Introduction to Bioinformatics Resources at NIH

Amy Stonelake, Ph.D.
NCI CCR Bioinformatics Training and Education Program (BTEP),
Program Manager

August 7, 2025

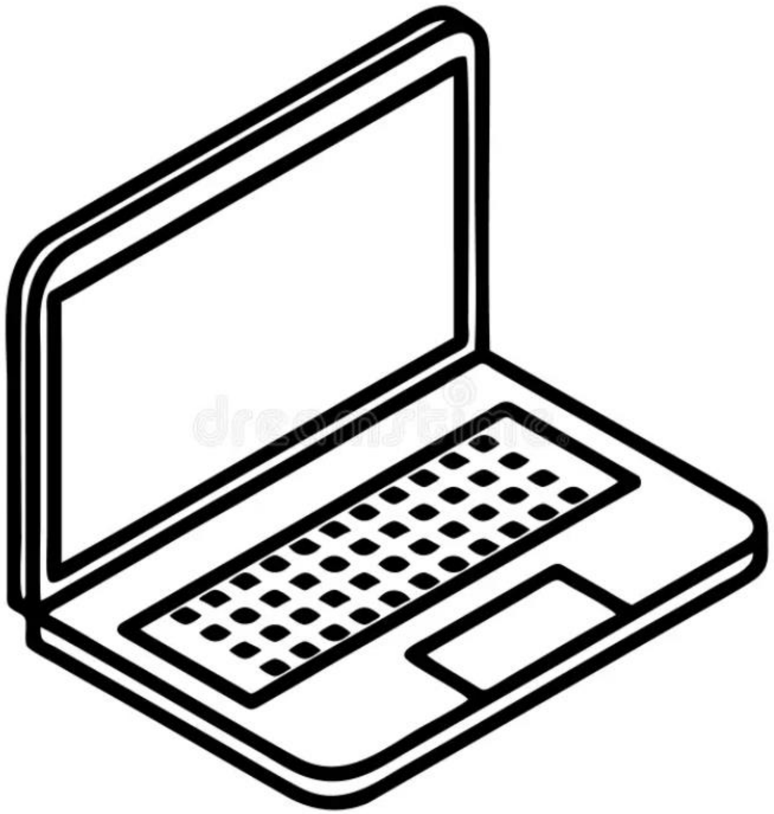
Bioinformatics Training and Education Program

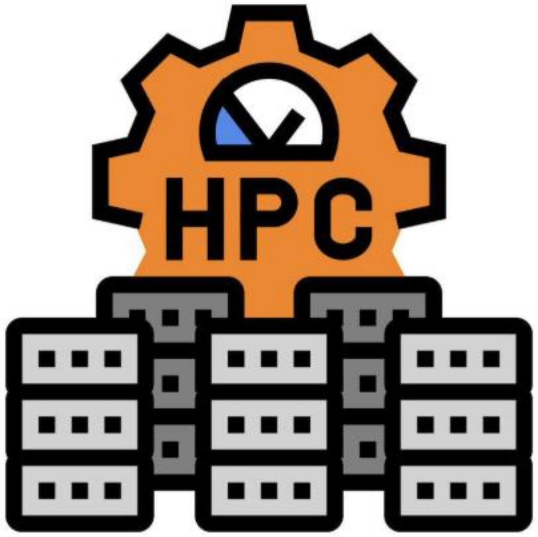
NIH NATIONAL CANCER INSTITUTE
Center for Cancer Research

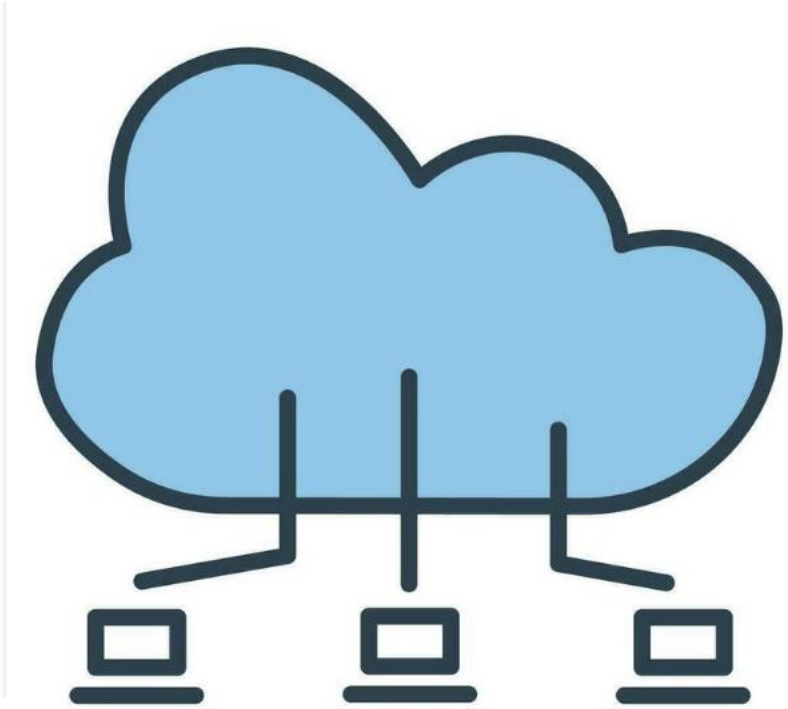
What we're going to cover.

- Local, HPC, and Cloud Resources
- The NCI Center for Cancer Research (CCR) Bioinformatics Training and Education Program (BTEP) and NIH Bioinformatics Calendar
- NIH High-Performance Compute Cluster (Biowulf)
- NCI – only Frederick-based Compute Cluster (FRCE)
- NIH Library – ~~Bioinformatics Workstations, Classes, Software, Expert Assistance,~~ AI, Data Science
- Cloud – NIH STRIDES Program Cloud Lab, NCI CRDC (CGC, ISB-Gateway), HTAN, AnVIL (NHGRI), NIGMS Sandbox
- Production Workflows (NCI) – CCBR Github, NIDAP
- Software Licenses – Partek Flow, Qiagen Pathway Analysis, SnapGene, Qlucore, etc.
- Free NIH-wide License – Coursera Learning Platform
- ~~NCI CCR – Dataquest licenses~~
- NIH List Servs and Teams

Compute Resources: From Laptops to HPC and Cloud





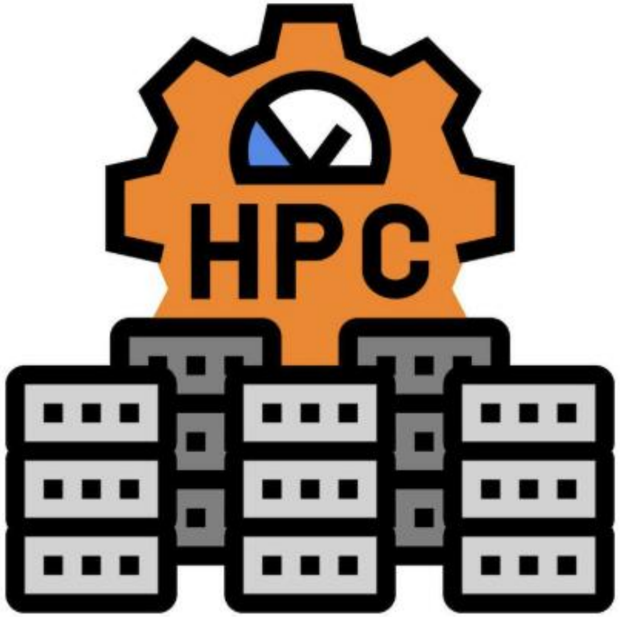


Using Local Machines for Bioinformatics

- Local machine is desktop/laptop
- Can run more than one OS, including Unix
- Gateway to HPC and/or Cloud
- Can run some commercial and open-source bioinformatics software
- Limited storage and compute power
- Lack of reproducibility in point-and-click workflows
- Machines with high CPU/GPU/RAM/SSD are expensive and require constant updating

The Power of High-Performance Computing (HPC)

- Computational power
- Hardware support
- Software maintenance and updates
- Parallel computing with multiple CPUs/GPUs
- Data management
- Pipelines and workflows installed
- Can perform computationally intensive tasks (genomic alignments, protein prediction, high resolution imaging data, AI/ML, multi-omics)



High Performance Computing (HPC)

