NIH Workshop on Cloud-Based Platforms Interoperability
October 30th and November 2nd, 2020

Welcome to the...

NIH Cloud Platforms Interoperability Fall 2020 Workshop

We'll be starting shortly!

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

AnVIL

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

Welcome & Introduction to Day 2

Adam Resnick
Children's Hospital of Philadelphia
Valerie Cotton
Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), NIH

Welcome & Introduction

Draft Roadmap in FunRetro
(from Day 1)

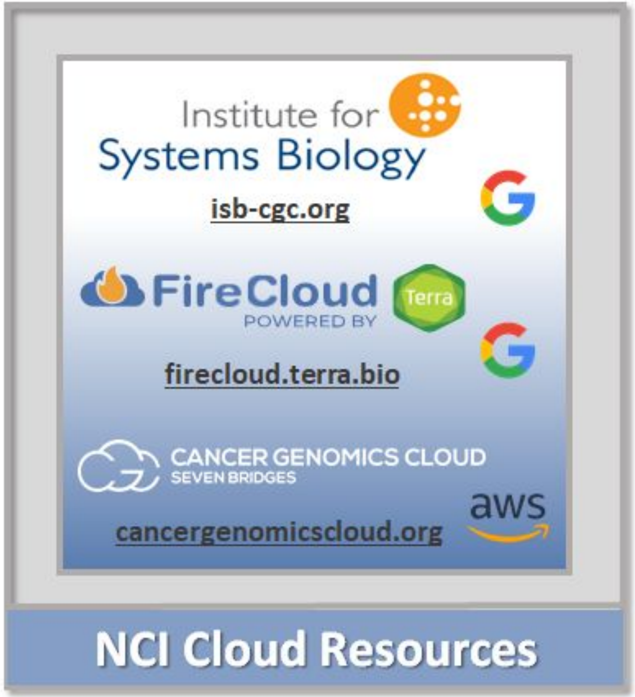
CRDC Cloud Costs: Current Practices

Tanja Davidsen
NCI Center for Biomedical Informatics and Information Technology (CBIIT)

Cloud Costs - NCI Cloud Resources - NCI CRDC

Connecting NCI data and compute in the cloud

- Access to large cancer data sets without need to download
- Access to workspaces, analysis tools, and pipelines
- Ability for researchers to bring their own data/tools



Cloud Costs - Compute - NCI CRDC

CRDC Cloud Resources Compute

- Each Cloud Resource provides $300 Credit to every new user
    - Some additional free credits beyond $300 can be requested via an application
    - Beyond the free credits users can use a credit card or a billing account
- Links to free credit information
    - Seven Bridges CGC
    - Broad FireCloud
    - Institute for Systems Biology CGC

Cloud Costs - Storage - NCI CRDC

CRDC Cloud Storage

- Copied on both Amazon and Google
- Mostly genomic, additional WGS, imaging, and proteomic data coming soon
- Size:
  - 3.5PB total and growing
  - 2-4PB additional data expected in FY21
- Costs:
  - 2PB free from STRIDES each: Amazon and Google
  - Additional FY21 budget for storage $1.5 million using STRIDES discount
- The future concerns:
  - Looking for funding opportunities as FY22 budgets will not cover storage needed
  - Looking for cost savings/sustainability: lower cost storage, compression, etc.

NHLBI BioData Catalyst Cloud Costs: Current Practices

Chip Schwartz
Senior Advisor, NHLBI

BioData Catalyst Cloud Credit Overview

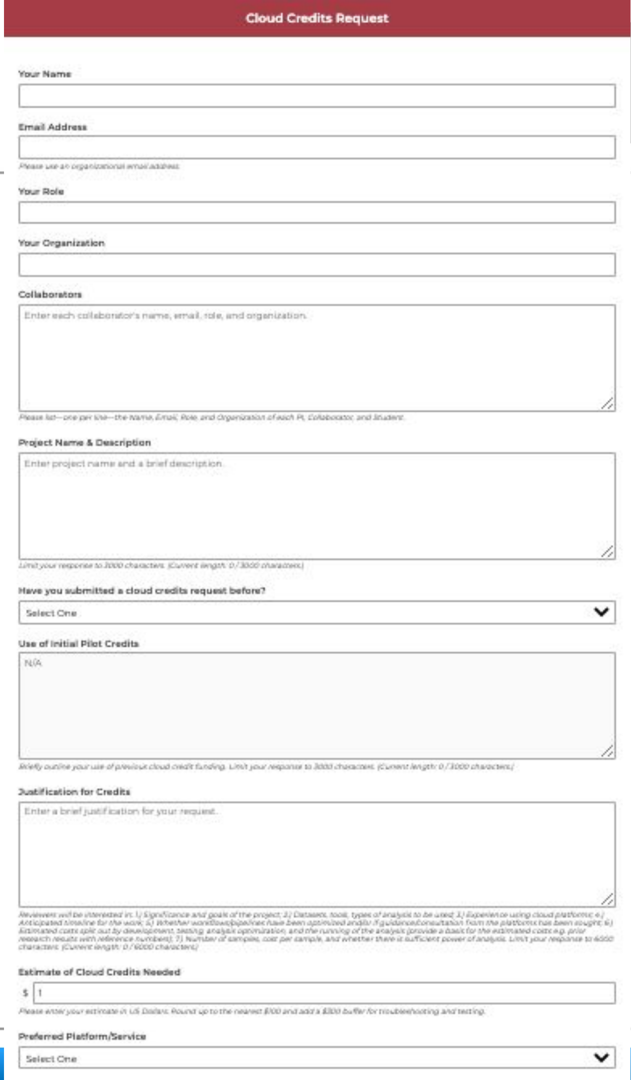
- Pilot Funding: The NHLBI currently provides **$500 in cloud credits** to new users of NHLBI BioData Catalyst via a billing group on either *BioData Catalyst Powered by Seven Bridges* or *BioData Catalyst Powered by Terra*.
- If the anticipated costs are in excess of $500:
  - Users can cover those costs using their own AWS and/or Google accounts which can be brought to BioData Catalyst
  - Users can apply for additional credits via the **NHLBI BioData Catalyst Cloud Credit Program**

More information at https://biodatacatalyst.nhlbi.nih.gov/resources/cloud-credits

BioData Catalyst Cloud Credits

- BioData Catalyst has a lightweight form for requesting additional credits
- Requests are reviewed and approved based on projected costs and scientific merit

More information at
https://biodatacatalyst.nhlbi.nih.gov/resources/cloud-credits



Cloud Credits Request

Your Name

Email Address

Please use an organizational email address.

Your Role

Your Organization

Collaborators

Enter each collaborator's name, email, role, and organization.

Please list—one per line—the Name, Email, Role, and Organization of each PI, Collaborator, and Student.

Project Name & Description

Enter project name and a brief description.

Limit your response to 3000 characters. (Current length: 0 / 3000 characters)

Have you submitted a cloud credits request before?

Select One

Use of Initial Pilot Credits

N/A

Briefly outline your use of previous cloud credit funding. Limit your response to 3000 characters. (Current length: 0 / 3000 characters)

Justification for Credits

Enter a brief justification for your request.

Reviewers will be interested in: 1) Significance and goals of the project; 2) Datasets, tools, types of analysis to be used; 3) Experience using cloud platforms; 4) Anticipated timeline for the work; 5) Whether workflows/pipelines have been optimized and/or if guidance/consultation from the platforms has been sought; 6) Estimated costs split out by development, testing, analysis optimization, and the running of the analysis (provide a basis for the estimated costs e.g. prior research results with reference numbers); 7) Number of samples, cost per sample, and whether there is sufficient power of analysis. Limit your response to 6000 characters. (Current length: 0 / 6000 characters)

Estimate of Cloud Credits Needed

$ 1

Please enter your estimate in US Dollars. Round up to the nearest $100 and add a $300 buffer for troubleshooting and testing.

Preferred Platform/Service

Select One

BioData Catalyst Cloud Credit Resources

- Currently, we offer these resources for **understanding cloud costs**
  - ○ Controlling your cloud costs (*BioData Catalyst Powered by Terra*)
  - ○ Cloud infrastructure pricing (*BioData Catalyst Powered by Seven Bridges*)
  - ○ Comprehensive tips for reliable and efficient analysis set-up (*BioData Catalyst Powered by Seven Bridges*)
- Next steps:
  - ○ White paper on estimating cloud cost is in the works

BioData Catalyst Cloud Credit Lessons Learned

- Cloud costs needs to be more accurately tracked and managed
  - We need to develop **better reporting** on cloud cost disbursement and usage. STRIDES dashboard should help resolve many reporting issues (beta coming soon)
  - We need to **QA/QC user pipelines/workflows** to improve performance and ensure maximum cost efficiency
- Cloud costs create a lot of **anxiety** for users
  - They can present a significant perceived barrier to entry
  - Evaluating costs takes time and can return varied results
- **Shared data storage** facilitates research collaborations and ultimately reduces costs

Kids First DRC Cloud Costs: Current Practices

David Higgins, PhD
Kids First Data Resource Center
Children's Hospital of Philadelphia

Kids First DRC Cloud Credit Overview

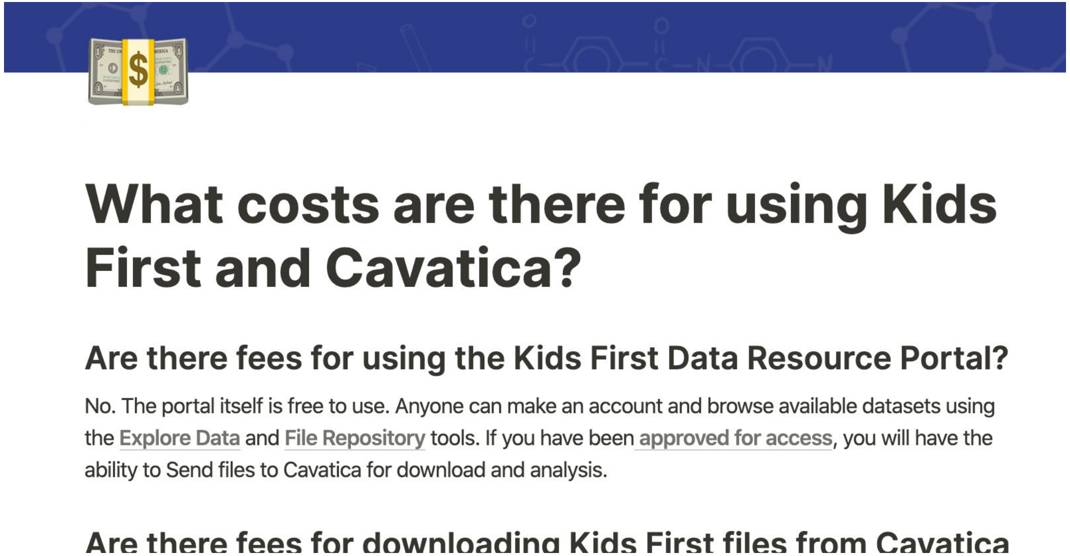
- Pilot Funding: All users receive $100 in pilot funds upon making an account in CAVATICA (Seven Bridges).

- Cloud Credits: Users can receive for more funds through NIH Common Fund.
  - Phase 1 (FY2020): Up to $5,000 upon approval of an application.
  - Phase 2 (FY2021):
    - All Kids First X01 PIs: receive $1,000 when they receive their data
    - Can then apply for additional funds to complete their analysis
    - KFDRC can track disbursement and usage through an admin account

Kids First DRC Cloud Credit Resources

What costs are there for using Kids First and Cavatica?

(FAQ on Kids First DRC Support Pages)



What costs are there for using Kids First and Cavatica?

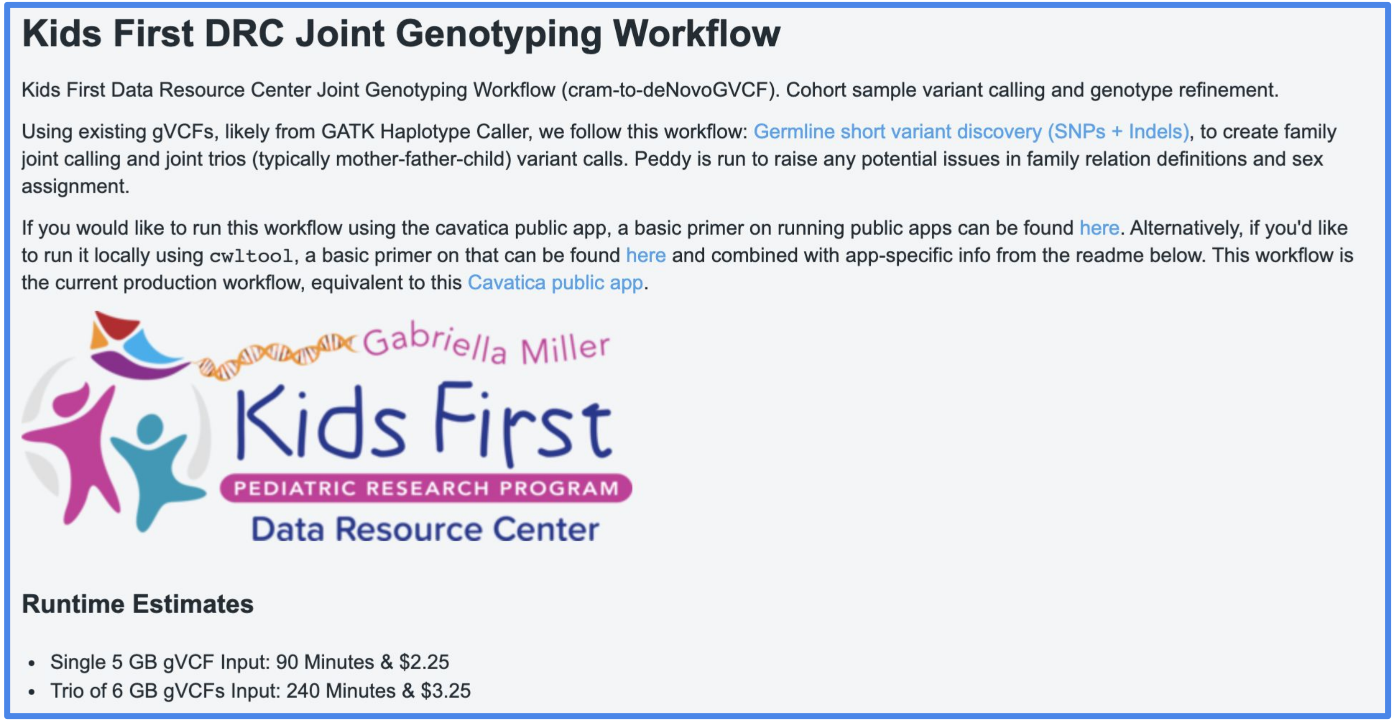
Are there fees for using the Kids First Data Resource Portal?

No. The portal itself is free to use. Anyone can make an account and browse available datasets using the Explore Data and File Repository tools. If you have been approved for access, you will have the ability to Send files to Cavatica for download and analysis.

Are there fees for downloading Kids First files from Cavatica

Kids First DRC Cloud Credit Resources

Benchmarking Statistics for Kids First Workflows on CAVATICA



Kids First DRC Joint Genotyping Workflow

Kids First Data Resource Center Joint Genotyping Workflow (cram-to-deNovoGVCF). Cohort sample variant calling and genotype refinement.

Using existing gVCFs, likely from GATK Haplotype Caller, we follow this workflow: Germline short variant discovery (SNPs + Indels), to create family joint calling and joint trios (typically mother-father-child) variant calls. Peddy is run to raise any potential issues in family relation definitions and sex assignment.

If you would like to run this workflow using the cavatica public app, a basic primer on running public apps can be found here. Alternatively, if you'd like to run it locally using `cwltool`, a basic primer on that can be found here and combined with app-specific info from the readme below. This workflow is the current production workflow, equivalent to this Cavatica public app.

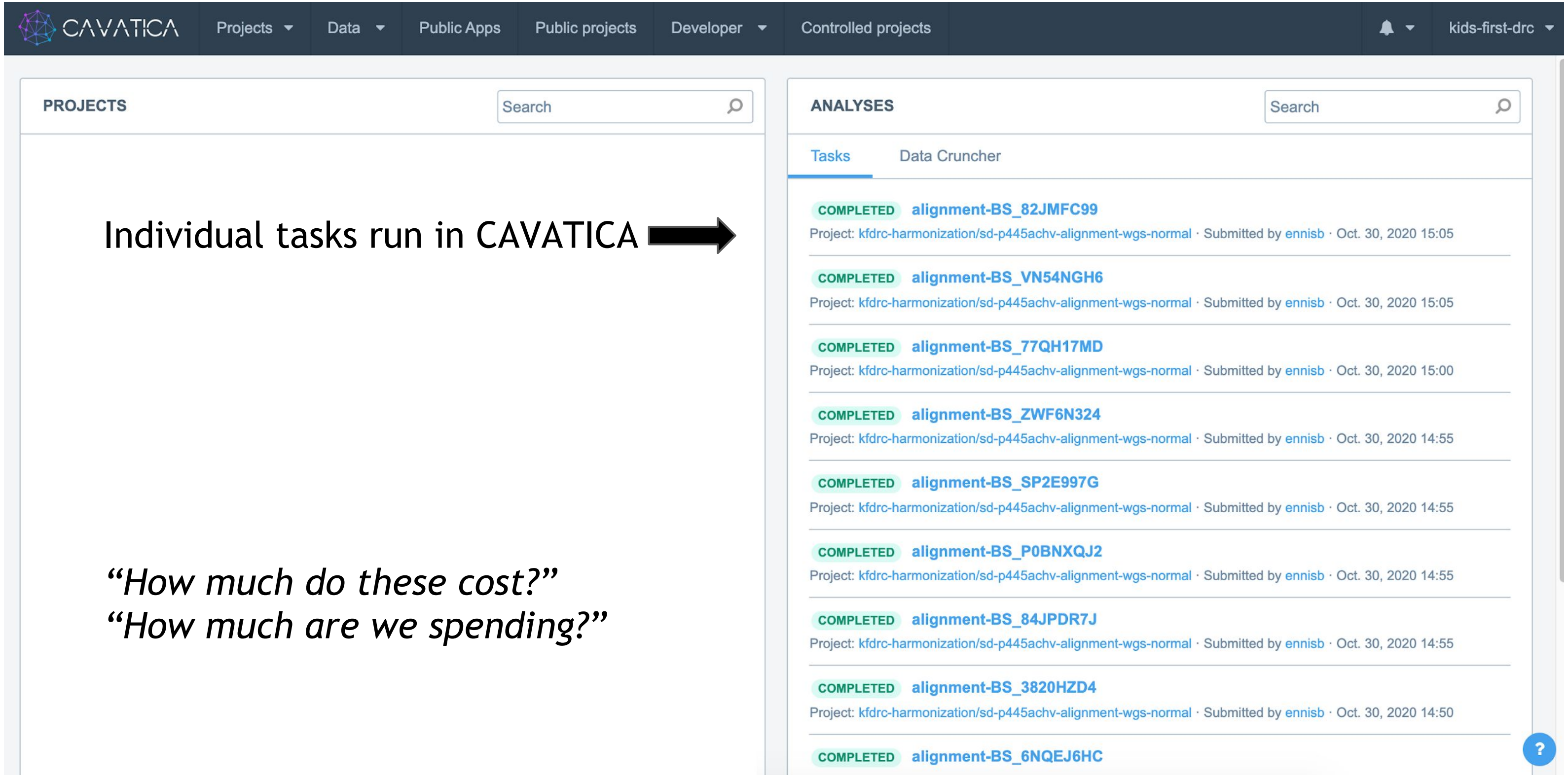
Runtime Estimates

- Single 5 GB gVCF Input: 90 Minutes & $2.25
- Trio of 6 GB gVCFs Input: 240 Minutes & $3.25

Kids First DRC Cloud Credit Lessons Learned

- In brief: users don’t have a strong sense of costs.
    - These might not be an issue on local clusters at their institutions.
    - Have access to spending, but no regular updates or alerts on balances.
- Adjustment from Pilot 1 to Pilot 2: give users more autonomy
    - Pilot 1: KFDRC manages financials as owners of the billing group
    - Pilot 2: User groups manage financials as owners of the billing group

Kids First DRC Cloud Credit Lessons Learned



Individual tasks run in CAVATICA →

COMPLETED alignment-BS_82JMFC99
Project: kfdrc-harmonization/sd-p445achv-alignment-wgs-normal · Submitted by ennisb · Oct. 30, 2020 15:05

COMPLETED alignment-BS_VN54NGH6
Project: kfdrc-harmonization/sd-p445achv-alignment-wgs-normal · Submitted by ennisb · Oct. 30, 2020 15:05

COMPLETED alignment-BS_77QH17MD
Project: kfdrc-harmonization/sd-p445achv-alignment-wgs-normal · Submitted by ennisb · Oct. 30, 2020 15:00

COMPLETED alignment-BS_ZWF6N324
Project: kfdrc-harmonization/sd-p445achv-alignment-wgs-normal · Submitted by ennisb · Oct. 30, 2020 14:55

COMPLETED alignment-BS_SP2E997G
Project: kfdrc-harmonization/sd-p445achv-alignment-wgs-normal · Submitted by ennisb · Oct. 30, 2020 14:55

COMPLETED alignment-BS_P0BNXQJ2
Project: kfdrc-harmonization/sd-p445achv-alignment-wgs-normal · Submitted by ennisb · Oct. 30, 2020 14:55

COMPLETED alignment-BS_84JPDR7J
Project: kfdrc-harmonization/sd-p445achv-alignment-wgs-normal · Submitted by ennisb · Oct. 30, 2020 14:55

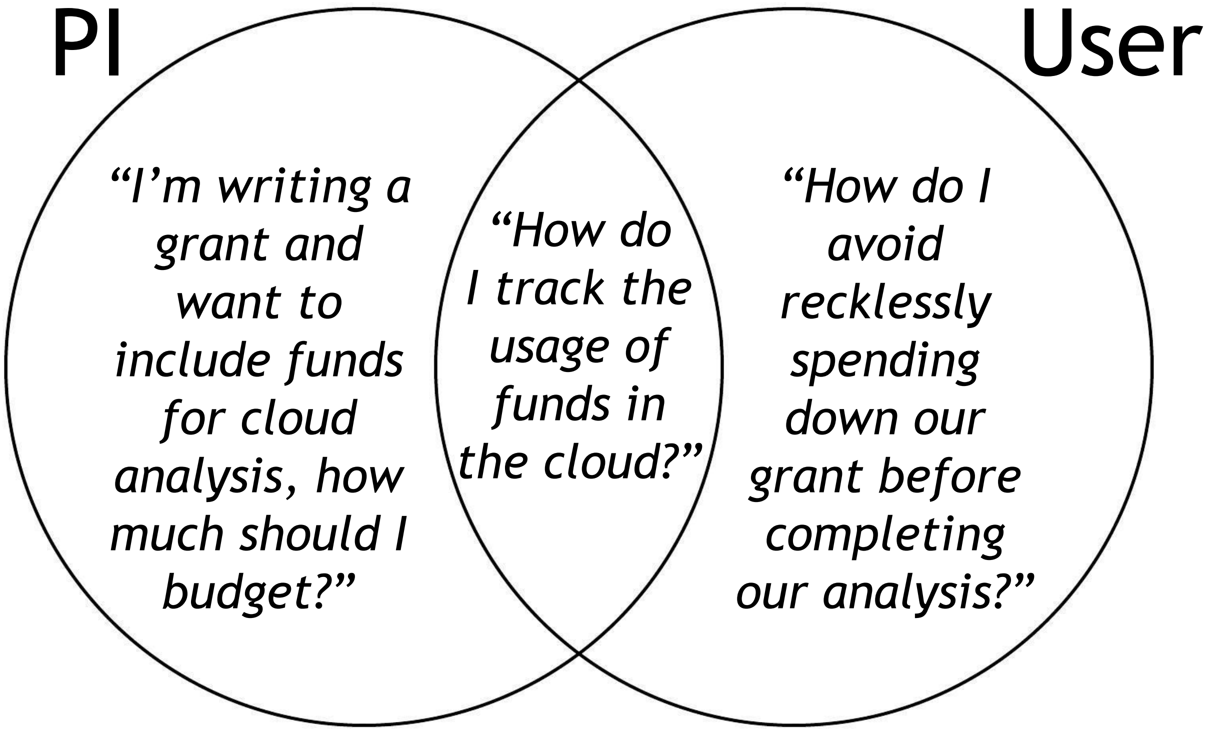
COMPLETED alignment-BS_3820HZD4
Project: kfdrc-harmonization/sd-p445achv-alignment-wgs-normal · Submitted by ennisb · Oct. 30, 2020 14:50

COMPLETED alignment-BS_6NQEJ6HC

“How much do these cost?”
“How much are we spending?”

