

NIH STRIDES Initiative

**Networking and Information Technology Research and
Development**

Middleware and Grid Interagency Coordination (MAGIC)

Valerie C. Virta, PhD

AAAS Science and Technology Policy Fellow

NIH STRIDES Initiative, Cloud Services Program

Center for Information Technology

National Institutes of Health

Outline

- Brief background on use of cloud at NIH and the NIH STRIDES Initiative
- High-level system architecture for NIH-managed cloud environments
- Benefits to this approach
- Next steps
- Success stories
- Questions and discussion

Turning Research Data Into Knowledge and Discovery



The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative

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

Partnerships with  ,  , and other commercial providers

Strategic Plan for Data Science: Goals and Objectives

Data Infrastructure

Optimize data storage and security

Connect NIH data systems

Modernized Data Ecosystem

Modernize data repository ecosystems

Support storage and sharing of individual datasets

Better integrate clinical and observational data into biomedical data science

Data Management, Analytics, and Tools

Support useful, generalizable, and accessible tools

Broaden utility of, and access to, specialized tools

Improve discovery and cataloging resources

Workforce Development

Enhance the NIH data science workforce

Expand the national research workforce

Engage a broader community

Stewardship and Sustainability

Develop policies for a FAIR data ecosystem

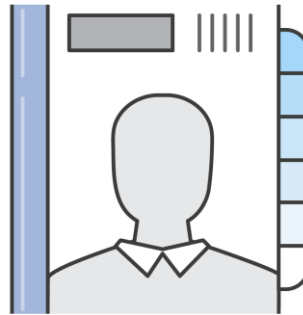
Enhance stewardship

STRIDES Benefits to Research Programs



Cost Discounts

Significant savings on full catalog of services, including compute, storage, and analytics



Professional Services & Support

Range of engagements, from consultations to custom-scoped collaborative development efforts



Training

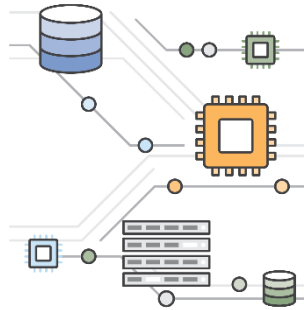
Inclusive of standard introductory content as well as customized training for biomedicine (in-person and online)

STRIDES Benefits in Aggregate to NIH



Additional Data Protections

More options to secure data and systems, using modern cyber-security capabilities



Ecosystem Development

Connecting data sets, tools, resources, and researchers in new ways



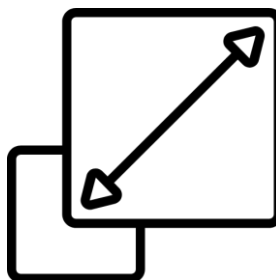
Trends & Insights

New insights into cost and usage of data sets and resources, to inform sustainability efforts

Cloud Computing Can Be Great for Research...



Always available, on-demand



Scales up and down easily



Pay as you go, for only what you use



Test and iterate



A rich and ever-growing set of tools and services

...But Getting There Can Be Difficult

- When NIH ICs want to use the cloud on behalf of their institutes and researchers, a host of complexities arise:
 - Setting up acquisition vehicles
 - Budgeting and paying for usage
 - Growing, securing, and maintaining prototype capabilities as more robust infrastructure, systems, and services
- Most ICs have established some projects in the cloud, but the number and maturity of these projects vary greatly by IC
 - STRIDES offers a standardized approach to accessing and paying for cloud services, primarily to accelerate our data science efforts—and this has surfaced additional needs and opportunities for managing NIH systems and data

Training is Critical

Regular trainings on cloud platforms, including on technical/infrastructure and applied scientific training



Results

- To date, STRIDES has held nearly 50 training sessions for over 1,200 NIH and extramural attendees
 - Multi-day, multi-modal (in-person and online)
- Training course examples:
 - Architecting with Google Cloud Platform
 - Big Data & Machine Learning on GCP
 - Developing on AWS
 - AWS Cybersecurity Overview

