

Data Analysis in Gen3 systems

Gen3 Community Forum
7 May 2025

The Agenda

- Introduction
- Using the Portable Format for Biomedical data (PFB) and the Data Library in the Biodata Catalyst Ecosystem to host and export multi-modal data from Gen3 to Terra and Seven Bridges - **Kyle Burton, CTDS, University of Chicago**
- Enabling collaborative environmental health research using ToxDataCommons - **Rance Nault, Michigan State University**
- Using the Task Execution Service (TES) in Gen3 for biological applications - **Pauline Ribeyre and Aarti Venkat, CTDS, University of Chicago**
- Toward AI-ready data commons: from computable data standards to interoperable AI models - **Jing Su, Indiana University**

Using the Portable Format for Biomedical data (PFB) and the Data Library in the Biodata Catalyst Ecosystem to host and export multi-modal data from Gen3 to Terra and Seven Bridges

Kyle Burton, CTDS, University of Chicago

NHLBI BioData Catalyst (BDC)



Launched in 2018, BDC is a cloud-based ecosystem that offers researchers data, analytic tools, applications, and workflows in secure workspaces.

Gen3 provides APIs for data queries and download, and providing cloud-based analysis workspaces by Velsera and Terra with rich tools and resources.



Portable Format for Bioinformatics (PFB)



Allows users to transfer the data, data model, and pointers to files in one package.

Data can be transferred while keeping the structure from the original source.

Consists of three parts:

1. Schema – Describes the properties in a JSON Data dictionary.
2. Metadata – Explains the links between nodes for each of the properties
3. Data – Values for the properties

Cohort builder for export and dynamic summary statistics display:

- Search facets leveraging harmonized variables.

Standardized Cohort Handoff support to move cohort to analysis workspaces in Broad's Terra system or Velsera's Seven Bridges system



The screenshot shows the BioData CATALYST interface, which is "Powered by Gen3". The top navigation bar includes links for NIH, BioData CATALYST (highlighted in red), Discovery, Dictionary, Exploration (which is underlined in red), Data Library, and Profile. Below the navigation is a toolbar with Data and File tabs, and links for Explorer Filters, Data Tools, Summary Statistics, and Table of Records. A prominent feature is a "Filters" section with a "Harmonized Variables" dropdown menu showing Project and Subject options, along with a "Collapse all" button and a search bar. Below this are buttons for Export to Seven Bridges, Export All to Terra, and Export to PFB. The main content area displays summary statistics: Subjects (1,130,330) and Projects (287). The entire interface has a clean, modern design with a white background and red accents for key features.

1. PFBs are generated dynamically for every handoff today
 - a. Many researchers prefer handing off an entire study / studies
2. We have many new datasets with different types of data coming in
 - a. The Gen3 Data Model can be expanded to include new searchable terms (this takes time)
3. The Gen3 Discovery Page provides source of truth dataset-level metadata, but UI PREVIOUSLY didn't support selection of datasets

If we improved 1,2,3: Researchers would have less friction in finding and getting data they're interested in handed off to analysis. Plus, we'd be able to provide more timely release of data.

Gen3 BDC Data Ingestion Pipeline creates **Whole Study PFBs** for each study in BDC, accessible by Gen3 Discovery and the Gen3 User Data Library.

Whole Study PFBs contain data for the entire study.

Whole Study PFBs data dictionary can be independent of the deployed Gen3 Data Dictionary

Whole-Study PFBs are available for transfer immediately.

PFB in BDC - Gen3 Discovery



Whole Study PFBs are available in BDC at Gen3 Discovery

Users select datasets from Discovery to add the study's PFBs to their **Data Library**

Gen3 Discovery also consists of study level FHIR data, DOI for the dataset, and other public metadata.

The screenshot shows a search results page for 'Framingham'. At the top, there are counts for 'STUDIES' (11) and 'TOTAL SUBJECTS' (48,777), a search bar containing 'Framingham', and buttons for 'Reset Selection' and 'Study Filters'. Below this is a table with columns: STUDY NAME, FULL NAME, NUMBER OF SUBJECTS, DBGAP ACCESSION NUMBER, RELEASED, and DATA AVAILABILITY. Three rows are listed:

STUDY NAME	FULL NAME	NUMBER OF SUBJECTS	DBGAP ACCESSION NUMBER	RELEASED	DATA AVAILABILITY
FHS_HMB-IRB-MDS_	Framingham Cohort	13070	phs000007.v31.p12.c1	Yes	
FHS_HMB-IRB-NPU-MDS_	Framingham Cohort	2079	phs000007.v31.p12.c2	Yes	
Framingham Heart Study (FHS) Imaging	Framingham Heart Study (FHS) Imaging	9122	phs003593.v1.p1.c1	Partially	

Each row has a detailed description below it, mentioning the Framingham Heart Study and its history. The last row also includes a note about the population-based nature of the study.

Gen3 User Data Library - BDC



Allows users to create lists of datasets.

- Lists contain whole study PFBs from Discovery

Lists are persisted across sessions.

- Researchers can always reference the datasets used in analysis

List items can be handed off to analysis platforms

- In BDC, these are Broad's Terra and Velsara's Seven Bridges.
- In Gen 3.2, support for export to PFB is expected in June.
- Extendable to any analysis platform or workspace that supports PFB parsing.

A screenshot of the BioData Catalyst Data Library interface. At the top, there is a navigation bar with the NIH logo, the BioData CATALYST logo, and links for Exploration, Discovery, Dictionary, Data Library, and Profile. Below the navigation bar, there is a section titled "Retrieve Selected" with a dropdown menu set to "Framingham c1". To the right of this, it shows "CREATED: May-1-2025 17:08:22" and "UPDATED: Apr-28-2025 09:47:24". Below this, there is a table with two rows. The first row has columns for "NAME", "ID", and "# FILES". It contains one item: "phs000007.v31.pl2.cl" with ID "2". The second row is titled "FILES" and has columns for "NAME" and "DESCRIPTION". It contains two items: one checked item with the description "Harmonized clinical data and subject-level sample file pointers. Harmonized to BDC's core standards." and one unchecked item with the description "Preharmonized files provided by the BioData Catalyst Data Management Core (DMC)". At the bottom, there is another section for "Framingham c2" with similar creation and update times.

Gen3 User Data Library - Analysis Handoff

- Users export whole study PFBs to Seven Bridges or Terra, where they are parsed and data is organized into data frames
- Files within the PFB that are accessible through GA4GH DRS are also pulled into their workspaces.
 - Example: CRAM, BAI, etc. files are available for workflows within these respective analysis systems.

The screenshot shows the BioData CATALYST interface. At the top, there are tabs for Exploration, Discovery, Dictionary, Data Library (which is selected), and Profile. Below the tabs, a message says "CREATED: Jan-30-2025 17:22:05 • UPDATED: Feb-25-2025 15:42:58". A "Retrieve Selected" button is visible. In the center, there's a table with columns: NAME, ID, # FILES, and ADDITIONAL DATA SOURCES. A modal window titled "Retrieve Data" is open, showing a list of selected files. One file is highlighted: "drg.712C/80215653-eb94-4c06-8ea5-8b6ff6e2c1fe Raw Serialized PFB created with drs guids from data-simulator GA4GH_DRS". The "Destination" dropdown menu shows "Export: Terra" and "Export: Seven Bridges".



The screenshot shows the Seven Bridges workspace interface. The top navigation bar includes DASHBOARD, DATA (selected), ANALYSES, WORKFLOWS, and SUBMISSION HISTORY. The main area has tabs for TABLES, REFERENCE DATA, and OTHER DATA. A large table lists various datasets, each with a "pfb:" prefix and a "state" column indicating "validated". An example row is "original_file (43678)" with "pfb: ga4gh.drs.uri" and "pfb: project_id". A search bar at the top right contains "Search" and a magnifying glass icon. At the bottom, there are pagination controls and an "Items per page" dropdown set to 100.

	pfb: ga4gh.drs.uri	pfb: project_id	pfb: state
original_file (43678)			
HG00096.cram	drs://drg.712C/drg.712C/291f273f-ec...	tutorial-synthetic_data_set_1	validated
HG00097.cram	drs://drg.712C/drg.712C/c98d1eb7-3...	tutorial-synthetic_data_set_1	validated
HG00099.cram	drs://drg.712C/drg.712C/24097e7b-...	tutorial-synthetic_data_set_1	validated
HG00100.cram	drs://drg.712C/drg.712C/15b46746-f...	tutorial-synthetic_data_set_1	validated
HG00101.cram	drs://drg.712C/drg.712C/9bcd72c6-7...	tutorial-synthetic_data_set_1	validated
HG00102.cram	drs://drg.712C/drg.712C/28ec8028-2...	tutorial-synthetic_data_set_1	validated
HG00103.cram	drs://drg.712C/drg.712C/616d7bcf-4...	tutorial-synthetic_data_set_1	validated
HG00105.cram	drs://drg.712C/drg.712C/4e43c739-9...	tutorial-synthetic_data_set_1	validated
HG00106.cram	drs://drg.712C/drg.712C/8490e2e7-2...	tutorial-synthetic_data_set_1	validated
HG00107.cram	drs://drg.712C/drg.712C/82b280ea-3...	tutorial-synthetic_data_set_1	validated

Benefits to Whole Study PFBs

1. Available for transfer immediately
 - a. No need to dynamically generate PFBs, reducing time and compute costs.
2. Reduces data model constraint on data ingestion
3. Supports other data types before data model adjustments
 - a. Data can be harmonized to other Gen3 DDs, rather than updating the data common's dictionary
4. Data is made available sooner
 - a. Data submitters can provide their own Gen3 DD representing their data.
5. Better data provenance, easier versioning and archiving in the future
 - a. As studies update, the previous whole study PFBs can be archived
6. Flexibility to support multi-model datasets going forward

- Gen3 User Data Library
 - Backend service, providing an API for users to manage lists
 - <https://github.com/uc-cdis/gen3-user-data-library>
- Gen3 Frontend Framework
 - Provides the user interface for a Gen3 data commons, including the data library
 - <https://github.com/uc-cdis/gen3-frontend-framework>
- Pypfb
 - Used to create Whole Study PFBs in BDC
 - <https://github.com/uc-cdis/pypfb>
 - Available in gen3sdk: <https://pypi.org/project/gen3/>
- More information about PFBs in BDC
 - <https://bdcatalyst.gitbook.io/biodata-catalyst-documentation/written-documentation/explore-available-data/gen3-discovering-data/pfb-files>

Enabling collaborative environmental health research using ToxDataCommons

Rance Nault, Michigan State University



Enabling collaborative environmental health research using ToxDataCommons

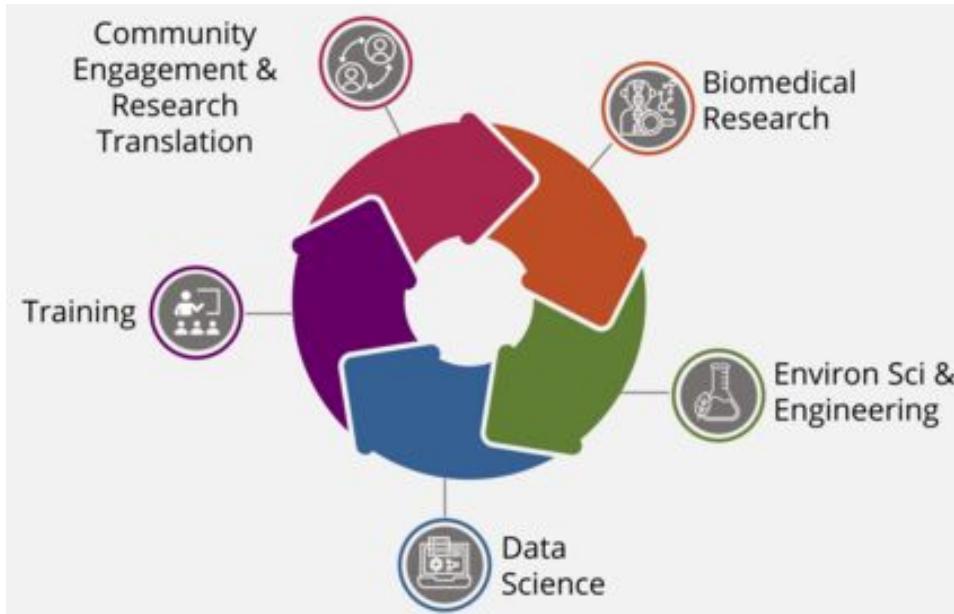
Gen3 Community Forum

Rance Nault

Department of Pharmacology & Toxicology, Institute for Integrative Toxicology
Michigan State University



MSU SUPERFUND RESEARCH CENTER



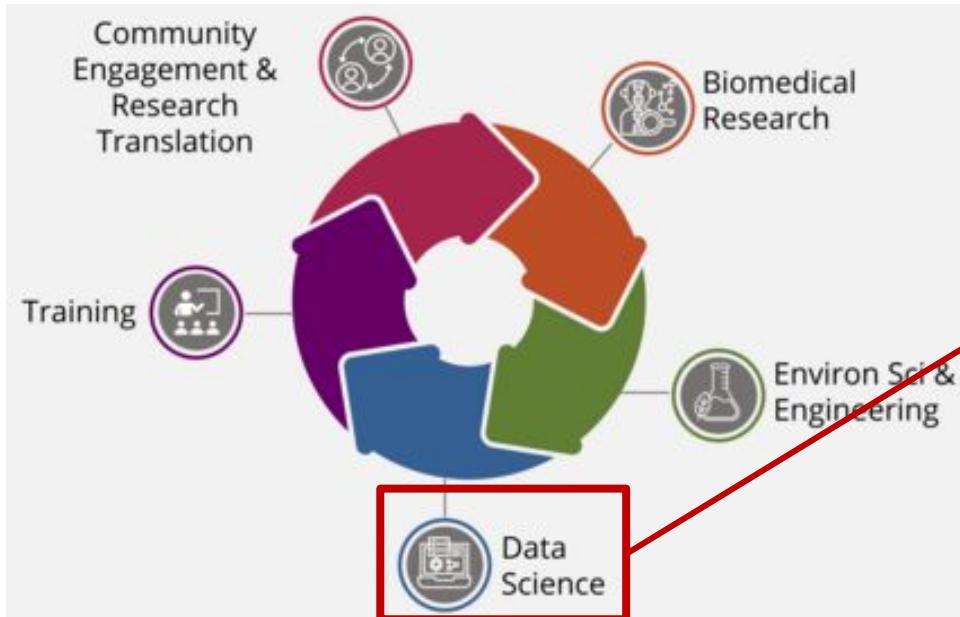
SRP provides practical, scientific solutions to protect health, the environment, and communities.

The **MSU SRC** is focused on the environmental contaminants that activate the aryl hydrocarbon receptor.

