

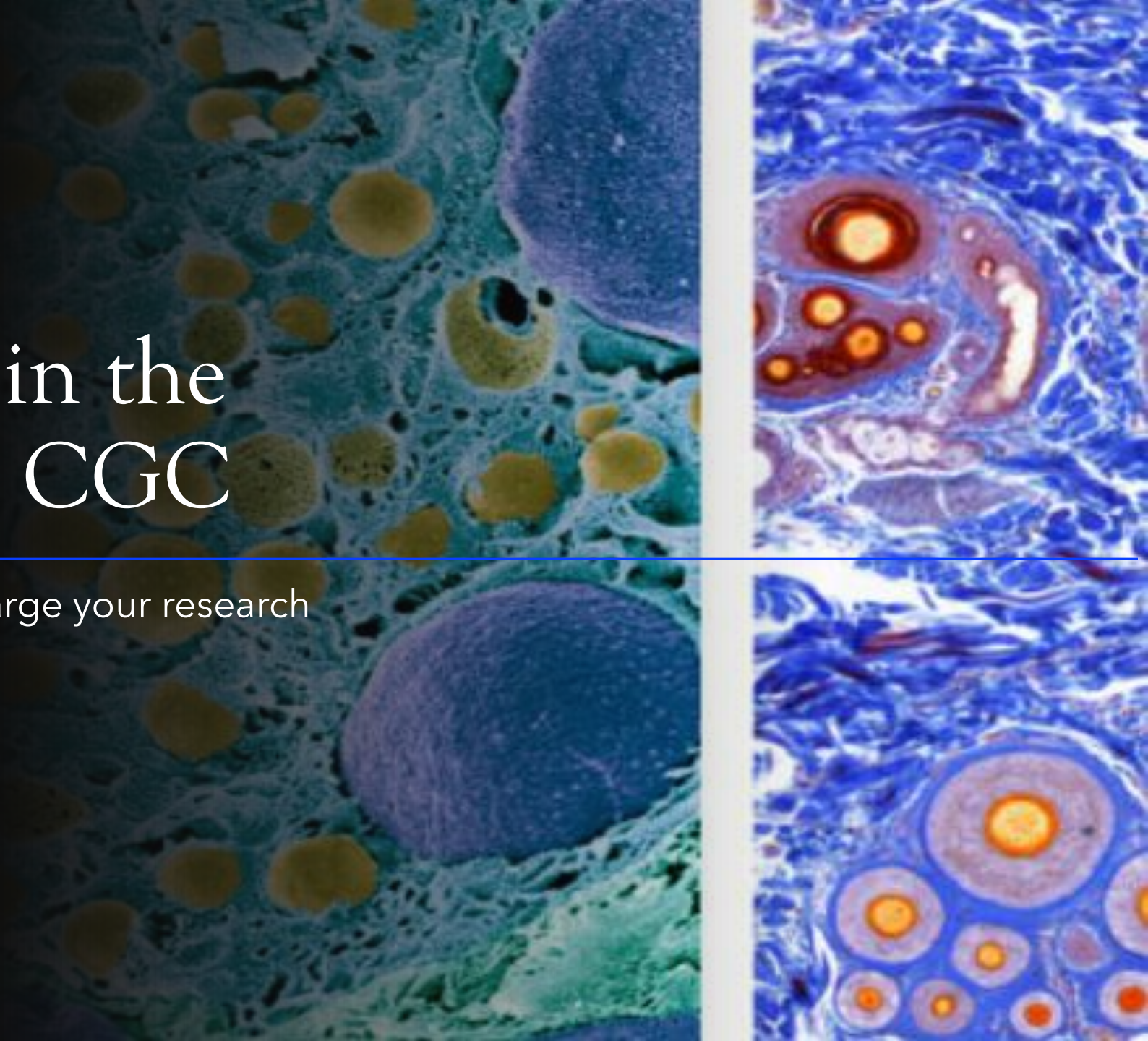
Bioinformatics in the Cloud with the CGC

Leveraging the cloud to turbocharge your research

BTEP webinar

07/21/2026

Jeni Cliffe



Mini-bio – Jenepher A. Cliffe (M.S. Astrophysics)



National Aeronautics and Space Administration
Goddard Space Flight Center

IMAGINE THE UNIVERSE!

[Home](#)
[Science](#)
[Observatories](#)
[Special Exhibits](#)
[Educators](#)
[Ask An Astrophysicist](#)
[Resources](#)
[Search](#)

What does

The best way to look at the universe is to wear the shoes of one, but this is the next best thing. We have the tools to solve many different things.

Science is a process. In these pages you are seeking.

At least one of the above methods will give you the correct answer. Which one do you want to try?

$$I = \frac{I_0}{r^2}$$

Click here to use M31's lightcurve and the $1/r^2$ relationship

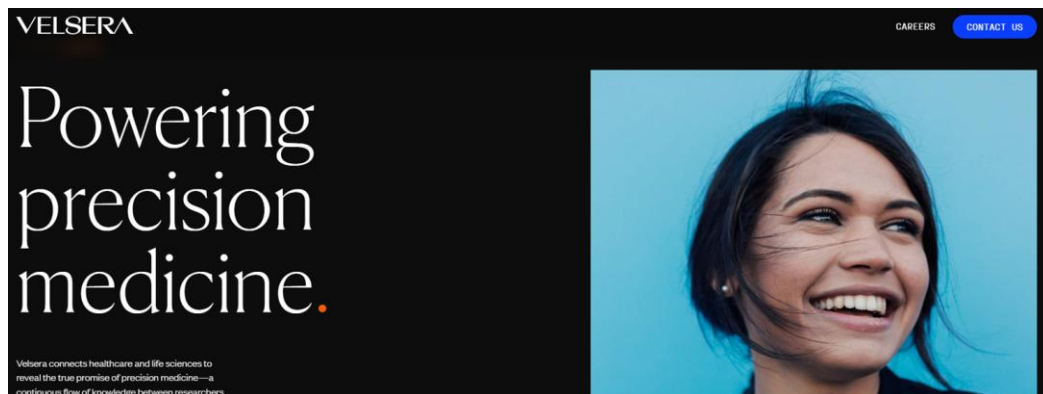


Click here to use M31's spectrum and the Doppler shift

$$H_0$$

Click here to try Hubble's Law

Community Engagement Manager, Velsera



- Make NIH research better by:
 - Helping scientists and doctors analyze their research data on the **Cloud**
 - Teaching people how to work with and manage **big data**
 - Giving people better **tools** so research happens faster

From exploding stars to –
humans?!

Agenda.♦

Bioinformatics in the Cloud with CGC

- 01** Introduction to the CGC and Cloud Computing

- 02** Accessing data and running tools (overview)

- 03** Variant analysis run-through

- 04** New feature highlights: OHIF viewer and 3D Slicer

Introduction to Cloud Computing with the Cancer Genomics Cloud

Let's take a poll!



How comfortable are you with
the cloud and cloud data?