Kids First DRC Cloud Credit Lessons Learned

Need two messages for two groups of people.



PI

User

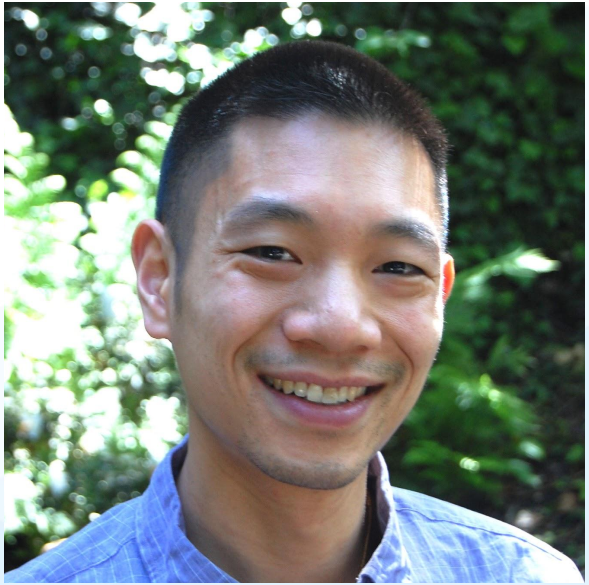
"I'm writing a grant and want to include funds for cloud analysis, how much should I budget?"

"How do I track the usage of funds in the cloud?"

"How do I avoid recklessly spending down our grant before completing our analysis?"

AnVIL
Cloud Costs:
Current Practices

Frederick Tan
Carnegie Institution / JHU



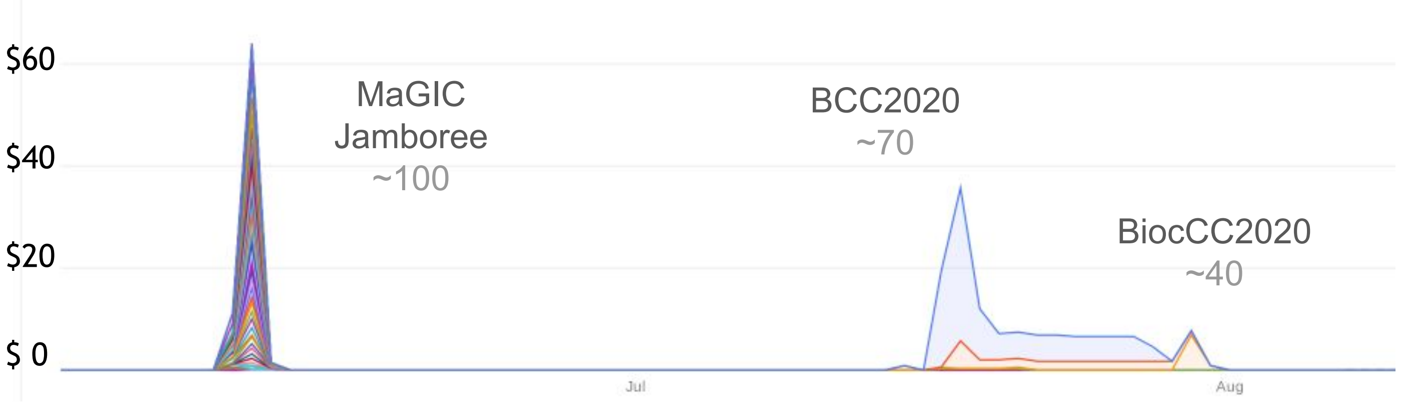
Barriers to Entry

Spending First Dollar

Estimating Costs

Reporting Charges

Writing Into Grants

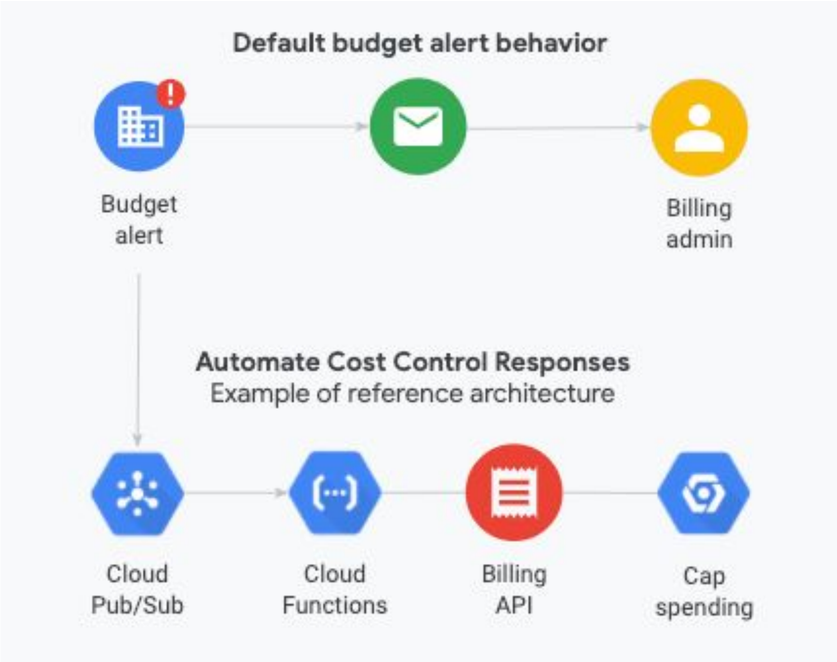


Cost for NIH STRIDES to fund workshops

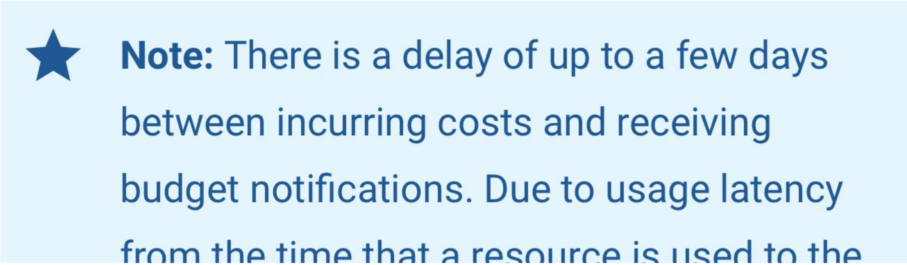
AnVIL

More information at datascience.nih.gov/strides

Estimating Costs - Disable Billing


Default budget alert behavior
Budget alert
Billing admin
Automate Cost Control Responses
Example of reference architecture
Cloud Pub/Sub
Cloud Functions
Billing API
Cap spending

```
# Disable billing
if cost_amount > budget_amount:
    body = {'billingAccountName': ''}
    projects.updateBillingInfo(
      name=project_name,
      body=body).execute()
```

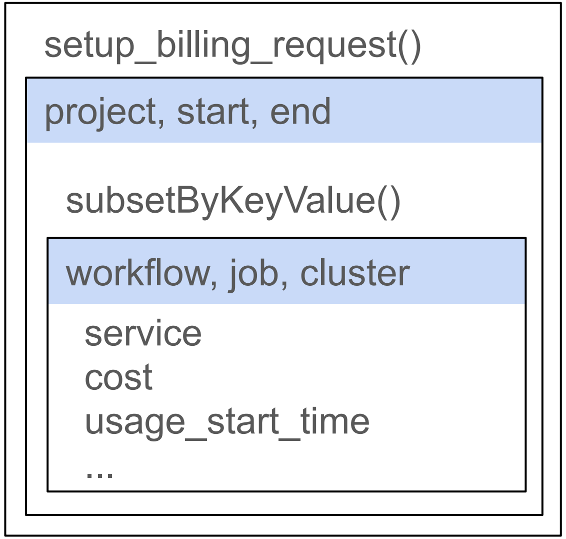


★ **Note:** There is a delay of up to a few days between incurring costs and receiving budget notifications. Due to usage latency from the time that a resource is used to the

More information at cloud.google.com/billing/docs/how-to/notify

Reporting Charges - AnVILBilling

Billing Data



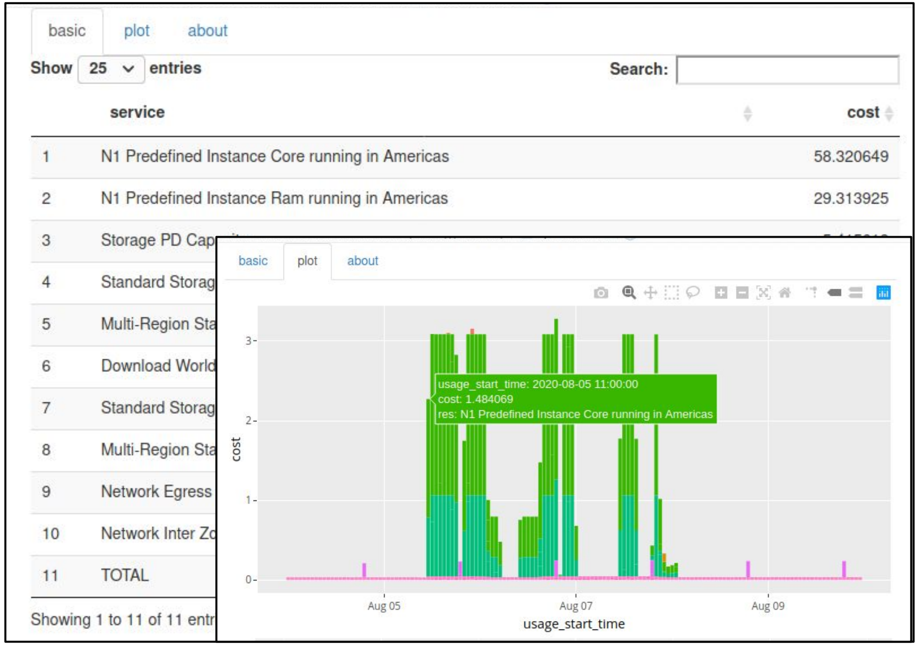
setup_billing_request()

project, start, end

subsetByKeyValue()

workflow, job, cluster

service
cost
usage_start_time
...



	service	cost
1	N1 Predefined Instance Core running in Americas	58.320649
2	N1 Predefined Instance Ram running in Americas	29.313925
3	Storage PD Cap	
4	Standard Storag	
5	Multi-Region Sta	
6	Download World	
7	Standard Storag	
8	Multi-Region Sta	
9	Network Egress	
10	Network Inter Zo	
11	TOTAL	

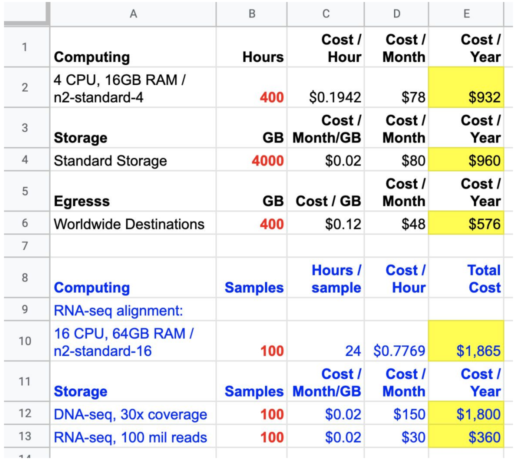
Showing 1 to 11 of 11 entr

usage_start_time: 2020-08-05 11:00:00
cost: 1.484069
res: N1 Predefined Instance Core running in Americas

More information at bioconductor.org/packages/AnVILBilling

Writing Into Grants - CC BY MOOC

Templates



	A	B	C	D	E
1	**Computing**	**Hours**	**Cost / Hour**	**Cost / Month**	**Cost / Year**
2	4 CPU, 16GB RAM / n2-standard-4	400	$0.1942	$78	$932
3	**Storage**	**GB**	**Cost / Month/GB**	**Cost / Month**	**Cost / Year**
4	Standard Storage	4000	$0.02	$80	$960
5	**Egresss**	**GB**	**Cost / GB**	**Cost / Month**	**Cost / Year**
6	Worldwide Destinations	400	$0.12	$48	$576
7					
8	**Computing**	**Samples**	**Hours / sample**	**Cost / Hour**	**Total Cost**
9	RNA-seq alignment:				
10	16 CPU, 64GB RAM / n2-standard-16	100	24	$0.7769	$1,865
11	**Storage**	**Samples**	**Cost / Month/GB**	**Cost / Month**	**Cost / Year**
12	DNA-seq, 30x coverage	100	$0.02	$150	$1,800
13	RNA-seq, 100 mil reads	100	$0.02	$30	$360

Estimate Costs
Budget Justification
Facilities & Equipment
IRB Application

Leanpub

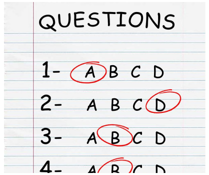
Cloud Computing
Cloud Costs
Estimating Costs
Monitoring Costs
Managing Costs
Funding Sources

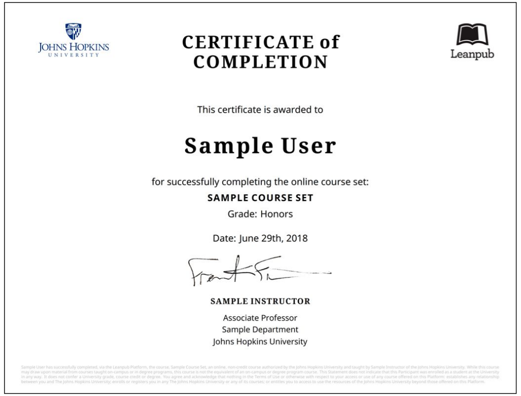
CC BY

Google Slides

Automate w/ Ari →

YouTube





JOHNS HOPKINS UNIVERSITY

Leanpub

CERTIFICATE of COMPLETION

This certificate is awarded to

Sample User

for successfully completing the online course set:

SAMPLE COURSE SET

Grade: Honors

Date: June 29th, 2018

SAMPLE INSTRUCTOR

Associate Professor
Sample Department
Johns Hopkins University

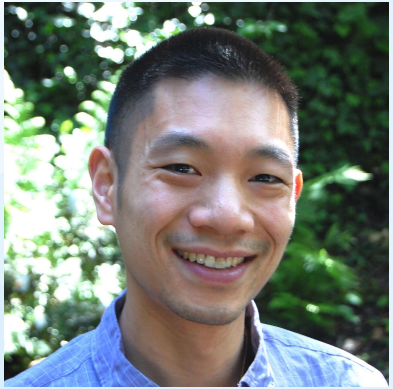
Sample User has successfully completed, via the Leanpub Platform, the course, Sample Course Set, an online, non-credit course authorized by the Johns Hopkins University and taught by Sample Instructor of the Johns Hopkins University. While this course may draw upon material from courses taught on-campus or in degree programs, this course is not the equivalent of an on-campus or degree program course. This Statement does not indicate that this Participant was enrolled as a student at the University in any way. It does not confer a University grade, course credit or degree. You agree and acknowledge that nothing in the Terms of Use or otherwise with respect to your access or use of any course offered on this Platform: establishes any relationship between you and The Johns Hopkins University; enrolls or registers you in any The Johns Hopkins University or any of its courses; or entitles you to access to use the resources of the Johns Hopkins University beyond those offered on this Platform.

AnVIL

More information at anvilproject.org/learn and tinyurl.com/JHUMOOCVIDEO

Cloud Costs: Near Term Improvements

Open Discussion
30 Minutes



Lunch Break

We will resume at 12:30 pm ET.

RAS & Data Sharing Between Platforms

Jiaqi Liu U. Chicago
Brian O'Connor Broad

Overview

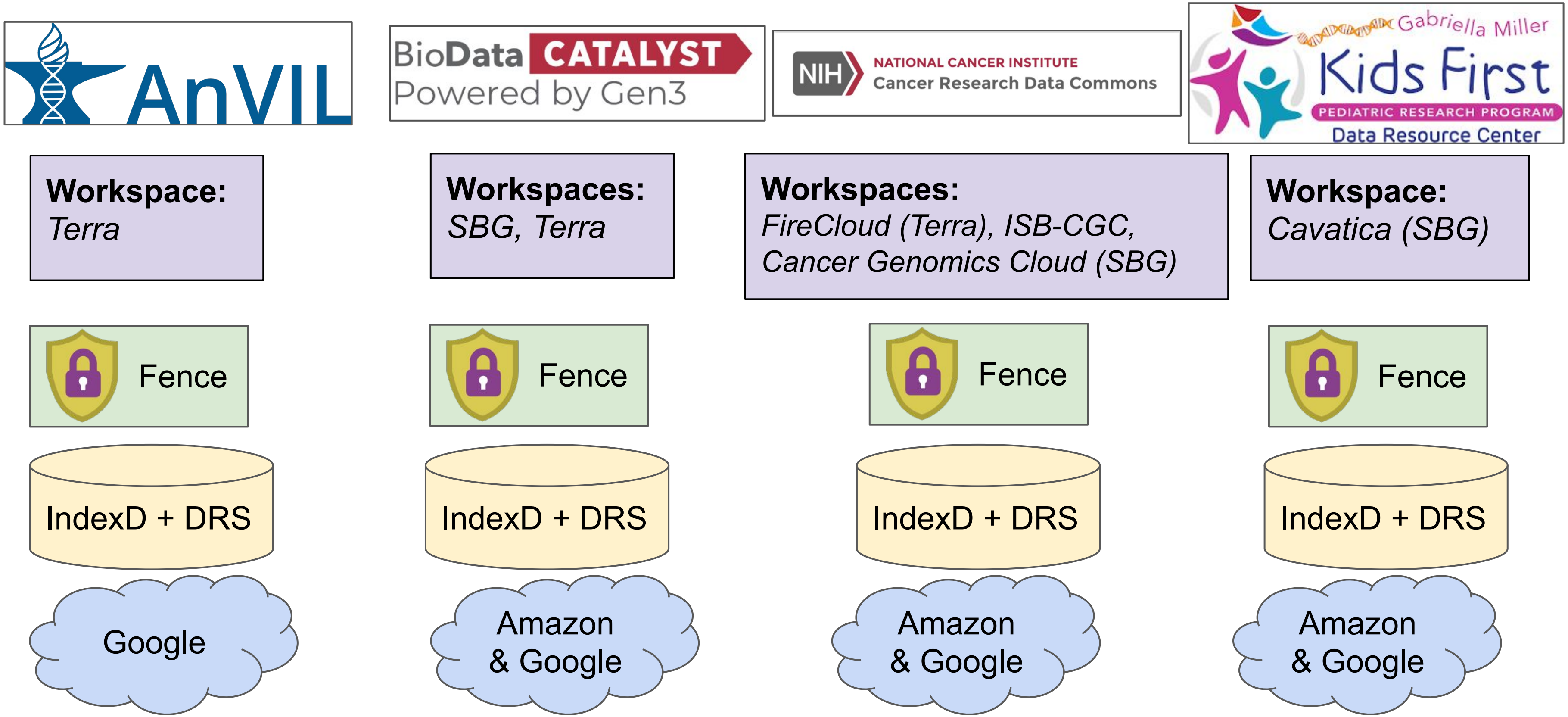
- 10 min - Jiaqi and Brian will give an overview of RAS and Data Sharing work
- 5 min - Group will brainstorm on areas of RAS & Data Sharing to focus on in 2021
- 25 min - Group discussion, deep dive on ~3 of the most popular topics

RAS Integration

Jiaqi Liu U. Chicago

Commonality of Framework Services

The commonality of AnVIL, BD Catalyst, CRDC, and Kids First framework services facilitates RAS adoption



AnVIL

Workspace: *Terra*

Fence

IndexD + DRS

Google

BioData CATALYST Powered by Gen3

Workspaces: *SBG, Terra*

Fence

IndexD + DRS

Amazon & Google

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

Workspaces: *FireCloud (Terra), ISB-CGC, Cancer Genomics Cloud (SBG)*

Fence

IndexD + DRS

Amazon & Google

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

Workspace: *Cavatica (SBG)*

Fence

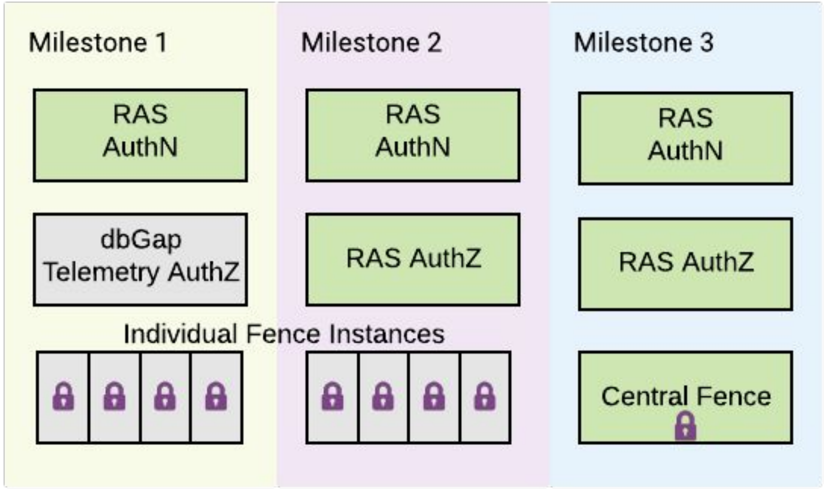
IndexD + DRS

Amazon & Google

Shared Technical RAS Plan - Work Divided into 3 Milestones

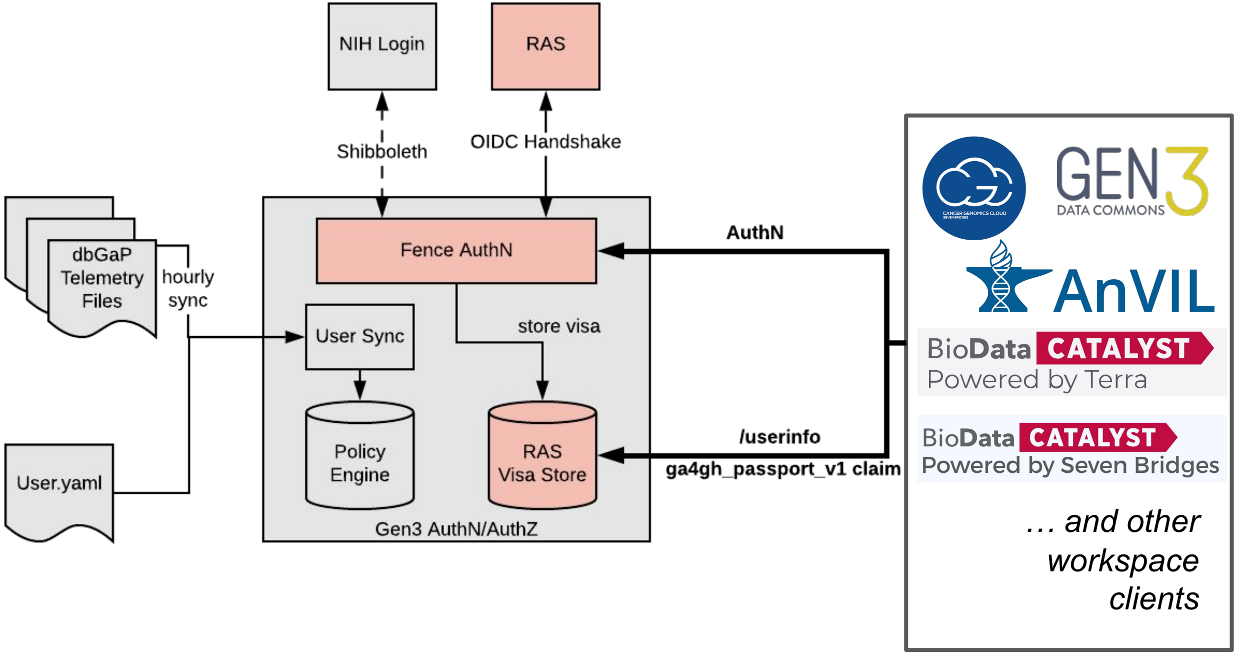
The ***U. Chicago****,* ***RAS****,* ***AnVIL, BDCat, CRDC, and Kids First****, and* ***GA4GH*** *teams have collaborated to bring RAS to our systems.*

