

Strategies & Trends on Cloud Computing at NIH

NIH Virtual Workshop on Broadening Cloud Computing Usage in Biomedical Research

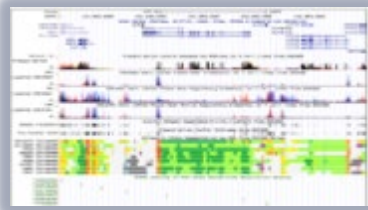
Nick Weber

Program Manager, Cloud Services | NIH STRIDES Initiative
NIH Center for Information Technology

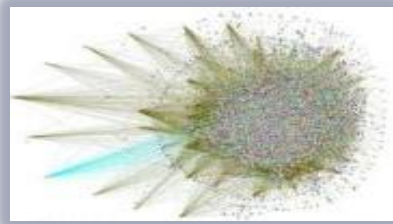
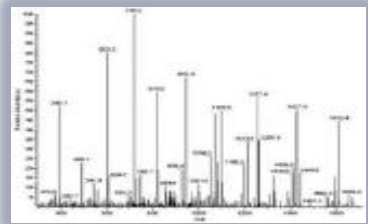
Topics for Discussion

- Research drivers for cloud computing
- What is cloud computing?
- Benefits and challenges of cloud computing for biomedical research
- Overview of NIH and NIH-funded cloud platforms
- Cost and security considerations of cloud computing
- Questions & Discussion

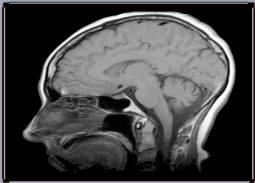
Research Drivers for Cloud Computing: Data-Intensive Science



Genomic



Other 'Omics



Imaging



Clinical & EHR



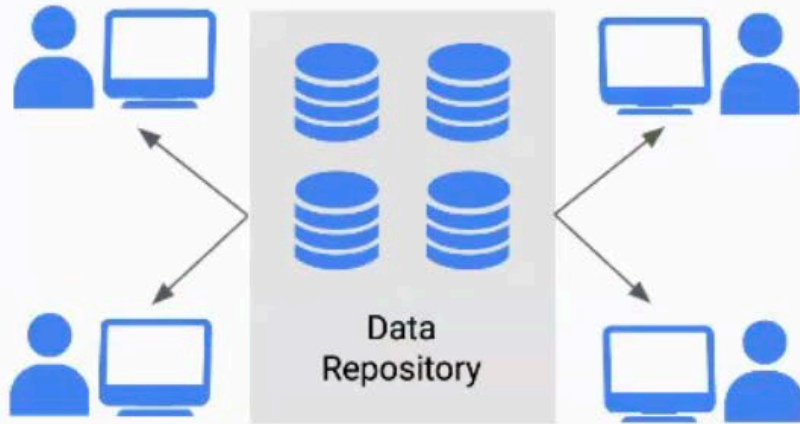
Environmental



Social

Inverting the Model of Data Sharing

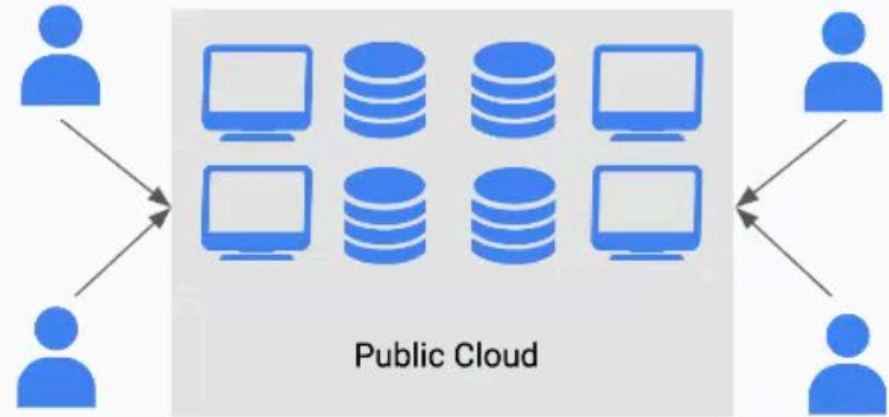
Old Way: Bring data to the researchers



Problems

Data sharing = data copying
Security
Accessibility
Inelastic

New Way: Bring researchers to the data



Solutions

Less Expensive
Audit Trails
Greater Accessibility
Elasticity

Expanding Access to TOPMed Genomics Data Using Cloud Services



Data generated by the TOPMed program are made available through **NHLBI's BioData Catalyst**, a **cloud-based platform** providing researchers with tools, applications, and workflows in secure workspaces.

The primary goal of BioData Catalyst is to **build a data science ecosystem that creates efficiencies for research** and ultimately results in discovery and scientific advancements.

KEY OUTCOMES OF USING CLOUD SERVICES AND TOOLS:

- **Share data easily from a central location**
- **Provide compute resources and access to data** to researchers who did not have access before
- **Extend the reuse of data generated** and stored by TOPMed projects

According to an interview with the University of Michigan team who works with the TOPMed and BioData Catalyst programs.





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.

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.

“

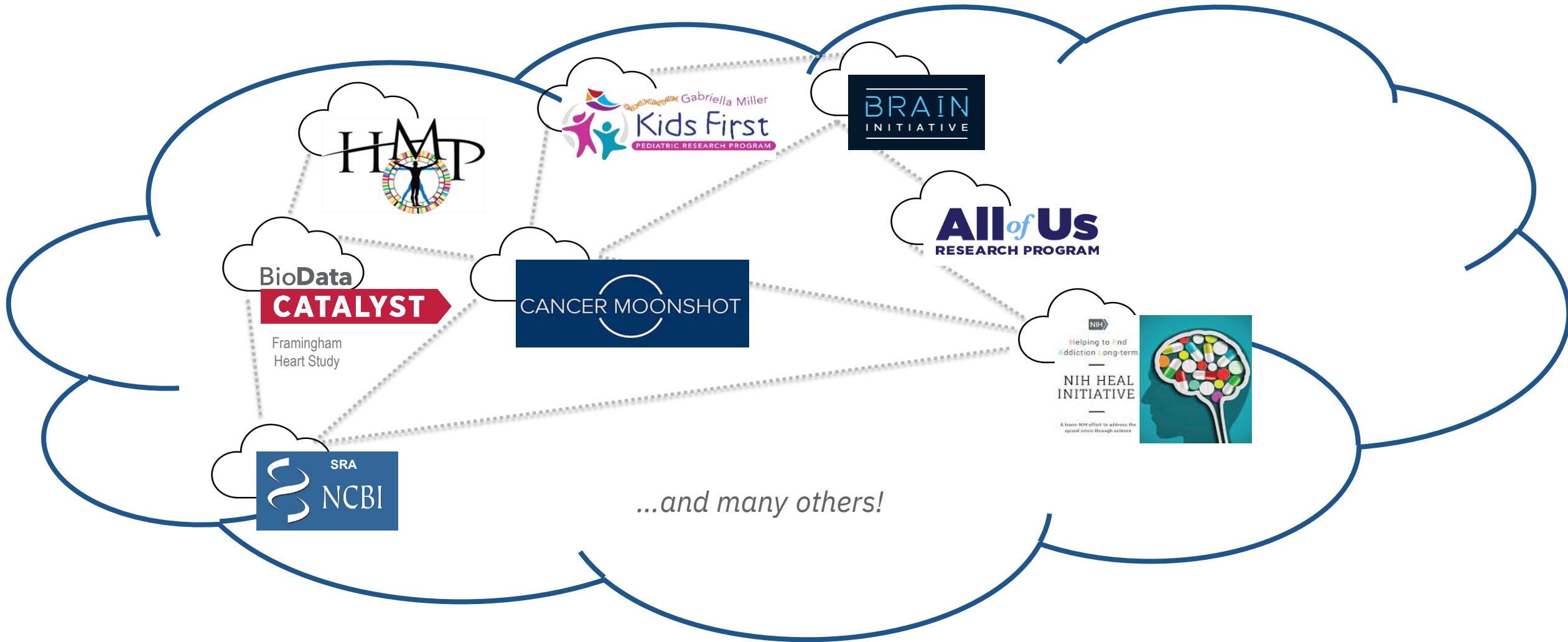
By moving to the cloud, we have been able to **compress a year's worth of data processing into a couple months.**

