NIH STRIDES Initiative Update

NIH DPCPSI Council of Councils Meeting

Andrea T. Norris

Chief Information Officer, National Institutes of Health
and
Director, Center for Information Technology

National Institutes of Health

May 20, 2022

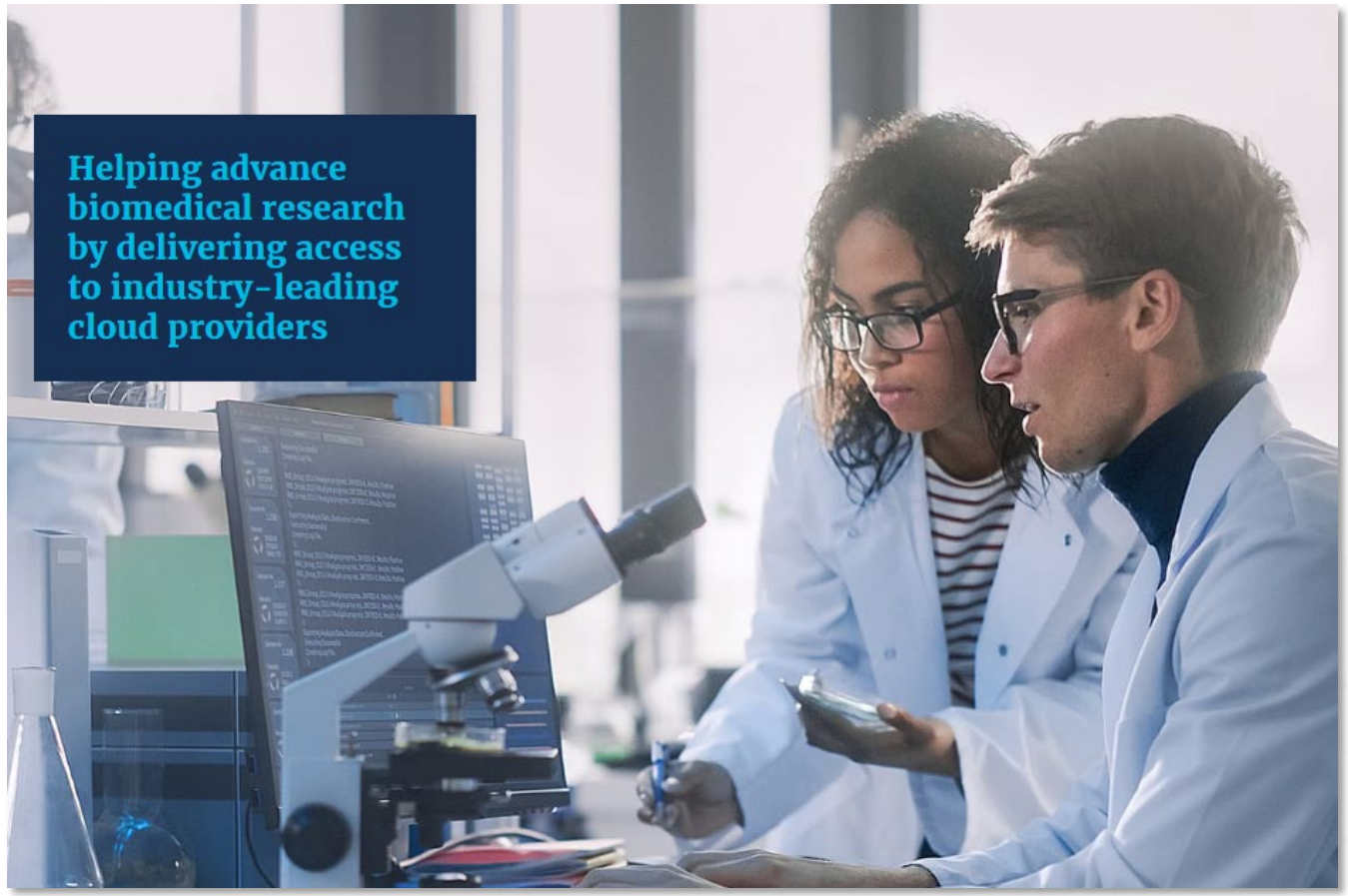
NIH STRIDES Initiative

The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability

- State-of-the-art data storage and computational capabilities
- Training and education for researchers
- Innovative technologies such as artificial intelligence and machine learning
- Professional engineering and technical support

