



# DATA BIOSPHERE

## *An Introduction*



UNIVERSITY OF CALIFORNIA  
**SANTA CRUZ**

Dr. Benedict Paten

UC Santa Cruz Genomics Institute



@BenedictPaten

<https://www.databiosphere.org/>



**Sage**Bionetworks

Dr. Brian O'Connor  
Sage Bionetworks



@boconnor



**BROAD**  
INSTITUTE

Dr. Tim Tickle

Broad Institute, Data Sciences Platform  
[ttickle@broadinstitute.org](mailto:ttickle@broadinstitute.org)



@timothy\_tickle

# A Data Biosphere for Biomedical Research



Benedict Paten Oct 16, 2017 · 5 min read

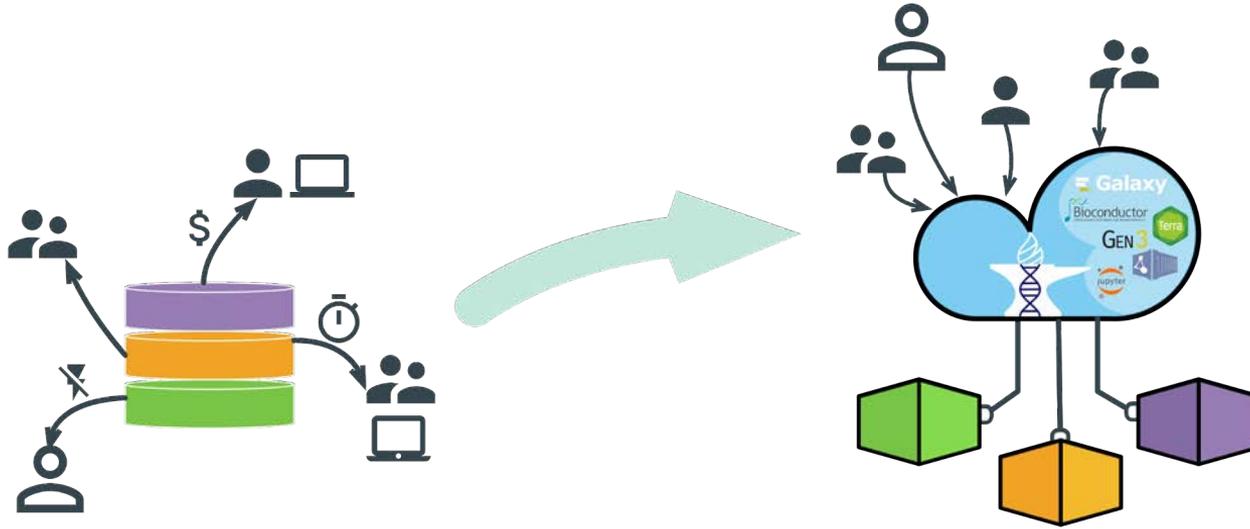


*We, the authors listed below, are privileged to be part of the growing global community bringing data and life science together. Our groups have been working together in overlapping combinations during the past two years to drive the creation of data commons to support flagship scientific initiatives. This document lays out our evolving vision for the next steps in that journey. Our hope is that others will join the effort to build momentum for an open, compatible, and secure approach to data within the larger research community. We welcome your feedback, and look forward to continuing this journey together.*

*Josh Denny (Vanderbilt), David Glazer (Verily Life Sciences), Robert L. Grossman (University of Chicago), Benedict Paten (University of California at Santa Cruz), Anthony Philippakis (Broad Institute)*

Problem: **data is getting too big**  
(to individually download and store)

# Data Biosphere: Invert the Model of Data Sharing



## *Traditional: Bring data to the researcher*

- Copying/moving data is costly
- Harder to enforce security
- Redundant infrastructure
- Siloed compute

## *Goal: Bring researcher to the data*

- Reduced redundancy and costs
- Active threat detection and auditing
- Greater accessibility
- Easier collaboration across institutions
- Elastic, shared, compute

# How Should a Data Biosphere be Structured?

<b>MODULAR</b>	Comprised of functional components with well-specified interface
<b>COMMUNITY FOCUSED</b>	Created by many groups to foster a diversity of ideas
<b>OPEN</b>	Open-source licenses, software, architecture to enable extensibility
<b>STANDARDS BASED</b>	Consistent with standards developed by coalitions such as GA4GH

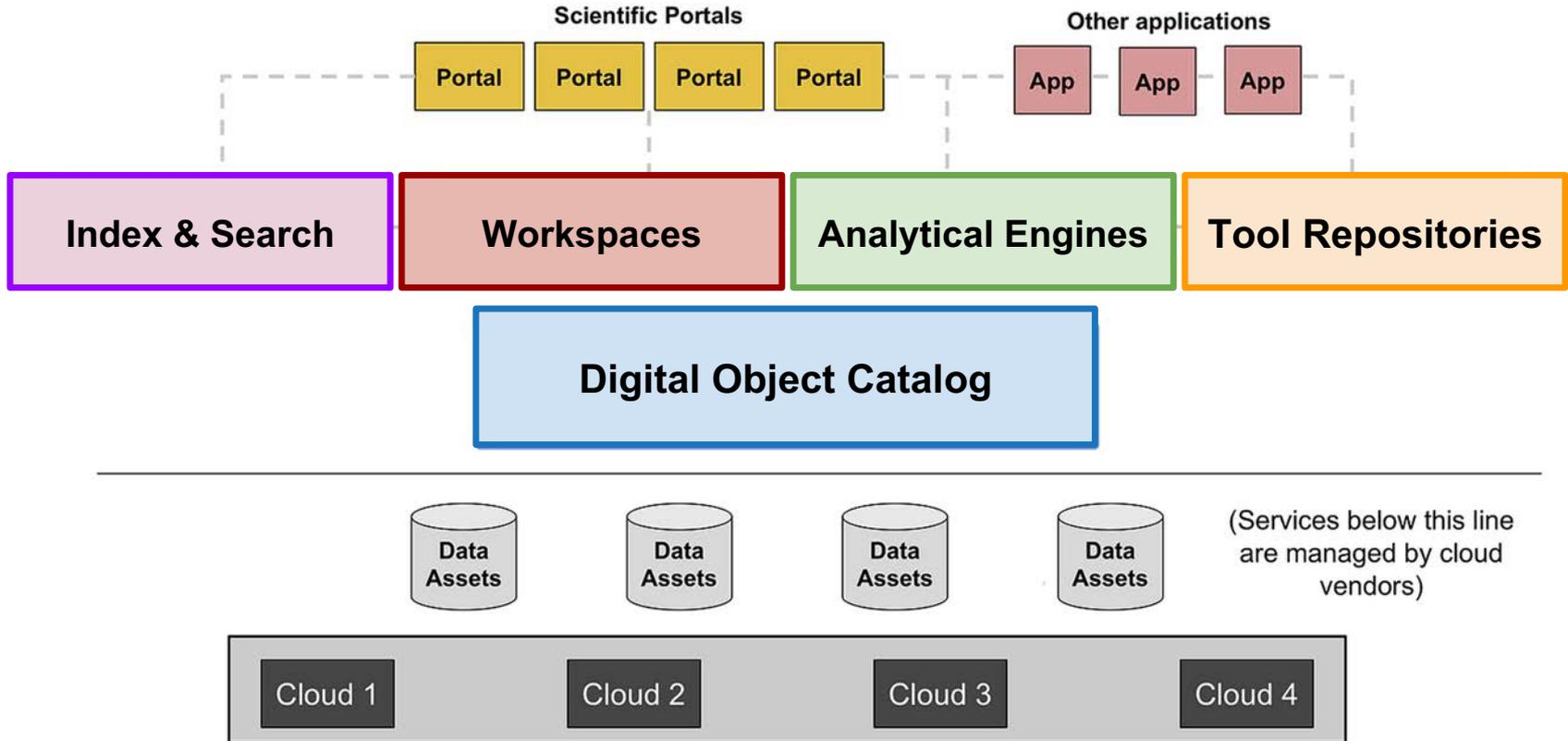


A Data Biosphere is...