- *We worked on a technical document to break down the work: CRDC RAS Integration Proposal v1.4*
- *CRDC-specific but highly applicable to* ***AnVIL****,* ***BD Catalyst****,* ***Kids First DRC***
- *We continue to coordinate in ad hoc meetings, various GA4GH working groups, and the NCPI Systems Interoperation Working Group*



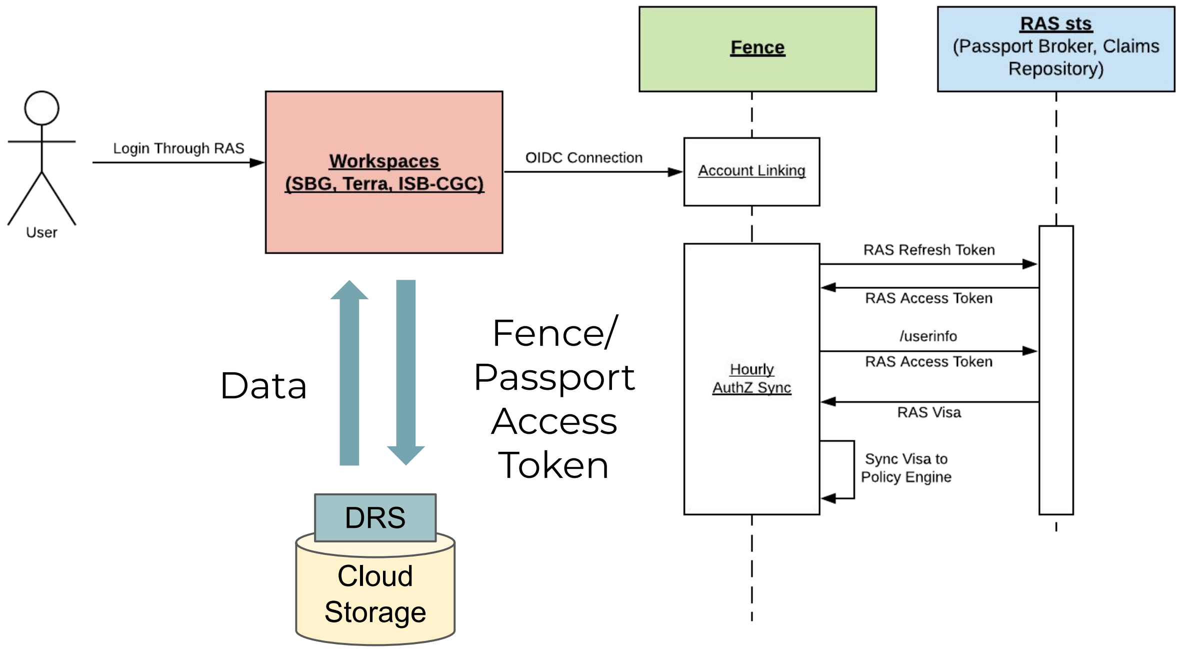
Work is Broken into Milestones:
M1, M2, and M3

RAS Integration - Milestone 1 (Now)



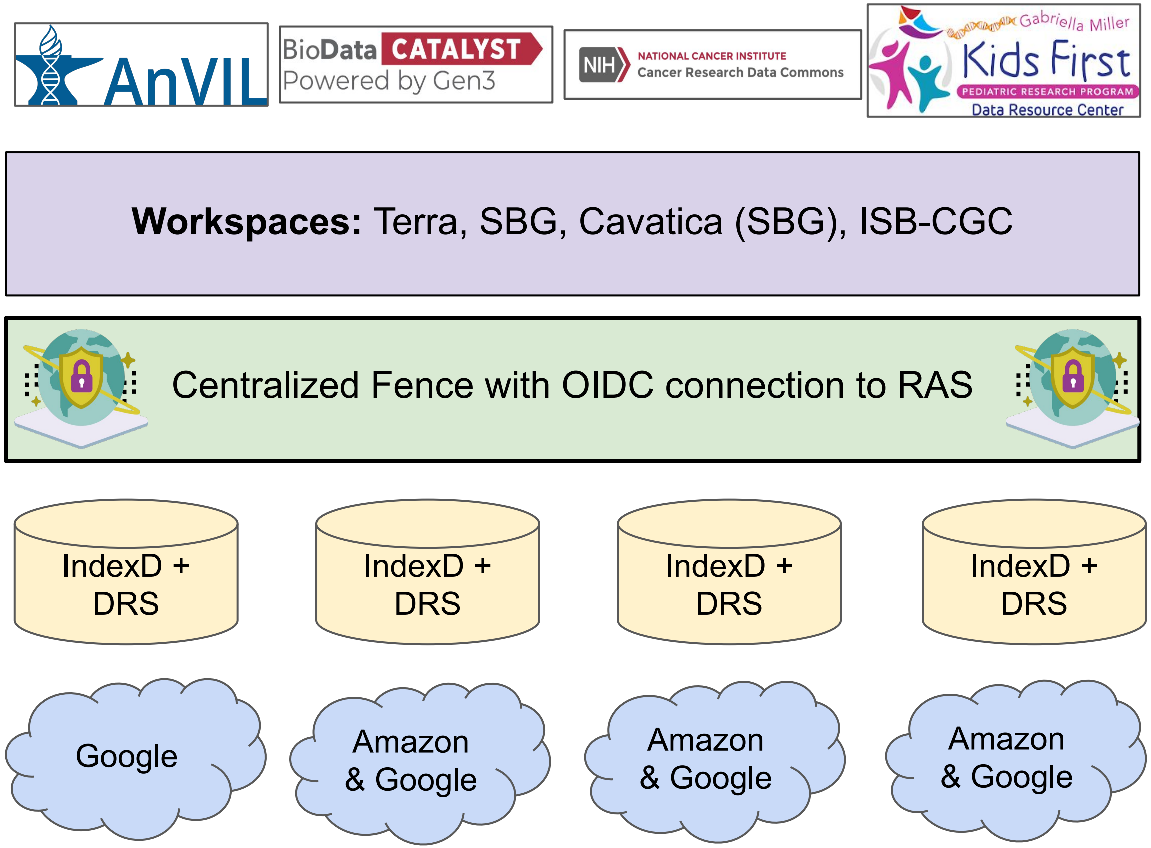
- Use RAS for user login (AuthN)
- Use Cases Enabled
  - Systems use Fence
  - SSO-like experience
  - Passport visas explored for securing derived results and data access UX

RAS Integration - Milestone 2



- Use RAS for Login and Authorization
  - Replace dbGaP telemetry files
- Use Cases Enabled
  - *Previous Milestone 1 use cases*
  - Researcher benefit from *realtime* AuthZ instead of dbGaP whitelists which can lag

RAS Integration - Milestone 3


AnVIL
BioData CATALYST
Powered by Gen3
NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons
Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

Workspaces: Terra, SBG, Cavatica (SBG), ISB-CGC

Centralized Fence with OIDC connection to RAS

IndexD + DRS

IndexD + DRS

IndexD + DRS

IndexD + DRS

Google

Amazon & Google

Amazon & Google

Amazon & Google

- Fully distributed Authorization
  - A“central” Fence, users can access any dataset via single linking event
- Use Cases Enabled
  - *Previous Milestones*
  - SSO with single consent
  - Data across stacks in single linking
  - Provide non-dbGaP/other project access lists via single Fence

RAS - Current Accomplishments

Completed Work:

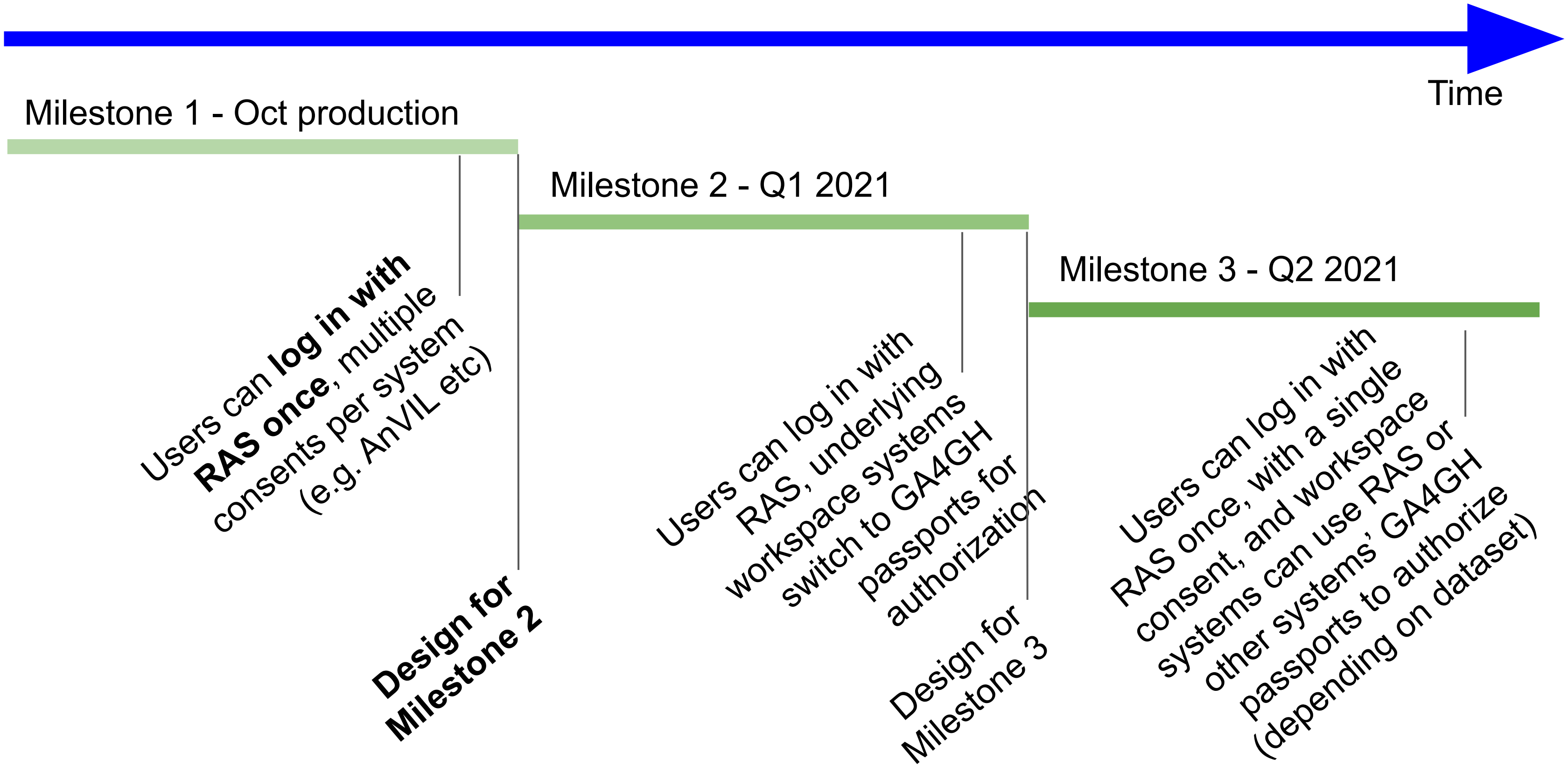
- **Implemented RAS AuthN Support:** As part of RAS Phase 1, U. Chicago has completed the code changes to Fence to use RAS for user login
- **Coordinated Deployment:** U. Chicago and partner stacks have adapted their systems to use the RAS login via Fence, production deployment is happening ***now***
- **Future Work Design:** U. Chicago, Broad, SBG, RAS, and the GA4GH continue to work together on the future milestones for RAS and GA4GH Passports

Continuing Work:

- **Authorization:** U. Chicago will switch to RAS passport visas for Authorization info
- **Continued Technical Milestones:** Milestone 2 and 3, simplifying the user experience
- **GA4GH & Standards:** Laying foundation for future GA4GH Passports, brokers beyond RAS

RAS Next Steps/Timeline

Goal is to have all cloud stacks coordinate and use RAS identically across systems



Time

Milestone 1 - Oct production

Users can **log in with RAS once**, multiple consents per system (e.g. AnVIL etc)

Design for Milestone 2

Milestone 2 - Q1 2021

Users can log in with RAS, underlying workspace systems switch to GA4GH passports for authorization

Design for Milestone 3

Milestone 3 - Q2 2021

Users can log in with RAS once, with a single consent, and workspace systems can use RAS or other systems' GA4GH passports to authorize (depending on dataset)

Data Sharing

Brian O’Connor Broad

What is Data Sharing?

There are many different facets of data sharing… many more than we could fit in a 40 minute conversation.

Some key areas of Data Sharing include:

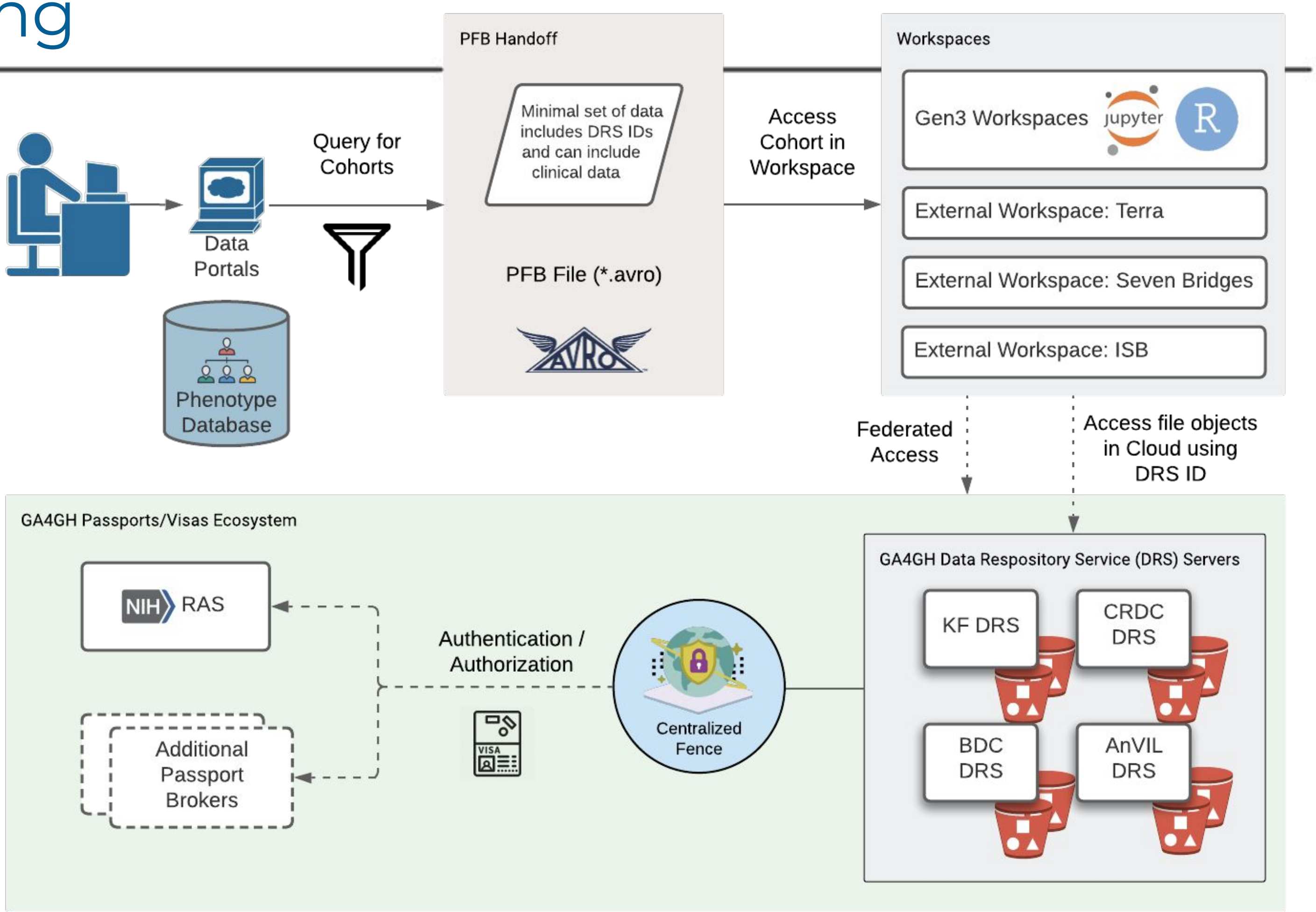
- **Sharing data files between systems**
- **Exchanging data models between systems**
- Common data models across systems
- Search of data and data models
  - Specialty access/search for specific data types (think Beacon API, htsget)
- AuthN/AuthZ, SSO
- BYOD and sharing derived results
- And many more…

I'm going to dive into the first two aspects and report back progress from NCPI Systems Interoperation group over the last year.

Open discussion at the end will help us frame what Data Sharing we should focus on in 2021

Data Sharing

DRS *facilities* ***data file sharing****,* ***PFB*** *facilitates* ***sharing data model + DRS URIs****,* ***RAS*** *gives us a common Auth system for* ***SSO*** *and* ***data access across systems***



Data Portals

Query for Cohorts

Phenotype Database

PFB Handoff

Minimal set of data includes DRS IDs and can include clinical data

PFB File (*.avro)

Access Cohort in Workspace

Workspaces

Gen3 Workspaces

External Workspace: Terra

External Workspace: Seven Bridges

External Workspace: ISB

Federated Access

Access file objects in Cloud using DRS ID

GA4GH Passports/Visas Ecosystem

NIH RAS

Additional Passport Brokers

Authentication / Authorization

Centralized Fence

GA4GH Data Respository Service (DRS) Servers

KF DRS

CRDC DRS

BDC DRS

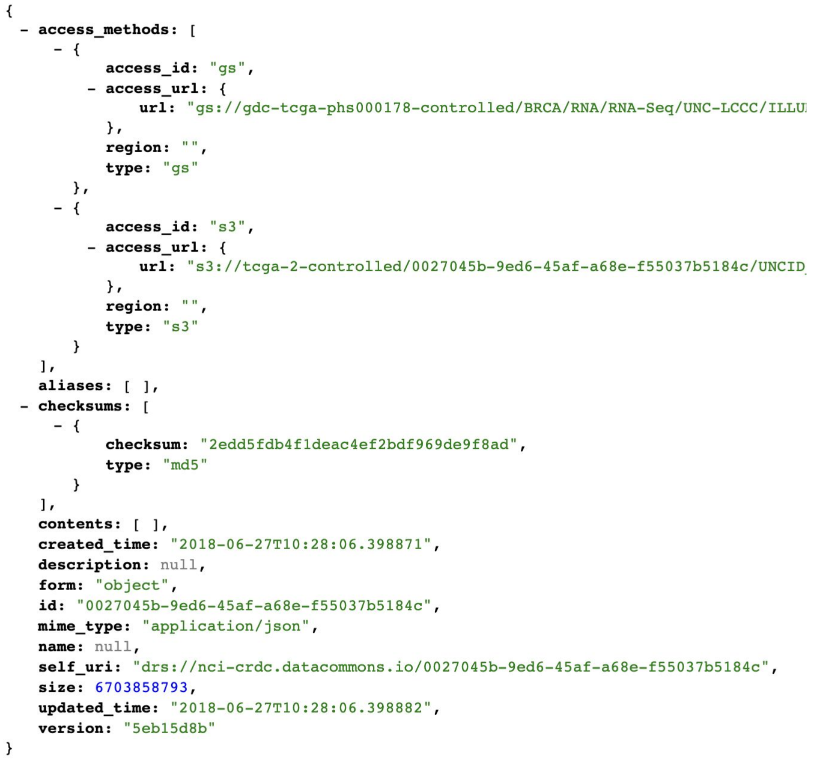
AnVIL DRS

From GA4GH Poster Session

GA4GH DRS 1.1

- The **Data Repository Service (DRS)** API provides a generic interface to data repositories so data consumers, including workflow systems, can access data objects in a single, standard way regardless of where they are stored and how they are managed.
- **DRS 1.1** was released in 2020 and it added support for **compact identifiers** which was key to the DRS servers supporting AnVIL, BD Catalyst, CRDC, and Kids First
- See [DRS 1.1 Transition within NCPI]()

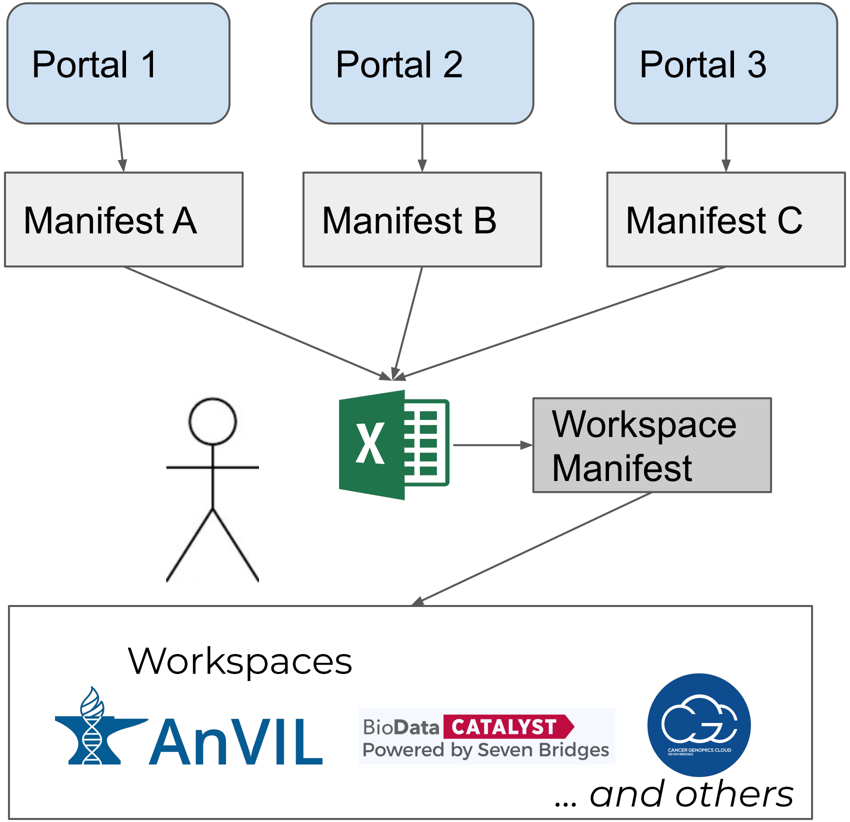
drs://dg.4DFC:0027045b-9ed6-45af-a68e-f55037b5184c



```
{
  - access_methods: [
      - {
            access_id: "gs",
          - access_url: {
                url: "gs://gdc-tcga-phs000178-controlled/BRCA/RNA/RNA-Seq/UNC-LCCC/ILLU
            },
            region: "",
            type: "gs"
        },
      - {
            access_id: "s3",
          - access_url: {
                url: "s3://tcga-2-controlled/0027045b-9ed6-45af-a68e-f55037b5184c/UNCID
            },
            region: "",
            type: "s3"
        }
    ],
    aliases: [ ],
  - checksums: [
      - {
            checksum: "2edd5fdb4f1deac4ef2bdf969de9f8ad",
            type: "md5"
        }
    ],
    contents: [ ],
    created_time: "2018-06-27T10:28:06.398871",
    description: null,
    form: "object",
    id: "0027045b-9ed6-45af-a68e-f55037b5184c",
    mime_type: "application/json",
    name: null,
    self_uri: "drs://nci-crdc.datacommons.io/0027045b-9ed6-45af-a68e-f55037b5184c",
    size: 6703858793,
    updated_time: "2018-06-27T10:28:06.398882",
    version: "5eb15d8b"
}
```