<https://iit.msu.edu/centers/superfund/>



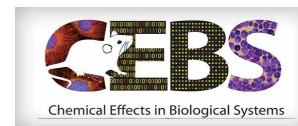
MSU SUPERFUND RESEARCH CENTER



SRP required Multiproject Center applicants to include a Data Management and Analysis Core (DMAC) to support the management and integration of data assets. The DMACs are intended to foster and enable the interoperability of data across the Center's projects and cores to accelerate the impact of the Center's research.

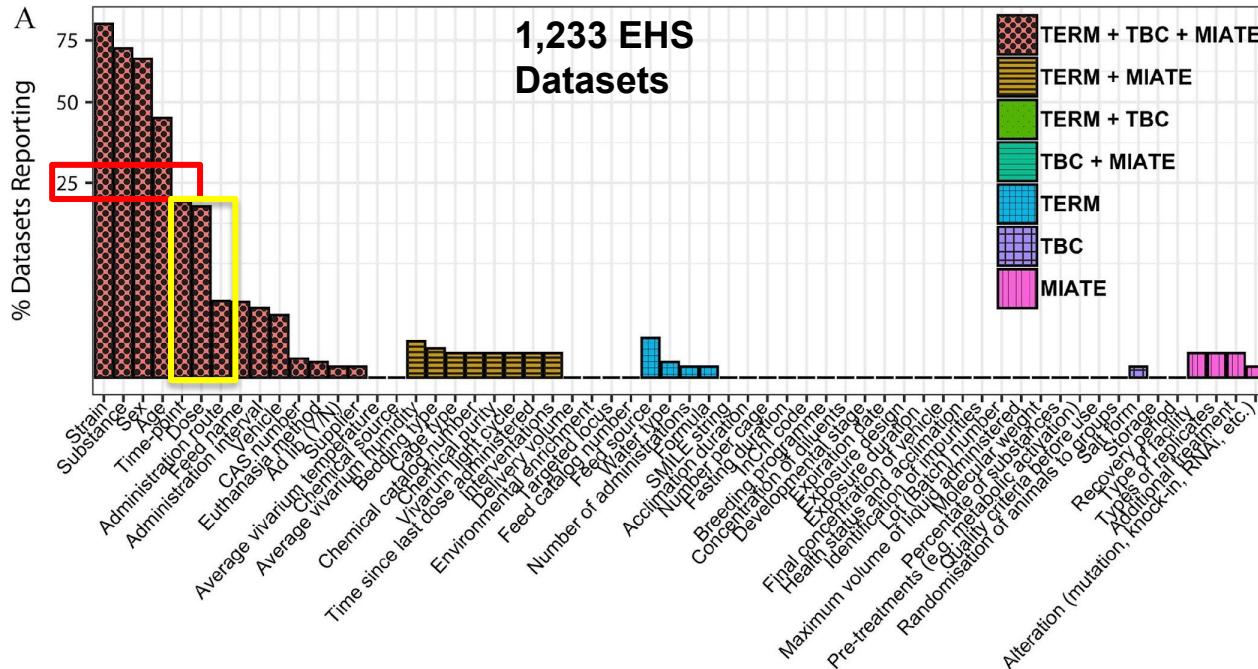
STATUS OF EHS DATA SHARING

Thousands of datasets are available through public databases and repositories



How many can be considered FAIR?

STATUS OF EHS DATA SHARING



Beyond challenges in finding relevant data, many are missing key information.

This makes use of AI/ML approaches challenging.



DEVELOPMENT OF ToxDATACOMMONS

Motivated by a critical gap in EHS data sharing but leveraged to accelerate collaborative research

The screenshot shows the Tox Data Commons web application. At the top, there is a dark green header bar with three items: "Tox Data Harmonizer", "Documentation", and "Login". Below the header is the main content area. In the center, there is a logo for "Superfund Research Center" next to a cloud icon containing the letters "DMAC". Below the logo, the text "Tox Data Commons" is centered. At the bottom, there is a horizontal navigation bar with six items: "Discovery" (with a magnifying glass icon), "Dictionary" (with an A-Z book icon), "Exploration" (with a circular arrow icon), "Query" (with a magnifying glass icon), "Example Analyses" (with a bar chart icon), and "Profile" (with a shield icon).



DEVELOPMENT OF ToxDATACOMMONS

Motivated by a critical gap in EHS data sharing but leveraged to accelerate collaborative research

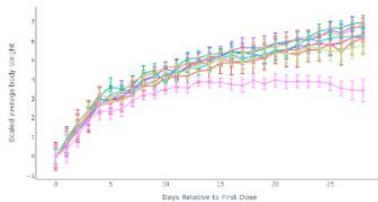
The screenshot shows the Tox Data Commons web application. At the top, there is a dark header bar with three items: "Tox Data Harmonizer", "Documentation", and "Login". Below the header, the "Superfund Research Center" logo is displayed next to a "DMAC" icon inside a cloud shape. The main content area has a title "Tox Data Commons" above several navigation links. These links include "Discovery" (with a magnifying glass icon), "Dictionary" (with an A-Z icon), "Exploration" (with a circular icon), "Query" (with a magnifying glass icon), "Example Analyses" (which is highlighted with a large red rectangular box), and "Profile" (with a shield icon). Each link has a small icon to its left.

COLLABORATIVE RESEARCH APPLICATIONS

Developing use cases that can be run anywhere

TDC Gross Pathology Visualization

This Jupyter notebook presents a use case for the toxdatacommons to plot basic endpoints such as body weight, food consumption, and terminal endpoints.



IN-HOUSE
MSU
HPCC
CLUSTER



TDC Metabolism-based graph neural network analysis

Not public yet



STANDARDIZED WORKFLOW

 Template based submissions

gen3 File Edit View Insert Format Data Tools Extension
N12 A B C
1 study_submitter_id euthanasia_date euthanasia_method eut
2 PRJ139
3 PRJ139
4 PRJ139
5 PRJ139
6 PRJ139
7 PRJ139
8 PRJ139
9 PRJ139
10 PRJ139
11 PRJ139

 Study-based Data Selection

Filters Subject Properties Study Properties Treatments
Collapse all
Submitter Id Select...
Sex male 367
female 164
Strain C57BL/6NCrl 453

 Batch dataset Downloads

```
> gen3-client  
download-multiple \  
--profile=ToxCDC \  
--manifest=manifest.json \  
--download-path=downloads
```

Finished
downloads/63af95d3-98c3-4d6d-a
6be-26398dbfc1d9 6723044 /
6723044 bytes (100%)

METADATA



CASE 1: INTEGRATING GROSS PATHOLOGY

Datasets were generated in at least 3 independent studies with several measured endpoints (weights, pathology, gene expression, ...).

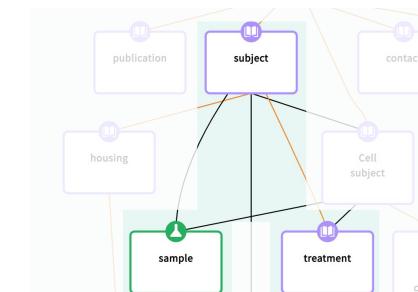
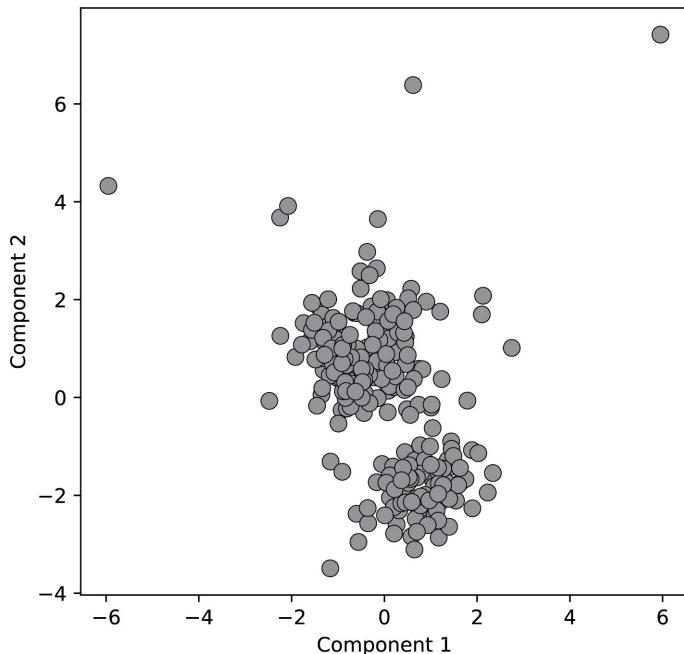


Can we identify informative patterns about chemical toxicity by integrating these studies?

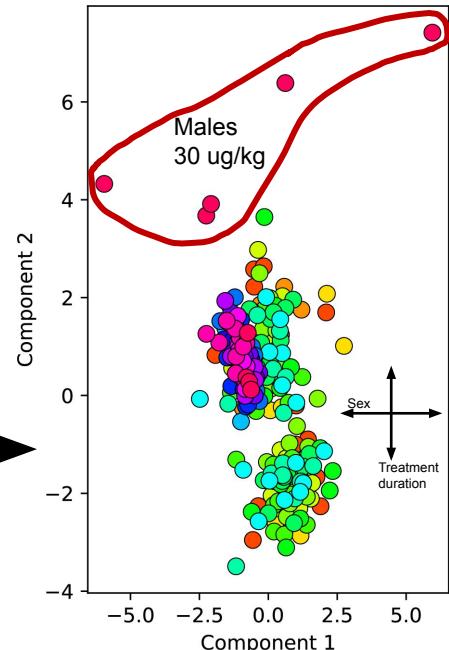


CASE 1: INTEGRATING GROSS PATHOLOGY

Dynamic time warping analysis of daily body weights of each *subject*.



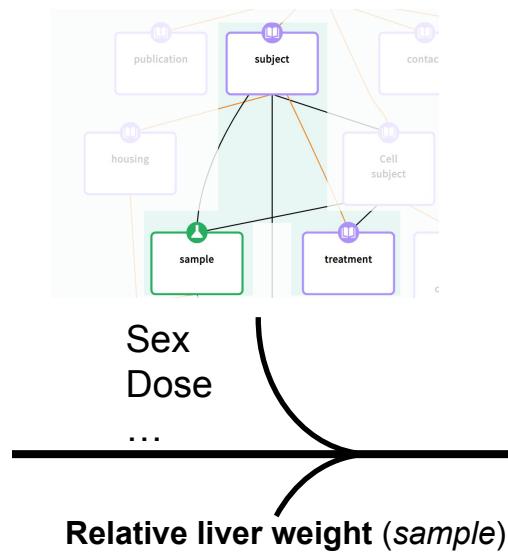
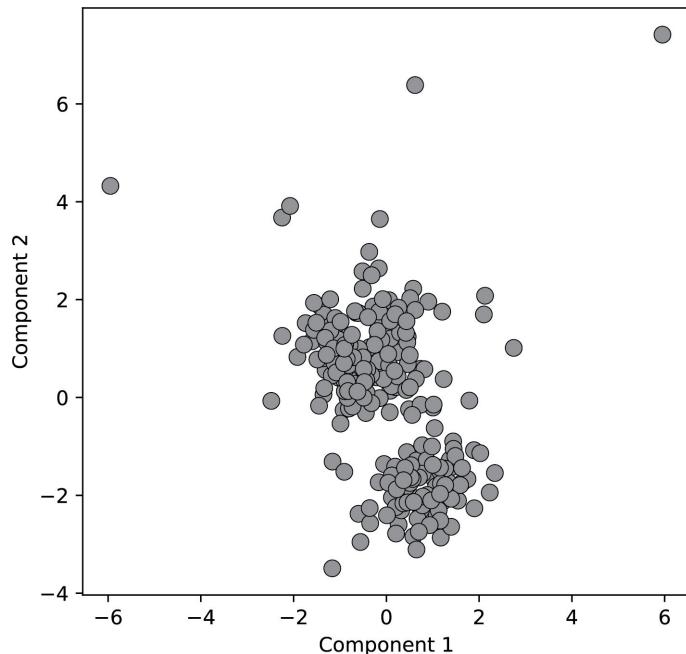
Metadata annotation reveals clustering



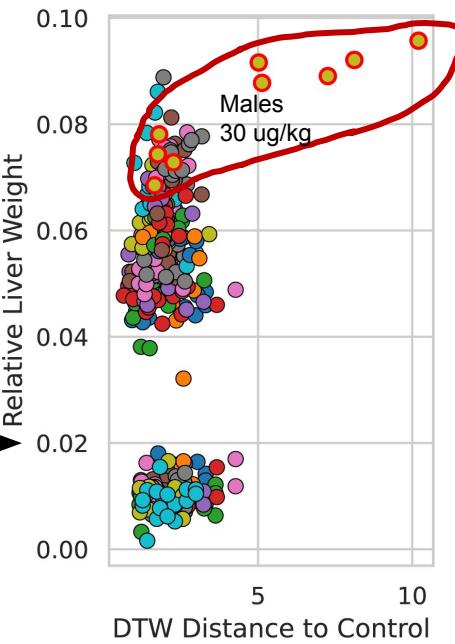


CASE 1: INTEGRATING GROSS PATHOLOGY

Dynamic Time Warping analysis of daily body weights of each *subject*.

