

How to Build a Data Mesh Using Gen3

Center for Translational Data Science (CTDS), University of Chicago

Department of Public Health Sciences, Biostatistics Laboratory, University of Chicago

The Agenda

- A Quick Introduction to Gen3 Data Meshes - Bob Grossman, CTDS, UChicago
- HEAL Data Mesh - Phil Schumm, Department of Public Health Sciences, UChicago
- Biomedical Research Hub (BRH) Data Mesh - Aarti Venkat, CTDS, UChicago
- How to Set Up a Gen3 Data Mesh - Sai Narumanchi, CTDS, UChicago
- Open Discussion
- Discussion on topics for next event

A Quick Introduction to Gen3 Data Meshes

Robert Grossman



THE UNIVERSITY OF
CHICAGO



Center for
Translational
Data Science

GEN3

Open Commons
Consortium

1. Why data meshes (aka data ecosystems)?

Data Meshes



The HEAL Data Platform enables search and discovery across multiple data repositories supporting the hundreds of projects that are part of the Helping to End Addiction Long-term (HEAL) Initiative.

Data Repositories: **9**



The Biomedical Research Hub enables search, discovery and the analysis of data from over 10 data commons from NIH Institutes, Centers and projects.

Data Repositories: **11**

Data Commons



83,709 Subjects
622 Attributes
52,141,509 Files
4.90 PB Total Size

107,418 Subjects
786 Attributes
16,824 Files
10.94 TB Total Size

658,278 Subjects
1,606 Attributes
468 Files
1.40 TB Total Size

2,096 Subjects
151 Attributes
10 Files
3.88 MB Total Size

1,499 Subjects
1,048 Attributes
3,820 Files
1.88 TB Total Size

1,390 Subjects
387 Attributes
6,555 Files
31.60 TB Total Size



237 Subjects
517 Attributes
391 Files
1.30 GB Total Size

265 Attributes
33,441,289 Files
99.20 TB Total Size

4,839 Subjects
888 Attributes
35,549 Files
34.57 TB Total Size

53,728 Subjects
1,464 Attributes
285,849 Files
117.64 TB Total Size

41,933 Subjects
551 Attributes
200,397 Files
803.96 TB Total Size

438,874 Subjects
770 Attributes
712,329 Files
3.88 PB Total Size

End to End Design Principle

What are the fewest number of services that can support the interoperability of data commons?

These are the data mesh (aka framework services).

These include Fence, Indexd, Gen3 Metadata Service, etc.

End-To-End Arguments in System Design

J. H. SALTZER, D. P. REED, and D. D. CLARK

Massachusetts Institute of Technology Laboratory for Computer Science

This paper presents a design principle that helps guide placement of functions among the modules of a distributed computer system. The principle, called the end-to-end argument, suggests that functions placed at low levels of a system may be redundant or of little value when compared with the cost of providing them at that low level. Examples discussed in the paper include bit-error recovery, security using encryption, duplicate message suppression, recovery from system crashes, and delivery acknowledgment. Low-level mechanisms to support these functions are justified only as performance enhancements.

CR Categories and Subject Descriptors: C.0 [General] Computer System Organization—*system architectures*; C.2.2 [Computer-Communication Networks]: Network Protocols—*protocol architecture*; C.2.4 [Computer-Communication Networks]: Distributed Systems; D.4.7 [Operating Systems]: Organization and Design—*distributed systems*

General Terms: Design

Additional Key Words and Phrases: Data communication, protocol design, design principles

1. INTRODUCTION

Choosing the proper boundaries between functions is perhaps the primary activity of the computer system designer. Design principles that provide guidance in this choice of function placement are among the most important tools of a system designer. This paper discusses one class of function placement argument that

2. Data Commons vs Data Meshes

Data Commons vs Data Ecosystems



Data Commons

Question: Given data from multiple projects within a particular discipline or domain, how do you curate and harmonize the data and make it available to the research community?

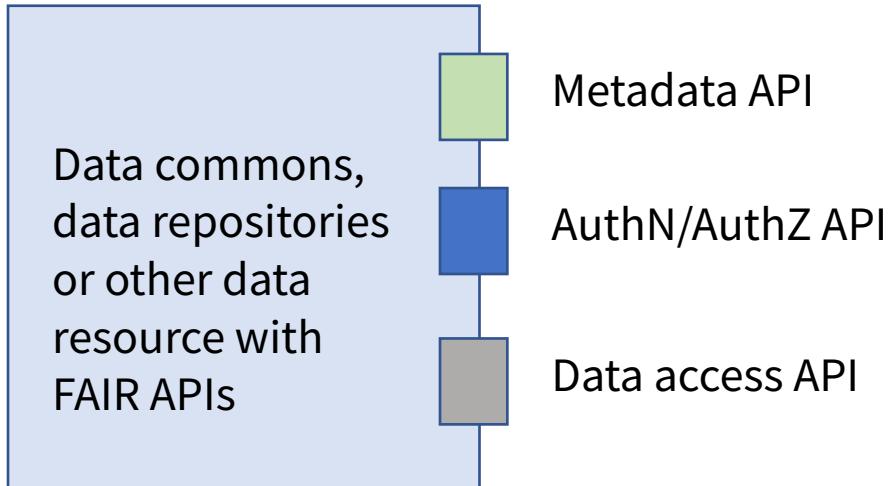
Assumptions: there is a common data model and data is curated and harmonized

Data Ecosystems

Question: Given multiple data repositories and data commons, how do you search for relevant data and bring it into a workspace to explore and analyze it?

Assumptions: there are multiple data models, but standard APIs for authN/authZ and data access; data is (generally) accessed at the dataset level or the data object level

Connect to a Gen3 Data Mesh

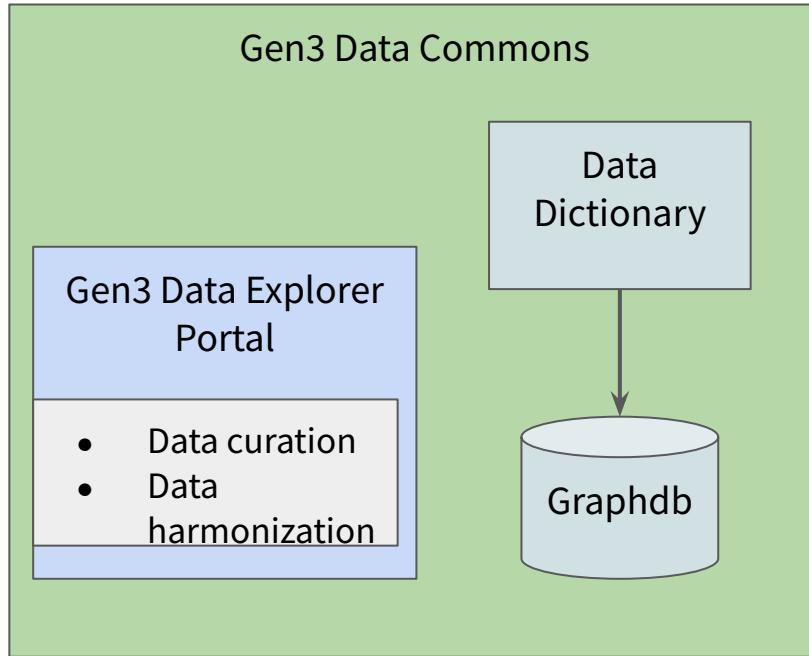


To be part of the data ecosystem a data commons or data resource must expose three APIs:

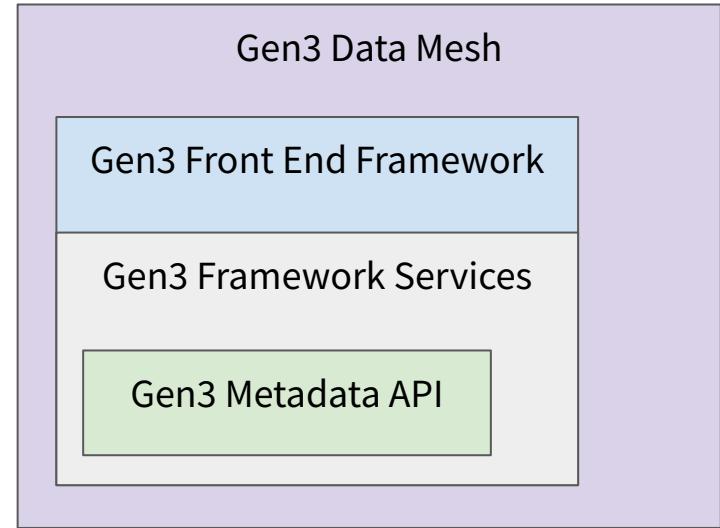
1. AuthN/AuthZ API
2. (FAIR) Metadata API
3. (FAIR) Data API

Gen3 data meshes discovery portals enable interactive data discovery and data exploration over all data commons, data repositories, and other cloud-based resources that expose FAIR APIs.