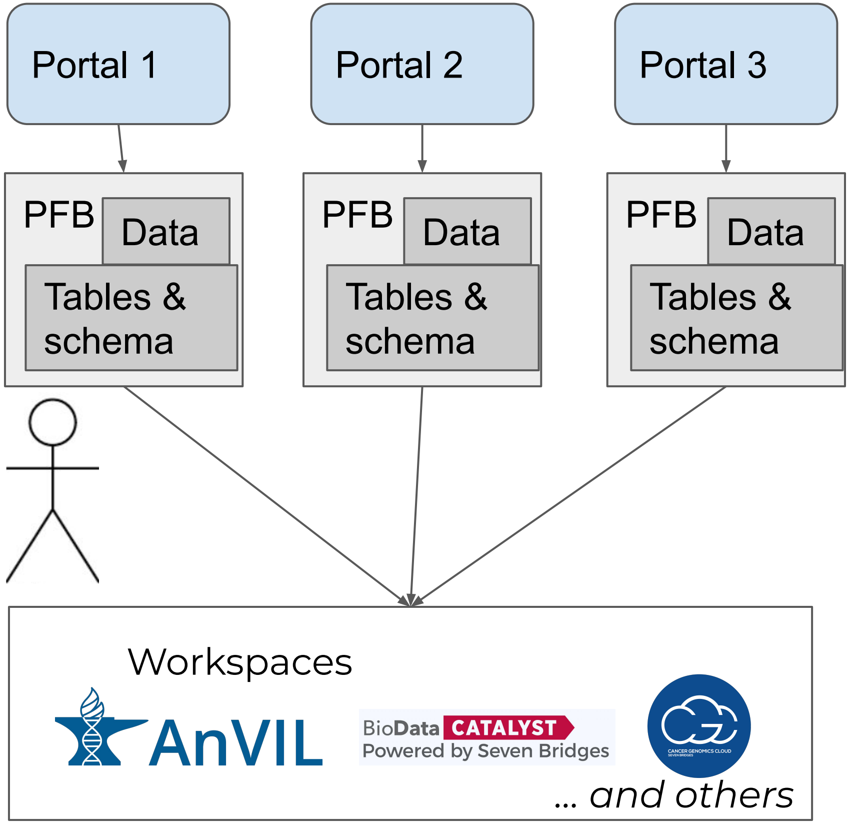
PFB

- Many portals, a smaller number of workspace environments
- DRS 1.1 useful for workspace environments to access data for compute
- But how do you find data in the multitude of portals and bring those results to a workspace environment?
- **Before PFB…**



PFB

- Many portals, a smaller number of workspace environments
- DRS 1.1 useful for workspace environments to access data for compute
- But how do you find data in the multitude of portals and bring those results to a workspace environment?
- **After PFB…**
  - PFB handoff description
  - PFB "light"



FunRetro for Discussion

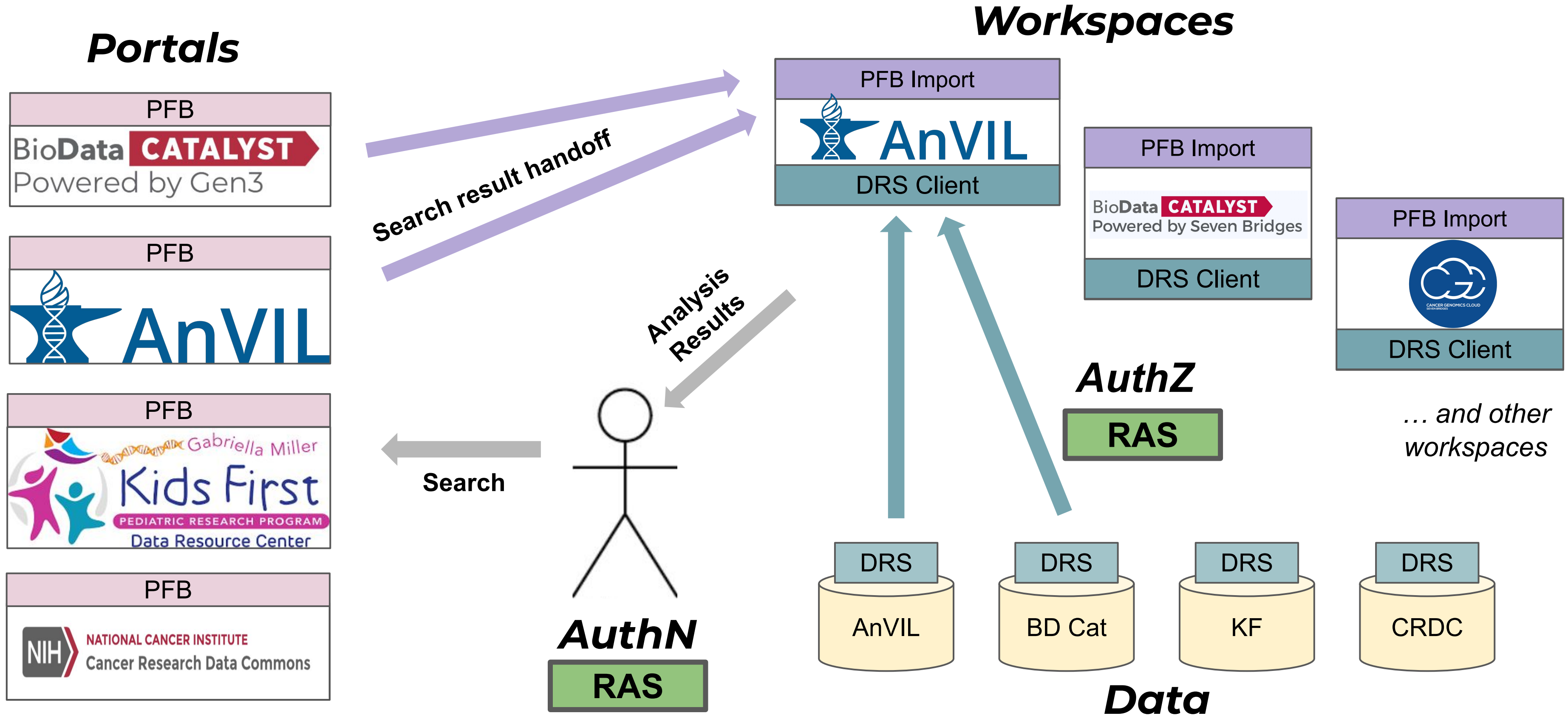
- 5 min - Group will brainstorm on areas of RAS & Data Sharing to focus on in 2021
  - *Let's add discussion ideas/vote in this **FunRetro Board***

- ***Top Topics:***
  - ***Search for data across stacks***
  - ***Policies needed for sharing data between IC stacks***
  - ***Handoff of search results, PFB, FHIR, or both? Something else?***
  - ***BYOD: how does "bring your own data" work across platforms? Is there any difference between sharing data between the IC stacks if its from a canonical project or if it's provided by a researcher upload?***

- 25 min - Group discussion, deep dive on ~3 of the most popular topics
  - What is the challenge?
  - What solutions should be examined?
  - Who needs to work on this topic?
  - When does the topic need to be solved by?
  - ...

Data Sharing

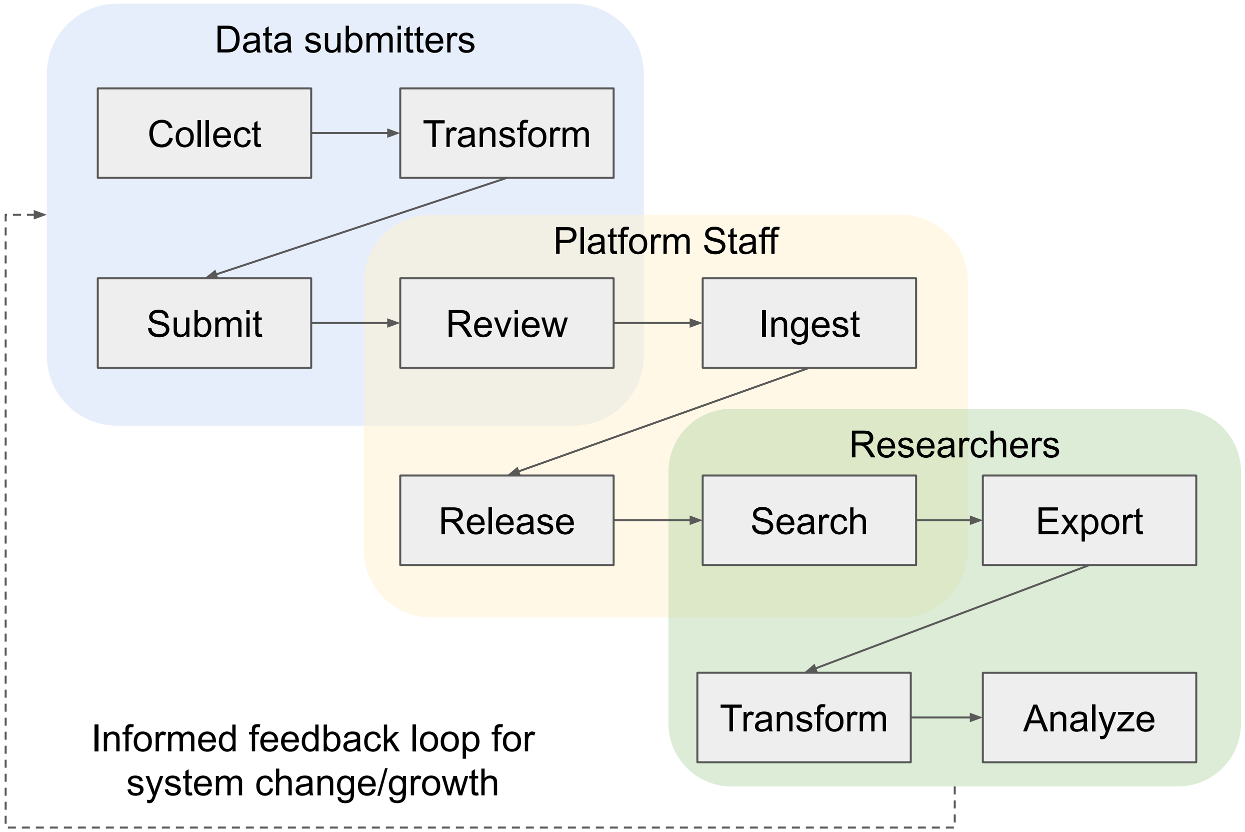
***DRS** facilities **data file sharing**, **PFB** facilitates **sharing data model + DRS URIs**,*
RAS** gives us a common Auth system for **SSO** and **data access across systems



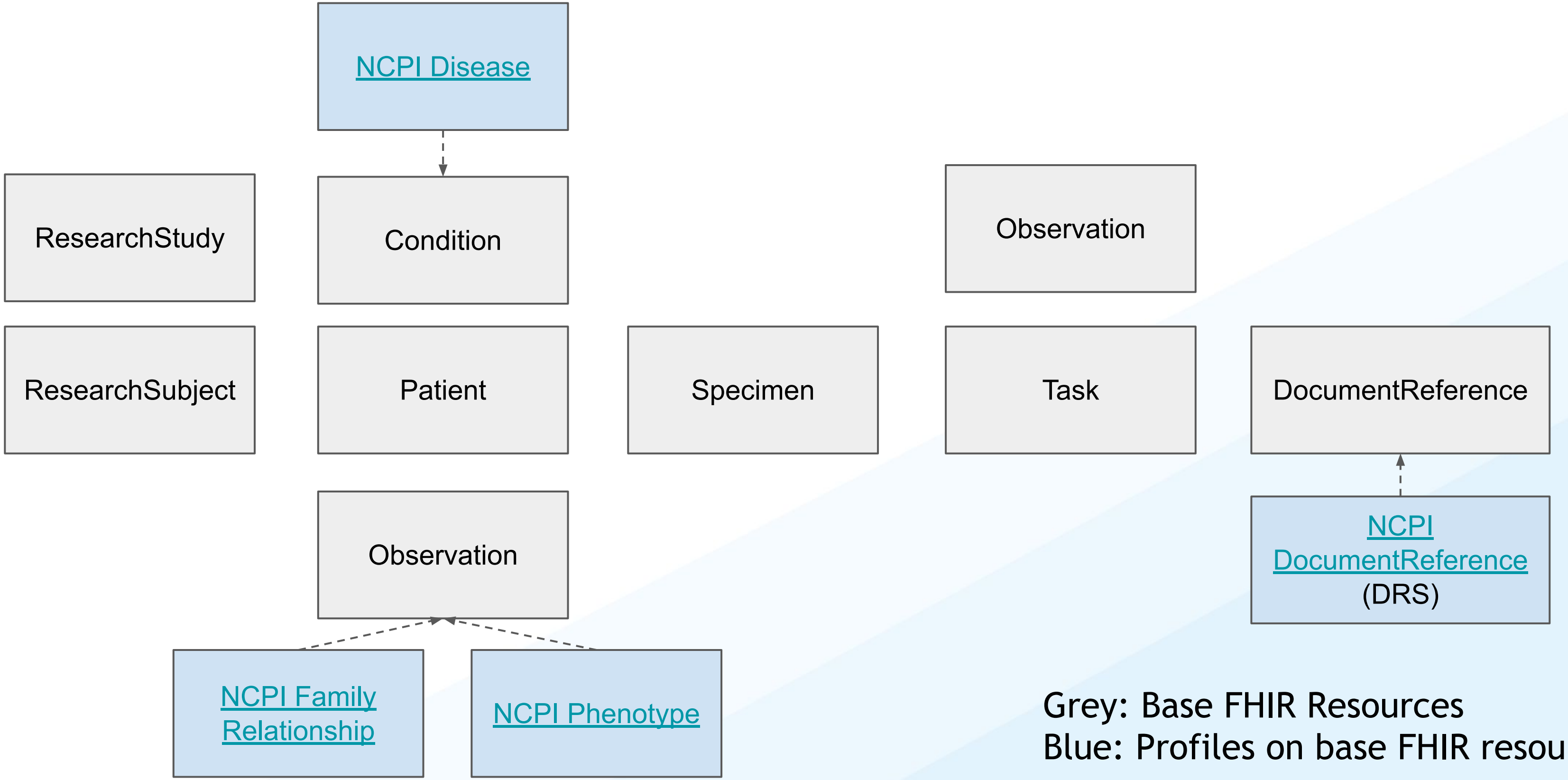
Clinical / Phenotypic Data: Representation, Search, and Handoff

Allison Heath, PhD
Director of Technology @ D3b, CHOP
Robert Carroll, PhD
Assistant Professor, VUMC

Clinical Data Flow Landscape

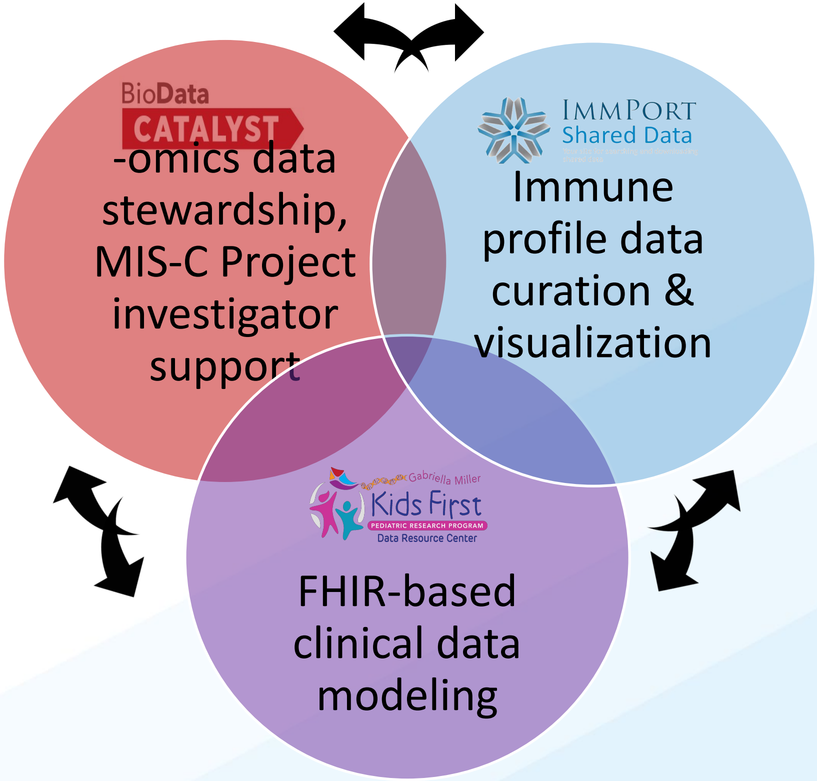


Representation Experience with FHIR WG



Grey: Base FHIR Resources
Blue: Profiles on base FHIR resources

MIS-C as Interoperability Use Case



BioData CATALYST

-omics data stewardship, MIS-C Project investigator support

IMMPORT Shared Data

Immune profile data curation & visualization

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

FHIR-based clinical data modeling

Ongoing Longitudinal data

NIAID/PRISM

- Determine the proportion with SARS-CoV-2 related death, rehospitalization or ongoing major complications at 12 months after presentation
- Determine immunologic mechanisms, immune signatures and predictive biomarkers associated with disease phenotypes

NICHD/POPS

- Study the influence of genetic factors, metabolic, protein profiles on therapeutic exposure and response
- Evaluate PK/PD of understudied drugs in hospitalized children with SARS-CoV-2 related illness
- Establish drug safety profile and adverse events with specific cardiac or neurologic impact

NHLBI/MUSIC

- Characterize the occurrence and time course of coronary artery involvement and ventricular dysfunction
- Characterize the occurrence and time course of non-cardiac organ dysfunction, inflammation and major medical events

Longitudinal

MIS-C Interoperability Across the Landscape

- Collection
  - CRFs and CDEs
  - EHR
- “Platform staff” is across platforms
  - How to view incoming/provisional data?
- Feeding multiple downstream platforms/tools
  - Initial handoffs?
  - Awareness of new data availability?
- Intersecting with other existing datasets
  - Known or search?

Quick Break

We will resume at 2:00 pm ET.

Genomic Analysis Use Cases & Working Groups

Jack DiGiovanna[1] & Michael Schatz[2,3]
[1]Program Director - Seven Bridges
[2]Program Director - AnVIL
[3]Bloomberg Distinguished Associate Professor - Johns Hopkins

Genomic Analysis

You have a great, testable hypothesis

You are authorized to use a large dataset

You have funding for yourself and cloud usage

You have access to a powerful cloud platform(s)

You already have the tools you need running locally

You may already be a winner and are moments away from analyzing tens of thousands of samples!

Is your pipeline wrapped in a workflow description language that the platform understands?

This has been a blocker for many researchers

Some users are comfortable to wrap tools, other users are not

(*before NCPI; but still somewhat today*) Certain datasets were only available on one platform, users with tools in *workflow_language_**a*** could only analyze that data with *workflow_language_**b.***

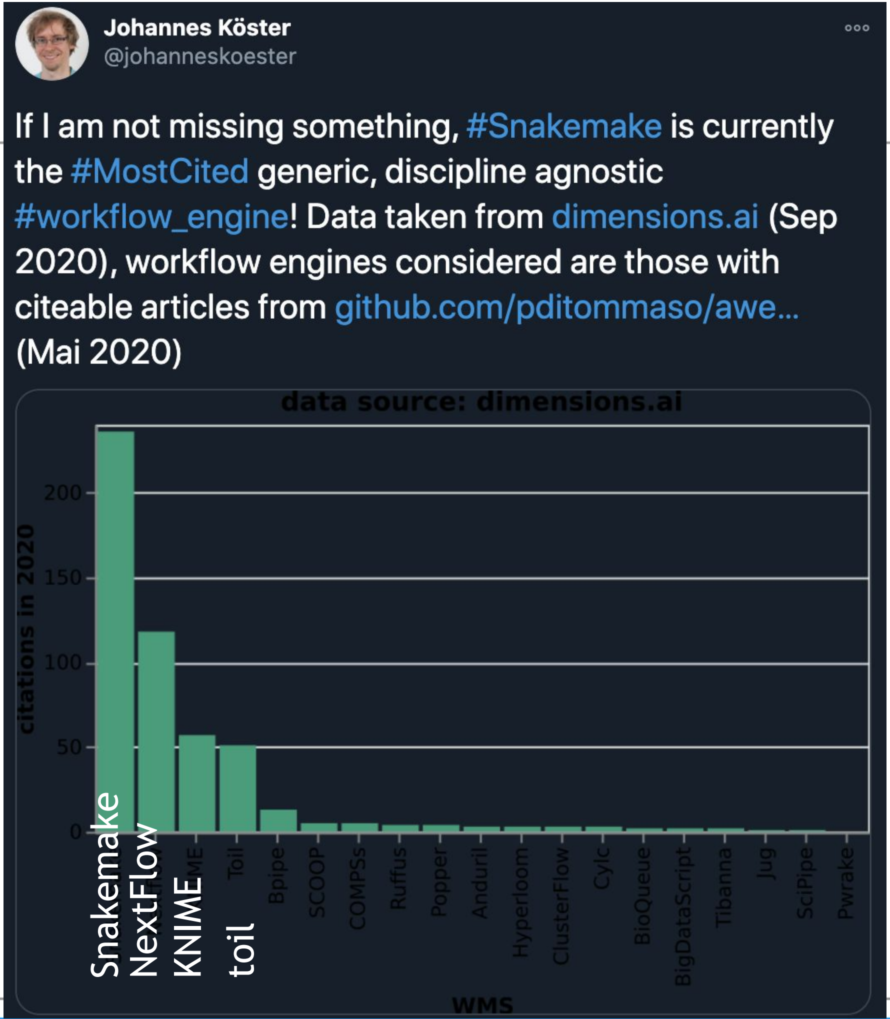
How to get your research done

1. Search the available pipelines on *Platform_of_Choice* (or *Platform_of_Necessity*)
2. Search pipeline repos, e.g. dockstore.org
3. Reach out to the Support teams at Platform
4. Reach out to your IC Program Officer, they know magic

We'd like the outcome of this session to be a concrete plan of how to leverage the NCPI work to get your research done. Some possibilities

5. Bring data over to your platform to process with *your_workflow_language*
6. Use existing tools on other platform in *other_workflow_language*

Do steps 5-6 work? What other strategies work? What other blockers do you face?



