Module 16

Exploring and Visualizing Large Datasets with Cloud Datalab

In this module we will:

- **Notebooks in the Cloud**
- Accessing BigQuery datasets from Cloud Datalab
- Visualizing Datasets in Charts
- Practice Reading ML Models

© 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

Increasingly, data analysis and machine learning are carried out in self-descriptive, shareable, executable notebooks

Share

Code

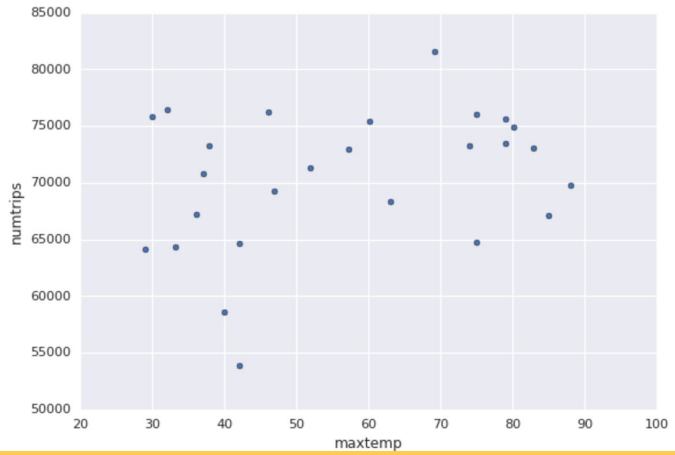
Output

Markup

Google Cloud Datalab demandforecast (autosaved)

Notebook Run Clear Reset python2

```
j = data[data['dayofweek'] == 7].plot(kind='scatter', x='maxtemp', y='numtrips')
```



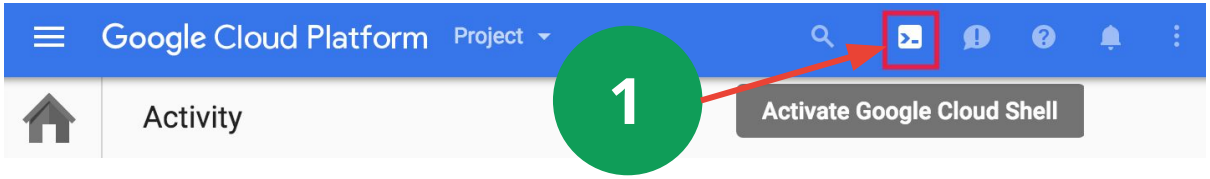
Removing the confounding factor does seem to reflect an underlying trend around temperature. But ... the data are a little sparse, don't you think? This is something that you have to keep in mind -- the more predictors you start to consider (here we are using two: day of week and maximum temperature), the more rows you will need so as to avoid *overfitting* the model.

A typical notebook contains code, charts, and explanations

© 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

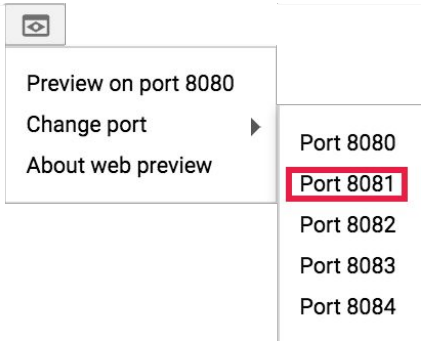
Starting Cloud Datalab in Cloud Shell is quite simple ...



2

```
datalab create my-datalab-vm \
              --machine-type  n1-highmem-8 \
              --zone us-central1-a
```