*Modular*

# Modular Components

*We designed the Data Biosphere around key components — each having discrete capabilities and clear rules of interaction*





A Data Biosphere is...  
*Community Focused*

# Projects using components of the Data Biosphere:



Nurses' Health Study

**All of Us**  
RESEARCH PROGRAM

**gp<sup>2</sup>** Global Parkinson's Genetics Program

**biobank<sup>uk</sup>**  
Improving the health of future generations

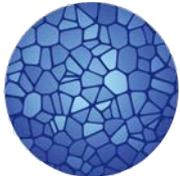


NHGRI AnVIL

**Single Cell**  
PORTAL

**LungMAP**  
Molecular Atlas of Lung Development Program

**BICCN**



**HUMAN CELL ATLAS**

**NHLBI**  
**TOPMed**



National Heart, Lung, and Blood Institute

**BioData**  
**CATALYST**

# GA4GH Provides Many Core Interoperability Standards for the Data Biosphere



*The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework.*



A Data Biosphere is...

*Modular*

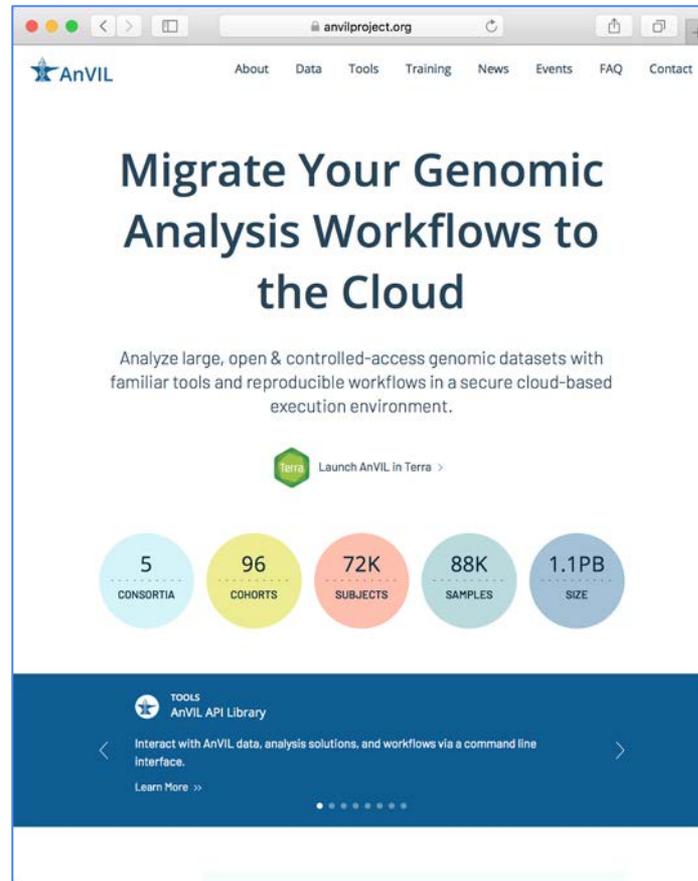


NHGRI AnVIL Shows Data Biosphere  
Modules in Action

# What is the NHGRI's AnVIL?

NHGRI funded the **Broad Institute**, **Johns Hopkins**, and multiple additional groups, including **UCSC** and **U. Chicago**, to build a platform inspired by the principles of the *Data Biosphere*

- **Cloud-based, scalable and interoperable computing resource**
- **Secure data access environment**
- **Collaborative computing environment for datasets and analysis workflows**
- <https://anvilproject.org>

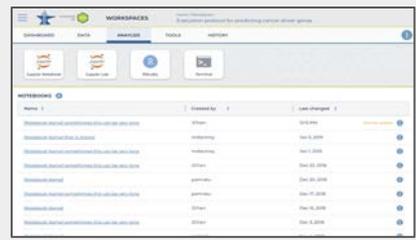


The screenshot shows the homepage of the AnVIL project website. The browser address bar displays 'anvilproject.org'. The navigation menu includes 'About', 'Data', 'Tools', 'Training', 'News', 'Events', 'FAQ', and 'Contact'. The main heading reads 'Migrate Your Genomic Analysis Workflows to the Cloud'. Below this, a sub-heading states: 'Analyze large, open & controlled-access genomic datasets with familiar tools and reproducible workflows in a secure cloud-based execution environment.' A 'Launch AnVIL in Terra >' button is visible. A row of five circular statistics is shown: 5 CONSORTIA (light blue), 96 COHORTS (yellow), 72K SUBJECTS (orange), 88K SAMPLES (teal), and 1.1PB SIZE (blue). At the bottom, a dark blue section features the 'TOOLS AnVIL API Library' logo and the text: 'Interact with AnVIL data, analysis solutions, and workflows via a command line interface. Learn More >>'.

# AnVIL Modules



**GEN3** Data Commons  
Data models,  
indexing, querying



**Terra** Workspaces,  
workflows, analysis



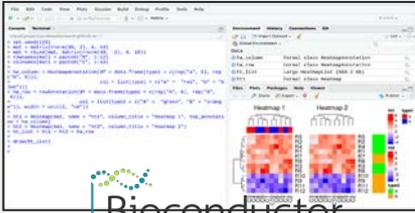
**Dockstore**  
Create, Share, Use  
Sharing containerized tools  
and workflows



