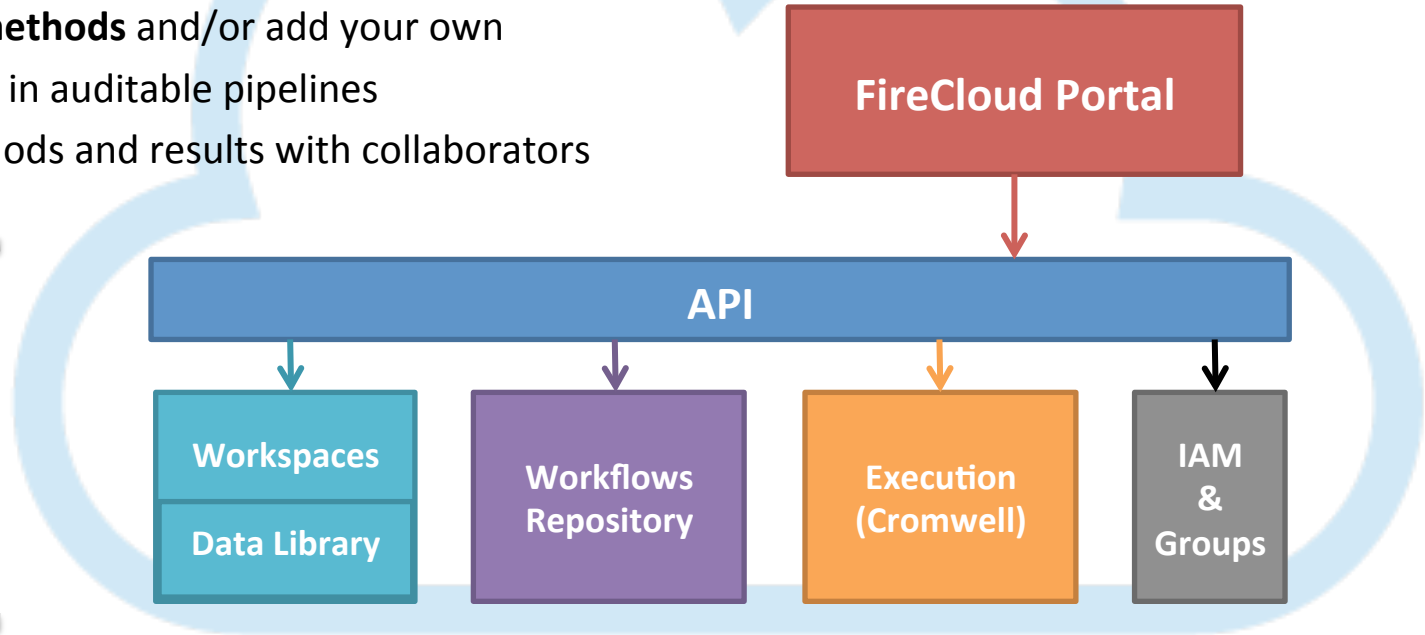
Pipelining with FireCloud

Overview

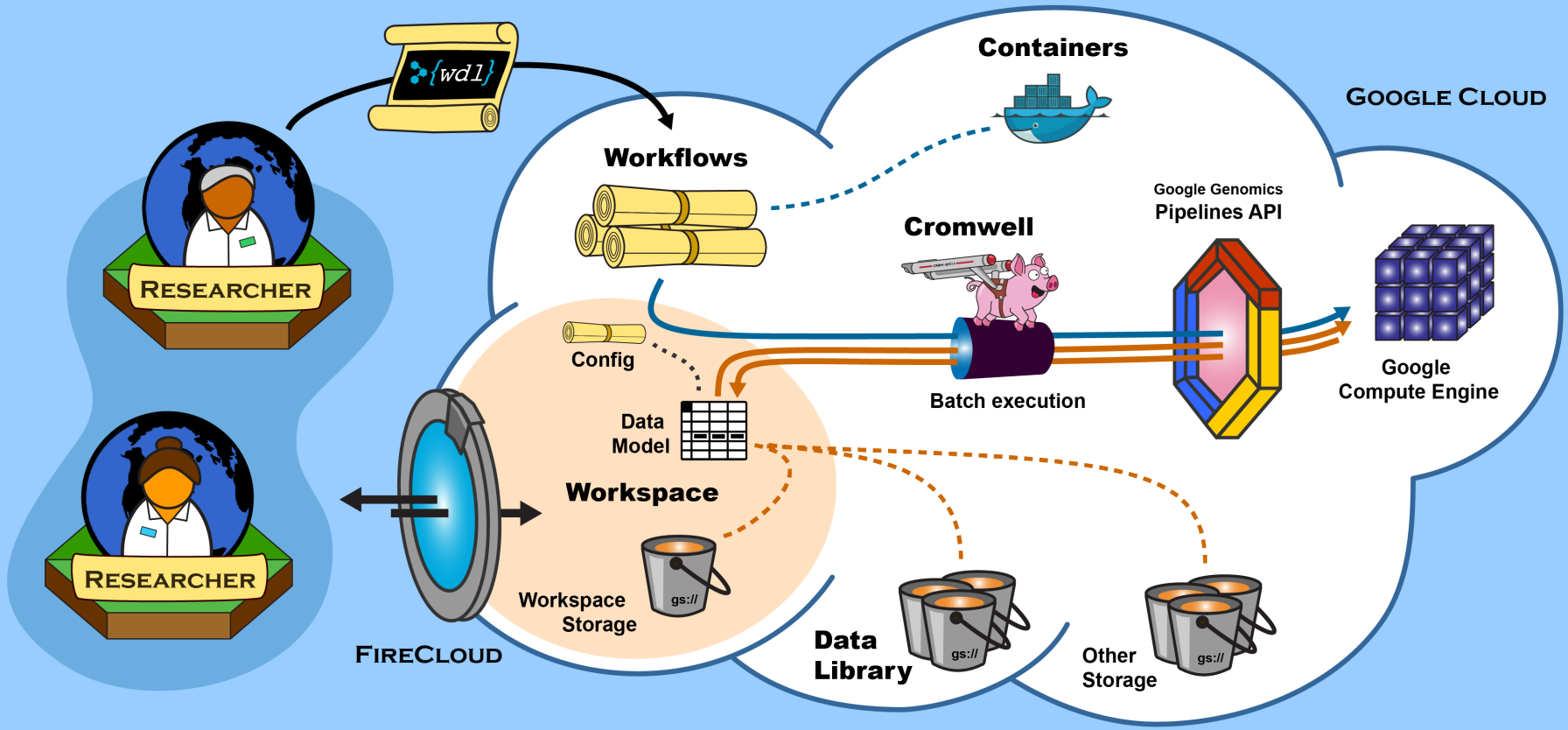
FireCloud: Share data and methods + Execute pipelines