”

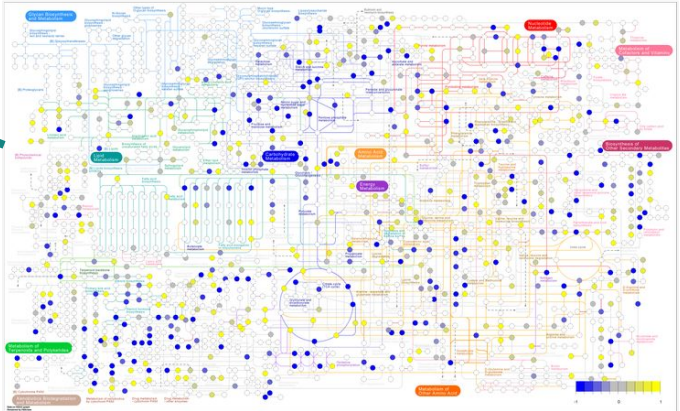
– **Jonathan LeFaive**, senior app programmer/analyst, Department of Biostatistics at the University of Michigan



Envisioning a Future of Interconnected Datasets



What will we discover when we can link individuals' electronic health care records with their personal data, alongside clinical and basic research data?

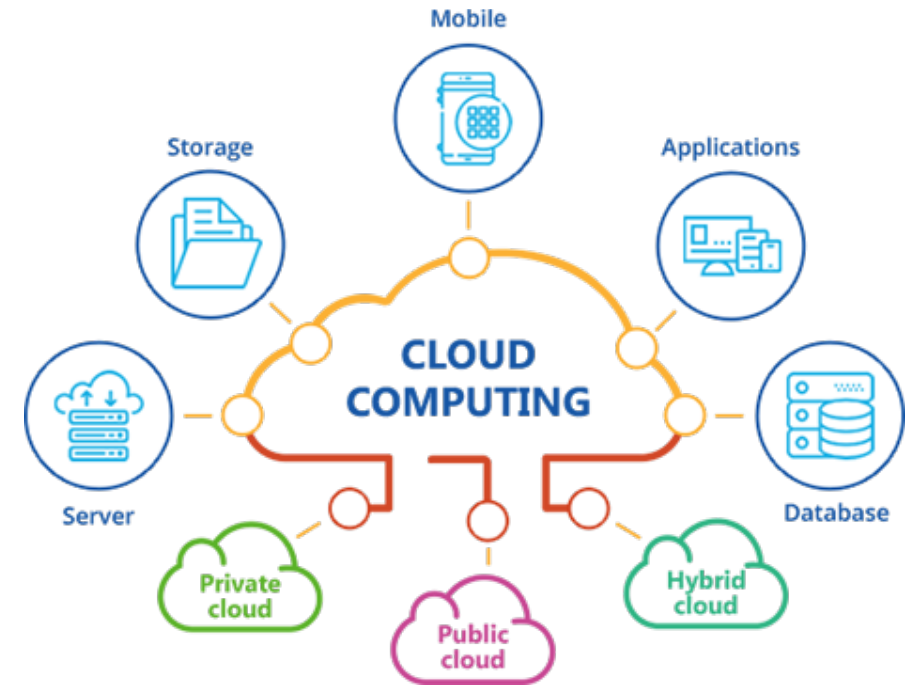


What is cloud computing, really?

What is cloud computing?







- Per NIST (2011):

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.



- 5 essential characteristics

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured service

 Facebook	Social	 Google Drive	Cloud Storage
 Google Gmail	Webmail	 iCloud	Cloud Storage
 YouTube	Consumer	 Slack	Collaboration

Common Cloud Infrastructure Providers



Deployment & Management

Application Services

- Amazon SQS
- Amazon ElasticTranscoder
- Amazon SES
- Amazon AppStream
- Amazon CloudSearch

Mobile Services

- Amazon Cognito
- Amazon Mobile Analytics
- Amazon SNS

Enterprise Applications

- Amazon WorkDocs
- Amazon WorkSpaces
- Amazon WorkMail

Application Services

Administration & Security

- AWS DirectoryService
- AWS IAM
- AWS Trusted Advisor
- AWS Config
- AWS CloudTrail
- Amazon CloudWatch

Deployment & Management

- Amazon CloudFormation
- AWS OpsWorks
- AWS CodeDeploy

Analytics

- Amazon Kinesis
- AWS Data Pipeline
- Amazon EMR

Foundation Services

Compute

- Amazon EC2
- AWS Lambda

Storage & Content Delivery

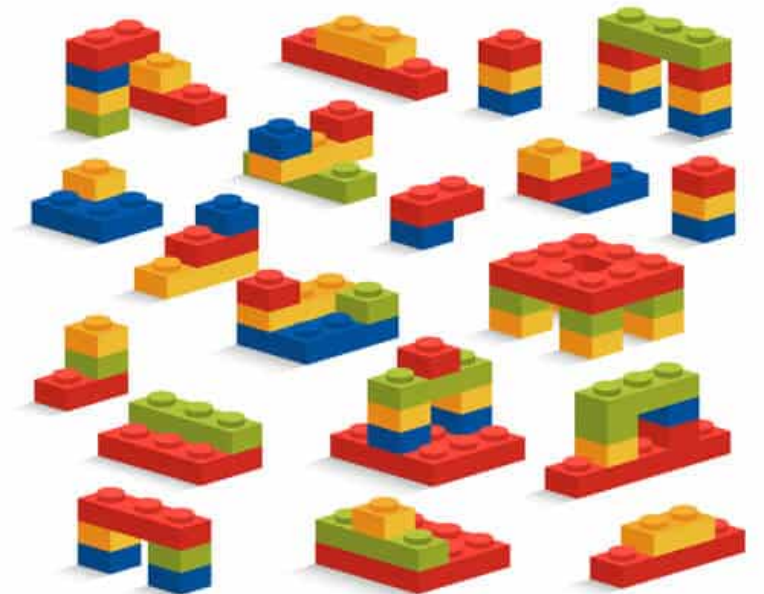
- Amazon CloudFront
- Amazon Glacier
- AWS Storage Gateway
- Amazon Content Delivery

