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

Susan Gregurick, Ph.D.

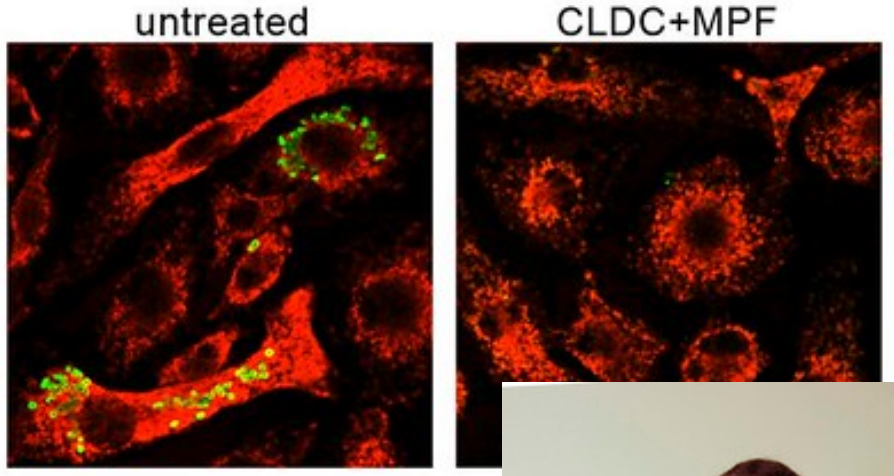
Associate Director for Data Science

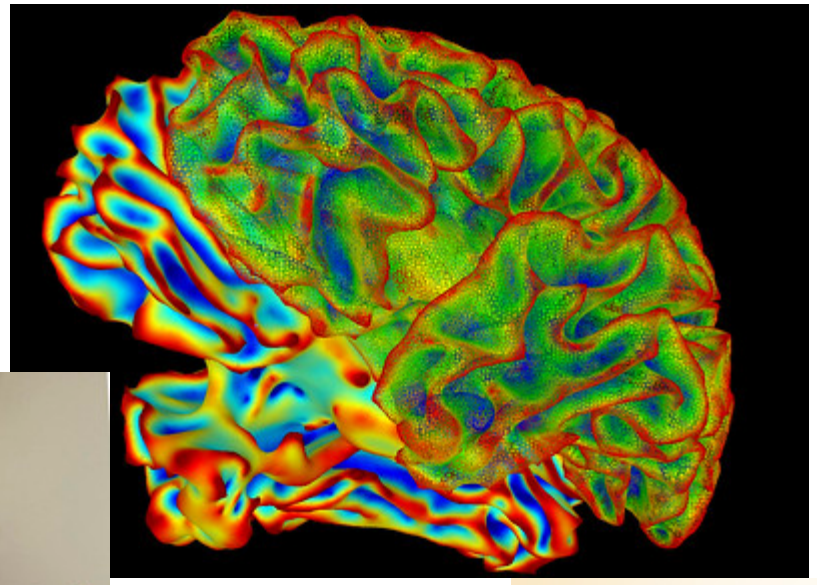
Office of Data Science Strategy

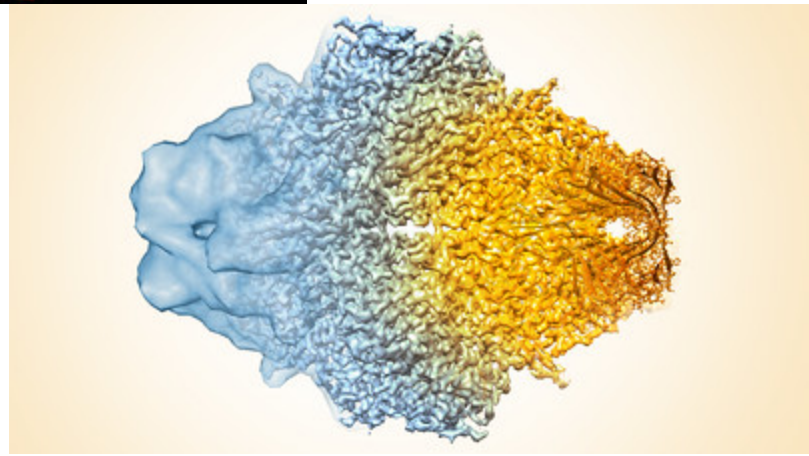
March 24th, 2020

National Institutes of Health
Office of Data Science Strategy

NIH supports many different biomedical research communities with diverse sets of data







