NCI Imaging Data Commons

Keyvan Farahani, PhD
Center for Biomedical Informatics and Information Technology

farahani@nih.gov
datascience.cancer.gov

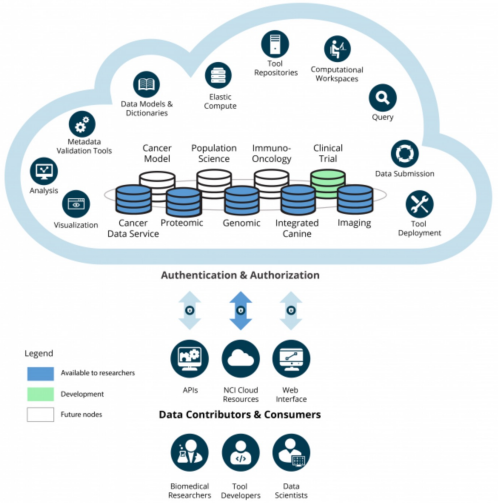
NIH NATIONAL CANCER INSTITUTE

NIH.AI / NVIDIA workshop, Oct 19, 2021

Cancer Research Data Commons (CRDC)

A data science infrastructure to connect repositories, analytical tools, and knowledge bases

- **Virtual, expandable, secure research infrastructure**
- **Storage and elastic compute**
- **Analysis, sharing, and archival of results**
- **Cross-domain analysis of large datasets**



The Cancer Research Data Commons (CRDC)

REPOSITORIES

Cancer Data Service (CDS)
Store and share NCI-funded data that are not hosted elsewhere to further advance scientific discovery across a broad range of research areas.

Clinical Trial Data Commons (CTDC)
Store and share data from NCI Clinical Trials. The resource is expected to launch in 2020.

Genomic Data Commons (GDC)
Share, analyze, and visualize harmonized genomic data, including TCGA, TARGET, and CPTAC.

Imaging Data Commons (IDC)
Share, analyze, and visualize multi-modal imaging data from both clinical and basic cancer research studies.

Integrated Canine Data Commons (ICDC)
Share data from canine clinical trials, including the PRE-medical Cancer Immunotherapy Network Canine Trials (PRECINCT) and the Comparative Oncology Program.

Proteomic Data Commons (PDC)
Share, analyze, and visualize proteomic data, such as CPTAC and The International Cancer Proteogenome Consortium (ICPC).

INFRASTRUCTURE

Cancer Data Aggregator (CDA)
Enables users to query and connect data distributed across the CRDC for integrative analysis.

Center for Cancer Data Harmonization (CCDH)
Provides semantic services and tools that facilitate interoperability of data across CRDC.

Data Commons Framework (DCF)
Provides secure user authentication and authorization and permanent digital object identifiers for data objects.

CLOUD RESOURCES

Broad Institute FireCloud
Access NCI-funded datasets TARGET and TCGA along with a rich collection of other datasets and collaborative projects that are part of the biomedical ecosystem. Run analysis tools at scale and collaborate securely on a scalable cloud environment.

ISB Cancer Gateway in the Cloud (ISB-CGC)
Access data sets using fully interactive web-based applications, including BigQuery, which is hosted on Google Cloud Platform.

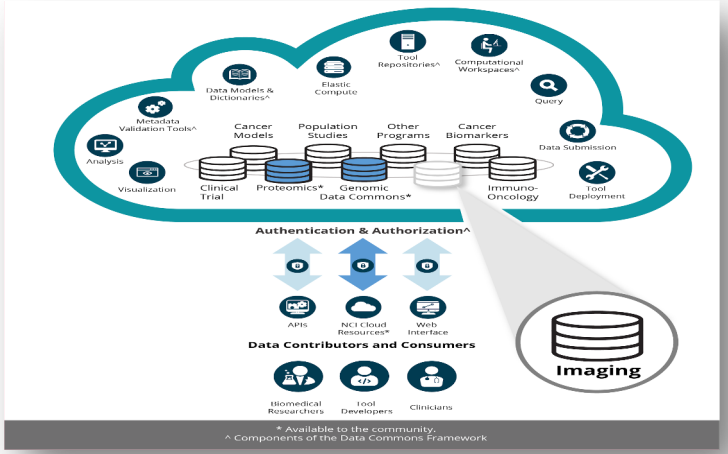
Seven Bridges Cancer Genomics Cloud (SB-CGC)
Explore and analyze large datasets alongside secure and scalable analytical resources for large-scale computational research.

