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

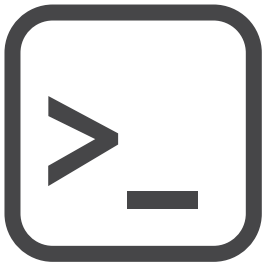
Goal: provide quality assurance reports for a patient’s precision cancer therapy

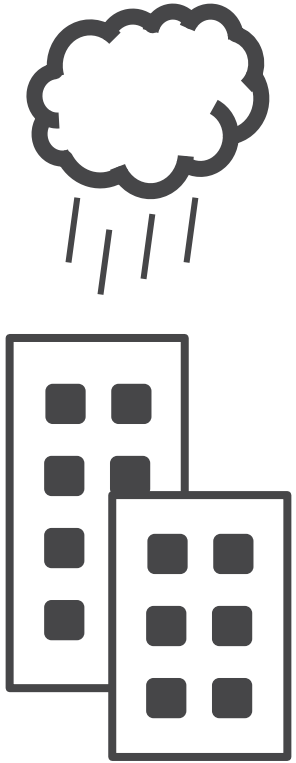
Process:

- 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

Galaxy PROJECT

Project plan

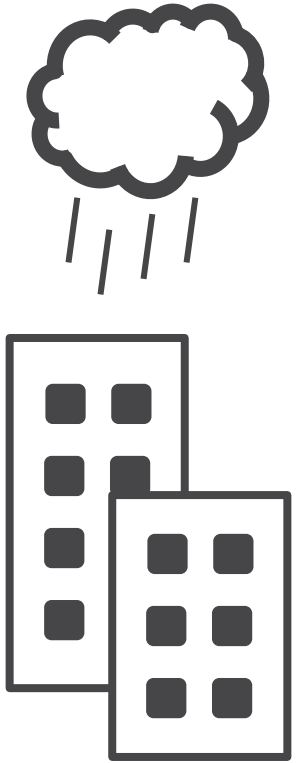




GPwhat?

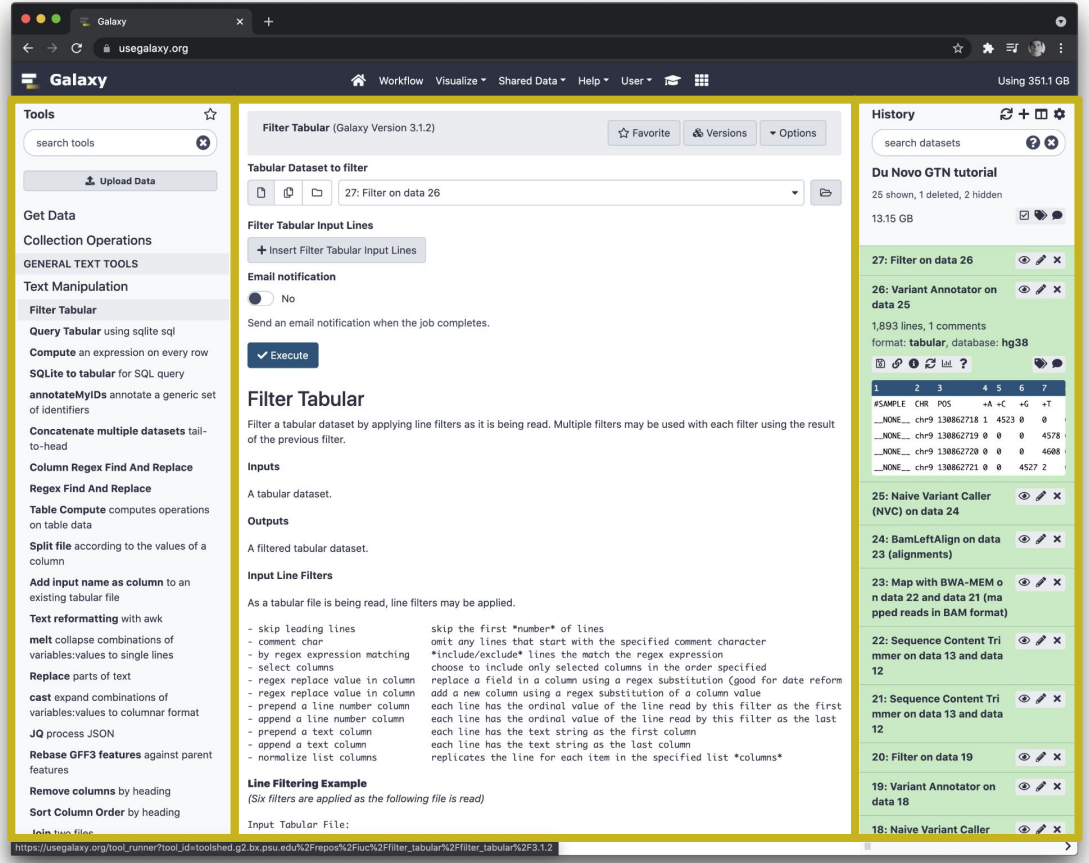
Galaxy PROJECT

Let’s lose the command prompt



GPwhat?

Galaxy?



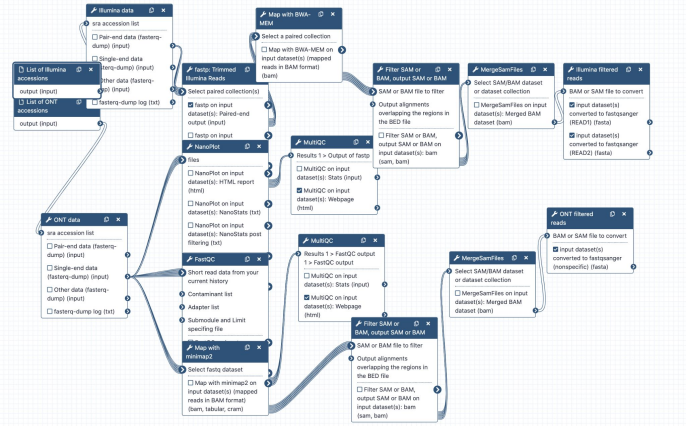
Tool panel

History

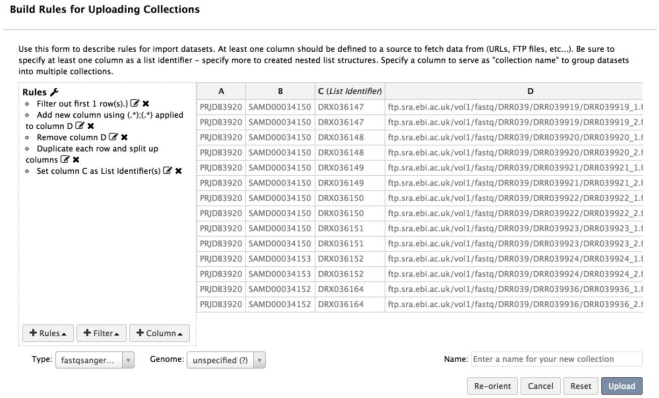
Tool form

Galaxy PROJECT

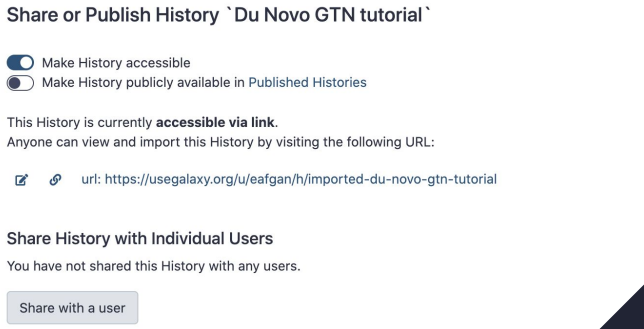
>8,000 tools integrated



Graphical workflow editor



Dataset collections



Sharing & publishing

Galaxy PROJECT

But what about Kat’s project?

IWC: Intergalactic Workflow Commission

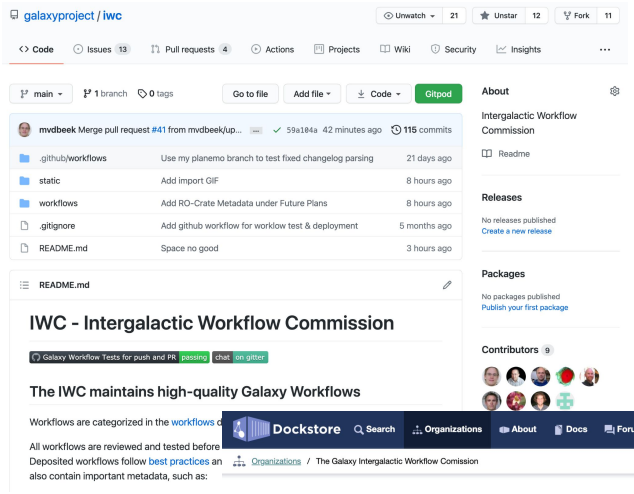
A community-driven effort to have:

