



# GATK Best Practices for Variant Discovery

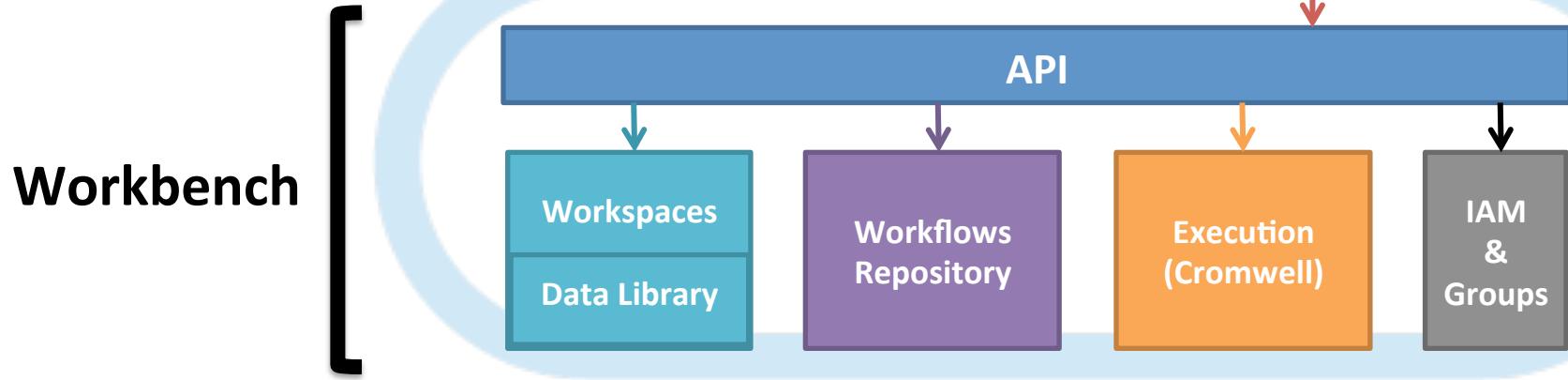


## Pipelining with FireCloud

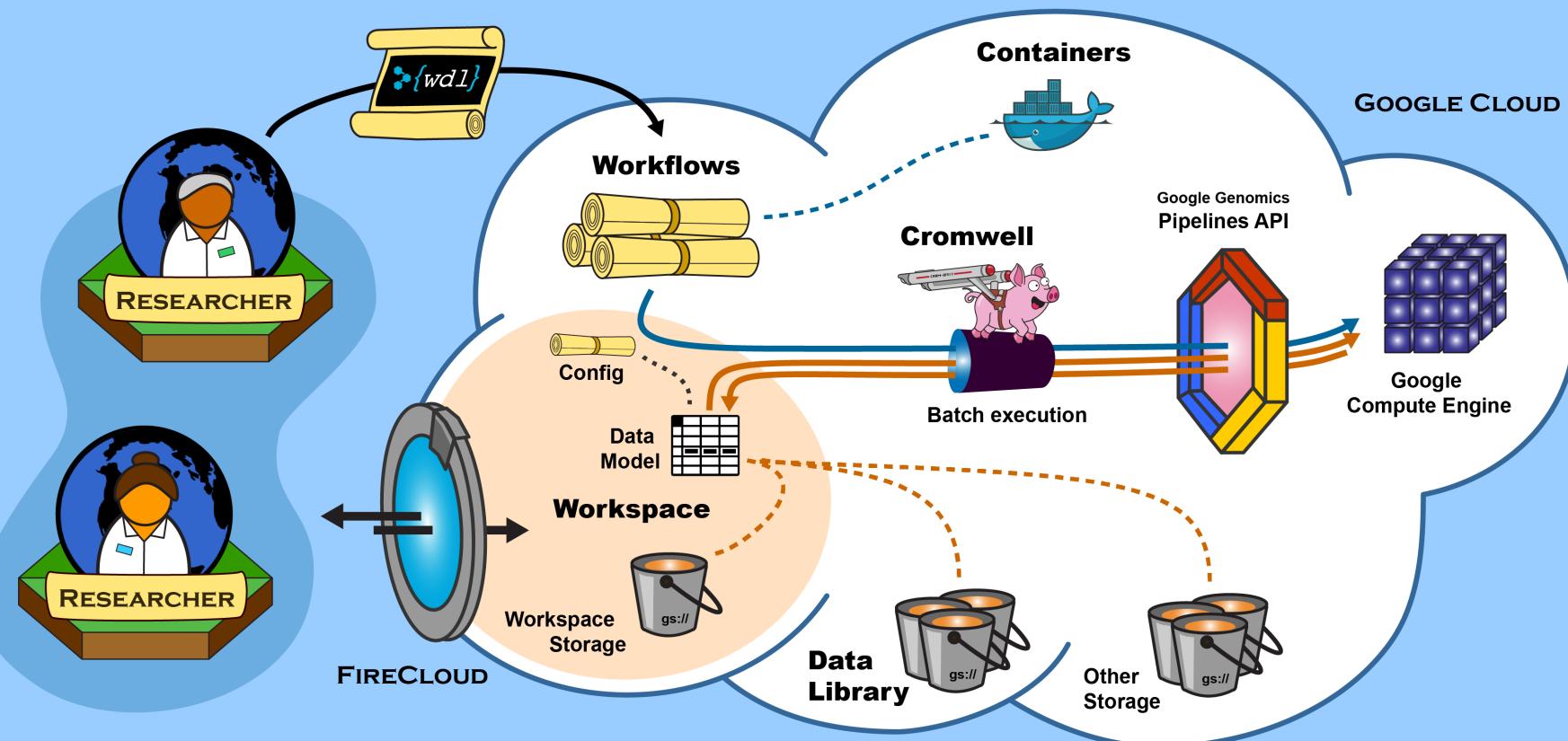
### Overview

# FireCloud: Share data and methods + Execute pipelines

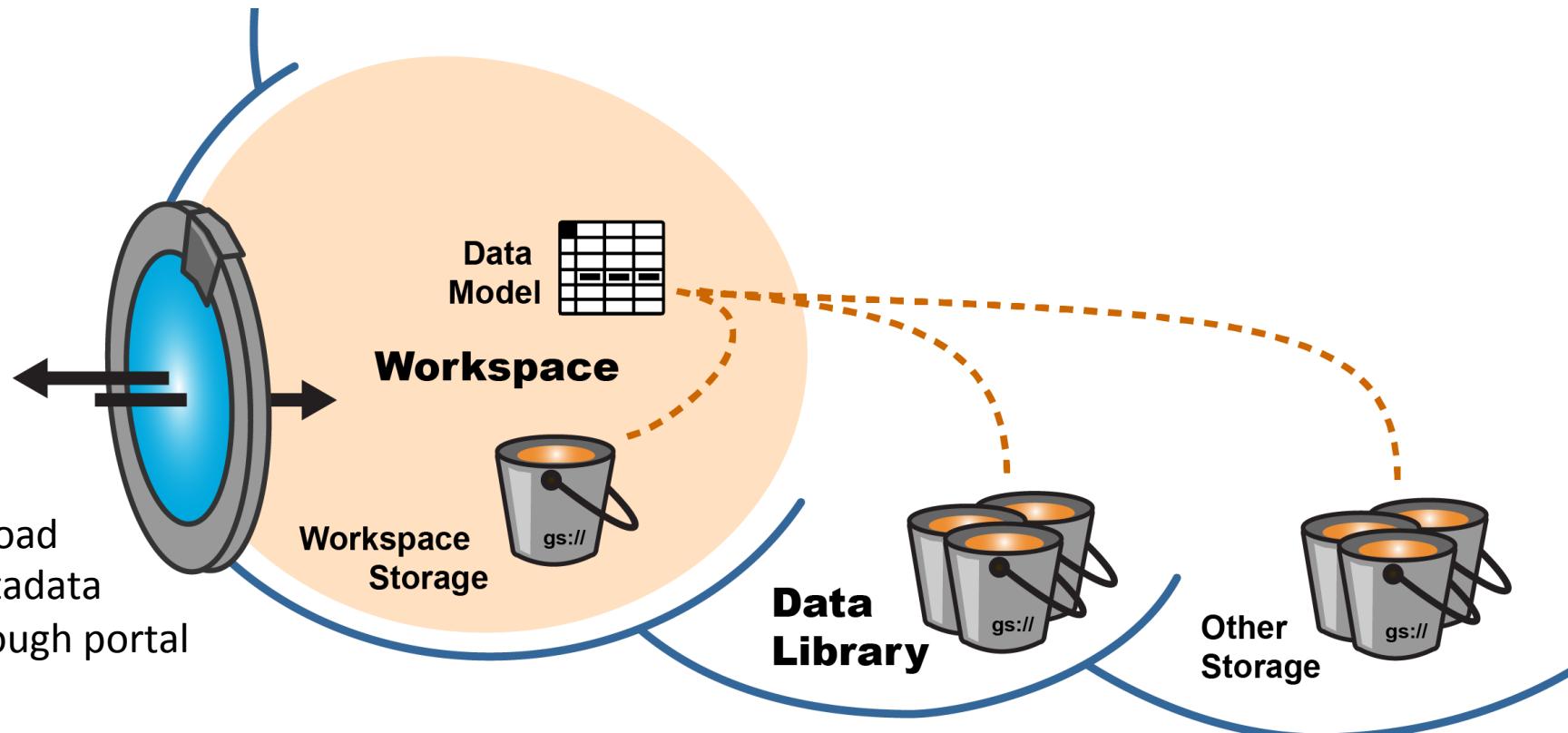
- Collaborative **cloud-based** analysis platform built on top of **Google Cloud Platform**
- **Free to access** / compute & storage charged by Google
- Access published **data** and/or add your own
- Access existing **methods** and/or add your own
- **Execute** analyses in auditable pipelines
- **Share**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



