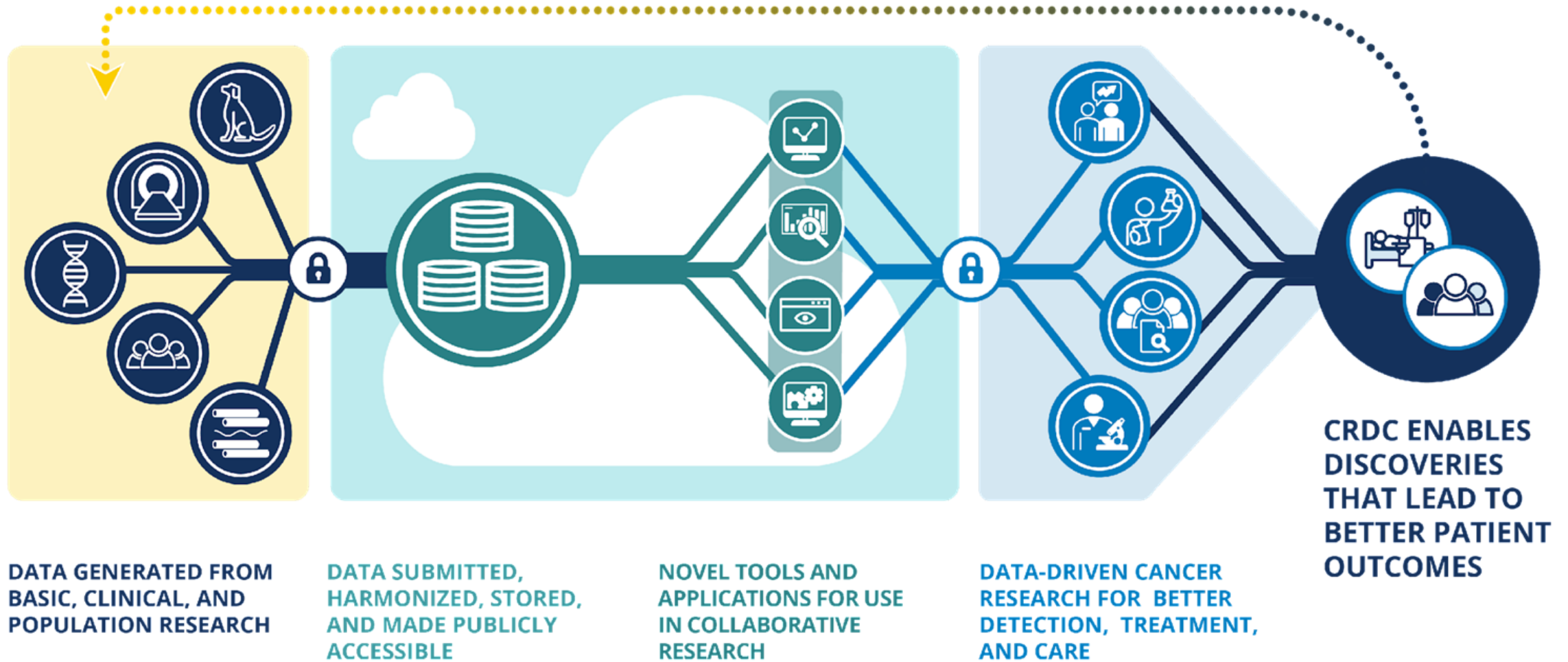


Cancer Research Data Commons (CRDC) Intro/Overview

Erin Beck
NCI CBIT, IDS

March 27, 2024

NCI's Cancer Research Data Commons (CRDC)



CRDC: By The Numbers



354 STUDIES



134K SUBJECTS



9.4PB+ DATA



2.4K+
YEARS OF COMPUTE



2K+ PUBLIC
TOOLS AND WORKFLOWS



82.3K+
UNIQUE USERS / YEAR



30K+
CRDC DATA CITATIONS

* As of Dec. 2023

AACR Cancer Research Series

A four-part invited series published online in March 2024 highlighting the CRDC's accomplishments from the past 10 years.

- ▶ LESSONS LEARNED AND FUTURE STATE
- ▶ RESOURCES TO SHARE KEY CANCER DATA
- ▶ CLOUD-BASED ANALYTICAL RESOURCES
- ▶ CORE STANDARDS AND SERVICES



Learn more about the series on the [CRDC Website](#).





