NIH’s Strategic Vision for Data Science: Enabling a FAIR-Data Ecosystem

Susan Gregurick, Ph.D.

Senior Advisor

Office of Data Science Strategy

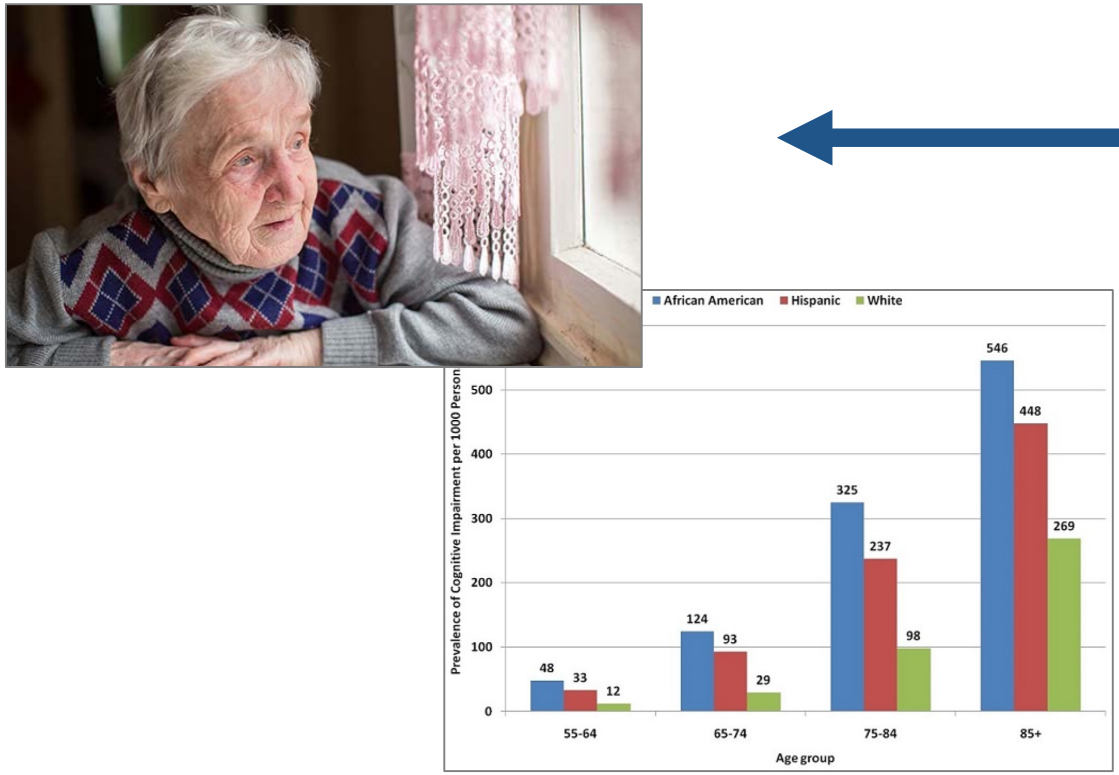
May 21, 2019

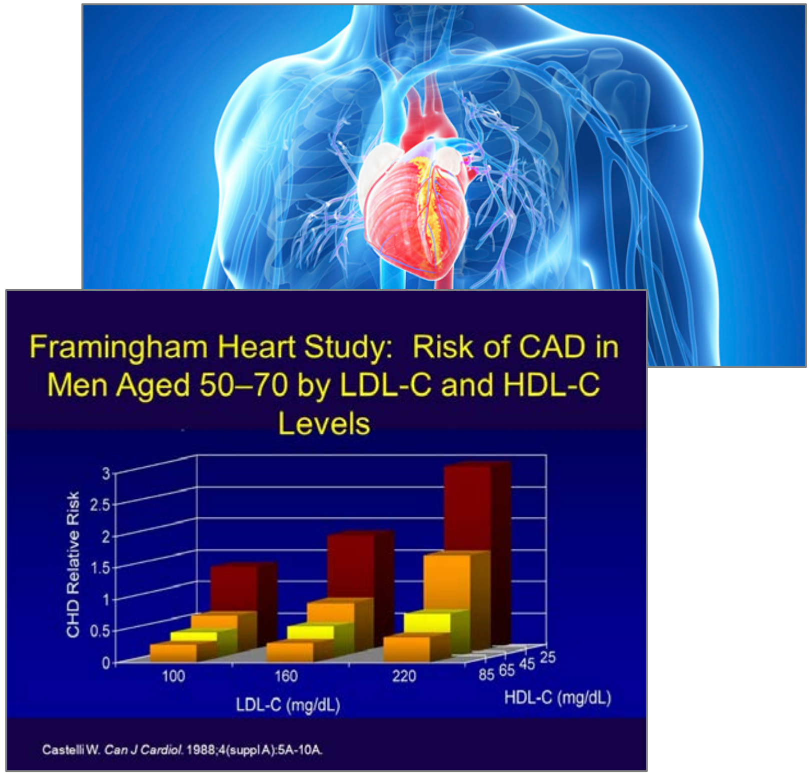
VISION

a **modernized, integrated, FAIR** biomedical data ecosystem

IMAGINE…

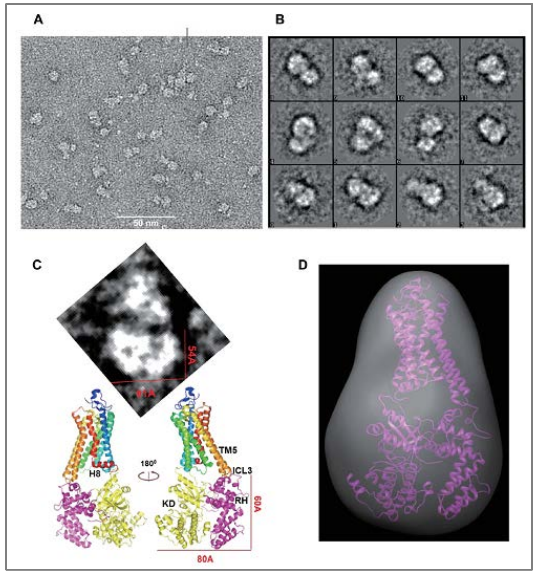
the ability to link data in the Framingham Heart Study (NHLBI) with Alzheimer’s health data (NIA) to understand correlative effects in cardiovascular health with aging and dementia.