# Running pipelines in FireCloud

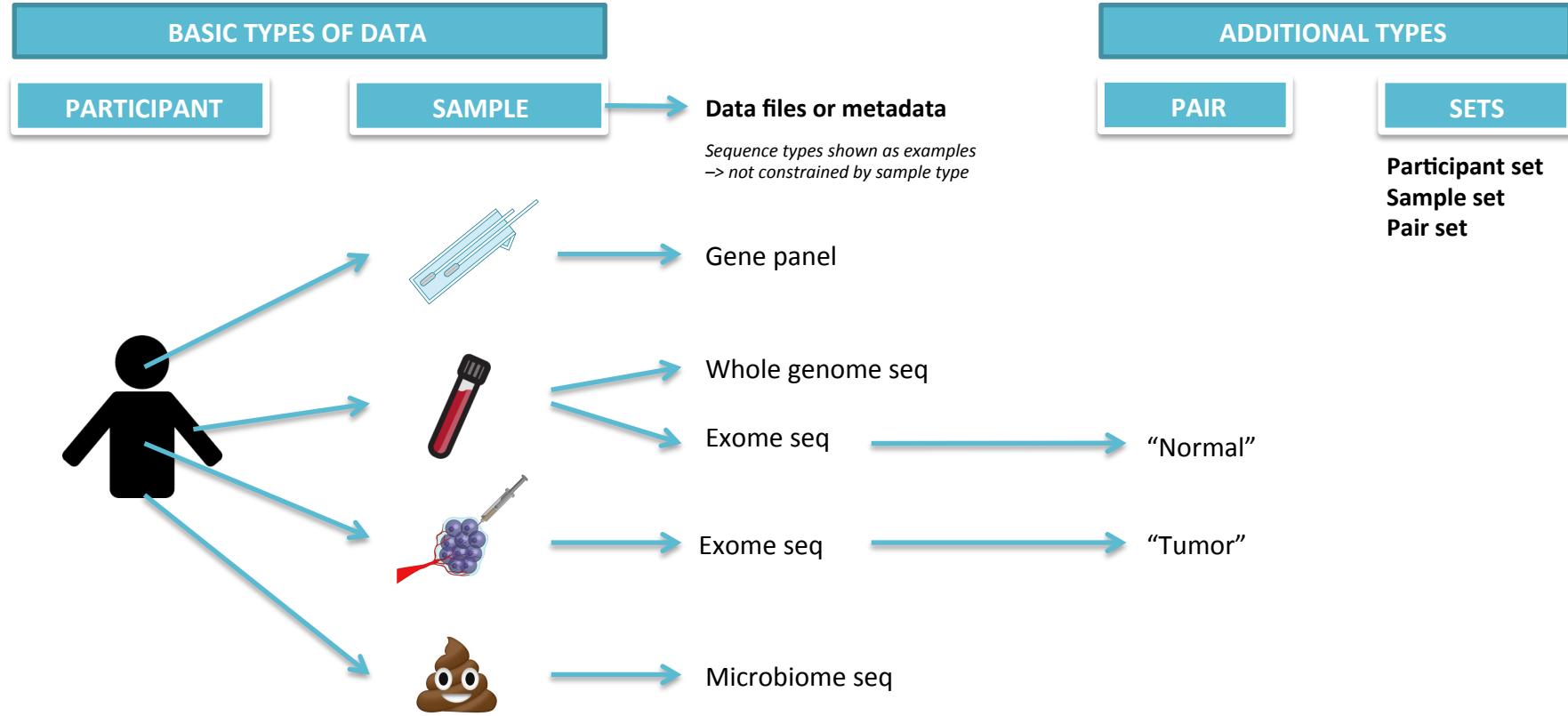


# DATA is stored in “buckets” on Google Cloud



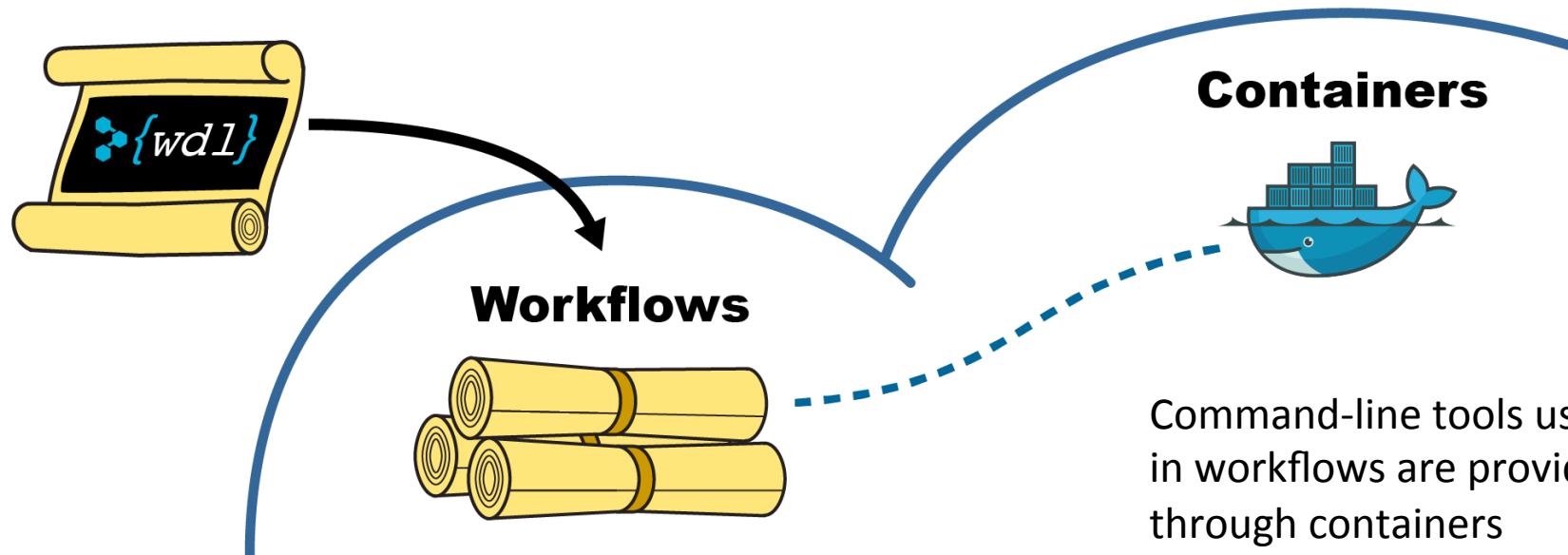
*Upload data itself to buckets through Google console or gsutil*

# Data Model = Where data is described / structured



# Workflows (=methods) are stored in a Method Repository

WDL = Workflow Definition Language