- Not everyone has access to an HPC
- Requires facility and staff
- Cost and maintenance
- Understanding of file structure, computational architecture
- Learning curve to interact at command line (Unix,Linux)

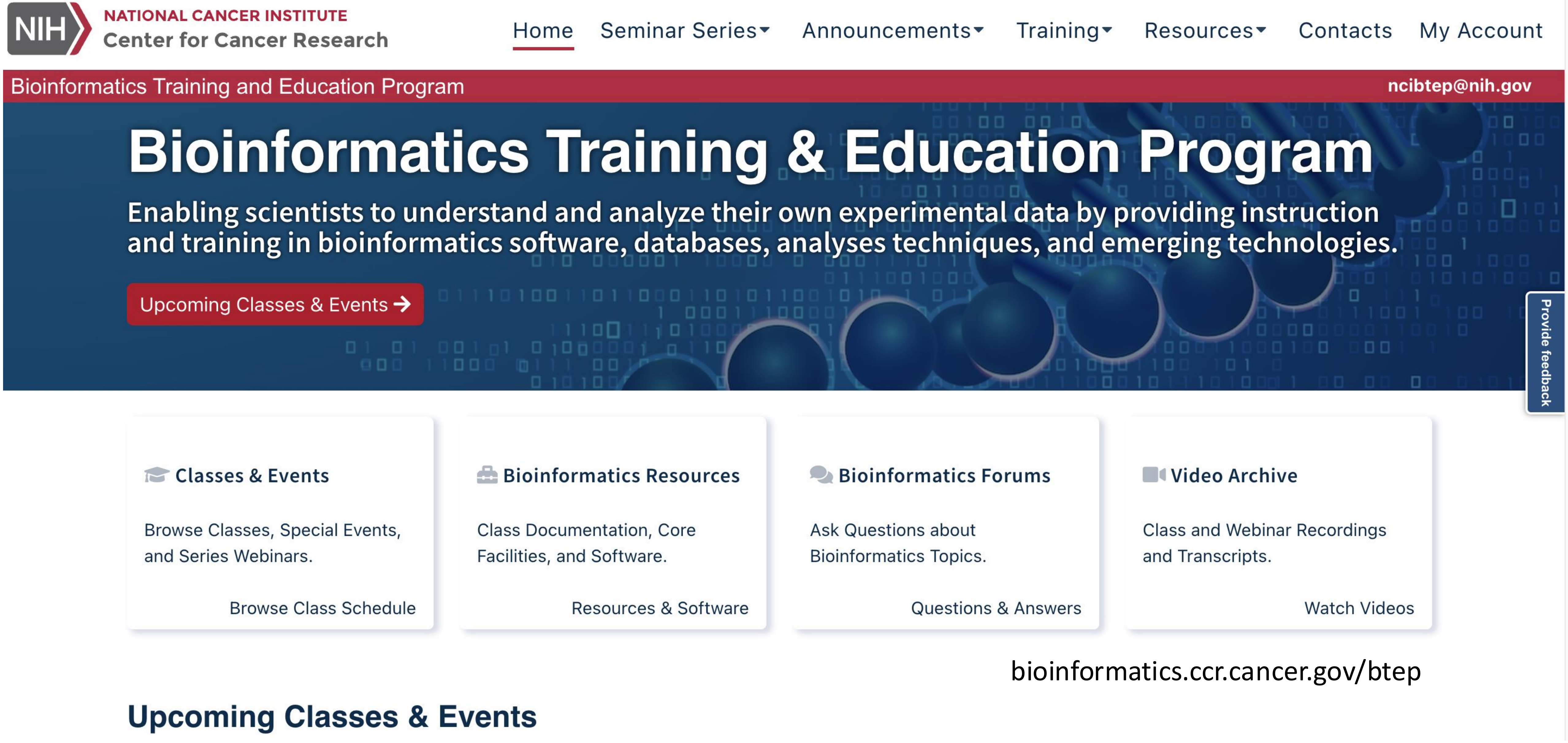
Cloud Resources

- Access to large, publicly available datasets
- Upload your own data
- Use only as much compute resources as needed
- Explore, analyze, and visualize data
- Bring tools to the data
- Collaboration via shared data and compute tools
- No need for onsite resources
- Data management and storage
- Cost

NCI CCR Bioinformatics Training and Education Program (BTEP)

Began in 2012 to provide training in bioinformatics software

In 2024, 80+ classes with over 2,000 attendees



Bioinformatics Training and Education Program ncibtep@nih.gov

Bioinformatics Training & Education Program

Enabling scientists to understand and analyze their own experimental data by providing instruction and training in bioinformatics software, databases, analyses techniques, and emerging technologies.

Classes & Events

Browse Classes, Special Events, and Series Webinars.

Browse Class Schedule

Bioinformatics Resources

Class Documentation, Core Facilities, and Software.

Resources & Software

Bioinformatics Forums

Ask Questions about Bioinformatics Topics.

Questions & Answers

Video Archive

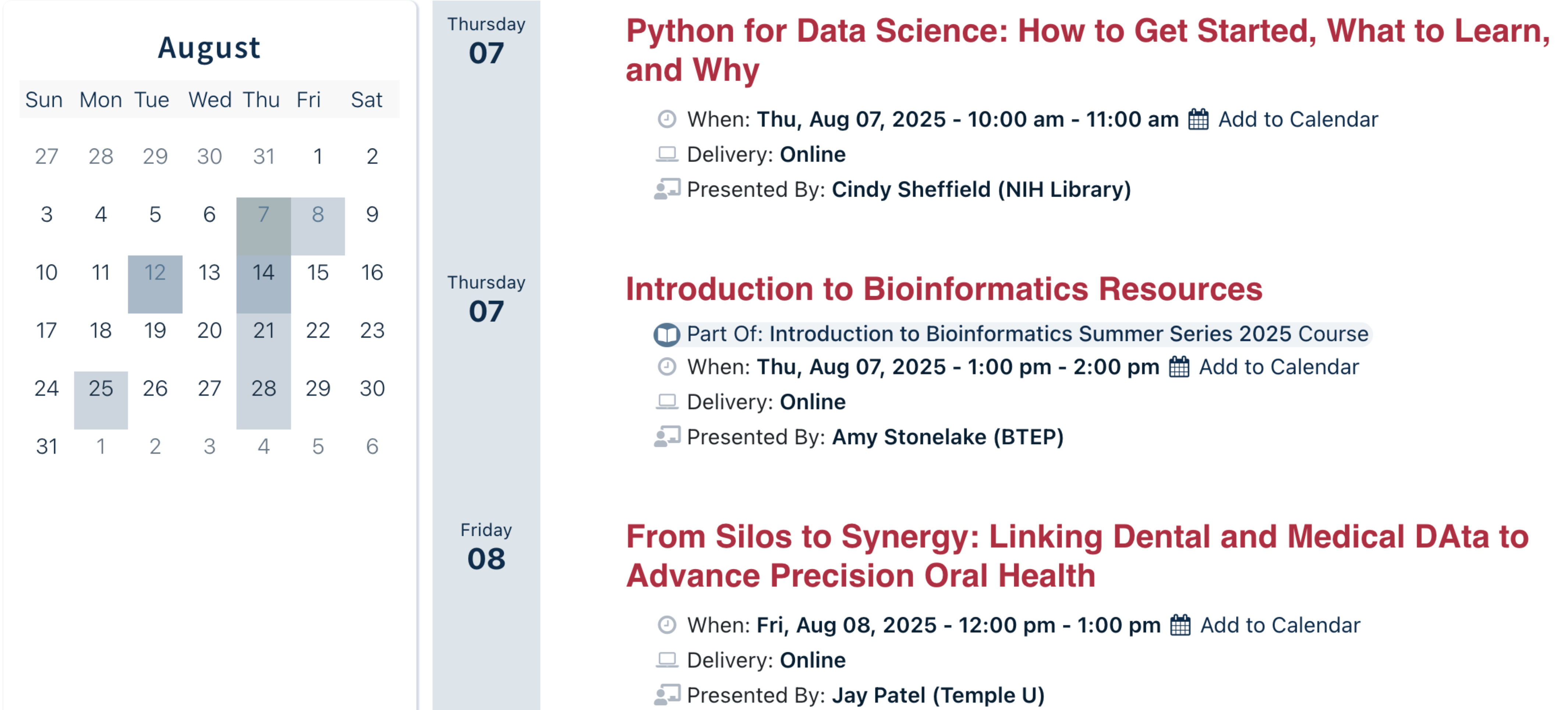
Class and Webinar Recordings and Transcripts.

