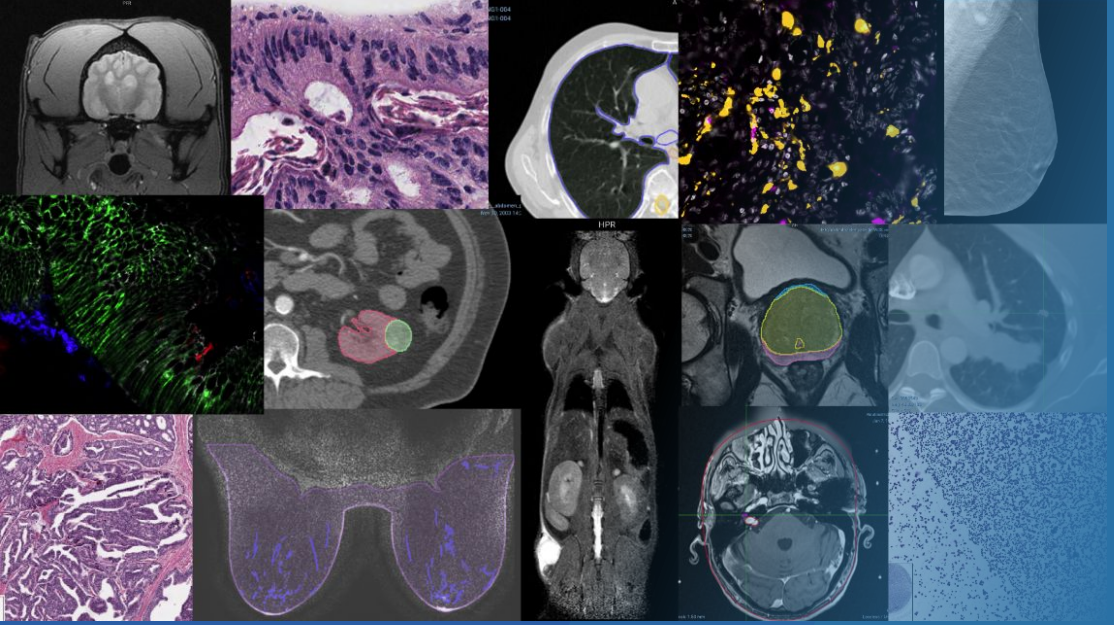
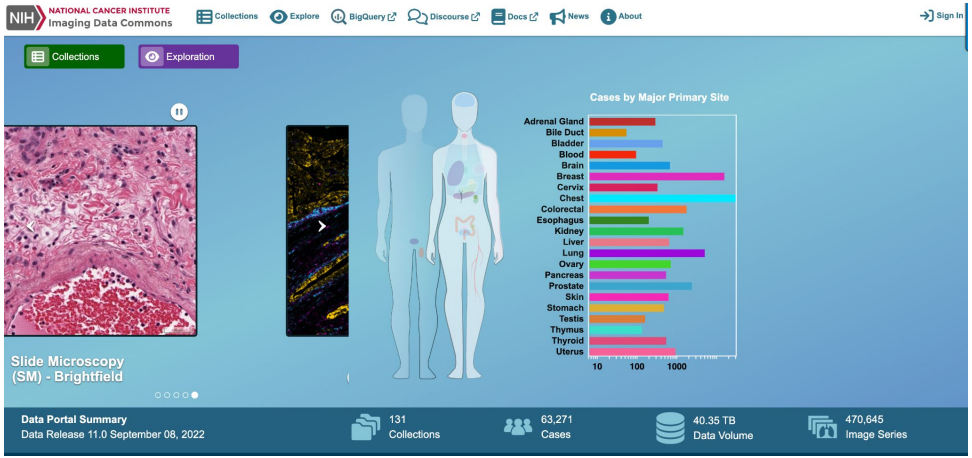


NCI Imaging Data Commons

Andrey Fedorov, PhD, on behalf of the IDC team
Brigham and Women's Hospital, Mass General Brigham, Boston
28 November 2022

National Cancer Institute (NCI) Imaging Data Commons (IDC)

NCI Imaging Data Commons (IDC) is a cloud-based repository of publicly available cancer imaging data co-located with the analysis and exploration tools and resources.