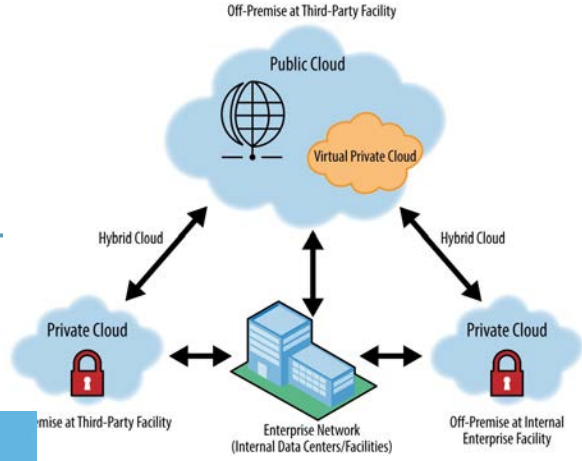
Data Science Drivers

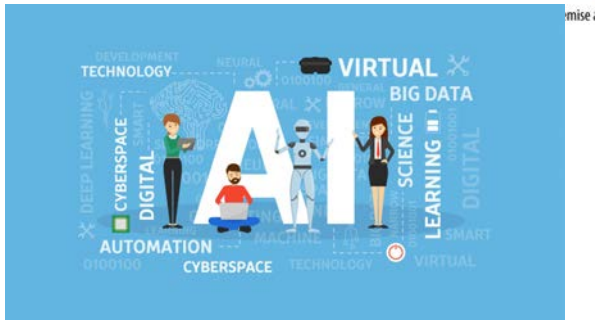
Democratization of computation, workflows and tools ➔ **more researchers can participate in data science**

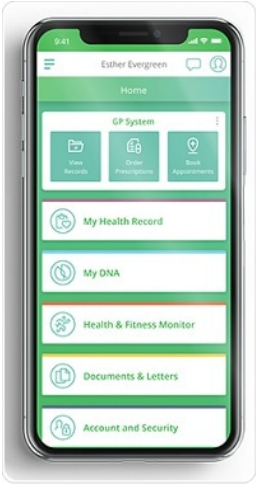
Rise of ML/DL/AI ➔ **well annotated data as a valuable resource**

Digital Data Governance ➔ governance and provenance technologies to **electronically manage data consent for appropriate reuse**

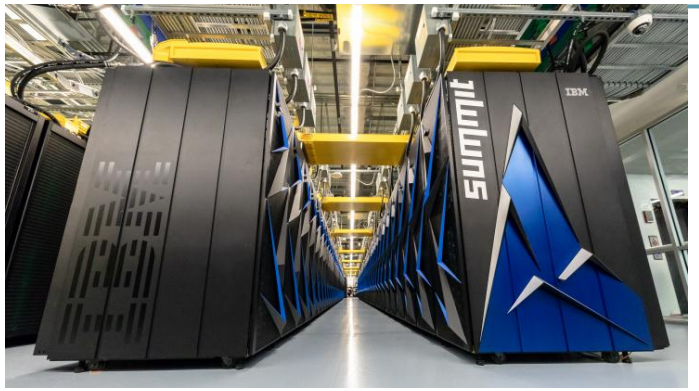
Large scale graph and analytics ➔ **data/information abstraction, on a large scale, is possible**







Computing Drivers



Heterogenous computing allows a convergence of physics based and data analytic based simulations

Integration of algorithms and computing, with microelectronics ➔ rise of edge/fog computing

Creating the first quantum internet

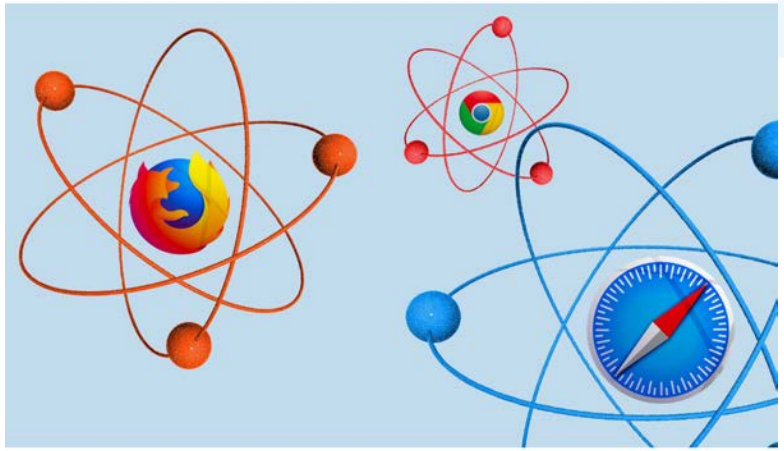


Illustration: Sara Grillo/Axios

Scientists in Chicago are trying to create the embryo of the first quantum internet.

Advances in Quantum Information, on a rapid timescale, unlocks new opportunities

Development of a computing-data ecosystem

This proliferation of data, and the accompanying computing resources and new algorithms, brings new opportunities for discovery, as well as new challenges

Genetic & dietary effects in COPD

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

Icon made by Roundicons from www.flaticon.com

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

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 and need guidance on how to put them into practice

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

However, there is significant confusion about what FAIR means in practice and it can be time consuming and the benefits to the data owner undertaking this 'extra work' can be unclear.