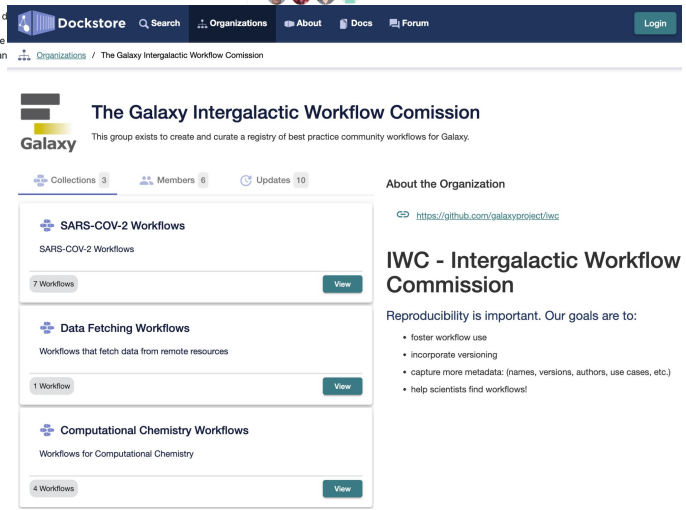
- A way to define and run workflow tests
- A (central) repository to collect workflows
- A way to define and run workflow tests
- A way to define workflow versions
- Conventions and standards for metadata
- Workflows

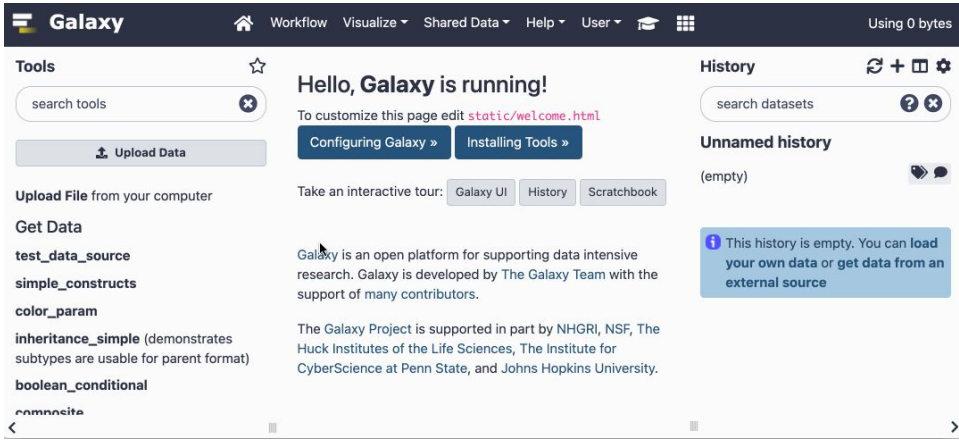
https://github.com/galaxyproject/iwc

Galaxy PROJECT

IWC workflows

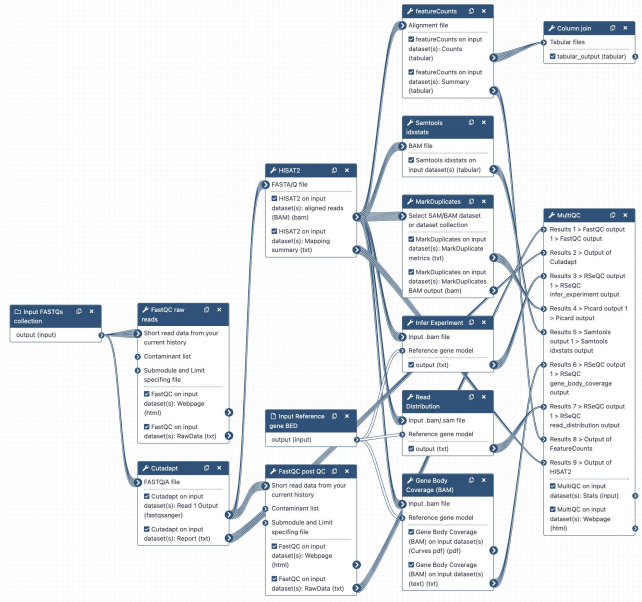






Galaxy PROJECT

A tuned RNA-Seq workflow



Start from an existing template

Iterate and make desired adjustments

FASTQ 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

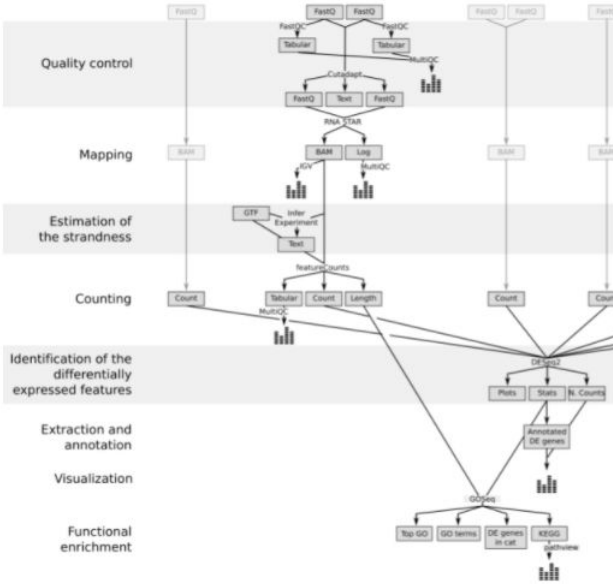
FASTQ Summary Statistics by column

Compute quality statistics

Draw nucleotides distribution chart

Switch to 2.1.0+ga
Switch to 2.1.0+ga
Switch to 2.1.0+ga
Switch to 2.1.0+ga
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



Galaxy PROJECT

And a report

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


```
# 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
```

Galaxy PROJECT

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

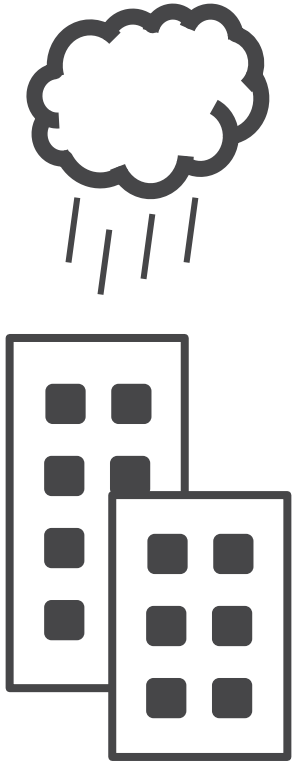
Vibrant community

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

Galaxy PROJECT

Working with private data

Galaxy



GPwhat?

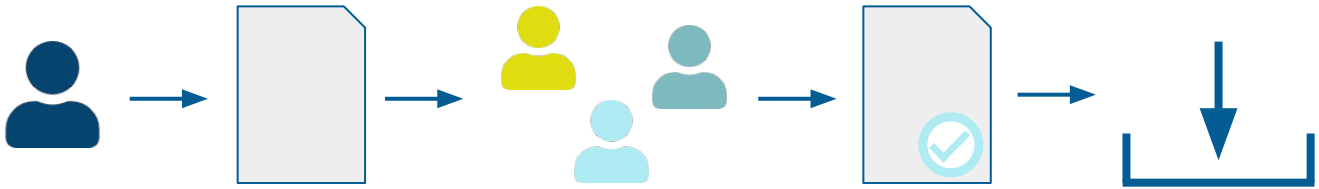
Galaxy PROJECT

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?

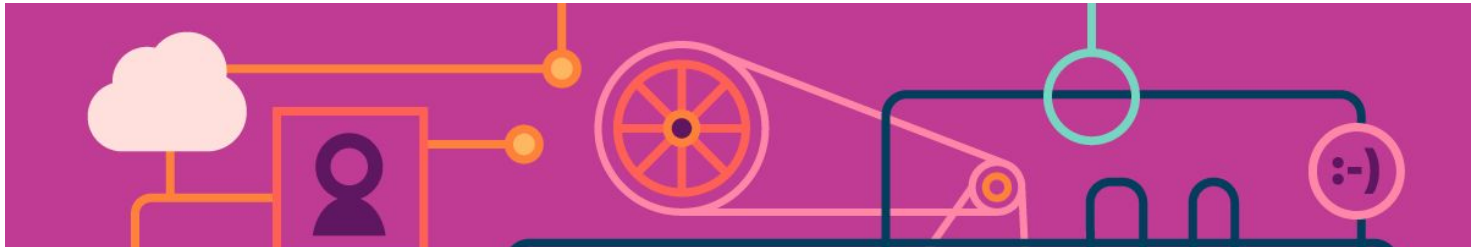
Working with sensitive data



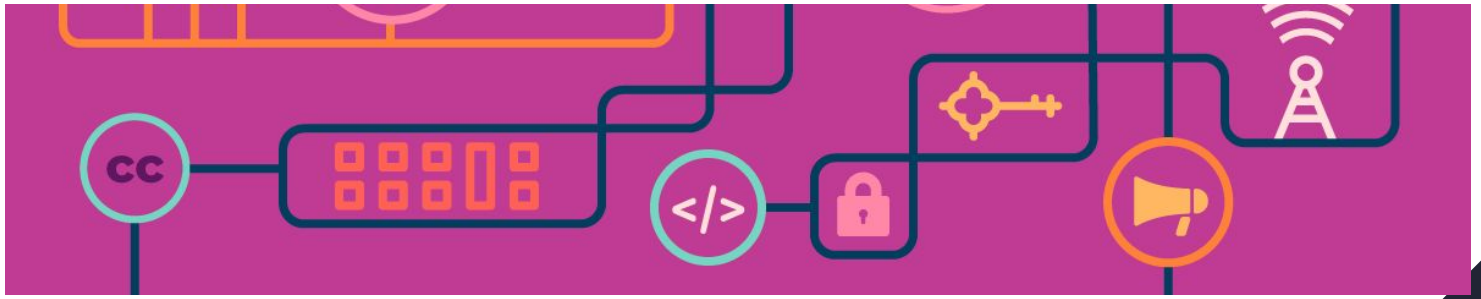
The **recipient institution is ultimately responsible** for maintaining the confidentiality, integrity, and availability of the data.