Database

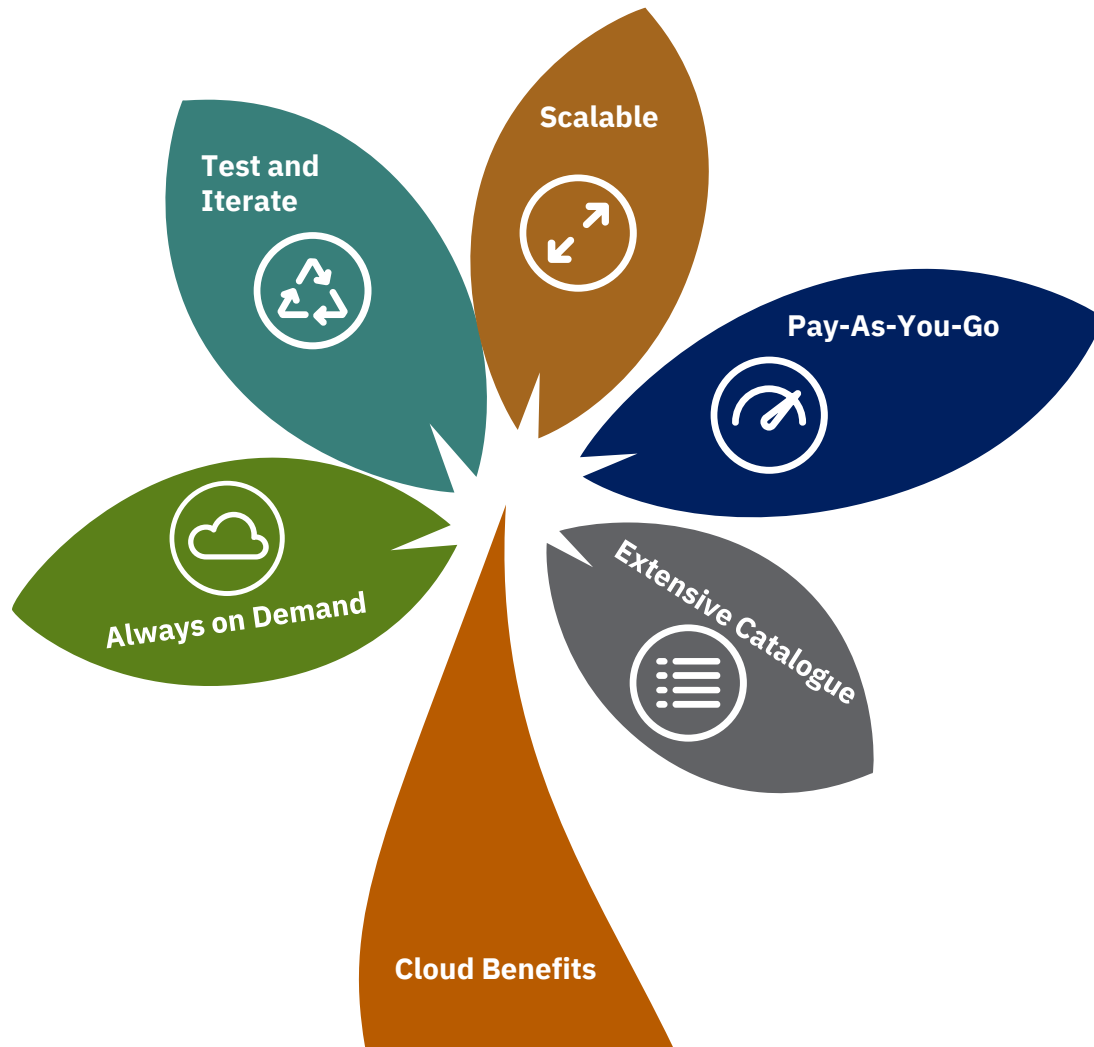
- Amazon Dynamo DB
- Amazon RDS
- Amazon Redshift
- Amazon Elastic Cache

Networking

- Amazon Route 53
- Amazon VPC
- AWS Direct Connect



Benefits of Cloud Computing for Research



- Always on demand**
Implement tools and services whenever and wherever they are needed
- Test and Iterate**
Leverage environments to test applications prior to deployment and make changes and collaborate as needed
- Scalable**
Easily meet the demand as it exists and automate elasticity
- Pay as you go**
Pay for what you use when you use it
- Extensive Catalogue**
of services and tools to leverage for research

Common Cloud Challenges

- Setting up acquisition vehicles
- Budgeting and paying for usage / optimizing for cost / preventing inadvertent cost overruns
- Learning new tools and new ways of working—for individual and organization
- Growing, securing, and maintaining easily-prototyped capabilities as robust infrastructure, systems, and services
- Many options and building blocks means (too?) many ways to do things



“I want you to find a bold and innovative new way to do everything exactly the same way we’ve been doing it for 25 years.”