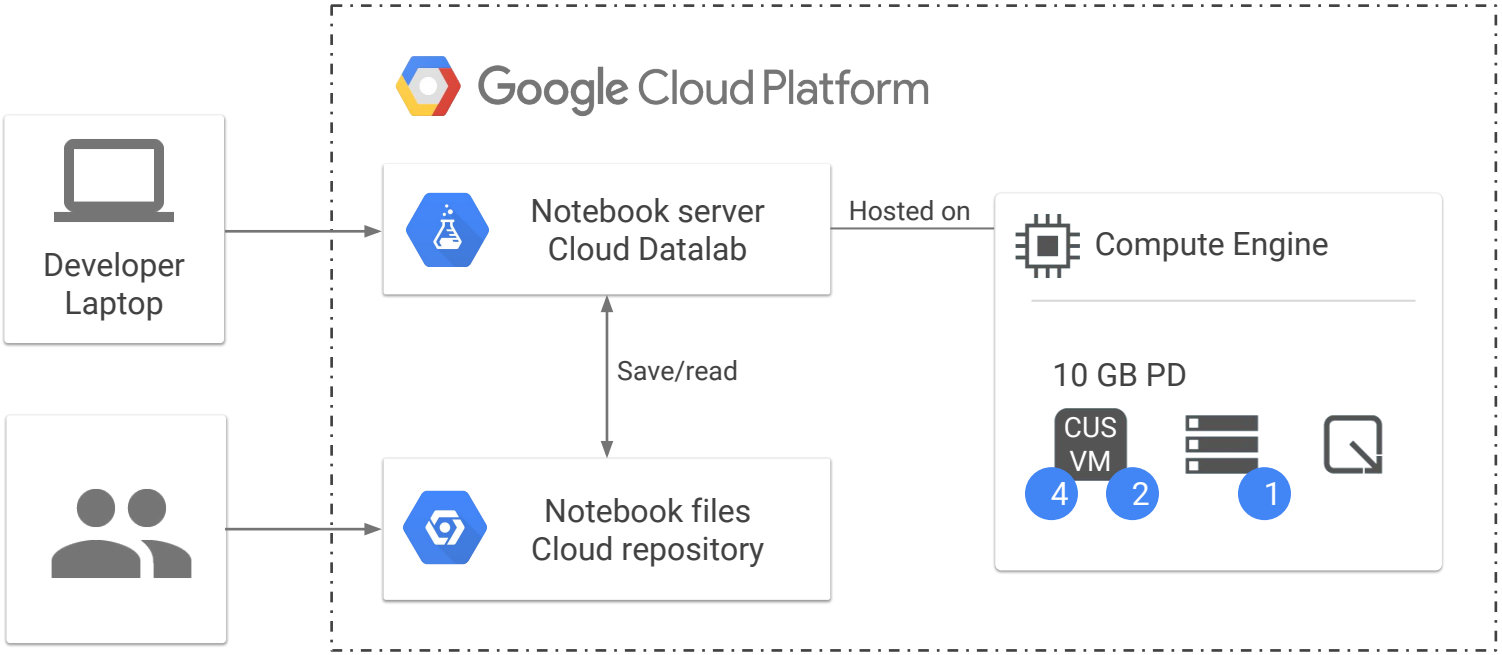
3



© 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

Datalab notebooks let you change the underlying hardware

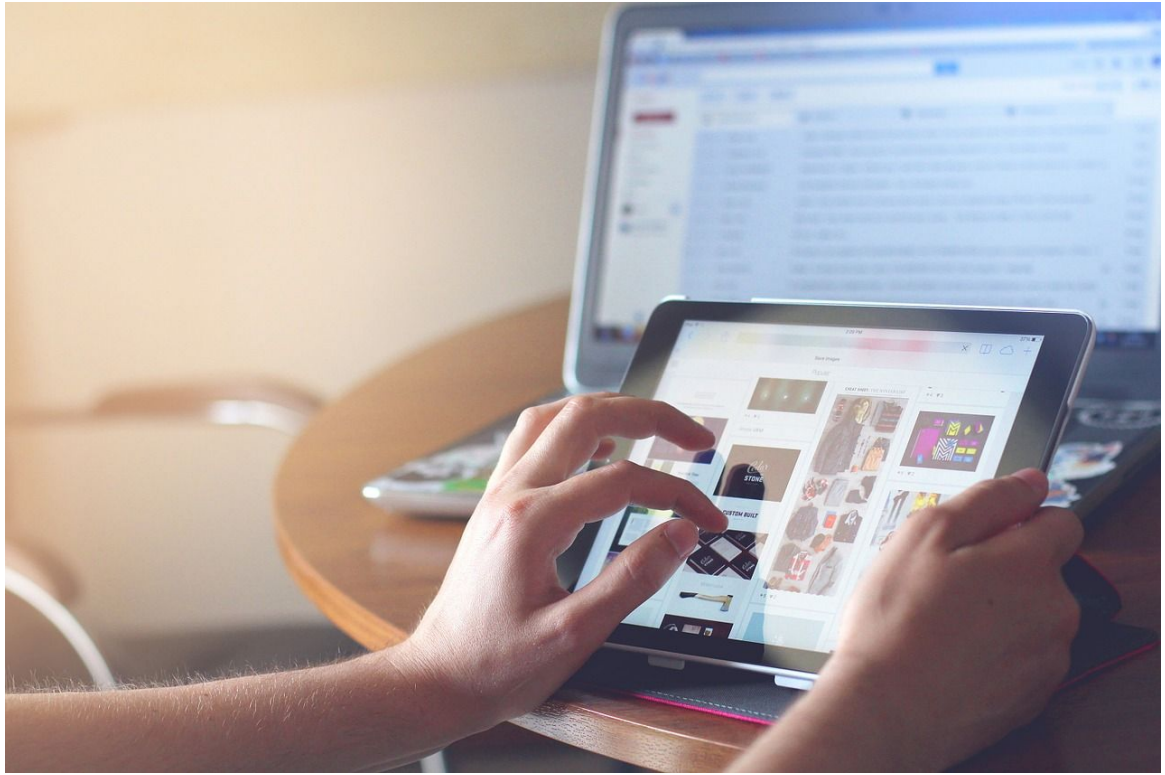


what happens to your work when you are through? You do want to stop paying for the datalab machine ...

© 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.

Demo: Analyzing ecommerce conversion data in Cloud Datalab

Track conversions and abandonments from home and product pages from an example ecommerce site



Google Analytics

© 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.

