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:

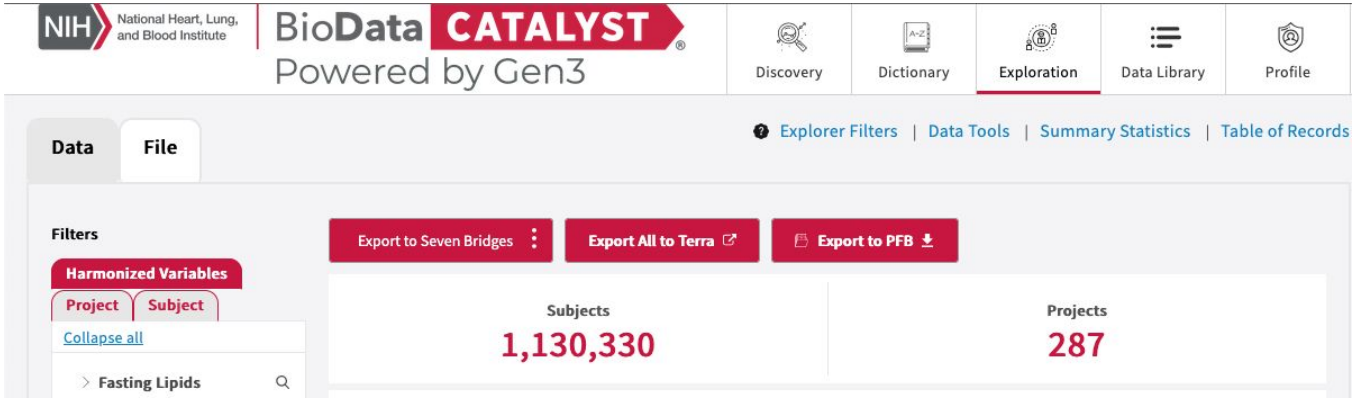
1. Schema – Describes the properties in a JSON Data dictionary.
2. Metadata – Explains the links between nodes for each of the properties
3. Data – Values for the properties

PFB in BDC - Gen3 Exploration

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



PFB in BDC - Past Observations

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.

PFB in BDC - Whole Study PFBs

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

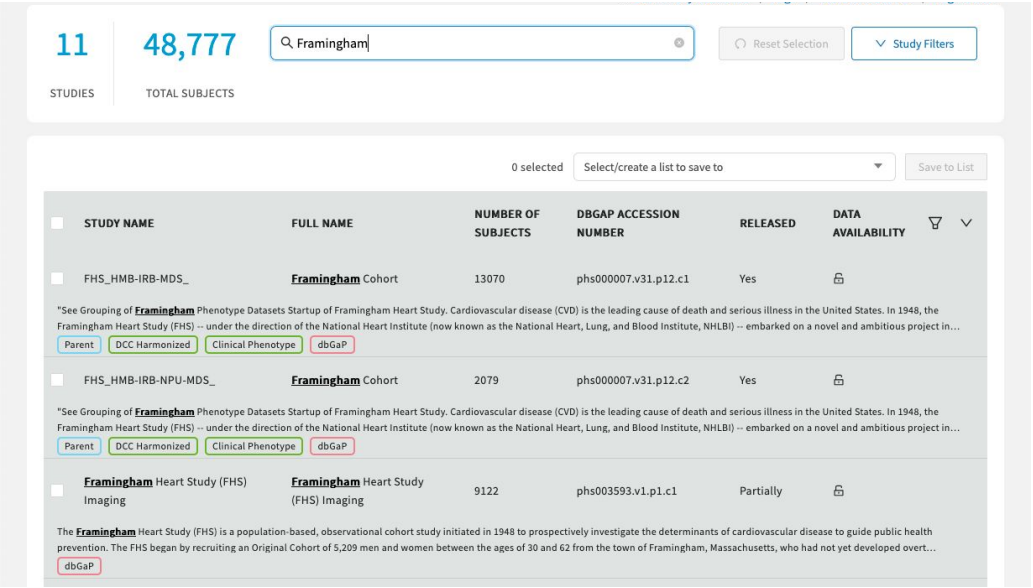
Whole-Study PFBs are available for transfer immediately.

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.



Gen3 User Data Library - BDC

Allows users to create lists of datasets.

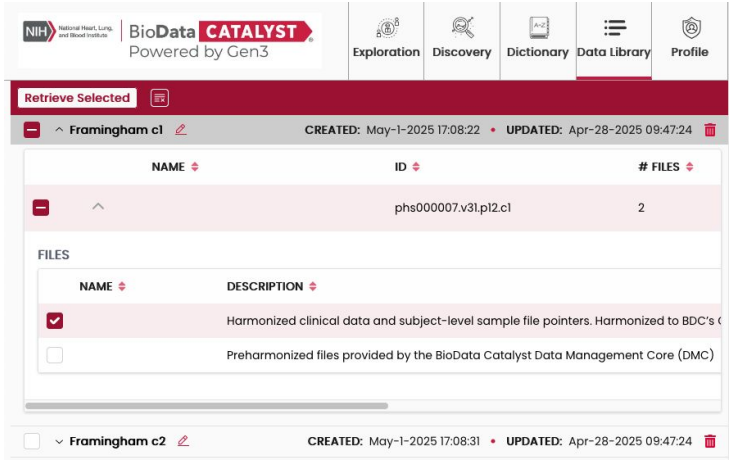
- Lists contain whole study PFBs from Discovery

Lists are persisted across sessions.

- Researchers can always reference the datasets used in analysis

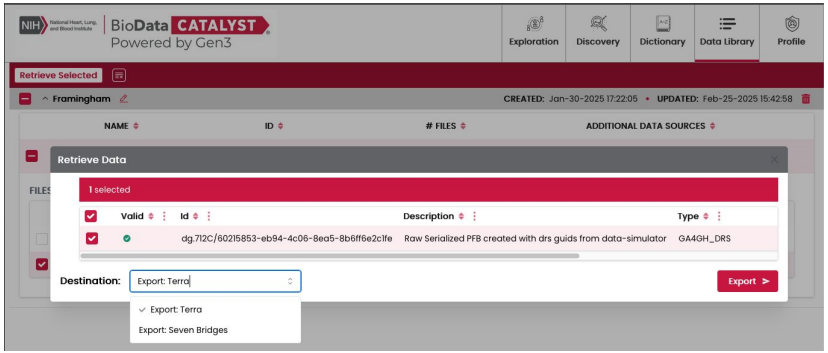
List items can be handed off to analysis platforms

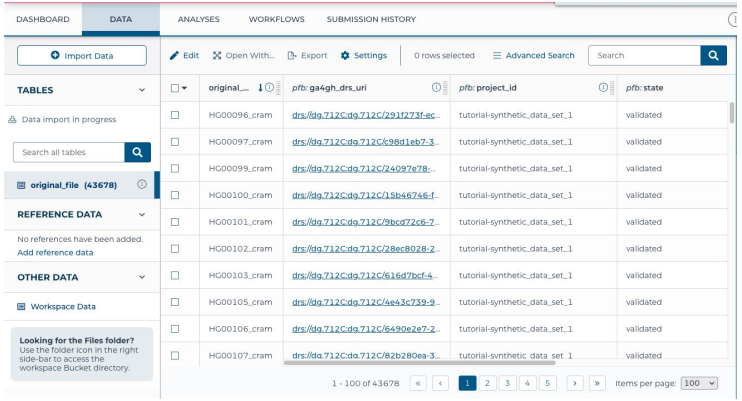
- In BDC, these are Broad's Terra and Velsera's Seven Bridges.
- In Gen 3.2, support for export to PFB is expected in June.
- Extendable to any analysis platform or workspace that supports PFB parsing.



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.





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

Links and Documentation

- 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

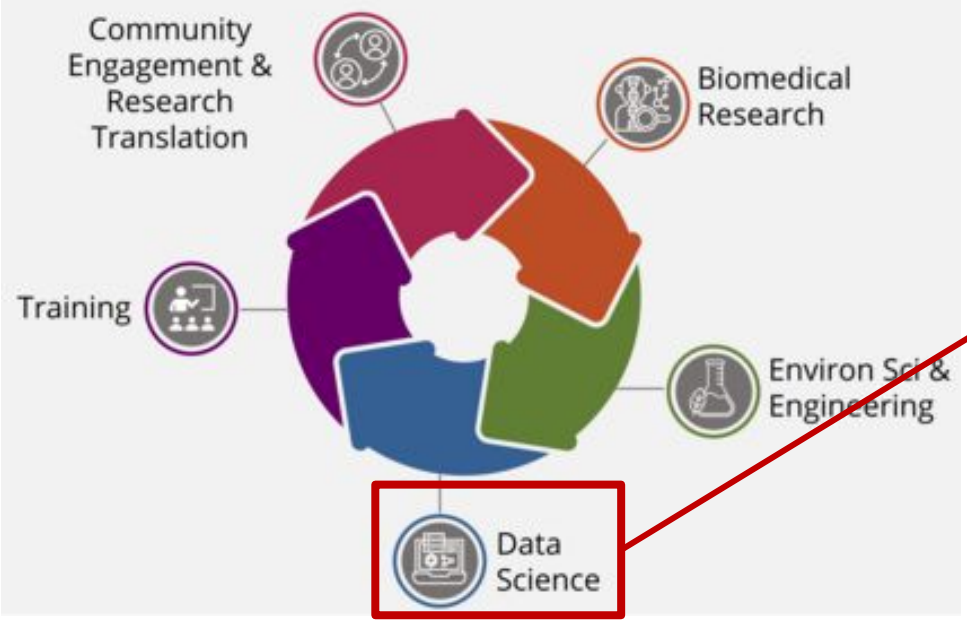
MSU SUPERFUND RESEARCH CENTER

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/

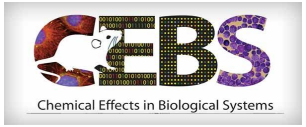
MSU SUPERFUND RESEARCH CENTER



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.

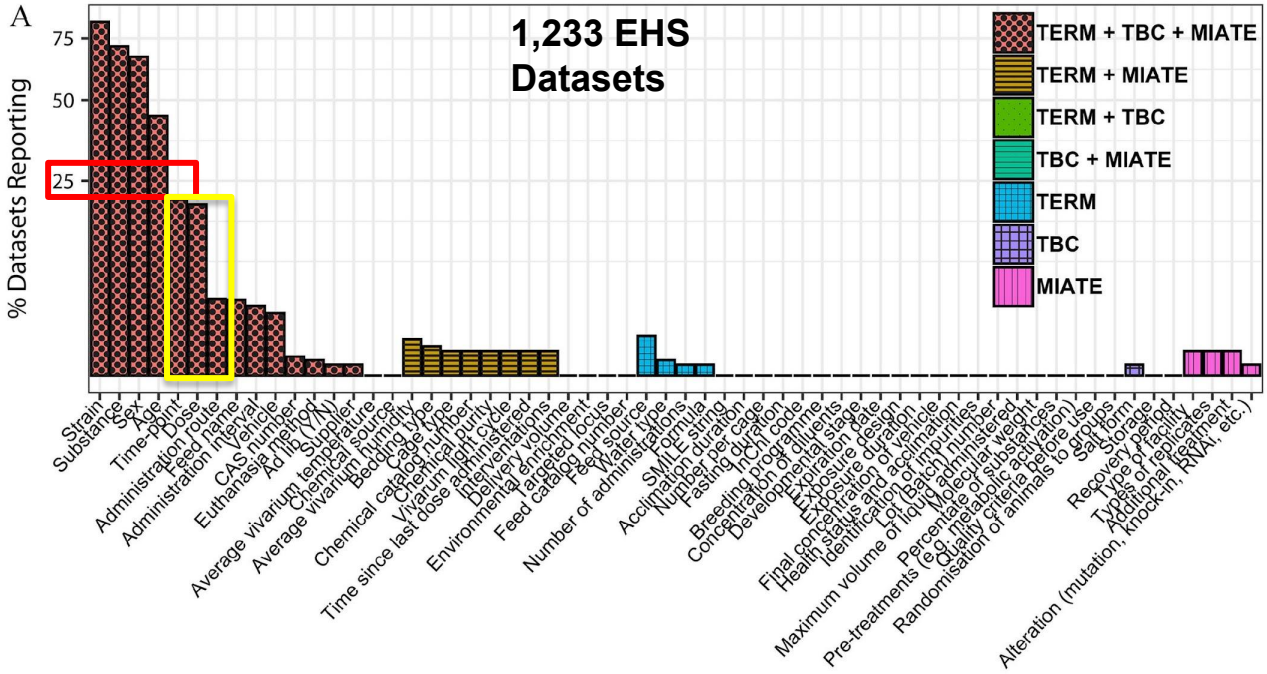
STATUS OF EHS DATA SHARING

Thousands of datasets are available through public databases and repositories



How many can be considered FAIR?

STATUS OF EHS DATA SHARING



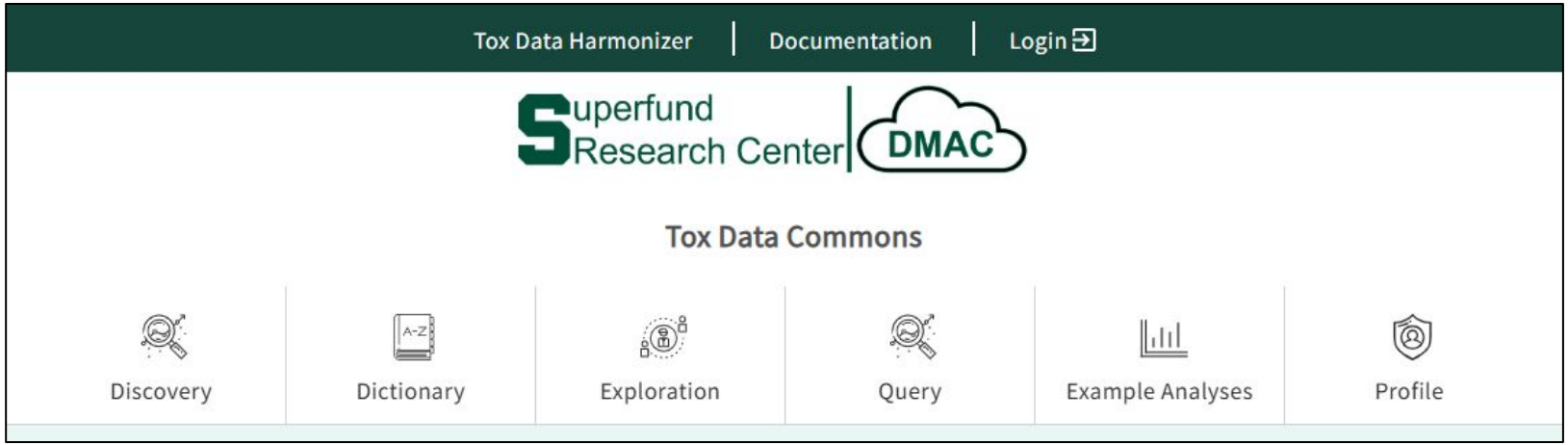
Beyond challenges in finding relevant data, many are missing key information.

This makes use of AI/ML approaches challenging.