NIH Strategic Plan for Data Science

Data resource ecosystem
and infrastructure
modernization

Data sharing, access,
and interoperability

EHR, clinical, and
observational data
availability enhancements



All while ensuring data confidentiality

Making Data FAIR

Findable

Data must have unique identifiers, effectively labeling it within searchable resources



Interoperable

Data should “use and speak the same language” through the use of standardized vocabularies



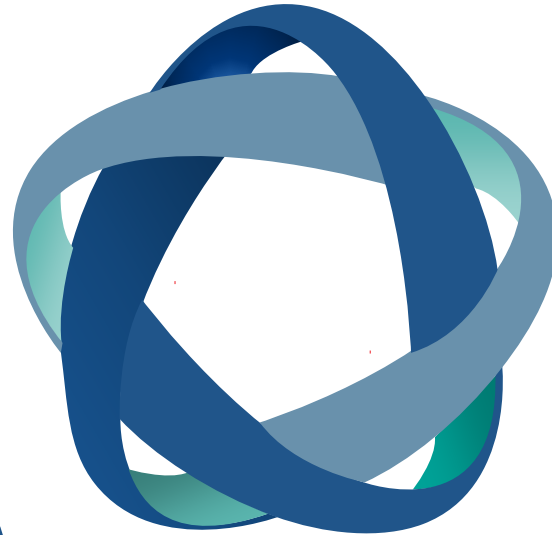
Accessible

Data must be easily retrievable through open systems, and require effective and secure authentication and authorization procedures



Reusable

Data must be adequately described to a new user, have clear information about data-usage licenses, and have a traceable “owner’s manual,” or provenance



NIH Cloud Ecosystem

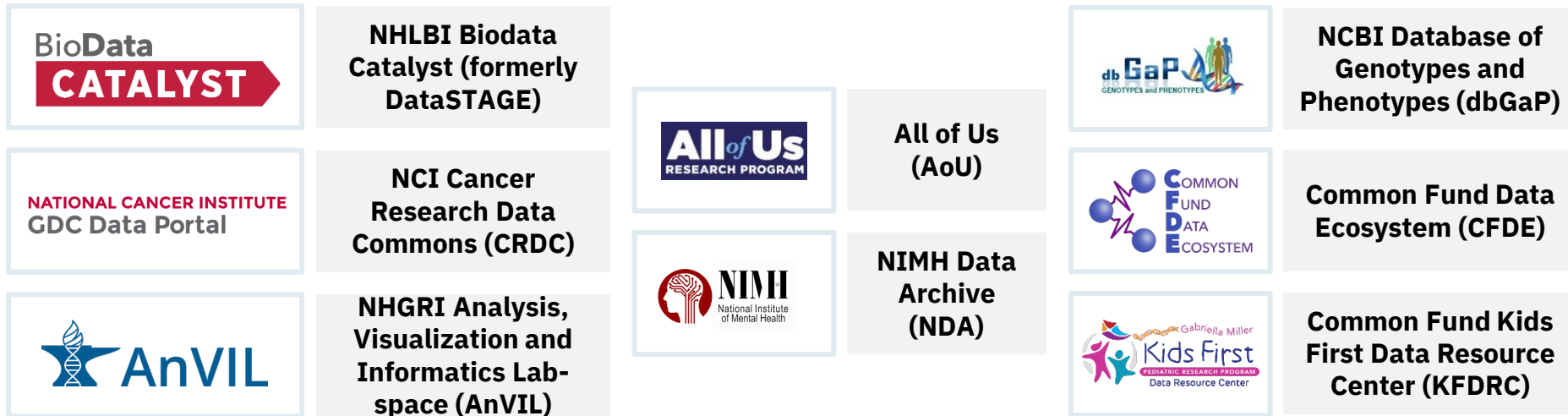
Cloud Research Platforms and Software Solutions



Platform Concept – Specialized Functionality



Evolving Ecosystem of Platforms and Capabilities



Cross-Platform Data Discovery

A dataset catalog for a "bird's eye view" of available datasets.

Generic Search Results Hand-off

A generic and universal hand-off mechanism so data portal users can further analyze search results on any analysis platform that supports the format.

NIH Researcher Authentication Service

A unified, efficient, and secure authentication and authorization service that enables streamlined access by researchers to NIH-funded data resources across multiple systems and provides standardized methods of logging and auditing such access.

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

Partnerships with



Google Cloud



Microsoft Azure

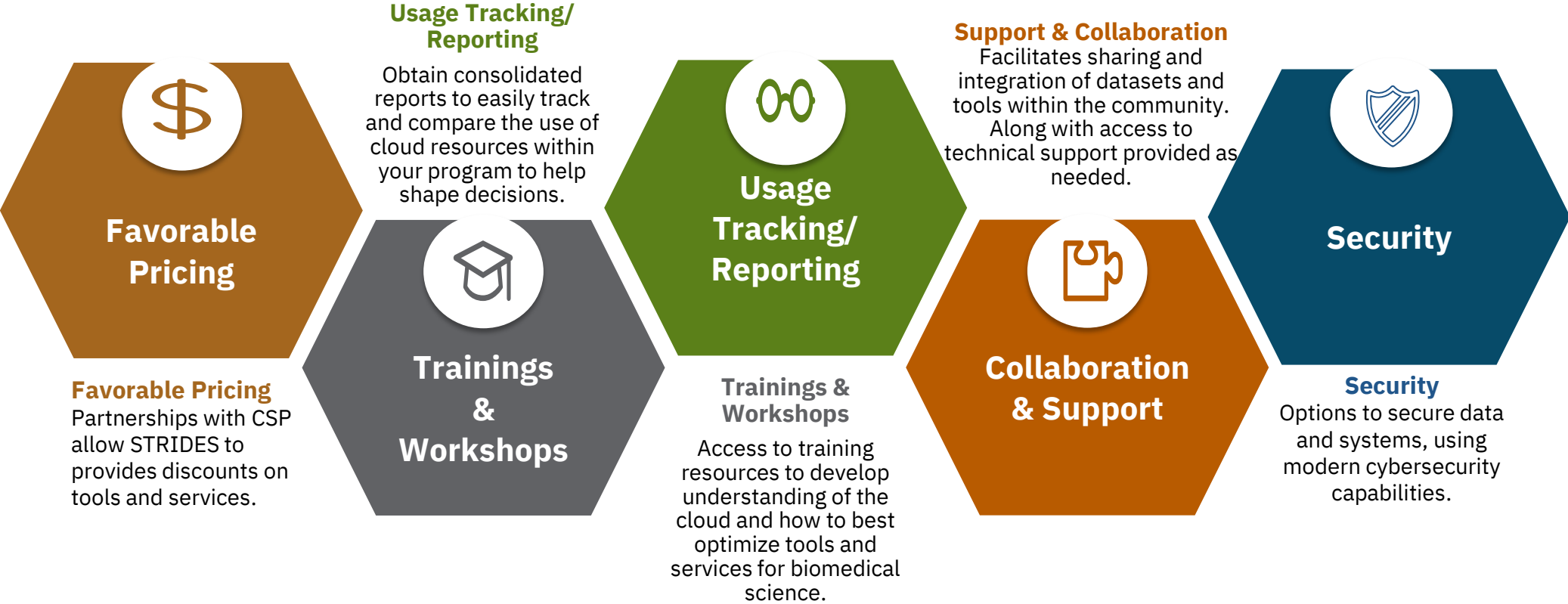
Core Motivations for STRIDES

- Democratization of computational research and data science
 - Leveling the playing field for traditionally underrepresented communities and institutions in biomedical research
- Cost savings and efficiencies for the research community at large
 - More usage begets more savings and greater overall discounts for all
- Strong partnerships with cloud providers
 - Resulting in collaborative R&D engagements and more direct focus and support on research



What STRIDES Provides to Research Programs

STRIDES provides several resources in addition to favorable pricing that may enhance and enable more efficient biomedical research



Favorable Pricing

Favorable Pricing
Partnerships with CSP allow STRIDES to provide discounts on tools and services.