https://datacommons.cancer.gov/

NCI Imaging Data Commons (IDC)

Cloud resource that* connects *researchers with:

- *Cancer image collections*
- *Robust infrastructure with imaging data, metadata and experimental metadata from disparate sources*
- *Resources for searching, identifying and viewing images*
- *Additional data types in other CRDC nodes*
- *Connectivity to NCI Cloud Resources for imaging and multi-modal cloud computations*

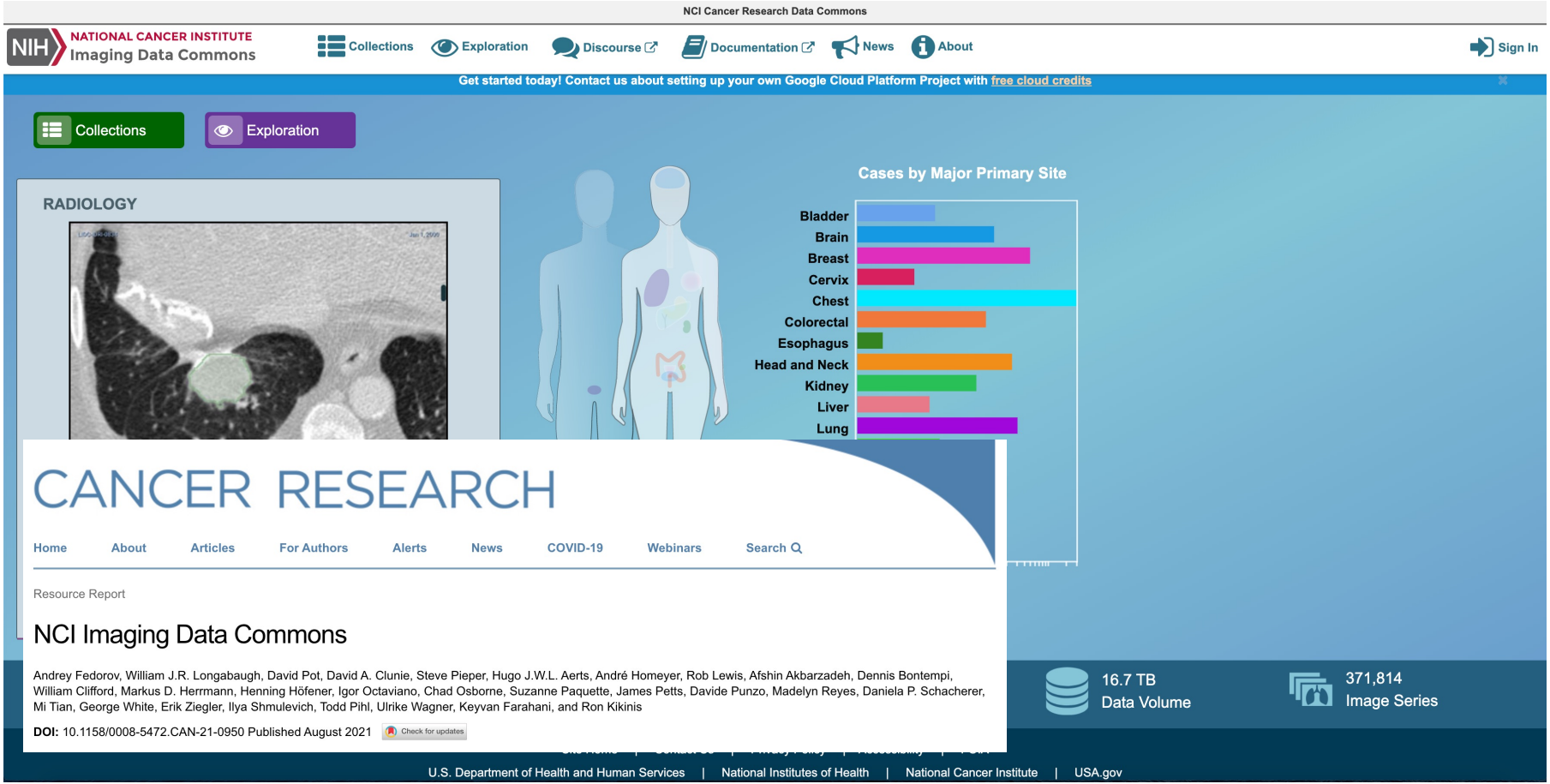


Implementation:

- Google Cloud Platform
- OHIF viewer
- Non-restrictive Open Source
- DICOM as prime standard

Production release: September 2021

https://github.com/ImagingDataCommons/IDC-WebApp



DOI: 10.1158/0008-5472.CAN-21-0950

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

NCI Cancer Research Data Commons

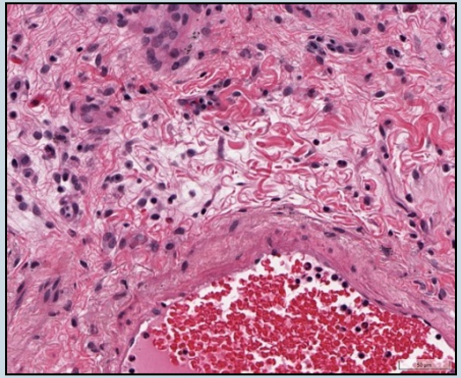
NIH NATIONAL CANCER INSTITUTE Imaging Data Commons

Collections | Exploration | Discourse | Documentation | News | About | Sign In

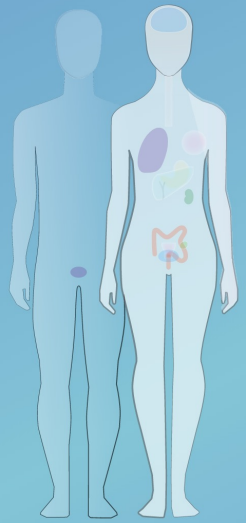
Get started today! Contact us about setting up your own Google Cloud Platform Project with free cloud credits

Collections | Exploration

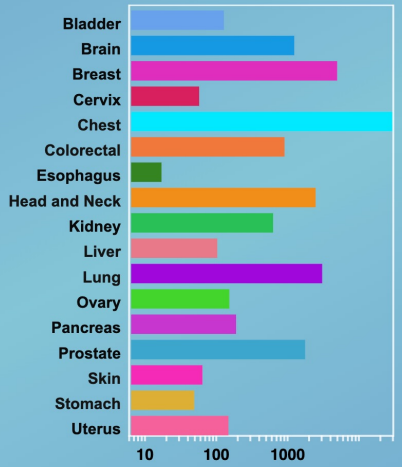
PATHOLOGY



Computed Tomography (CT) | Magnetic Resonance (MR) | Positron Emission Tomography (PET) | Slide Microscopy (SM)



