

Galaxy and software containers: a recipe for success

Enis Afgan
Research Scientist at Johns Hopkins University

March 11, 2022

NCI's Containers and Workflows Interest Group Webinar Series

Let's follow Kat, a wet lab biologist from the 2nd floor

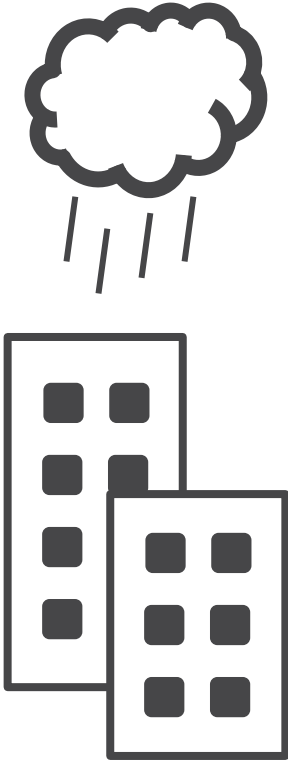
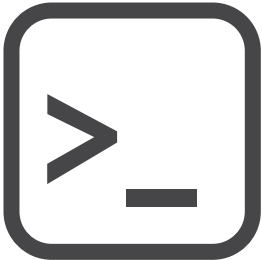
Goal: provide quality assurance reports for a patient's precision cancer therapy

Process:

- Tune and run an internal lab pipeline for processing patient's RNA-Seq data
- Use protected TCGA data to compute gene expression signatures
- Compare the computed signature to patient's data
- Produce interactive and PDF reports for dissemination
- Version the analysis workflow for approval process
- Scale the analysis to a growing number of samples

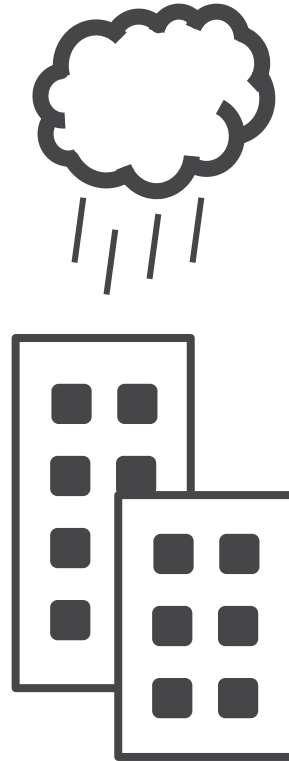


Project plan



GPwhat?

Let's lose the command prompt



GPwhat?

Galaxy?

T
O
O
L
P
A
N
E
L

The screenshot shows the Galaxy web interface with the 'Filter Tabular' tool form open. The interface includes a top navigation bar with 'Galaxy', 'Workflow', 'Visualize', 'Shared Data', 'Help', 'User', and 'Using 351.1 GB'. The left sidebar contains a 'Tools' panel with a search bar and a list of tool categories: 'Get Data', 'Collection Operations', 'GENERAL TEXT TOOLS', 'Text Manipulation', 'Filter Tabular', 'Query Tabular using sqlite sql', 'Compute an expression on every row', 'SQLite to tabular for SQL query', 'annotateMyIDs', 'Concatenate multiple datasets', 'Column Regex Find And Replace', 'Regex Find And Replace', 'Table Compute', 'Split file', 'Add input name as column', 'Text reformatting with awk', 'melt', 'Replace parts of text', 'cast', 'JQ process JSON', 'Rebase GFF3 features', 'Remove columns by heading', and 'Sort Column Order'. The main tool form is titled 'Filter Tabular (Galaxy Version 3.1.2)' and includes a 'Tabular Dataset to filter' dropdown, 'Filter Tabular Input Lines' section with an 'Execute' button, an 'Email notification' section, and a detailed description of the tool's function. Below the description is an 'Input Line Filters' section with a list of filter options and their descriptions. The right sidebar shows a 'History' panel with a list of recent datasets and their associated tools.

Tools

search tools

Upload Data

Get Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter Tabular

Query Tabular using sqlite sql

Compute an expression on every row

SQLite to tabular for SQL query

annotateMyIDs

Concatenate multiple datasets

Column Regex Find And Replace

Regex Find And Replace

Table Compute

Split file

Add input name as column

Text reformatting with awk

melt

Replace parts of text

cast

JQ process JSON

Rebase GFF3 features

Remove columns by heading

Sort Column Order

Input Line Filters

Filter Tabular (Galaxy Version 3.1.2)

Tabular Dataset to filter

Filter Tabular Input Lines

Email notification

Execute

Filter Tabular

Input Line Filters

Line Filtering Example

Input Tabular File:

History

Du Novo GTN tutorial

27: Filter on data 26

26: Variant Annotator on data 25

25: Naive Variant Caller (NVC) on data 24

24: BamLeftAlign on data 23 (alignments)

23: Map with BWA-MEM on data 22 and data 21 (mapped reads in BAM format)

22: Sequence Content Trimmer on data 13 and data 12

21: Sequence Content Trimmer on data 13 and data 12

20: Filter on data 19

19: Variant Annotator on data 18

18: Naive Variant Caller

H
I
S
T
O
R
Y

Tool form

```

positional arguments:
command
  clean          Remove unused packages and caches.
  compare       Compare packages between conda envirs
  config        Modify configuration values in .cond after the git config command. Writes file (Observes conda) by default
  create        Create a new conda environment from packages.
  help          Displays a list of available conda c strings.
  info          Display information about current co
  init          Initialize conda for shell interact
  install       Installs a list of packages into a s environment.
  list          Lists linked packages in a conda env
  package       Low-level conda package utility. (EX
  remove        Remove a list of packages from a spe
  uninstall     Alias for conda remove.
  run           Run an executable in a conda environ
  search        Search for packages and display asso
  update        Updates conda packages to the latest
  upgrade       Alias for conda update.

optional arguments:
  -h, --help  Show this help message and exit.
  -v, --version Show the conda version number and ex

```

```

1 <tool id="fastqc" name="FastQC" version="0.72">galaxy
2 <description>Read Quality reports</description>
3 <requirements>
4 <requirement type="package" version="0.11.8">
5 </requirements>
6 <stdio>
7 <exit_code range="1:" level="fatal" description
8 <exit_code range="-1:" level="fatal" description
9 <regex match="There is insufficient memory for the
10 source="stdout"
11 level="fatal" description="Out of memory error occurred"
12 </stdio>
13 <command>[[[DATA
14 #import re
15 #set input_name = re.sub('[^\w-\s]', '_', str
16
17 #if $input_file.ext.endswith('.gz'):
18 #set input_file_sl = $input_name + '.gz'
19 #elif $input_file.ext.endswith('.bz2'):
20

```

Split file according to the values of a column (Galaxy Version 0.2)

File to select

Column: 1

Email notification

Send an email notification when the job completes.

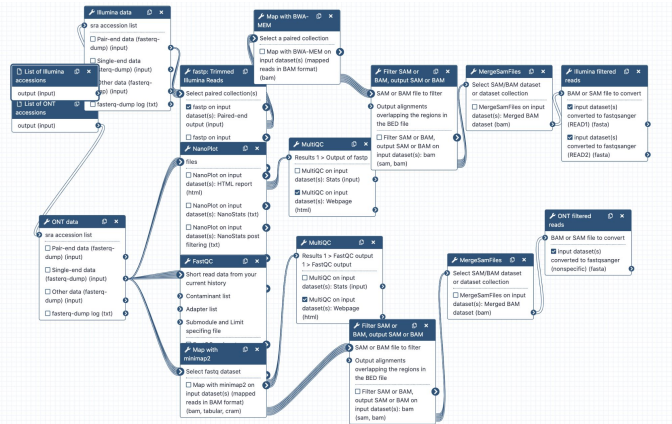
What it does

This tool splits a file into different smaller files using a specific column. It will own file.

Example

Splitting on column 5 from this:

>8,000 tools integrated



Graphical workflow editor

Build Rules for Uploading Collections

Use this form to describe rules for import datasets. At least one column should be defined to source to fetch data from (URLs, FTP files, etc...). Be sure to specify at least one column as a list identifier – specify more to created nested list structures. Specify a column to serve as "collection name" to group datasets into multiple collections.

Rules

- Filter out first 1 row(s)
- Add new column using L*(,*) applied to column D
- Remove column D
- Duplicate each row and split up columns
- Set column C as List Identifier(s)

	A	B	C (List Identifier)	D
PRJDB3920	SAMD00034150	DRX036147	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039919/DRR039919_1_1	
PRJDB3920	SAMD00034150	DRX036147	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039919/DRR039919_2_1	
PRJDB3920	SAMD00034150	DRX036148	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039920/DRR039920_1_1	
PRJDB3920	SAMD00034150	DRX036148	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039920/DRR039920_2_1	
PRJDB3920	SAMD00034150	DRX036149	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039921/DRR039921_1_1	
PRJDB3920	SAMD00034150	DRX036149	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039921/DRR039921_2_1	
PRJDB3920	SAMD00034150	DRX036150	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039922/DRR039922_1_1	
PRJDB3920	SAMD00034150	DRX036150	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039922/DRR039922_2_1	
PRJDB3920	SAMD00034150	DRX036151	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039923/DRR039923_1_1	
PRJDB3920	SAMD00034150	DRX036151	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039923/DRR039923_2_1	
PRJDB3920	SAMD00034153	DRX036152	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039924/DRR039924_1_1	
PRJDB3920	SAMD00034153	DRX036152	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039924/DRR039924_2_1	
PRJDB3920	SAMD00034152	DRX036164	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039936/DRR039936_1_1	
PRJDB3920	SAMD00034152	DRX036164	ftp.sra.ebi.ac.uk/vol1/fastq/DRR039/DRR039936/DRR039936_2_1	

Type: FastqSanger... Genome: unspecified (7)

Name: Enter a name for your new collection

Buttons: Re-orient, Cancel, Reset, Upload

Dataset collections

Share or Publish History `Du Novo GTN tutorial`

- Make History accessible
- Make History publicly available in Published Histories

This History is currently **accessible via link**. Anyone can view and import this History by visiting the following URL:

url: <https://usegalaxy.org/u/efafgan/h/imported-du-novo-gtn-tutorial>

Share History with Individual Users

You have not shared this History with any users.

Share with a user

Sharing & publishing

But what about Kat's project?

IWC: Intergalactic Workflow Commission

A community-driven effort to have:

A way to define and run workflow tests

A (central) repository to collect workflows

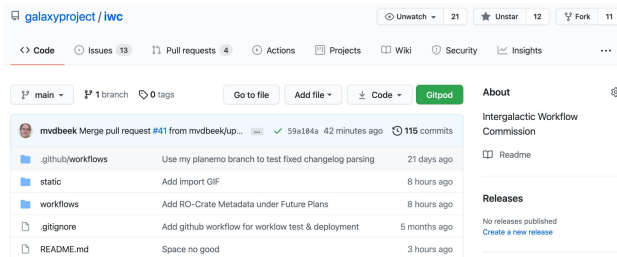
A way to define and run workflow tests

A way to define workflow versions

Conventions and standards for metadata

Workflows

IWC workflows



galaxyproject / iwc

Code Issues 13 Pull requests 4 Actions Projects Wiki Security Insights

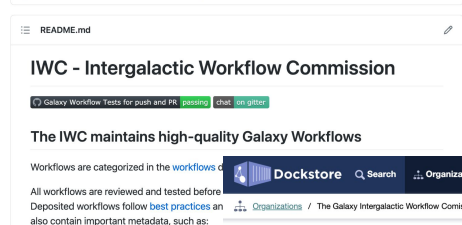
main 1 branch 0 tags Go to file Add file Code Gitpod About

Intergalactic Workflow Commission

Readme

Releases

Contributors

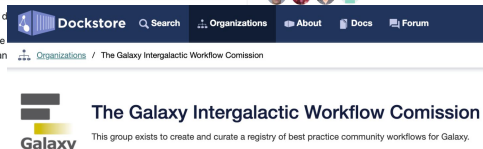


IWC - Intergalactic Workflow Commission

The IWC maintains high-quality Galaxy Workflows

Workflows are categorized in the [workflows](#) collection

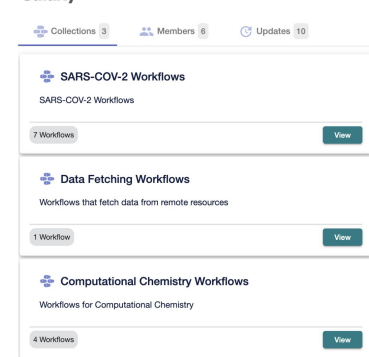
All workflows are reviewed and tested before deposited workflows follow [best practices](#) and also contain important metadata, such as:



The Galaxy Intergalactic Workflow Commission

This group exists to create and curate a registry of best practice community workflows for Galaxy.