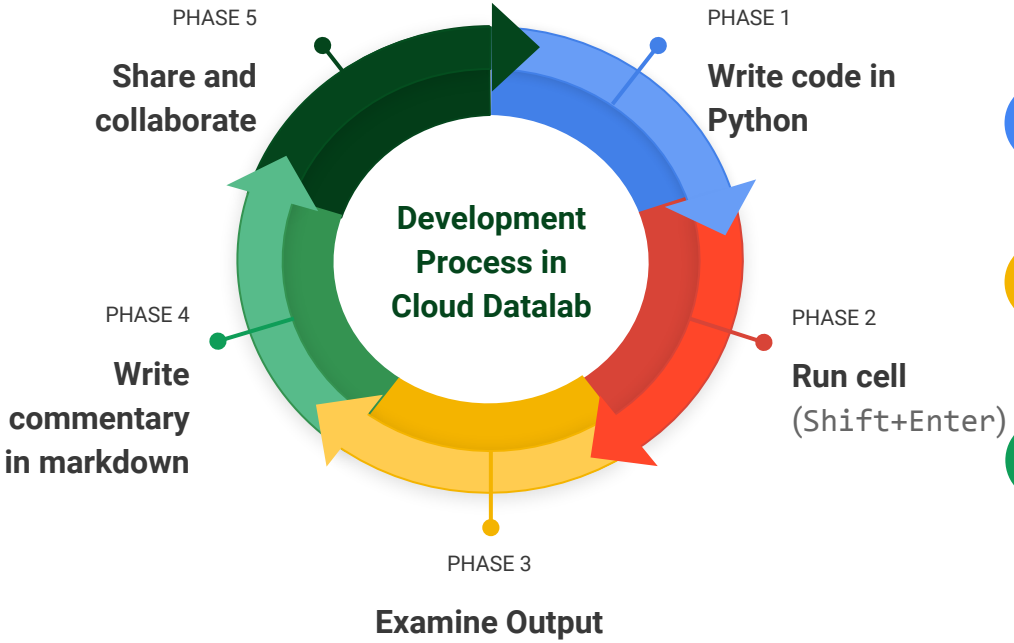
Google Cloud

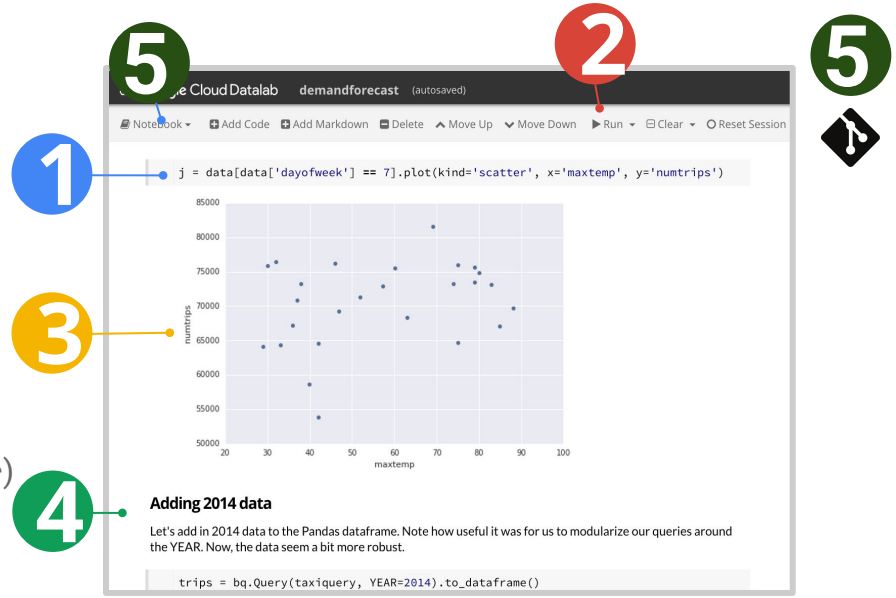
Demo of Cloud Datalab

© 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

Datalab notebooks are developed in an iterative, collaborative process





© 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

Users of the same GCP project can simply connect to your VM

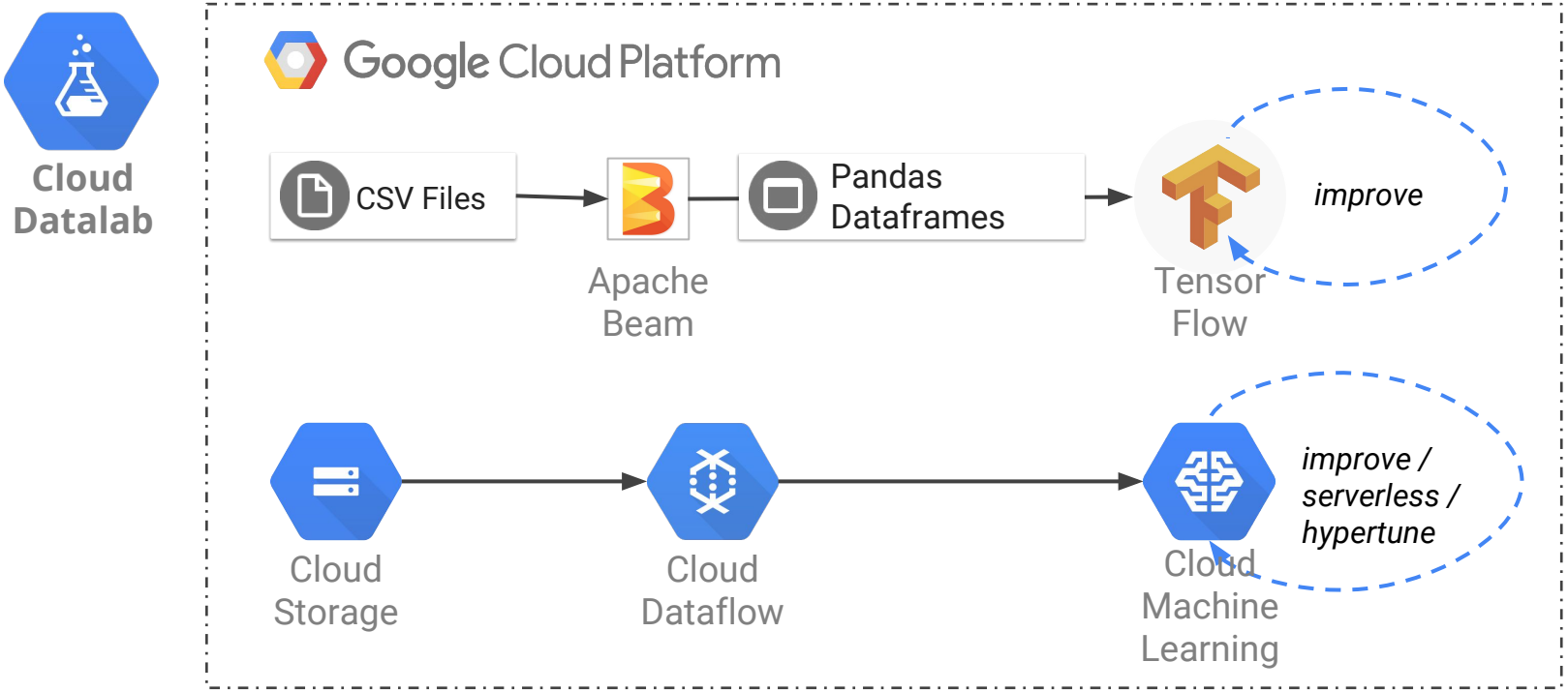
data-insights-evanjones

```
Welcome to Cloud Shell! Type "help" to get started.
jonesevan007@data-insights-evanjones:~$ datalab connect jonesevan007
```

Consider also pushing your notebook to a version control repository for sharing

© 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.

You can develop locally with Datalab and then scale out data processing to the cloud



© 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

Datalab integrates well with Google Cloud Platform products

Exploring and Analyzing	BigQuery, Google Cloud Storage
Machine Learning and Modeling	TensorFlow and GCML
Visualizing	Google Charts or Plotly or matplotlib
Seamless product combination	CMLE, Dataflow, CloudStorage
Integration	authentication and code source control

© 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.

Module 16

Exploring and Visualizing Large Datasets with Cloud Datalab

In this module we will:

- Notebooks in the Cloud
- **Accessing BigQuery datasets from Cloud Datalab**
- Visualizing Datasets in Charts
- Practice Reading ML Models

© 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

Import BigQuery as a library into Cloud Datalab to use it later

```
import google.datalab.bigquery as bq
```

© 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.

