GATK Best Practices for Variant Discovery

Introduction to Pipelining Platforms

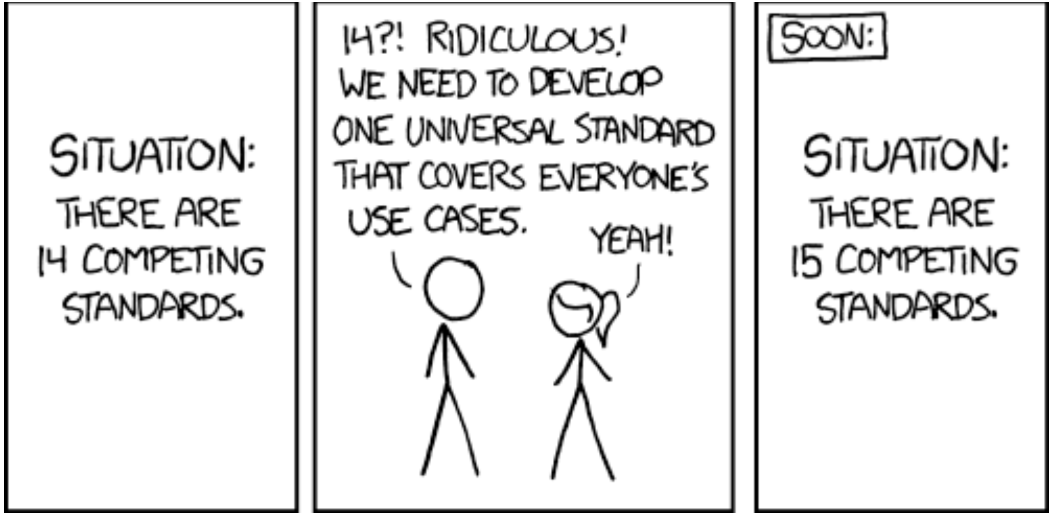
How we run workflows

BROAD INSTITUTE

http://software.broadinstitute.org/gatk/

gatk

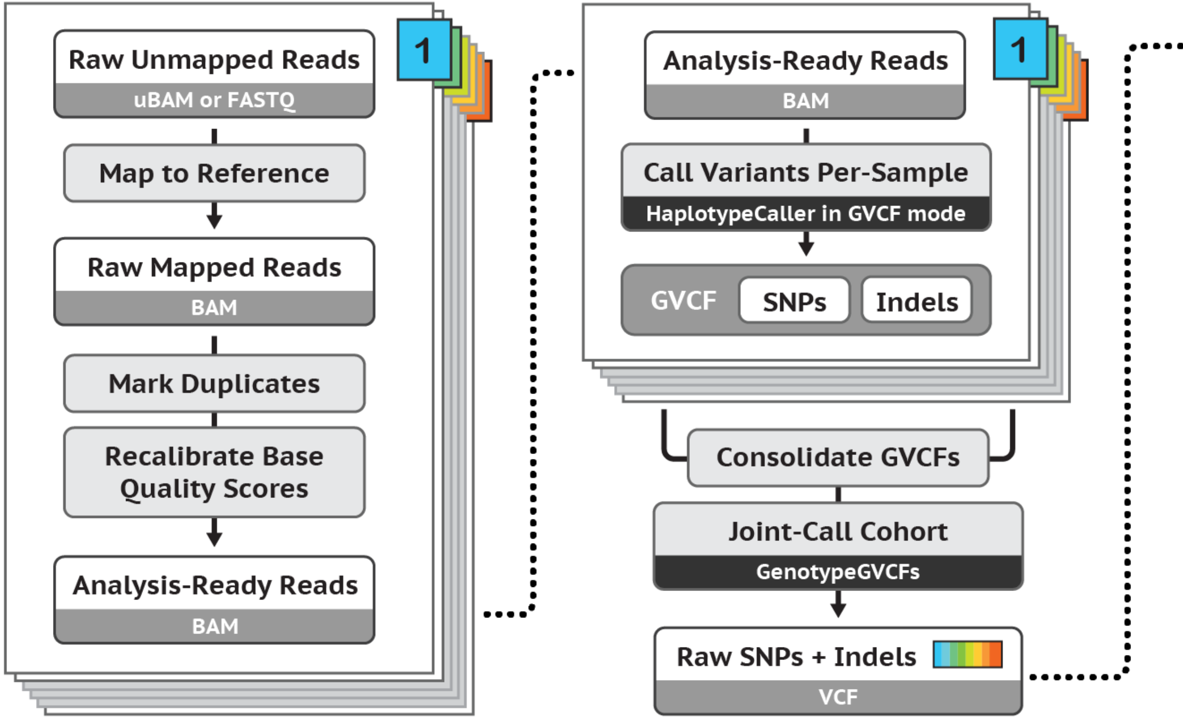
Pipelining solutions tend to proliferate