Galaxy PROJECT

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**.



Galaxy PROJECT

Designing the necessary security perimeter

1. NIST National Institute of Standards and Technology

2. 

FISMA COMPLIANCE

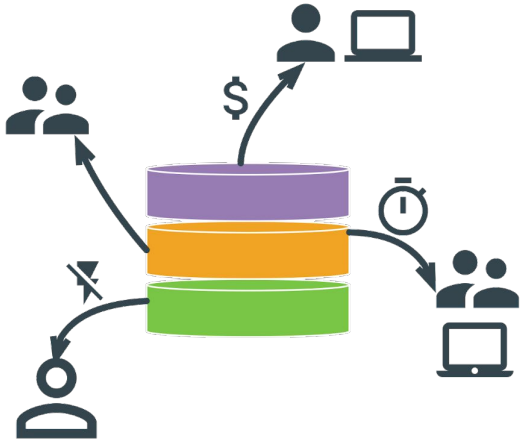
HIPAA COMPLIANT

FedRAMP

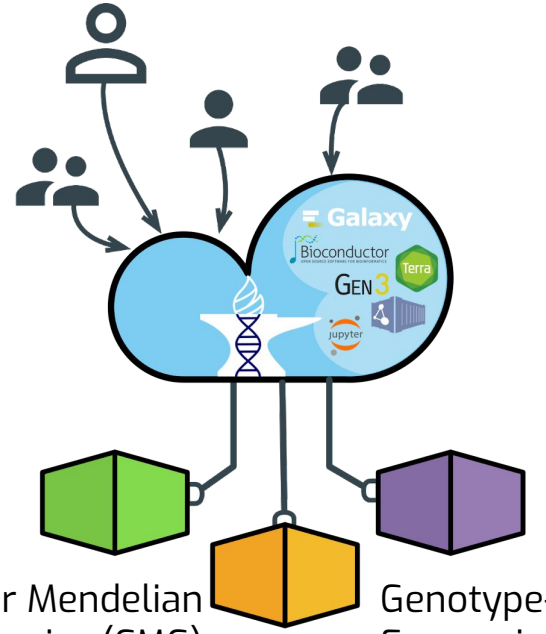
3. Conclusion: out of reach for individual investigators and (most) labs.

Galaxy PROJECT

NHGRI AnVIL overview



FedRAMP



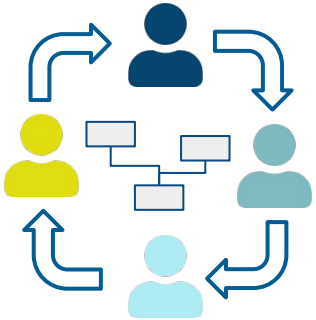
http://anvilproject.org

Galaxy PROJECT

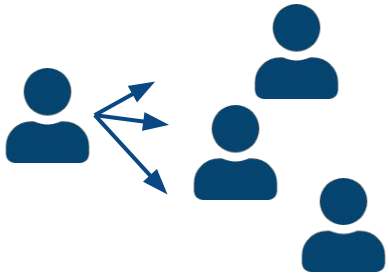
Architecting Galaxy for protected datasets