**jupyter**  
Live code, equations,  
visualizations and narratives



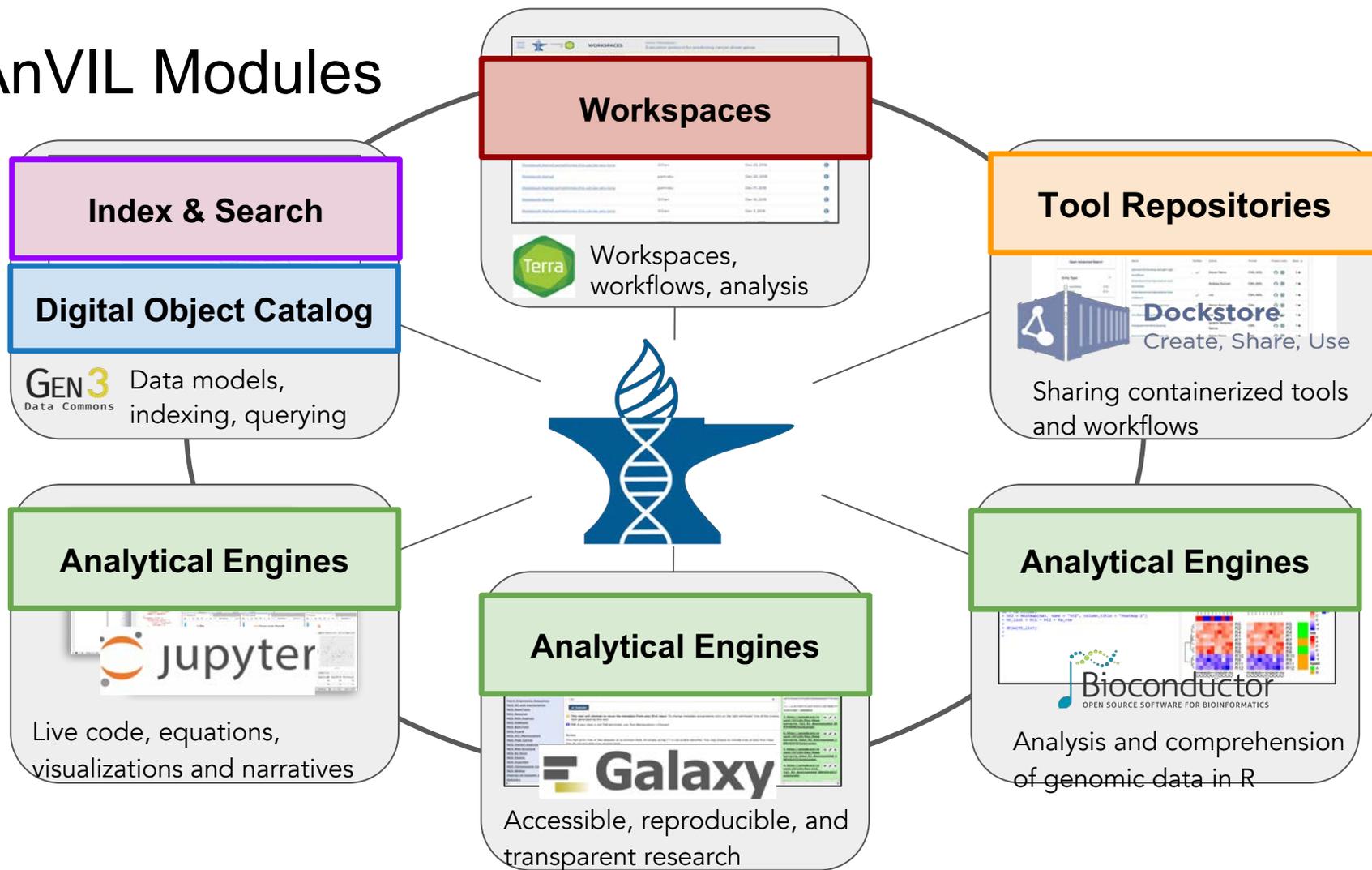
**Galaxy**  
Accessible, reproducible, and  
transparent research



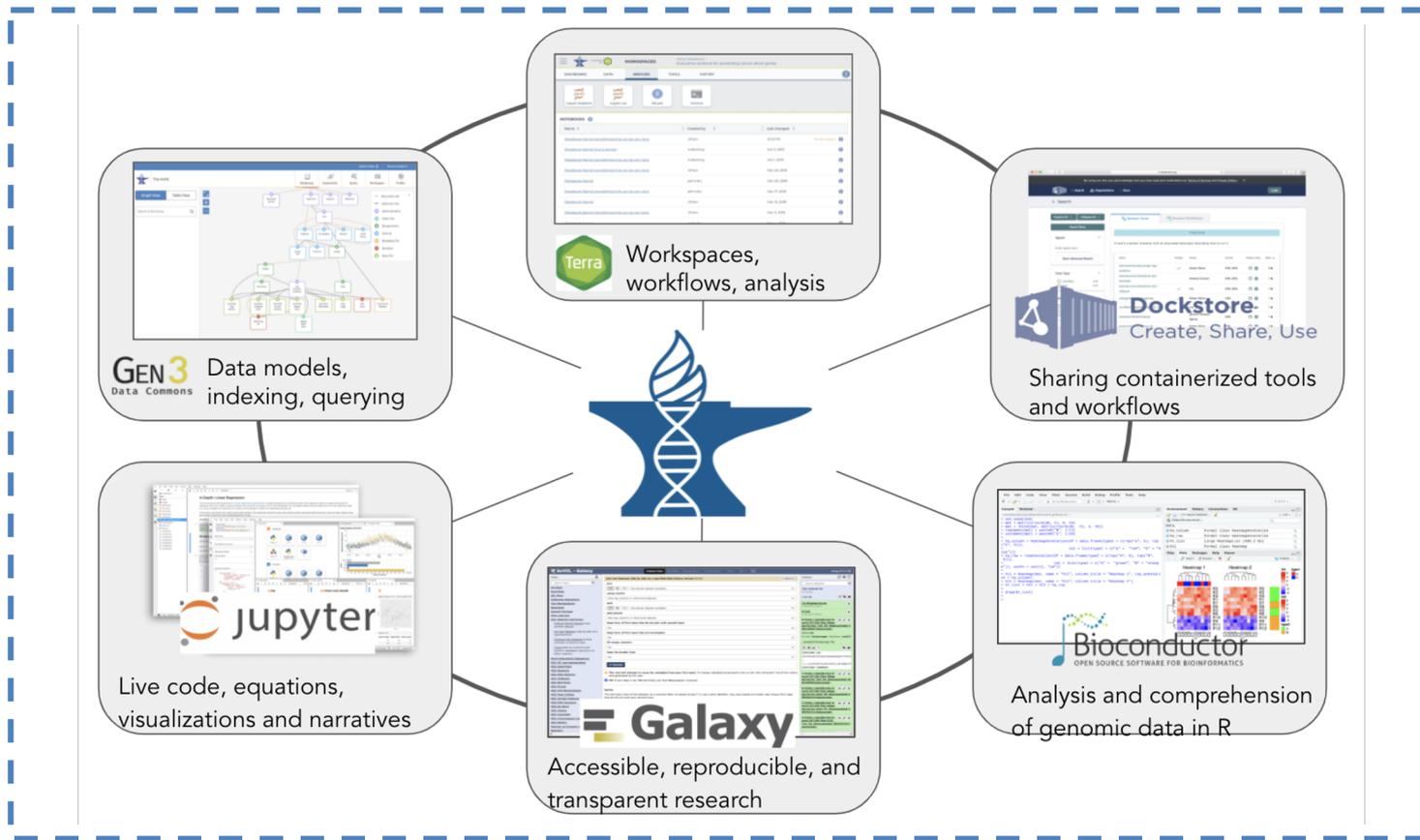
**Bioconductor**  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS  
Analysis and comprehension  
of genomic data in R



# AnVIL Modules



# AnVIL Modules Deployed in FISMA Moderate Environment



FISMA Moderate  
2 ATOs

**Terra recently  
achieved FedRAMP**

# Data Biosphere Modules Power Platforms

- A Data Biosphere is *not just about standalone modules*
- **AnVIL** is a great illustration of the various Data Biosphere Modules in a federated environment
- AnVIL is important because it illustrates *how modules can be assembled to form a Federated Data Biosphere Platform*
- **Terra** is an underlying, fully-formed platform built with Data Biosphere-inspired modules



<https://terra.bio>



A Data Biosphere is...  
*Community Focused*

# Advancing access to TOPMed data

BioData Catalyst provides one point of entry to the most TOPMed datasets, including Freeze 8 data.