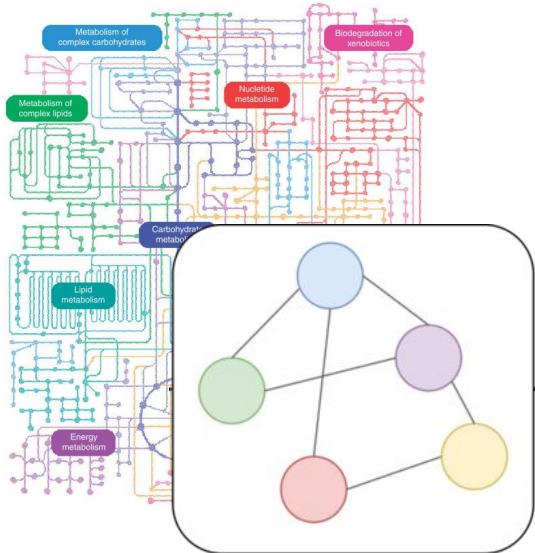


Metadata annotation reveals clustering

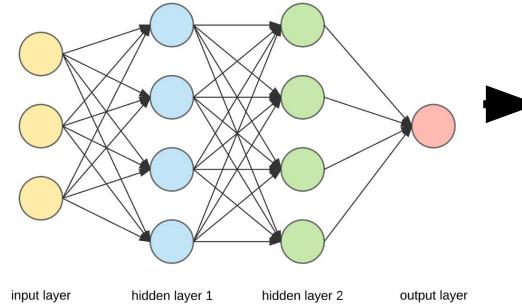




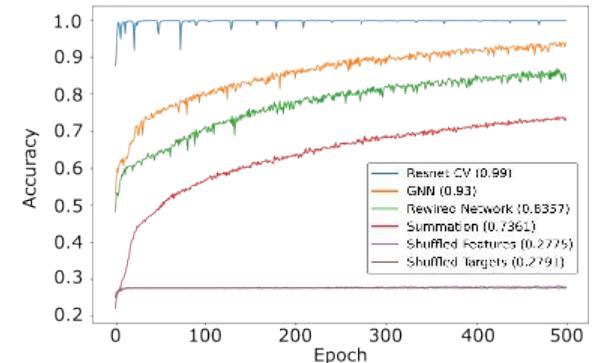
CASE 2: METABOLIC REACTION GNN



Metabolism is well suited for graph neural network approaches



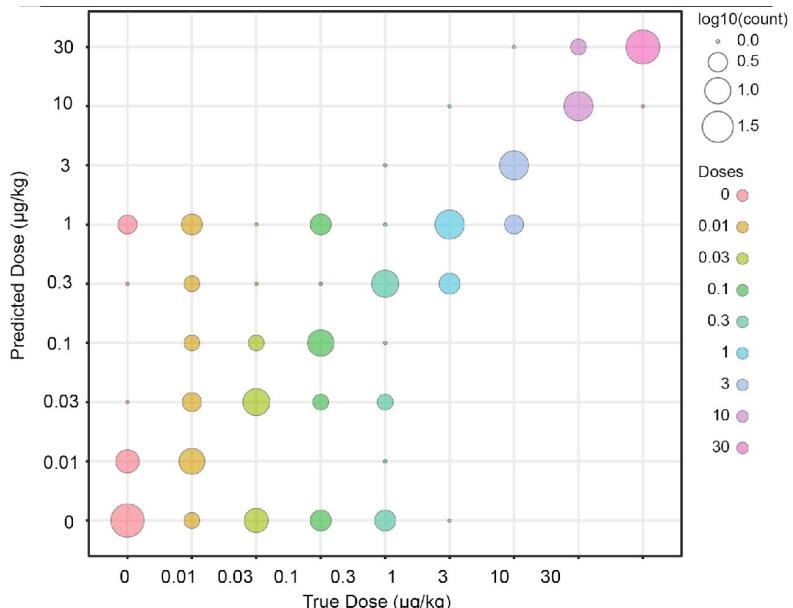
Only ~4% of samples were used, a significant portion excluded due to missing metadata



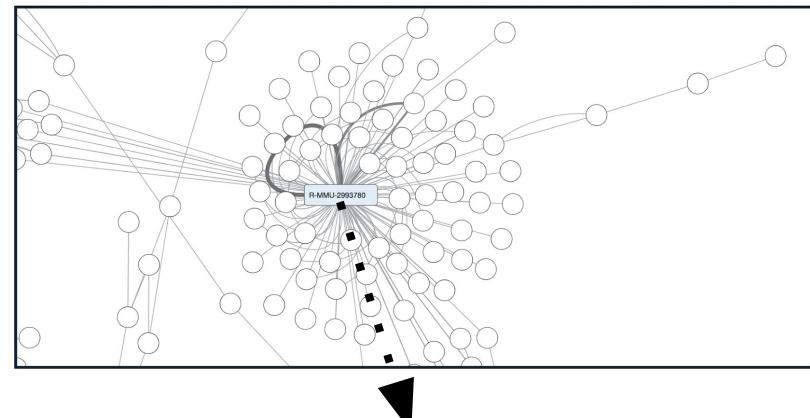
GNN training strategy originally described in:
Berkhout, J. G., et al., (2023).
<https://doi.org/10.1016/j.patter.2023.100758>

CASE 2: METABOLIC REACTION GNN

Study datasets are processed through standardized workflow and GNN model



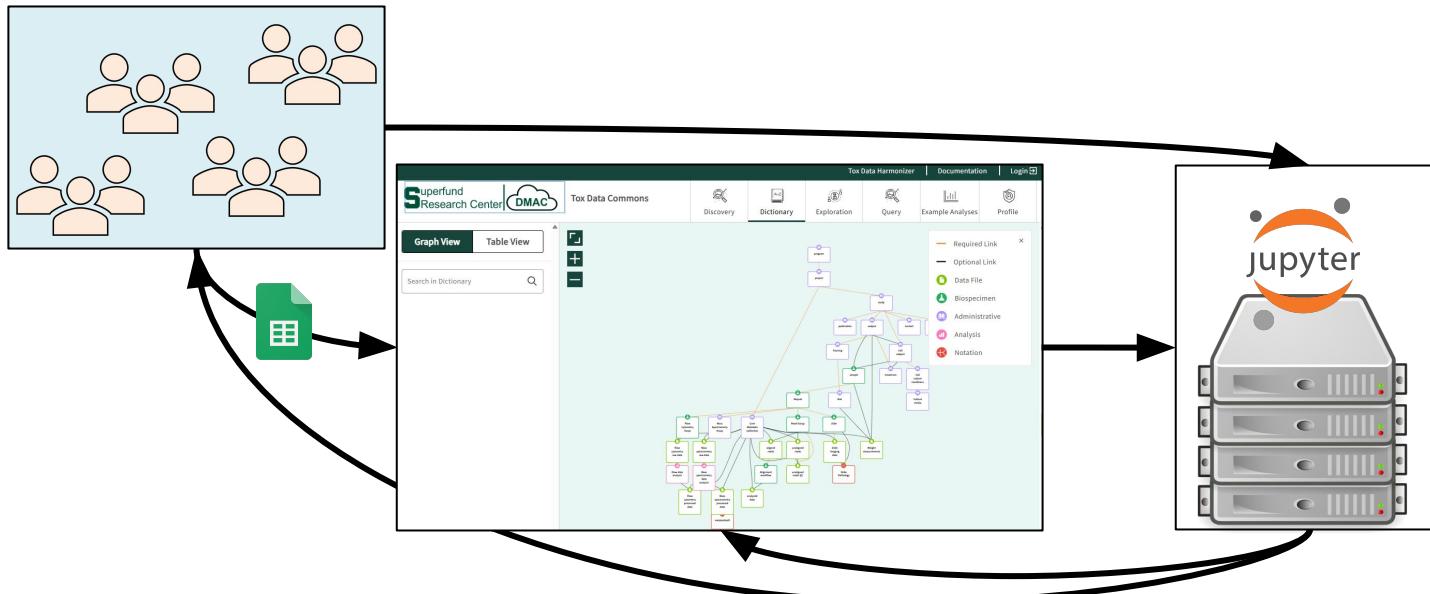
Identify metabolic reactions which are not found using more traditional methods.



“Transfer of SUMO1 from E1 to UBE2I (UBC9)” is a reaction predictive of treatment group



SUMMARY: GEN3 ENABLES COLLABORATION



Standardizing metadata plays a crucial role in advancing collaborative science

FUTURE DIRECTION

- Containerization of analyses to support reproducibility across HPCC/Cloud resources.
- Develop and automate novel multi-model AI/ML tools.
(gross pathology, histopathology, transcriptomic, clinical chemistry, ...)
- Implement the tools and resources in a scalable manner beyond the MSU Superfund Research Program (especially metadata collection tool)



ACKNOWLEDGEMENTS

MSU Superfund Research Center

Keji Yuan
Tim Zacharewski
Giovan Cholico
Eric Kasten
Jonathan Babbage
Todd Hall

CTDS:

Chris Meyer
Ed Malinowski
Jawad Qureshi



National Institute of
Environmental Health Sciences
Superfund Research Program



PVAT PPG

Funding:

NIEHS SRP P42 ES004911
NHLBI P01 HL152951

Using the GA4GH Task Execution Service (TES) in Gen3 for biological applications

**Pauline Ribeyre and Aarti Venkat, CTDS,
University of Chicago**

Typical use cases for biological research and discovery



- Bioinformatics or data science workflows (nextflow, or other workflow language)
 - RNASeq, Variant calling, Copy number inference, Methylation and others
 - Training and testing AI/ML models for prediction or classification task
 - Federated learning applications
- Data
 - Small batch of data locally for testing and development
 - Remote, S3 bucket
 - Bring your own data
 - Data from a data commons
- Containers
 - Single or multiple for each workflow
 - E.g. fastqc, salmon, deepvariant etc

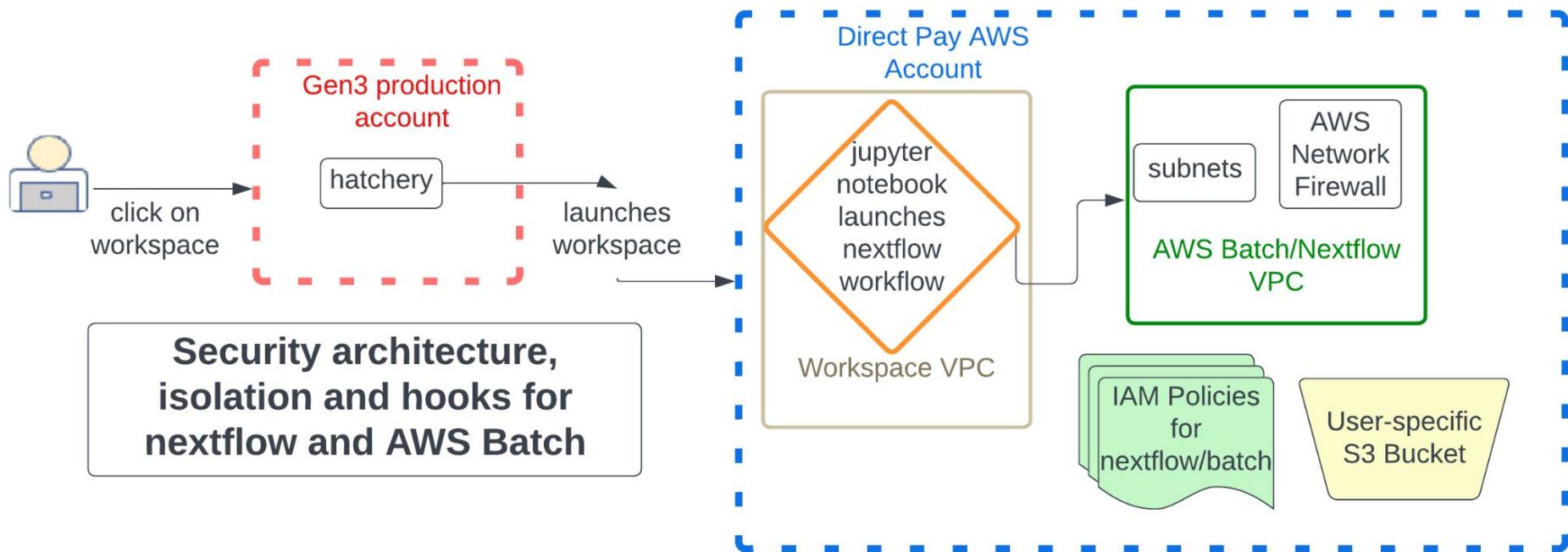
Typical use cases for biological research and discovery



- Bioinformatics or data science workflows (nextflow, or other workflow language)
 - RNASeq, Variant calling, Copy number inference, Methylation and others
 - Training and testing AI/ML models for prediction or classification task
 - Federated learning applications
- Data
 - Small batch of data locally for testing and development
 - Remote, S3 bucket
 - Bring your own data
 - Data from a data commons
- Containers
 - Single or multiple for each workflow
 - E.g. fastqc, salmon, deepvariant etc

We are building a mechanism to run containerized workflows in Gen3 in a secure, isolated and scalable manner

Our previous proof-of-concept solution (v1)



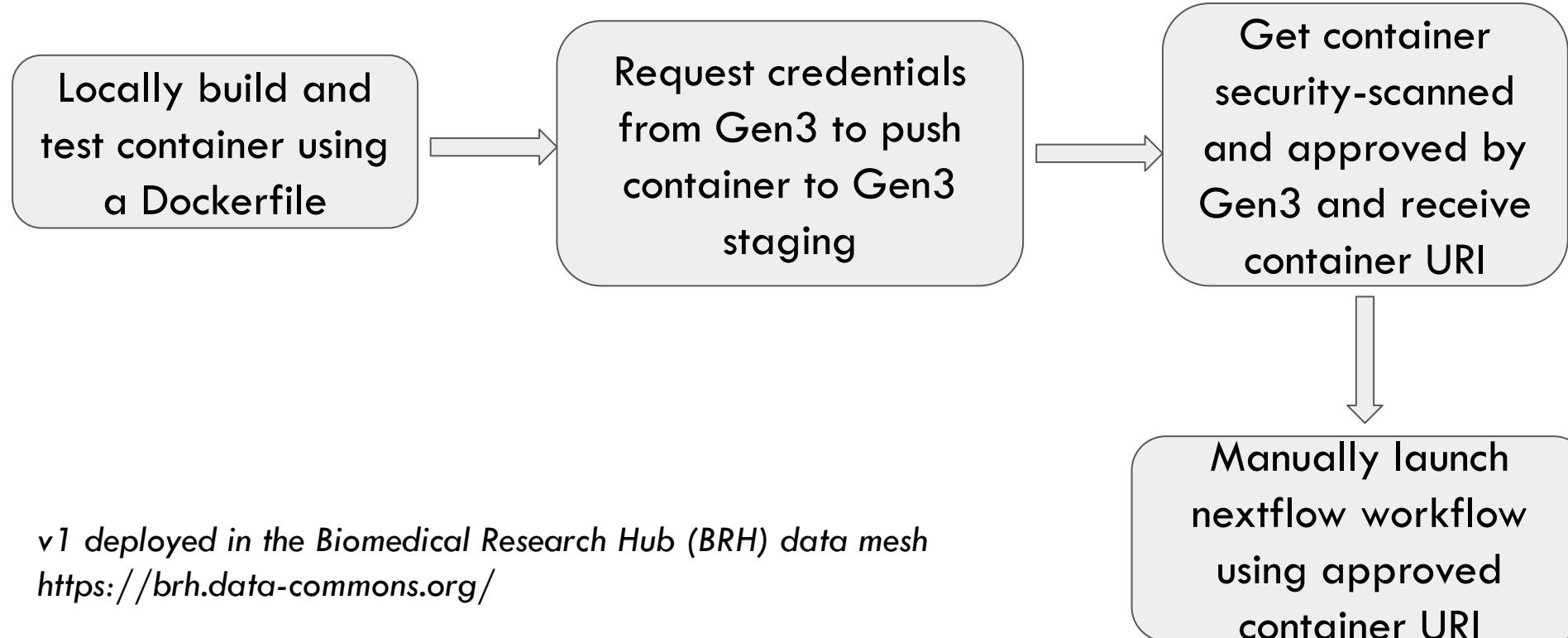
v1 solution enabled launching workflows from a JupyterLab notebook in Gen3 workspace

