Introduction to Workflow Management Solutions for Public Health Bioinformatics

Model for Distributed Public Health Bioinformatics: Week 3 – Connecting WDL Workflows with Terra.Bio

Tuesday May 11th, 2022
Kevin G. Libuit, MS | Theiagen Genomics

Training Workshop Overview

Communication and Support

- Slack workspaces:
  - **Terra-US-PHL; #wdl-writing**
  - **StaPH-B; #workflow-management, #cromwell_noobz**
- Email: **support@terrapublichealth.zendesk.com**
- Weekly Office Hours:
  - **Mountain Region - Friday 9-10AM (PDT)**
  - **North East Region - Friday 10-11AM (PDT)**

Training Workshop Overview

Main Course Objective

Learn how to use **workflow management systems** to **develop accessible**, **interoperable**, **and reproducible** public health bioinformatics solutions

{wdl} nextflow

Last Week's Content: WDL Tasks & Workflows

Major Takeaways (WDL Task Files):

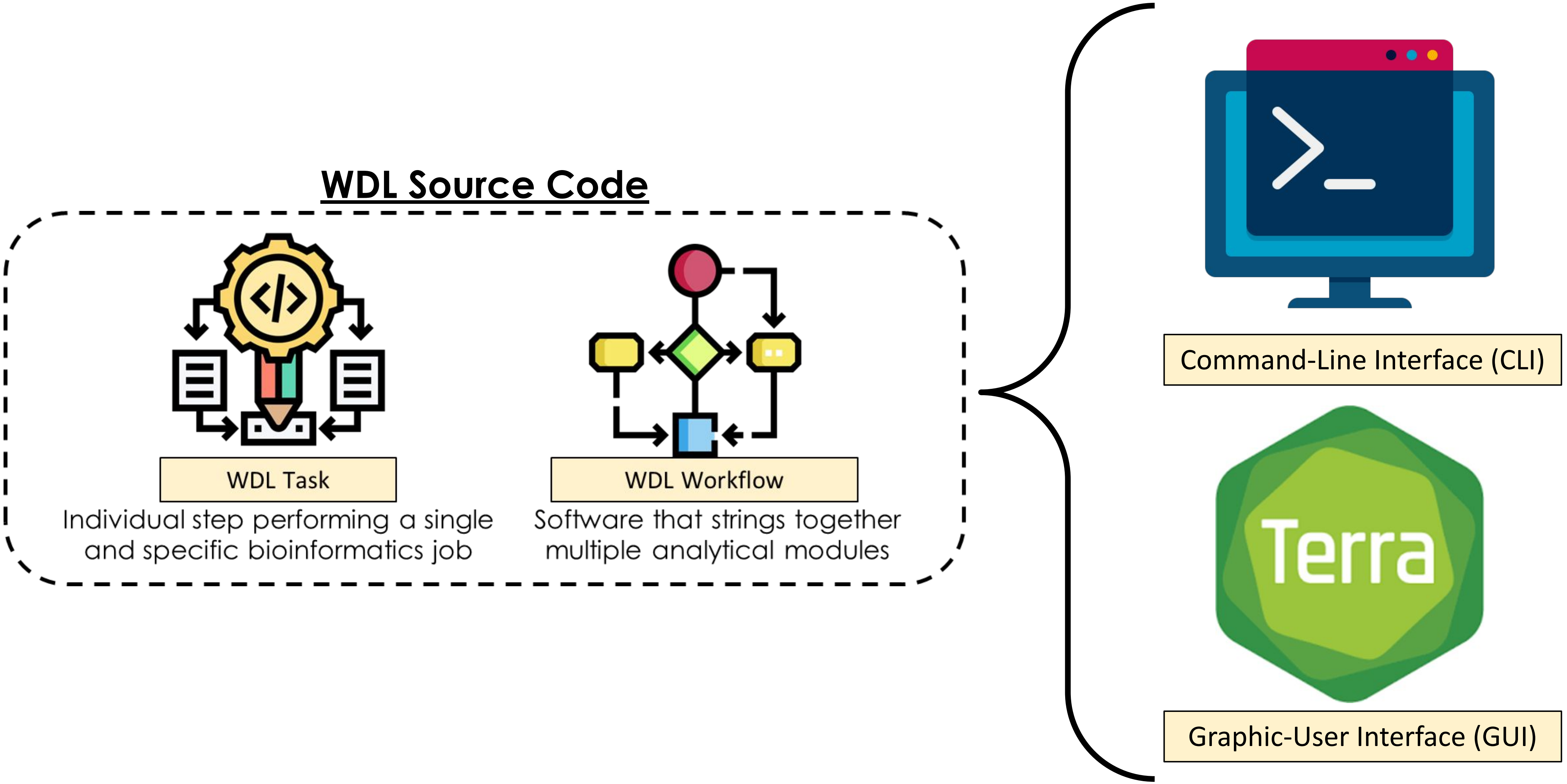
- **Input Section**
  - **Obligate Inputs:** Required for the task to run successfully
  - **Optional Inputs:** Not required for the task to run successfully
  - **Declared Inputs:** Default values of obligate or optional inputs; can be overridden when task is called in workflow
- **Command Section**
  - Two Options Available to Define a Command Element: <<< >>> or { }
- **Output Section**
  - **Obligate Outputs:** Required for the task to run successfully
  - **Optional Outputs:** Not required for the task to run successfully
- **Runtime Section**
  - Recognized runtime attributes depend on **both workflow engine & compute backend**

Last Week's Content: WDL Tasks & Workflows

Major Takeaways (WDL Workflow Files):

- **Input Section**
    - **Obligate Inputs:** Required for the workflow to run successfully
    - **Optional Inputs:** Not required for the workflow to run successfully
    - **Declared Inputs:** Default values of obligate or optional inputs; can be overridden when workflow is run
- **Call Section**
    - **Task Inputs:** Declared for each task called – even if no input values are required
    - **Task Aliases:** Helpful when a single task is called multiple times
- **Output Section**
    - **Obligate Outputs:** Required for the task to run successfully
    - **Optional Outputs:** Not required for the task to run successfully

Executing WDL Workflows



WDL Source Code

WDL Task

Individual step performing a single and specific bioinformatics job

WDL Workflow

Software that strings together multiple analytical modules

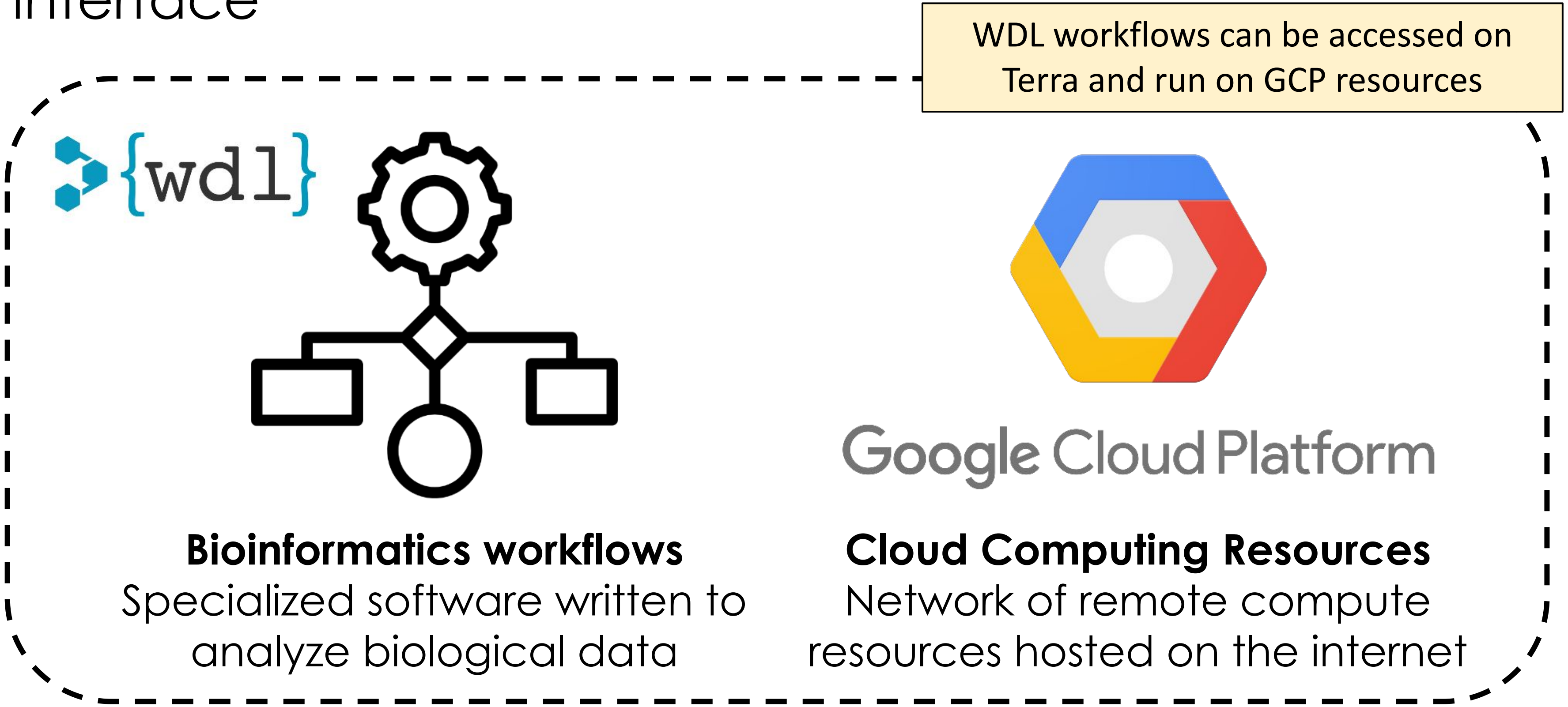
Command-Line Interface (CLI)

Graphic-User Interface (GUI)

Introduction to the Terra Platform

Terra is a **bioinformatics web application** that connects users to bioinformatics workflows (WDL) & dynamic cloud computing resources (GCP) through a clean and intuitive user interface

Terra compatibility expanding:
Will support Nextflow workflows and Azure backend



WDL workflows can be accessed on Terra and run on GCP resources

Bioinformatics workflows
Specialized software written to analyze biological data

Cloud Computing Resources
Network of remote compute resources hosted on the internet

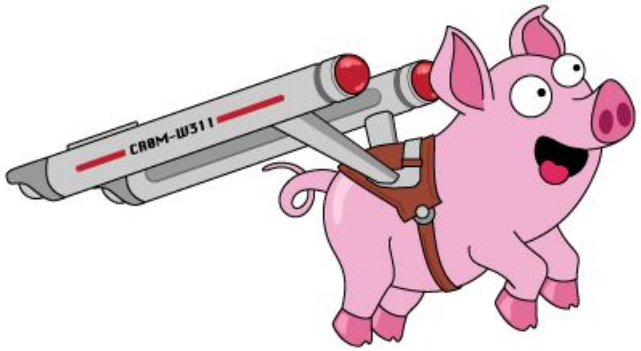
Introduction to the Terra Platform

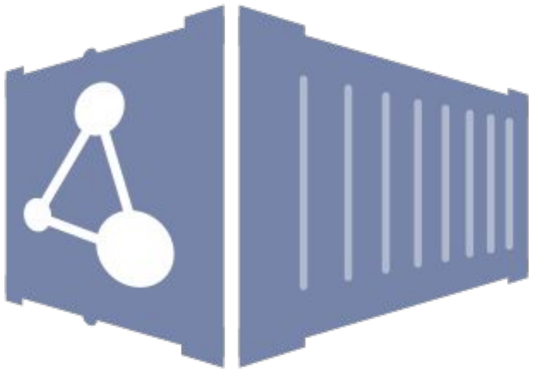
Terra users can access WDL workflows hosted on Dockstore

- Workflow repository that enables "*researchers and developers to **share and reuse** analytical **workflows and tools** in a way that makes them **machine readable and runnable** in a **variety of environments***"