Data Commons Portal vs Data Ecosystem Browser



1. Set up a data model
2. Harmonize data at the subject level
3. Ingest and curate data
4. Build a front end



1. Leverage dataset metadata
2. Use framework services to assign digital IDs (GUIDs) and metadata
3. Use front end framework to build ecosystem browser.
4. Select datasets export to workspace

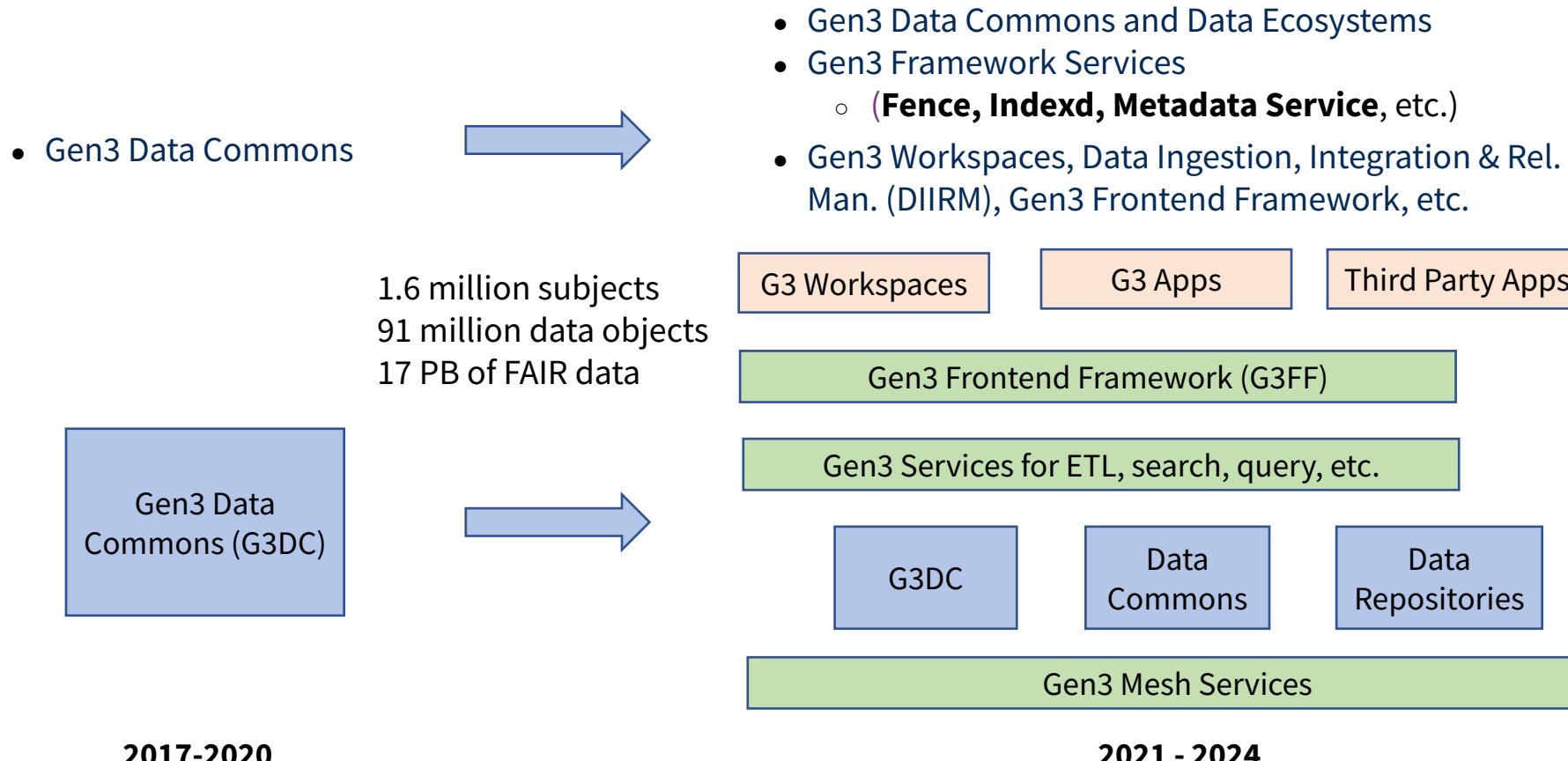
Gen3 Data Commons vs Data Meshes



- **Data commons** software platforms that co-locate: 1) curated data, 2) cloud-based computing infrastructure, and 3) commonly used software applications, tools and services to create a governed resource for managing, analyzing and sharing data with a research community.
- **Data meshes** (aka data ecosystems) integrate multiple data commons, computational platforms, and other cloud-based resources operated by different organizations, along with a hybrid governance framework, and enable the management, discovery, analysis and sharing of data.
- **Data Mesh Services** (aka Data Commons Framework Services) are a set of services to develop and operate data commons and data meshes.
- **Gen3** is an open-source data platform for building data commons and data meshes over a set of data mesh services.

Source: - Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234. arxiv.org/abs/1809.01699 PMID: 30691868 PMCID: PMC6474403
- Craig Barnes, Binam Bajracharya, ..., and Robert L. Grossman, The Biomedical Research Hub: A Federated Platform for Patient Research Data, Journal of the American Medical Informatics Association, 2021, doi:10.1093/jamia/ocab247.

Gen3 Data Platform (2023 Version)

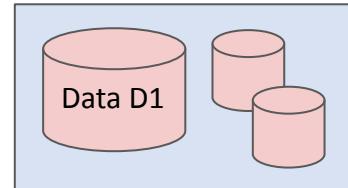


2017-2020

2021 - 2024

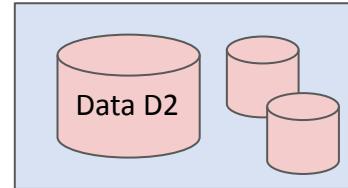
3. Security, Compliance & Governance, for Gen3 Data Meshes

Gen3 Security & Compliance Follows NIST SP 800-53 Moderate



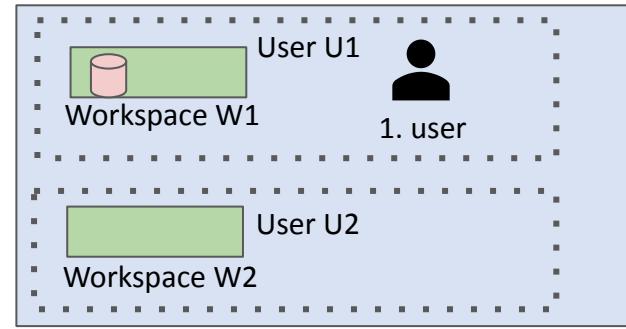
- Metadata API
- AuthN/AuthZ API
- Data access API

Data Commons A



- Metadata API
- AuthN/AuthZ API
- Data access API

Data Commons B



Cloud Platform C (authorized environment)



Cloud Platform D

Gen3 Data Mesh (aka Framework) Services



Cloud platform
boundary



Security and
compliance boundary



Workspace
for user



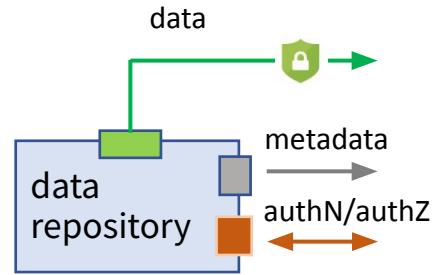
Cloud platform

Hybrid Governance Model

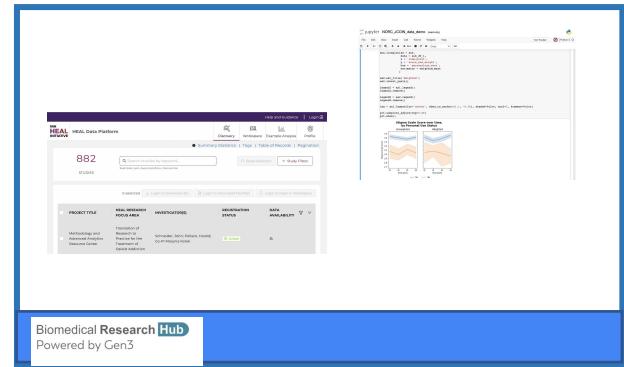
Data Commons Governance



Shared Governance (between commons & mesh platform)



Data Mesh Governance (mesh platform)



- DUA agreements between data submitters & repositories
- Required metadata, CDE, etc.
- Any data curation, etc.

- Data Mesh Services
- FAIR APIs
- Interoperating AuthN/AuthZ
- System “Interoperability Agreement”

- Which data repositories to connect to
- Governance rules for authorizing workspaces

Source: Craig Barnes, Binam Bajracharya, ..., and Robert L. Grossman, The Biomedical Research Hub: A Federated Platform for Patient Research Data, Journal of the American Medical Informatics Association, 2021, doi:10.1093/jamia/ocab247.

4. Examples of Gen3 Data Meshes

The Biomedical Research Hub (BRH)



The screenshot shows the BRH Discovery Portal interface. At the top, it displays '345 STUDIES' and '583,526 TOTAL SUBJECTS'. Below this is a 'TAGS BY CATEGORY' sidebar with options like 'Genomics', 'Proteome', 'Proteome Data Commons', 'Behavioral Data Commons', 'Cancer Imaging Data Commons', 'Rare Disease', 'Atlas', and 'Show 2 more tags'. The main content area features a search bar ('Search studies by keyword...') and an 'ADVANCED SEARCH >' link. A table lists two datasets: 'Opioid Environment Policy Scan (OEPS)' and 'Amerispeak Brief Stigma Survey (JCDSN 020)'. Each dataset row includes columns for 'STUDY NAME', 'FULL NAME', 'NUMBER OF SUBJECTS', 'ID NUMBER', 'DATA COMMONS', 'TAGS', and 'DATA AVAILABILITY'. The 'OEPS' row shows '1,022,000' subjects, ID '1U2CD00350098-02_b', and 'JCDSN' as the data common. The 'Amerispeak' row shows '1,000' subjects, ID '1U2CD00350098-02_a', and 'JCDSN' as the data common. Below the table, a note states: 'The Opioid Environment Policy Scan provides access to data at multiple spatial scales to help characterize the multi-dimensional risk environment impacting opioid use in justice populations across the United States.' Another note below the second dataset says: 'The Amerispeak Brief Stigma Survey provides access to data at multiple spatial scales to help characterize the multi-dimensional risk environment impacting opioid use in justice populations across the United States.'

