Cancer Research Data Commons (CRDC) Ecosystem Overview

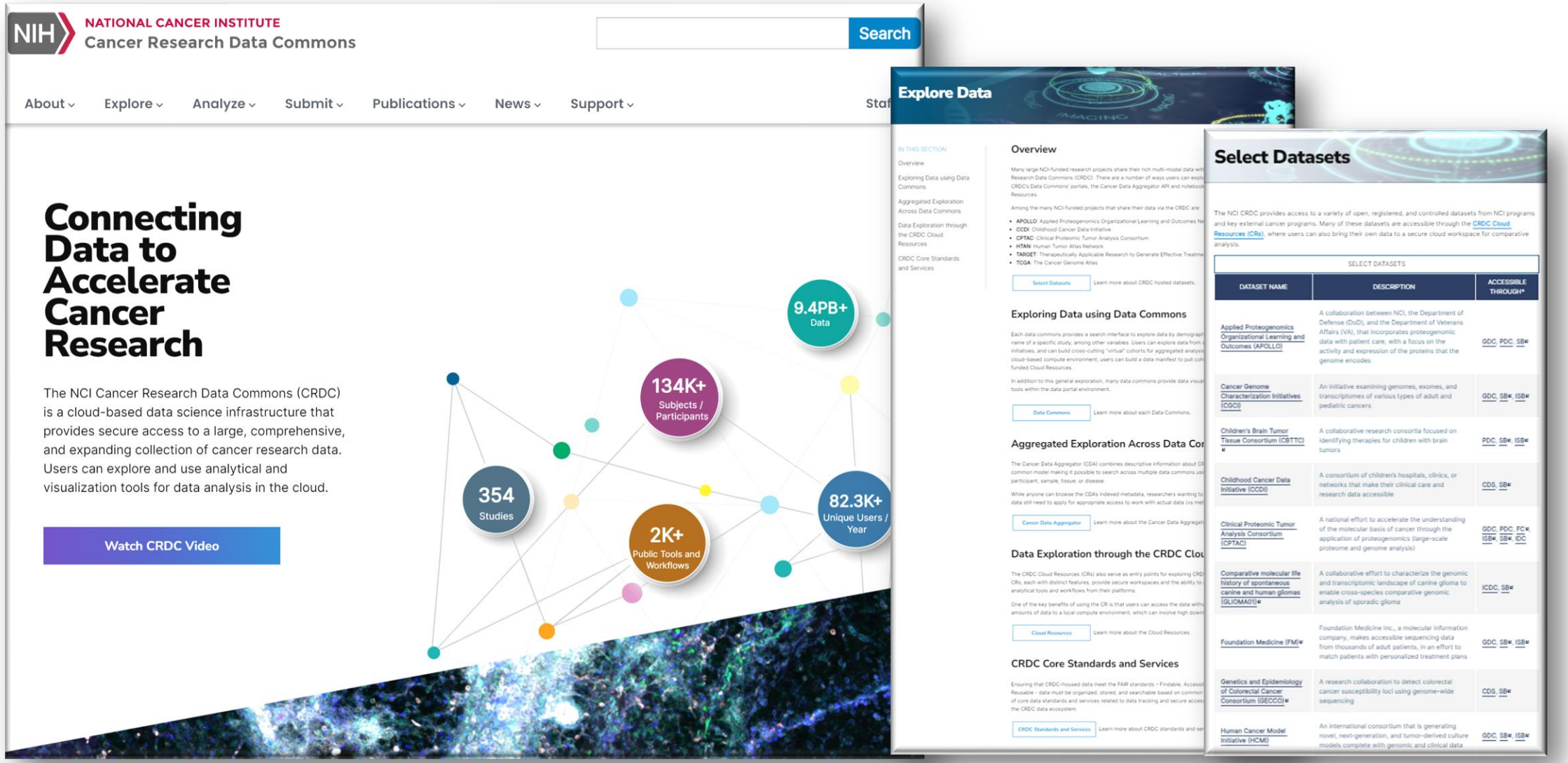
Tanja Davidsen, Ph.D.

October 17, 2024

NIH NATIONAL CANCER INSTITUTE

CRDC Website: datacommons.cancer.gov

Resources: https://datacommons.cancer.gov/resources



CRDC Insights: Quarterly Newsletter

https://datacommons.cancer.gov/crdc-insights

NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

NIH Data Management and Sharing Policy: CRDC's Role

The new policy is in effect, and applies to new grant applications, competitive renewals, or competitive revisions. In brief:

- Data sharing now pertains to all researchers with no budget minimum.
- Applications, renewals, or revisions must include a data management and sharing plan.
- Data must be shared at time of publication or by the end of the performance period, whichever is sooner.

The Cancer Research Data Commons (CRDC) is home to a collection of data commons and cloud resources that host datasets from NCI-funded research, and make those datasets accessible to the research community. Learn more about submitting and accessing data, and using CRDC tools for your research. {link to data landing page}

CRDC Resources in the Classroom

Faculty members and data scientists with Purdue University and Velsara (Seven Bridges) teamed up to produce a four-part online workshop that introduces the Cancer Genomics Cloud. The series also provides hands-on lessons in bulk- and single-cell RNA-seq analysis using datasets provided by Purdue researchers. *Read the full story.*

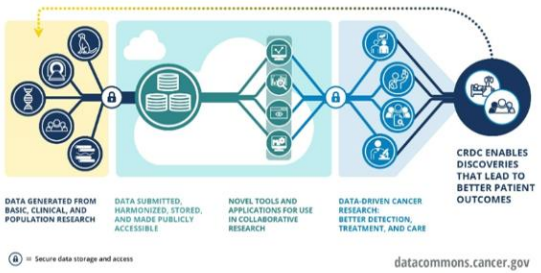
HTAN: Methods Workshop at AACR 2023 Annual Meeting

The Human Tumor Atlas Network (HTAN) is working closely with CRDC to ensure long-term legacy and reuse of HTAN data, and to share data through NCI's Cloud Resources. This methods workshop will demonstrate how to access, query, use data within the cloud environment, and visualize HTAN data derived from a variety of assay types. Read more about this workshop on the AACR Annual Meeting website.

Announcement: Funding Opportunities

The Office of XYZ has released a RFP/grant solicitation regarding data interoperability. Find more information on their Interoperability Initiative page.

CRDC: Empowering the Scientific Community to Make New Discoveries



A new infographic illustrates how CRDC supports the work of cancer researchers. This is available for use in presentations. Contact our general email box below.

In the News

NCI Director Monica Bertagnolli was recently interviewed by National Public Radio (NPR) about the work of the NCI and its impact on patients and families. She also discussed her own cancer diagnosis and her commitment to participating in research trials. Listen here through the NPR website.

About the Cancer Research Data Commons

The NCI Cancer Research Data Commons (CRDC) is a cloud-based data science infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data. Users can explore and use analytical and visualization tools for data analysis in the cloud.

Subscribe to this newsletter here: LINK TO SUBSCRIBE button on our news page.

Quick Links

Data Releases: Updated March 2023

A round-up of new datasets available through our data commons and cloud resources.

Datasets, Access, and Submission

Aggregated information across CRDC data commons and cloud resources about filing a data submission request, and how to access data currently housed in a CRDC data commons or on a cloud resource.

Getting Started

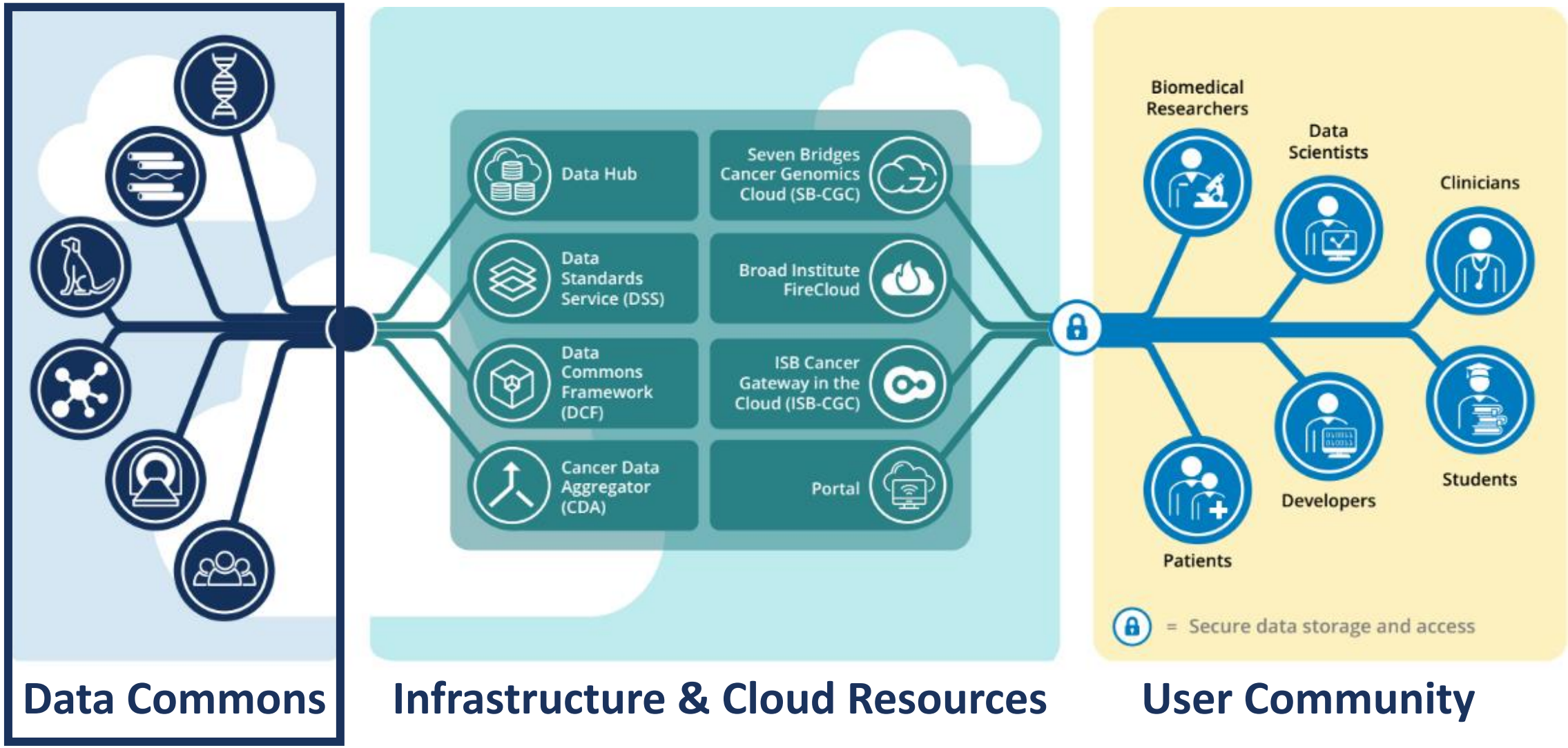
Aggregated listings of user manuals, tutorials, and virtual office hours, across CRDC data commons and cloud resources.