Watch Videos

Upcoming Classes & Events

bioinformatics.ccr.cancer.gov/btep

NIH Bioinformatics Calendar



August

Sun	Mon	Tue	Wed	Thu	Fri	Sat
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

Thursday
07

Python for Data Science: How to Get Started, What to Learn, and Why

When: **Thu, Aug 07, 2025 - 10:00 am - 11:00 am** Add to Calendar
Delivery: **Online**
Presented By: **Cindy Sheffield (NIH Library)**

Thursday
07

Introduction to Bioinformatics Resources

Part Of: Introduction to Bioinformatics Summer Series 2025 Course
When: **Thu, Aug 07, 2025 - 1:00 pm - 2:00 pm** Add to Calendar
Delivery: **Online**
Presented By: **Amy Stonelake (BTEP)**

Friday
08

From Silos to Synergy: Linking Dental and Medical DAta to Advance Precision Oral Health

When: **Fri, Aug 08, 2025 - 12:00 pm - 1:00 pm** Add to Calendar
Delivery: **Online**
Presented By: **Jay Patel (Temple U)**

bioinformatics.ccr.cancer.gov/btep

NCI Center for Cancer Research Bioinformatics Training and Education Program (BTEP)

- Bioinformatics for Beginners (Unix, RNA-Seq, Pathway)
- Distinguished Speakers Seminar Series
- Single Cell Seminar Series
- Coding Clubs
- BTEP Bulletin
- Class Documentation and Resource Pages
- NIH Bioinformatics Calendar
- In-person/virtual help sessions
- Training/Events/Classes
- Video Archive
- Programming (Python, R, Unix)
- Microbiome
- Topics in Bioinformatics

Although we are a NCI/CCR resource…

- Most of our events are open to all at NIH
- Our website is open to the world
- We answer questions within and outside of NIH
- Video Archive is open and available
- Class documentation and bioinformatics resources info are open
- We advertise our events throughout NIH

Where should I do my work?

LOCAL

HPC

CLOUD

Working on local machine

Your NIH laptop or desktop

What can you do on your local machine?

- Run Data Analysis Software (CLC Genome Workbench, SnapGene, Qiagen Pathway Analysis)
- Connect to HPC (high-performance cluster) via SSH or GUI
- Access the cloud

HPC (High-Performance Compute) at NIH

Biowulf (NIH)

FRCE (Frederick)

Why you should get to know Biowulf

BIOWULF IS THE HIGH PERFORMANCE CLUSTER (HPC) AT NIH.

IT CAN HOLD A LOT MORE DATA THAN YOUR PERSONAL COMPUTER.

IT HAS MUCH MORE COMPUTE RESOURCES THAN YOUR PERSONAL COMPUTER.

IT CAN HELP YOU ANALYZE “BIG DATA”.

IT IS AVAILABLE TO ALL NIH RESEARCHERS

Working on Biowulf – Things to Know

- hpc.nih.gov
- Minimal, fixed monthly cost
- Thousands of scientific applications (software) available
- Reference Data (NCBI db, BLAST db, genomic alignment data)
- Login/head node
- Learn some beginner Unix (Command Line)
- Your home directory and your data directory
- Start an interactive node (sinteractive)
- Setting up swarm jobs (repetitive jobs)
- Running batch jobs (scripts/programs)
- Moving big data (Globus)
- https://bioinformatics.ccr.cancer.gov/docs/resources-for-bioinformatics/Biowulf/
- HPC on Demand GUI (RStudio, VSCode, JupyterNotebook, IGV, Matlab)

Frederick Research Computing Environment (FRCE)
NCI only

- Available within NCI at no cost, optimized for Frederick researchers
- 250+ scientific applications installed
- Next Gen Sequencing, structural biology, cryogenic electron microscopy, imaging and AI, text mining
- Log in ssh or GUI, file transfers with Globus, batch and interactive jobs
- https://ncifrederick.cancer.gov/staff/frce/

Cloud

- Cloud Resources (NCI, NHGRI)
- NIH STRIDES Initiative – NIH Cloud Lab

How is the Cloud different from HPC?

- Cost based on usage
- Flexible compute power
- Analyze publicly available data

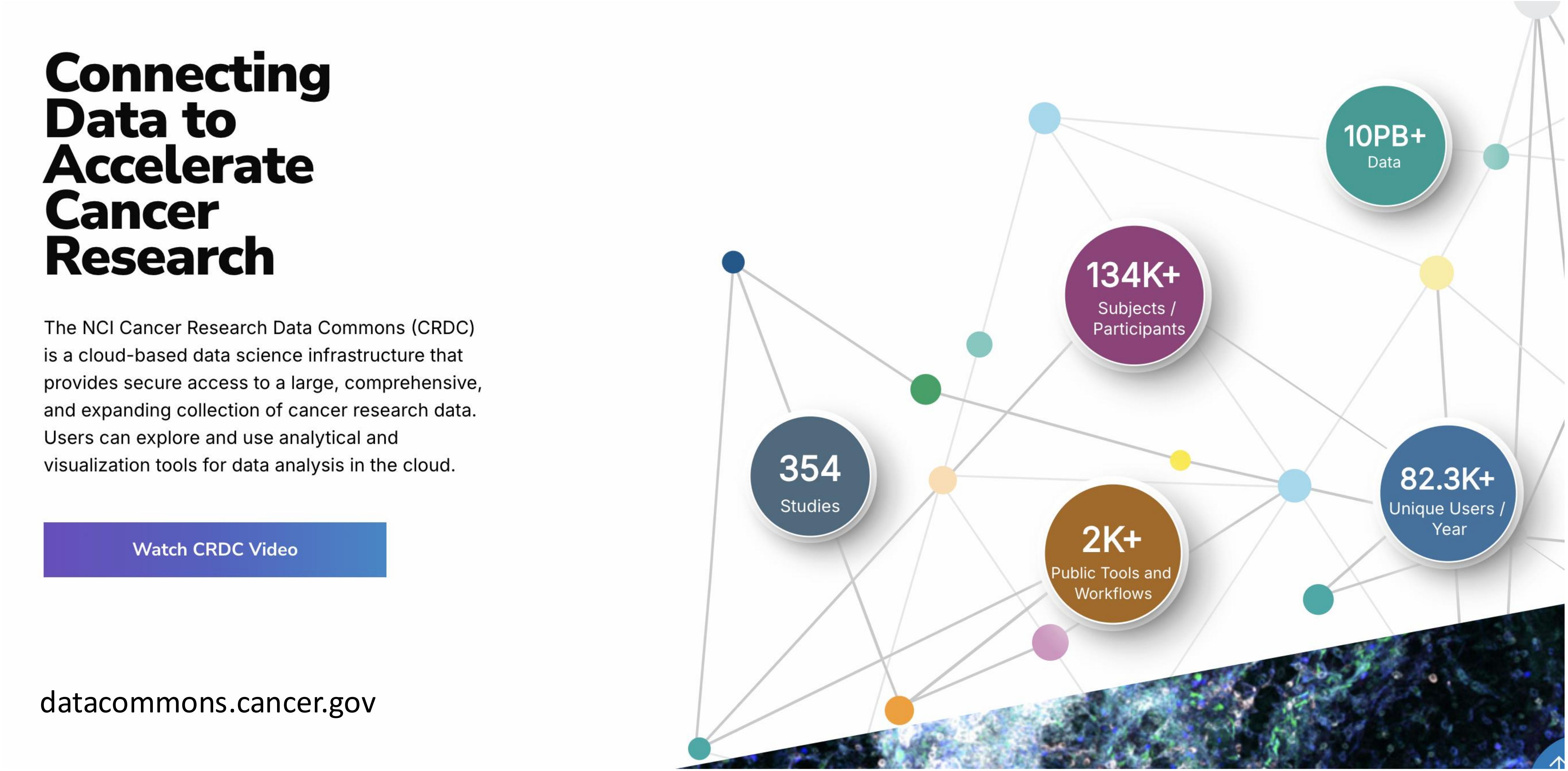
How is Cloud similar to HPC?

