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

Susan Gregurick, Ph.D.

Associate Director for Data Science

Office of Data Science Strategy

February 10, 2020

National Institutes of Health
Office of Data Science Strategy

Genetic & dietary effects in COPD

Chronic Obstructive Pulmonary Disease 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

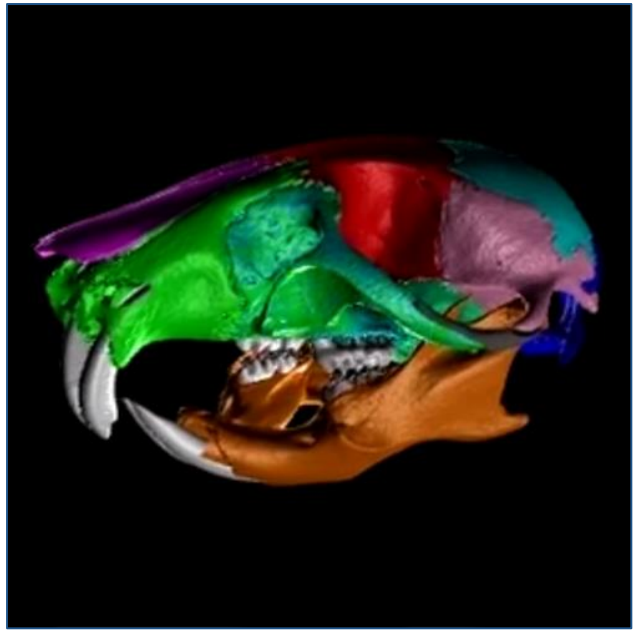
Obtaining access to all the relevant datasets so they can be analyzed

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

Dental, Oral and Craniofacial (DOC) Research

DOC research requires datasets from a wide variety of sources, combining dental record systems with clinical and research datasets, with unique challenges of facial data



Mouse skull Image from 3D rendering created by USC Center for Craniofacial Molecular Biology

Advances in facial imaging have lead to rich sources of imaging and quantitative data for the craniofacial region. In addition, research in model organisms can support DOC research in humans.

But Data access and consent are significant concerns with such inherently identifiable data types.

Challenges

Joining dental and clinical health records in order to integrate datasets from these two parts of the health system

Necessary datasets are spread across multiple repositories

Addressing the ELSI of facial data and DNA-based facial identification

Machine Learning in life sciences

Machine learning (ML) has great potential to help scientists make sense of the vast quantities of data being generated by modern instruments

Techniques like genomics, imaging, and others are generating vast amounts of data, that have to be processed and could be integrated with other data sources such as patient data

ML is becoming increasingly attractive as a way to analyze these types of datasets and find features of interest that may have mechanistic or diagnostic value

Challenges

Access to datasets of sufficient quality and quantity to support machine learning applications

Moving the large datasets across the network to local or to the cloud

Access to training and/or staff to develop new algorithms

Obtaining access to specialized hardware to run the analysis software

Global studies

Icon made by Roundicons from www.flaticon.com

The International epidemiologic Databases to Evaluate AIDS (IeDEA) network assembles data from 7 large regional research cohorts- representing over 500 HIV clinics around the world

NIH is supporting global research projects like IeDEA that have to collect and integrate data from hundreds of clinics worldwide.

This presents significant data management challenges to ensure that each clinic collects data in a compatible fashion and records the data in a compatible format.

International data access requirements can significantly impact a researcher's ability to access data collected in another country

Challenges

Harmonizing data from multiple sources so that it can be integrated and analyzed together.

Data access regulations to govern who can access data and what they can do with it are complex.

National regulations may preclude the use of US-based cloud resources limiting access to data and infrastructure that could support these projects

FAIR and data sharing

Researchers understand the concepts behind FAIR but need guidance on how to put FAIR into practice

The FAIR principles (Findable, Accessable, Interoperable and Reusable) are familiar to many.

However, there is confusion about what FAIR means in practice.

It can be time consumingto create FAIR datasets.

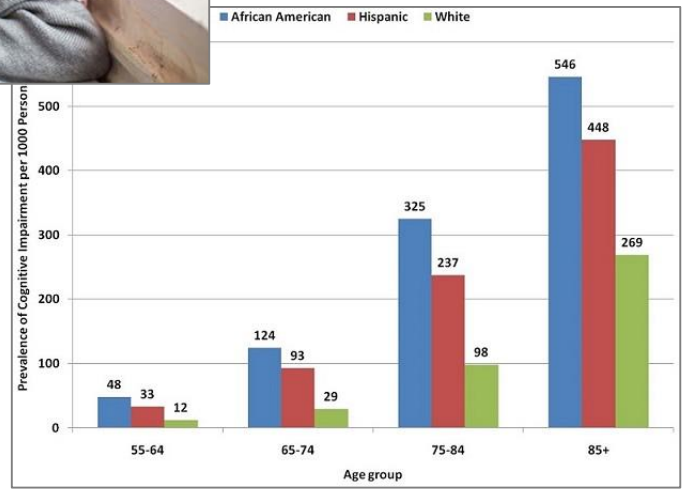
Challenges

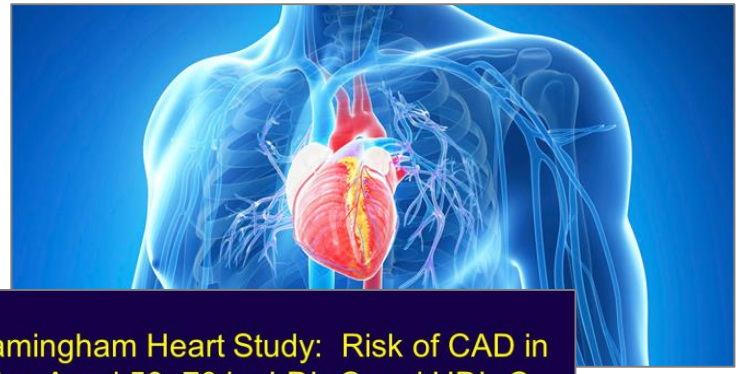
Prioritizing dataset annotation and curation when it is time consuming and perceived as an added burden

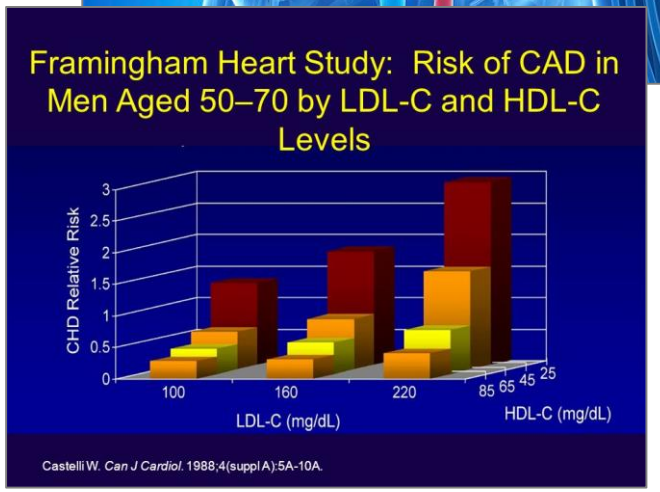
Selecting metadata to annotate their data that is compatible with other datasets and tools in the ecosystem

Where to put the data so it can be stored for the long term and securely accessed by authorized users (as appropriate)

IMAGINE… the ability to link data in the Framingham Heart Study with Alzheimer's health data to understand associations in, for example, cardiovascular health with aging and dementia.