Collections 3 Members 6 Updates 10



- SARS-COV-2 Workflows**
SARS-COV-2 Workflows
7 Workflows [View](#)
- Data Fetching Workflows**
Workflows that fetch data from remote resources
1 Workflow [View](#)
- Computational Chemistry Workflows**
Workflows for Computational Chemistry
4 Workflows [View](#)

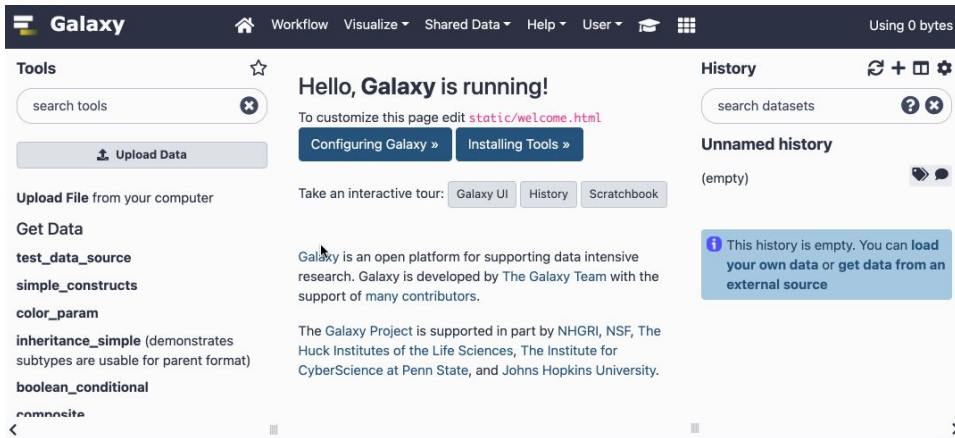
About the Organization

<https://github.com/galaxyproject/iwc>

IWC - Intergalactic Workflow Commission

Reproducibility is important. Our goals are to:

- foster workflow use
- incorporate versioning
- capture more metadata: (names, versions, authors, use cases, etc.)
- help scientists find workflows!



Galaxy

Workflow Visualize Shared Data Help User

Using 0 bytes

Tools

search tools

Upload Data

Upload File from your computer

Get Data

- test_data_source
- simple_constructs
- color_param
- inheritance_simple (demonstrates subtypes are usable for parent format)
- boolean_conditional
- composite

Hello, Galaxy is running!

To customize this page edit [static/welcome.html](#)

[Configuring Galaxy »](#) [Installing Tools »](#)

Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

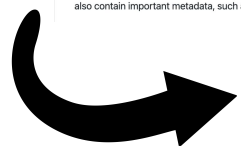
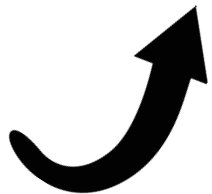
History

search datasets

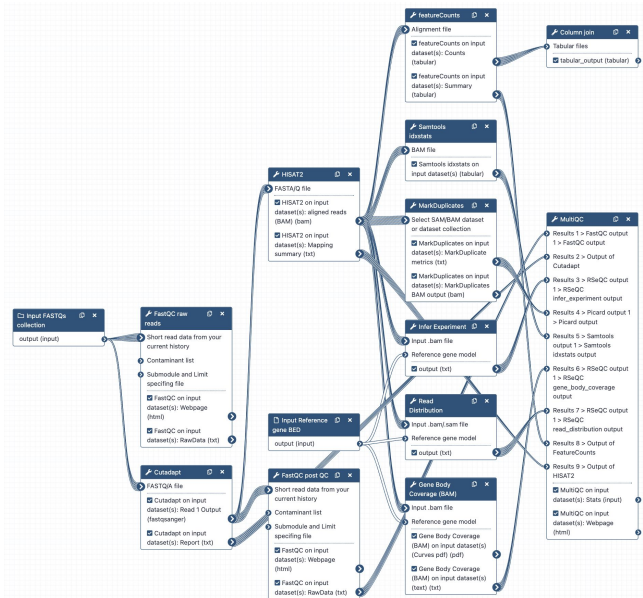
Unnamed history

(empty)

This history is empty. You can load your own data or get data from an external source



A tuned RNA-Seq workflow



Start from an existing template

Iterate and make desired adjustments

FASTQC Quality Control

Trimmomatic flexible read trimming tool for Illumina NGS data

MultiQC aggregate results from bioinformatics analyses into a single report

FASTQE visualize fastq files with emoji's 🤓

PRINSEQ to process quality of sequences

Draw quality score boxplot

FastQC Read Quality reports

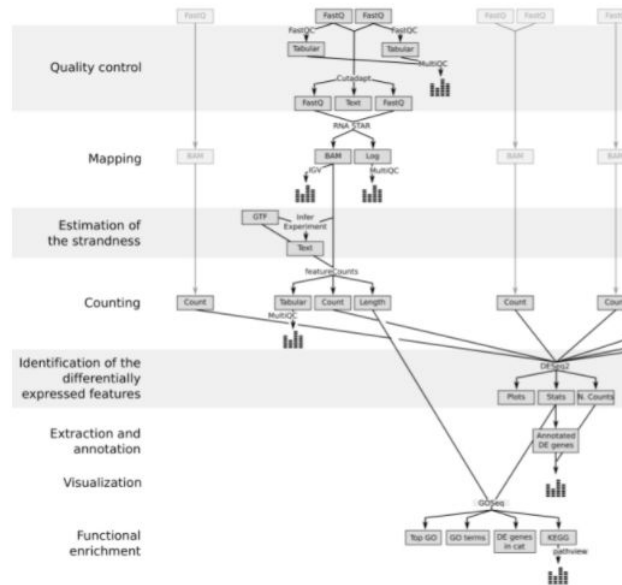
FASTQ Summary Statistics by column

Compute quality statistics

Draw nucleotides distribution chart

- Switch to 2.1.0+g
- Switch to 2.1.0+g
- Switch to 2.1.0+g
- Switch to 2.1.0
- Switch to 2.0.5.2
- Switch to 2.0.5.1
- Switch to 2.0.3
- Switch to 1.0.1
- Switch to 1.0.0

Finalized workflow



And a report

- Templated workflow invocation report
- Automatically generated for each invocation
- Interactive
- Coming soon: export as PDF

Title: de novo RNAseq (13NOV19.1)
Username: eafgan

Galaxy Report [🔗](#)

Created with Galaxy 22.01 on March 11, 2022, 4:25 PM

Identifier e83efa06d06970ca

Workflow Execution Summary of de novo RNAseq (13NOV19.1)

Workflow Inputs

Input Dataset: https://zenodo.org/record/583140/files/G1E_rep1_forward_read_%28SRR549355_1%29

```
# Workflow Execution Summary of de novo RNAseq (13NOV19.1)
```

```
FASTQ inputs files in `fastqsanger` format.
```

```
### Workflow Inputs
```

```
`` galaxy
invocation_inputs()
````
```

```
Quality control
```

```
We trim the reads to get rid of low quality bases at the read
```

Dataset: [https://zenodo.org/record/583140/files/G1E\\_rep1\\_forward\\_read\\_%28SRR549355\\_1%29](https://zenodo.org/record/583140/files/G1E_rep1_forward_read_%28SRR549355_1%29)