2024 AACR Annual Meeting: San Diego, CA



Presentations

- **Impact of the Cancer Research Data Commons (CRDC)**
 - Sunday, April 7 - 1:00pm – 2:00pm
- **NCI Artificial Intelligence (AI) Programs and Resources for Advancing Cancer Research**
 - Wednesday, April 10 - 10:15am -11:15am



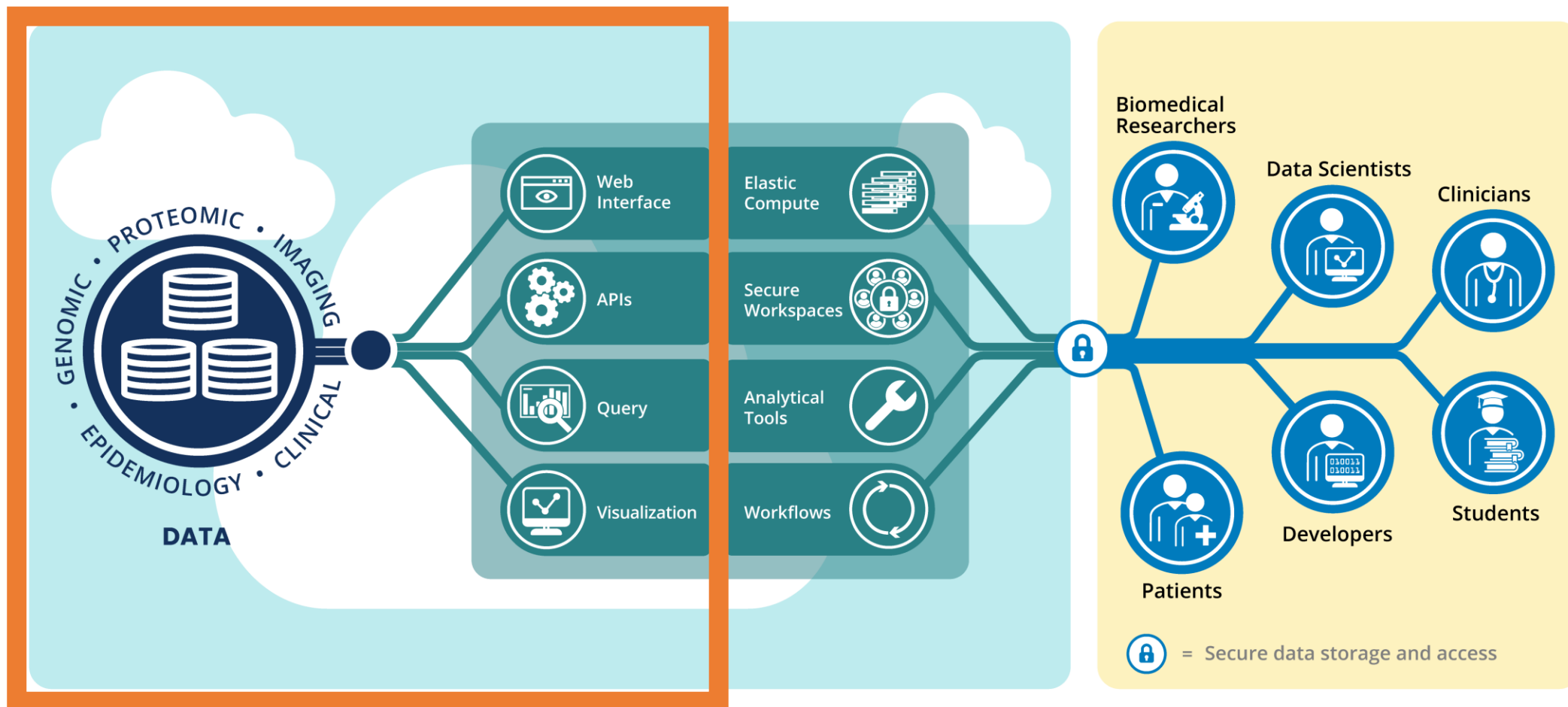
Posters

- **ISB – Cancer Gateway in the Cloud**
 - Monday, April 8
- **CRDC – Sustainability Implementation Planning**
 - Monday, April 8
- **Velsera – Seven Bridges, Cancer Genomics Cloud**
 - Monday, April 8
 - Tuesday, April 9
 - Wednesday, April 10
- **Broad – FireCloud (Terra)**
 - Wednesday, April 10



View the [AACR Program](#) for more details.