BRH Discovery Portal. Each data commons or data resource in the BRH data mesh exposes metadata about its datasets through FAIR APIs. The Gen3 Discovery Portal uses the metadata to power search. Data can then be explored and analyzed in workspaces. BRH is a joint project between the Center for Translational Data Science at the University of Chicago, OCC and AWS.

- The Biomedical Research Hub (BRH) is a data platform operated by the Center for Translational Data Science (CTDS) at the University of Chicago at a FISMA Moderate security & compliance level with an ATO from NIH.
- The BRH is part of the NIH STRIDES program.
- Projects can set up their own data commons within the BRH and use the BRH for their data mesh services and their security and compliance services.
- Researchers can use the BRH Discovery Portal to find datasets of interest and use secure and compliant workspaces to access and analyze their data.
- BRH uses a hybrid governance model between the projects that operate the data commons and CTDS that operates the mesh services.

The HEAL Data Platform is a Gen3 Mesh for the NIH HEAL Initiative

GEN3
DATA COMMONS

- FAIR API for metadata
 - FAIR API for data (e.g. GA4GH DRS)
 - NIH RAS, GA4GH Visas and Passpo

**FAIR = Findable,
Accessible,
Interoperable &
Reusable**



HEAL Data generators and providers

Multiple data repositories

HFAI Platform

GEN3

HEAL uses Gen3
for the Platform
and the FAIR
data services

Notebooks in secure workspaces supporting interactive data analysis

Researcher

HEAL Data Platform

Phil Schumm

HEAL Data Mesh

Phil Schumm, Department of Public Health Sciences, University of Chicago (with Bob Grossman and the HEAL Platform Team at CTDS)

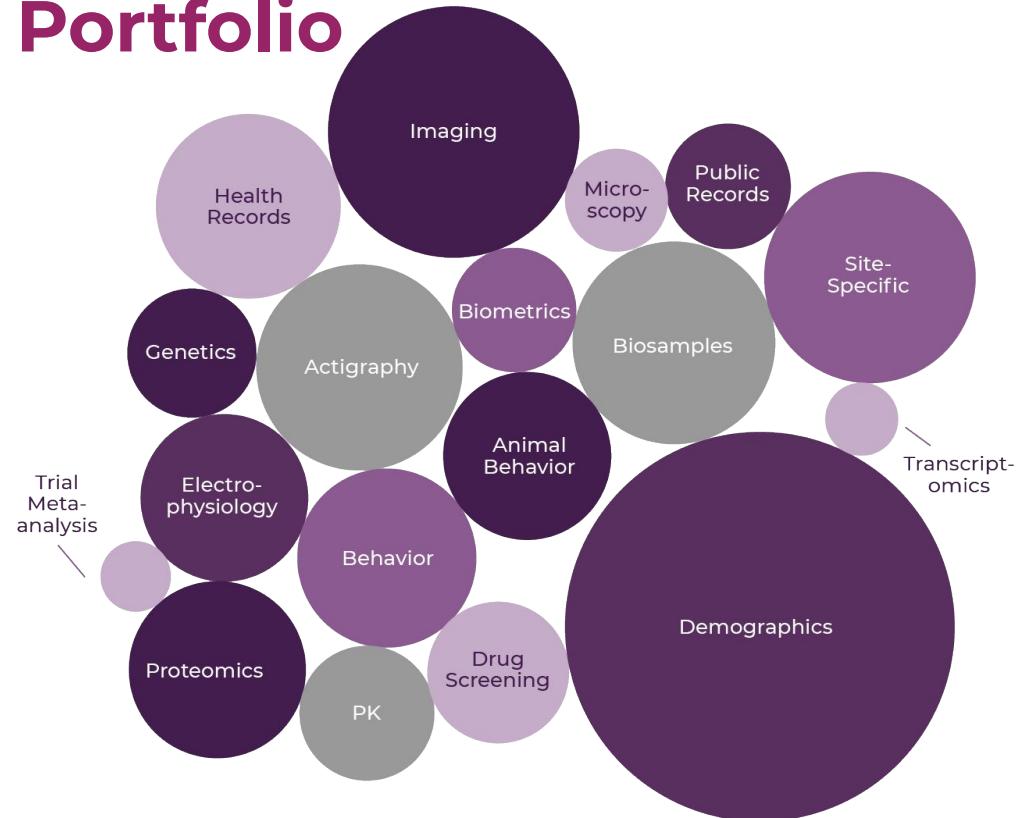
NIH's Helping to End Addiction Long-term (HEAL) Initiative

- Trans-agency effort to address the U.S. national opioid crisis
- 800+ NIH-funded studies within two broad areas:
 - Improving Prevention and Treatment for Opioid Misuse and Addiction
 - Enhancing Pain Management
- Strong data sharing mandate

HEAL Studies Comprise a Remarkably Diverse Research Portfolio

~800
Studies

20+ data
types

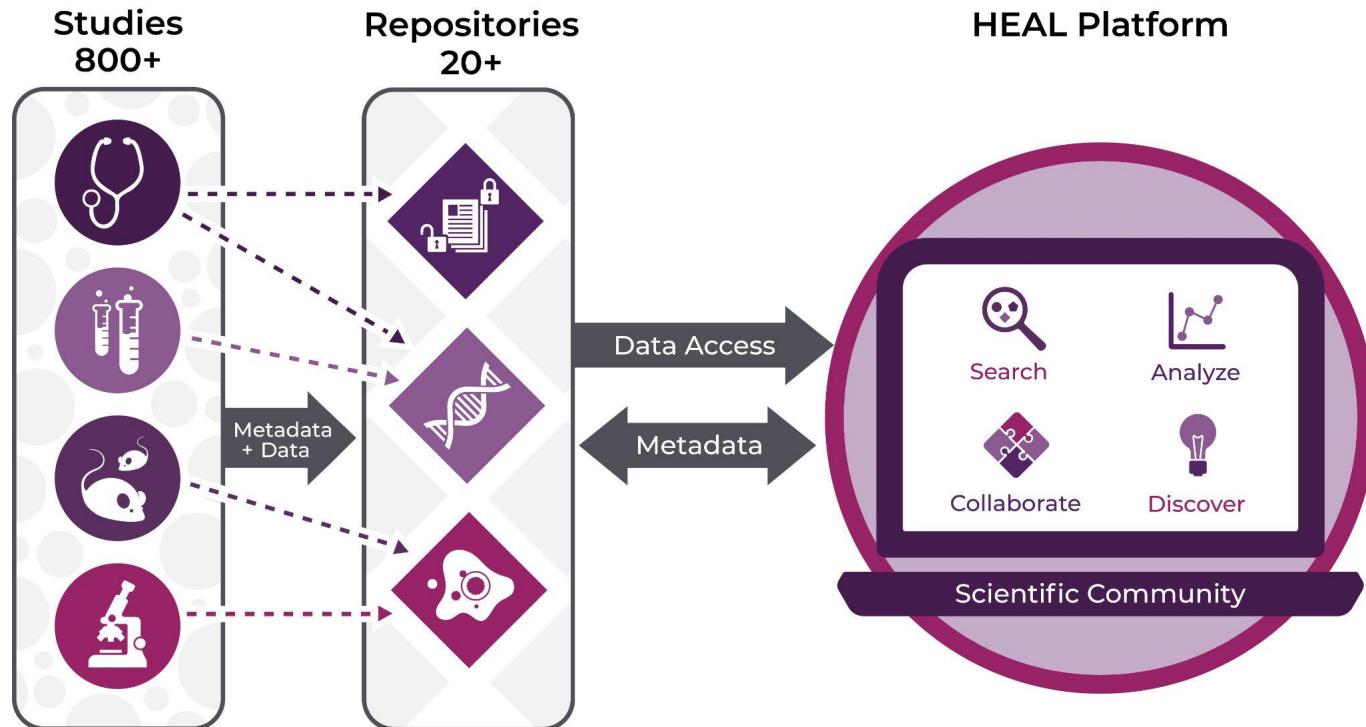


*This figure represents a small sampling of the 800 HEAL-funded studies

Rationale for a Data Mesh

- Extremely broad range of disciplines, study designs, and data types favor specialized repositories
- NIH's desire to reuse existing repository resources (including procedures for requesting and granting data access)
- Need for rapid sharing and collaboration, including across disciplines, to speed scientific discovery

HEAL Data Ecosystem



HEAL Data Platform



Discovery Workspace Example Analysis Profile

HEAL Data Platform

The **HEAL Data Platform** is a single web interface which allows visitors to **discover, access and analyze data** generated by HEAL funded, as well as HEAL relevant, studies within a distributed ecosystem. Making HEAL data easily findable enables secondary, cross-study analyses, promotes dissemination of the HEAL Initiative's research and accelerates new discoveries.

