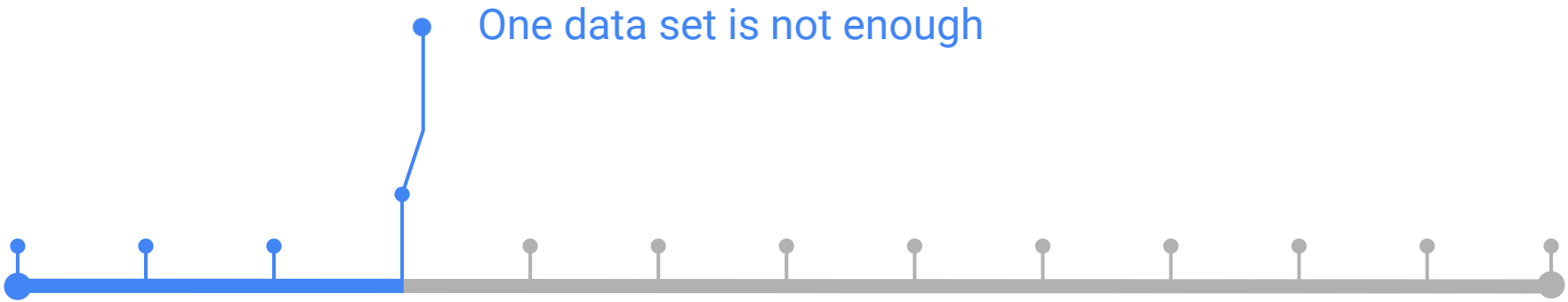
Deep Learning Walkthrough - 03

Cassie Kozyrkov
Chief Decision Scientist, Google Cloud
GitHub: kozyrkov; Twitter: @quaesita

Google Cloud

Step 3 | Split your data



Chief actors:

Engineer

Why split?

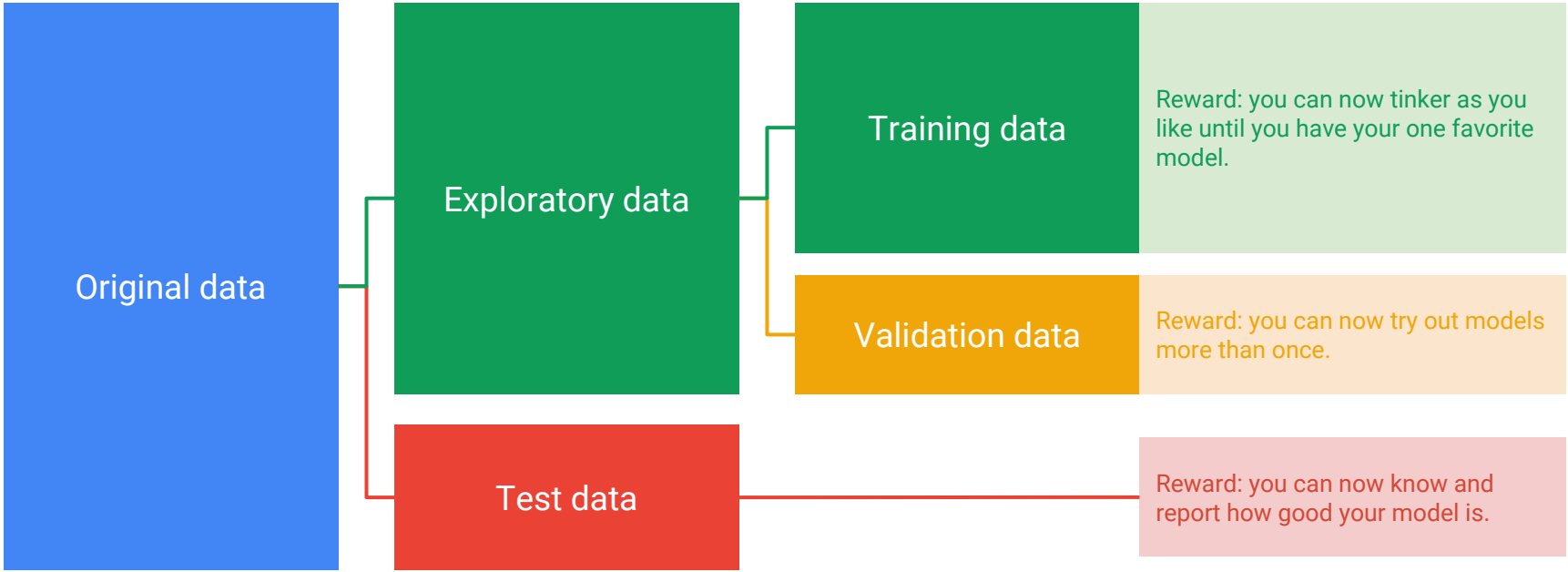
Your ML system is no good to you if it can't deal with new data