Command-line tools used  
in workflows are provided  
through containers

# WDL is designed to be easy to read and write

```
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"
```

```
        }
```

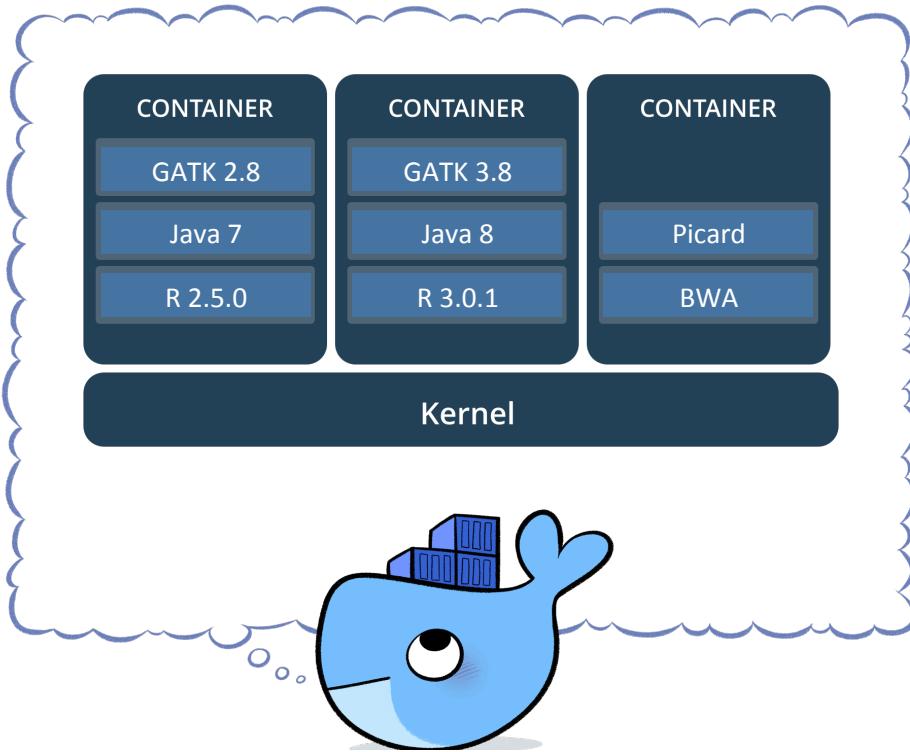
```
        output {
```

```
            File out= "${id}.ext"
```

```
        }
```

```
}
```