Johannes Köster
@johanneskoester

If I am not missing something, #Snakemake is currently the #MostCited generic, discipline agnostic #workflow_engine! Data taken from dimensions.ai (Sep 2020), workflow engines considered are those with citeable articles from github.com/pditommaso/awe... (Mai 2020)

Not on the graph
Cromwell (**WDL**)
Most **CWL** executors
Galaxy

CWL, WDL, Snakemake & Galaxy WF

The Common Workflow Language (CWL) is an open standard for describing analysis workflows and tools in a way that makes them portable and scalable across a variety of software and hardware environments, from workstations to cluster, cloud, and high performance computing (HPC) environments. https://www.commonwl.org/

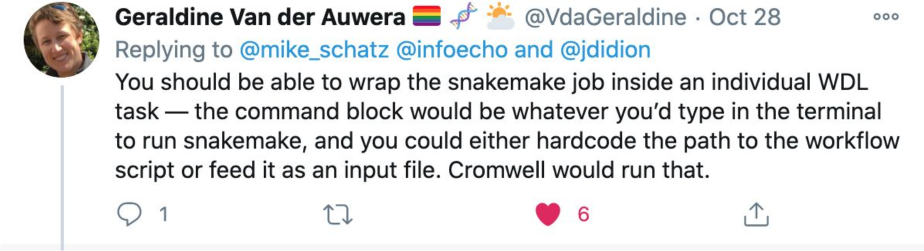
The Workflow Description Language (WDL) is a way to specify data processing workflows with a human-readable and -writeable syntax. WDL makes it straightforward to define analysis tasks, chain them together in workflows, and parallelize their execution. The language makes common patterns simple to express, while also admitting uncommon or complicated behavior; and strives to achieve portability not only across execution platforms, but also different types of users. https://openwdl.org/

The Snakemake workflow management system is a tool to create reproducible and scalable data analyses. Workflows are described via a human readable, Python based language. They can be seamlessly scaled to server, cluster, grid and cloud environments, without the need to modify the workflow definition. Finally, Snakemake workflows can entail a description of required software, which will be automatically deployed to any execution environment. https://snakemake.readthedocs.io/en/stable/

A Galaxy workflow is a series of tools and dataset actions that run in sequence as a batch operation. Workflows can be generated quickly from the analysis already completed in a history. Workflow can be reused over and over, not only reducing tedious work, but enhancing reproducibility by applying the same exact methods to all of your data.
https://galaxyproject.org/learn/advanced-workflow/

Workflow Interoperability

- Searching, storing, and publishing using multiple workflow languages (e.g. Dockstore)
- Single node solutions for launching a workflow written in language X within workflow engine Y?



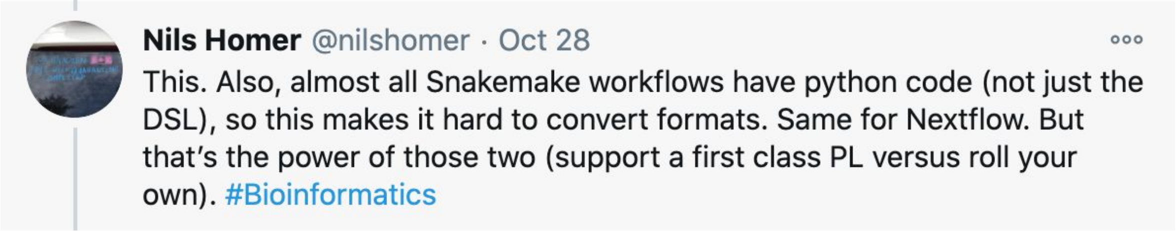
Geraldine Van der Auwera 🏳️‍🌈 🧬 ⛅ @VdaGeraldine · Oct 28

Replying to @mike_schatz @infoecho and @jdidion

You should be able to wrap the snakemake job inside an individual WDL task — the command block would be whatever you'd type in the terminal to run snakemake, and you could either hardcode the path to the workflow script or feed it as an input file. Cromwell would run that.

💬 1 🔁 ❤ 6

- Converter from Workflow X to Workflow Y?

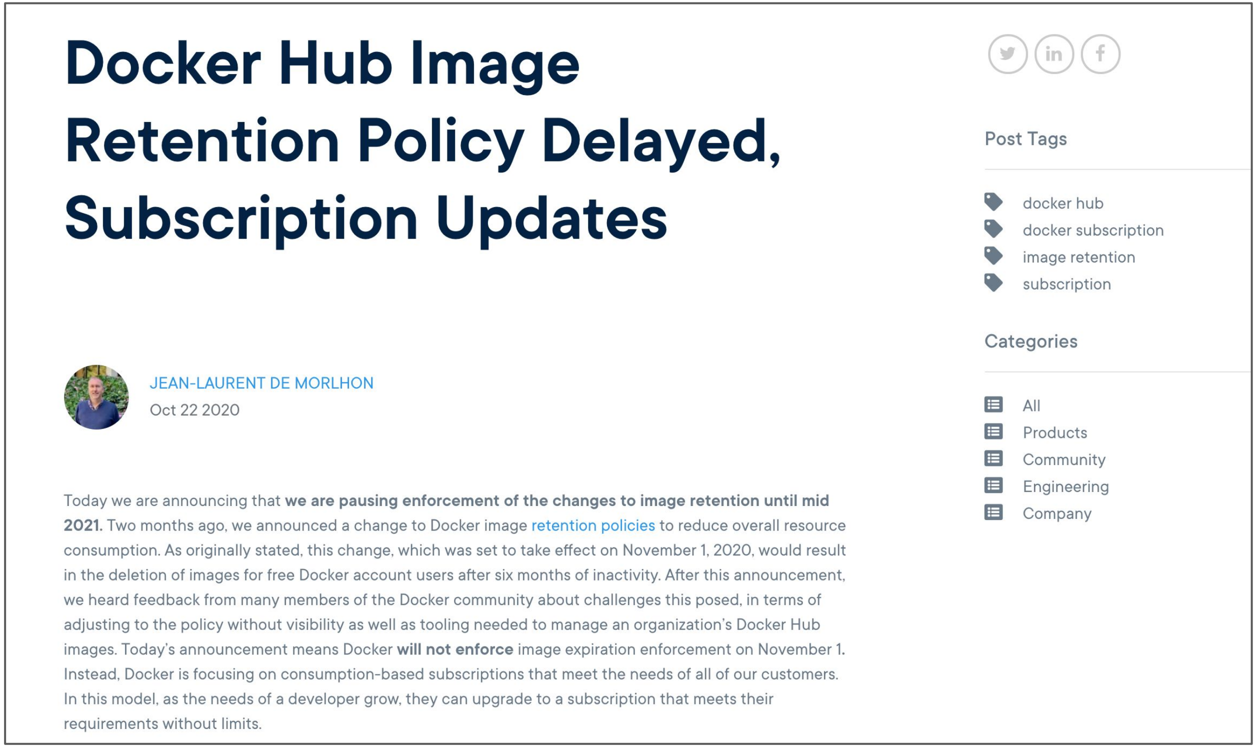


Nils Homer @nilshomer · Oct 28

This. Also, almost all Snakemake workflows have python code (not just the DSL), so this makes it hard to convert formats. Same for Nextflow. But that's the power of those two (support a first class PL versus roll your own). #Bioinformatics

- Better GUI/Lint/Debugger support

Containers & Binary Management



Docker Hub Image Retention Policy Delayed, Subscription Updates

JEAN-LAURENT DE MORLHON

Oct 22 2020

Today we are announcing that **we are pausing enforcement of the changes to image retention until mid 2021.** Two months ago, we announced a change to Docker image retention policies to reduce overall resource consumption. As originally stated, this change, which was set to take effect on November 1, 2020, would result in the deletion of images for free Docker account users after six months of inactivity. After this announcement, we heard feedback from many members of the Docker community about challenges this posed, in terms of adjusting to the policy without visibility as well as tooling needed to manage an organization's Docker Hub images. Today's announcement means Docker **will not enforce** image expiration enforcement on November 1. Instead, Docker is focusing on consumption-based subscriptions that meet the needs of all of our customers. In this model, as the needs of a developer grow, they can upgrade to a subscription that meets their requirements without limits.

Post Tags

- docker hub
- docker subscription
- image retention
- subscription

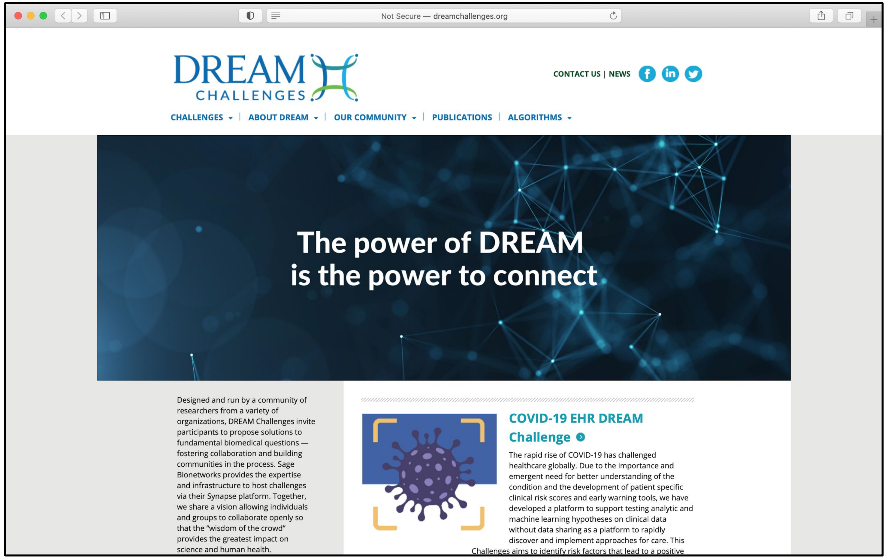
Categories

- All
- Products
- Community
- Engineering
- Company

While our immediate issues have been pushed back until mid-2021, these issues could resurface at any time and on any product.

Should NCPI work towards developing an alternative resource for hosted binary & container management?

Defining Best Practices



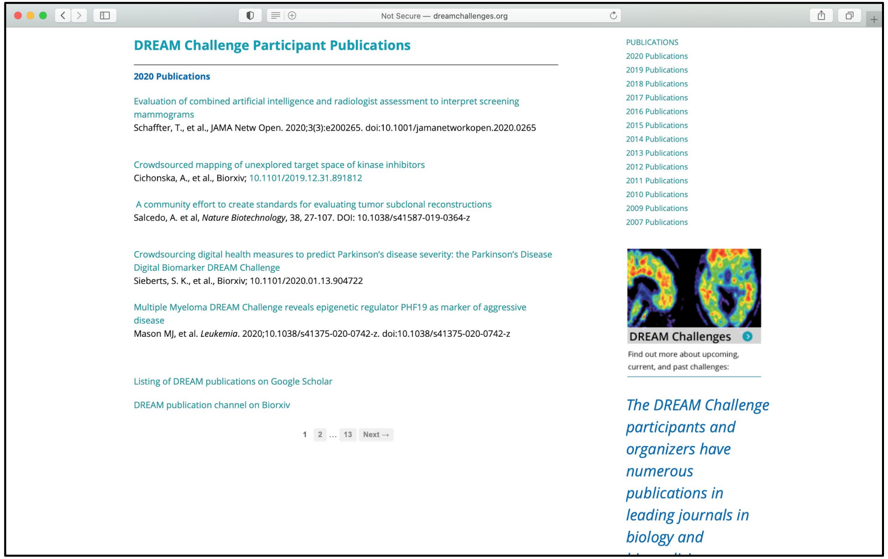
Not Secure — dreamchallenges.org
DREAM CHALLENGES
CONTACT US | NEWS
CHALLENGES | ABOUT DREAM | OUR COMMUNITY | PUBLICATIONS | ALGORITHMS

The power of DREAM
is the power to connect

Designed and run by a community of researchers from a variety of organizations, DREAM Challenges invite participants to propose solutions to fundamental biomedical questions — fostering collaboration and building communities in the process. Sage Bionetworks provides the expertise and infrastructure to host challenges via their Synapse platform. Together, we share a vision allowing individuals and groups to collaborate openly so that the "wisdom of the crowd" provides the greatest impact on science and human health.

COVID-19 EHR DREAM Challenge

The rapid rise of COVID-19 has challenged healthcare globally. Due to the importance and emergent need for better understanding of the condition and the development of patient specific clinical risk scores and early warning tools, we have developed a platform to support testing analytic and machine learning hypotheses on clinical data without data sharing as a platform to rapidly discover and implement approaches for care. This Challenges aims to identify risk factors that lead to a positive



DREAM Challenge Participant Publications

2020 Publications

Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms
Schaffter, T., et al., JAMA Netw Open. 2020;3(3):e200265. doi:10.1001/jamanetworkopen.2020.0265

Crowdsourced mapping of unexplored target space of kinase inhibitors
Cichonska, A., et al., Biorxiv; 10.1101/2019.12.31.891812

A community effort to create standards for evaluating tumor subclonal reconstructions
Salcedo, A. et al, *Nature Biotechnology*, 38, 27-107. DOI: 10.1038/s41587-019-0364-z

Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge
Sieberts, S. K., et al., Biorxiv; 10.1101/2020.01.13.904722

Multiple Myeloma DREAM Challenge reveals epigenetic regulator PHF19 as marker of aggressive disease
Mason MJ, et al. *Leukemia*. 2020;10.1038/s41375-020-0742-z. doi:10.1038/s41375-020-0742-z

Listing of DREAM publications on Google Scholar

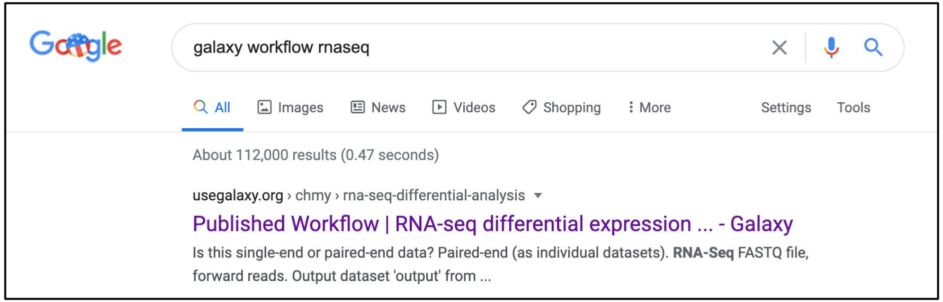
DREAM publication channel on Biorxiv

Find out more about upcoming, current, and past challenges:

The DREAM Challenge participants and organizers have numerous publications in leading journals in biology and

- Should NCPI organize DREAM-like challenges for interoperability technologies?
- Should NCPI organize DREAM-like challenges for genomic analysis?
- Where would we start?

Planning for Obsolescence



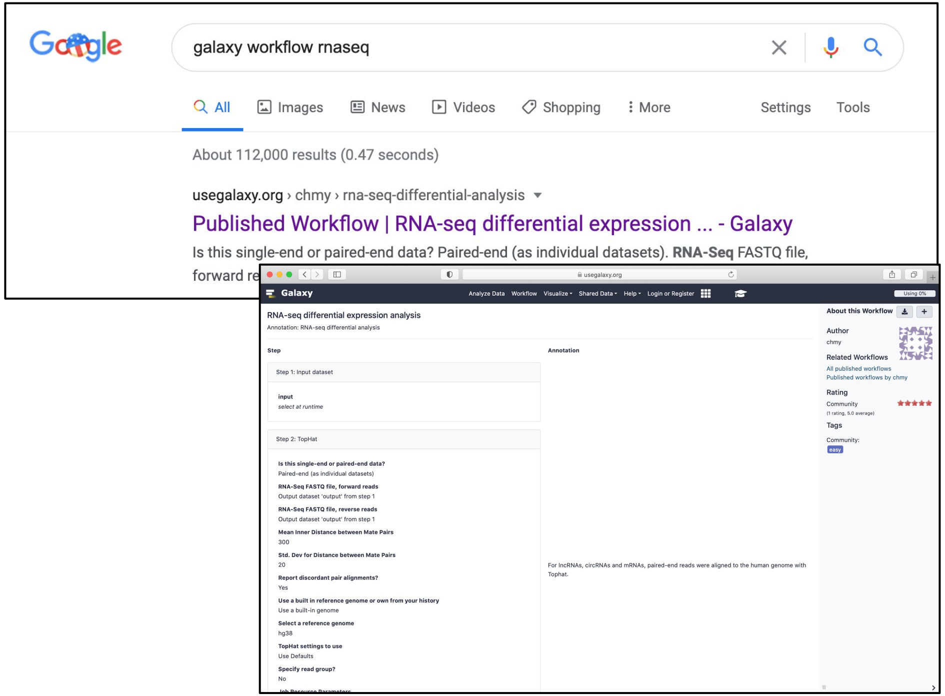
About 112,000 results (0.47 seconds)

usegalaxy.org › chmy › rna-seq-differential-analysis

Published Workflow | RNA-seq differential expression ... - Galaxy

Is this single-end or paired-end data? Paired-end (as individual datasets). **RNA-Seq** FASTQ file, forward reads. Output dataset 'output' from ...

Planning for Obsolescence



galaxy workflow rnaseq

About 112,000 results (0.47 seconds)

usegalaxy.org › chmy › rna-seq-differential-analysis

Published Workflow | RNA-seq differential expression ... - Galaxy

Is this single-end or paired-end data? Paired-end (as individual datasets). **RNA-Seq** FASTQ file, forward re

RNA-seq differential expression analysis

Annotation: RNA-seq differential analysis

Step	Annotation
Step 1: Input dataset **input** *select at runtime*	
Step 2: TopHat **Is this single-end or paired-end data?** Paired-end (as individual datasets) **RNA-Seq FASTQ file, forward reads** Output dataset 'output' from step 1 **RNA-Seq FASTQ file, reverse reads** Output dataset 'output' from step 1 **Mean Inner Distance between Mate Pairs** 300 **Std. Dev for Distance between Mate Pairs** 20 **Report discordant pair alignments?** Yes **Use a built in reference genome or own from your history** Use a built-in genome **Select a reference genome** hg38 **TopHat settings to use** Use Defaults **Specify read group?** No **Job Resource Parameters**	For lncRNAs, circRNAs and mRNAs, paired-end reads were aligned to the human genome with Tophat.

About this Workflow

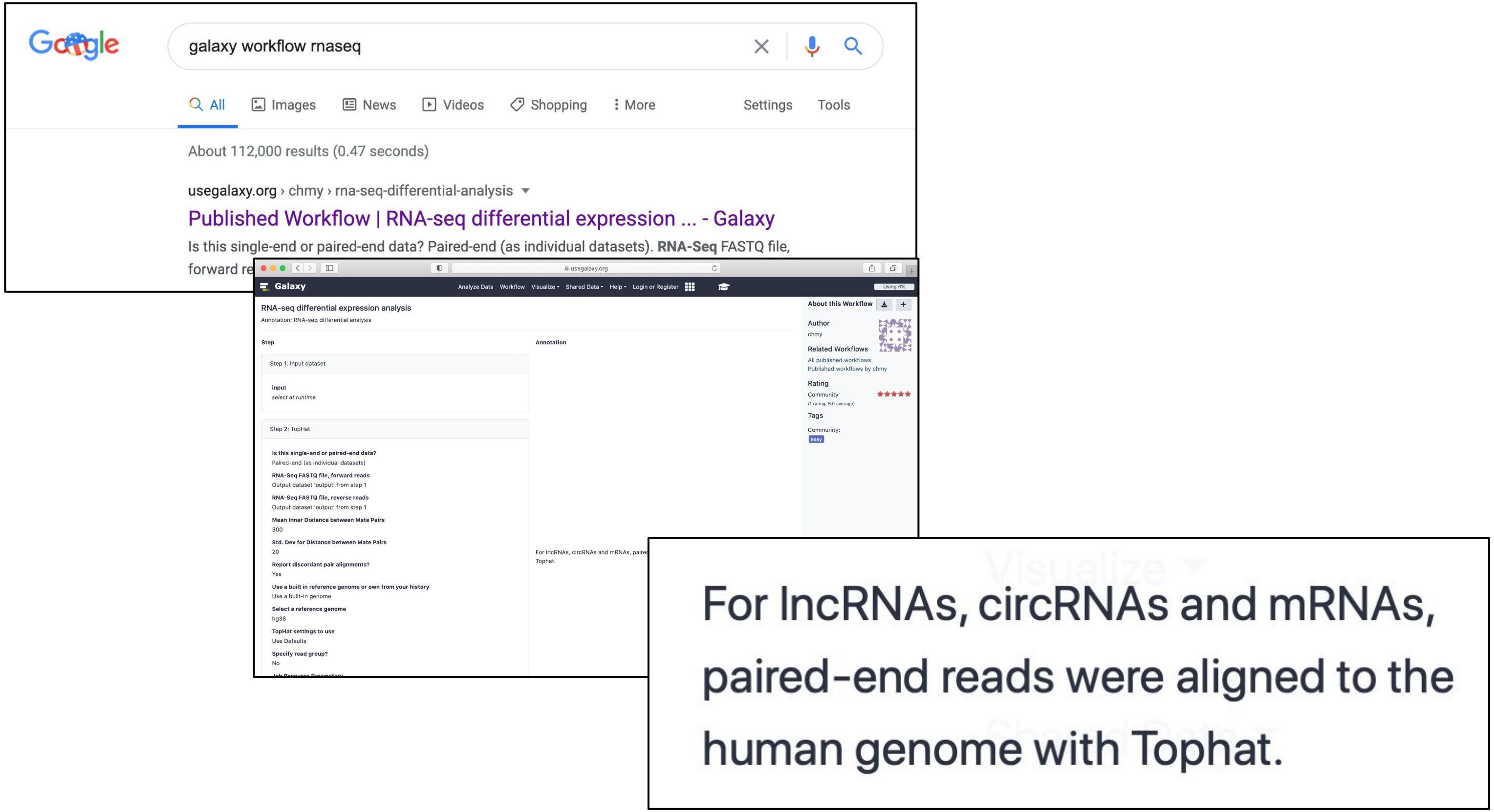
Author
chmy

Related Workflows
All published workflows
Published workflows by chmy

Rating
Community ★★★★★
(1 rating, 5.0 average)

Tags
Community:
easy

Planning for Obsolescence



galaxy workflow rnaseq

About 112,000 results (0.47 seconds)

usegalaxy.org › chmy › rna-seq-differential-analysis

Published Workflow | RNA-seq differential expression ... - Galaxy

Is this single-end or paired-end data? Paired-end (as individual datasets). **RNA-Seq** FASTQ file, forward re

RNA-seq differential expression analysis

Annotation: RNA-seq differential analysis

Step

Step 1: Input dataset

input
select at runtime

Step 2: TopHat

Is this single-end or paired-end data?
Paired-end (as individual datasets)

RNA-Seq FASTQ file, forward reads
Output dataset 'output' from step 1

RNA-Seq FASTQ file, reverse reads
Output dataset 'output' from step 1

Mean Inner Distance between Mate Pairs
300

Std. Dev for Distance between Mate Pairs
20

Report discordant pair alignments?
Yes

Use a built in reference genome or own from your history
Use a built-in genome

Select a reference genome
hg38

TopHat settings to use
Use Defaults

Specify read group?
No

Job Resource Parameters

Annotation

For IncRNAs, circRNAs and mRNAs, paire
Tophat.

About this Workflow

Author
chmy

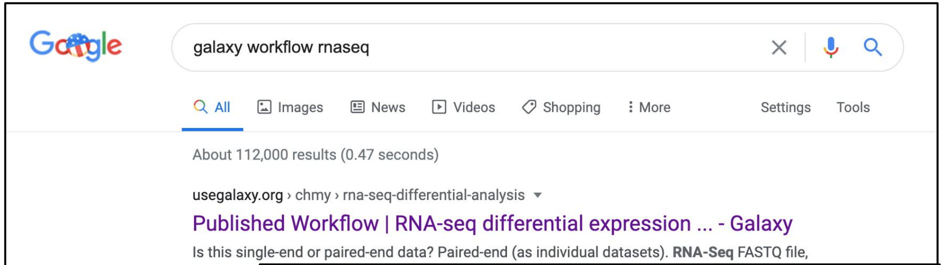
Related Workflows
All published workflows
Published workflows by chmy

Rating
Community ★★★★★
(1 rating, 5.0 average)

Tags
Community:
easy

For IncRNAs, circRNAs and mRNAs, paired-end reads were aligned to the human genome with Tophat.