Be afraid, be very afraid

It's too easy to build a system that's really good at old data but fails miserably on new

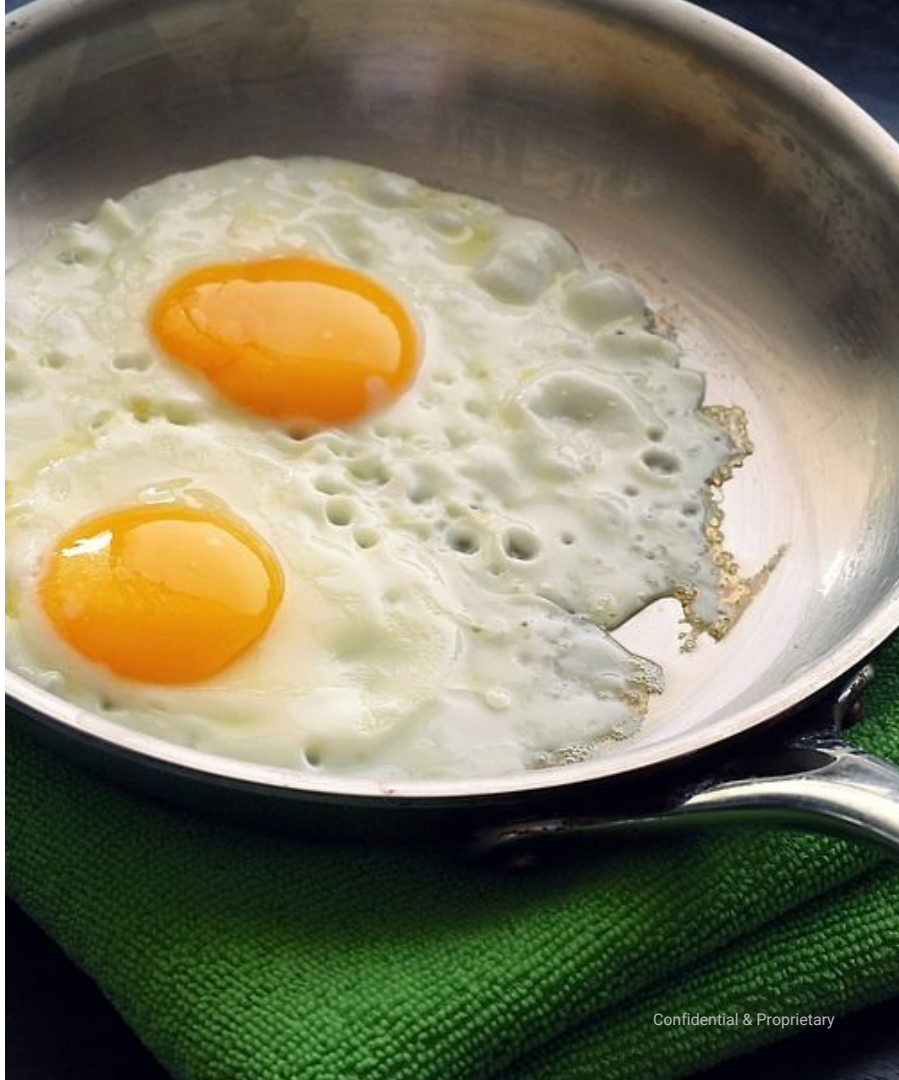
Set some data aside for careful evaluation