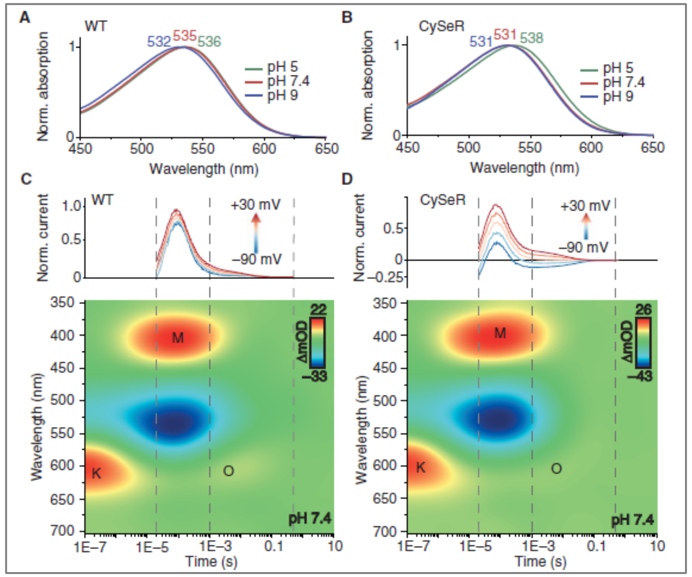

Framingham Heart Study: Risk of CAD in Men Aged 50–70 by LDL-C and HDL-C Levels

Castelli W. *Can J Cardiol.* 1988;4(suppl A):5A-10A.

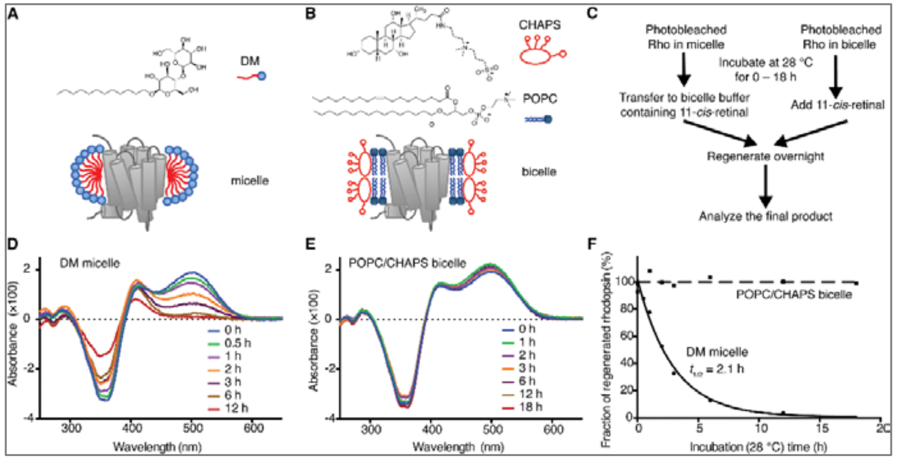
IMAGINE… the ability to quickly obtain access to data, and related information, from published articles.



Negative stain EM reveals the principal architecture of the rhodopsin/GRK5 complex. (Image by Van Andel Research Institute)



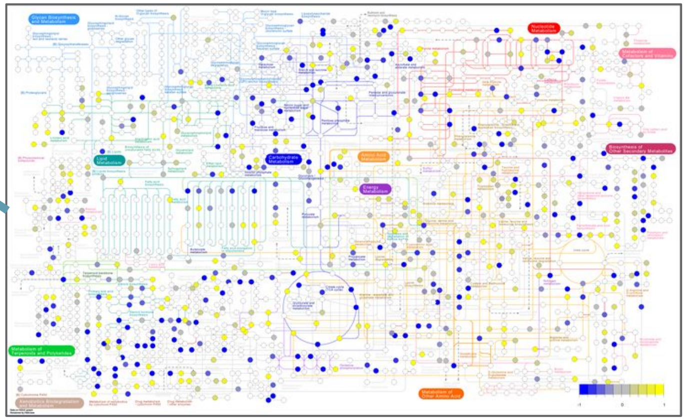
Absorption spectra of purified CsR-WT (A) and CySeR (B) at pH 5 (green), pH 7.4 (red), and pH 9 (blue). R. Fudim, e al, Science Signaling, 2019



Energetics of Chromophore Binding in the Visual Photoreceptor of Rhodopsin, H. Tian et al, Biophysical Journal, 2017.

IMAGINE… the new capabilities that artificial intelligence and advanced technologies offer medical research, treatment, and prevention.

IMAGINE… the ability to link electronic health care records with personal data and with clinical and basic research data.



This is the promise of the *NIH Strategic Plan for Data Science*

…and here’s how we will get there.

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

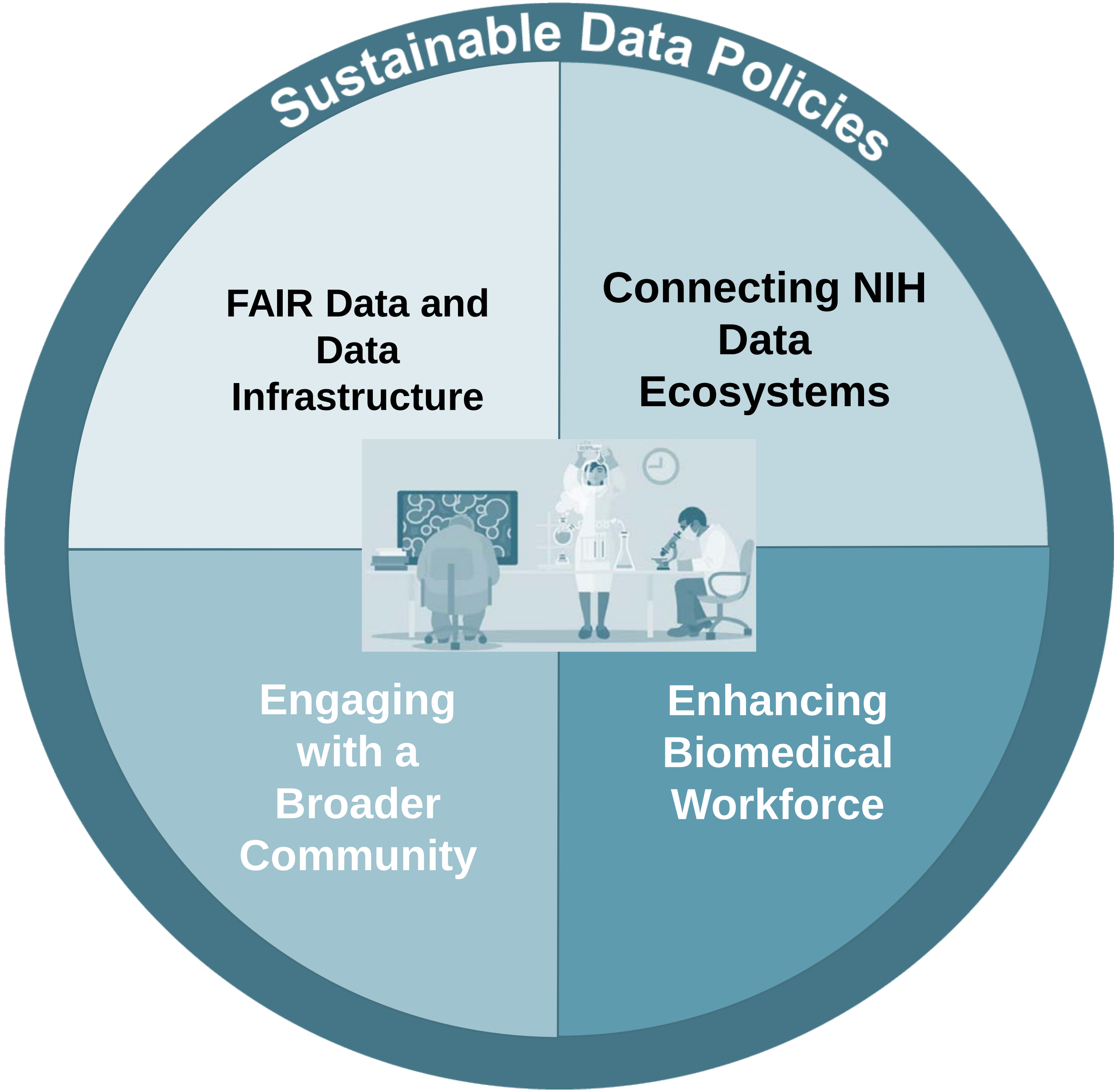
- Enhance the NIH data science workforce
- Expand the national research workforce
- Engage a broader community