- 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

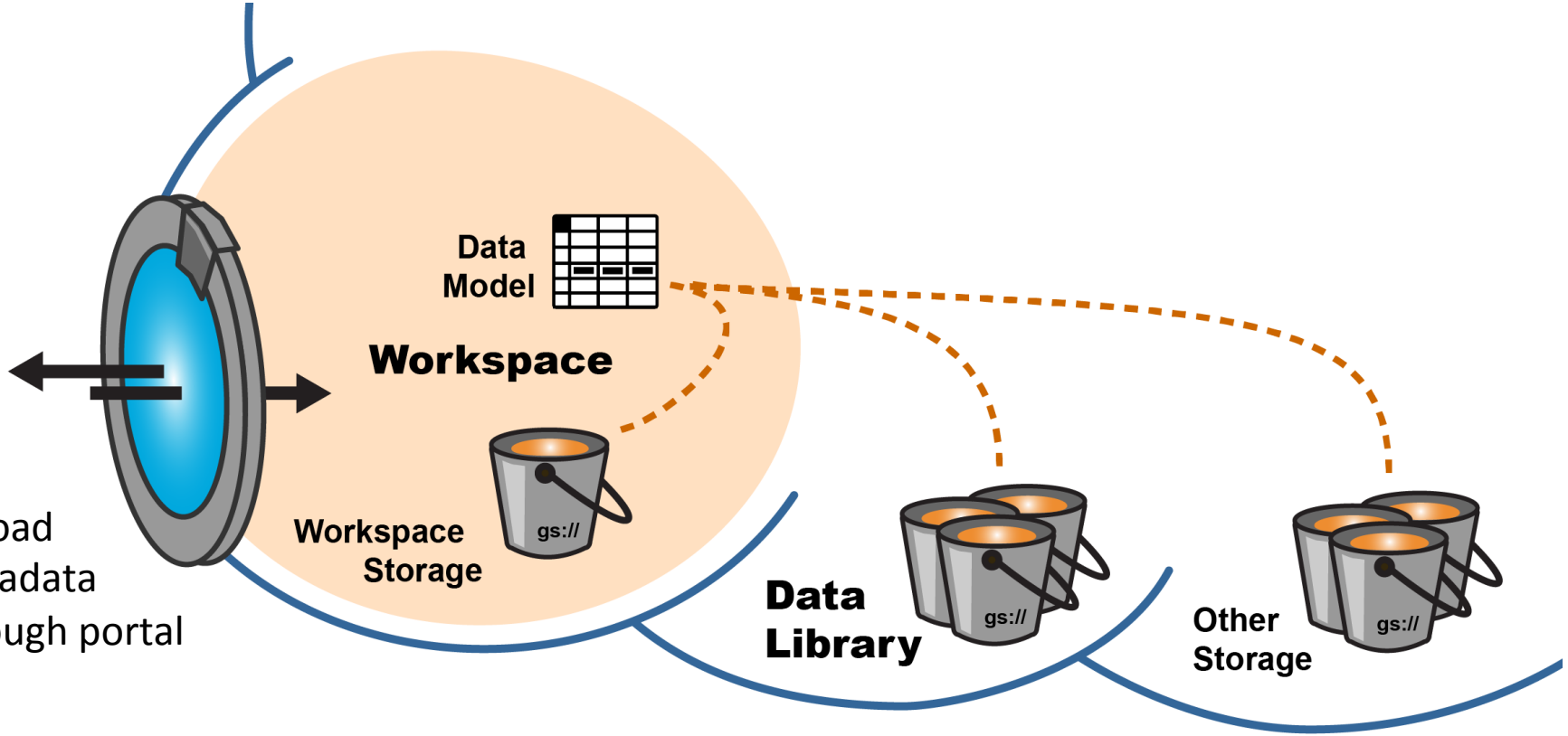
Workbench



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

BASIC TYPES OF DATA

PARTICIPANT

SAMPLE → **Data files or metadata**

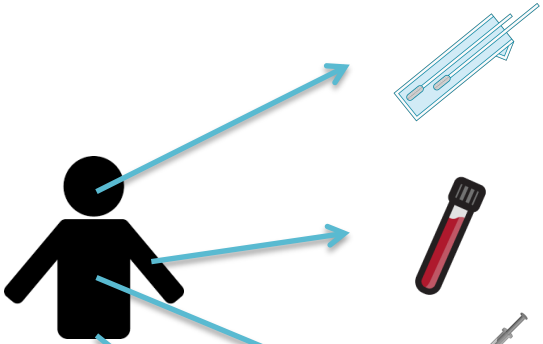
Sequence types shown as examples
–> not constrained by sample type

ADDITIONAL TYPES

PAIR

SETS

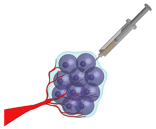
Participant set
Sample set
Pair set



→ Gene panel

→ Whole genome seq

→ Exome seq → “Normal”

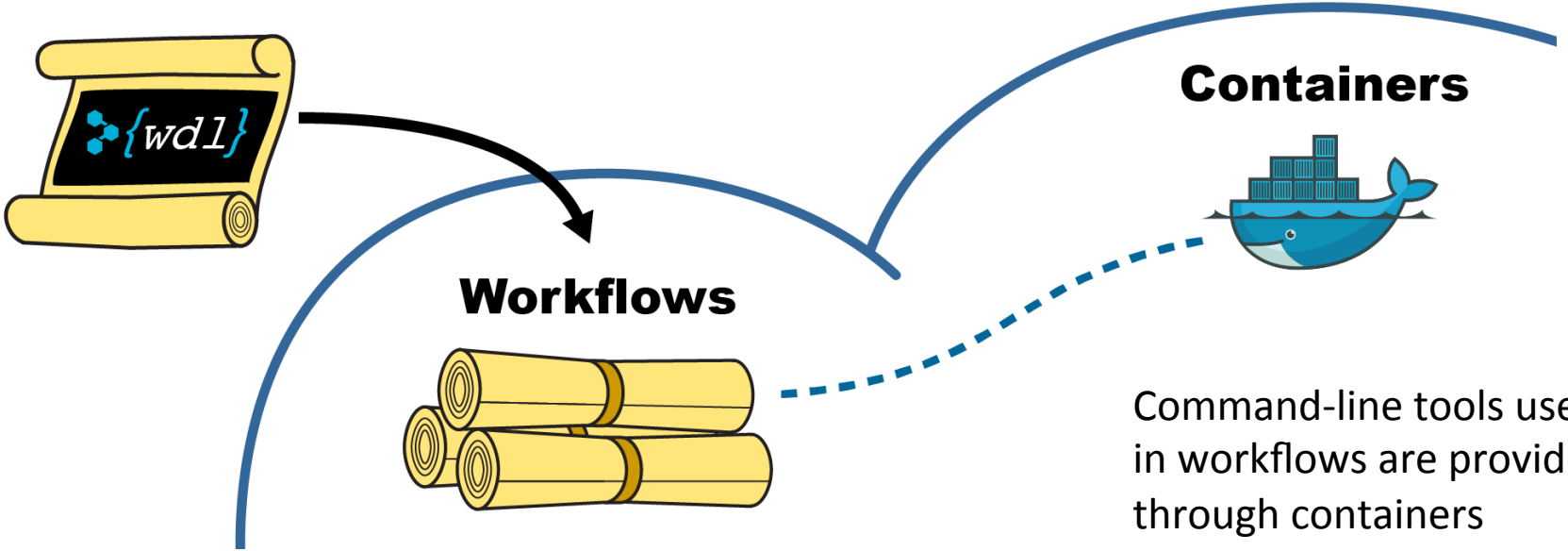


→ Exome seq → “Tumor”