1



Not
comfortable
*I've used it very little,
or never*

2



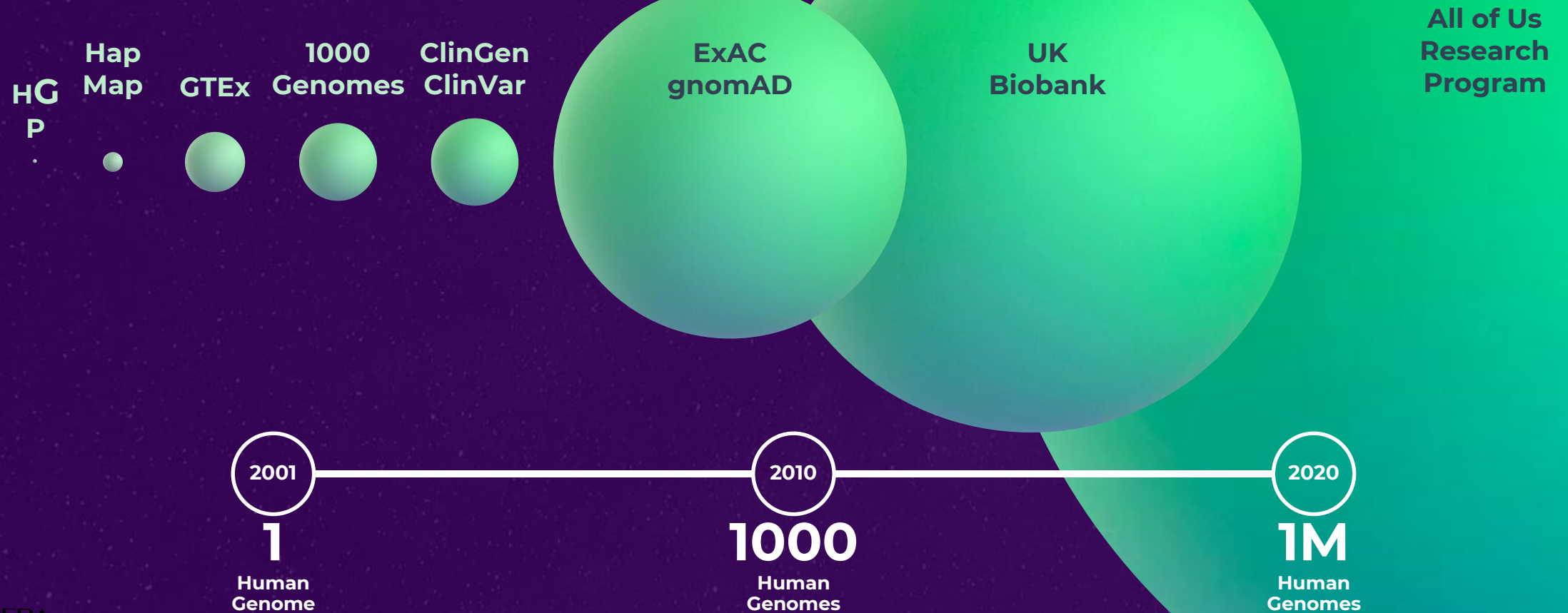
Somewhat
comfortable
I've used it before

3



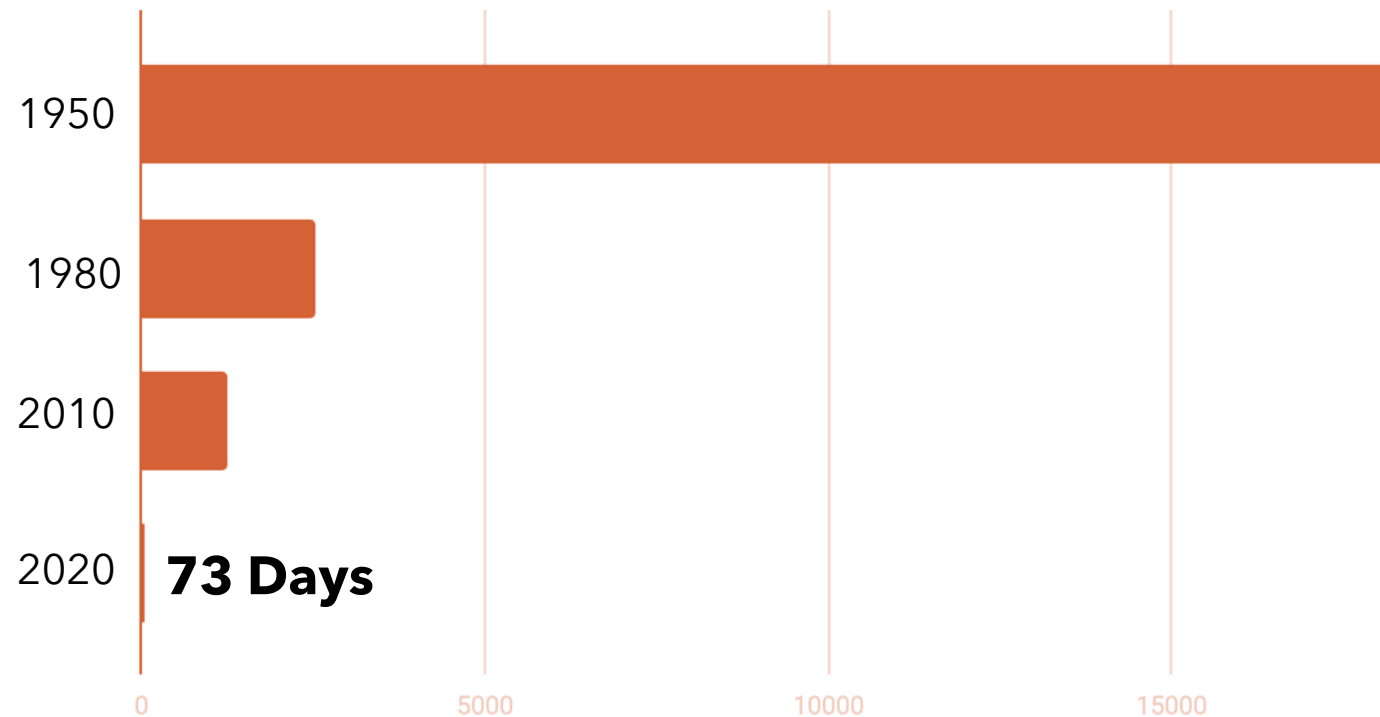
Very comfortable
I use it all the time

Data generation is on a roll...