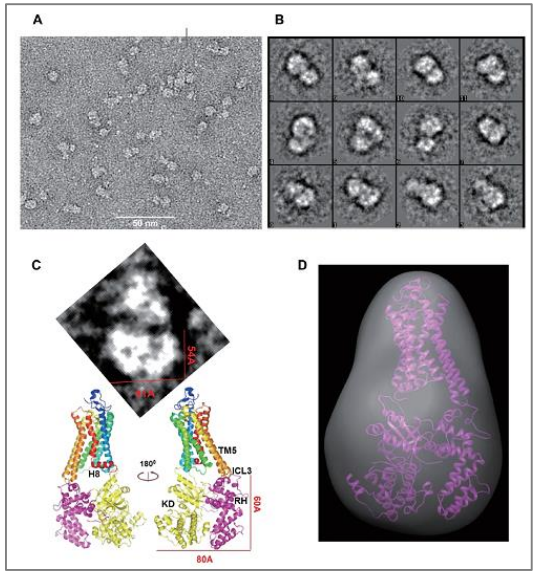




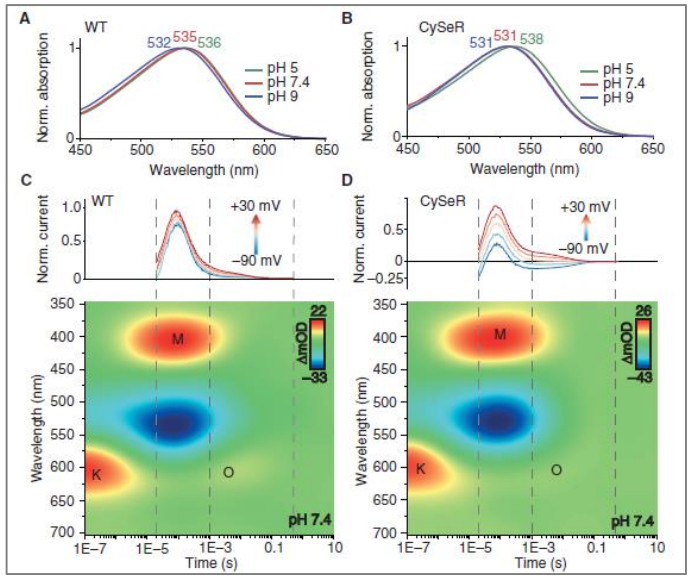


IMAGINE…

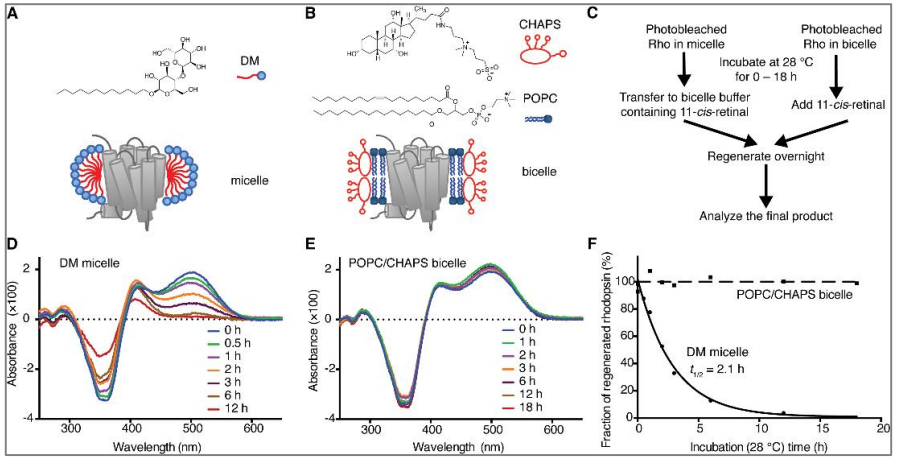
the ability to quickly obtain access to data, and related information, from published articles.



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



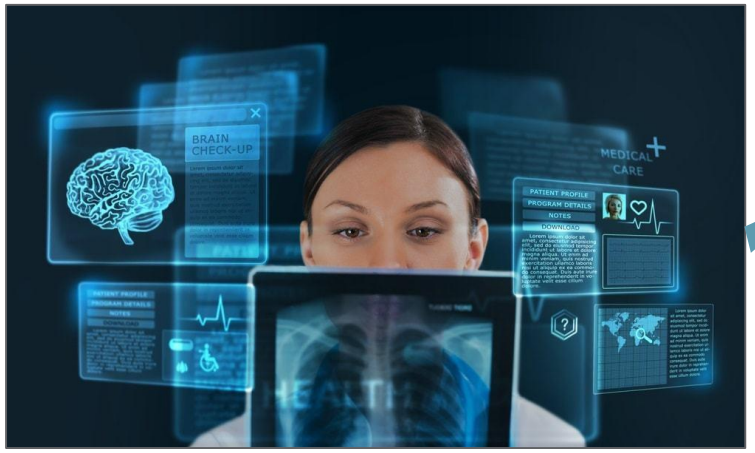
Absorption spectra of purified CsR-WT (A) and CySeR (B) at pH 5 (green), pH 7.4 (red), and pH 9 (blue). R. Fudim, e al, Science Signaling, 2019

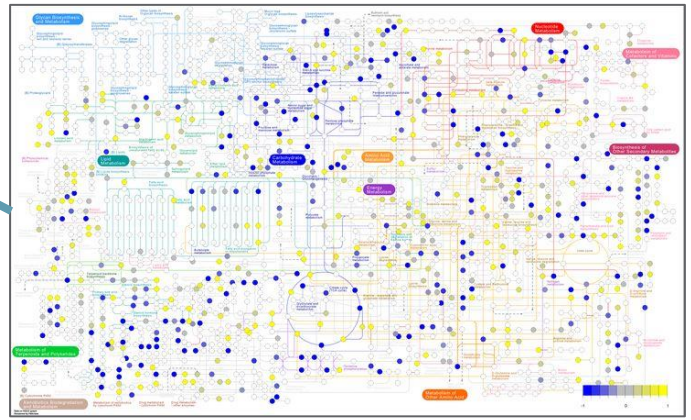


Energetics of Chromophore Binding in the Visual Photoreceptor of Rhodopsin, H. Tian et al, Biophysical Journal, 2017.

IMAGINE…

the 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.

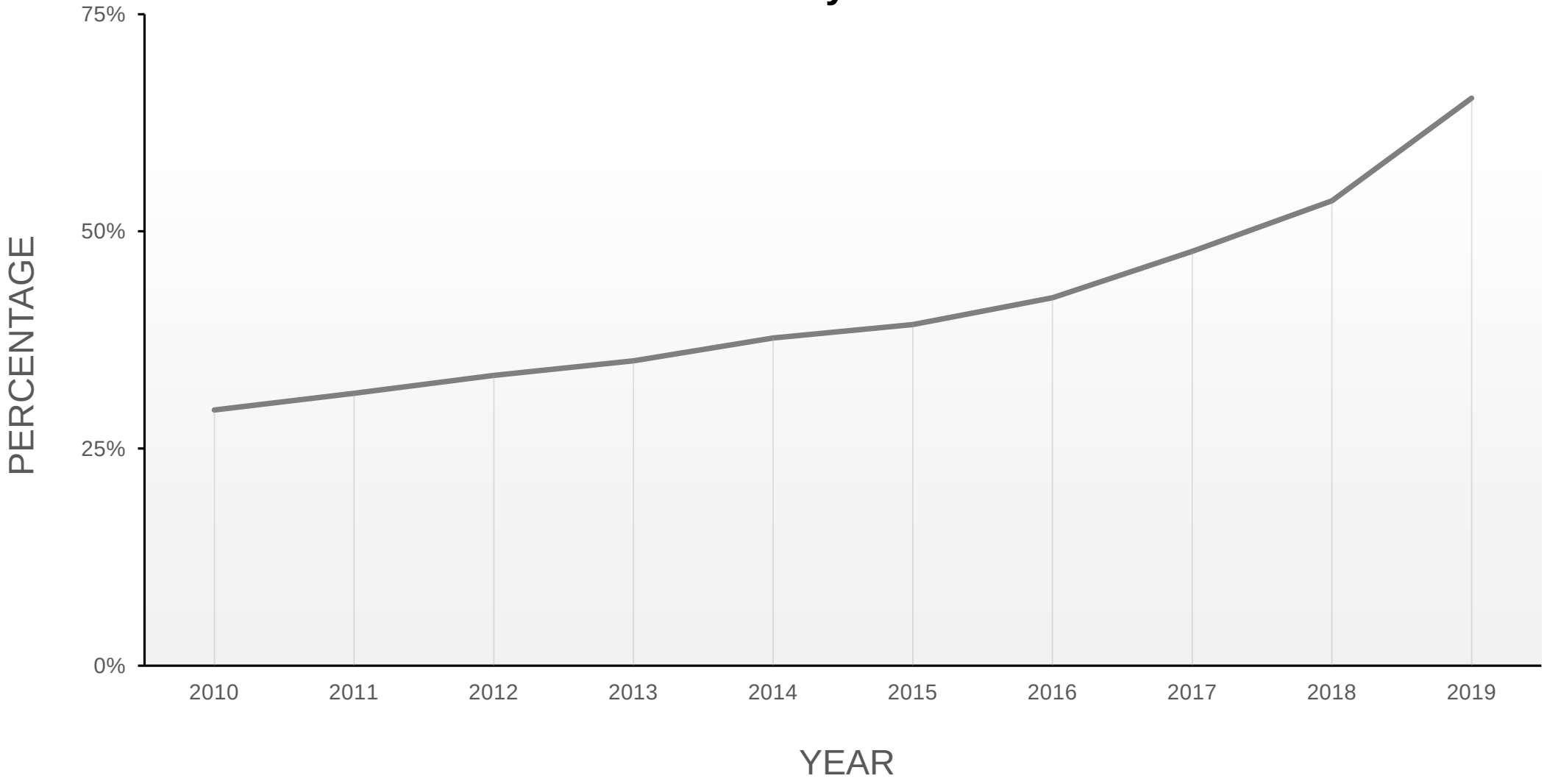
Strategic Plan for Data Science: Goals and Objectives