Partnerships with aws Google Cloud Microsoft Azure

The Impact of STRIDES



Impact as of March 31, 2022

170+	PETABYTES OF DATA
215M+	COMPUTE HOURS
740+	RESEARCH PROGRAMS
$31M+	COST SAVINGS
4200+	PEOPLE TRAINED

NIH Center for Information Technology

Major Research Institutions Participating

Major Research Programs Supported

All of Us RESEARCH PROGRAM

Framingham Heart Study — Three Generations of Dedication 1948 1971 2002

National Ophthalmic Disease Genotyping Network eyeGENE® A program of the National Institutes of Health

GTEx Portal

PRIMED consortium

HuBMAP Human BioMolecular Atlas Program

AMP PD

HEAL NIH · Helping to End Addiction Long-term

NHLBI TOPMed

Metabolomics Workbench

Common Fund Data Ecosystem

SPARC

AnVIL

MoTrPAC The Molecular Transducers of Physical Activity Consortium

UniProt

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Glycoscience

HMP

DK net

eRA

IDG ILLUMINATING the DRUGGABLE GENOME

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

NIMH Data Archive NDA | ABCD | CCF | NIAAA$_{DA}$ | OAI

NIA IMPACT COLLABORATORY TRANSFORMING DEMENTIA CARE

NATIONAL CANCER INSTITUTE GENOMIC DATA COMMONS Access the Data #NCIGDC

NIH Center for Information Technology

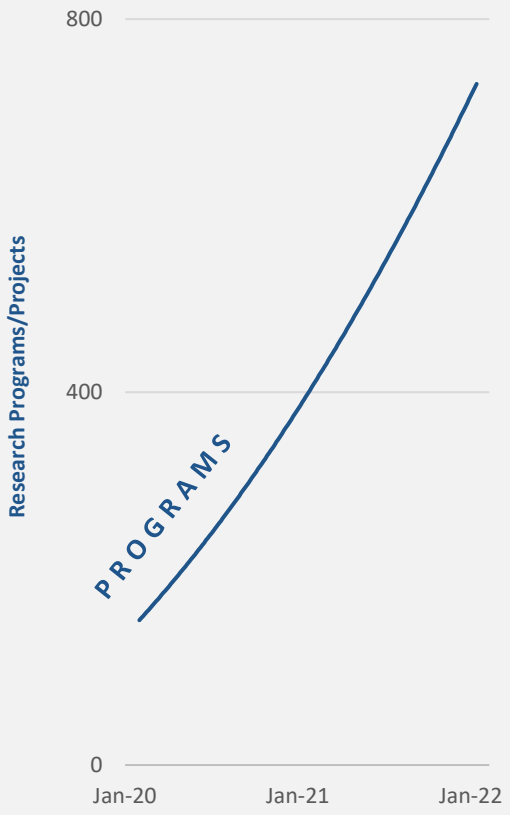
Galaxy PROJECT

SenNet

SRA Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®.