Stewardship and Sustainability

- Develop policies for a FAIR data ecosystem
- Enhance stewardship

Strategic Plan for Data Science: Goals and Objectives

- ***FAIR Data and Data Infrastructure***
- ***Connecting NIH Data Ecosystems***
- ***Engaging with a Broader Community***
- ***Enhancing Biomedical Workforce***
- ***Sustainable Data Policies***



Sustainable Data Policies

FAIR Data and Data Infrastructure

Connecting NIH Data Ecosystems

Engaging with a Broader Community

Enhancing Biomedical Workforce

New: Office of Data Science Strategy

The NIH **Office of Data Science Strategy (ODSS)** in the Office of the Director:

- Provides leadership and coordination on the strategic plan for data science, in collaboration with the ICOs.
- Helps develop and implement NIH’s vision for a **modernized** and **integrated** biomedical data ecosystem.
- Develops a diverse and talented data science workforce.
- Coordinates with trans-NIH governance committees.
- In coordination with the CIO, builds strategic partnerships to develop and disseminate advanced technologies and methods.

Implementation Progress:
Oct. 2018 – Present

- **FAIR Data and Data Infrastructure**
- Sustainable Data Policies
- Connecting NIH Data Ecosystems
- Engaging with a Broader Community
- Enhancing Biomedical Workforce

Making Data *FAIR*

Findable	□ must have unique identifiers, effectively labeling it within searchable resources.
Accessible	□ must be easily retrievable via open systems and effective and secure authentication and authorization procedures.
Interoperable	□ should “use and speak the same language” via use of standardized vocabularies.
Reusable	□ must be adequately described to a new user, have clear information about data-usage licenses, and have a traceable “owner’s manual,” or provenance.

Overview of Sharing Publication and Related Data

NIH strongly encourages
open access Data Sharing Repositories
as a first choice.

https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

Options of scaled implementation for sharing datasets

Datasets up to **2 gigabytes**

PubMed Central

- PMC stores publication-related supplemental materials and datasets directly associated publications. Up to 2 GB.
- Generate Unique Identifiers for the stored supplementary materials and datasets.

Datasets up to **20*gigabytes**

Use of commercial and non-profit repositories

- Assign Unique Identifiers to datasets associated with publications and link to PubMed.
- Store and manage datasets associated with publication, up to 20* GB.

High Priority Datasets **petabytes**

STRIDES Cloud Partners

- Store and manage large scale, high priority NIH datasets. (Partnership with STRIDES)
- Assign Unique Identifiers, implement authentication, authorization and access control.

The TRUST Principles for Data Repositories

Transparency

- is achieved by providing publicly accessible evidence of the services that a repository can and can not offer.

Responsibility

- is a commitment to provide high technical quality data services.

User community

- is the focus on the uses and potential uses of the data and services offered.

Sustainability

- is the capability to support long-term data preservation and use.

Technology

- is the infrastructure and capabilities to support the repository operations.

Develop Characteristics for Open Access Data Sharing Repositories

Trans-NIH
BioMedical Informatics Coordinating Committee (BMIC)

- Characteristics drafted, includes provisions for repositories with human data
- Developed and reviewed in trans-NIH process
- Planned Community Input: Request for Information (RFI)

Develop criteria for open-access
NIH data sharing repositories

Optimized Funding for NIH Data Repositories and Knowledgebases

- Data resources are important research tools
- Historically funded through research grants
- Funding mechanism should be optimal for type of resource
- **End goal:** researcher confident in data and information integrity

- **Solution: New Funding Announcement** for data repositories and knowledgebases
- Resource plan requirement