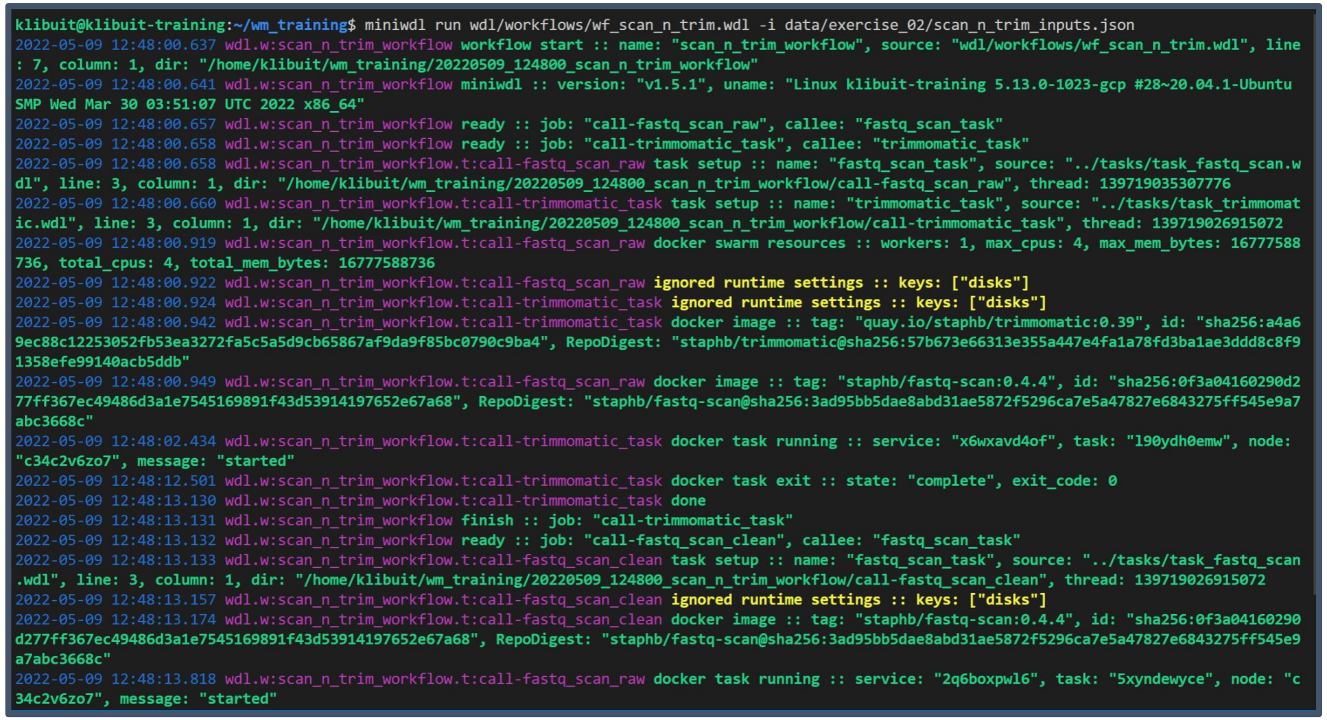
From Terra, workflows are executed using the **Cromwell engine**

- Execution on GCP backend



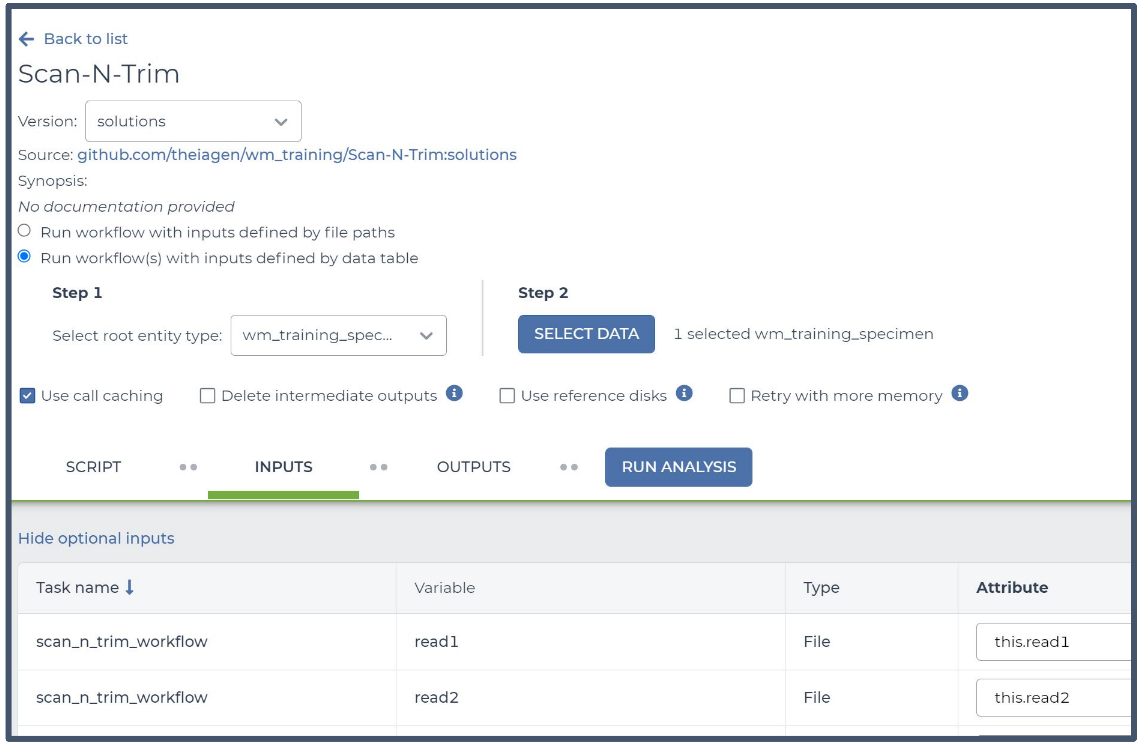


Running Scan-N-Trim Workflow



```
klibuit@klibuit-training:~/wm_training$ miniwdl run wdl/workflows/wf_scan_n_trim.wdl -i data/exercise_02/scan_n_trim_inputs.json
2022-05-09 12:48:00.637 wdl.w:scan_n_trim_workflow workflow start :: name: "scan_n_trim_workflow", source: "wdl/workflows/wf_scan_n_trim.wdl", line
: 7, column: 1, dir: "/home/klibuit/wm_training/20220509_124800_scan_n_trim_workflow"
2022-05-09 12:48:00.641 wdl.w:scan_n_trim_workflow miniwdl :: version: "v1.5.1", uname: "Linux klibuit-training 5.13.0-1023-gcp #28~20.04.1-Ubuntu
SMP Wed Mar 30 03:51:07 UTC 2022 x86_64"
2022-05-09 12:48:00.657 wdl.w:scan_n_trim_workflow ready :: job: "call-fastq_scan_raw", callee: "fastq_scan_task"
2022-05-09 12:48:00.658 wdl.w:scan_n_trim_workflow ready :: job: "call-trimmomatic_task", callee: "trimmomatic_task"
2022-05-09 12:48:00.658 wdl.w:scan_n_trim_workflow.t:call-fastq_scan_raw task setup :: name: "fastq_scan_task", source: "../tasks/task_fastq_scan.w
dl", line: 3, column: 1, dir: "/home/klibuit/wm_training/20220509_124800_scan_n_trim_workflow/call-fastq_scan_raw", thread: 139719035307776
2022-05-09 12:48:00.660 wdl.w:scan_n_trim_workflow.t:call-trimmomatic_task task setup :: name: "trimmomatic_task", source: "../tasks/task_trimmomat
ic.wdl", line: 3, column: 1, dir: "/home/klibuit/wm_training/20220509_124800_scan_n_trim_workflow/call-trimmomatic_task", thread: 139719026915072
2022-05-09 12:48:00.919 wdl.w:scan_n_trim_workflow.t:call-fastq_scan_raw docker swarm resources :: workers: 1, max_cpus: 4, max_mem_bytes: 16777588
736, total_cpus: 4, total_mem_bytes: 16777588736
2022-05-09 12:48:00.922 wdl.w:scan_n_trim_workflow.t:call-fastq_scan_raw ignored runtime settings :: keys: ["disks"]
2022-05-09 12:48:00.924 wdl.w:scan_n_trim_workflow.t:call-trimmomatic_task ignored runtime settings :: keys: ["disks"]
2022-05-09 12:48:00.942 wdl.w:scan_n_trim_workflow.t:call-trimmomatic_task docker image :: tag: "quay.io/staphb/trimmomatic:0.39", id: "sha256:a4a6
9ec88c12253052fb53ea3272fa5c5a5d9cb65867af9da9f85bc0790c9ba4", RepoDigest: "staphb/trimmomatic@sha256:57b673e66313e355a447e4fa1a78fd3ba1ae3ddd8c8f9
1358efe99140acb5ddb"
2022-05-09 12:48:00.949 wdl.w:scan_n_trim_workflow.t:call-fastq_scan_raw docker image :: tag: "staphb/fastq-scan:0.4.4", id: "sha256:0f3a04160290d2
77ff367ec49486d3a1e7545169891f43d53914197652e67a68", RepoDigest: "staphb/fastq-scan@sha256:3ad95bb5dae8abd31ae5872f5296ca7e5a47827e6843275ff545e9a7
abc3668c"
2022-05-09 12:48:02.434 wdl.w:scan_n_trim_workflow.t:call-trimmomatic_task docker task running :: service: "x6wxavd4of", task: "190ydh0emw", node:
"c34c2v6zo7", message: "started"
2022-05-09 12:48:12.501 wdl.w:scan_n_trim_workflow.t:call-trimmomatic_task docker task exit :: state: "complete", exit_code: 0
2022-05-09 12:48:13.130 wdl.w:scan_n_trim_workflow.t:call-trimmomatic_task done
2022-05-09 12:48:13.131 wdl.w:scan_n_trim_workflow finish :: job: "call-trimmomatic_task"
2022-05-09 12:48:13.132 wdl.w:scan_n_trim_workflow ready :: job: "call-fastq_scan_clean", callee: "fastq_scan_task"
2022-05-09 12:48:13.133 wdl.w:scan_n_trim_workflow.t:call-fastq_scan_clean task setup :: name: "fastq_scan_task", source: "../tasks/task_fastq_scan
.wdl", line: 3, column: 1, dir: "/home/klibuit/wm_training/20220509_124800_scan_n_trim_workflow/call-fastq_scan_clean", thread: 139719026915072
2022-05-09 12:48:13.157 wdl.w:scan_n_trim_workflow.t:call-fastq_scan_clean ignored runtime settings :: keys: ["disks"]
2022-05-09 12:48:13.174 wdl.w:scan_n_trim_workflow.t:call-fastq_scan_clean docker image :: tag: "staphb/fastq-scan:0.4.4", id: "sha256:0f3a04160290
d277ff367ec49486d3a1e7545169891f43d53914197652e67a68", RepoDigest: "staphb/fastq-scan@sha256:3ad95bb5dae8abd31ae5872f5296ca7e5a47827e6843275ff545e9
a7abc3668c"
2022-05-09 12:48:13.818 wdl.w:scan_n_trim_workflow.t:call-fastq_scan_raw docker task running :: service: "2q6boxpwl6", task: "5xyndewyce", node: "c
34c2v6zo7", message: "started"
```

Command-Line Interface (CLI)



Scan-N-Trim

Version: solutions

Source: github.com/theiagen/wm_training/Scan-N-Trim:solutions

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

● Run workflow(s) with inputs defined by data table

Step 1

Select root entity type: wm_training_spec...