**Over 80% of
trainings
have waitlists!**

20-25 spots available in each course; typical demand is about 30 per course



National Institutes of Health

Highlights: STRIDES by the Numbers*

18

NIH ICs
participating

148 programs

230 accounts
onboarded

35

extramural
institutions
participating

>1,200

people trained

71

petabytes stored

>20M

compute hours

\$40M contributed

\$5.8M saved (avoided)

* As of 6/24/20

Intramural Onboarding Summary*

AWS Fully Onboarded Programs: 122 (+64 in 2020; +7 last month)

- 3 currently in progress

Google Fully Onboarded Programs: 65 (+18 in 2020; +4 last month)

- 1 currently in progress

CSP	Onboarding Participant Type / Status by Phase	# of IC with account(s)	# Provisioned / Billing Acct Assigned
AWS	Intramural (<i>including credit accounts</i>)	14	122
GCP	Intramural (<i>including credit accounts</i>)	12	65

*as of 6/10/20

Extramural Onboarding Summary

AWS Enrolled Institutions: 14 (+8 in 2020; +2 last month)

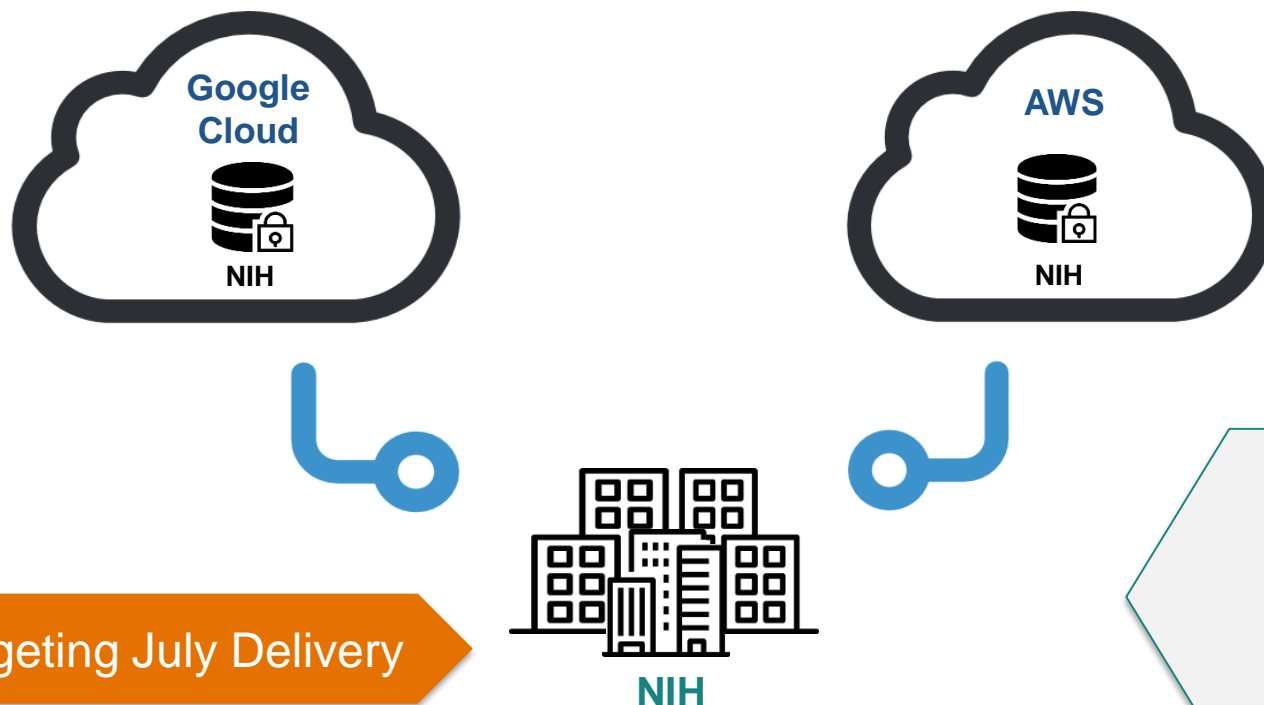
- **University of Michigan, University of Massachusetts Medical Center, Johns Hopkins University, Mayo Clinic, Roswell Park Cancer Center, UC Santa Cruz, University of Chicago, University of Pittsburgh**
 - 30 currently in discussions