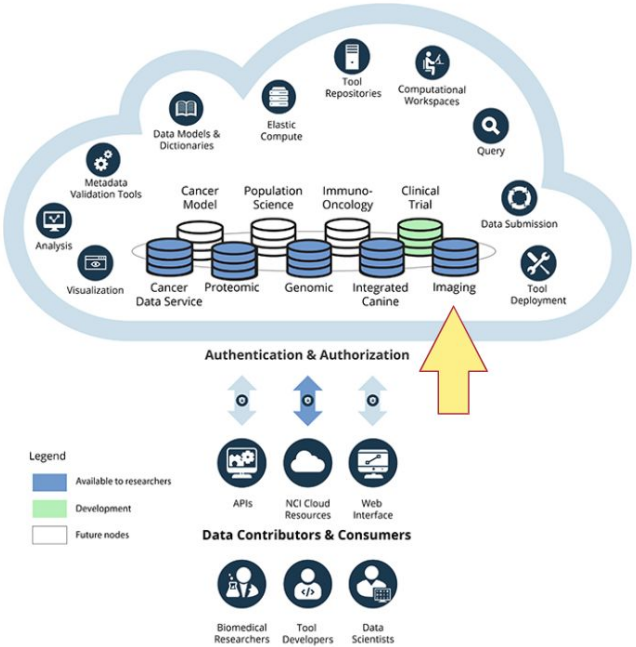
- Public DICOM imaging datasets
- Images, image-derived and image-related (clinical) data
- Radiology, digital pathology, and more
- Tools for search, visualization, exploration of data

Data Portal Summary
Data Release 12.0 October 11, 2022

128 Collections | 63,316 Cases | 40.37 TB Data Volume | 470,850 Image Series

https://imaging.datacommons.cancer.gov

NCI Imaging Data Commons (IDC)

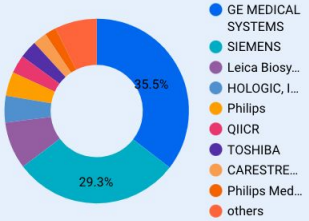


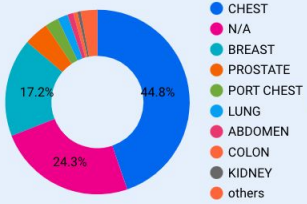
IDC is a component within the broader NCI Cancer Research Data Commons (CRDC) infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data.

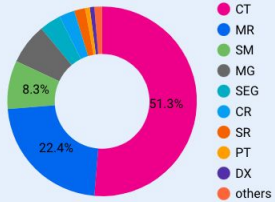
https://datacommons.cancer.gov/

NCI Imaging Data Commons in numbers

Collections	DOIs	Size on disk	Cases	Studies	Series	Instances	Distinct StudyDate
131	124	44.4T	63,271	137,978	470,645	41.9M	8,296



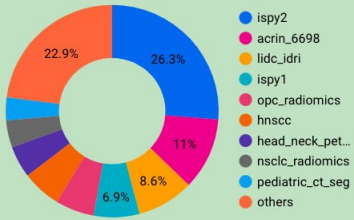


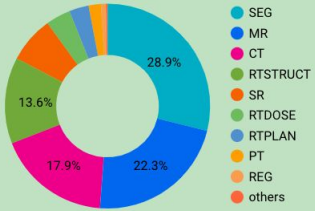


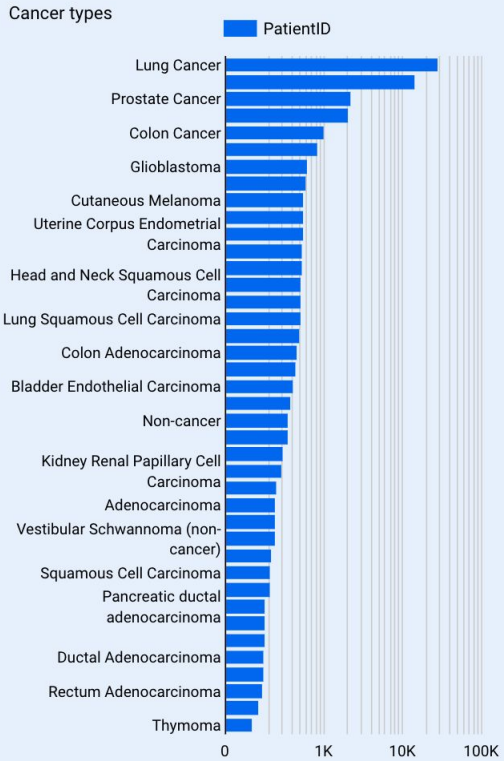
Manufacturer	BodyPartExamined	Modality
87	84	21