- Large compute resources
- Pre-installed software and tools
- Bioinformatics and data science analyses
- Analyze your own data

Publicly Available NCI Cloud Resources

- Cancer Research Data Commons (CRDC)
- Cancer Genomics Cloud/7Bridges/Velsera
- ISB Gateway in the Cloud/CGC
- Human Tumor Atlas Network (HTAN)

NCI Cancer Research Data Commons

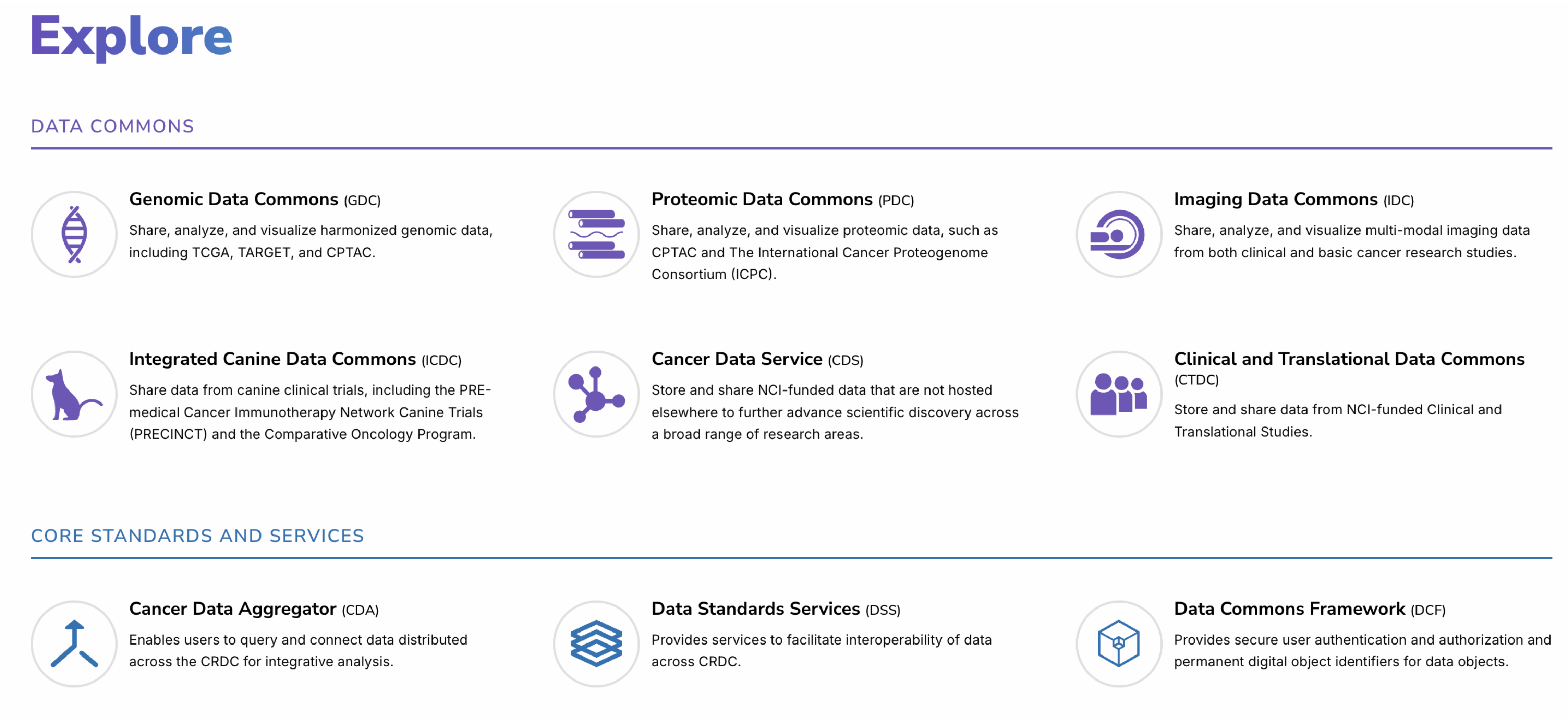


Connecting Data to Accelerate Cancer Research

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.

datacommons.cancer.gov

NCI Cancer Research Data Commons



Explore

DATA COMMONS

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

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

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.

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 and Translational Data Commons (CTDC)
Store and share data from NCI-funded Clinical and Translational Studies.

CORE STANDARDS AND SERVICES

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

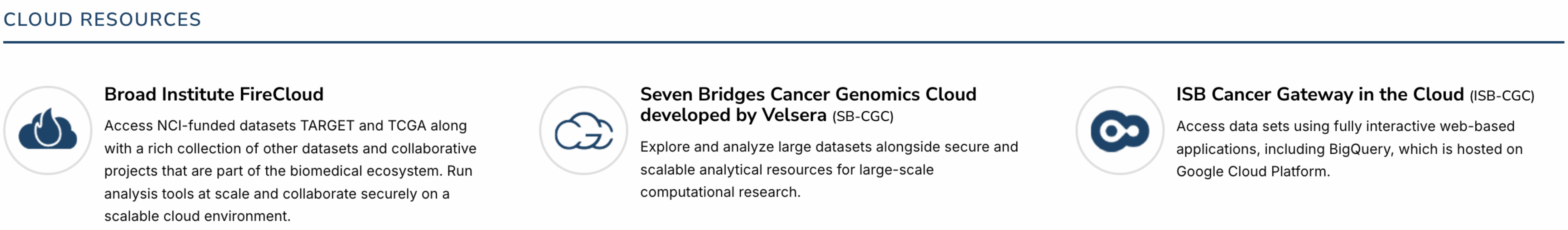
Data Standards Services (DSS)
Provides services to facilitate interoperability of data across CRDC.

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

datacommons.cancer.gov

NCI Cancer Research Data Commons

datacommons.cancer.gov



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.

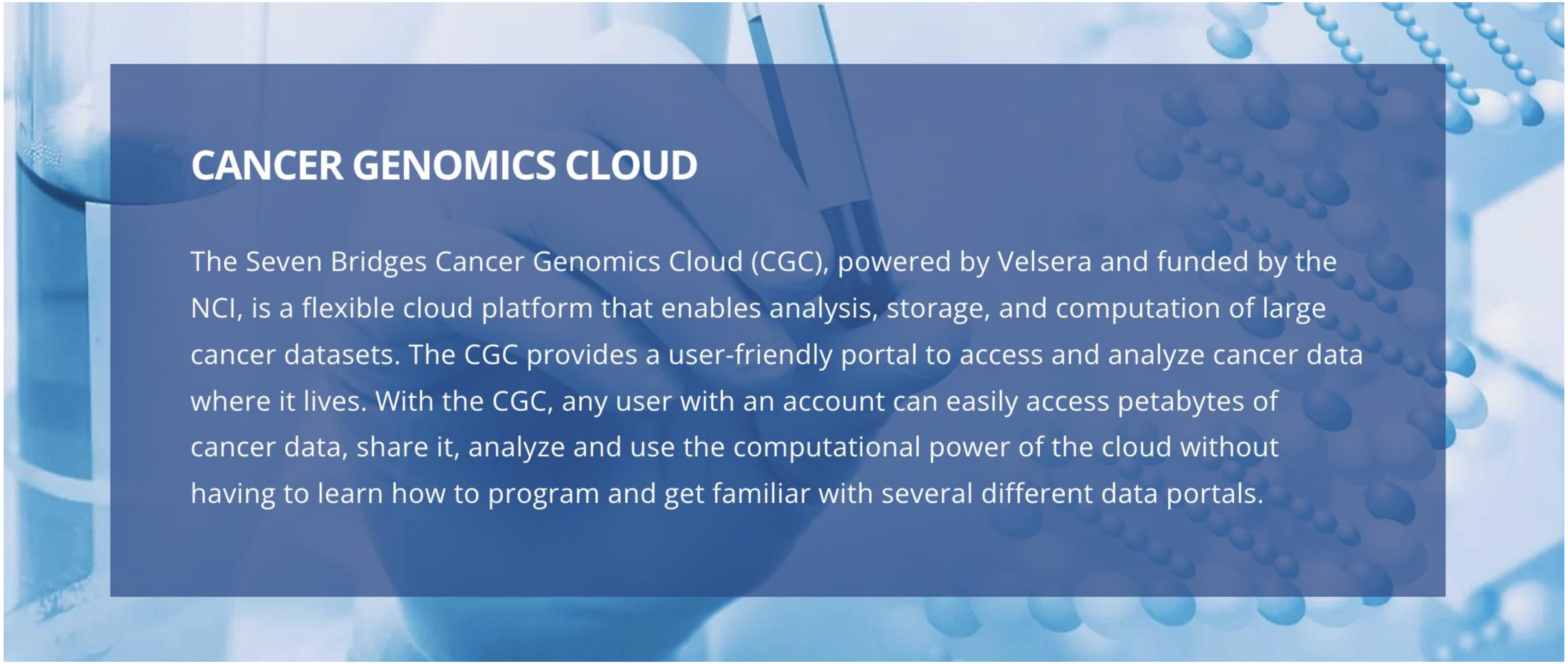
Seven Bridges Cancer Genomics Cloud developed by Velsera (SB-CGC)

Explore and analyze large datasets alongside secure and scalable analytical resources for large-scale computational research.