Accessibility

Reproducibility



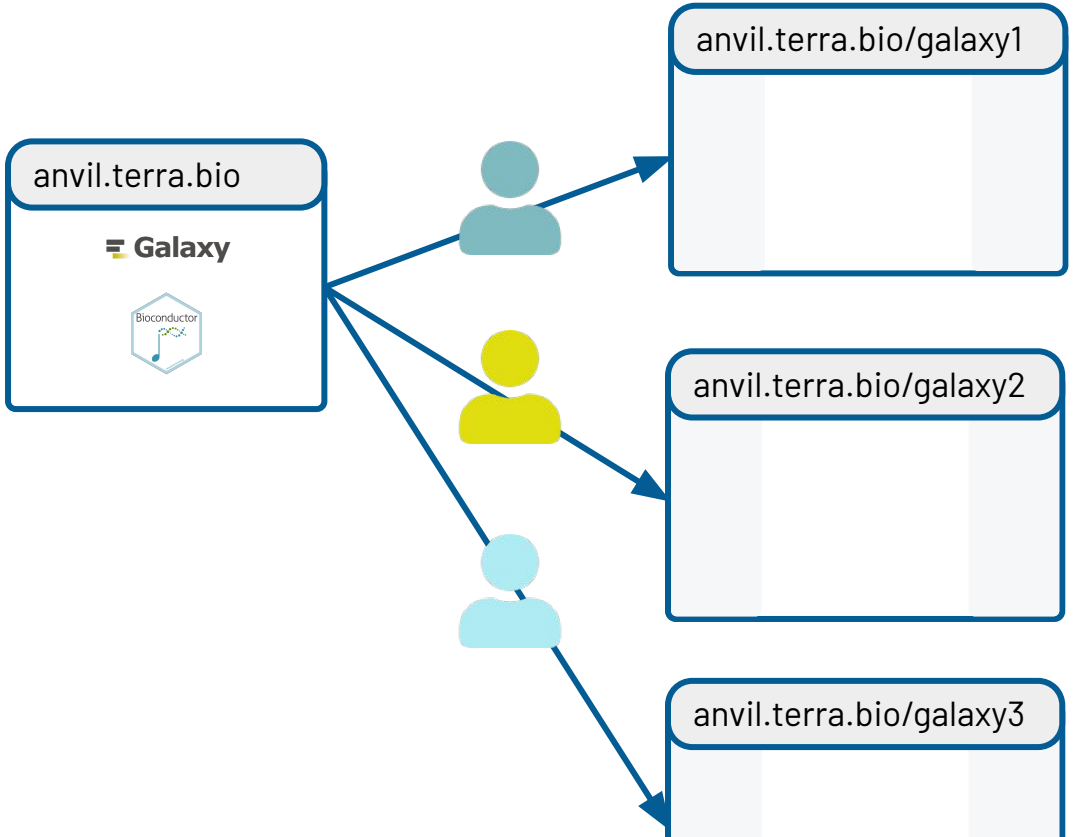
Sharing



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

Galaxy PROJECT

Replicating accessibility



kubernetes

HELM

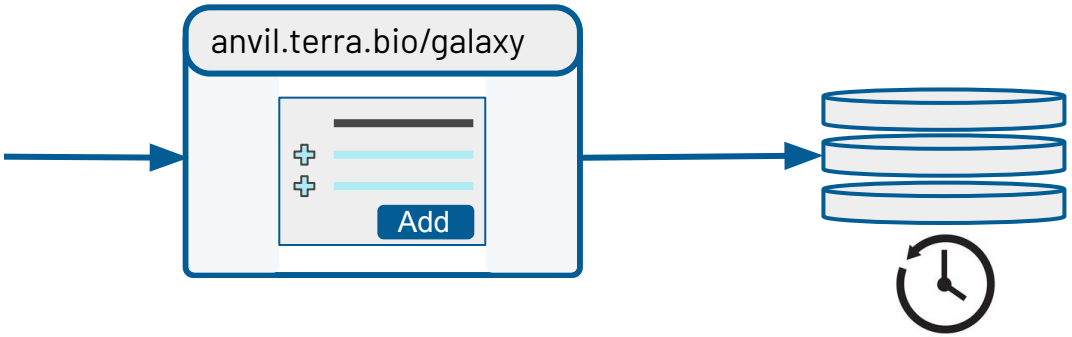
Galaxy PROJECT

Replicating reproducibility

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

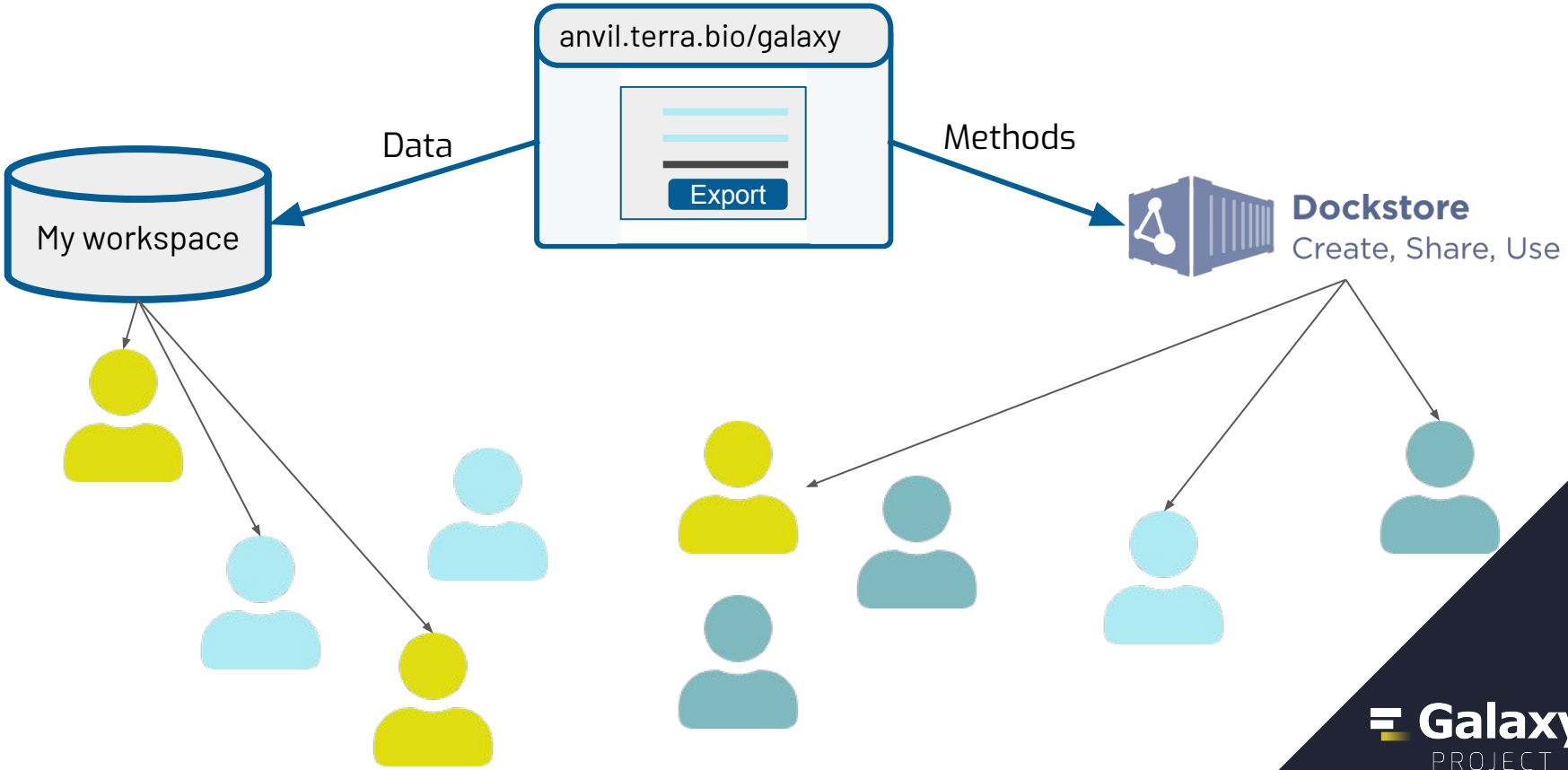
Search Summary

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

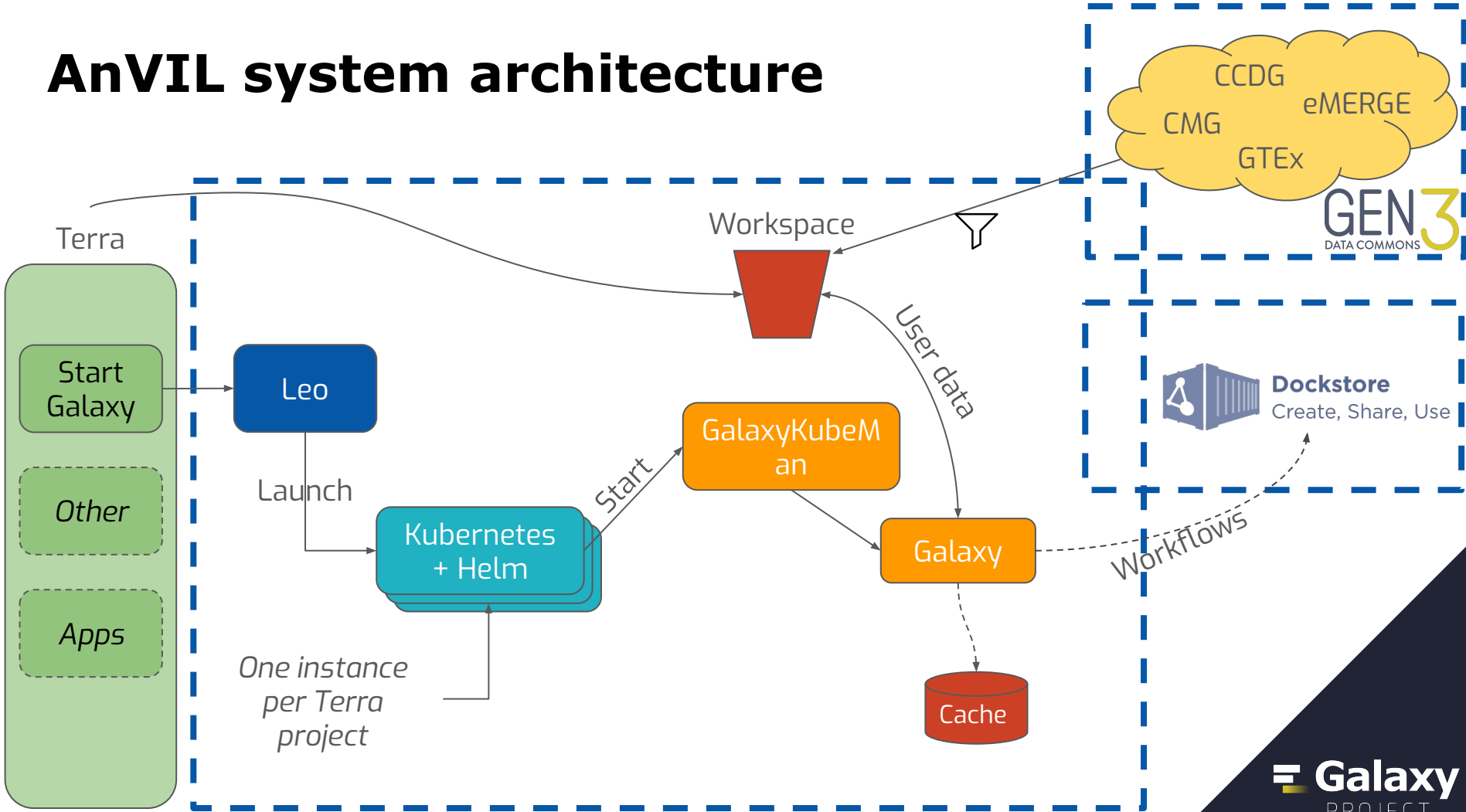


Galaxy PROJECT

Replicating sharing



AnVIL system architecture



Galaxy & Terra

Galaxy integration with Terra reaches beyond AnVIL

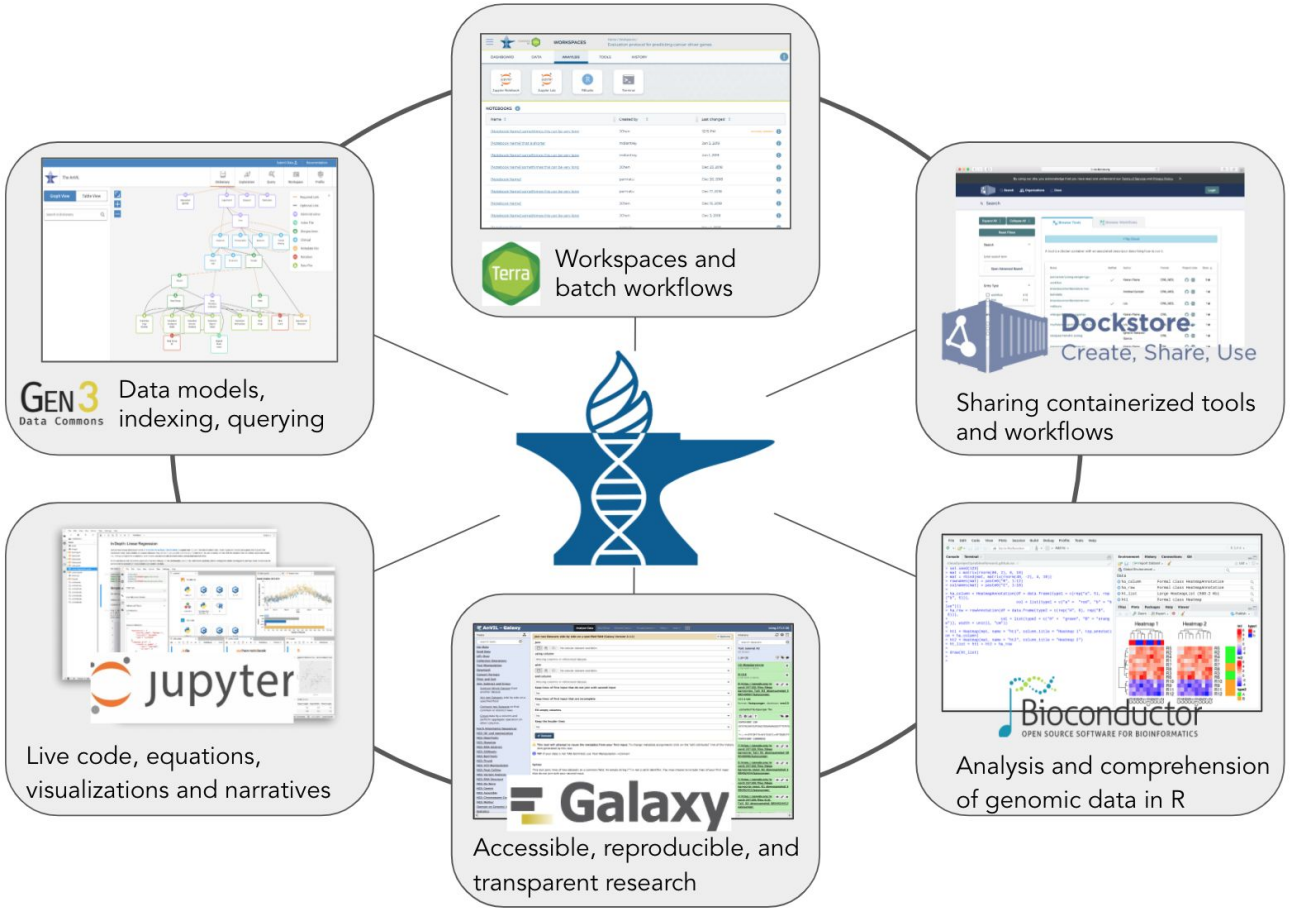
Multiple FireCloud-backed Terra instances now offer Galaxy