Overview of steps

Run a user-developed nextflow workflow with container (v1)



Containers are run in FedRAMP moderate environment with security compliance



v1 deployed in the Biomedical Research Hub (BRH) data mesh
<https://brh.data-commons.org/>

BRH workspace

← → G brh.data-commons.org/workspace 🔍 ⭐ A

Documentation Email Support Cite BRH aartiv@uchicago.edu Logout

Biomedical Research Hub Powered by GEN3

Discovery Workspace Example Analysis Profile

Account Information

Account	Workspace Account Manager Edit
Direct Pay ▼	
Total Charges (USD)	Spending Limit (USD)
0.00	225.00

To analyze all data to which you have access, please authorize external data resources in the [Profile](#) page. X



(Beta) Nextflow with CPU instances

4.0CPU, 10Gi memory

[Launch](#)



(Beta) Nextflow with GPU instances

4.0CPU, 10Gi memory

[Launch](#)



(Generic) Jupyter Lab Notebook with R Kernel

2.0CPU, 8Gi memory

[Launch](#)



(Tutorials) Example Analysis Jupyter Notebooks

2.0CPU, 8Gi memory

[Launch](#)

Launch workflow from JupyterLab notebook

The screenshot shows a JupyterLab interface with the following components:

- File Browser:** On the left, there is a sidebar with a file tree. The current path is shown as `/ pd`. The tree lists several files and directories:
 - chip_data (17 hours ago)
 - chip_results (1 minute ago)
 - sdk_data (14 days ago)
 - chip_template.ipynb (19 hours ago)
 - chip.ipynb (3 minutes ago) - This file is selected.
 - main.nf (5 minutes ago)
 - midrc_gpu_batch_test.... (14 days ago)
 - nextflow.config (5 minutes ago)
- Terminal:** In the center, there is a terminal window titled `jovyan@:~/pd` running a Python 3 (ipykernel) session. The terminal output shows the execution of a Nextflow pipeline:

```
executor > awsbatch (2)
[a9/9c9b2d] process > identify_chip_variants (1) [ 0%] 0 of 2

executor > awsbatch (2)
[a9/9c9b2d] process > identify_chip_variants (1) [ 0%] 0 of 2

executor > awsbatch (2)
[1b/d6412f] process > identify_chip_variants (2) [ 50%] 1 of 2

executor > awsbatch (2)
[1b/d6412f] process > identify_chip_variants (2) [ 50%] 1 of 2

executor > awsbatch (2)
[a9/9c9b2d] process > identify_chip_variants (1) [100%] 2 of 2 ✓
Completed at: 21-Mar-2024 22:03:56
Duration : 3m 11s
CPU hours : (a few seconds)
Succeeded : 2
```
- Bottom Navigation:** At the bottom, there are buttons for "Terminate Workspace" (orange), "Exit Fullscreen" (blue), and workspace status indicators (Python 3 (ipykernel) | Idle).

Launch workflow from JupyterLab notebook

The screenshot shows a JupyterLab workspace interface. On the left, there is a file browser with a sidebar for account information. The main area contains three tabs: 'nextflow-welcome.html', 'chip.ipynb' (which is currently active), and 'jovyan@:~/pd'. The 'chip.ipynb' tab displays a command-line log of a Nextflow workflow execution:

```
executor > awsbatch (2)
[a9/9c9b2d] process > identify_chip_variants (1) [ 0%] 0 of 2

executor > awsbatch (2)
[a9/9c9b2d] process > identify_chip_variants (1) [ 0%] 0 of 2

executor > awsbatch (2)
[1b/d6412f] process > identify_chip_variants (2) [ 50%] 1 of 2

executor > awsbatch (2)
[1b/d6412f] process > identify_chip_variants (2) [ 50%] 1 of 2

executor > awsbatch (2)
[a9/9c9b2d] process > identify_chip_variants (1) [100%] 2 of 2 ✓
Completed at: 21-Mar-2024 22:03:56
Duration : 3m 11s
CPU hours : (a few seconds)
Succeeded : 2
```

At the bottom, there are buttons for 'Terminate Workspace' and 'Exit Fullscreen'.

Low throughput submission and testing, limited to AWS Batch, cost tracking with direct pay

THEME ARTICLE: CONVERGED COMPUTING: A BEST-OF-BOTH
WORLDS OF HPC AND CLOUD

The GA4GH Task Execution Application Programming Interface: Enabling Easy Multicloud Task Execution

Alexander Kanitz , University of Basel, 4056, Basel, Switzerland

Matthew H. McLoughlin , Microsoft Research and AI, Redmond, WA, 98052, USA

Liam Beckman , Oregon Health and Science University, Portland, OR, 97239, USA

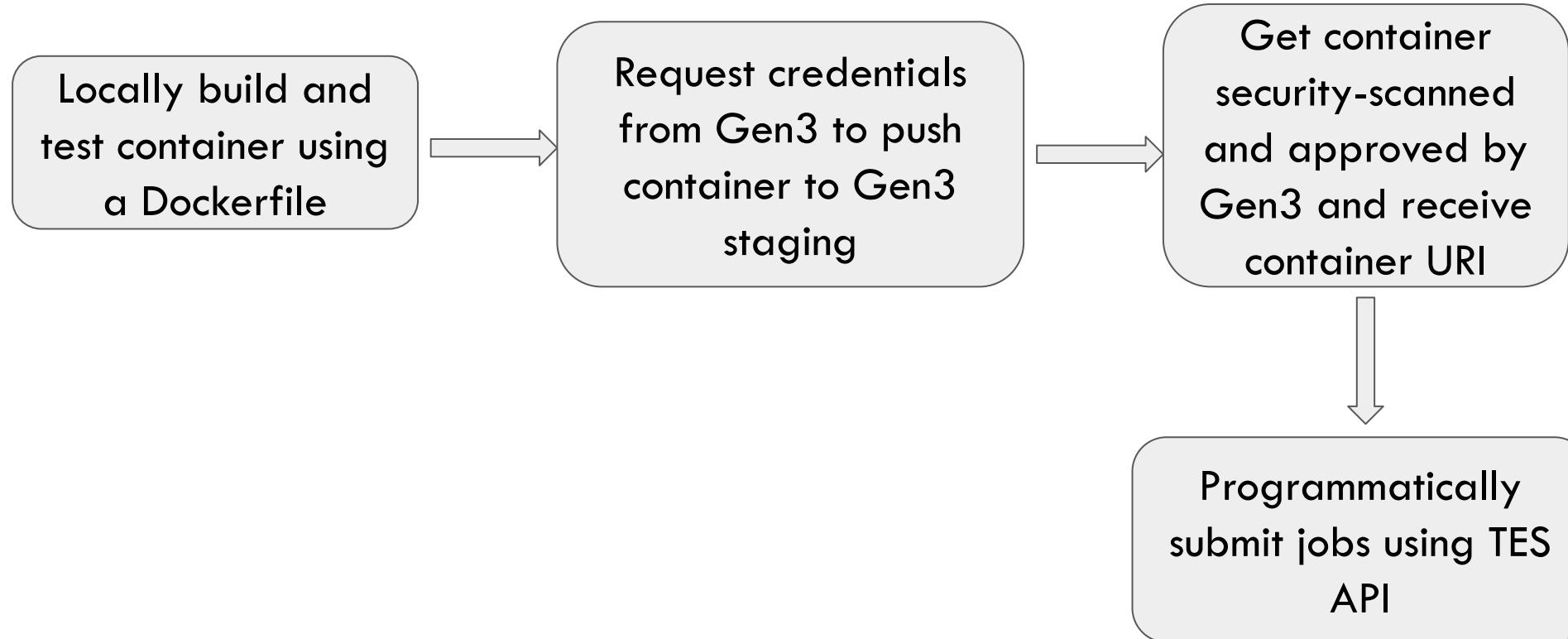
The GA4GH Cloud Workstream

Venkat S. Malladi , Microsoft Research and AI, Redmond, WA, 98052, USA

Kyle Elliott , Oregon Health and Science University, Portland, OR, 97239, USA

Overview of steps to run a user-developed workflow container (v2)

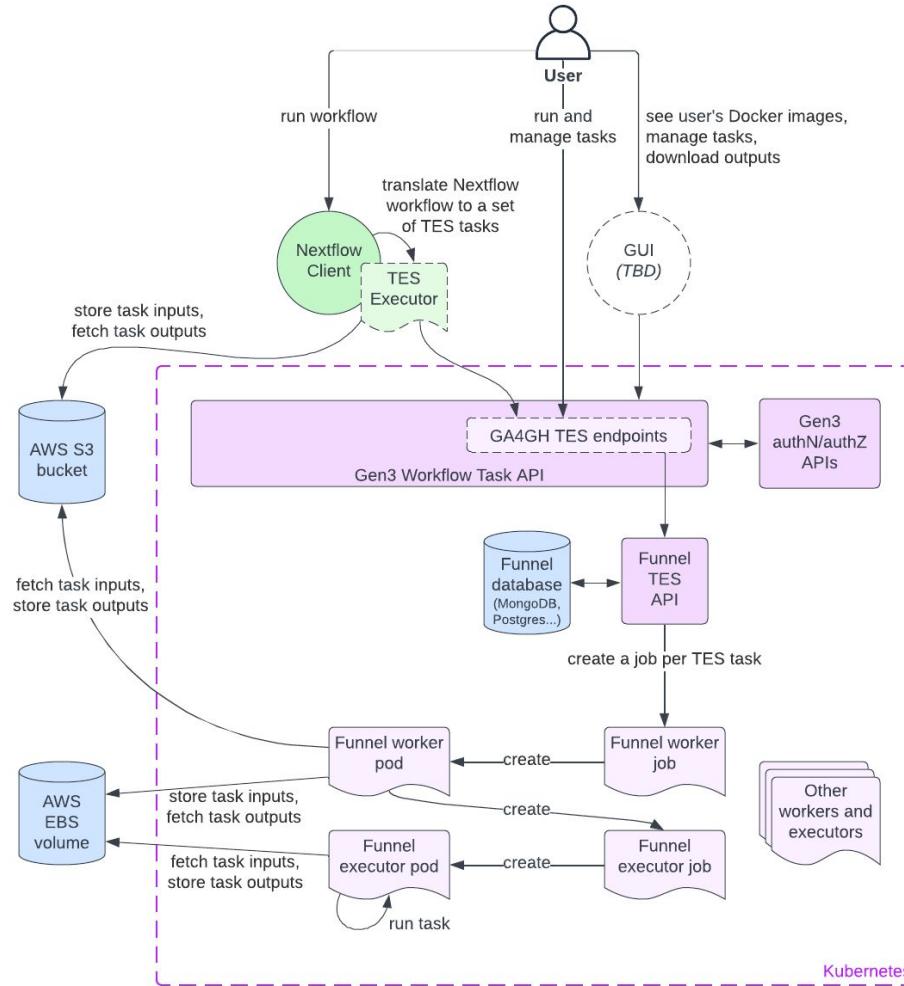
Containers are run in FedRAMP moderate environment with security compliance



Architecture

TES Backend: Funnel

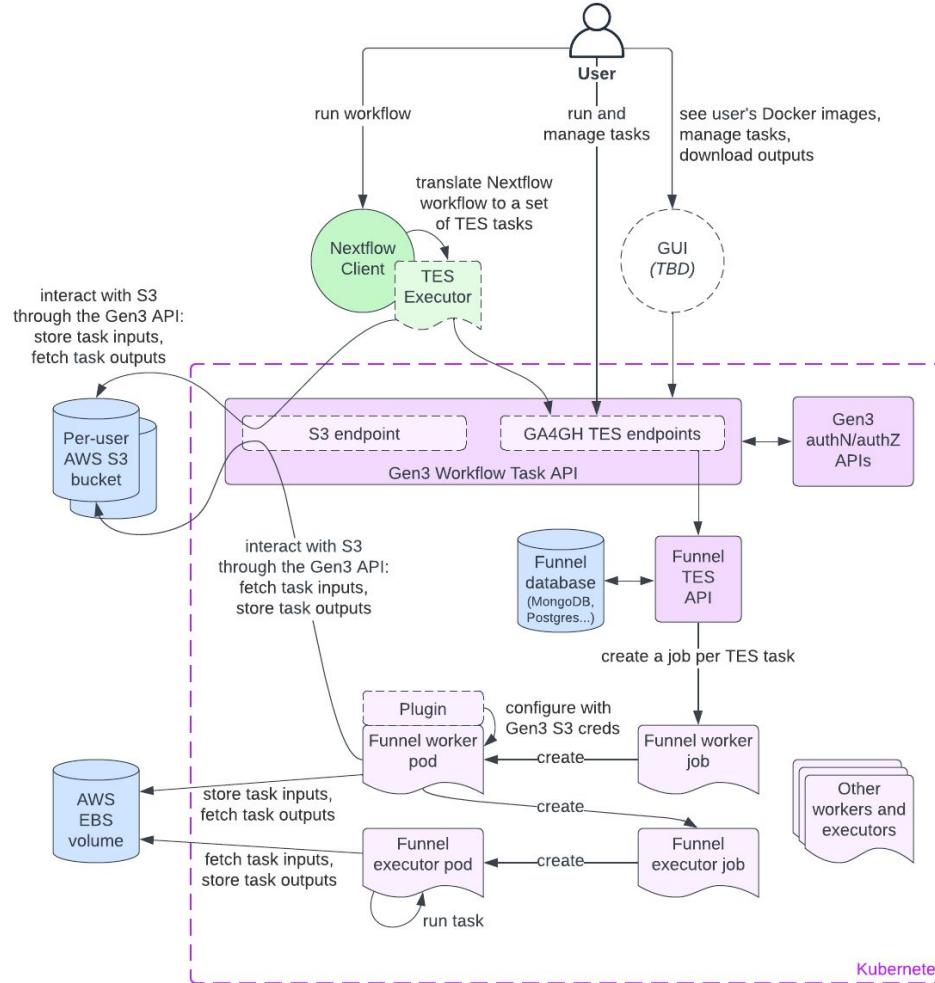
- Tool for distributed task execution
- Developed by the Oregon Health & Science University