Contact us: NCICRDC@mail.nih.gov

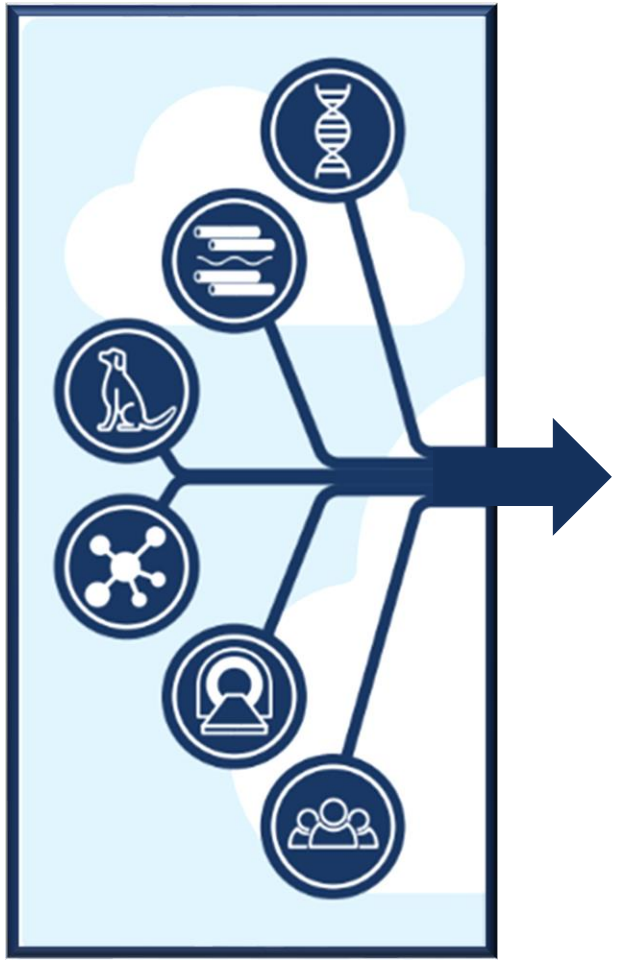
And subscribe to this newsletter on our CRDC Insights page.

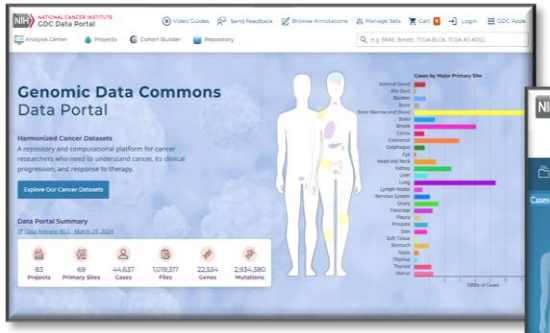
CRDC Ecosystem: Data Commons

CRDC Ecosystem: Data Commons



CRDC Ecosystem: Data Commons

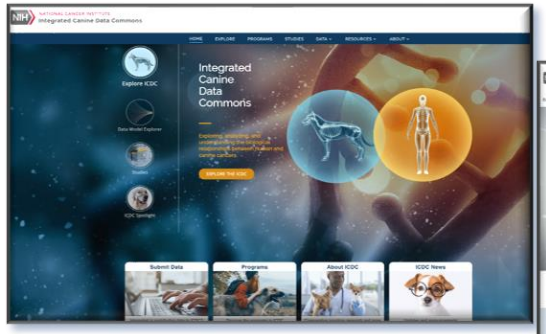




Genomic Data Commons (GDC)

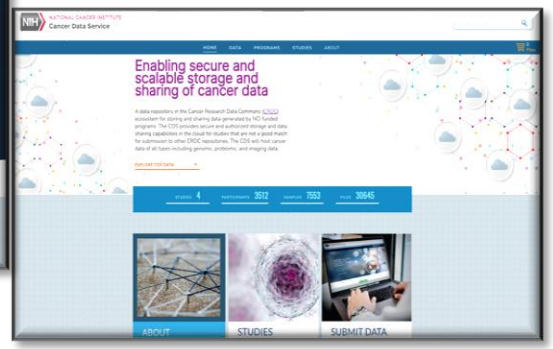
Proteomic Data Commons (PDC)

Imaging Data Commons (IDC)



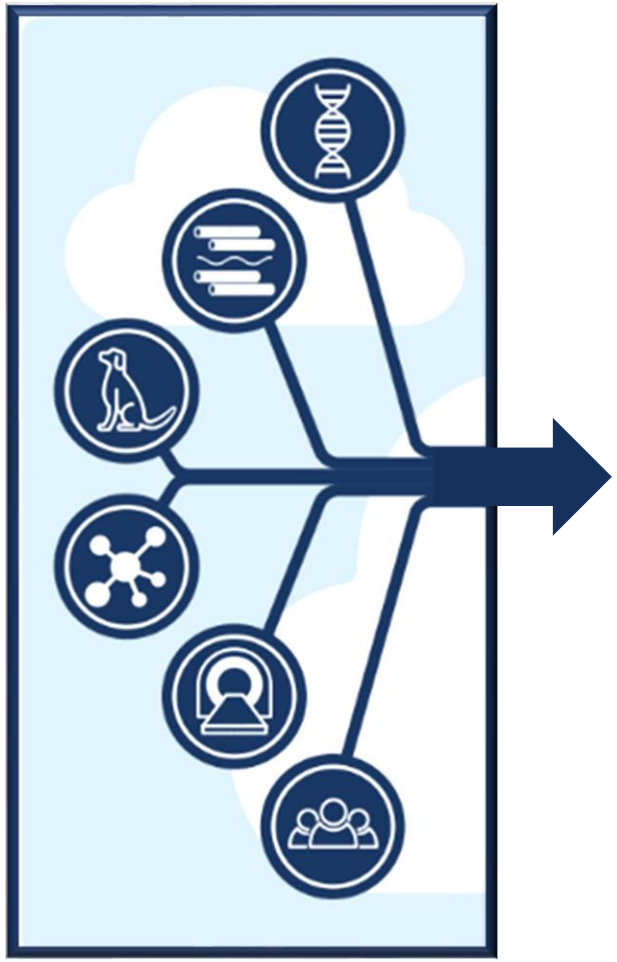
Integrated Canine DC (ICDC)

Clinical & Translational DC (CTDC)



Cancer Data Service (CDS)

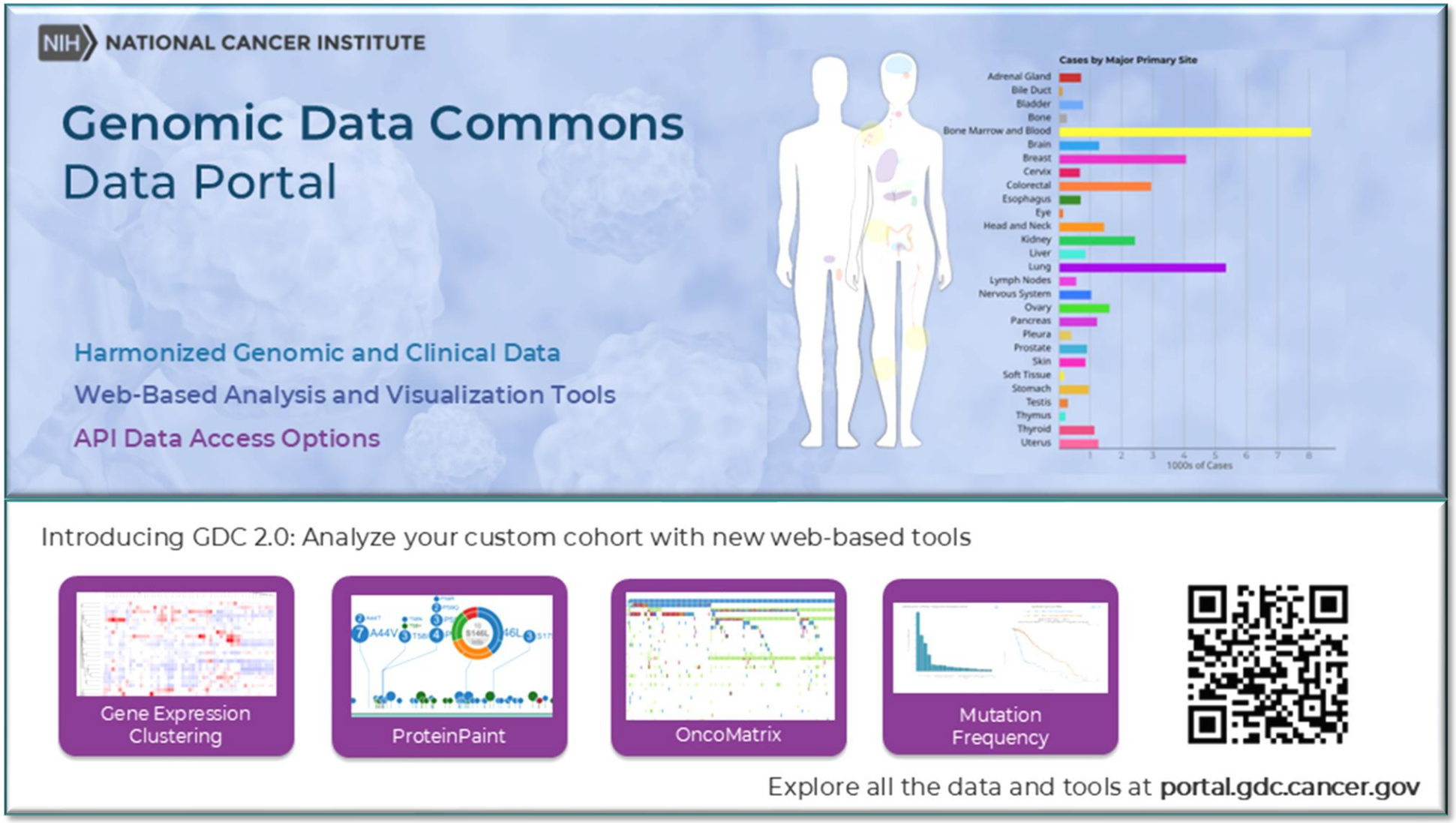
CRDC Ecosystem: Data Commons



Coming Soon…
Population Science Data Commons currently in development and testing