The Era of Big Data in Health

Doubling Time of Health Knowledge

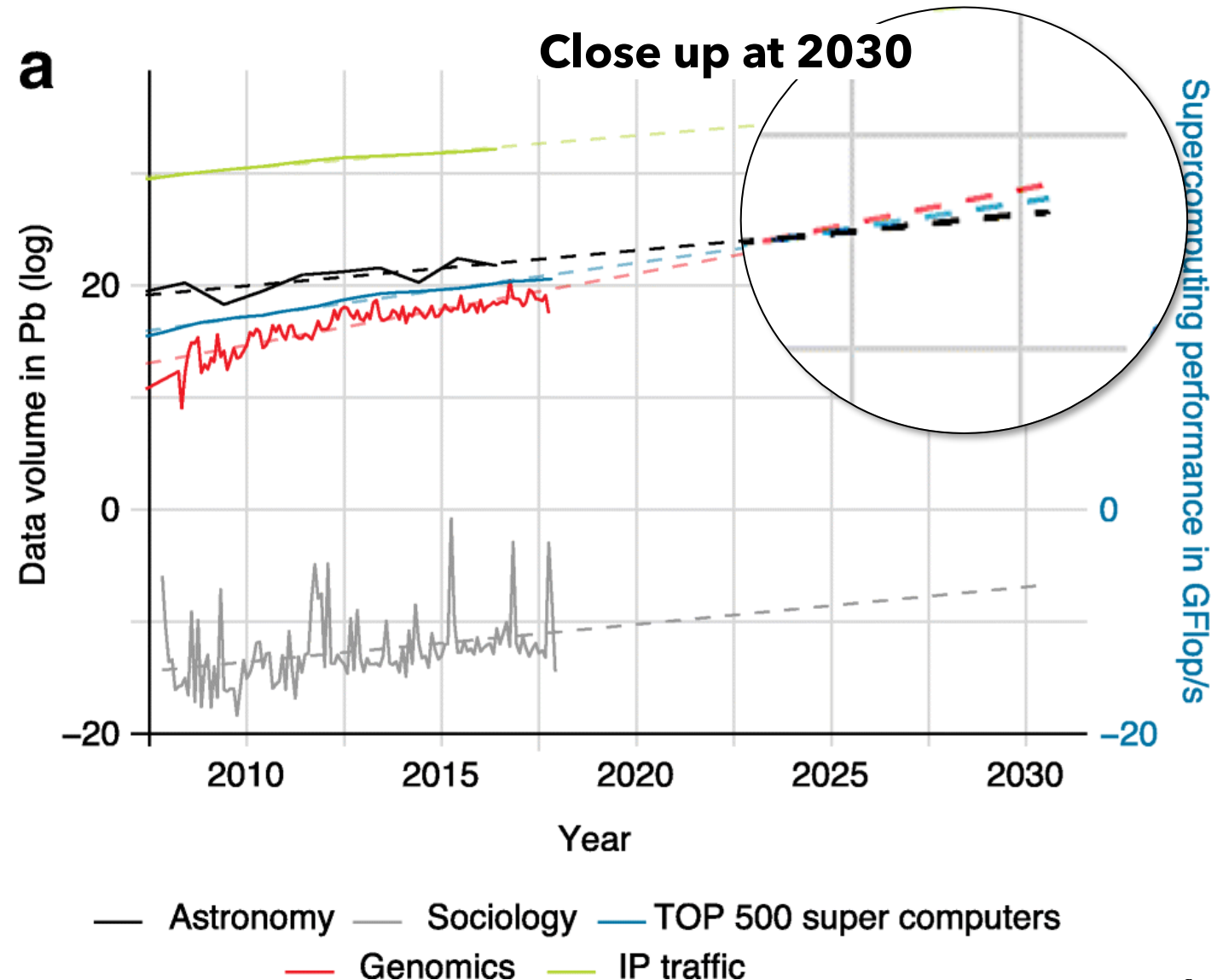


More biomedical data were generated in 2021 than all previous years **combined**

More different **kinds** of data – gene and genome sequencing, imaging, sensors (i.e. Fitbit), cellular measurement data, and more

How much data is it?

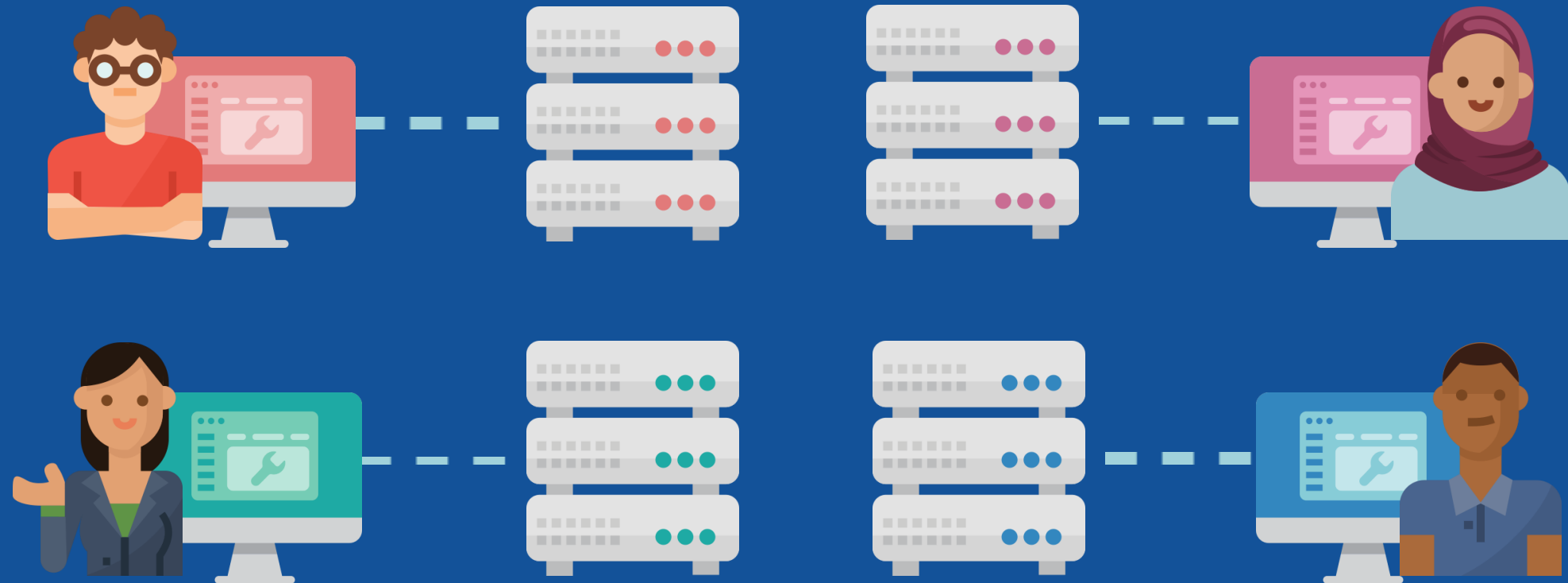
- Estimated between 2 and 40 exabytes of new data per year (Stephens *et al.* 2015)
- Genomics data is growing **faster than *the internet***
- Analysis-to-Data Paradigm – analyze data “where it lives”



How do we share and analyze datasets?