```
@SRR549355.100000091/1
TGGGTATCGGGTGGTGTGTGACGGCTGGGTGCCATGGTAGCTAGGCCCGGCGCAGGGACTGCCGCTTCCATCTTCATGCTCTCGGAGAGTGACA
+
CCCGFFFFHHHHICFGHIGIJJGJJJJIFHJIGIIIGGGIJJFIIJJHFFDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDCCDEDDDDDBDB-><4><CD>
@SRR549355.100000119/1
AGCGGTCGTTGAGCTCCGAGCTCTCTTTCTCTGACGGCGGACAGCGCGTGGGCTGAGCGCGTGCGGGGGGGGTGGCGCGCTGCCCGCC
+
BB@FFFFDGHGHJJJJJJJJJJJJJJJJJJJJJJIIJJJJIHFFDDDDDBDDDDDDDDDDDDDBDB@BDBDD@B#####
@SRR549355.100000232/1
TGGGAAGCTGGTGGCTCTCAGGAGAGAGGAGTCCGACTTCTGTTTGGTGACCTGGAGACTCTGGGCGCACTGTTGCTCAGGACAGGGCTGTCCCTT
+
@@@DDFFDHFDFH@GGGHIGIGHCF;E@CFHGIIJ6D@DHGHBFHGCHIJI@CEHCEE->?DEFD>>C? :?BDC@>C>@C>?3AA??<?A58ACCC:
@SRR549355.100000329/1
GGTTGGCAGGCATGGCTCCATGGCTGGGTGTGAGATCTGGAGGTGAGCGGGGACAGCGCGTCCAGCGGGCTGGCGGGCCCGGAGAGCAGGGAG
+
CCBFFFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJHIGHJJJJJJJJIIJJJJJJJJHFDSSDDDBDB@BDDDCDDDDDDDDDBDDDDDDDDDDDD<
@SRR549355.100000548/1
TGAGAGTAAGAGACACAGGAGAGGCCGGTACGGCGGGGCCCGCGGCGGCGTTCCAGGGCGCGGGGTACCACGGGAGGGGGGGGGGCGAGGGAG
```

# So why Galaxy?

# 1

## Versatility

- Many popular tools and visualizations available
- Built-in graphical workflow editor
- Expansive sharing and versioning

# 2

## Accessibility

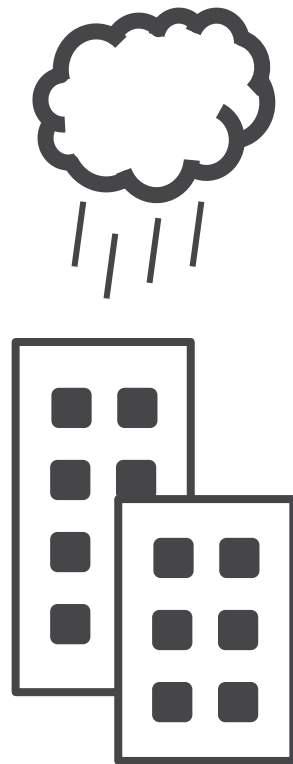
- Free managed services available for exploration
- Scalable cloud deployments
- Integration with data commons and datasets

# 3

## Vibrant community

- Training and help available
- Current with ongoing research topics
- Vetted workflows
- Trusted, with >10,000 citations

# Working with private data



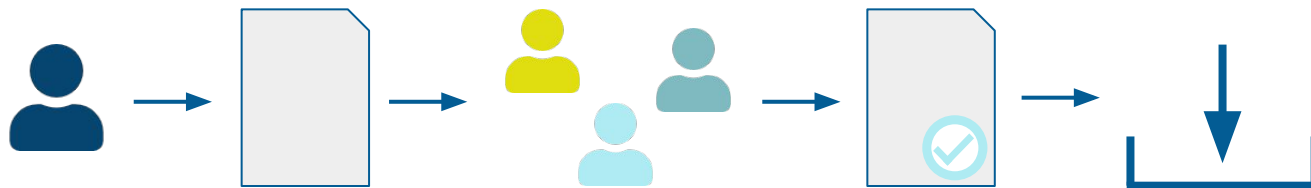
**GPwhat?**

# Combining private & protected data

- Kat has some local patient data that is regulated by the hospital and/or patient consent agreements
- Protected TCGA data lives in a data commons

In what kind of environment can this data be combined and analyzed?

# Working with sensitive data



The **recipient institution is ultimately responsible** for maintaining the confidentiality, integrity, and availability of the data.

# Working with sensitive data



The single most important element for maintaining the security of controlled access data is to **design security into the chosen environment.**





# Designing the necessary security perimeter

1.

**NIST**  
National Institute of  
Standards and Technology

---

2.

**NIST 800-53 Compliance Best Practices**

**Discover and Classify**  
Sensitive Data

**Map Data and Permissions**

**Manage**  
Access Control

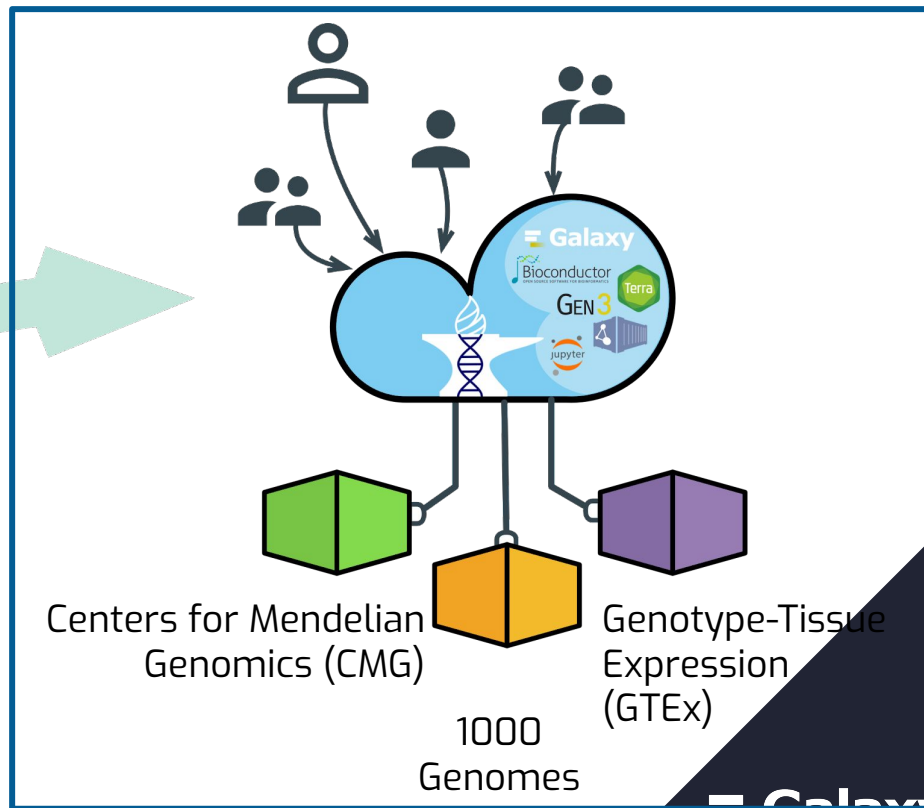
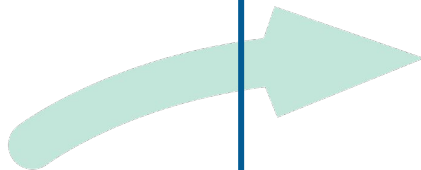
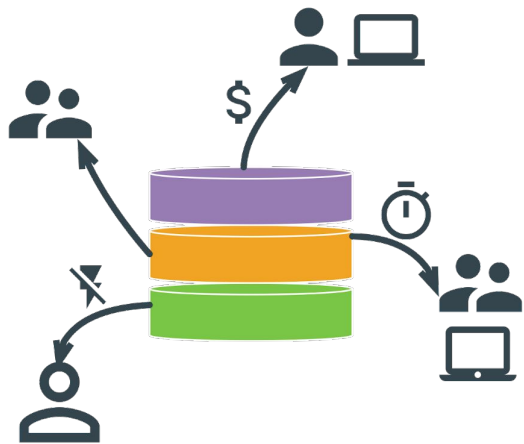
**Monitor Data,**  
File Activity, and  
User Behavior



3.

Conclusion: out of reach for individual investigators and (most) labs.

# NHGRI AnVIL overview

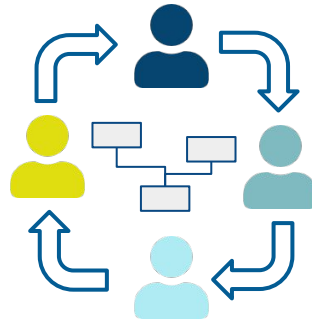


# Architecting Galaxy for protected datasets

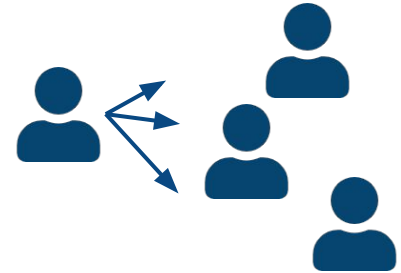
## Accessibility



## Reproducibility

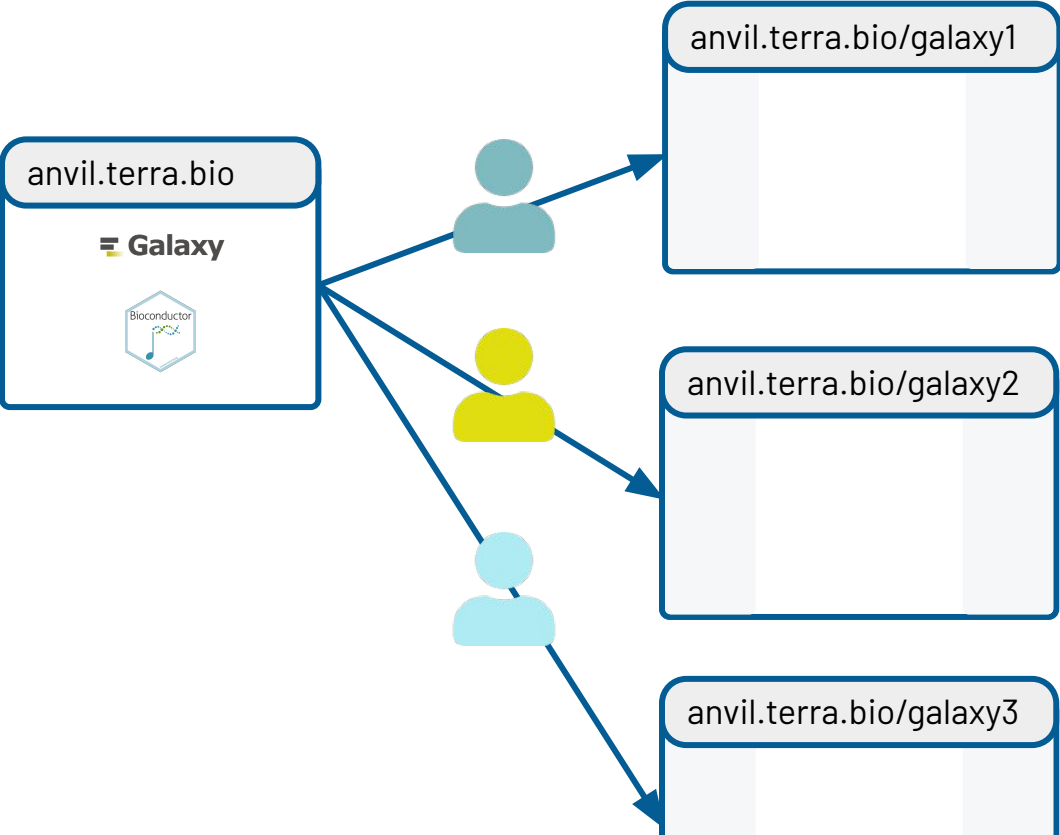


## Sharing



How do we maintain, or replicate, these capabilities when working with protected data?

# Replicating accessibility



kubernetes

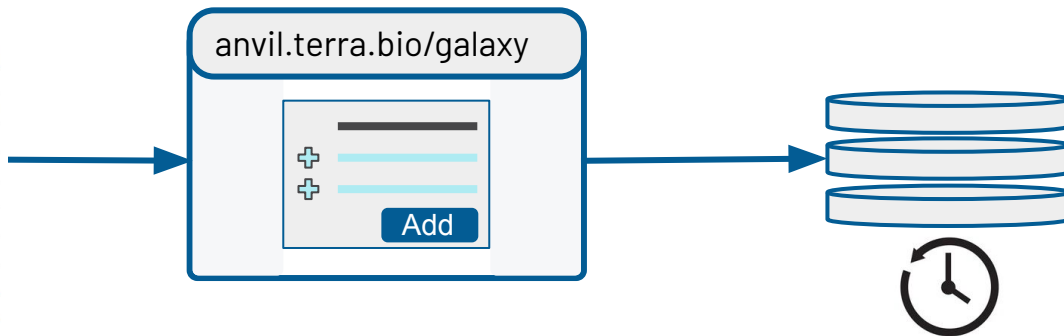


# Replicating reproducibility

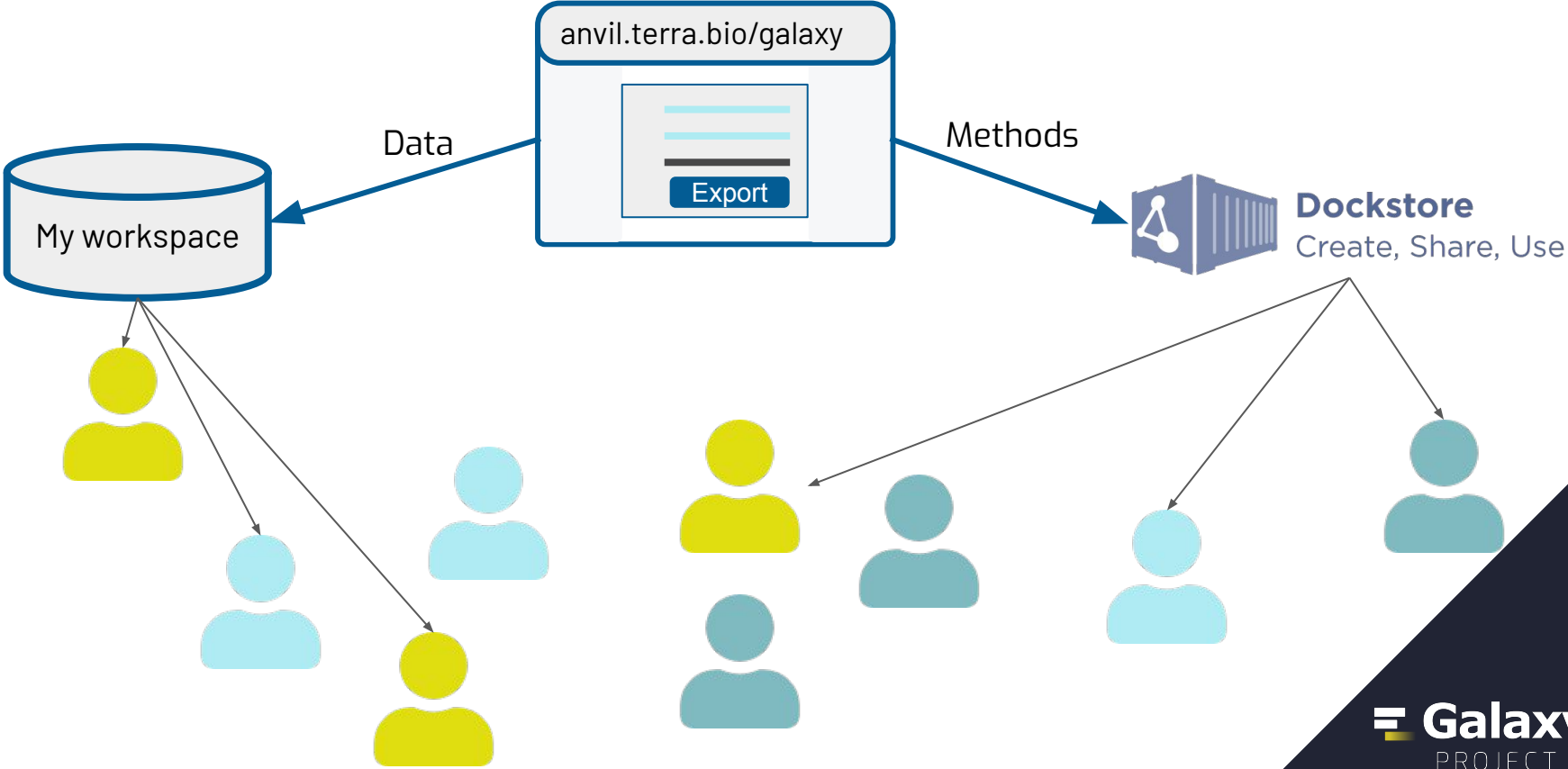
The history panel tracks complete provenance of a dataset but what if the instances are transient?

## Search Summary

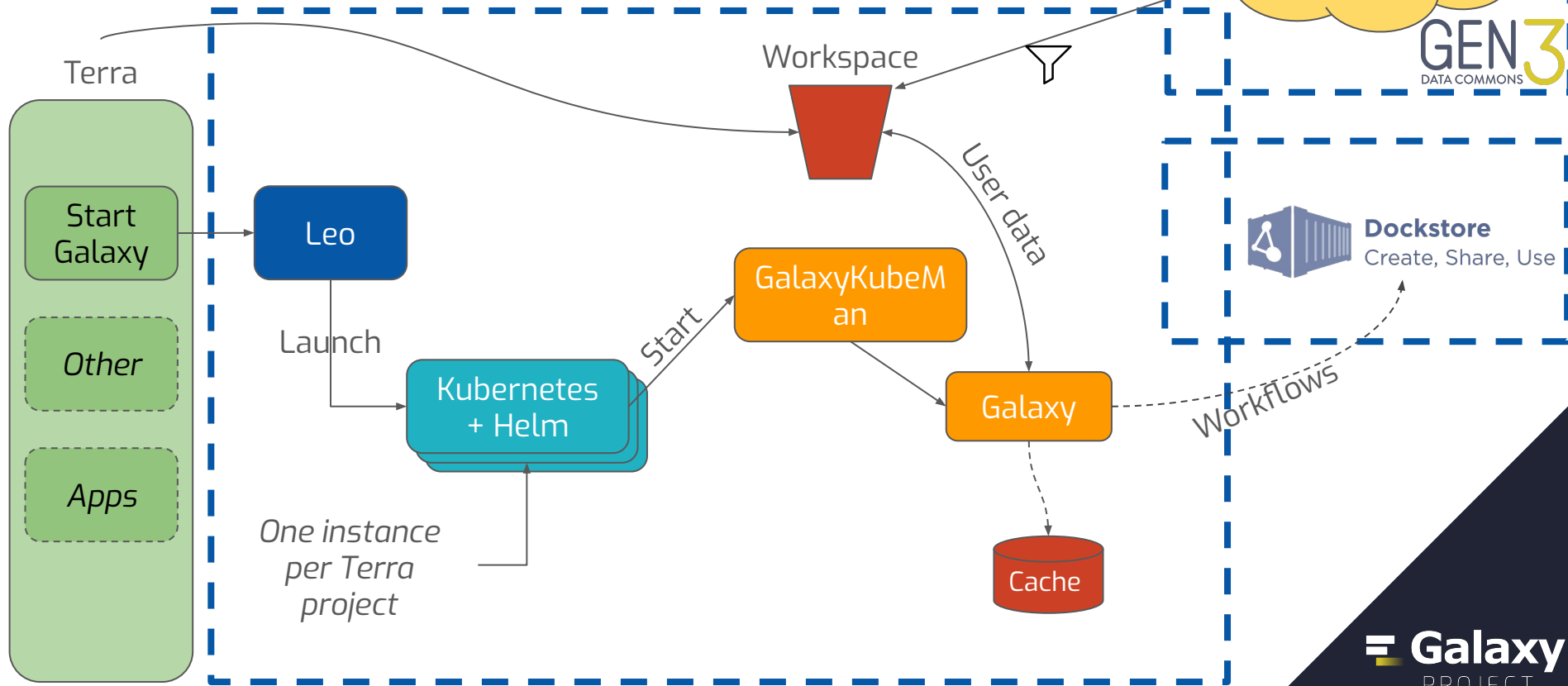
| Consortium       | Cohorts | Samples | Subjects | Size (TB) |
|------------------|---------|---------|----------|-----------|
| 1000 Genomes     | 1       | 3,202   | 3,202    | 73.00     |
| CCDG             | 199     | 269,875 | 256,426  | 2,582.41  |
| CMG              | 36      | 10,268  | 10,063   | 76.05     |
| Convergent Neuro | 2       | 304     | 304      | 5.32      |
| GTEx (v8)        | 1       | 17,382  | 979      | 182.00    |
| HPRC             | 1       | 57      | 47       | 149.00    |
| PAGE             | 4       | 690     | 690      | 16.99     |
| WGSPD1           | 5       | 1,504   | 9,943    | 176.85    |
| Totals           | 249     | 303,282 | 281,654  | 3,261.62  |



# Replicating sharing



# AnVIL system architecture



# Galaxy & Terra

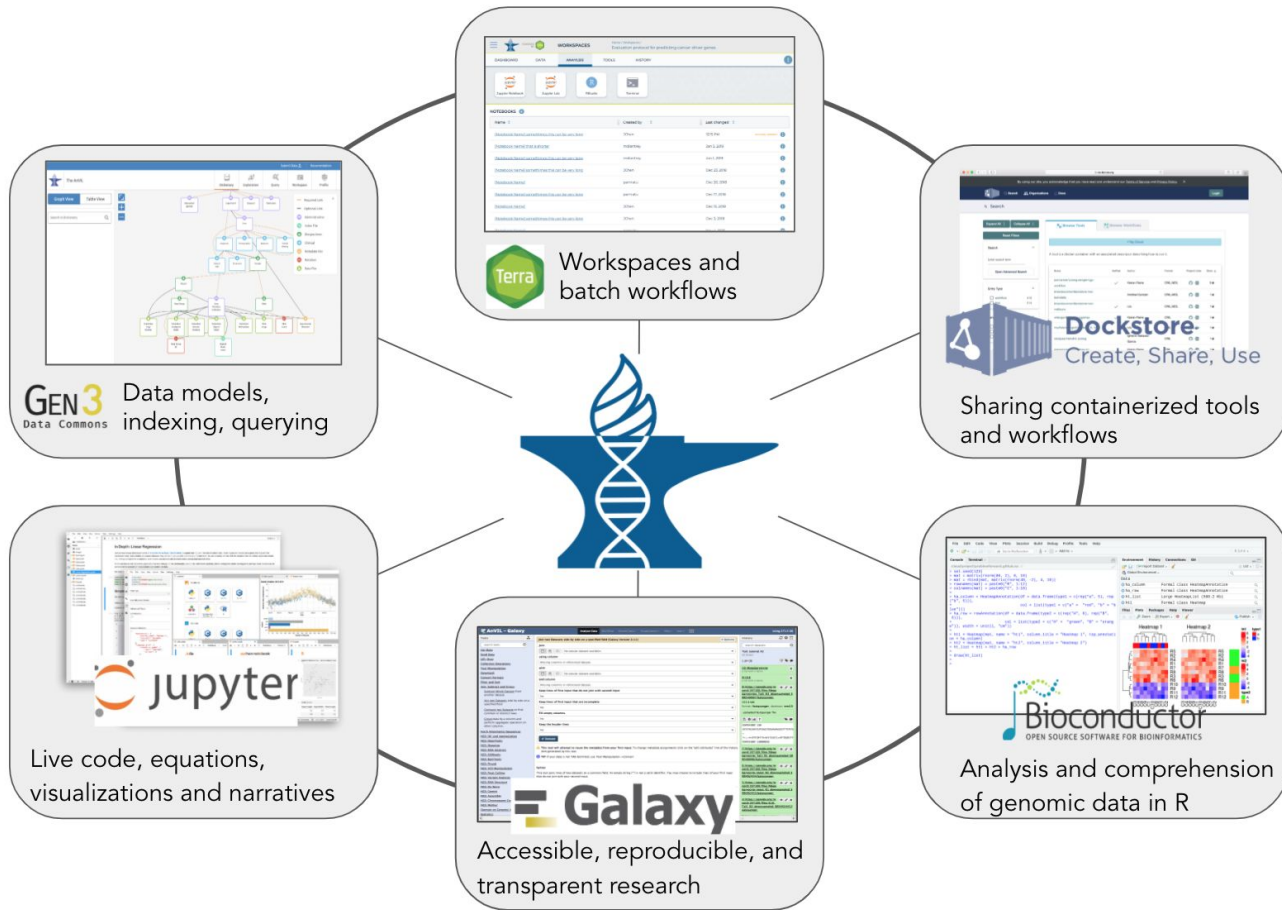
Galaxy integration with Terra reaches beyond AnVIL

Multiple FireCloud-backed Terra instances now offer Galaxy

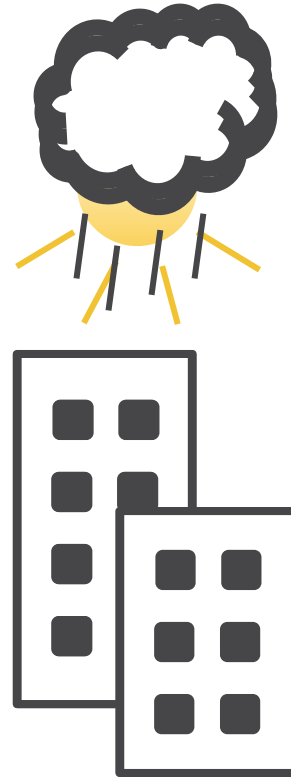
- <https://anvil.terra.bio>
- <https://app.terra.bio>
- <https://firecloud.terra.bio>
- <https://terra.biodatacatalyst.nhlbi.nih.gov>
- <https://workbench.researchallofus.org>



# AnVIL is a suite of applications



# Approachable cloud computing

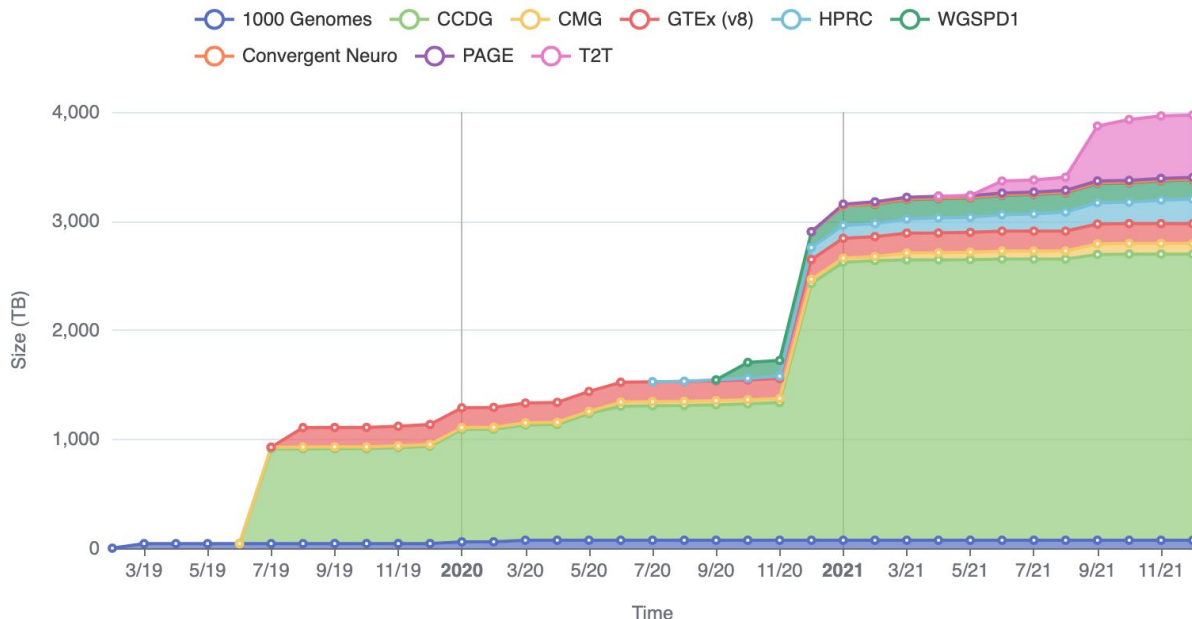


**GPwhat?**

# Useful data is often on multiple commons

- Kat is comparing genetic signatures in tumors and in normal cells
- This requires datasets that have matching samples of DNA and RNA data
- TCGA focused on the tumor samples with the number of normal samples in the TCGA dataset being fairly small
- The GTEx dataset has many more normal samples for the same tissue
- Combine the two to better understand gene expression activity

# Cloud data available on AnVIL



## Current Consortia

- CCDG
- CMG
- GTEx
- 1000 Genomes
- eMERGE
- PAGE
- T2T/HPRC

## Planned Consortia

- GREGOR
- PRIMED
- IGVF
- Covid19hg
- CSER
- NIA, NIMH, UDN

Almost 4 petabytes, 300,000 genomes and growing.  
Currently engaged with over 20 consortia!

# One collection of data is good. More is better.



*How do we go about multiplying the benefit of these now-accessible data?*

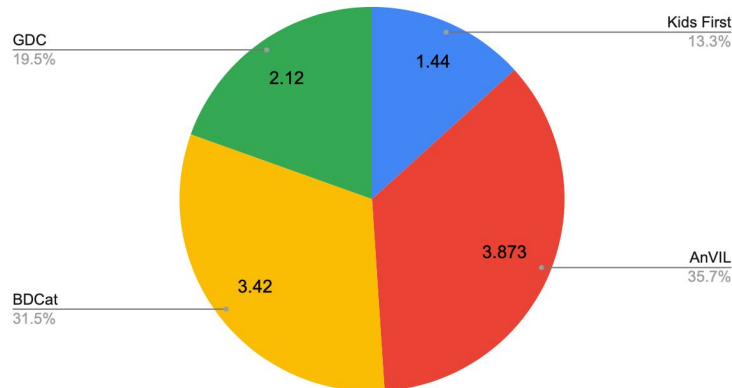


- If we are successful, we will catalyze the creation of an open and federated data ecosystem.
  - Others have done it before (SWIFT, the internet, the web).
- If we fail, we will degenerate into a collection of monolithic data silos
  - Others have done this before too (medical records in US hospitals)...<sup>11</sup>

# NIH Cloud Platform Interoperability Effort (NCPI)



Data Size (PB)



Researcher Auth Service



Data Repository Service



Fast Healthcare Interoperability Resources

11Pb / 689k participants and growing!

Cross-platform accessibility through several key technologies

# Ongoing NCPI efforts

Data discovery

- <https://anvilproject.org/ncpi/data>

Hand-off of search results from portals to workspace environments

- <https://youtu.be/YGZTxDdaWqk>

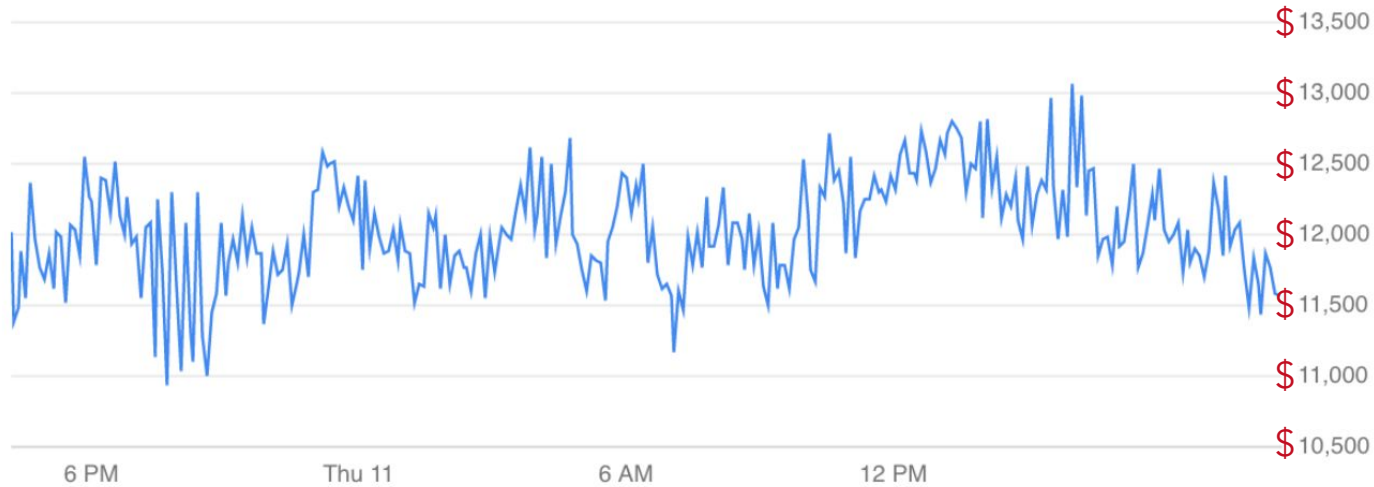
Single Sign-On with NIH RAS

Join a working group! <https://anvilproject.org/ncpi#working-groups>

# T2T Analysis on Google Cloud Platform

Preview

1 hour 4 hours 1 day



dollars/hour



# Cloud Costs are complicated

## E2 standard machine types

The following table shows the calculated cost for standard predefined machine types in the E2 machine family. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

Standard machine types have 4 GB of memory per vCPU.

| Iowa (us-central1) <span>Monthly</span> <input checked="" type="radio"/> <span>Hourly</span> |                                                                                                                                                                                                      |        |             |                         |
|----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|-------------|-------------------------|
| Machine type                                                                                 | Virtual CPUs                                                                                                                                                                                         | Memory | Price (USD) | Preemptible price (USD) |
| e2-standard-2                                                                                | 2                                                                                                                                                                                                    | 8GB    | \$0.067006  | \$0.020102              |
| e2-standard-4                                                                                | 4                                                                                                                                                                                                    | 16GB   | \$0.134012  | \$0.040204              |
| e2-standard-8                                                                                | 8                                                                                                                                                                                                    | 32GB   | \$0.268024  | \$0.080408              |
| e2-standard-16                                                                               | 16                                                                                                                                                                                                   | 64GB   | \$0.536048  | \$0.160816              |
