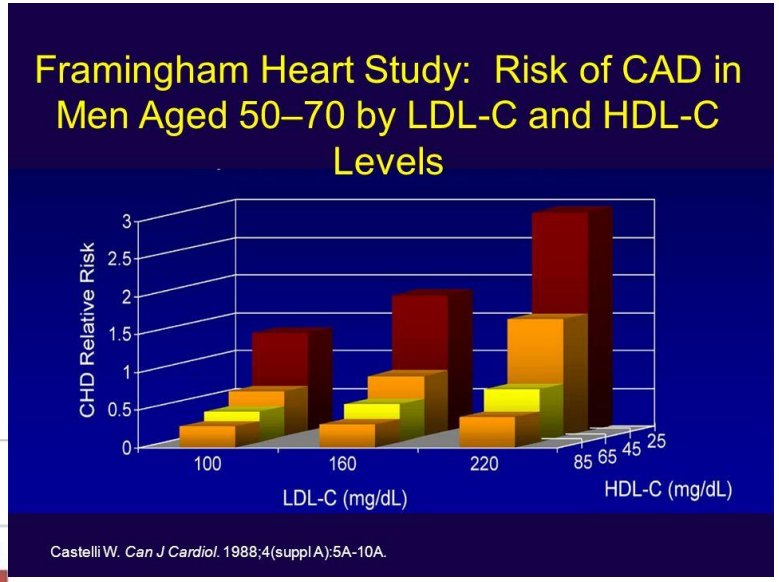
Enabling a FAIR-Data Ecosystem

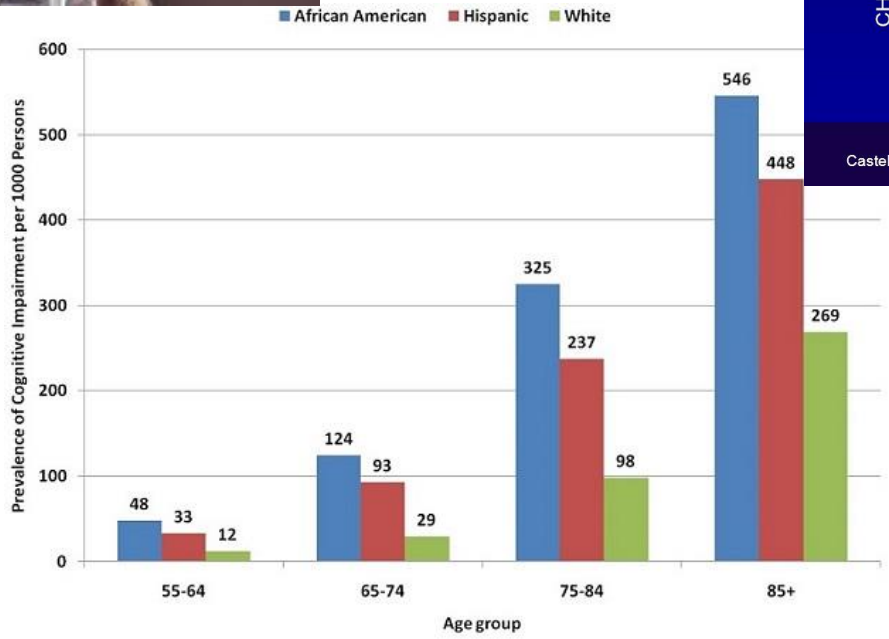
National Institutes of Health

April 8, 2019

IMAGINE…

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