Genomic Data Commons (GDC)

https://portal.gdc.cancer.gov



Genomic Data Commons (GDC)

https://portal.gdc.cancer.gov

- **Data:** Filter, query, visualize, analyze & download
- **Harmonized** to same genome standard and variant calling pipeline
- **New:** WGS Variant Callers, Copy Number Variation, new single cell RNA-Seq data, new whole slide images
- **Programs:** TCGA, TARGET, CPTAC, MATCH, more
- **GDC 2.0 Launched June 2024:** Empowers researchers with a cohort-centric approach
  - **Custom Cohort Creation & Data Download**
  - **Advanced Analysis Tools**
  - **SDK for Tool Integration**

Users: >90K+ unique users per month
Citations: 2,500+

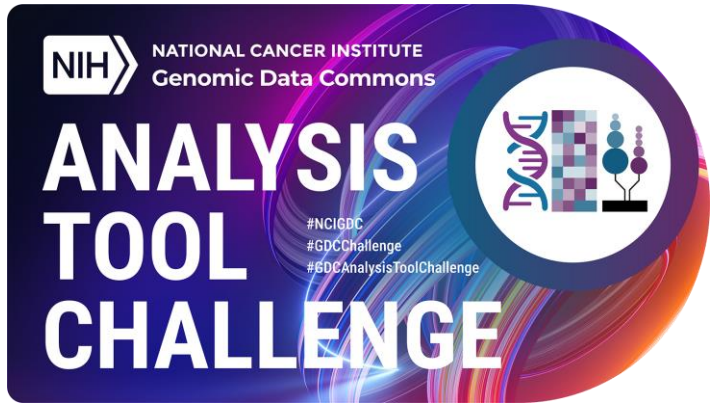
Downloads: >4PBs data downloaded per month

Global Access: Users from >90 different countries

GDC Analysis Tool Challenge

Join the GDC Analysis Tool Challenge and help shape the future of cancer research by integrating a novel analysis tool to explore GDC data

- **Requirements:**
  - **Scientific Need & Innovation:** Integrate a tool that addresses a scientific need and goes beyond existing GDC Analysis Tools
  - **GDC Integration:** Utilize GDC data and leverage the GDC SDK for tool integration
  - **Open-source:** Integrated tools must support open-source distribution



Register today!

Closes October 31, 2024

https://www.challenge.gov/?challenge=nci-gdc-analysis-tool-challenge

Proteomic Data Commons (PDC)

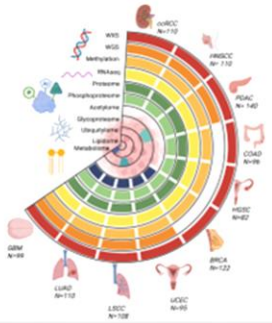
https://pdc.cancer.gov

Proteomic Data Commons (PDC)

https://pdc.cancer.gov

- **Data:** Filter, query, visualize & download
- **Data Harmonization:** for uniform analysis of all PDC data
- **Programs:** CPTAC, ICPC, APOLLO, TCGA
- **New:** Now supports metabolomics & lipidomics data
- **PDC By The Numbers:**
  - 163 Studies
  - +130K Data Files
  - 43 TB Data
  - >1PB Download
- Data Analysis in the cloud facilitated by the NCI Cloud Resources

CPTAC's Pan-Cancer Multi-omic Papers & Data



Annual Clinical Data Updates

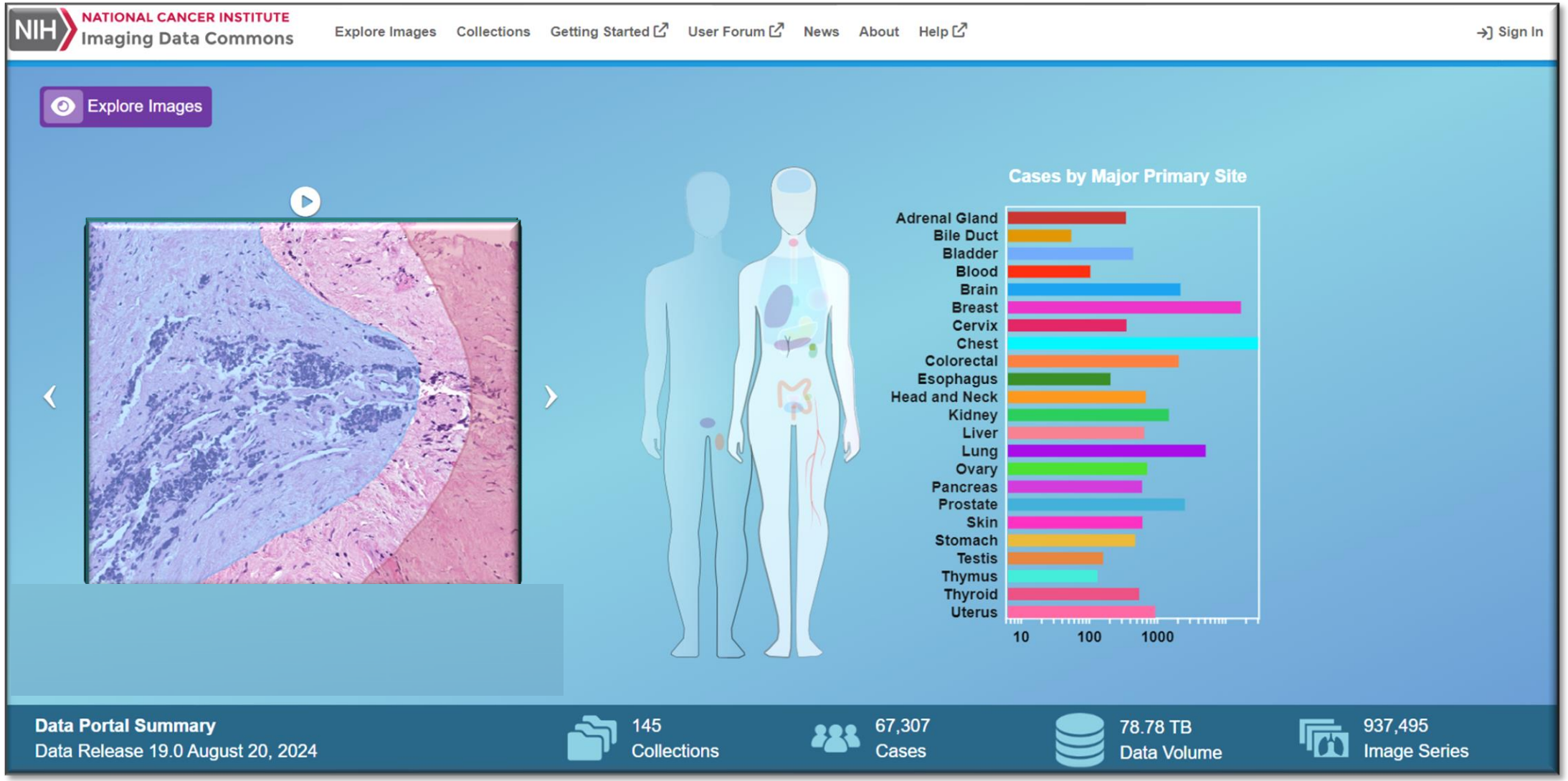
Cancer Research Communications
@CRC_AACR

Recently published:
NCI's Proteomic Data Commons: A Cloud-Based Proteomics Repository Empowering Comprehensive Cancer Analysis Through Cross-Referencing with Genomic & Imaging Data, by Ratna R. Thangudu et al.

https://doi.org/10.1158%2F2767-9764.CRC-24-0243

Imaging Data Commons (IDC)

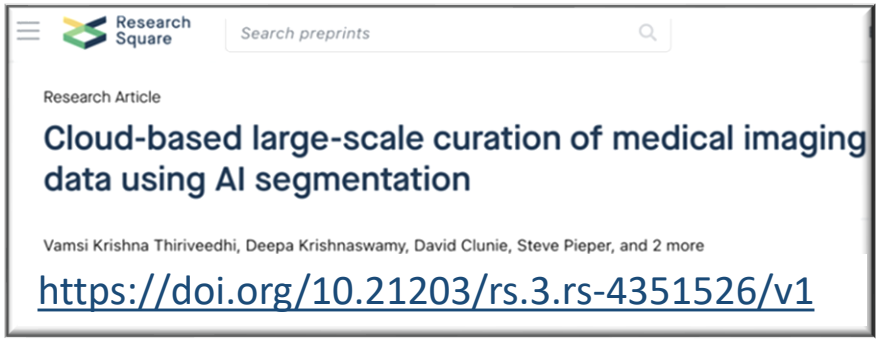
https://portal.imaging.datacommons.cancer.gov/