[EXPLORE DATA](#)

[REGISTER YOUR STUDY](#)

The Helping to End Addiction Long-term Initiative, or **NIH HEAL Initiative®**, is an aggressive, trans-agency effort to speed scientific solutions to stem the national opioid and pain public health crises.

[LEARN MORE](#)





Search HEAL studies and related datasets for download or analysis in a workspace.

[DISCOVER](#)



Explore Tutorials and Example Analysis.

[ANALYZE](#)



View answers to frequently asked questions.

[FAQS](#)



Watch tutorial videos to learn how to interact with the HEAL Platform.

[TUTORIALS](#)



Explore helpful resources for Prevention, Treatment and Support related to Opioid Use Disorder.

[RESOURCES](#)

HEAL Data Platform

- Metadata for 800+ HEAL-funded studies ingested from NIH RePORTER
- 160+ HEAL investigators have registered their studies
- Aggregates additional study-level metadata from ClinicalTrials.gov and individual data repositories
- Investigators enter and update study-level and variable-level metadata for their own studies
- Investigators can request access to cloud-based workspaces
- Working on establishing interoperability with HEAL-recommended repositories

Pre-collection of Metadata

- **Study-level** metadata include information about your objectives and execution
 - Provided via a structured form in CEDAR
- **Variable-level** metadata (sometimes referred to as data dictionaries) include information about each of the variables you are collecting
 - Extracted automatically from datasets using tools we provide
 - May be further annotated to include descriptions and **linkages to Common Data Elements (CDEs), vocabularies and ontologies**

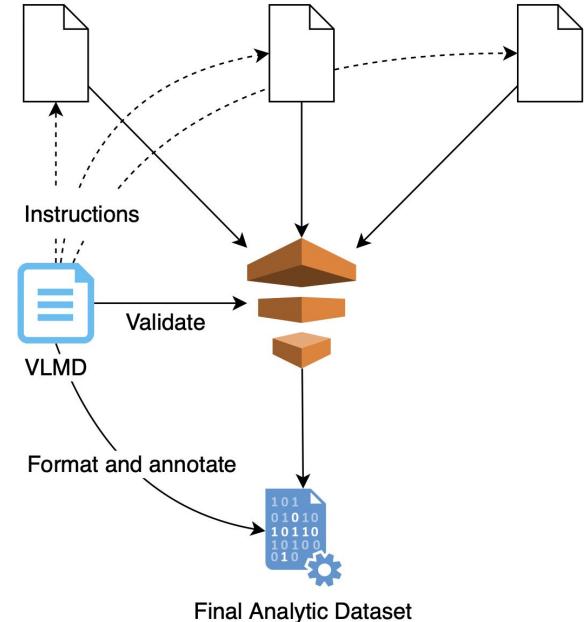
Category or Type/Stage of Study Research
Is the study conducting primary or secondary research?
Is the study conducting observational or experimental research?

	A	B	C
1	name	description	type
2	participant_id	Unique identifier for participant	string
3	race	Self-reported race	integer
4	age	What is your age? (age at enrollment)	integer
5	hispanic	Are you of Hispanic, Latino, or Spanish origin?	boolean
6	sex_at_birth	Sex of the participant at birth	string
7			
8			

Metadata Can be Used to Plan and Facilitate Meta-analyses and Collaborative Research

- Identify candidate studies for inclusion based on characteristics of study and the data collected
- Variable-level metadata provide a formal, standardized way to communicate data requirements among collaborators and help data harmonization
- Variable-level metadata can be used automatically to validate data submissions and format data for analysis

Each of these can be done **prior to submission of data to a repository.**



Heterogeneity Across Repositories

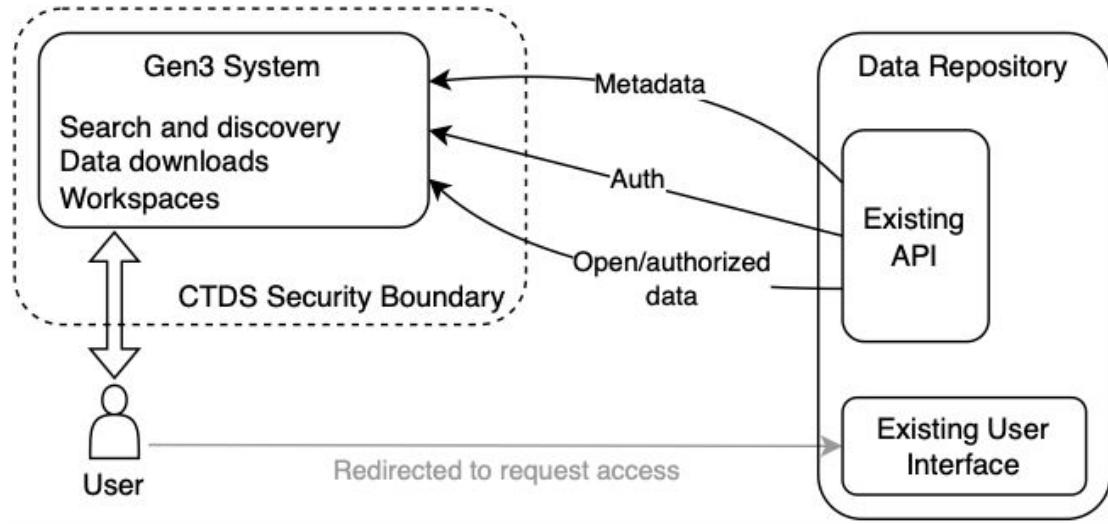
- Different disciplines and cultures
- Different degrees of openness and mechanisms for requesting and approving access
- Different API capabilities
- Exposure to interoperability

Gen3's microservice architecture, openness and adherence to standards provides the flexibility needed.

Example 1. Data repository or commons with fully functional API

Ideal for:

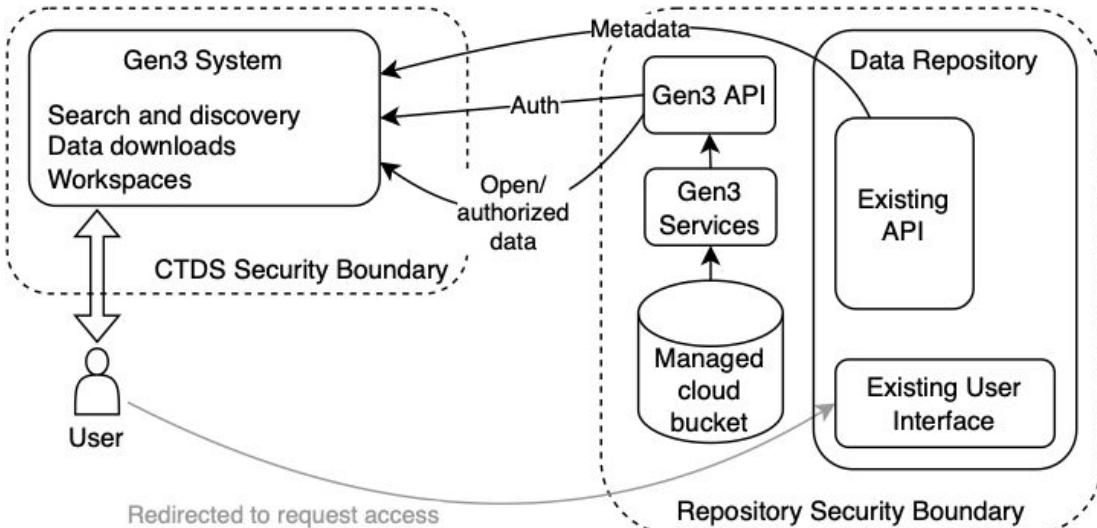
- Repositories for which all metadata and data are openly accessible and which have at least minimal APIs permitting access to metadata and data
- Repositories which contain restricted access data but which have an API permitting secure authentication and authorization



Example 2. Repositories able to manage Gen3 FAIR services-enabled cloud bucket

Ideal for:

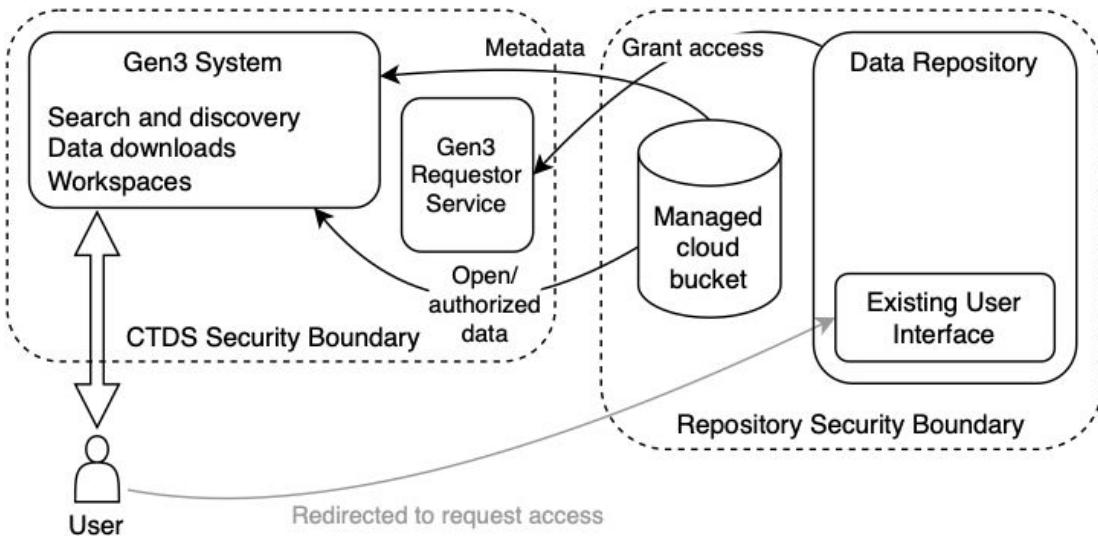
- Repositories that currently have only a partially-functional API (e.g., API for metadata only)
- Repositories that do not currently have an API but are planning to develop one
- Repositories that have no current plans to develop an API but are interested in trying out one or more Gen3 components