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

https://terra.bio/a-galaxy-of-tools-at-your-fingertips/

Galaxy PROJECT

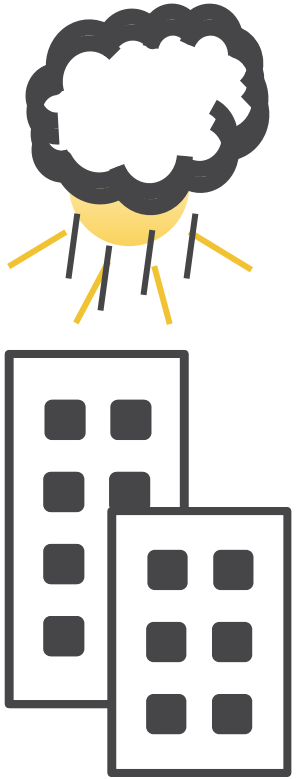
AnVIL is a suite of applications



Galaxy PROJECT

Approachable cloud computing

Galaxy



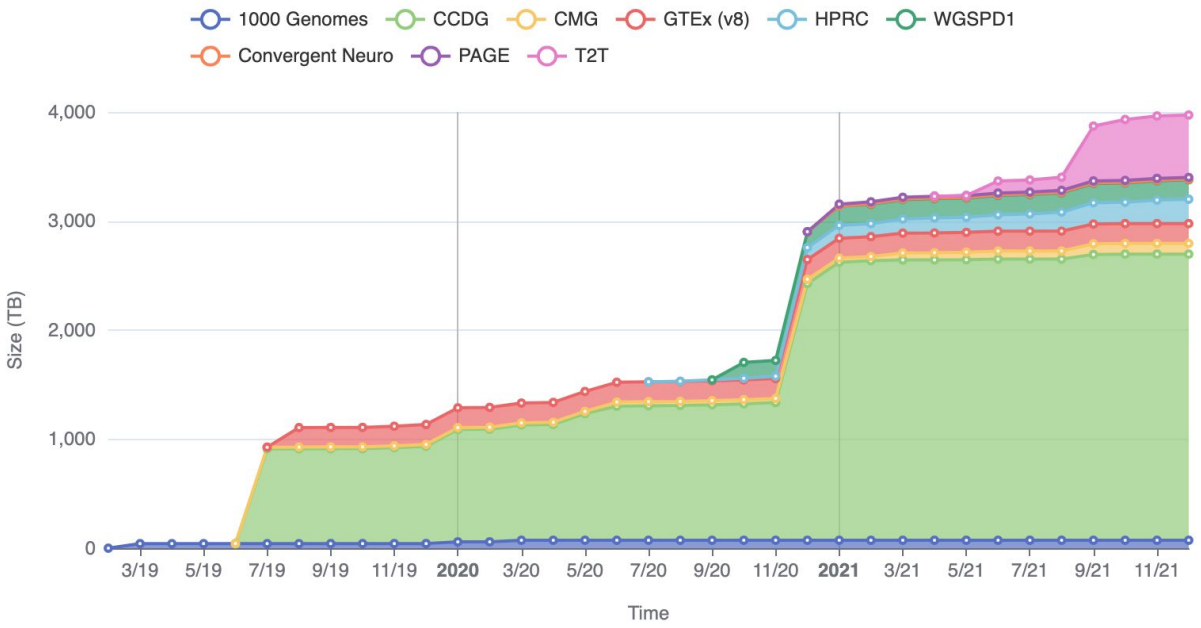
GPwhat?

Galaxy PROJECT

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!

Galaxy PROJECT

One collection of data is good. More is better.

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

NIH National Heart, Lung, and Blood Institute | BioData CATALYST

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

NCBI
National Center for Biotechnology Information

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

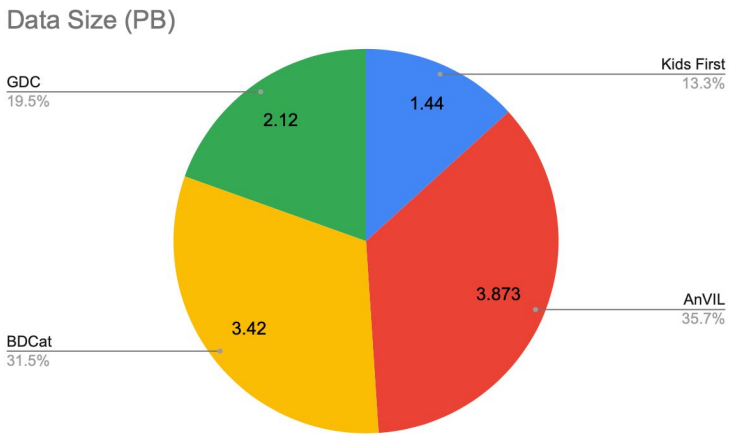
- If we are successful, we will catalyze the creation of an open and federated data ecosystem.
  - Others have done it before (SWIFT, the internet, the web).

- If we fail, we will degenerate into a collection of monolithic data silos
  - Others have done this before too (medical records in US hospitals)...

https://anvilproject.org/ncpi

Galaxy PROJECT

NIH Cloud Platform Interoperability Effort (NCPI)



Researcher Auth Service

Data Repository Service

Fast Healthcare Interoperability Resources

11Pb / 689k participants and growing!
Cross-platform accessibility through several key technologies

Galaxy PROJECT

Ongoing NCPI efforts

Data discovery

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

Hand-off of search results from portals to workspace environments

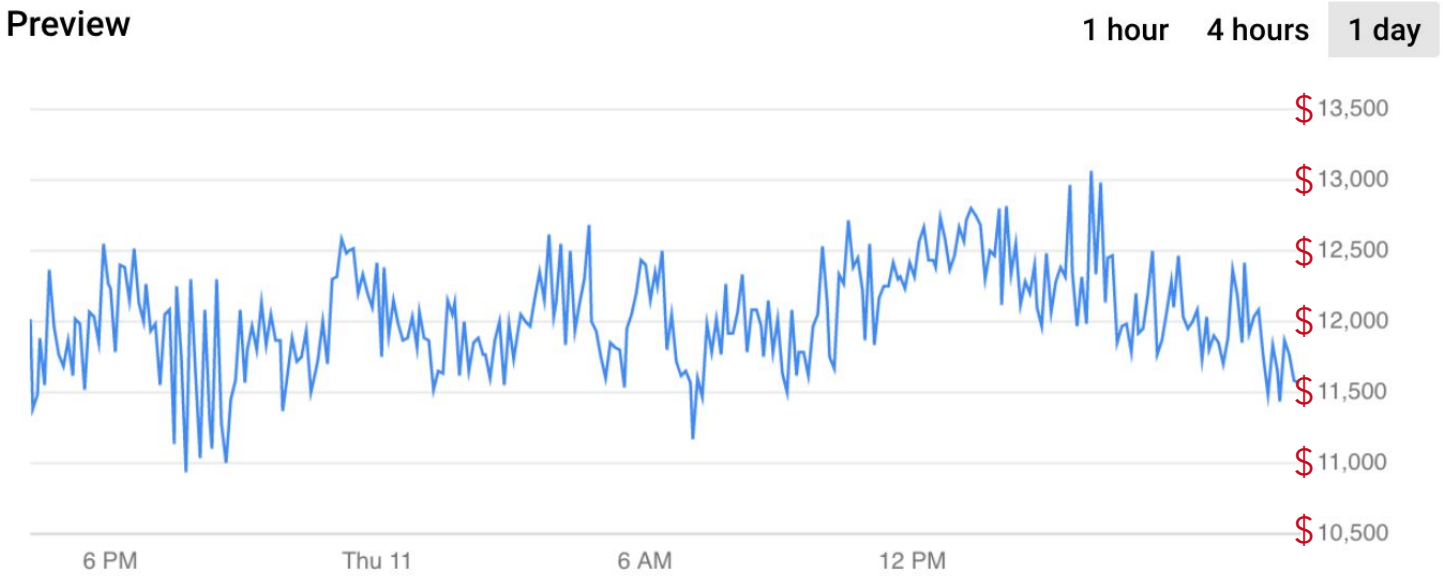
- https://youtu.be/YGZTxDdaWqk

Single Sign-On with NIH RAS

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

Galaxy PROJECT

T2T Analysis on Google Cloud Platform



dollars/hour

Galaxy PROJECT

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) — Monthly / Hourly

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 E2 custom vCPUs and memory.			