Unprecedented Growth

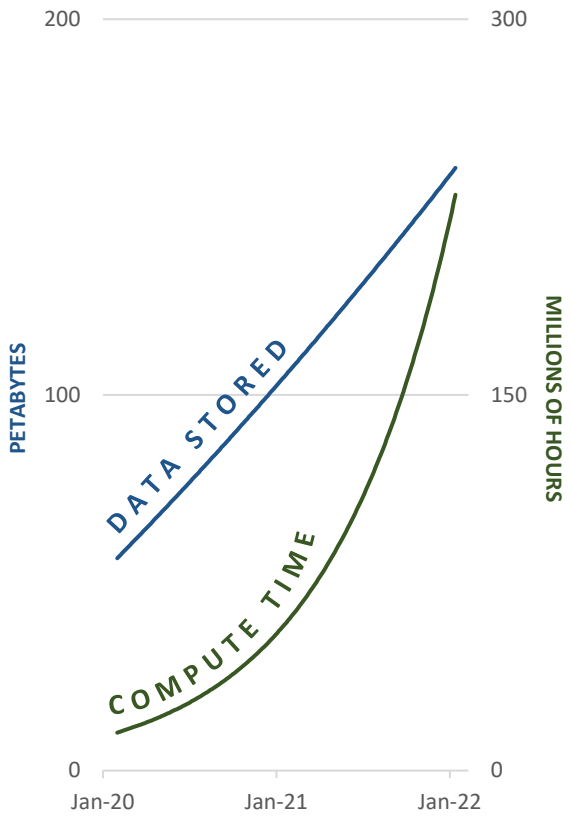
Participation



+382%

PERCENT INCREASE

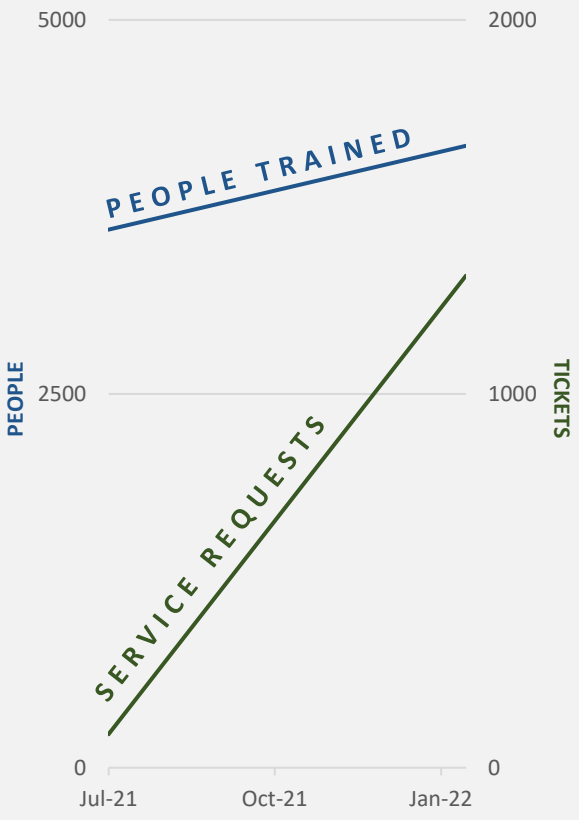
Utilization



+864%

AVERAGE PERCENT INCREASE

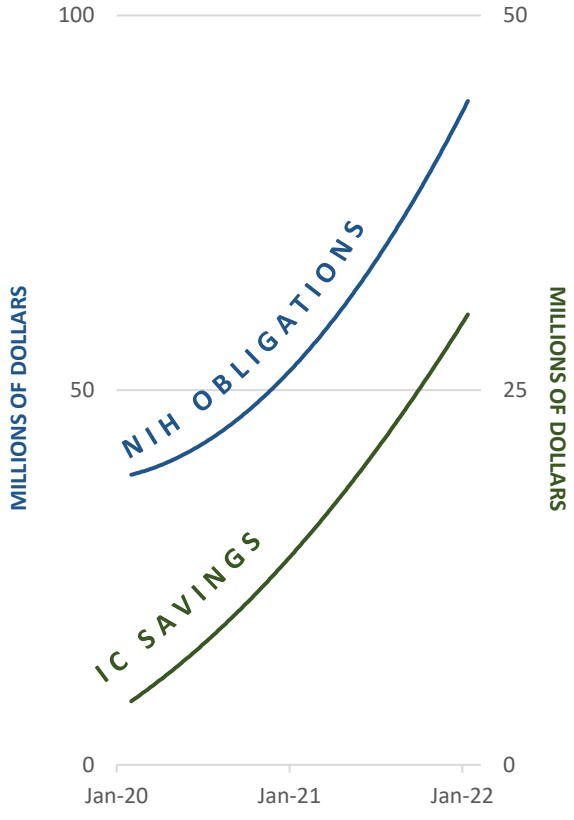
Support



+561%

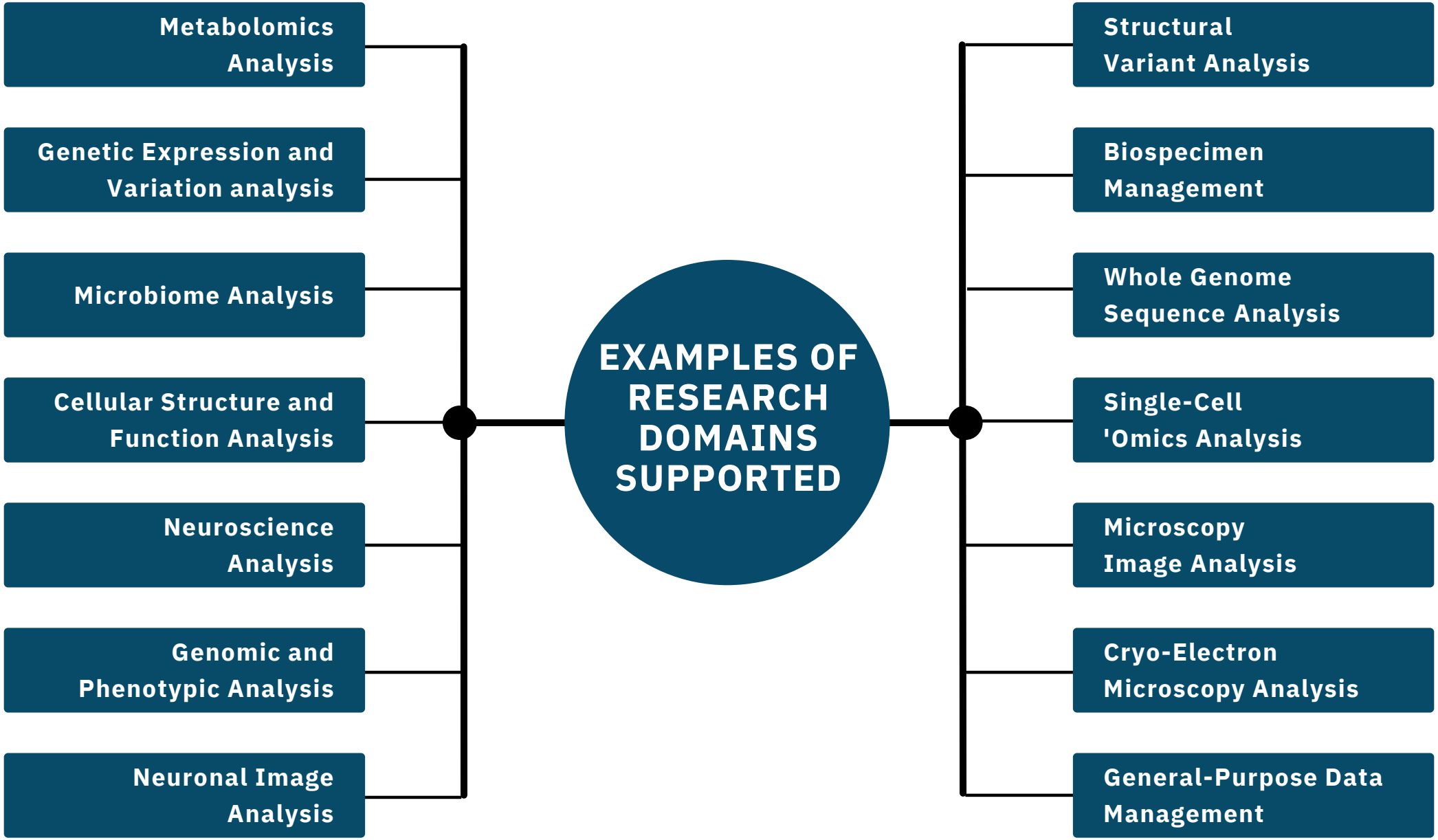
AVERAGE PERCENT INCREASE

Financial Impact



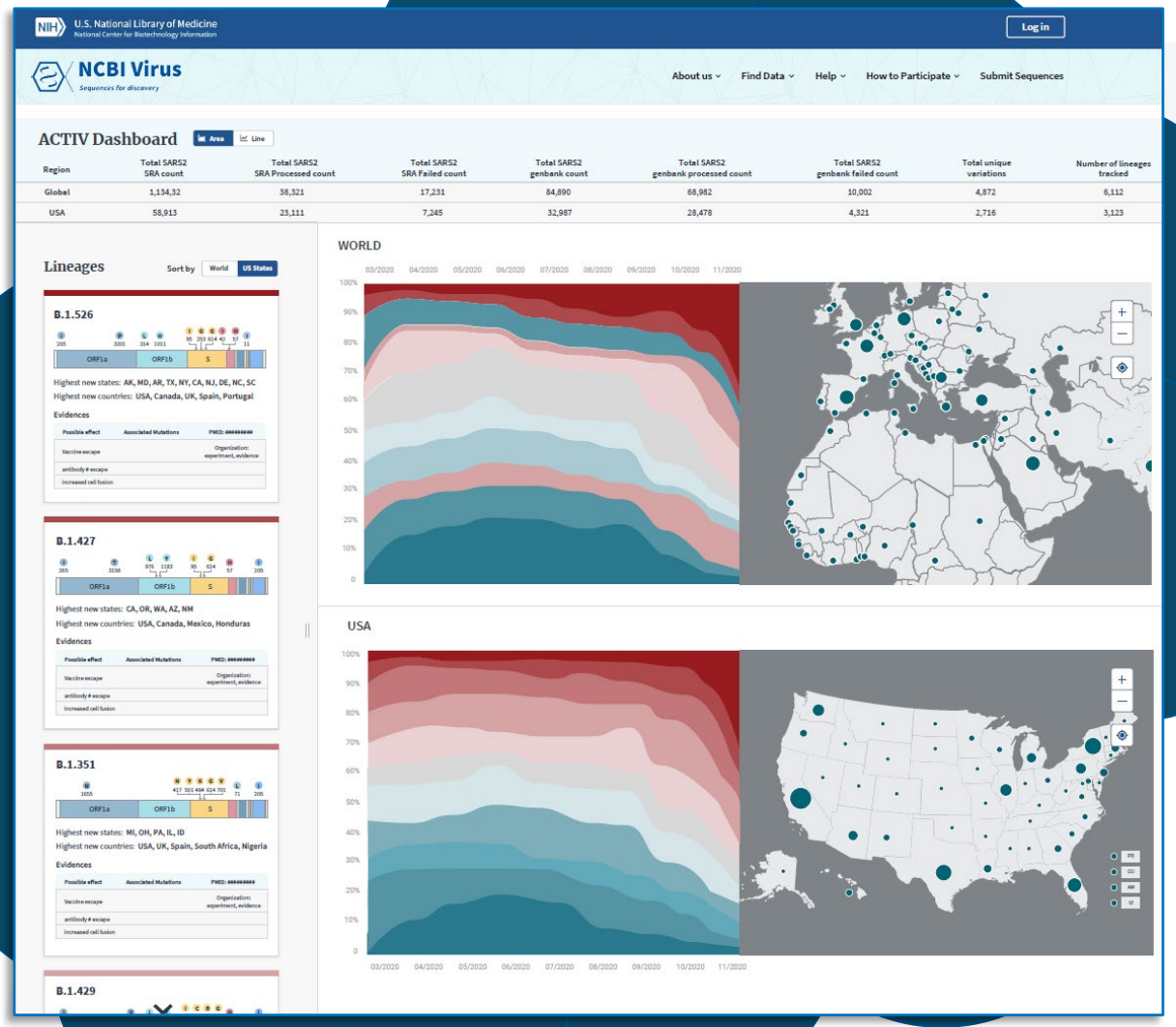
+325%

AVERAGE PERCENT INCREASE