Example 3. Repository does not run additional services

Ideal for:

- Repositories that do not have a current API nor plans to develop one, and do not have the resources to manage additional services



Biomedical Research Hub

Aarti Venkat

Biomedical Research **Hub**
Powered by Gen3

How to search and discover effectively across multiple data commons?



stats.gen3.org

GEN3

1,458,879 Total Subjects 22,937,136 Total Files 15.57 PB Total File Size



AccessClinicalData@NIAID

2,096 Subjects
151 Attributes
10 Files
Total Size 3.88 MB



BLOOD PROFILING ATLAS IN CANCER

4,839 Subjects
888 Attributes
35,549 Files
Total Size 34.57 TB



Veterans Affairs
Data Commons

658,278 Subjects
1,606 Attributes
224 Files
Total Size 1.21 TB



BDGC

107,418 Subjects
786 Attributes
16,495 Files
Total Size 6.53 TB



BioData CATALYST

240,460 Subjects
770 Attributes
667,328 Files
Total Size 3.74 PB



Open Access Data Commons
1,366 Subjects
1,452 Attributes
1,598 Files
Total Size 13.77 TB



the Melanoma Genetics Consortium
1,390 Subjects
387 Attributes
6,555 Files
Total Size 31.6 TB



The AnVIL
26,636 Subjects
551 Attributes
187,134 Files
Total Size 502.27 TB



MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER
21,465 Subjects
590 Attributes
3,421,096 Files
Total Size 6.63 TB



Kids First
CHILDREN'S MILESTONES
DATA RESOURCE CENTER
21,833 Subjects
776 Attributes
786,021 Files
Total Size 6.78 PB



Justice Community Opioid
Innovation Network (JCOIN)



237 Subjects
517 Attributes



1,499 Subjects
1,048 Attributes



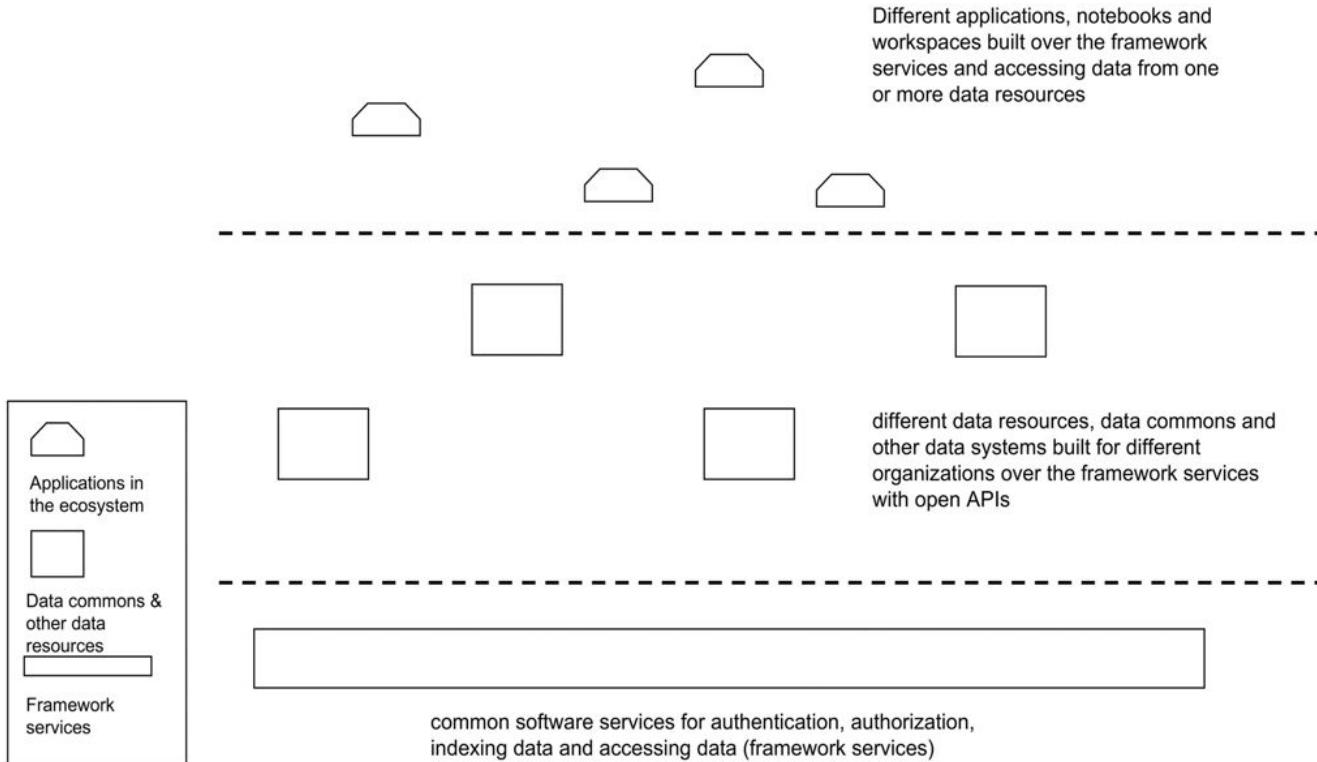
163,695 Subjects
1,606 Attributes



NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

83,709 Subjects
622 Attributes

Architectural overview of BRH



Broad range of study types and resulting data

- Imaging
- Transcriptomics
- Genetics and genomics
- Single cell
- Proteomics
- Clinical trials/phenotypes
- Demographics
- Public records

brh.data-commons.org

Biomedical Research Hub
Powered by Gen3

516 STUDIES 604,204 TOTAL SUBJECTS

Search studies by keyword...

Data Commons

BioData Catalyst

CRDC Cancer Imaging Data Commons

CRDC Genomic Data Commons

CRDC Integrated Canine Data Commons

CRDC Proteomic Data Commons

IBD Commons

JCOIN

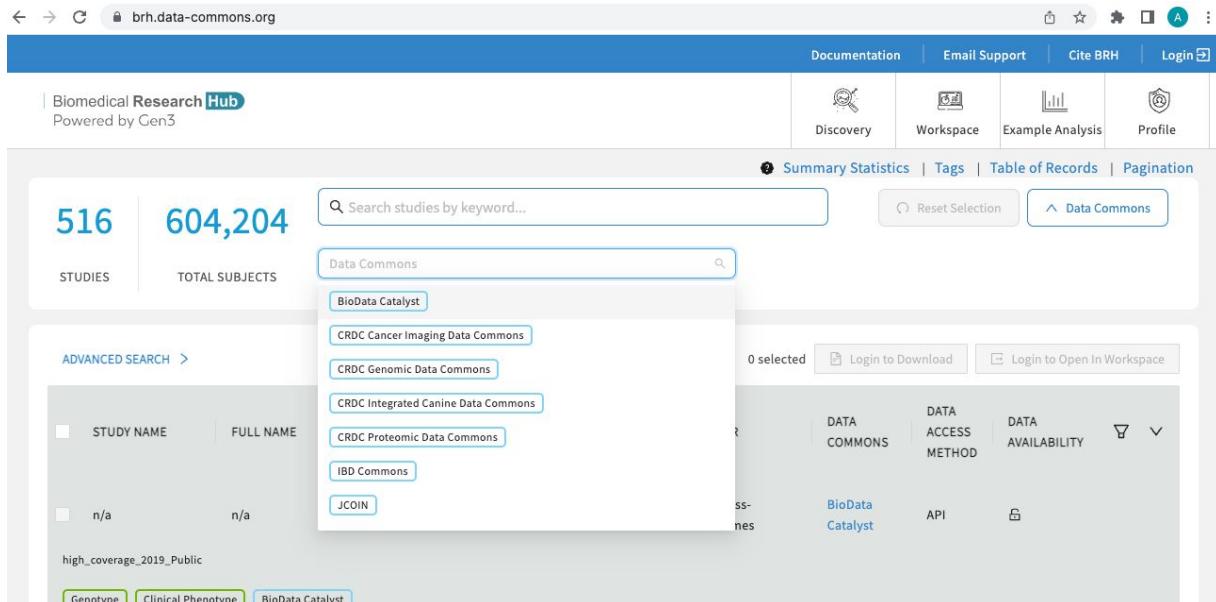
0 selected

Login to Download Login to Open In Workspace

STUDY NAME	FULL NAME	DATA COMMONS	DATA ACCESS METHOD	DATA AVAILABILITY
n/a	n/a	BioData Catalyst	API	

high_coverage_2019_Public

Genotype Clinical Phenotype BioData Catalyst



Special features of BRH: Search, Data access, Data availability

brh.data-commons.org/discovery

Discovery Homepage Example Inquiry Help

Summary Statistics | Tags | Table of Records | Pagination

6 25,211

framingham

Reset Selection Data Commons

STUDIES TOTAL SUBJECTS Data Commons

ADVANCED SEARCH > 0 selected Download Open In Workspace

STUDY NAME	FULL NAME	NUMBER OF SUBJECTS	ID NUMBER	DATA COMMONS	DATA ACCESS METHOD	DATA AVAILABILITY
n/a	Framingham Cohort	13,070	phs000007.v31.p12.c1	BioData Catalyst	API	