→ Microbiome seq

Workflows (=methods) are stored in a Method Repository

WDL = Workflow Definition Language



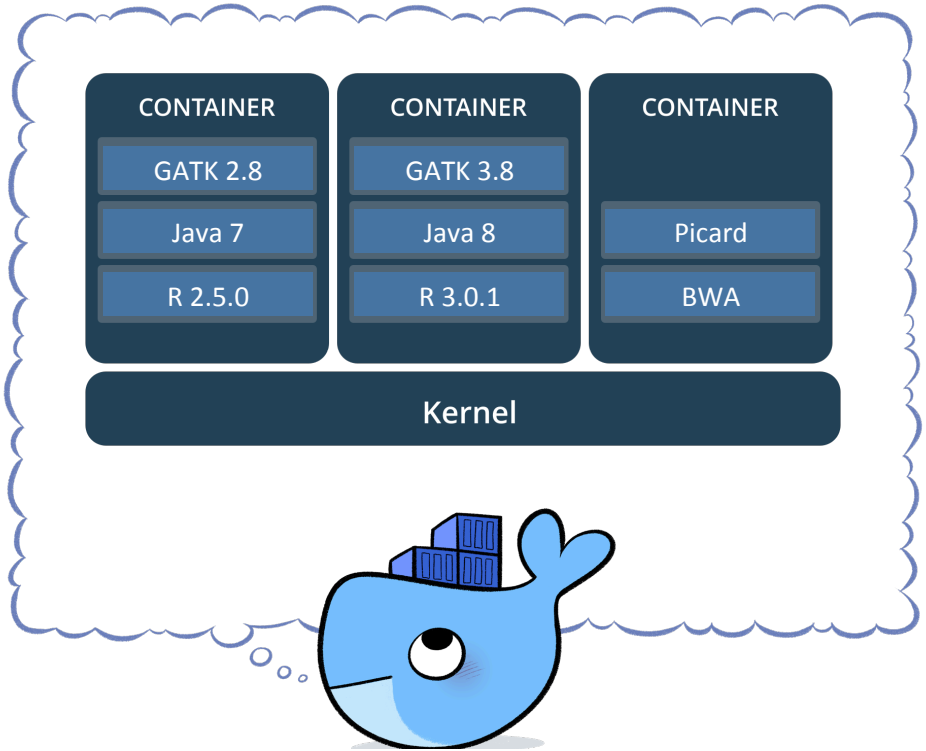
WDL is designed to be easy to read and write

```
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"
    }
}
```

http://www.openwdl.org

Containerized tools ensure portability and reproducibility



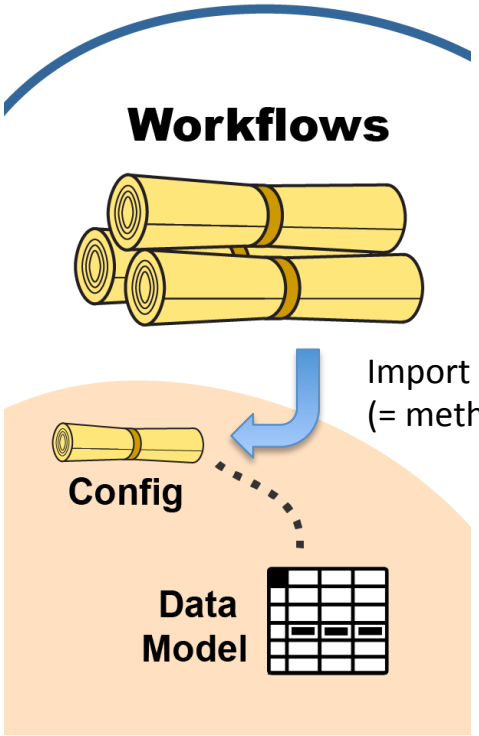
A container encapsulates
all the software dependencies
associated with running a program

Takes the guesswork out of running pipelines on different platforms!

- Portability & reproducibility !

