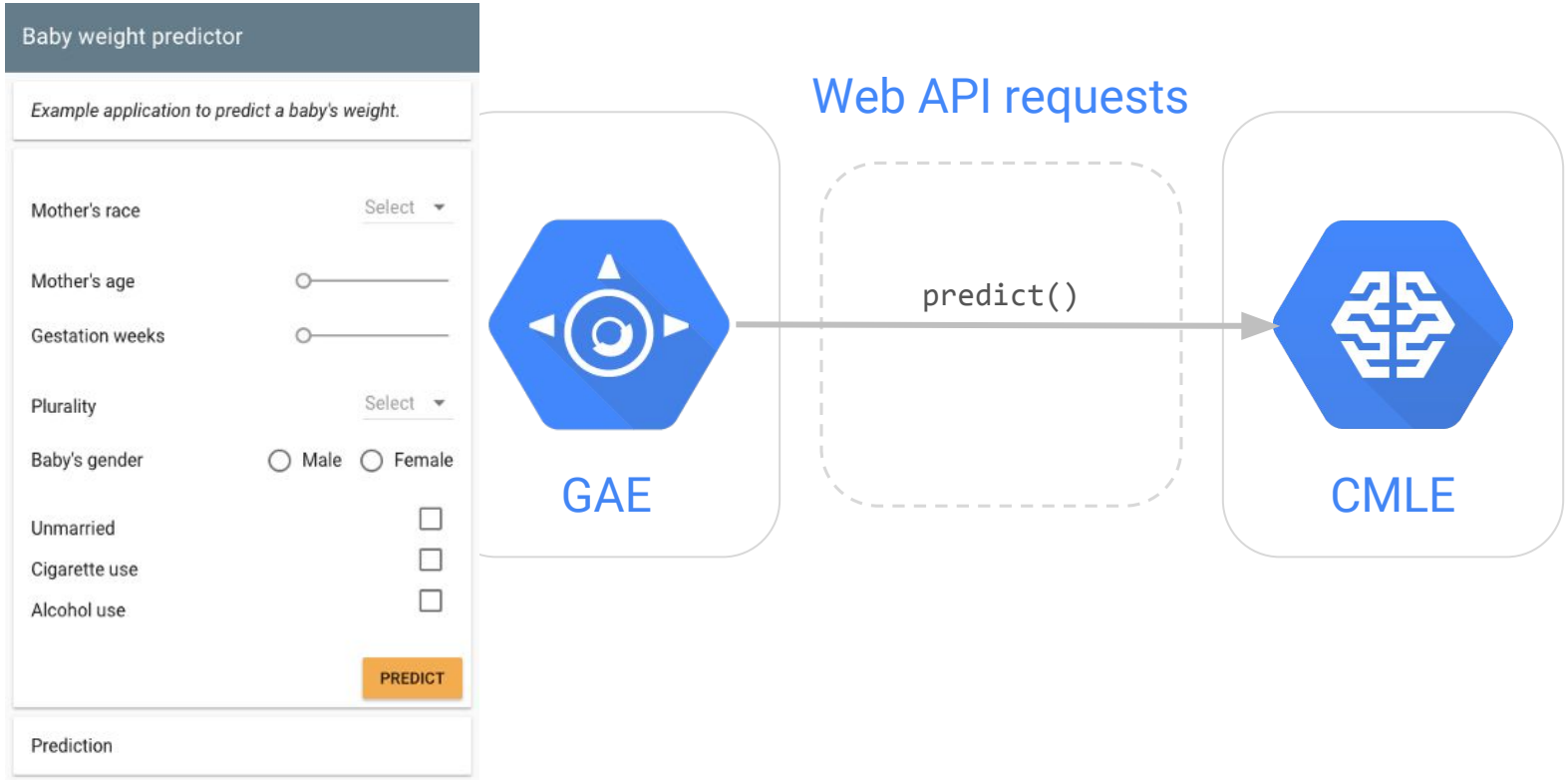
codelabs.developers.google.com/cloud-quest-scientific-data

github.com/GoogleCloudPlatform/training-data-analyst/blob/master/blogs/babyweight/