Image collections accompanied by annotations

Collections	Cases	Studies
38	6,480	10,233





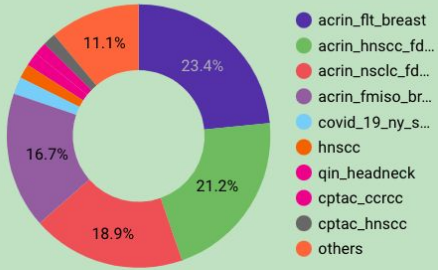


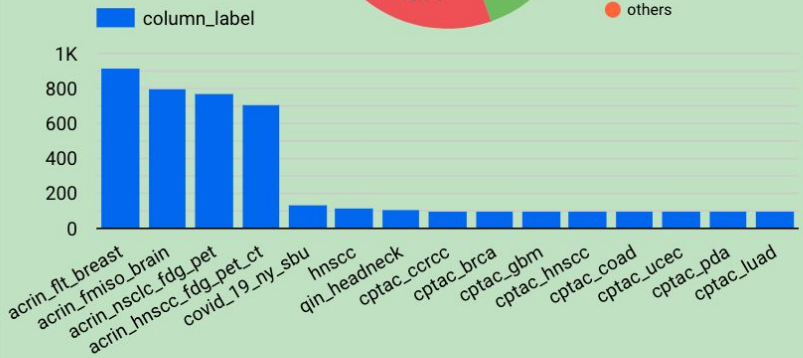
Live dashboard at https://datastudio.google.com/u/0/reporting/04cf5976-4ea0-4fee-a749-8bfd162f2e87/page/p_s7mk6eybqc

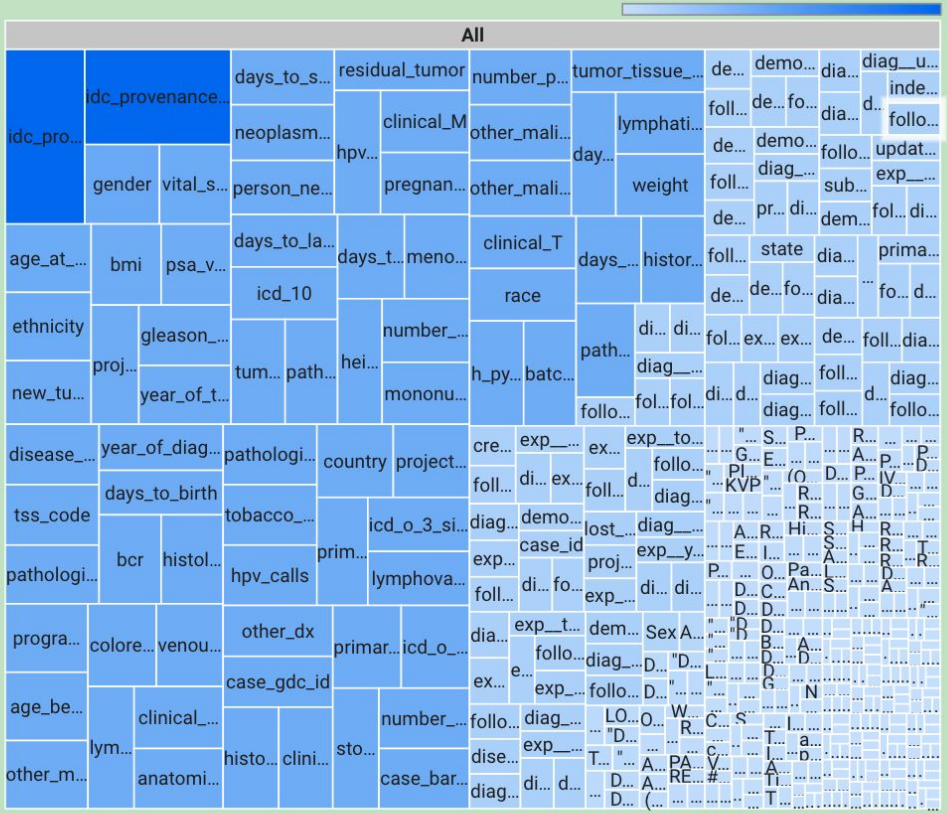
Image collections accompanied by clinical data

Collections	Tables	Columns	Patients
78	189	5,371	20,935

Right: Summary of the number of distinct columns (dictionary values) per collection.

