

# NIH IT Ecosystem

---

**Susan Gregurick, Ph.D.**

**Associate Director for Data Science**

**Office of Data Science Strategy**

*December 13<sup>th</sup>, 2019*



**National Institutes of Health**  
*Office of Data Science Strategy*

# Creating an NIH IT Ecosystem

---

- What are some of the **challenges** that our researchers face?
- What are we **doing now** to make it easier to access NIH data resources?
  - STRIDES
- What **else can we do** to help researchers use data & tools across NIH data platforms?
  - Single sign-on

# Genetic & dietary effects in COPD



Icon made by Roundicons from [www.flaticon.com](http://www.flaticon.com)

Chronic Obstructive Pulmonary Disease (COPD) is a significant cause of death in the US, genetic and dietary data are available that could be used to further understand their effects on the disease

Separate studies have been done to collect genomic and dietary data for subjects with COPD.

**Researchers know that many of the same subjects participated in the two studies.**

Linking these datasets together would allow them to examine the combined effects of genetics and diet using the subjects present in both studies. However, different identifiers were used to identify the subjects in the different studies.

## Challenges

**Obtaining access** to all the relevant datasets so they can be analyzed

**Understanding consent for each study** to ensure that data usage limitations are respected

**Connecting data from the same subject across different datasets** so that the genetic and dietary data from the same subjects can be linked and studied

# Pediatric Oncology



Image by mcmurryjulie from Pixabay

Rare diseases like pediatric cancers are especially challenging because no single source has enough data to allow identification of causative variants on their own

Being able to aggregate as much data as possible is a core requirement for rare disease studies due to the limited number of patients that exhibit a given disease.

These use cases touch on some of the most challenging aspects of privacy, security and ethics, as well as the technical components of authentication, authorization, data management and more.

## Challenges

**Securely querying data across multiple resources** is a prerequisite to allow relevant datasets to be identified

**Applying for, and managing data access to multiple datasets housed at different resources** can be remarkably challenging.

**Time is of the essence** for many clinical applications but is particularly acute in a pediatric setting

# Cardiovascular Genomics



Icon made by Roundicons from [www.flaticon.com](http://www.flaticon.com)

Researchers investigating the genetic components of cardiovascular disease need to integrate data from multiple repositories to inform their studies

NHLBI's Trans-Omics for Precision Medicine (TOPMed) program has extensive data relating to Heart, Lung, Blood and Sleep phenotypes from ~144,000 participants from over 80 different studies. This data is available through the NHLBI DataSTAGE platform.

**Sequence data for the same subjects is available via NHGRI's AnVIL platform, however, the two platforms do not talk to each other...**

## Challenges

**Accessing data** so that more people can participate

**Data Access** for users who do not belong to traditional academic research organizations

**Combining data** housed in two separate data repositories

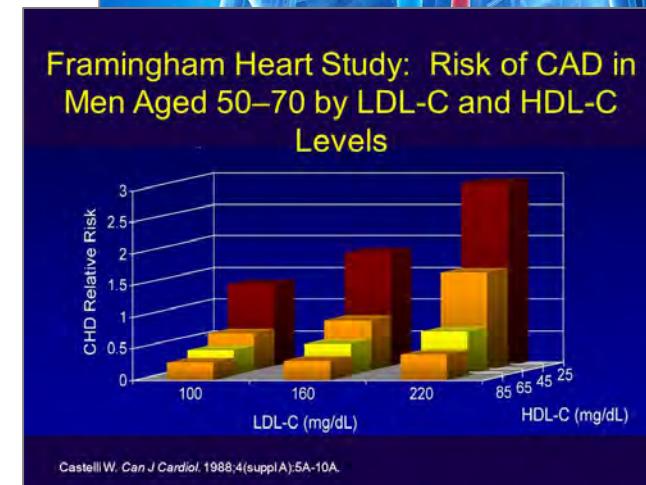
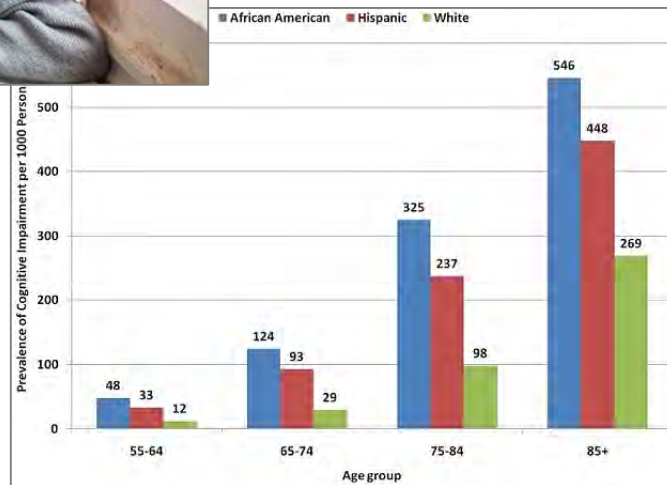
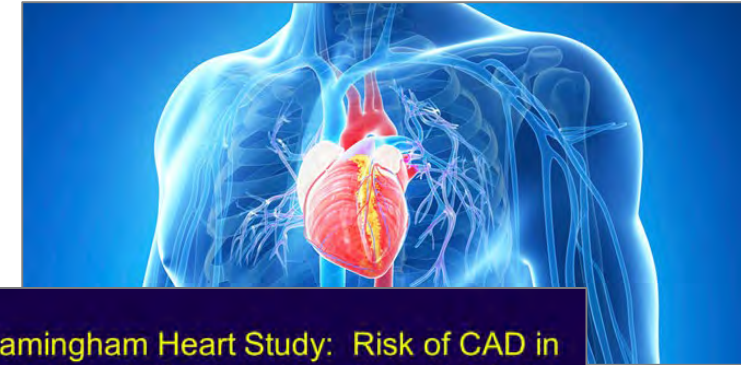
# Common Themes From Researcher Stories