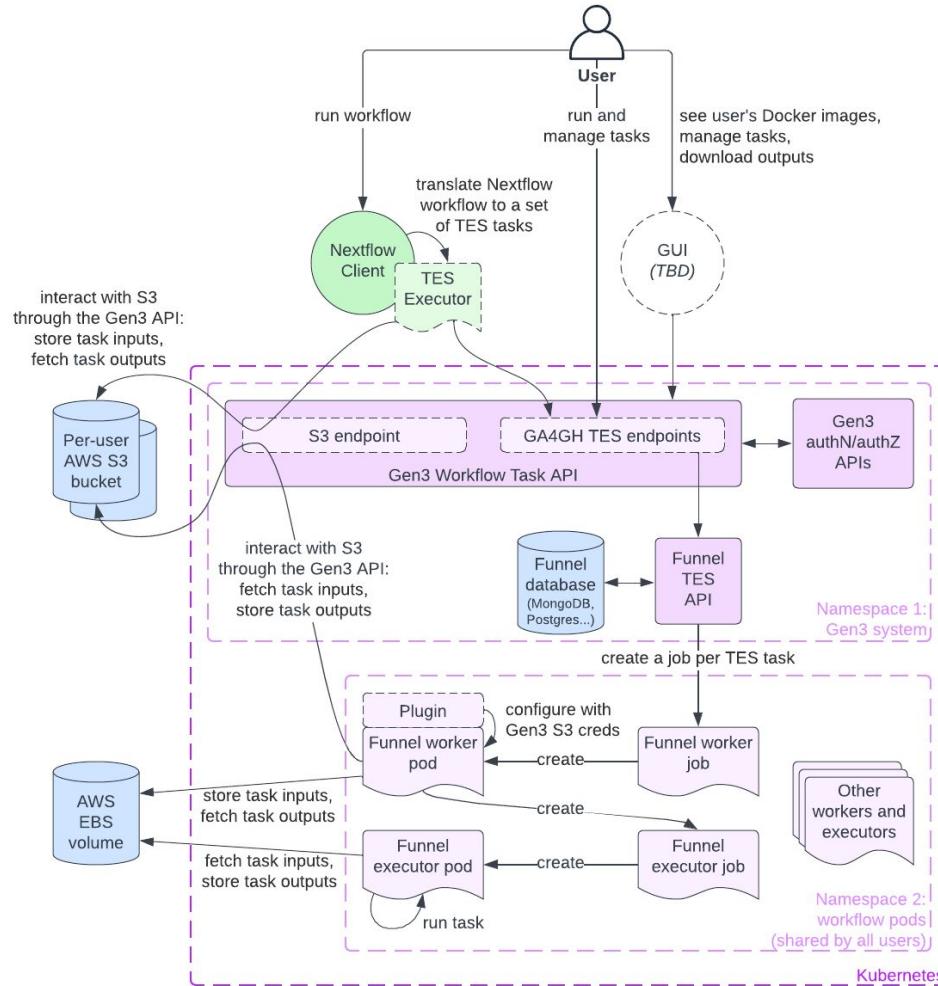
Architecture

Gen3 S3 endpoint

- The user and the Funnel worker authenticate with Gen3
- One S3 bucket per user
- Abstracts AWS credentials away from the user and the Funnel worker



Architecture



Nextflow and GA4GH TES demo

The screenshot displays a dual-pane interface for managing cloud storage and executing API requests.

Left Panel (AWS S3 Bucket View):

- URL:** us-east-1.console.aws.amazon.com/s3/buckets/gen3wf-pauline-planx-pla-net-16
- Bucket Name:** gen3wf-pauline-planx-pla-net-16
- Actions:** Objects, Metadata, Properties, Permissions, Metrics, Management, Access Points.
- Objects (0):** No objects present.
- Actions:** Create Folder, Upload (highlighted).
- Information:** Objects are fundamental entities stored in Amazon S3. You can use Amazon S3 Inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more.
- Search:** Find objects by prefix.
- Table Headers:** Name, Type, Last modified, Size, Storage class.
- Message:** No objects. You don't have any objects in this bucket.
- Buttons:** Upload.

Right Panel (Gen3 API Request):

- User:** bruno
- Environment:** No Environment
- Request:** GET https://pauline.planx-pla.net/workflows/storage/info
- Method:** GET
- Headers:** None
- Body:** None
- Auth:** None
- Vars:** None
- Script:** None
- Assert:** None
- Tests:** None
- Docs:** None
- Params:** None
- Icon:** A paper airplane icon.
- Buttons:** Send Request (Cmd + Enter), New Request (Cmd + B), Edit Environments (Cmd + E).

Summary and next steps

Current efforts

- Demo use cases for different types of workflows, inputs and outputs, including easy way to retrieve outputs from workflows
- Track compute costs per user

Using v2 researchers will be able to:

- Locally develop and test containers (e.g. using their terminal)
- Develop and perform a small-scale test of workflows that run containers using languages such as nextflow
- Scalable submission of workflows using the TES API to compute over 1000s of files in a parallelized/scalable manner

Toward AI-ready data commons: from computable data standards to interoperable AI models

Jing Su, Indiana University

Toward AI-ready data commons: from computable data standards to interoperable AI models

Jing Su

Associate Professor

Biostatistics and Health Data Science
Indiana University School of Medicine

Introduction

Biomed Info Lab
Graph AI



1. PI of Biomedical Informatics Lab
2. Director, Data Management Services team at Biostatistics and Health Data Science
3. Associate Director of Real-world Data, Biostatistics and Data Management Core at Indiana University Health Simon Comprehensive Cancer Center



The Data Management Services Team: AI-ready data, infrastructure, and implementation



Team: 30 members



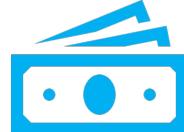
Grants: 66
Billables: 45



Publications: 90+



Pending Proposals: 46



Annual Budget: \$3.0M



Research Networks:
Global, National, & Regional



New norm of complex data in clinical studies

1. Modern clinical studies

- Real-world data: EMR, medical claims, etc.
- Clinical research data
- Multiomics data
- Medical imaging and pathological imaging
- Clinical notes
- Data from various sources/programs

2. ML/AI on complex and heterogeneous data



Oncology Research
Information Exchange
Network (ORIEN)



The Alcoholic Hepatitis Network
Integrated therapies for alcohol use
and ALD (ITAALD) Network



The Future of
Health Begins
With You

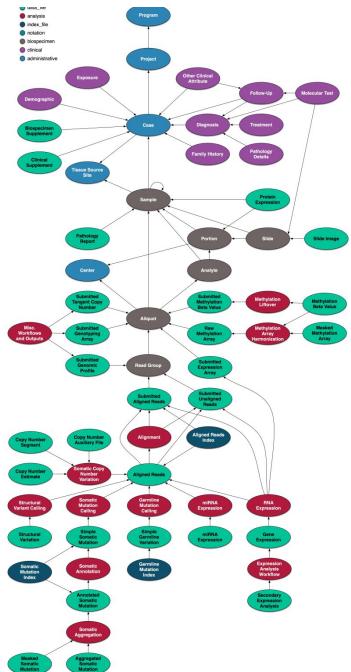


Graph data model Graph AI Research data commons



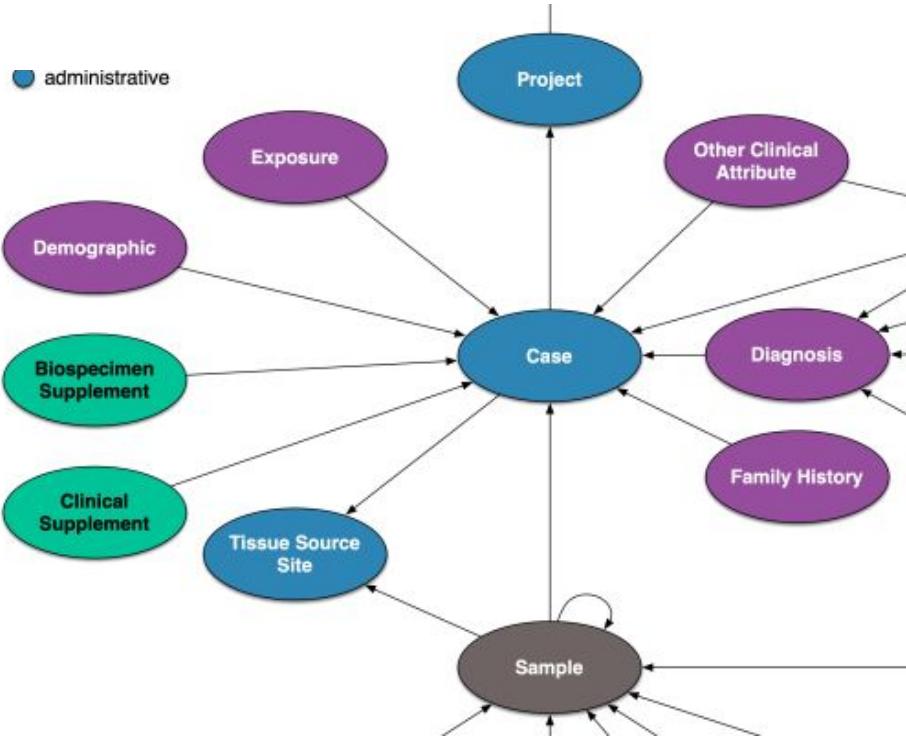
Graph data models: graphs to organize data artifacts

NCI Genomics Data Common (GDC) Data Model



Node types

- data_file
 - analysis
 - index_file
 - notation
 - biospecimen
 - clinical
 - administrative



Graph data modeling enables AI in research

Graph ontology and common data elements:

1. Define the integration and harmonization of real-world and research data
2. Enable new clinical study designs
3. Support AI/ML-readiness of biomedical big data
4. Widely used in real-world data infrastructures



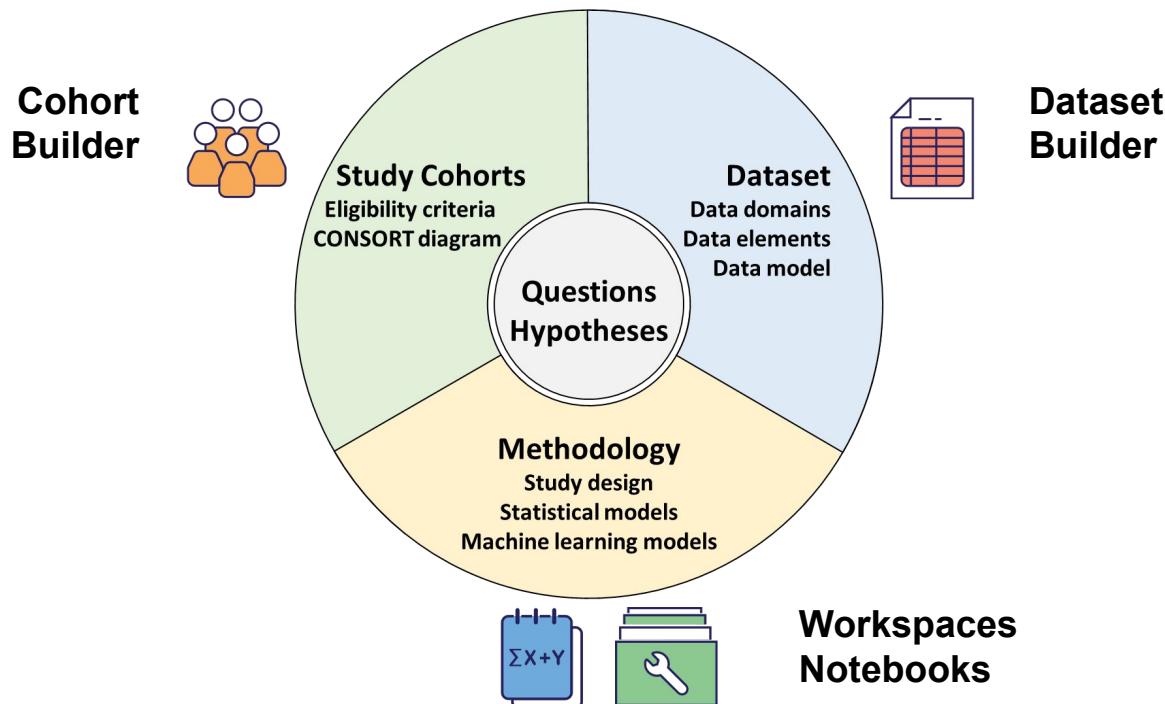
ARDaC: Alcohol Research Data Commons

an AI/ML-ready platform for clinical trials and observational studies

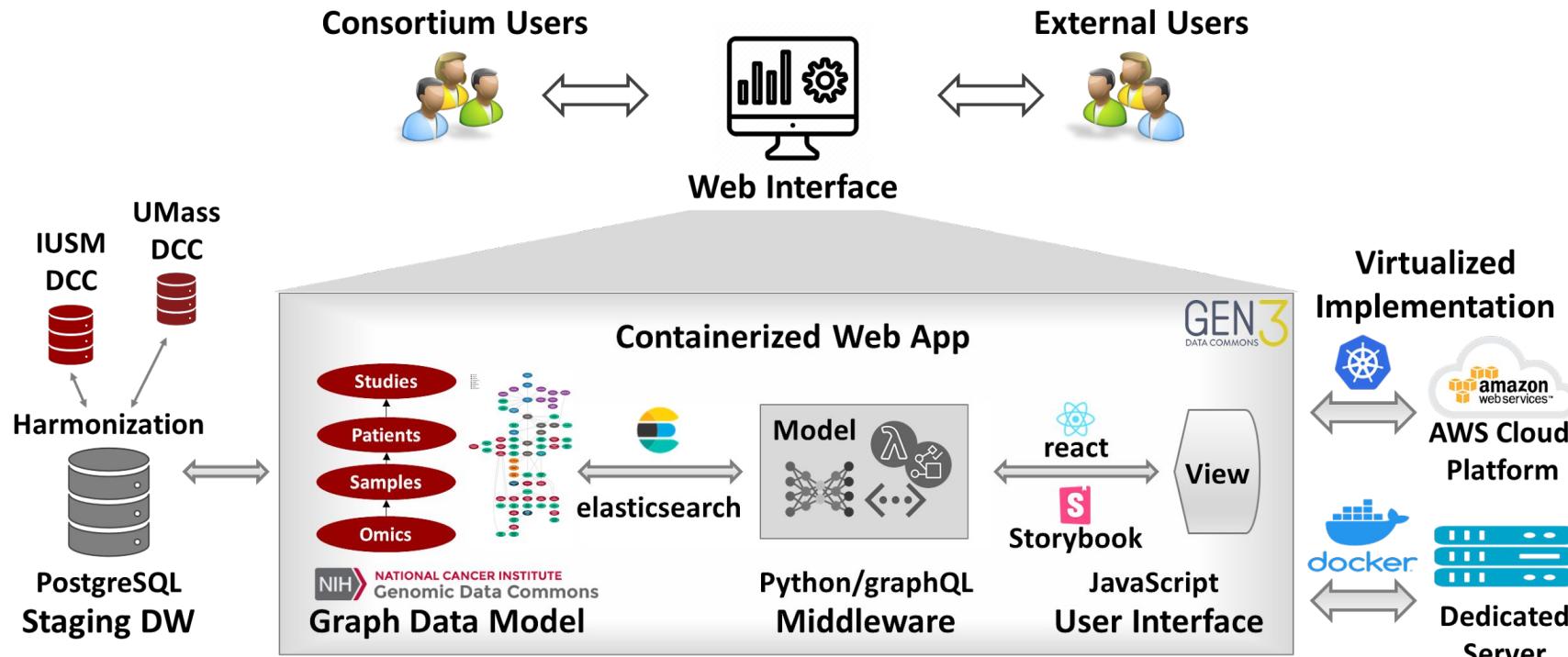
NIAAA: 2U24AA026969



RDC provides essential functionalities for clinical studies



ARDaC: Architecture



The AlcHepNet Research Data Commons

portal.ardac.org

Browse Data | Documentation | Login

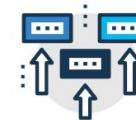
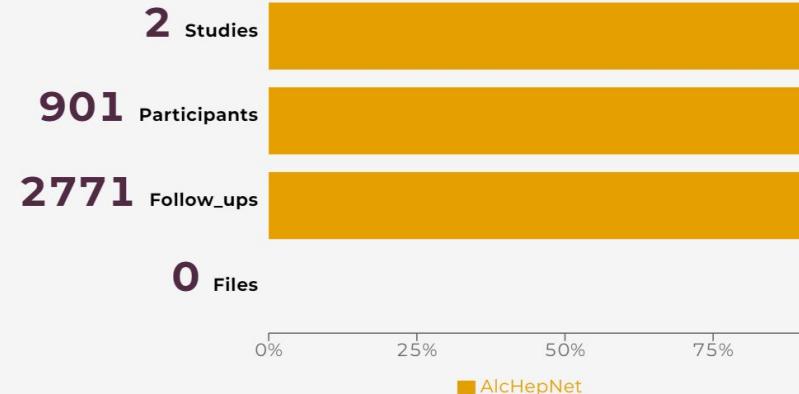
AlcHepNet