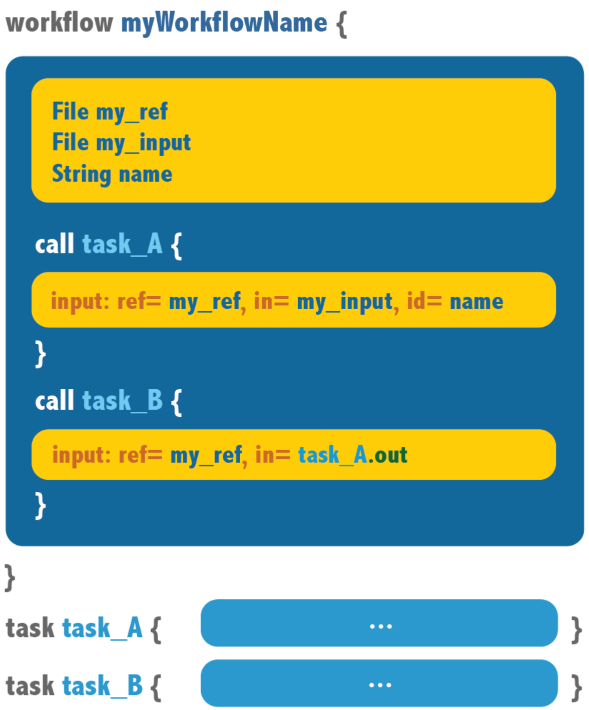
Randall Munroe, XKCD
https://www.xkcd.com/927/

Use scripted workflows to automate processing

- Automate repetitive tasks
- Increase auditability and reproducibility
- Reduce human error
- Reduce time spent re-implementing the wheel

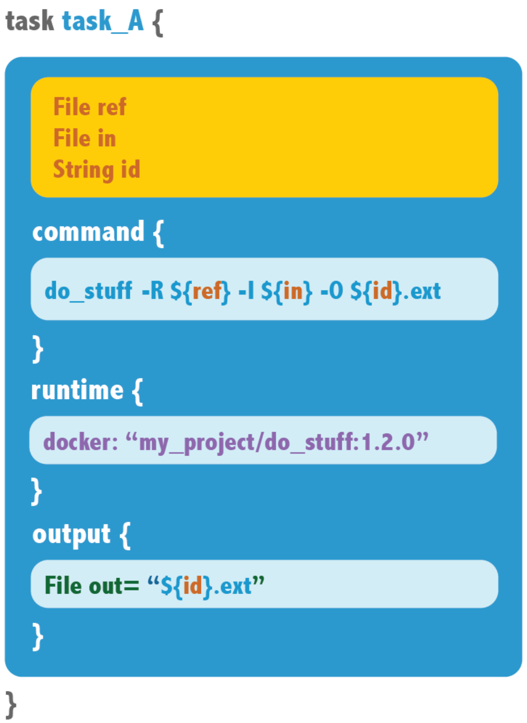


Use a workflow language that humans can understand



```
workflow myWorkflowName {
    File my_ref
    File my_input
    String name

    call task_A {
        input: ref= my_ref, in= my_input, id= name
    }
    call task_B {
        input: ref= my_ref, in= task_A.out
    }
}
task task_A {   ...   }
task task_B {   ...   }
```

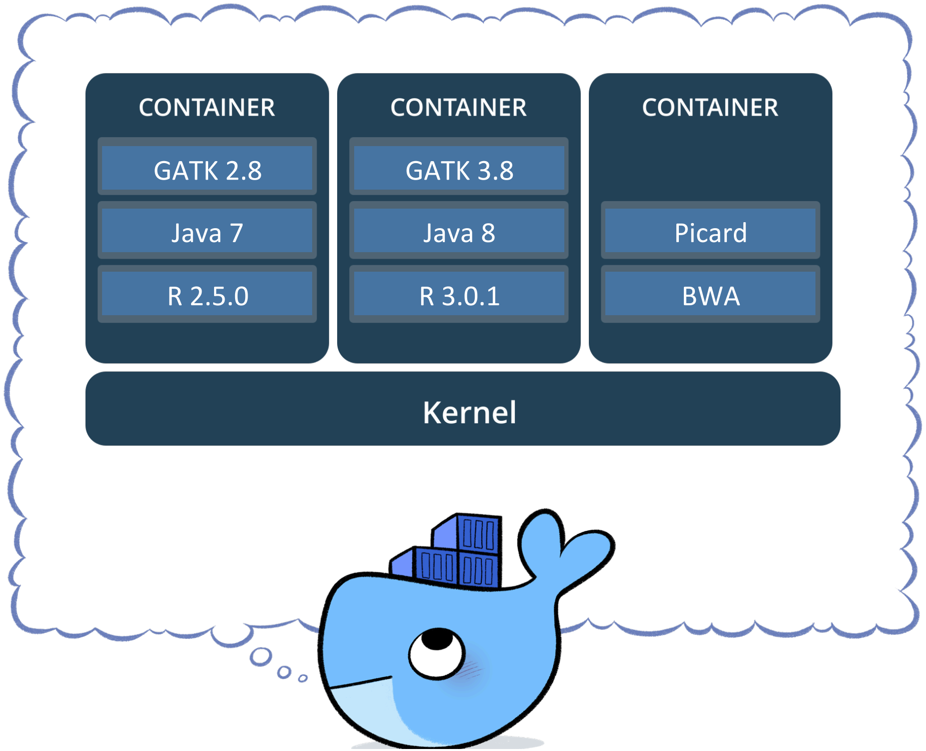


```
task task_A {
    File ref
    File in
    String id

    command {
        do_stuff -R ${ref} -I ${in} -O ${id}.ext
    }
    runtime {
        docker: "my_project/do_stuff:1.2.0"
    }
    output {
        File out= "${id}.ext"
    }
}
```