Step 2

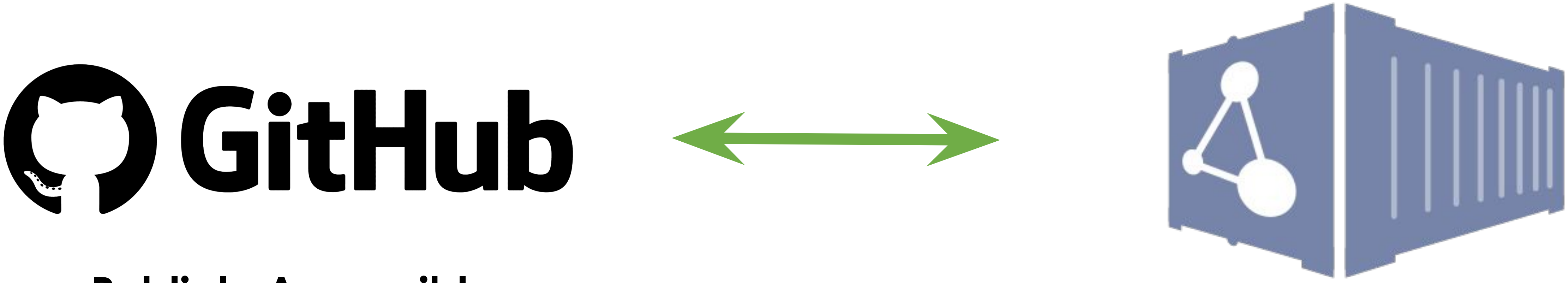
1 selected wm_training_specimen

☑ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks ☐ Retry with more memory

Task name ↓	Variable	Type	Attribute
scan_n_trim_workflow	read1	File	this.read1
scan_n_trim_workflow	read2	File	this.read2

Graphic-User Interface (GUI)

Connecting WDL Workflows to Terra.Bio



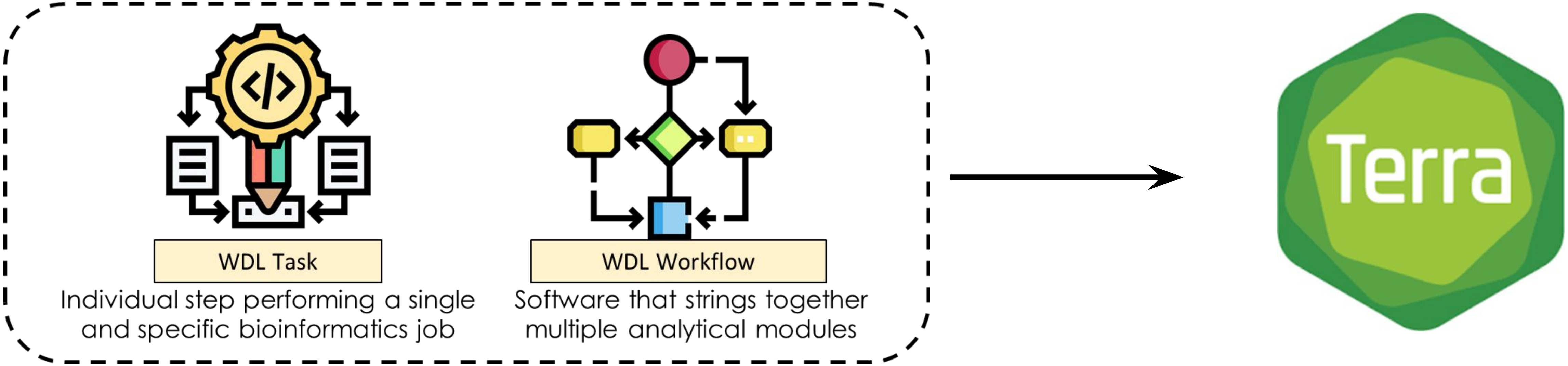
Publicly Accessible Github Repository

Containerized WDL Workflows

Linked Dockstore Account

Can continuously configure on GitHub repository after initial setup

Connecting WDL Workflows to Terra.Bio



WDL Source Code

WDL Task

Individual step performing a single and specific bioinformatics job

WDL Workflow

Software that strings together multiple analytical modules

Hosting a Terra-Accessible WDL Workflow on Dockstore

1. Host WDL source code on publicly accessible GitHub repository
2. Create a Dockstore account
3. Sync GitHub repo with Dockstore account
4. Create a .dockstore.yml file into the top level of GitHub repository
5. Publish workflow from Dockstore

Running WDL Workflows on Terra.Bio

{wdl}

Terra

Terra Platform – Data Tables & Workflows

Terra Data Table

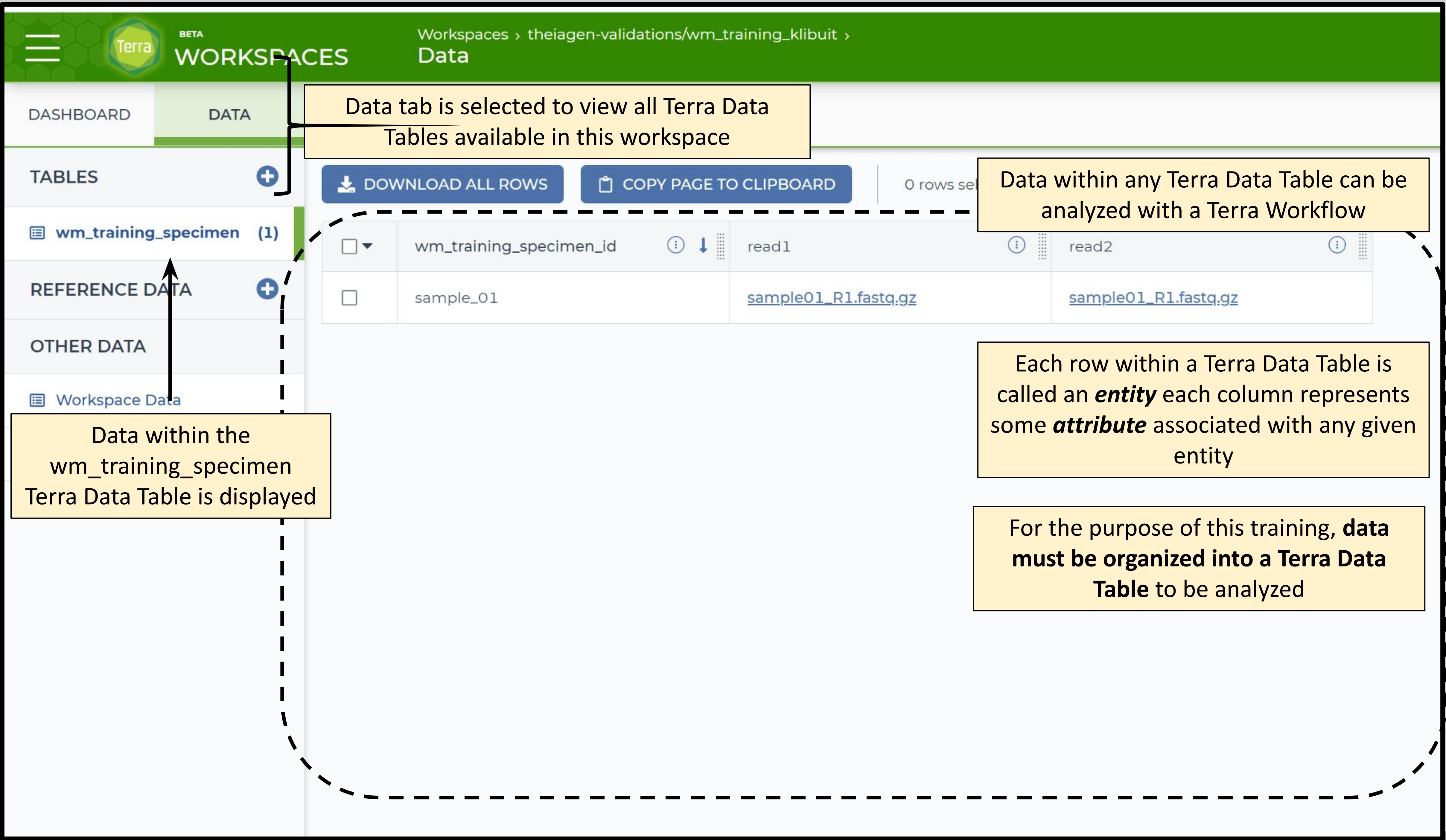
How data is organized within the Terra platform

- Data within a Terra Data Table can be uploaded by the user or generated by a Terra Workflow

Terra Workflows

WDL workflows executable from the Terra platform

- Can take inputs from a Terra Data Table
- Populates outputs to a Terra Data Table



□	wm_training_specimen_id	read1	read2
□	sample_01	sample01_R1.fastq.gz	sample01_R1.fastq.gz

Data tab is selected to view all Terra Data Tables available in this workspace

Data within any Terra Data Table can be analyzed with a Terra Workflow

Each row within a Terra Data Table is called an ***entity*** each column represents some ***attribute*** associated with any given entity

Data within the wm_training_specimen Terra Data Table is displayed

For the purpose of this training, **data must be organized into a Terra Data Table** to be analyzed

Terra Platform – Data Tables & Workflows

Once data has been **uploaded to Terra** and **organized into a Terra Data Table**, we can analyze it with a Terra Workflow

Terra Data Table

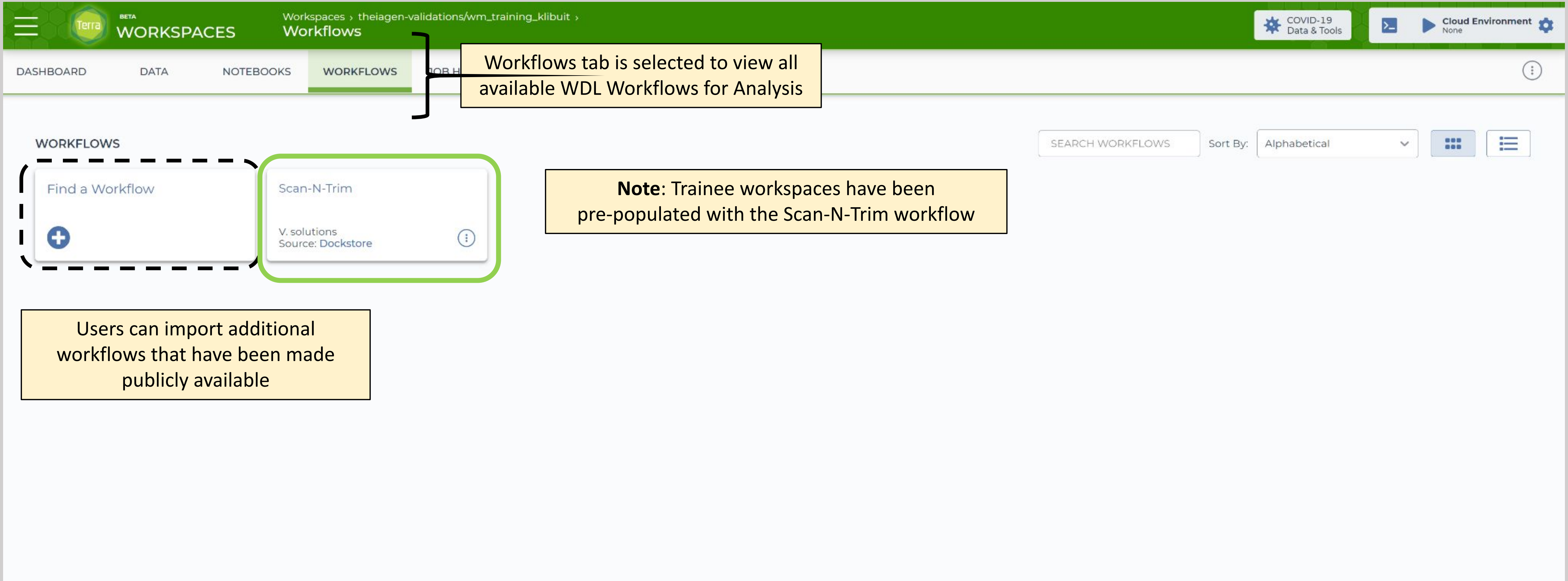
How data is organized within the Terra platform

- Data within a Terra Data Table can be uploaded by the user or generated by a Terra Workflow