**406,853**

Participants

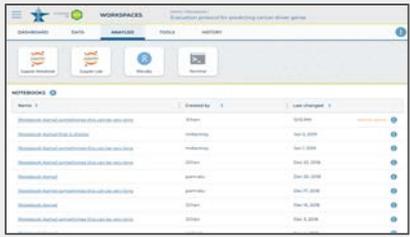
**3.42**

Petabytes of Data

**Access biomedical data  
when you need it and how  
you need it**



<https://biodatacatalyst.nhlbi.nih.gov>



**Terra** Workspaces,  
workflows, analysis



**Dockstore**  
Create, Share, Use

Sharing containerized tools  
and workflows



**GEN3** Data models,  
indexing, querying

**NIH** National Heart, Lung,  
and Blood Institute

**BioData**  
**CATALYST**

**SevenBridges**

The Seven Bridges workspace  
environment (CWL)



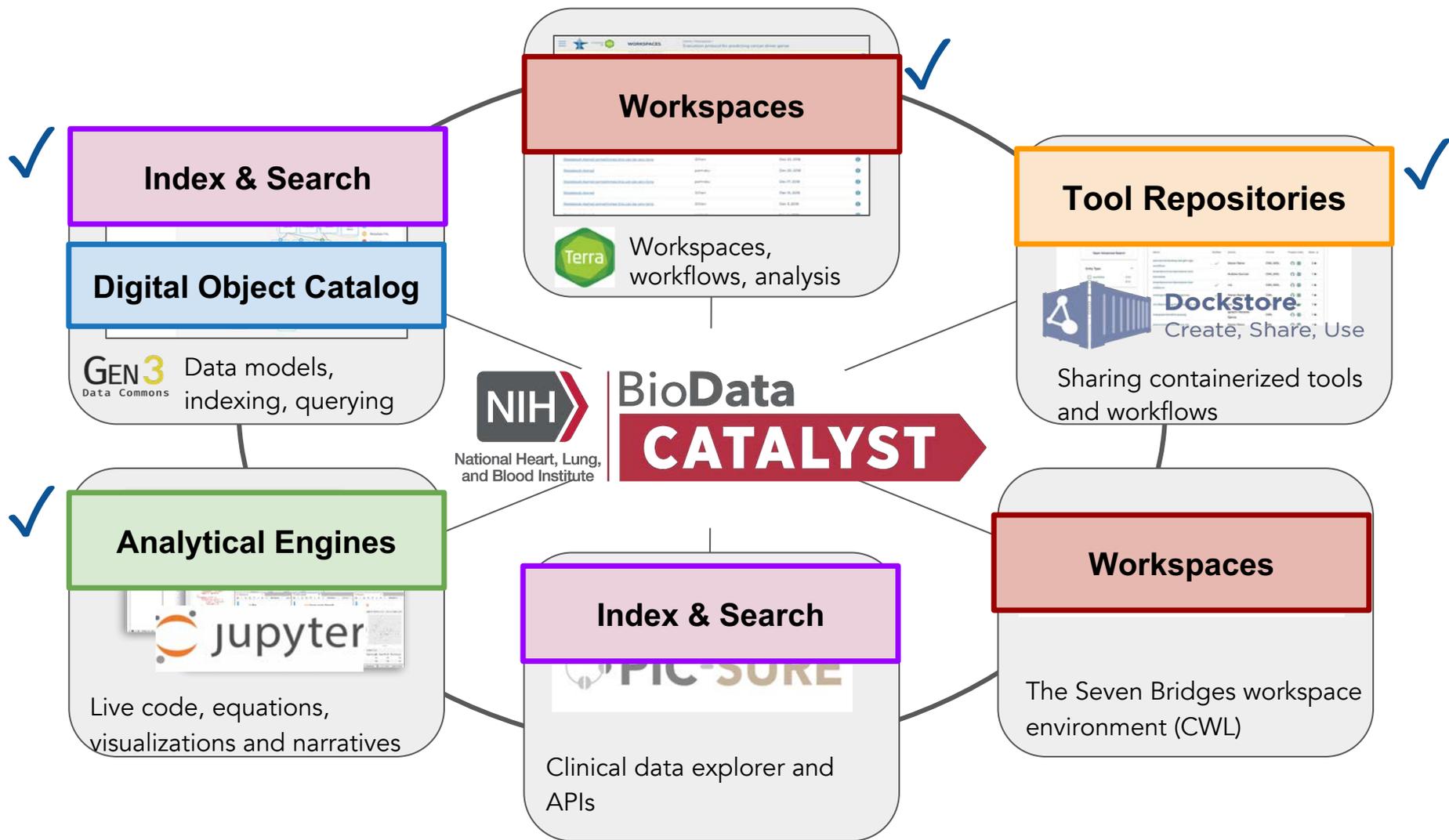
**PIC-SURE**

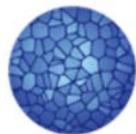
Clinical data explorer and  
APIs



**jupyter**

Live code, equations,  
visualizations and narratives

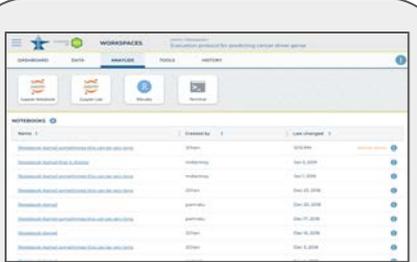




# HUMAN CELL ATLAS DATA PORTAL



UNIVERSITY OF CALIFORNIA  
**SANTA CRUZ** Data models,  
indexing, querying



 Storage, workspaces,  
workflows, analysis



 **Dockstore**  
Create, Share, Use

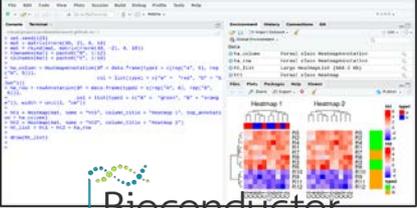
Sharing containerized tools  
and workflows