Usage Tracking/Reporting

Obtain consolidated reports to easily track and compare the use of cloud resources within your program to help shape decisions.



Usage Tracking/Reporting

Trainings & Workshops
Access to training resources to develop understanding of the cloud and how to best optimize tools and services for biomedical science.



Trainings & Workshops

Support & Collaboration

Facilitates sharing and integration of datasets and tools within the community. Along with access to technical support provided as needed.



Collaboration & Support



Security

Security
Options to secure data and systems, using modern cybersecurity capabilities.

**Helping advance
biomedical research
by delivering access
to industry-leading
cloud providers.**



The STRIDES Initiative aims to help NIH and its institutions accelerate biomedical research by reducing barriers in utilizing commercial cloud services. This initiative aims to harness the power of the cloud to accelerate biomedical discovery. NIH and NIH-funded researchers can take advantage of STRIDES benefits.

Enroll Now

strides@nih.gov

Gain access to

- Discounts on partner services
- Professional services consultations
- Access to training
- Potential collaborative engagements

>110
petabytes
stored

>100M
compute
hours

>600
Program /
project
accounts
onboarded

>\$20M
cost savings
for NIH

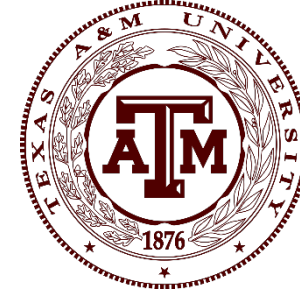
>4000
people trained



Stanford University



EMORY



MASSACHUSETTS GENERAL HOSPITAL



Major Research Institutions Enrolled



COLUMBIA UNIVERSITY IRVING MEDICAL CENTER



Penn UNIVERSITY of PENNSYLVANIA



Yale University



BROWN



HARVARD MEDICAL SCHOOL



WAYNE STATE UNIVERSITY



Caltech



WISCONSIN UNIVERSITY OF WISCONSIN-MADISON



Georgetown University



DANA-FARBER CANCER INSTITUTE



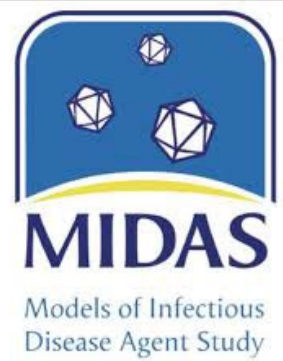
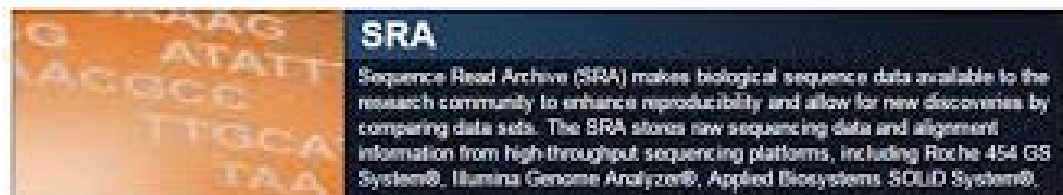
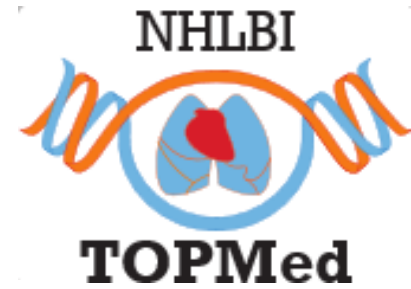
Icahn School of Medicine at Mount Sinai



NATIONAL CANCER INSTITUTE
GENOMIC DATA COMMONS



Major
Research
Programs
Supported



STRIDES Training



- Incredible **demand for cloud training** (nearly all courses have waitlists)
 - Course offerings range from fundamentals, to research support/technical topics (e.g., security, networking), to applied topics (e.g., big data and machine learning)
 - Modalities include in-person, virtual, scheduled, on-demand, at NIH, at major research institutions, associated with conferences/events— with a high demand for all
- Based on feedback from participants, 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

STRIDESTraining@nih.gov

Cloud Cost & Security Considerations

Typical Cloud Costs and How They're Accrued

How Cloud Costs Are Accrued:

Cloud costs are charged based on services used. These costs are typically metered costs, meaning you only pay for the actual services consumed, which largely fall into major categories of consumption.

Cloud Resources by Consumption Categories

Compute

- Virtual Machines
- Container Execution
- Kubernetes Clusters
- Spark & Hadoop Clusters
- Pipelines
- Query Engines
- ML/AI Services
- Serverless Compute

Storage

- Block Storage
- File Storage
- Object Storage
- SQL Storage
- NoSQL Storage
- DICOM/EHR/FHIR Storage

Analytics

- Sagemaker
- Quicksight
- Kinesis
- Big Query/Big Query Storage API

Networking

- Ingress - Generally Free
- External IPs
- Direct Connect
- Egress – Internal (within CSP)
- Egress – External (outside CSP)
- Virtual Private Cloud

Data Transfer

- Data Transfer, CloudFront, DataSync, Snowball

Management

- Logging Services, CloudWatch

Messaging

- Pub/Sub, queuing, email

Monitoring

- Security Hub, Inspector, Security Command Center

Simple Example of How Cloud Costs Accrue

Building Cloud Infrastructure - DICOM Imaging

This application shares DICOM medical images in a protected environment by leveraging the Google Healthcare API. Using GCP resources the application launches open-source software that allows for the viewing and analysis of medical images.

Resource Type	Resource Used	Size
Compute	Virtual Machine (VM)	1 VM
	Kubernetes Cluster	3 VMs
Storage	Cloud Storage Buckets (Staging)	3.4 TBs
	DICOM Store	3.4 TBs
	SQL Instances	1 Instance
Networking	External IPs	2 IPs
	Data Ingress	No cost
3 rd Party Software	Open-Source Software	No cost
Total Cost* per Month		\$285

**Cost derivations on next slide*

Cloud Cost Analysis – Cost Orientation Examples

These screenshots show the GCP pricing estimates, prior to applying STRIDES discounts, for the DICOM imaging analysis.

Reference: DICOM Imaging Analysis Cost Calculator

Google Cloud Pricing Calculator

COMPUTE ENGINE | APP ENGINE | KUBERNETES ENGINE | CLOUD RUN | VMWARE ENGINE | CLOUD STORAGE | NETWORKING EGRESS | CLOUD LOAD BALANCING | INTEL & CL

Search for a product you are interested in.