Nault, R., *et al.* (2023). *Environmental health perspectives*. https://doi.org/10.1289/EHP11484

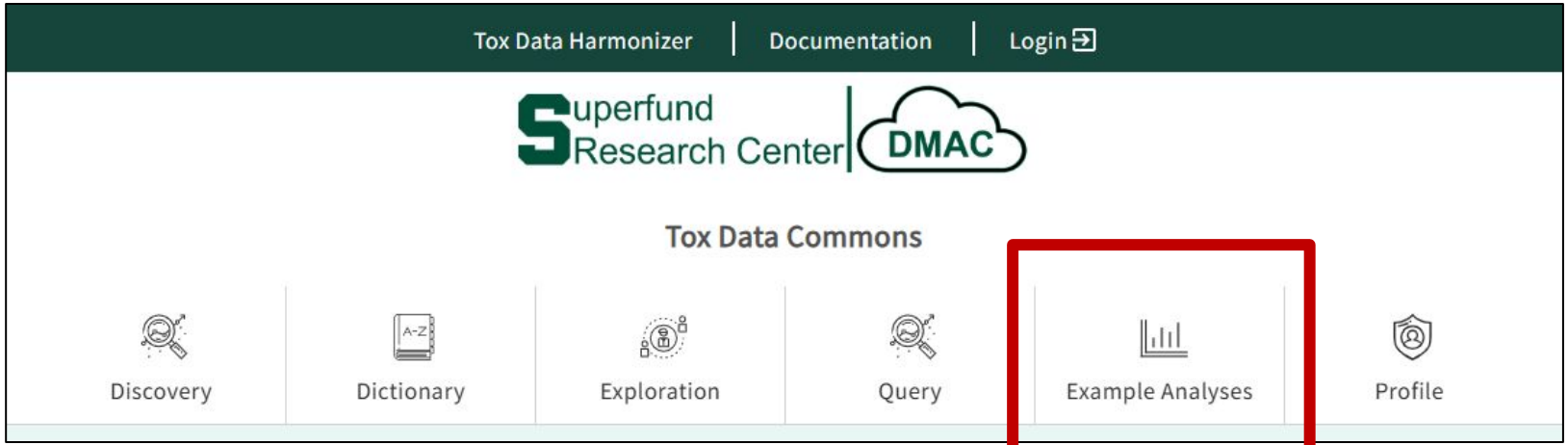
DEVELOPMENT OF TOXDATACOMMONS

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



DEVELOPMENT OF TOXDATACOMMONS

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

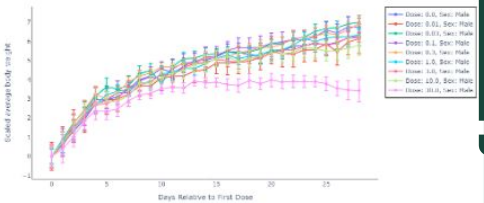


COLLABORATIVE RESEARCH APPLICATIONS

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

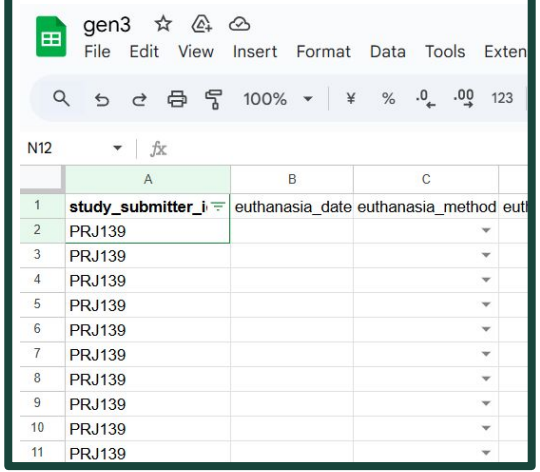
ICER

TDC Metabolism-based graph neural network analysis

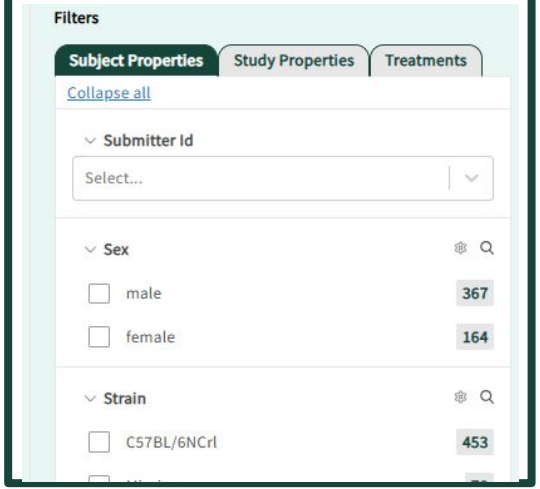
Not public yet

STANDARDIZED WORKFLOW

Template based submissions



Study-based Data Selection



Batch dataset Downloads

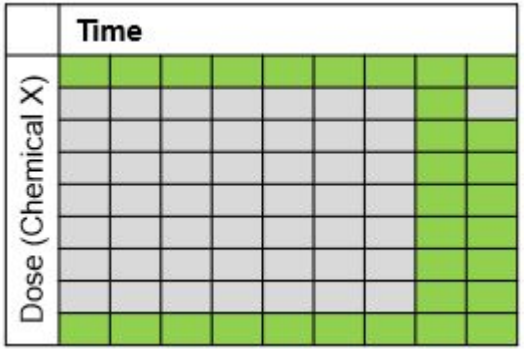
```
> gen3-client
download-multiple \
--profile=ToxDC \
--manifest=manifest.json \
--download-path=downloads

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

METADATA

CASE 1: INTEGRATING GROSS PATHOLOGY

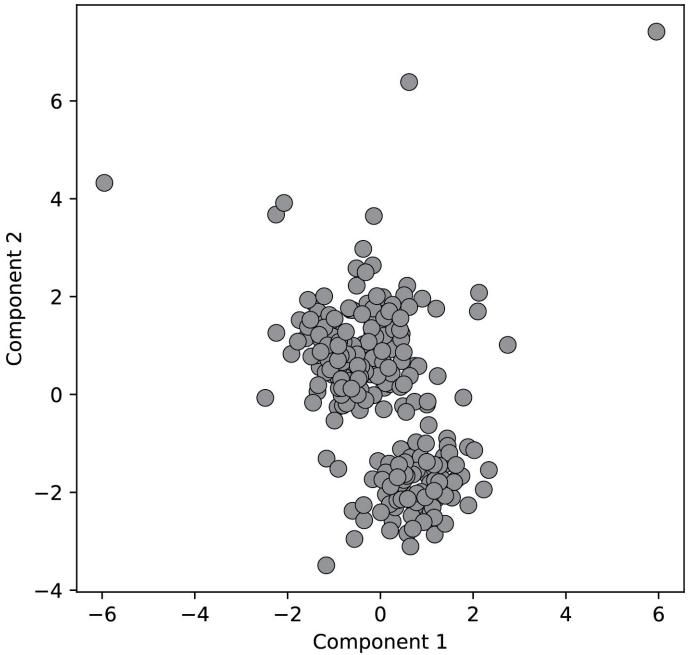
Datasets were generated in at least 3 independent studies with several measured endpoints (weights, pathology, gene expression, …).

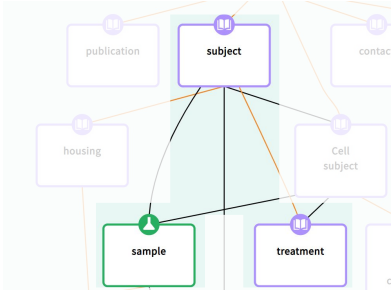


Can we identify informative patterns about chemical toxicity by integrating these studies?

CASE 1: INTEGRATING GROSS PATHOLOGY

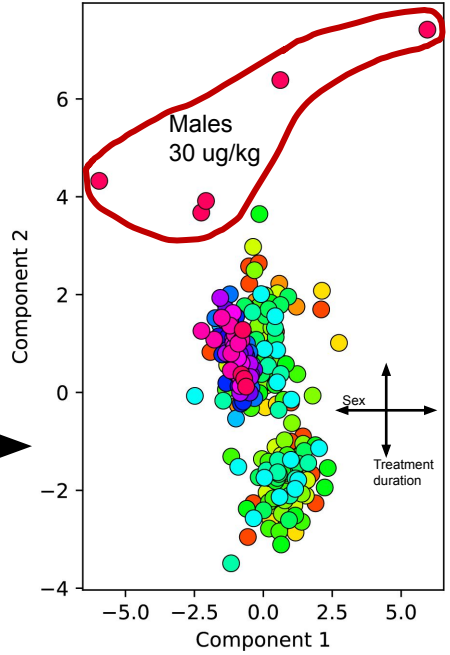
Dynamic time warping analysis of daily body weights of each *subject*.





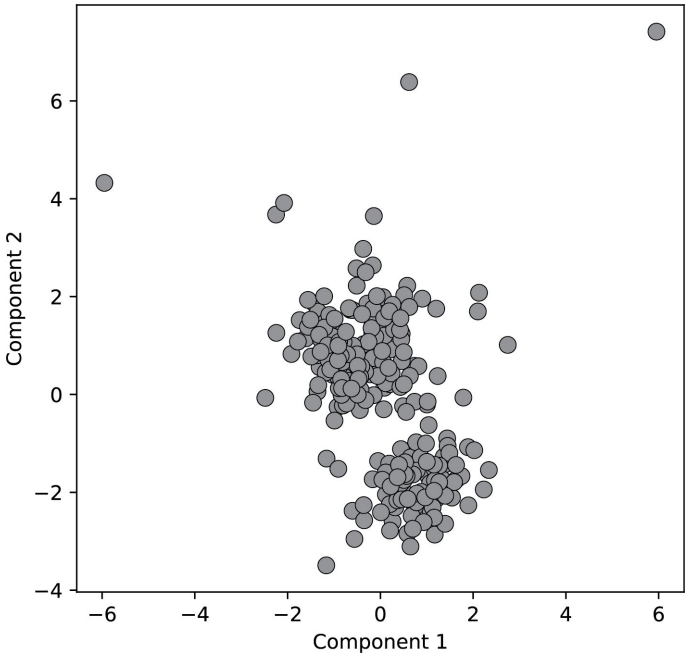
Sex
Dose
…

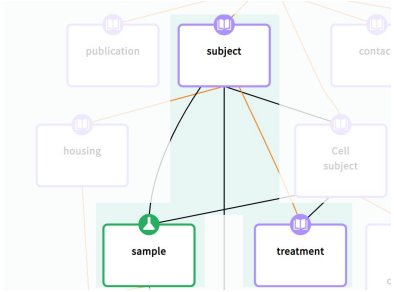