CRDC: Ecosystem

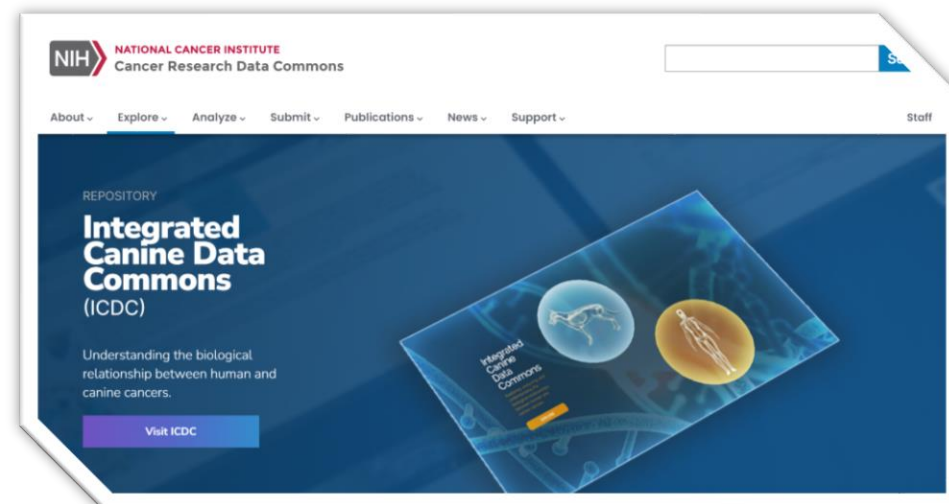
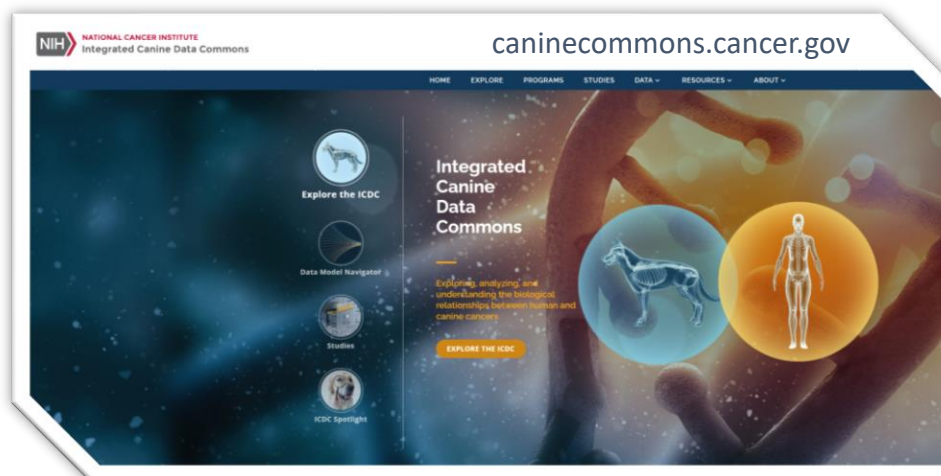


