Experiences in the Data Commons: Bringing together Data, Community, and Compute to Improve Patient Outcomes

12:00 to 12:30 p.m.
Hubert H. Humphrey Building - East Wing

RTI INTERNATIONAL

Data Modernization Leadership

Rebecca Boyles, MSPH

Founding Director, Center for Data Modernization Solutions

RTI International

rboyles@rti.org

Rashonda Lewis, JD, MHA

Data Governance & Privacy Specialist

RTI International

rmlewis@rti.org

Taking Lessons from Research to Move the Dial in Public Health

- RTI overview
  - What is RTI and what do we offer?
  - The role of a data commons approach on the path to the North Star
  - How the National Institutes of Health's (NIH's) vision provides lessons for the Centers for Disease Control and Prevention's (CDC's) data modernization strategy
  - Data commons as a core and proven approach
  - Examples from Data Commons Efforts at NIH and what they offer the public health ecosystem
    - National Heart, Lung, and Blood Institute (NHLBI) BioData Catalyst
    - Helping to End Addiction Long-term® (HEAL) Initiative Data Ecosystem
    - NIH Cloud Platform Interoperability (NCPI)
- Lessons learned and an approach to the future

delivering **the promise of science**
for global good

RTI International is an independent, nonprofit research institute dedicated to improving the human condition. We combine scientific rigor and technical expertise in social and laboratory sciences, engineering, and international development to deliver solutions to the critical needs of clients worldwide.

Leading the Way

RTI Center for Data Modernization Solutions

Providing fit-for-purpose solutions to our clients to enable use of data for the public good. We bridge the research-information technology gap, applying a data ecosystem perspective that enables clients to maximize the value of data.

Data Governance and Strategy • Advanced Analytics • Infrastructure and Security • Software, Systems, and Maintenance • Operations Support

Moving Science Forward Requires an Integrated Approach to Data

Data Infrastructure	Data storage and security System interoperability
Modernized Data Ecosystem	Modernize existing silos Enable data integration across domains
Data Management, Analytics, and Tools	Reusable and accessible tools Improve discovery of resources Reduce barrier to use
Workforce Development	Upskill the workforce Engage a broader and more diverse community
Stewardship and Sustainability	Strategy Governance Sustainability models

https://datascience.nih.gov/nih-strategic-plan-data-science

Moving to the North Star in Public Health Also Requires an Integrated Approach to Data

Build the Right Foundation	Expand foundational infrastructure Modernize and connect key surveillance systems and sources Transform legacy systems, processes, and activities
Accelerate Data into Action	Increase interoperability through data standards Advance forecasting and predicting analytics Promote health equity
Support and Extend partnerships	Ensure partner alignment and collaboration Support policies for data exchange
Develop a State-of-the-Art Workforce	Identify workforce needs Increase data science capacity Facilitate state, tribal, local and territorial (STLT) data science upskilling
Manage Change and Governance	Govern policies, planning, and resources Manage culture change Streamline acquisition processes

https://www.cdc.gov/surveillance/data-modernization/priorities/index.html

Vision for Biomedical Research Data

- A modernized, integrated, FAIR, TRUST-ed biomedical data ecosystem
- E-infrastructure to support FAIR data and break down silos while leveraging legacy investments

The concept of a data commons is central to the solution.

- Fosters the development of a digital ecosystem and associated community
- Democratizes compute resources, tool development, and analysis
- Empowers previously under-resourced groups
- Enables users to focus on analytics research but removes the compute barrier

RTI Partners with NIH to Apply the Data Commons Approach

- We work across domain areas of public health.
- Technologists work shoulder-to-shoulder with subject matter experts and top scientists to develop strategies for future-proof systems that are modular, interoperable, and extensible.
- We provide a data ecosystem perspective on needs from data governance, harmonization, and security to systems integration, interoperability, and sustainability.
- We are committed to standards of quality and transparency (ISO and CMMI), so our clients trust us and know that we are delivering the best possible solutions.