 Live code, equations,  
visualizations and narratives

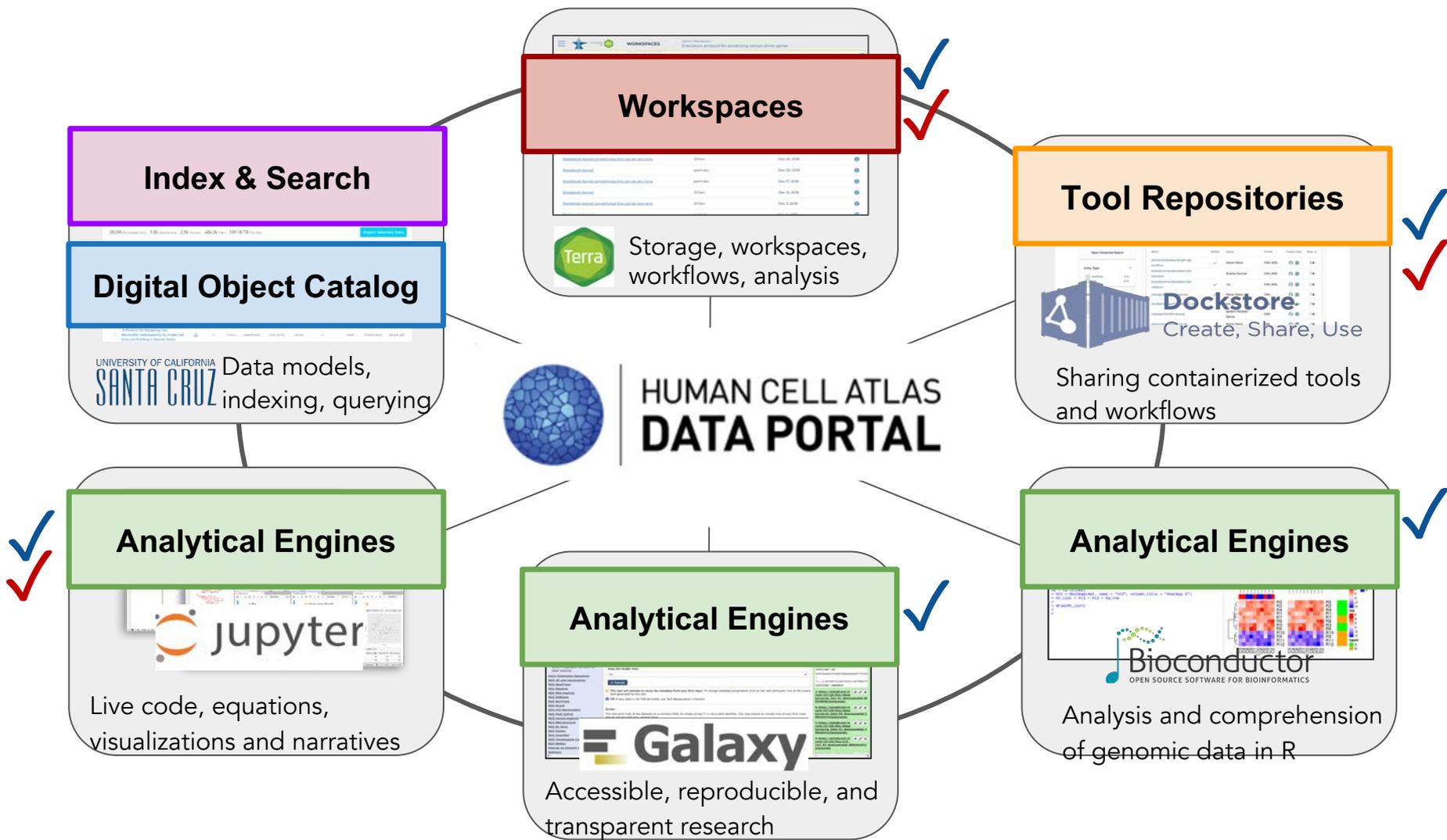


 Accessible, reproducible, and  
transparent research



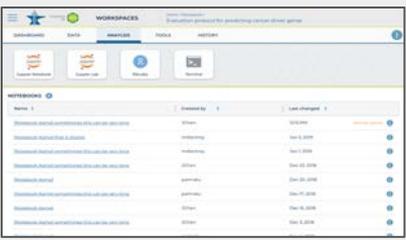
 **Bioconductor**  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Analysis and comprehension  
of genomic data in R





UNIVERSITY OF CALIFORNIA  
**SANTA CRUZ** Data models,  
indexing, querying



**Terra** Storage, workspaces,  
workflows, analysis



**Dockstore**  
Create, Share, Use  
Sharing containerized tools  
and workflows



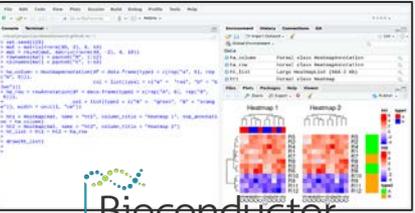
**LungMAP**  
Molecular Atlas of Lung  
Development Program



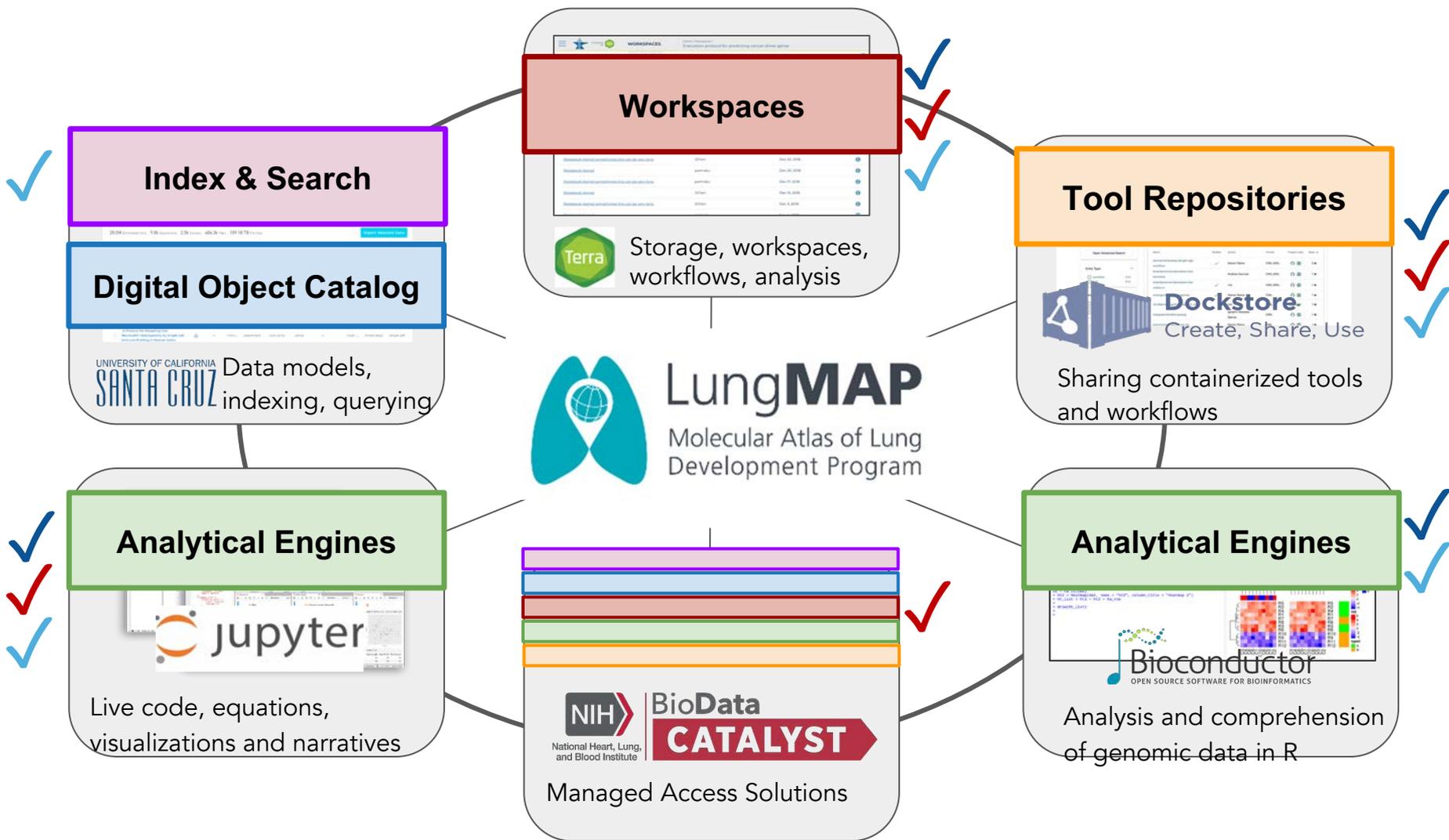
**jupyter**  
Live code, equations,  
visualizations and narratives



**BioData  
CATALYST**  
Managed Access Solutions



**Bioconductor**  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS  
Analysis and comprehension  
of genomic data in R





A Data Biosphere is...

*Open*

# Open Source Throughout



## Data Biosphere

We are creating a vibrant ecosystem of interoperable modules and data environments for the biomedical community.  
<https://www.databiosphere.org/>

[Overview](#) [Repositories 121](#) [Packages](#) [People 20](#) [Projects](#)

### Popular repositories

**toil** Public

A scalable, efficient, cross-platform (Linux/macOS) and easy-to-use workflow engine in pure Python.

Python 777 223

**dsub** Public

Open-source command-line tool to run batch computing tasks and workflows on backend services such as Google Cloud.

Python 204 38

**terra-ui** Public

Web user interface for the Terra platform

JavaScript 40 13

**leonardo** Public

Notebook service

Scala 27 15

**job-manager** Public

Job Manager API and UI for interacting with asynchronous batch jobs and workflows.