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 and securely accessed by authorized users (as appropriate)

The Rime of the Ancient Mariner, Samuel Taylor Coleridge

(excerpted)

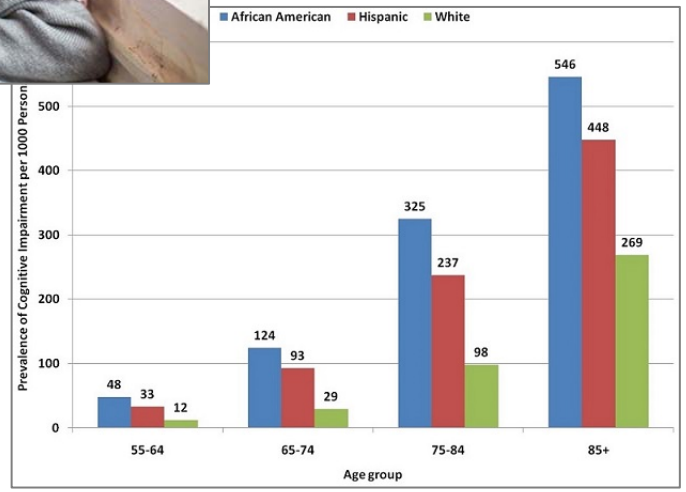
Day after day, day after day,
We stuck, nor breath nor motion;
As idle as a painted ship
Upon a painted ocean.

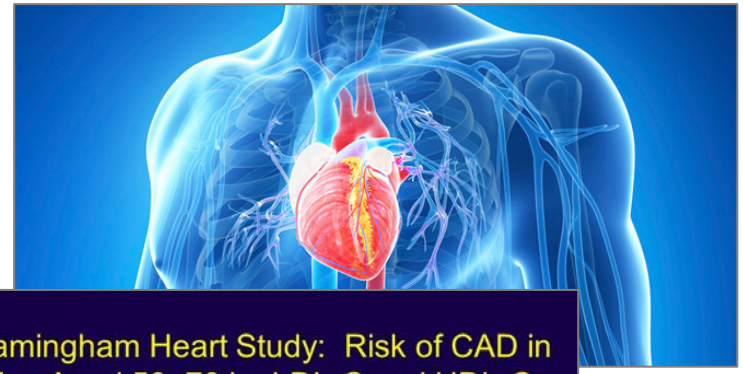
Water, water, every where,
And all the boards did shrink;
Water, water, every where,
Nor any drop to drink.

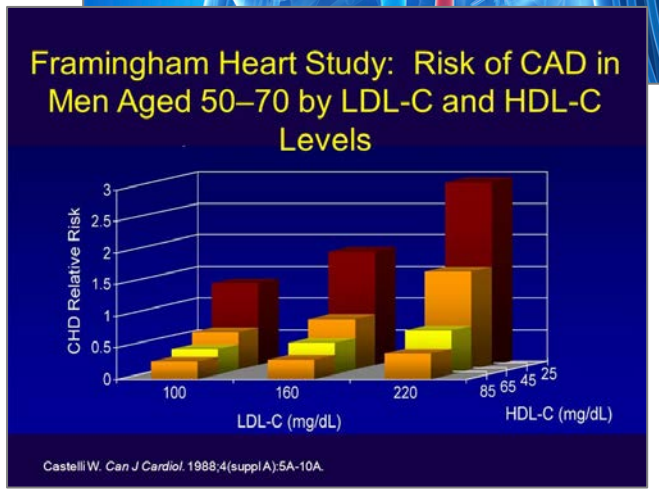
But what if we could…..

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.







Research articles

Survival of retinal ganglion cells after damage to the occipital lobe in humans is activity dependent

Colleen L. Schneider, Emily K. Prentiss, Ania Busza, Kelly Matmati, Nabil Matmati, Zoë R. Williams, Bogachan Sahin and Bradford Z. Mahon

Published: 27 February 2019 | https://doi.org/10.1098/rspb.2018.2733

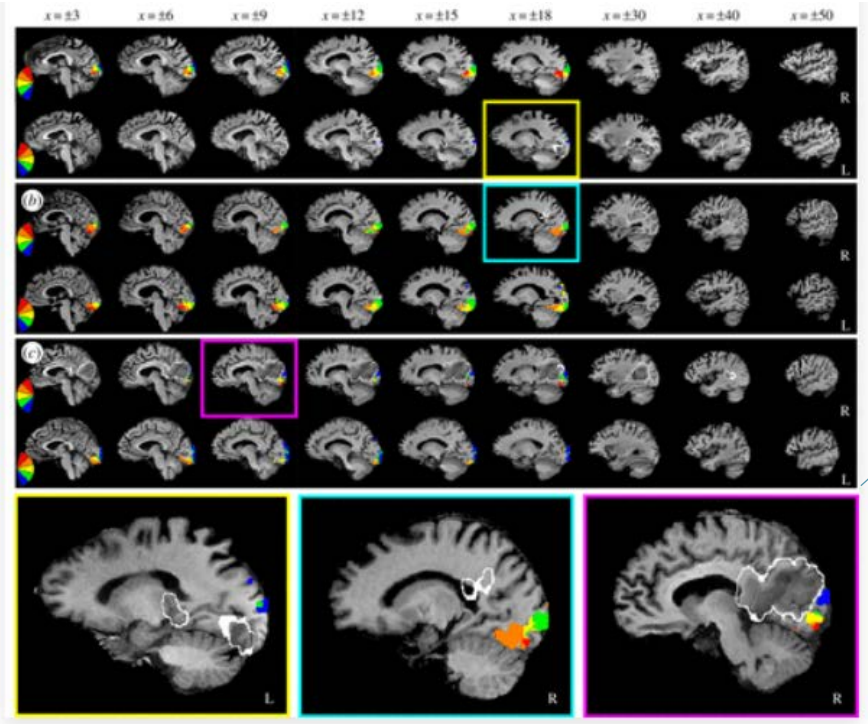


Figure 1. Overview of key measures. (*a*) Example measures from participant 5 collected at the final time point. Winner map of fMRI activity to flickering checkerboard wedges (stimulus example shows random order, lesion outlined from clinical T2 FLAIR or diffusion-weighted image *DWI shown in white; left panel), GCC thickness averaged over both eyes