Google Enrolled Institutions: 8 (+4 in 2020; +1 May; +3 June to date)

- Georgetown University, Johns Hopkins University, Harvard School of Public Health, University of Chicago, University of Pittsburgh, Penn State University, Dana-Farber
 - 33 currently in discussions

CSP	Onboarding Participant Type / Status by Phase	# of Reseller Agreements Signed	# of STRIDES Agreement Signed	# Provisioned / Billing Acct Assigned
AWS	Extramural (including Pilot 40, NIH Invoiced & non-edu managed)		14	9
GCP	Extramural (including Pilot 40, NIH Invoiced & non-edu managed)	4	4	8

Operationalizing Cloud for NIH-Managed Data



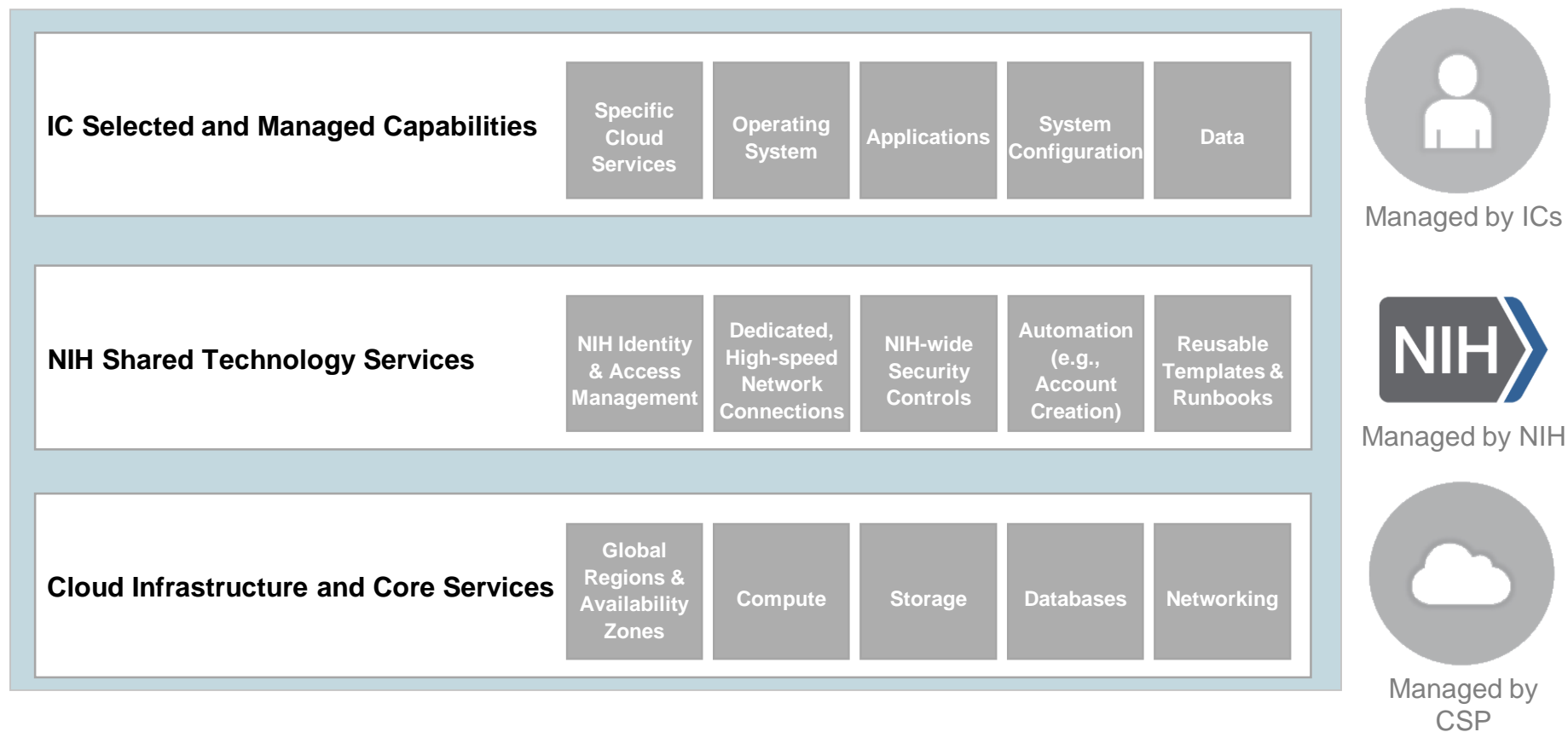
Targeting July Delivery

GOAL:

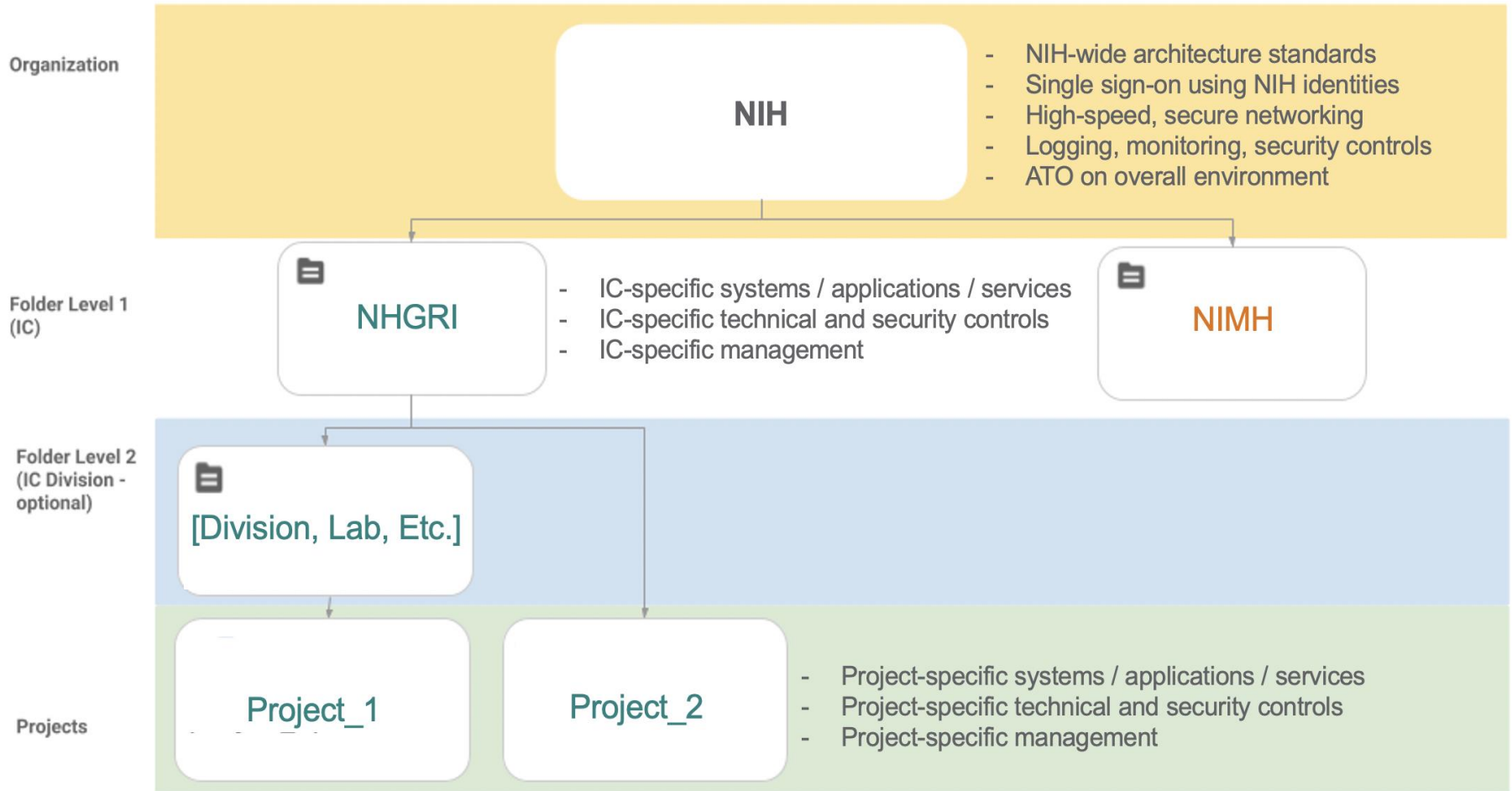
Robust cloud capabilities with standardized cybersecurity controls and extensive ATO inheritance