Lab #7: Build an AppEngine app to serve ML predictions

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

In this lab, we will use AppEngine to invoke ML predictions

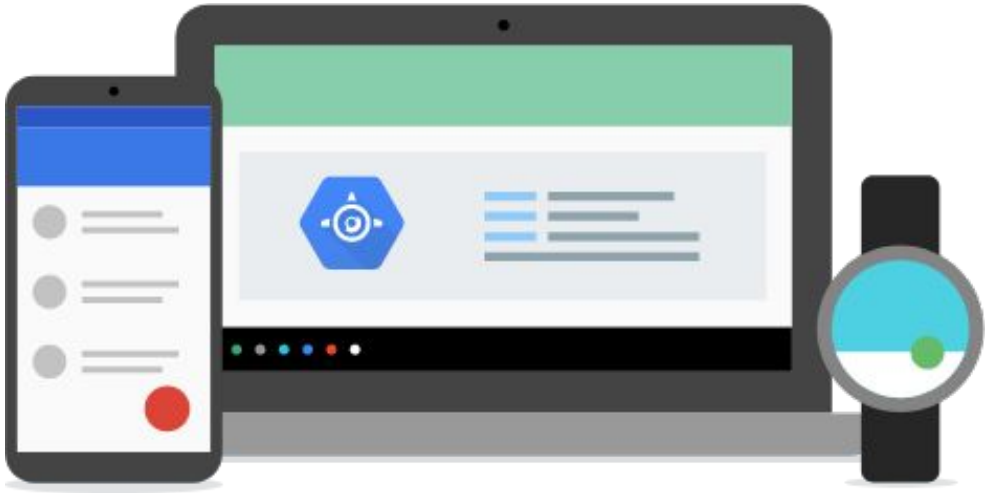


© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

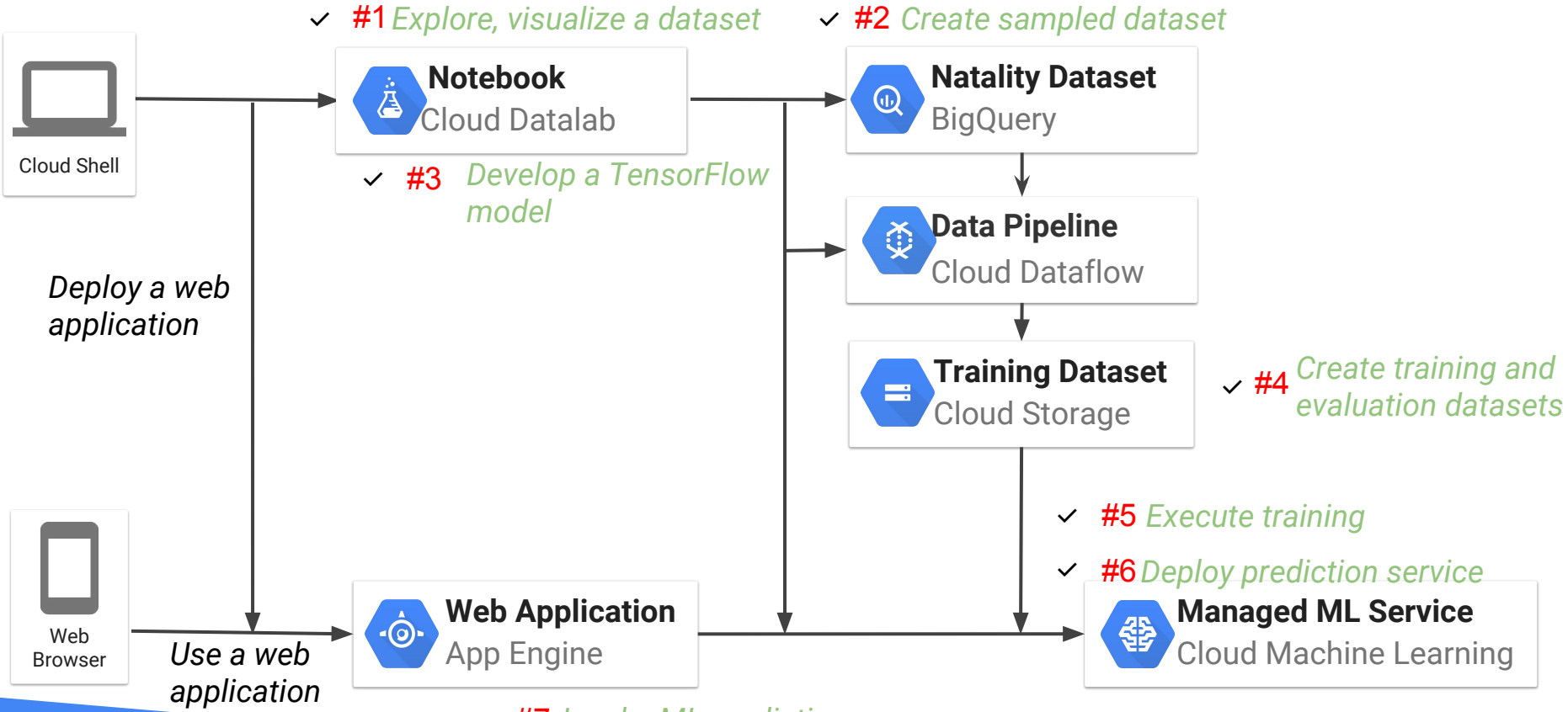
Google App Engine is a fully-managed service for building web backends

Supports Java, Node.js, Ruby, C#, Go, Python, and PHP



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Summary: An end-to-end process to operationalize ML models



= Lab Exercises

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

To learn more about TensorFlow on Google Cloud

Hands-on Tensorflow (next session)

goo.gl/nqsCo5

Serverless Machine Learning with TensorFlow bootcamp in Amsterdam

Thursday, October 19 (9:00 – 17:00) at GoDataDriven, Wibautstraat 202
www.godatafest.com/google

Serverless Machine Learning with TensorFlow on Coursera

https://www.coursera.org/learn/serverless-machine-learning-gcp/

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.


cloud.google.com
Images by Connie Zhou