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

Configuration = specify inputs and parameters



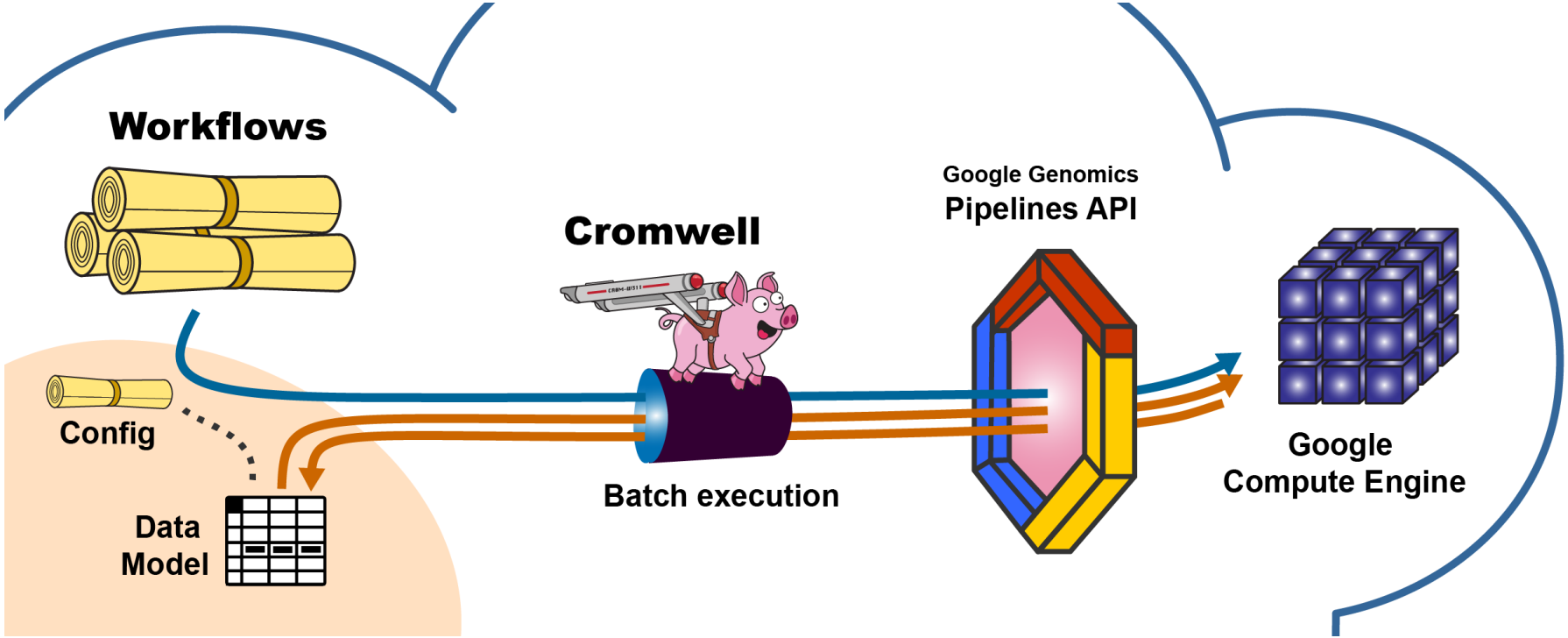
The “config” is equivalent to the WDL’s inputs JSON file