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) — Monthly / Hourly

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 Custom vCPU and memory.			

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) — Monthly / Hourly

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 E2 custom vCPUs and memory.			

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) — Monthly / Hourly

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 Custom vCPU and memory.			

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) — Monthly / Hourly

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 E2 custom vCPUs and memory.			

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.

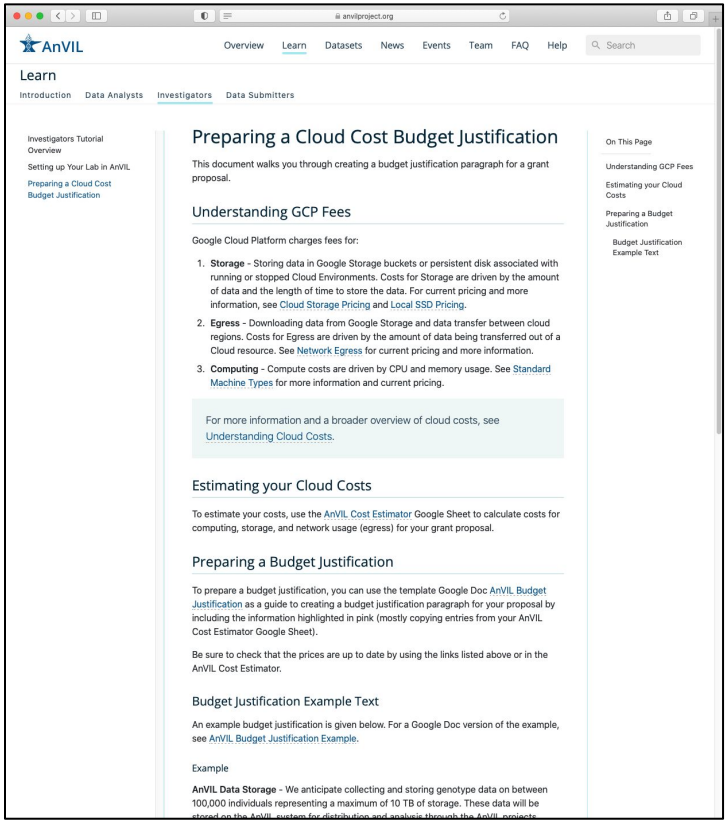
High-CPU machine types have one vCPU for every 1 GB of memory. High-CPU machine types are ideal for tasks that require moderate memory configurations for the needed vCPU count.

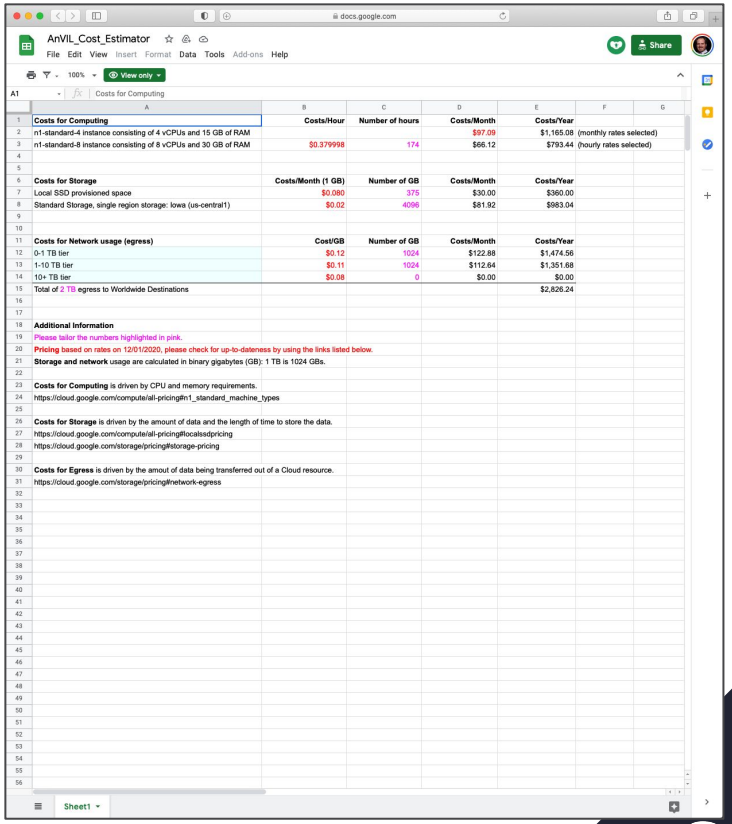
Iowa (us-central1) — Monthly / Hourly

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 Custom vCPUs and memory.			

https://cloud.google.com/compute/vm-instance-pricing

AnVIL Cloud Cost Budget Templates





https://anvilproject.org/learn/investigators/budget-templates

Galaxy PROJECT

Understanding costs

AnVIL project · anvilproject.org

Shared · Published (unlisted) Feb 25 · 1 like

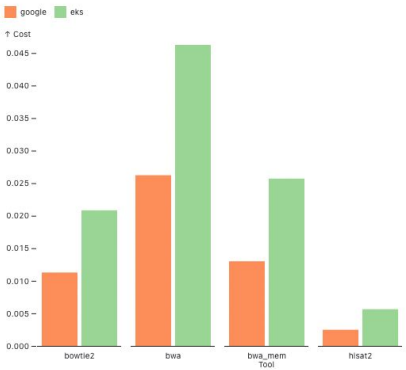
Genomics tool benchmarking for cloud resource cost estimation

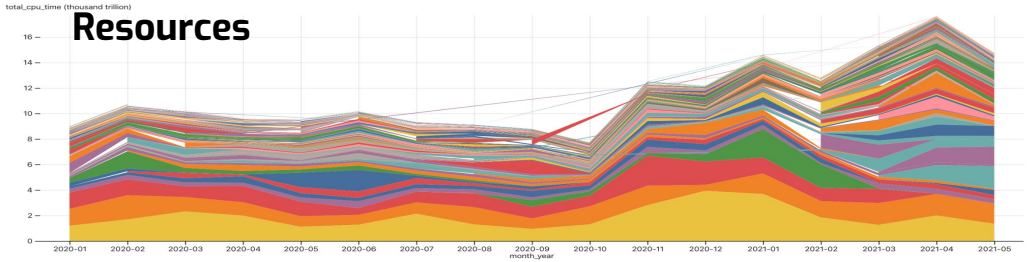
▶ Table of Contents

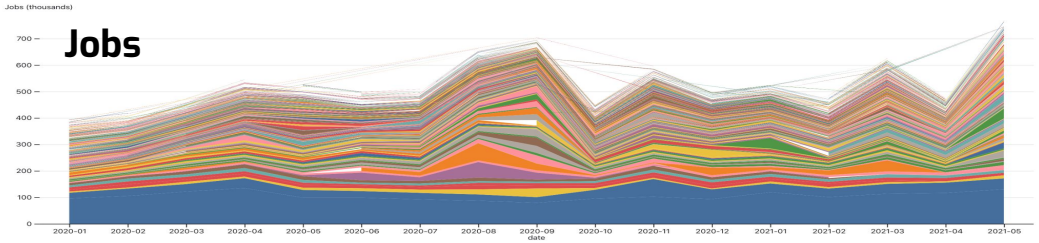
How much will it cost?

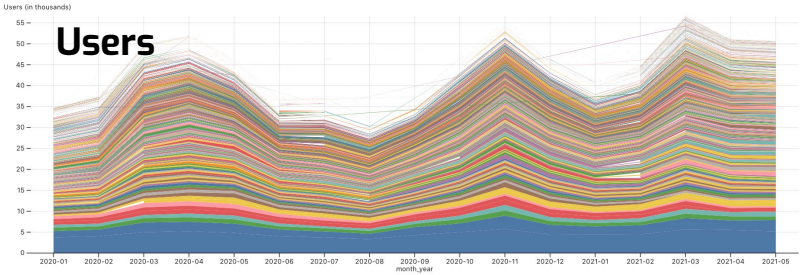
Cloud resources offer great conveniences and capabilities but how much will it cost to run a given analysis or a tool? The following diagram offers a high-level representation of the anticipated workload costs as a function of data. The data is represented in terms of file size. If you are not sure about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

Cost









Galaxy PROJECT

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

Galaxy PROJECT

NIH/ODSS STRIDES Initiative

https://datascience.nih.gov/strides

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

Galaxy PROJECT

Funding your cloud experience

But ultimately - you gotta write them grants!

