Center for Translational Data Science
AT THE UNIVERSITY OF CHICAGO

OPEN COMMONS CONSORTIUM

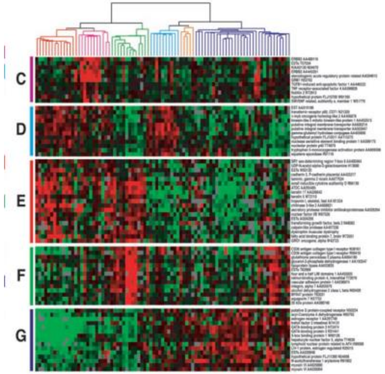
Introduction to the Gen3 Platform for Data Commons and Data Ecosystems

Phillis Tang

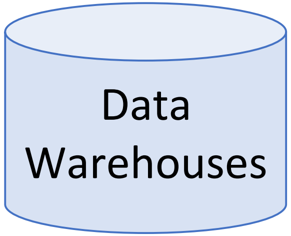
Center for Translational Data Science

University of Chicago

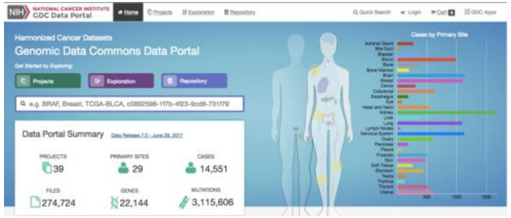
& Open Commons Consortium



Databases organize the data around a **project**.



Data warehouses organize the data for an **organization** (and are enabled by enterprise computing)



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

e.g. BRAF, Breast, TCGA-BLCA, c0892598-1f7b-4f23-9cd8-731f79

Data Portal Summary Data Release 7.0 - June 29, 2017

PROJECTS	PRIMARY SITES	CASES
39	29	14,551
FILES	GENES	MUTATIONS
274,724	22,144	3,115,606

Data Commons organize the data for a scientific discipline, **community**, or field and are enabled by large scale cloud computing.

Project

Databases

1982 - present

- Data repository
- Data catalogs
- Download data

(Virtual) Organization

OSDC OPEN SCIENCE DATA CLOUD

A Petabyte-scale Scientific Community Cloud

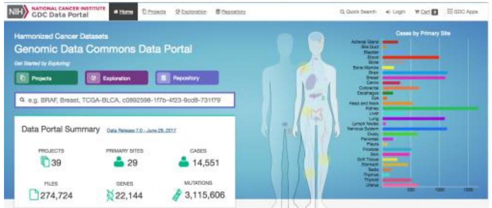
The OSDC enables scientific researchers to easily manage, share, and analyze large datasets.

Data Clouds

2010 - 2020

- Supports large data & data intensive computing with **cloud computing**
- Researchers can analyze data with collaborative tools (**workspaces**) – so data does **not** have to be downloaded)

Discipline



NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

e.g. BRAF, Breast, TCGA-BLCA, c0892598-1f7b-4f23-9cd8-731f79

Data Portal Summary Data Release 7.0 - June 29, 2017

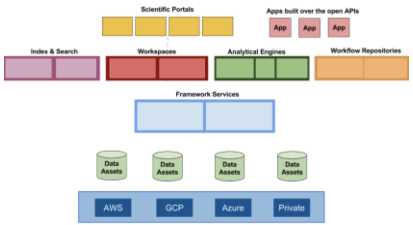
PROJECTS	PRIMARY SITES	CASES
39	29	14,551
FILES	GENES	MUTATIONS
274,724	22,144	3,115,606

Data Commons

2014 - 2024

- Supports large data
- Workspaces
- **Common data models**
- **Core data services**
- **Data & Commons Governance**
- **Harmonized data**
- **Data sharing**
- **Reproducible research**

Multi-Discipline



Data Ecosystems

2018 - 2028

- Interoperates **multiple data commons, databases, knowledge bases**, and other resources
- Supports **ecosystem of commons, portals, notebooks, applications & simulations** across multiple disciplines