| e2-standard-32                                                                               | 32                                                                                                                                                                                                   | 128GB  | \$1.072096  | \$0.321632              |
| Custom machine type                                                                          | If your ideal machine shape is in between two predefined types, using a custom E2 machine type could save you as much as 40%. For more information, see <a href="#">E2 custom vCPUs and memory</a> . |        |             |                         |

## N2 standard machine types

The following table shows the calculated costs for standard predefined machine types in the N2 machine family. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

Standard machine types have 4 GB of memory per vCPU.

| Iowa (us-central1) <span>Monthly</span> <input checked="" type="radio"/> <span>Hourly</span> |                                                                                                                                                                                               |        |             |                         |
|----------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|-------------|-------------------------|
| Machine type                                                                                 | Virtual CPUs                                                                                                                                                                                  | Memory | Price (USD) | Preemptible price (USD) |
| n2-standard-2                                                                                | 2                                                                                                                                                                                             | 8GB    | \$0.097118  | \$0.02354               |
| n2-standard-4                                                                                | 4                                                                                                                                                                                             | 16GB   | \$0.194236  | \$0.04708               |
| n2-standard-8                                                                                | 8                                                                                                                                                                                             | 32GB   | \$0.388472  | \$0.09416               |
| n2-standard-16                                                                               | 16                                                                                                                                                                                            | 64GB   | \$0.776944  | \$0.18832               |
| n2-standard-32                                                                               | 32                                                                                                                                                                                            | 128GB  | \$1.553888  | \$0.37664               |
| n2-standard-48                                                                               | 48                                                                                                                                                                                            | 192GB  | \$2.330832  | \$0.56496               |
| n2-standard-64                                                                               | 64                                                                                                                                                                                            | 256GB  | \$3.107776  | \$0.75328               |
| n2-standard-80                                                                               | 80                                                                                                                                                                                            | 320GB  | \$3.88472   | \$0.9416                |
| Custom machine type                                                                          | If your ideal machine shape is in between two predefined types, using a custom machine type could save you as much as 40%. For more information, see <a href="#">Custom vCPU and memory</a> . |        |             |                         |

## E2 high-memory machine types

The following table shows the calculated cost for the E2 high-memory predefined machine types. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

High-memory machine types have 8 GB of memory per vCPU. High-memory instances are ideal for tasks that require more memory relative to virtual CPUs.

| Iowa (us-central1) <span>Monthly</span> <input checked="" type="radio"/> <span>Hourly</span> |                                                                                                                                                                                                      |        |             |                         |
|----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|-------------|-------------------------|
| Machine type                                                                                 | Virtual CPUs                                                                                                                                                                                         | Memory | Price (USD) | Preemptible price (USD) |
| e2-highmem-2                                                                                 | 2                                                                                                                                                                                                    | 16GB   | \$0.09039   | \$0.027118              |
| e2-highmem-4                                                                                 | 4                                                                                                                                                                                                    | 32GB   | \$0.18078   | \$0.054236              |
| e2-highmem-8                                                                                 | 8                                                                                                                                                                                                    | 64GB   | \$0.36156   | \$0.108472              |
| e2-highmem-16                                                                                | 16                                                                                                                                                                                                   | 128GB  | \$0.72312   | \$0.216944              |
| Custom machine type                                                                          | If your ideal machine shape is in between two predefined types, using a custom E2 machine type could save you as much as 40%. For more information, see <a href="#">E2 custom vCPUs and memory</a> . |        |             |                         |

## N2 high-memory machine types

The following table shows the calculated cost for the N2 high-memory predefined machine types. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

High-memory machine types have 8 GB of memory per vCPU. High-memory instances are ideal for tasks that require more memory relative to virtual CPUs.

| Iowa (us-central1) <span>Monthly</span> <input checked="" type="radio"/> <span>Hourly</span> |                                                                                                                                                                                               |        |             |                         |
|----------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|-------------|-------------------------|
| Machine type                                                                                 | Virtual CPUs                                                                                                                                                                                  | Memory | Price (USD) | Preemptible price (USD) |