Metadata annotation reveals clustering



CASE 1: INTEGRATING GROSS PATHOLOGY

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

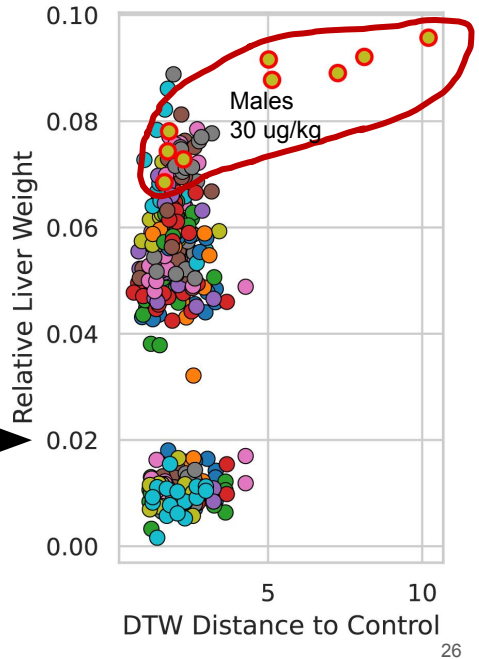




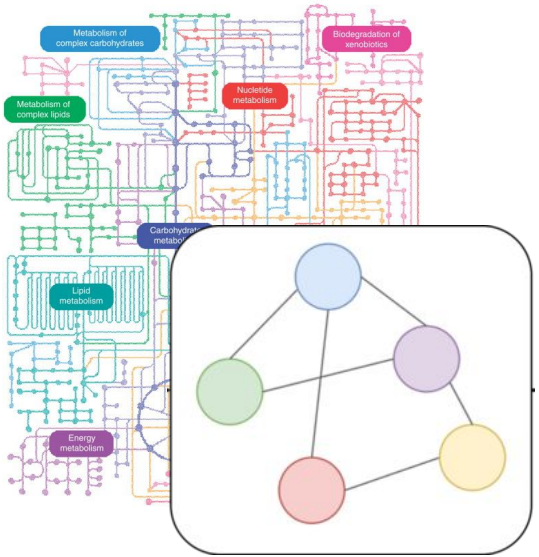
Sex
Dose
…

Relative liver weight (*sample*)

Metadata annotation reveals clustering



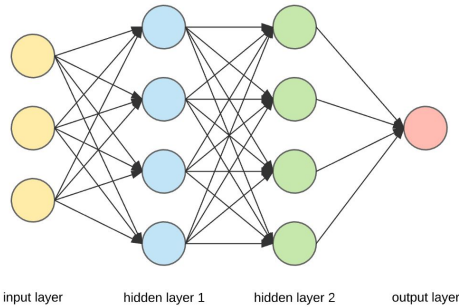
CASE 2: METABOLIC REACTION GNN



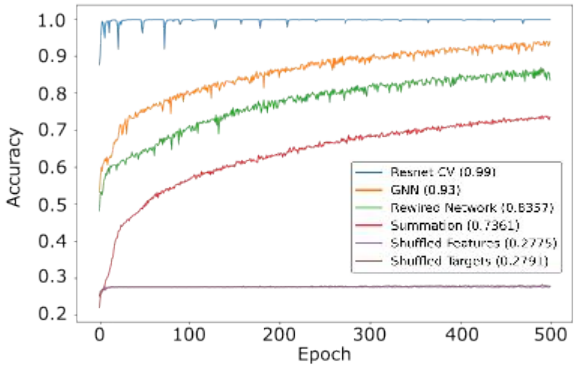
Metabolism is well suited for graph neural network approaches

recount3

GEO
Gene Expression Omnibus



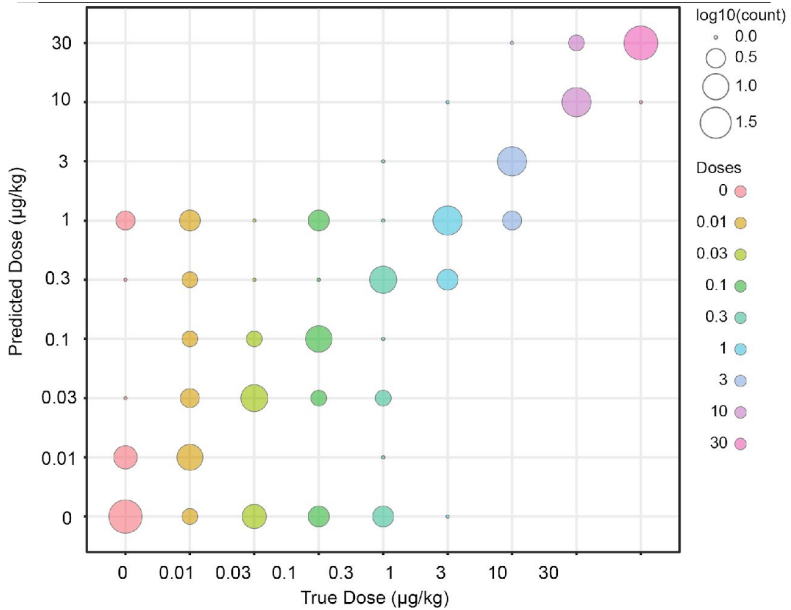
Only ~4% of samples were used, a significant portion excluded due to missing metadata



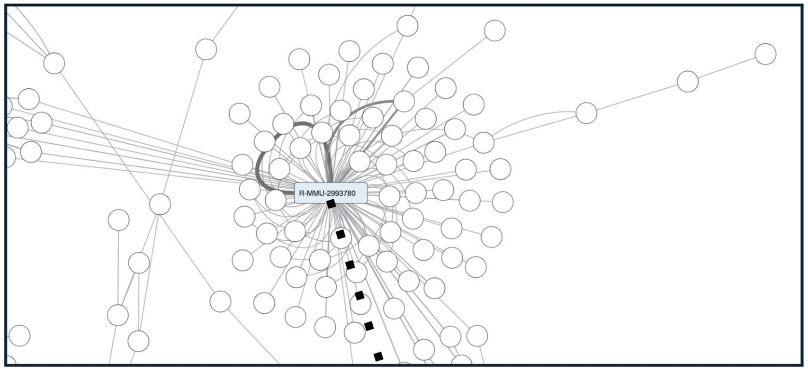
GNN training strategy originally described in: Burkhart, J. G., et al., (2023). https://doi.org/10.1016/j.patter.2023.100758

CASE 2: METABOLIC REACTION GNN

Study datasets are processed through standardized workflow and GNN model

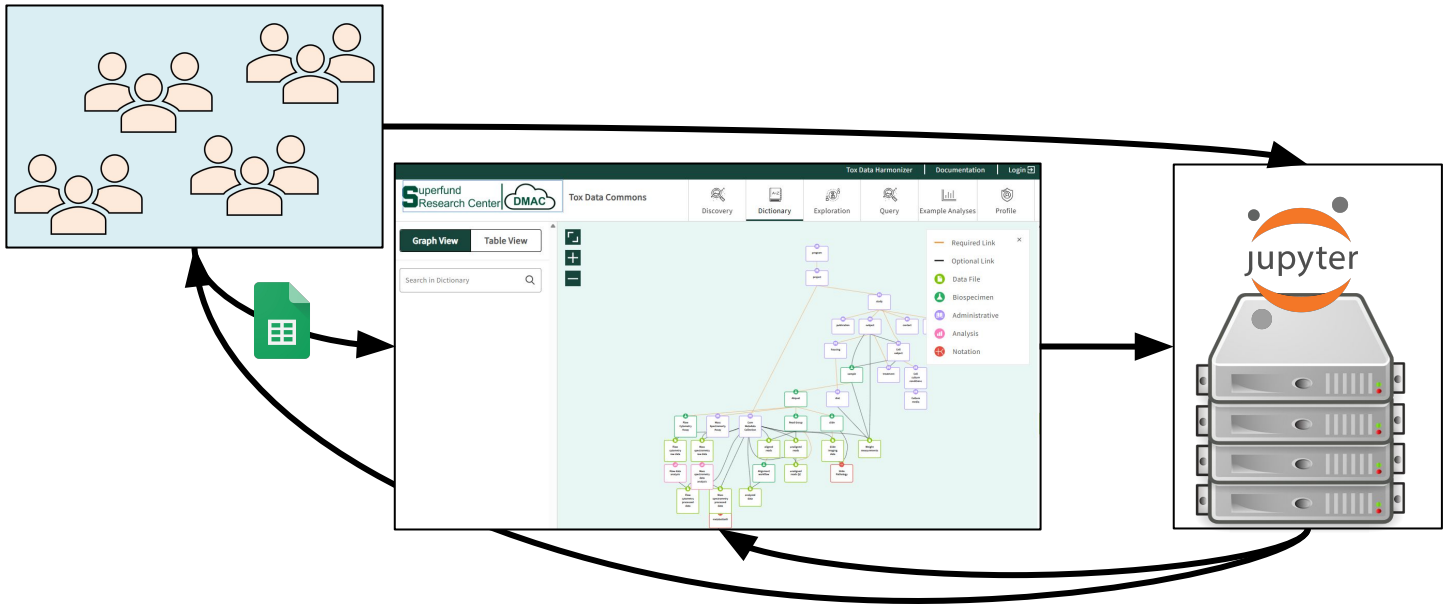


Identify metabolic reactions which are not found using more traditional methods.



"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

KRUMWARE

NIH 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

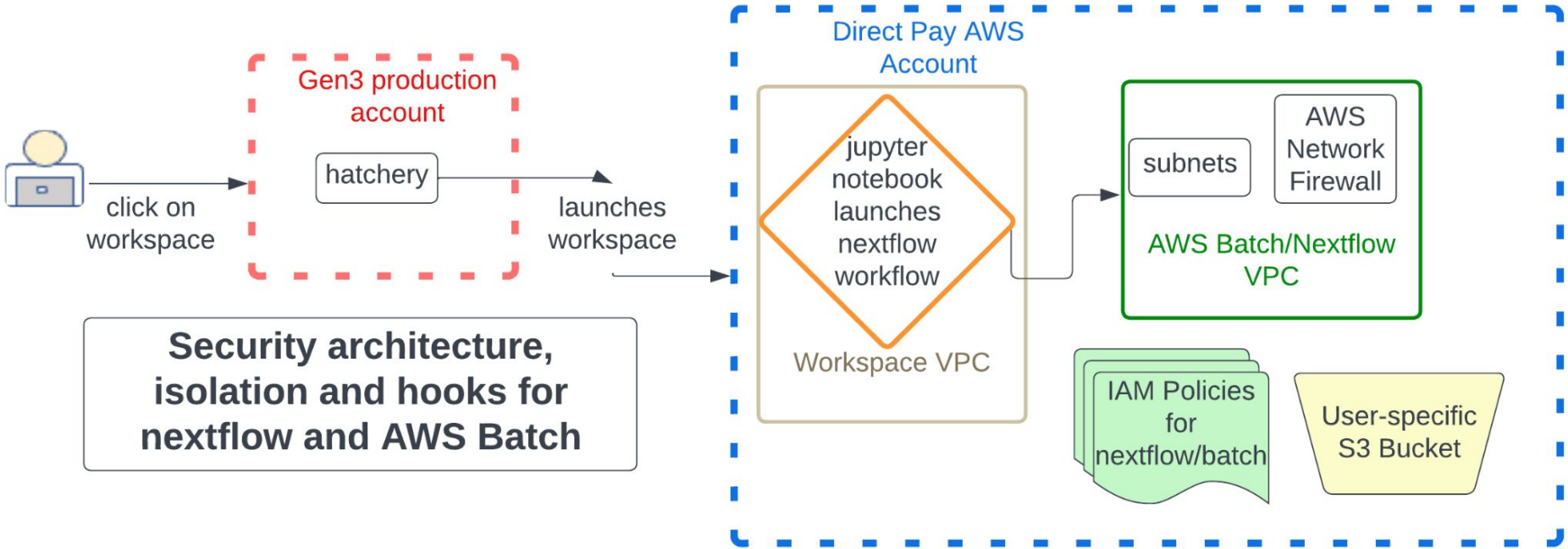
- Bioinformatics or data science workflows (nextflow, or other workflow language)
  - RNASeq, Variant calling, Copy number inference, Methylation and others
  - Training and testing AI/ML models for prediction or classification task
  - Federated learning applications
- Data