Traditional Approach

Bring data to researchers



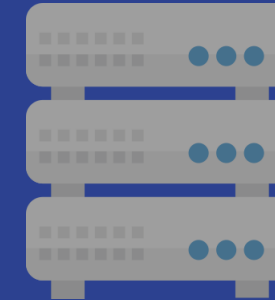
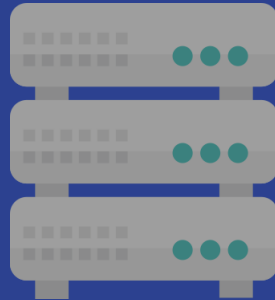
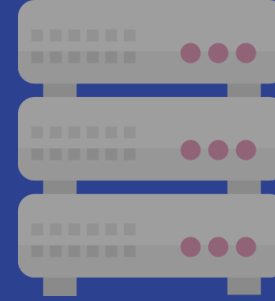
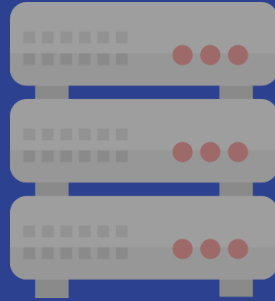
What's wrong with this traditional approach? ?

Data is
copied, not
shared

High storage &
infrastructure
costs

Hard to
share/
reproduce

Security on
individuals



A new cloud-native vision for biomedical research



A new cloud-native vision for biomedical research

~~Legacy
copies, not
shared~~

~~High storage &
infrastructure
costs~~

~~Hard to
share/
reproduce~~

~~Security on
individuals~~

True
immediate
data sharing

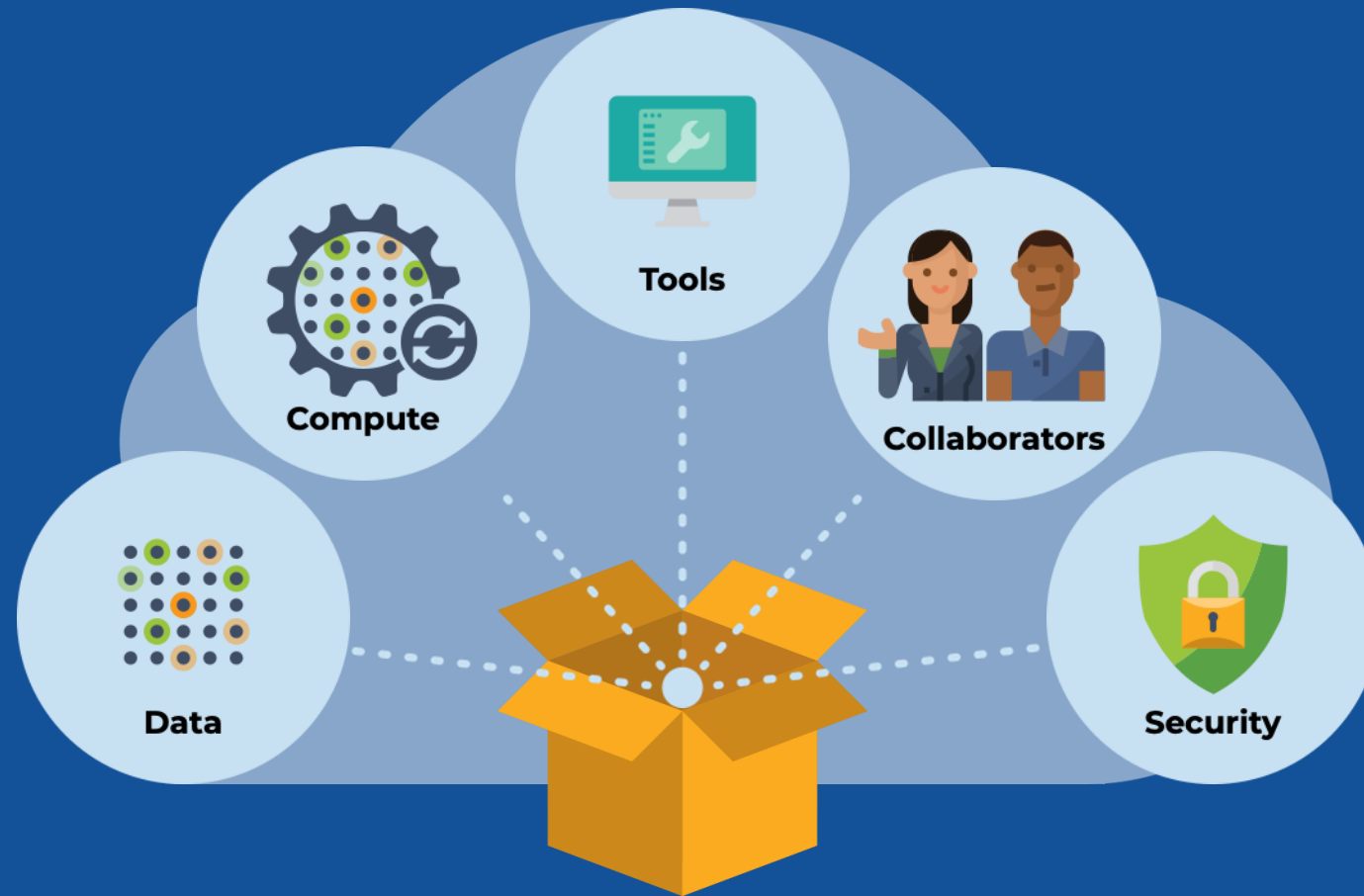
Single storage,
many can
access. Elastic
storage &
compute (pay
for your use)

Shareable,
reproducible
tools and
standardized
environments

Centralized
security

A new cloud-native vision for bioinformatics

**Migrating to the cloud
can be challenging!!**



CANCER GENOMICS CLOUD

SEVEN BRIDGES



cancergenomicscloud.org

A new cloud-native vision for biomedical research

The screenshot shows the GENESIS Public Project interface. The top navigation bar includes links for Home, Projects, Data, Public Apps, Public Projects, and Developer. The main header displays the project title "JAC_copy_WGS Data Analysis with GENESIS Public Project" and options for Interactive Browsers, Settings, and Notes. The left sidebar contains a Description section with a search bar and a list of analyses. The right sidebar shows a Members section with a search bar and a list of analyses. Three callout boxes are overlaid on the interface:

- Access**: Immediately & seamlessly (including controlled data)
- Analyze**: Easy to use and find curated tools and resources
- Share**: As easily as a Google doc, with granular permissions for control

Description

WGS Data Analysis with G

This project is designed to introduce the user to the (SeqArray, SeqVarTools, and SNPRelate) used to perform WGS data. It consists of interactive analyses with example code that is used in GENESIS public apps, prepared with the results. Also, there are several example task apps for performing the analysis that utilize the code. Also covers use of the STAAR Pipeline public application.

The code in this project was developed as a series of exercises for the Bruce Weir Summer Institute in Statistical Genetics (SISG), and is also available on GitHub: https://uw-gac.github.io/SISG_2024.