- 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

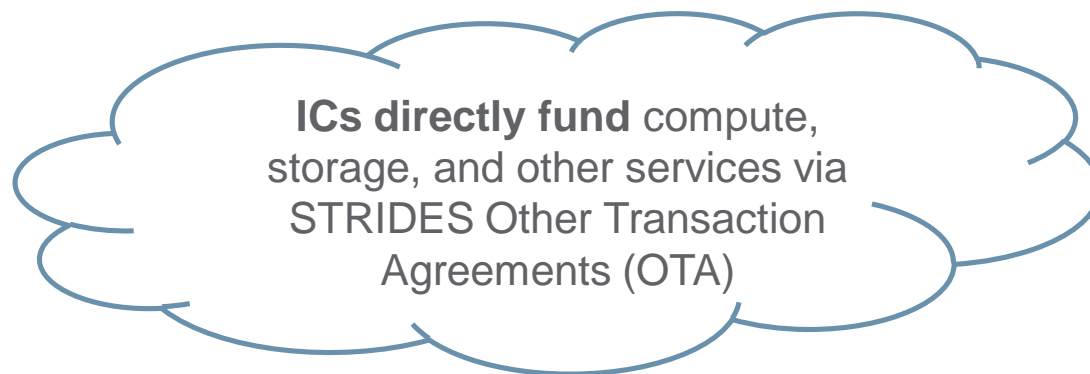
NIH-Managed Cloud Environments—Service Layers



NIH-Managed Cloud Environments—Overall Architecture



Funding Model



NIH Shared Technology Services

NIH Identity
& Access
Management

Dedicated,
High-speed
Network
Connections

NIH-wide
Security
Controls

Automation
(e.g.,
Account
Creation)

Reusable
Templates &
Runbooks

Shared technology service costs are **centrally funded**

How Researchers and NIH Can Benefit



Consistent use of NIH identities
and login credentials



Institute and Center funding and
expense management,
with NIH billing/payment service support