EXAMPLES OF RESEARCH DOMAINS SUPPORTED
Metabolomics Analysis
Genetic Expression and Variation analysis
Microbiome Analysis
Cellular Structure and Function Analysis
Neuroscience Analysis
Genomic and Phenotypic Analysis
Neuronal Image Analysis
Structural Variant Analysis
Biospecimen Management
Whole Genome Sequence Analysis
Single-Cell 'Omics Analysis
Microscopy Image Analysis
Cryo-Electron Microscopy Analysis
General-Purpose Data Management

SARS-CoV-2 analysis data is accessible & **freely available**

- Standardized “aligned read” files increase utility and impact of data
- All files and analysis data are hosted by AWS ODP and GCP Public Datasets
- Interactive NCBI COVID web resource that integrates sequence analysis data with metadata
- Encourages use by other data platforms

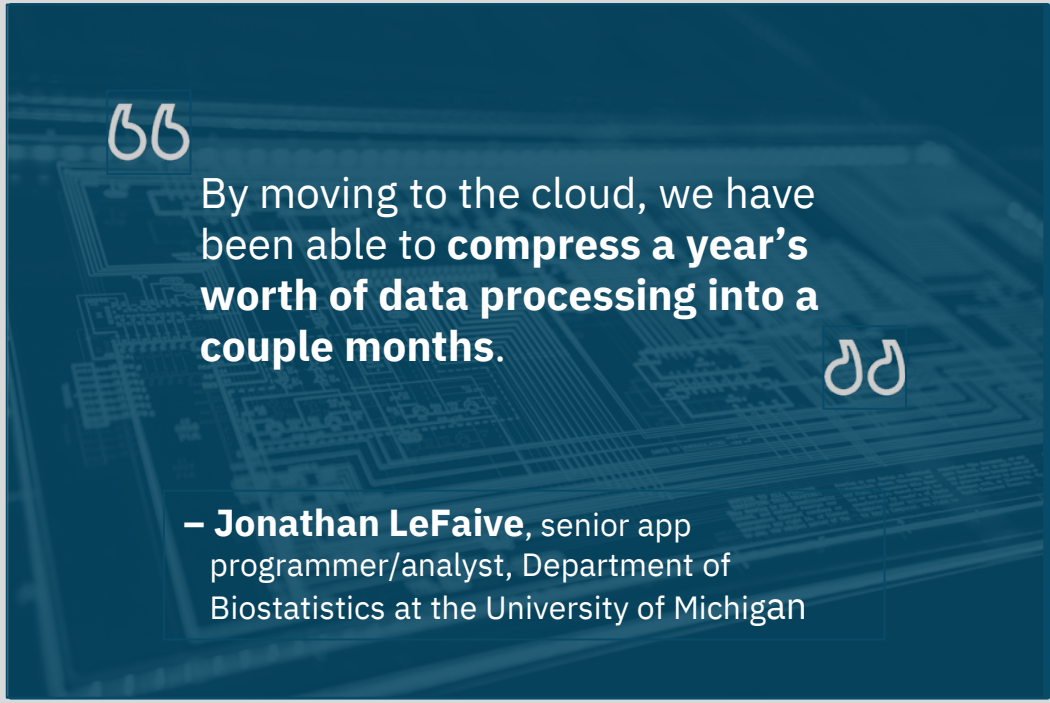


NIH National Heart, Lung, and Blood Institute | BioData CATALYST

The **TOPMed Imputation Server**, which leverages TOPMed's ethnically diverse data, was immediately popular among the research community since it launched in May 2020. As of today, **the server has imputed over 21.4 million genomes**.

The **STRIDES Initiative made this possible**, as it provided **access to favorable pricing and excellent engineering support** from the STRIDES Initiative partners.