Planning for Obsolescence



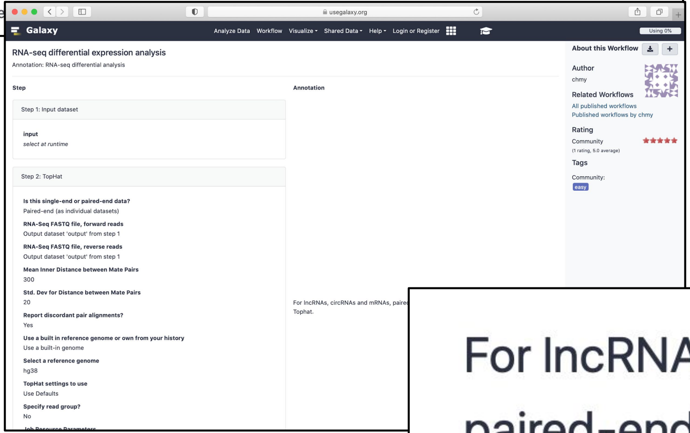
galaxy workflow rnaseq

About 112,000 results (0.47 seconds)

usegalaxy.org › chmy › rna-seq-differential-analysis

Published Workflow | RNA-seq differential expression ... - Galaxy

Is this single-end or paired-end data? Paired-end (as individual datasets). **RNA-Seq** FASTQ file, forward re



RNA-seq differential expression analysis

Annotation: RNA-seq differential analysis

Step

Annotation

Step 1: Input dataset

input
select at runtime

Step 2: TopHat

Is this single-end or paired-end data?
Paired-end (as individual datasets)

RNA-Seq FASTQ file, forward reads
Output dataset 'output' from step 1

RNA-Seq FASTQ file, reverse reads
Output dataset 'output' from step 1

Mean Inner Distance between Mate Pairs
300

Std. Dev for Distance between Mate Pairs
20

Report discordant pair alignments?
Yes

Use a built in reference genome or own from your history
Use a built-in genome

Select a reference genome
hg38

TopHat settings to use
Use Defaults

Specify read group?
No

Job Resource Parameters

For lncRNAs, circRNAs and mRNAs, paire
Tophat.

About this Workflow

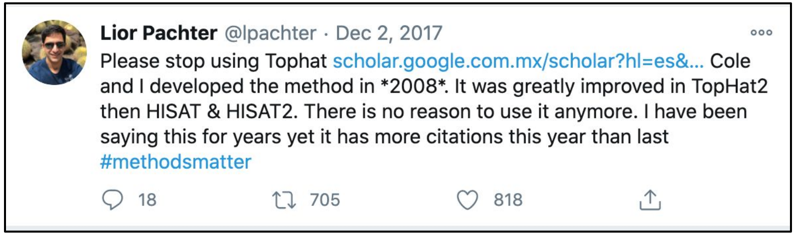
Author
chmy

Related Workflows
All published workflows
Published workflows by chmy

Rating
Community ★★★★★
(1 rating, 5.0 average)

Tags
Community:
easy

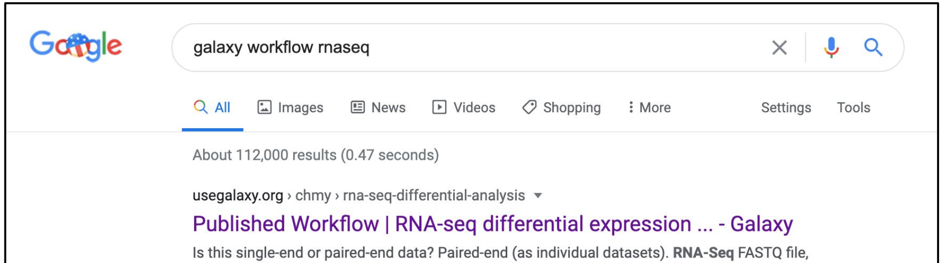
For lncRNAs, circRNAs and mRNAs, paired-end reads were aligned to the human genome with Tophat.



Lior Pachter @lpachter · Dec 2, 2017

Please stop using Tophat scholar.google.com.mx/scholar?hl=es&... Cole and I developed the method in *2008*. It was greatly improved in TopHat2 then HISAT & HISAT2. There is no reason to use it anymore. I have been saying this for years yet it has more citations this year than last #methodsmatter

Planning for Obsolescence



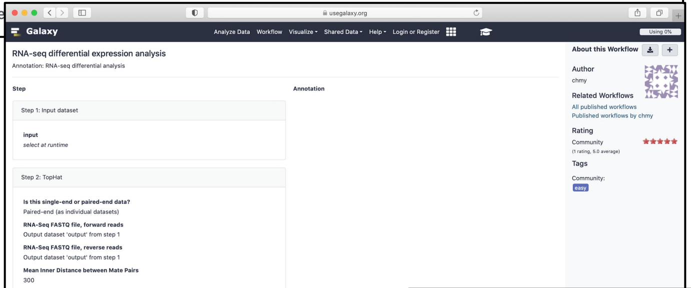
galaxy workflow rnaseq

About 112,000 results (0.47 seconds)

usegalaxy.org › chmy › rna-seq-differential-analysis

Published Workflow | RNA-seq differential expression ... - Galaxy

Is this single-end or paired-end data? Paired-end (as individual datasets). **RNA-Seq** FASTQ file, forward re



RNA-seq differential expression analysis

Annotation: RNA-seq differential analysis

Step

Annotation

Step 1: Input dataset

input
select at runtime

Step 2: TopHat

Is this single-end or paired-end data?
Paired-end (as individual datasets)

RNA-Seq FASTQ file, forward reads
Output dataset 'output' from step 1

RNA-Seq FASTQ file, reverse reads
Output dataset 'output' from step 1

Mean Inner Distance between Mate Pairs
300

Std. Dev for Distance between Mate Pairs
20

Report discordant pair alignments?
Yes

Use a built in reference genome or own from your history
Use a built-in genome

Select a reference genome
hg38

TopHat settings to use
Use Defaults

Specify read group?
No

Job Resource Parameters

For IncRNAs, circRNAs and mRNAs, paire
Tophat.

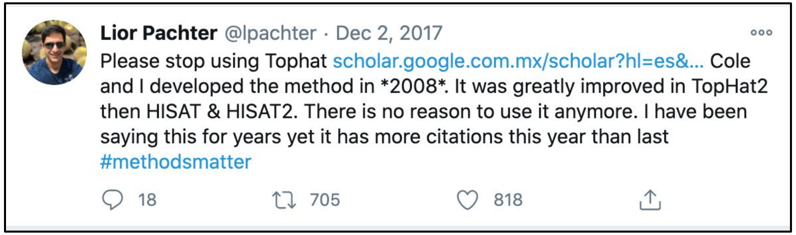
About this Workflow

Author
chmy

Related Workflows
All published workflows
Published workflows by chmy

Rating
Community ★★★★★
(1 rating, 5.0 average)

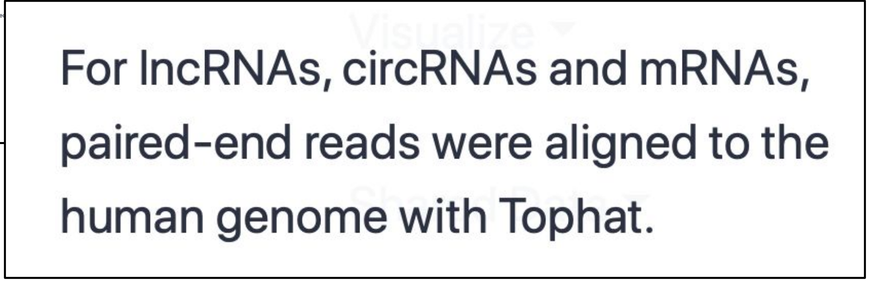
Tags
Community:
easy



Lior Pachter @lpachter · Dec 2, 2017

Please stop using Tophat scholar.google.com.mx/scholar?hl=es&... Cole and I developed the method in *2008*. It was greatly improved in TopHat2 then HISAT & HISAT2. There is no reason to use it anymore. I have been saying this for years yet it has more citations this year than last #methodsmatter

Our metadata will be our longest-lasting artifacts



For IncRNAs, circRNAs and mRNAs, paired-end reads were aligned to the human genome with Tophat.

Discussion

What **existing** workflow interop strategies do you know about today?

What **ideal** workflow interop strategy would be most impactful for your research today?

Is workflow obsolescence a topic we should address today or in the future?

Imaging Data

Ashok Krishnamurthy[1] & Steve Pieper[2]
1: RENCI, UNC-Chapel Hill and BioData Catalyst
2: Isomics, Inc. and NCI Imaging Data Commons

Imaging Data

You will have 40 minutes.

Notes:
https://docs.google.com/document/d/1TSnNqeP_FtQ2MqiEsZU4DnDFqjkf6_vyFeQnHNBodLw/edit

Three Imaging Use Cases

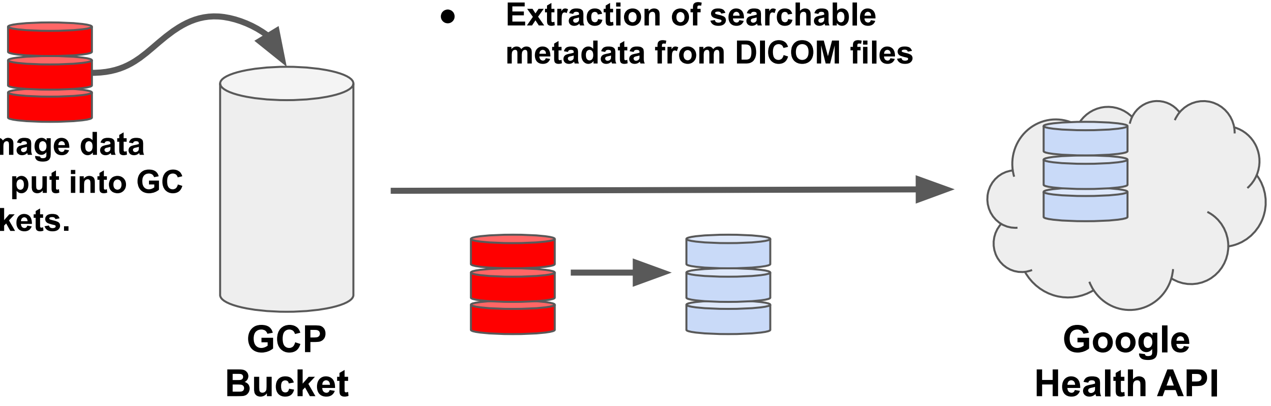
- Image Data Ingestion Workflow
- Clinical/Research Radiologist Workflow
- Imaging/AI/Machine Learning Workflow

BioData Catalyst Image Ingestion

- BDCatalyst to date has largely focused on ingestion of clinical and genomics data
- Images present unique challenges
  - Scale of image files
    - COPDGene Phase 1 has 22M DICOM files
  - Unique PHI issues
    - DICOM metadata
    - Embedded images (eg pacemaker with serial no.)
  - DICOM protocol standards

Team Helium Rapid Image Ingestion Proposal

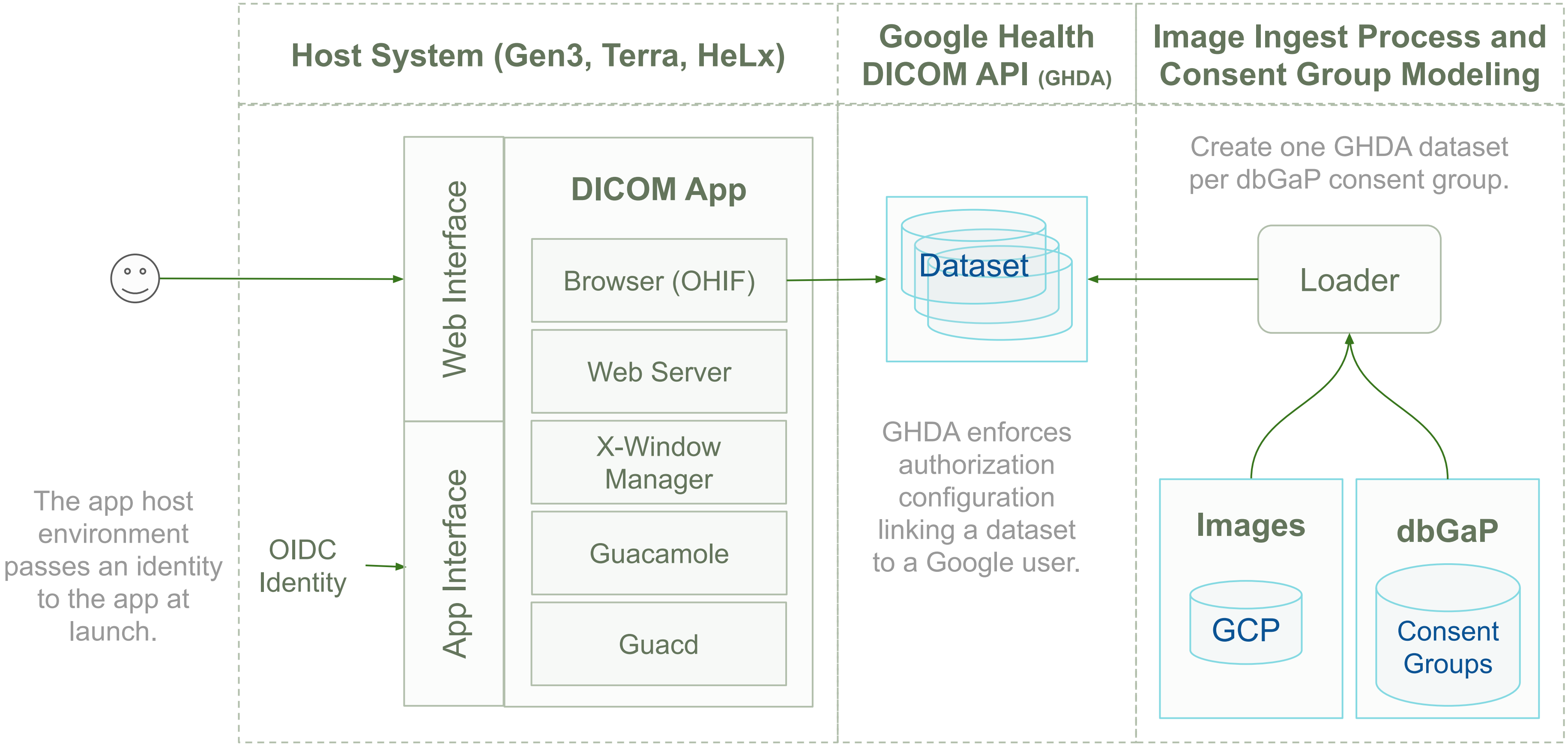
- Image de-identification
- Images in bucket by dbGaP Consent groups
- Extraction of searchable metadata from DICOM files



New image data will be put into GC in buckets.

- Ingest images into Google Health API
- Images are segregated by consent groups

BDC Image Ingest, DICOM Viewer App, and Authorization Proposal



Host System (Gen3, Terra, HeLx)

Google Health DICOM API (GHDA)

Image Ingest Process and Consent Group Modeling

The app host environment passes an identity to the app at launch.

OIDC Identity

Web Interface

App Interface

DICOM App

Browser (OHIF)

Web Server

X-Window Manager

Guacamole

Guacd

Dataset

GHDA enforces authorization configuration linking a dataset to a Google user.

Create one GHDA dataset per dbGaP consent group.

Loader

Images

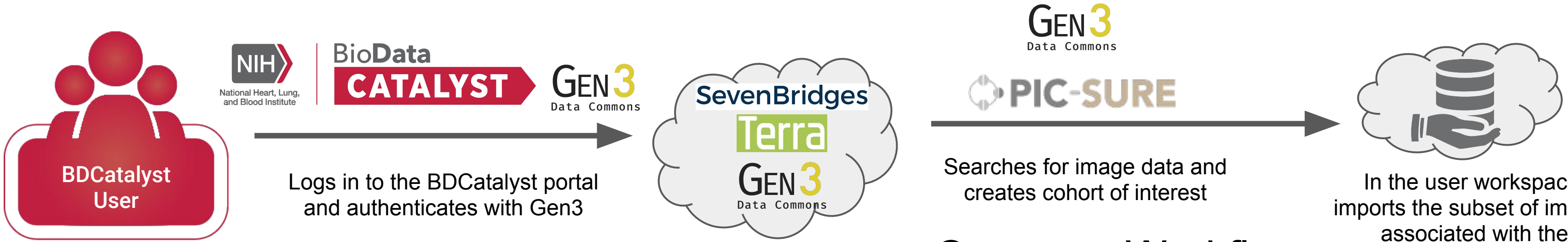
GCP

dbGaP

Consent Groups

This assumes that (a) the user authenticates to Google in the DICOM App or (b) a mapping from eRACommons provided by Gen3/Fence to Google identity is implemented.

Image Workflow Proposals



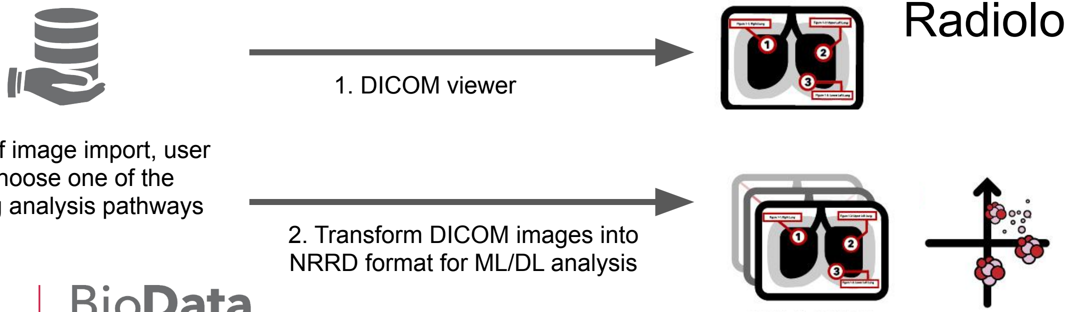
Logs in to the BDCatalyst portal and authenticates with Gen3

Creates a workspace for data analysis

Searches for image data and creates cohort of interest

In the user workspace, imports the subset of images associated with the GUIDs/Subject IDs

Common Workflow



As part of image import, user can choose one of the following analysis pathways

1. DICOM viewer

Radiologist Workflow

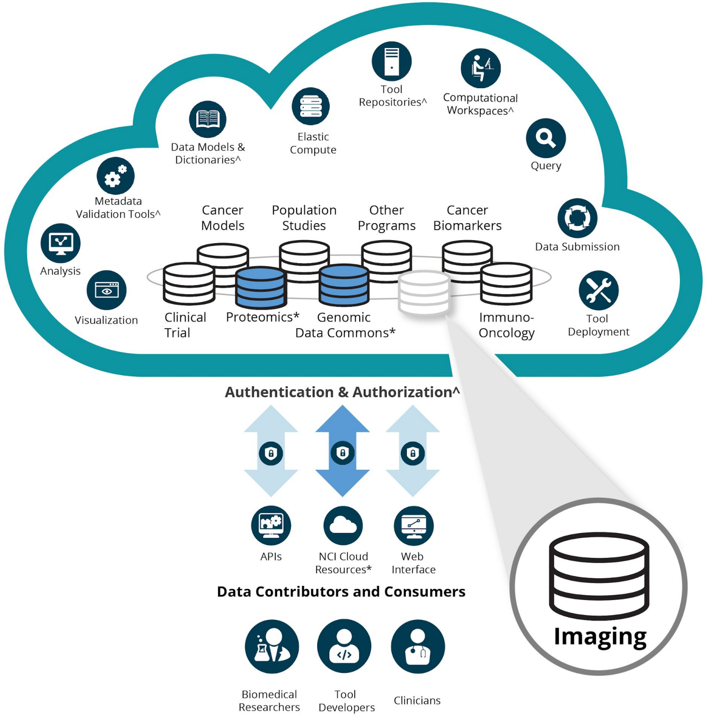
2. Transform DICOM images into NRRD format for ML/DL analysis

Imaging/AI/ML Workflow

Cancer Research Data Commons (CRDC) Imaging Data Commons (IDC)

*The NCI Imaging Data Commons will be a cloud-based resource that **connects** researchers with*

1. *cancer image collections*
2. *a robust infrastructure that contains imaging data and metadata*
3. *resources for searching, identifying and viewing images*
4. *links to other Cancer Research Data Commons nodes.*

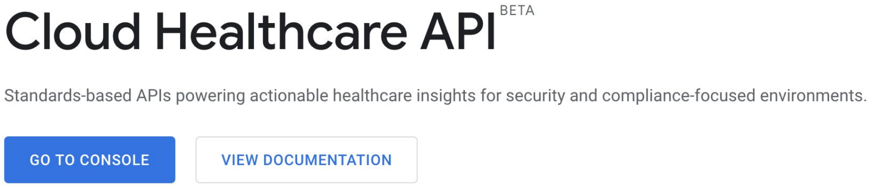


* Available to the community.
^ Components of the Data Commons Framework

IDC Uses Google Healthcare

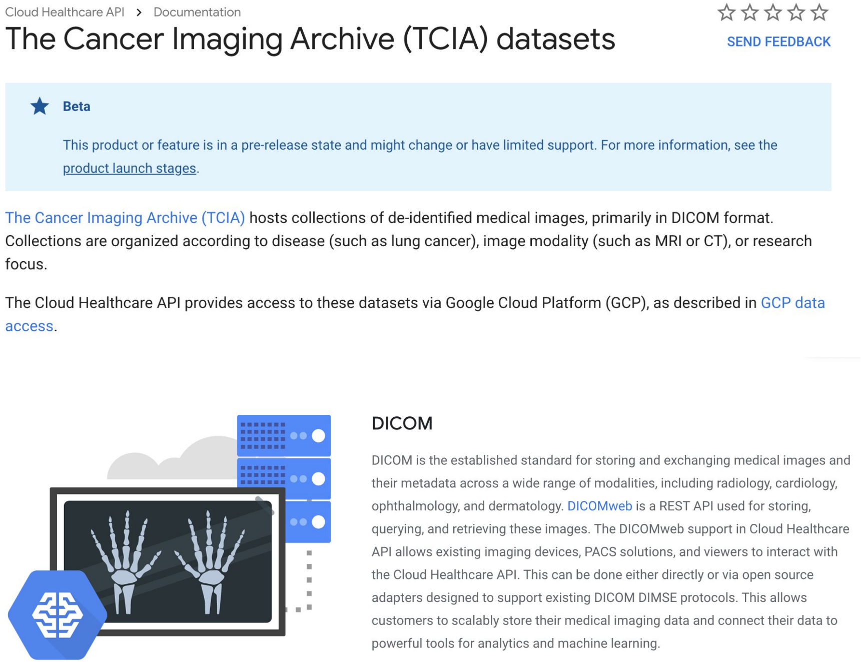
- Scalable DICOM service
- BigQuery (SQL) for DICOM headers
- Authentication, data security, compute, GPU, notebooks …

Goal is to use Google to make workable site, but use open standards so the same methodology can work anywhere



Cloud Healthcare API BETA

Standards-based APIs powering actionable healthcare insights for security and compliance-focused environments.



The Cancer Imaging Archive (TCIA) datasets

★ **Beta**

This product or feature is in a pre-release state and might change or have limited support. For more information, see the product launch stages.

The Cancer Imaging Archive (TCIA) hosts collections of de-identified medical images, primarily in DICOM format. Collections are organized according to disease (such as lung cancer), image modality (such as MRI or CT), or research focus.