  - Small batch of data locally 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

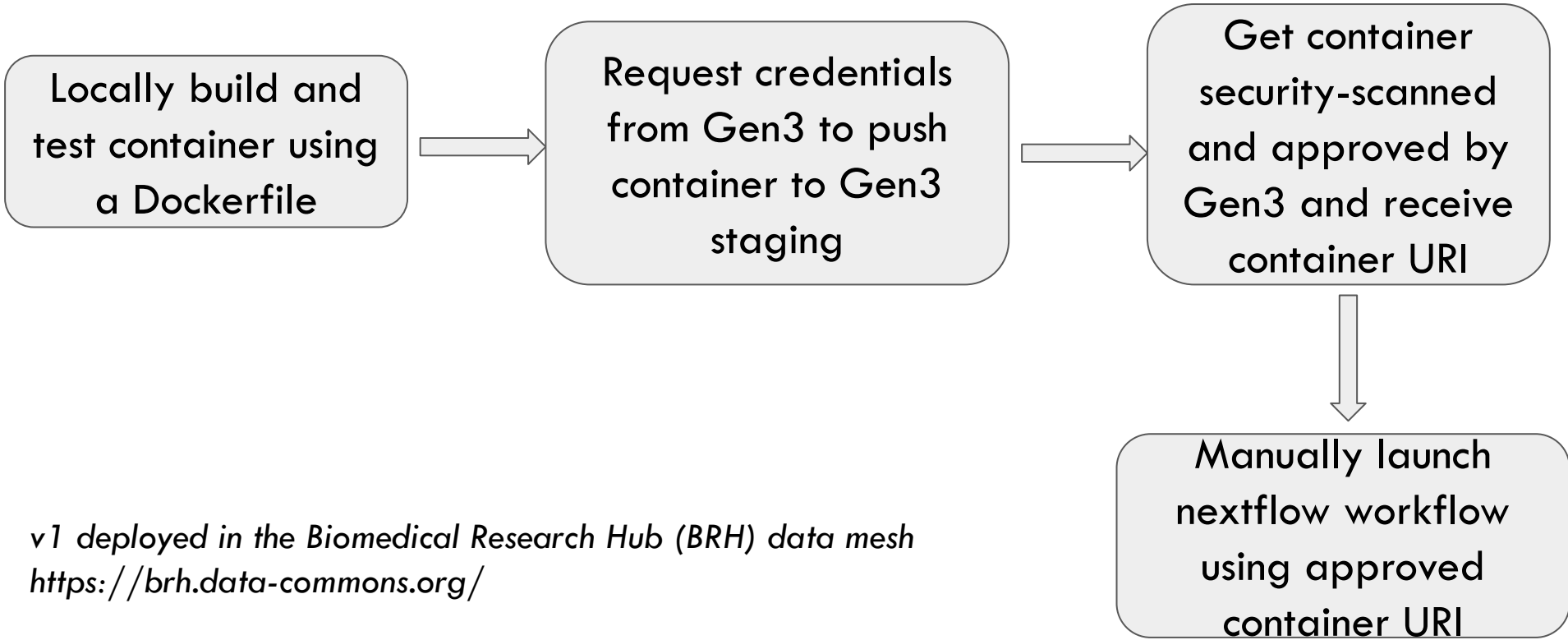
Our previous proof-of-concept solution (v1)



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

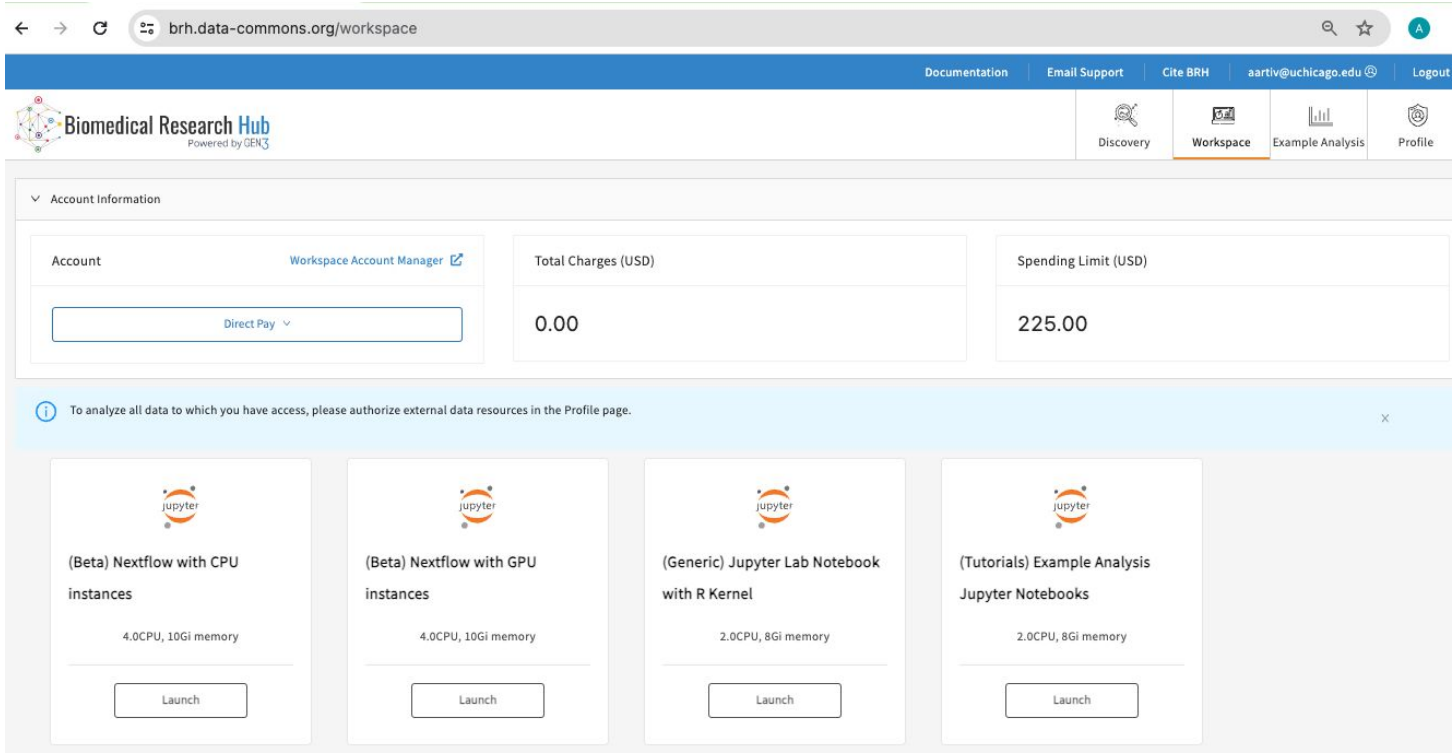
Overview of steps
Run a user-developed nextflow workflow with container (v1)

Containers are run in FedRAMP moderate environment with security compliance

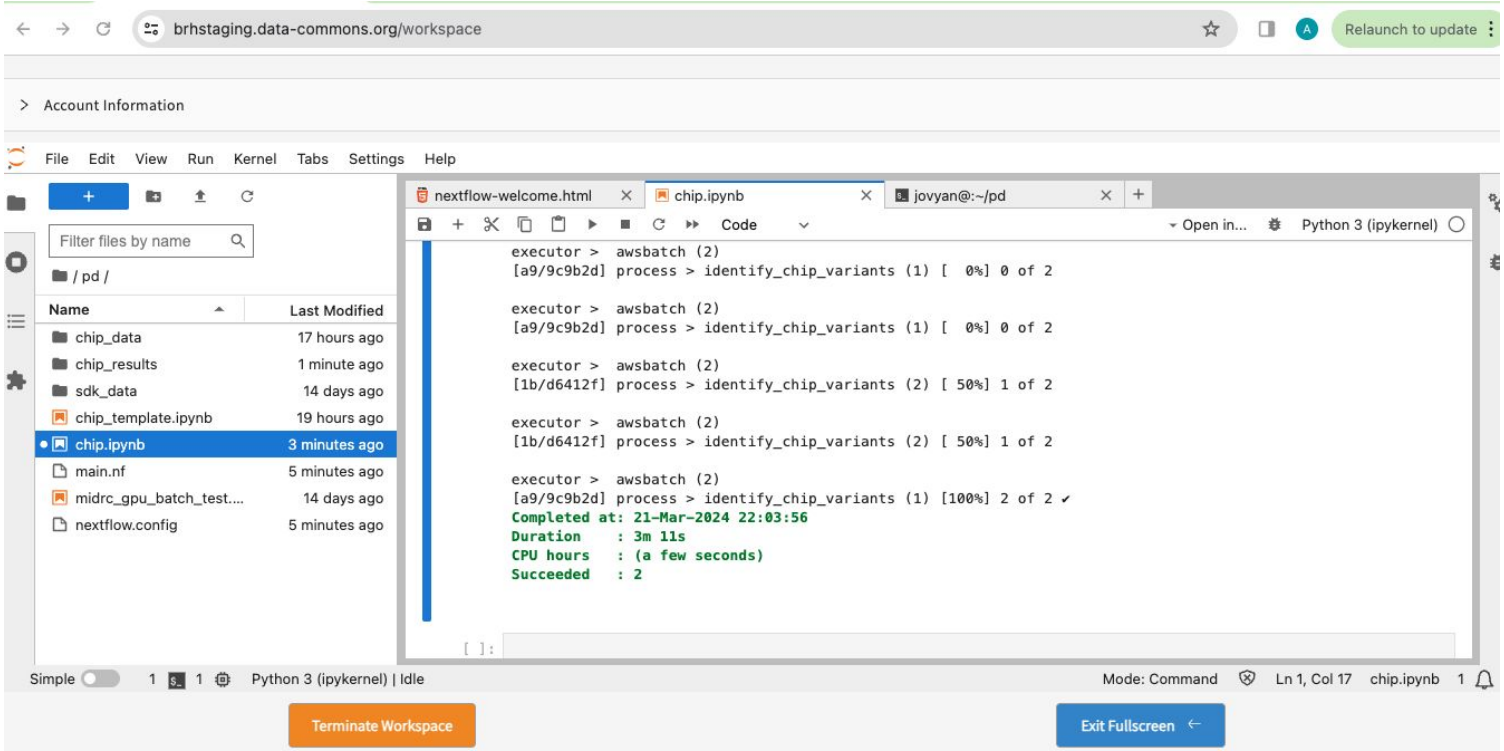


v1 deployed in the Biomedical Research Hub (BRH) data mesh
https://brh.data-commons.org/

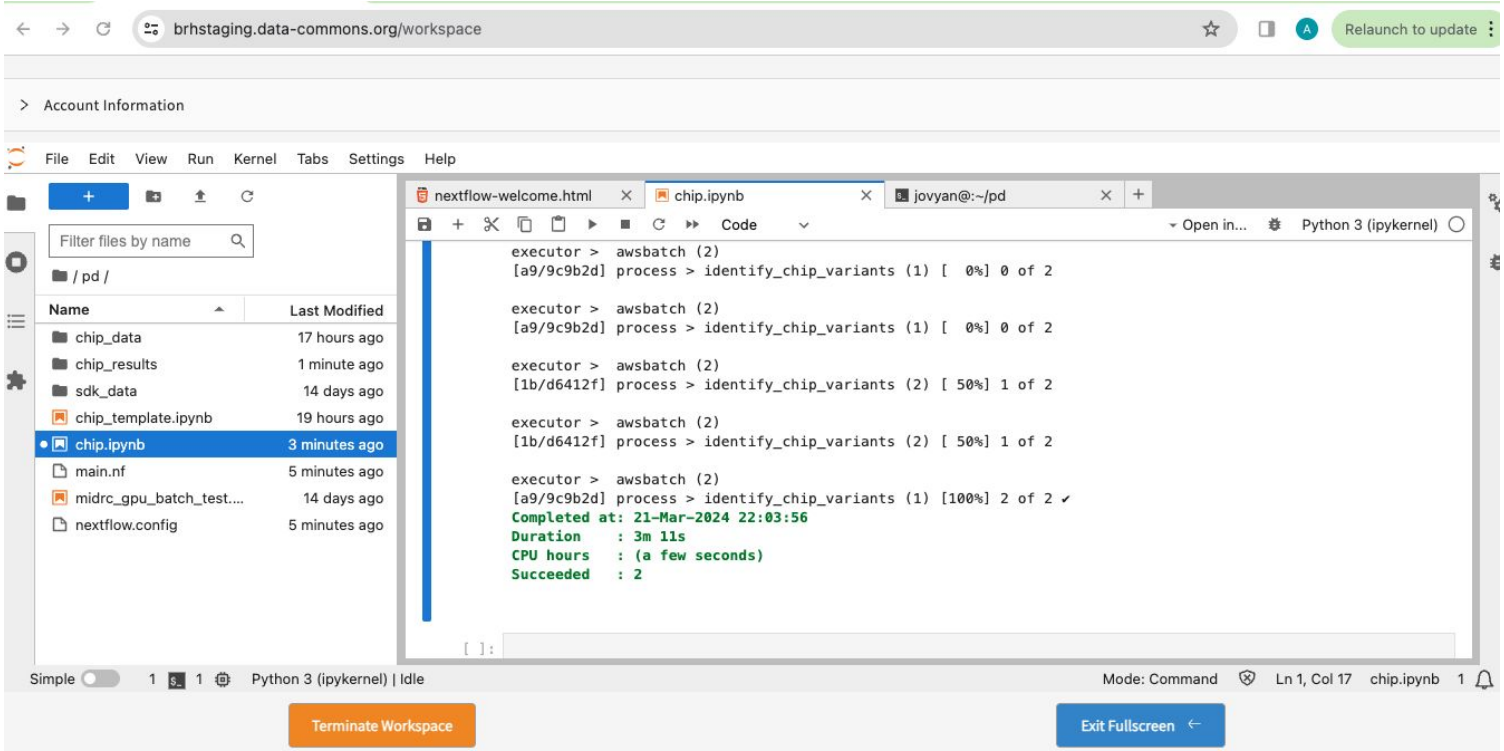
BRH workspace



Launch workflow from JupyterLab notebook



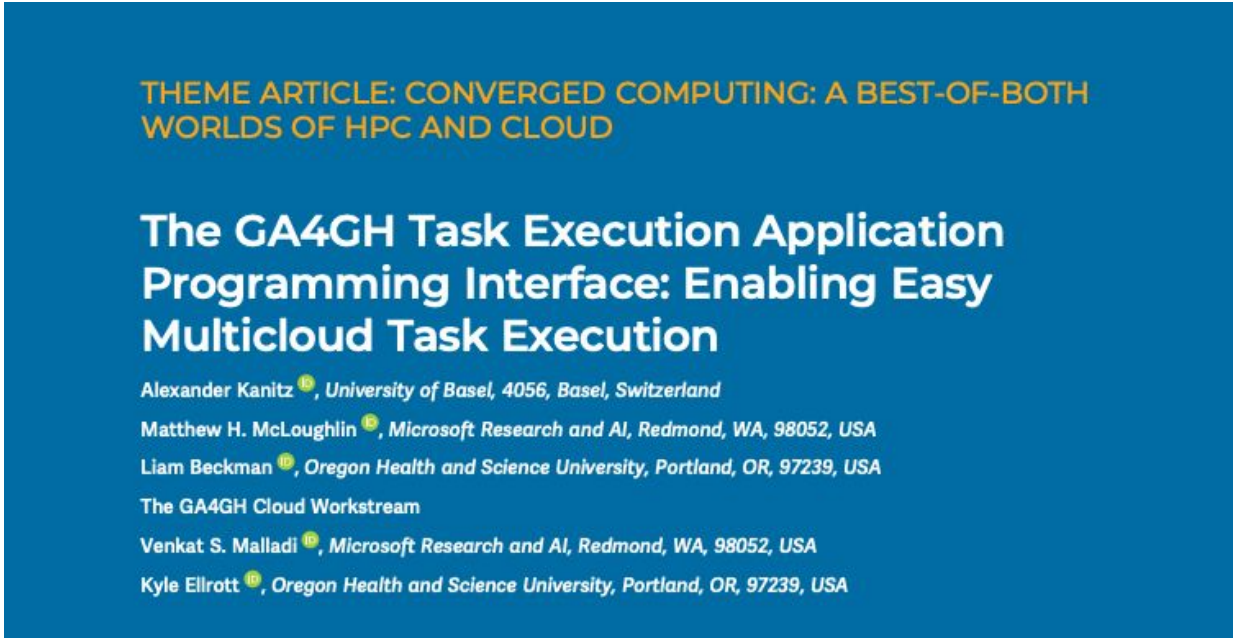
Launch workflow from JupyterLab notebook



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

Current focus (v2)
Programmatic submission of workflows using TES API

GEN3 DATA COMMONS



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

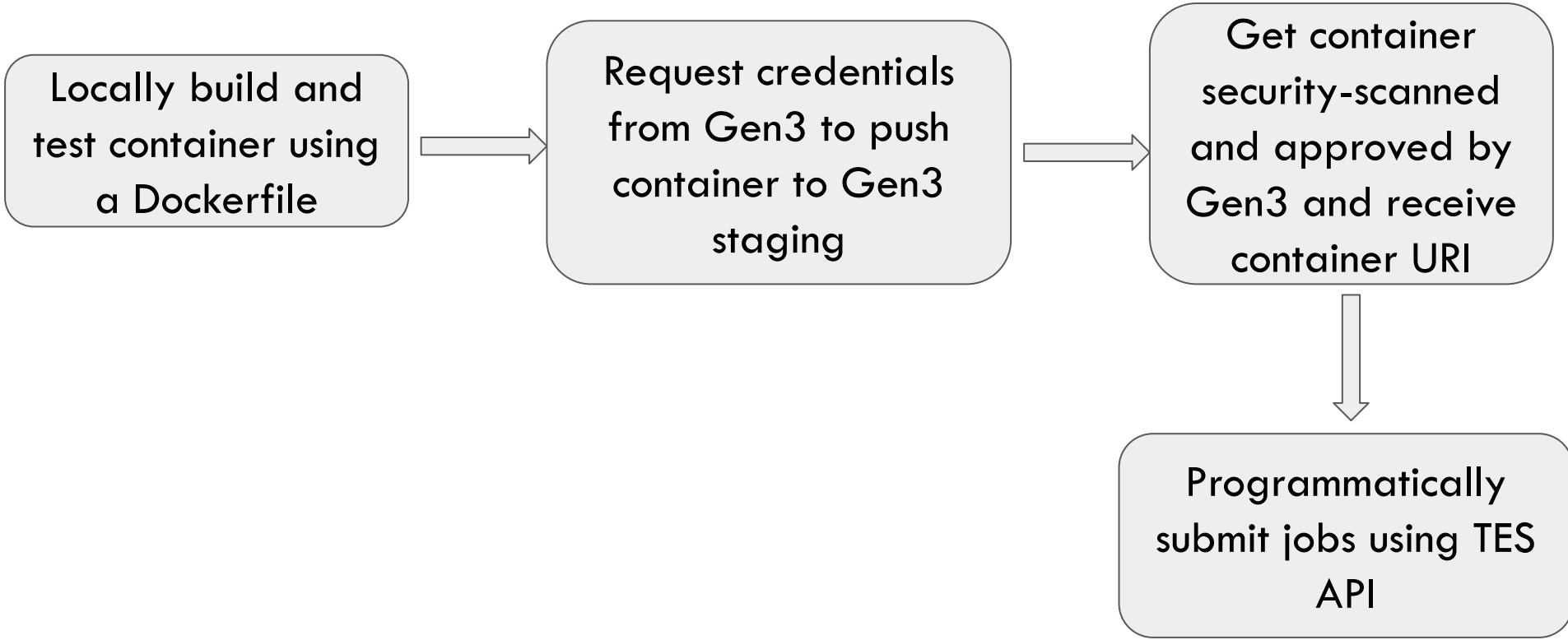
Venkat S. Malladi, *Microsoft Research and AI, Redmond, WA, 98052, USA*

Kyle Ellrott, *Oregon Health and Science University, Portland, OR, 97239, USA*

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10564576

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

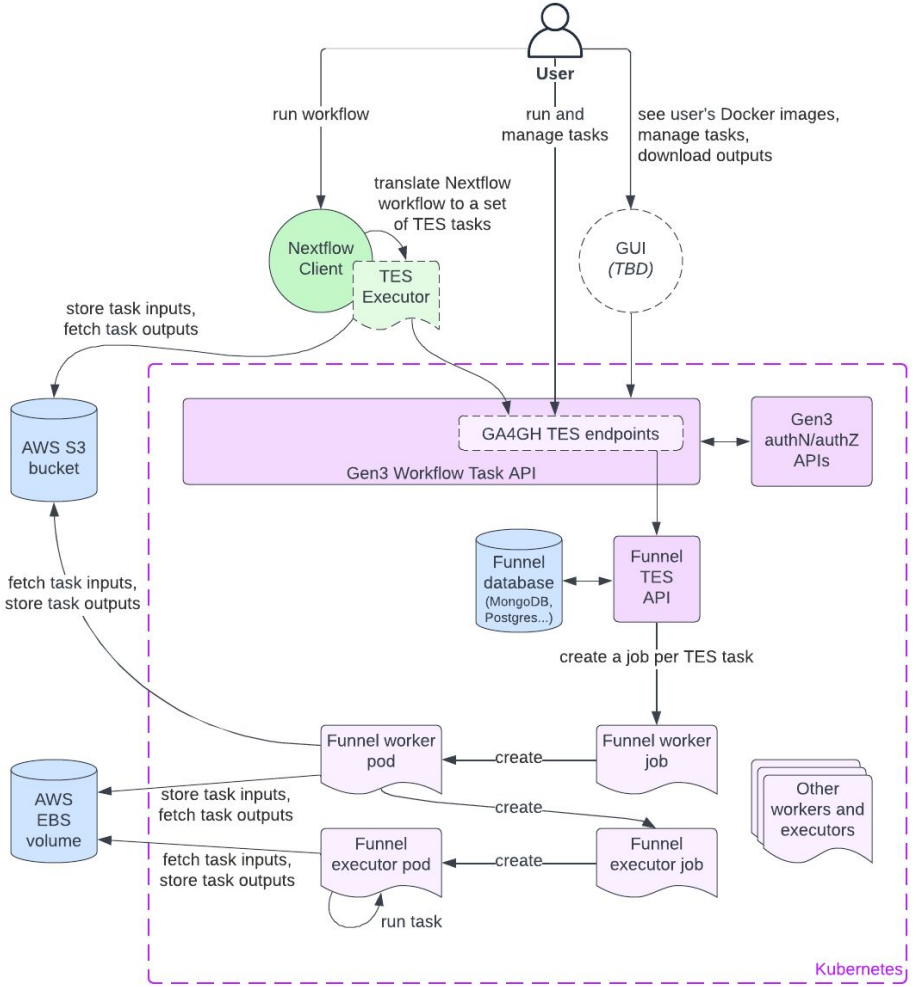
Containers are run in FedRAMP moderate environment with security compliance



Architecture

TES Backend: Funnel

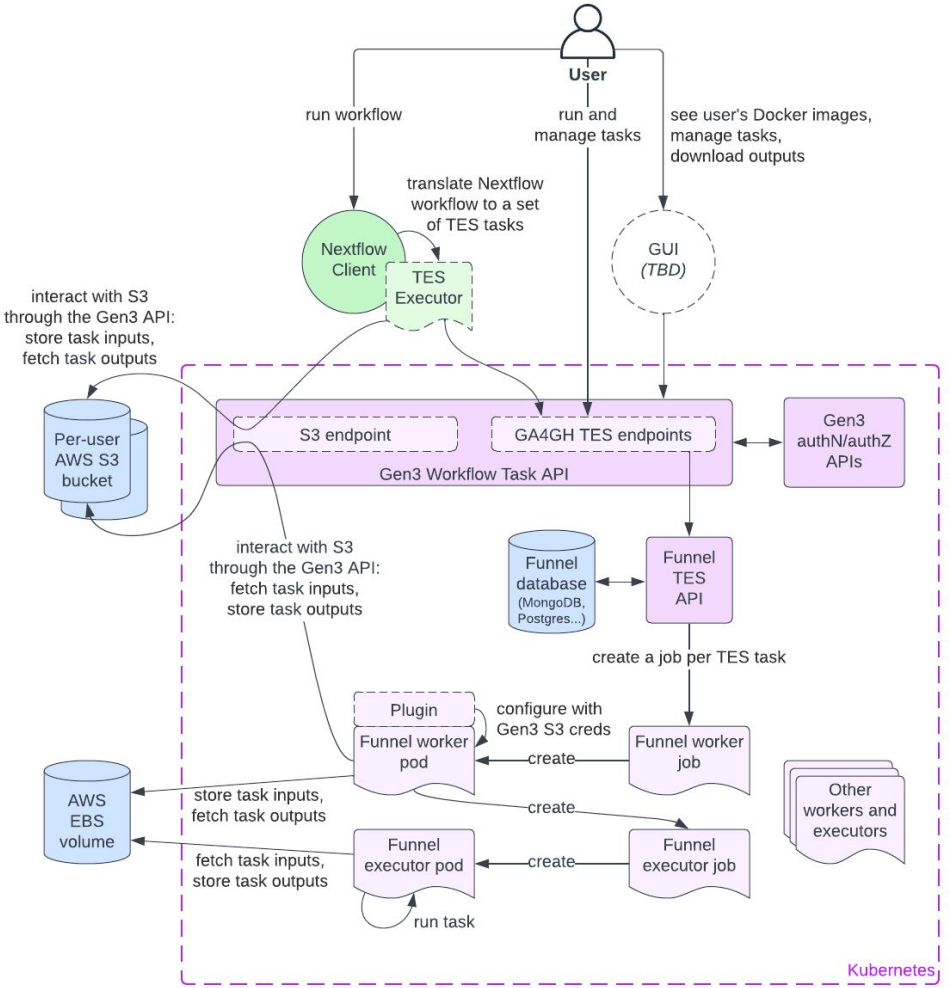
- Tool for distributed task execution
- Developed by the Oregon Health & Science University