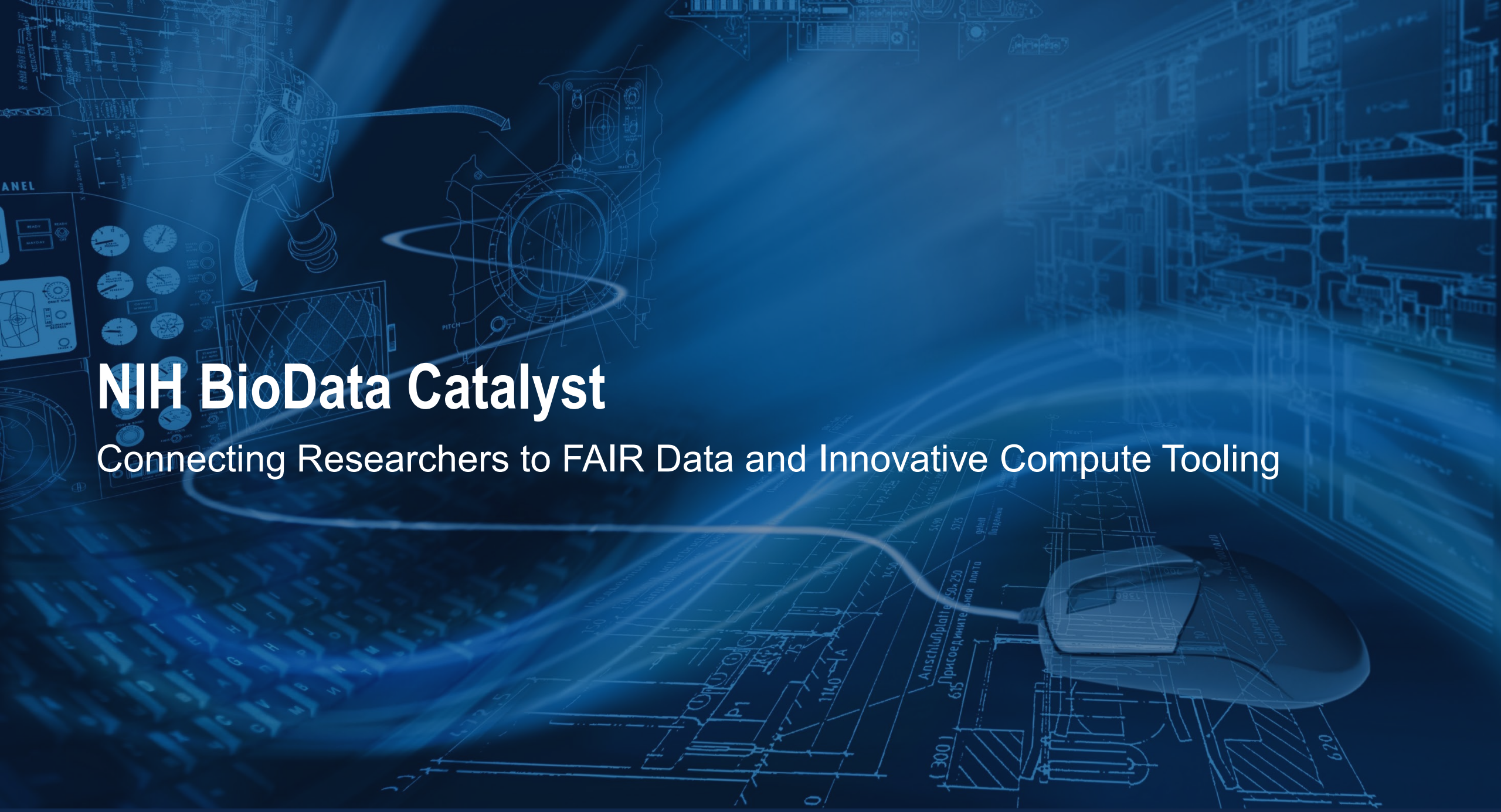
What's In Our Toolkit?

- Data governance and strategy
- Data-centric portfolio analysis
- Data management consulting service
- Software and system governance
- User engagement, outreach, and training
- Use case-driven standards adoption
- Cloud and microservice architecture transformation
- Secure interoperability solutions

Acknowledgements

Support for this work was provided by the National Institutes of Health through the Data Commons program (award 1OT3OD025646-01), NHLBI BioData Catalyst (award 3OT3HL147154-01), NIH HEAL (award 3OT2OD031940-01), and NCPI (1OT2OD034190-01 award).

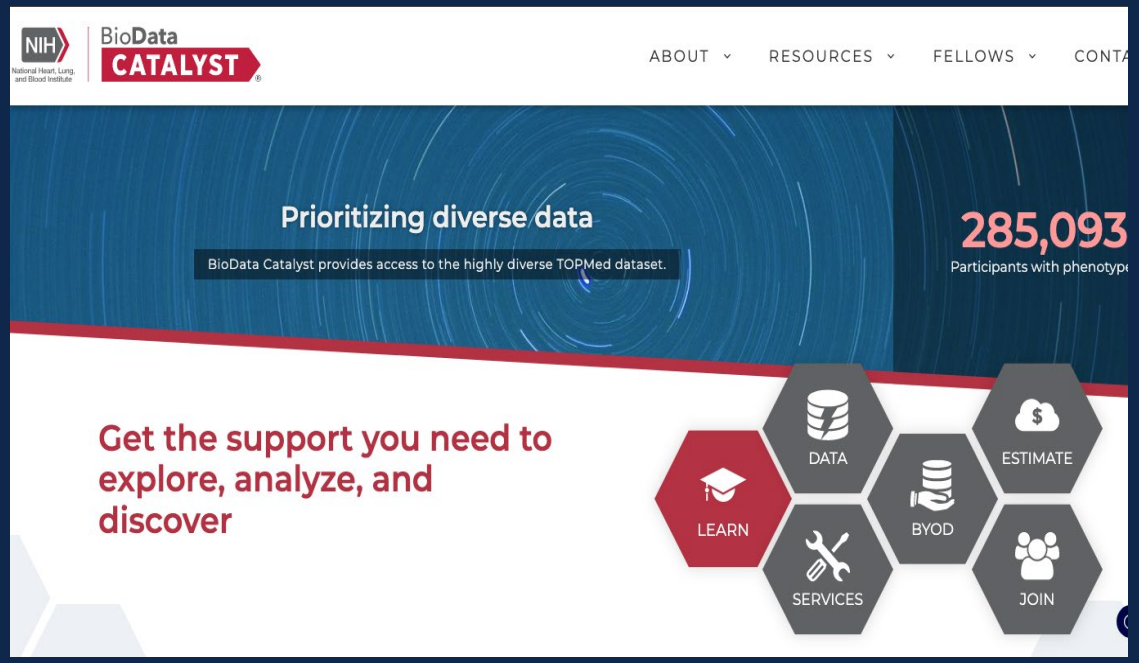
Any opinions expressed in this document are solely our own and do not necessarily reflect the views of NIH or affiliated organizations and institutions.