Data splitting



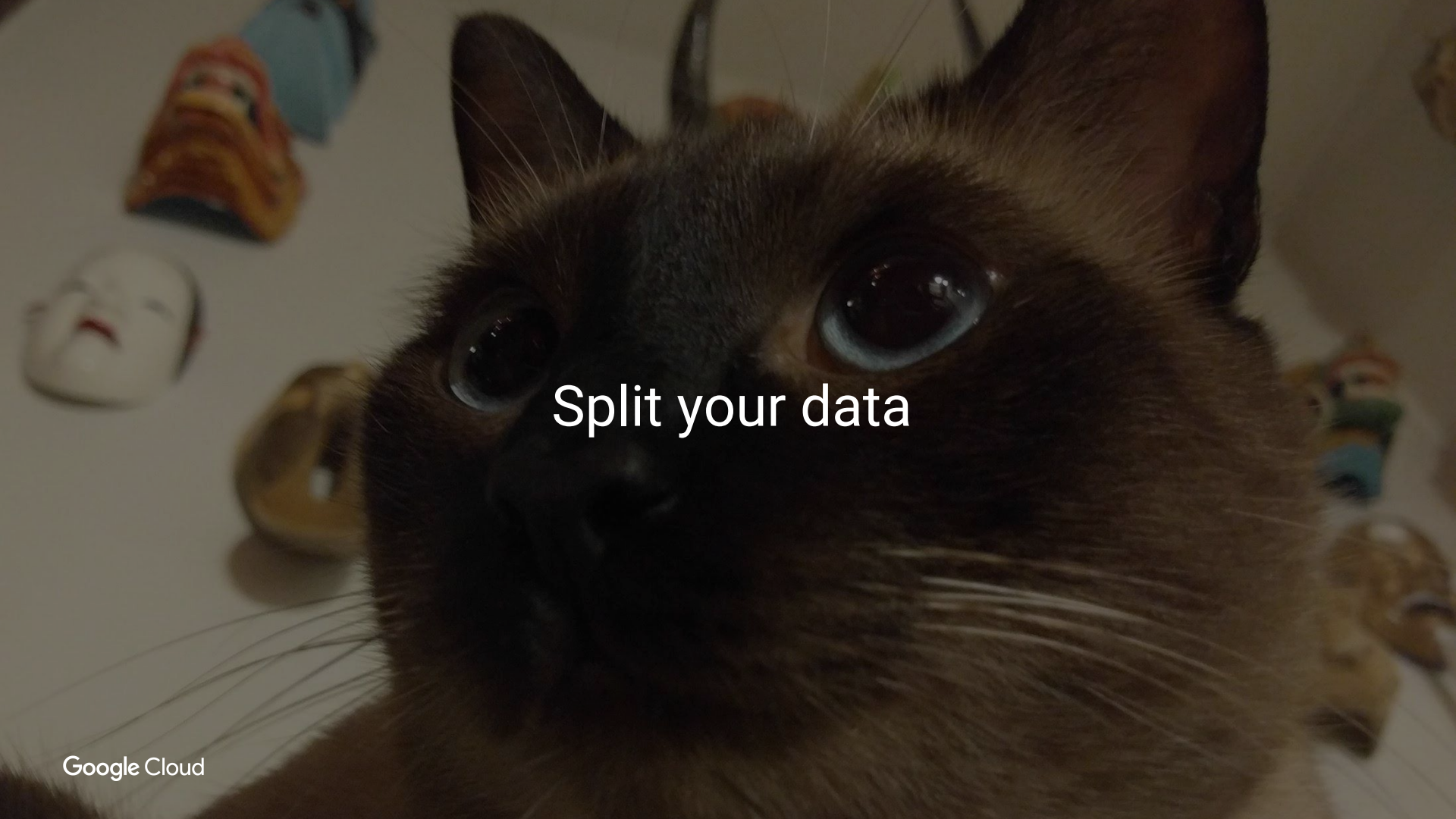
In practice

Time can split your data.