Terra Workflows

Bioinformatics Workflows executable from the Terra platform

- Takes input from a Terra Data Table
- Populates outputs to a Terra Data Table



Workspaces › theiagen-validations/wm_training_klibuit ›
Workflows

WORKFLOWS

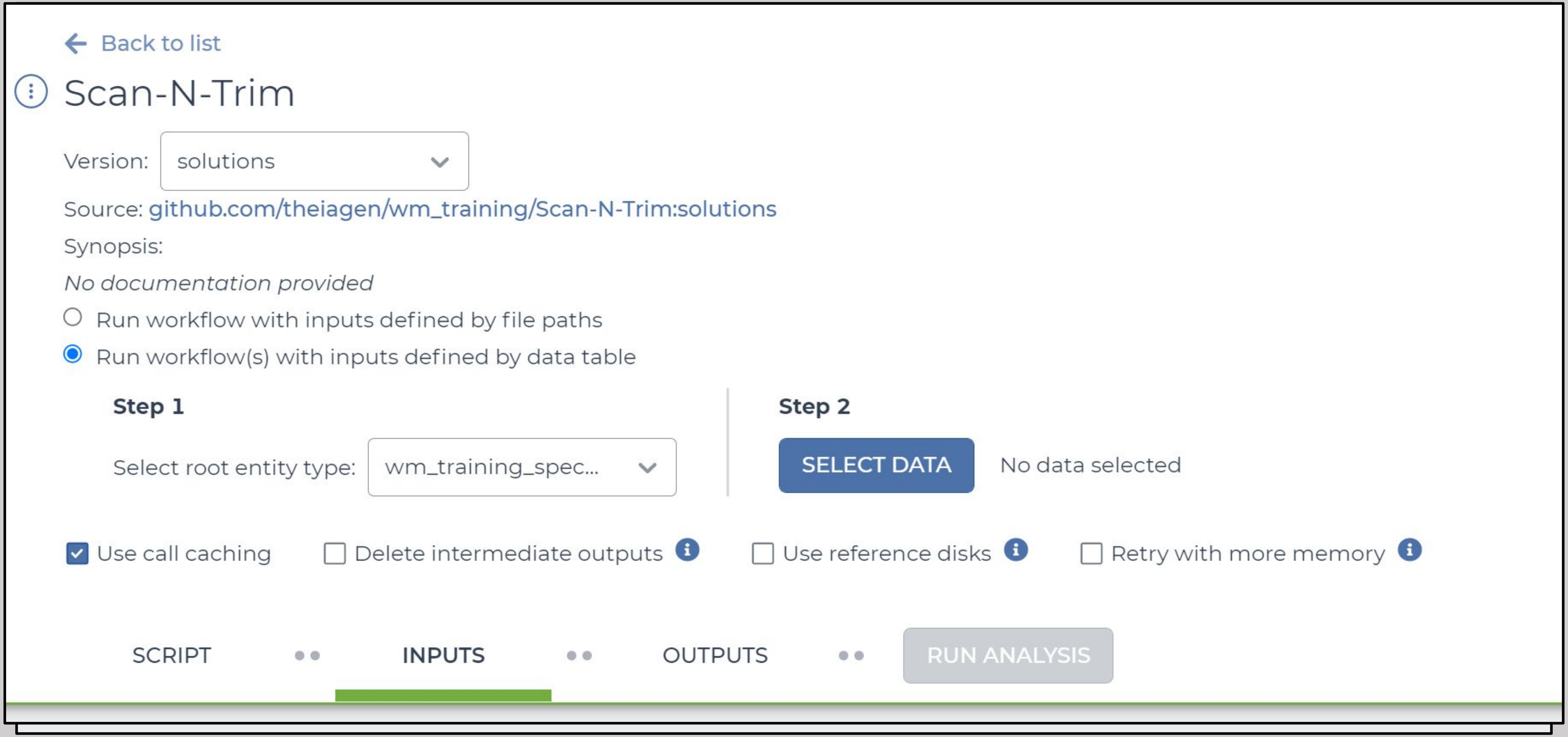
Sort By: Alphabetical

Scan-N-Trim
V. solutions
Source: Dockstore

Workflows tab is selected to view all available WDL Workflows for Analysis

Note: Trainee workspaces have been pre-populated with the Scan-N-Trim workflow

Users can import additional workflows that have been made publicly available



Scan-N-Trim

Version: solutions

Source: github.com/theiagen/wm_training/Scan-N-Trim:solutions

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

● Run workflow(s) with inputs defined by data table

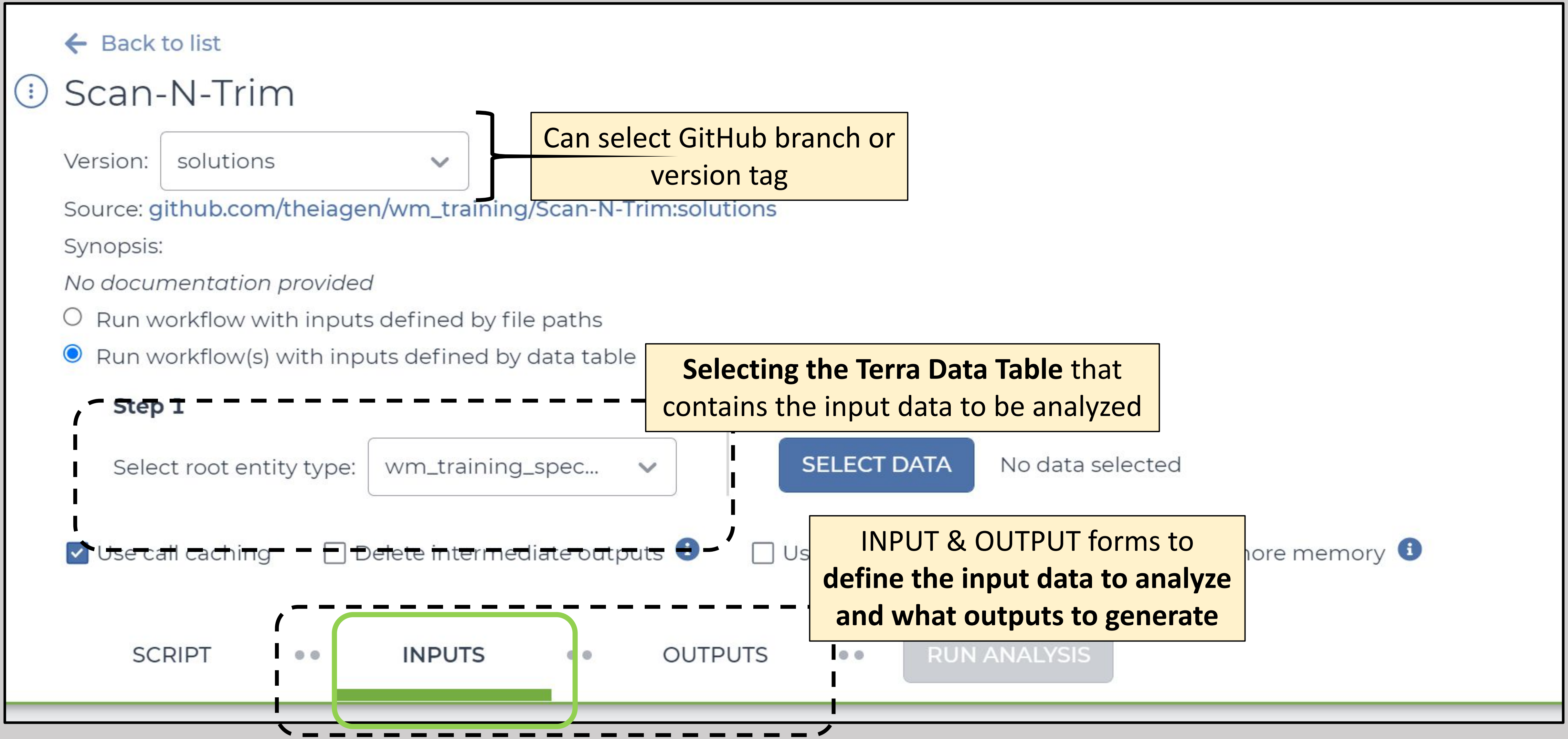
Step 1

Select root entity type: wm_training_spec...

Step 2

No data selected

☑ Use call caching □ Delete intermediate outputs □ Use reference disks □ Retry with more memory



← Back to list

Scan-N-Trim

Version: solutions

Source: github.com/theiagen/wm_training/Scan-N-Trim:solutions

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

● Run workflow(s) with inputs defined by data table

Step 1

Select root entity type: wm_training_spec...