Use %%bq to execute BigQuery commands

```
%%bq tables describe -n "google.com:analytics-bigquery.LondonCycleHelmet.ga_sessions_20130910"
```

name	type	mode	description
visitorId	INTEGER	NULLABLE	
visitNumber	INTEGER	NULLABLE	
visitId	INTEGER	NULLABLE	
visitStartTime	INTEGER	NULLABLE	
date	STRING	NULLABLE	
totals	RECORD	NULLABLE	
trafficSource	RECORD	NULLABLE	
device	RECORD	NULLABLE	
customDimensions	RECORD	REPEATED	
hits	RECORD	REPEATED	
fullVisitorId	STRING	NULLABLE	
▸ **totals**			
▸ **trafficSource**			

Describing tables is a great way to explore schemas within Cloud Datalab

© 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

Write and execute queries and view results inside the notebook

```
%%bq query -n sessions
SELECT fullVisitorId, visitId, hit.hitNumber as hitNumber, hit.page.pagePath as path
FROM `google.com:analytics-bigquery.LondonCycleHelmet.ga_sessions_20130910`
  CROSS JOIN UNNEST(hits) as hit
ORDER BY visitStartTime, hitNumber
```

```
%bq execute --query sessions
```

	fullVisitorId	visitId	hitNumber	path
1	2879713562608983525	1,378,803,173	1	/
2	2879713562608983525	1,378,803,173	2	/vests/

Use -n to provide a name for the query as you can have more than one query in a single notebook

Consider validating complex queries inside the BigQuery web UI if you are getting syntax errors

© 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

BigQuery operations have defined parameters

%%bq datasets {list | create | delete}
%%bq tables
%%bq query
%%bq execute
%%bq extract
%%bq sample
%%bq dryrun
%%bq udf
%%bq load

```
%%bq datasets create --name "irs_990"
```

```
%%bq datasets list
```

- data-insights-evanjones.irs_990

© 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.