Instances

Number of instances *
1

What are these instances for?
Operating System / Software
Free: Debian, CentOS, CoreOS, Ubuntu, or other User Provided OS

Machine Class
Regular

Machine Family
General purpose

Series
E2

Machine type
e2-standard-2 (vCPUs: 2, RAM: 8GB)

Datacenter location
Northern Virginia (us-east4)

Instances using ephemeral public IP
Instances using static public IP
1

Committed usage
None

Average hours per day each server is running *
24 hours per day

Average days per week each server is running *
7

ADD TO ESTIMATE

Reference: DICOM Imaging Analysis Cost Estimate

Estimate

Compute Engine

1 x

730 total hours per month

VM class: regular

Instance type: e2-standard-2

Region: Northern Virginia

Ephemeral public IP 730 hours: USD 2.92

Estimated Component Cost: USD 58.01 per 1 month

Persistent Disk

Northern Virginia

Zonal standard PD: 228 GiB

USD 10.03

Google Kubernetes Engine

3 x Applications

2,190 total hours per month

Instance type: e2-medium

Region: Northern Virginia

Machine Class: REGULAR

GCE Instance Cost: USD 82.64

GKE Cluster Management Fee: USD 0.00

Estimated Component Cost: USD 82.64 per 1 month

Cloud SQL for PostgreSQL

User Information

of instances: 1

Instance type: CP-DB-PG-CUSTOM-1-3.75

Location: Northern Virginia

730.0 total hours per month

SSD Storage: 10.0 GiB

Backup: 0.0 GiB

USD 54.57

Cloud Healthcare API

Montréal, Canada

Standard Requests: 100,000

Multi-operation requests: 100,000

Standard Blob Storage: 3,420.16 GiB

USD 79.23

Total Estimated Cost: USD 284.48 per 1 month

Estimate Currency
USD - US Dollar

EMAIL ESTIMATE | SAVE ESTIMATE

NIH Cloud Lab Pilot

Cloud sandboxes can be incredibly useful for those with varying levels of cloud experience, allowing researchers to create and work through their own research scenarios in a safe space. We are developing **minimum viable product (MVP) Cloud Lab environments on AWS and GCP for NIH research use, likely to be available for beta testing in early 2022.**

What can Cloud Lab sandboxes be used for?



Exploring/Comparing the Various Cloud

Researchers can gain an understanding of the differences between cloud platforms before they commit to a given one



Supplementing Cloud Training

Researchers can use sandboxes to practice and reinforce what they learned in training, even using their own data



Experimenting with Simple Cloud Solutions

Researchers interested in solutions for specific scientific tasks can use the sandbox to build proofs of concept or other simple solutions to understand level of effort to build and maintain over time



Benchmarking Costs

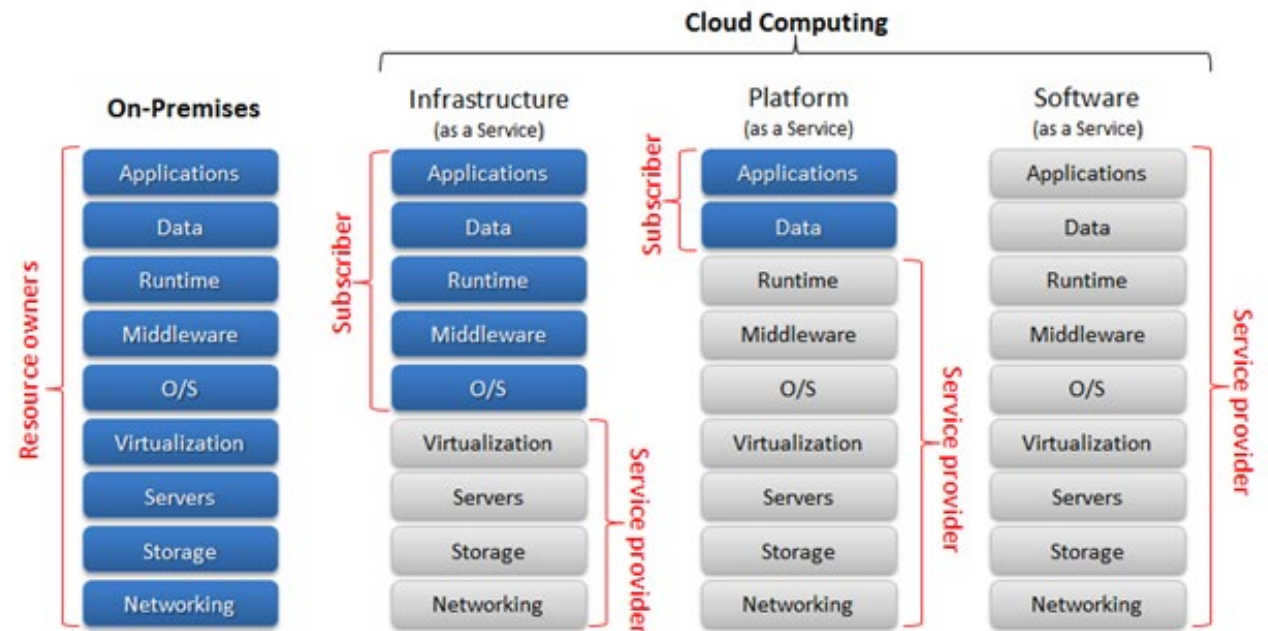
Testing out different tools and configurations (compute instance types/sizes, data storage services, analytical tools, etc.) to optimize research analyses

Security Considerations

- We're all aware of—and probably have been affected by—security breaches
- Cybersecurity isn't just the security or IT person's responsibility... it's everyone's
- It costs a lot more to try to “bolt security on” at the end as opposed to “bake it in” throughout