http://www.openwdl.org

Use containers for portability & reproducibility

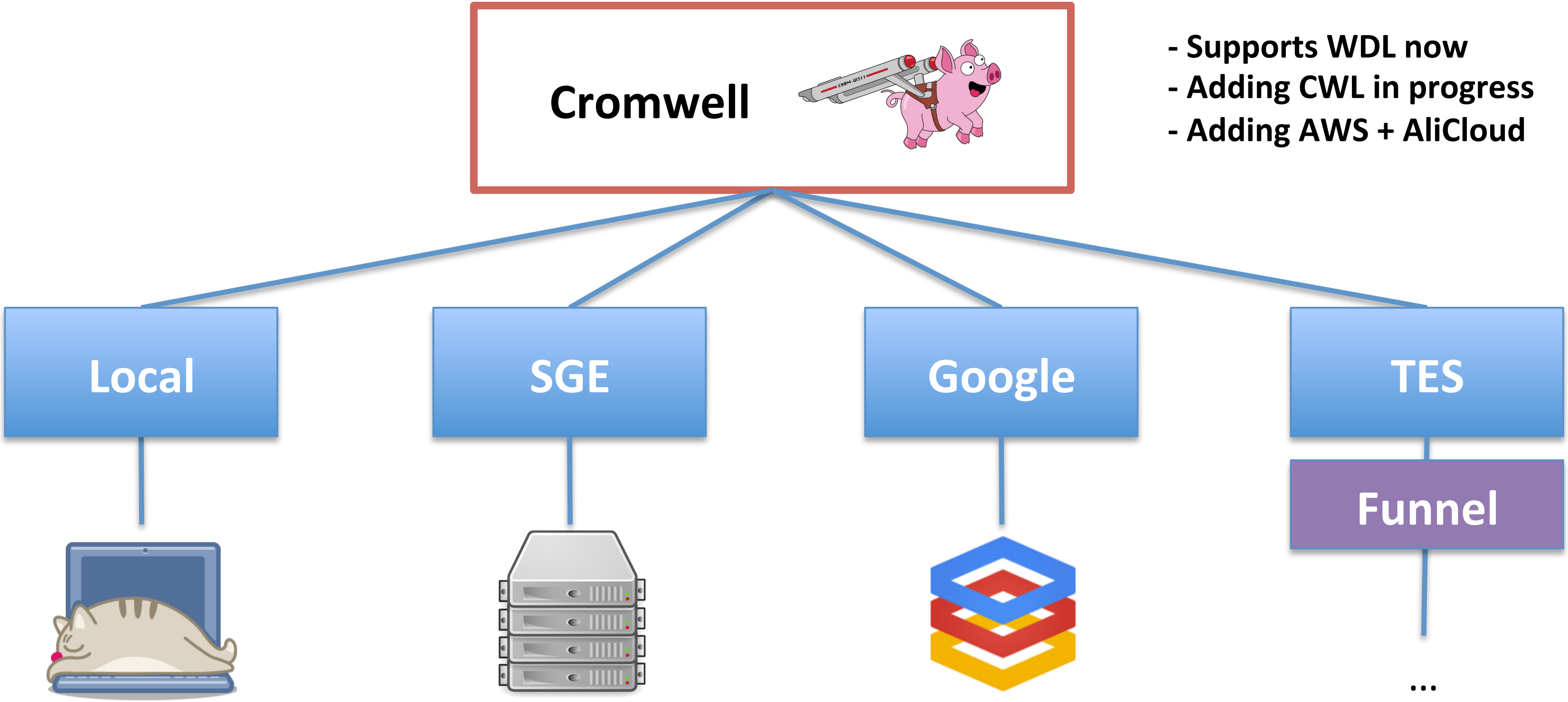


A container encapsulates **all the software dependencies** associated with running a program

Takes the guesswork out of running pipelines on different platforms!

Modified from https://www.docker.com/what-container

Use a workflow execution engine that runs anywhere

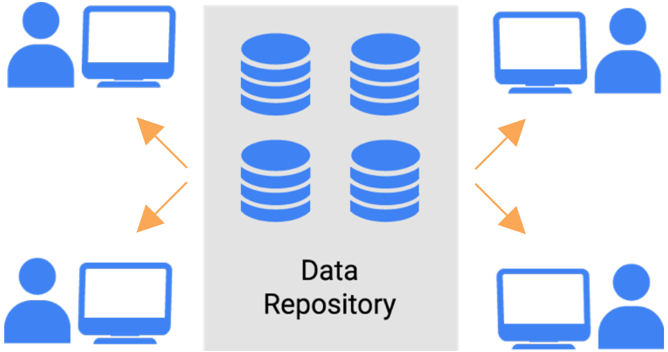


- Supports WDL now
- Adding CWL in progress
- Adding AWS + AliCloud

https://github.com/broadinstitute/cromwell

Use the Cloud to share methods & data

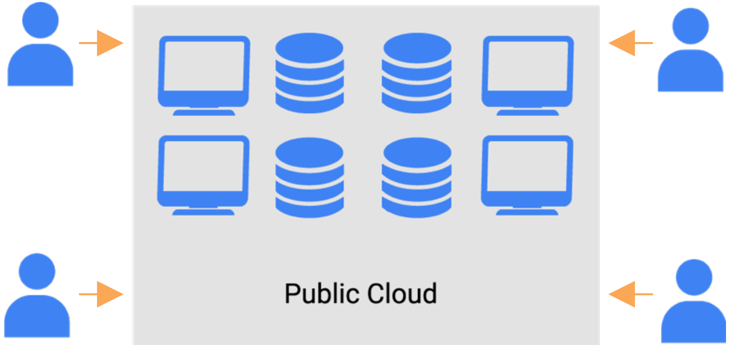
Traditional Way: Bring data to the researchers