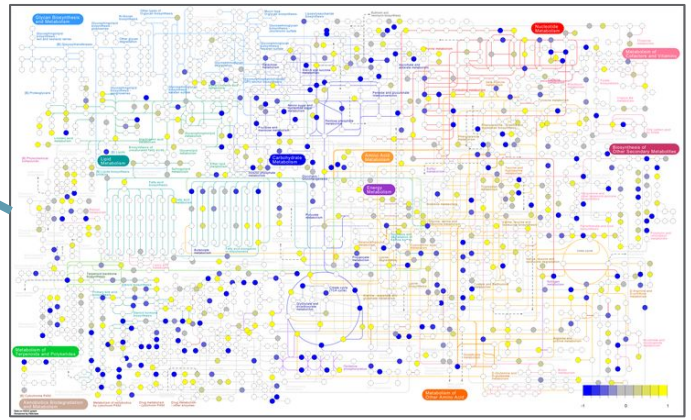
ParticipantID	fMRI	Behavior	wedge	Patient_Age	TimePoint	deltaTofscan	nVoxTC_cont	deltaTofOCT	sector	MacularT	deltaTofHum	sensitivity	total_dev
1	365	86	1	55	2	NaN	NaN	NaN	5	NaN	63	20.17	-9.5
1	365	86	2	55	2	NaN	NaN	NaN	4	NaN	63	25.5	-5.1666667
1	365	86	3	55	2	NaN	NaN	NaN	3	NaN	63	26.17	-3.3333333
1	365	86	4	55	2	NaN	NaN	NaN	2	NaN	63	28.67	-2.5
1	365	86	5	55	2	NaN	NaN	NaN	1	NaN	63	27.5	-3.6666667
1	365	86	6	55	2	NaN	NaN	NaN	12	NaN	63	26	4.8333333

What if:

- Journal articles could link to repository data sets
- Metadata were computable so that a search for similar datasets was possible
- Analysis tools were linked to datasets, via Github, Bioconductor, Galaxy or other….