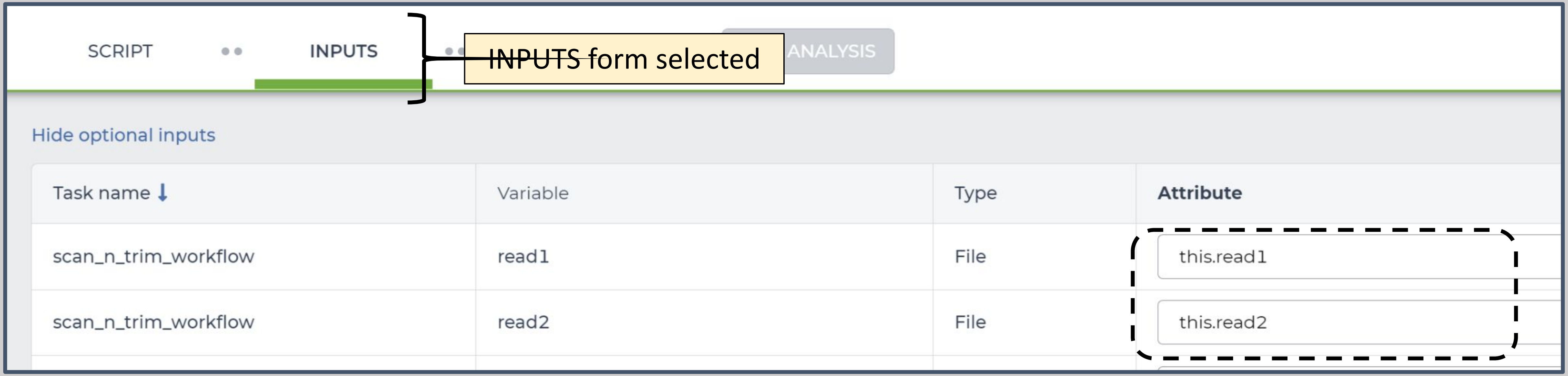
SELECT DATA No data selected

☑ Use call caching □ Delete intermediate outputs □ Us nore memory

Can select GitHub branch or version tag

Selecting the Terra Data Table that contains the input data to be analyzed

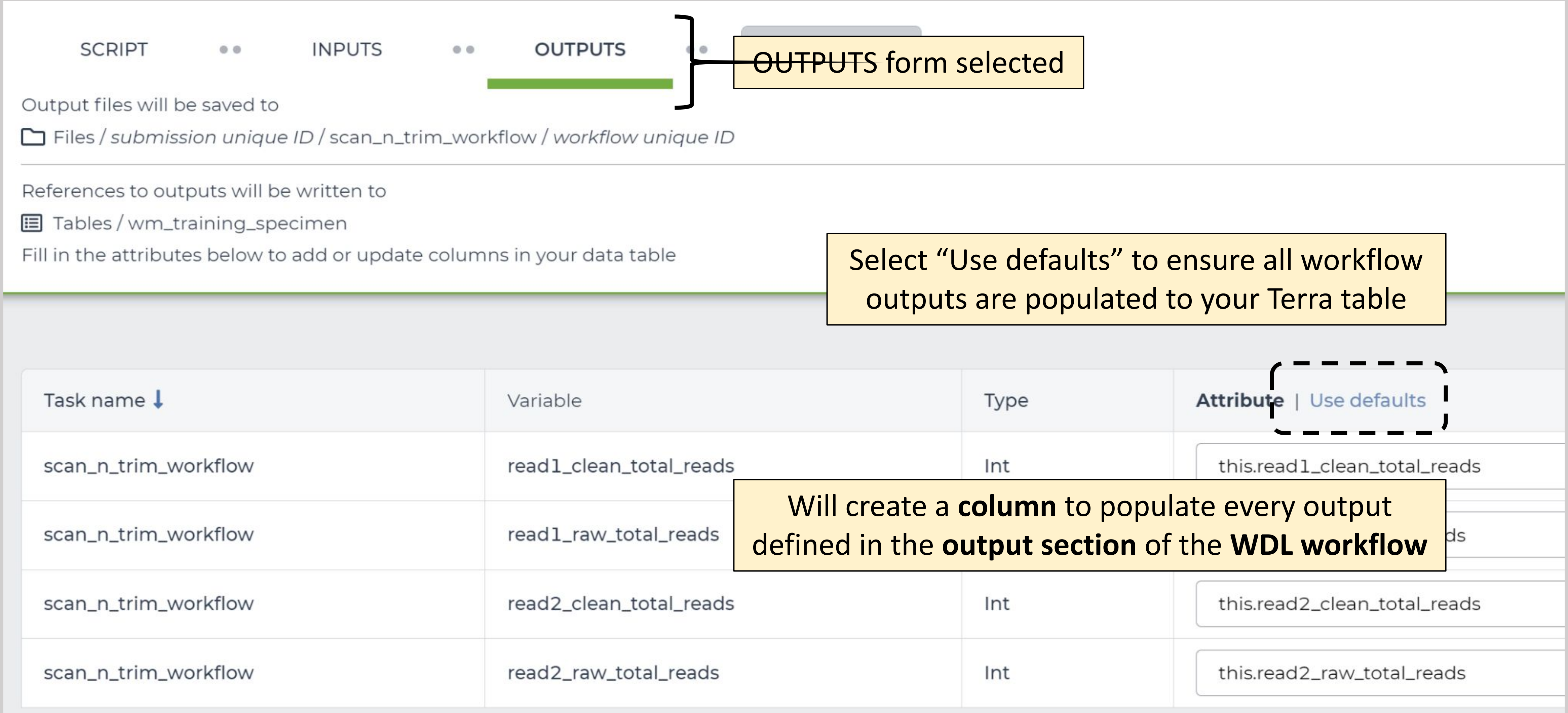
INPUT & OUTPUT forms to **define the input data to analyze and what outputs to generate**



INPUTS form selected

Task name	Variable	Type	Attribute
scan_n_trim_workflow	read1	File	this.read1
scan_n_trim_workflow	read2	File	this.read2

Inputs defined using “*this.{attribute}*” notation
this = selected root entity
{attribute} = attribute within proceeding item
this.reads = reads attribute within the root entity



OUTPUTS form selected

Output files will be saved to

Files / *submission unique ID* / scan_n_trim_workflow / *workflow unique ID*

References to outputs will be written to

Tables / wm_training_specimen

Fill in the attributes below to add or update columns in your data table

Select “Use defaults” to ensure all workflow outputs are populated to your Terra table

Task name ↓	Variable	Type	**Attribute** | Use defaults
scan_n_trim_workflow	read1_clean_total_reads	Int	this.read1_clean_total_reads
scan_n_trim_workflow	read1_raw_total_reads		ds
scan_n_trim_workflow	read2_clean_total_reads	Int	this.read2_clean_total_reads
scan_n_trim_workflow	read2_raw_total_reads	Int	this.read2_raw_total_reads

Will create a **column** to populate every output defined in the **output section** of the **WDL workflow**

Connecting WDL Workflows with Terra.Bio

Major Takeaways:

- Terra is a **bioinformatics web application** that connects users to bioinformatics workflows (WDL) & dynamic cloud computing resources (GCP) through a **clean and intuitive user interface**
  - Terra compatibility expanding to support **Nextflow workflows** and **Azure backend**
- Terra runs **WDL workflows** using the **Cromwell engine** and executes on **GCP resources**
  - Terra account setup requires GCP account
- WDL workflows must be **hosted on Dockstore** to enable Terra access
  - Requires synchronizing a public GitHub repository with Dockstore account

Lecture Exercise: Running the Scan-N-Trim WDL Workflow From Terra

Scan_N_Trim:

- Navigate to the Terra workspace provided to you for this training
- From the Workflows tab, select the Scan-N-Trim workflow
- Run this workflow with inputs defined by data table
    - ○ Select the wm_training_specimen as the root entity
    - ○ Define the required inputs and use defaults for the outputs
    - ○ Run this workflow on sample_01

Complete this exercise during our 20m session break;
we will begin again at 12:08PM (PT)

Lecture Exercise: Linking Your WDL Workflow to Terra

We will regroup again at 12:58AM (PT)

For Trainees that Have Shared a GitHub Username:

- Create a dev branch on the wm_training repository
- Merge your work onto this branch and push all commits to GitHub
- Write a .dockstore.yml file to host your Week 2 solution to Dockstore and push commit
  - Workflow name must be set as `Scan-N-Trim`
- Run this workflow from your Terra workspace

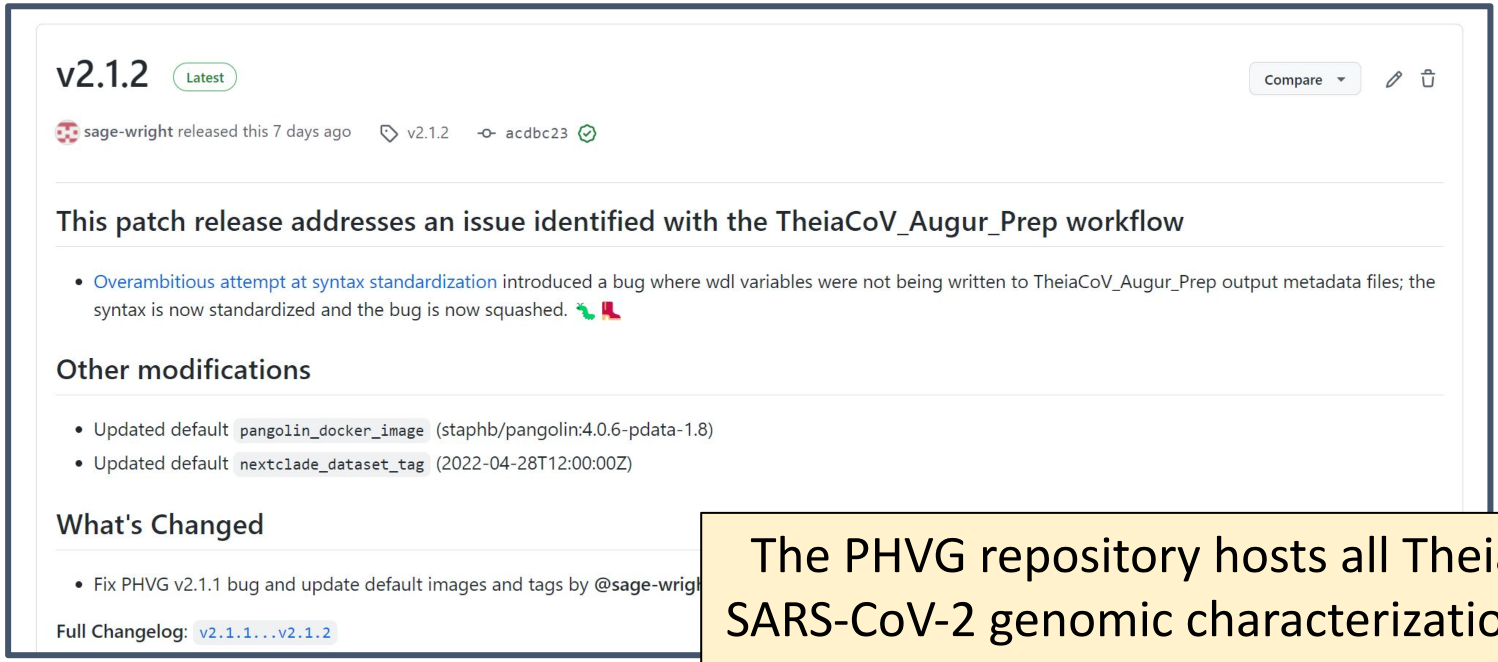
For Trainees that Have Not Shared a GitHub Username:

- Modify your Week 2 solution to capture the cleaned read files
- Add a task to your Week 2 solution to:
  - Perform Shovill assembly using these cleaned read files
  - Assess the quality of this assembly using QUAST

Terra Workflow Versions

Linked to GitHub Version Releases and Branches

- **Version Release**: Static iteration of a GitHub repository
  - Ideally a validated release of a code base that will not change



v2.1.2

sage-wright released this 7 days ago · v2.1.2 · acdbc23 ✓

This patch release addresses an issue identified with the TheiaCoV_Augur_Prep workflow

- Overambitious attempt at syntax standardization introduced a bug where wdl variables were not being written to TheiaCoV_Augur_Prep output metadata files; the syntax is now standardized and the bug is now squashed. 🐛👢

Other modifications

- Updated default `pangolin_docker_image` (staphb/pangolin:4.0.6-pdata-1.8)
- Updated default `nextclade_dataset_tag` (2022-04-28T12:00:00Z)

What's Changed

- Fix PHVG v2.1.1 bug and update default images and tags by @sage-wrigh

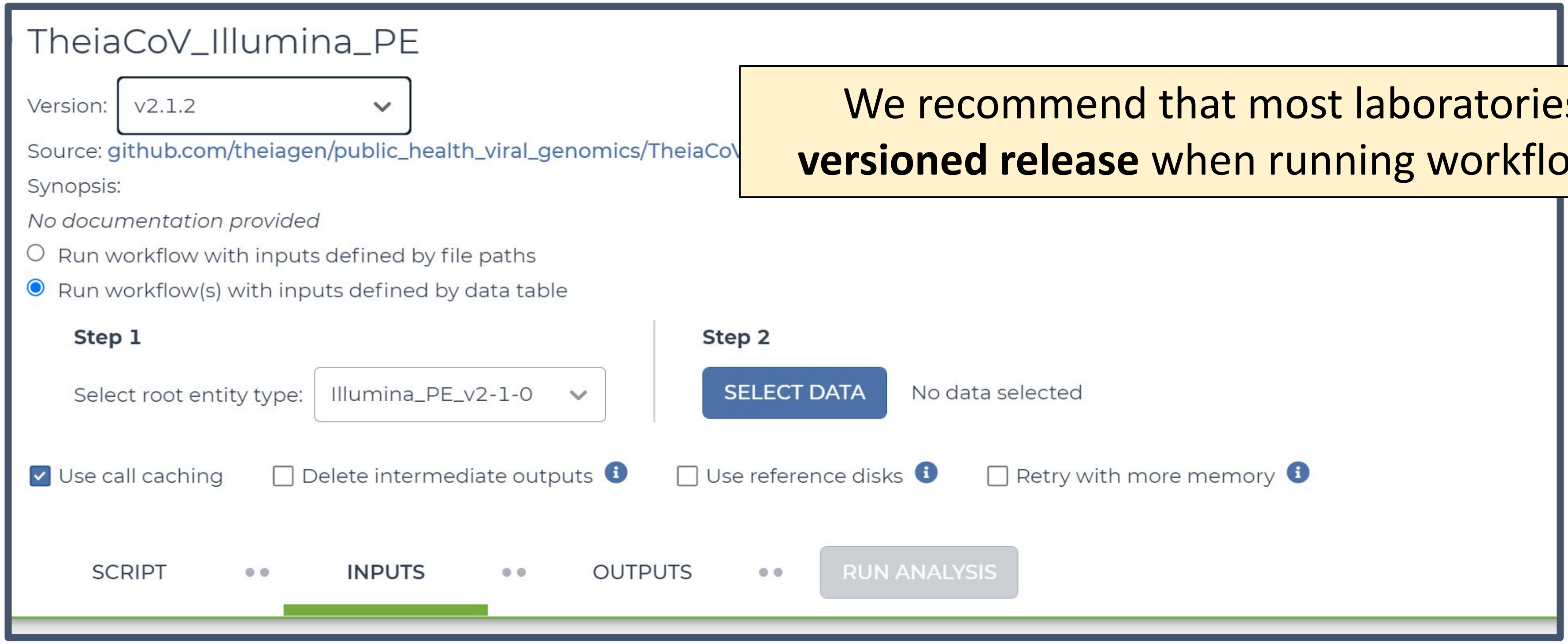
Full Changelog: `v2.1.1...v2.1.2`

The PHVG repository hosts all Theiagen workflows for SARS-CoV-2 genomic characterization, epidemiology and submission preparation