Danger! Pitfall alert

The world represented by your training data is the only world you can expect to succeed in!



Google Cloud

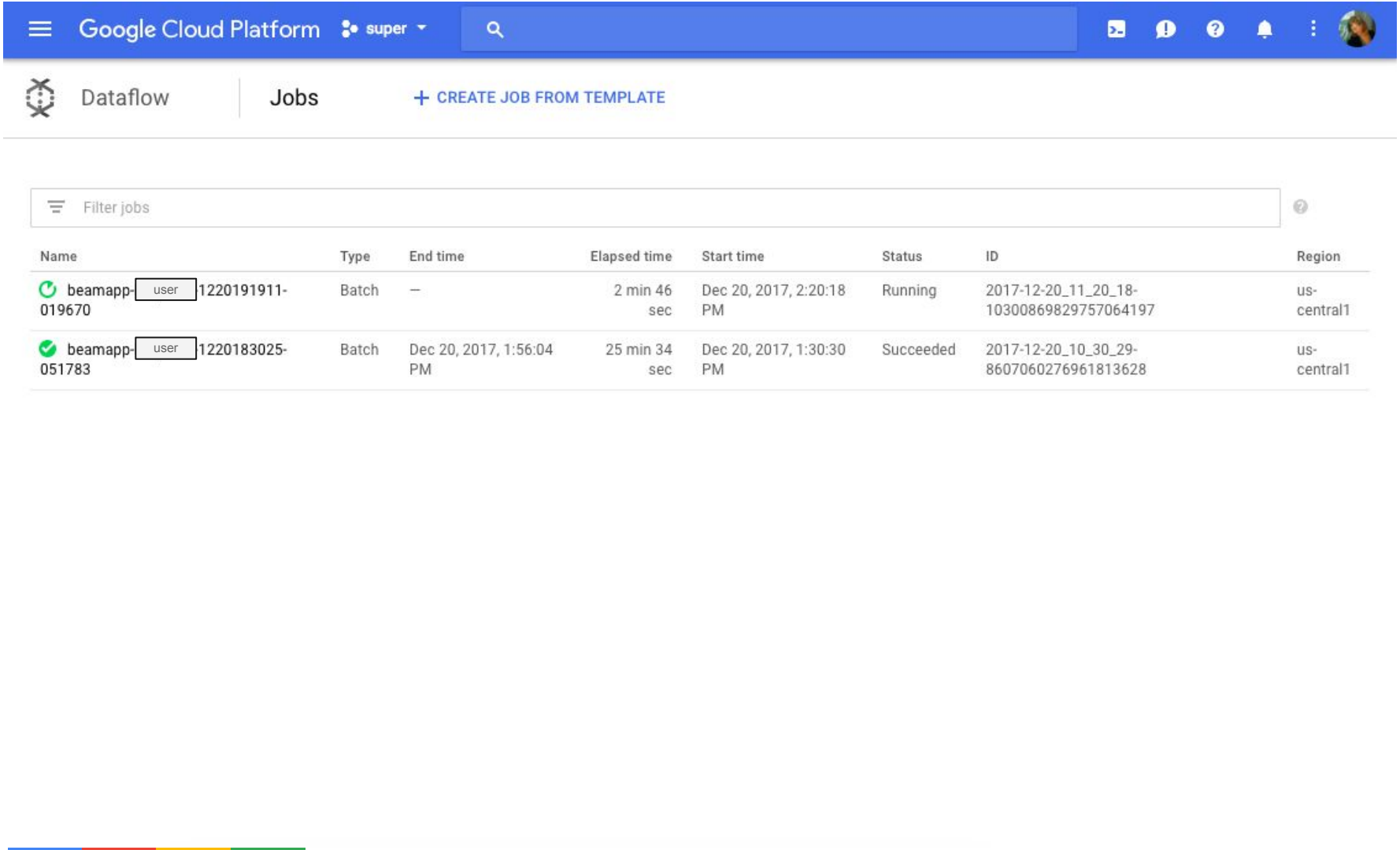
Secure | https://ssh.cloud.google.com/projects/super-188716/zones/us-east1-c/instances/super-188716-compute-instance?authus...

```
(env)  user  !super-188716-compute-instance:~/sc17/cats$ sh run_step_3_split_images.sh $PROJECT $BUCKET catimages
No handlers could be found for logger "oauth2client.contrib.multistore_file"
/home/  user  /env/local/lib/python2.7/site-packages/apache_beam/io/gcp/gcsio.py:122: DeprecationWarning: object()
takes no parameters
  super(GcsIO, cls).__new__(cls, storage_client))
INFO:oauth2client.transport:Attempting refresh to obtain initial access_token
INFO:oauth2client.transport:Attempting refresh to obtain initial access_token
INFO:oauth2client.transport:Attempting refresh to obtain initial access_token
INFO:root:Starting training_images(0.5) | validation_images(0.3) | test_images(0.2) split from images with source f
ile pattern gs://super-188716-bucket/catimages/all_images/*
INFO:root:Destination parent directory: gs://super-188716-bucket/catimages/
INFO:root:Starting the size estimation of the input
INFO:oauth2client.transport:Attempting refresh to obtain initial access_token
INFO:root:Finished the size estimation of the input at 1 files. Estimation took 0.889538049698 seconds
INFO:root:Starting the size estimation of the input
INFO:oauth2client.transport:Attempting refresh to obtain initial access_token
INFO:root:Finished computing size of: 10000 files
INFO:root:Finished computing size of: 20000 files
INFO:root:Finished computing size of: 30000 files
INFO:root:Finished computing size of: 40000 files
INFO:root:Finished computing size of: 50000 files
```