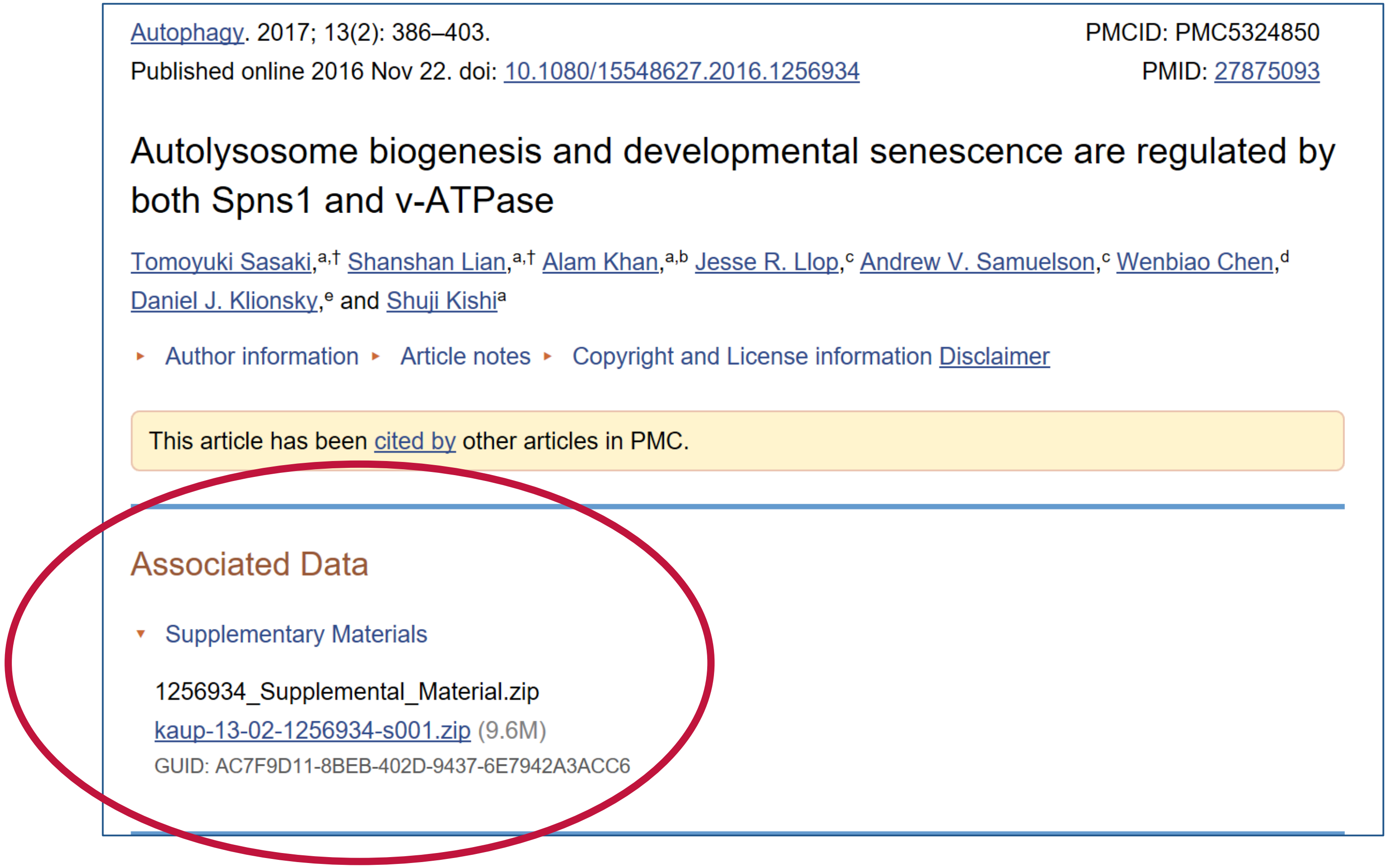
Scientific Impact

1.Community Engagement

1.Quality of Data and Services and Efficiency of Operations

Governance

Sharing Datasets as Supplementary Materials



Autophagy. 2017; 13(2): 386–403.
Published online 2016 Nov 22. doi: 10.1080/15548627.2016.1256934

PMCID: PMC5324850
PMID: 27875093

Autolysosome biogenesis and developmental senescence are regulated by both Spns1 and v-ATPase

Tomoyuki Sasaki,[a,†] Shanshan Lian,[a,†] Alam Khan,[a,b] Jesse R. Llop,[c] Andrew V. Samuelson,[c] Wenbiao Chen,[d] Daniel J. Klionsky,[e] and Shuji Kishi[a]

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer

This article has been cited by other articles in PMC.

Associated Data

▾ Supplementary Materials

1256934_Supplemental_Material.zip
kaup-13-02-1256934-s001.zip (9.6M)
GUID: AC7F9D11-8BEB-402D-9437-6E7942A3ACC6

✓ FAIR Data: Linking datasets to publications (PubMed)

Piloting a Repository to Make Research Data Citable, Sharable, and Discoverable Using Figshare

Data is openly accessible

Documented with customizable, discipline-specific metadata

Authors can link grant information to data

All data is associated with a license

Self-publish any data type in any file format

Assign institutionally (NIH) branded DOI

Indexed in Google and discoverable across search engines

Ability to embargo data assets

Usage metrics tracked openly

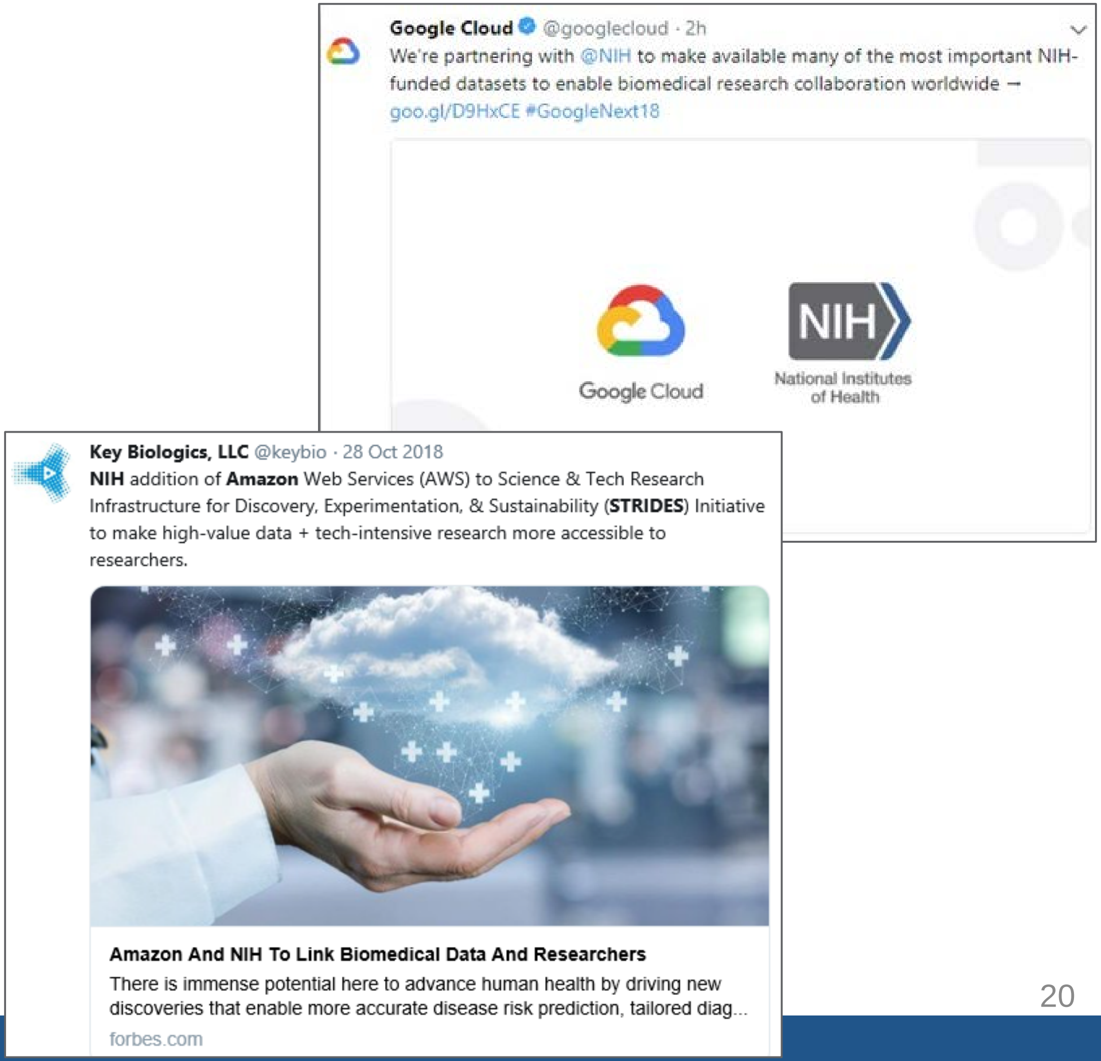
FAIR implementation

✓ Providing FAIR-enabled, open-access options for datasets

Science & Tech Research Infrastructure for Discovery, Experimentation and Sustainability Initiative