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

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.

Percentage of NIH Supported PMC publications with data availability statement



NIH Data Management and Sharing Policy Development

- **Researchers** with NIH-funded or conducted research projects resulting in the generation of scientific data will be required to submit a Plan
- **Plans** should explain how scientific data generated by a research study will be managed and which of these scientific data will be shared

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**	Datasets up to **20*gigabytes**	High Priority Datasets **petabytes**
PubMed Central	**Use of commercial and [illegible]**	**STRIDES Cloud Partners**
• 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.	• Assign Unique Identifiers to datasets associated with publications and link to PubMed. • Store and manage datasets associated with publication, up to 20* GB.	• Store and manage large scale, high priority NIH datasets. (Partnership with STRIDES) • Assign Unique Identifiers, implement authentication, authorization and access control.

https://datascience.nih.gov/data-ecosystem/biomedical-data-repositories-and-knowledgebases

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	**Community Engagement**
Quality of Data and Services and Efficiency of Operations	**Governance**

Optimized Funding for NIH Data Repositories and Knowledgebases

Funding Opportunities

- NIH released two funding opportunities on Jan. 17 to support biomedical data repositories and knowledgebases:
- Biomedical Data Repository (PAR-20-089)
- Biomedical Knowledgebase (PAR-20-097)

Scientific Impact

Community Engagement

Quality of Data and Services and Efficiency of Operations

Governance

NIH supports many repositories for biomedical data sharing

DASH Data and Specimen Hub

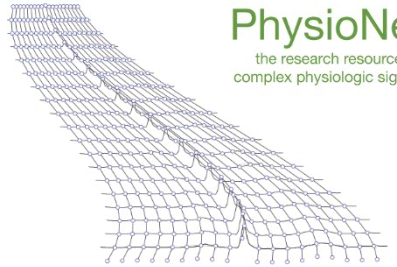
NIAGADS

CELL IMAGE LIBRARY

RCSB PDB PROTEIN DATA BANK

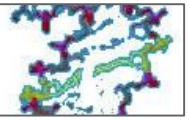
GEO Gene Expression Omnibus

PhysioNet
the research resource for complex physiologic signals



AphasiaBank

dbSNP
Short Genetic Variations



AMP-AD Knowledge Portal

OpenfMRI

IMMPORT
BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY

BioLINCC
Biologic Specimen and Data Repository Information Coordinating Center
Web Access to NHLBI Biospecimens and Data

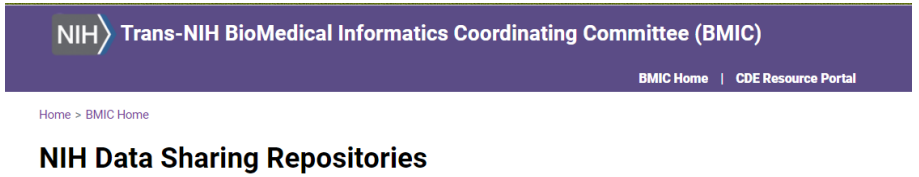
FITBIR
Federal Interagency Traumatic Brain Injury Research
INFORMATICS SYSTEM

NIMH
National Institute of Mental Health

How to find Data Repositories?

- **BMIC Data Repository Listing**

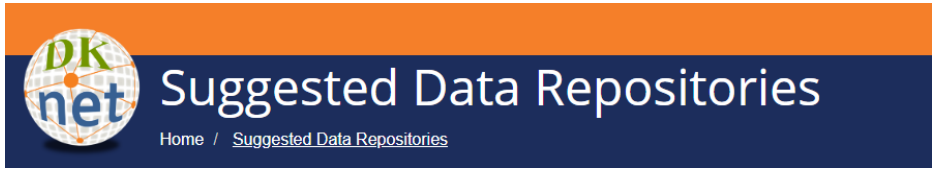
https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html



- **SciCruch/dkNET**

  Organized by repository type and scientific area.

https://dknet.org/about/Suggested-data-repositories



- **FAIRsharing**

https://fairsharing.org/

- **DataMed**

https://datamed.org/

“The goal of the Virtual Workshop is to understand metrics and different use cases for how the metrics can be used to assess the value and impact of data repositories and their datasets.”

Session 1: Evaluating and measuring data use and utility
Session chair: Daniella Lowenberg

Session 2: Stakeholder use cases for data usage and utility metrics
Session chair: Dr. Warren Kibbe

Technology support:

Pre-workshop Workshop Post-Workshop

slido →
Realtime polls and discussions

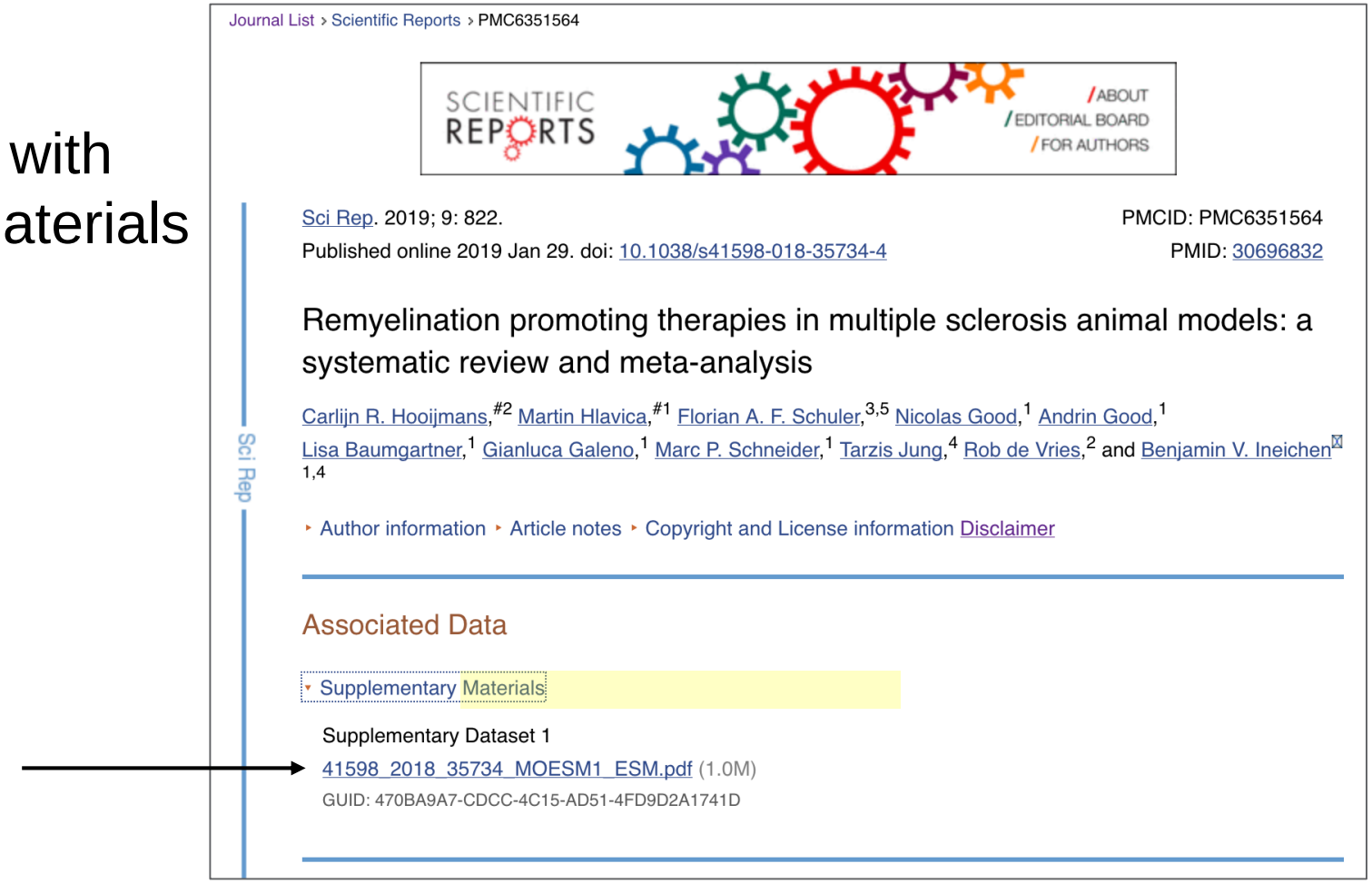
powered by ideascale →
Idea generations and discussions for long term community engagement