ARDaC: AlcHepNet Research Data Commons

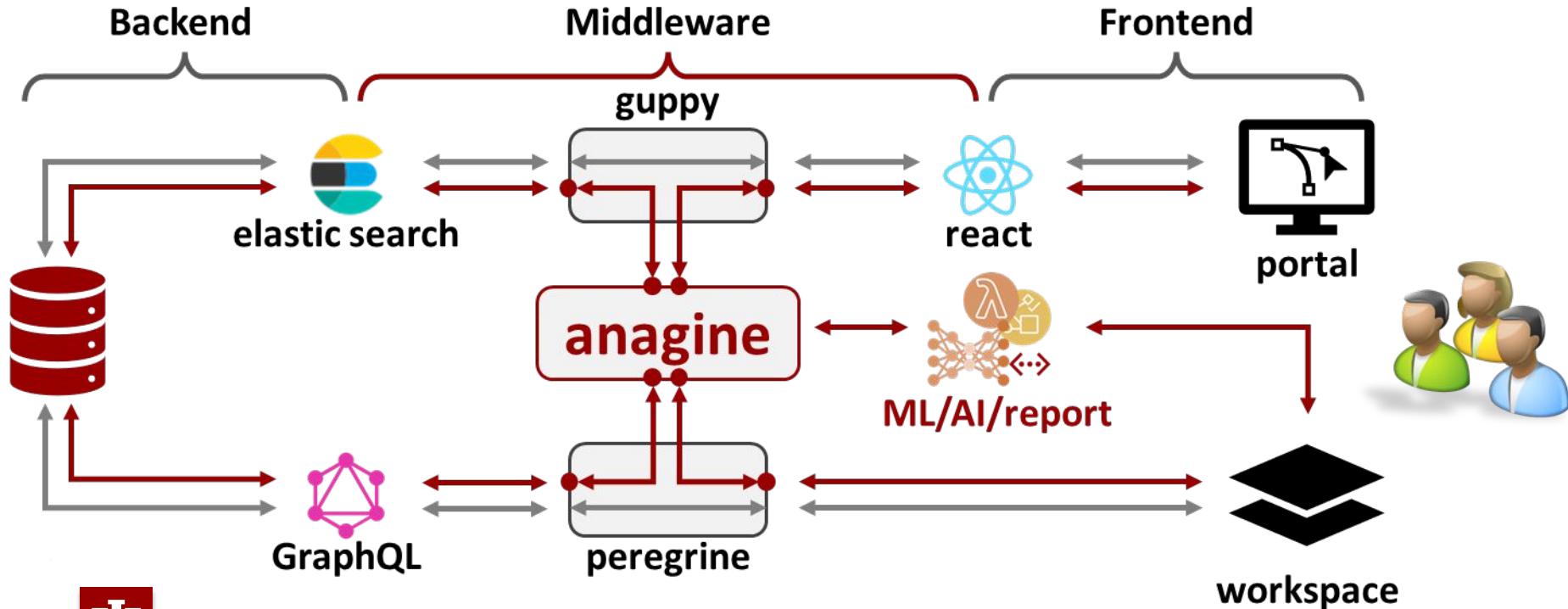
Dictionary Exploration Files Query Workspace Profile

AlcHepNet Research Data Commons

The AlcHepNet Research Data Commons, or ARDaC, is an integrative environment for exploring the clinical and omics data generated by AlcHepNet and related translational studies, and for investigators within the network to access data. ARDaC is sponsored by National Institute on Alcohol Abuse and Alcoholism.

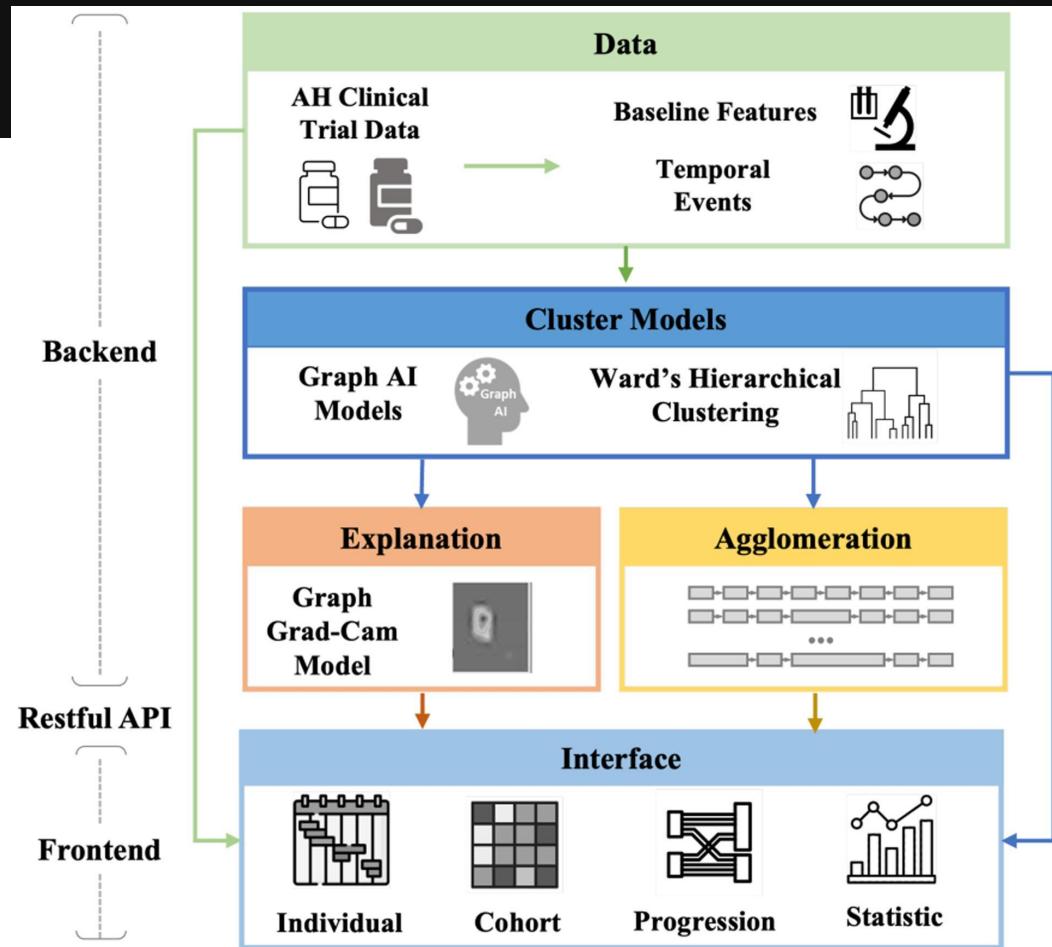


anagine: the AI/ML analysis engine of ARDaC

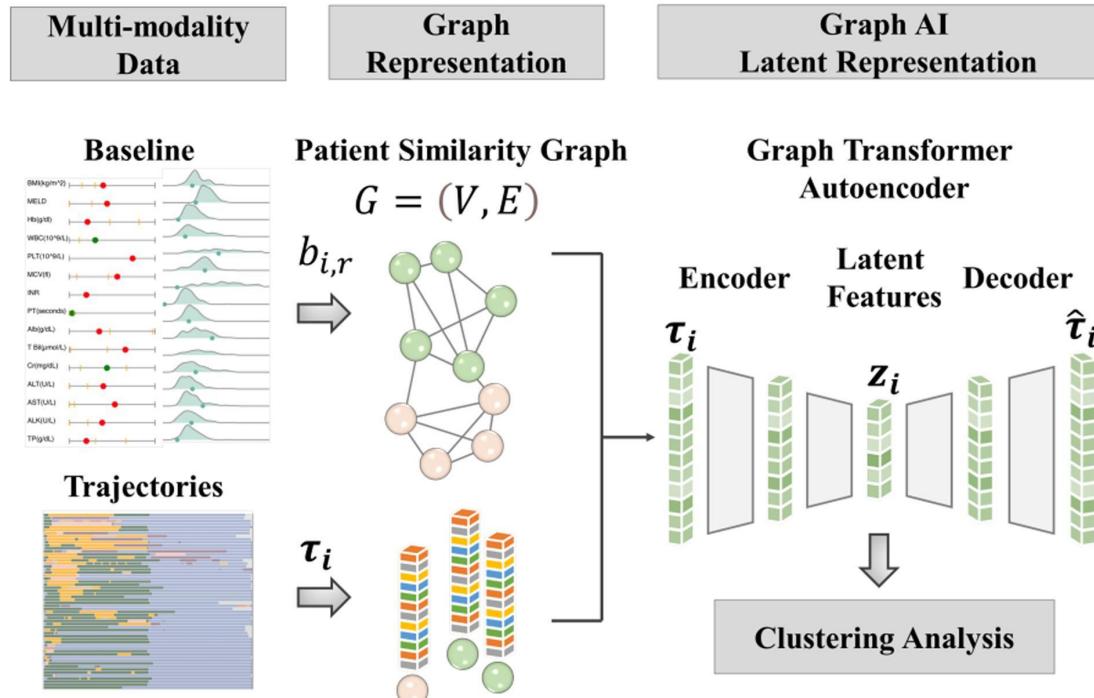


TrialView: ML/AI on ARDaC

Powered by ARDaC
graph data model



Trajectory learning with graph AI



Alcoholic Hepatitis Clinical Data Explorer

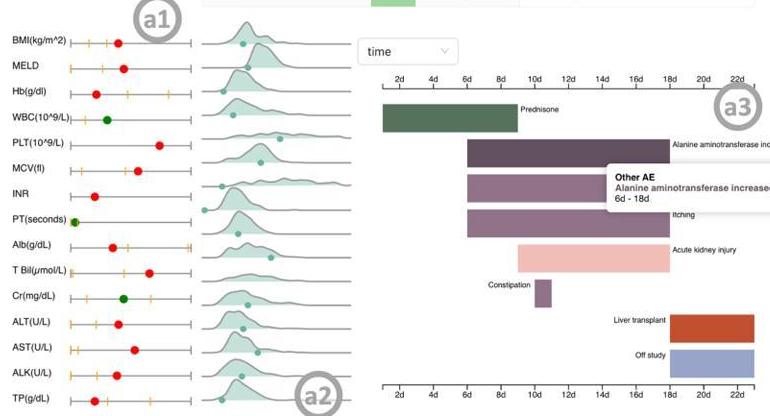
Individual View



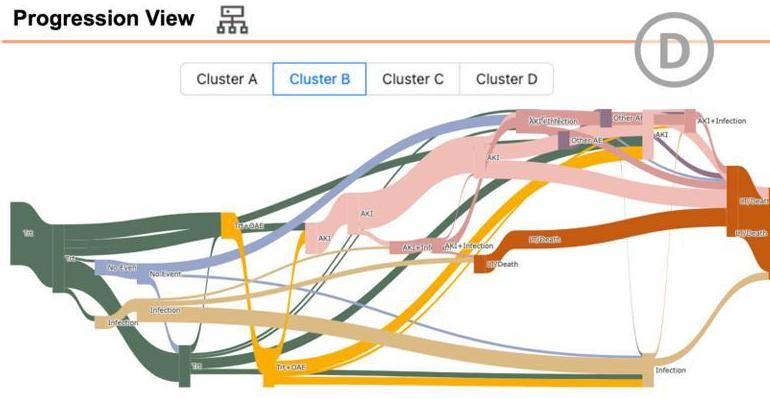
Select a case here

80036

Sex	Male	Race	White	# of Drinks in 30 days	126
# of drinking days out of 30 days	21	RCT Arm	Treatment A	Days on Treatment	8
Death	No	End Study Reason	Liver Transplant		