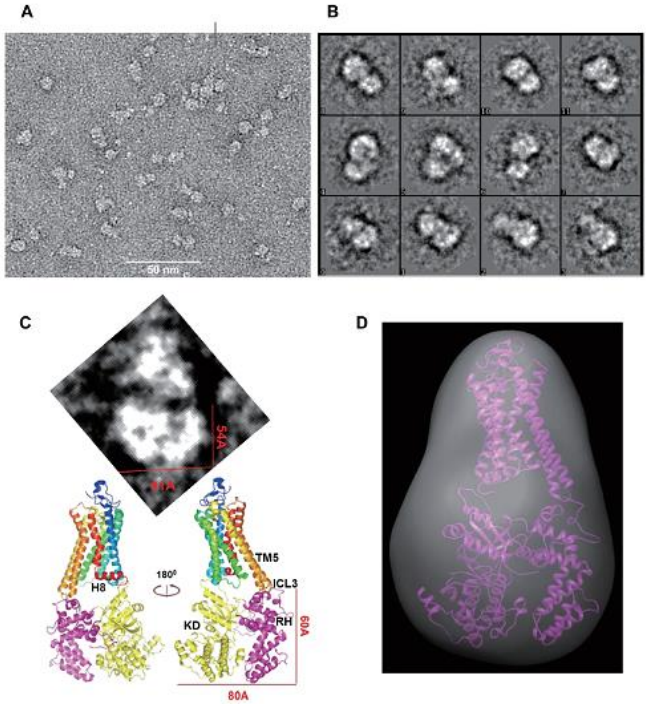




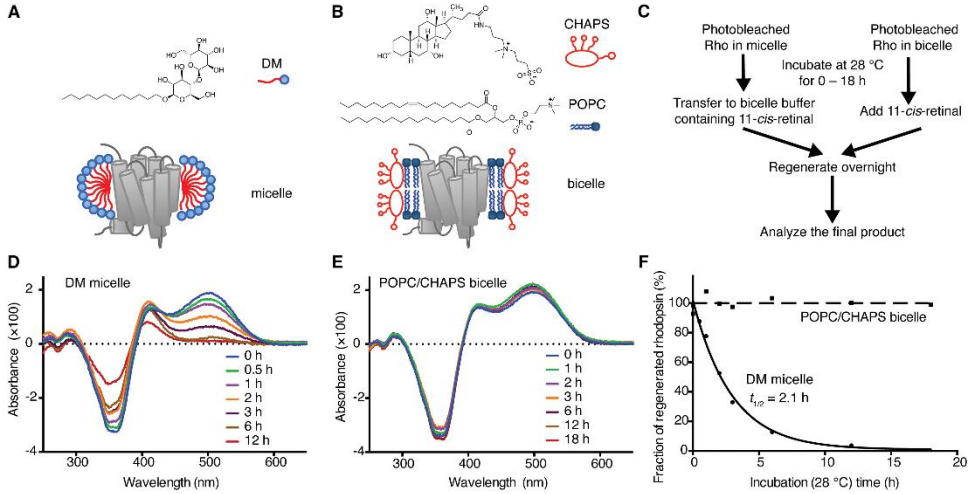
National Institutes of Health

IMAGINE…

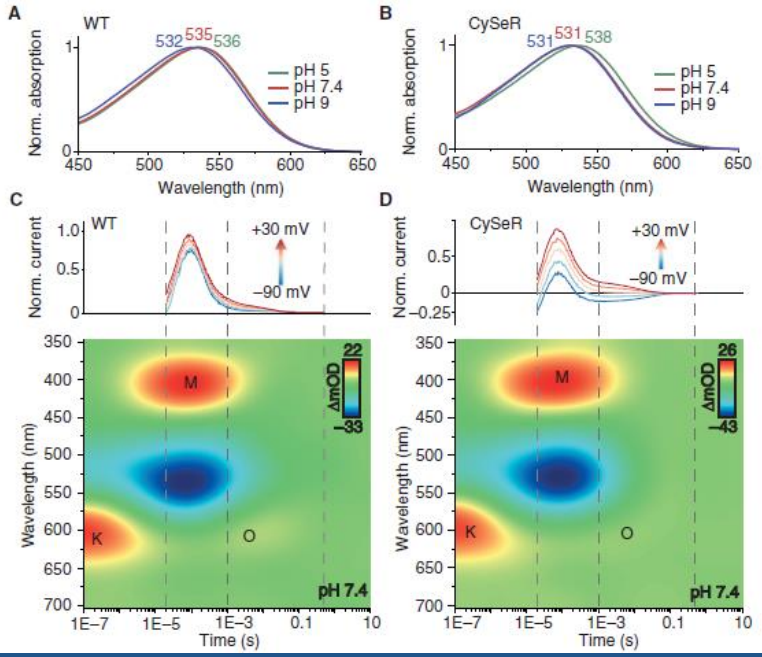
the ability to quickly obtain access to data, and related information, from published articles