Problems

Data sharing = data copying
Requires big infrastructure at each site
Largely fixed compute
Individual security implementations

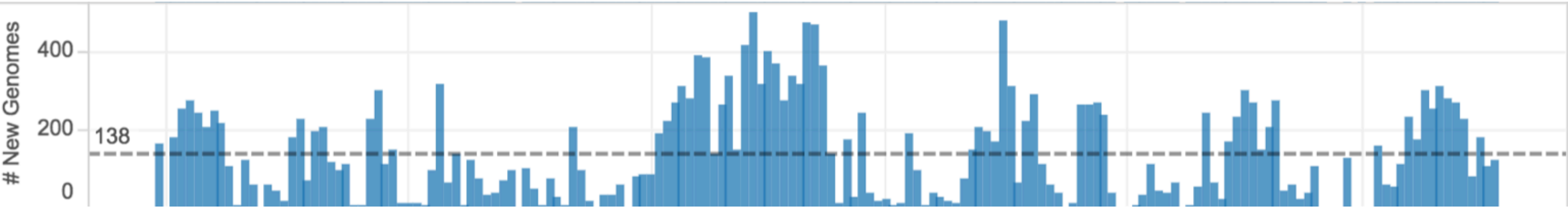
Cloud Way: Bring researchers to the data



Solutions

True data sharing
Cloud provides the infrastructure
Elastic compute and storage
Centralized security implementation

Clouds are elastic!



Genome processing requests per day over the last several months in the cloud

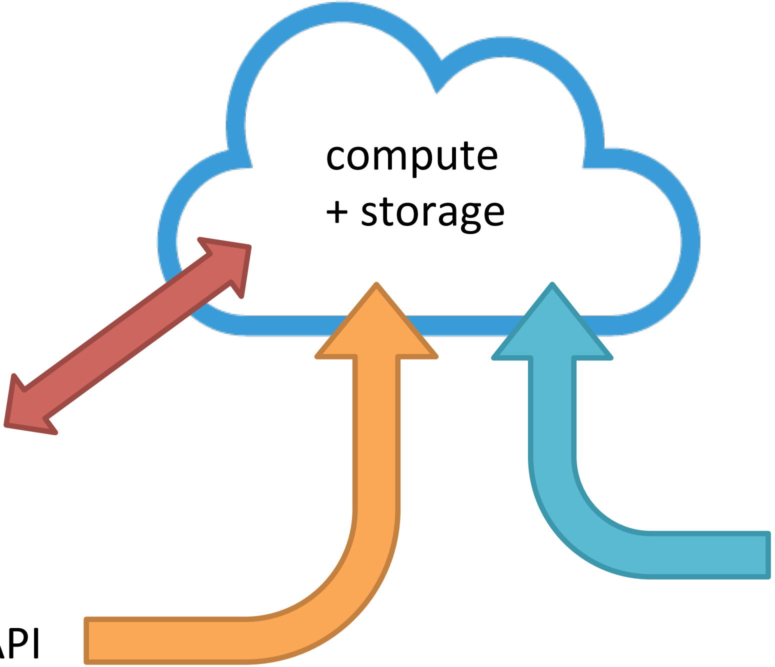
Genome processing requested by the Genomics Platform has been “spiky”

- We haven’t needed to pay for compute power when we weren’t using it
- We can easily tolerate the spikes without being forced to maintain a backlog of “things to process once everything calms down”

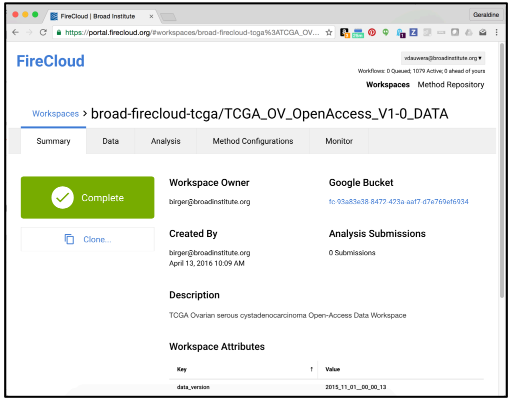
Several paths to running GATK pipelines on Cloud

COMMAND-LINE

- **Directly from GATK4 with Spark**
- **Pipelining service**
  Google Genomics API



WEB GUI / FRONT END



https://portal.firecloud.org/#workspaces/broad-firecloud-tcga%3ATCGA_OV...