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/Velsera



CANCER GENOMICS CLOUD

The Seven Bridges Cancer Genomics Cloud (CGC), powered by Velsera and funded by the NCI, is a flexible cloud platform that enables analysis, storage, and computation of large cancer datasets. The CGC provides a user-friendly portal to access and analyze cancer data where it lives. With the CGC, any user with an account can easily access petabytes of cancer data, share it, analyze and use the computational power of the cloud without having to learn how to program and get familiar with several different data portals.

cancergenomicscloud.org

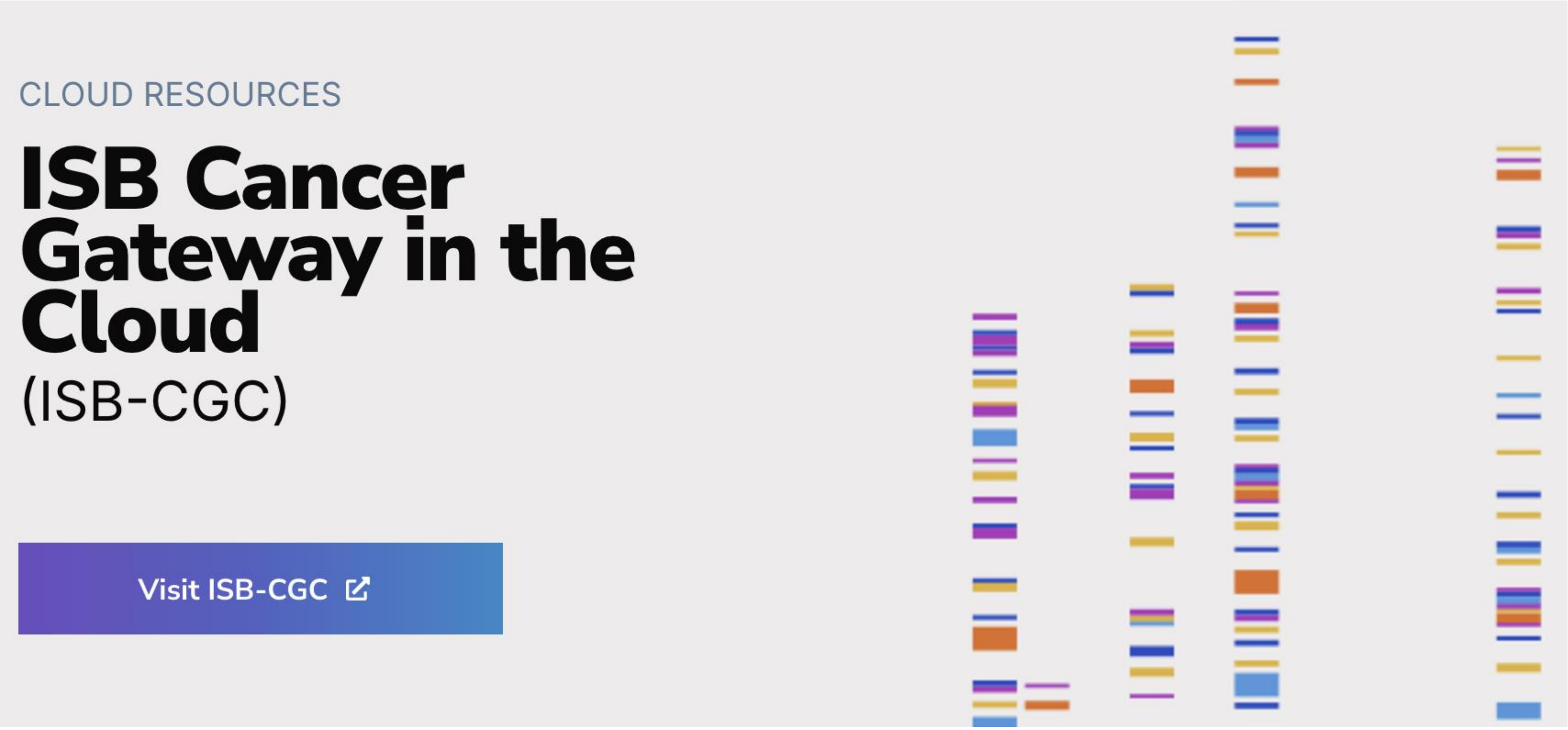
Cancer Genomics Cloud (CGC) SBR/Velsera

- Users can access:
    - Genomics data
    - Proteomics data
    - TCGA data
    - Data from multiple species
    - Their own data
    - Tool Library (850+)
    - Common Workflow Language (CWL)
    - Bring your own pipeline or tools
    - Jupyter Lab and Rstudio available

cancergenomicscloud.org

ISB Cancer Gateway in the Cloud (ISB-CGC)