- First **STRIDES** agreement: Google Cloud (July 2018)
- Second **STRIDES** agreement: Amazon Web Services (Oct. 2018)
- Other Transaction mechanism
- Additional partnerships anticipated

https://datascience.nih.gov/strides



Google Cloud @googlecloud · 2h

We're partnering with @NIH to make available many of the most important NIH-funded datasets to enable biomedical research collaboration worldwide → goo.gl/D9HxCE #GoogleNext18

Key Biologics, LLC @keybio · 28 Oct 2018

NIH addition of **Amazon** Web Services (AWS) to Science & Tech Research Infrastructure for Discovery, Experimentation, & Sustainability (**STRIDES**) Initiative to make high-value data + tech-intensive research more accessible to researchers.

Amazon And NIH To Link Biomedical Data And Researchers

There is immense potential here to advance human health by driving new discoveries that enable more accurate disease risk prediction, tailored diag...

forbes.com

✓ FAIR Data: Move/Access to high priority data sets in cloud service providers

Overview of Sharing Publication and Related Data

NIH strongly encourages
open access Data Sharing Repositories
as a first choice.

https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

Options of scaled implementation for sharing datasets

Datasets up to **2 gigabytes**

PubMed Central

- PMC stores publication-related supplemental materials and datasets directly associated publications. Up to 2 GB.
- Generate Unique Identifiers for the stored supplementary materials and datasets.

Datasets up to **20*gigabytes**

Use of commercial and non-profit repositories

- Assign Unique Identifiers to datasets associated with publications and link to PubMed.
- Store and manage datasets associated with publication, up to 20* GB.

High Priority Datasets **petabytes**

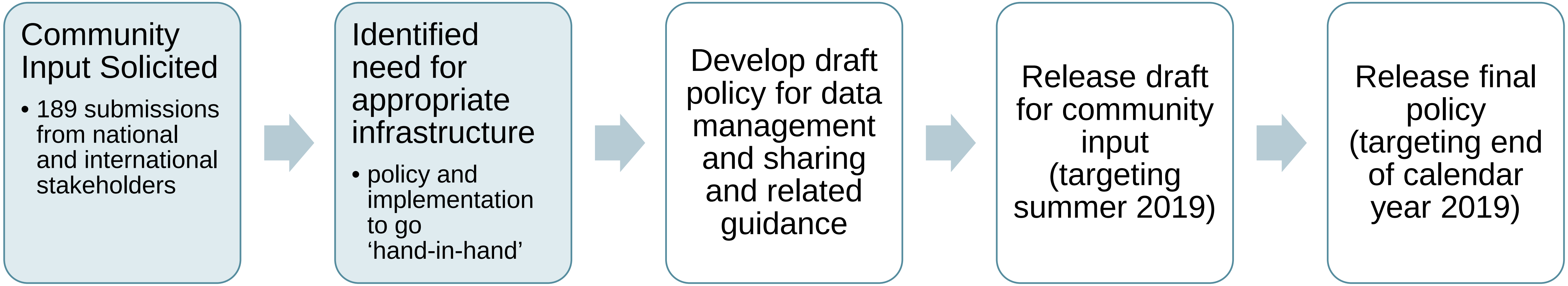
STRIDES Cloud Partners

- Store and manage large scale, high priority NIH datasets. (Partnership with STRIDES)
- Assign Unique Identifiers, implement authentication, authorization and access control.

Implementation Progress: Oct. 2018 – Present

- FAIR Data and Data Infrastructure
- **Sustainable Data Policies**
- Connecting NIH Data Ecosystems
- Engaging with a Broader Community
- Enhancing Biomedical Workforce

Planning for an NIH Data Management and Sharing Policy



Community Input Solicited

- 189 submissions from national and international stakeholders

Identified need for appropriate infrastructure

- policy and implementation to go ‘hand-in-hand’

Develop draft policy for data management and sharing and related guidance

Release draft for community input (targeting summer 2019)

Release final policy (targeting end of calendar year 2019)

✓ Sustainable data management and sharing policy

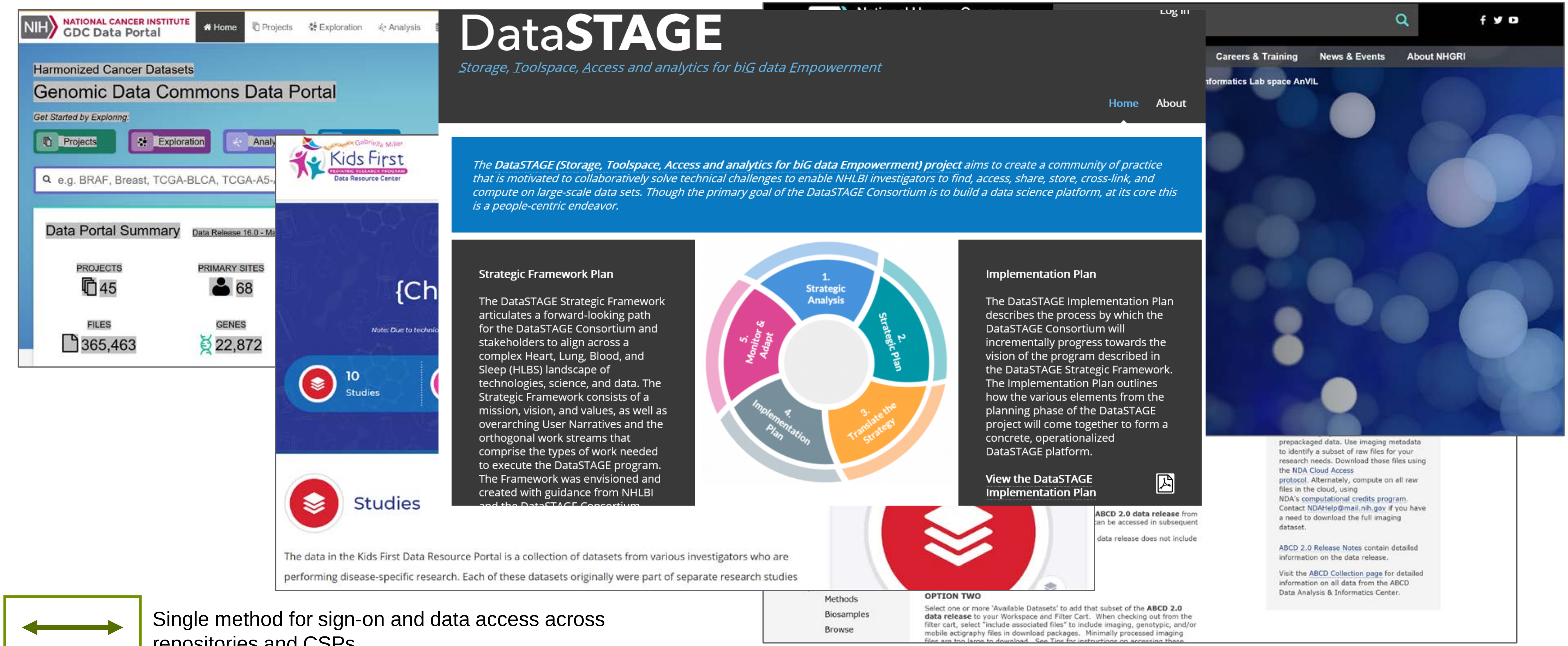
Implementation Progress:
Oct. 2018 – Present