sh run_step_2b_get_images.sh $PROJECT dataset catinfo $BUCKET catimages

--split-names training_images validation_images test_images

--split-fractions 0.5 0.3 0.2


Google Cloud Platform
super
Dataflow
Jobs
+ CREATE JOB FROM TEMPLATE
Filter jobs
Name
Type
End time
Elapsed time
Start time
Status
ID
Region
beamapp- user 1220191911-019670
Batch
—
2 min 46 sec
Dec 20, 2017, 2:20:18 PM
Running
2017-12-20_11_20_18-10300869829757064197
us-central1
beamapp- user 1220183025-051783
Batch
Dec 20, 2017, 1:56:04 PM
25 min 34 sec
Dec 20, 2017, 1:30:30 PM
Succeeded
2017-12-20_10_30_29-8607060276961813628
us-central1

beamapp-[user]-1220191911-019670 LOGS

Job

Job summary

Job name	beamapp-[user]-1220191911-019670
Job ID	2017-12-20_11_20_18-10300869829757064197
Region	us-central1
Job status	Running Stop job
SDK version	Apache Beam SDK for Python 2.2.0
Job type	Batch
Start time	Dec 20, 2017, 2:20:18 PM
Elapsed time	3 min 11 sec

Autoscaling

Workers	20
Current state	Autoscaling: Raised the number of workers to 20 based on the rate of progress in the currently running step(s). Dec 20, 2017 2:20 PM