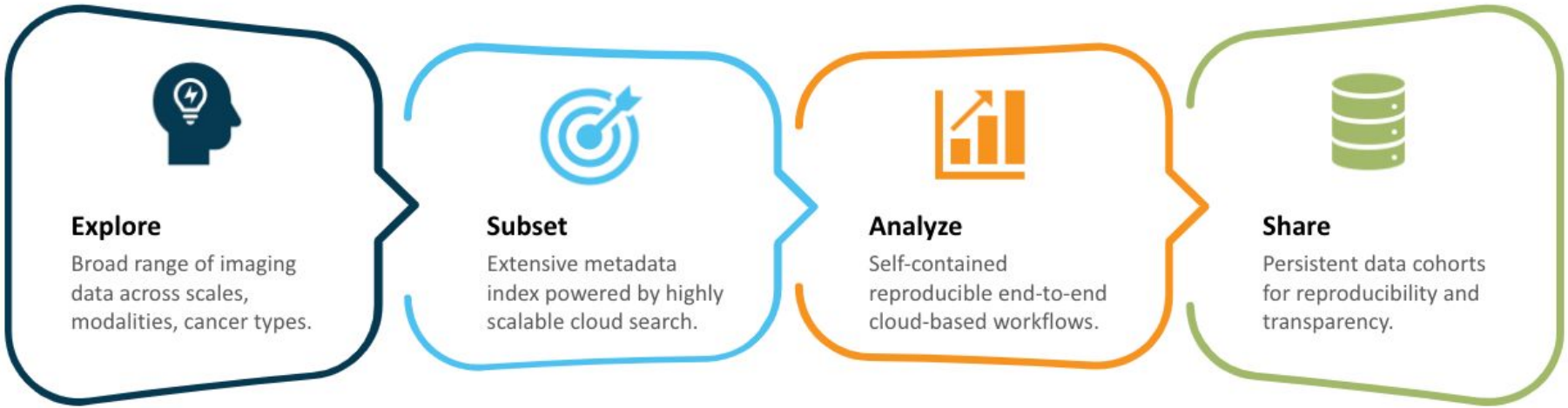






Left: Count of distinct dictionary values per collection. Right: Distinct dictionary values across all collections with corresponding clinical data (mouse over individual items to see full column label text). Size of the rectangle corresponds to the number of collections where specific column label was encountered.

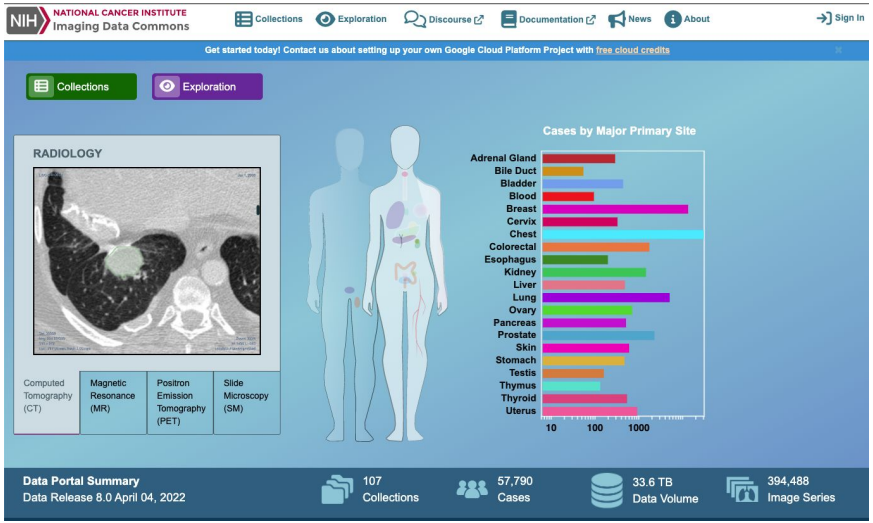
Live dashboard at https://datastudio.google.com/u/0/reporting/04cf5976-4ea0-4fee-a749-8bfd162f2e87/page/p_s7mk6eybqc

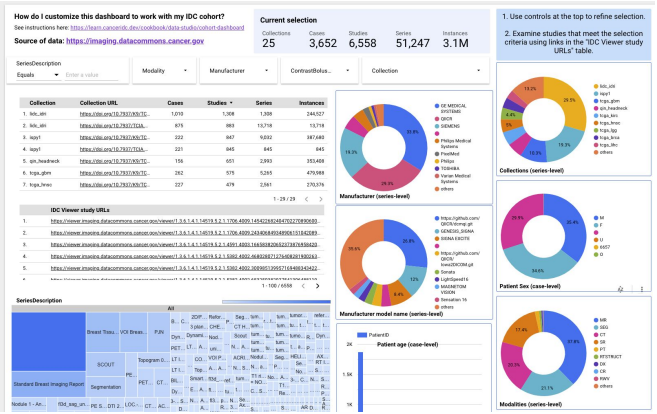

Explore
Broad range of imaging data across scales, modalities, cancer types.
Subset
Extensive metadata index powered by highly scalable cloud search.
Analyze
Self-contained reproducible end-to-end cloud-based workflows.
Share
Persistent data cohorts for reproducibility and transparency.