NIH BioData Catalyst

Connecting Researchers to FAIR Data and Innovative Compute Tooling

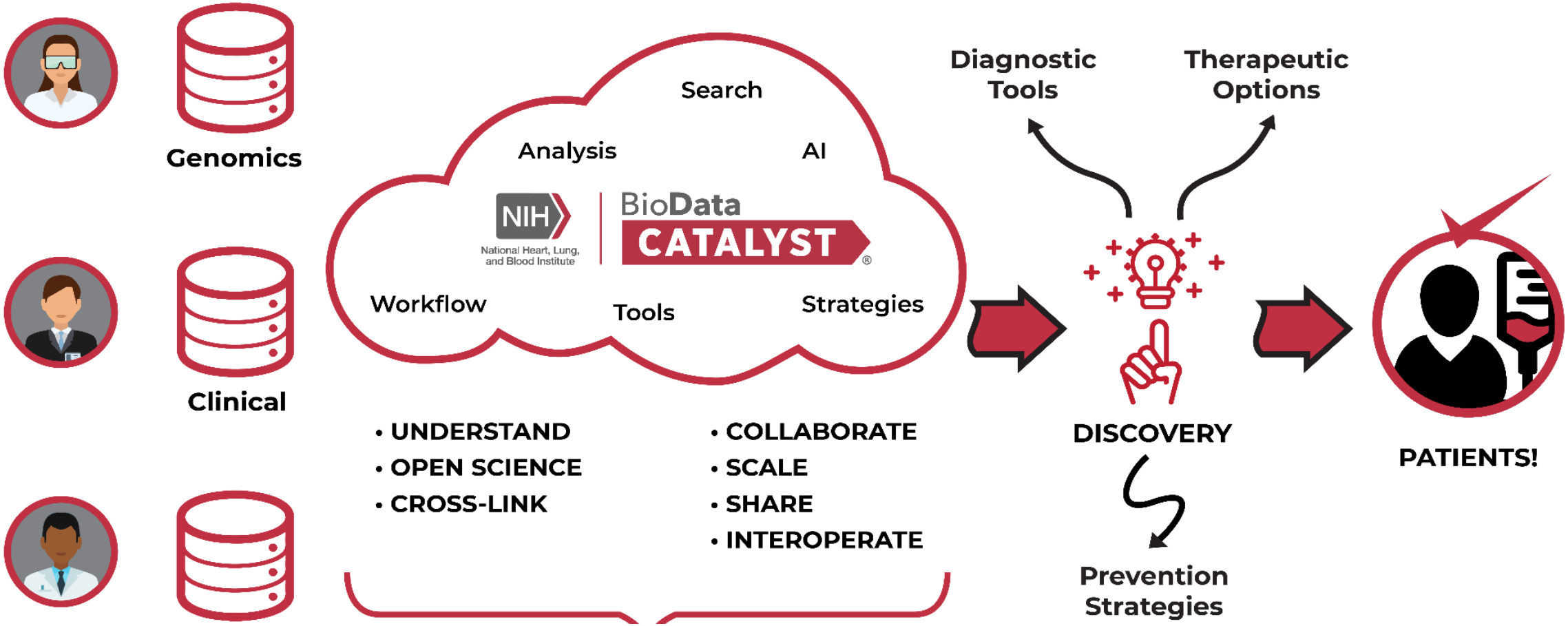
What is BioData Catalyst?



- A community-driven ecosystem where researchers can find, access, share, store, and compute on large-scale data sets
- Secure cloud-based workspaces, search tools and workflows, and applications to address community needs
- Exploratory data analysis, genomic and imaging tools, tools for reproducibility, and adoption of open APIs to enable data exchange

Community-Driven Mission

The ***mission*** of BioData Catalyst is to develop and integrate advanced cyberinfrastructure, leading edge tools, and FAIR data to support the NHLBI research community.



Data Strategy and Implementation Planning

- Transparent coordination of multiple software development teams
- Implementation of openAPIs and microservice architecture
- Defining and facilitating (meta)data standard and collection
- Support for community development efforts

Software and System Governance

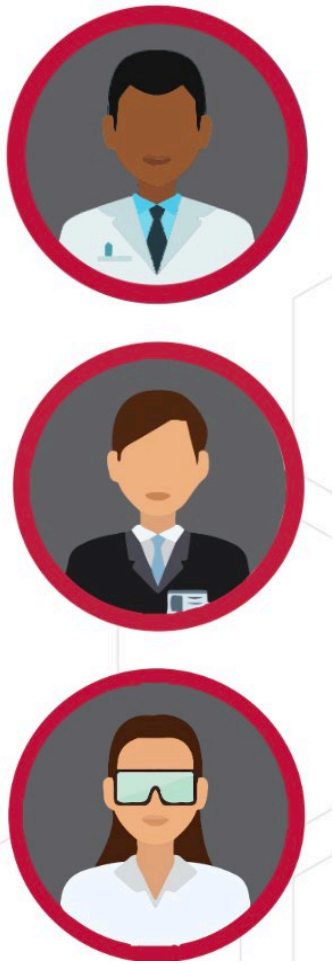
- Data and software release management
- Change control management
- Cloud cost modeling

User Engagement & Training

- Early Career Fellows program
- White glove consulting for data providers and users
- Helpdesk and user support

We Are Achieving the Mission by Applying Tools from the RTI Toolkit

This Approach Adds Extraordinary Value to Data and Brings People Together



- Access to **hundreds of terabytes** of heart, lung, blood, and sleep data in the cloud (based on data access approvals)
- **Team collaboration** on data
- Users bring data, tools, and workflows **to the data**
- Hundreds of optimized plug-and-play tools, allowing researchers to **focus on science**, not technology
- **Democratizes** access to data and compute resources
- More done in less time: **faster computing** in the cloud
- **Support**, tutorials, and documentation to help users navigate the system
- **Discounted cloud** to new users of the system
- Continues to evolve according to **user community needs**!