# Containerized tools ensure portability and reproducibility

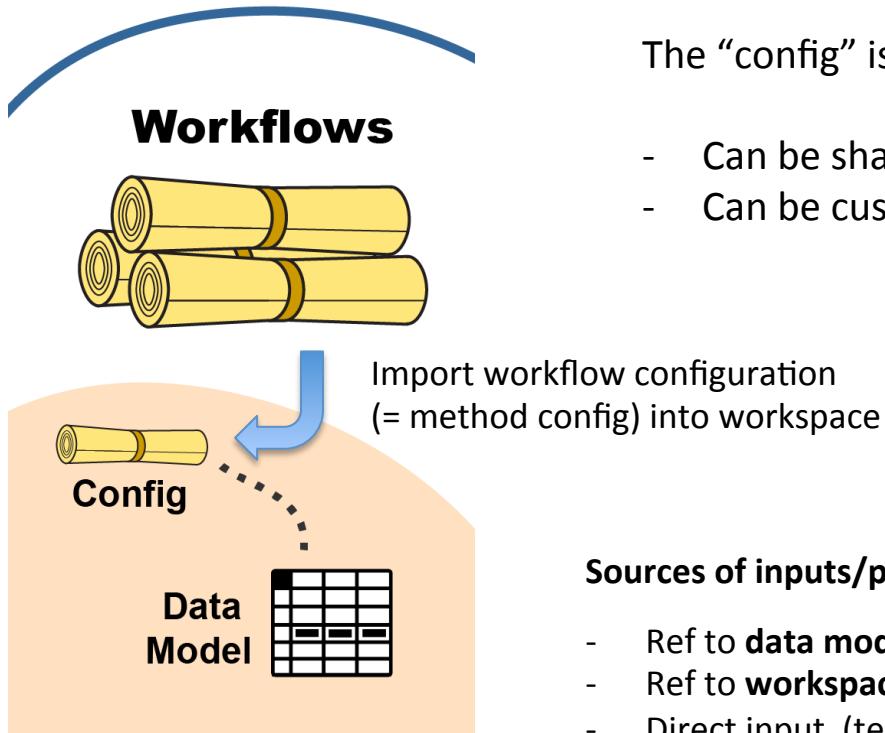


A container encapsulates  
**all the software dependencies**  
associated with running a program

Takes the guesswork out of running  
pipelines on different platforms!

➤ Portability & reproducibility !