Imaging Data Commons (IDC)

https://portal.imaging.datacommons.cancer.gov/

- **Share, Analyze, Visualize:** de-identified, harmonized, multi-modal, open access imaging data
- **Programs:** CCDI, CMB, HTAN, GTEx, TCGA, CPTAC, more
- **Data:** Images and annotations
  - ~80TB data, ~1M images
  - DICOM standard
  - Interoperable & Searchable
- **Quantitative features:**
  - Lower barriers for analysis
  - Enable cross-omics studies
- **Continuous Data Enrichment:**
  - Expert and AI-based curation
  - Collaborations, benchmarking, comparison studies
- **Cloud-based computing at scale:**
  - Manual annotation not practical
  - Unparalleled ability to scale AI-based automatic curation
  - Enabled by Cloud Resources


Research Square

Search preprints

Research Article

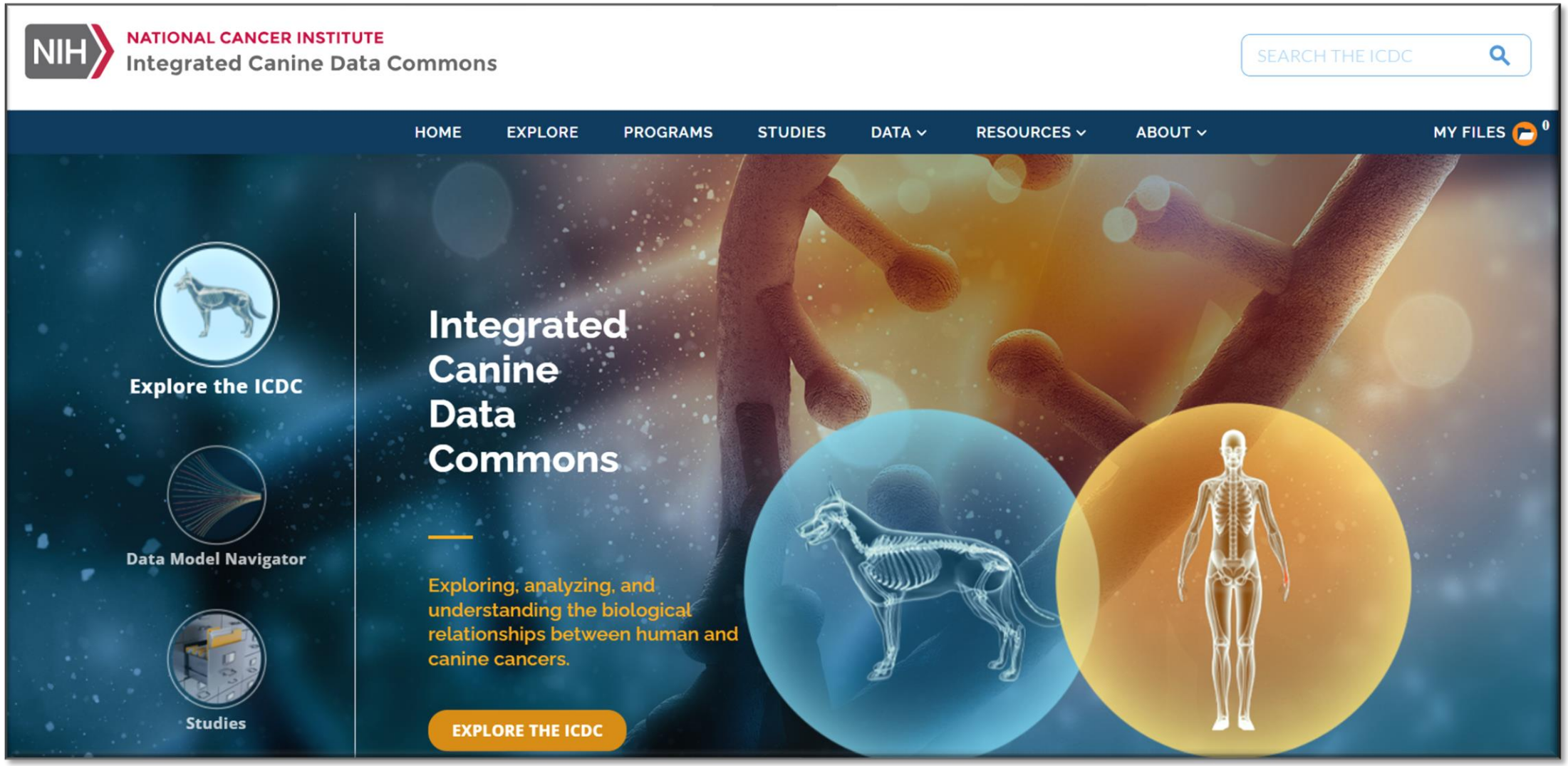
Cloud-based large-scale curation of medical imaging data using AI segmentation

Vamsi Krishna Thiriveedhi, Deepa Krishnaswamy, David Clunie, Steve Pieper, and 2 more

https://doi.org/10.21203/rs.3.rs-4351526/v1

Integrated Canine Data Commons (ICDC)

https://caninecommons.cancer.gov/



Integrated Canine Data Commons (ICDC)
https://caninecommons.cancer.gov/

- **Canine Clinical Trial Data:**
  - PRE-medical Cancer Immunotherapy Network Canine Trials (PRECINCT)
  - Comparative Oncology Program
- **Recent Accomplishments:**
  - "Leading the pack: Best practices in comparative canine cancer genomics to inform human oncology"
  - Direct export of genomic files to Seven Bridges Cancer Genomics Cloud (SB-CGC)
- **Future Directions:**
  - Database of annotated biospecimen locations, biobank contact information, and data from other sources like veterinary foundations
  - Longitudinal data
  - Include CDEs in Data Model
  - New studies: PRECINCT01, COTC021, OSA02, OSA04
  - **Data Submission:** via CRDC Centralized Data Submission Portal

Clinical & Translational Data Commons (CTDC)

https://clinical.datacommons.cancer.gov/

Clinical & Translational Data Commons (CTDC)

https://clinical.datacommons.cancer.gov/

- **Launched:** September 2024
- **Datatypes:** clinical, genomic, molecular, PRO, pharmacological and more
- **Access Levels:** Hosting open, registered, controlled access
- **Accepting Data:** from NCI-funded clinical studies (not just trials) including immuno-oncology data
- **First Dataset:** Cancer Moonshot Biobank
- **Future Datasets:** to include
  - Molecular Analysis for Therapy Choice (MATCH) Trial
  - Cancer Immune Monitoring and Analysis Centers (CIMAC)
- **Data Submission:** via CRDC Centralized Data Submission Portal

Cancer Data Service (CDS)

https://dataservice.datacommons.cancer.gov/

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

SEARCH CDS

HOME DATA PROGRAMS STUDIES ABOUT

0 Files

Enabling secure, scalable storage and sharing of cancer data

A data repository in the Cancer Research Data Commons (CRDC) ecosystem for storing and sharing data generated by NCI funded programs. The CDS Portal provides secure and authorized storage and data sharing capabilities in the cloud for studies that are not a good match for submission to other CRDC repositories. The CDS Portal will host cancer data of all types including genomic, proteomic, and imaging data.

EXPLORE CDS PORTAL

STUDIES	PARTICIPANTS	SAMPLES	FILES
30	49892	80230	329972

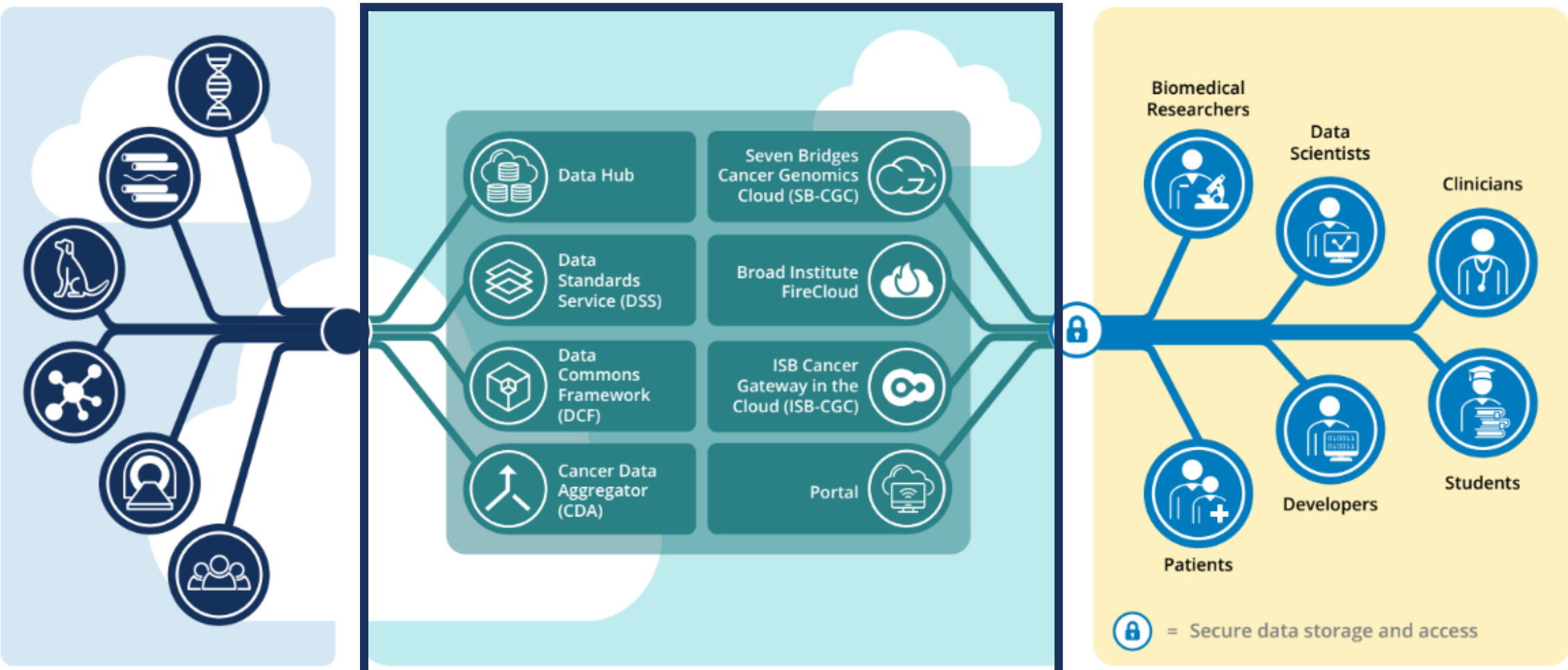
Cancer Data Service (CDS)

https://dataservice.datacommons.cancer.gov/

- **Cloud-Based Repository:** secure and scalable storage for open or controlled access data
- **Datatype Agnostic:** CDS has a flexible data model for storing and sharing a variety of datatypes
  - Genomic, proteomic, imaging data, others on request
- **Since January 2022:**
  - 2.5 PB data submitted
  - 1.998 PB data released
- **Supporting compliance** with NIH data sharing policies
- **Data Discovery:** CDS Portal allows users to explore cancer research open harmonized (meta)data
- **Cloud-Based Analysis Tools:** one click manifest export to SB-CGC
  - Seamless analysis in the cloud - bring your own tools and data
- **Programs:**
  - Human Tumor Atlas Network (HTAN), Childhood Cancer Data Initiative (CCDI), CPTAC, PDXNet
- **Data Submission:** via CRDC Centralized Data Submission Portal