Gen3 Data Commons

BRAIN Commons

Total Files: 16,722
Total Size: 270.14 GB

BloodPAC
BLOOD PROFILING ATLAS IN CANCER

Total Files: 1,825
Total Size: 242.77 GB

NHLBI
DATA STAGE

Total Files: 71,368
Total Size: 344.03 TB

ACCOUNT

Total Files: 1,952
Total Size: 3.77 TB

NIAID DATA HUB

Total Files: 2,553
Total Size: 271.52 GB

GenoMEL
the Melanoma Genetics Consortium

Total Files: 4,008
Total Size: 20.77 TB

Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

Total Files: 113,319
Total Size: 2.08 PB

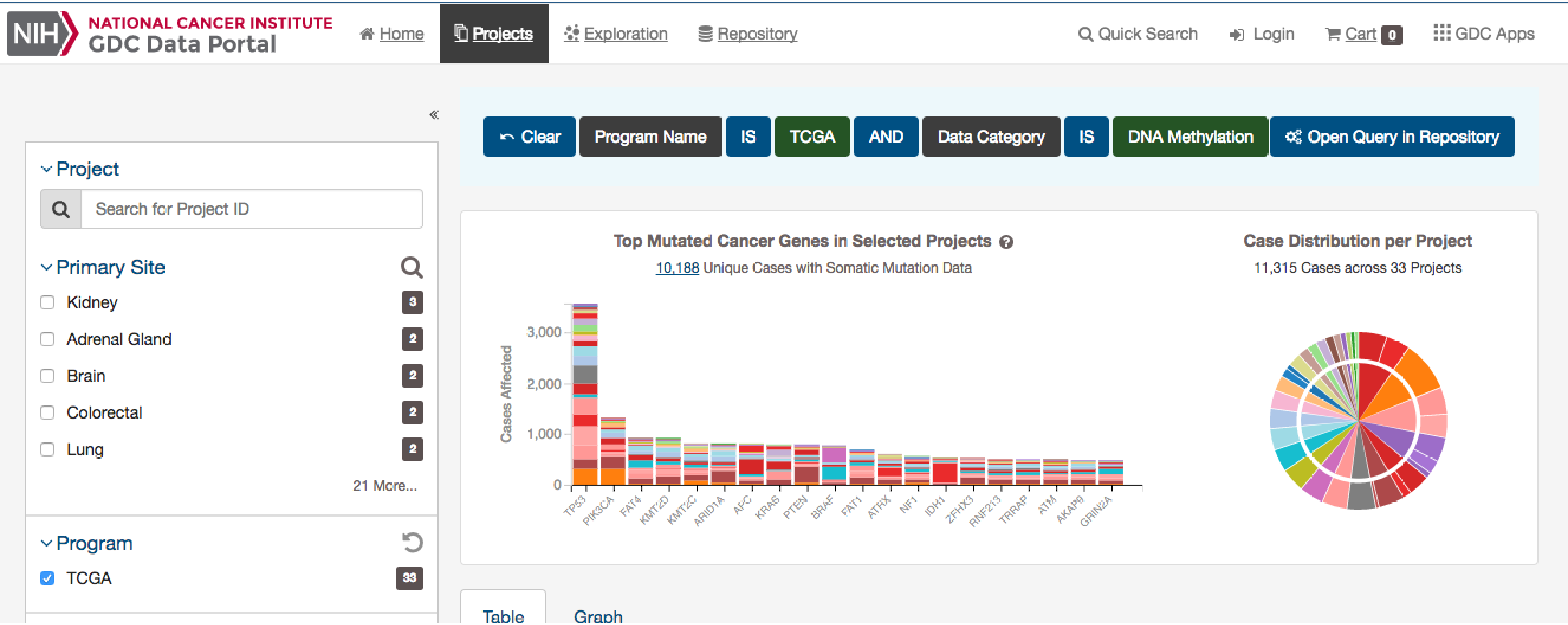
NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Total Files: 1,156,065
Total Size: 1.58 PB