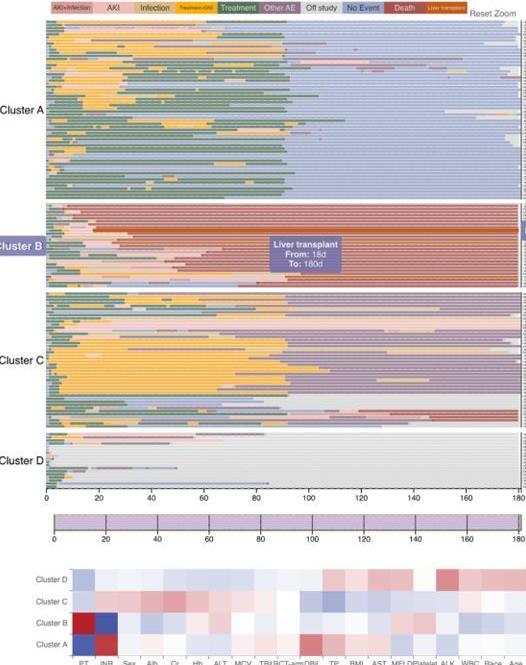
Progression View



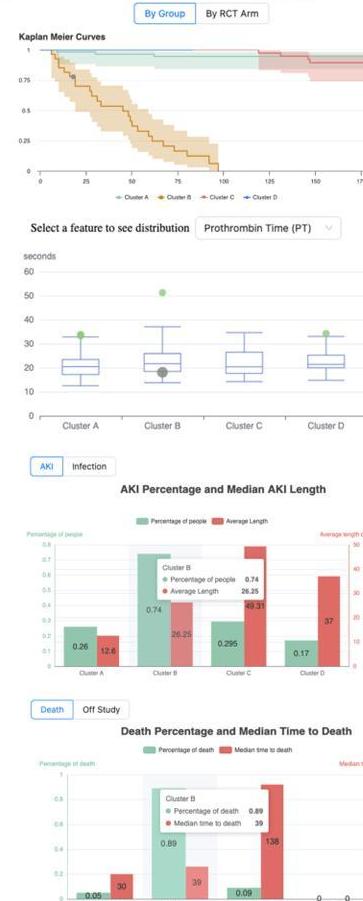
Cohort View



Cluster Method **Ward** Graph Transformer Layout Method **Original** Even



Statistics View





IPO: Indiana Precision Oncology Research Data Commons

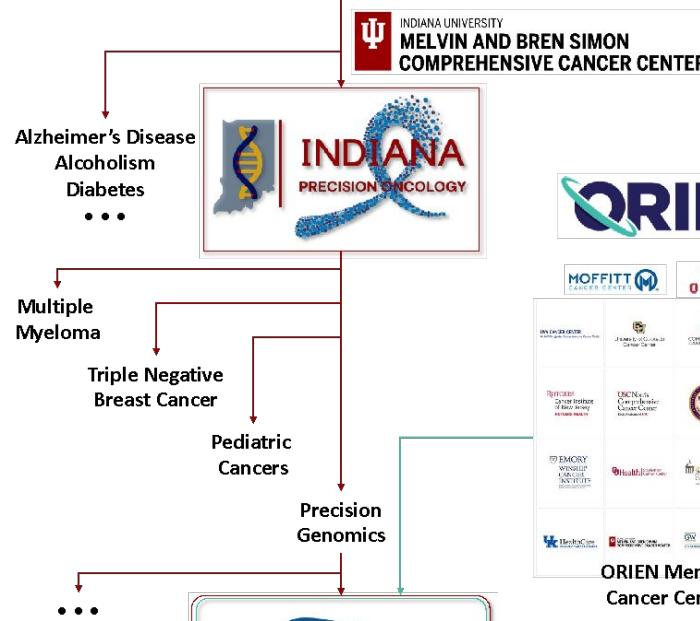


INDIANA UNIVERSITY SCHOOL OF MEDICINE

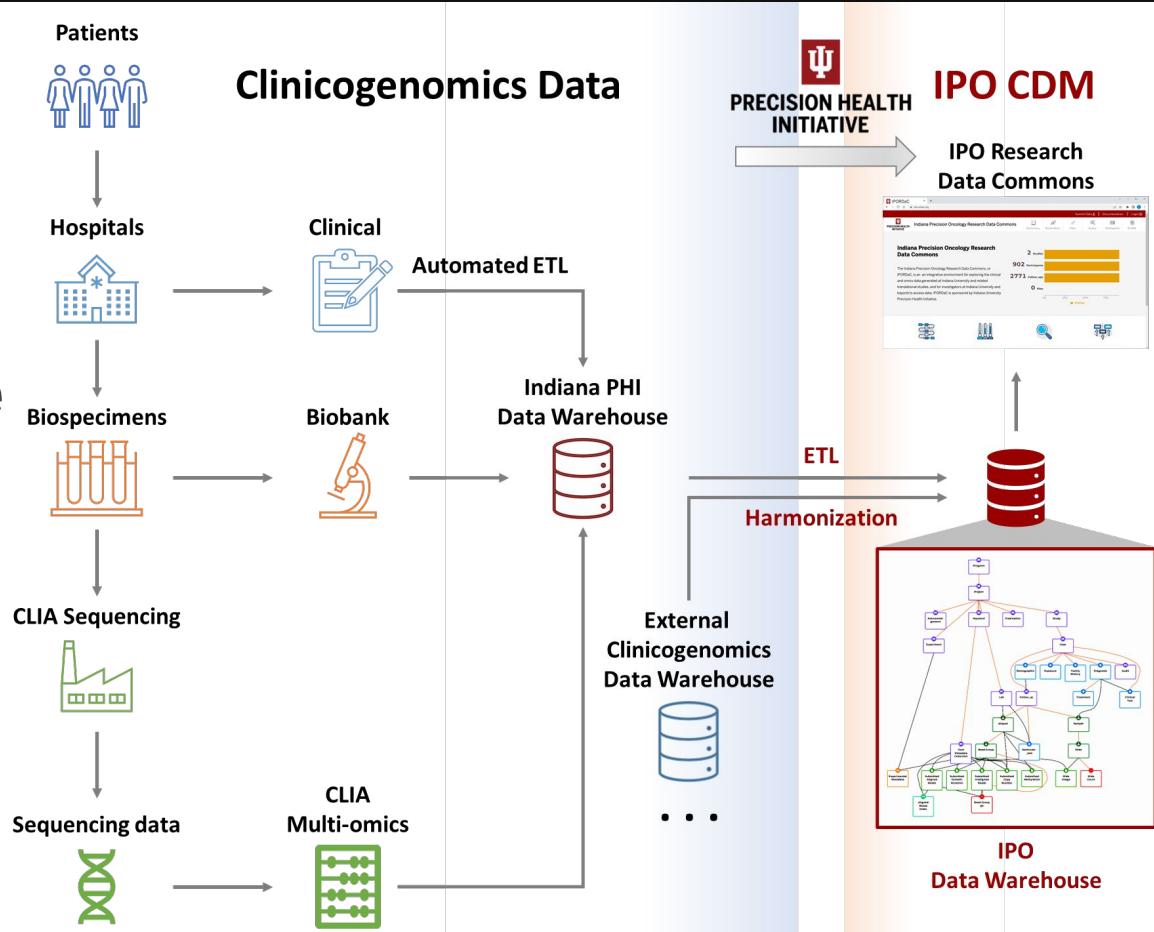
The clinicogenomics ecosystem

Indiana Precision Health Initiative

ORIEN: Oncology Research Information Exchange Network



IPO clinicogenomics data architecture



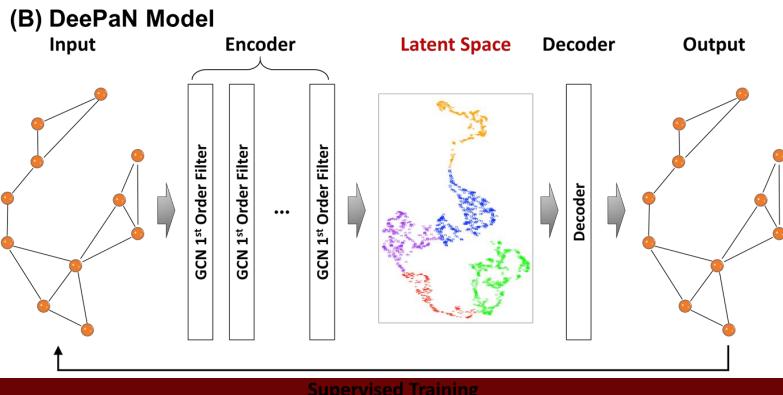
GAIPO: graph AI for precision oncology

NCI: 3P30CA082709-25S1

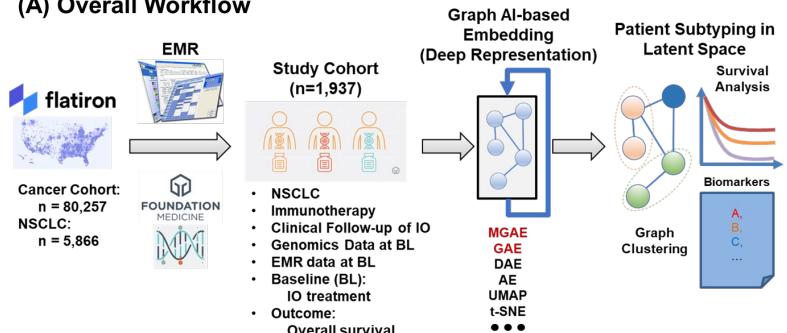


Graph AI models generated at IU

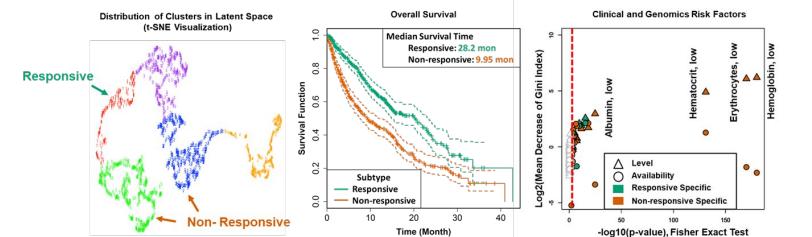
DeePaN, Su's lab
 Genomics and clinical data
<https://www.nature.com/articles/s41746-021-00381-z>



(A) Overall Workflow

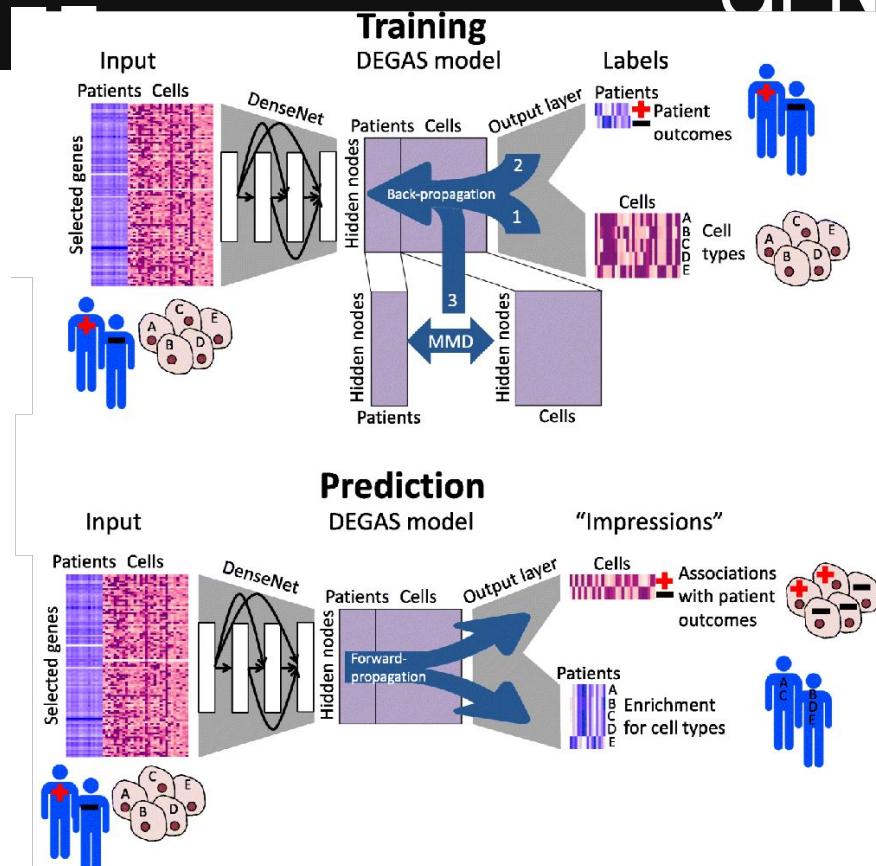


(C) Modeling Results and Interpretation



Graph AI models generated at IU

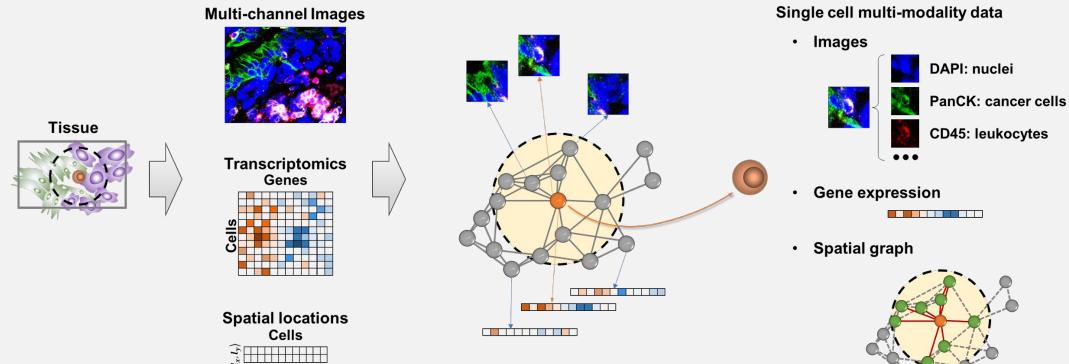
DEGAS, Huang's lab
Genomics and Clinical data
<https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-022-01012-2>



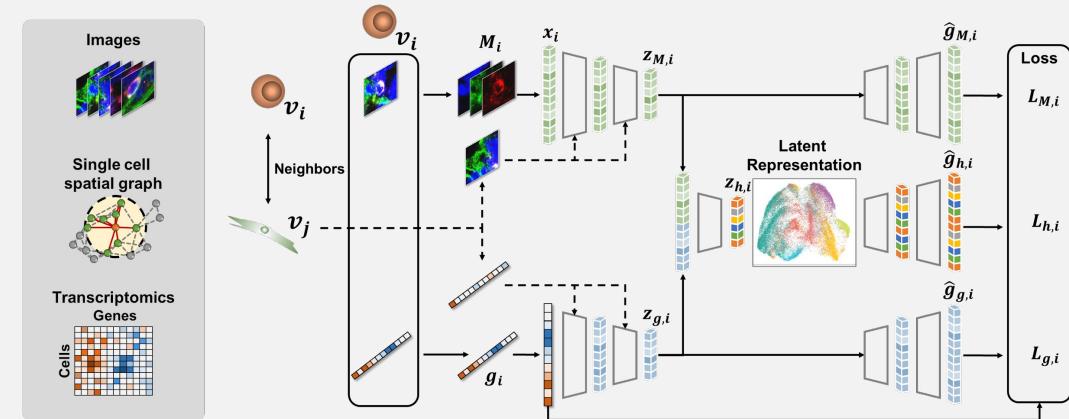
Graph AI models generated at IU

SiGra, Su's lab
 Single-cell spatial data
<https://www.nature.com/articles/s41467-023-41437-w>