# Configuration = specify inputs and parameters



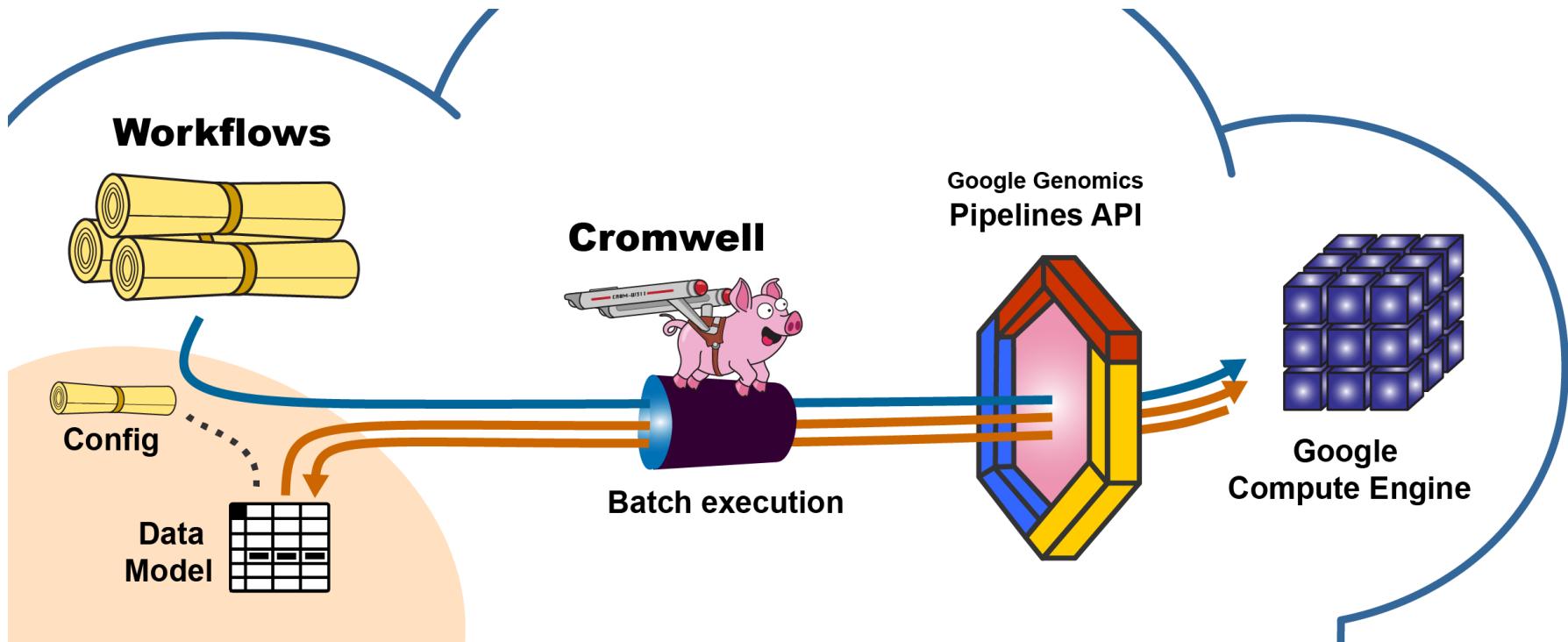
The “config” is equivalent to the WDL’s inputs JSON file

- Can be shared in the method repository
- Can be customized within the workspace

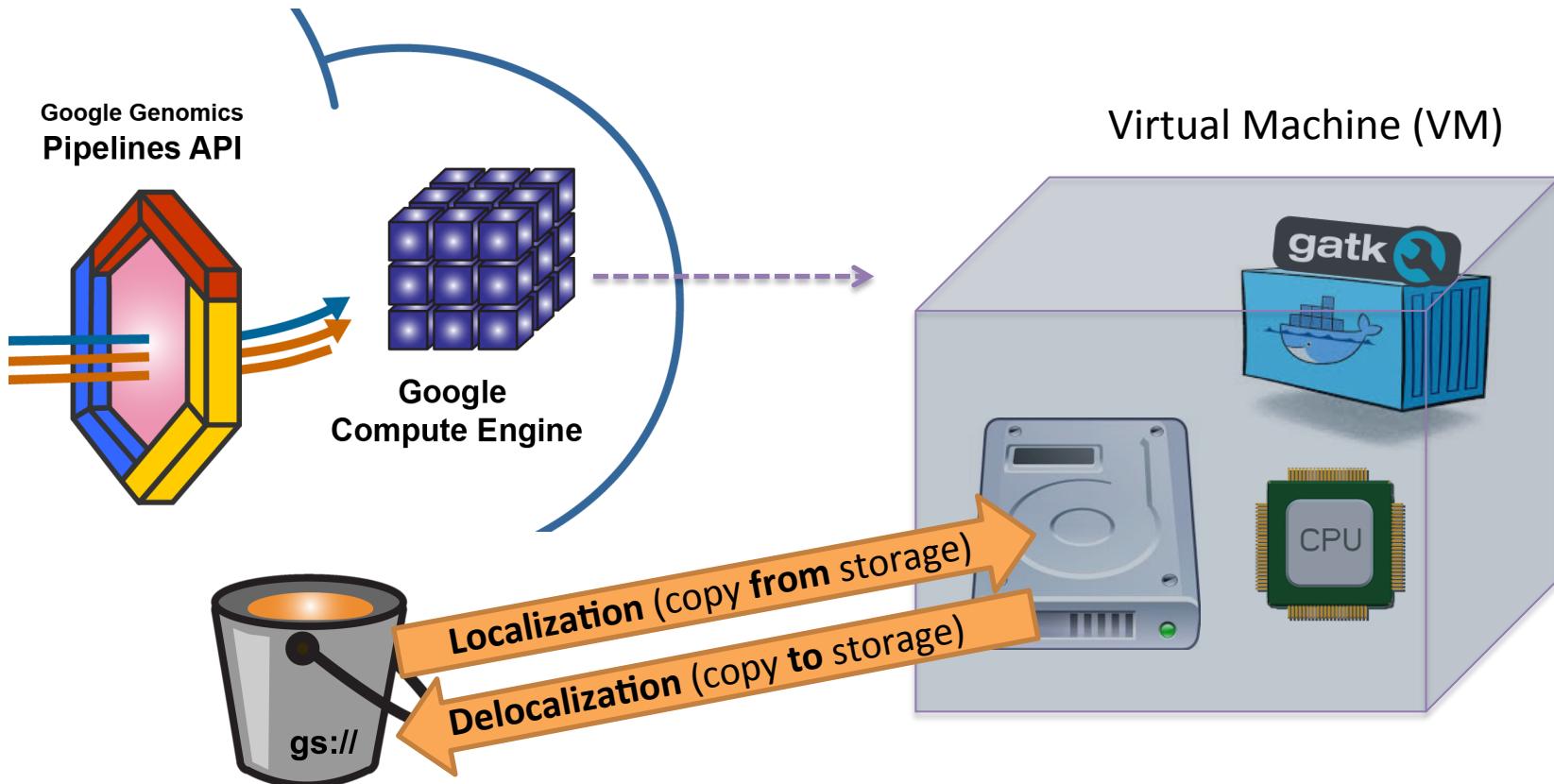
## Sources of inputs/parameters:

- Ref to **data model** (incl. paths to data files)
- Ref to **workspace attributes** (=global variables)
- Direct input (text, numbers, filepaths etc)

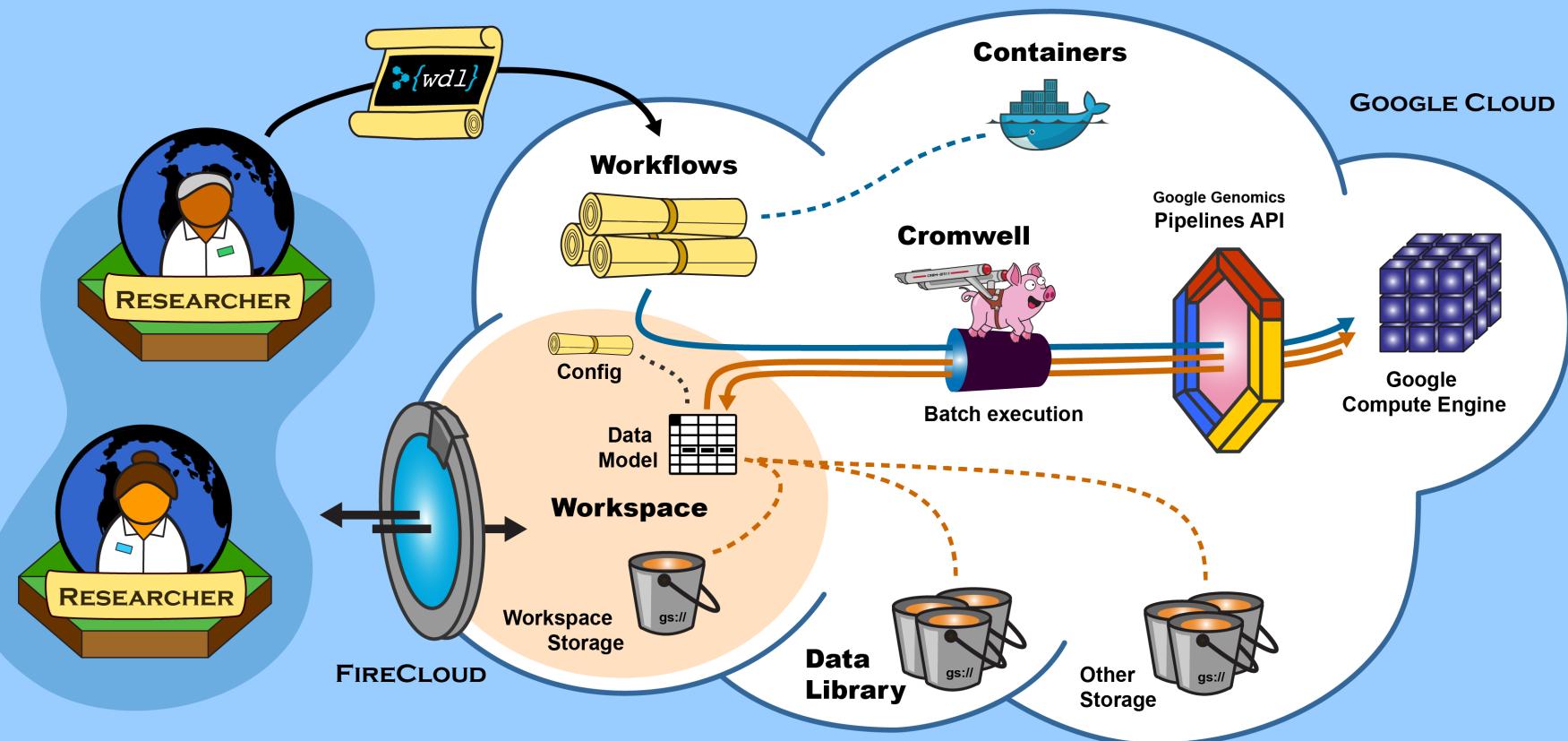
# Workflow + config are sent to Cromwell for execution