The University of Michigan team manages the TOPMed Imputation Server.



SUCCESS STORY

CRDC Radiogenomics: Machine Learning (ML) Research in the Cloud

Goal: Use deep learning and radiomics to predict mutation status of gliomas from pre-operative MRI scans.

> The days when a researcher could download data to the computer under their desk are rapidly fading. The NCI Imaging Data Commons, with its connections to the other data types (genomics, proteomics, clinical) in the Cancer Research Data Commons, provides an **efficient means to solve important multimodal AI problems using cloud-scale resources** that will advance biomedical science and the care of patients.
>
> –**Bradley Erickson**, MD, PhD, Professor of Radiology and Medical Director of AI at Mayo Clinic

IDC
- Imaging Data Commons (IDC)
- Cohort exploration
- Imaging data preparation and QA

GDC
- Genomics Data Commons (GDC)
- Obtain mutation status
- Obtain demographics

GCP
- Google Cloud Platform (GCP)
- Match imaging & genomic data
- ML model development & evaluation

Slide courtesy of Conte and Erickson (Mayo Clinic)

STRIDES Training

Incredible **demand for cloud training** (nearly all courses have waitlists)

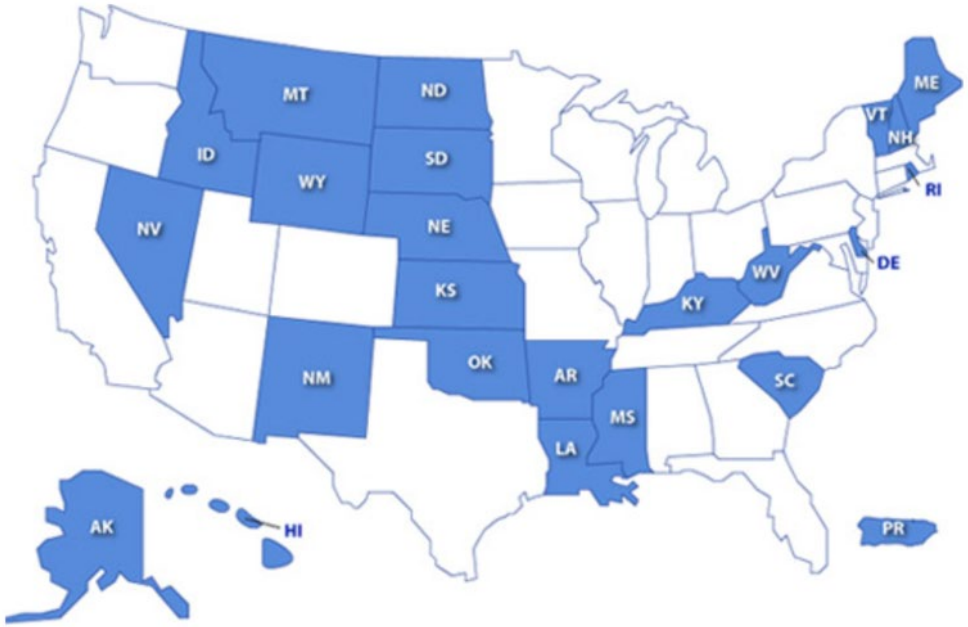
- Course offerings range from fundamentals, to research support/technical topics

Cloud providers have developed training courses with **content and examples specific to biomedical research**, meant to address researcher needs and challenges

Codeathons are regularly offered to provide a hands-on way for researchers, data scientists, and others to interact with the cloud platforms to solve specific problems

Support for Major NIH Diversity & Capacity Building Programs

- Collaborative R&D projects with cloud providers to support minority serving institutions (MSI) and Institutional Development Award (IDeA) states
  - Proteomics pipeline development with University of Arkansas for Medical Sciences
  - RNA-seq workflow & training with University of Maine system
- Targeted engagement and training efforts at MSIs, including Historically Black Colleges and Universities (HBCU) and Tribal Colleges and Universities (TCU)
  - NIH Virtual Workshop on Broadening Cloud Computing Usage in Biomedical Research
- Special research credits from cloud providers to jumpstart programs from institutions underrepresented in computational-/data-intensive research



IDeA is a congressionally mandated program that builds research capacity in states that historically have had low levels of NIH funding.