vdauwera@broadinstitute.org ▾

Workflows: 0 Queued; 1079 Active; 0 ahead of yours

Workspaces › broad-firecloud-tcga/TCGA_OV_OpenAccess_V1-0_DATA

Workspace Owner
birger@broadinstitute.org

Google Bucket
fc-93a83e38-8472-423a-aaf7-d7e769ef6934

Created By
birger@broadinstitute.org
April 13, 2016 10:09 AM

Analysis Submissions
0 Submissions

Description
TCGA Ovarian serous cystadenocarcinoma Open-Access Data Workspace

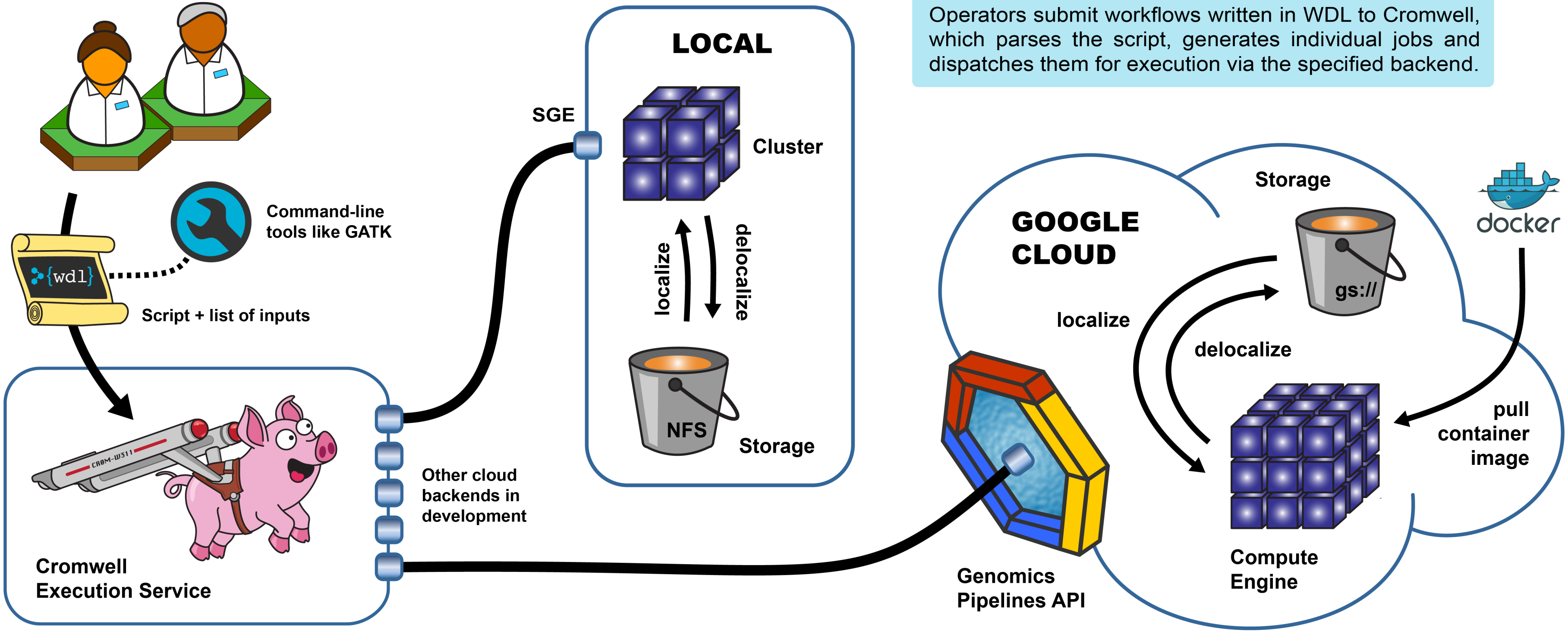
Workspace Attributes

Key ↑	Value
data_version	2015_11_01__00_00_13

- **FireCloud**
  Portal to Workbench
  http://www.firecloud.org

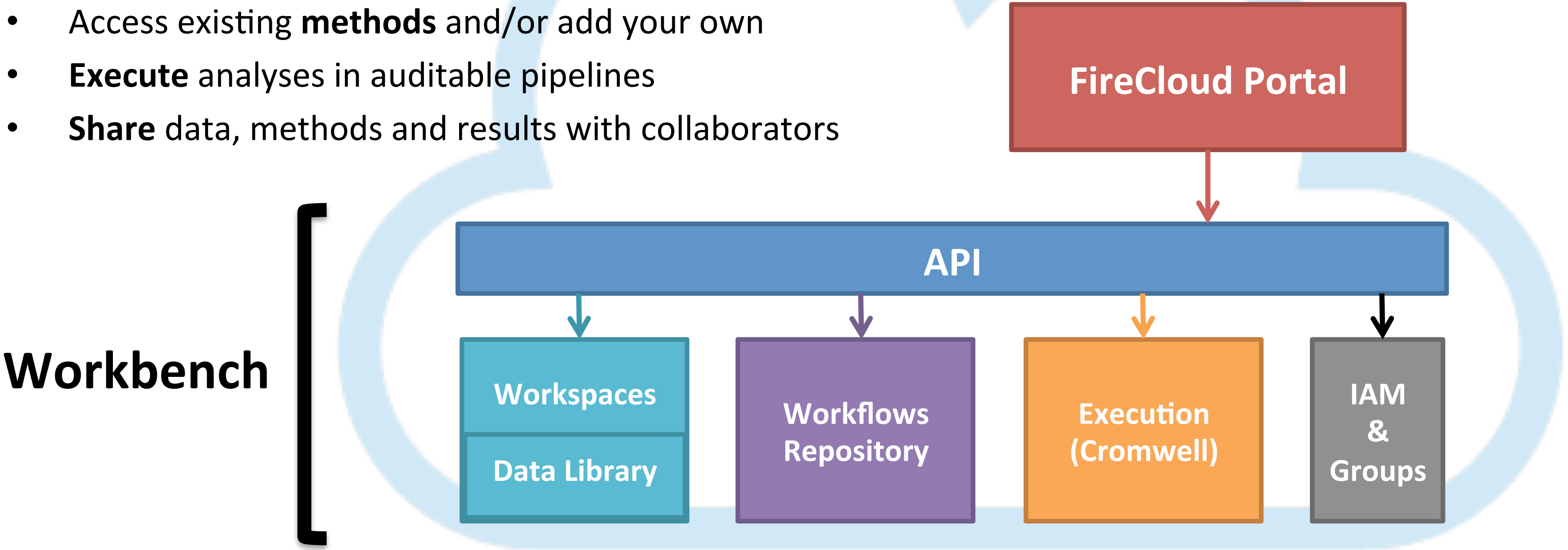
Our setup: some local development / all production on Cloud

Operators submit workflows written in WDL to Cromwell, which parses the script, generates individual jobs and dispatches them for execution via the specified backend.