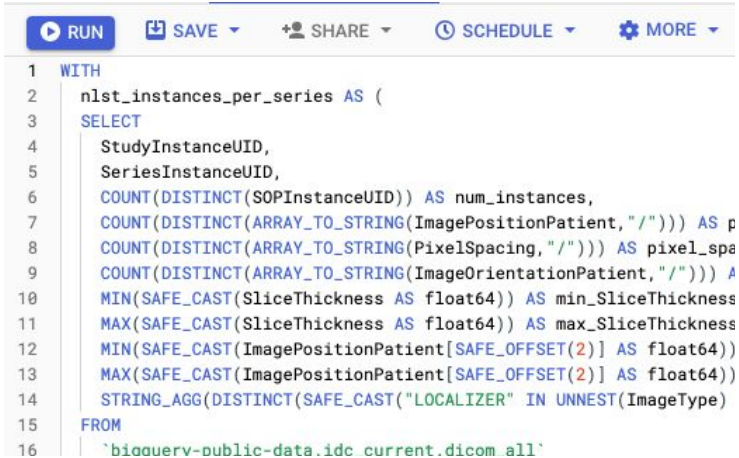
What you get

Without downloading data or installing any software on your computer:

- **Search and explore data**



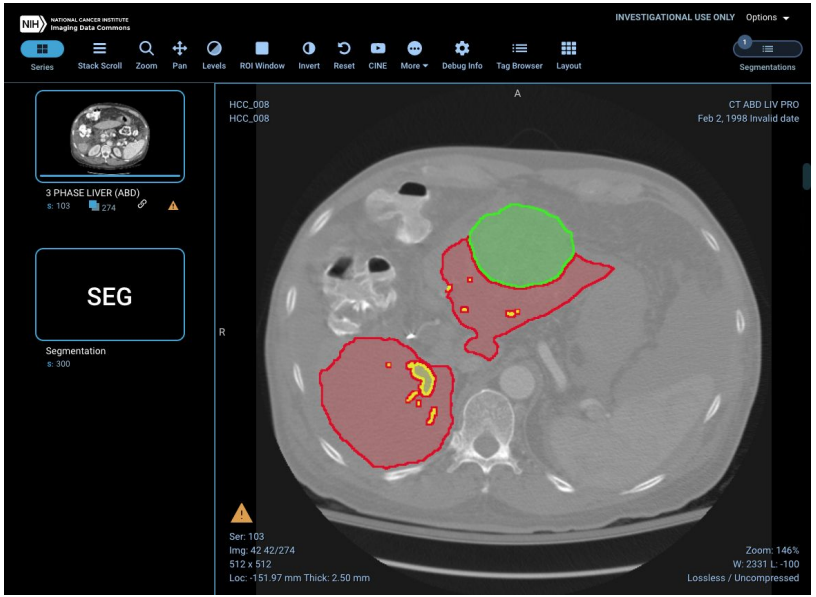