Developing the Workforce: Relevance for Public Health

For the Researcher:

- Offers early-career researchers **funding for novel and innovative research**
- Address a **scientific topic that can be answered** using BioData Catalyst
- Conducts analysis toward **publication**

For the System:

- **Improves the ecosystem** based on Fellow feedback
- **Contributes to the functionality** of the ecosystem
- Contributes to **diversity** across fields of study, institutions, geography, and investigators

Alexander Bick, MD, PhD

CHIP expansion and CVD

Melissa Cline, PhD

Genetics of Cardiomyopathies

Jacqueline Dron, PhD

Genetics CAD Lifecourse

Einat Granot-Hershkovitz, PhD

Ancestry-enriched Variants and CVD

Jamie Murkey, MPH

Psychosocial stress and CVD

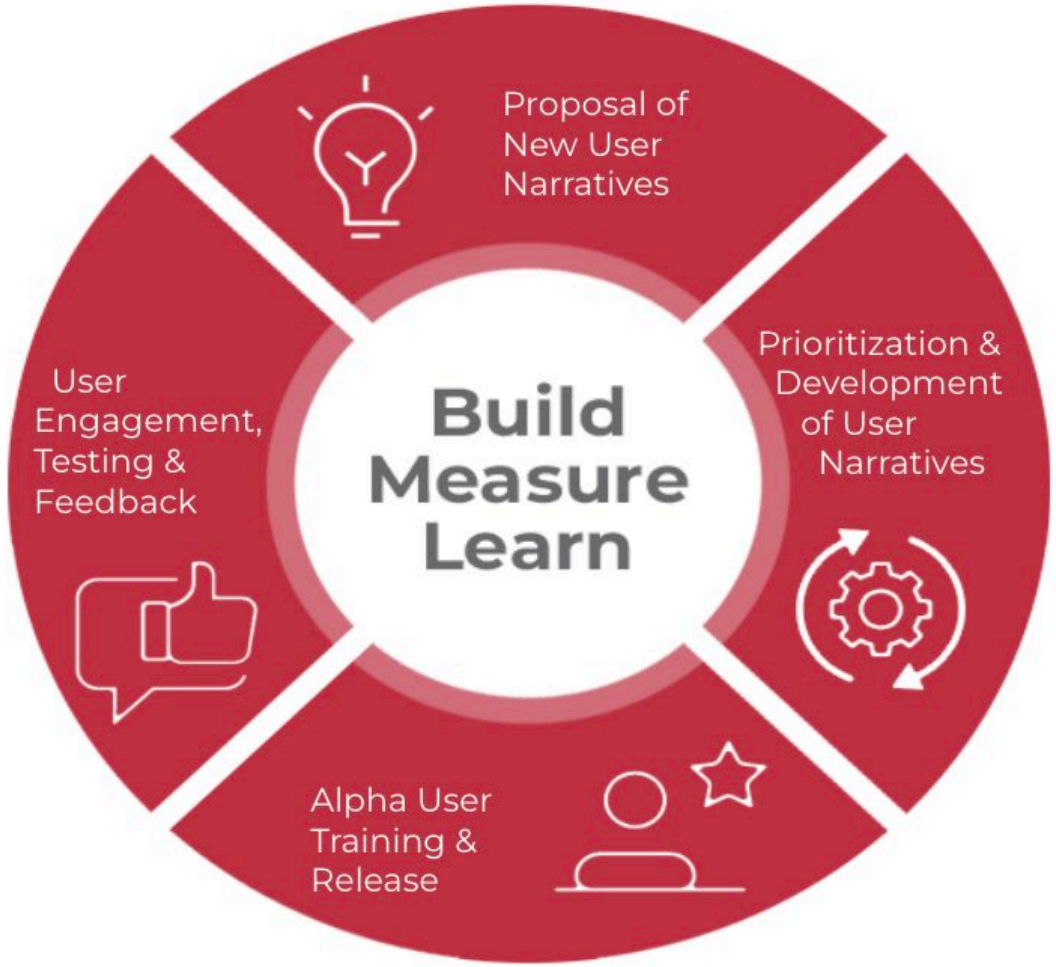
Yaling Tang, MD

Transcriptomics in Heart Failure

Xuefang Zhao, PhD

Structural Variants and Lipids

Continuous Learning with Our Partners Leads to Engagement and Success




NIH HEAL Data Stewards
Building the data infrastructure to address the opioid epidemic

RTI is Collaborating with UNC and NIH to Build the HEAL Data Ecosystem

- Supporting the $2B HEAL Initiative
- Over 1,000 awards
- Generating a large and diverse array of research data
- Accelerating sharing of HEAL-generated data and results among the broader community
- Drive clinical implementation
- Making data FAIR (findable, accessible, interoperable, and reusable)

Building the Data Infrastructure to Address the Opioid Epidemic

Data Strategy and Governance

- Review of legacy data repositories
- Define and facilitate (meta)data standards and collection