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

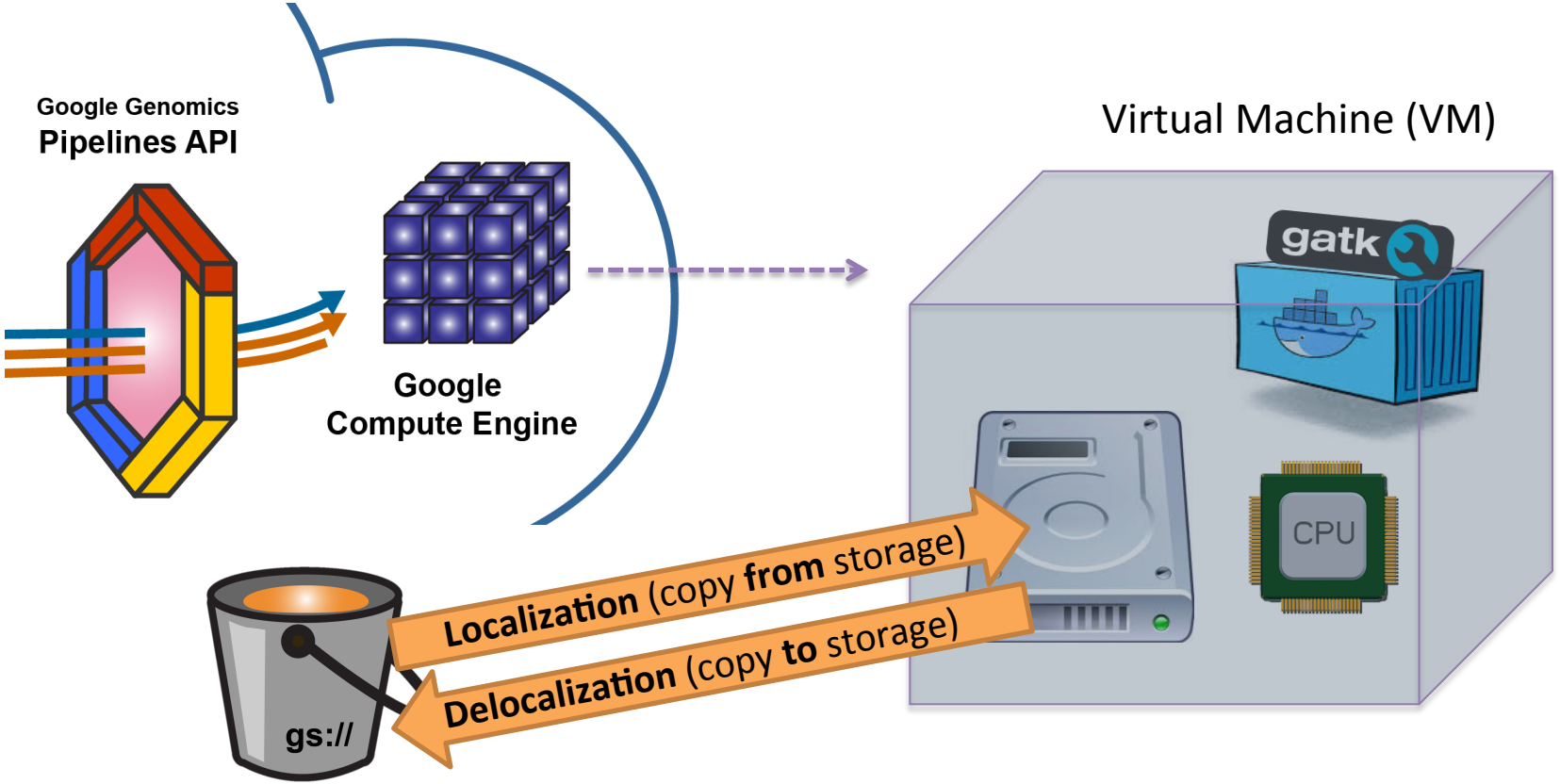
Sources of inputs/parameters:

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

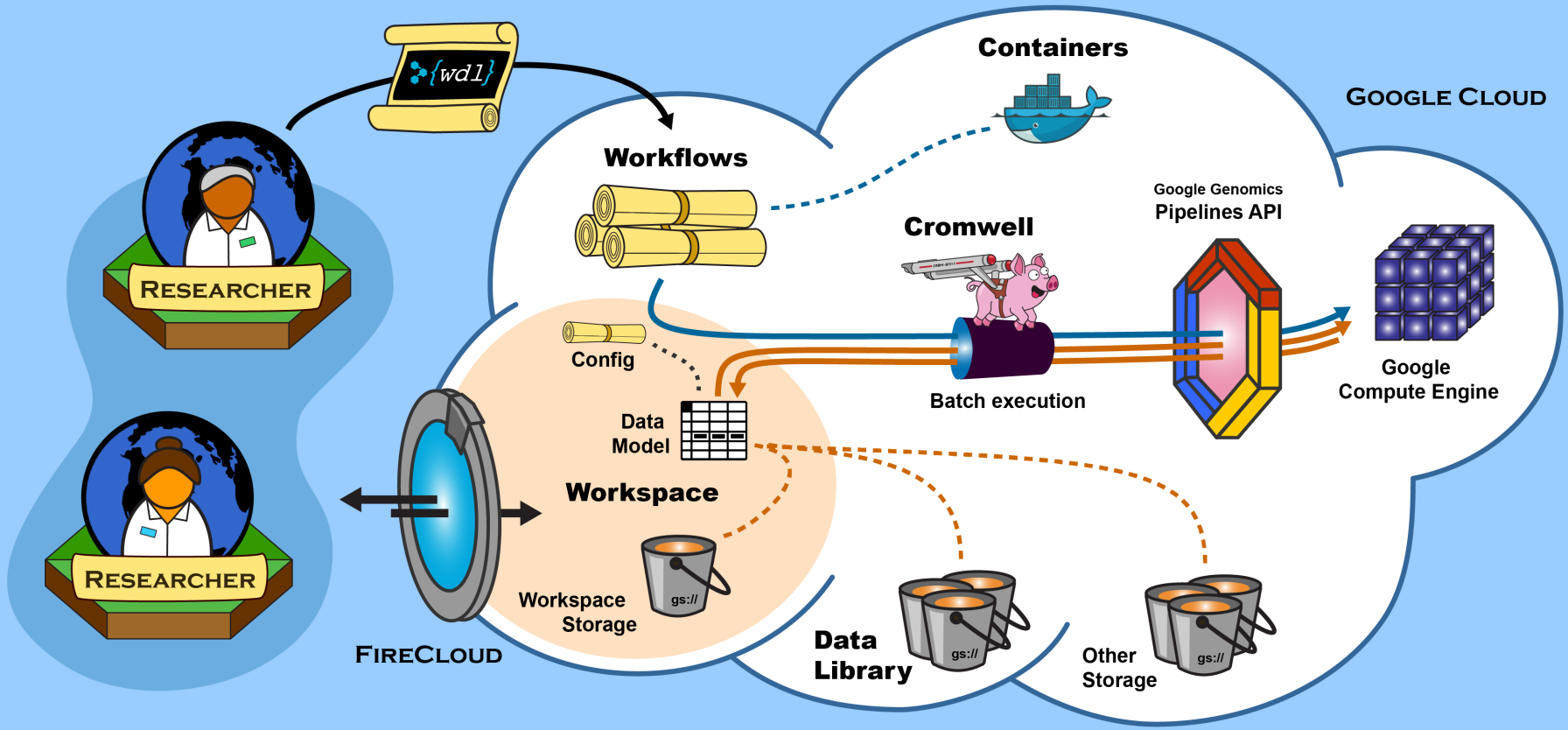
Workflow + config are sent to Cromwell for execution