Security controls
and guardrails



Dedicated, secure connectivity
for data/file transfer



Monthly reports and dashboard

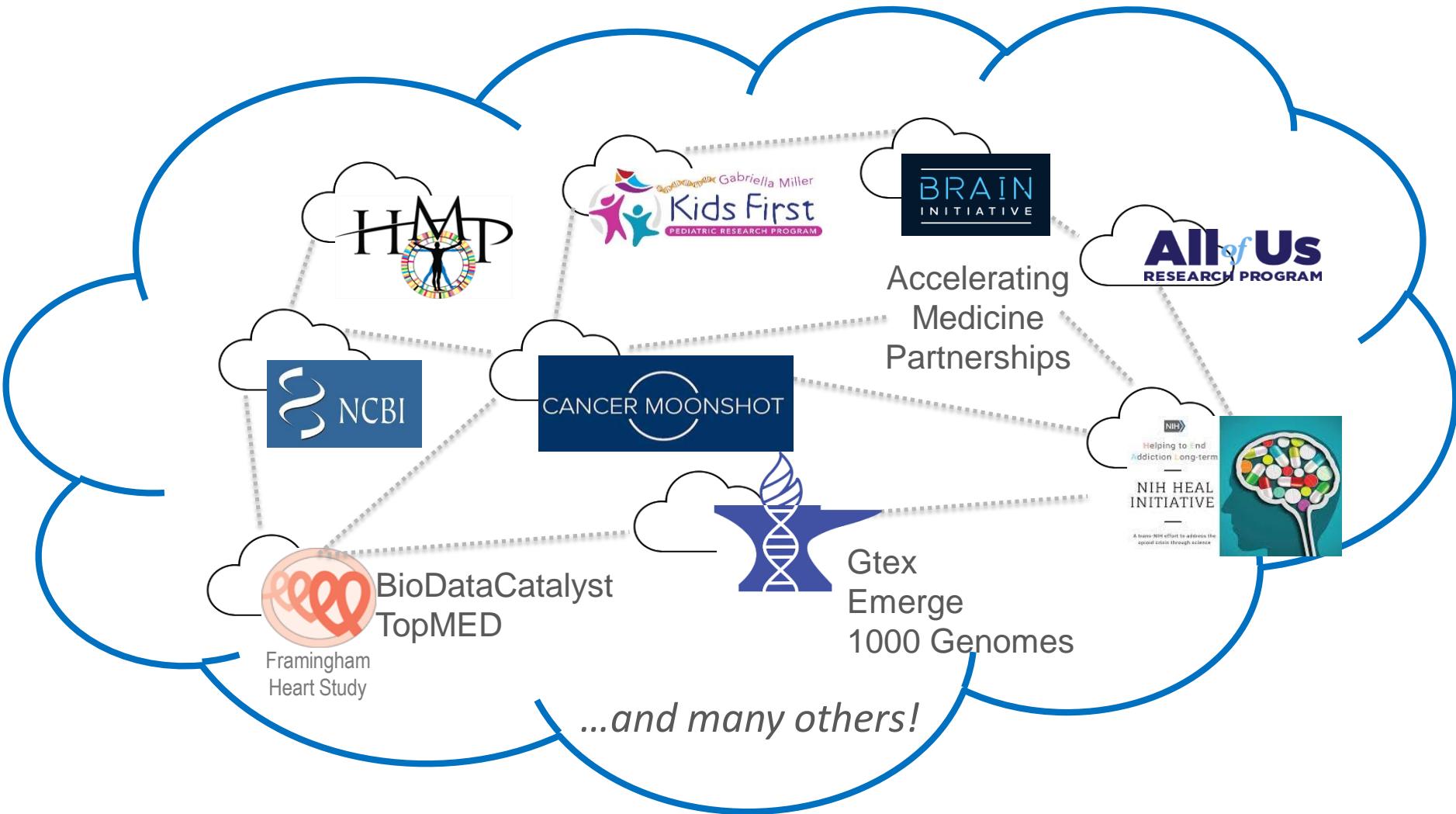


Monitoring and threat protection

How a Researcher Uses the NIH-Managed Cloud Environments

1. Decide on the cloud platform best suited for the research (based on features, cost, and/or other factors)
2. Participate in NIH-provided trainings
3. Provide project and funding documentation and co-fund
4. Receive access to the managed cloud environment
5. Architect, construct, and configuration resources (e.g., storage, compute, database, etc.) to support the research project
6. Get direct technical and troubleshooting support
7. Make research discoveries, share data, and publish findings

Success Stories for a Future of Interconnected Data Sets



Success Stories: Sequence Read Archive



The National Library of Medicine's National Center of Biotechnology Information **Sequence Read Archive (SRA)** is the largest-growing repository of molecular data, archives of raw sequencing data, and alignment information from high-throughput sequencing platforms.

SRA data is used by more than 100,000 researchers every month. As of September 2019, five petabytes of public SRA data has been moved to the cloud.