Data-Centric Portfolio Analysis

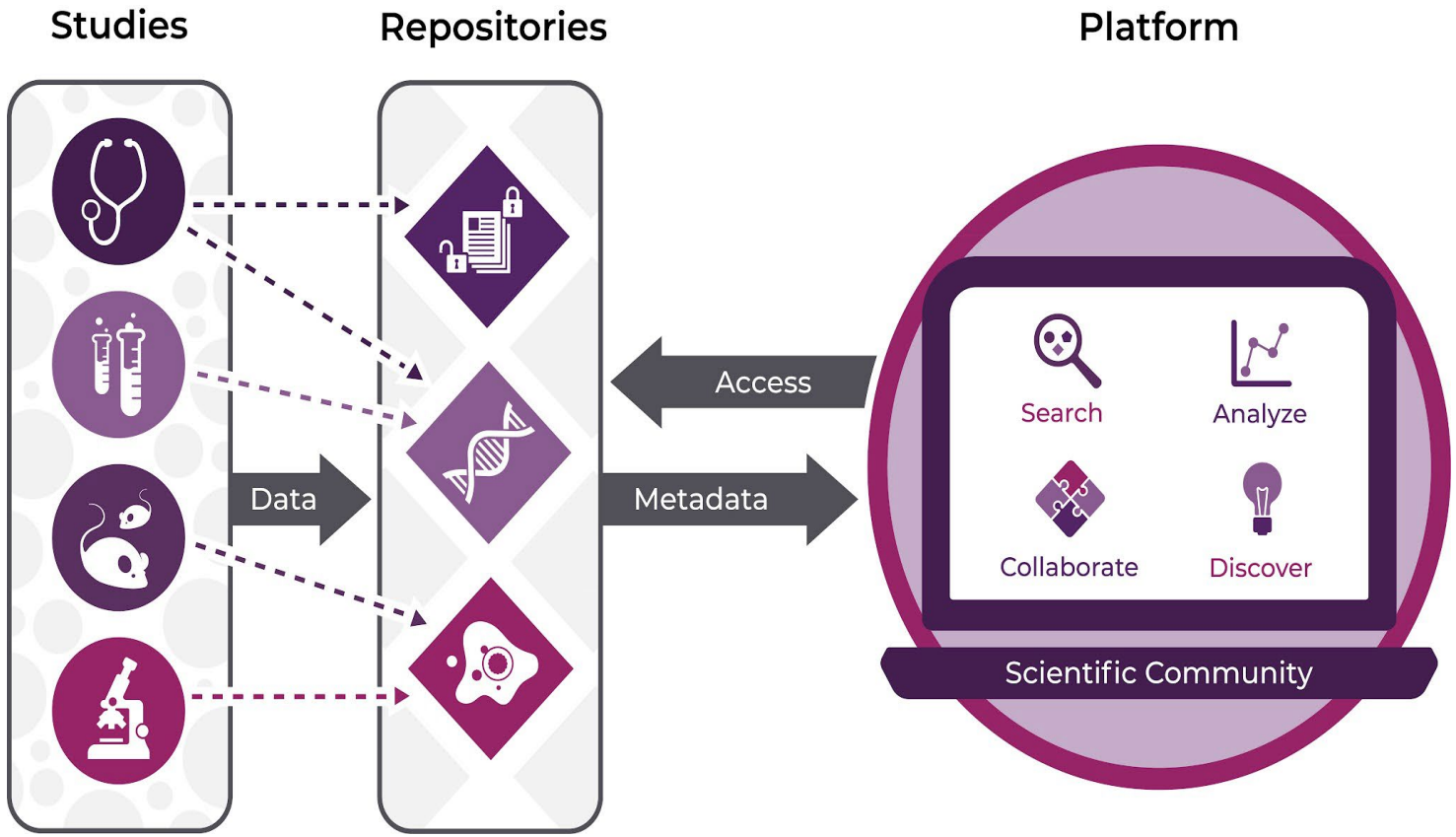
- Data Asset Inventory
- Portfolio dashboard
- Artificial Intelligence (AI) tools to search

Data Management Consulting Service

- White glove consulting for data providers and users
- Data privacy and ethical reuse
- Fresh FAIR webinar series to upskill workforce

Developing a Data Strategy

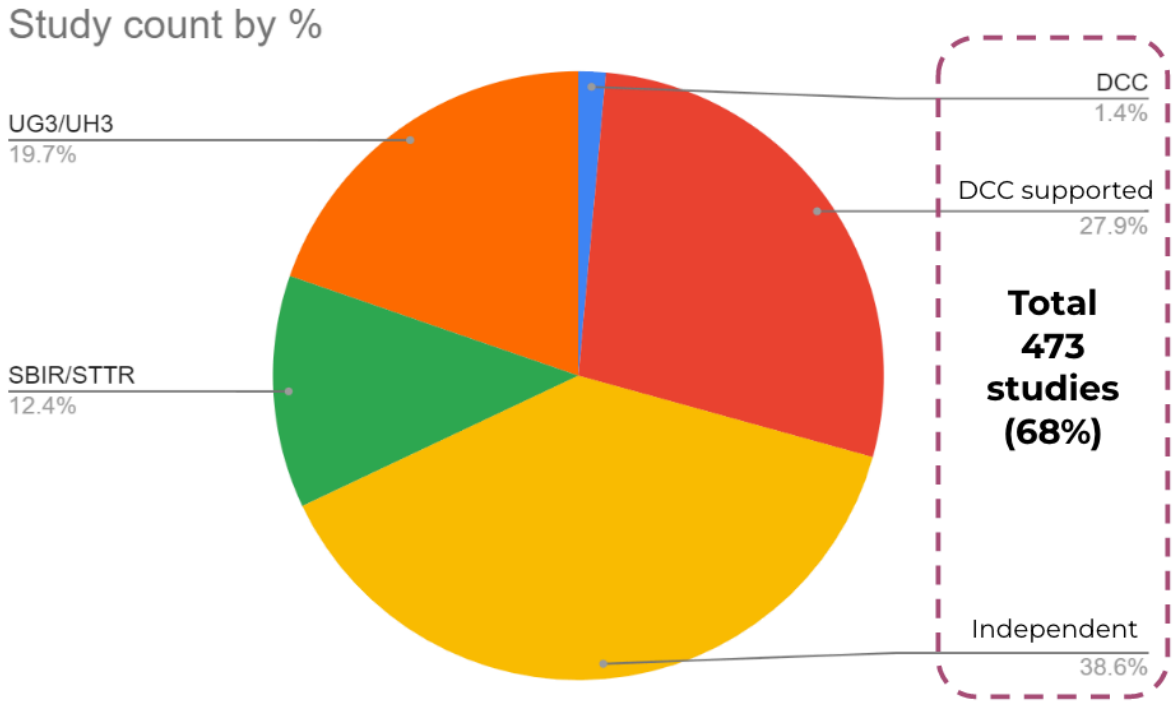
- Data distributed across 26 HEAL-compliant repositories
- Findable and Accessible through HEAL Data Platform metadata catalog
- HEAL Stewards work to provide solutions for managing the diverse data across the HEAL Data ecosystem



Data-Centric Portfolio Analysis

Capturing the landscape of data providers enables NIH to understand the current and future state of data assets.