- GoogleBigQuery – query across multiple data tables
- TCGA and TARGET data



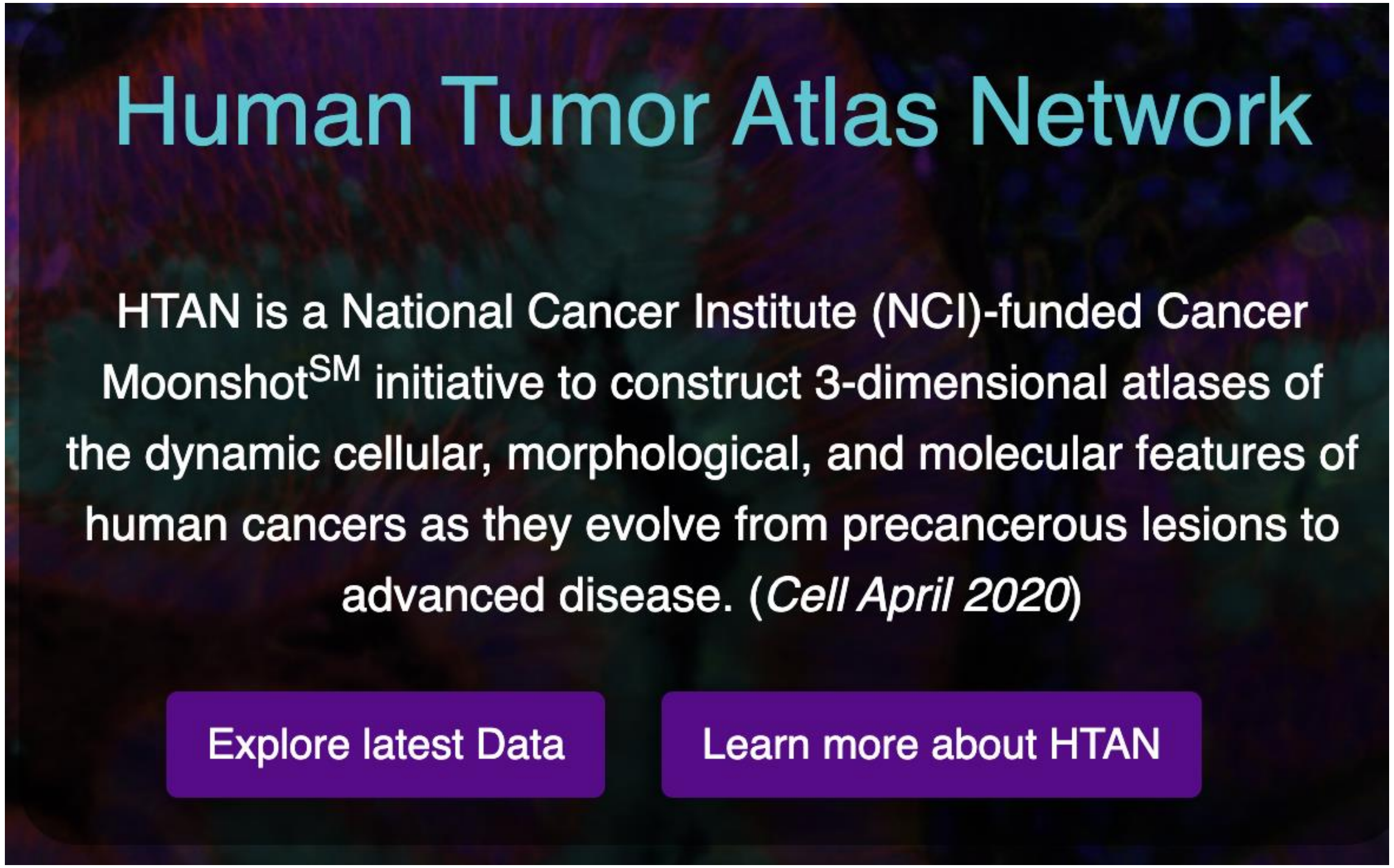
CLOUD RESOURCES

ISB Cancer Gateway in the Cloud

(ISB-CGC)

https://datacommons.cancer.gov/analytical-resource/isb-cancer-gateway-cloud

Human Tumor Atlas Network



Human Tumor Atlas Network

HTAN is a National Cancer Institute (NCI)-funded Cancer Moonshot^{SM} initiative to construct 3-dimensional atlases of the dynamic cellular, morphological, and molecular features of human cancers as they evolve from precancerous lesions to advanced disease. (*Cell April 2020*)

humantumoratlas.org

NHGRI Analysis Visualization and Informatics Lab-space (AnVIL)

anvilproject.org

Free trials available

Tools

Dockstore – create and share docker-based workflows

NCPI – interoperate with other NIH data commons

Bioconductor

Galaxy

Jupyter (python, R)

Data sources

Telomere-to-telomere genome (T2T)

1000G thousand genomes

Genotype-tissue expression project (GTEx)

NIH STRIDES Iniative

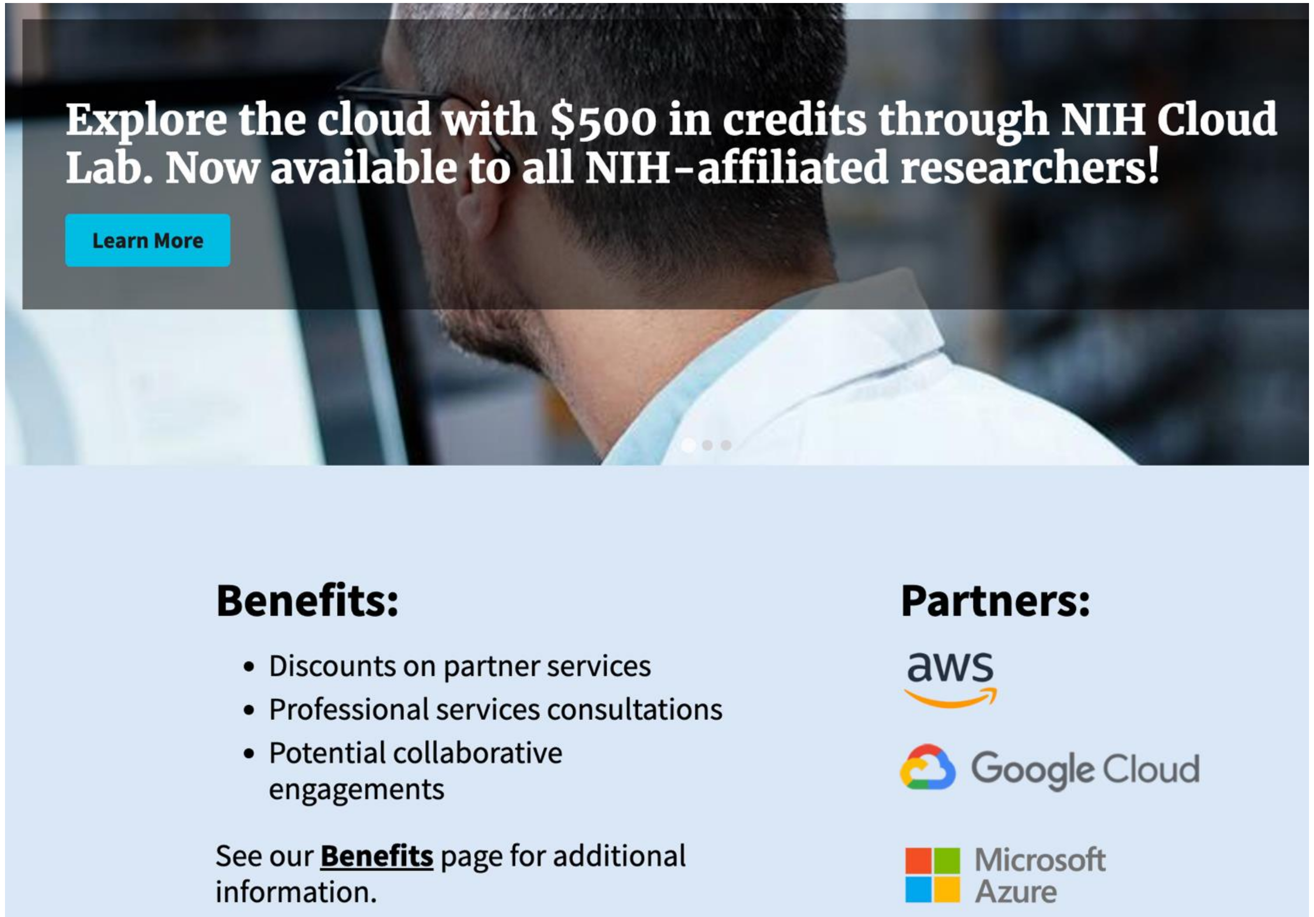
Accelerating Biomedical Research

NIH Office of Data Science Strategy (ODSS)

cloud.nih.gov

NIH Cloud Lab

NIH STRIDES Cloud Lab



Explore the cloud with $500 in credits through NIH Cloud Lab. Now available to all NIH-affiliated researchers!

Benefits:

- Discounts on partner services
- Professional services consultations
- Potential collaborative engagements

See our **Benefits** page for additional information.

Partners:

cloud.nih.gov

Workflows (pipelines)

CCR Collaborative Bioinformatics Resource

Bioinformatics assistance to further CCR researchers' goals.

NIDAP Training

Online training for interactive CCBR workflows for bioinformatics analyses on NIDAP. Currently released workflows include: Bulk RNA-seq, Single-cell RNA-seq, and Digital Spatial Profiling (DSP).

Project Support

Learn how CCBR can assist with CCR Researchers with their projects.

Pipelines & Workflows

Workflows & Pipeline development. View whole exome & genome, single cell RNA-Seq, and ChIP-Seq pipline examples.

bioinformatics.ccr.cancer.gov/ccbr

CCBR Pipelines and NIDAP (NIH Integrated Data Analysis Platform)

NGS pipelines

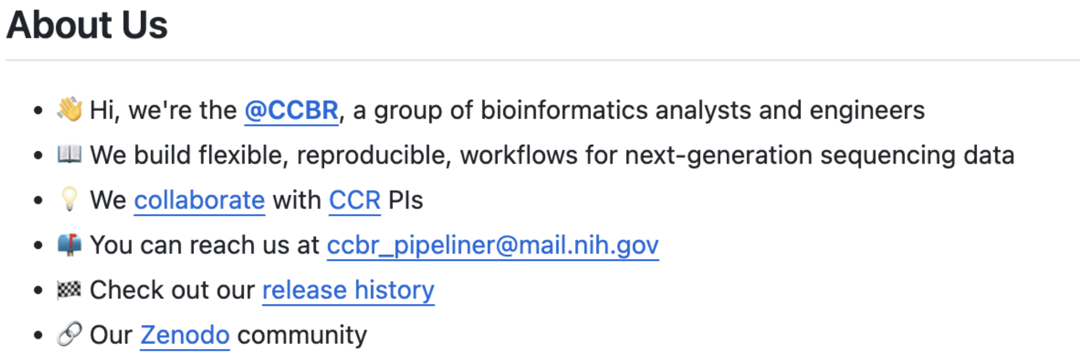
Available to all at NIH (GitHub)