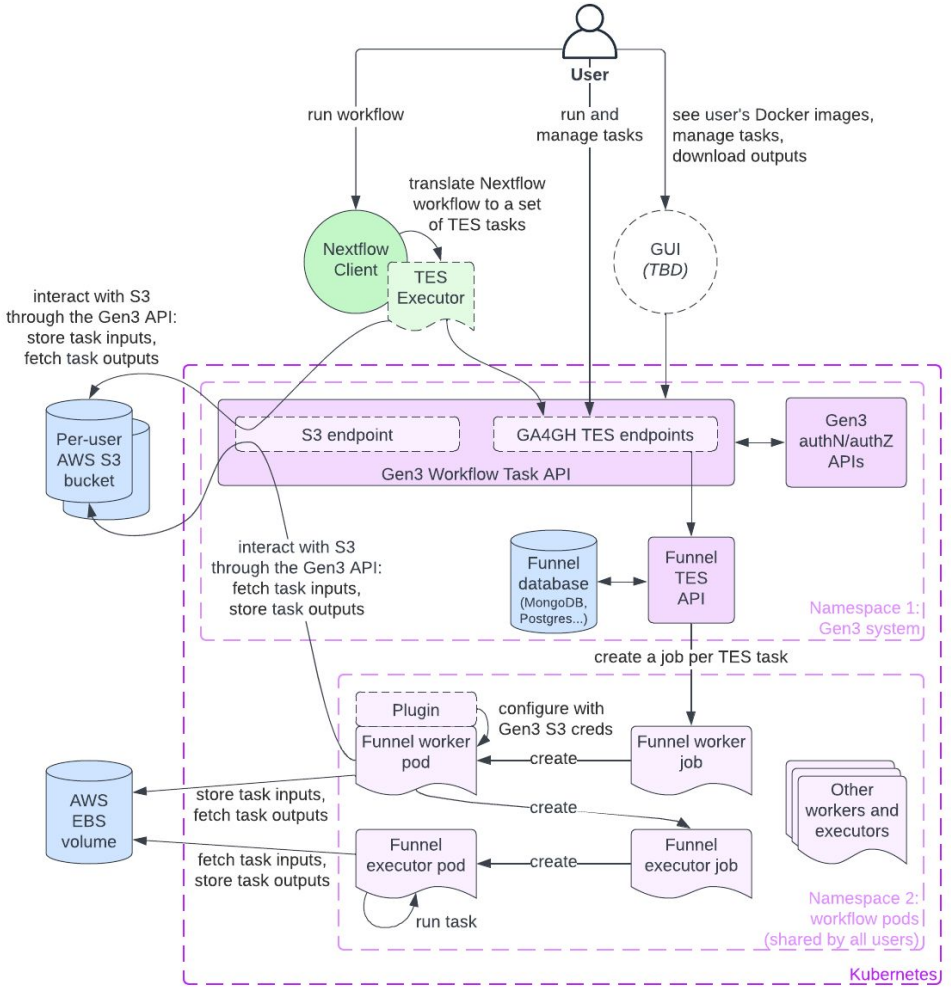
Architecture

Gen3 S3 endpoint

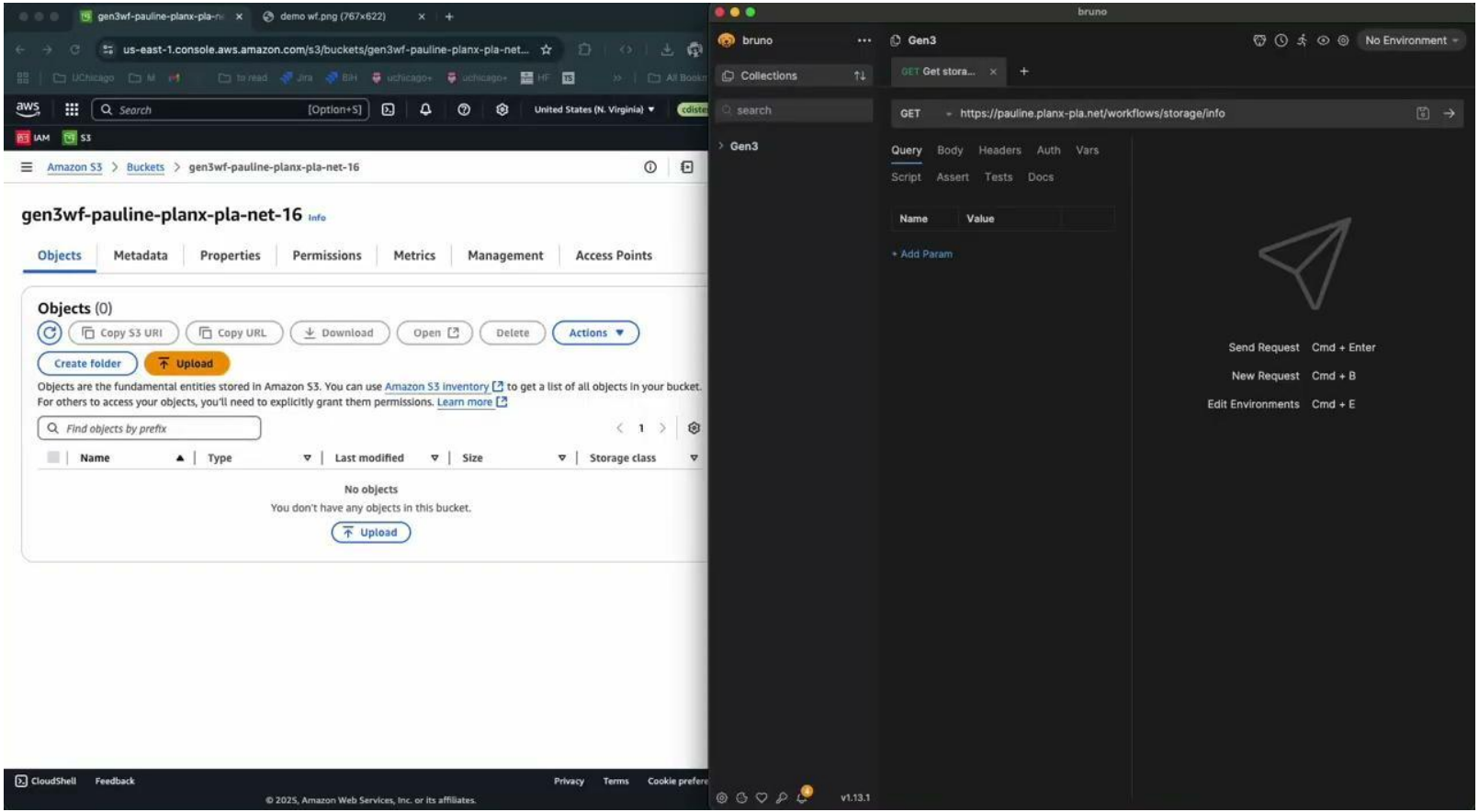
- The user and the Funnel worker authenticate with Gen3
- One S3 bucket per user
- Abstracts AWS credentials away from the user and the Funnel worker



Architecture



Nextflow and GA4GH TES demo



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:

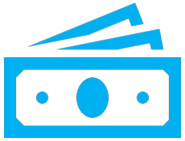
AI-ready data, infrastructure, and implementation

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

ORIEN

Oncology Research Information Exchange Network (ORIEN)

AlcHepNet

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

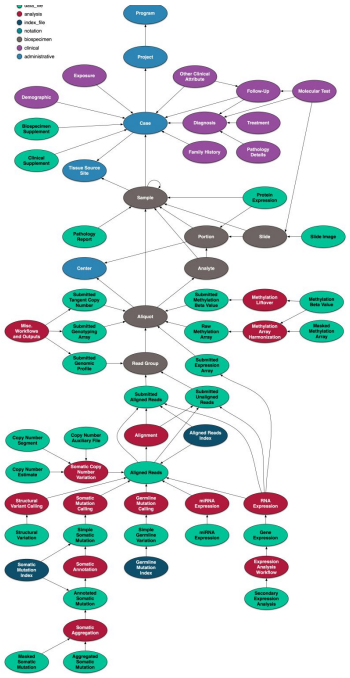
All of Us RESEARCH PROGRAM

The Future of Health Begins With You

Graph data model
Graph AI
Research data commons

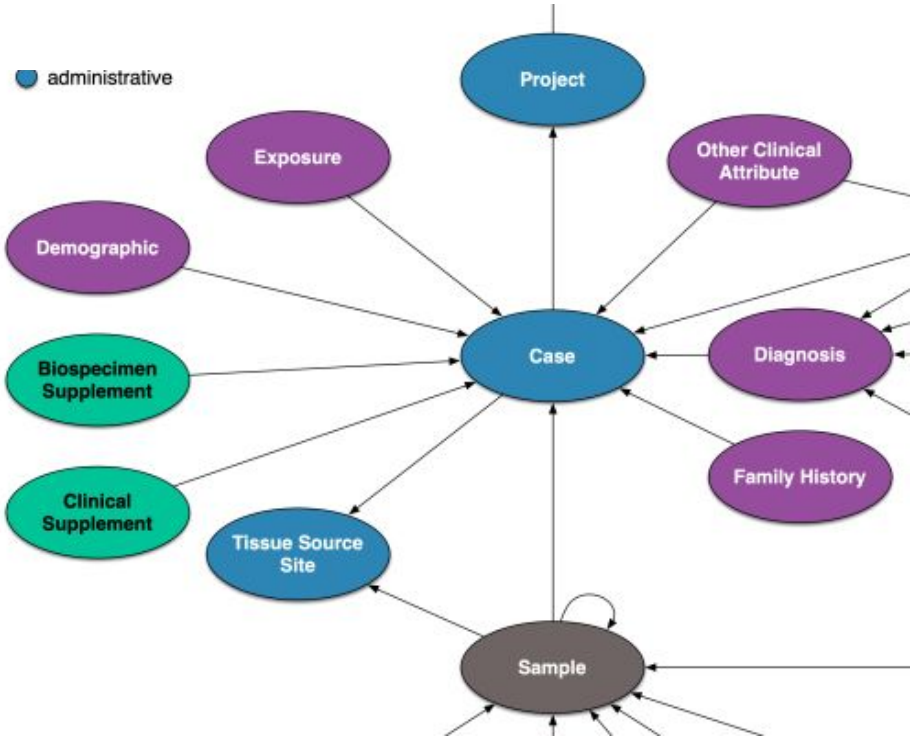
Graph data models: graphs to organize data artifacts

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:

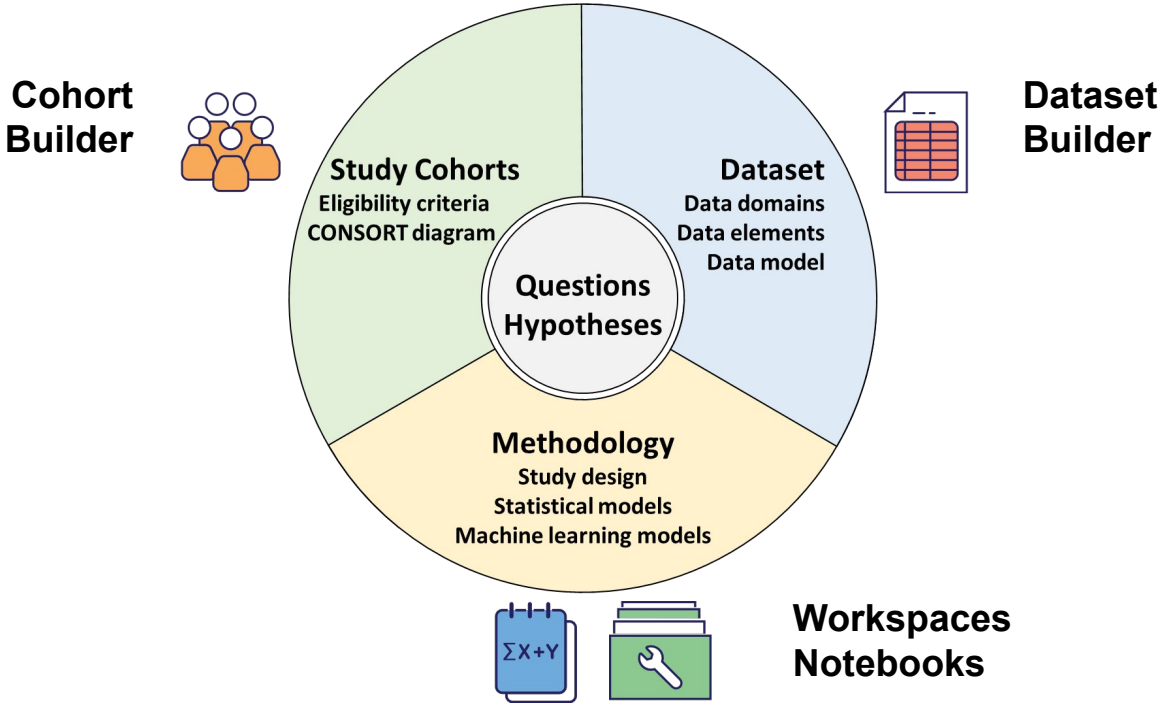
1. Define the integration and harmonization of real-world and research data
2. Enable new clinical study designs
3. Support AI/ML-readiness of biomedical big data
4. Widely used in real-world data infrastructures

GEN3 DATA COMMONS

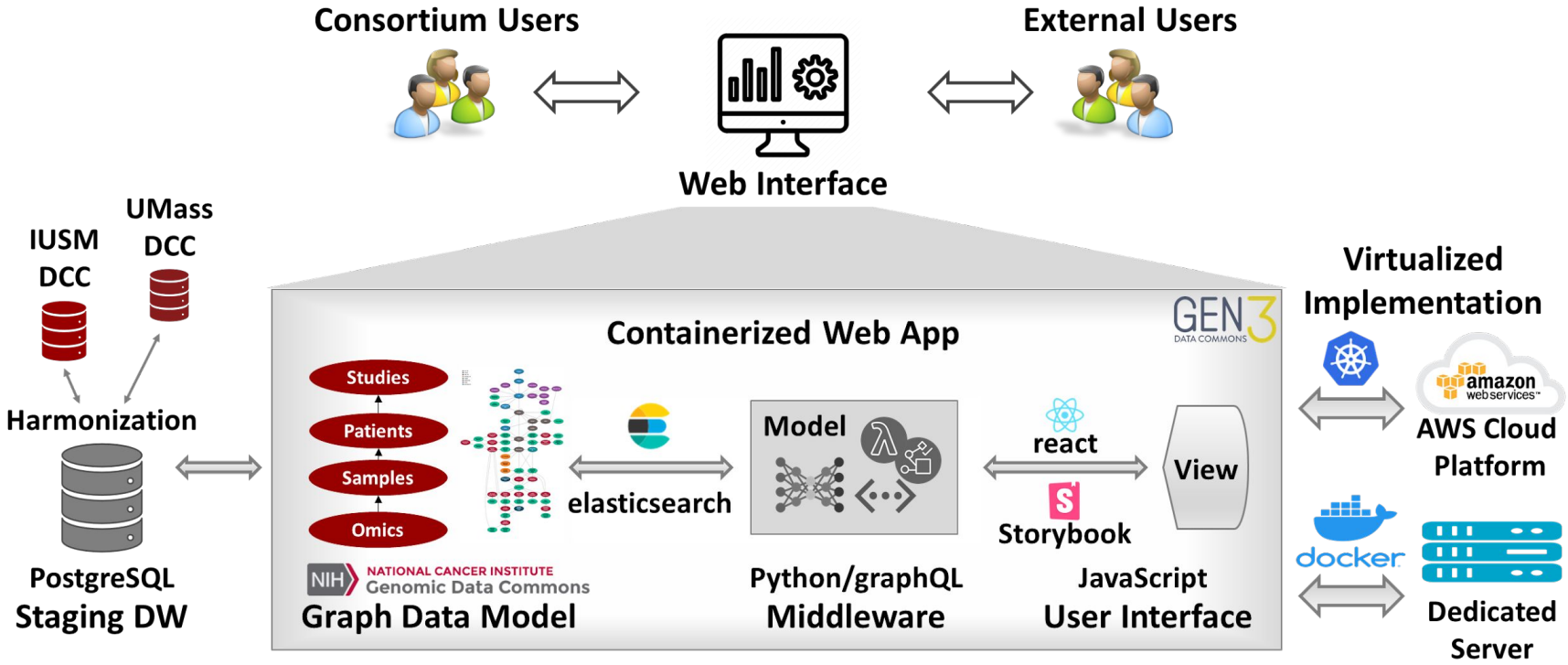
ARDaC: Alcohol Research Data Commons an AI/ML-ready platform for clinical trials and observational studies

NIAAA: 2U24AA026969

RDC provides essential functionalities for clinical studies



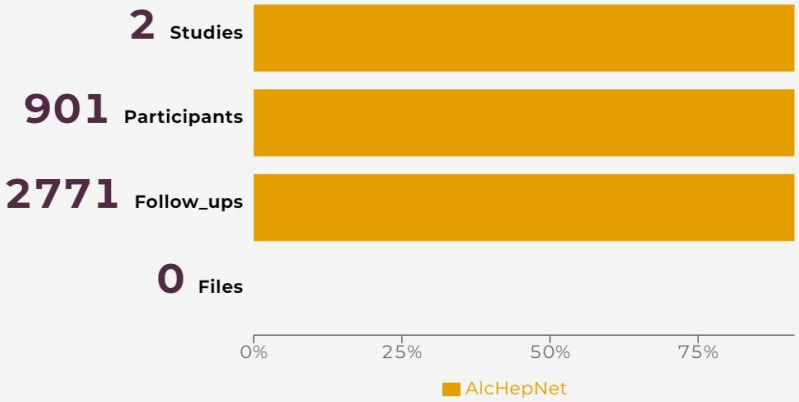
ARDaC: Architecture