Analyses covered in the interactive GENESIS Tutorial 2024

01_gds_intro.Rmd notebook covers:

- Converting a VCF to GDS

Members

renee_sears
Copy, Write, Execute

Search

COMPLETED GDS Genotype Extractor run - 09-30-24 11:57:17
Submitted by: jenepher_cliffe_era · Jul 20, 2026 12:24

COMPLETED STAARexercise_example_STAARpipelineSummary_VarSet_genebased
Submitted by: jenepher_cliffe_era · Jul 20, 2026 12:24

COMPLETED STAARexercise_example_STAARpipelineSummary_IndVar_gene_centric
Submitted by: jenepher_cliffe_era · Jul 20, 2026 12:24



CANCER GENOMICS CLOUD

SEVEN BRIDGES

Powerful, yet easy to use interfaces to empower cancer researchers to draw new insights from petabyte scale data.

Stable, secure, and highly customizable cloud storage and computing platform

3+

Petabytes
Public Data

2.4M+

Years of
Compute

1300+

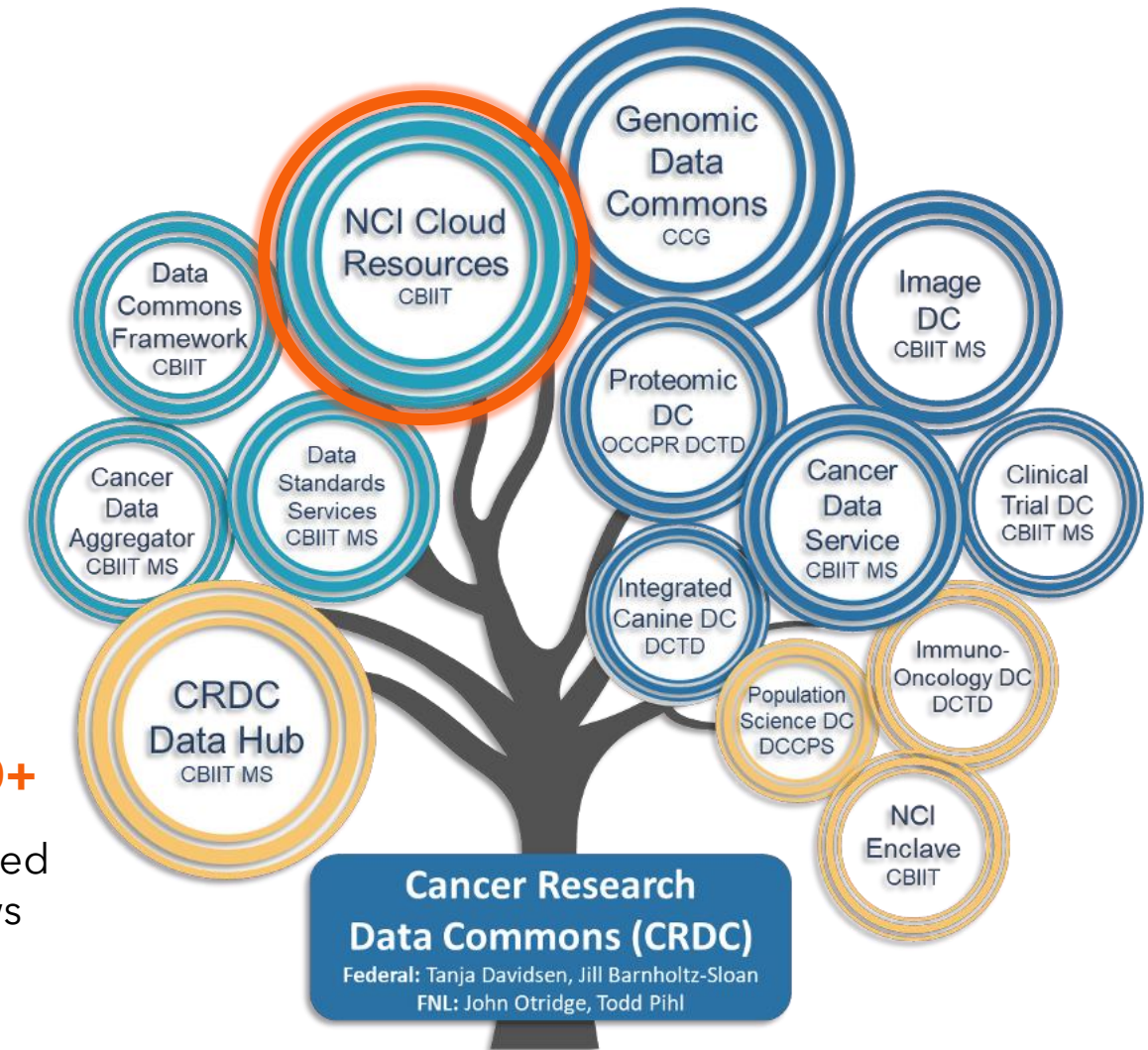
Public Tools
& Workflows

12,500+

Users

109,000+

User-Created
Workflows



Data Commons



Data Interoperability and Analysis



Future Components

Who are CGC Users?

The CGC is designed to serve a wide range of scientists and users with varying skill sets



BIOINFORMATICIANS

- Store, manage, and share data
- Access public and proprietary datasets
- Query, build, and investigate cohorts of interest
- Access optimized tools and workflows
- Create, optimize, maintain, and distribute new tools and workflows
- Create push-button automation solutions
- Analyze data at scale with tools and workflows
- Conduct interactive exploratory analyses
- Explore/visualize results/gather insights
- Easily collaborate with stakeholders
- Integrate with external systems



BENCH SCIENTISTS

- Store, manage, and share data
- Run optimized tools/workflows at scale
- Conduct defined analyses via push-button solutions
- Investigate/visualize results
- Easily collaborate with other stakeholders



CLINICIANS

- Conduct validated analyses via push-button solutions
- Query, build, and investigate cohorts of interest
- Create reports
- Investigate/visualize results
- Easily collaborate with other stakeholders



DEVELOPERS

- Create, optimize, and maintain new tools and workflows
- Create push-button automation solutions
- Create custom interfaces for specific use cases
- Distribute proprietary tools/workflows
- Integrate upstream/downstream



ADMINISTRATORS

- Manage and control users
- Monitor and control institutional assets
- Manage and monitor projects
- Monitor and control costs
- Create reports

Navigating the Cancer Genomics Cloud

“Projects” are the main unit of Organization

The screenshot shows the 'AIDR - Project Template' page in the CGC SB Genomics application. The interface includes a top navigation bar with links for Projects, Data, Public Apps, Public Projects, and Developer. A secondary navigation bar contains Dashboard, Files, Apps, Tasks, and Data Studio. The main content area is titled 'AIDR - Project Template' and includes a description of the challenge submission guidelines. A callout box highlights that projects can contain files, apps, workflows, collaborators, interactive sessions, and notes. The user profile 'new_cgic_user_demo' with 'ADMIN' privileges is visible in the top right. The bottom section shows a search bar and a message about executions.