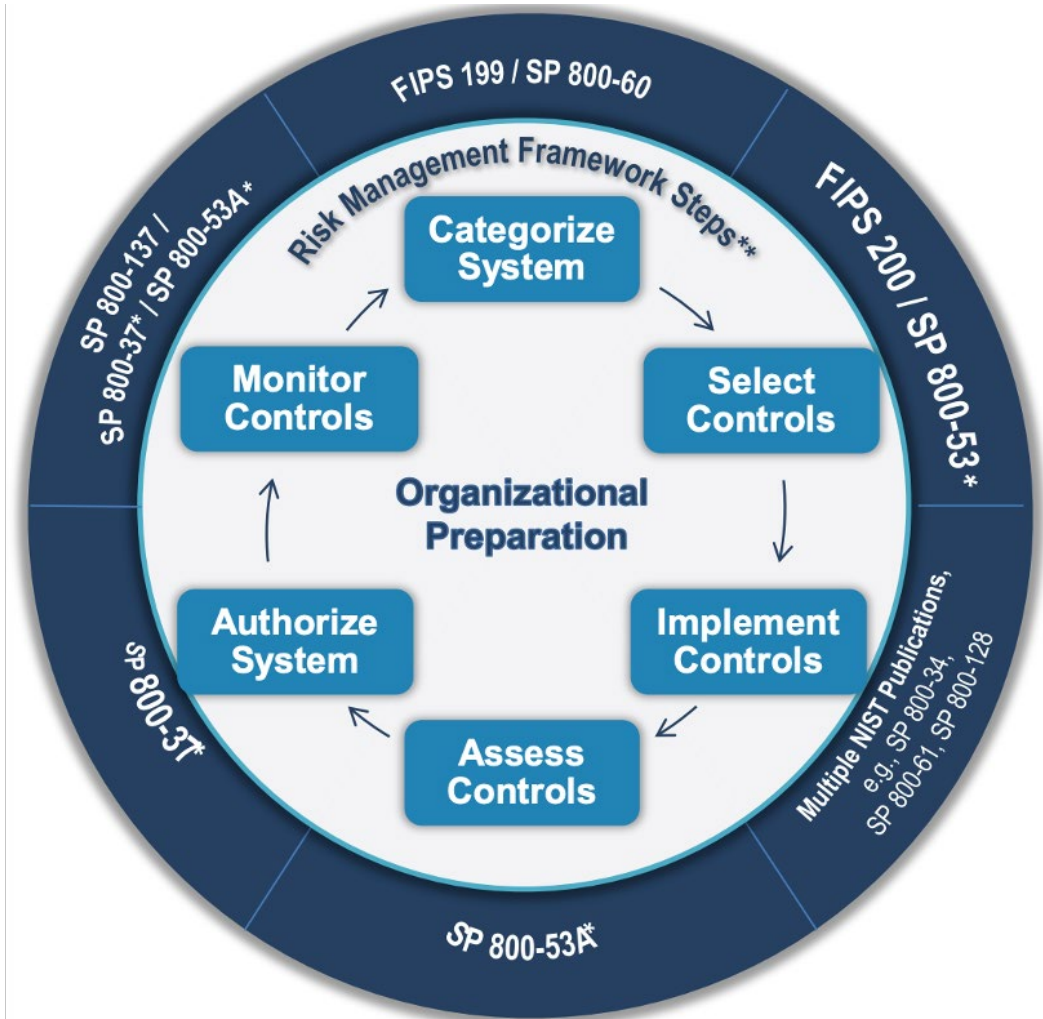


Separation of Responsibilities



FISMA & FedRAMP

- National Institute of Standards and Technology (NIST) outlines a Risk Management Framework for information systems
- Federal Information Security Management Act (FISMA) requires federal systems to have authority to operate (ATO) to protect confidentiality, integrity, and availability of systems and data
- Federal Risk and Authorization Management Program (FedRAMP) applies FISMA controls for cloud services, under a “authorize once and reuse many times” approach
 - 3rd party assessments, with systems having current FedRAMP authorizations listed at <https://marketplace.fedramp.gov/>



NIST SP 800-171 for Universities with Federal Grants/Contracts

NIST 800-171:

Some examples of organizations that would need to comply with NIST 800-171:

- Universities supported by federal grants
- Manufacturers supplying goods to federal agencies
- Service providers for federal agencies

Here's a simple table with a quick recap of the two publications:

NIST 800-171	NIST 800-53
Non-federal organizations	Federal organizations and companies with direct network connections
14 Security Control Families	18 Security Control Families

Resources Exist!

NIST SP 800-171 Compliance Template

Friday, September 27, 2019 | Briefs, Case Studies, Papers, Reports

Sources(s): **Community**

Access Control, Compliance, Cybersecurity, Cybersecurity Policy, Data Security, Security Management

Abstract


Higher education institutions continue to refine their understanding of the impact of NIST Special Publication 800-171 on their IT **systems and the data they receive from the federal government**. This compliance template will help institutions map the **NIST SP 800-171 requirements to other common security standards used in higher education**, and provides suggested responses to controls listed in NIST SP 800-171.

The NIST SP 800-171 Compliance Template was prepared by Common Solutions Group (<http://stonesoup.org/>) members. Its purpose is to provide a starting point for NIST SP 800-171 compliance. It is published by EDUCAUSE with the permission of the Common Solutions Group Steering Committee. **The template was updated September 2019.**

Related Resources

- [An Introduction to NIST SP 800-171 for Higher Education Institutions](#)
- [NIST SP 800-171 & CUI with Ron Ross Webinar](#)

DOWNLOAD RESOURCES

Download File 

Filter by title

- Microsoft compliance offerings
 - Microsoft compliance offerings
 - Global
 - US Government
 - CJIS
 - CNSSI 1253
 - DFARS
 - DoD IL2
 - DoD ILS
 - DoE 10 CFR Part 810
 - EAR
 - FedRAMP
 - FIPS 140-2
 - IRS 1075
 - ITAR
 - NDAA Section 889
 - NIST 800-161
 - NIST 800-171**
 - NIST 800-53



U.S. | GOVERNMENT AND PUBLIC SECTOR

NIST 800-171

NIST SP 800-171

08/23/2021 • 5 minutes to read • 

About NIST SP 800-171

The US National Institute of Standards and Technology (NIST) promotes and maintains measurement standards and guidelines to help protect the information and information systems of federal agencies. In response to Executive Order 13556 on managing controlled unclassified information (CUI), it published [NIST SP 800-171](#), *Protecting Controlled Unclassified Information In Nonfederal Information Systems and Organizations*. CUI is defined as information, both digital and physical, created by a government (or an entity on its behalf) that, while not classified, is still sensitive and requires protection.

NIST SP 800-171 was originally published in June 2015 and has been updated several times since then in response to evolving cyberthreats. It provides guidelines on how CUI should be securely accessed, transmitted, and stored in nonfederal information systems and organizations; its requirements fall into four main categories:

- Controls and processes for managing and protecting
- Monitoring and management of IT systems
- Clear practices and procedures for end users
- Implementation of technological and physical security measures