Keep the Cloud Datalab command reference handy

Google Cloud Datalab

Search docs

- google.datalab Module
- google.datalab.bigquery Module
- google.datalab.data Module
- google.datalab.ml Module
- google.datalab.stackdriver.monitoring Module
- google.datalab.storage Module
- google.datalab Commands
- mltoolbox.classification.dnn
- mltoolbox.classification.linear
- mltoolbox.regression.dnn
- mltoolbox.regression.linear
- mltoolbox.image.classification
- datalab.bigquery Module
- datalab.context Module
- datalab.data Module
- datalab.stackdriver.monitoring Module
- datalab.storage Module
- datalab Commands

Docs » google.datalab Commands

google.datalab Commands

%bq

```
usage: %bq [-h]
           {datasets,tables,query,execute,extract,sample,dryrun,udf,datasour
           ...

Execute various BigQuery-related operations. Use "%bq <command> -h" for help
on a specific command.

positional arguments:
  {datasets,tables,query,execute,extract,sample,dryrun,udf,datasource,load}
                        commands
    datasets            Operations on BigQuery datasets
    tables              Operations on BigQuery tables
    query               Create or execute a BigQuery SQL query object,
                        optionally using other SQL objects, UDFs, or externa
                        datasources. If a query name is not specified, the
                        query is executed.
    execute             Execute a BigQuery SQL query and optionally send the
                        results to a named table. The cell can optionally
                        contain arguments for expanding variables in the
                        query.
    extract             Extract a query or table into file (local or GCS)
    sample              Display a sample of the results of a BigQuery SQL
                        query. The cell can optionally contain arguments for
                        expanding variables in the query, if -q/--query was
                        used, or it can contain SQL for a query.
    dryrun              Execute a dry run of a BigQuery query and display
                        approximate usage statistics
    udf                 Create a named Javascript BigQuery UDF
    datasource          Create a named Javascript BigQuery external data
                        source
    load                Load data from GCS into a BigQuery table. If creatin
                        a new table, a schema should be specified in YAML or
                        JSON in the cell body, otherwise the schema is
```

In Scope:
BQ commands
Charting commands
ML Toolbox commands (pre-built)

Out of Scope: (covered in Data Engineering)
ML / TensorFlow commands

Access it here:

http://googledatalab.github.io/pydatalab/

© 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

BigQuery commands summary

- Having your dataset and table creation
- Running multiple queries in succession in the same window
- Quickly manipulate and export CSV files to GCS
- Running more than just BigQuery commands (charting, ML)
- Clearly and cleanly document your work with markdown
- Version control and share your notebooks
- Option to develop locally in datalab (e.g. Pandas) then uplevel to cloud (e.g. BigQuery, Dataflow) when ready

© 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.

Module 16

Exploring and Visualizing Large Datasets with Cloud Datalab

In this module we will:

- Notebooks in the Cloud
- Accessing BigQuery datasets from Cloud Datalab
- **Visualizing Datasets in Charts**
- Practice Reading ML Models

© 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

Visualize tables or query results with charting commands

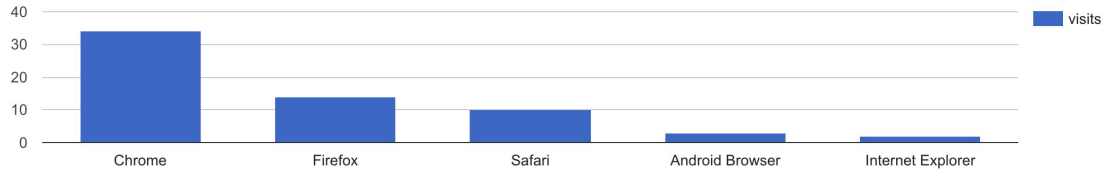
```
%%bq query --name visits_by_browser
SELECT
  SUM(totals.visits) AS visits,
  device.browser AS browser
FROM
  `google.com:analytics-bigquery.LondonCycleHelmet.ga_sessions_20130910`
GROUP BY
  browser
ORDER BY
  visits DESC
```

```
%%bq execute --query visits_by_browser
```

visits	browser
34	Chrome
14	Firefox
10	Safari
3	Android Browser
2	Internet Explorer

(rows: 5, time: 1.1s, 1KB processed, job: job_FFciA5SL_0-wqQwFpt9ol6tLF43l)

```
%%chart columns --data visits_by_browser --fields browser,visits
```



Get charting data ready in a table, query, dataframe, or list

Specify chart type, data source, and fields to visualize

© 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

Each chart type may have different required parameters

%%chart area
%%chart bars
%%chart bubbles
%%chart columns
%%chart heatmap
%%chart histogram
%%chart line
%%chart treemap
%%chart scatter
....and many more

Use --help to view parameters

```
%%chart bars --help
```

```
usage: %%chart bars [-h] [-f FIELDS] -d DATA

Generate a bars chart.

optional arguments:
  -h, --help            show this help message and exit
  -f FIELDS, --fields FIELDS
                        The field(s) to include in the c
  -d DATA, --data DATA  The name of the variable referen
                        Query to chart
```

© 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.