---

- Researchers know that many of **the same subjects participated in the different NIH studies**, but it is extremely time consuming to find and link that data across our platforms.
- **Securely querying data across multiple resources** is a prerequisite to allow relevant datasets to be identified.
- **Combining the same type of data** housed in separate data repositories increases the power of the data analysis.
- **Harmonizing data** from multiple sources so that it can be integrated and analyzed together is still a challenge.

IMAGINE...

the ability to link data in the Framingham Heart Study (NHLBI) with Alzheimer's health data (NIA) to understand correlative effects in cardiovascular health with aging and dementia.



# Harnessing the Power of the Cloud for Biomedical Research

---

Cloud computing offers multiple opportunities NIH can leverage to advance biomedical research, including:

- Computation on biomedical data at an **unprecedented scale**
- Broad access to **cutting-edge cloud technology** with, for example, industry-leading **security** tools
- Storage of **large, diverse data** in a way that enables easier sharing, access, and reuse of data with other researchers
- A **community-driven approach** to data science that breaks down disciplinary silos
- Adopt and develop **cloud-based tools** from industry or academia for biomedical research



# The STRIDES Initiative: Enabling NIH Enterprise-Wide Access to the Cloud

---

- The STRIDES (Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability) Initiative establishes innovative partnerships with cloud service providers to offer NIH and NIH-funded institutions:
- **Discounts** on computing, storage, and related cloud services.
  - **Access to tailored professional cloud consultations and cloud technical support services** from the STRIDES Initiative partners.
  - **Discounted in-person and online training** for researchers, data owners, and others interested in learning about cloud computing. Some of the online course offerings are free.
  - **Opportunities to explore methods and approaches** that may advance biomedical research through collaborations.
  - **Access to emerging technologies, (e.g. machine learning and artificial intelligence)** for data preparation and analysis.

# The STRIDES Initiative: Enabling NIH Enterprise-Wide Access to the Cloud (continued)

---

- Working with targeted institutions in a pilot approach, seeking to streamline and automate onboarding process
- Institutions continue to handle day-to-day ownership and management of research data and financial responsibilities

# The STRIDES Initiative: Catalyzing NIH Dataset Migration to the Cloud

- NHLBI Framingham Heart Study
  - All of Us Research Program
  - NCI Cancer Research Data Commons
  - NCBI data resources (12 PB!)
  - NHLBI Trans-Omics for Precision Medicine (TOPMed) Program
  - Gabriella Miller Kid's First
  - NHGRI AnVIL
- And Many More**
- **Moved over 30 PB of data into Google and AWS**
    - Largest biomedical data set available for biomedical research
    - 1 PB is equivalent to over 4,000 digital photos per day, over your entire life
  - Next year we anticipate up to 50 PB of data in the cloud

# The STRIDES Initiative: Next Steps to Expand Access and Impact

---

## Need to:

- Lower costs for NIH-funded institutions
- Increase training in cloud computing
- Add additional STRIDES partners
- Common controls and protection standards

## End goal Researchers should be able to:

- Seamless STRIDES accounts through universities and institutions
- Increase researcher data science skills
- Greater analytic and data management capabilities for NIH funded programs
- Greater assurance in the confidentiality, integrity, and availability of data