Past solicitations

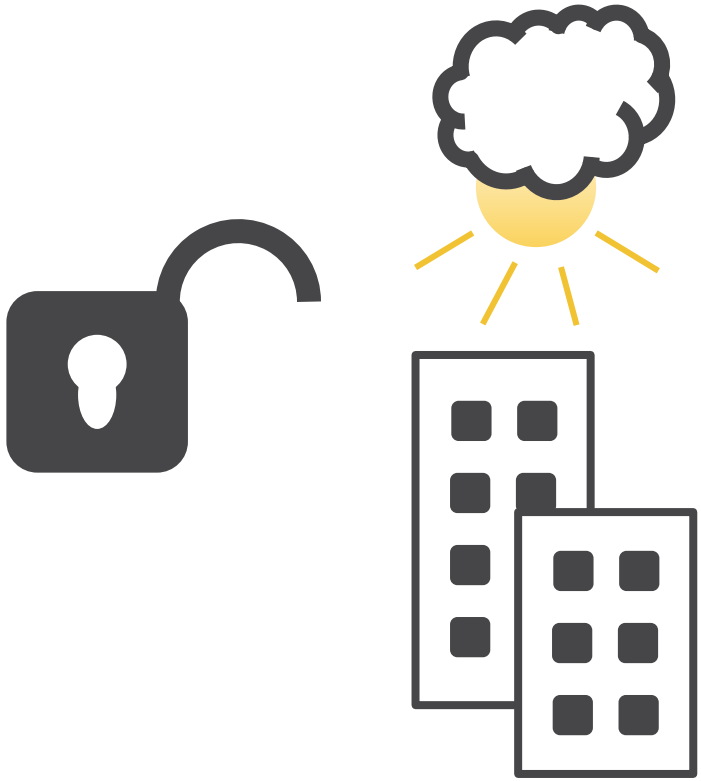
- Genomic Variation and Function Data and Administrative Coordinating Center (RFA-HG-20-046)
- Polygenic Risk Score (PRS) Methods and Analysis for Populations of Diverse Ancestry (RFA-HG-20-002)
- Mendelian Genomics Research Centers (RFA-HG-20-007)

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)

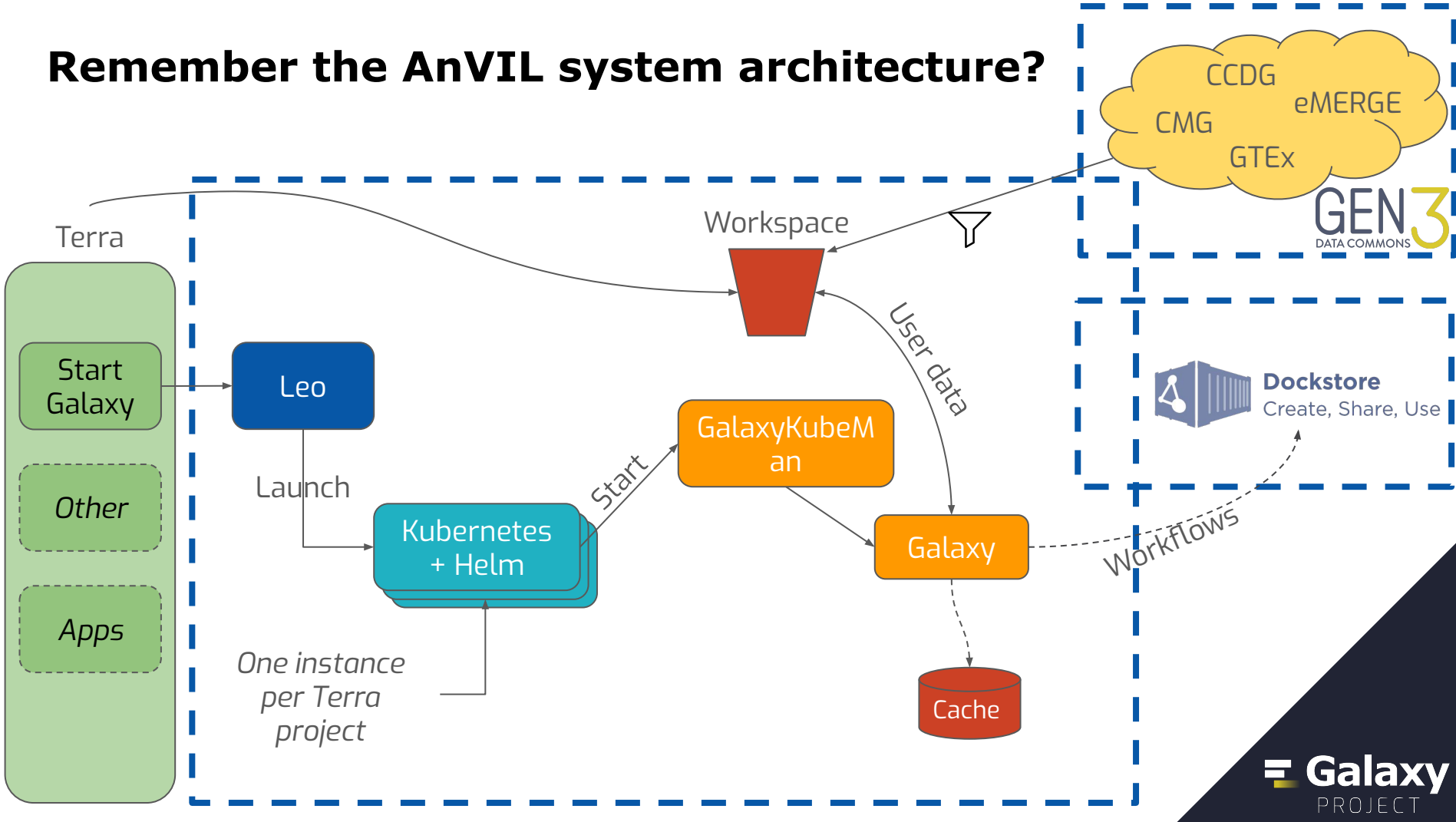
How to design the cloud for scalability?

Galaxy

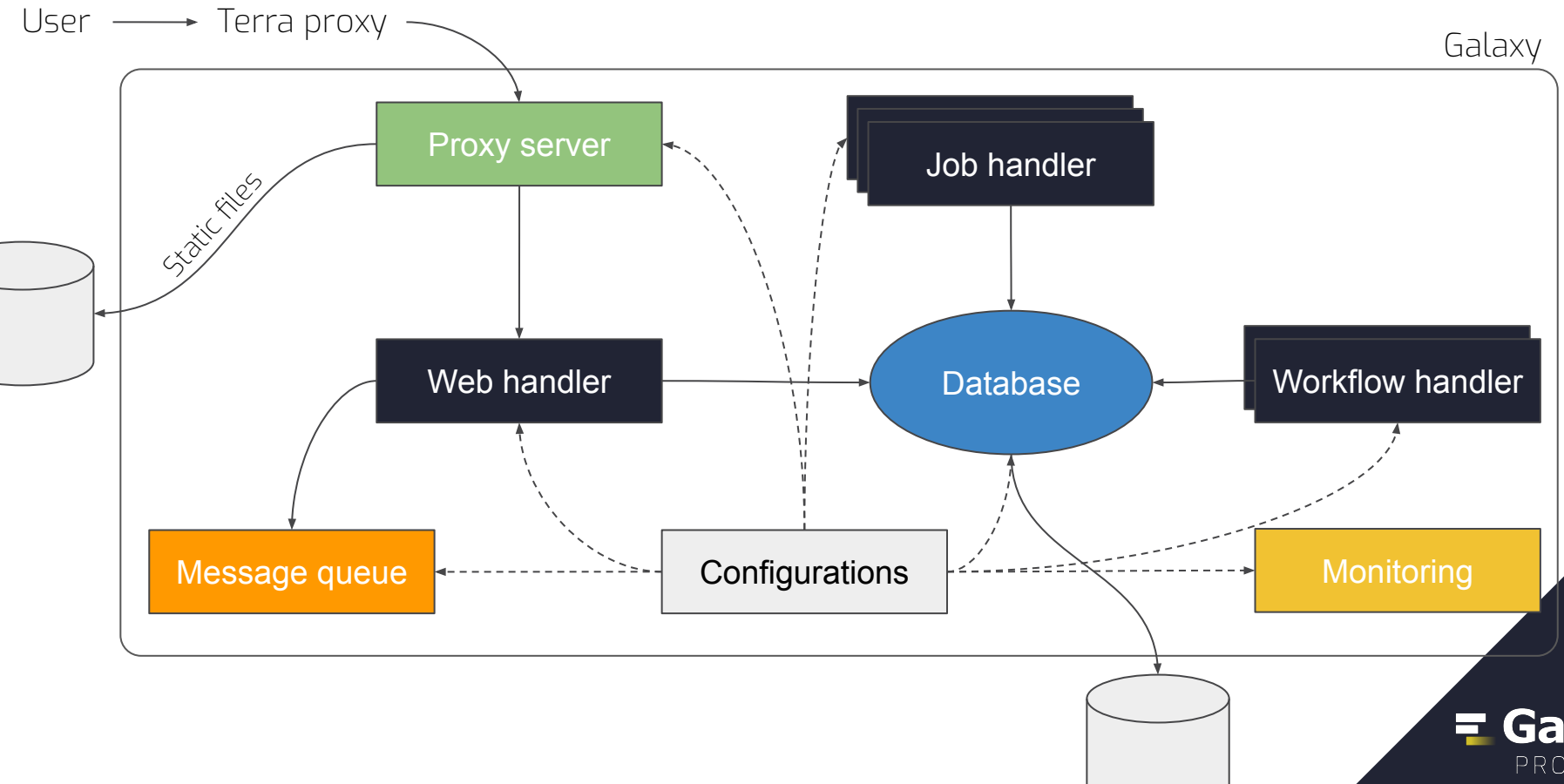


GPU what?