Module 16

Exploring and Visualizing Large Datasets with Cloud Datalab

In this module we will:

- Notebooks in the Cloud
- Accessing BigQuery datasets from Cloud Datalab
- Visualizing Datasets in Charts
- **Practice Reading ML Models**

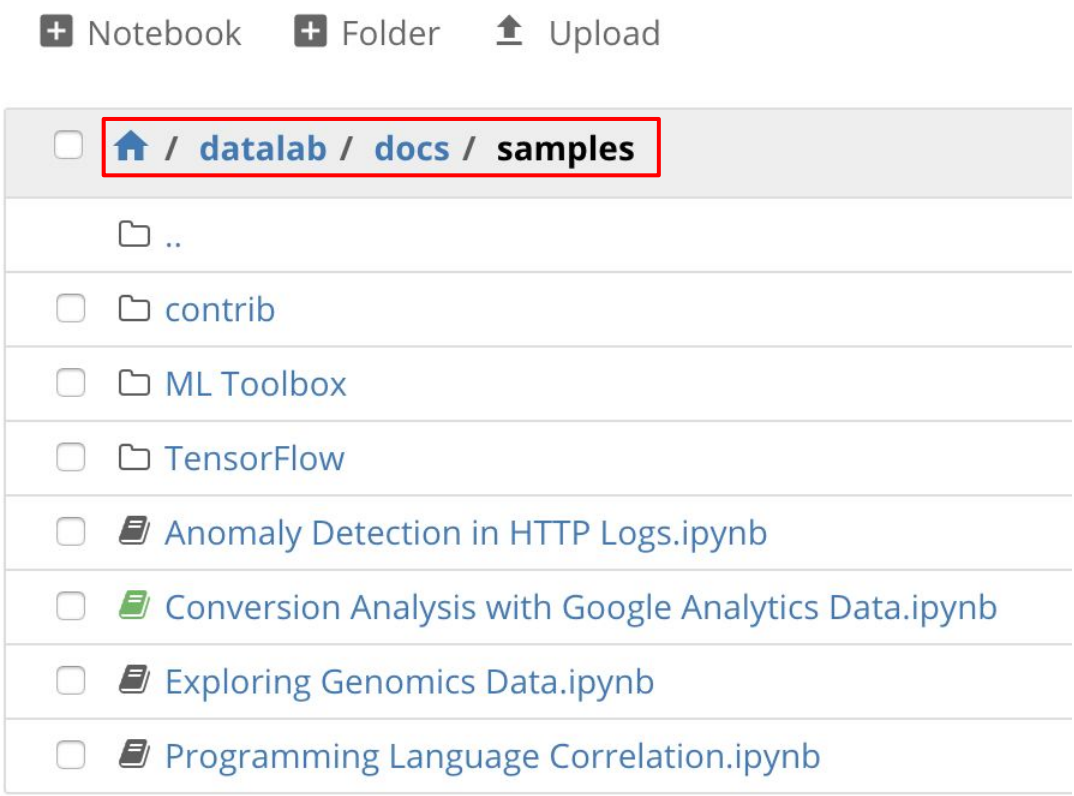
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Your Cloud Datalab instance comes preloaded with samples

Access them under docs and samples

Here is our Google Analytics example →



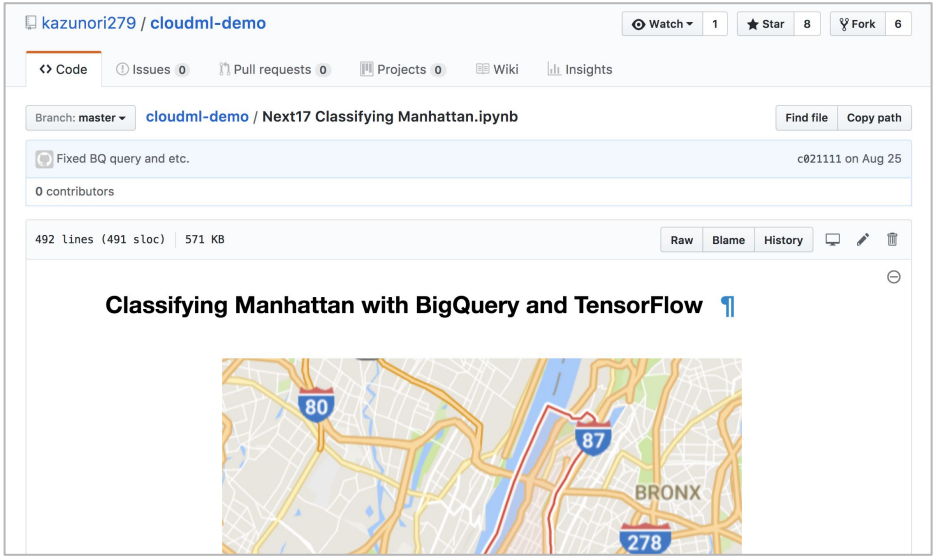
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