IMAGINE…

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

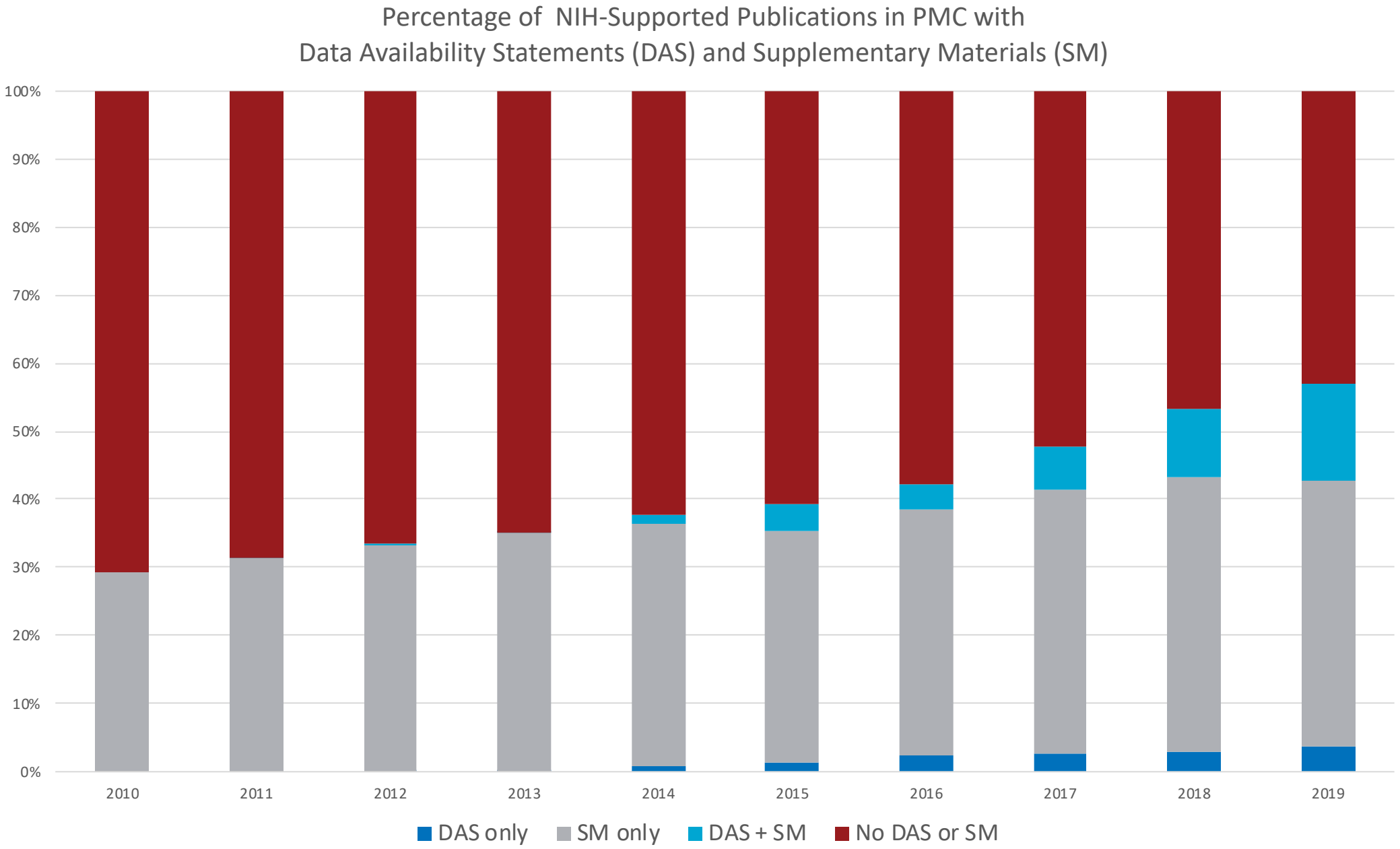


This is the promise of the *NIH Strategic Plan for Data Science*

…and here's how we will get there.

Making Data *FAIR*

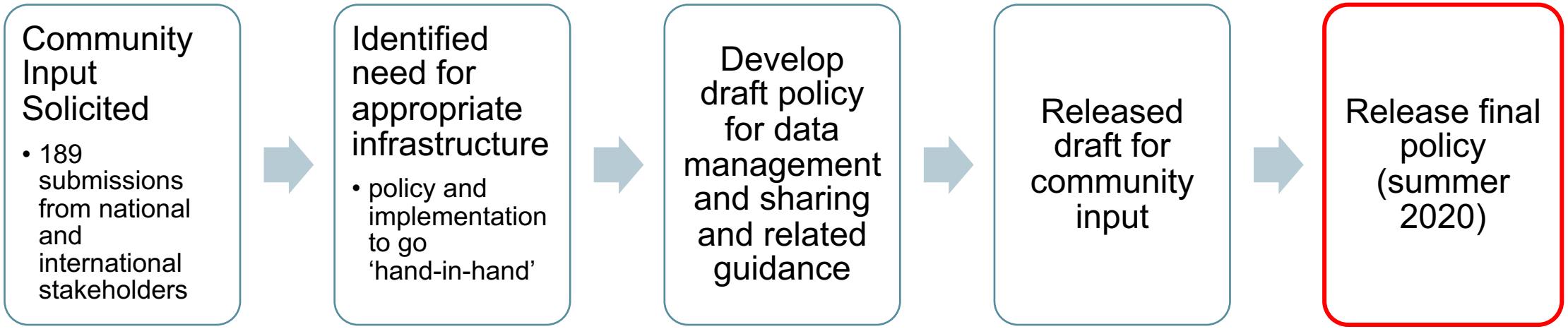
Findable	• must have unique identifiers, effectively labeling it within searchable resources.
Accessible	• must be easily retrievable via open systems and effective and secure authentication and authorization procedures.
Interoperable	• should “use and speak the same language” via use of standardized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



slide courtesy of Kathryn Funk

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**

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.

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

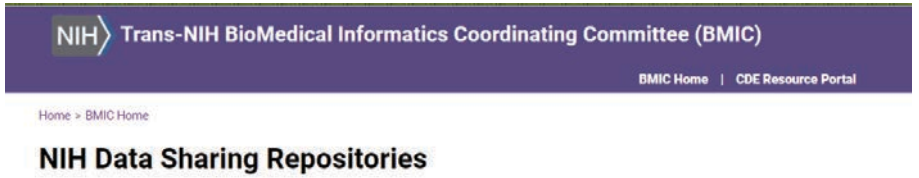
FITBIR
Federal Interagency Traumatic Brain Injury Research
INFORMATICS SYSTEM