CRDC: Data Commons

datacommons.cancer.gov

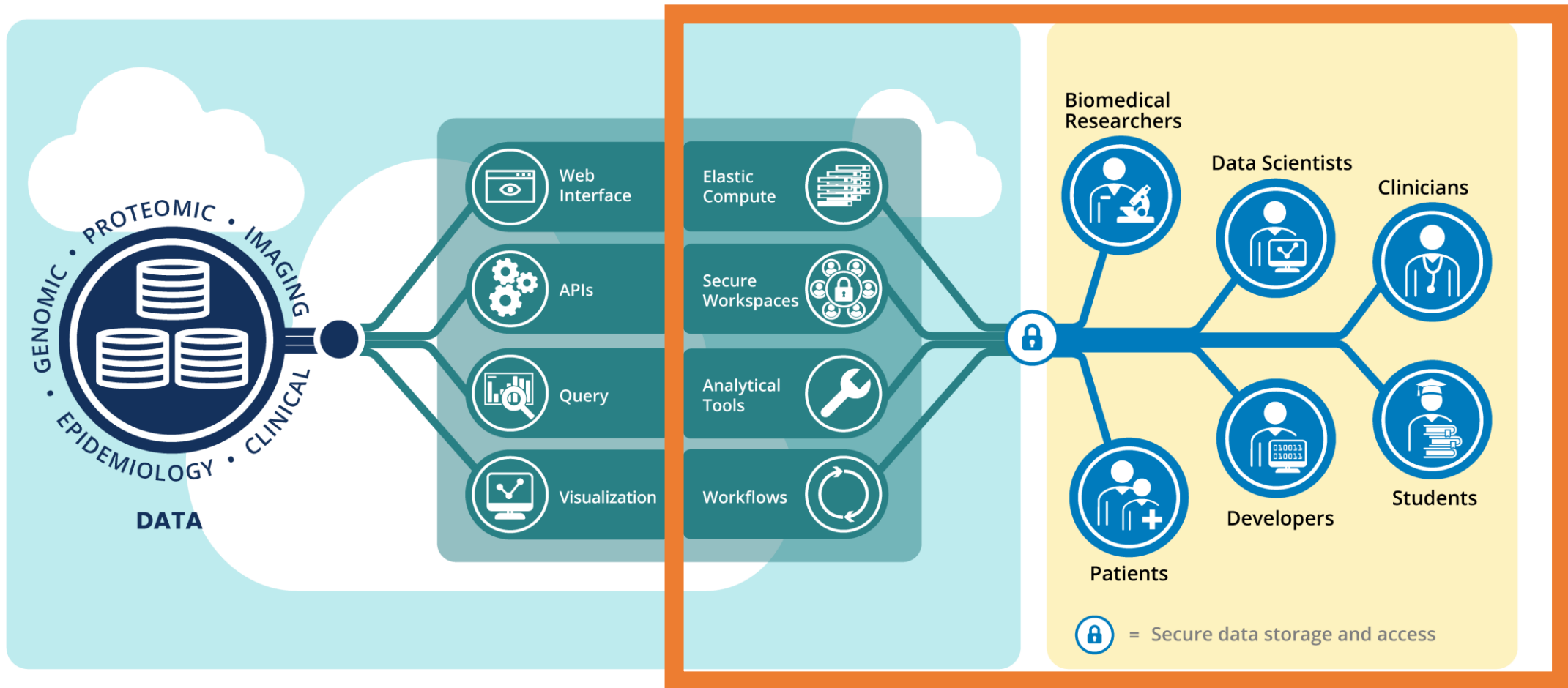


DATA COMMON	KEY FEATURES
 <p>Cancer Data Service (CDS)</p>	The CDS houses and shares data type-agnostic data that are not a fit for other CRDC data commons. Data go through QC but are not harmonized. The CDS includes both open and controlled access data.
 <p>Clinical and Translational Data Commons (CTDC)</p>	The CTDC is designed to share clinical, biospecimen, and molecular characterization data from clinical trials and other studies. Users can explore data through the CTDC portal.
 <p>Genomic Data Commons (GDC)</p>	The GDC is designed to share harmonized genomic data, including WGS, WXS, RNAseq, miRNA-seq, scRNAseq, ATAC-seq, and DNA methylation data. The GDC supports free data downloading (both raw sequencing data and derived data), and hosts both open and controlled access data. The GDC Data Portal supports free online data exploration and analysis of custom cohorts.
 <p>Imaging Data Commons (IDC)</p>	The IDC shares de-identified imaging data, including both radiology and pathology slide images. All images are harmonized using DICOM standards. All data in the IDC are open access.
 <p>Integrated Canine Data Commons (ICDC)</p>	The ICDC shares data from the veterinary records of pet dogs that naturally developed tumors. Key data types include WXS, WGS, RNA-Seq, and DNA Methylation. All data (including raw sequence data) are open access.
 <p>Proteomic Data Commons (PDC)</p>	PDC primarily shares mass spectrometry-based proteomic data. The PDC portal supports online data exploration and visualization. All data (both raw and derived data) are open access.