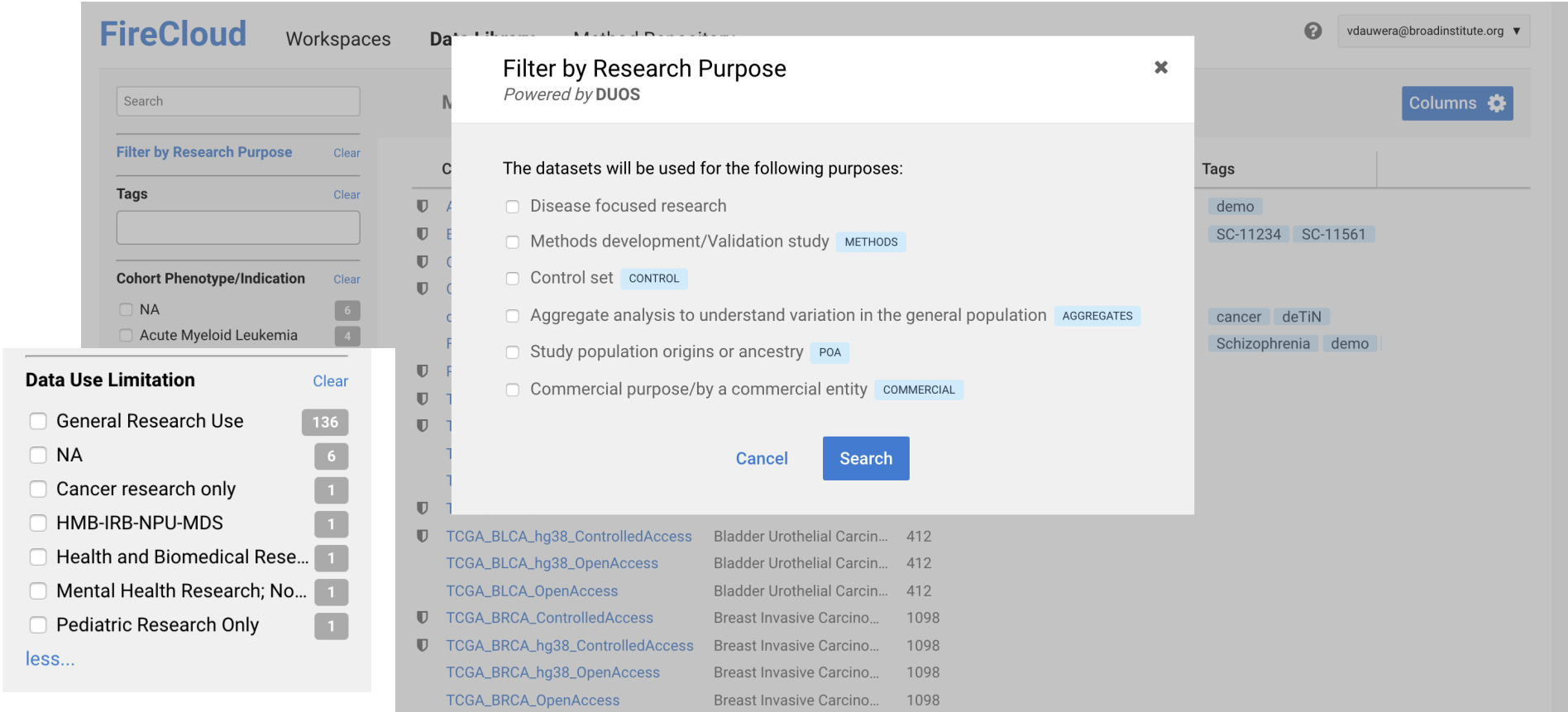
Each workflow task is executed on its own VM