You can load any iPython notebook into Cloud Datalab

Fork and copy data science notebooks that you want to practice with

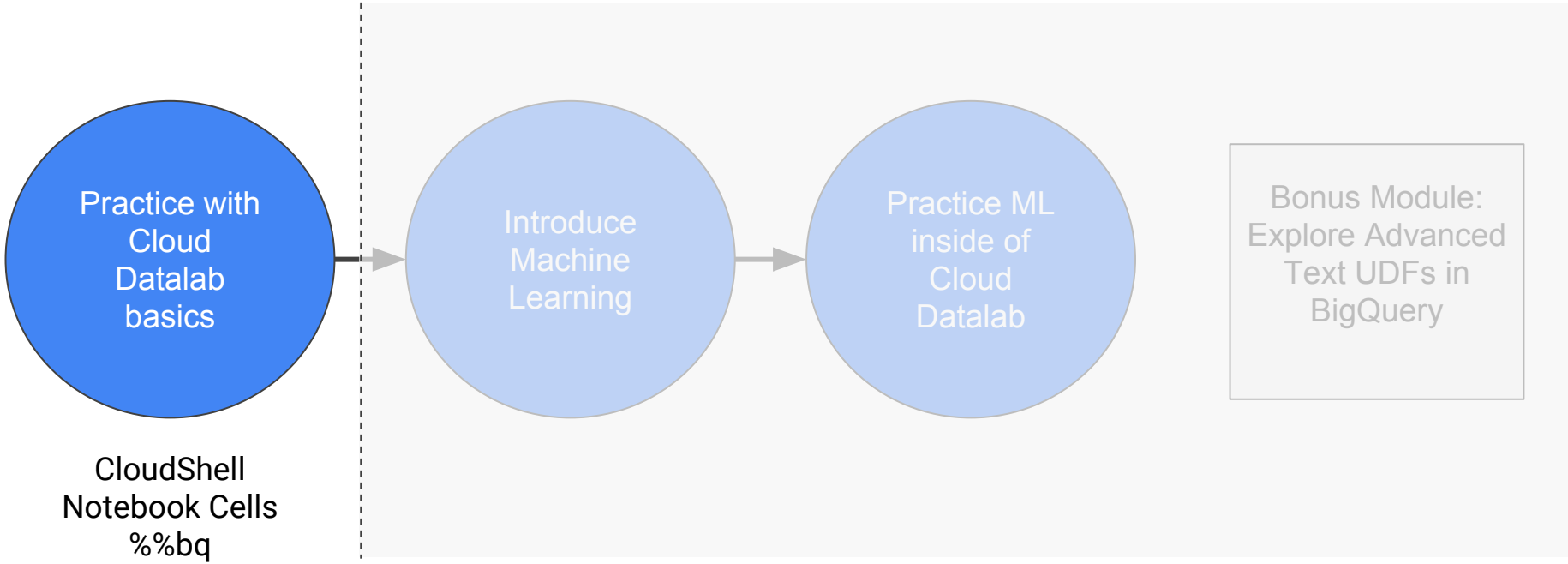
GitHub is a good starting place to find iPython notebooks



