Deep Learning Walkthrough - 03

Code in github.com/google-aai/sc17

Cassie Kozyrkov

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



Google Cloud

Step 3 | Split your data



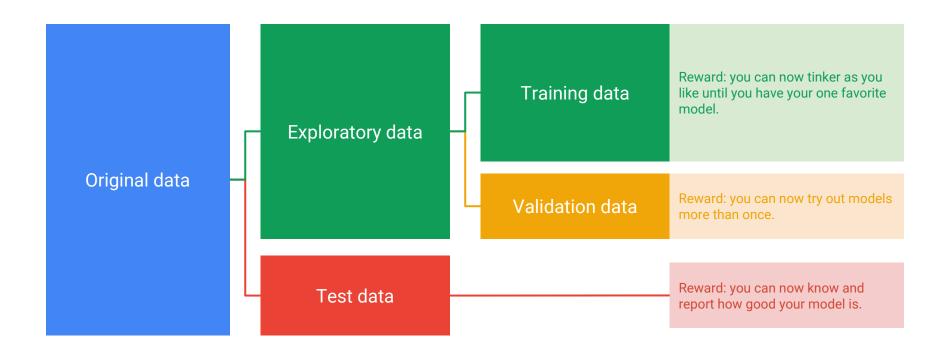
Why split?

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



Ensure you have new data for evaluation by setting some aside

Data splitting



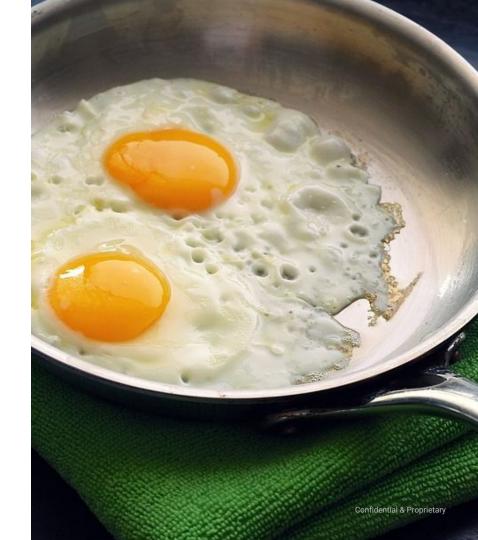
Google Cloud

In practice



Time can split your data for you.

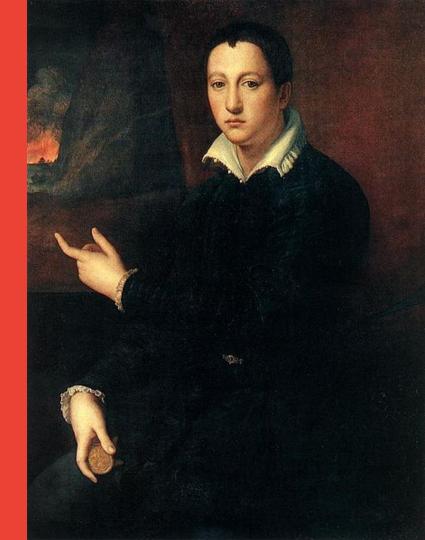
It's okay to explore everything you've gathered so far as long as you will collect more data when it's time to evaluate.

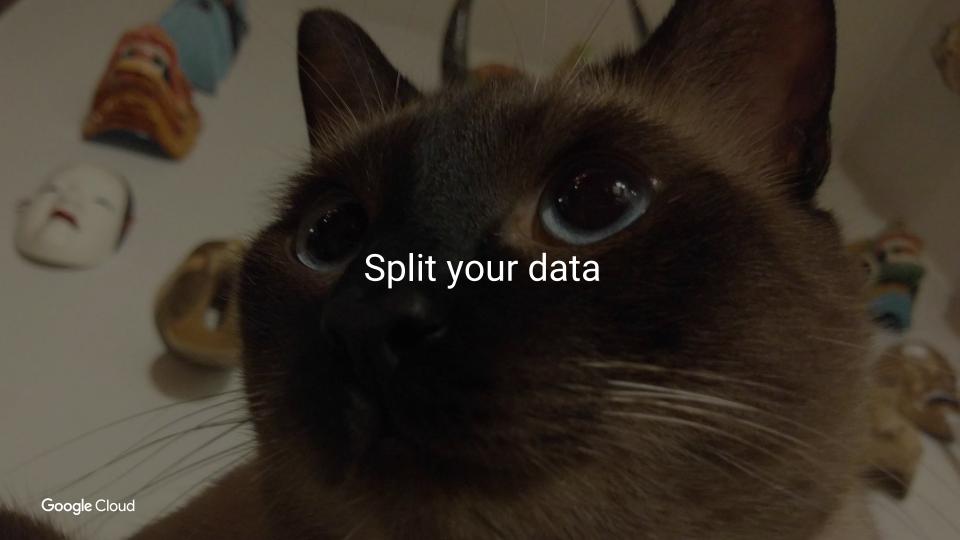


Danger! Pitfall alert

Ensure that all your datasets are relevant and representative of the environment you'll be using your ML system in.