"See Grouping of [Framingham Phenotype Datasets](#) Startup of Framingham Heart Study. Cardiovascular disease (CVD) is the leading cause of death and serious illness in the United States. In 1948, the Framingham Heart Study (FHS) -- under the direction of the National Heart Institute (now known as the National Heart, Lung, and Blood Institute, NHLBI) -- embarked on a novel and ambitious project in health research. At the time, little was known about the causes of heart disease and stroke. The study has since become one of the most important sources of information about heart disease and its risk factors. The study has been followed by three generations of participants, and continues to provide valuable insights into the prevention and treatment of cardiovascular diseases." [Read more](#)

Parent DCC Harmonized Clinical Phenotype dbGaP BioData Catalyst

STUDY NAME	FULL NAME	NUMBER OF SUBJECTS	ID NUMBER	DATA COMMONS	DATA ACCESS METHOD	DATA AVAILABILITY
n/a	Framingham Cohort	2,079	phs000007.v31.p12.c2	BioData Catalyst	API	

"See Grouping of [Framingham Phenotype Datasets](#) Startup of Framingham Heart Study. Cardiovascular disease (CVD) is the leading cause of death and serious illness in the United States. In 1948, the Framingham Heart Study (FHS) -- under the direction of the National Heart Institute (now known as the National Heart, Lung, and Blood Institute, NHLBI) -- embarked on a novel and ambitious project in health research. At the time, little was known about the causes of heart disease and stroke. The study has since become one of the most important sources of information about heart disease and its risk factors. The study has been followed by three generations of participants, and continues to provide valuable insights into the prevention and treatment of cardiovascular diseases." [Read more](#)

Special features of BRH: Compute workspaces



brh.data-commons.org/workspace

Documentation Email Support Cite BRH aartiv@uchicago.edu Logout

Biomedical Research Hub
Powered by Gen3

Discovery Workspace Example Analysis Profile

> Account Information

jupyter GDC_TCGA-CHOL_RNA_analysis (read only)

File Edit View Insert Cell Kernel Help

Not Trusted Python 3 (ipykernel) O

Markdown

Gene Expression Analysis of Project TCGA-CHOL

Please note: This notebook uses open access data

Qiong Liu

April 7th, 2022

Cholangiocarcinoma (CCA) is aggressive cancer found in the slender tubes that carry the digestive fluid bile through the liver. The Cancer Genome Atlas (TCGA) program contains abundant molecular profilings of over 20,000 primary cancer and matched normal samples spanning 33 cancer types. In this notebook, we demonstrated how to retrieve RNA expression data of project TCGA-CHOL from [Genomic Data Commons \(GDC\) data portal](#), and perform data

Terminate Workspace Make Fullscreen

Special features of BRH: Track spending limits

Documentation | Email Support | Cite BRH | aartiv@uchicago.edu | Logout

Biomedical Research Hub
Powered by Gen3

Discovery | Workspace (selected) | Example Analysis | Profile

Account Information

Account	Apply for an account
Trial Workspace	0.00
Total Charges (USD)	Spending Limit (USD)
0.00	0.00

jupyter GDC_TCGA-CHOL_RNA_analysis (read only)

File Edit View Insert Cell Kernel Help

Not Trusted | Python 3 (ipykernel)

Gene Expression Analysis of Project TCGA-CHOL

Please note: This notebook uses open access data

Qiong Liu

Special features of BRH: AggMDS

Aggregate metadata service (aggMDS) caches the metadata from two or more metadata sources to provide a unified view of the commons on the discovery page

Aggregate metadata service

Elastic search

Gen3 MDS
adapter for GDC

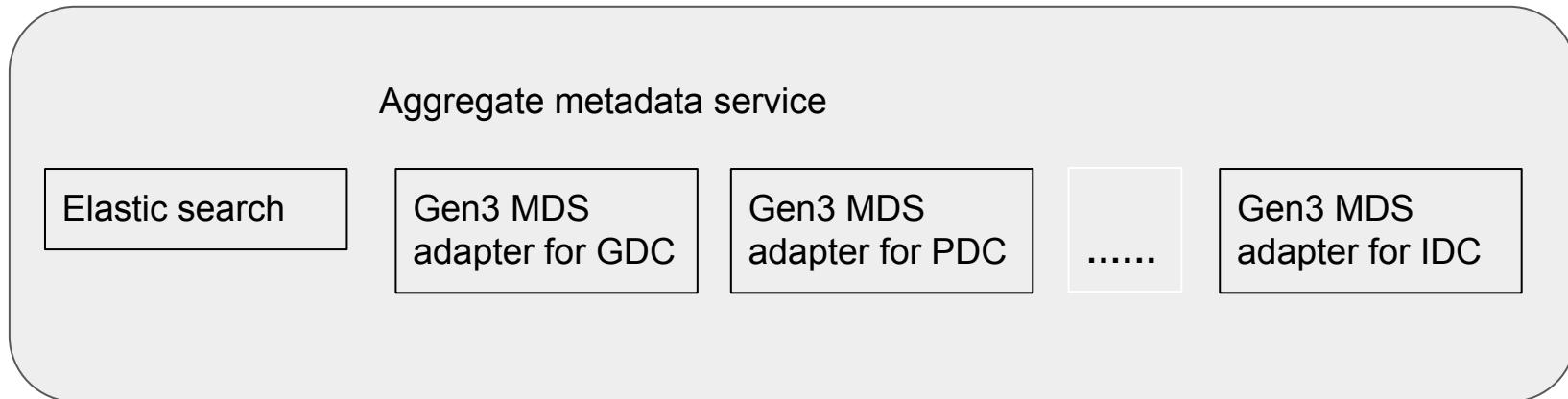
Gen3 MDS
adapter for PDC

.....

Gen3 MDS
adapter for IDC

Special features of BRH: AggMDS

Aggregate metadata service (aggMDS) caches the metadata from two or more metadata sources to provide a unified view of the commons on the discovery page



How to update aggMDS with an additional data commons and power portal?

Special features of BRH: Workspace token service

brh.data-commons.org/discovery

SUBJECTS	COMPLEXITY	METHOD	API	MANIFEST	LOCK
n/a Methodology and Advanced Analytics Resource Center 1,000 1U2CDA050098-01_a JCOIN			API, Manifest		🔒
<p>Abstract: Tracking changes in stigma associated with OUD is important, for as stigma intensifies over time it might make it more difficult to find positive treatment effects in JCOIN. Given the critical importance of public opinion towards policy making and approaches used in the justice system, this project will measure public support for OUD treatment in the general public, assess stigma associated with OUD, and perceptions of...</p> <p>Community Opioid Use Practice Quality Racial and Ethnic Populations Justice-involved Populations Observational Clinical Pain Treatment JCOIN</p>					
n/a Methodology and Advanced Analytics Resource Center n/a 1U2CDA050098-01_b JCOIN			API, Manifest		🔒
<p>The Opioid Environment Policy Scan (OEPS) is a free, open-source data warehouse centered on the multi-dimensional risk environment impacting opioid use and health outcomes, particularly for justice communities, across the United States. The OEPS consolidates cleaned and processed data spanning Medications for Opioid Use Disorder (MOUD) Access, Health, Built Environment, Economic, and other contextual data at the Census...</p> <p>Community Opioid Use Practice Quality Racial and Ethnic Populations Justice-involved Populations JCOIN</p>					
n/a Prescription Drug Abuse Policy System (NIDA) n/a NIDA_PDAPS_1a JCOIN			API, Single File		🔒
<p>PDAPS is funded by the National Institute on Drug Abuse to track key state laws related to prescription drug abuse. Samaritan 911 Immunity, Medical Marijuana, Opioid Related Controls, Prescription Drug Monitoring Program, and...</p> <p>You do not have access to this study.</p> <p>You don't have read access to /programs/parent/projects/CARDIA_HMB-IRB_...</p> <p>Opioid Use JCOIN</p>					
n/a Coronary Artery Risk Development in Young Adults (CARDIA) Study - Cohort 3,111 phs000285.v3.p2.c1 BioData Catalyst			API		🔒
<p>"CARDIA is a study examining the etiology and natural history of cardiovascular disease beginning in young adulthood. In 1985-1986, a cohort of 5115 healthy black and white men and women aged 18-30 years were selected to have approximately the same number of people in subgroups of age (18-24 and 25-30), sex, race, and education (high school or less and more than high school) within each of four US Field Centers. These sam...</p>					

Special features of BRH: Workspace token service

← → ⌂ brh.data-commons.org/identity

Documentation | Email Support | Cite BRH | aartiv@uchicago.edu ⓘ | Logout ☰

Biomedical Research Hub
Powered by Gen3

Discovery Workspace Example Analysis Profile

Link accounts from external data resource(s)

JCOIN Google Login
IDP: jcoin-google
Provider URL: <https://jcoin.datacommons.io> Refresh JCOIN Google Login
Status: expires in 20 days

NCI-CRDC NIH Login
IDP: crdc-nih
Provider URL: <https://nci-crdc.datacommons.io> Authorize NCI-CRDC NIH Login
Status: not authorized

MIDRC NIH Login
IDP: midrc-nih
Provider URL: <https://data.midrc.org> Authorize MIDRC NIH Login
Status: not authorized

How to Set up a Mesh

Sai Narumanchi

Agenda

- **Data Commons and Meshes**
- Key features for Data Meshes
- Aggregate Metadata Sync
- Token Service (Workspace Token service)

Data Mesh (aka “Data Ecosystem”)

- Contain **multiple data commons and/or data repositories** and cloud computing infrastructure
- Use the same framework services as a data commons for the fundamental behavior.
 - Fence – used for AuthN and AuthZ services utilizing OpenID connect flow.
 - Arborist – an Attribute Based Access control policy engine.
 - Windmill (aka Data Portal) – a Front end interactive website.
- Also has a few additional services that allow connecting and interacting with multiple data commons.