February 19, 2020
https://www.scgcorp.com/Metrics2020

335 registered as of 2/4/2020

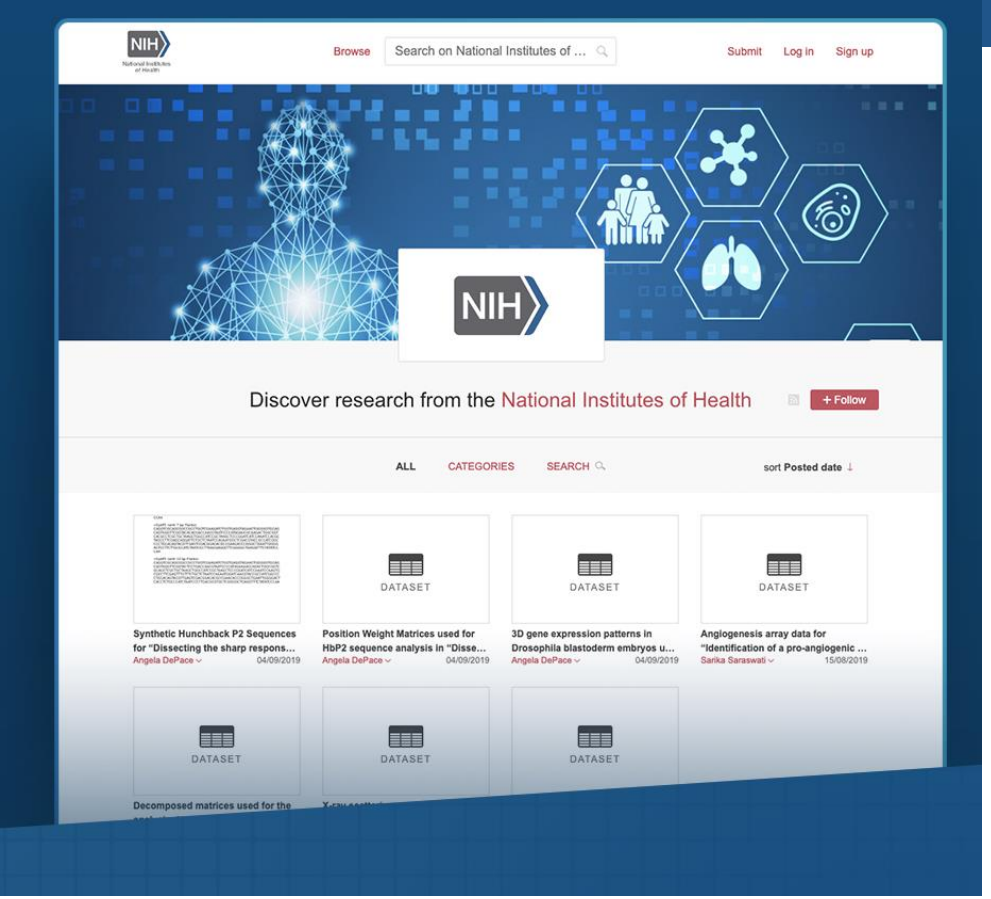
Datasets in PubMed Centeral

PubMed now links datasets, with DOIs, in the supplemental materials of publications



Providing FAIR-enabled, open-access options for datasets

Generalist Repository Pilot: NIH Figshare



https://nih.figshare.com/f/faq

Share

- Self-publish any data type and file format
- Link grant information
- Bulk-upload with API
- 100GB storage per user

Discover

- Access open, de-identified data
- Search and filter on metadata
- Indexed in Google
- Track usage metrics

Cite

- Assign a DOI
- Attach a license
- Ability to embargo
- Secure storage on FedRAMP AWS S3

Make your research FAIR in a few easy steps

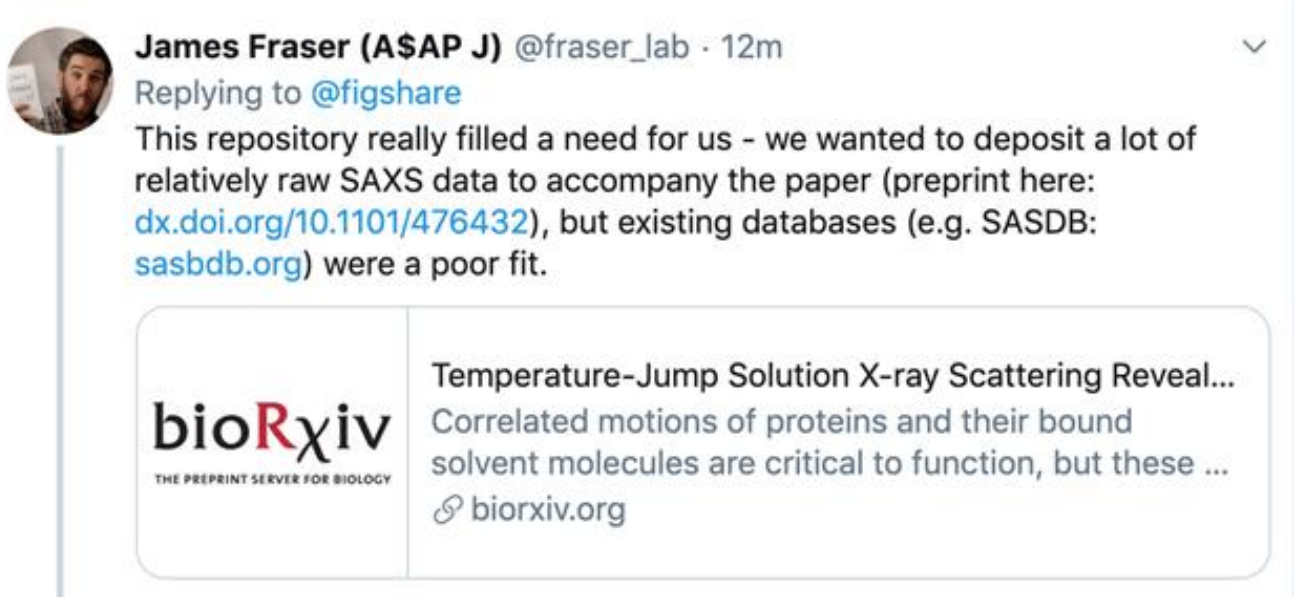
1. Create an account at nih.figshare.com.
2. Create a new item.
3. Assign important metadata to your dataset to help provide context for reuse, link to relevant funding information or associated publications, and make your research more discoverable.
4. Publish! Once your content is published, it'll go into a review queue to be checked for metadata completeness and ensure all submitted content adheres to NIH policies.
5. Once live, Figshare will track all attention and potential impact around your research. All published research receives a DOI, which will help with data citation.

For more information about the NIH Figshare pilot or to share your questions, ideas, or suggestions, please email datascience@nih.gov. For technical support, please email nihsupport@figshare.com.