Cases by Major Primary Site



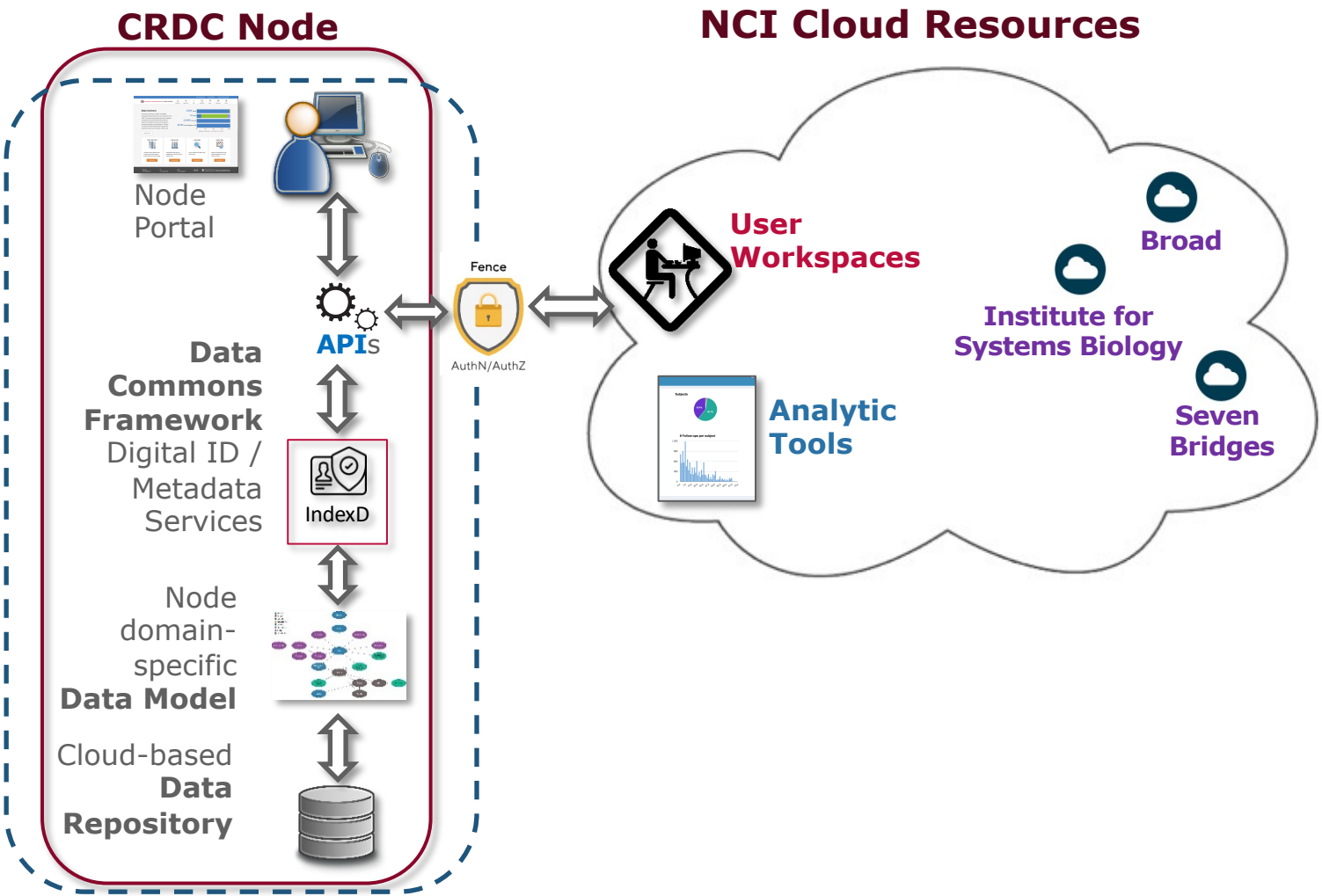
Data Portal Summary Data Release 4.0 September 27, 2021

113 Collections | 43,428 Cases | 16.7 TB Data Volume | 371,814 Image Series

Site Home | Contact Us | Privacy Policy | Accessibility | FOIA

U.S. Department of Health and Human Services | National Institutes of Health | National Cancer Institute | USA.gov

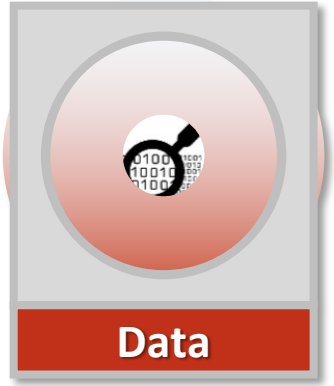
https://portal.imaging.datacommons.cancer.gov/


CRDC Node
NCI Cloud Resources
Node Portal
APIs
Data Commons Framework
Digital ID / Metadata Services
IndexD
Node domain-specific Data Model
Cloud-based Data Repository
Fence
AuthN/AuthZ
User Workspaces
Analytic Tools
Broad
Institute for Systems Biology
Seven Bridges

The NCI Cloud Resources

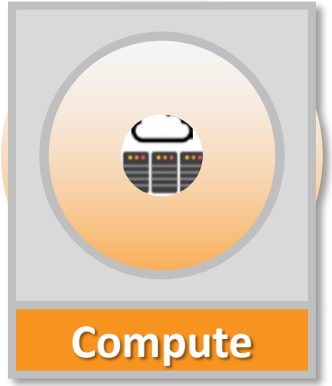
Three resources connecting NCI data and compute in the cloud

- Access to large cancer data sets without need to download
- Access to workspaces, analysis tools, and pipelines
- Ability for researchers to bring their own data and tools



Data

- Access and analyze data from a dozen genomics, proteomics, and imaging datasets without downloading
- Upload your data to the cloud