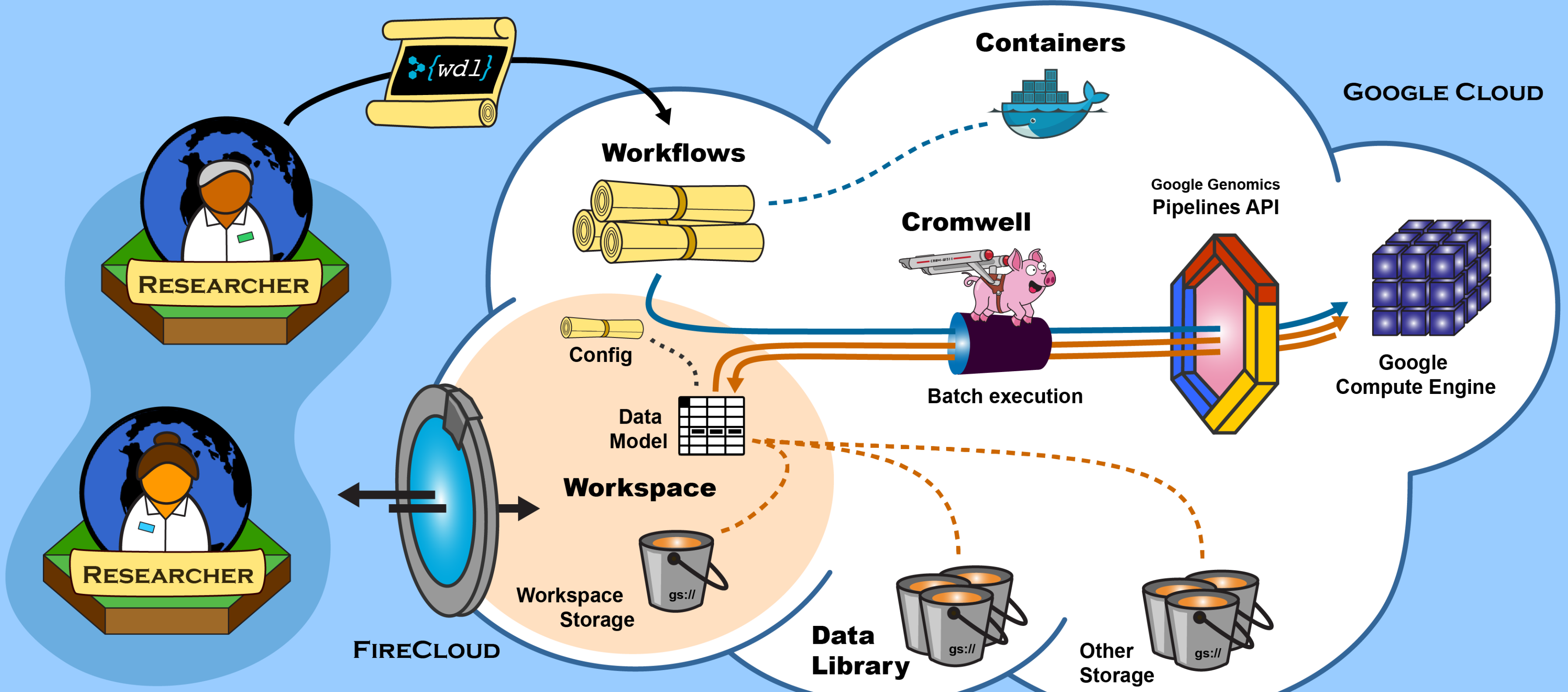


Self-service analysis for all: Workbench + FireCloud

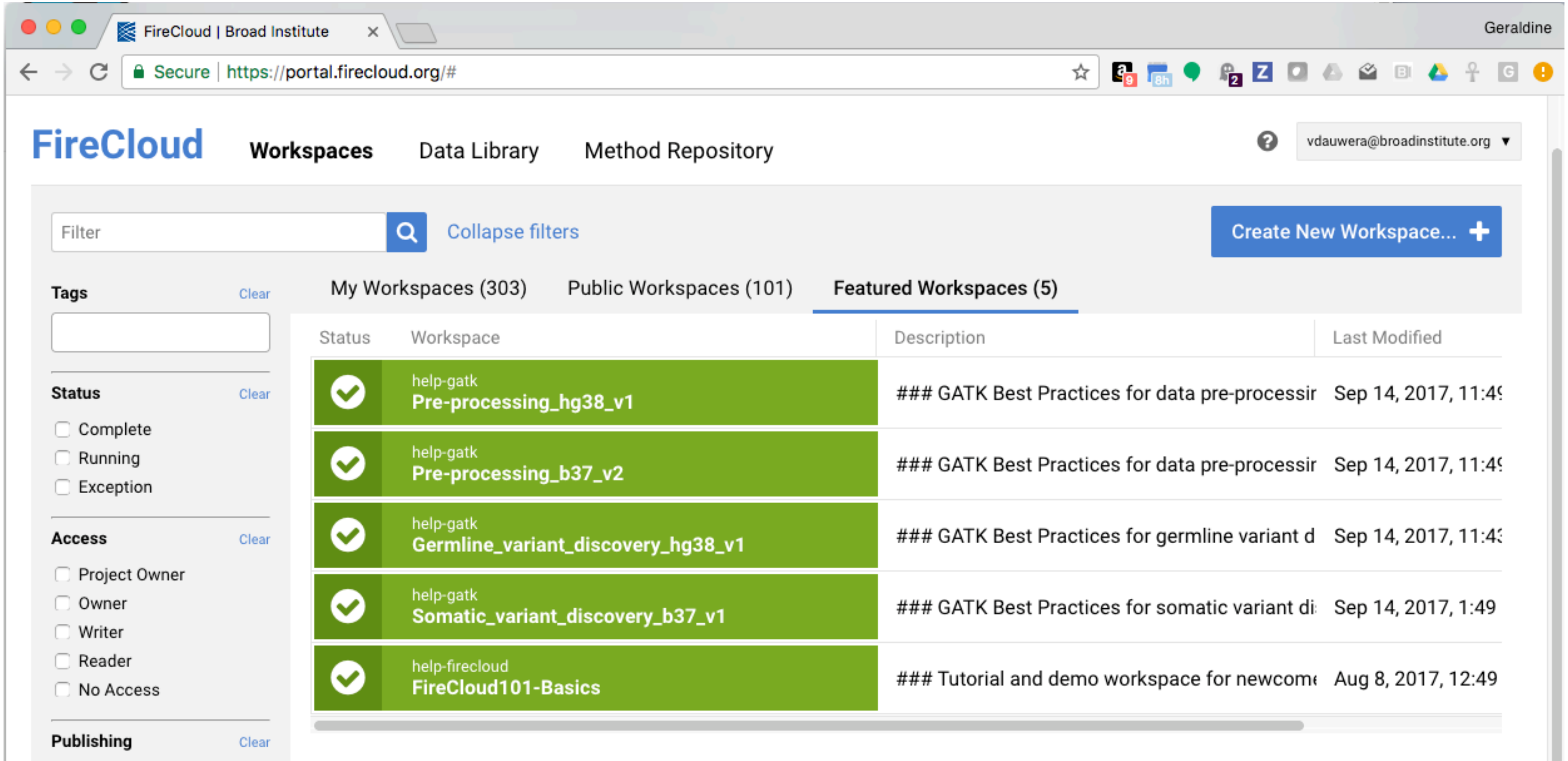
- Collaborative **cloud-based** analysis platform built on top of **Google Cloud Platform**
- **Free to access** / compute & storage charged by Google
- Access published **data** and/or add your own
- Access existing **methods** and/or add your own
- **Execute** analyses in auditable pipelines
- **Share** data, methods and results with collaborators