https://cgc.sbgenomics.com/u/cfisher92/aidr-project-template

Projects ▾ Data ▾ Public Apps ▾ Public Projects Developer ▾

new_cgic_user_demo ▾

Dashboard Files Apps Tasks Data Studio

AIDR - Project Template ⓘ

Interactive Browsers Settings Notes

Email notifications

new_cgic_user_demo ADMIN
Copy, Write, Execute, Admin

ve project

Search

Tasks Data Studio

Your executions will appear here. Before you start, [learn more about them.](#)

Projects can contain

- Files
- Apps and workflows
- Collaborators
- Interactive sessions
- Notes

Description

AIDR **submission**

Guidelines for Challenge Submission

These are taken from the "How to Enter" tab at the Challenge challenge: [Challenge.Gov](#)

A. Analyze Data and assess its AI Readiness

- Qualitative assessment (through a manual review of documentation and governance metrics that outputs a description of the current landscape)
- Quantitative assessment (through a script that outputs some metrics measures such as data completeness and consistency)

B. Process the data to improve its readiness and build an AI/ML model

- Resolve any data issues and present data in a higher AI/ML readiness as needed (through a "ml_model" script)
- Build an AI/ML model that takes the processed data and produce a model (through the same "ml_model" script)

Access Public Data, or Use Your Own

 Projects ▾ Data ▾ Public Apps ▾ Public Projects Developer ▾  ▾ rowan_beck_era ▾

Dashboard Files Apps Tasks Data Studio **Demonstration: Building an App**  Interactiv... browsers Settings Notes

 Files

Extension: All ▾ Sample ID: All ▾ Task ID: All ▾ Tags: All ▾ + Clear filters

☐ ▾ ▲ Name	Task ID	Created on	Extension	Size
☐ G20479.HCC1143.2_1Mreads.tar.gz TEST	-	Aug. 10, 2023 11:44	TAR.GZ	115.0 MiB
☐ G20479.HCC1143.2_1Mreads_pe_1.fastq	8f6e866e-b767-45af-...	Aug. 10, 2023 11:46	FASTQ	232.2 MiB
☐ G20479.HCC1143.2_1Mreads_pe_2.fastq	8f6e866e-b767-45af-...	Aug. 10, 2023 11:46	FASTQ	232.2 MiB



New folder

+ Add files ▾

Case Explorer and Data Browser

Public Files

Projects

Your Computer

FTP / HTTP

GA4GH Data Repository Service (DRS)

Data Tools

Volumes

Import from a manifest file

Files and file management

The screenshot shows the Velsera web interface. The top navigation bar includes 'Home', 'Projects', 'Data', 'Public Apps', and 'Public Projects'. Below this, a secondary bar has 'Dashboard', 'Files' (selected), 'Apps', 'Tasks', 'Data Studio', and 'Interactive Apps'. On the right, there are links for 'Interactive Browsers', 'Settings', and 'Notes', along with a user profile 'JC'. The main content area shows a file browser for 'Root / input_data'. It includes a search bar, filters for 'Extension' and 'Tags', and a '+ Add files' button. A sidebar on the left shows a tree view with 'Files' and 'input_data' (selected). The main list displays two files with checkboxes and metadata tags like 'WES', 'Tumor sample', and 'GRCh37'. A purple callout box with a blue gradient border is overlaid on the right side of the interface.

The Files tab has

- Folders
- Text search
- Filter by metadata
- Filter by Tags

LEARN

File Management:
<https://youtu.be/Z4ygBEBunuY>
Knowledge Center:
[Files and metadata](https://cancer.genomicscloud.org/files-and-metadata)
(cancer.genomicscloud.org)

No coding required to run an analysis

The screenshot displays the Velsera web interface for configuring a task. The top navigation bar includes links for Projects, Data, Public Apps, Public Projects, and Developer, along with a user profile for rowan_beck_era. The main header shows the task name "Copy of Bulk RNA-Seq Transcription Profiling of HSV-1 Infecte..." and various utility links like Dashboard, Files, Apps, Tasks, Data Studio, Interactive Browsers, Settings, and Notes. The task is currently in "DRAFT" status and is titled "Differential Expression - Salmon + DESeq2 run - 11-30-23 17:12:19". It was last updated by rowan_beck_era on Nov. 30, 2023 at 12:12. The app is identified as "Differential Expression - Salmon + DESeq2" with revision 1. Below the header, there are tabs for "Task Inputs" and "Execution Settings". The "Task Inputs" tab is active, showing three main sections: Inputs, App Settings, and Output Settings. The Inputs section includes a "Batching" toggle set to "Off", a list of "FASTQ read files" (SRR9058997_1.fastq, SRR9058993_1.fastq, SRR9058992_2.fastq, SRR9058992_1.fastq, SRR9058991_2.fastq, and 25 more items), a "GTF annotation" (GRCh38ERCC.ensembl95.gtf), and a "Genome FASTA" field. The App Settings section includes "DESeq2" parameters: "Covariate of interest" set to "Genotype", "Factor level - reference" set to "WT", and "Factor level - test" set to "KD". The Output Settings section lists various output options, all of which are currently set to "No value". An "Activity Monitor" button is located at the bottom right of the configuration area.

Inputs

Batching ☐ Off

FASTQ read files * [Change selection](#)

- SRR9058997_1.fastq
- SRR9058993_1.fastq
- SRR9058992_2.fastq
- SRR9058992_1.fastq
- SRR9058991_2.fastq
- ...and 25 more items

GTF annotation * [Change selection](#)

- GRCh38ERCC.ensembl95.gtf

Genome FASTA [Select file\(s\)](#)

No files selected

App Settings

[Edit parameters](#) [Show editable](#)

DESeq2 (#deseq2_1_26_0)

Covariate of interest *

Factor level - reference

Factor level - test

Output Settings

DESeq2 analysis results. ?	No value
Expression matrix genes ?	No value
Expression matrix transcripts ?	No value
Gene-level quantification ?	No value
HTML report ?	No value
HTML reports ?	No value
Normalized counts ?	No value
RData file ?	No value
Report zip ?	No value
Salmon Quant archive ?	No value
Salmon quant log ?	No value
Transcript-level quantification ?	No value
pheno out	No value ?

Activity Monitor

Over 1300 curated tools and workflows



High quality apps & documentation



Customized user tools/workflows



Updated regularly



Secure



Optimized to run on cloud

Variant Filtration
SAM/BAM-Processing
Copy Number Analysis
Quantification
Structural Variant Calling
Quality Control
Proteomics
Gene Fusion
FASTQ Processing
FASTA Processing
BED Processing
Filtering
RNA-Seq
Imaging
ChIP-Seq
Differential Splicing
Long Reads
Other Variant-Calling
Epigenetics
Single Cell
Annotation
Indexing
Alignment
Variant Calling
Metagenomics
File Format
Immune Epitopes
VCF Processing
sam/BAM Processing
RNA-Seq
ChIP-Seq
Differential Splicing
Long Reads
Other Variant-Calling
Epigenetics
Single Cell
Annotation
Indexing
Alignment
Variant Calling
Metagenomics
File Format

Apps and Workflows:

<https://youtu.be/ayCKOafCGB0>

Knowledge Center:

[Workflows and tools
\(cancer-genomics-cloud.org\)](https://cancer-genomics-cloud.org)

Data Studio – Interactive Data Analysis

Data Studio

Derive new insights using interactive analysis environments with Data Studio environments. Code in Python, SAS, and R to record and share your analyses.