Microsoft and NIST SP 800-171

AWS Public Sector Blog

Get Your University Ready for NIST 800-171

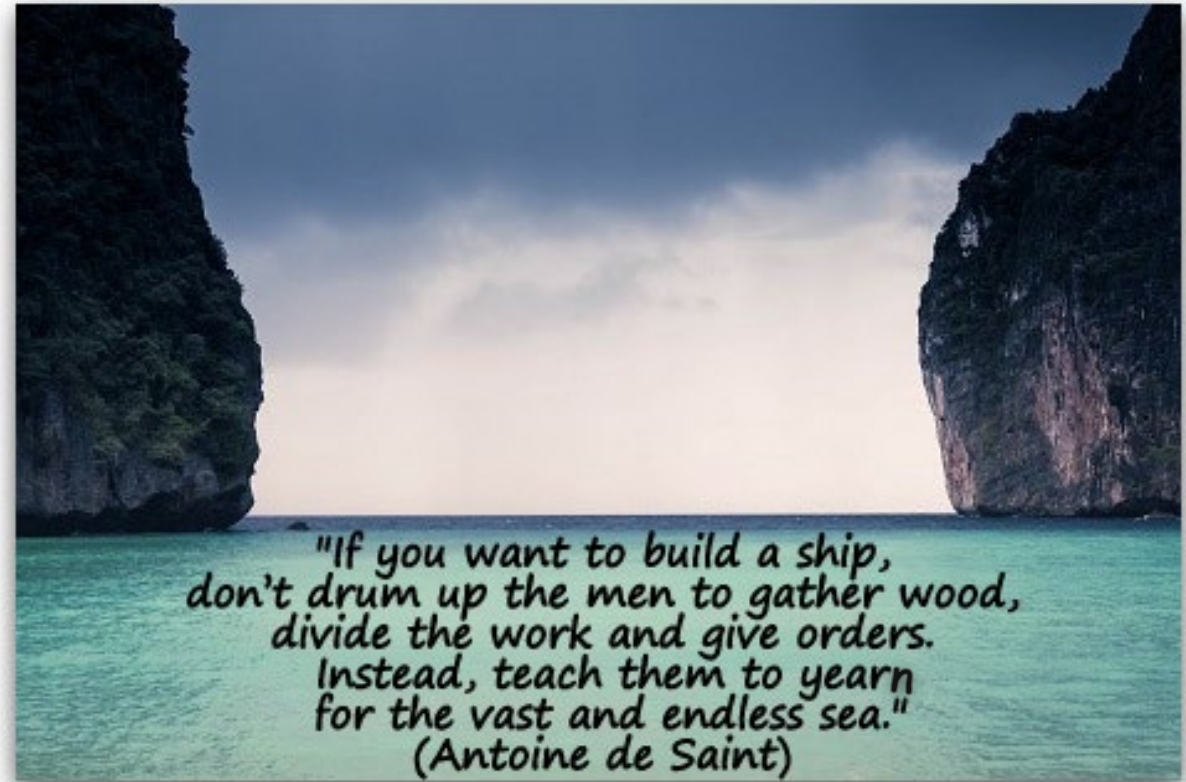
by [AWS Public Sector Blog Team](#) | on 19 SEP 2017 | in [Education](#), [Public Sector](#) | [Permalink](#) | [Share](#)

The deadline to implement National Institute of Standards and Technology (NIST) Special Publication 800-171 is fast approaching. Beginning in January 2018, you may miss out on government funding that stipulates its implementation if you have not taken action.

In 2015, [NIST published Special Publication \(SP\) 800-171](#) – Protecting Controlled Unclassified Information in Non-federal Information Systems and Organizations – introducing the standards for non-federal entities, such as academic institutions working under a government contract. NIST 800-171 was meant to take the security controls from a larger NIST publication, [NIST 800-53](#), and assist non-federal agencies to apply controlled unclassified information (CUI) controls to their environments. When NIST 800-171 was published, it specified a grace period that ends on December 31, 2017. Therefore, compliance with the framework is mandatory beginning in 2018.

Thank You & Closing Thoughts

- Cloud-based research is a team sport!
 - Researchers
 - IT/Research IT staff
 - Security professionals
 - Financial/administrative support
- **Many thanks** to the NIH Cloud Services (aka STRIDES) team and our supporting cast, to ODSS, and to the meeting organizers, moderators, & scribes
- *And thanks to you all for coming and participating in this workshop!*



Questions?

Nick Weber

nick.weber@nih.gov

Supplementary Slides

Availability of the entire Sequence Read Archive (SRA) on AWS & GCP



- **36.4 PB of public and controlled-access SRA data**, hosted by the National center for Biotechnology Information at the National Library of Medicine, is now available on Amazon Web Services (AWS) and Google Cloud Platform (GCP).
- Simplifying computational access to these data in conjunction with collaborative partnerships with open source & data analysis platforms **accelerates genomics research and discovery in the management of COVID-19 and beyond.**

KEY OUTCOMES OF USING CLOUD SERVICES AND TOOLS:

- A mechanism for **faster access to vital large datasets**
- Ability to **share data easily from a central location**
- **Availability of compute resources and access to data** for researchers
- **Reproducibility** of analytical processes and datasets generated

Problem: While public sequence data represents a major opportunity for viral discovery, its exploration has been inhibited by a lack of efficient methods for searching this petabyte-scale data which is growing exponentially!

Solution: Serratus- a new cloud computing architecture tailored for ultra-high throughput sequence alignment at the petabase scale! The **goal** of Serratus was **to identify and share every coronavirus sequence from over 10 years of data collected by the global research community.**

Outcome: Identification of **tens of thousands of coronavirus and coronavirus-like viral alignments and family identifications made freely available to the research community to catalyze a new era of viral discovery.**



We can now do this in 3-4 days instead of 12+ months directly **as a result of the SRA data being available in the cloud.** This means we can share this data with the CoV researchers today, when it can make a difference, not a year from now. This is **important for COVID-19 now and will be important in response to the next pandemic.**



– **Artem Babaian**, Lead Developer at Serratus and corresponding author for publication.

Additional Cloud Benefits

- Renting (OpEx) is more flexible than buying (CapEx) — though it can also be more expensive, at least at the *unit level*
- *Overall*, cloud can significantly reduce cost, e.g., with automated failover, tiered storage / storage lifecycle automation, spinning up/shutting down instances on demand, use of reserved and spot/preemptible instances, etc.
- There is a reduced reliance local data center resources and staffing, with ability to shift focus to more mission-centric activities
- Software-defined nature of cloud (infrastructure as code) → DevSecOps practices → increased process consistency / software quality for organizations
- Cloud offers robust and expanding set of security tools used by many organizations, and risk is shared between provider and user

What **STRIDES** is versus What **STRIDES** is Not

There are many benefits that come with enrollment within STRIDES, however, we will discuss what is not within the scope of the Initiative

STRIDES enables biomedical research, providing a vast number of services

STRIDES **does not** directly **make decisions** for your program; however, **it does provide support** for decision making

<i>STRIDES is ...</i>	<i>STRIDES is not...</i>
<ul style="list-style-type: none">• An NIH program, and part of NIH’s data science portfolio• A mechanism for NIH and NIH-funded researchers to access and use cloud compute, storage, and related services• One method for using the cloud to support biomedical research• Encouraged by NIH	<ul style="list-style-type: none">• A destination (i.e., there is no “STRIDES cloud”)• A service for researchers to store or analyze research data• The only method for using the cloud to support biomedical research• Required by NIH