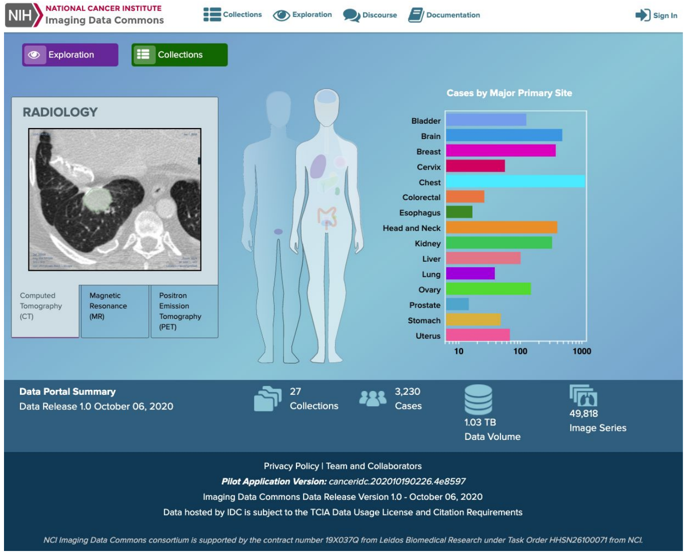
The Cloud Healthcare API provides access to these datasets via Google Cloud Platform (GCP), as described in GCP data access.

DICOM

DICOM is the established standard for storing and exchanging medical images and their metadata across a wide range of modalities, including radiology, cardiology, ophthalmology, and dermatology. DICOMweb is a REST API used for storing, querying, and retrieving these images. The DICOMweb support in Cloud Healthcare API allows existing imaging devices, PACS solutions, and viewers to interact with the Cloud Healthcare API. This can be done either directly or via open source adapters designed to support existing DICOM DIMSE protocols. This allows customers to scalably store their medical imaging data and connect their data to powerful tools for analytics and machine learning.

What’s inside IDC pilot release

- Data in requester pays buckets
- Metadata in BigQuery tables
- Exploration portal
- Viewer (interfacing data via proxy)
- Documentation
- Video tutorials
- Forum (Discourse)
- Example integration with tools (Colab Notebooks, DataStudio)
- Analytics use cases - under development



Data Portal Summary
Data Release 1.0 October 06, 2020

27 Collections

3,230 Cases

1.03 TB Data Volume

49,818 Image Series

Privacy Policy | Team and Collaborators

Pilot Application Version: *canceridc.202010190226.4e8597*

Imaging Data Commons Data Release Version 1.0 - October 06, 2020

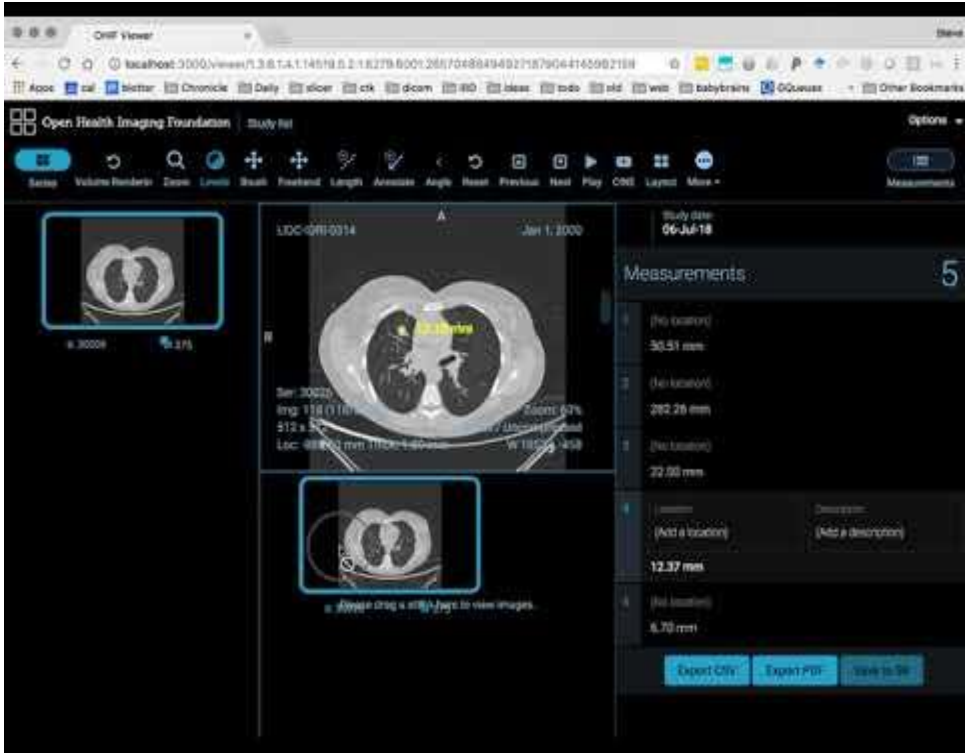
Data hosted by IDC is subject to the TCIA Data Usage License and Citation Requirements

NCI Imaging Data Commons consortium is supported by the contract number 19X037Q from Leidos Biomedical Research under Task Order HHSN26100071 from NCI.

Image viewer: OHIF

Open Health Imaging Foundation

- Free open source, browser-based (zero install!)
  - Open source, modern Javascript
  - DICOM standard images, segmentations, annotations
  - Professional design
- DICOMweb supported by Google, Siemens, and open source servers
- VTK.js WebGL visualization
- Pathology plugin development
  - DICOM Whole Slide Imaging
  - Efficient DICOMweb pyramid access



Contributors: Steve Pieper, James Petts, Erik Ziegler, Trinity Urban, Gordon Harris (OHIF), Markus Herrmann (BWH/MGH CCDS)

https://github.com/OHIF/Viewers

Kids First Imaging Pilots

1. Began piloting imaging data generation in response to user community requests for such data.

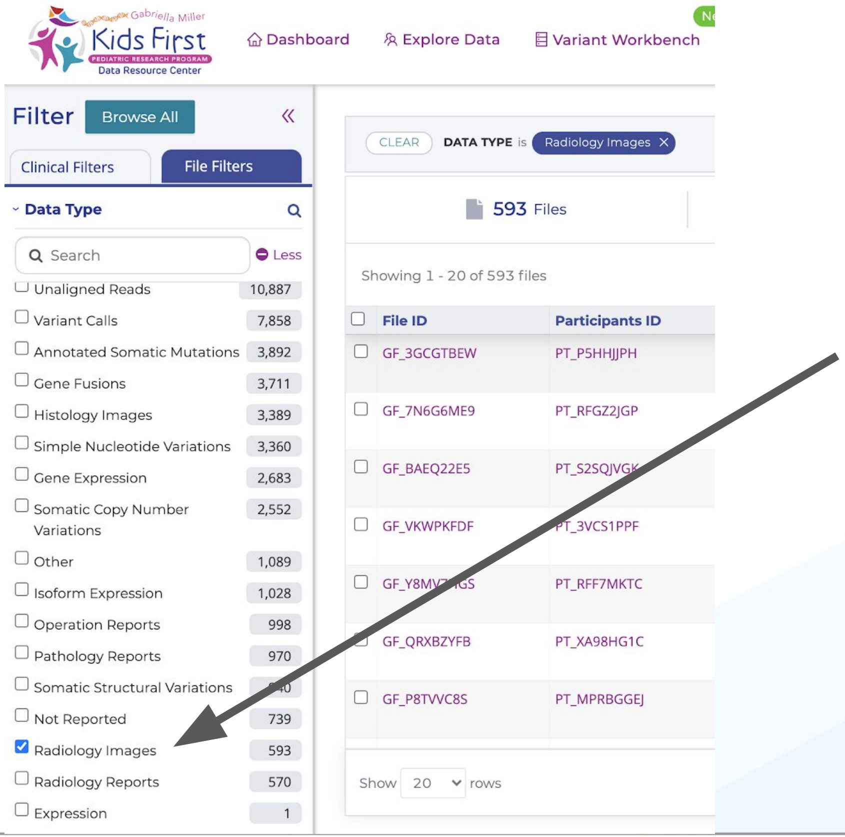
2. As the Kids First Datasets expanded, it became apparent imaging studies are core elements of clinical data collection efforts across X01 investigators in both congenital/birth defects and cancer contexts.

3. Many investigators have access to and/or collect imaging data, but have yet to fully leverage multimodal analysis that spans imaging/clinical/genomic data.

4. Growing and emerging needs for interoperating with MIS-C and INCLUDE data.

Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

Kids First Imaging Pilots



Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

Filter Browse All

Clinical Filters | File Filters

Data Type

Search ● Less

□ Unaligned Reads	10,887
□ Variant Calls	7,858
□ Annotated Somatic Mutations	3,892
□ Gene Fusions	3,711
□ Histology Images	3,389
□ Simple Nucleotide Variations	3,360
□ Gene Expression	2,683
□ Somatic Copy Number Variations	2,552
□ Other	1,089
□ Isoform Expression	1,028
□ Operation Reports	998
□ Pathology Reports	970
□ Somatic Structural Variations	940
□ Not Reported	739
☑ Radiology Images	593
□ Radiology Reports	570
□ Expression	1

CLEAR DATA TYPE is Radiology Images ✕

593 Files

Showing 1 - 20 of 593 files

□	File ID	Participants ID
□	GF_3GCGTBEW	PT_P5HHJJPH
□	GF_7N6G6ME9	PT_RFGZ2JGP
□	GF_BAEQ22E5	PT_S2SQJVGK
□	GF_VKWPKFDF	PT_3VCS1PPF
□	GF_Y8MV[illegible]GS	PT_RFF7MKTC
□	GF_QRXBZYFB	PT_XA98HG1C
□	GF_P8TVVC8S	PT_MPRBGGEJ

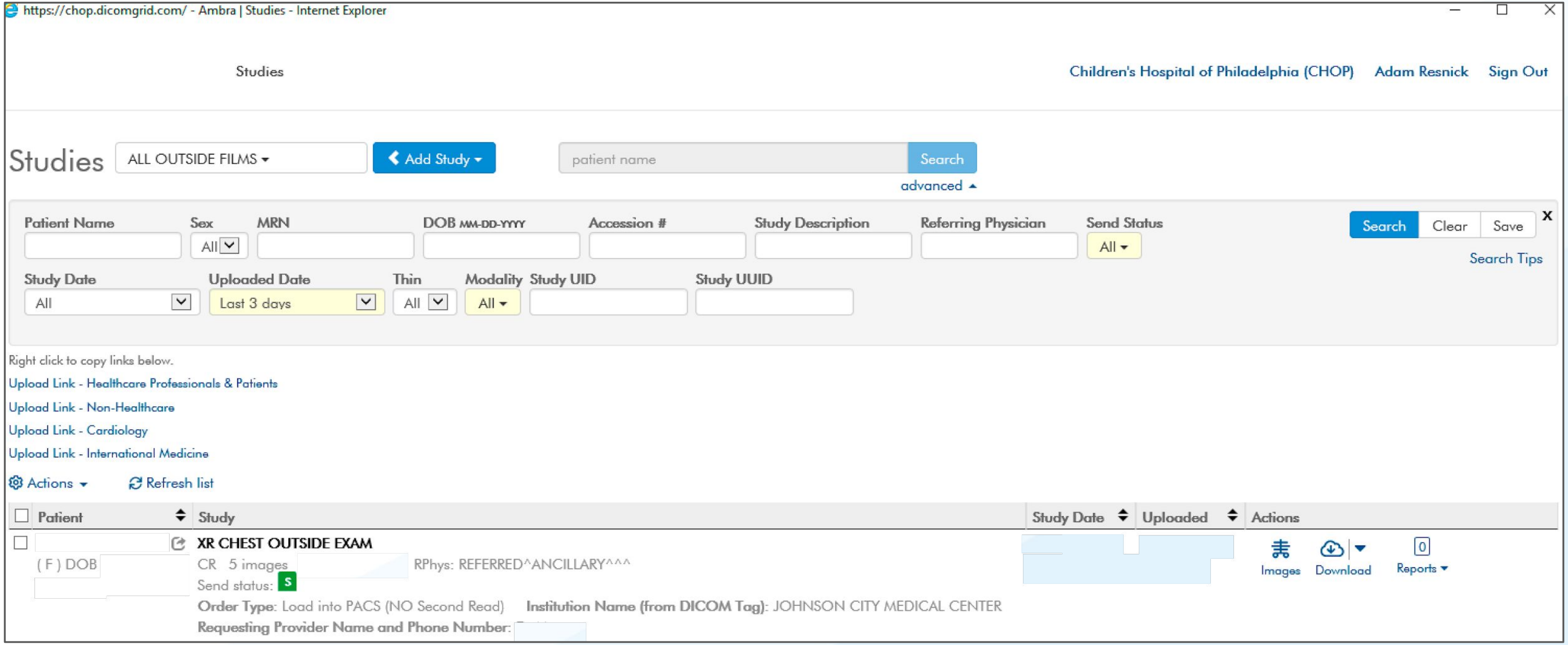
Show 20 rows

However, unlike bam/cram files -- getting to a DICOM file listing is often insufficient for investigators interested in using these data to make decisions on use.

The Cohort creation process for imaging datasets requires a dedicated imaging context, supporting both “human” and “machine” use/review/analysis setting.

Kids First Imaging -- Data workflows/oportunities

Unlike genomics, imaging is a clinical standard with **imaging data “movement” as an inherent feature of most hospital system (even if not optimized).**



https://chop.dicomgrid.com/ - Ambra | Studies - Internet Explorer

Studies

Children's Hospital of Philadelphia (CHOP) Adam Resnick Sign Out

Studies ALL OUTSIDE FILMS ▾ ❮ Add Study ▾ patient name Search

advanced ▴

Patient Name:

Sex: All

MRN:

DOB MM-DD-YYYY:

Accession #:

Study Description:

Referring Physician:

Send Status: All ▾

Search Clear Save X

Search Tips

Study Date: All

Uploaded Date: Last 3 days

Thin: All

Modality: All ▾

Study UID:

Study UUID:

Right click to copy links below.

Upload Link - Healthcare Professionals & Patients

Upload Link - Non-Healthcare

Upload Link - Cardiology

Upload Link - International Medicine

Actions ▾ Refresh list

□	Patient	Study	Study Date	Uploaded	Actions
□	(F) DOB	XR CHEST OUTSIDE EXAM CR 5 images RPhys: REFERRED^ANCILLARY^^^ Send status: S Order Type: Load into PACS (NO Second Read) Institution Name (from DICOM Tag): JOHNSON CITY MEDICAL CENTER Requesting Provider Name and Phone Number:			Images Download ▾ Reports ▾

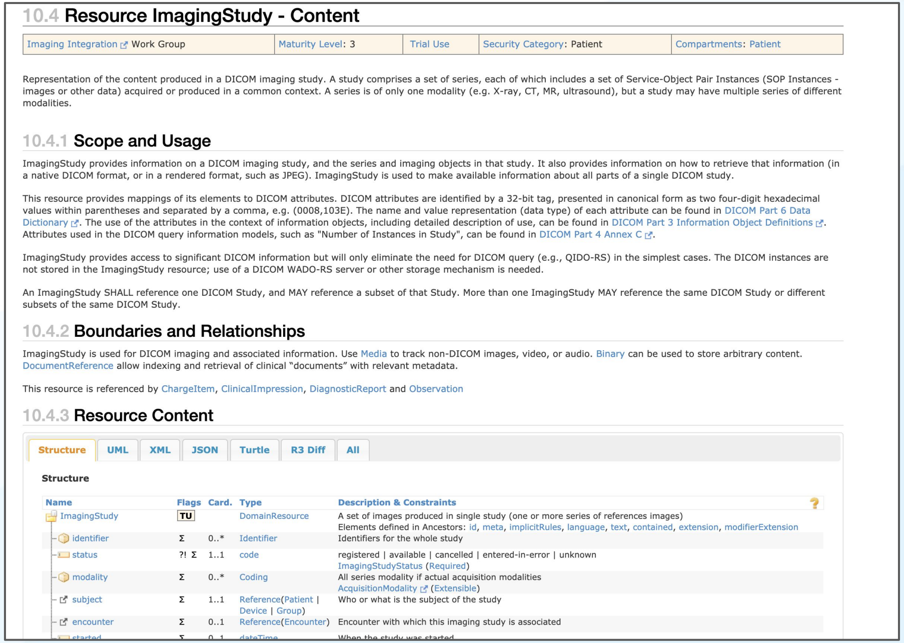
Kids First Imaging -- Data workflows/opportunities

Unlike genomics, imaging is a clinical standard with imaging data “movement” as an inherent feature of most hospital system (even if not optimized).

De-identification workflows is a key need!

Like genomics - multi-cloud needs . . .

We shouldn’t “throw away” FHIR-based structuring defined for imaging



10.4 Resource ImagingStudy - Content

Imaging Integration Work Group	Maturity Level: 3	Trial Use	Security Category: Patient	Compartments: Patient

Representation of the content produced in a DICOM imaging study. A study comprises a set of series, each of which includes a set of Service-Object Pair Instances (SOP Instances - images or other data) acquired or produced in a common context. A series is of only one modality (e.g. X-ray, CT, MR, ultrasound), but a study may have multiple series of different modalities.

10.4.1 Scope and Usage

ImagingStudy provides information on a DICOM imaging study, and the series and imaging objects in that study. It also provides information on how to retrieve that information (in a native DICOM format, or in a rendered format, such as JPEG). ImagingStudy is used to make available information about all parts of a single DICOM study.

This resource provides mappings of its elements to DICOM attributes. DICOM attributes are identified by a 32-bit tag, presented in canonical form as two four-digit hexadecimal values within parentheses and separated by a comma, e.g. (0008,103E). The name and value representation (data type) of each attribute can be found in DICOM Part 6 Data Dictionary. The use of the attributes in the context of information objects, including detailed description of use, can be found in DICOM Part 3 Information Object Definitions. Attributes used in the DICOM query information models, such as "Number of Instances in Study", can be found in DICOM Part 4 Annex C.

ImagingStudy provides access to significant DICOM information but will only eliminate the need for DICOM query (e.g., QIDO-RS) in the simplest cases. The DICOM instances are not stored in the ImagingStudy resource; use of a DICOM WADO-RS server or other storage mechanism is needed.

An ImagingStudy SHALL reference one DICOM Study, and MAY reference a subset of that Study. More than one ImagingStudy MAY reference the same DICOM Study or different subsets of the same DICOM Study.

10.4.2 Boundaries and Relationships

ImagingStudy is used for DICOM imaging and associated information. Use Media to track non-DICOM images, video, or audio. Binary can be used to store arbitrary content. DocumentReference allow indexing and retrieval of clinical “documents” with relevant metadata.

This resource is referenced by ChargeItem, ClinicalImpression, DiagnosticReport and Observation

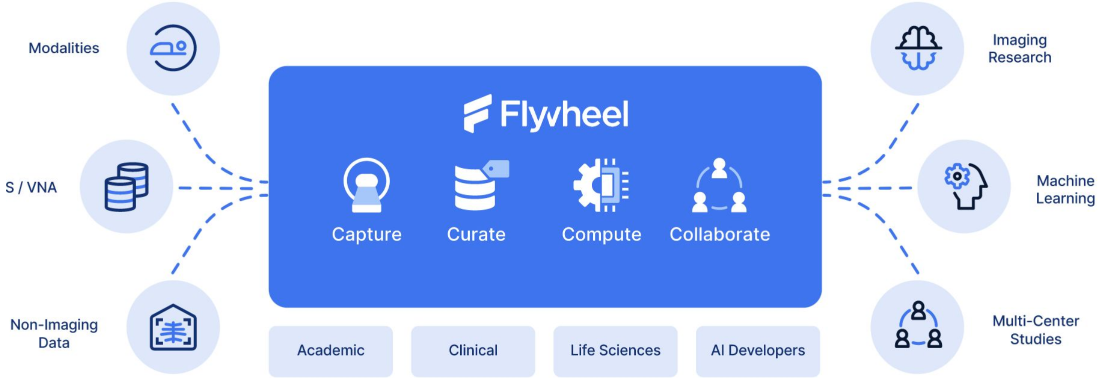
10.4.3 Resource Content

Structure | UML | XML | JSON | Turtle | R3 Diff | All

Structure

Name	Flags	Card.	Type	Description & Constraints
ImagingStudy	TU		DomainResource	A set of images produced in single study (one or more series of references images) Elements defined in Ancestors: id, meta, implicitRules, language, text, contained, extension, modifierExtension
identifier	Σ	0..*	Identifier	Identifiers for the whole study
status	?! Σ	1..1	code	registered | available | cancelled | entered-in-error | unknown ImagingStudyStatus (Required)
modality	Σ	0..*	Coding	All series modality if actual acquisition modalities AcquisitionModality (Extensible)
subject	Σ	1..1	Reference(Patient | Device | Group)	Who or what is the subject of the study
encounter	Σ	0..1	Reference(Encounter)	Encounter with which this imaging study is associated
started	Σ	0..1	dateTime	When the study was started