beamapp-[user]-1220191911-019670 LOGS

Job

Job summary

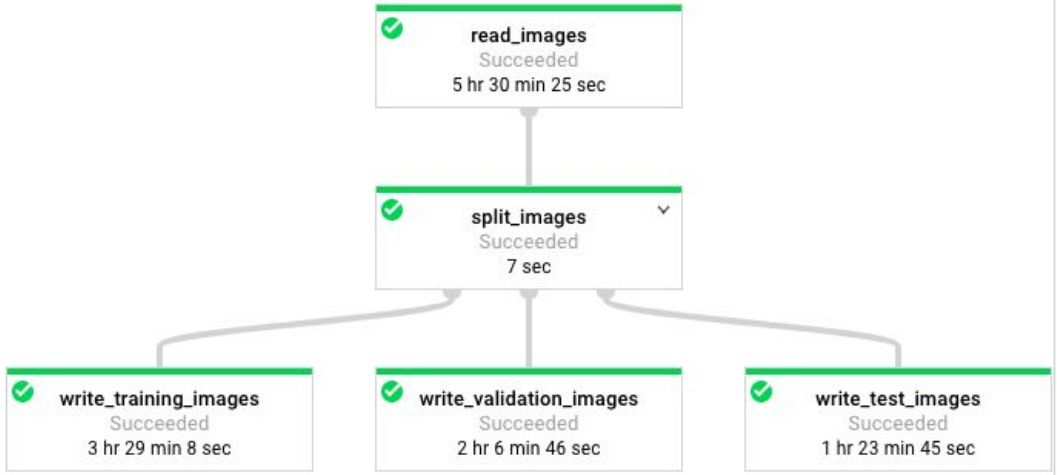
Job name	beamapp-[user]-1220191911-019670
Job ID	2017-12-20_11_20_18-10300869829757064197
Region	us-central1
Job status	Running Stop job
SDK version	Apache Beam SDK for Python 2.2.0
Job type	Batch
Start time	Dec 20, 2017, 2:20:18 PM
Elapsed time	8 min 43 sec

Autoscaling

Workers	20
Current state	Autoscaling: Raised the number of workers to 20 based on the rate of progress in the currently running step(s). Dec 20, 2017 2:20 PM