- Plan for new data types and analysis needs
- Align resource allocation
- Upskill and train in advance
- Identify barriers early



AI Powered Search and Inferencing: A Biological Lens on Data

Researcher

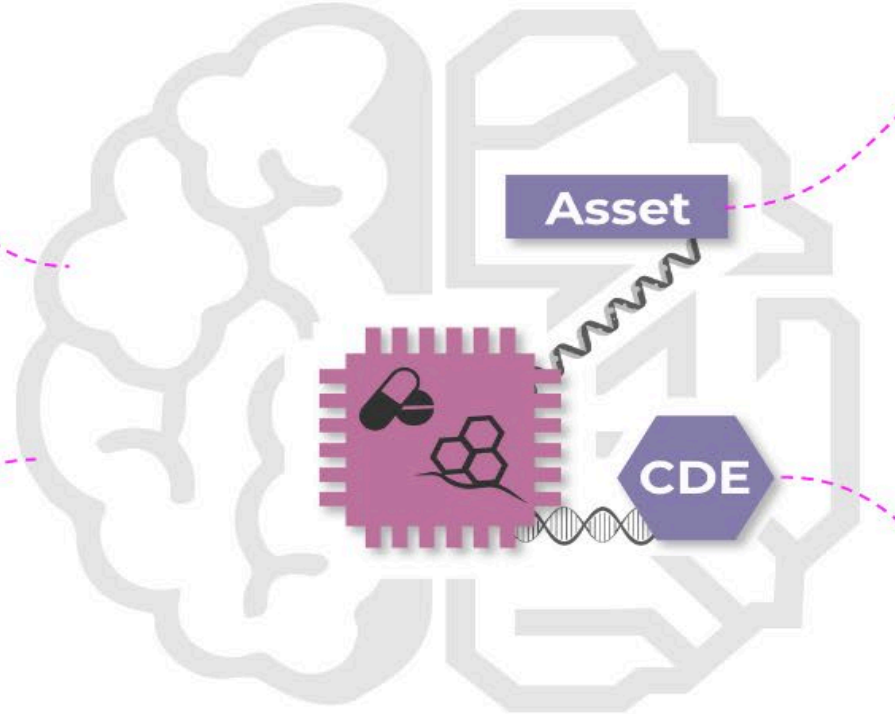
What exists?

How is it related?

Can it be harmonized?

Data Steward

Think



Analyze

Harmonize

We Meet People Where They Are

DATA MANAGEMENT CONSULTANCY

Identify **data access and user support challenges and opportunities**, including statistical analysis.

Through **specialized consulting,** understand the diversity of data and integrate appropriately into overall management workflows, protocols, and plans.

Create **data sustainability plan** in collaboration with HEAL researchers to assist with making data FAIR.

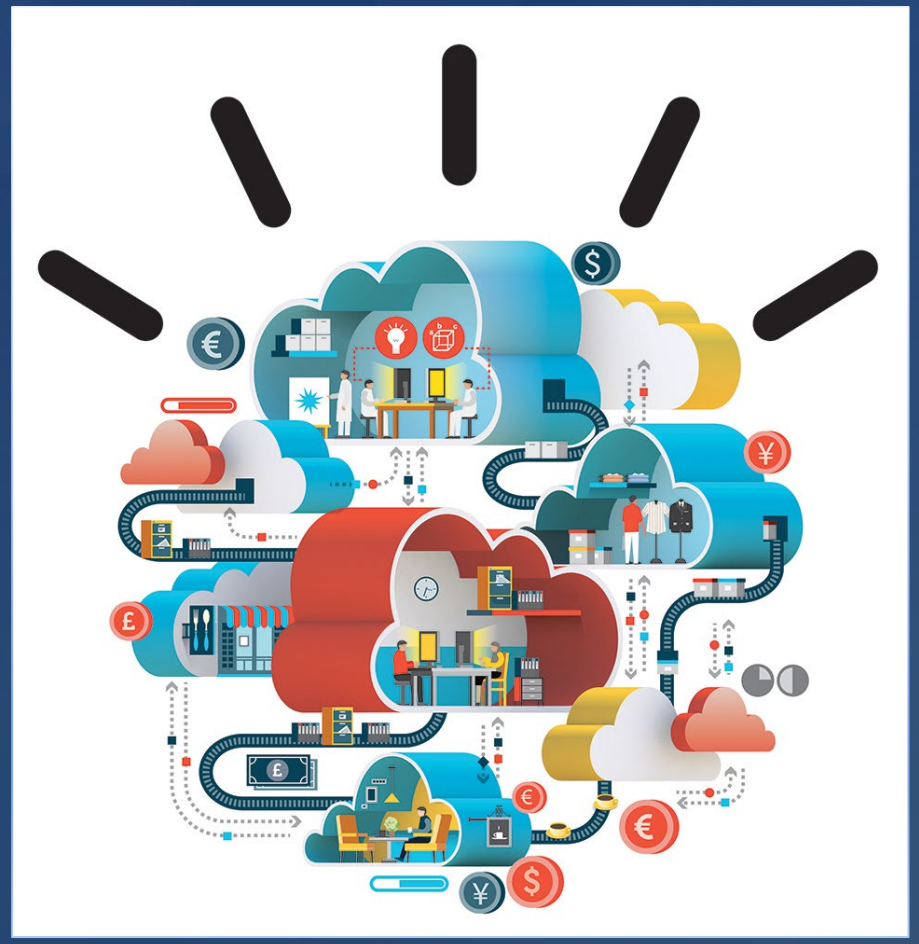
NIH Cloud Platform Interoperability

Implementing cross-platform, cross-domain interoperability

De-Siloing: Integrated Cloud Ecosystem

The NIH NCPI effort aims to establish and implement guidelines and technical standards to empower end-user analyses across participating NIH cloud platforms.