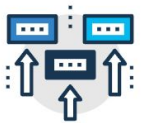
AlcHepNet is a clinical and translational research initiative funded by the National Institute on Alcohol Abuse and Alcoholism (NIAAA), a division of NIH.

AlcHepNet Research Data Commons

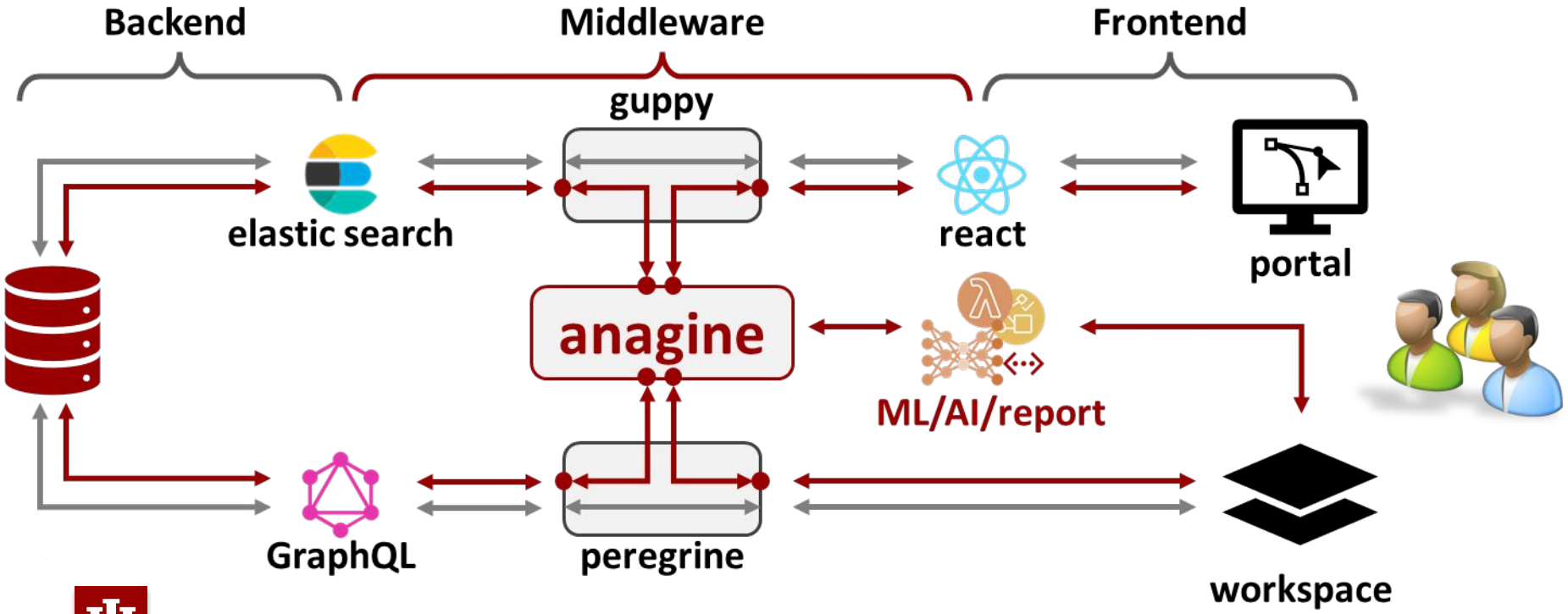
The AlcHepNet Research Data Commons, or ARDaC, is an integrative environment for exploring the clinical and omics data generated by AlcHepNet and related translational studies, and for investigators within the network to access data. ARDaC is sponsored by National Institute on Alcohol Abuse and Alcoholism.





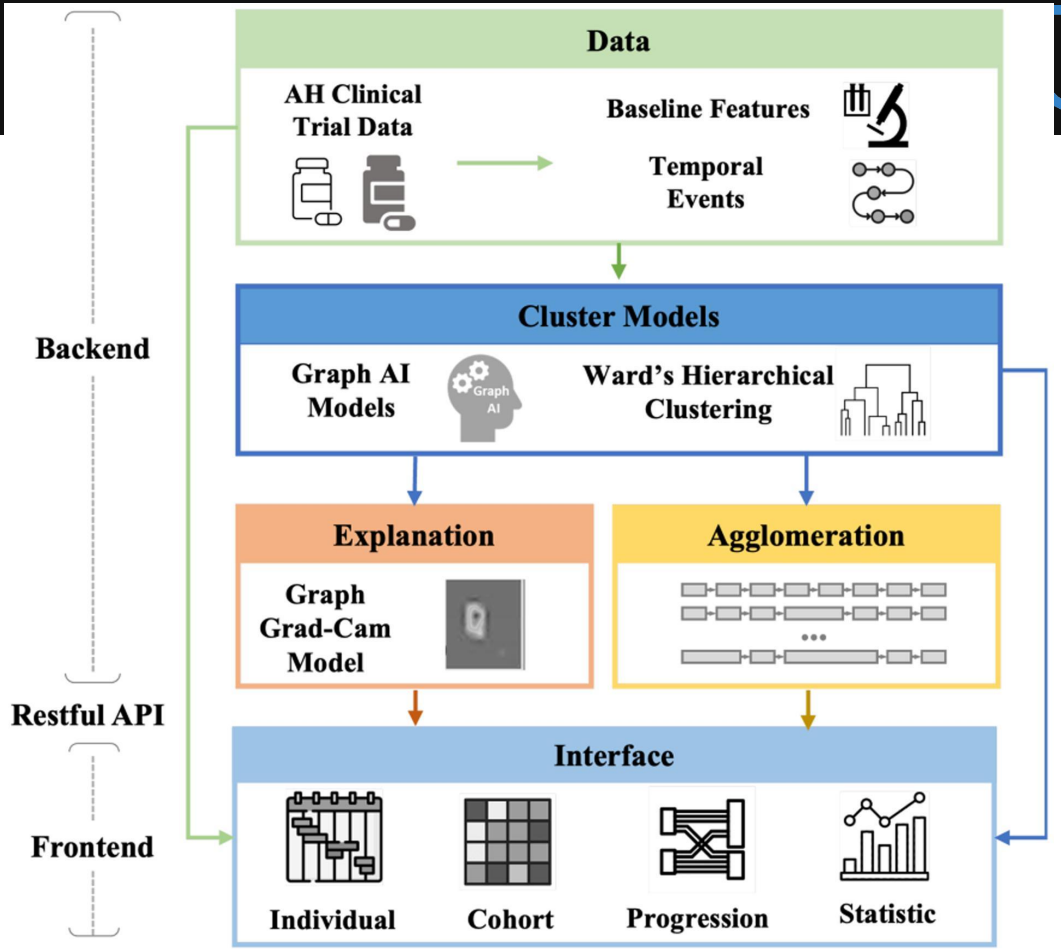



anagine: the AI/ML analysis engine of ARDaC

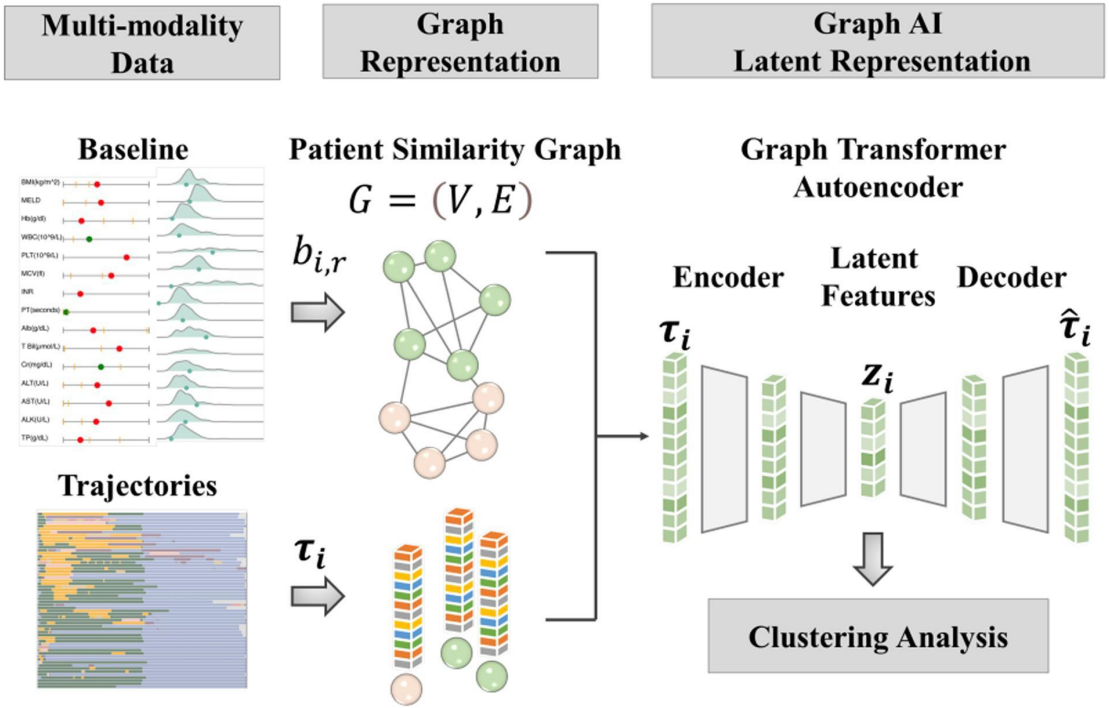


TrialView: ML/AI on ARDaC

Powered by ARDaC graph data model



Trajectory learning with graph AI



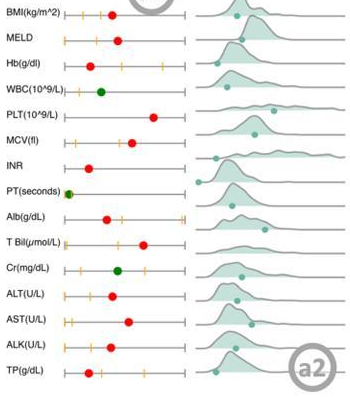
Alcoholic Hepatitis Clinical Trial Data Explorer

Individual View

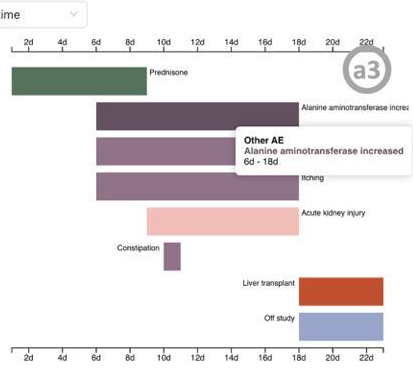
Select a case here

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		



a1

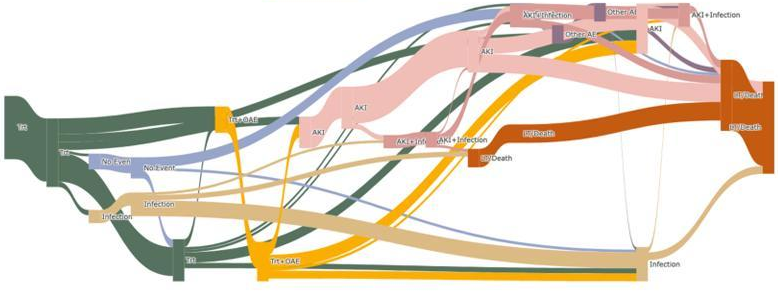


a3

a2

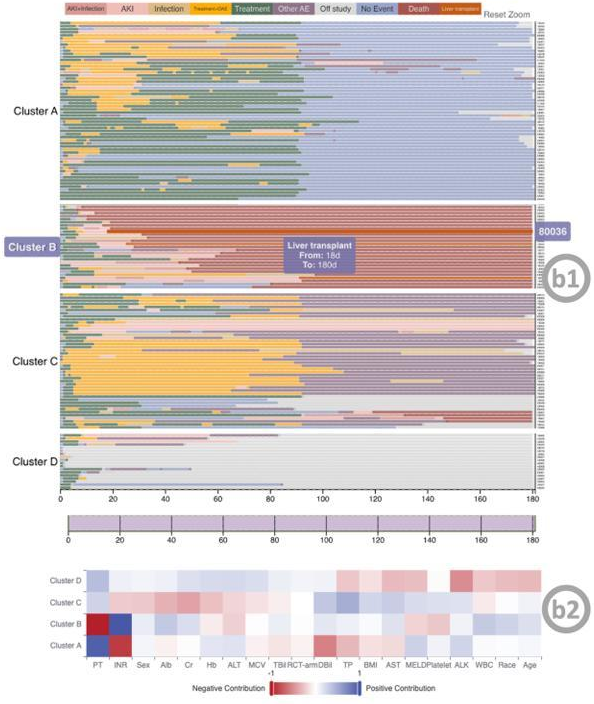
Progression View

Cluster A | Cluster B | Cluster C | Cluster D



Cohort View

Cluster Method: Ward | Graph Transformer Layout Method: Original | Even



b1

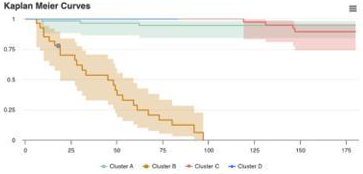
b2

RCT arm and Outcome Distribution

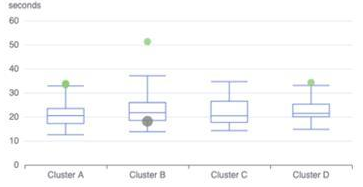
Sex and Outcome Distribution

Statistics View

By Group | By RCT Arm

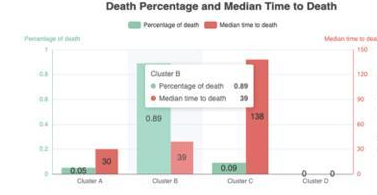