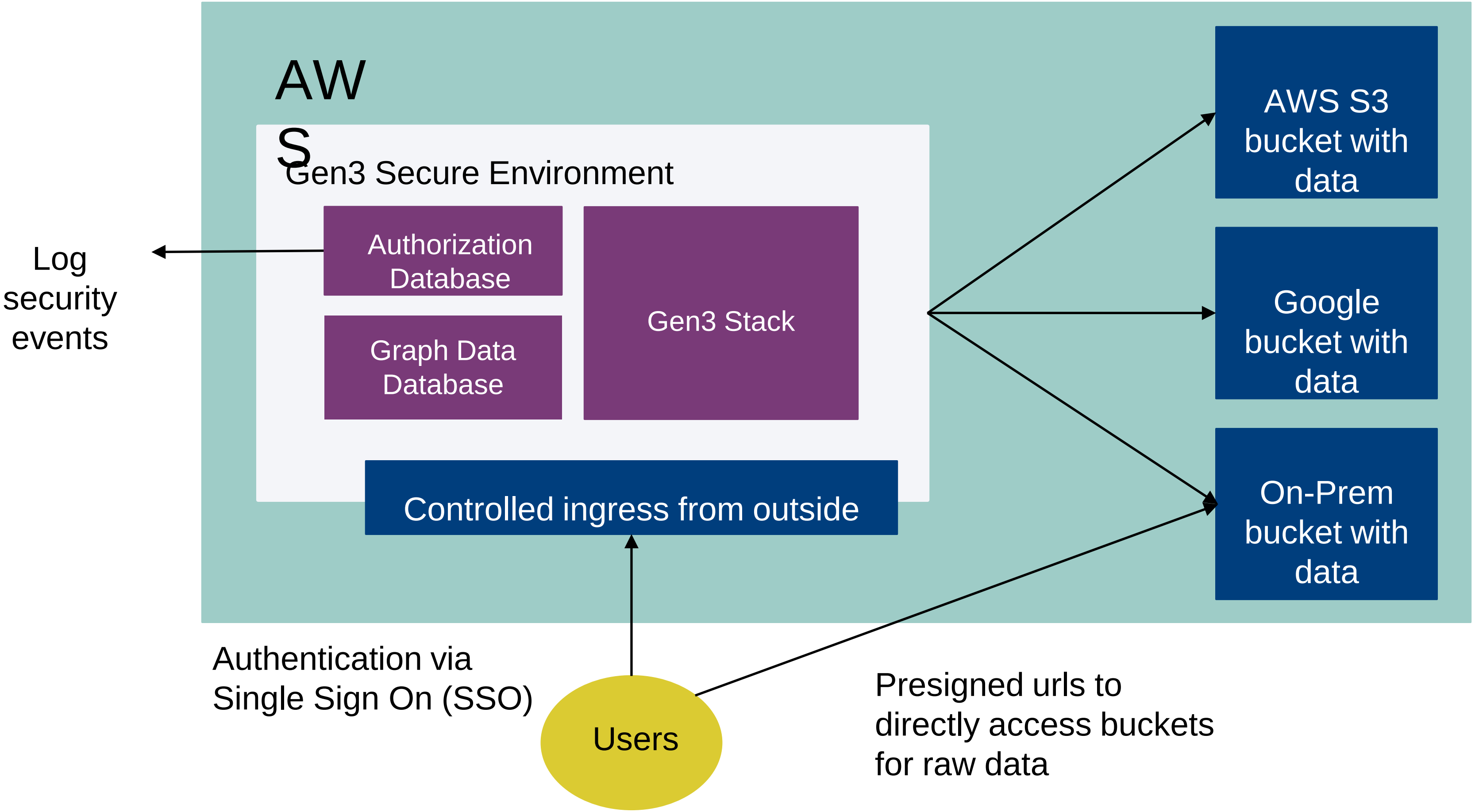
OCC Environmental Data Commons

Total Files: 9,402,905
Total Size: 22.25 TB

Genomic Data Commons - data exploration



GEN3 Data Commons



Authentication via Single Sign On (SSO)

Presigned urls to directly access buckets for raw data

Data Access Control

Cloud Bucket With Data

- Bucket policy prevents access by unauthorized users
- Data access is logged for auditing and compliance

Gen3 Auth

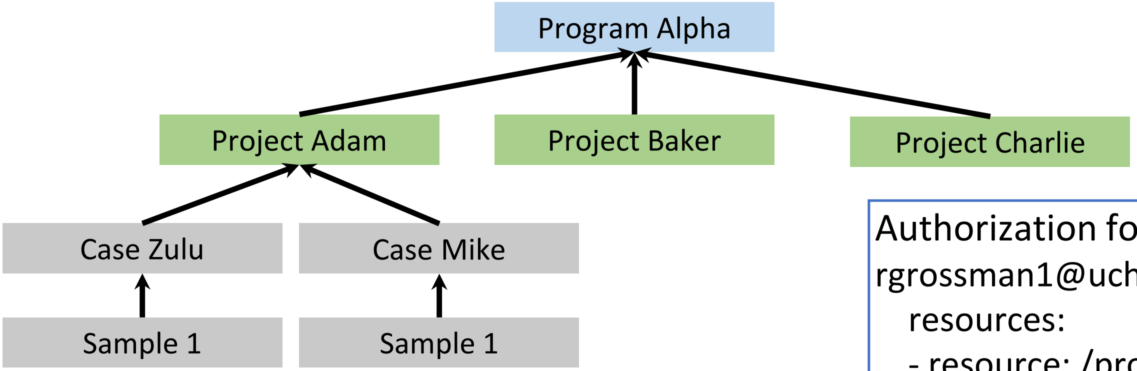
- Gen3 Auth(Fence) provides Authentication and Authorization, and Data Access.
- Gen3 Auth works with multiple identify providers (IdP) including Google, and easily adaptable for any support OIDC provider
- This enables Single Sign On (SSO) compatibility with most systems
- Authorization for data access via internal Access Control List specified by the stakeholders