# NIH's Data Platforms are Rich

The image displays a collage of NIH data platforms. At the top left is the **NIH NATIONAL CANCER INSTITUTE GDC Data Portal** with navigation links for Home, Projects, Exploration, Analysis, and Repository. Below it is the **Genomic Data Commons Data Portal** with a search bar and filters for Projects, Exploration, and Analysis. In the center is the **All of Us RESEARCH PROGRAM** dashboard. To the right is the **NIH BioData CATALYST** logo, with the NIH logo and the text "National Heart, Lung, and Blood Institute". Below the All of Us logo is the **Kids First** Data Resource Center, featuring a navigation menu and a search bar. The main content of the Kids First page is titled "About the Research {Childhood Cancer + Structural Birth Defects}" and includes a statistics bar with the following data:

10 Studies	8,018 Participants	2,831 Families	11,700 Samples	34,471 Files	904.37 TB Size
------------	--------------------	----------------	----------------	--------------	----------------

Below the statistics bar is a section titled "Studies" with a description: "The data in the Kids First Data Resource Portal is a collection of datasets from various investigators who are performing disease-specific research. Each of these datasets originally were part of separate research studies".

On the right side of the collage, there is a snippet of a webpage with a "Learn more" button and a "Page 1 of 1" indicator. Below this is a section titled "Genomic Analysis, Visualization, Informatics space (Ar)" with a large red icon of a stack of papers.

# NIH Data Platforms

---

- Portals that **provide intuitive interface** to data, computational, and analytical resources
- **Access to high value biomedical data** spanning multi-data domains and disease areas
- Rich suites of **computational resources** and tools to explore, analyze, and visualize data
- Individual and group **workspaces to enable researchers** to upload or access data, create experiments and conduct analysis, and store or share results
- Approaches for data access that are fit for purposes and that ensure **data remain safe, secure and private**

# Create greater interoperability between NIH supported cloud based, high value data platforms

---

## Need to:

- Standardizing aspects of the User Experiences (e.g. buttons, overall layout, etc)
- Uniform, efficient data access
- Data & patient identifiers that are resolvable across platforms
- Ability to search and explore data across platforms
- Allow for workflows across platforms and across cloud service providers

## End goal Researchers should be able to:

- Data access quicker and easier
- Find all data related to a patient, study or data collection
- Data and metadata harmonized in such a way that it can be integrated and analyzed
- Analyze data using tools across platforms and across clouds

# NIH Researcher Auth Service: Towards Single 'Sign-on' Across NIH Data Resources

---

---

Streamline login for authorization of controlled-access data

---

Make use of industry standard technology (web tokens)

---

Enforce multi-factor authentication for security

---

Keep flexible for different NIH needs: 'do no harm to existing systems'