| n2-highmem-2                                                                                 | 2                                                                                                                                                                                             | 16GB   | \$0.131014  | \$0.03178               |
| n2-highmem-4                                                                                 | 4                                                                                                                                                                                             | 32GB   | \$0.262028  | \$0.06356               |
| n2-highmem-8                                                                                 | 8                                                                                                                                                                                             | 64GB   | \$0.524056  | \$0.12712               |
| n2-highmem-16                                                                                | 16                                                                                                                                                                                            | 128GB  | \$1.048112  | \$0.25424               |
| n2-highmem-32                                                                                | 32                                                                                                                                                                                            | 256GB  | \$2.096224  | \$0.50848               |
| n2-highmem-48                                                                                | 48                                                                                                                                                                                            | 384GB  | \$3.144336  | \$0.76272               |
| n2-highmem-64                                                                                | 64                                                                                                                                                                                            | 512GB  | \$4.192448  | \$1.01696               |
| n2-highmem-80                                                                                | 80                                                                                                                                                                                            | 640GB  | \$5.24056   | \$1.2712                |
| Custom machine type                                                                          | If your ideal machine shape is in between two predefined types, using a custom machine type could save you as much as 40%. For more information, see <a href="#">Custom vCPU and memory</a> . |        |             |                         |

## E2 high-CPU machine types

The following table shows the calculated cost for E2 high-CPU predefined machine types. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

High-CPU machine types have one vCPU for every 1 GB of memory. High-CPU machine types are ideal for tasks that require moderate memory configurations for the needed vCPU count.

| Iowa (us-central1) <span>Monthly</span> <input checked="" type="radio"/> <span>Hourly</span> |                                                                                                                                                                                                      |        |             |                         |
|----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|-------------|-------------------------|
| Machine type                                                                                 | Virtual CPUs                                                                                                                                                                                         | Memory | Price (USD) | Preemptible price (USD) |
| e2-highcpu-2                                                                                 | 2                                                                                                                                                                                                    | 2GB    | \$0.049468  | \$0.01484               |
| e2-highcpu-4                                                                                 | 4                                                                                                                                                                                                    | 4GB    | \$0.098936  | \$0.02968               |
| e2-highcpu-8                                                                                 | 8                                                                                                                                                                                                    | 8GB    | \$0.197872  | \$0.05936               |
| e2-highcpu-16                                                                                | 16                                                                                                                                                                                                   | 16GB   | \$0.395744  | \$0.11872               |
| e2-highcpu-32                                                                                | 32                                                                                                                                                                                                   | 32GB   | \$0.791488  | \$0.23744               |
| Custom machine type                                                                          | If your ideal machine shape is in between two predefined types, using a custom E2 machine type could save you as much as 40%. For more information, see <a href="#">E2 custom vCPUs and memory</a> . |        |             |                         |

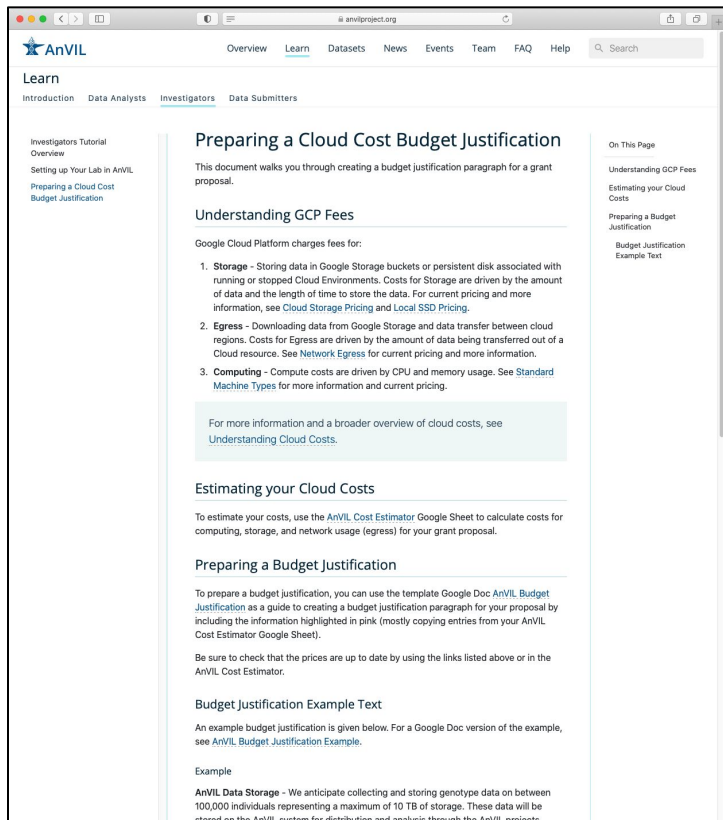
## N2 high-CPU machine types

The following table shows the calculated cost for N2 high-CPU predefined machine types. The vCPUs and memory from each of these machine types are billed by their individual [predefined vCPU and memory prices](#), but these tables provide the cost that you can expect using a specific machine type.

High-CPU machine types have one vCPU for every 1 GB of memory. High-CPU machine types are ideal for tasks that require moderate memory configurations for the needed vCPU count.

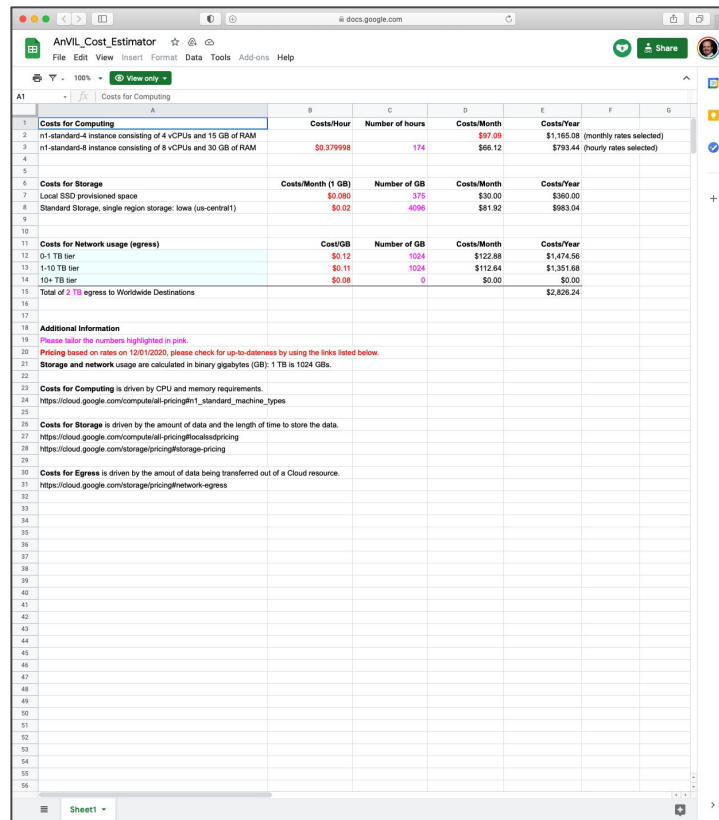
| Iowa (us-central1) <span>Monthly</span> <input checked="" type="radio"/> <span>Hourly</span> |                                                                                                                                                                                                |        |             |                         |
|----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|-------------|-------------------------|
| Machine type                                                                                 | Virtual CPUs                                                                                                                                                                                   | Memory | Price (USD) | Preemptible price (USD) |
| n2-highcpu-2                                                                                 | 2                                                                                                                                                                                              | 2GB    | \$0.071696  | \$0.01736               |
| n2-highcpu-4                                                                                 | 4                                                                                                                                                                                              | 4GB    | \$0.143392  | \$0.03472               |
| n2-highcpu-8                                                                                 | 8                                                                                                                                                                                              | 8GB    | \$0.286784  | \$0.06944               |
| n2-highcpu-16                                                                                | 16                                                                                                                                                                                             | 16GB   | \$0.573568  | \$0.13888               |