CRDC Ecosystem: Infrastructure and Cloud Resources for Interoperability

CRDC Ecosystem: Infrastructure & Cloud Resources



Data Commons

Infrastructure & Cloud Resources

User Community

CRDC: Data Commons Framework

A re-usable, expandable framework

Core principles and structures

Modular components

Fence

Centralized Authentication & Authorization

IndexD

Centralized Indexing

Cloud Storage

CRDC Data in Amazon and Google

CRDC: NCI Cloud Resources

Democratizing access to cancer research data

- Access to **large cancer data sets** without need to download or move data
- Access to **workspaces, analysis tools, and workflows/pipelines**
- **Bring your own data and tools**: collaborative pre-publication workspaces
- Funds to get you started as a new user: **$300-$10,000**

ISB's Cancer Gateway in the Cloud

Great for command-line, BigQuery, Specialty DBs

Broad's FireCloud

Great for running production pipelines

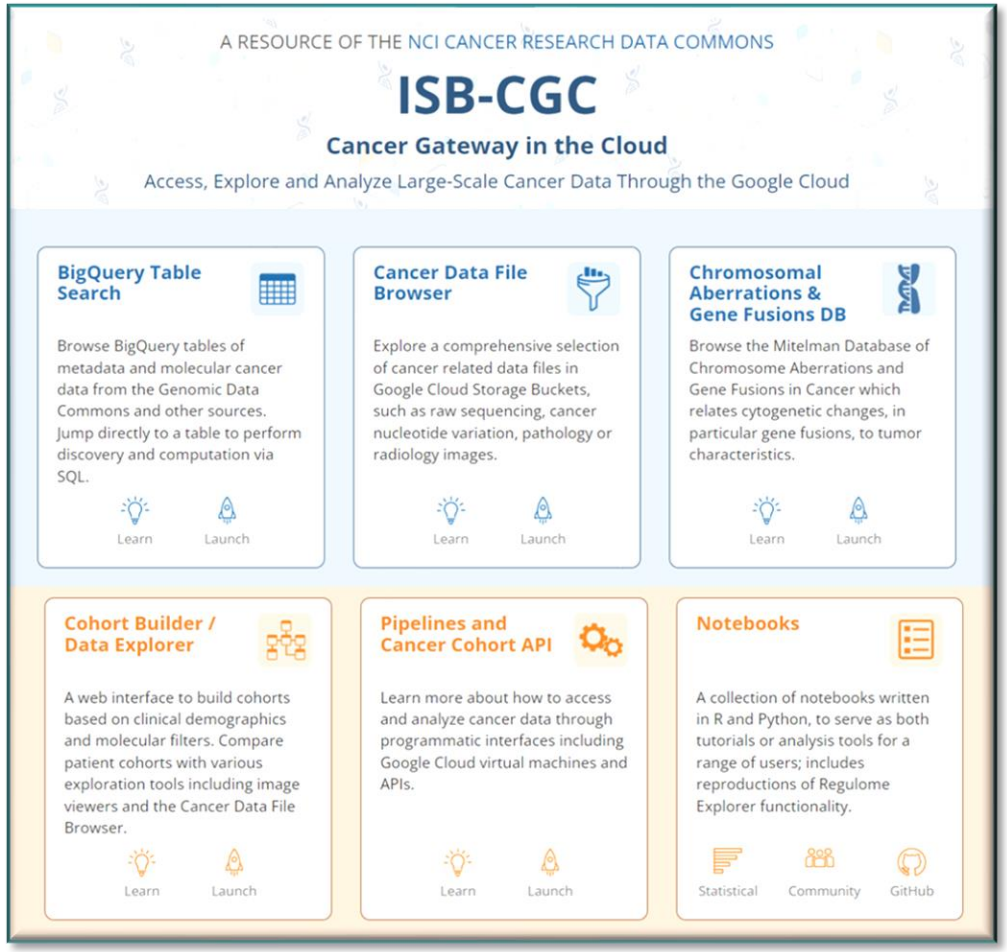
Seven Bridges' Cancer Genomics Cloud

Great for non-technical user Interface, visual displays

CRDC: NCI Cloud Resources

Institute for Systems Biology Cancer Gateway in the Cloud (ISB-CGC)

- **ISB-CGC's Cohort Builder:** combined with Cancer Data Aggregator can now generate cross-data-common cohorts
- **Google BigQuery:** Tabular CRDC data stored and browsable through BigQuery search tool
- **ISB-CGC-hosted Mitelman Database of Chromosomal Aberrations:** includes search optimizations and integrated data visualizations
- **Bioinformatics Notebooks:** ISB-CGC's growing collection includes guidance on how to use multiple datatypes

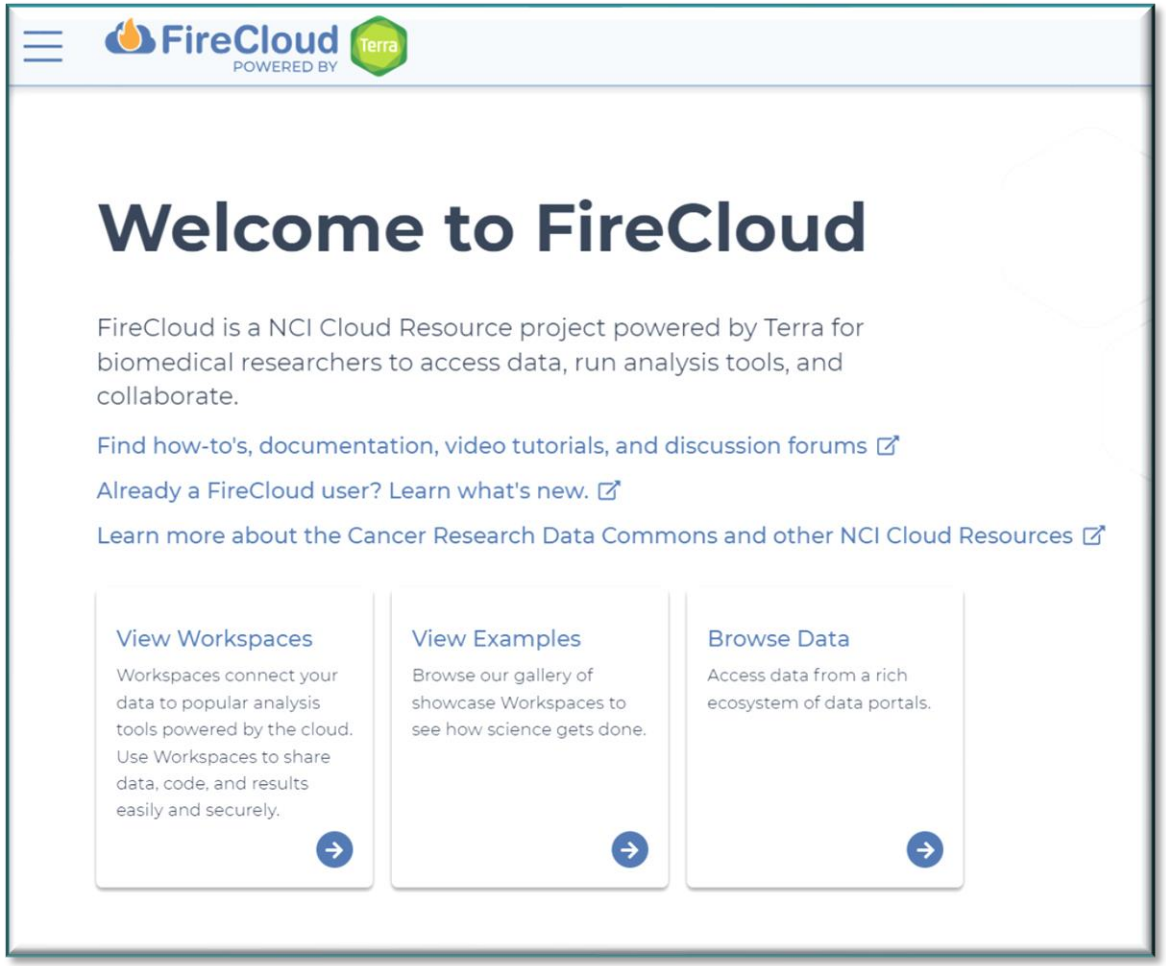


https://portal.isb-cgc.org/

CRDC: NCI Cloud Resources

Broad's FireCloud, powered by Terra

- FireCloud allows data integrations with GDC, PDC, and IDC
- FireCloud is a cloud workbench - you can use CPU and GPUs to power your analysis, and scale as needed
- The FireCloud infrastructure is used to power the curation of >250,000 WGS (and growing) concurrently for the All of Us resource