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



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



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

National Institutes of Health

How can we facilitate data & information sharing (FAIR)

- ✓ Link datasets to publications (PubMed)
- Develop criteria for open access NIH Data Sharing Repositories
- Provide FAIR-enabled, open access for datasets that underly a publication resulting from NIH funded research

Measuring the Impact of Digital Repositories

Interagency Workshop (2017) to identify current assessment metrics, tools, and methodologies that are effective in measuring the impact of digital data repositories

Key Takeaways:

- A group with broad expertise and experience able to formulate and recommend best practices for data sharing and reuse.
- A data citation system that treats data as first-class objects comparable to publications in the research life cycle.
- **Data repository certification that is understandable and usable across a broad range of repositories.**
- New methods to assess economic impacts and opportunity costs when a repository is maintained or eliminated.
- A suite of strategies that repositories can use to achieve financial sustainability.

https://www.nitrd.gov/pubs/BD-IWG-Digital-Repository-Workshop-Report-2018.pdf

Open Access Data Sharing Repositories: Criteria

Goal: *Develop a set of core criteria to guide support, evaluation, and other considerations of biomedical data repositories*

NIH BioMedical Informatics Coordinating Committee (BMIC)

- Improve communication, collaboration and coordination around biomedical informatics, data and open science
- Cross-cutting issues of interest to all NIH
- Members appointed by ICO director
- Chartered in 2007; reports to Scientific Data Council

Sources to inform good criteria

- Clarivate Analytics
- CODATA
- CoreTrustSeal
- Data Seal of Approval
- Digital Curation Center
- Earth System Science Data
- FAIR Principles
- Interagency Input
- Nature Scientific Data
- PLOS
- Research Data Alliance
- Science Europe
- Smithsonian
- US Geological Survey

Criteria Mapped To:

- Dryad
- FigShare
- Mendeley Data

Next Steps

- Criteria drafted, includes provisions for repositories with human data
- Criteria reviewed by BMIC, Scientific Data Council and Data Science Policy Council
- **Community Input**: Request for Information (RFI)

Overview of Sharing Publication & 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 & access control

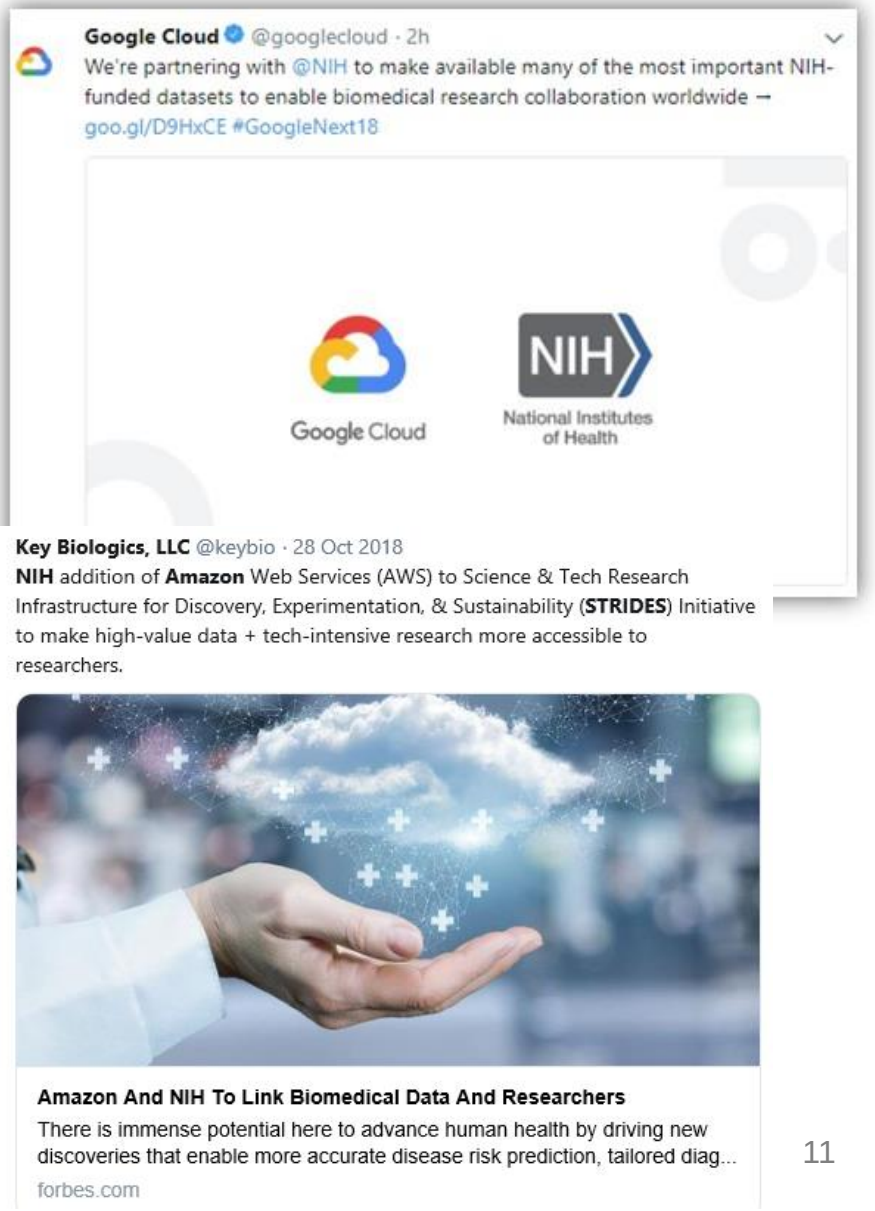
Could we do more? Connecting NIH Data Ecosystems