CCBR NIDAP (NIH)

Free to use

Next Gen Seq Workflows/Pipelines
CCR Collaborative Bioinformatics Resource (CCBR)

github.com/ccbr



About Us

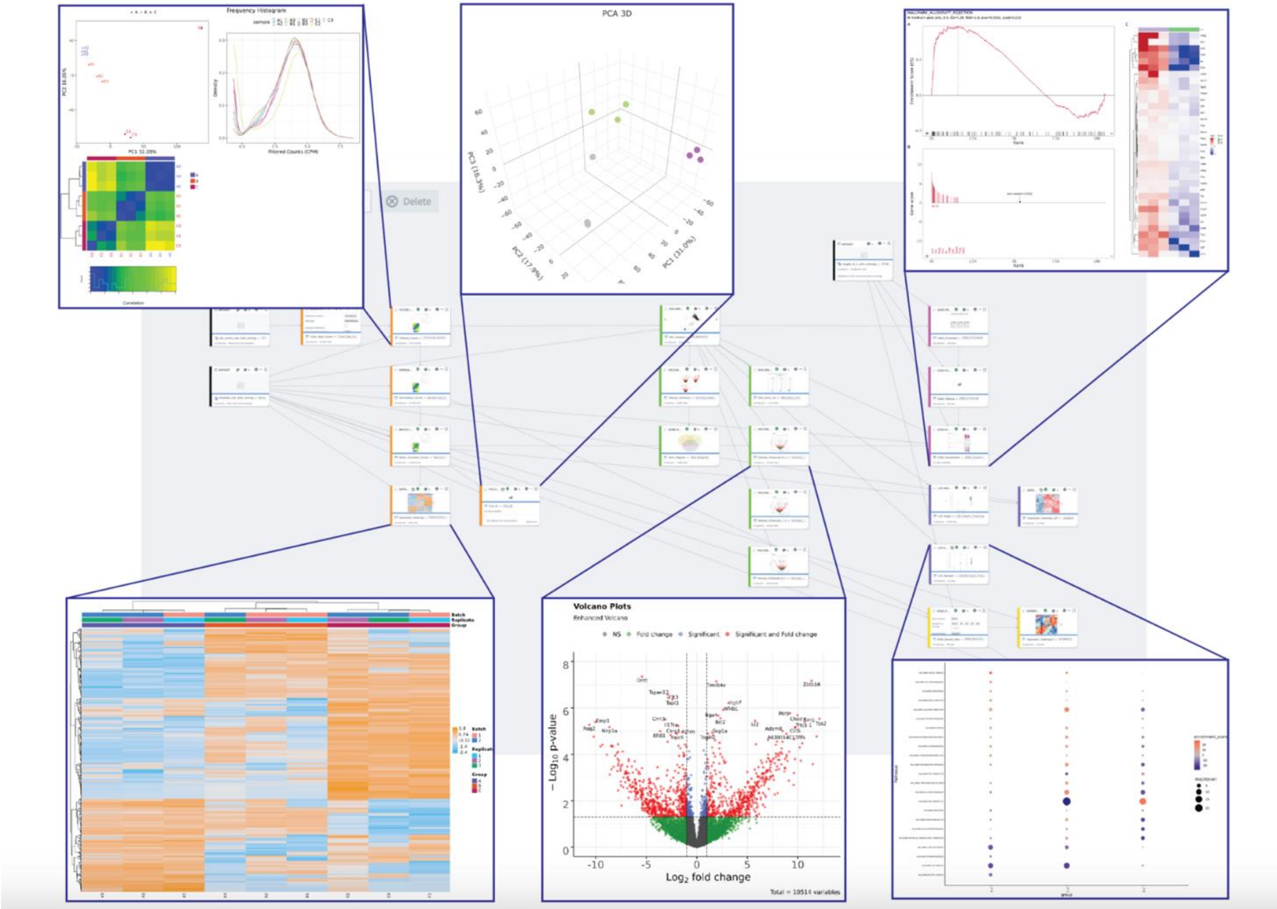
- 👋 Hi, we're the **@CCBR**, a group of bioinformatics analysts and engineers
- 📖 We build flexible, reproducible, workflows for next-generation sequencing data
- 💡 We collaborate with CCR PIs
- 📫 You can reach us at ccbr_pipeliner@mail.nih.gov
- 🏁 Check out our release history
- 🔗 Our Zenodo community

Available NGS Pipelines/Workflows

Here is a list of our prominent pipelines and their release schedule on BIOWULF:

Data Type	Pipeline Name	CLI* availability date	GUI* availability date
RNASeq[1]	RENEE snakemake	July 3rd 2023	July 14th 2023
WESSeq[2]	XAVIER snakemake	July 21th 2023	Sep 1st 2023
ATACSeq[3]	ASPEN snakemake	November 30th 2023	TBD
ChIPSeq[4]	CHAMPAGNE nextflow	October 15th 2023	TBD
CRISPRSeq[5]	CRISPIN nextflow	September 31st 2023	TBD
CUT&RunSeq[6]	CARLISLE snakemake	October 31st 2023	TBD
EV-Seq[10]	ESCAPE snakemake	March 26th, 2024	TBD
circRNASeq[7]	CHARLIE snakemake	*Jul 31st 2024*	TBD
scRNASeq[8]	SINCLAIR nextflow	*Sep 30th 2024*	TBD
WGSSeq[9]	LOGAN nextflow	*Sep 30th 2024*	TBD
spatialSeq[11]	SPENCER nextflow	TBD	TBD

Bulk RNA-Seq Analysis on NIDAP
(NIH Integrated Data Analyis Portal)



https://bioinformatics.ccr.cancer.gov/ccbr/education-training/nidap-training/

Commercial Software Available (NCI)

- Partek Flow – RNA-Seq, CITE-Seq, ATAC-Seq
- Qiagen Ingenuity Pathway Analysis (IPA) – from differential expression analysis (DEA) to pathways, biomarkers, drug targets
- CLC Genomics Workbench
- Qlucore Omics Explorer – RNA-Seq, visualizations, statistical analysis
- Request access at service.cancer.gov (NCI)
- SnapGene – molecular biology (alignment, PCR primers, cloning)