Persistent Identifiers and Tracking Attention, Use, and Reuse

- All submissions have a DOI
  - Supports data citation
  - Usage and citation statistics
  - Other alternative metrics

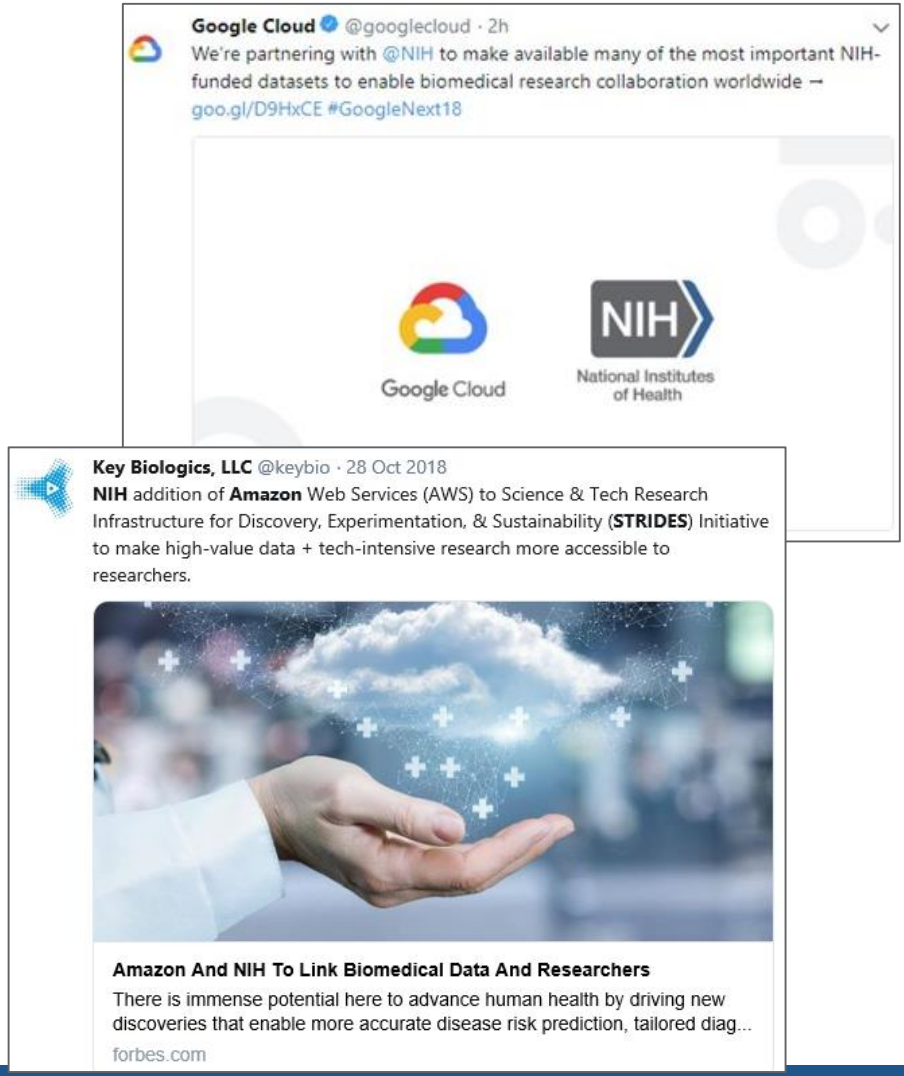
- Platform and dataset statistics and metrics
  - Openly available
  - Exported to other NIH systems using the API



Providing FAIR-enabled, open-access options for datasets

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

- **Discounts on STRIDES Initiative partner services**—Favorable pricing on computing, storage, and related cloud services for NIH Institutes, Centers, and Offices (ICOs) and **NIH-funded researchers through their institutions**.
- **Professional services**—Access to professional service consultations and technical support from the STRIDES Initiative partners.
- **Training**—Access to training for researchers, data owners, and others to help ensure optimal use of available tools and technologies.
- **Potential collaborative engagements**—Opportunities to explore methods and approaches that may advance NIH's biomedical research objectives



https://datascience.nih.gov/strides

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

GET STARTED WITH THE STRIDES INITIATIVE

Explore the Use of Cloud Environments at Your Institution

1 SELECT A STRIDES INITIATIVE PARTNER

Learn from the NIH STRIDES Initiative team about the different industry-leading STRIDES Initiative partners available to your institution.

2 ENROLL IN THE STRIDES INITIATIVE

Engage with your preferred STRIDES Initiative partner to begin the enrollment process for your institution and start experiencing the available services and benefits.

3 SET UP YOUR ACCOUNT

Your NIH-funded researchers with active grants will work with your selected STRIDES Initiative partner to set up a STRIDES Initiative account.

4 CONFIGURE YOUR CLOUD ENVIRONMENT

Your STRIDES Initiative partner will equip your NIH-funded researchers with the resources they need to set up their cloud environment.

5 USE THE CLOUD FOR BIOMEDICAL RESEARCH

Your NIH-funded researchers can now begin using the cloud environment for computation, storage, and data analysis.

Cloud training opportunities and professional services are available on an ongoing basis.

Examples of Datasets in the STRIDES Cloud

- NHLBI Framingham Heart Study
- All of Us Research Program
- NCI Genomic Data Commons
- **NCBI data resources (12 PB!)**
- NHLBI Trans-Omics for Precision Medicine (TOPMed) Program
- Gabriella Miller Kid's First

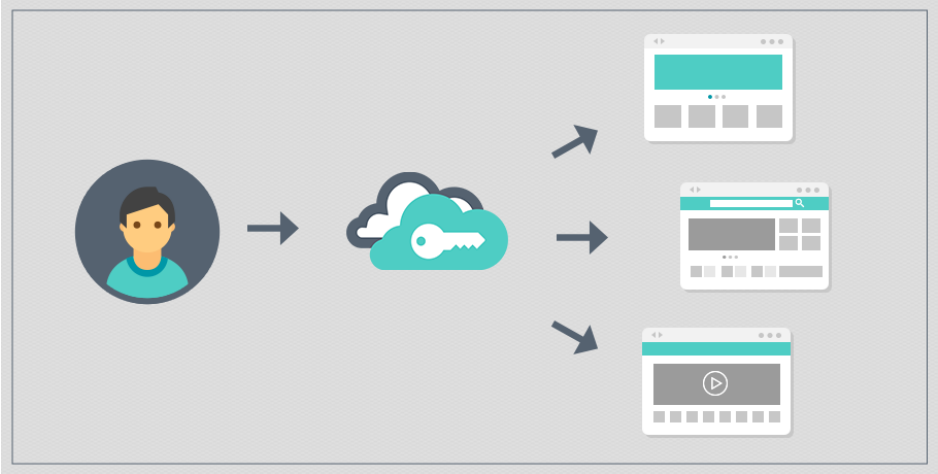
And Many More

- **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
  - We can search across this amount of data using advance AI algorithms

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



Researcher Authentication Service hits a first milestone

- Successfully integrated Globus' login functionality with a new NIH Login capability that uses OpenID Connect (OIDC) for electronic Research Administration (eRA) Commons accounts.
- Globus customers can login in with eRA Commons
- This integration represents a foundational part of the RAS project.
- OIDC can be rapidly **adopted and extended** to support other integration partners (e.g., Google, Terra, and Login.gov)