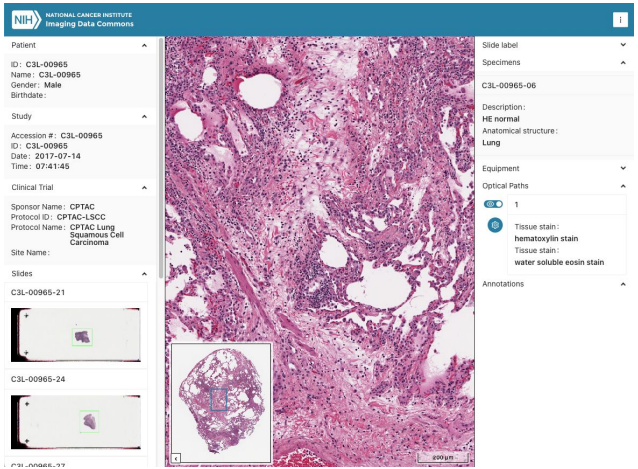


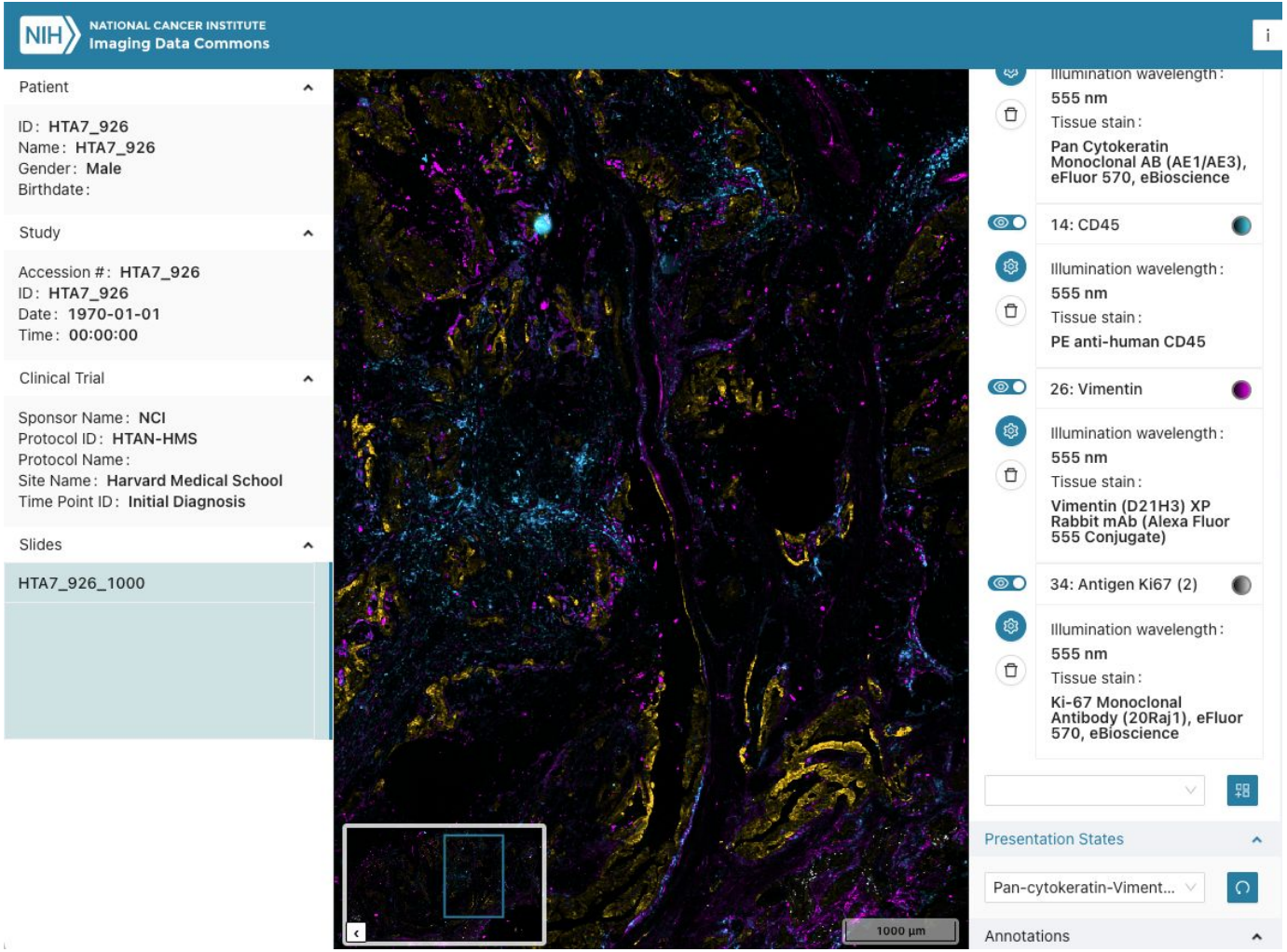
What you get

Without downloading data or installing any software on your computer:

- Search and explore data
- **Visualize images and derived data**





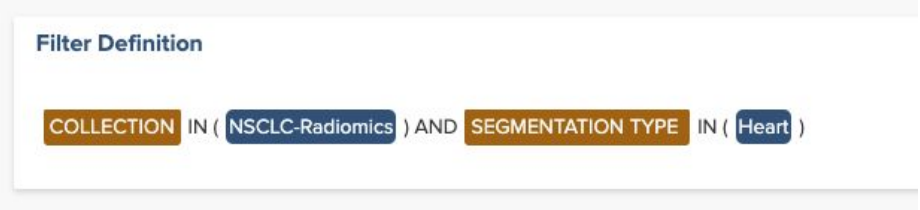

NIH
NATIONAL CANCER INSTITUTE
Imaging Data Commons
Patient
ID: HTA7_926
Name: HTA7_926
Gender: Male
Birthdate:
Study
Accession #: HTA7_926
ID: HTA7_926
Date: 1970-01-01
Time: 00:00:00
Clinical Trial
Sponsor Name: NCI
Protocol ID: HTAN-HMS
Protocol Name:
Site Name: Harvard Medical School
Time Point ID: Initial Diagnosis
Slides
HTA7_926_1000
Illumination wavelength:
555 nm
Tissue stain:
Pan Cytokeratin Monoclonal AB (AE1/AE3), eFluor 570, eBioscience
14: CD45
Illumination wavelength:
555 nm
Tissue stain:
PE anti-human CD45
26: Vimentin
Illumination wavelength:
555 nm
Tissue stain:
Vimentin (D21H3) XP Rabbit mAb (Alexa Fluor 555 Conjugate)
34: Antigen Ki67 (2)
Illumination wavelength:
555 nm
Tissue stain:
Ki-67 Monoclonal Antibody (20Raj1), eFluor 570, eBioscience
Presentation States
Pan-cytokeratin-Viment...
Annotations
1000 µm

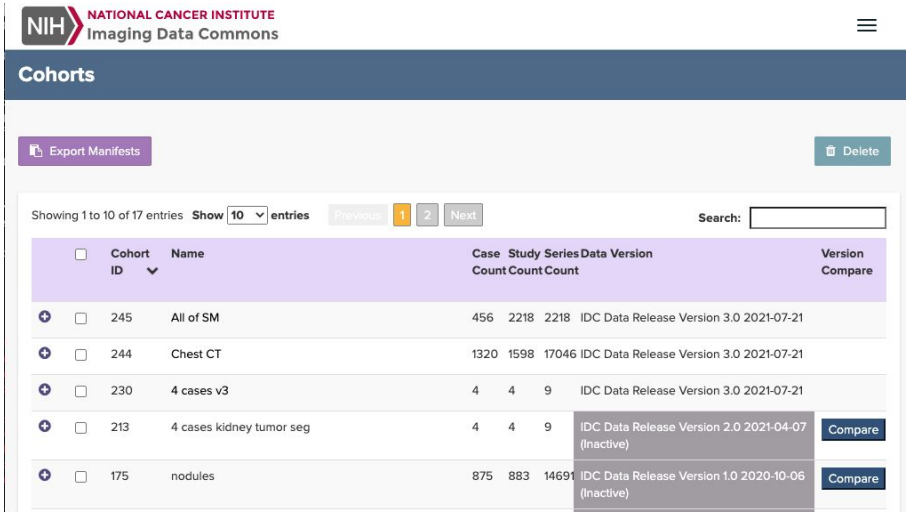
What you get

Without downloading data or installing any software on your computer:

- Search and explore data
- Visualize images and derived data
- **Define and save cohorts**

```
SELECT
  StudyInstanceUID,
  gcs_url
FROM
 `bigquery-public-data.idc_current.dicom_all`
WHERE
 Modality = "CT"
 AND collection_id = "nlst"
 AND SAFE_CAST(SliceThickness AS float64) < 1
```