Data Access Control

- Gen3 auth has a Role Based Access Control (RBAC) engine

Gen3 Auth

The RBAC engine understands the hierarchical nature of a users permissions, and can be used to determine if the user has access to a specific piece of data



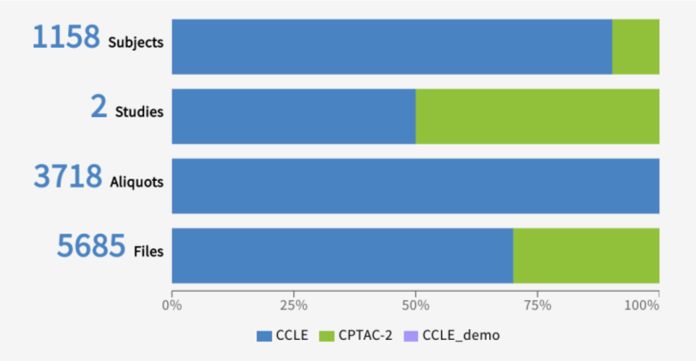
Authorization for a user would then be stored as:

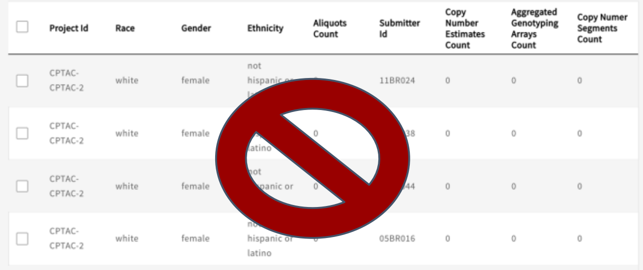
```
rgrossman1@uchicago.edu:
   resources:
   - resource: /programs/alpha/projects/baker
     privilege: [create, read, read-storage, write-storage]
   - resource: /programs/alpha/projects/adam/cases/zulu
     privilege: [read, read-storage]
```

Giving write (submission) access to the Baker project and all nodes underneath it, while read access to only the Zulu case in the Adam project

Data Access Control

Query Gateway





□	Project Id	Race	Gender	Ethnicity	Aliquots Count	Submitter Id	Copy Number Estimates Count	Aggregated Genotyping Arrays Count	Copy Numer Segments Count
□	CPTAC-CPTAC-2	white	female	not hispanic or latino	0	11BR024	0	0	0
□	CPTAC-CPTAC-2	white	female	[illegible] latino	0	[illegible]38	0	0	0
□	CPTAC-CPTAC-2	white	female	not hispanic or [illegible]	0	[illegible]44	0	0	0
□	CPTAC-CPTAC-2	white	female	not hispanic or latino	0	05BR016	0	0	0

- Query gateway provides the potential to limit the queries that users can perform and control when results are returned.

Examples of queries:
Query1: StandardDeviation(variable) where STUDENTS_GENDER is MALE