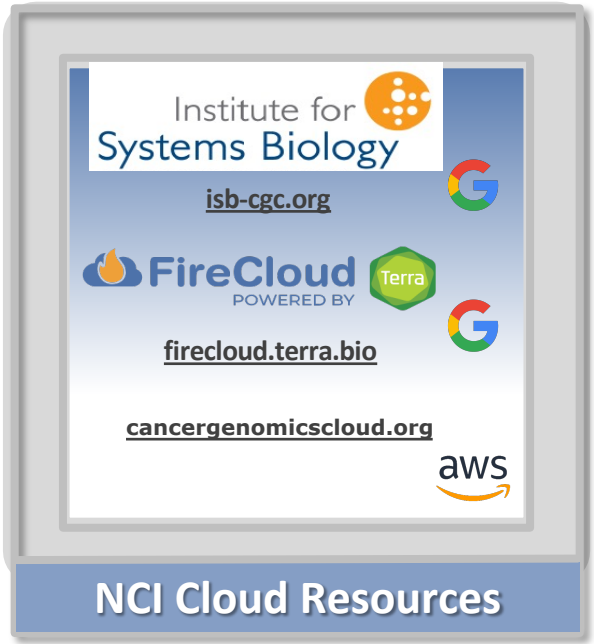


Compute

- Perform large scale analysis using the elastic compute of commercial cloud platforms
- Upload your tools to the cloud, create your own workflows

Security

- dbGaP-authorized users can connect to controlled access datasets
- Systems meet strict Federal security guidelines



NCI Cloud Resources

#NCICloud

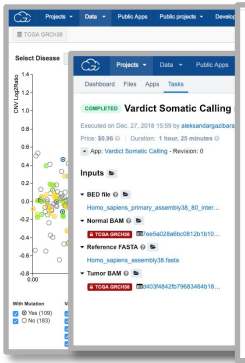
Why Three Cloud Resources?

Great for running production pipelines

FireCloud POWERED BY Terra

Great for non-technical users
User friendly UI, visual displays

CANCER GENOMICS CLOUD
SEVEN BRIDGES



CREDITS TO SUPPORT YOUR RESEARCH

An important goal of the CGC is to understand how researchers can use cloud computing resources to analyze their data. Unlike traditional models, where the cost of compute and data storage is paid up front, on the CGC you only incur costs as you run an analysis.

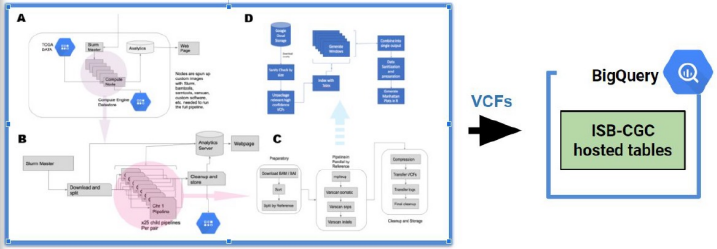
However, we've found that it can be daunting to try to learn a system while worrying about analysis costs. For this reason, the NCI has generously provided substantial funds to support your compute and storage on the CGC as you are getting familiar with the platform. When you create an account on the CGC, you'll be automatically granted **$300** in credits. Have a big project in mind? Submit a collaborative project request to access up to **$10,000** in credits.

Of course, once you've used the credits, you can contact us to create a billing group that can be supported via a credit card or purchase order.