E.g., BRH, HEAL

Agenda

- Data Commons and Meshes
- **Key features for Data Meshes**
- Aggregate Metadata Sync
- Token Service (Workspace Token service)

Key features for Data Meshes

- To allow a Data mesh to connect and interact with data from multiple data commons we need to make sure the following are possible
 - a. Fetch metadata from connected commons.
 - b. Access appropriate studies
 - c. View and provide authorization connected commons

Agenda

- Data Commons and Meshes
- Key Services and Jobs for Data Meshes
- **Aggregate Metadata Sync**
- Token Service (Workspace Token service)

Aggregate Metadata Sync



- **Aggregate Metadata Service**
- AggMDS sync job – copies metadata from multiple data commons into a single data store.
- New `/aggregate` endpoint of the Metadata Service.
- JSON based configuration defines information about
 - Data source and Data Adapter information
 - Normalizing data fields
 - Adding optional individual overrides
- Gen3 data adapter and adding new Custom Adapters.

Aggregate Metadata Service

The Metadata Service can be configured to aggregate metadata from multiple other Metadata Service instances.

Aggregation APIs

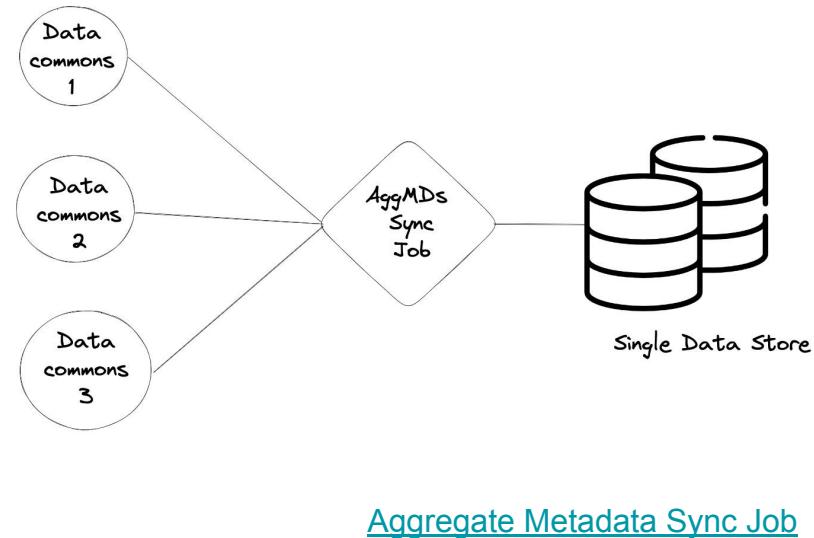
The aggregated MDS APIs and scripts copy metadata from one or many metadata services into a single data store. This enables a metadata service to act as a central API for browsing Metadata using clients such as the Ecosystem browser.

The aggregate metadata APIs and migrations are disabled by default unless `USE_AGG_MDS=true` is specified. The `AGG_MDS_NAMESPACE` should also be defined for shared Elasticsearch environments so that a unique index is used per-instance.

The aggregate cache is built using Elasticsearch. See the `docker-compose.yaml` file (specifically the `aggregate_migration` service) for details regarding how aggregate data is populated.

Being able to fetch metadata from all the connected data commons.

- Aggregate Metadata Service
- **AggMDS sync job – copies metadata from multiple data commons into a single data store.**
- New `/aggregate` endpoint of the Metadata Service.
- JSON based configuration defines information about
 - Data source and Data Adapter information
 - Normalizing data fields
 - Adding optional individual overrides
- Gen3 data adapter and adding new Custom Adapters.



Being able to fetch metadata from all the connected data commons.



- Aggregate Metadata Service
- AggMDS sync job – copies metadata from multiple data commons into a single data store.
- **New `/aggregate` endpoint of the Metadata Service.**
- JSON based configuration defines information about
 - Data source and Data Adapter information
 - Normalizing data fields
 - Adding optional individual overrides
- Gen3 data adapter and adding new Custom Adapters.

Aggregate

GET	/aggregate/commons	Get Commons
GET	/aggregate/info/{what}	Get Commons Info
GET	/aggregate/metadata	Get Aggregate Metadata
GET	/aggregate/metadata/guid/{guid}	Get Aggregate Metadata Guid
GET	/aggregate/metadata/{name}	Get Aggregate Metadata For Commons
GET	/aggregate/metadata/{name}/info	Get Aggregate Metadata Commons Info
GET	/aggregate/tags	Get Aggregate Tags

[`/aggregate` API endpoint](#)

Being able to fetch metadata from all the connected data commons.



- Aggregate Metadata Service
- AggMDS sync job – copies metadata from multiple data commons into a single data store.
- New ` /aggregate` endpoint of the Metadata Service.
- **JSON based configuration defines information about**
 - Data source and Data Adapter information
 - Normalizing data fields
 - Adding optional individual overrides
- Gen3 data adapter and adding new Custom Adapters.

```
"ICPSR": {  
    "mds_url": "https://www.icpsr.umich.edu/icpsrweb/neutral/oai/studies",  
    "commons_url": "https://www.icpsr.umich.edu",  
    "adapter": "icpsr",  
    "filters": {  
        "study_ids": [30122, 37887, 37833, 37842, 37841, 35197]  
    },  
    "field_mappings": {  
        "tags": [],  
        "sites": "",  
        "year": "2020",  
        "shortName": "study_name",  
        "location": "path:coverage[0]",  
        "summary": {  
            "path": "description",  
            "filters": ["strip_html"],  
            "default_value": "N/A"  
        },  
        ...  
    },  
    "per_item_values": {  
        "10.3886/ICPSR30122.v5": {  
            "_manifest": [  
                {  
                    "md5sum": "7cf87ce47b91e3a66332222bc22d098",  
                    "file_name": "example1.zip",  
                    "file_size": 23334,  
                    "object_id": "dg.XXXX/208f4c52-771e-409a-c920-4bcba3c03c51",  
                    "commons_url": "externalaldata.commonns1.org"  
                }  
            ],  
            "data_availability": "available",  
            "authz": "/programs/open",  
            ...  
        }  
    }  
}
```

Being able to fetch metadata from all the connected data commons.