- FAIR Data and Data Infrastructure
- Sustainable Data Policies
- **Connecting NIH Data Ecosystems**
- Engaging with a Broader Community
- Enhancing Biomedical Workforce

Examples of Datasets Moving to the STRIDES Cloud

- NHLBI Framingham Heart Study
- All of Us Research Program
- NCI Genomic Data Commons
- NCBI data resources
- NHLBI Trans-Omics for Precision Medicine (TOPMed) Program
- NCI Proteomics Data Commons and Imaging Data Commons
- NIMH Data Archive
- Gabriella Miller Kids First Pediatric Research Program
- Transformative CryoEM Program
- **And many others!**

NIH’s Data Environments are Rich, but Siloed



NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home · Projects · Exploration · Analysis

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects · Exploration · Analy

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-

Data Portal Summary Data Release 16.0 - Ma

PROJECTS	PRIMARY SITES
45	68
FILES	**GENES**
365,463	22,872

Gabriella Miller Kids First Pediatric Research Program Data Resource Center

{Ch

Note: Due to technic

10 Studies

Studies

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

DataSTAGE

Storage, Toolspace, Access and analytics for biG data Empowerment

Log in

Home · About

The DataSTAGE (Storage, Toolspace, Access and analytics for biG data Empowerment) project aims to create a community of practice that is motivated to collaboratively solve technical challenges to enable NHLBI investigators to find, access, share, store, cross-link, and compute on large-scale data sets. Though the primary goal of the DataSTAGE Consortium is to build a data science platform, at its core this is a people-centric endeavor.

Strategic Framework Plan

The DataSTAGE Strategic Framework articulates a forward-looking path for the DataSTAGE Consortium and stakeholders to align across a complex Heart, Lung, Blood, and Sleep (HLBS) landscape of technologies, science, and data. The Strategic Framework consists of a mission, vision, and values, as well as overarching User Narratives and the orthogonal work streams that comprise the types of work needed to execute the DataSTAGE program. The Framework was envisioned and created with guidance from NHLBI and the DataSTAGE Consortium

Implementation Plan

The DataSTAGE Implementation Plan describes the process by which the DataSTAGE Consortium will incrementally progress towards the vision of the program described in the DataSTAGE Strategic Framework. The Implementation Plan outlines how the various elements from the planning phase of the DataSTAGE project will come together to form a concrete, operationalized DataSTAGE platform.

View the DataSTAGE Implementation Plan

National Human Genome

Careers & Training · News & Events · About NHGRI

nformatics Lab space AnVIL

prepackaged data. Use imaging metadata to identify a subset of raw files for your research needs. Download those files using the NDA Cloud Access protocol. Alternately, compute on all raw files in the cloud, using NDA’s computational credits program. Contact NDAHelp@mail.nih.gov if you have a need to download the full imaging dataset.

ABCD 2.0 Release Notes contain detailed information on the data release.

Visit the ABCD Collection page for detailed information on all data from the ABCD Data Analysis & Informatics Center.

ABCD 2.0 data release from
an be accessed in subsequent

data release does not include

Methods
Biosamples
Browse

OPTION TWO

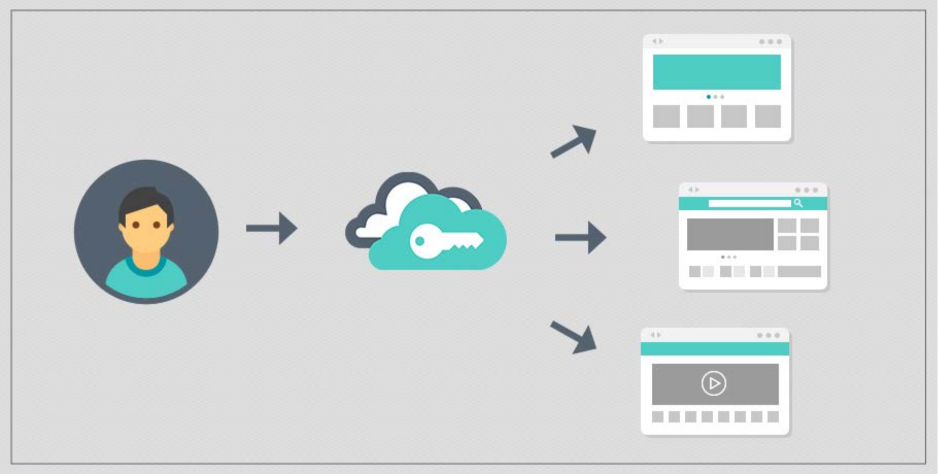
Select one or more ‘Available Datasets’ to add that subset of the **ABCD 2.0 data release** to your Workspace and Filter Cart. When checking out from the filter cart, select "include associated files" to include imaging, genotypic, and/or mobile actigraphy files in download packages. Minimally processed imaging files are too large to download. See Tips for instructions on accessing these

↔ Single method for sign-on and data access across repositories and CSPs

Single ‘Sign-on’ Across NIH Data Resources

- Streamlined login for authorization of controlled-access data
- Make use of industry standard technology (web tokens)
- 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