datacommons.cancer.gov

CRDC: Ecosystem



CRDC: Cloud Resources

Table 2. **Tool Availability**

Tool Category	Tools	 Broad FireCloud	 ISB-CGC	 SB-CGC
Workflows	CWL Workflow Support	✓	✓	✓
	WDL Workflow Support	✓	✓	✓
	Nextflow Workflow Support	Coming soon	✓	✓
	Publicly Available Workflows from Dockstore	✓	✓	✓
Analysis Types	Existing workflows and tools used by community	Variant calling (long and short reads), GWAS, RNA, ML, Epigenomics, fusion detection	Variant calling (short reads), RNAseq, ML, CNV, Epigenomics, correlations using BigQuery derived datasets	Variant calling (long and short reads), GWAS, Bulk RNAseq, Single-Cell RNAseq, ML, Epigenomics, Multiomics, Proteomics, Fusion Detection, Imaging Analysis
Tutorials	Example Tool Analysis Projects	✓	✓	✓
Interactive Applications	Jupyter	✓	✓	✓
	RStudio	✓	✓	✓
	RShiny Apps	Coming soon	✓	✓
	Galaxy	✓		✓
	SAS			Coming soon
	Command Line Sessions	Coming soon	✓	✓
	Interactive Querying (BigQuery, etc)		✓	✓
User Driven Content	User Written Workflow Support	✓	✓	
	User Created Interactive Apps	Coming soon	✓	
	User Defined Project Resources		✓	✓
	APIs for scripting	✓	✓	✓
Analytic Workspaces	Bring your own data	✓	✓	✓
	Access Controlled Data	✓	✓	✓
Cloud Native Tool Support	Billing	Cloud-specific	Cloud-specific	Integrated
	Command Line Tools, e.g. gsutil	✓	✓	via Python / R
	Make use of Cloud specific tools such as TensorFlow, BigQuery, etc	✓	✓	✓
	STRIDES support	✓	✓	Coming soon



Broad FireCloud

- Built on Terra Platform
- Google



The ISB Cancer Gateway in the Cloud (ISB-CGC):

- Google
- Big-query tables



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

- Amazon

CRDC Questions:

- Put in chat
- erin.beck2@nih.gov

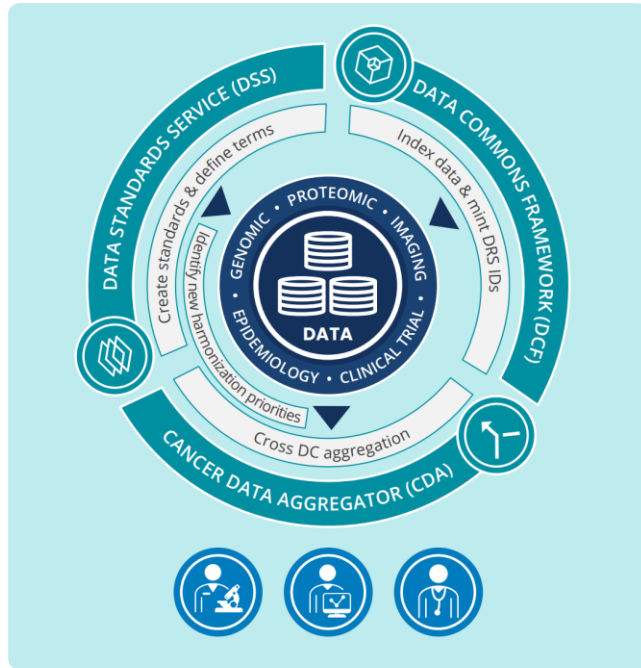
Next Up:

- ISB – Cancer Gateway in the Cloud



CRDC: Interoperability

CRDC INFRASTRUCTURE DELIVERS
STANDARDIZED, AGGREGATED, SEARCHABLE DATA



- **Data Standards Services**
 - Common metadata/semantics & data model
- **Data Commons Framework**
 - Manages authN & authZ
 - Indexing of all data files
- **Cancer Data Aggregator**
 - Federated search via query, APIs & GUI (coming soon)
- **Cloud Resources**
 - Access data without download
 - Secure workspaces
 - Elastic compute
 - Analytical tools & Workflows

file manifest, GUID (DRS ID)