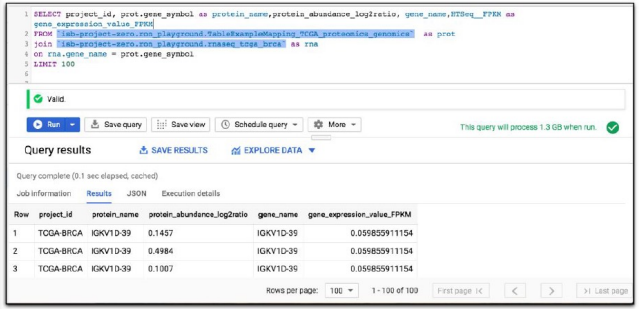
Great for command-line users, BigQuery

Institute for Systems Biology

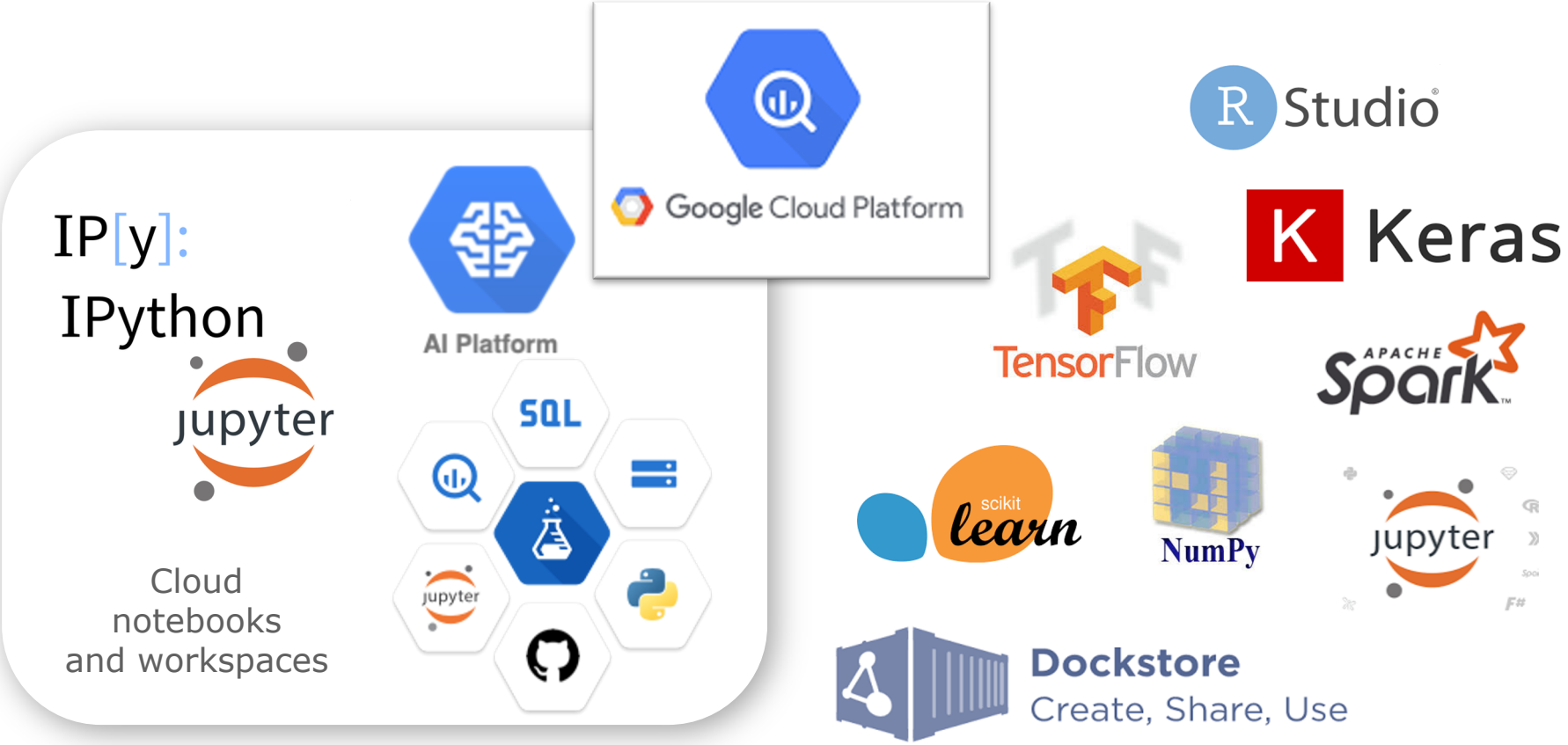
Users can transform results from their own workflows/pipelines to Google BigQuery tables