Select a feature to see distribution: Prothrombin Time (PT)



AKI | Infection

Death | Off Study

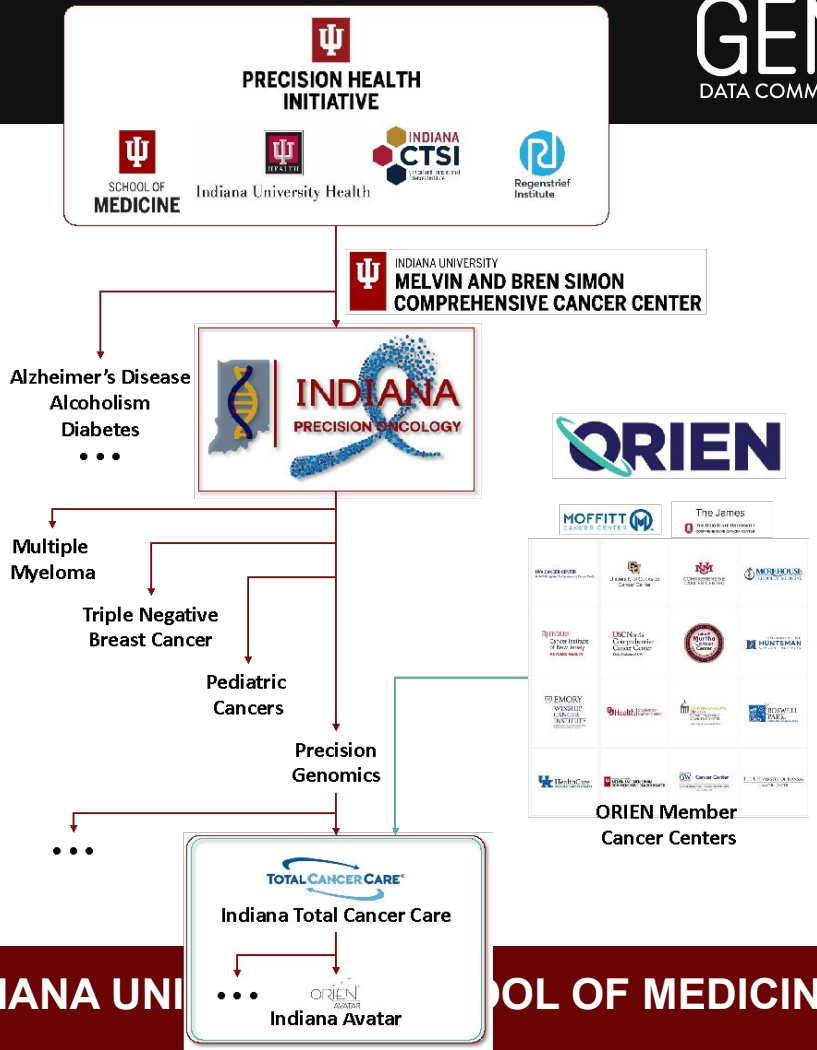


IPO: Indiana Precision Oncology Research Data Commons

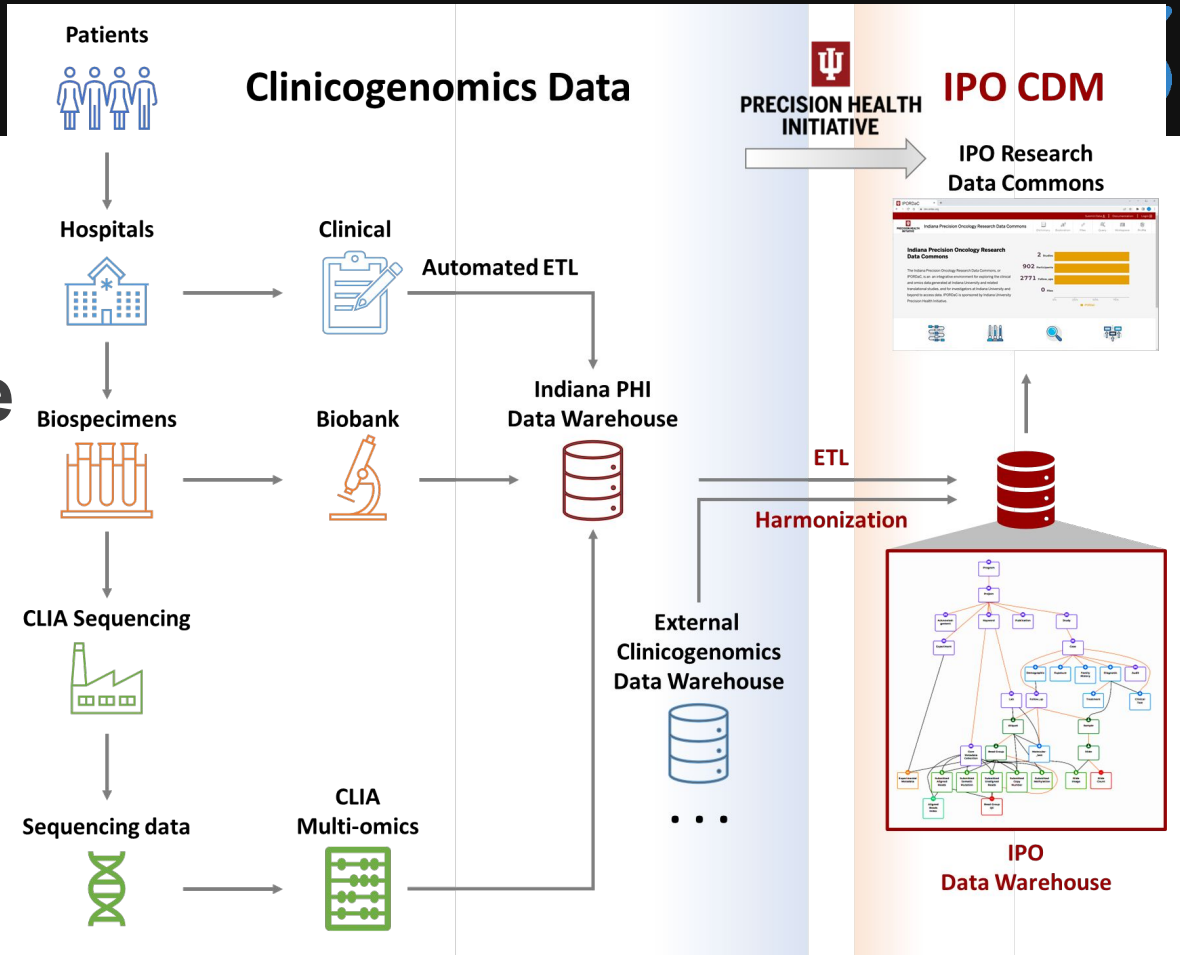
The clinicogenomics ecosystem

Indiana Precision Health Initiative

ORIEN: Oncology Research Information Exchange Network



IPO clinicogenomics data architecture



GEN3 DATA COMMONS

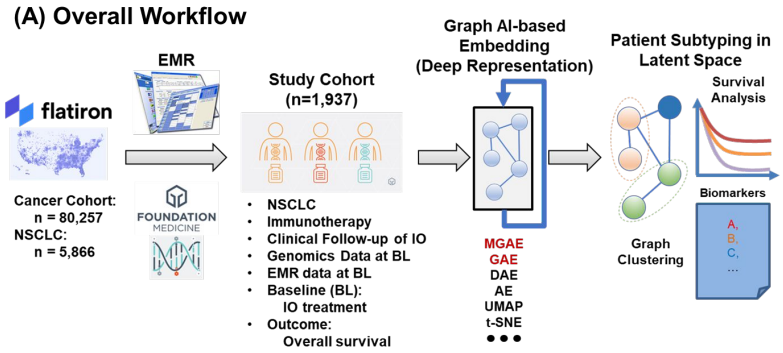
GAIPO: graph AI for precision oncology

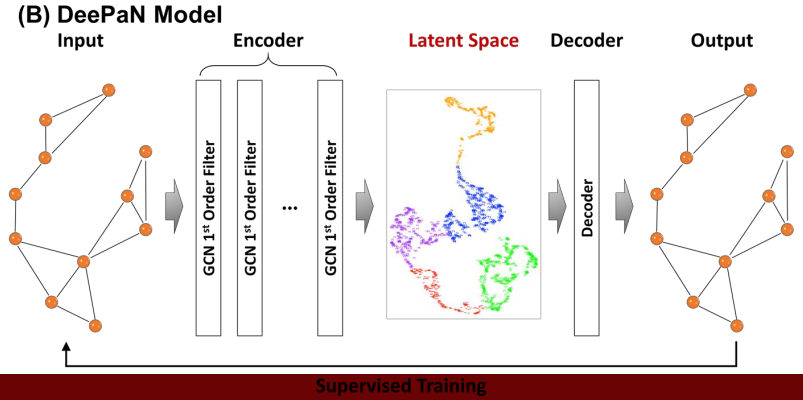
NCI: 3P30CA082709-25S1

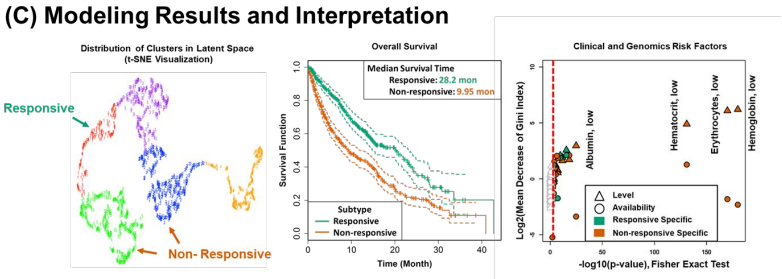
INDIANA UNIVERSITY SCHOOL OF MEDICINE

Graph AI models generated at IU

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

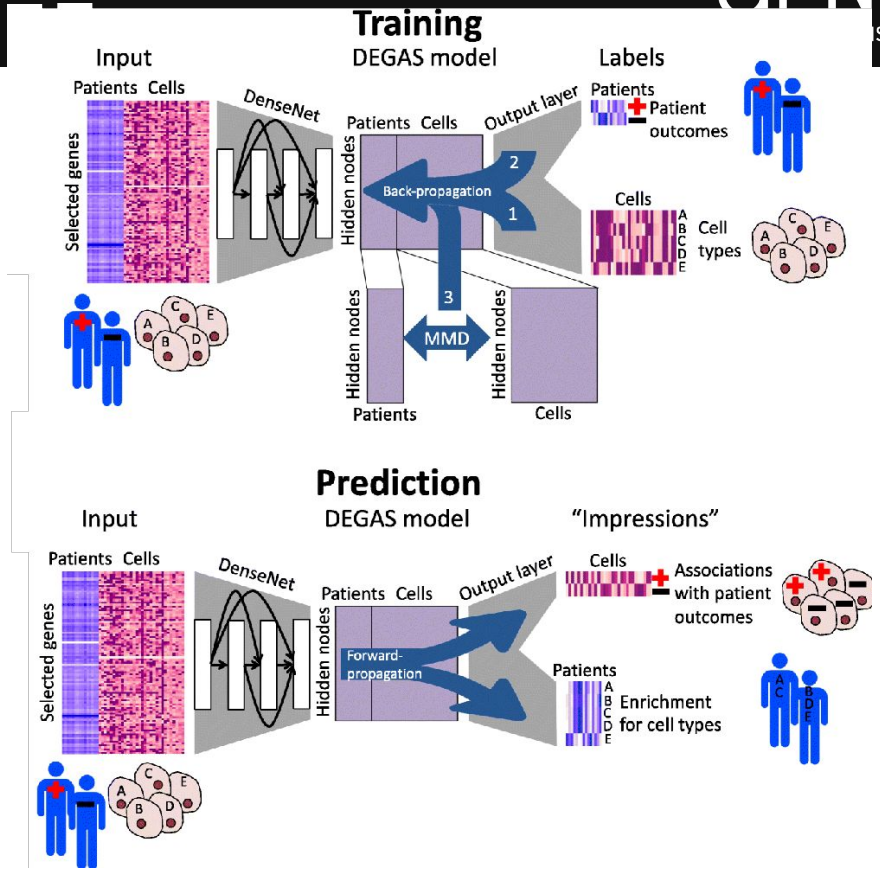






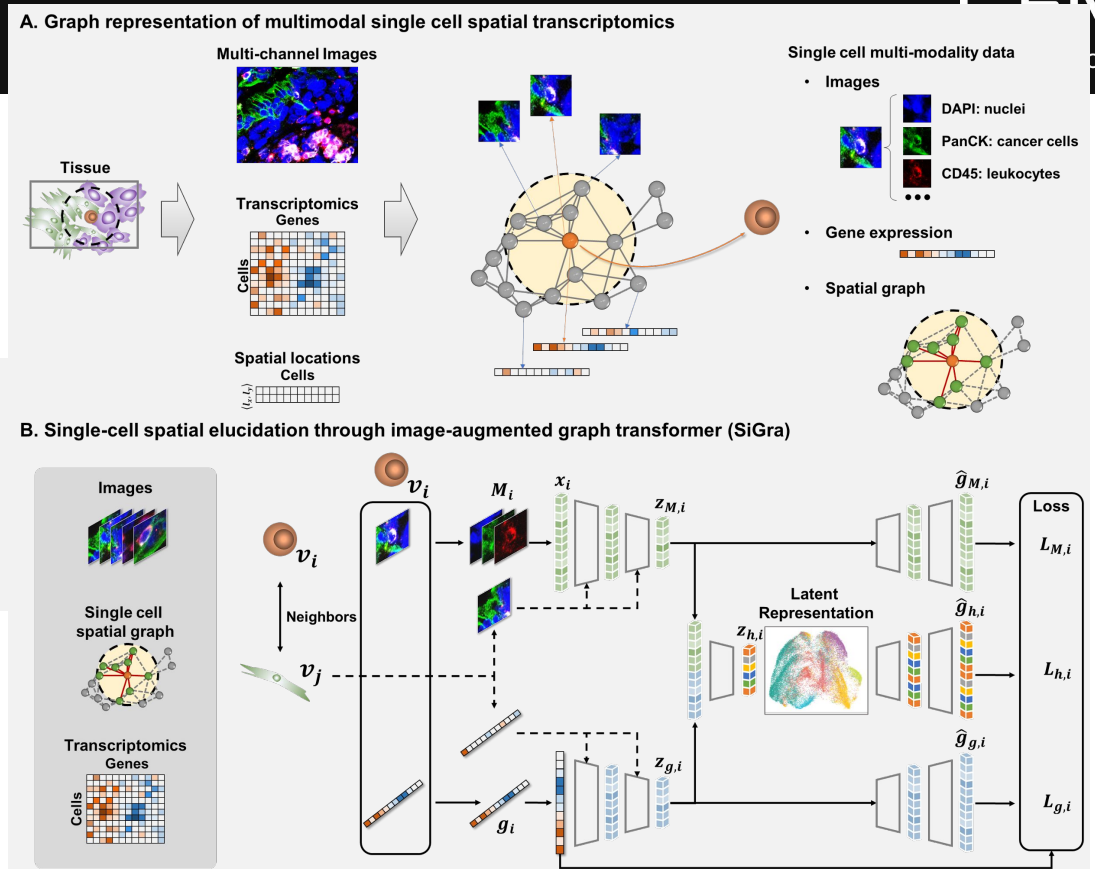
Graph AI models generated at IU

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



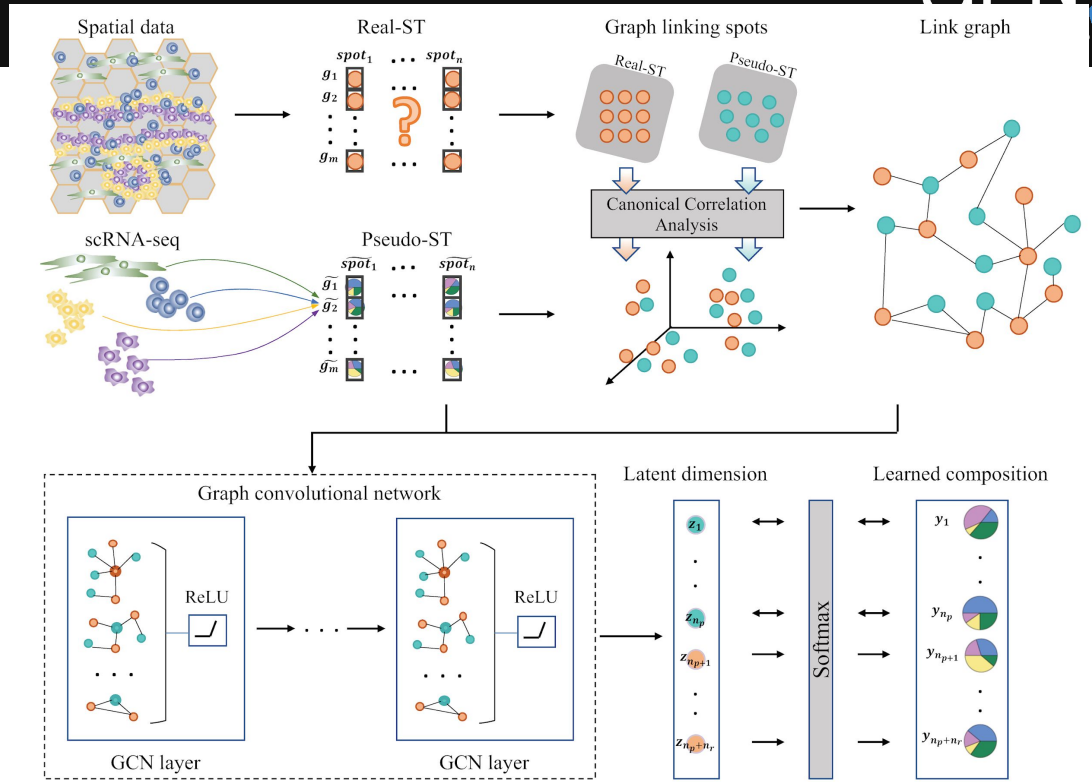
Graph AI models generated at IU

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



Graph AI models generated at IU

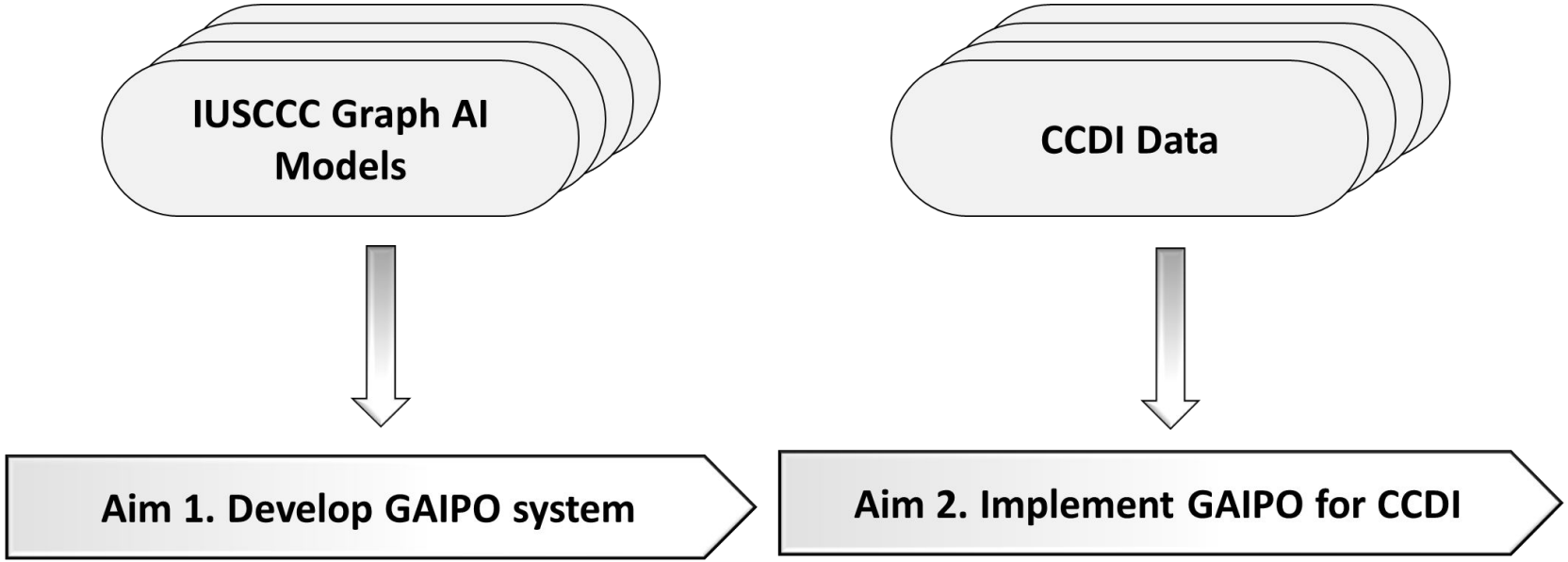
DSTG, Su's lab
Single-cell and spatial data
https://academic.oup.com/bib/article/22/5/bbaa414/6105942



Childhood Cancer Clinical Data Commons (C3DC)

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

Aims of GAIPO



CCDI's data sources and data elements

Cancer Types
- leukemias
- central nervous system neoplasms
- lymphomas
- neuroblastoma

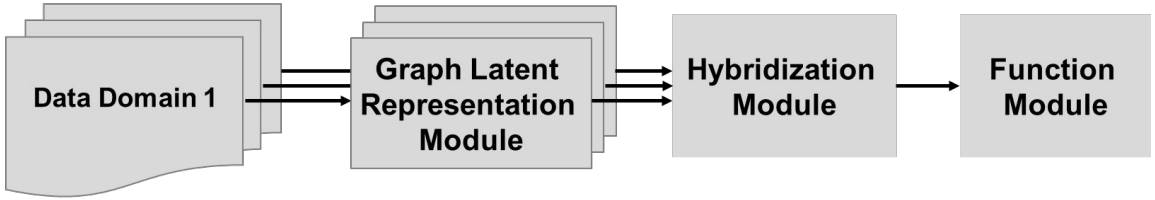