Google Cloud Platform super

beamapp-[user]-1220191911-019670 LOGS



Job

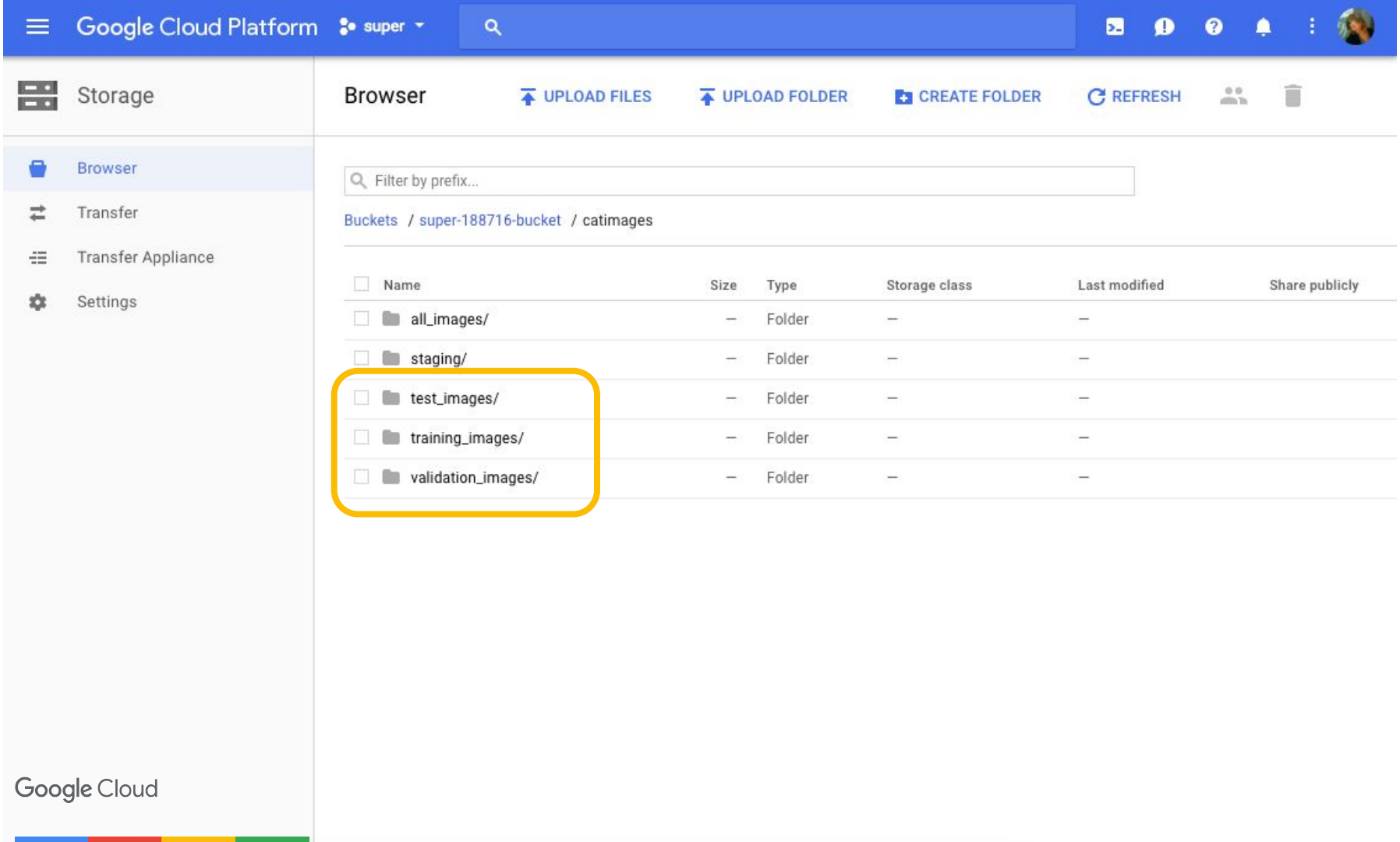
Job summary

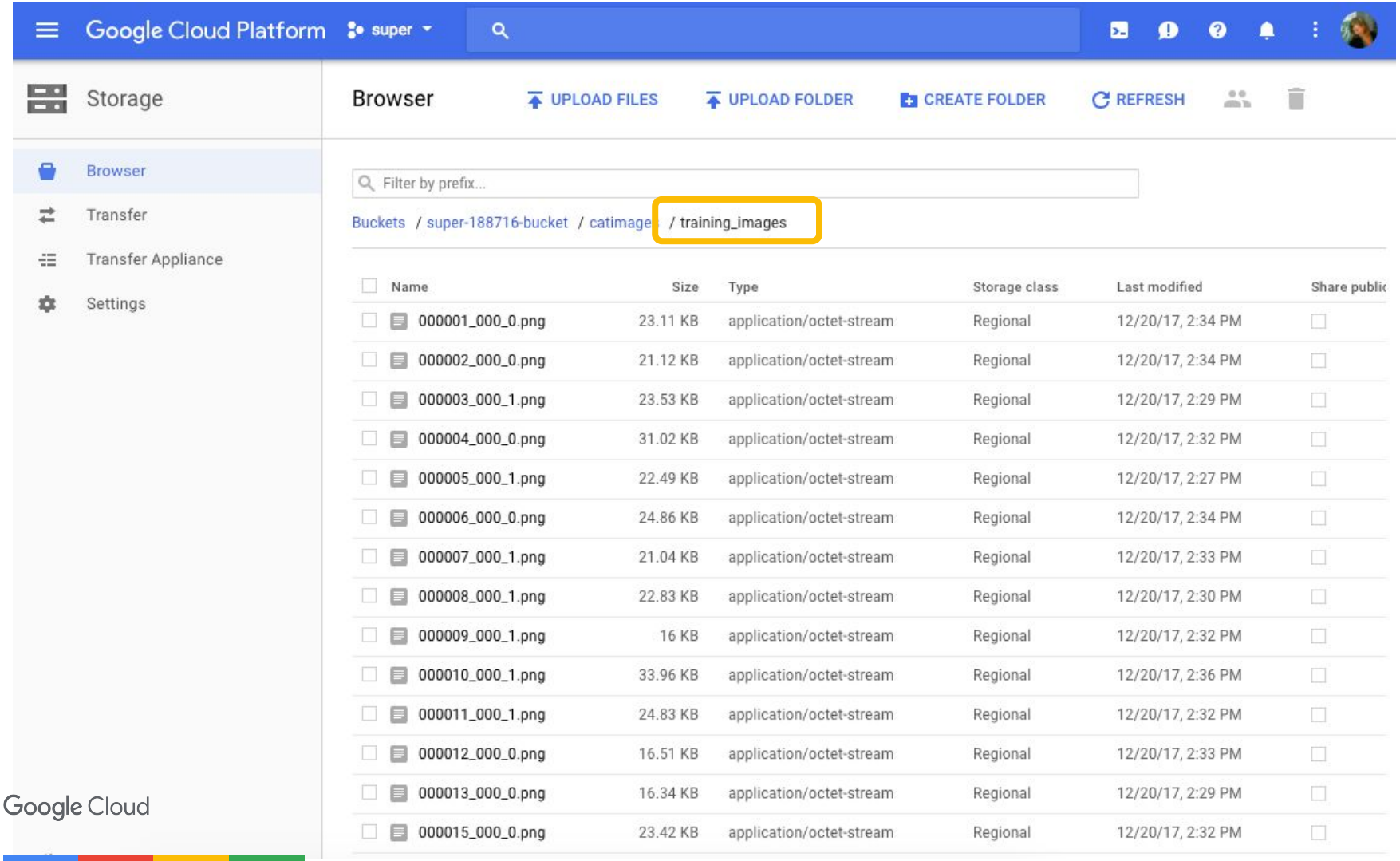
Job name	beamapp-[user]-1220191911-019670
Job ID	2017-12-20_11_20_18-10300869829757064197
Region	us-central1
Job status	Succeeded
SDK version	Apache Beam SDK for Python 2.2.0
Job type	Batch
Start time	Dec 20, 2017, 2:20:18 PM
Elapsed time	17 min 52 sec