TypeScript 21 5

**azul** Public

Metadata indexer and query service used for HCA and CGP

Python 17 5

### People



### Top languages

Python Java JavaScript Scala Shell

### Most used topics

pipeline workflow

[Repositories](#)

# Example: Single cell transcriptomics

- [Cumulus workflows on Dockstore](#)

- Generate counts matrices
- Demultiplex hashed nuclei
- Single cell/single nucleus analysis

- [Cumulus documentation](#)

**AnVIL** Analysis, Visualization and Informatics Lab-space /  
**Cumulus**  
Cloud-based single-cell/single-nucleus genomics analysis workflows.

[github.com/klarman-cell-observatory/cumulus/Count](https://github.com/klarman-cell-observatory/cumulus/Count)  
Last updated May 31, 2020

[github.com/klarman-cell-observatory/cumulus/Cumulus\\_subcluster](https://github.com/klarman-cell-observatory/cumulus/Cumulus_subcluster)  
Last updated May 31, 2020

[github.com/klarman-cell-observatory/cumulus/Smart-Seq2\\_create\\_reference](https://github.com/klarman-cell-observatory/cumulus/Smart-Seq2_create_reference)  
Last updated May 31, 2020

[github.com/klarman-cell-observatory/cumulus/Cellranger\\_atac\\_aggr](https://github.com/klarman-cell-observatory/cumulus/Cellranger_atac_aggr)  
Last updated May 31, 2020

**About Cumulus**

Cumulus is a cloud-based framework, which aims to achieve a scalable, comprehensive, cost-effective, and user-friendly analysis solution on single-cell/single-nucleus genomics. It consists of a series of workflows in WDL, which covers from sequencer output extraction to downstream analysis, and across different protocols and omics assays.

**Tutorial**

- Cumulus featured workspace on Terra.
- Example of cell-hashing and CITE-Seq analysis using Cumulus.
- Tutorials on Downstream analysis using Pegasus.
- Cumulus tutorial videos on Youtube.

**Documentation**

Please refer to <https://cumulus.readthedocs.io> for detailed documentation on Cumulus.

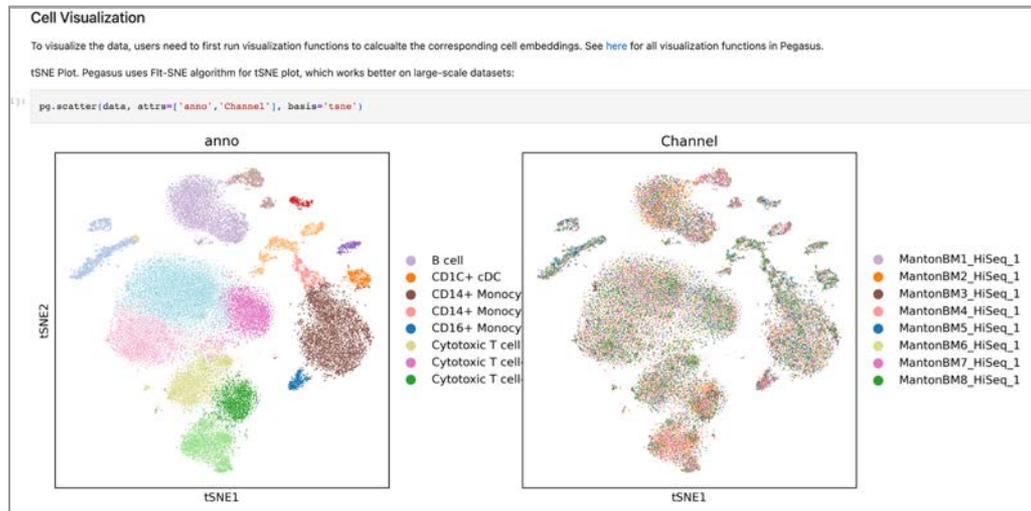
Cumulus uses Pegasus for downstream analysis. Pegasus is a Python package which can be used separately, and its documentation website is: <https://pegasus.readthedocs.io>.

Contributed by: Bo Li & Yiming Yang (Cumulus Team, Genentech)

# Example: Single cell transcriptomics

- [Cumulus tutorial](#) on Terra
- FASTQ to a normalized counts matrix
- Differential expression analysis
- Clustering analysis
- Visualize data using multiple algorithms

[Nature Communications Vol. 10: 2907 \(2019\)](#)



Contributed by: Bo Li & Yiming Yang (Cumulus Team, Genentech)



A Data Biosphere is...  
*Standards Based*

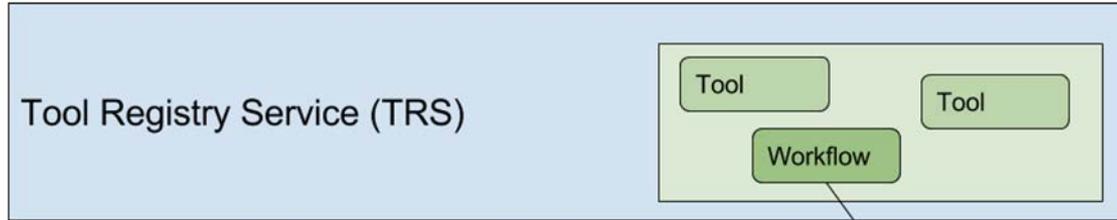
# GA4GH Provides Many Core Interoperability Standards for the Data Biosphere



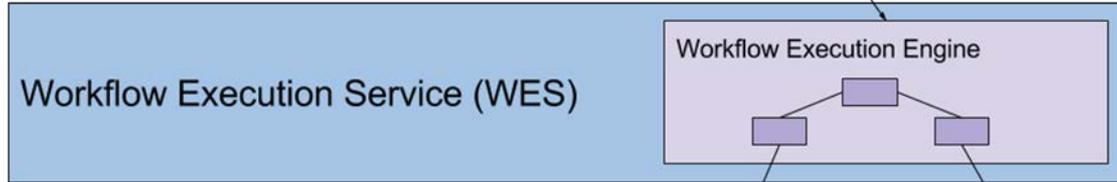
*The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework.*

# Key GA4GH Cloud Interoperability Standards

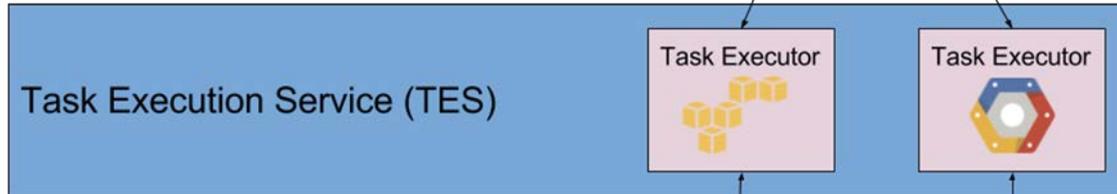
**Sharing Tools  
and Workflows**



**Executing  
Workflows**



**Executing  
Individual  
Tasks**



**Accessing Data**



# NCPI Effort - Breaking Down Data Silos in NIH

*The **NIH Cloud Platform Interoperability (NCPI)** effort empowers end-users to analyze data across participating platforms.*

*It facilitates the realization of a **trans-NIH, federated data ecosystem** by establishing and implementing guidelines and technical standards.*