Remember the AnVIL system architecture?

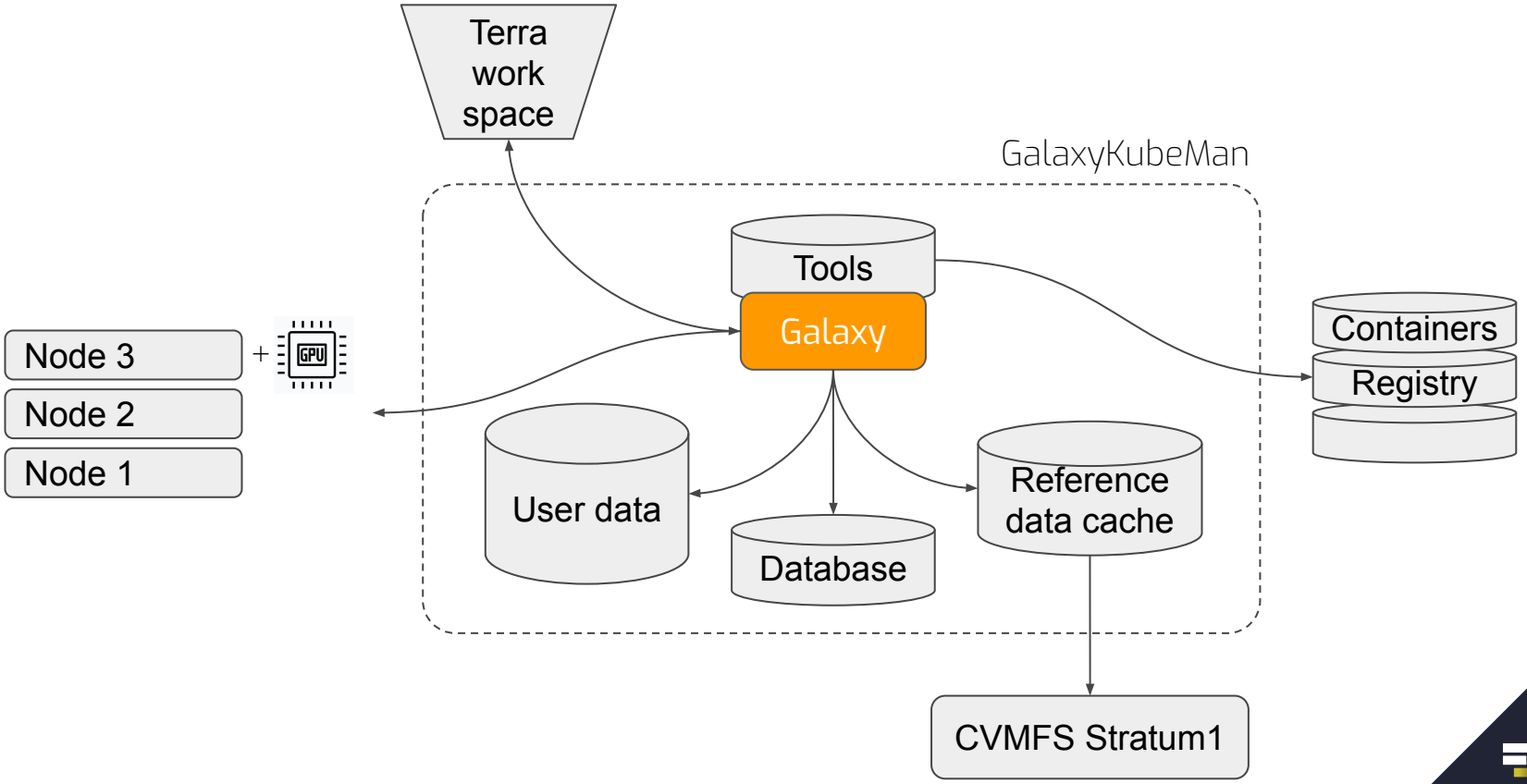


Galaxy server components



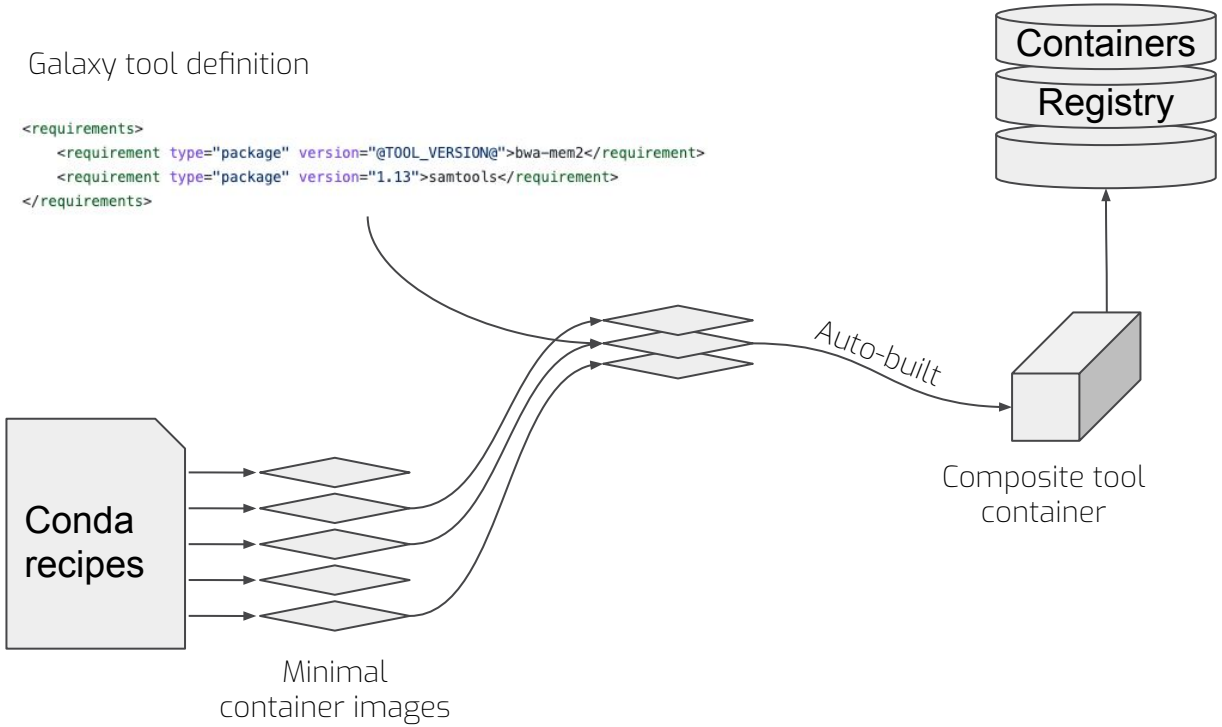
Galaxy PROJECT

Architecture for the overall Galaxy server



Galaxy PROJECT

Tools and Biocontainers



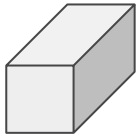
- Versioned tools, inlc. dependencies
- Nearly 10,000 images hosted
- Usable outside the Galaxy ecosystem

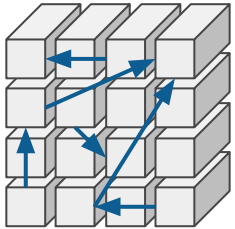
Galaxy PROJECT

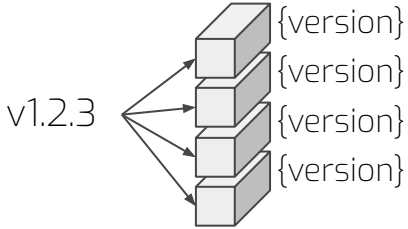
Change and service management

- 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







docker

kubernetes

HELM

Galaxy PROJECT

Galaxy Helm chart

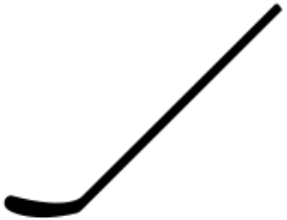
```
helm install galaxy
```

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

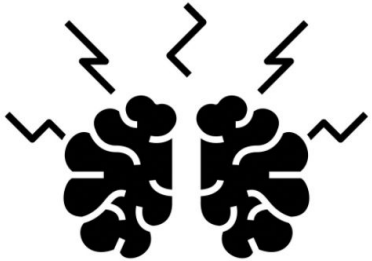
https://github.com/galaxyproject/galaxy-helm

Galaxy PROJECT

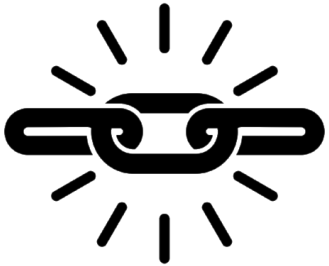
Our experience with K8s



There is definitely a learning curve



Debugging is tough. No other way to say it.



Codified. Portable. Robust.

Powerful and very well designed.

Galaxy PROJECT

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

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/

Galaxy PROJECT

Acknowledgments

https://galaxyproject.org/

NIH National Institutes of Health
NSF
CZI
Johns Hopkins University
The University of Melbourne
Broad Institute
PennState
Oregon Health & Science University
OHSU