Success Stories: Kids First

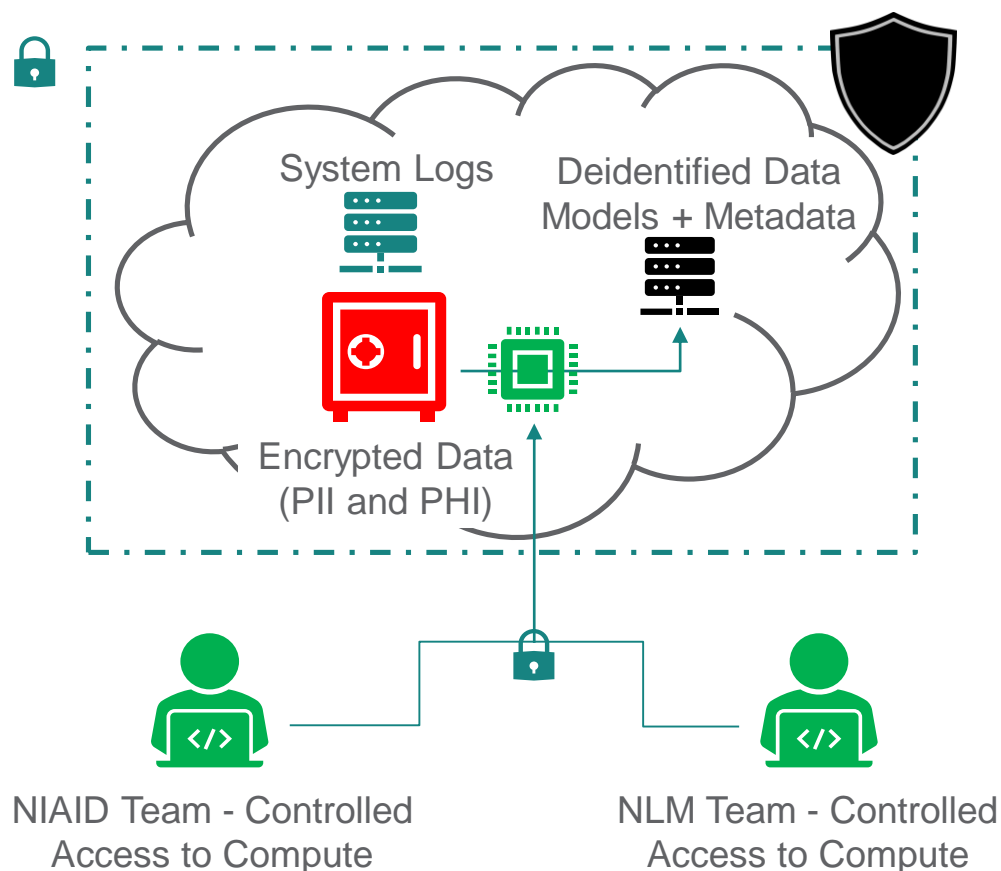


In 2018, **Kids First** launched the Gabriella Miller Kids First Data Resource Portal, a large-scale database of clinical and genetic data from patients with childhood cancers and structural birth defects and their families.

The program has sequenced more than 20,000 samples from childhood cancer and structural birth defects cohorts since 2015.

Success Stories: PII-Secured AWS Computing Environment (PACE)

The **PII-Secured AWS Computing Environment (PACE)** at the National Institute for Allergies and Infectious Diseases (NIAID) leverages the AWS cloud for collaborative patient record text mining and candidate gene prioritization.



The project uses machine learning in collaboration with researchers at the National Library of Medicine (NLM).

The STRIDES Team



Nick Weber
STRIDES Program
Manager



James Davis
Business Analyst



Allissa Dillman
Training Strategist



Dana Gaffney
Business Analyst



Matt Gieseke
Training
Coordinator



**Djamil Lakhdar-
Hamina**
Cloud Support
Engineer



Eric Mensah
Cloud Architect



Joel Mills
AWS Account
Manager



Anteju Nuhanovic
Cloud Architect



Joel Peterson
Cloud Architect



Todd Reilly
Onboarding Lead



Tom Shaw
Google Accounts



Michelle Speir
Communications
Specialist



Joshua Stultz
Cloud Architect



Jared Taylor
Program Analyst



Valerie Virta
AAAS Science &
Tech Policy Fellow



Annie Wang
Civic Digital Fellow



Sherika Wynter
Product Manager



Bob Zhao
Civic Digital Fellow

Questions?

Contact:

Valerie Virta, AAAS S&TP Fellow | valerie.virta@nih.gov

STRIDES general inquiries | STRIDES@nih.gov

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development
(NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674,
Fax: 202-459-9673, Email: nco@nitrd.gov, Website: <https://www.nitrd.gov>