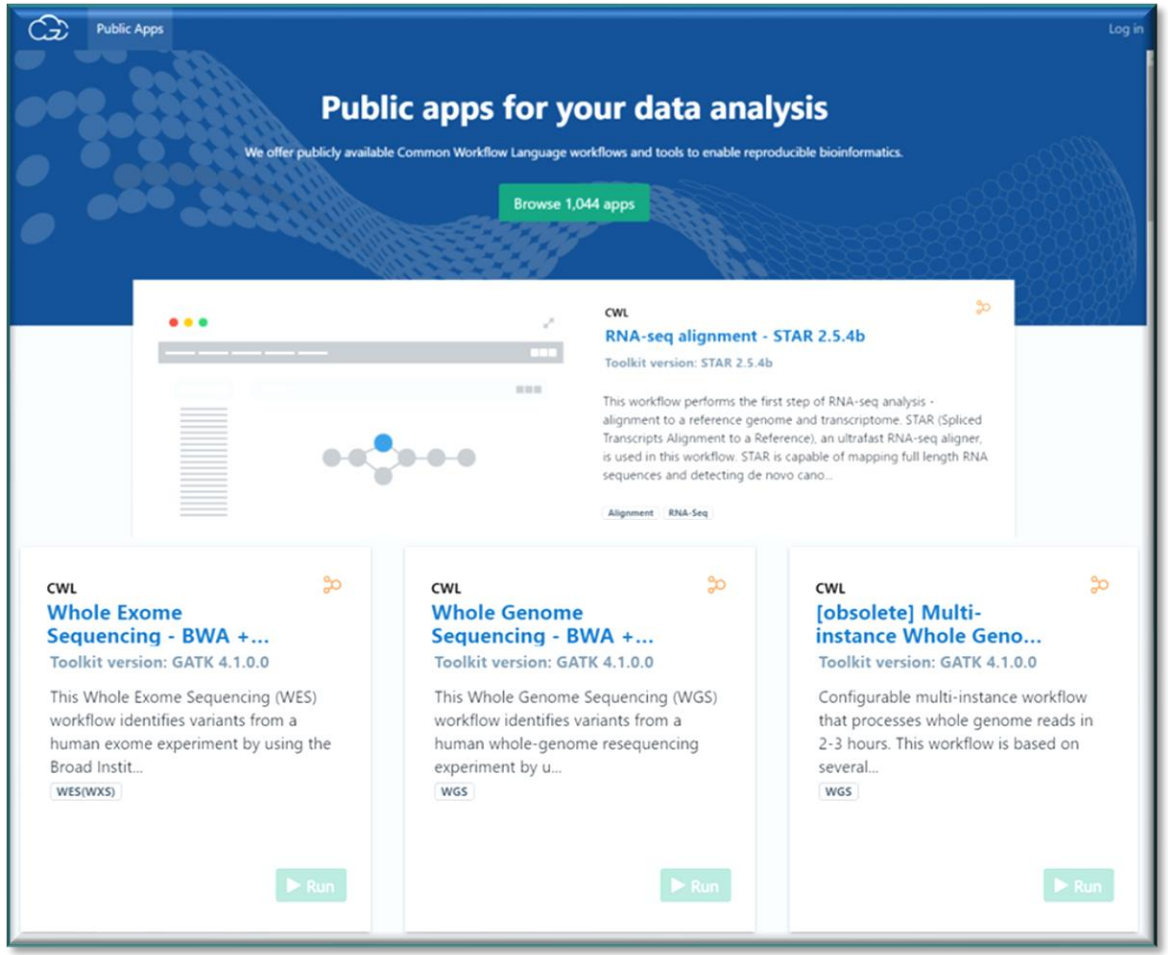


https://firecloud.terra.bio/

CRDC: NCI Cloud Resources

Seven Bridges Cancer Genomics Cloud (SB-CGC), powered by Velsera

- **Easy to Use:** point-and-click interface, >1,000 informatics workflows & tools
- **Extensive Support:** tutorials, public projects, office hours, and onboarding videos for all user types
- **Cutting-edge Research Support:** capitalizing on advances in ML & genAI
- **Interoperability/FAIR:** supporting multiple workflow languages
- **Citations:** >130 total
- **Training:** Committed to training the next generation of scientists



https://www.cancergenomicscloud.org/

CRDC: NCI Cloud Resources

Training Tomorrow's Data Scientists

The CGC Community

CGC as a Teaching and Training platform

The National Institutes of Health (NIH) has made several four
funding basic and translational research which generates ma
as well as data analysis ecosystems or platforms to empower
community to analyze these large datasets. Over the years, t
Genomics Cloud platform has not only enabled researchers t
research, but it has been used as an effective teaching platform for training th
generation of national research workforce. Several universities and research
institutes are effectively utilizing the CGC for their teaching and training progra

"It has been an honor and pleasure working with your group, and the participants are excited about using CGC for their research projects."

"Without the CGC, this would have been impossible."

"I cannot explain the level of support I would get [from the Velsera staff]. They got me through the last moments of my PhD."

"Just do it."

"...I'm really grateful that that, you know I got the chance to work with you guys in creating this course. So, if you are thinking about it, just do it."

"Over 700 analytical tools w
CGC platform that can be u
non-bioinformaticians. This
very little to no codi

If of course you're a bioinf
savvy, you have that choice to
the code but most people I k
want to deal with tha

SUCCESS STORIES

1. TEACHING USING THE CGC - GEORGETOWN UNIVERSITY

We interviewed our long term collaborators Dr. Yuriy Gusev and Ms. Krithika Bhuvaneshwar from Georgetown University who have been using the CGC platform for the past 3 years to train the next generation of data scientists in their Masters in Health Inf
interview:

2. TEACHING USING THE CGC - UCSD

Professor Jeremy Chien from UC Davis used CGC to teach a streamlined, hands-on course in cancer genomics to undergraduates. The platform eliminated the lengthy setup and steep learning curve usually associated with bioinformatics, allowing students to dive directly into real-world data analysis. The results? Engaged students who gained

3. TEACHING USING THE CGC - PURDUE UNIVERSITY

We partnered with Min Zhang at Purdue University to deliver a four-part series on RNAseq as part of their STAT-581 course. Each class consisted of a lecture on an RNAseq topic and hands-on training on implementing an analysis on the CGC. We used data in the Sequence Read Archive, providing a real world example for the students to follow. All students were able to accomplish the training in each class. Here is some of the written feedback we received:

CRDC: NCI Cloud Resources

Training Tomorrow's Data Scientists

CRDC Insights

Updates from the Cancer Research Data Commons:
Empowering the Scientific Community to Make New Discoveries

Sign up for the newsletter *

Sign up

Vaidhyanathan Mahaganapathy, Ph.D., Reflects on Support Received from NCI Cloud Resources

October 14, 2024

Vaidhyanathan (Vaidhy) Mahaganapathy, Ph.D., is a Computational Biologist focused on genomics research for the Applied AI team at the Ellison Institute of Technology (EIT). He recently spoke with CRDC Insights, reflecting on the support he received during his dissertation research from the team at Seven Bridges-Cancer Genomic Cloud (SB-CGC), powered by Velsera. He also commented on the complexity of working with huge datasets and trends in cancer research data science, and offered advice to graduate students looking to start a research career.

Dr. Mahaganapathy's dissertation focused on detecting clonal hematopoiesis (CH), a process prevalent in older people, which has an impact on immune function and inflammation, and is related to many cancers and diseases of aging. For his dissertation, he analyzed 17,000 whole exome sequences in the TCGA dataset, housed through the CRDC, to look across thousands of subjects for mutations consistent with CH.

Other Features from CRDC Insights

- CRDC 2024 Fall Symposium - Oct 16 & 17 - Registration is Open
- The CRDC's New Data Submission Portal
- Recommendations from NCI CRDC AI Data Readiness Challenge Winners
- Vaidhyanathan Mahaganapathy, Ph.D., Reflects on Support Received from NCI Cloud Resources
- The National Cancer Institute (NCI) Medical Image De-Identification Benchmark

CRDC: Internal Interoperability Projects

Challenge: Access comprehensive datasets like TCGA/CPTAC from multiple commons for integrative analysis

- Use common standards

Data Standards Services (DSS)

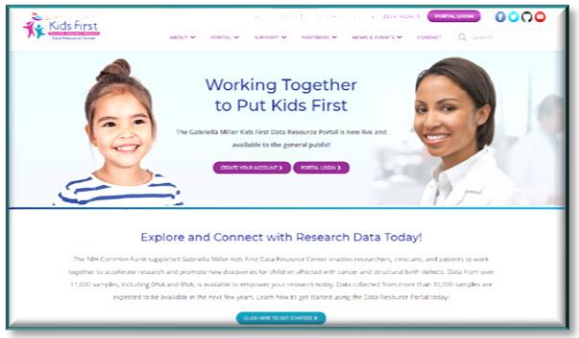
- Semantic harmonization and shared data model across CRDC
- Leverage existing standards