- Aggregate Metadata Service
- AggMDS sync job – copies metadata from multiple data commons into a single data store.
- New `/aggregate` endpoint of the Metadata Service.
- JSON based configuration defines information about
 - Data source and Data Adapter information
 - Normalizing data fields
 - Adding optional individual overrides
- **Gen3 data adapter and adding new Custom Adapters.**

```
class RemoteMetadataAdapter(ABC):

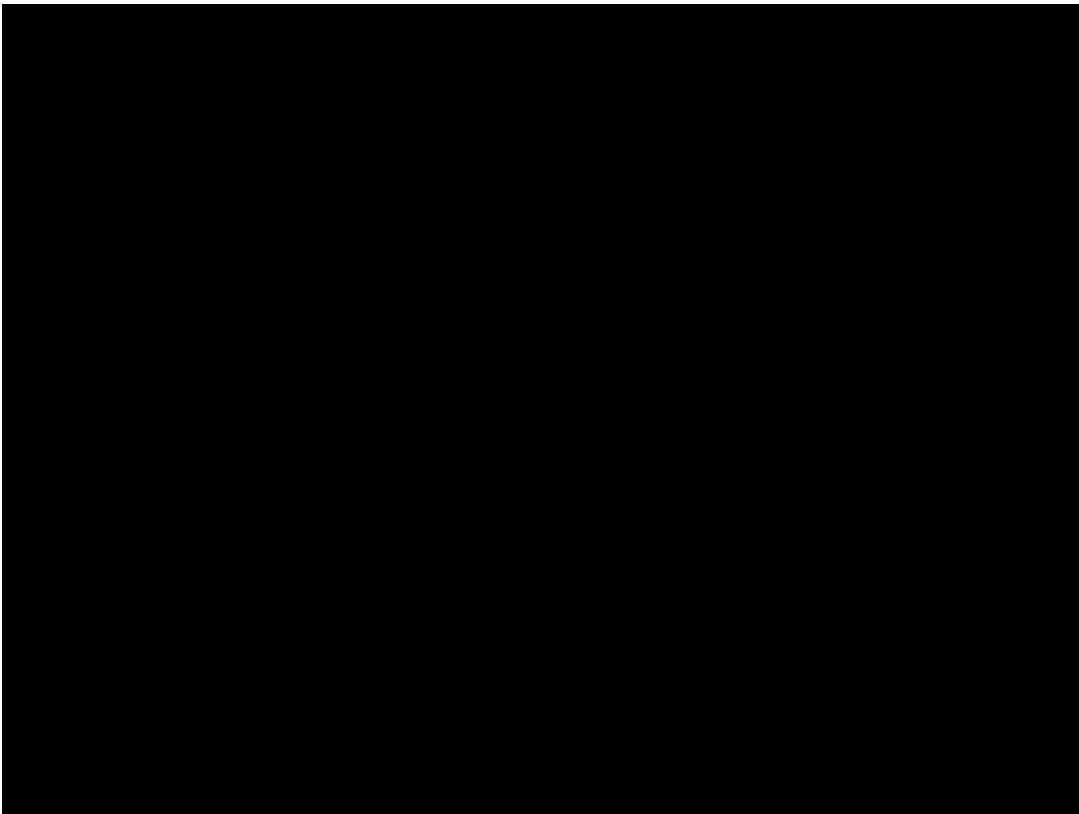
    @abstractmethod
    def getRemoteDataAsJson(self, **kwargs) -> Tuple[Dict, str]:
        """ needs to be implemented in derived class """

    @abstractmethod
    def normalizeToGen3MDSFields(self, data, **kwargs) -> Dict:
        """ needs to be implemented in derived class """

    @staticmethod
    def mapFields(item: dict, mappings: dict, global_filters: list = []) -> dict:
        """
        maps fields from the remote field name to the normalized, or
        standardized version. Unless you need special processing this funct:
        parameters:
            * item: metadata entry to be mapped
            * mappings: a dictionary of the remote field to normalize, this
                passed in from the configuration file_name
            * global_filters to apply
        """

    @staticmethod
    def setPerItemValues(item: dict, perItemValues: dict) -> None:
        """
        Overrides the item field values with those in perItemsValues.
        parameters:
            * item: metadata entry to override
            * perItemValues: a dictionary of field names to values
        """
```

Demo - AggMDS



View & Search Studies on Discovery Page

Biomedical Research Hub
Powered by Gen3

Discovery Workspace Example Analysis Profile

Summary Statistics | Tags | Table of Records | Pagination

Reset Selection Data Commons

516 STUDIES 604,204 TOTAL SUBJECTS

Search studies by keyword...

0 selected Login to Download Login to Open In Workspace

STUDY NAME FULL NAME

ADVANCED SEARCH >

IBD Commons

JCOIN

n/a n/a

high_coverage_2019_Public

Genotype Clinical Phenotype BioData Catalyst

n/a n/a 695 tutorial-biolincc_camp BioData Catalyst API

No description has been provided for this study.

Tutorial Clinical Phenotype BioData Catalyst

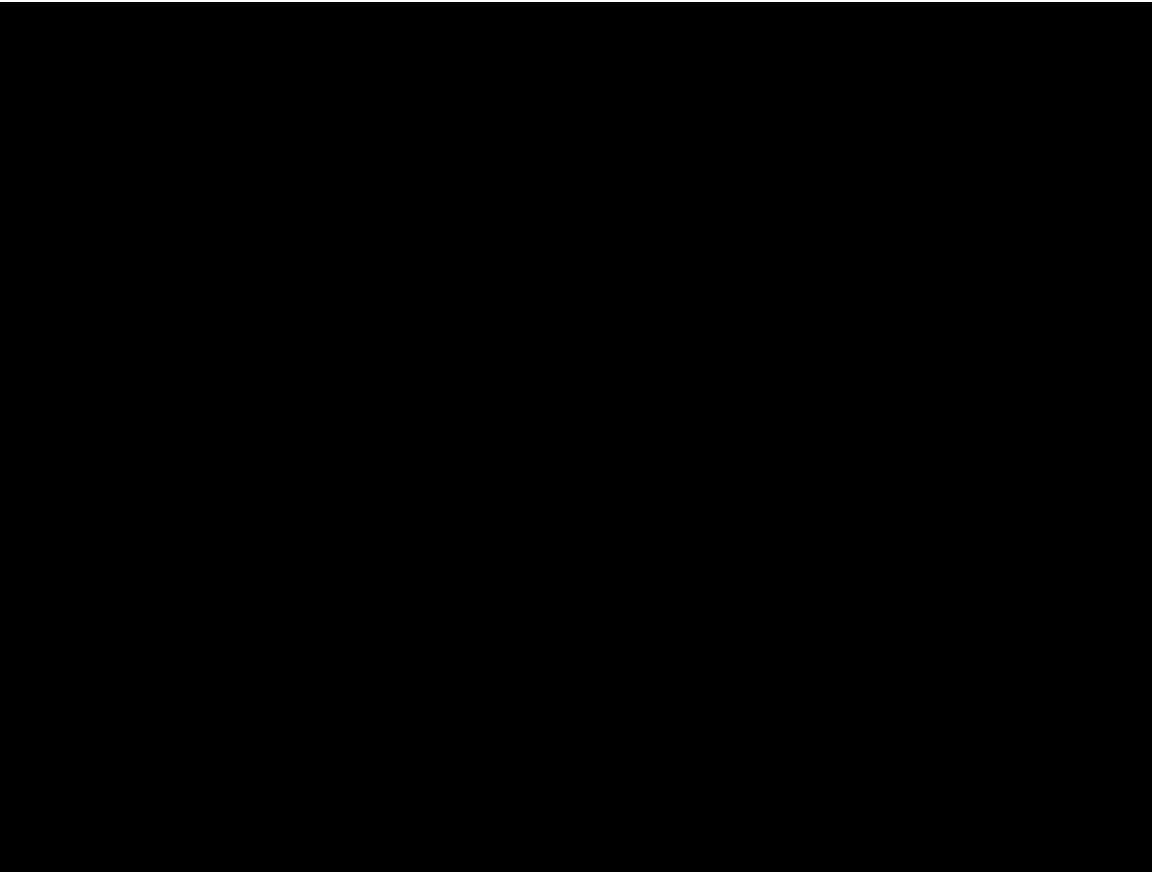
n/a n/a 4,434 tutorial-biolincc_framingham BioData Catalyst API, Single File

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. The study began in 1948 and 5,209 subjects were initially enrolled in the study. Participants...

Agenda

- Data Commons and Meshes
- Key Services and Jobs for Data Meshes
- Aggregate Metadata Sync
- **Token Service (Workspace Token service)**

Demo - Using WTS in a Data Mesh



Token Service (Workspace Token service)

- Acts as an additional OIDC client to interact with Fence on behalf of the user.
- Configuring additional data commons and updating **external_oidc** block.
 - Creating fence-clients on the external commons to generate **client-id** and **client-secret**
- New **/aggregate** endpoint of WTS allows users to fetch authz mapping of all the connected data commons.

Configuring additional data commons

- **external_oidc** field in appcreds.json
- Create this WTS as a fence-client in the other commons' fence.
 - Store the **client-id** and **client-secret** in the **external_oidc** block.
- **GET** request to **/wts/external_oidc**

Token Service (Workspace Token service)

/aggregate endpoint of WTS

- For users to be able to access the studies based on the access control policies they have in the respective connected commons, we need to have a mechanism to fetch the access control policies of the user in all the connected commons.
- This can be achieved by fetching the response from **/authz/mapping** from each of the connected commons.
- To get this , one needs to add **/authz/mapping** in the **aggregate_endpoint_allowlist** in appcreds.json
- Then send a GET request to the **/wts/aggregate/authz/mapping** in the data mesh

Connect Profile to External Data Resources



Help and Guidance | blarrick@uchicago.edu @ | Logout

NIH HEAL INITIATIVE HEAL Data Platform

Discovery Workspace Example Analysis Profile

Link accounts from external data resource(s)

JCOIN Google Login
IDP: jcoin-google
Provider URL: <https://jcoin.datacommons.io>

Status: not authorized

FAIR Repository Google Login
IDP: externaldata-google
Provider URL: <https://externaldata.healdata.org>

Status: not authorized

Create API key P

You don't have any API key. Please create one!

You have the following API key(s)

API key(s)	Expires
------------	---------

You have access to the following resource(s)

Resource(s)	Method(s)
/argocd	access
/cedar	*, create
/dashboard	access
/data_file	file_upload

Acknowledgements



- Gen3 Forum Steering Committee
 - Robert Grossman, Center for Translational Data Science, University of Chicago
 - Steven Manos, Australian BioCommons
 - Claire Rye, New Zealand eScience Infrastructure
 - Plamen Martinov, Open Commons Consortium
 - Michael Fitzsimons, Center for Translational Data Science, University of Chicago
- Speakers
 - Robert Grossman, Center for Translational Data Science, University of Chicago
 - Sai Narumanchi, Center for Translational Data Science, University of Chicago
 - Aarti Venkat, Center for Translational Data Science, University of Chicago
 - Phil Schumm, Biostatistics Laboratory, Department of Public Health, University of Chicago

Open Discussion

Topic Ideas for Gen3 Community Events