FireCloud puts GATK4 workflows in everyone's hands

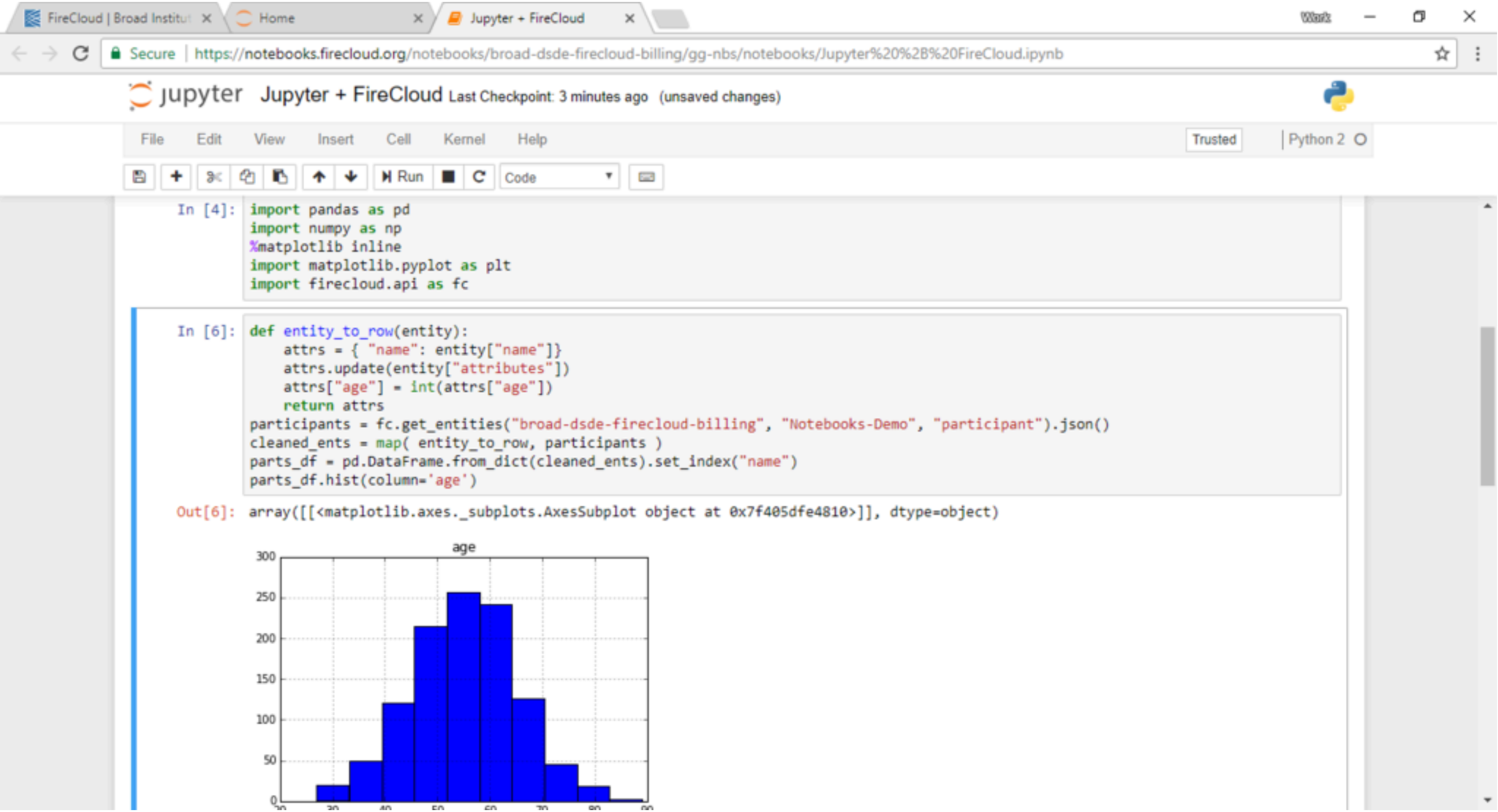


GATK4 workflows preloaded in FireCloud workspaces



Status	Workspace	Description	Last Modified
✓	help-gatk **Pre-processing_hg38_v1**	### GATK Best Practices for data pre-processir	Sep 14, 2017, 11:4
✓	help-gatk **Pre-processing_b37_v2**	### GATK Best Practices for data pre-processir	Sep 14, 2017, 11:4
✓	help-gatk **Germline_variant_discovery_hg38_v1**	### GATK Best Practices for germline variant d	Sep 14, 2017, 11:4
✓	help-gatk **Somatic_variant_discovery_b37_v1**	### GATK Best Practices for somatic variant di	Sep 14, 2017, 1:49
✓	help-firecloud **FireCloud101-Basics**	### Tutorial and demo workspace for newcom	Aug 8, 2017, 12:49

Coming next: Interactive analysis with Notebooks

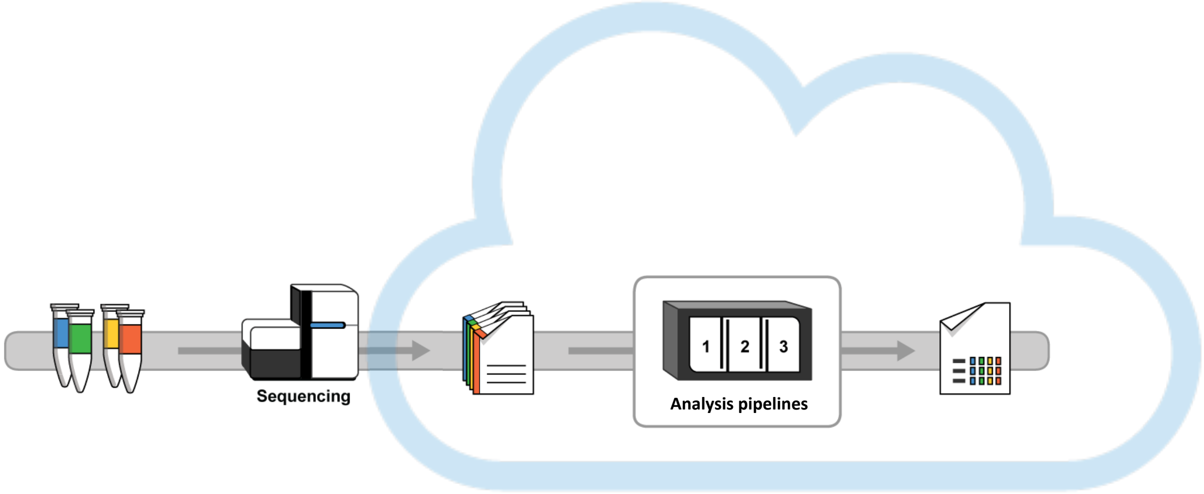


FireCloud | Broad Institu
Home
Jupyter + FireCloud
Work
Secure
https://notebooks.firecloud.org/notebooks/broad-dsde-firecloud-billing/gg-nbs/notebooks/Jupyter%20%2B%20FireCloud.ipynb
jupyter
Jupyter + FireCloud Last Checkpoint: 3 minutes ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Help
Trusted
Python 2
Run
Code
age
300
250
200
150
100
50
0
20
30
40
50
60
70
80
90

```
In [4]: import pandas as pd
        import numpy as np
        %matplotlib inline
        import matplotlib.pyplot as plt
        import firecloud.api as fc
```

```
In [6]: def entity_to_row(entity):
            attrs = { "name": entity["name"]}
            attrs.update(entity["attributes"])
            attrs["age"] = int(attrs["age"])
            return attrs
        participants = fc.get_entities("broad-dsde-firecloud-billing", "Notebooks-Demo", "participant").json()
        cleaned_ents = map( entity_to_row, participants )
        parts_df = pd.DataFrame.from_dict(cleaned_ents).set_index("name")
        parts_df.hist(column='age')
```

```
Out[6]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f405dfe4810>]], dtype=object)
```



SEE YOU ON THE CLOUD!