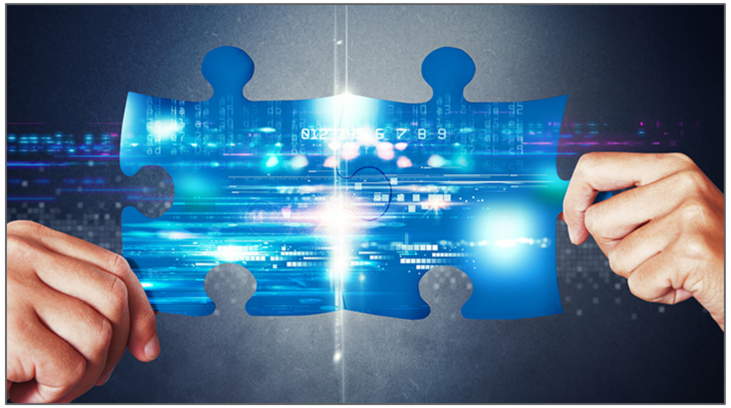


Implementation Progress:
Oct. 2018 – Present

- FAIR Data and Data Infrastructure
- Sustainable Data Policies
- Connecting NIH Data Ecosystems
- **Engaging with a Broader Community**
- Enhancing Biomedical Workforce

Leverage, Develop, and Extend Methods and Tools from Broader Communities

- Partner with other federal agencies (e.g. National Science Foundation) on data science activities
- Leverage SBIR/STTR to bring in industry expertise
- Engage a broader community through codeathons, citizen science, and challenges
- Improve software sustainability, efficiency and utility



COMING SOON

✓ Engaging a broader community

Implementation Progress: Oct. 2018 – Present

- FAIR Data and Data Infrastructure
- Sustainable Data Policies
- Connecting NIH Data Ecosystems
- Engaging with a Broader Community
- **Enhancing Biomedical Workforce**

Enhance the Biomedical Workforce

Coding it Forward

CIVIC DIGITAL FELLOWSHIP

- 10 undergraduate fellows for 2019 placed in admin or funding offices for 10-week summer program
- Student-led non-profit places tech-savvy students in federal agencies
- ODSS will coordinate on-campus networking opportunities for fellows

https://www.codingitforward.com/

Graduate Data Science Summer Program

- 13 master’s-level interns for 2019
- Pilot driven by discussion with local universities consortium
  - UVA, George Mason, George Washington, UMD, University of Delaware/Georgetown, Johns Hopkins
- Open to students from any university

https://www.training.nih.gov/data_science_summer

NIH Data Science Senior Fellowships

- One- or two-year **national service sabbatical** in high-impact NIH programs
- Seeking **data science and technology** experts
- Work with large volumes of biomedical research data, impact public health, gain policy exposure

- Expecting 5+ fellows in first cohort, starting late 2019
- Program evaluation in 2024

COMING SOON

Improve Data Science-Related Training through T Grants, F and K Awards

- Expand expectations for development of quantitative and computational skills for students and postdoctoral fellows supported by NIH training (T) grants
  - NIGMS T32/T34, Neuroscience T32, or NLM T15 FOAs
- Disseminate across all training mechanism and ICs

- Launch data science-focused training programs in specific biomedical research areas of high need
  - Biomedical behavioral and social science
  - Neuroscience

  ✓ RFA-OD-19-011

Improving R&R and RCR and Evaluating Efficacy of Interventions

- Support development of training modules to fill in gaps in rigor and reproducibility in data science

- Support training modules on responsible conduct of research in data science

- Improvement and expansion of K25 program (Mentored Quantitative Research Development – PA-18-396)

COMING SOON

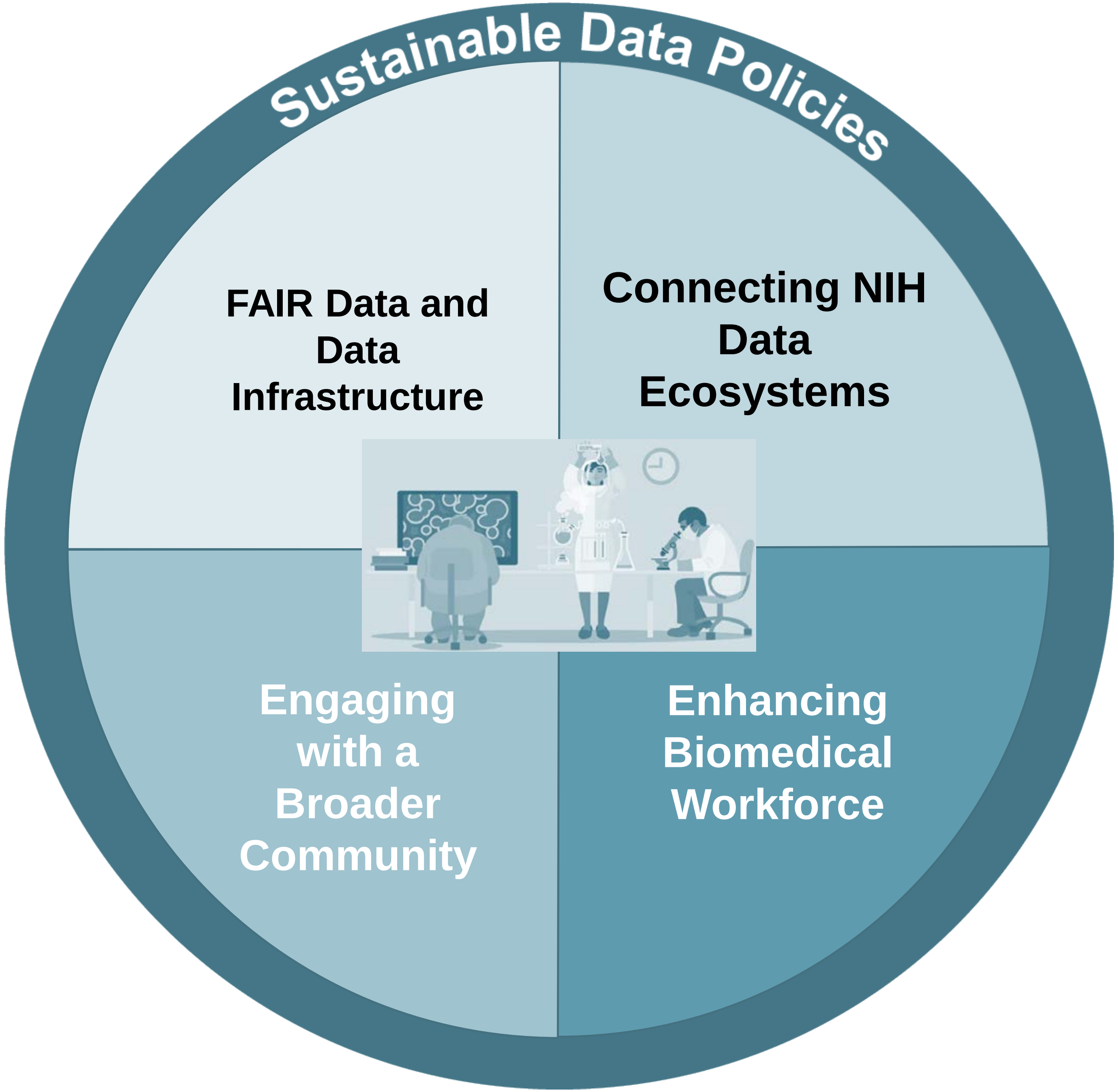
Building Diversity in Biomedical Data Science