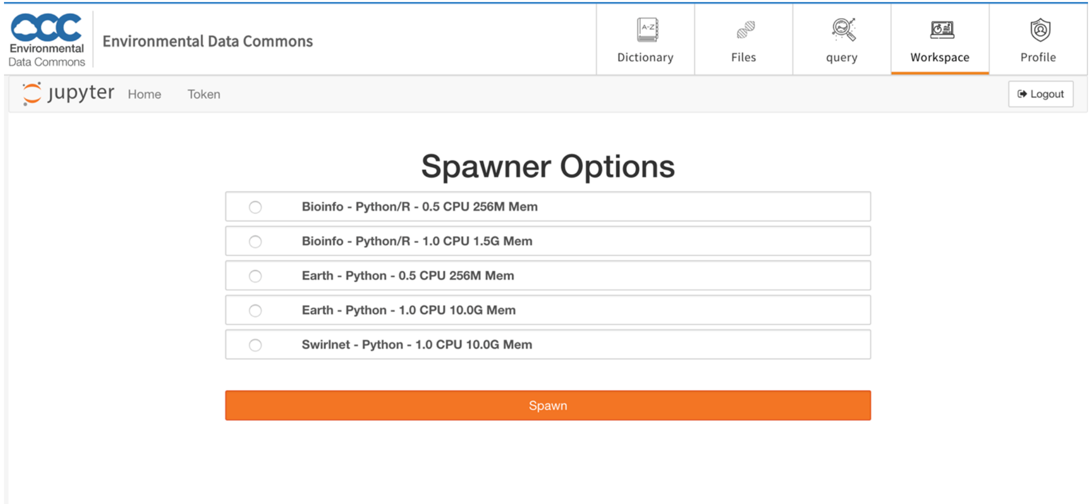
Blue = querying user can specify

Results returned only when # of students represented in the query > a threshold. *I.e. only return standard deviations when the query is computing it for at least 10 students.*

Jupyter Notebooks

Jupyter

- Jupyter Notebooks are powerful tools for creating custom analysis over datasets
- Gen3 runs Jupyter Notebooks in a secure cloud environment helping to reduce the need to download data to laptops, etc.



Spawner Options

○ Bioinfo - Python/R - 0.5 CPU 256M Mem

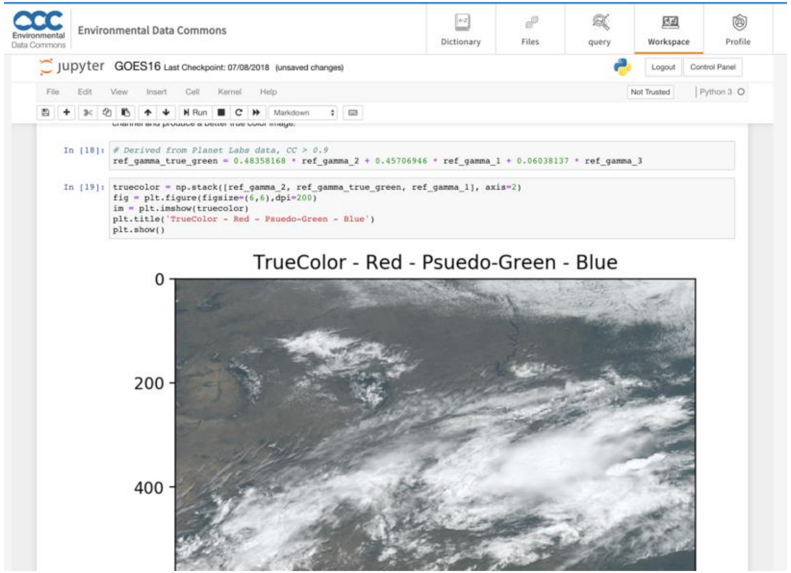
○ Bioinfo - Python/R - 1.0 CPU 1.5G Mem

○ Earth - Python - 0.5 CPU 256M Mem

○ Earth - Python - 1.0 CPU 10.0G Mem

○ Swirlnet - Python - 1.0 CPU 10.0G Mem

Spawn



GOES16 Last Checkpoint: 07/08/2018 (unsaved changes)

channel and produce a better true color image.