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/

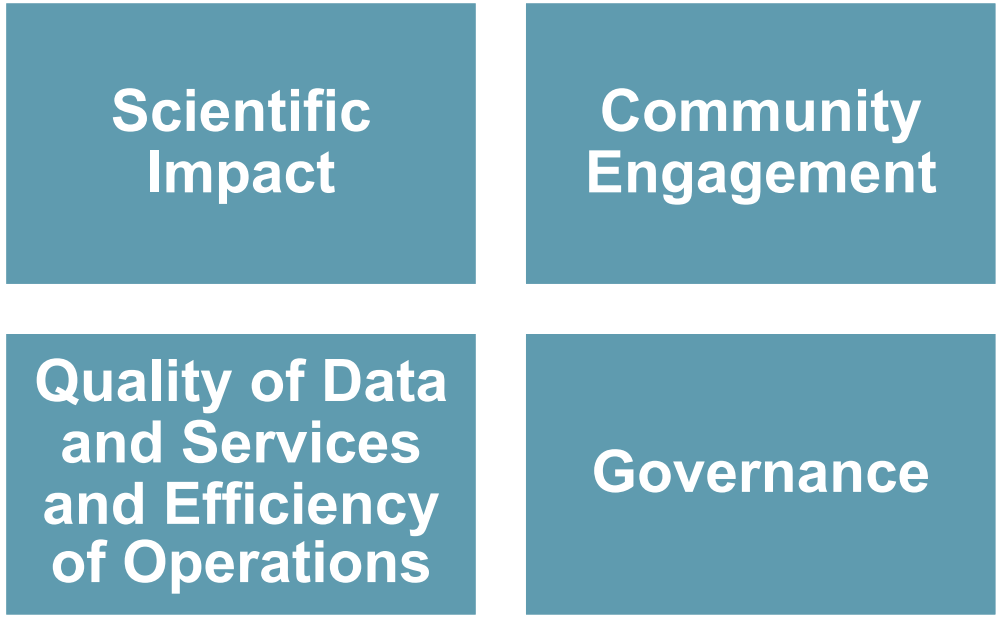
- **DataMed**

https://datamed.org/

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)



Co-Chairs: Maryann Martone, University of California, San Diego & *Shelley Stall*, American Geophysical Union

- Learn how generalist repositories see themselves in the larger biomedical data repository landscape.
- Understand how institutional data repositories are creating suites of solutions for their researchers and how they see generalist repositories fitting into this landscape.
- Consider desired characteristics of data repositories and how they relate to institutional expectations of data storage and preservation solutions.
- Explore adoption of common infrastructure, standards, and federated search solutions to enable greater discoverability of NIH research data across federated data repositories.
- Address the role of data curators in ensuring that data and metadata are sufficiently well curated to enhance discovery and enable reuse.

A few takeaways

- “Coopetition” – repositories collaborate on common goals while still maintaining competition on specific features – working together “below the value line”
- Discoverability requires both descriptive metadata and PIDs, with knowledge graphs being potentially powerful ways to discover objects at scale
- Librarians are well-situated and well-trained to advocate for, build awareness, educate, and provide tools to help w reproducibility
- Success requires culture change, not just new tech

Virtual Workshop on Data Metrics

- More than 300 people from around the world joined rich discussion focused on metrics
  - Co-chaired by Daniella Lowenberg, California Digital Library, and Warren Kibbe, Chief Data Officer at Duke Cancer Institute

- Next steps:
  - Continue engaging with community via Ideascale platform
  - Summary Report

Harnessing the power of the cloud

Turning Research Data Into Knowledge and Discovery

The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative

- 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 aws Google Cloud and other commercial 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.

STRIDES highlights by the numbers

13
NIH ICs participating

16
Extramural Participating Institutions

124
Programs/projects onboarded

>400
People trained

45
Petabytes stored

7.1M
Total compute hours

$3.9M
Dollars saved (cost avoidance)

Supplements to Enhance Software Tools for Open Science

Notice of Special Interest (NOSI) NOT-OD-20-073.

- Supplements to Support Enhancement of Software Tools for Open Science
- Up to $150K direct costs. Supplement applications **due May 15**
- Must address data science goals but in scope of parent grant
- 20 Institutes and 3 centers (FIC, NCATS, NCCIH)
- R01, R03, R21, U01, R00, R33, R35, R37, R61
- *Contact*: Jess Mazerik, ODSS

Goals

- Enhance software engineering of valuable scientific tools
  - New collaborations between biomedical researchers & software engineers
- Make research tools “cloud-ready”
  - Working with STRIDES initiative is encouraged but not required

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
- NHGRI AnVIL Platform
- Gabriella Miller Kid's First