Interact with files directly using our OHIF image viewer and Galaxy.

Videos:

R Studio on the CGC: https://youtu.be/_f8DWRTS0rE

JupyterLab on the CGC: <https://youtu.be/iycWVHhqgXg>

Documentation:

About Data Studio (cancergenomicscloud.org) | <https://docs.cancergenomicscloud.org/docs/about-data-cruncher>

Apps – Bulk Processing Analysis

Command Tools and Workflows

- Convert any R or Python or bash script into an app that can be optimized and automated.
- Automated and optimize to scale up.



Videos:

CGC Onboarding:

<https://youtube.com/playlist?>

Public Apps Gallery:

<https://youtu.be/ayCKOafCGB0?si=ps0O3ihGvgNvanB>

Documentation:

Apps on the CGC

(cancergenomicscloud.org) |

<https://docs.cancergenomicscloud.org/docs/apps-on-the-cgc>

CGC Democratizes Complex Analyses



Easy data
management



Secure
collaboration &
managed billing



Flexible & fully
reproducible
methods



Optimized
bioinformatics
algorithms



Scalable
computation



Extensible &
developer
friendly tools

A FAIR data ecosystem

- **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (FAIR)
- Immediate scaling (“Rent” vs “Buy”)
- Levels the playing field across institutions
- Many researchers can access data - no duplicate copies.
- Reproducibility - data and methods together.



HANDS-ON DEMO

cgc.SBGenomics.com

Variant Calling

Coming soon to the Cancer Genomics Cloud

Quantitative QC for Lung CT: 3D slicer

Dashboard Files Apps Tasks Data Studio Interactive Apps Copy of COVID-19 Image Segmentation with Deep... Interactive Browsers Notes

Description

This project provides deep learning image segmentation notebook, along with a pretrained model for segmentation images. The model was trained on a publicly available dataset. Analysis files include:

- A Jupyter notebook walking through the model. This notebook can be repurposed for other applications by replacing the image preprocessing code and datasets.
- A Keras model based off of a UNet architecture with an encoder pretrained with ImageNet weights.
- Images and masks in NIfTI (.nii) format, including masks for ground glass opacities and consolidation.
- Images and lung masks in NumPy array format. The images are shuffled from the order in the NIfTI files.
- A plot of loss history over the course of training. This will be overwritten if a new model is trained.
- Utility and dependency files.

To use the tools, select **Interactive Analysis > Data Cruncher > JupyterLab**. The notebook can be run using CPU, single-GPU, or multi-GPU instances in the **SB Machine Learning environment**. To use the model as is for predictions,

What it does

- Computes spatial overlap and boundary metrics
- Quantifies volume bias and over-segmentation
- Categorizes masks into pass, review, or fail
- Generates auditable segmentation quality records

🔔 Email notifications

yachna

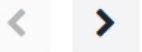
Copy, Write, Execute

yachnajain

Copy, Write, Execute

jenepher_cliffe_era_login

Copy, Write, Execute



🔍 Search

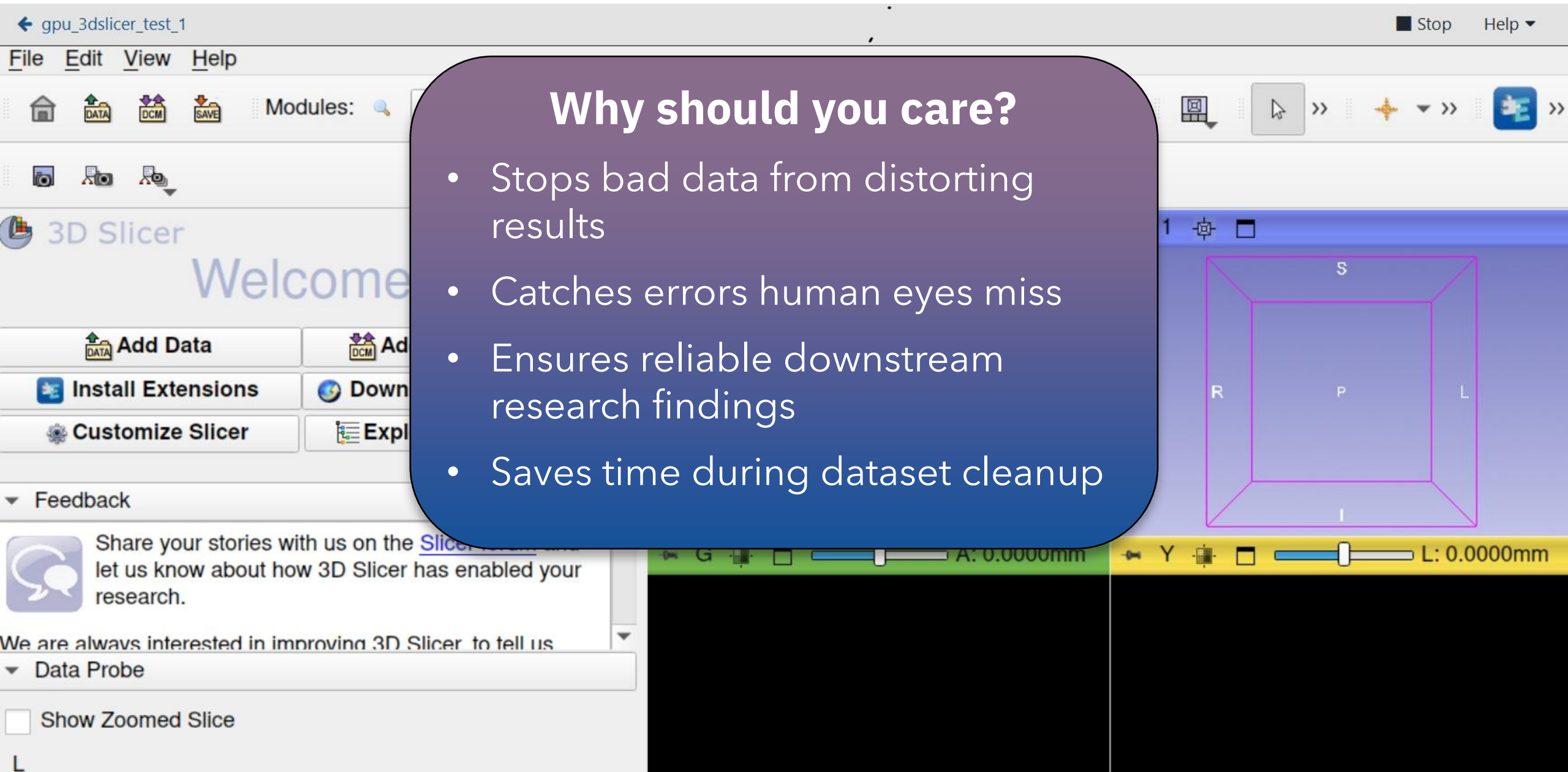
Tasks Data Studio

COMPLETED

SBG Compressor CWL1.0 run - 06-29-26 07:48:48

Submitted by: yatishp · Jun 29, 2026 3:48

Quantitative QC for Lung CT: 3D slicer



Why should you care?