UAMS
UNIVERSITY OF ARKANSAS FOR MEDICAL SCIENCES

Maine's Public Universities
UNIVERSITY OF MAINE SYSTEM

New Award Supplements: Cloud Module Development IDeA State Institutions

Collaborative R&D engagements with STRIDES cloud partners to develop new biomedical capabilities in the cloud

Module	Awardee*
Fundamentals of Bioinformatics (configuration, data manipulation, genome assembly)	NH INBRE
DNA Methylation Sequencing Data Analysis (sequence processing and analysis)	HI INBRE
Consensus Pathway Analysis (high-throughput data processing, differential gene expression, gene set enrichment, consensus analysis and visualization)	NV INBRE
Assay for Transposase Accessible Chromatin (ATAC-seq) to identify open/accessible regions of the genome	NE INBRE
AI/ML Development (Python/Pytorch, BigData, Deep Learning, Hadoop and Map Reduce, Image analysis)	AR INBRE
Biofilm-Microbiome Composition [16S and Metagenomics], Diversity, and Function	SD INBRE
Data Science for Biology (Introduction to R and R Studio, creating plots, statistical model)	SFSU
Transcriptome Analysis (QC, preprocessing, normalization, assembly, annotation, differential expression)	ME INBRE
Biomarker Discovery from Proteomics, Metabolomics, and Transcriptome data	RI INBRE
Integrating Multi-Omics (Transcriptome, Epigenetics, and Proteomics datasets)	ND INBRE

**INBRE = IDeA Networks of Biomedical Research Excellence*

NIH Cloud Lab Overview

A cloud testbed allowing researchers to “try before they buy”

Primary Cloud Lab Use Cases

Exploring the Cloud Consoles
Researchers can gain an understanding of the look and feel of cloud environments before they jump into a full STRIDES account for research

Supplementing Cloud Training
Researchers can use the sandbox to strengthen their understanding of cloud training or follow along with training content in a separate environment.

Experimenting with Simple Cloud Solutions
Researchers interested in solutions for specific scientific tasks can use the sandbox to build proof of concept or other simple solutions to understand LOE and other details for production.

Benchmarking Costs
Testing out different tools and configurations (instance types, sizes, etc.) to optimize research analyses

STRIDES for NIH: Enterprise Cloud Platforms

National Institutes of Health

Enterprise Cloud Platforms for NIH-Wide Use

STRIDES Enterprise: Features

- **Secure, dedicated network connectivity** from NIH to cloud platforms
- **Applied and inheritable cybersecurity controls** and authority to operate (ATO)
- **Optimized cloud environments** with assistance from the cloud vendors
- **Extended federated login** and Identity and Access Management services

Common Researcher Requests

- An easy path to cloud
- A place to start learning the cloud
- Help in understanding how cloud can benefit them
- Advice on how to do things differently in the cloud
- Hearing about how others do things in the cloud
- Guidance and best practices for cloud
- Help building their workloads
- Customized shelf-ready services and services
- Help defining their specific security controls and documentation
- Help designing the architecture for their systems
- Help when something doesn't work as expected
- Help optimizing how they are using cloud resources

Looking Ahead

Continue to increase the adoption and use of STRIDES

- Include STRIDES in all relevant NIH funding opportunities as encouraged but not required
- Emphasize minority-serving and historically unrepresented institutions
- Establish new complementary advisory groups for Enterprise Cloud Platforms & STRIDES Extramural Adoption

Explore expansion of partnerships to include widely used biomedical software and platforms

Scale adoption of enterprise cloud platforms and support within NIH

Develop more scalable, sustainable cloud training strategies

- Partner with longstanding, NIH-funded extramural training centers