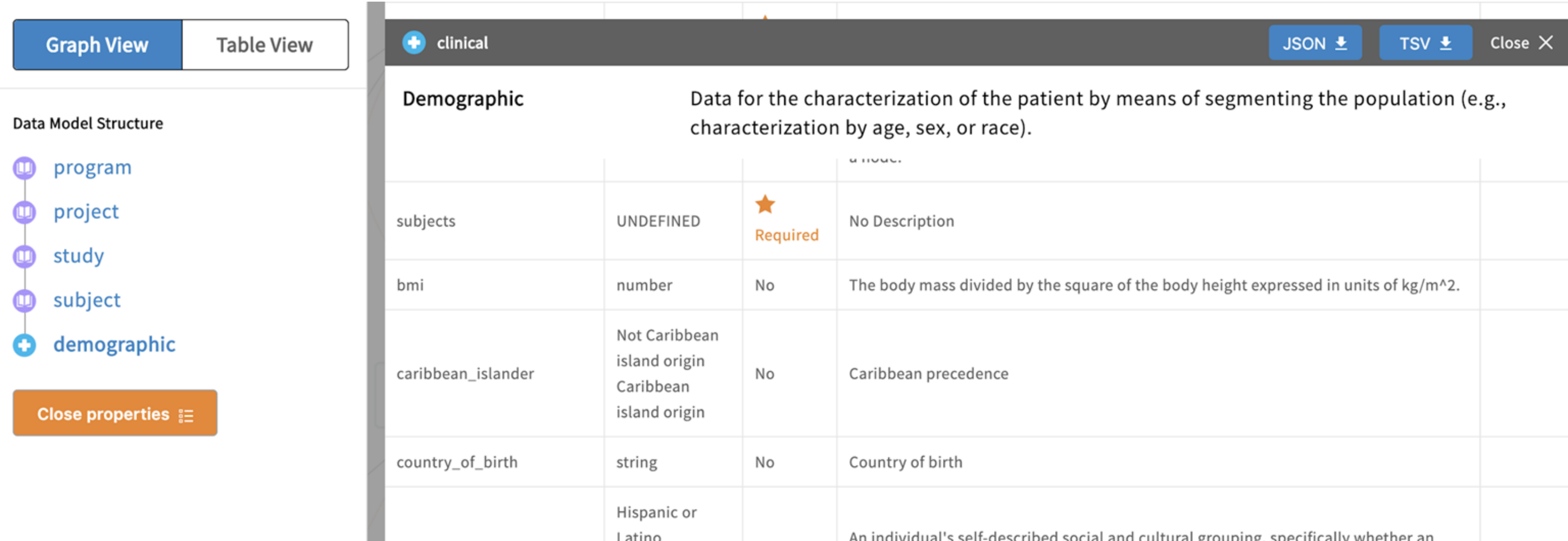
```
In [18]: # Derived from Planet Labs data, CC > 0.9
         ref_gamma_true_green = 0.48358168 * ref_gamma_2 + 0.45706946 * ref_gamma_1 + 0.06038137 * ref_gamma_3

In [19]: truecolor = np.stack([ref_gamma_2, ref_gamma_true_green, ref_gamma_1], axis=2)
         fig = plt.figure(figsize=(6,6),dpi=200)
         im = plt.imshow(truecolor)
         plt.title('TrueColor - Red - Psuedo-Green - Blue')
         plt.show()
```

Data Ontologies

Dictionary viewer

- Gen3 dictionary viewer allows browsing data vocabularies within a particular data commons



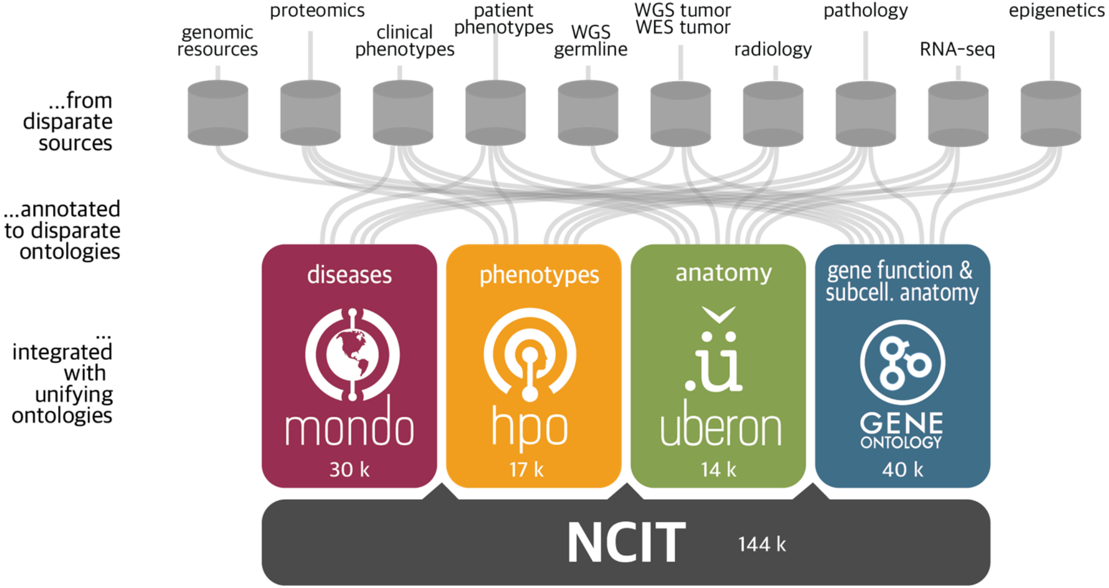
Demographic Data for the characterization of the patient by means of segmenting the population (e.g., characterization by age, sex, or race).