FHIR: Making EHR Interoperable

Fast

Healthcare

Interoperability

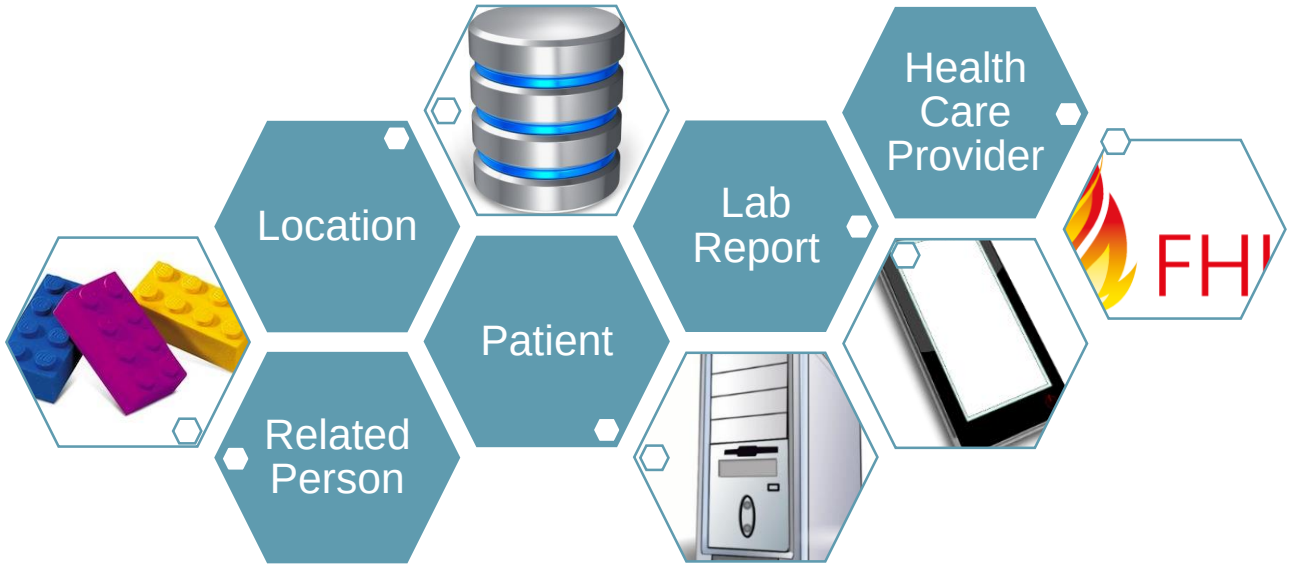
Resources

- Developed by Health Level Seven International (HL7), a non-profit organization
- Designed specifically for exchanging electronic health care record data
- For patients and providers, it can be applied to mobile devices, web-based applications, and cloud services
- FHIR is already widely used in hundreds of applications across the globe for the benefit of providers, patients and payers

HL7 FHIR

FHIR Servers and FHIR APIs

- Small logically discrete units of “information”
- Reusable “components” that can be assembled into working systems
- FHIR consists of approximately 100+ resources:
  - Foundational: e.g., Data Types
  - Supporting: e.g., Security, Privacy
  - Administrative: e.g., Patient
  - Healthcare: e.g., Allergy
  - Clinical Reasoning: e.g., Measure



NIH Efforts Regarding the Use of FHIR® Standard

- **Advancing Sharing of Phenotypic Information through FHIR**
  - Will **increase the availability of high-quality standardized phenotypic information for genomic research and genomic medicine** by extracting relevant data from electronic health record (EHR) systems using FHIR

- **Development and Testing of FHIR Tools for Researchers**
  - Will fund development and testing of tools and resources that enable clinical researchers to **extract clinical data from an EHR system for research** and **map their research data to the FHIR standard to be deposited in FHIR servers**

Examples of FHIR-Based Projects at NIH

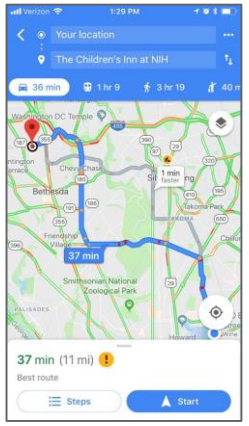
dbGaP on FHIR: development of FHIR API to retrieve standardized phenotypic data from several large population cohort studies

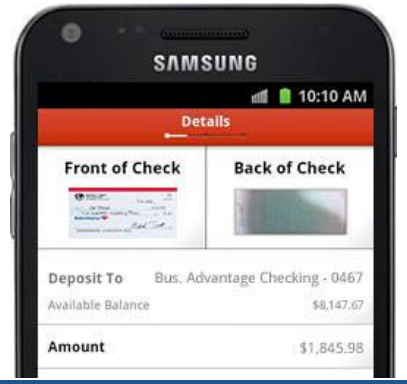
eMERGE on FHIR: will release a FHIR-based schema for returning structured genetic test results identified in a clinical report

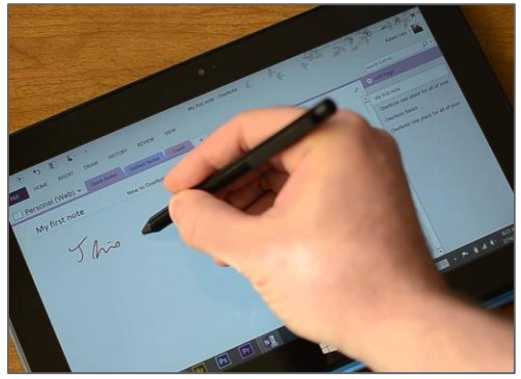
EHRs to FHIR: developing prototype of shared infrastructure that uses FHIR to enable authorized users to retrieve bulk data from partner organizations

AI is All Around Us…

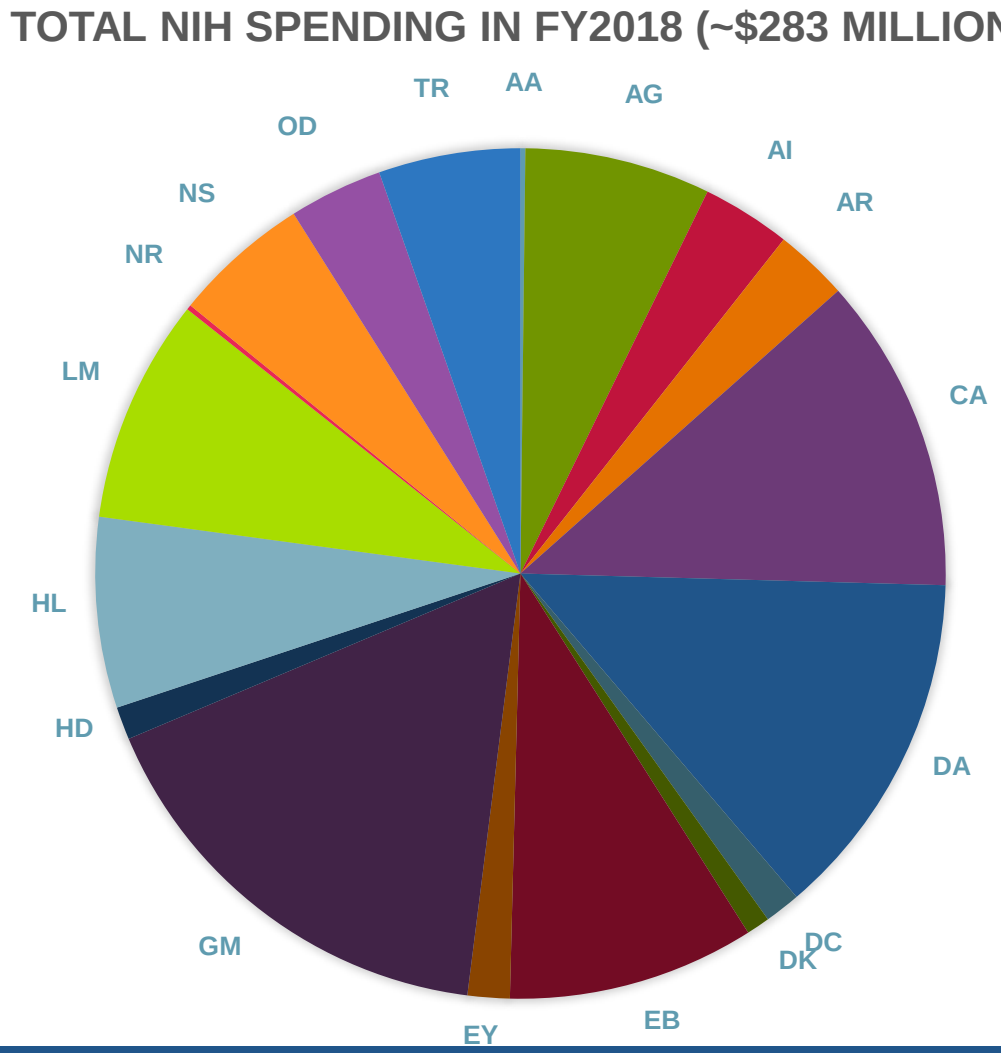
A Convergence of Technology, Computing, and Artificial Intelligence Algorithms