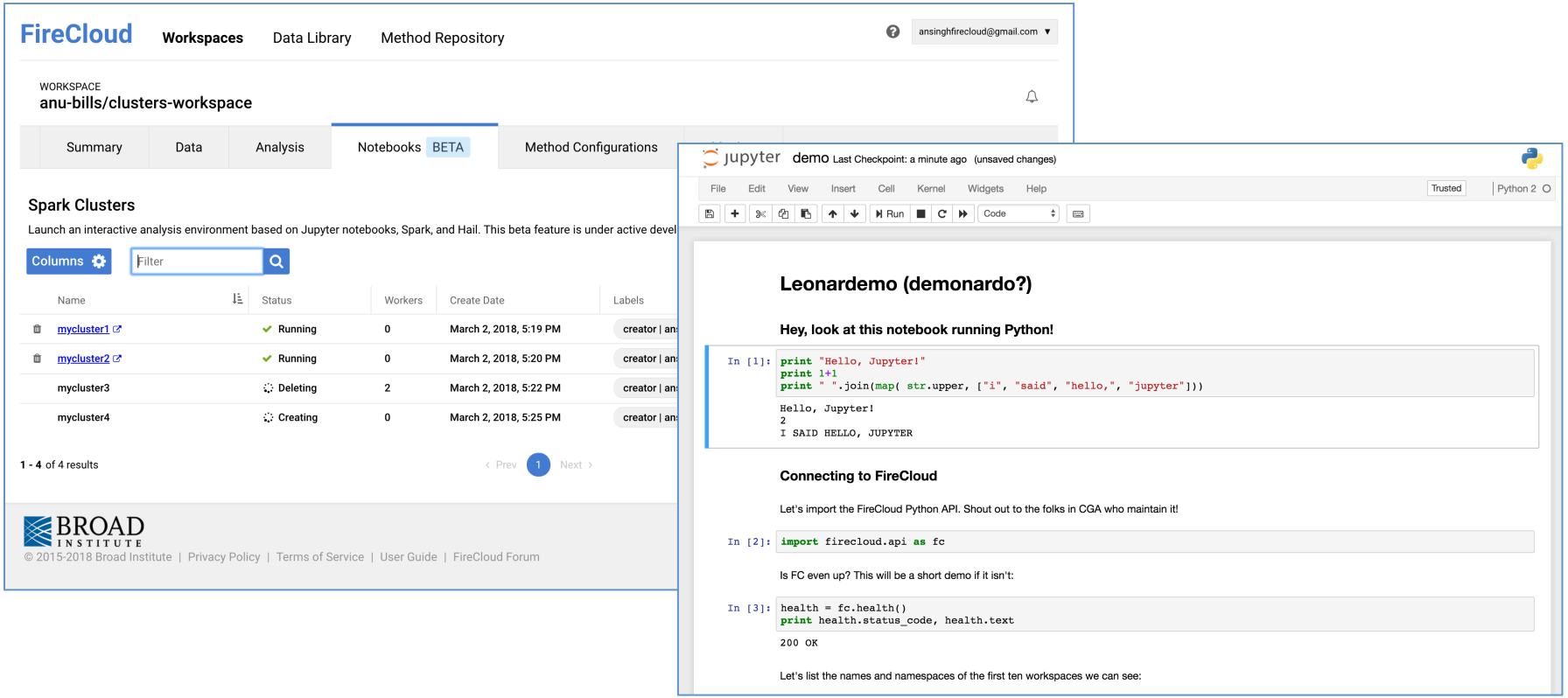
Running pipelines in FireCloud



Also: Data Library with data use-aware search



Coming soon: interactive analysis with Notebooks



https://portal.firecloud.org

GUIDED TOUR