Autoscaling

Workers	0
Current state	Autoscaling: Raised the number of workers to 20 based on the rate of progress in the currently running step(s).

Dec 20, 2017 2:20 PM

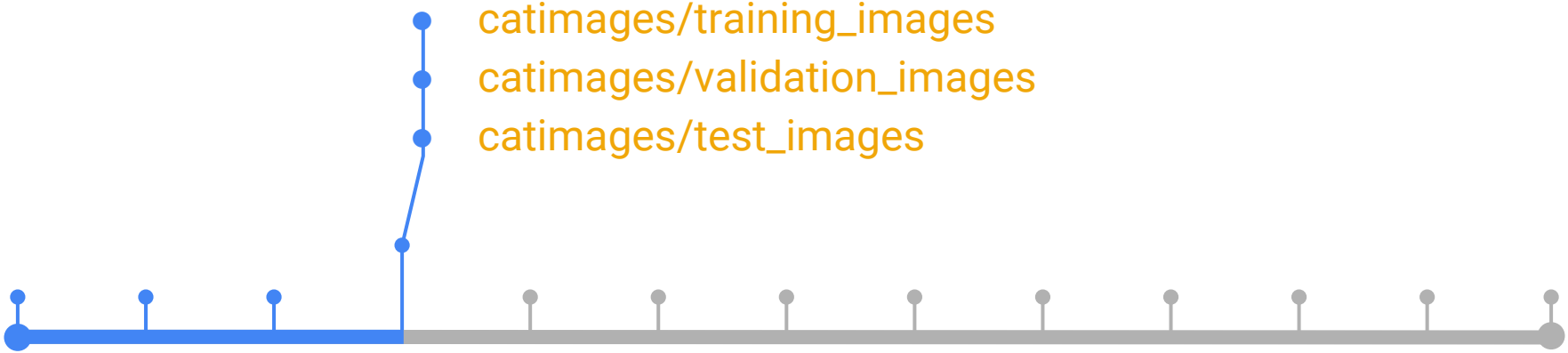

Google Cloud Platform
super
Storage
Browser
Transfer
Transfer Appliance
Settings
Browser
UPLOAD FILES
UPLOAD FOLDER
CREATE FOLDER
REFRESH
Filter by prefix...
Buckets / super-188716-bucket / catimages
Name
Size
Type
Storage class
Last modified
Share publicly
all_images/ – Folder – –
staging/ – Folder – –
test_images/ – Folder – –
training_images/ – Folder – –
validation_images/ – Folder – –
Google Cloud



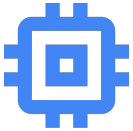
Google Cloud

Step 3 is finished | You now have IDs of instances that go in:

Step 3 is finished | You now have IDs of instances that go in:



Chief actors:



Engineer