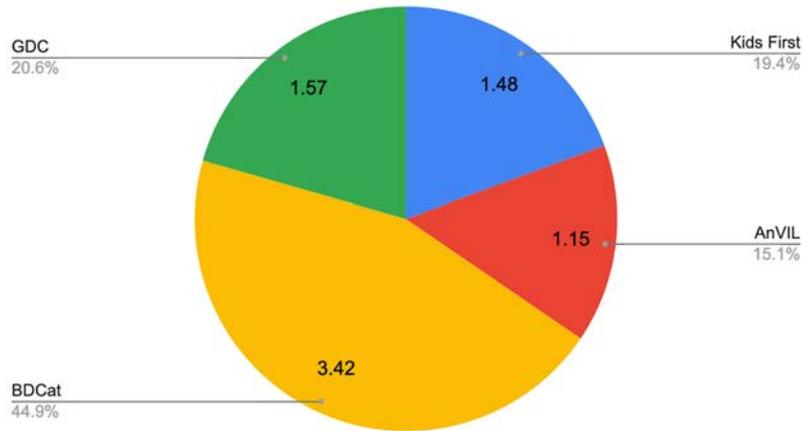


<https://anvilproject.org/ncpi>

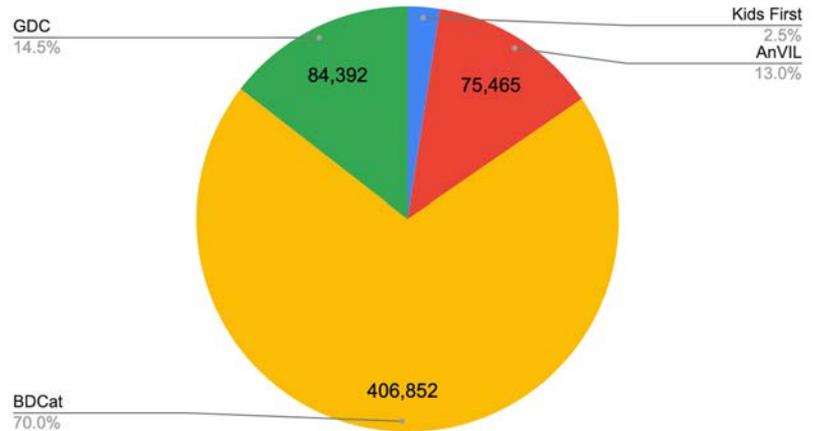
# Challenges & Opportunities of Data Growth

*Extraordinary growth of data... in just 4 NIH platforms (AnVIL, BioData Catalyst, CRDC, and GMKF) we see ~8PB of data accessible covering ~600K participants*

Data Size (PB)