- FHIR and other APIs
- Cross-platform governance
- Cross-platform security
- User training and workforce development



Critical Data Modernization Efforts

Data Governance

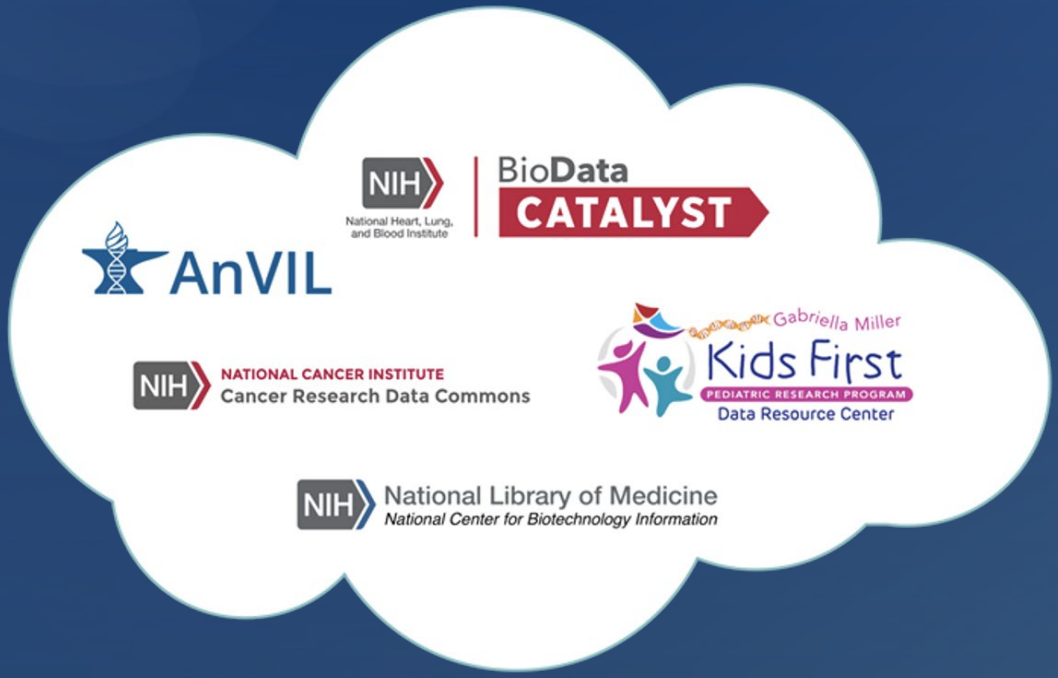
- Craft data governance framework
- Identify and address policy barriers to trans-platform data sharing
- Create principles for interoperability across security boundaries

Use Case-Driven Standards Adoption

- HL7 FHIR for exchange of clinical and phenotype data
- Common AAI approach
- Genomic GA4GH API adoption
- User testing for acceptance
- Federated semantic search

User Outreach and Training

- Public dataset catalog
- Cloud cost and FHIR trainings
- Early Career Fellows program

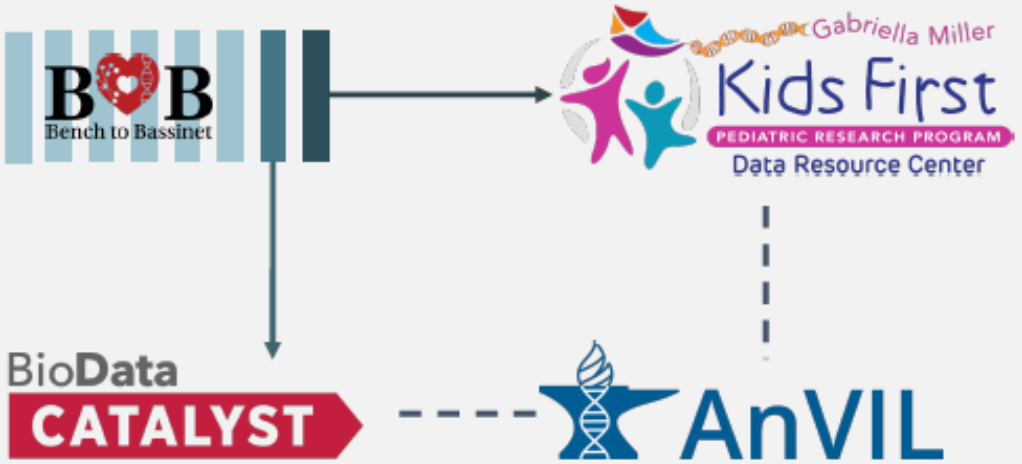


Toward an Integrated Ecosystem

PROBLEM STATEMENT

Pediatric Cardiac Genetics Consortium (PCGC)
an observational study of participants with congenital heart defects (CHD).

Gabriella Miller Kids First Pediatric Research Program (GMKF) **stores a subset of the PCGC** project in the Kids First Data Resource Center (KFDRC).



TOPMed program includes an additional subset of the PCGC project, representing up to 3230 participants which is hosted in BioData Catalyst Ecosystem (BDCatalyst).

With Anvil, this work enables data access and analysis across three data platforms/ ecosystems, with the goal of bringing together the PCGC data for the first time in the cloud for researchers.

EARLY ACCOMPLISHMENTS

Collaboration

The team has been collaborating on an approach to index and provide this data through the KFDR while a user is in BDCatalyst platform & vis versa.

Index Retained

TOPMed PCGC (phs001735) and GMKF PCGC data continue to reside in respective cloud systems and each program maintains their own index of the data.