| n2-highcpu-32                                                                                | 32                                                                                                                                                                                             | 32GB   | \$1.147136  | \$0.27776               |
| n2-highcpu-48                                                                                | 48                                                                                                                                                                                             | 48GB   | \$1.720704  | \$0.41664               |
| n2-highcpu-64                                                                                | 64                                                                                                                                                                                             | 64GB   | \$2.294272  | \$0.55552               |
| n2-highcpu-80                                                                                | 80                                                                                                                                                                                             | 80GB   | \$2.86784   | \$0.6944                |
| Custom machine type                                                                          | If your ideal machine shape is in between two predefined types, using a custom machine type could save you as much as 40%. For more information, see <a href="#">Custom vCPUs and memory</a> . |        |             |                         |

# AnVIL Cloud Cost Budget Templates



The screenshot shows the AnVIL website's 'Learn' section, specifically the 'Investigators' page. The main heading is 'Preparing a Cloud Cost Budget Justification'. The page content includes:

- Understanding GCP Fees**: A section explaining that Google Cloud Platform charges fees for storage, egress, and computing. It lists three categories: Storage, Egress, and Computing, each with a brief description of what they cover and where to find pricing information.
- Estimating your Cloud Costs**: A section stating that users should use the AnVIL Cost Estimator Google Sheet to calculate costs for computing, storage, and network usage.
- Preparing a Budget Justification**: A section explaining that the template Google Doc is intended as a guide for creating budget justification paragraphs for proposals.
- Budget Justification Example Text**: A section providing an example of budget justification text, including a note about data storage: 'AnVIL Data Storage - We anticipate collecting and storing genotype data on between 100,000 individuals representing a maximum of 10 TB of storage. These data will be stored on the AnVIL system for distribution and analysis through the AnVIL projects.'



The screenshot shows the 'AnVIL\_Cost\_Estimator' Google Sheet. The table below summarizes the costs for computing, storage, and network usage.

|                                         | A                                                                                                                                                         | B                         | C                      | D                  | E                 | F                        | G |
|-----------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------|------------------------|--------------------|-------------------|--------------------------|---|
| <b>Costs for Computing</b>              |                                                                                                                                                           |                           |                        |                    |                   |                          |   |
| 1                                       |                                                                                                                                                           | <b>Costs/Hour</b>         | <b>Number of hours</b> | <b>Costs/Month</b> | <b>Costs/Year</b> |                          |   |
| 2                                       | n1-standard-4 instance consisting of 4 vCPUs and 15 GB of RAM                                                                                             | \$97.09                   |                        | \$97.09            | \$1,165.08        | (monthly rates selected) |   |
| 3                                       | n1-standard-8 instance consisting of 8 vCPUs and 30 GB of RAM                                                                                             | \$0.379998                | 174                    | \$66.12            | \$793.44          | (hourly rates selected)  |   |
| <b>Costs for Storage</b>                |                                                                                                                                                           |                           |                        |                    |                   |                          |   |
| 6                                       |                                                                                                                                                           | <b>Costs/Month (1 GB)</b> | <b>Number of GB</b>    | <b>Costs/Month</b> | <b>Costs/Year</b> |                          |   |
| 7                                       | Local SSD provisioned space                                                                                                                               | \$0.060                   | 375                    | \$30.00            | \$360.00          |                          |   |
| 8                                       | Standard Storage, single region storage; iowa (us-central1)                                                                                               | \$0.02                    | 4096                   | \$81.92            | \$983.04          |                          |   |
| <b>Costs for Network usage (egress)</b> |                                                                                                                                                           |                           |                        |                    |                   |                          |   |
| 11                                      |                                                                                                                                                           | <b>Cost/GB</b>            | <b>Number of GB</b>    | <b>Costs/Month</b> | <b>Costs/Year</b> |                          |   |
| 12                                      | 0-1 TB tier                                                                                                                                               | \$0.12                    | 1024                   | \$122.88           | \$1,474.56        |                          |   |
| 13                                      | 1-10 TB tier                                                                                                                                              | \$0.11                    | 1024                   | \$112.64           | \$1,351.68        |                          |   |
| 14                                      | 10+ TB tier                                                                                                                                               | \$0.08                    | 0                      | \$0.00             | \$0.00            |                          |   |
| 15                                      | <b>Total of 2 TB egress to Worldwide Destinations</b>                                                                                                     |                           |                        |                    | <b>\$2,826.24</b> |                          |   |
| <b>Additional Information</b>           |                                                                                                                                                           |                           |                        |                    |                   |                          |   |
| 18                                      | Please enter the numbers highlighted in pink.                                                                                                             |                           |                        |                    |                   |                          |   |