---

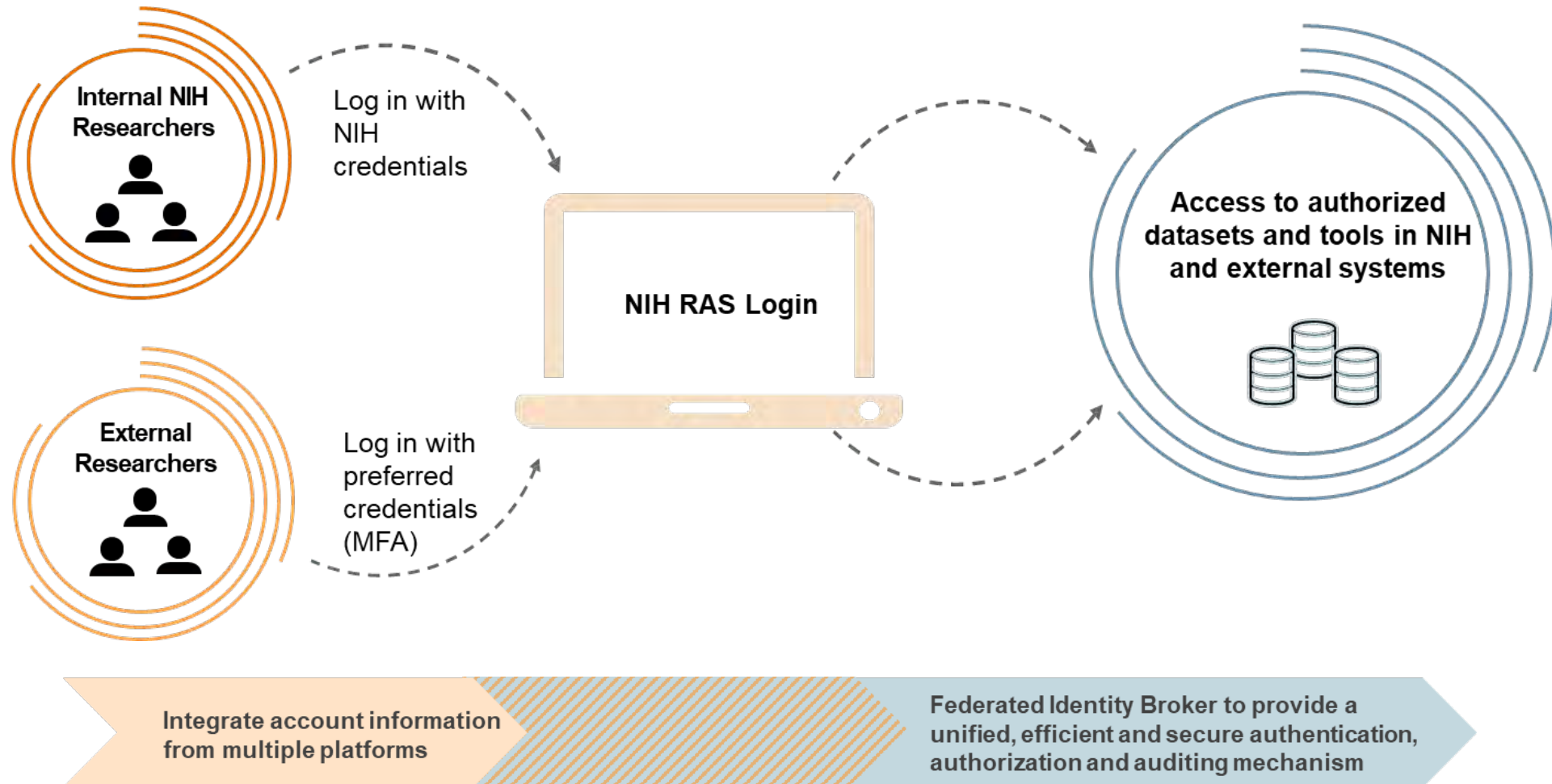


## End goal:

NIH-wide system for a consistent method to access data across NIH data resources



# NIH Researcher Auth Service: High-level Workflow



# NIH Researcher Auth Service: First Partner Systems



**NCBI Database of Genotypes  
and Phenotypes (dbGaP)**



**NHGRI Analysis,  
Visualization and Informatics  
Lab-space (AnVIL)**



**Common Fund Kids First  
Data Resource Center**



**NHLBI Biodata Catalyst  
(formerly DataSTAGE)**



**All of Us (AoU)**

**NATIONAL CANCER INSTITUTE  
GDC Data Portal**

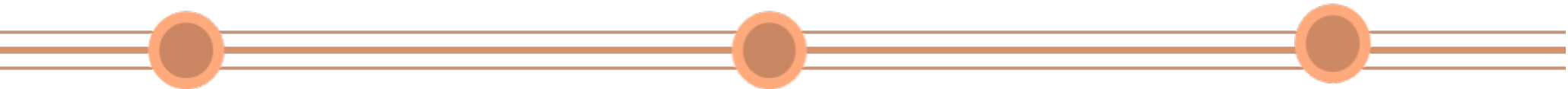
**NCI Cancer Research Data  
Commons (CRDC)**



**NIMH Data Archive (NDA)**

# NIH Researcher Auth Service: First Use Cases

---



Log into an analytic platform one time, to seamlessly analyze data stored in one or multiple repositories

Log into a system with ORCID credentials, to access data requested with an eRA Commons ID

Access audit logs for a given dataset, to rapidly respond to a data management incident

# NIH Researcher Auth Service: First Integration Milestone

---

- Globus customers can login in with eRA Commons accounts to access Globus data transfer services.
- This integration uses OpenID Connect (OIDC), the modern standard to exchange access information between systems.
- This integration represents a foundational part of the RAS project.
- The OIDC service can be rapidly **adopted and extended** to support other integration partners (e.g., Google, Terra, and Login.gov).



# Creating an NIH IT Ecosystem: Important Goals for 2020



## Enhancing the STRIDES Initiative

- Additional Partners
- Enhance cloud-based workflows & analysis

## Improving Researcher Experience

- Easier login to NIH resources
- Improve access to data
- Improve federated data search

## Maintaining Security and Assurances

- Standard audit & trace logs
- Adhere to and adopt community standards

# Special Thanks

---

**STRIDES: Andrea Norris, Nick Weber**, Tom Shaw, James Davis, Nigel Horne, **Todd Reilly**, Antej Nuhanovic, Matt Gieseke, Joel Peterson, Valerie Virta, Sherika Wynter, Michelle Speir

**Researcher Authentication Services:** Regina Bures, Ishwar Chandramouliswaran, Tanja Davidsen, Valentina Di Francesco, **Jeff Erickson**, Tram Huyen, **Rebecca Rosen**, Steve Sherry, Alastair Thomson, Greg Farber, Dylan Klomparens, Charles Schmitt, Susan, Wright, Ken Wiley, Kristofor Langlais, James Coulomb, Lora Kutkat, Nick Weber, Deloitte and BioTeam