			a node.
subjects	UNDEFINED	★ Required	No Description
bmi	number	No	The body mass divided by the square of the body height expressed in units of kg/m^2.
caribbean_islander	Not Caribbean island origin Caribbean island origin	No	Caribbean precedence
country_of_birth	string	No	Country of birth
	Hispanic or Latino		An individual's self-described social and cultural grouping, specifically whether an

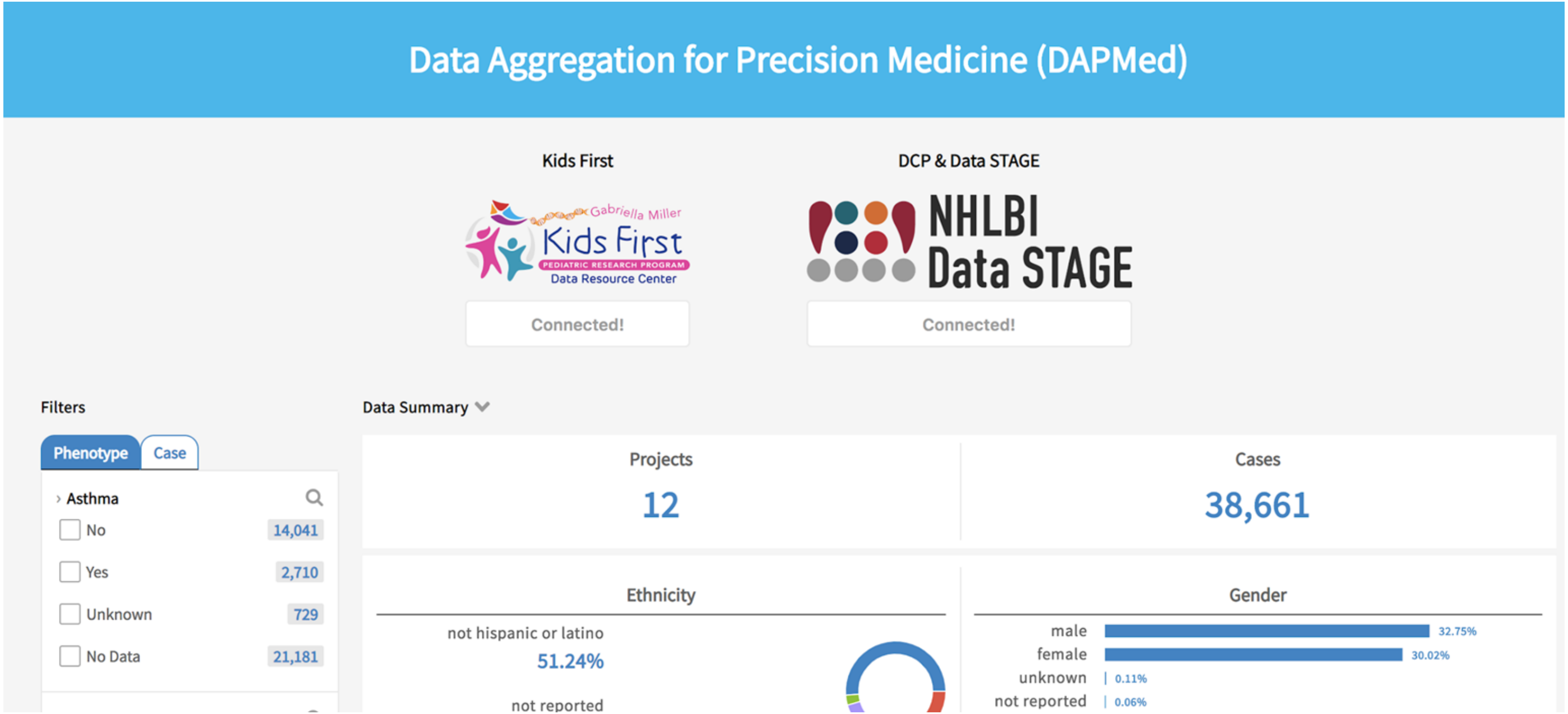
Data Ontologies

PFB

- Ontologies contain controlled vocabulary developed by a standards body.
- Data dictionaries contain references to the ontology terms allowing harmonization of differing data dictionaries



Data Aggregation



Data Aggregation for Precision Medicine (DAPMed)

Kids First

Connected!

DCP & Data STAGE

Connected!