# Each workflow task is executed on its own VM



# Running pipelines in FireCloud



# Also: Data Library with data use-aware search

FireCloud Workspaces Data Library Methods Development Tags

vdauwera@broadinstitute.org Columns

Search

Filter by Research Purpose Clear

Tags Clear

Cohort Phenotype/Indication Clear

NA 6

Acute Myeloid Leukemia 4

Data Use Limitation Clear

General Research Use 136

NA 6

Cancer research only 1

HMB-IRB-NPU-MDS 1

Health and Biomedical Rese... 1

Mental Health Research; No... 1

Pediatric Research Only 1

less...

Filter by Research Purpose  
Powered by DUOS

The datasets will be used for the following purposes:

- Disease focused research
- Methods development/Validation study METHODS
- Control set CONTROL
- Aggregate analysis to understand variation in the general population AGGREGATES
- Study population origins or ancestry POA
- Commercial purpose/by a commercial entity COMMERCIAL

Cancel Search

| Dataset                         | Description                  | Count |
|---------------------------------|------------------------------|-------|
| TCGA_BLCA_hg38_ControlledAccess | Bladder Urothelial Carcin... | 412   |
| TCGA_BLCA_hg38_OpenAccess       | Bladder Urothelial Carcin... | 412   |
| TCGA_BLCA_OpenAccess            | Bladder Urothelial Carcin... | 412   |
| TCGA_BRCA_ControlledAccess      | Breast Invasive Carcino...   | 1098  |
| TCGA_BRCA_hg38_ControlledAccess | Breast Invasive Carcino...   | 1098  |
| TCGA_BRCA_hg38_OpenAccess       | Breast Invasive Carcino...   | 1098  |
| TCGA_BRCA_OpenAccess            | Breast Invasive Carcino...   | 1098  |

# Coming soon: interactive analysis with Notebooks

**FireCloud** Workspaces Data Library Method Repository

ansinghfirecloud@gmail.com ▾

WORKSPACE  
anu-bills/clusters-workspace

Summary Data Analysis Notebooks **BETA** Method Configurations

**Spark Clusters**  
Launch an interactive analysis environment based on Jupyter notebooks, Spark, and Hail. This beta feature is under active development.

Columns Filter

| Name       | Status     | Workers | Create Date            | Labels              |
|------------|------------|---------|------------------------|---------------------|
| mycluster1 | ✓ Running  | 0       | March 2, 2018, 5:19 PM | creator   anu-bills |
| mycluster2 | ✓ Running  | 0       | March 2, 2018, 5:20 PM | creator   anu-bills |
| mycluster3 | ⌚ Deleting | 2       | March 2, 2018, 5:22 PM | creator   anu-bills |
| mycluster4 | ⌚ Creating | 0       | March 2, 2018, 5:25 PM | creator   anu-bills |

1 - 4 of 4 results < Prev 1 Next >

**BROAD INSTITUTE**  
© 2015-2018 Broad Institute | Privacy Policy | Terms of Service | User Guide | FireCloud Forum

jupyter demo Last Checkpoint: a minute ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 2 O

### Leonardemo (demonardo?)

Hey, look at this notebook running Python!

```
In [1]: print "Hello, Jupyter!"  
print l+1  
print " ".join(map(str.upper, ["i", "said", "hello", "jupyter"]))
```

Hello, Jupyter!  
2  
I SAID HELLO, JUPYTER

### Connecting to FireCloud

Let's import the FireCloud Python API. Shout out to the folks in CGA who maintain it!

```
In [2]: import firecloud.api as fc
```

Is FC even up? This will be a short demo if it isn't:

```
In [3]: health = fc.health()  
print health.status_code, health.text
```

200 OK

Let's list the names and namespaces of the first ten workspaces we can see:

<https://portal.firecloud.org>

## GUIDED TOUR