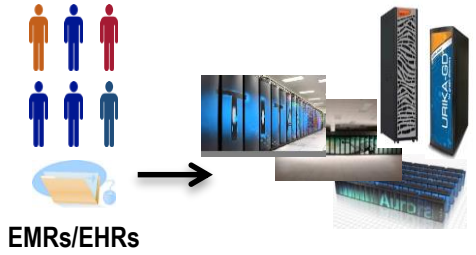


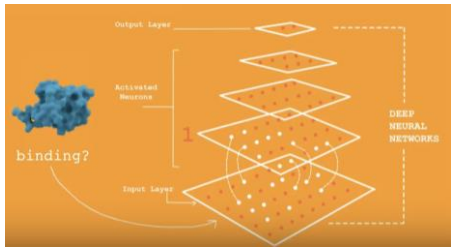


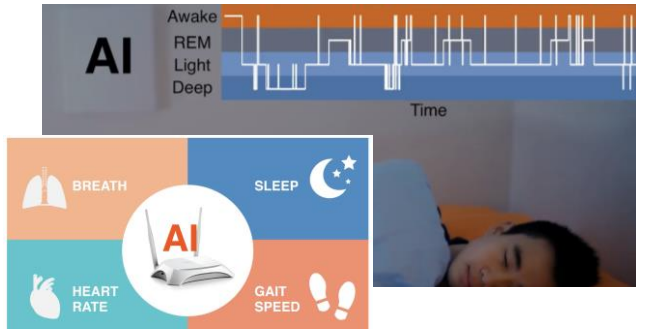


AI Investments Reach Across NIH









Input
Chest X-Ray Image

CheXNet
121-layer CNN

Output
Pneumonia Positive (85%)

NIH Advisory Committee to the Director AI Working Group

Rediet Abebe
Cornell University

Greg Corrado
Google

Kate Crawford
AI Now Institute

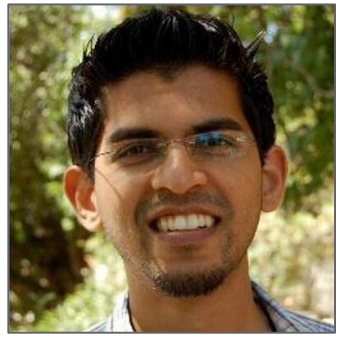
Barbara Engelhardt
Princeton University

David Glazer
Verily Life Sciences
(Co-Chair)

David Haussler
UC Santa Cruz

Dina Katabi
MIT

Daphne Koller
Insitro



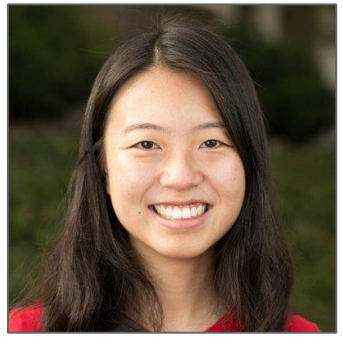
Anshul Kundaje
Stanford University

Eric Lander
Broad Institute

Jennifer Listgarten
UC Berkeley

Michael McManus
Intel Corporation

Lawrence Tabak
NIH
(Co-Chair)



Serena Yeung
Stanford University

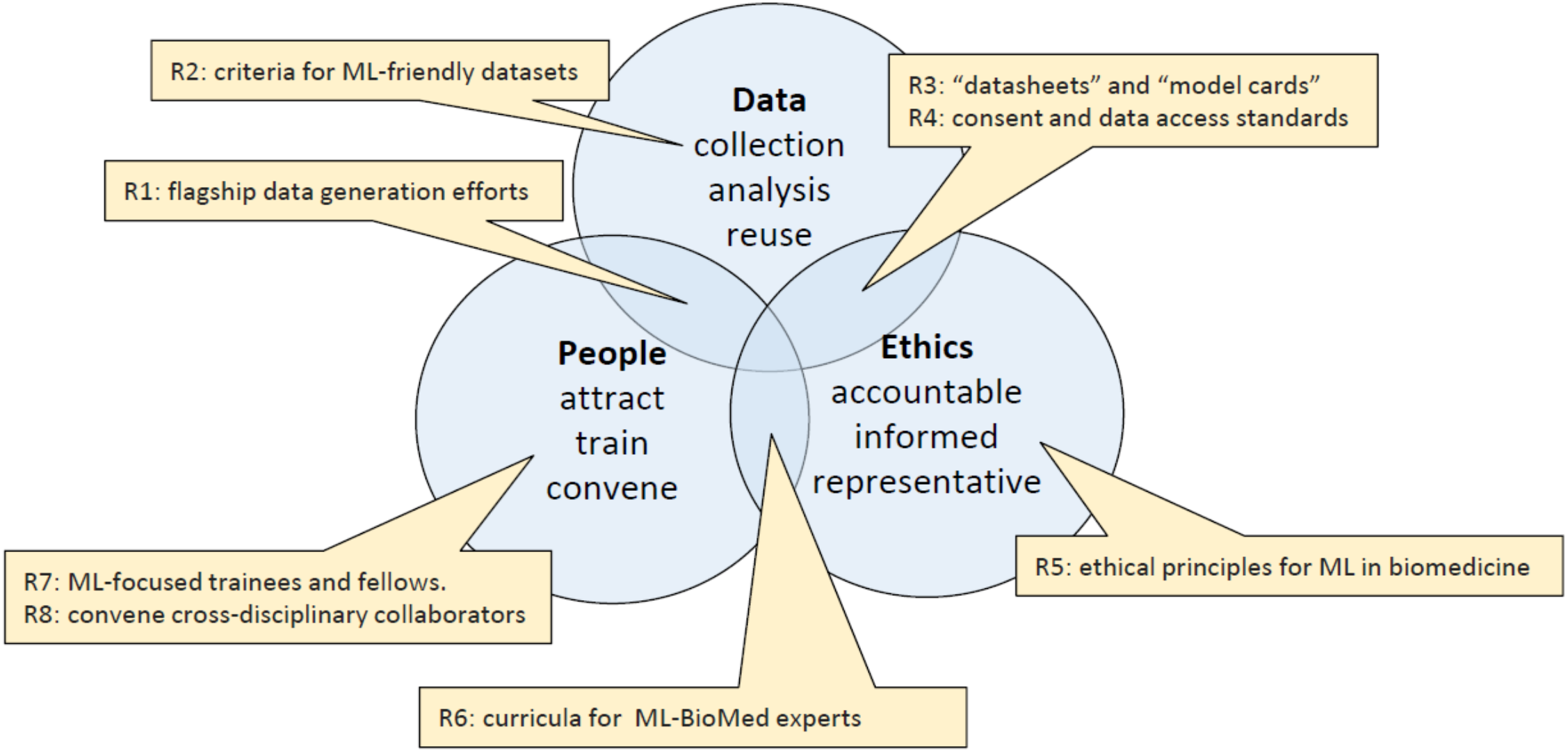
Charged December 2018; Interim Report June 2019

Charge to the AI Working Group (December 14, 2018)

- Are there opportunities for cross-NIH effort in AI? How could these efforts reach broadly across biomedical topics and have positive effects across many diverse fields?
- How can NIH help build a bridge between the computer science community and the biomedical community?
- What can NIH do to facilitate training that marries biomedical research with computer science?
  - Computational and biomedical expertise are both necessary, but careers may not look like traditional tenure track positions that follow the path from PhD to post-doc to faculty
- Identify the major ethical considerations as they relate to biomedical research and using AI/ML/DL for health-related research and care, and suggest ways that NIH can build these considerations into its AI-related programs and activities

Recommendations

https://acd.od.nih.gov/documents/presentations/12132019AI.pdf



Enhance the Biomedical Workforce

Graduate Data Science Summer Program

- Master's-level internships at NIH
- 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

Enhance the Biomedical Workforce

Coding it Forward

- Undergraduate fellowships

- Fellows spend 10 weeks at NIH channeling their computational expertise toward hands-on experience with biomedical data-related challenges