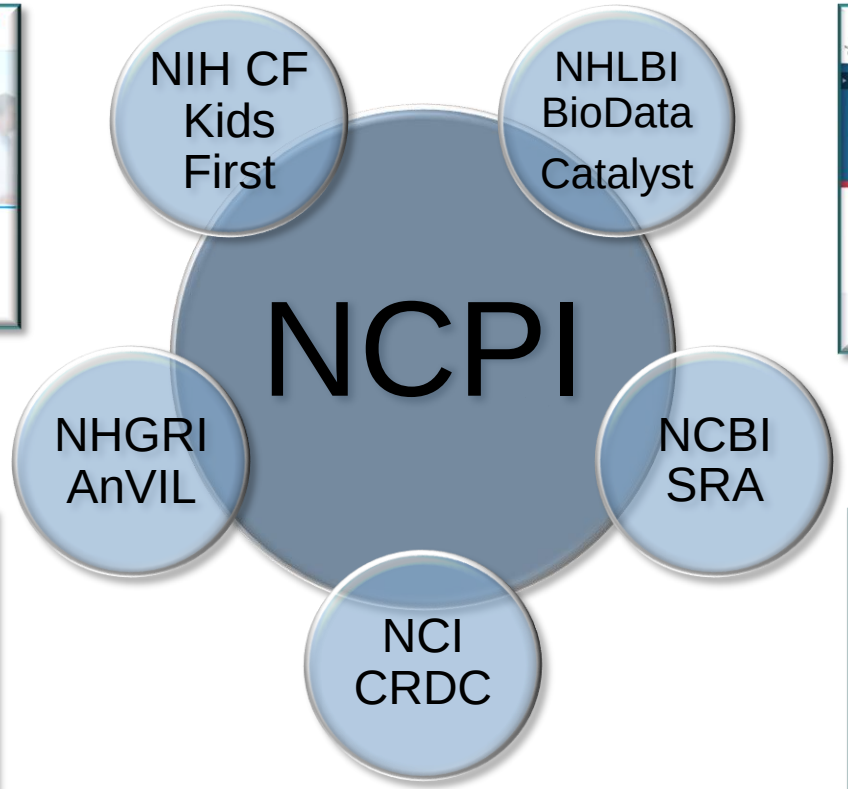
Cancer Data Aggregator (CDA)

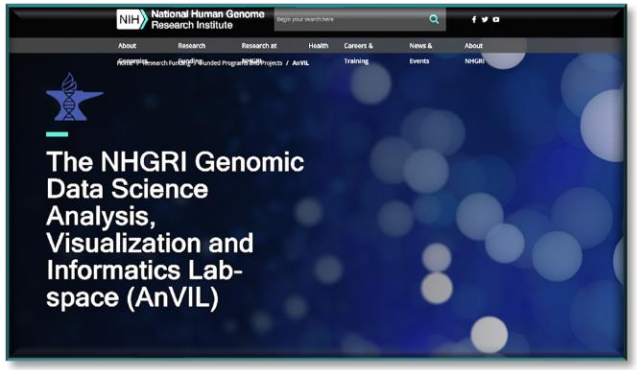
- API layer for query/aggregation of data cross-CRDC
- Database for biospecimen, clinical, and phenotypic metadata from CRDC datasets

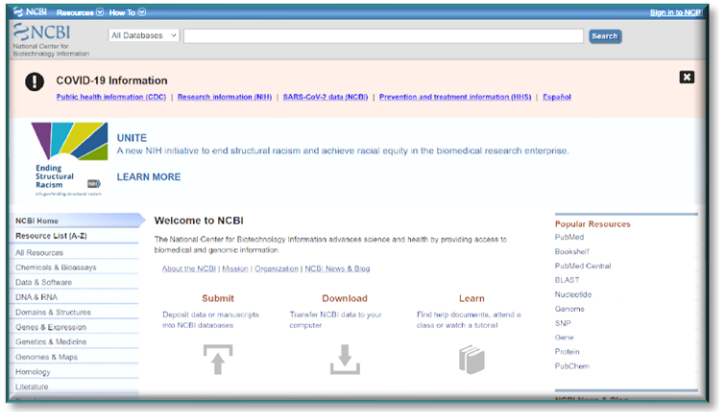
NCPI: NIH Cloud Platforms for Interoperability

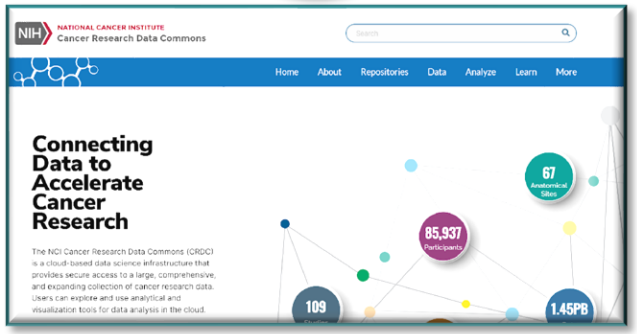
Connecting with a Greater Data Ecosystem











RAS: NIH Researcher Auth Services

Towards Single "Sign on" Across NIH Data Resources

AUTHENTICATION
No matter what preferred credentials researchers enter, their account identity and attributes are transferred

AUTHORIZATION
Researchers have access to the datasets they need and expect in ways they understand; web of trust

AUDITING
Trace and log data in a standard way to protect staff and intellectual property

OBJECTIVES

- Researchers leverage a simplified login across all of the participating data resources
- Build trust and a healthy, collaborative relationship
- Lower the barriers for new systems to plug into the RAS
- Systems leverage a standardized auditing service

NIH STRATEGIC PLAN FOR DATA SCIENCE

GOAL 1: SUPPORT A HIGHLY EFFICIENT AND EFFECTIVE BIOMEDICAL RESEARCH DATA INFRASTRUCTURE

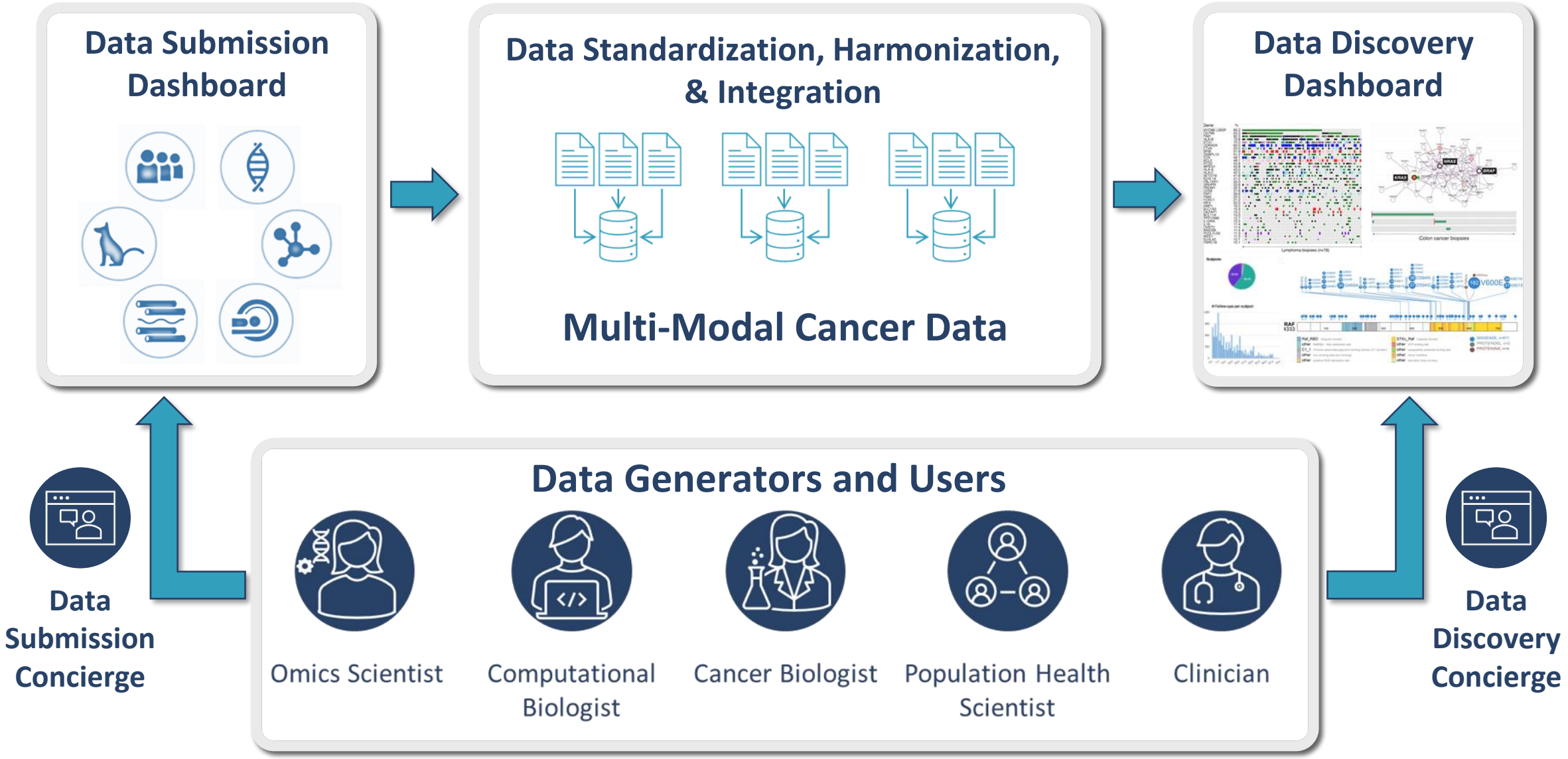
Objective 1-1: Optimize Data Storage and Security

1.1.3 Support technical and infrastructure needs for data security, authorization of use, and unique identifiers to index and locate data