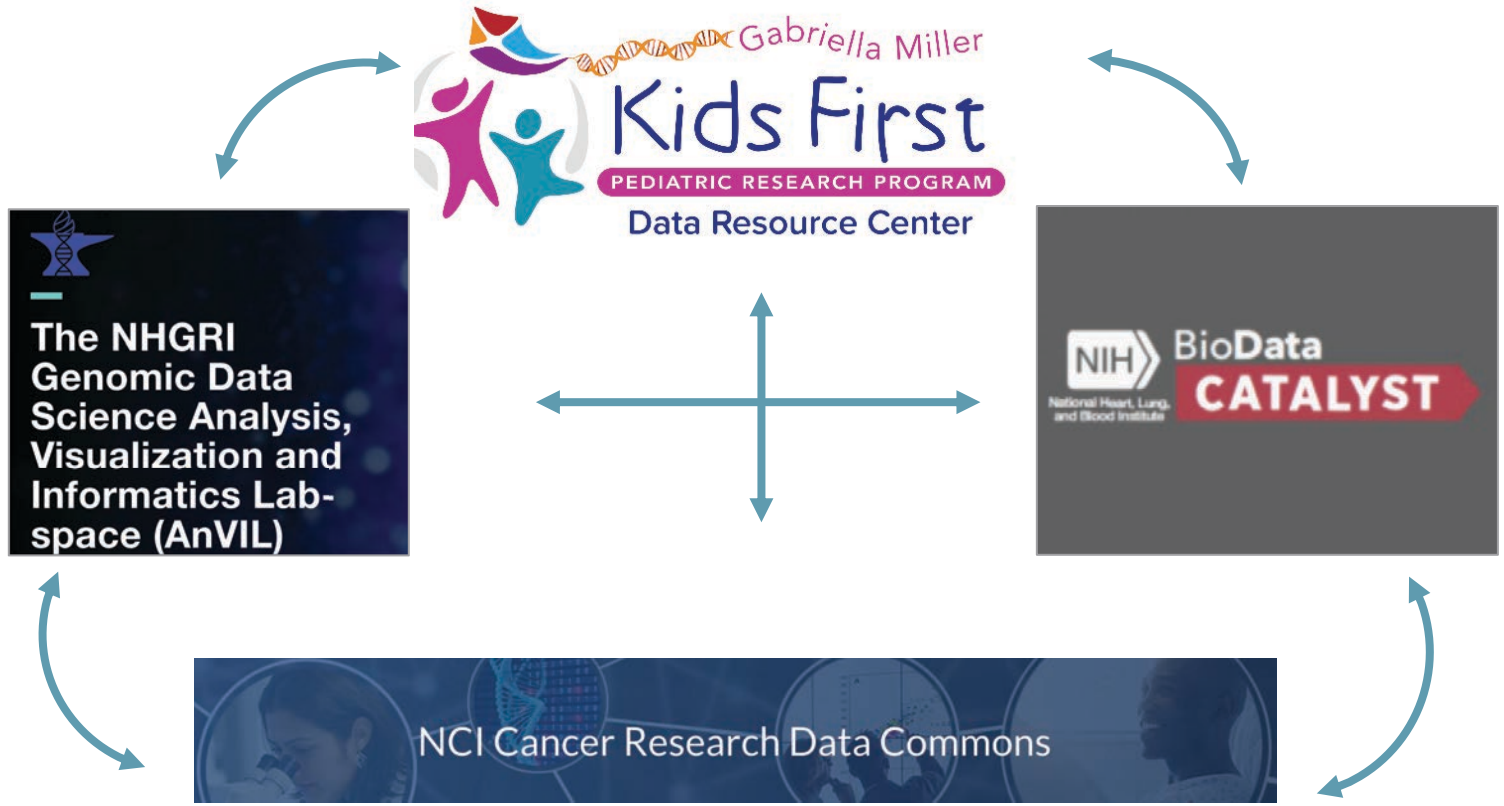
And Many More

- **Moved over 45 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 well over 50 PB of data in the cloud
  - We can search across this amount of data using advance AI algorithms

Connecting our data platform resources

NIH leverages Platform Technologies

NIH creates infrastructure that integrates data storage, data management, and computational tools in a single cloud environment, taking full advantage of commercial platform providers who are expert in this space



DNAnexus

SevenBridges

CYVERSE

Terra

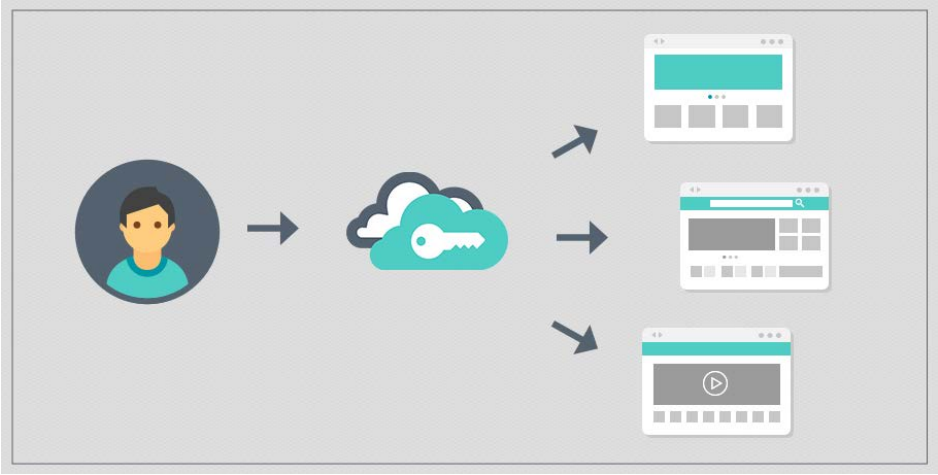
GEN3 Data Commons

Palantir

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



NIH Researcher Auth Service: GOALS

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

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.