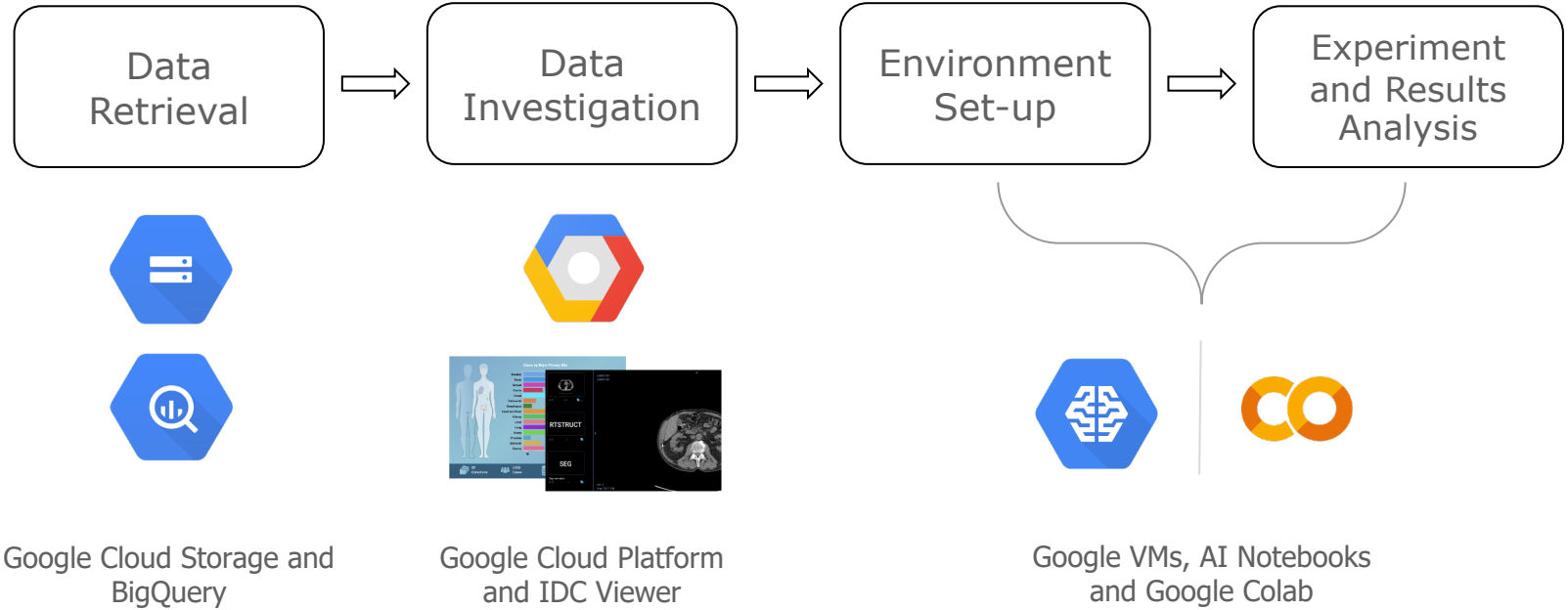
Use SQL to join tables in BigQuery for cross-omics & cross-node data integration



Cancer Cloud Resources



IDC AI workflow



Courtesy of Hugo Aerts (BWH)

IDC's potential for imaging AI

- IDC can play a central role by providing data to enable end-to-end transparent and reproducible AI pipelines for cancer imaging.
- Easy access to high quality, standardized, de-identified imaging and metadata in IDC that can be combined with fully reproducible AI pipelines in cloud based environments.
- Empower AI researchers to reproduce published results, provide materials for research, training and education purposes, as well as guide overall developments of the IDC platform.
- Selected AI use cases for several clinical scenarios in cancer imaging are being developed by IDC and collaborators to highlight these capabilities.

IDC Use Cases

- Essential utilization of IDC/CRDC infrastructure and standards toward:
  - Development of novel AI/ML tools:
    - Applications in imaging – detection, diagnosis, and treatment planning/monitoring
    - Promote transparency, reproducibility and reusability
- Cloud-credits are available to support novel developments

Acknowledgements

Mass General Brigham
Ron Kikinis
Andrey Fedorov
Hugo Aerts
Markus Hermann

Institute for Systems Biology
William Longabaugh

General Dynamics IT
David Pot

Isomics
Steve Piper

Pixelmed
David Clunie

Fredrick National Laboratory for Cancer Research
Todd Pihl
Ulli Wagner

Center for Biomedical Informatics and Information Technology (NCI)
Tanja Davidsen
Allen Dearry

farahani@nih.gov

datascience.cancer.gov

NATIONAL CANCER INSTITUTE

www.cancer.gov

www.cancer.gov/espanol