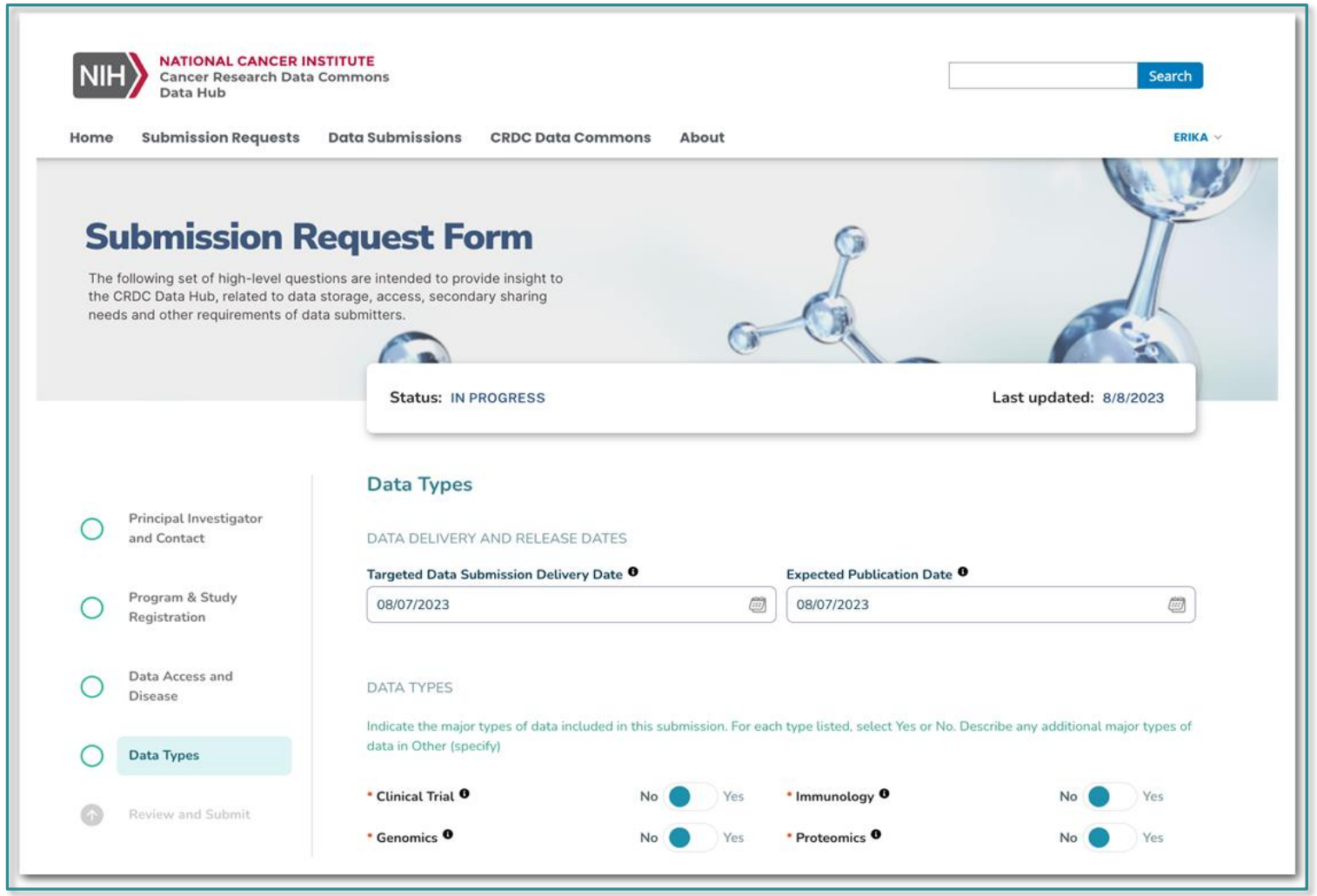
NIH RAS is sponsored by NIH ODSS and lives in NIH CIT IAM

CRDC Ecosystem: Future State

CRDC Data Portals



CRDC Data Portal: Centralized Data Submission

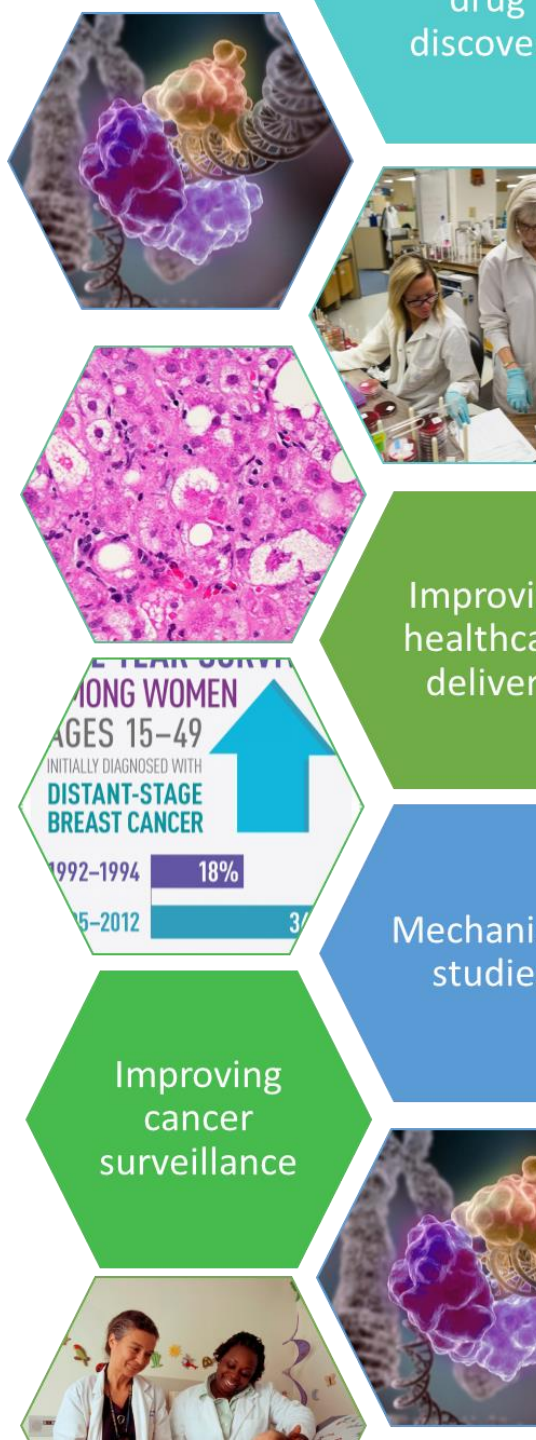


- **Step 1 - Submission Request:** All CRDC Data Commons accepting submission requests through the new Submission Portal
- **Step 2 – Submit Data:**
  - **Currently Accepting Data through the Portal:** Cancer Data Service, Clinical and Translational DC, Integrated Canine DC
  - **Future Integration:** Genomic DC, Proteomic DC, Imaging DC

Data Submission Concierge

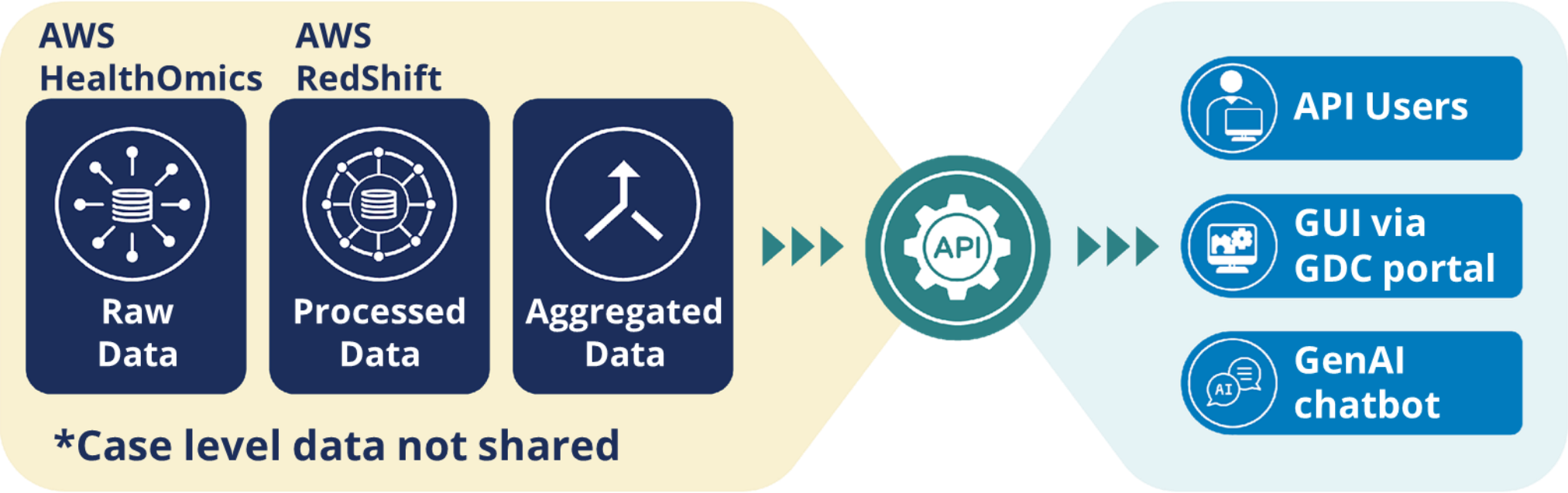
CRDC for Enabling AI in Cancer Research

- **AI in Imaging**
  - AI used in automated de-identification of images
  - Model Hub AI repository
- **AI Data Readiness (AIDR)**
  - Expands FAIR principles to make data accessible for use in AI future applications
  - Request for Information
  - Community Challenge



NCI Genomic Enclave

Data Donation from Tempus: 3300 cases of genomic profiling data, DNA and RNA sequencing



ARPA-H Biomedical Data Fabric (BDF) Toolbox

In collaboration with the National Cancer Institute, ARPA-H will advance the next-generation of tools to synthesize and speed use of health research data, **starting with cancer.**

Make biomedical research data easier to use

Reduce effort for data integration

Develop new data fabric capabilities & tools

Build health science models to realize the potential of the Cancer Data Ecosystem

https://arpa-h.gov/research-and-funding/programs/arpa-h-bdf-toolbox

Acknowledgements

- NCI's CRDC Program & Partners:
  - CBIIT Informatics and Data Science (IDS) Program
    - Data Ecosystems Branch (DEB)
    - Clinical & Translational Research Informatics Branch (CTRIB)
    - Computational Genomics & Bioinformatics Branch (CGBB)
  - Center for Cancer Genomics (CCG): GDC
  - DCTD, Office of Cancer Clinical Proteomics Research (OCCPR): PDC
  - DCTD, Developmental Therapeutics Program (DTP): ICDC
  - Division of Cancer Control and Population Sciences (DCCPS): PSDC
  - DCTD, Cancer Imaging Program (CIP): The Cancer Imaging Archive
  - DCTD, Cancer Therapy Evaluation Program (CTEP): NCTN Data Archive
- NCI Frederick National Lab Team (FFRDC)
- All CRDC Subcontractor teams
- All Partners throughout NHI/NCI and data contributors