| 20                                      | Pricing based on rates on 12/01/2020, please check for up-to-dateness by using the links listed below.                                                    |                           |                        |                    |                   |                          |   |
| 21                                      | Storage and network usage are calculated in binary gigabytes (GB): 1 TB is 1024 GBs.                                                                      |                           |                        |                    |                   |                          |   |
| 23                                      | Costs for Computing is driven by CPU and memory requirements.                                                                                             |                           |                        |                    |                   |                          |   |
| 24                                      | <a href="https://cloud.google.com/compute/ai/pricing#n1_standard_machine_types">https://cloud.google.com/compute/ai/pricing#n1_standard_machine_types</a> |                           |                        |                    |                   |                          |   |
| 25                                      | Costs for Storage is driven by the amount of data and the length of time to store the data.                                                               |                           |                        |                    |                   |                          |   |
| 27                                      | <a href="https://cloud.google.com/compute/ai/pricing/locations/pricing">https://cloud.google.com/compute/ai/pricing/locations/pricing</a>                 |                           |                        |                    |                   |                          |   |
| 28                                      | <a href="https://cloud.google.com/storage/pricing/storage-pricing">https://cloud.google.com/storage/pricing/storage-pricing</a>                           |                           |                        |                    |                   |                          |   |
| 29                                      |                                                                                                                                                           |                           |                        |                    |                   |                          |   |
| 30                                      | Costs for Egress is driven by the amount of data being transferred out of a Cloud resource.                                                               |                           |                        |                    |                   |                          |   |
| 31                                      | <a href="https://cloud.google.com/storage/pricing/network-egress">https://cloud.google.com/storage/pricing/network-egress</a>                             |                           |                        |                    |                   |                          |   |

# Understanding costs



AnVIL project | anvilproject.org

RR Shared Published (unlisted) Feb 25 1 like

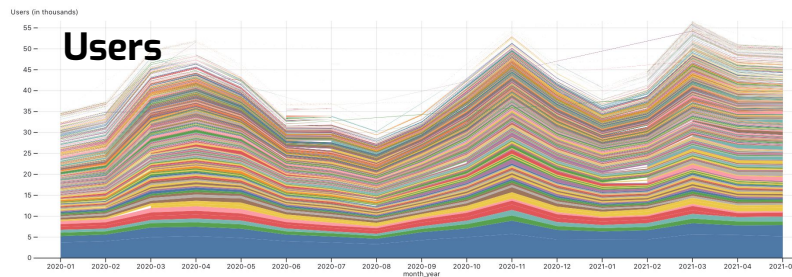
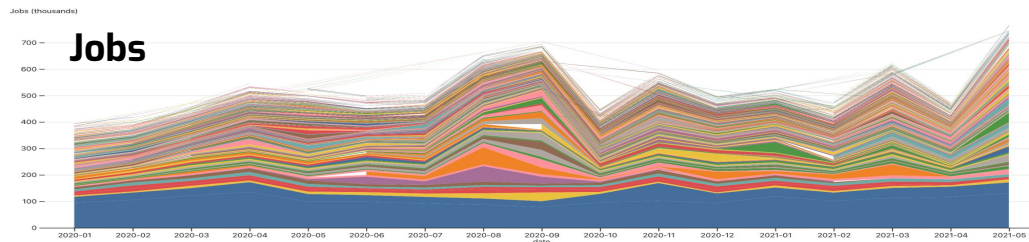
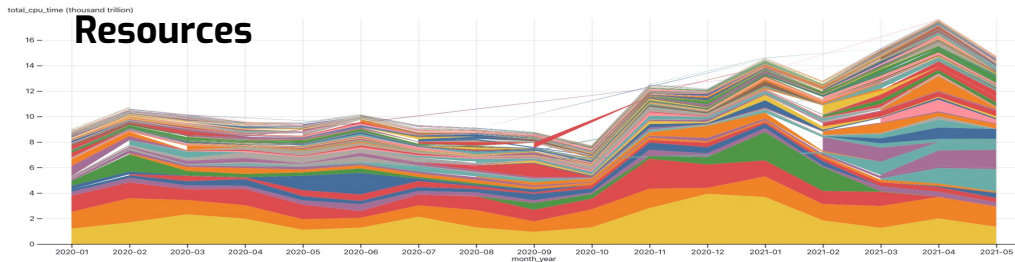
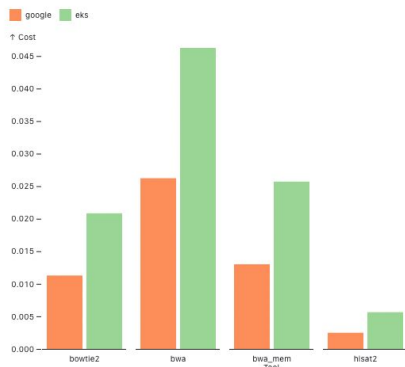
## Genomics tool benchmarking for cloud resource cost estimation

### ► Table of Contents

### How much will it cost?

Cloud resources offer great conveniences and capabilities but how much will it cost to run a given analysis or a tool? The following diagram offers a high-level representation of the anticipated workload costs as a function of data. The data is represented in terms of file size. If you are not sure about the size of your data, we can estimate it from the type and number of samples you are working with.

### Cost



**Stay tuned**

**<https://observablehq.com/@anvilproject/cost-estimation>**

# Helping with the AnVIL cloud costs

## AnVIL Cloud Credits Continued Program (AC3)

- Competitive program for supporting use of AnVIL
- Training track and research track
- Apply via:

<https://anvilproject.org/news/2021/11/22/announcing-anvil-cloud-credits-continued-program>

# NIH/ODSS STRIDES Initiative

U.S. Department of Health & Human Services | National Institutes of Health | Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI)

NIH National Institutes of Health  
Office of Data Science Strategy

Home Strategic Plan Resources Research Funding News & Events About

COVID-19

- Public health information from CDC
- Research information from NIH | Español
- NIH staff guidance on coronavirus (NIH Only)
- NIH and other federal agencies have made COVID-19 data available through several Open-Access Data and Computational Resources

STRIDES Initiative

About Cloud Preparing to Use the Cloud Partner Offerings Success Stories

## About the STRIDES Initiative

The **NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative** allows NIH to explore the use of cloud environments to streamline NIH data use by partnering with commercial providers. NIH's STRIDES Initiative aims to modernize biomedical research by reducing economic and process barriers in utilizing commercial cloud services. These partnerships enable access to rich datasets and advanced computational infrastructure, tools, and services.

The STRIDES Initiative is one of many NIH-wide efforts to implement the **NIH Strategic Plan for Data Science**, which provides a roadmap for modernizing the NIH-funded biomedical data science ecosystem. Data generated via biomedical research continues to outpace the ability to process, store, and analyze in many local environments.

By leveraging the STRIDES Initiative, NIH and NIH-funded institutions can begin to create a robust, interconnected ecosystem that breaks down silos related to generating, analyzing, and sharing research data. NIH-funded researchers with an active NIH award may take advantage of the STRIDES Initiative for their NIH-funded research projects. Eligible researchers include NIH intramural researchers and awardees of NIH contracts, other transaction agreements, grants, cooperative agreements, and other agreements.

Benefits of using the STRIDES Initiative include:

- Professional services**—Access to professional service consultations and technical support from the STRIDES Initiative partners.
- Training**—Access to training for researchers, data owners, and others to help ensure optimal use of available tools and technologies.
- Discounts on STRIDES Initiative partner services**—Favorable pricing on computing, storage, and related cloud services for NIH Institutes, Centers, and Offices (ICOs) and NIH-funded institutions.
- Potential collaborative engagements**—Opportunities to explore methods and approaches that may advance NIH's biomedical research objectives (with scope and milestones of engagements agreed upon separately).

## STRIDES Benefits

- **Discounts** (typically 10%-25%) on computing, storage, and related cloud services for NIH Institutes, Centers, and Offices (ICOs) and NIH-funded institutions & investigators.
- **Professional services** — Access to professional service consultations and technical support from the STRIDES Initiative partners.
- **Training** — Access to training for researchers, data owners, and others to help ensure optimal use of available tools and technologies.
- **Potential collaborative engagements** — Opportunities to explore methods and approaches that may advance NIH's biomedical research objectives

The first STRIDES Initiative partnership was established with **Google Cloud** in July 2018; a second partnership was established with **Amazon Web Services (AWS)** in September 2018. **Microsoft Azure** was established as a third partner in July 2021.

Read more

Get started with the STRIDES Initiative

# Funding your cloud experience

But ultimately - you gotta write them grants!

## Past solicitations

- Genomic Variation and Function Data and Administrative Coordinating Center ([RFA-HG-20-046](#))
- Polygenic Risk Score (PRS) Methods and Analysis for Populations of Diverse Ancestry ([RFA-HG-20-002](#))
- Mendelian Genomics Research Centers ([RFA-HG-20-007](#))

## Currently open

- Administrative Supplements to Support Enhancement of Software Tools for Open Science ([NOT-OD-22-068](#))
- CZI Essential Open Source Software for Science ([Cycle 5](#))

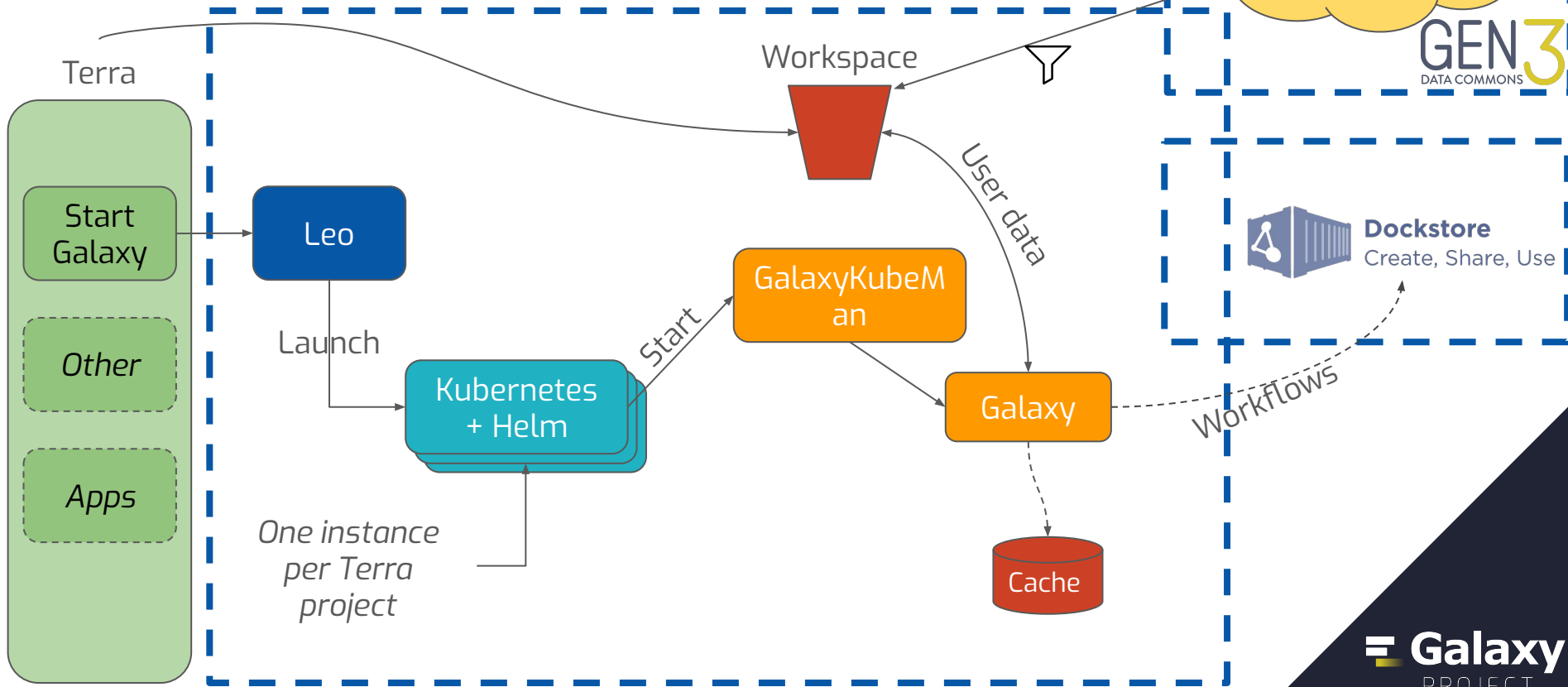
# How to design the cloud for scalability?



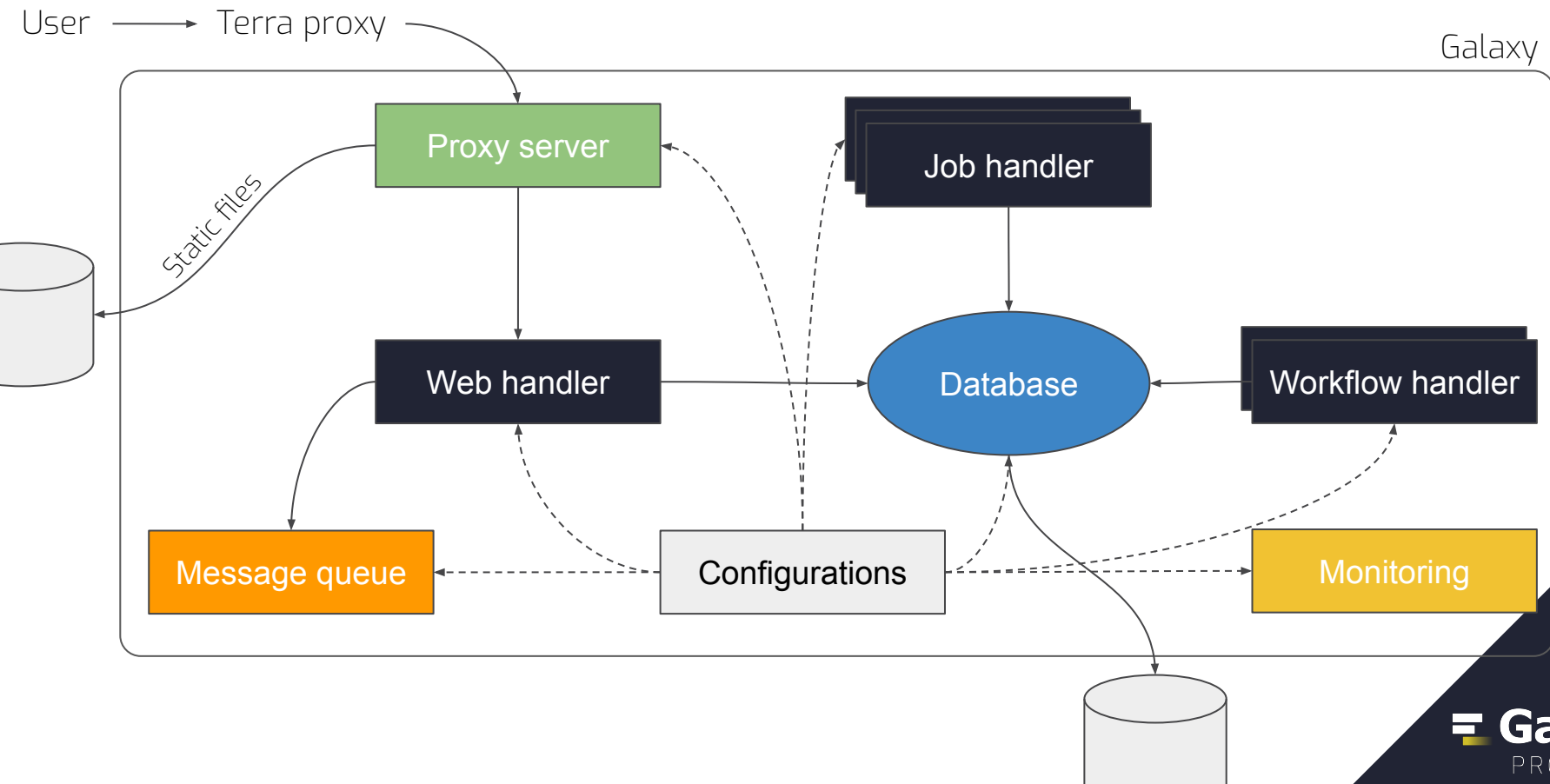
**GPWhat?**



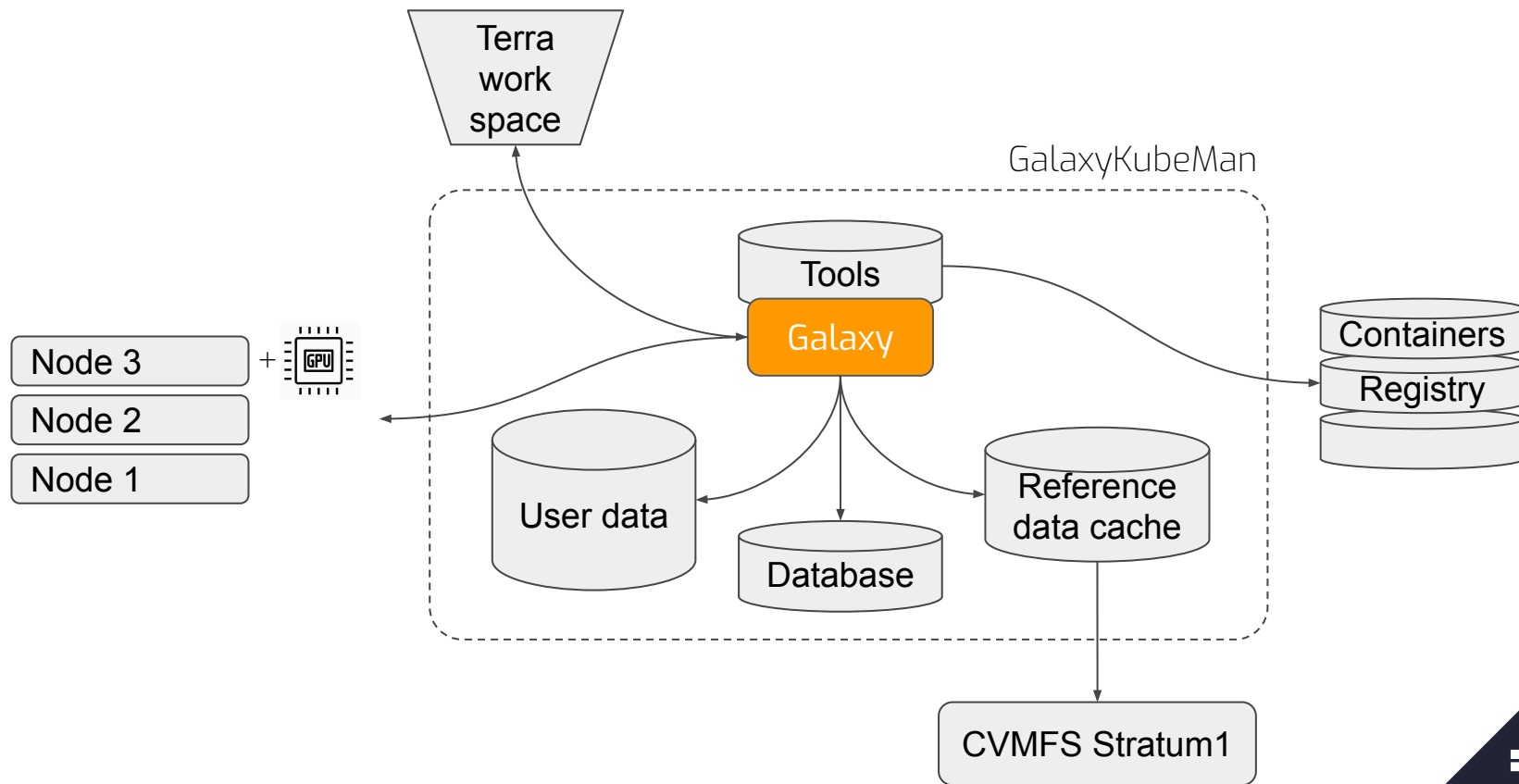
# Remember the AnVIL system architecture?



# Galaxy server components



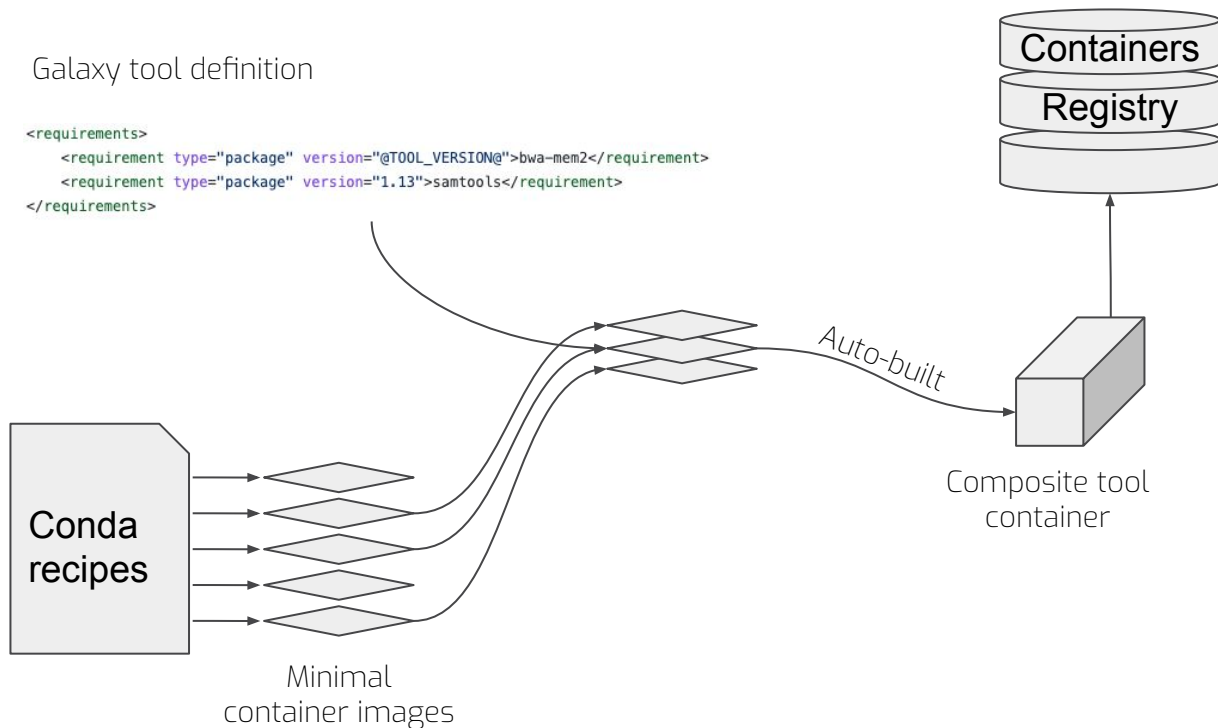
# Architecture for the overall Galaxy server



# Tools and Biocontainers

Galaxy tool definition