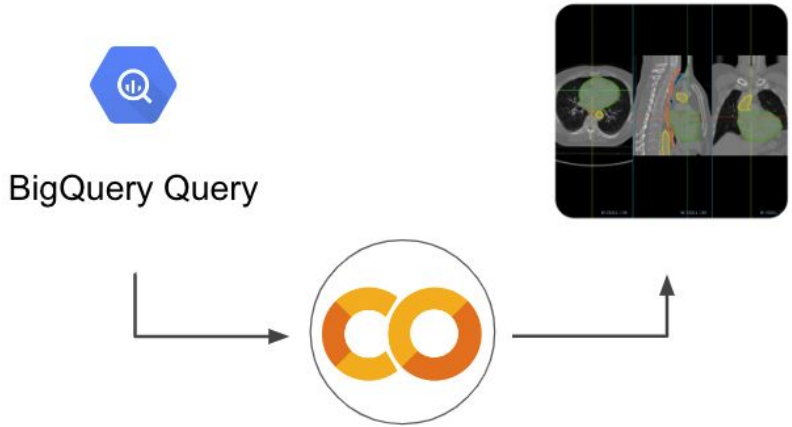


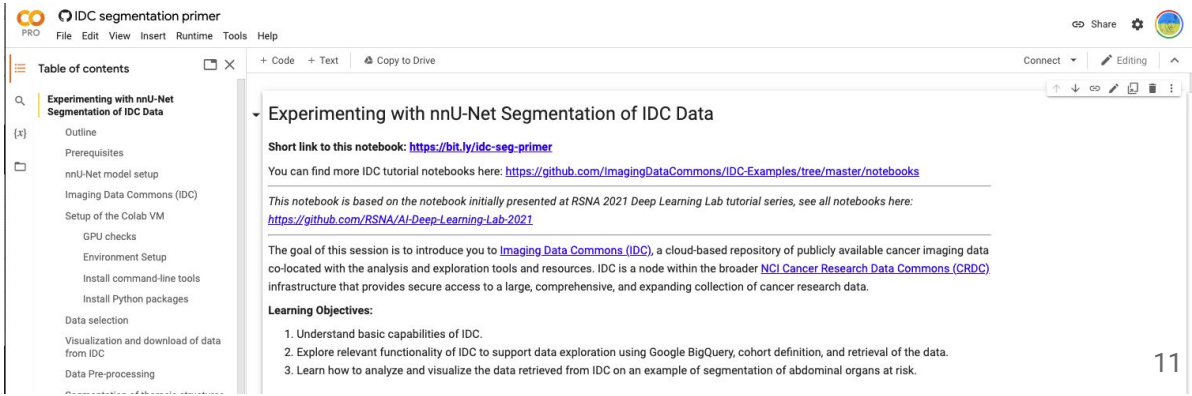


What you get

Without downloading data or installing any software on your computer:

- Search and explore data
- Visualize images and derived data
- Define and save cohorts
- **Develop/share analysis notebooks**



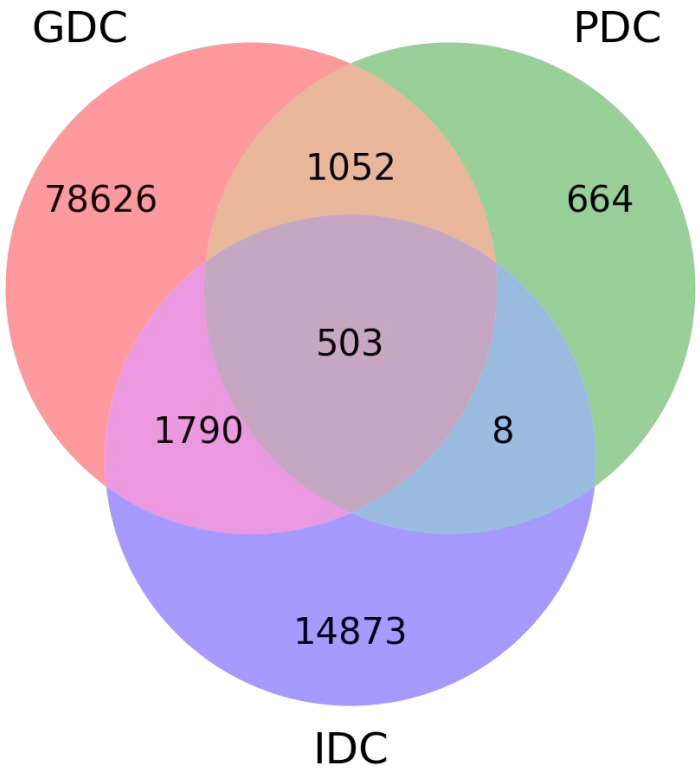


What you get

Without downloading data or installing any software on your computer:

- Search and explore data
- Visualize images and derived data
- Define and save cohorts
- Develop/share analysis notebooks
- **Find matching non-imaging data**

https://datacommons.cancer.gov/cancer-data-aggregator



Number of cases that have matching data between IDC, Genomic Data Commons (GDC) and Proteomic Data Commons (PDC)

Figure courtesy Fabian Seidl, ISB-CGC

What you get

Without downloading data or installing any software on your computer:

- Search and explore data
- Visualize images and derived data
- Define and save cohorts
- Develop/share analysis notebooks
- Find matching non-imaging data

Can I just download IDC data from the cloud and move on?

- Yes!

Downloading data

> You will need to complete prerequisites described in **Getting started with GCP** in order to be able to follow the instructions below!

IDC does not have an interactive point-and-click download application! If you want to download data from IDC you will need to use command line interface (Terminal on Mac/Linux or Command prompt on Windows).

Download of data from IDC is a 2-step process covered on this page:

- **Step 1:** create the manifest - the list of files defined by the Google Storage `gs://` URLs;
- **Step 2**: given that list of files, download files to your computer or to a cloud VM.

If you are analyzing IDC data in Google Colab, check out our Colab cookbook notebook that includes examples of how to query and download IDC data!

https://learn.canceridc.dev/data/downloading-data

What you get

FOR FREE

Without downloading data or installing any software on your computer:

- Search and explore data
- Visualize images and derived data
- Define and save cohorts
- Develop/share analysis notebooks
- Find matching non-imaging data

Can I just download IDC data from the cloud and move on?

- **Yes!**
  - **no out of cloud egress fees**