Terra Workflow Versions

Linked to GitHub Version Releases and Branches

- **Version Release**: Static iteration of a GitHub repository
  - Ideally a validated release of a code base that will not change



TheiaCoV_Illumina_PE

Version: v2.1.2

Source: github.com/theiagen/public_health_viral_genomics/TheiaCo

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

● Run workflow(s) with inputs defined by data table

Step 1

Select root entity type: Illumina_PE_v2-1-0

Step 2

No data selected

☑ Use call caching □ Delete intermediate outputs □ Use reference disks □ Retry with more memory

We recommend that most laboratories **utilize a versioned release** when running workflows on Terra

Terra Workflow Versions

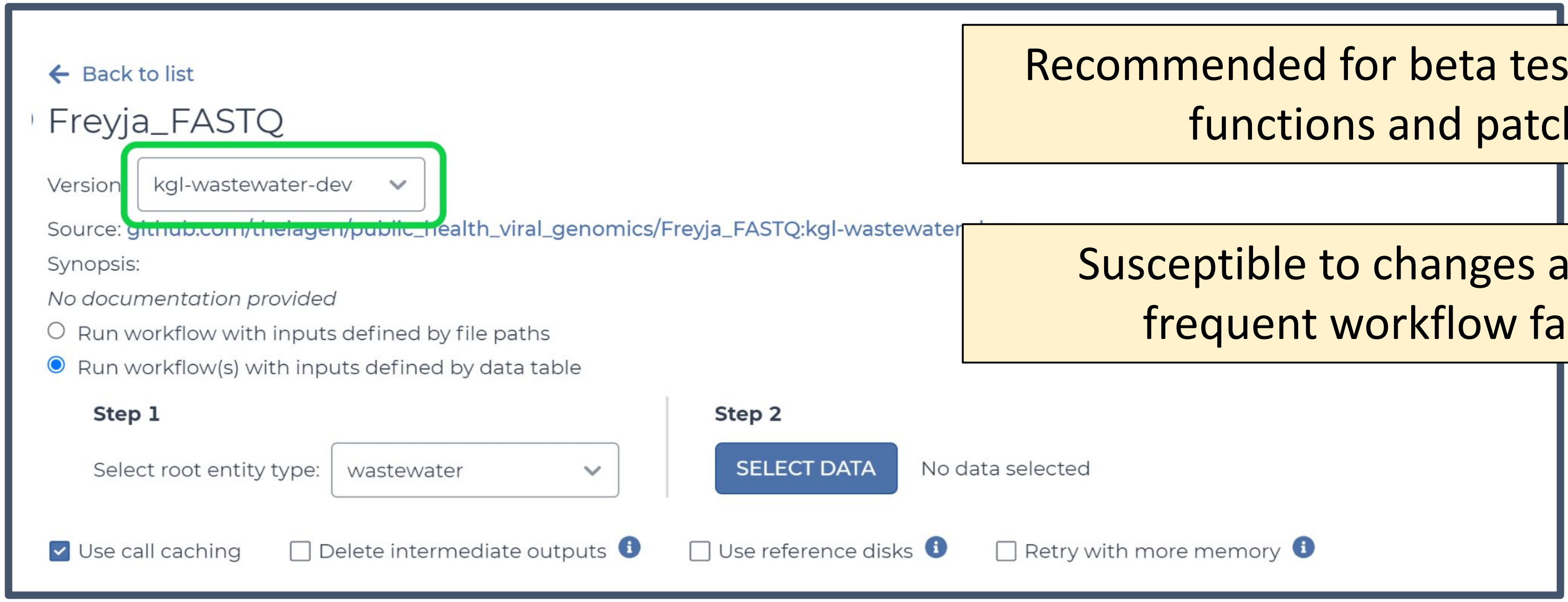
Linked to GitHub Version Releases and Branches

- **Branches**: Dynamic iteration of a GitHub repository
    - Unvalidated and under active development; susceptible to more regular code changes for optimization and testing
    - Utilized to validate patches or new workflow functions before including into a versioned release
    - Theiagen practice:
        - Designate the name of the branch with the developer, new function/patch, and “*dev*”, e.g. “*kgl-wastewater-dev*”

Terra Workflow Versions

Linked to GitHub Version Releases and Branches

- **Branches**: Dynamic iteration of a GitHub repository



← Back to list

Freyja_FASTQ

Version: kgl-wastewater-dev

Source: github.com/theiagen/public_health_viral_genomics/Freyja_FASTQ:kgl-wastewater

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

● Run workflow(s) with inputs defined by data table

Step 1

Select root entity type: wastewater

Step 2

No data selected

☑ Use call caching □ Delete intermediate outputs □ Use reference disks □ Retry with more memory

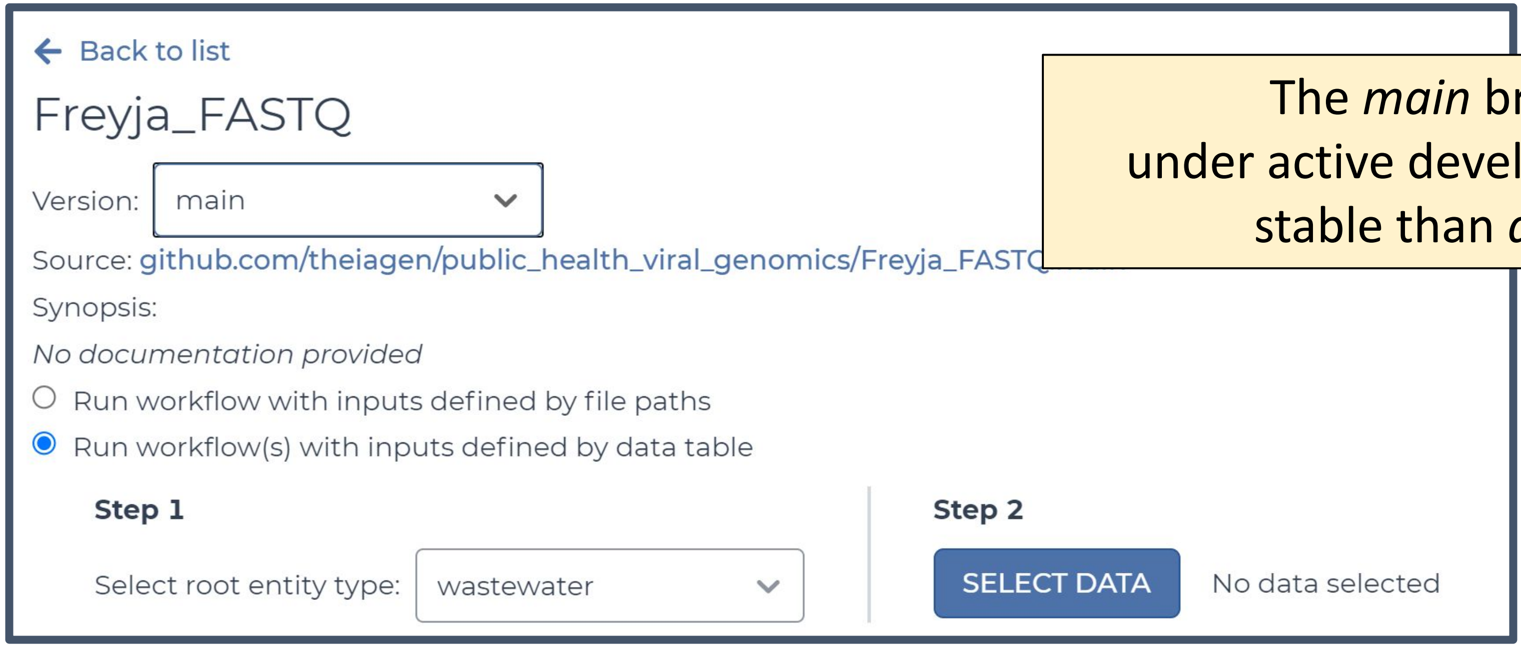
Recommended for beta testers of new functions and patches

Susceptible to changes and more frequent workflow failures

Terra Workflow Versions

Linked to GitHub Version Releases and Branches

- **Branches**: Dynamic iteration of a GitHub repository



Freyja_FASTQ

Version: main

Source: github.com/theiagen/public_health_viral_genomics/Freyja_FASTQ

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

● Run workflow(s) with inputs defined by data table

Step 1

Select root entity type: wastewater

Step 2

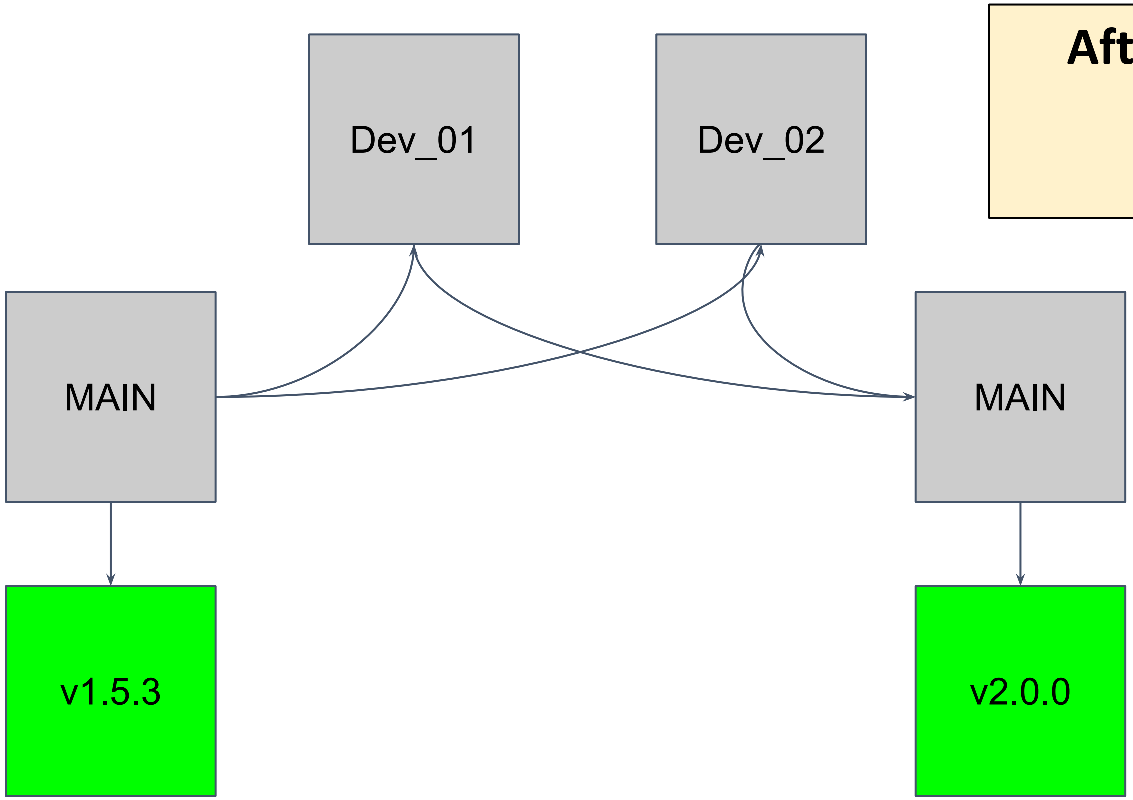
No data selected

The *main* branch is also under active development, but more stable than *dev* branches