```
<requirements>
 <requirement type="package" version="@TOOL_VERSION@">bwa-mem2</requirement>
 <requirement type="package" version="1.13">samtools</requirement>
</requirements>
```

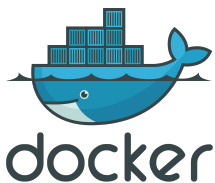
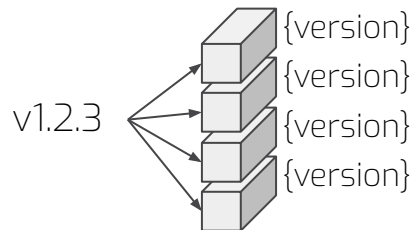
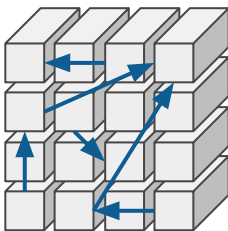
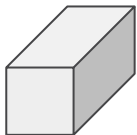


- Versioned tools, inc. dependencies
- Nearly 10,000 images hosted
- Usable outside the Galaxy ecosystem

# Change and service management

- Tracking configuration changes
  - Roll back to a working revision
  - Codify all values
- Server scaling
- Zero downtime upgrades
- Maintaining service uptime and robustness
  - Replication
  - Cattle vs. pet approach

# Tech enabling these deployments

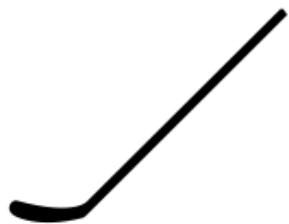


# Galaxy Helm chart

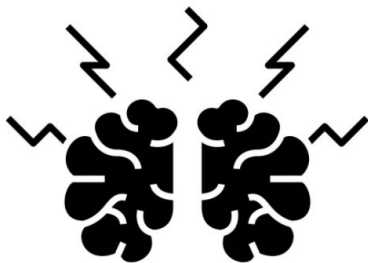
```
helm install galaxy
```

- All software components get deployed for a production Galaxy server
- Service management capabilities built in
- Support for codified change management

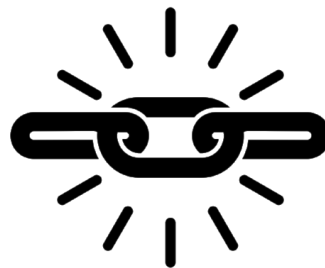
# Our experience with K8s



There is definitely a learning curve



Debugging is tough. No other way to say it.



Codified.  
Portable.  
Robust.



Powerful and very well designed.



# Final announcements

The Gallantries, Galaxy Training Network & Galaxy Community are happy to announce

More than 2,200  
people registered!

# GTN Smörgåsbord 2

# 14-18 March 2022

Save the date!

[bit.ly/smorgasbord2](https://bit.ly/smorgasbord2)

Join a **free, global**, week-long Galaxy Training event covering everything from RNA-Seq, Single Cell, Proteomics, SARS-CoV-2 *and more!* This year will include Galaxy Admin Training.



 @gxytraining @Gallantries\_EU



With the support  
of the  
European Union



[Home](#) | [Key Dates](#) | [Conduct](#) | [Abstracts](#) | [Schedule](#) | [Training](#) | [CoFest](#) | [Sponsors](#) | [Travel](#) | [Register](#) | [Childcare](#) | [Organizers](#)

## 2022 Galaxy Community Conference (GCC2022)

July 17-23, 2022

University of Minnesota, Twin Cities  
Minneapolis, Minnesota, United States

[#UseGalaxy2022](#)

**Abstracts due April 12**



University of Minnesota



Minneapolis-St. Paul

<https://galaxyproject.org/events/gcc2022/>

# Acknowledgments



<https://galaxyproject.org/>