Data sources
- National Childhood Cancer Registry
- CCDI Molecular Characterization Initiative
- CCDI OncoKids
- NCI CCSG CCDI Supplement Additional Genomic Submission
- Pediatric Cancer Knowledge Base (histological images)

Clinical data
- demographics
- pathological diagnosis
- treatments
- relapse and death

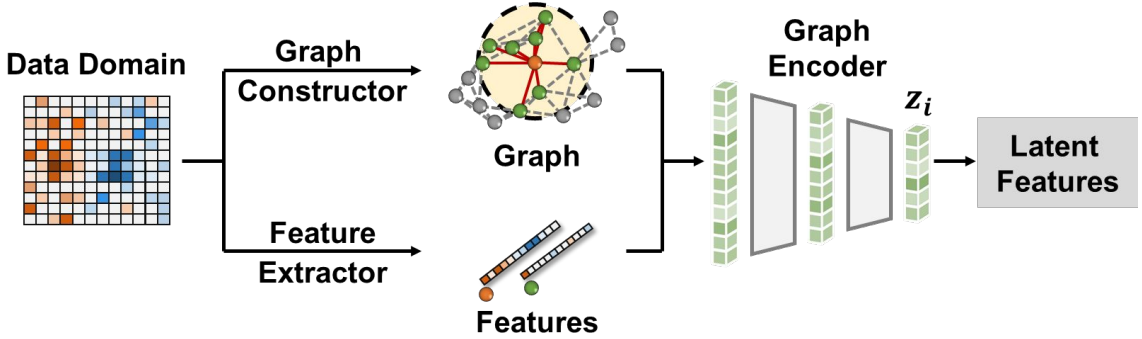
Omics data
- RNA-seq
- whole genome sequencing
- whole exome sequencing
- emerging spatial omics data

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

GEN3 DATA COMMONS

ARDaC Design

Wanzhu Tu, PhD
PI

Jing Su, PhD
Director

Data Modeling & Management

Carla Kettler
Data Management

Nanxin Jin
Data models

System Implementation

Alan Walsh, PhD
Cloud Implementation

George Nitsos
IT & Security

System Development & Visualization

Zuotian Li
Analytic Visualization

Dr. Xiang Liu
Development

Hao Wang
Analytic Visualization

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

AlcHepNet

AlcHepNet is a clinical and translational research initiative funded by the National Institute on Alcohol Abuse and Alcoholism (NIAAA), a division of NIH.

NIH National Institute on Alcohol Abuse and Alcoholism
NIH...Turning Discovery Into Health®

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

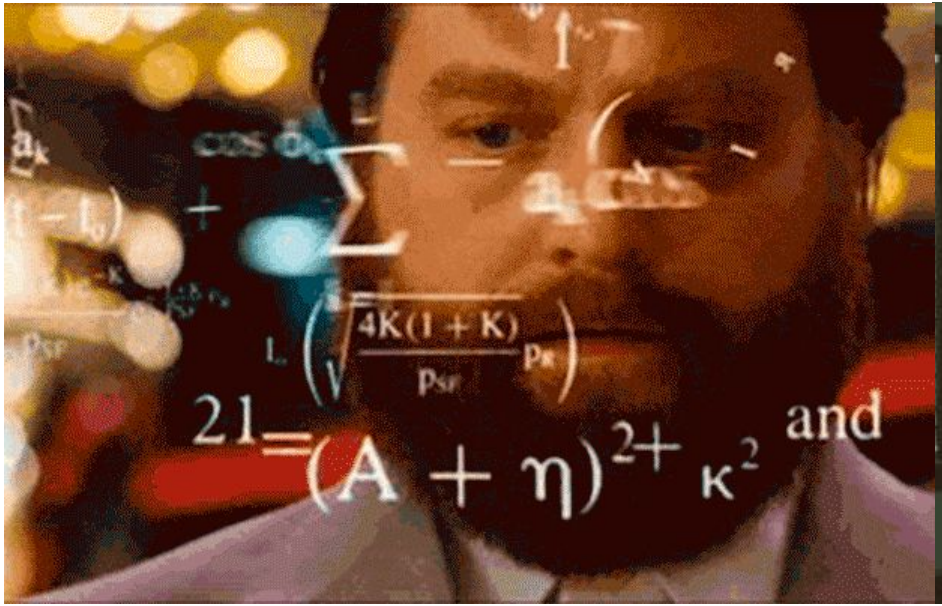
Nanxin Jin
Data models

Hao Wang
Analytic
Visualization

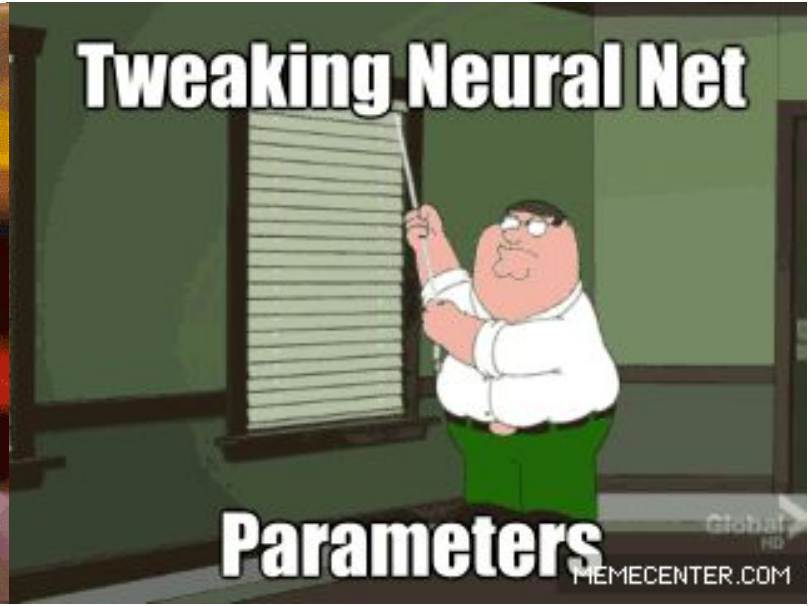
George Nitsos
IT & Security

Questions?

Design …



Tuning …



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