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





```
Secure https://ssh.cloud.google.com/projects/super-188716/zones/us-east1-c/instances/super-188716-compute-instance?authus...
              super-188716-compute-instance:~/sc17/cats$ sh run step 3 split images.sh $PROJECT $BUCKET catimages
(env)
No handlers could be found for logger "oauth2client.contrib.multistore file"
              /env/local/lib/python2.7/site-packages/apache beam/io/gcp/gcsio.py:122: DeprecationWarning: object()
/home/
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 qs://super-188716-bucket/catimages/all images/*
INFO:root:Destination parent directory: qs://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

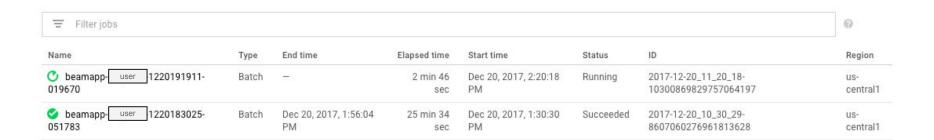
--split-fractions 0.5 0.3 0.2

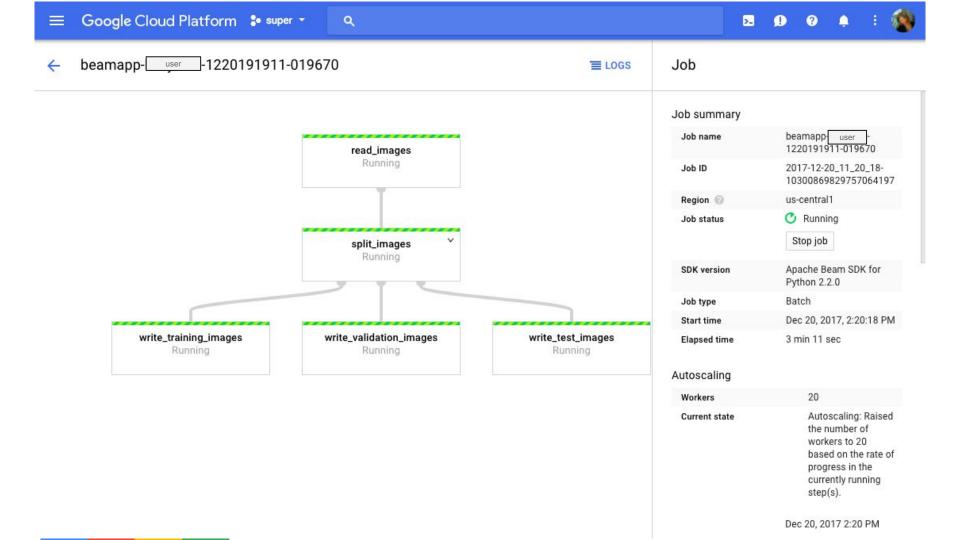




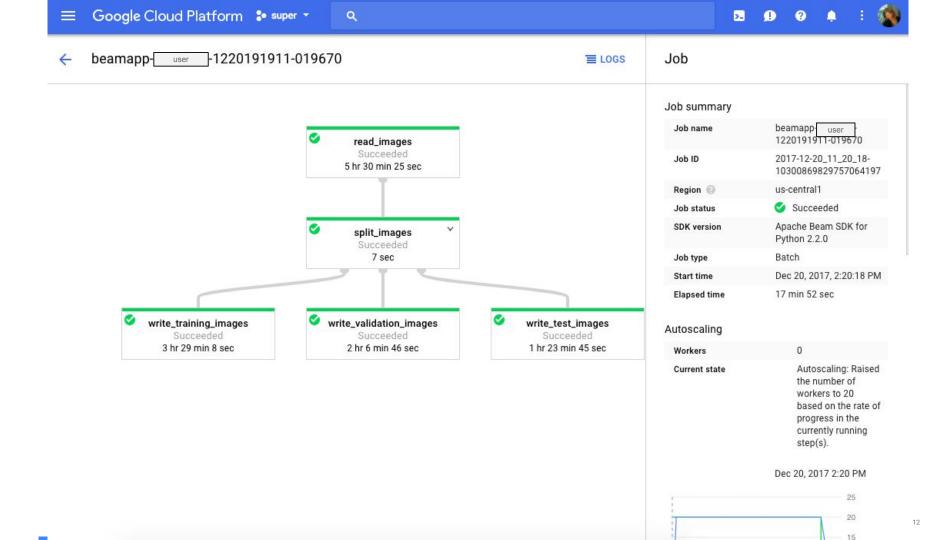
Jobs

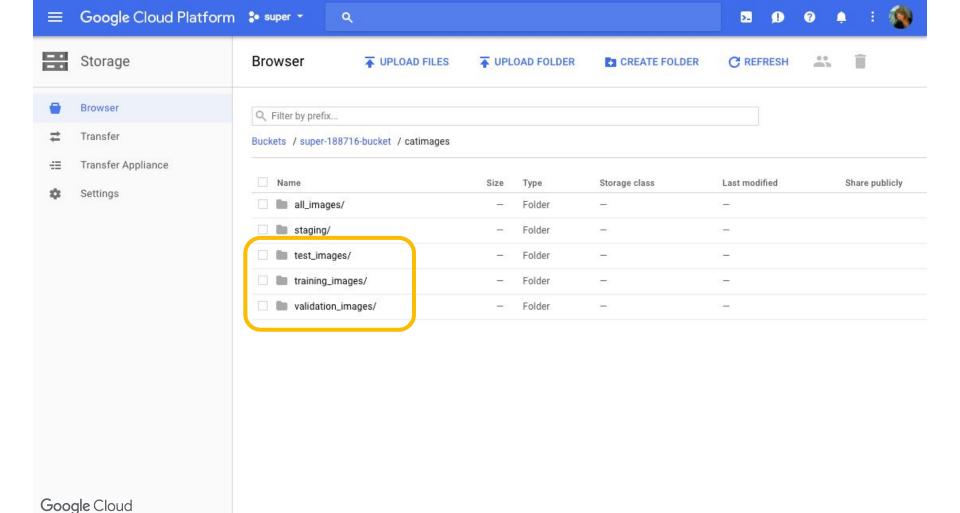
+ CREATE JOB FROM TEMPLATE

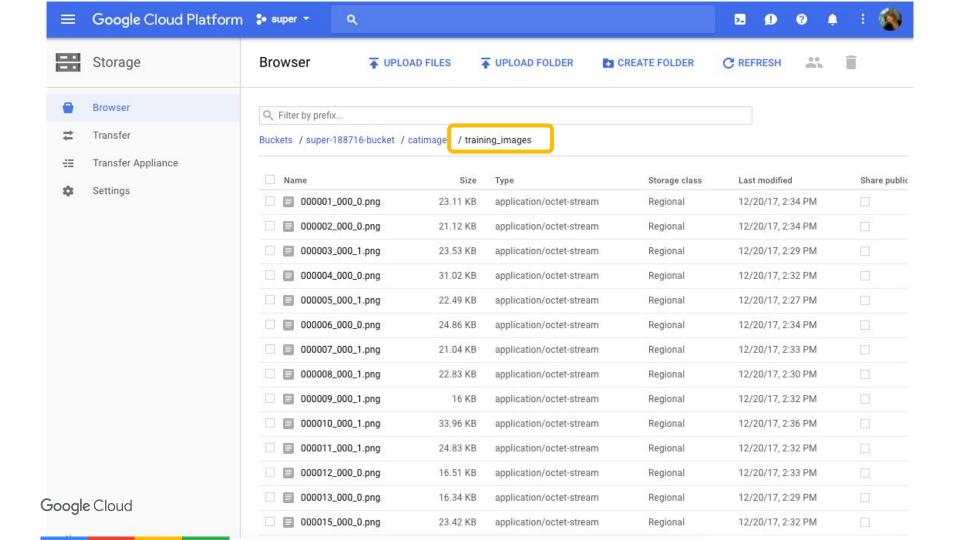




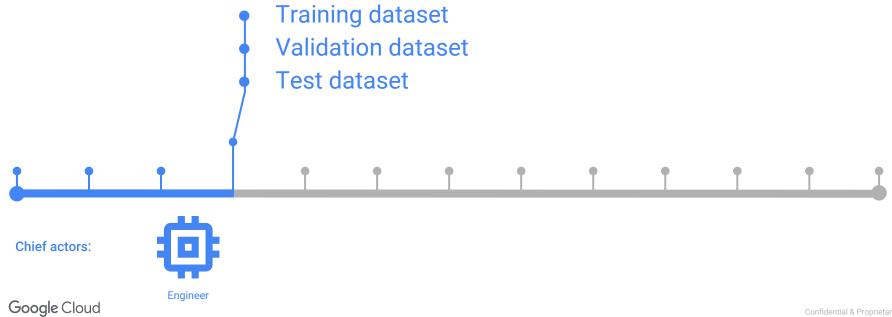








Step 3 is finished | You now have IDs of instances that go in:



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