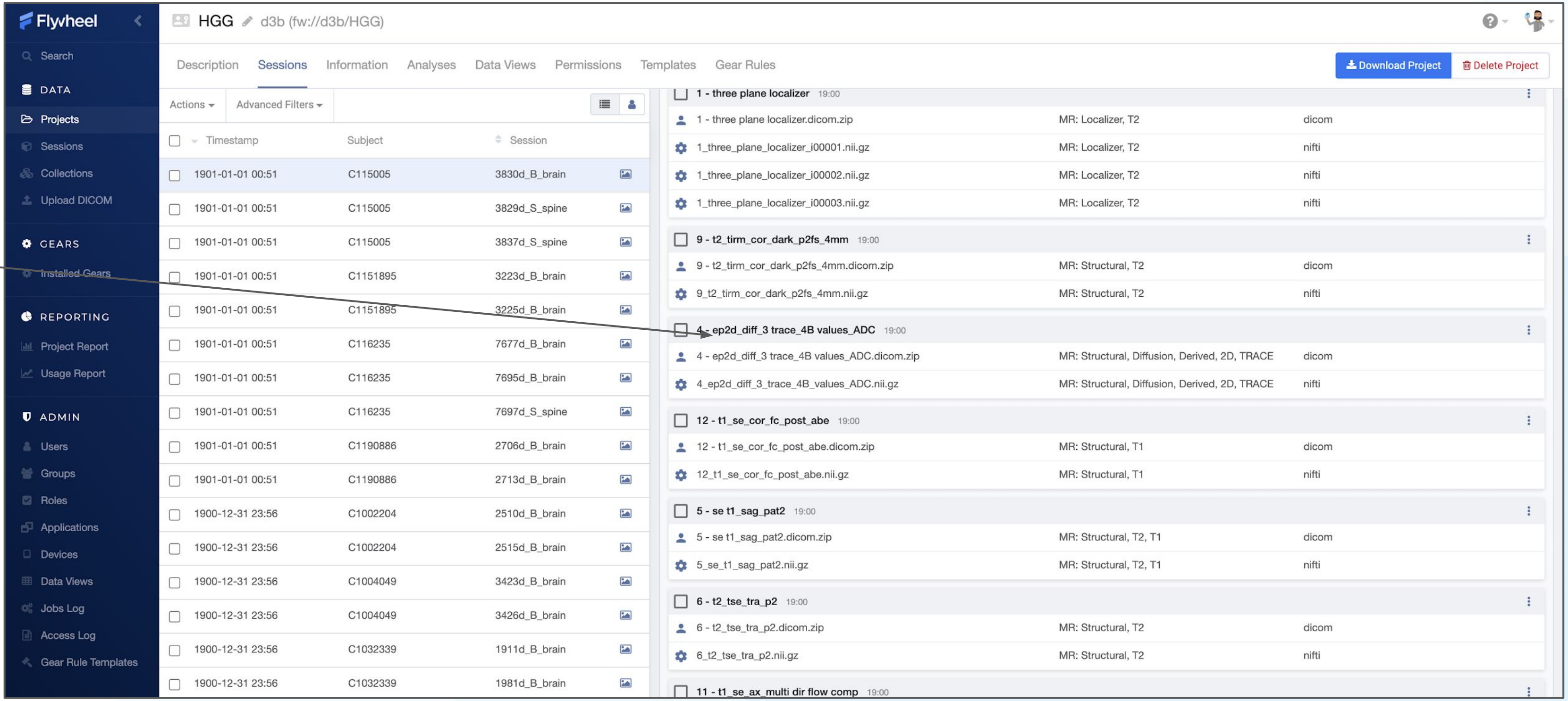
Kids First Imaging -- Flywheel Pilot



Flywheel

Kids First Imaging -- Flywheel Pilot

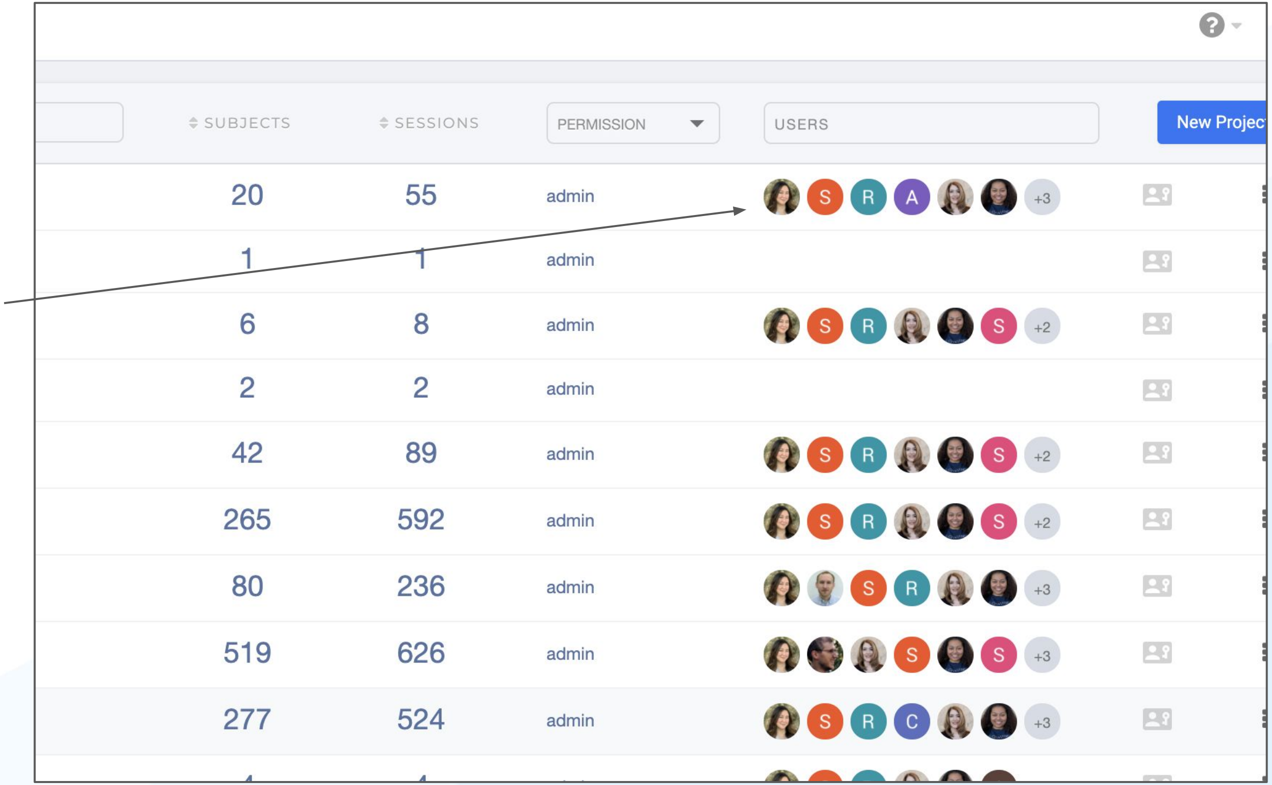
“See” inside DICOMs



Timestamp	Subject	Session
1901-01-01 00:51	C115005	3830d_B_brain
1901-01-01 00:51	C115005	3829d_S_spine
1901-01-01 00:51	C115005	3837d_S_spine
1901-01-01 00:51	C1151895	3223d_B_brain
1901-01-01 00:51	C1151895	3225d_B_brain
1901-01-01 00:51	C116235	7677d_B_brain
1901-01-01 00:51	C116235	7695d_B_brain
1901-01-01 00:51	C116235	7697d_S_spine
1901-01-01 00:51	C1190886	2706d_B_brain
1901-01-01 00:51	C1190886	2713d_B_brain
1900-12-31 23:56	C1002204	2510d_B_brain
1900-12-31 23:56	C1002204	2515d_B_brain
1900-12-31 23:56	C1004049	3423d_B_brain
1900-12-31 23:56	C1004049	3426d_B_brain
1900-12-31 23:56	C1032339	1911d_B_brain
1900-12-31 23:56	C1032339	1981d_B_brain

1 - three plane localizer 19:00		
1 - three plane localizer.dicom.zip	MR: Localizer, T2	dicom
1_three_plane_localizer_i00001.nii.gz	MR: Localizer, T2	nifti
1_three_plane_localizer_i00002.nii.gz	MR: Localizer, T2	nifti
1_three_plane_localizer_i00003.nii.gz	MR: Localizer, T2	nifti
9 - t2_tirm_cor_dark_p2fs_4mm 19:00		
9 - t2_tirm_cor_dark_p2fs_4mm.dicom.zip	MR: Structural, T2	dicom
9_t2_tirm_cor_dark_p2fs_4mm.nii.gz	MR: Structural, T2	nifti
4 - ep2d_diff_3 trace_4B values_ADC 19:00		
4 - ep2d_diff_3 trace_4B values_ADC.dicom.zip	MR: Structural, Diffusion, Derived, 2D, TRACE	dicom
4_ep2d_diff_3_trace_4B_values_ADC.nii.gz	MR: Structural, Diffusion, Derived, 2D, TRACE	nifti
12 - t1_se_cor_fc_post_abe 19:00		
12 - t1_se_cor_fc_post_abe.dicom.zip	MR: Structural, T1	dicom
12_t1_se_cor_fc_post_abe.nii.gz	MR: Structural, T1	nifti
5 - se t1_sag_pat2 19:00		
5 - se t1_sag_pat2.dicom.zip	MR: Structural, T2, T1	dicom
5_se_t1_sag_pat2.nii.gz	MR: Structural, T2, T1	nifti
6 - t2_tse_tra_p2 19:00		
6 - t2_tse_tra_p2.dicom.zip	MR: Structural, T2	dicom
6_t2_tse_tra_p2.nii.gz	MR: Structural, T2	nifti
11 - t1_se_ax_multi dir flow comp 19:00		

Kids First Imaging -- Flywheel Pilot



	SUBJECTS	SESSIONS	PERMISSION	USERS	New Projec
	20	55	admin	S R A +3	
	1	1	admin		
	6	8	admin	S R S +2	
	2	2	admin		
	42	89	admin	S R S +2	
	265	592	admin	S R S +2	
	80	236	admin	S R +3	
	519	626	admin	S S +3	
	277	524	admin	S R C +3	

Project-level management

Kids First Imaging -- Flywheel Pilot

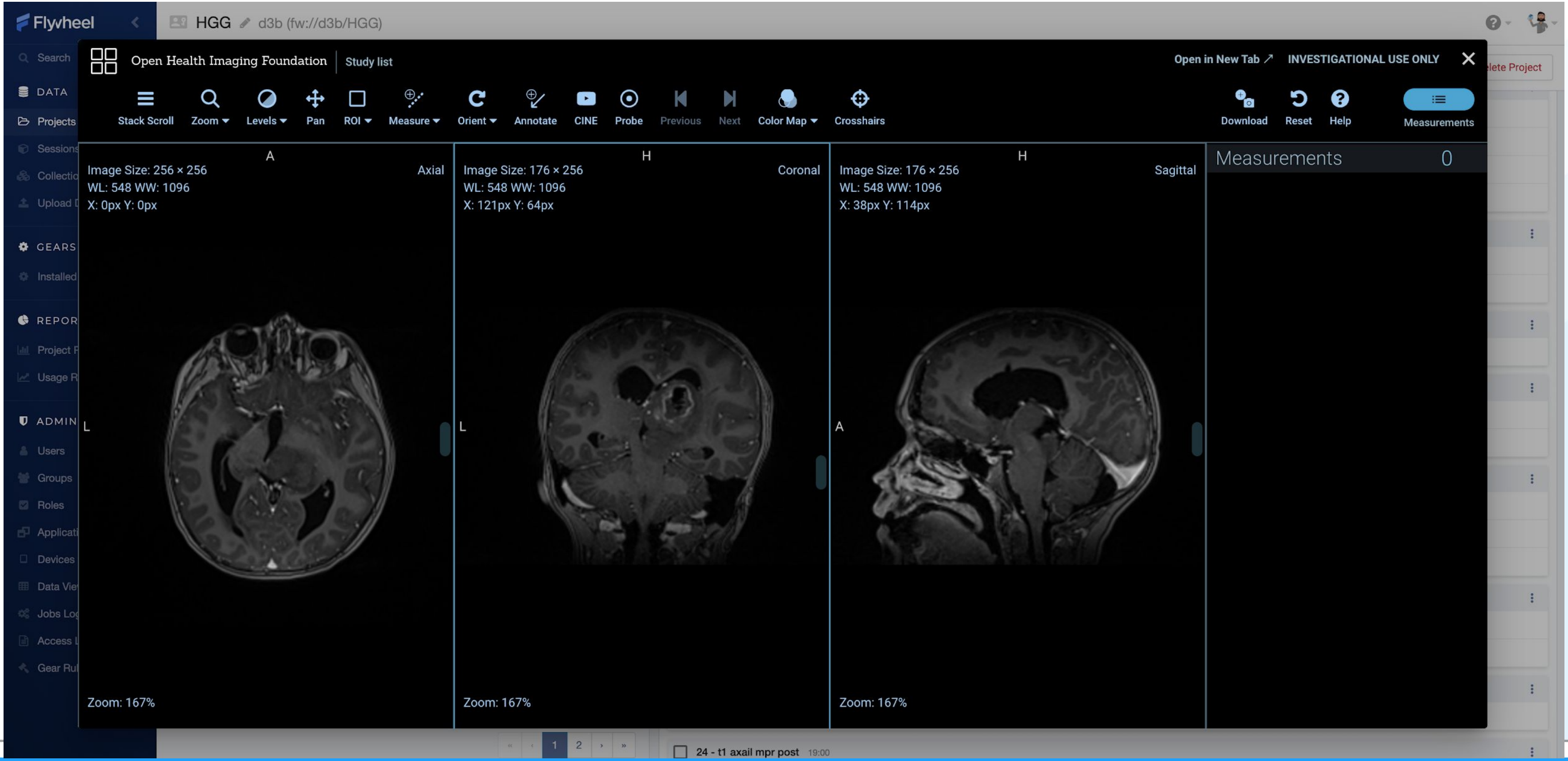


Image Size: 256 × 256
WL: 548 WW: 1096
X: 0px Y: 0px
Axial
Zoom: 167%

Image Size: 176 × 256
WL: 548 WW: 1096
X: 121px Y: 64px
Coronal
Zoom: 167%

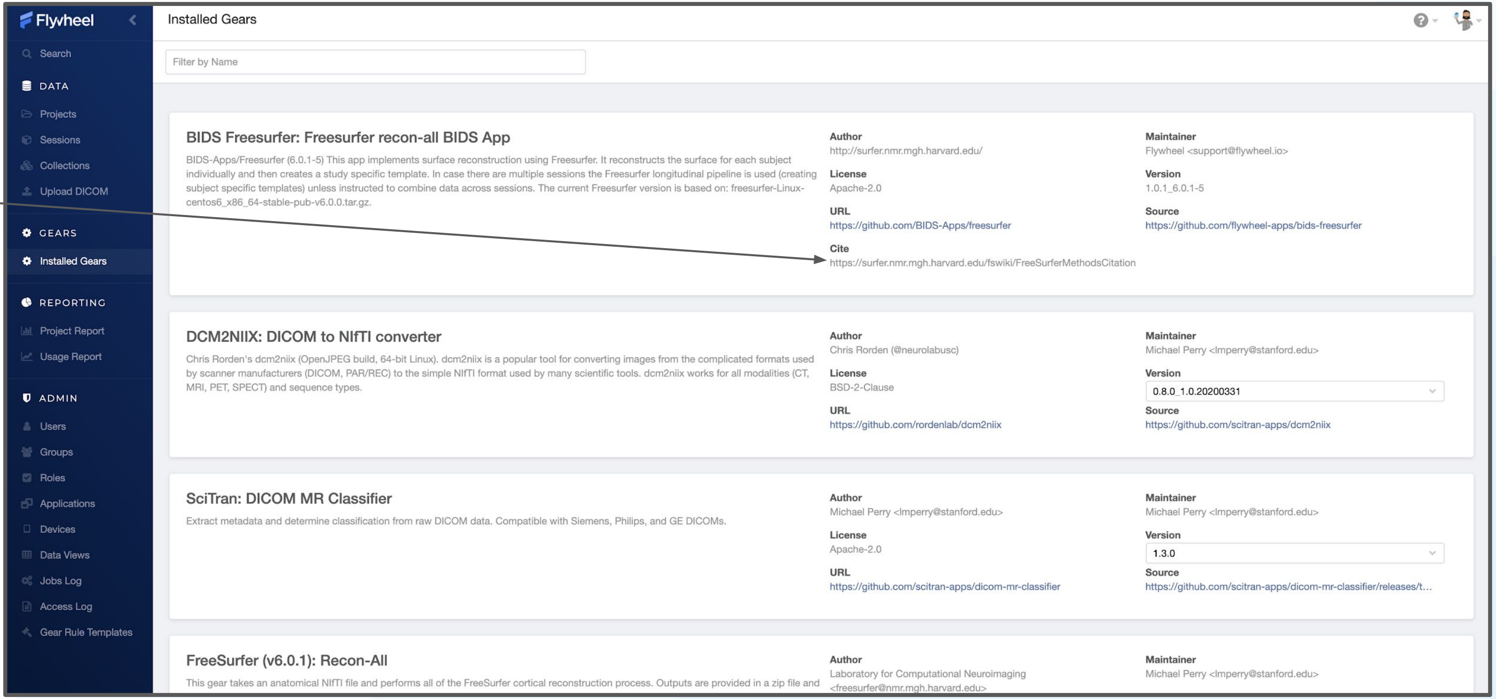
Image Size: 176 × 256
WL: 548 WW: 1096
X: 38px Y: 114px
Sagittal
Zoom: 167%

Measurements 0

24 - t1 axail mpr post 19:00

Kids First Imaging -- Flywheel Pilot

Shareable Workflows



Installed Gears

BIDS Freesurfer: Freesurfer recon-all BIDS App

BIDS-Apps/Freesurfer (6.0.1-5) This app implements surface reconstruction using Freesurfer. It reconstructs the surface for each subject individually and then creates a study specific template. In case there are multiple sessions the Freesurfer longitudinal pipeline is used (creating subject specific templates) unless instructed to combine data across sessions. The current Freesurfer version is based on: freesurfer-Linux-centos6_x86_64-stable-pub-v6.0.0.tar.gz.

Author
http://surfer.nmr.mgh.harvard.edu/

License
Apache-2.0

URL
https://github.com/BIDS-Apps/freesurfer

Cite
https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferMethodsCitation

Maintainer
Flywheel <support@flywheel.io>

Version
1.0.1_6.0.1-5

Source
https://github.com/flywheel-apps/bids-freesurfer

DCM2NIIX: DICOM to NIfTI converter

Chris Rorden's dcm2niix (OpenJPEG build, 64-bit Linux). dcm2niix is a popular tool for converting images from the complicated formats used by scanner manufacturers (DICOM, PAR/REC) to the simple NIfTI format used by many scientific tools. dcm2niix works for all modalities (CT, MRI, PET, SPECT) and sequence types.

Author
Chris Rorden (@neurolabusc)

License
BSD-2-Clause

URL
https://github.com/rordenlab/dcm2niix

Maintainer
Michael Perry <lmperry@stanford.edu>

Version
0.8.0_1.0.20200331

Source
https://github.com/scitran-apps/dcm2niix

SciTran: DICOM MR Classifier

Extract metadata and determine classification from raw DICOM data. Compatible with Siemens, Philips, and GE DICOMs.

Author
Michael Perry <lmperry@stanford.edu>

License
Apache-2.0

URL
https://github.com/scitran-apps/dicom-mr-classifier

Maintainer
Michael Perry <lmperry@stanford.edu>

Version
1.3.0

Source
https://github.com/scitran-apps/dicom-mr-classifier/releases/t...

FreeSurfer (v6.0.1): Recon-All

This gear takes an anatomical NIfTI file and performs all of the FreeSurfer cortical reconstruction process. Outputs are provided in a zip file and

Author
Laboratory for Computational Neuroimaging <freesurfer@nmr.mgh.harvard.edu>

Maintainer
Michael Perry <lmperry@stanford.edu>

Discussion

- Do the 3 use cases capture what is needed?
- How far are we from Imaging data being an important driver of interoperability
- Should we set up an Imaging Interoperability Workgroup?

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

Quick Break

We will resume at 3:30 pm ET.

Group Discussion
Refining a Road Map

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

AnVIL

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

What We See From Here

Draft Roadmap in FunRetro
(from Day 1)

RAS and Data Sharing in FunRetro
(from Day 2)

Prioritizing Interoperability: Datasets and Initiatives

Can we leverage and define specific datasets and initiatives as the “canvas” against which we prioritize our interoperability efforts?

GTEx

👍 5 💬 1

Common Fund Data Ecosystem (CFDE) - goal is to build interoperability at the metadata and data level for 8 + Common Fund programs

👍 8 💬 0

Cancer Data Aggregator & CCDH for discovery and query across NCI datasets

👍 11 💬 1

Workflow interoperability?

Cloud vendor interoperability?

Imaging use case review?

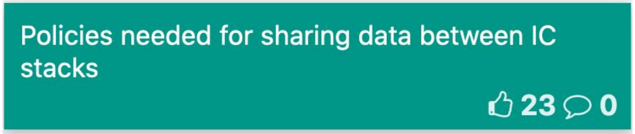
Training and outreach?

FHIR Use Cases?

2021 Outline by End of November

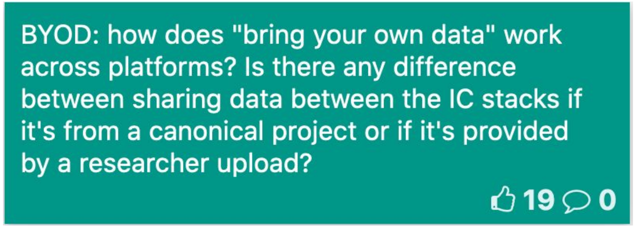
Governance WG

- Recap of current next steps



Policies needed for sharing data between IC stacks

👍 23 💬 0



BYOD: how does "bring your own data" work across platforms? Is there any difference between sharing data between the IC stacks if it's from a canonical project or if it's provided by a researcher upload?

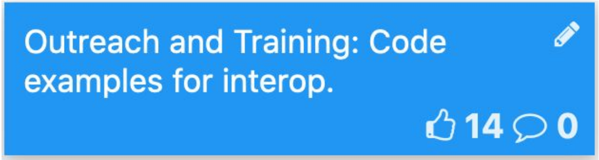
👍 19 💬 0

Need to also consider intersections of governance and implementations, e.g.:

- Cross-stack workflow execution, meta-data association,
- Fully executable policies

Outreach WG

- Recap of current next steps
- Tighter integration with FHIR group
- Exploring GA4GH Discovery API



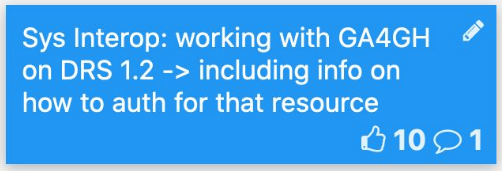
Outreach and Training: Code examples for interop.

👍 14 💬 0

2021 Outline by End of November

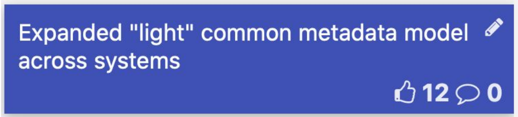
System Interop WG

- Recap of current next steps



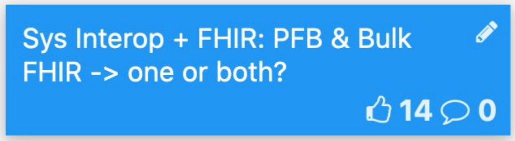
Sys Interop: working with GA4GH on DRS 1.2 -> including info on how to auth for that resource

👍 10 💬 1



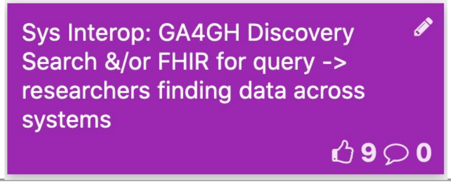
Expanded "light" common metadata model across systems

👍 12 💬 0



Sys Interop + FHIR: PFB & Bulk FHIR -> one or both?

👍 14 💬 0



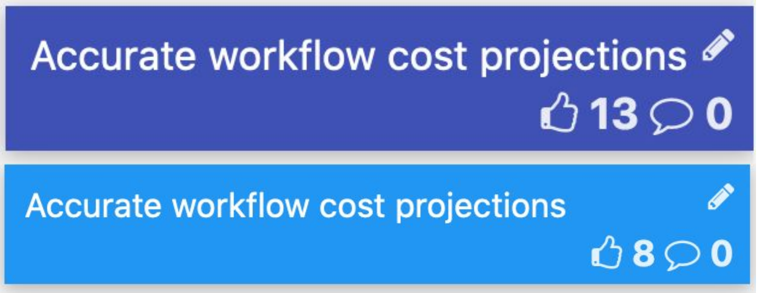
Sys Interop: GA4GH Discovery Search &/or FHIR for query -> researchers finding data across systems

👍 9 💬 0

FHIR WG

- Document current best practices for NCPI FHIR Model
- Limited scope pilot of prod data, access, and tools
- Community engagement
- Identify path to providing spanning set of data and metadata
- Continue to develop on new projects and data, especially emerging studies, eg MIS-C

Cross WG Items

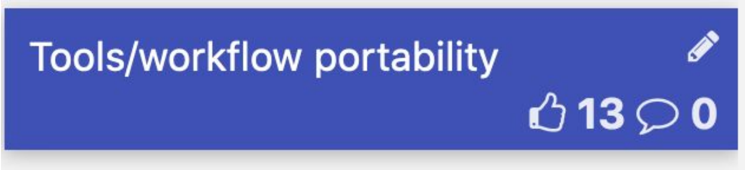


Accurate workflow cost projections ✎

👍 13 💬 0

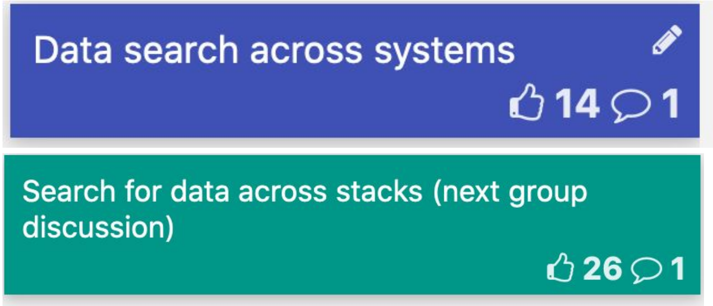
Accurate workflow cost projections ✎

👍 8 💬 0



Tools/workflow portability ✎

👍 13 💬 0



Data search across systems ✎

👍 14 💬 1

Search for data across stacks (next group discussion)

👍 26 💬 1

NIH Workshop on Cloud-Based Platforms Interoperability
October 30th and November 2nd, 2020

Thank You for Attending!

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

AnVIL

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

Other Template Slides

Feel Free to Copy/Paste as Needed

This Is Where the Title or Headline Goes.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

This Is Where the Title or Headline Goes.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

This Is Where the Title or Headline Goes.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

This Is Where the Title or Headline Goes.

Compared Subject #1

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Compared Subject #2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

This Is Where the Title or Headline Goes.

Compared Subject #1

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Compared Subject #2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.