User Experience

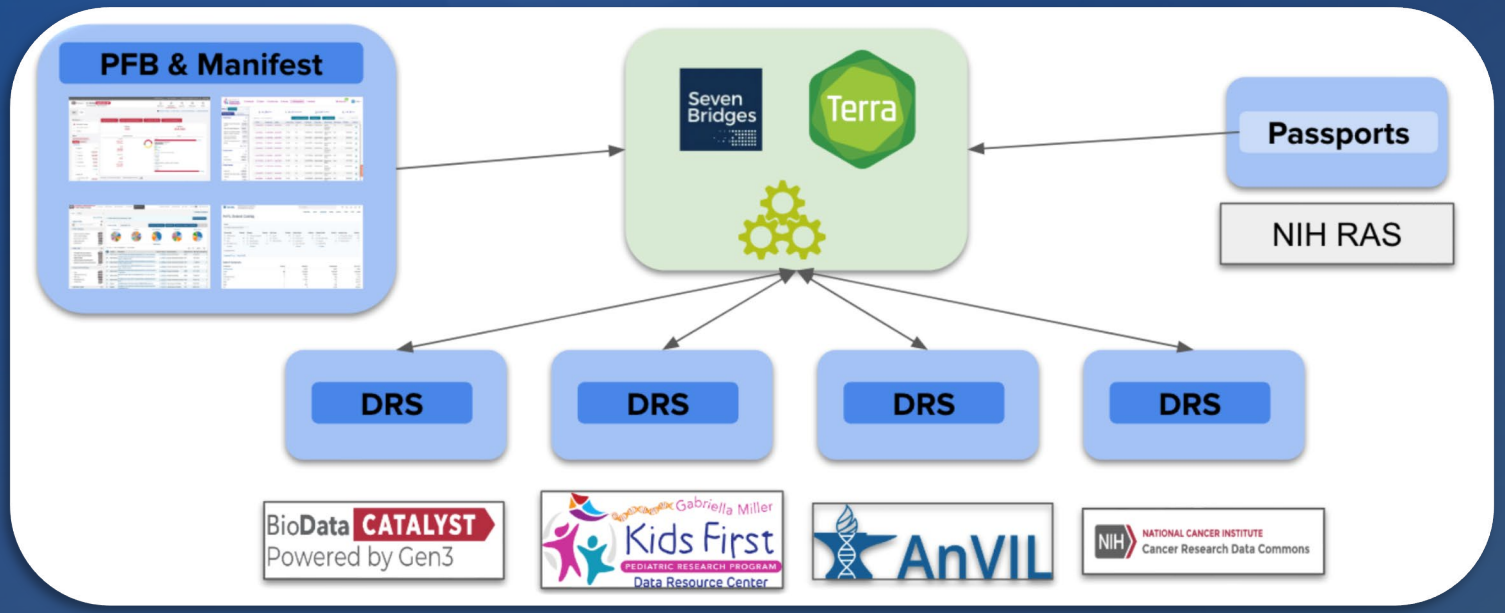
Users will be able to see both TOPMed PCGC and Kids First PCGC **in either portal interface**

Search and Export

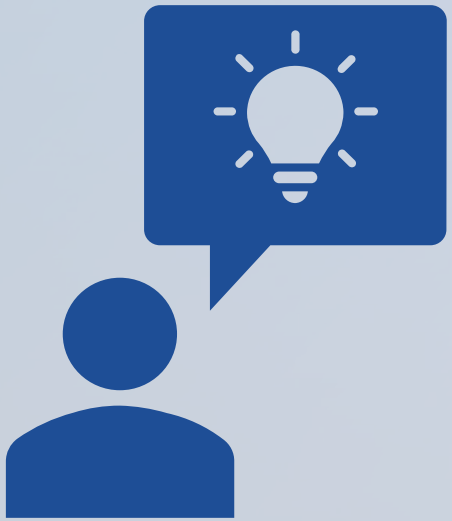
Data can be searched and exported via the PFB on FHIR convention with GA4GH DRS used for **data access and analysis in either KFDRC the BDCatalyst workspace environments** (Terra and SevenBridges).

Dr Wilson analyzes sex chromosomes across all participating systems in a distributed analysis.

Data is not copied; ownership is retained by the original system.

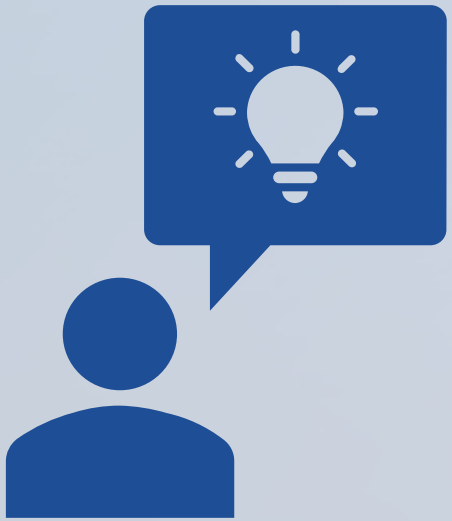


Lessons from the Data Commons can inform **public health data ecosystem development.**



- At its heart, this work is **people-centric.** The technology succeeds when there is a dedicated, cross-domain community around and supporting a Commons.
- Coordinating distributed networks of contributions to an ecosystem requires radical transparency and clear lines of decision-making.
- Successful efforts engage national and international forums, even if that slows down implementation.

Lessons from the Data Commons can inform **public health data ecosystem development.**



- A collaborative and clearly articulated data governance framework is essential.
- System and interoperability approaches must be use case-driven.
- Alpha users need to be compensated for their time and effort.
- Cost and ease of use are common drivers.

Future Direction

- Continue people-first, tech-forward approach.
- Continue partnering with the scientific community to create efficient, modular, connected, and scalable data ecosystems that meet users' need on demand and in real time.
- Build partnerships with public health practitioners, applying lessons learned and early methods to create a foundation for a successful DMI initiative.
  - Data Asset Inventory
  - Portfolio Analysis
  - Use Case Collection
  - Stakeholder engagement, including STLTs