Terra Workflow Versions

Linked to GitHub Version Releases and Branches

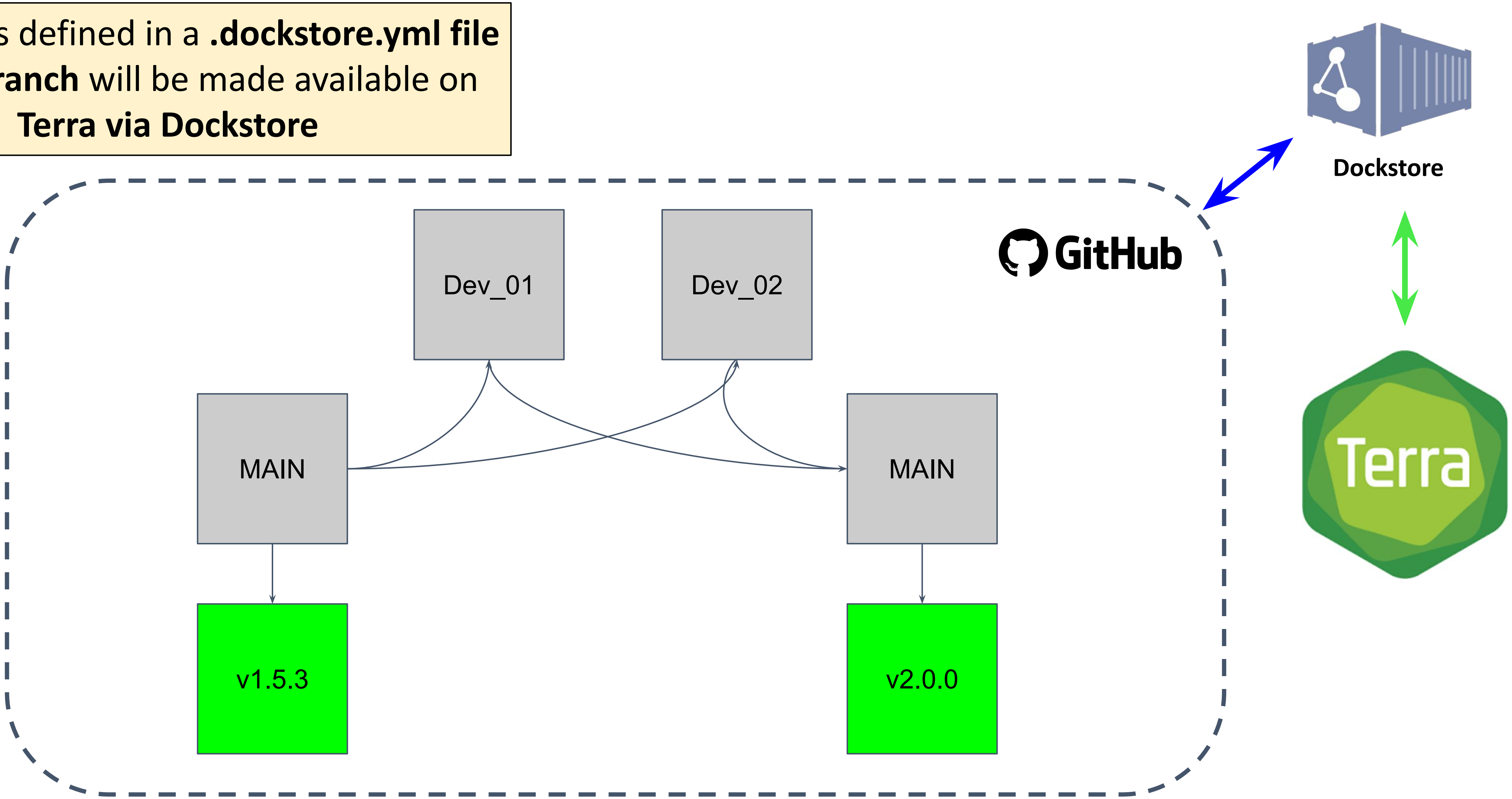
- **Branches**: Dynamic iteration of a GitHub repository



After validating the functionality of a **dev branch**, it gets **merged** it into the **main branch**

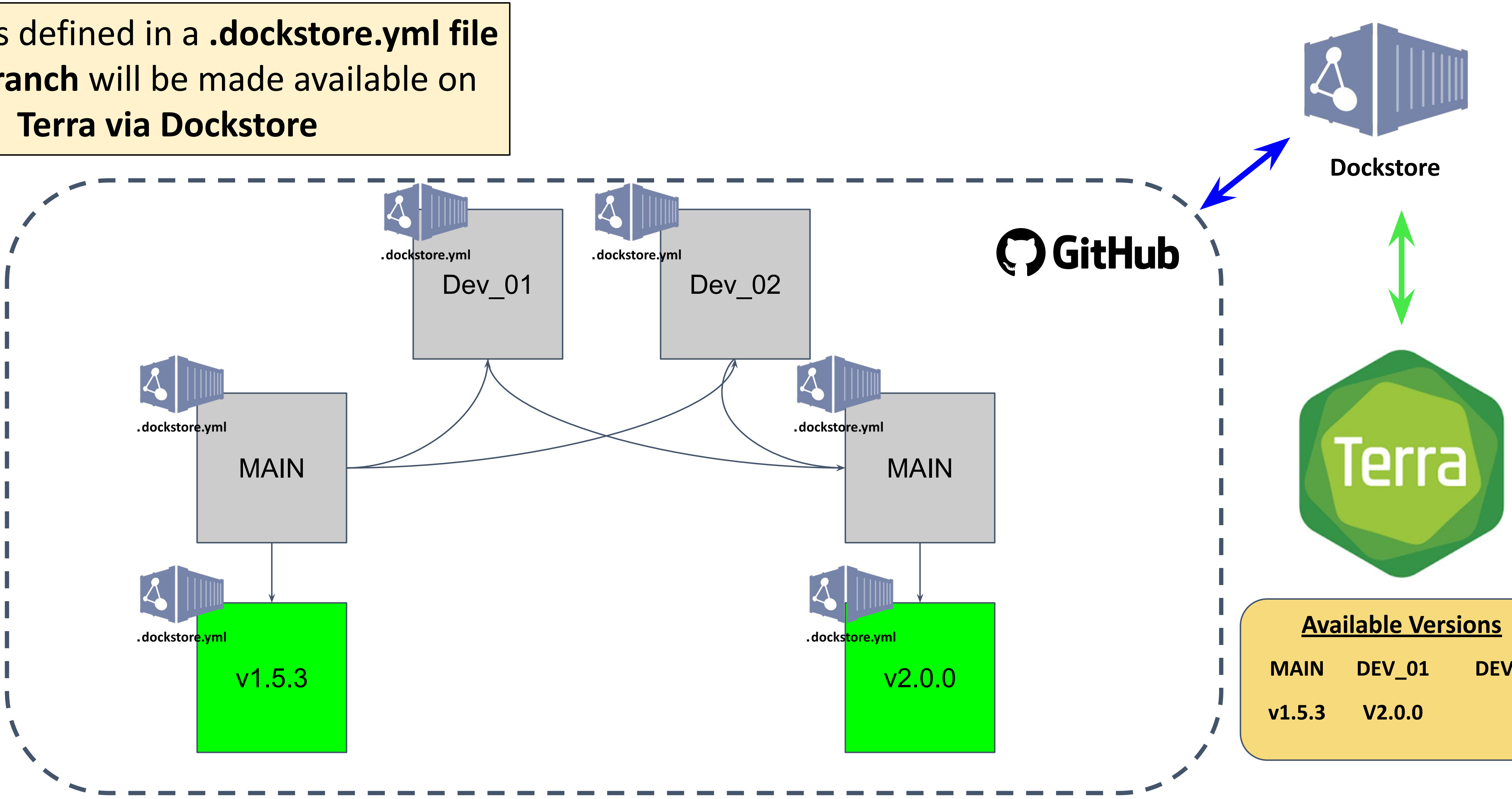
Terra Workflow Versions

Workflows defined in a **.dockstore.yml file** on **any branch** will be made available on **Terra via Dockstore**



Terra Workflow Versions

Workflows defined in a **.dockstore.yml file** on **any branch** will be made available on **Terra via Dockstore**



Available Versions

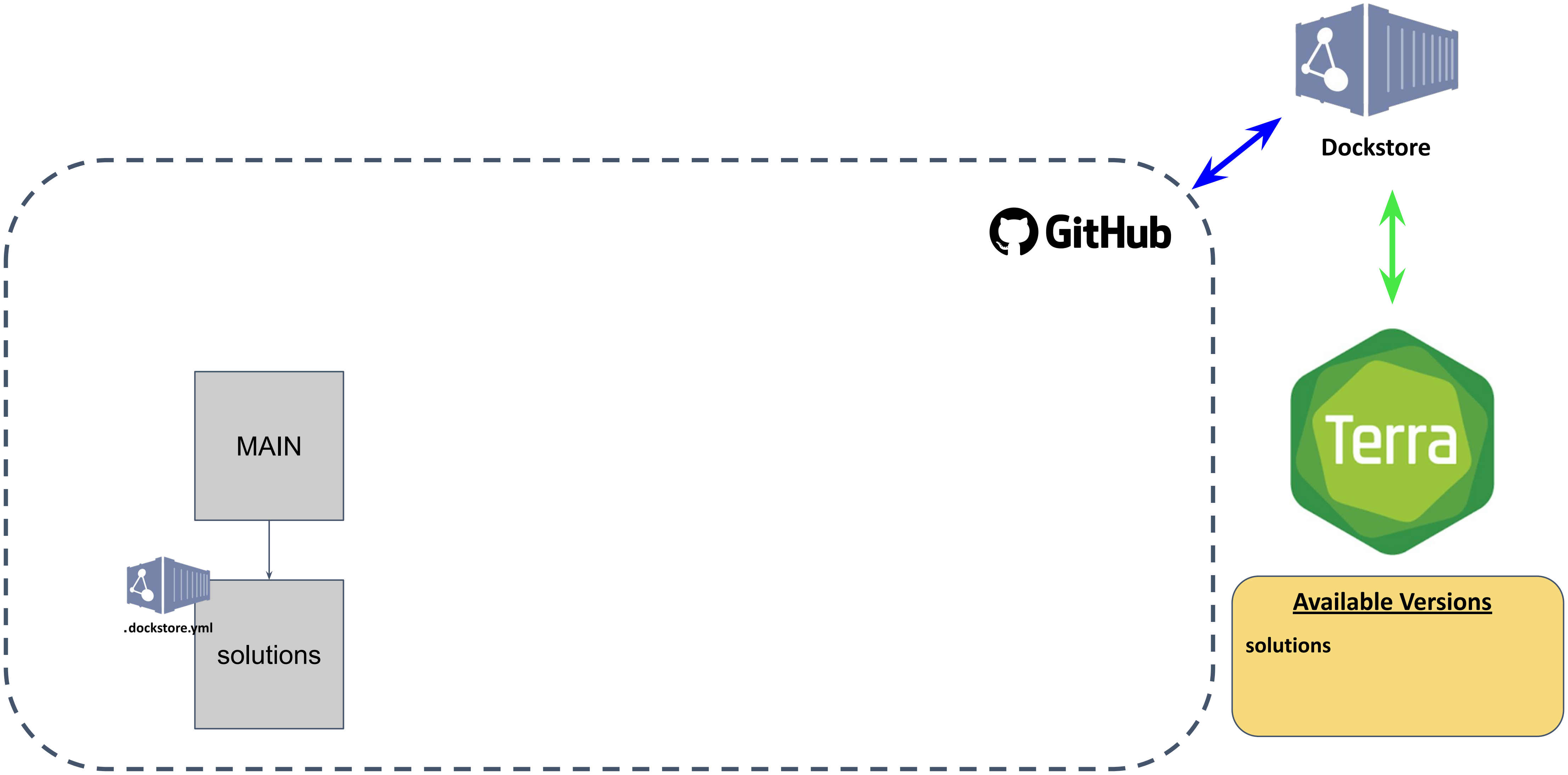
MAIN	DEV_01	DEV_02
v1.5.3	V2.0.0	

Lecture Exercise: Linking Your WDL Workflow to Terra

For Trainees that Have Shared a GitHub Username:

- Create a dev branch on the wm_training repository
- Merge your work onto this branch and push all commits to GitHub
- Write a .dockstore.yml file to host your Week 2 solution to Dockstore and push commit
    - Workflow name must be set as `Scan-N-Trim`
- Run this workflow from your Terra workspace

Terra Workflow Versions



Dockstore

GitHub

MAIN

solutions

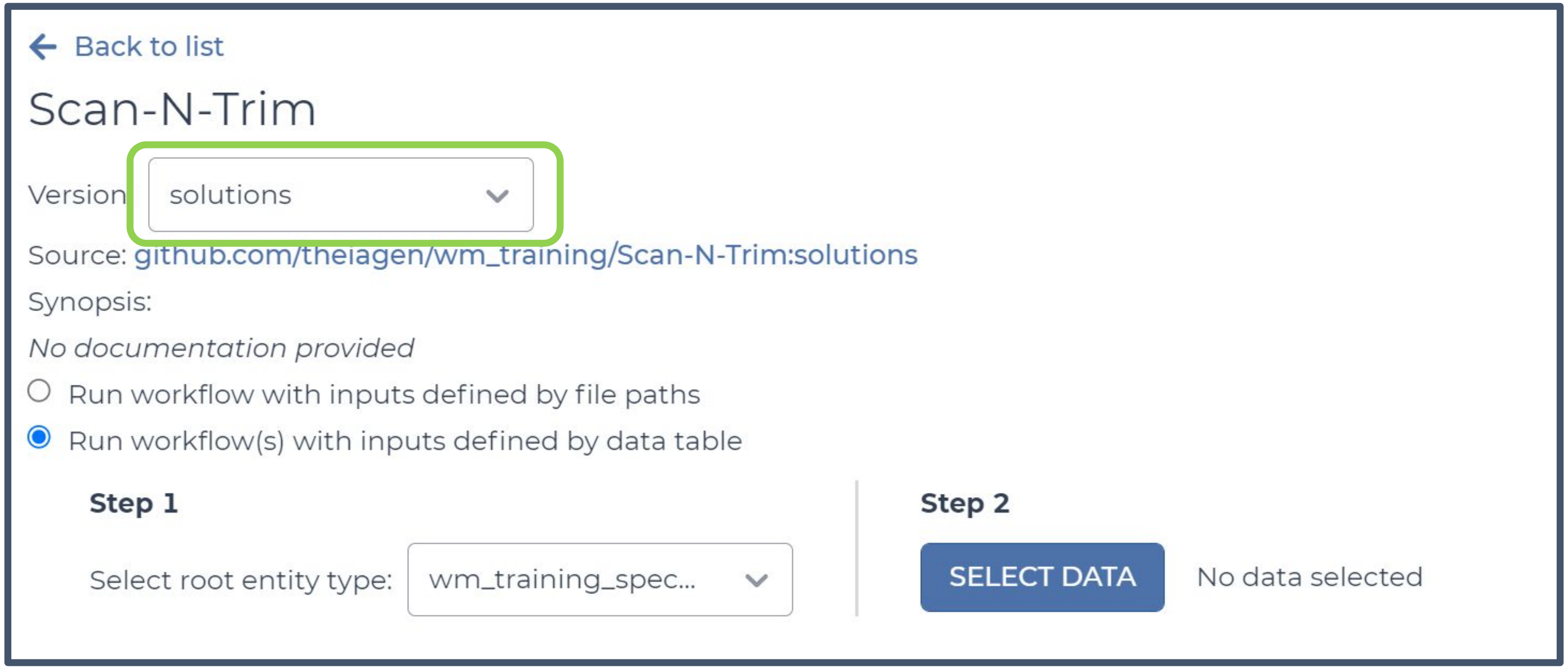
.dockstore.yml

Terra

Available Versions

solutions

From Terra Workspace



← Back to list

Scan-N-Trim

Version solutions

Source: github.com/theiagen/wm_training/Scan-N-Trim:solutions

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

● Run workflow(s) with inputs defined by data table

Step 1

Select root entity type: wm_training_spec...

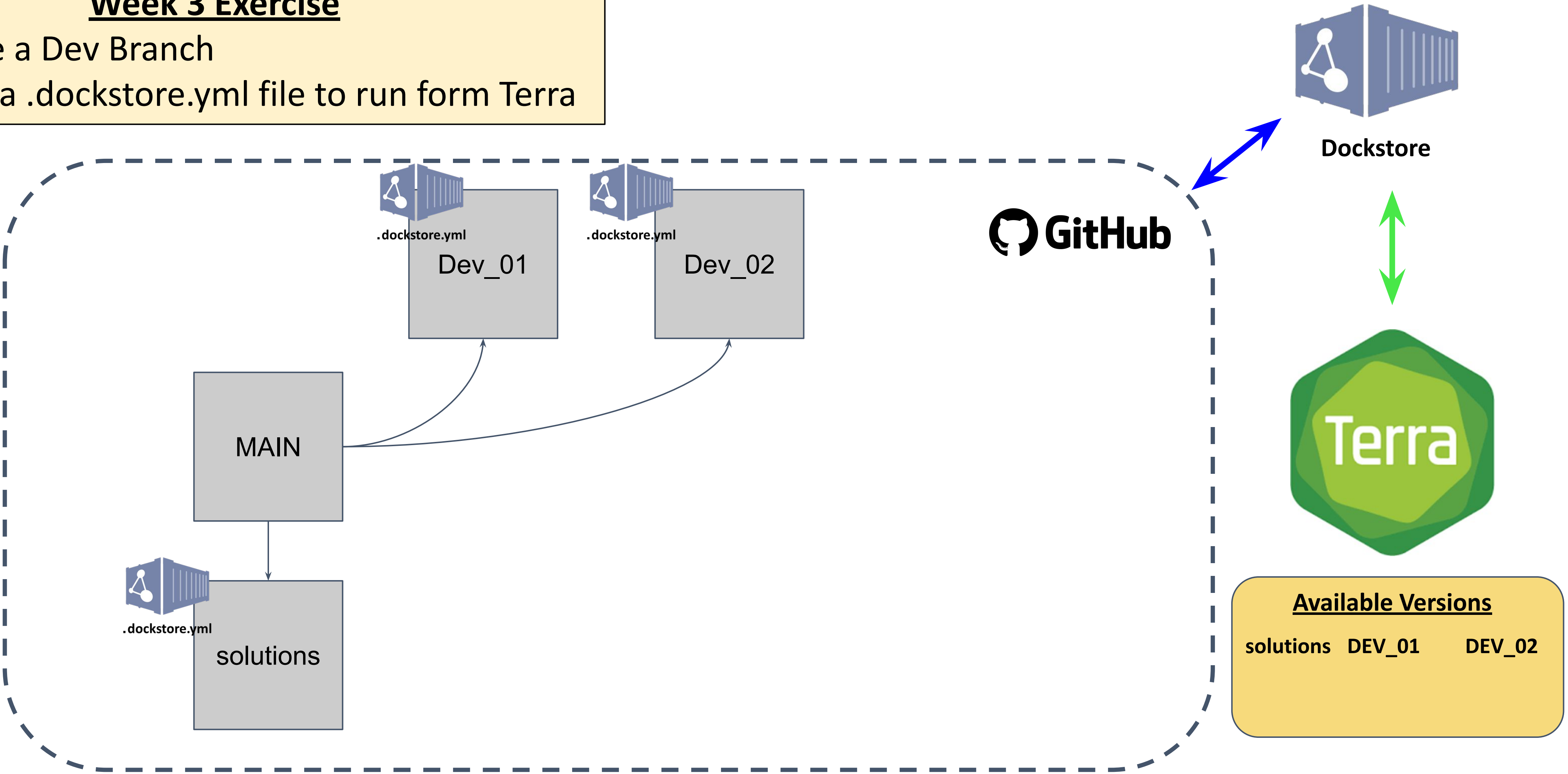
Step 2

No data selected

Terra Workflow Versions

Week 3 Exercise

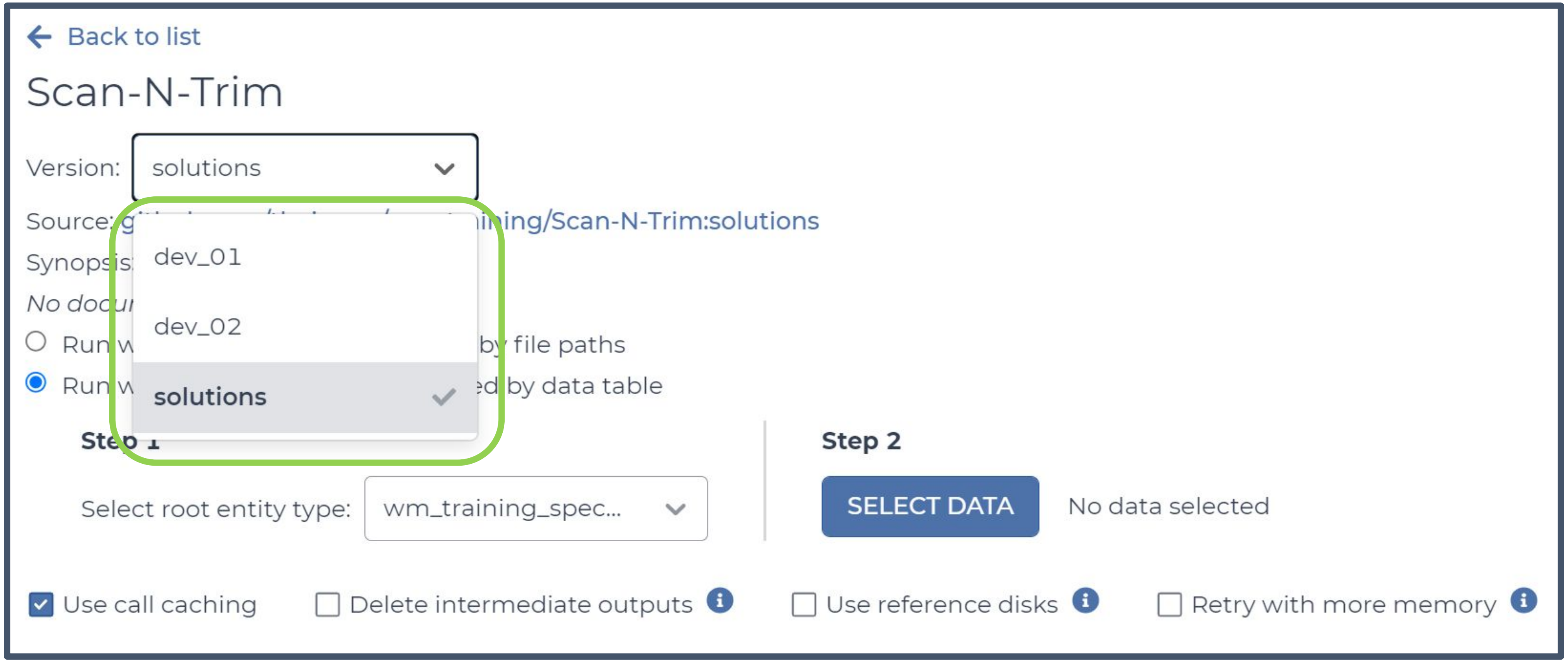
1. Create a Dev Branch
2. Write a .dockstore.yml file to run form Terra



Available Versions

solutions DEV_01 DEV_02

From Terra Workspace



← Back to list

Scan-N-Trim

Version: solutions

Source: g ... ining/Scan-N-Trim:solutions

Synopsis:

No docu

○ Run w ... by file paths

● Run w ... ed by data table

dev_01

dev_02

solutions ✓

Step 1

Select root entity type: wm_training_spec...

Step 2

No data selected

☑ Use call caching □ Delete intermediate outputs □ Use reference disks □ Retry with more memory

Workflow Management Training

10m to begin Exercise Part 2

We will regroup again at 10:58AM (PT)