- Boot camps/short training programs for diverse cohorts of biomedical research trainees
  - Includes STRIDES CSP and professional societies

- Increase emphasis on quantitative and computational skills development in existing diversity programs
  - E.g. new language in NIGMS FOAs (already in place)

VISION

a **modernized, integrated, FAIR** biomedical data ecosystem



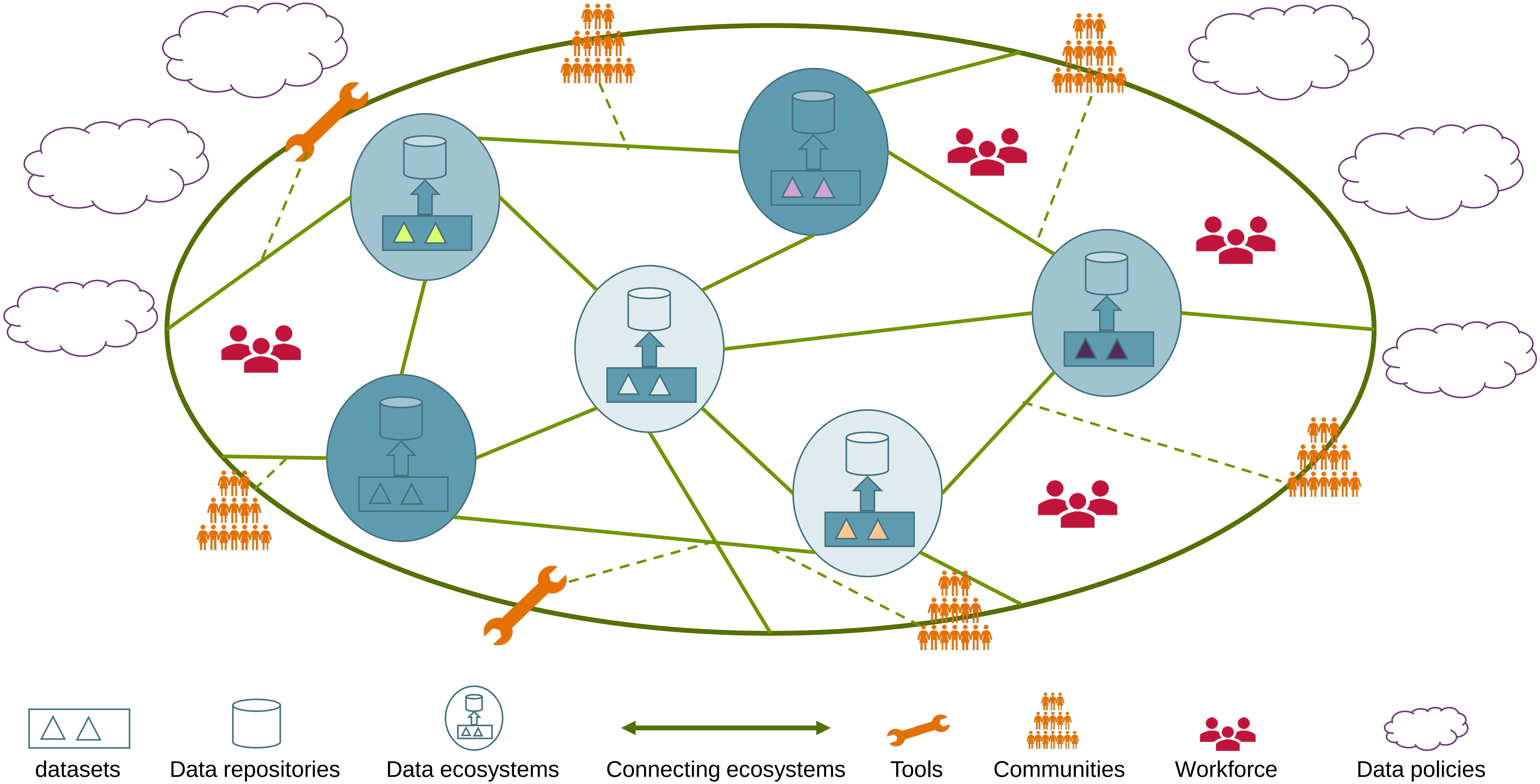
Sustainable Data Policies

FAIR Data and Data Infrastructure

Connecting NIH Data Ecosystems

Engaging with a Broader Community

Enhancing Biomedical Workforce



datasets

Data repositories

Data ecosystems

Connecting ecosystems

Tools

Communities

Workforce

Data policies

Special Thanks

- **STRIDES:** Andrea Norris, Nick Weber and NMDS team
- **Connecting NIH Data Resources:** Vivien Bonazzi, Regina Bures, Ishwar Chandramouliswaran, Tanja Davidsen, Valentine Di Francesco, Jeff Erickson, Tram Huyen, Rebecca Rosen, Steve Sherry, Alastair Thomson, Nick Weber, and BioTeam
- **Linking Publications to Datasets:** Jim Ostell and NCBI Implementation Team
- **Data Repository and Knowledgebase Resources**: Valentina di Francesco, Ajay Pillai, Qi Duan, Dawei Lin, Christine Colvis, and James Coulombe
- **Trustworthy Data Repositories**: Dawei Lin, Kim Pruitt, Jennie Larkin, Elaine Collier, Christine Melchior, Minghong Ward, and Matthew McAuliffe
- **Criteria for Open Access Data Sharing Repositories:** Mike Huerta, Dawei Lin**,** Maryam Zaringhalam, Lisa Federer and BMIC Team
- **Pilot for Scaled Implementation for Sharing Datasets:** Ishwar Chandramouliswaran and Jennie Larkin
- **Coding-it-Forward Fellows Summer Program:** Jess Mazerik
- **Data Science Training**: Valerie Florance, Jon Lorsch, Kay Lund, Kenny Gibbs, Shoshana Kahana, Erica Rosemond, Carol Shreffler
- **Diversity in Biomedical Data Science**: Valerie Florance, Jon Lorsch, Hanna Valantine, Roger Stanton, Charlene Le Fauve, Ravi Ravichandran, Zeynep Erim, Derrick Tabor, Rick Ikeda

Stay Connected

@NIHDataScience

/NIH.DataScience

www.datascience.nih.gov

NIH National Institutes of Health
Office of Data Science Strategy