© 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

Recap: This module covered the basics of Cloud Datalab

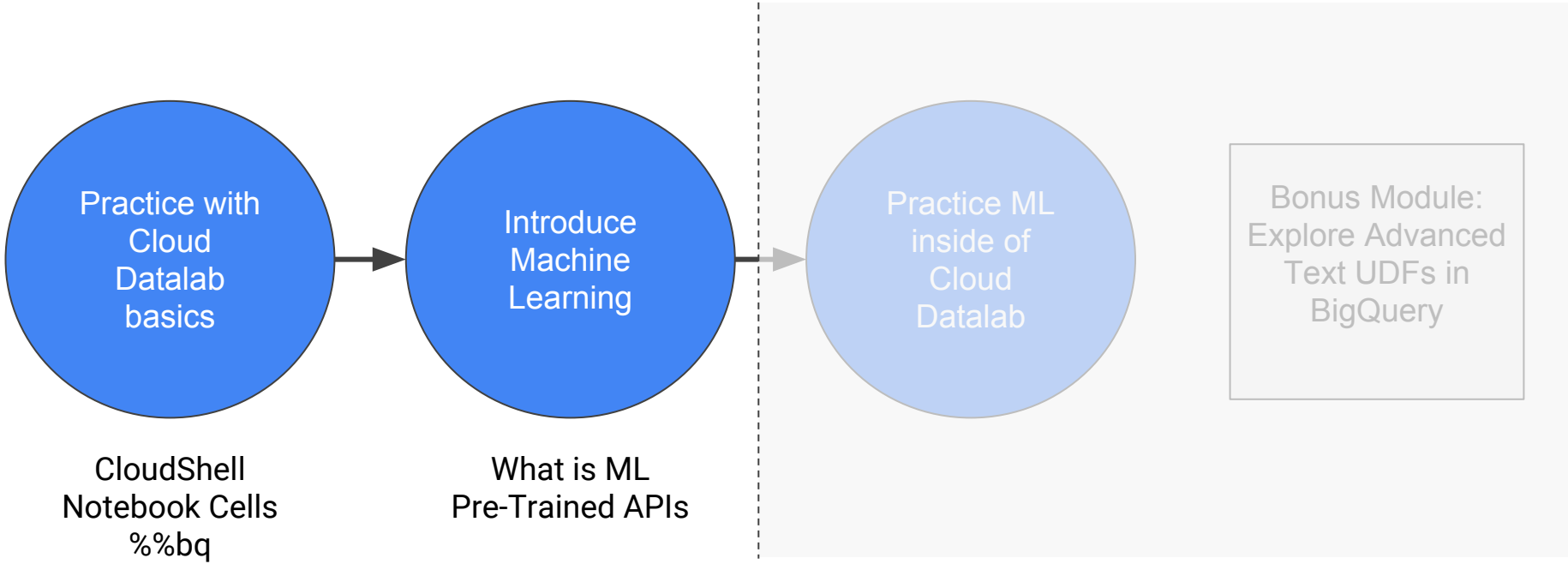


CloudShell
Notebook Cells
%%bq
%%chart

© 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

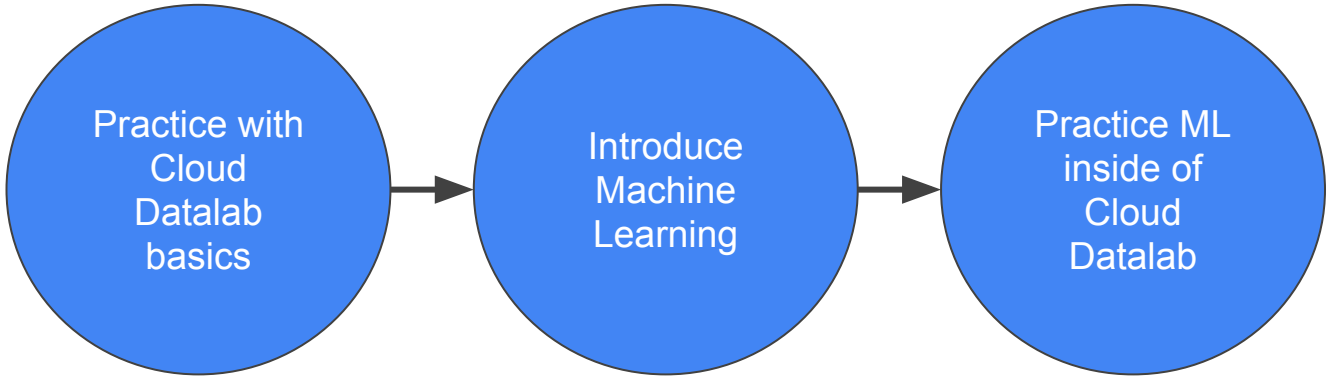
The next module covers the basics of machine learning



© 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

Then, the next module you practice ML within Cloud Datalab



CloudShell
Notebook Cells
%%bq
%%chart

What is ML
Pre-Trained APIs

ML Toolbox
ML Model Phases
%%ml

Bonus Module: Explore Advanced Text UDFs in BigQuery

© 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

ML Preview: Practice Reading an ML notebook

Preprocess in BigQuery

Preprocess the training data on BigQuery

In [11]:
```
%%sql --module nyc_collisions
SELECT
  IF(borough = 'MANHATTAN', 1, 0) AS is_mt,
  latitude,
  longitude
FROM
  `bigquery-public-data.new_york.nypd_mv_collisions`
WHERE
  LENGTH(borough) > 0
  AND latitude IS NOT NULL AND latitude != 0.0
  AND longitude IS NOT NULL AND longitude != 0.0
  AND borough != 'BRONX'
ORDER BY
  RAND()
LIMIT
  10000
```

Leverage ML models

Define a neural network

In [15]:
```
import tensorflow as tf
tf.logging.set_verbosity(tf.logging.ERROR) # supress warning messages

# define two feature columns with real values
feature_columns = [tf.contrib.layers.real_valued_column("", dimension=2)]

# create a neural network
dnnc = tf.contrib.learn.DNNClassifier(
  feature_columns=feature_columns,
  hidden_units=[20, 20, 20, 20],
  n_classes=2)

dnnc
```

© 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 16
Connecting BigQuery to Cloud Datalab

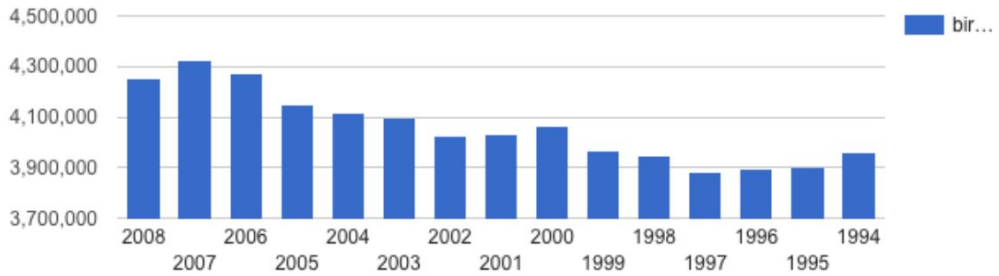
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Lab: Connecting BigQuery to Cloud Datalab

Flex your BigQuery skills inside of Cloud Datalab and learn how you can analyze, visualize, and document your findings all in a single notebook

```
%chart columns --data total_births --fields year,birth_count
```

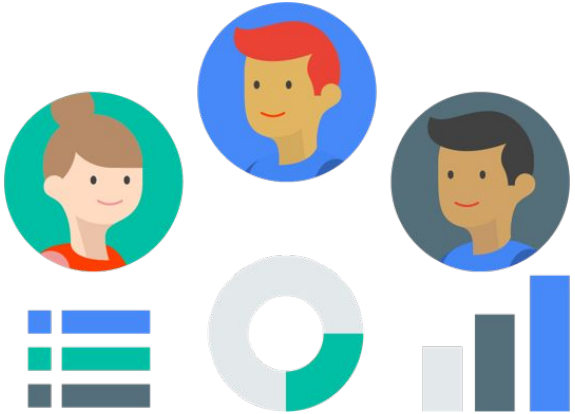


© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Cloud Datalab notebooks are collaborative and built for data

Access and Process your Data with BigQuery



Visualize Query Results in Charts and Graphs

Document, Share, and Collaborate

© 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.