- Stops bad data from distorting results
- Catches errors human eyes miss
- Ensures reliable downstream research findings
- Saves time during dataset cleanup

OHIF viewer: Data-Driven QC for Brain MRI

[Dashboard](#) [Files](#) [Apps](#) [Tasks](#) [Data Studio](#) [Interactive Apps](#)

the-cancer-imaging-archive-tcia ⓘ

[Interactive Browsers](#) [Notes](#)

Description

The Cancer Imaging Archive and Public Project

This project contains radiological imaging data from The Cancer Imaging Archive (TCIA), focusing on collection of data from The Cancer Genome Atlas (TCGA). This data is made available to the public and is focused on connecting genotypes to phenotypes. The dataset includes radiological imaging data stored in the standard DICOM format and is identified to ensure that they are free of bias and they are available for public use.

Use the data in this project

The TCIA data files in this project are ready to be explored by using the file filters to select files based on their specific extensions or by modality using the "Modality" metadata field. For example, if you are interested in MRI images from patients with Glioblastoma Multiforme, you can filter for files with "MR" in the "Modality" metadata field.

What it does

- Standardizes MRI quality control workflows
- Validates DICOM metadata against physical measurements
- Measures tumor dimensions and volume
- Quantifies image noise and artifacts
- Exports machine-readable DICOM Structured Reports

 [Email notifications](#)

yachnajain

Copy, Write, Execute

yatishp

Copy, Write, Execute

jenepher_cliffe_era_login

Copy, Write, Execute



 Search

COMPLETED

FSL Brain Segmentation run - 07-01-26 10:03:39

Submitted by: anushuab · Jul 1, 2026 7:52

OHIF viewer: Data-Driven QC for Brain MRI

Dashboard

Files

Apps

Tasks

Data Studio

Interactive Apps

the-cancer-imaging-archive-tcia ⓘ

Interactive Browsers

Notes

Search names and description

Cost Estimator: All ▾

Create app

+ Add apps ▾

Name ▲

FSL Brain Segmentation

This workflow is used to analyze brain M

SBG TCIA Archive Viewer

The TCIA Archive Viewer consists of two s

Why should you care?

- Prevents silent model training failures
- Ensures reproducible segmentation dataset quality
- Eliminates inconsistent multi-site image data
- Automates downstream quality audit trails
- Saves manual dataset review time

difi...

Modifi...

huab

Jul 01, 202...

Run ...

huab

Jul 01, 202...

Run ...

Showing 1 - 2 of 2 < >

VELSERA

33



INTEGRATED TOOLS ENABLING

- Streamlined data access
- Reproducible
- High-throughput gating for AI pipelines
(garbage in, garbage out protection)
- Democratized computational oncology

Resources for Working on the Cancer Genomics Cloud

You are not alone!

Onboarding resources

- \$300 pilot cloud credits ([how to apply](#))
- [Getting Started playlist](#) on YouTube
- Extensive [Knowledge Center](#)



Public apps and projects

- Curated tutorials and templates to adapt to your data.
- Vetted and tested apps and tools

New Public Projects (Template projects)	37
Public Apps (Published via projects)	33
Public Shiny-based Apps (PUBLIC_WEB_APP)	15

Get The Support You Need

FAILED Task 1 - StringTie run [Get support](#) [View stats & logs](#)

Executed on June 12, 2020 07:53 by [sevenbridges](#)

Spot Instances: **On** | Memoization (WorkReuse): **Off** | Price: **\$0.01** | Duration: **5 minutes**

App: **StringTie** - Revision: 0

Error:
This task ran into a problem during execution and did not finish.
[Show details](#)

Inputs

- Aligned reads**
[HCC1143-CCLE-RNASeq-subset01.genome_aligned...](#)
- Reference annotation file**
[Homo_sapiens.GRCh38.84.gtf](#)

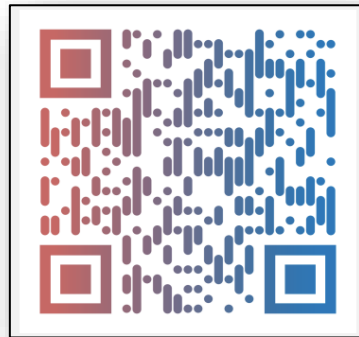
App Settings [Show all](#)

Create input files for Ballgown and DESeq2	No value
Disable trimming	True
Ignore alignments on the specified sequence	No value
Keep annotated transcripts only	No value
Maximum fraction of multiply mapped reads	0.95
Minimum anchor length for junctions	10
Minimum isoform abundance	0.1
Minimum isoform length	200
Minimum junction coverage	1
Minimum locus gap separation value	50
Minimum read coverage	2.5
Number of threads	2
Output covered reference transcripts	False
Output gene abundance	False
Transcripts name prefix	STRG

Output Settings

Archived ballgown input tables	No value
Assembled transcripts	No value
Covered reference transcripts	No value
DESeq2 gene count matrix	No value
DESeq2 transcript count matrix	No value
Gene abundance estimation	No value

Get The Support You Need



Every Week:

- 10:00 am ET Tuesday
- 2:00 pm ET Thursday

NCI Funding Is Available on the CGC

Pilot Credit Funds

- **\$300 of cloud credits**
- Free for new CGC users
- Easy to request when signing up (see [Getting Started – Cancer Genomics Cloud](#))
- Fast approval



Collaborative Projects

- Cost estimation, optimization, and planning support
- Great for researchers new to bioinformatics and cloud approaches
- **Up to \$10k compute/storage costs**
- Fast, rolling applications
- To date > 60 projects
- Apply at [Collaborative Projects – Cancer Genomics Cloud](#)



Support@Velsera.Com

<https://meet.google.com/kbs-ojnj-dcg>



Every Week:

- 10:00 am ET Tuesday
- 2:00 pm ET Thursday

Stay in touch.



Jeni Cliffe

Community Engagement Manager
Jenepher.Cliffe@velsera.com



Michelle Engle, PhD

Community Engagement Manager
Michelle.Engle@velsera.com

Thank you.♦

Learn more at
CancerGenomicsCloud.org

VELERA



COMING SOON!
cgc.SBGenomics.com

OHIF viewer and 3D slicer