Filters

Phenotype | Case

› Asthma

□ No	14,041
□ Yes	2,710
□ Unknown	729
□ No Data	21,181

Data Summary

Projects	Cases
12	38,661

Ethnicity

not hispanic or latino

51.24%

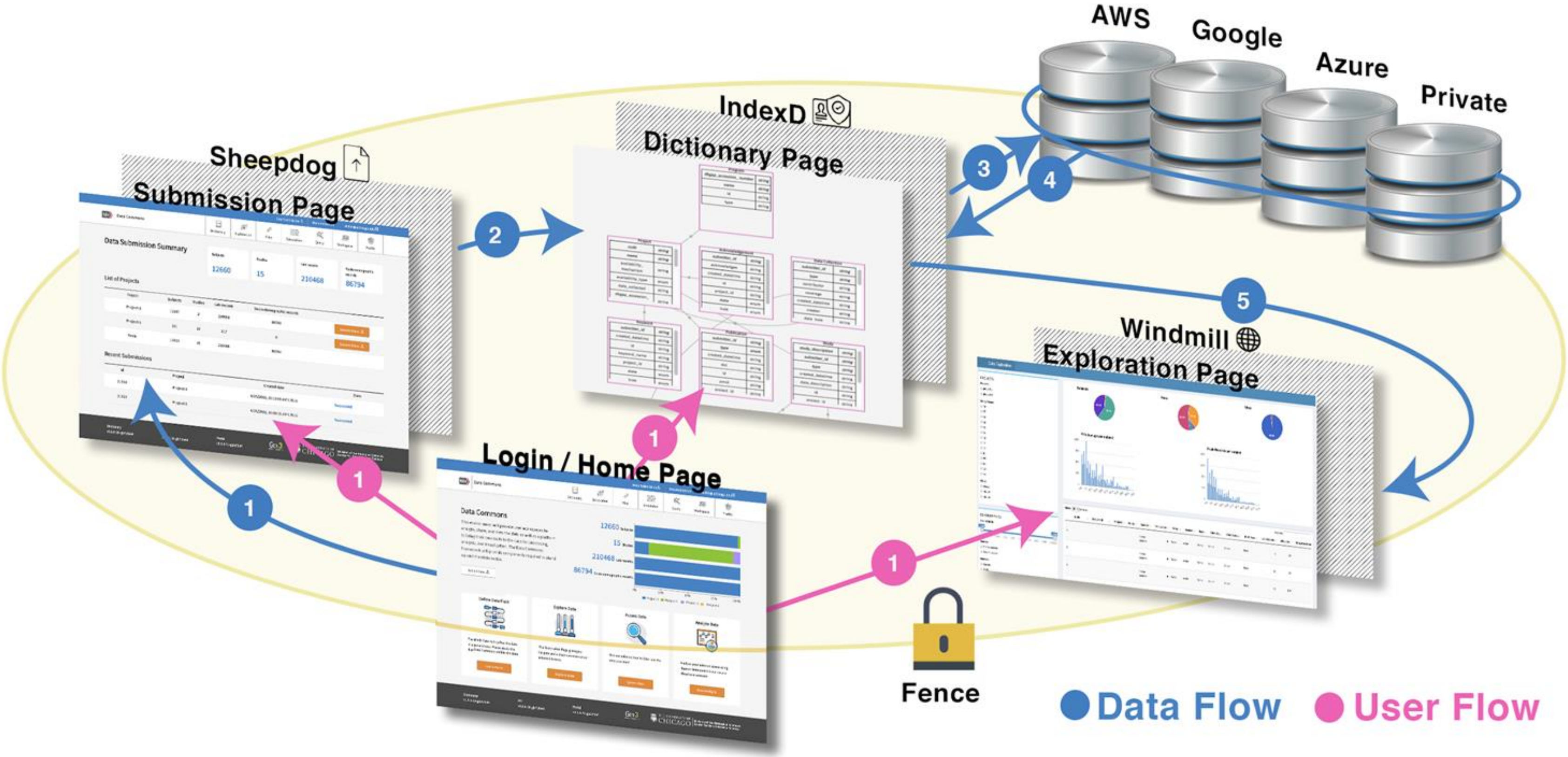
not reported

Gender

male	32.75%
female	30.02%
unknown	0.11%
not reported	0.06%

Data & User Flow with Gen3

GEN3 Data Commons


AWS
Google
Azure
Private
IndexD
Dictionary Page
Sheepdog
Submission Page
Data Submission Summary
12660
15
210468
86794
Windmill
Exploration Page
Login / Home Page
Data Commons
12660
15
210468
86794
Fence
1
1
1
1
2
3
4
5
Data Flow
User Flow