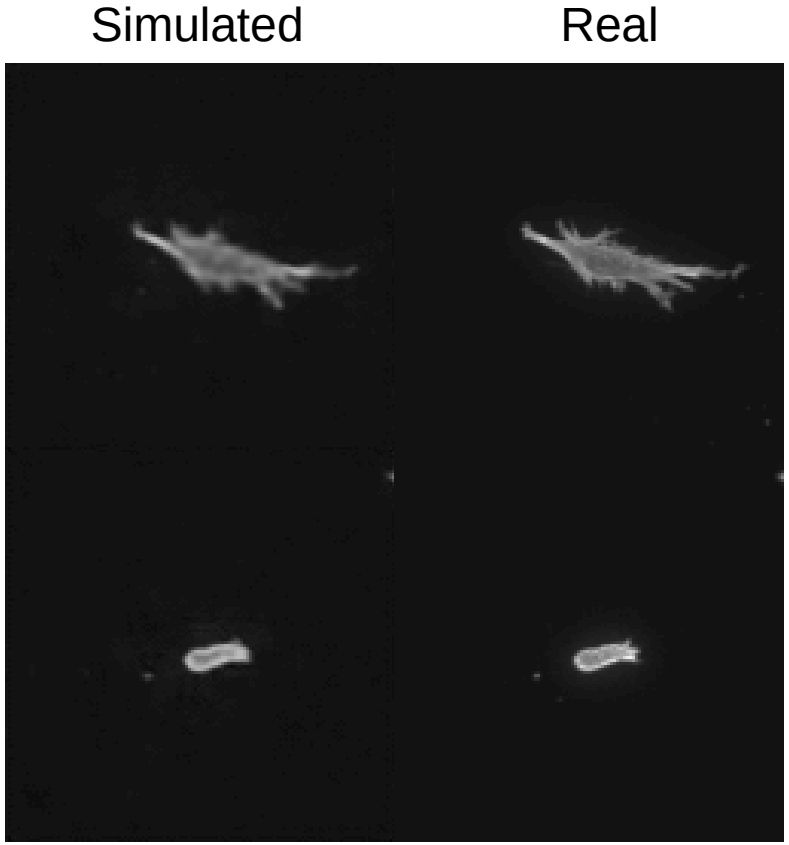
https://www.codingitforward.com/

Civic Digital Fellow @ NIDCR

- Developed machine learning models to predict migration paths and morphology of fibroblast cells in extra-cellular medium
- What would machine learning consider “interesting features”?
- Applications to cancer research by looking at the difference in what's interesting between healthy and cancerous cells

Mentor: John Prue, CIO

Fellow: Isaac Robinson, Computer Science and Music @ Yale



Example predictions of cell motion for a short subsequence.

NIH Data and Technology Advancement (DATA) National Service Scholar Program

National Institutes of Health
Office of Data Science Strategy

- One- or two-year **national service program** with high-impact NIH projects
- Seeking **industry data and computer scientists**, experts from related fields
- Expecting 5+ fellows in first cohort, starting in summer 2020

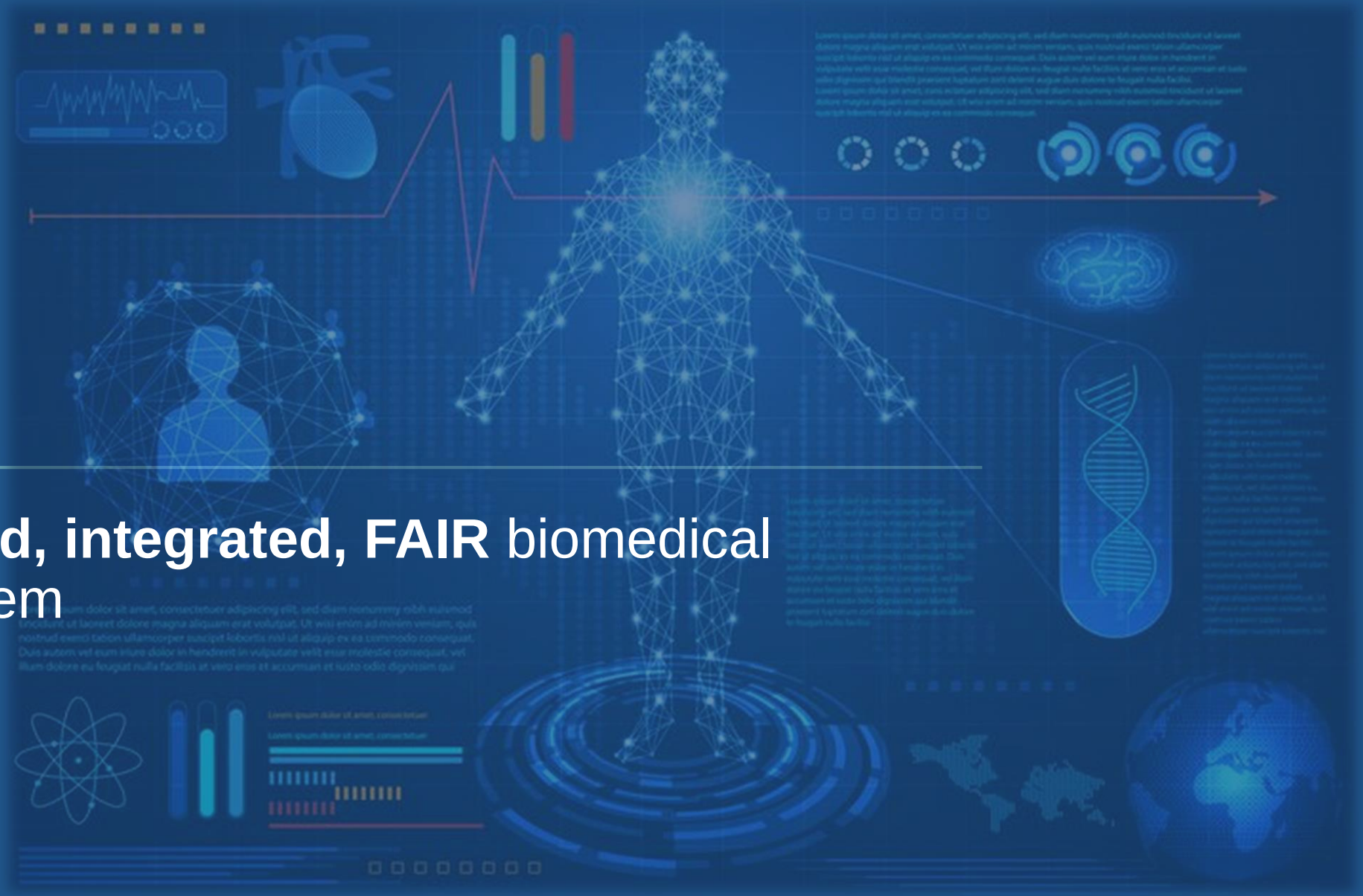
Applications due April 30

- Submit **CV** and **cover letter** including vision statement and projects of interest to datascience@nih.gov**.**
- Eligibility: doctoral degree (required) and **industry experience** (strongly preferred)
- Women and individuals from underrepresented groups **are encouraged to apply.**

https://datascience.nih.gov/data-scholars

VISION

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



Office of Data Science Strategy

- Provide leadership and coordination on the NIH strategic plan for data science.
- Develop and implement NIH's vision for a **modernized** and **integrated** biomedical data ecosystem.
- Enable a diverse and talented data science workforce.
- In coordination with the CIO, build strategic partnerships for advanced technologies and methods.

Special Thanks

- **STRIDES:** Andrea Norris, Nick Weber and NMDS team
- **Connecting NIH Data Resources:** Regina Bures, Ishwar Chandramouliswaran, Tanja Davidsen, Valentine 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
- **Linking Publications to Datasets:** Jim Ostell and NCBI Implementation Team
- **Data Repository and Knowledgebase Resources**: Valerie Florance, Valentina di Francesco, Ajay Pillai, Qi Duan, Dawei Lin, Christine Colvis, Jennie Larkin, Ravi Ravichandran, and James Coulombe
- **FHIR Pilots**: Teresa Zayas-Caban and Belinda Seto
- **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, Wynn Meyer
- **Graduate Data Science Summer Program**: Sharon Milgram and Phil Ryan (OITE)
- **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

National Institutes of Health
Office of Data Science Strategy