Check with your IC for software availability

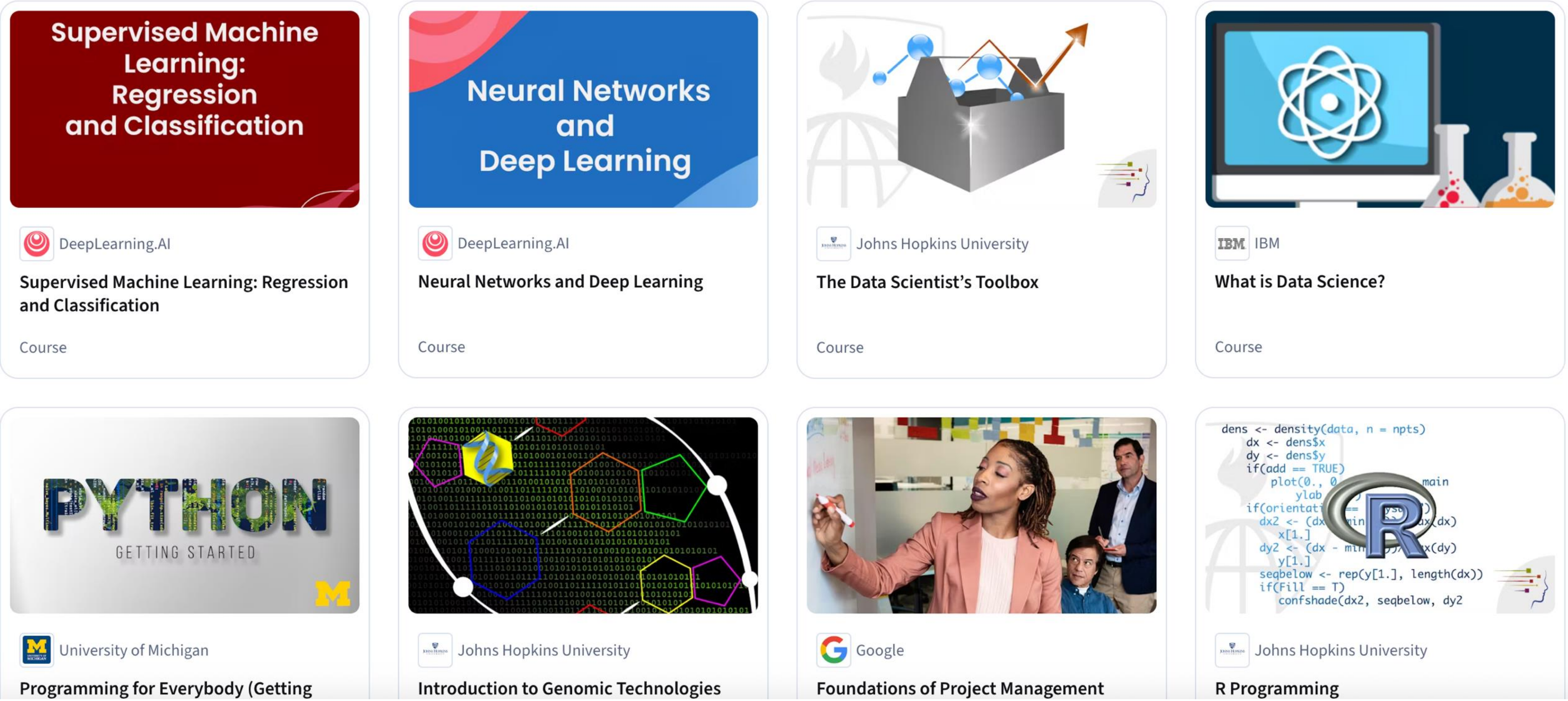
Free licenses for all at NIH

Coursera - learning platform

Training: Coursera licenses are available to all NIH, provided by NIH ODSS

- Free for everyone at NIH
- Video lectures
- Work at your own pace
- Large, worldwide, online classes (MOOCS)
- Courses, specializations, and guided projects
- Earn certificates for your resume/CV
- So many courses (thousands) available: Programming (R, Unix, Python), Genomics, Bioinformatics, Data Science, Language learning, AI
- https://bioinformatics.ccr.cancer.gov/btep/self-learning/

Coursera Most Popular at NIH



DeepLearning.AI

Supervised Machine Learning: Regression and Classification

Course

DeepLearning.AI

Neural Networks and Deep Learning

Course

Johns Hopkins University

The Data Scientist's Toolbox

Course

IBM

What is Data Science?

Course

University of Michigan

Programming for Everybody (Getting

Johns Hopkins University

Introduction to Genomic Technologies

Google

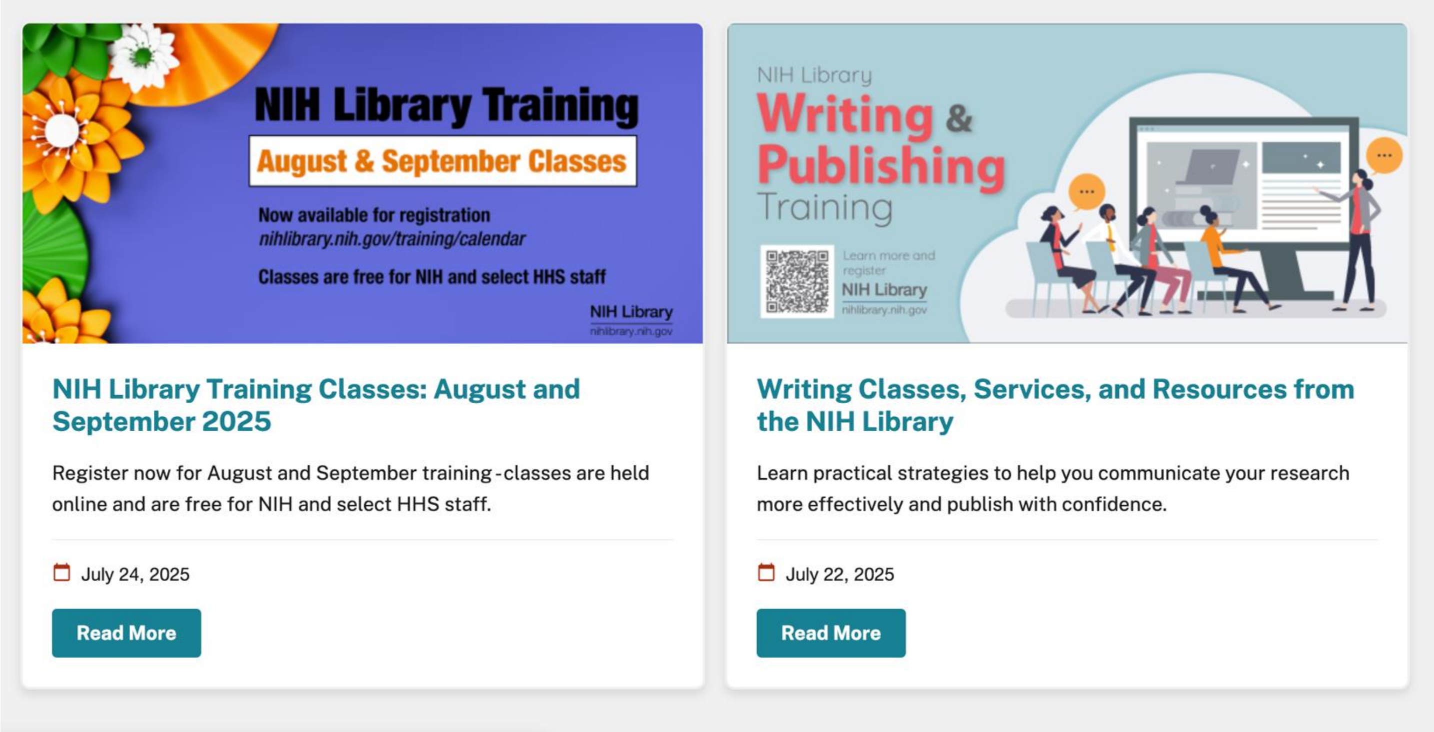
Foundations of Project Management

Johns Hopkins University

R Programming

Dataquest Licenses no Longer Available

More Resources - NIH Library (Bldg 10)



NIH Library Training Classes: August and September 2025

Register now for August and September training - classes are held online and are free for NIH and select HHS staff.

July 24, 2025

Writing Classes, Services, and Resources from the NIH Library

Learn practical strategies to help you communicate your research more effectively and publish with confidence.

July 22, 2025

Aug 8	NIH Library Resources for NIH Staff
Aug 12	How to Write a Research Paper: Part 2
Aug 13	How to Create an Effective ORCiD Profile
Aug 21	Resources for Finding and Sharing Research Data
Aug 26	Medical Imaging Workflows in MATLAB
Aug 28	Advanced PubMed

nihlibrary.nih.gov

Listservs, Teams, and Interest Groups

AI Interest Group

Ryan O’Neill and Samar Samarjeet

ARTIFICIAL-INTELLIGENCE@LIST.NIH.GOV

AI Club –weekly sessions in NIH Library

NIH AI Symposium (May, 2025) – one day, in-person, Bethesda campus

Sends ”Weekly Roundup” newsletter

NIH GenAI Community of Practice

- Nick Weber and Diane Babski (co-chairs), Rochelle Rayos
- nih-genaicommunity@groups.nih.gov
- 1200+ members
- Holds monthly meetings and monthly Office Hours
- Learn about Chirp (NIH GenAI)

We send events to (list.nih.gov)

- BIOINFORMATICS-SIG-L
- DATA-SCIENCE
- STAFF SCIENTISTS
- FELLOW
- SINGLECELLGENOMICS (AND SPATIAL TRANSCRIPTOMICS)L
- SPATIALBIOLOGY@LIST.NIH.GOV

Bioinformatics Training & Education Program

Amy Stonelake, Ph.D.
Program Manager

Joe Wu, Ph.D.
BTEP Trainer

Alex Emmons, Ph.D.
BTEP Trainer

ncibtep@nih.gov
https://bioinformatics.ccr.cancer.gov/btep


End of Presentation
ANY QUESTIONS?