A. Graph representation of multimodal single cell spatial transcriptomics



B. Single-cell spatial elucidation through image-augmented graph transformer (SiGra)

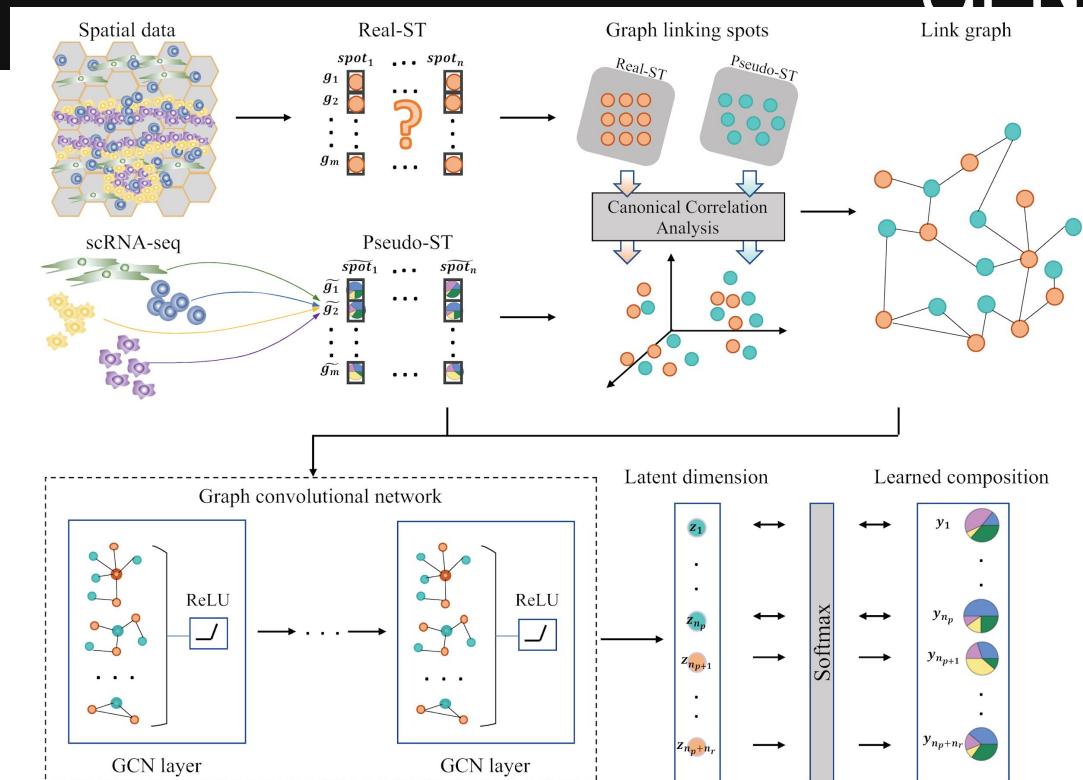


Graph AI models generated at IU

DSTG, Su's lab

Single-cell and spatial data

<https://academic.oup.com/bib/article/22/5/bbaa414/6105942>



Childhood Cancer Clinical Data Commons (C3DC)

← → ⌂ clinicalcommons.ccdi.cancer.gov/explore

This repository is under review for potential modification in compliance with Administration directives.

An official website of the United States government

NIH NATIONAL CANCER INSTITUTE
Childhood Cancer Clinical Data Commons

Home Explore Cohort Analyzer Studies Data Model Resources About

search C3DC Search

373 DIAGNOSES 13,597 PARTICIPANTS 21 STUDIES

Race

7,959 Participants

White

Sex at Birth

Sex at Birth	Count
Male	~7,500
Female	~6,500
Not Reported	~1,000
Unknown	~100

Diagnosis

1 Participants

8650/0 : Interstitial cel...

STUDY

DEMOGRAPHICS

Participant ID Search

UPLOAD PARTICIPANTS SET

SEX AT BIRTH

RACE

DIAGNOSIS

TREATMENT

TREATMENT RESPONSE

SURVIVAL

Anatomic Site

2,351 Participants

C71.9 : Posterior cranial...

Age at Diagnosis (years)

Age Group	Count
0 - 4	~4,000
5 - 9	~3,000
10 - 14	~2,500
15 - 19	~2,200
20 - 29	~500
> 29	~100

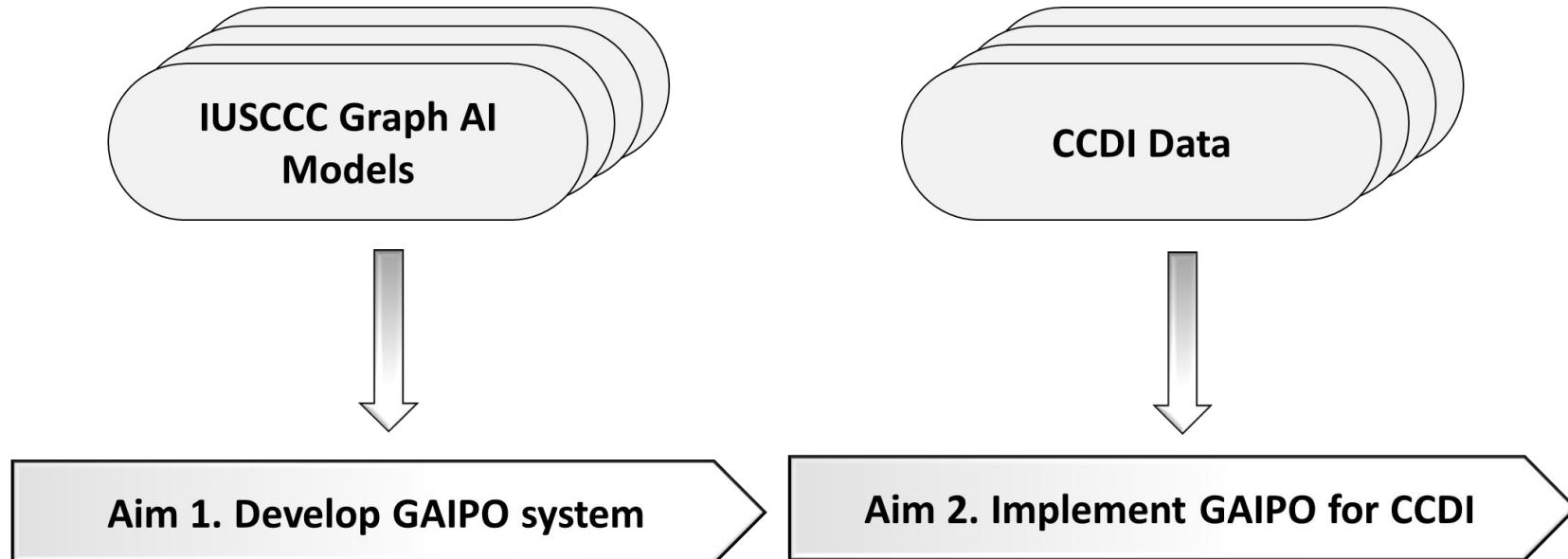
Treatment Type

2,317 Participants

Surgical Procedure

<https://clinicalcommons.ccdi.cancer.gov/>

Aims of GAIGO



CCDI's data sources and data elements

Cancer Types

- leukemias
- central nervous system neoplasms
- lymphomas
- neuroblastoma

Data sources

- National Childhood Cancer Registry
- CCDI Molecular Characterization Initiative
- CCDI OncoKids
- NCI CCSG CCDI Supplement Additional Genomic Submission
- Pediatric Cancer Knowledge Base (histological images)

Clinical data

- demographics
- pathological diagnosis
- treatments
- relapse and death

Omics data

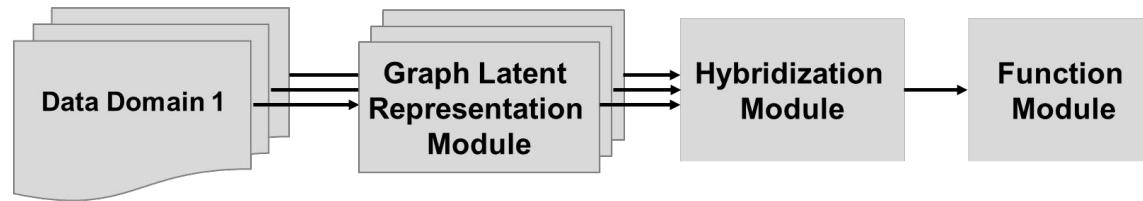
- RNA-seq
- whole genome sequencing
- whole exome sequencing
- emerging spatial omics data

Pathological imaging data

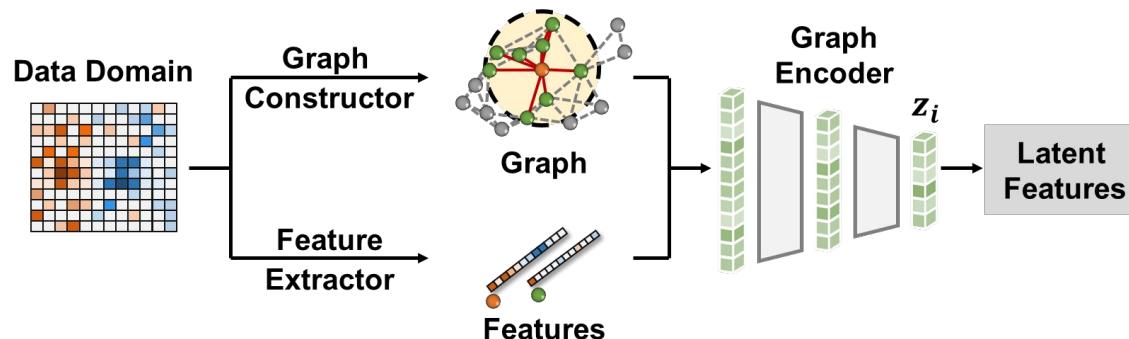
- Histological slide images



GAIPO's framework



Graph Latent Representation Module



GAIPO's functionalities

Data domains

- pathological images
- omics data
- clinical data

Graph constructor

- similarity
- adjacency
- mutual neighboring

Feature extractor

- VGG16
- highly variable genes
- availability
- dummy coding

Graph layer

- graph transformer
- GCN
- GAT
- GNN
- graphSAGE

Hybridization

- concatenation
- tensor

Function

- decoder
- classifier
- survival
- clustering



Conclusion: graph common data model and research data commons

1. The foundation of clinicogenomics research data commons
2. Enable AI/ML-ready collaborative research ecosystems
3. National data hubs across projects and institutes
4. Nexus of research and collaboration
5. Engine of novel research



The ARDaC Development Team

ARDaC Design



**Wanzhu Tu, PhD
PI**

Data Modeling & Management



**Carla Kettler
Data Management**

System Implementation



**Alan Walsh, PhD
Cloud Implementation**

System Development & Visualization



**Zuotian Li
Analytic
Visualization**



**Dr. Xiang Liu
Development**

Collaborations



**Qianqian Song, PhD
Assistant Professor
Health Outcome and
Biomedical Informatics
University of Florida**



**Robert Grossman, Ph.D.
Professor of Medicine
and Computer Science
University of Chicago**



**Michael Fitzsimons, Ph.D.
Director of Research
Programs and Scientific
Outreach
University of Chicago**

The GAIPO Development Team



Kelvin Lee, MD
PI



Kun Huang
PhD



Karen Pollok
PhD



Zanyu Shi
Development



Xiang Liu, PhD
Development



Alan Walsh, PhD
Cloud Implementation



Jing Su, PhD
Director



Waqas Amin
PhD



**Netsanet
Gebregziabher**
Data management



Nanxin Jin
Data models



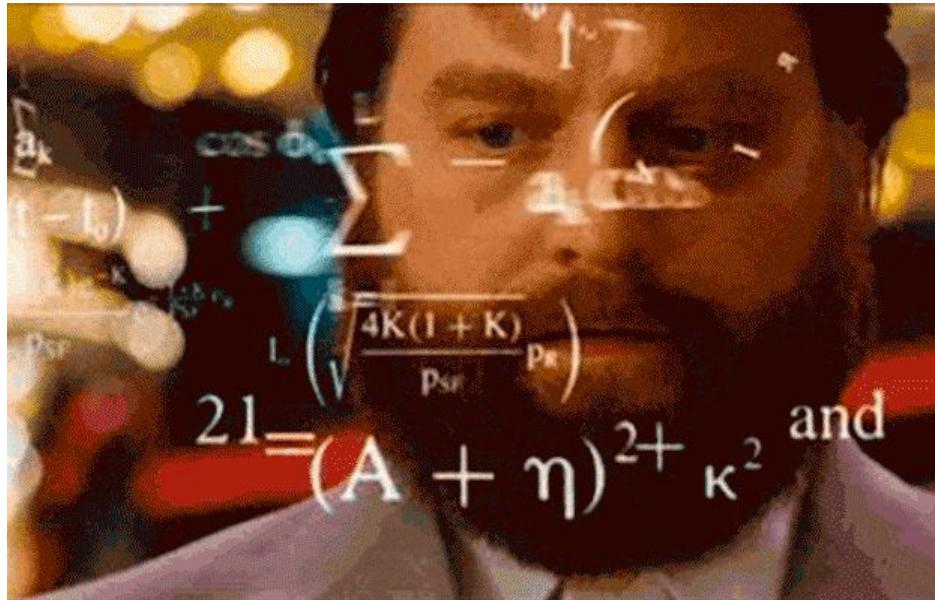
Hao Wang
**Analytic
Visualization**



George Nitsos
IT & Security

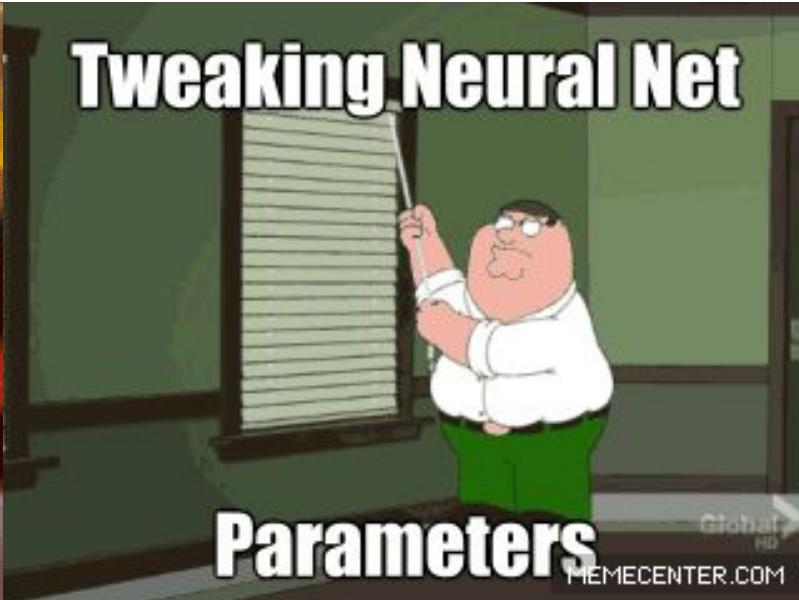
Questions?

Design ...



Tuning ...

Tweaking Neural Net



Parameters

HUMECENTER.COM

Acknowledgements



- **Speakers**
 - Kyle Burton - Center for Translational Data Science, University of Chicago
 - Rance Nault - Michigan State University
 - Aarti Venkat and Pauline Ribeyre - Center for Translational Data Science, University of Chicago
 - Jing Su - Indiana University
- **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