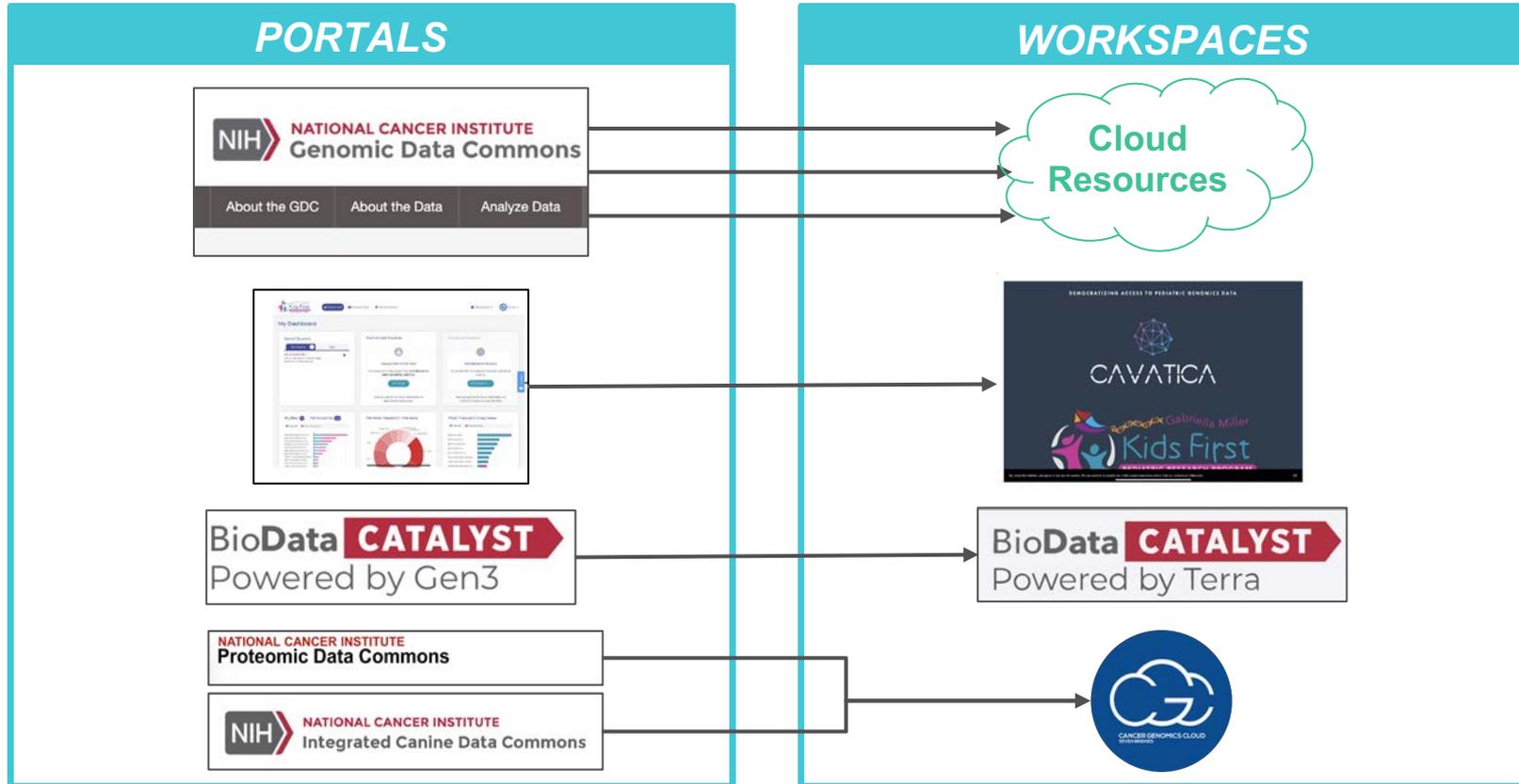


Participants



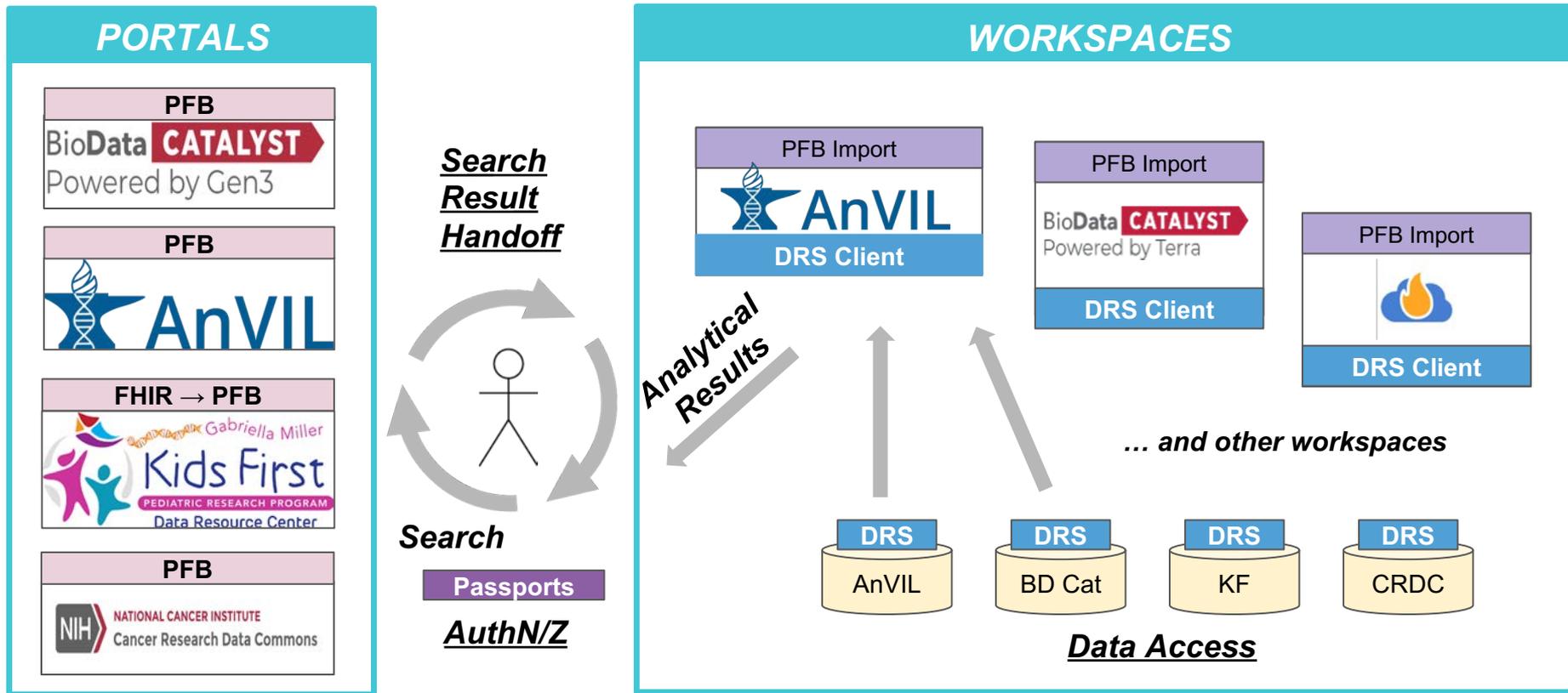
# Data Silos & FAIR Systems Interoperation

*Data portals connect (intra-IC) with analysis systems (workspaces)*



# NCPI Vision for FAIR Systems Interop

Data portals connect to any *workspaces (inter-IC)*, workspace access *data (inter-IC)*



# NCPI Systems Interop by the Numbers

Collectively, we have achieved improved interoperability in 2020-21 across multiple systems through **PFB/manifests**, **GA4GH DRS**, and **GA4GH Passports (RAS)**.

## Mid-2021 Results

- Search Handoff: PFB, manifests

4 portals,  
**~581K subjects**



- Data Access: GA4GH DRS 1.1

4 DRS Servers  
**~7.6PB of data**



- Auth: RAS for AuthN

**RAS GA4GH Passports**



## Supported Platforms

BioData **CATALYST**  
Powered by Terra

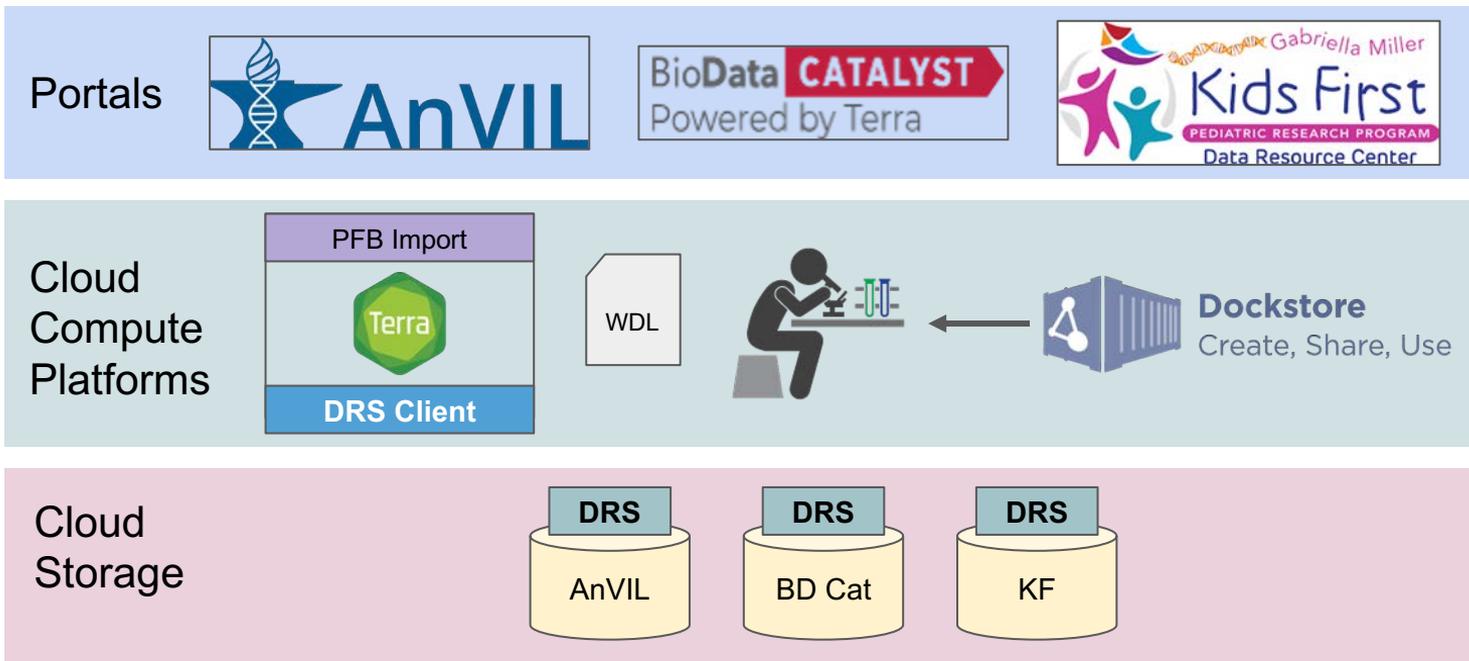


**NIH** NATIONAL CANCER INSTITUTE  
Cancer Research Data Commons



# Researcher Use Cases - An NCPI Success Story

Researchers are using NCPI systems through GA4GH standards e.g. [Use Case #7](#): Tim Majarian's cross dataset analysis for Congenital Heart Disease



Data Biosphere encourages the use of GA4GH API standards to facilitate work across Data Biosphere implementations



# DATA BIOSPHERE

We now see an ecosystem of platforms that support the next generation of biomedical research using Data Biosphere principles:

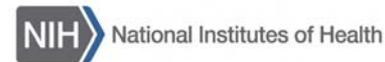
**Modularity:** Many components exist and build platforms like Terra

**Community Focus:** Many groups collaborate together

**Openness:** Many projects use Open Source approaches

**Standards Adoption:** Many projects use interoperability standards

# Thank You!



## Special thanks to:

- Anthony Philippakis
- Robert Grossman
- John Marioni
- Timothy Tickle
- Joshua Denny
- David Glazer
- Elizabeth Sheets
- Timothy Harris
- Helen Parkinson



<https://www.databiosphere.org>



# For More Information...

<https://www.databiosphere.org>

Projects using Data Biosphere principles:

- [AnVIL Portal](#)
- [BioData Catalyst Portal](#)
- [HCA Data Portal - Human Cell Atlas](#)
- [LungMAP2 DCC](#)

Tools:

- [Dockstore](#)
- [Broad Methods Repository](#)

Workspaces:

- [Terra Featured Workspaces](#)