- ✓ Move/Access to high priority data sets in cloud service providers
- ✓ Single method for sign-on & data access across repositories and across cloud service providers
- Adaptative interface(s) & smart(er) data display/directory(ies)
- Workspace, access to NIH & community tools
- Computability on cloud service providers (and others…)
- Ability to connect to biomedical, clinical and healthcare research communities

Science & **T**ech **R**esearch **I**nfrastructure for **D**iscovery, **E**xperimentation and **S**ustainability

- 1st STRIDES agreement: Google Cloud (July ‘18)
- 2nd STRIDES agreement: Amazon Web Services (Oct ‘18)
- Other Transaction mechanism used
- Additional partnerships anticipated

https://commonfund.nih.gov/strides



Planning for a trans-NIH Data Management and Sharing Policy

- NIH solicited community input for policy foundation
  - **Oct. 2018**: Input requested on key provisions for potential data management and sharing policy (NOT-OD-19-014)
  - Two public webinars with ~800 participants (combined)
  - 189 submissions from national and international stakeholders

- Identified need for appropriate infrastructure; policy and implementation to go 'hand-in-hand'

- **<u>Next steps</u>**
  - Develop draft policy for data management and sharing and related guidance
  - Release draft for community input (target date summer 2019)
  - ***Release final policy target date end of 2019 calendar year***

Strategic Plan for Data Science: Goals and Objectives

Data Infrastructure	Modernized Data Ecosystem	Data Management, Analytics, and Tools	Workforce Development	Stewardship and Sustainability
• Optimize data storage and security • Connect NIH data systems	• Modernize data repository ecosystem • Support storage and sharing of individual datasets • Better integrate clinical and observational data into biomedical data science	• Support useful, generalizable, and accessible tools and workflows • Broaden utility of and access to specialized tools • Improve discovery and cataloging resources	• Enhance the NIH data-science workforce • Expand the national research workforce • Engage a broader community	• Develop policies for a FAIR data ecosystem • Enhance stewardship

https://datascience.nih.gov

Thank you, Questions

STRIDES
Andrea Norris, **Nick Weber** & 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 & NCBI Implementation team

Trustworthy Workshop
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

Making Data FAIR

Findable

Accessible

Interoperable

Reusable

- To be **Findable**, data must have unique identifiers, effectively labeling it within searchable resources.
- To be **Accessible**, data must be easily retrievable via open systems and effective and secure authentication and authorization procedures.
- To be **Interoperable**, data should “use and speak the same language” via use of standardized vocabularies.
- To be **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.