Integrating healthcare data and clinical data

FHIR® Standard and Application Program Interface

Fast

Healthcare

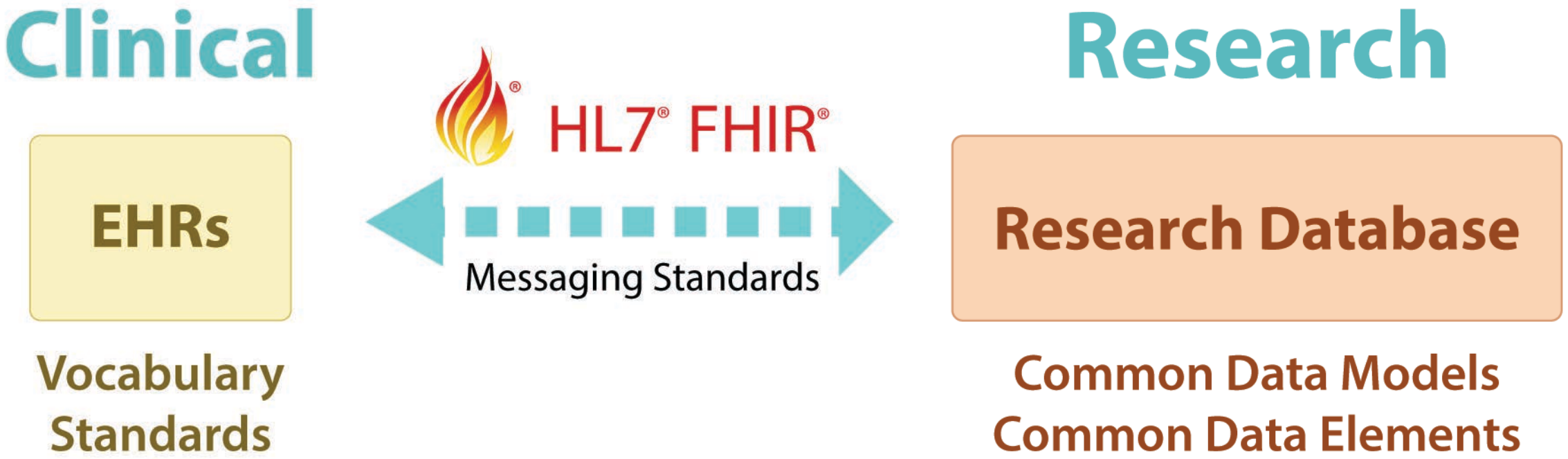
Interoperability

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

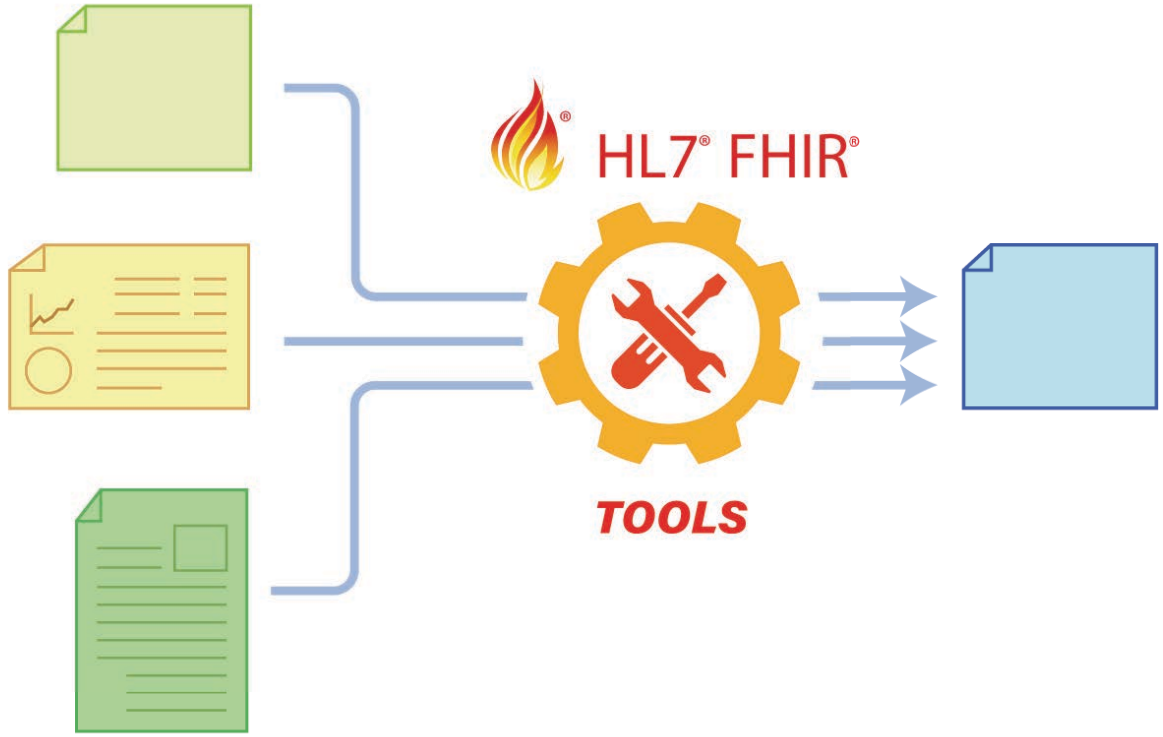
HL7 FHIR

Sharing Clinical Data for Research Using FHIR



Increasing Availability of High-Quality Data Using FHIR

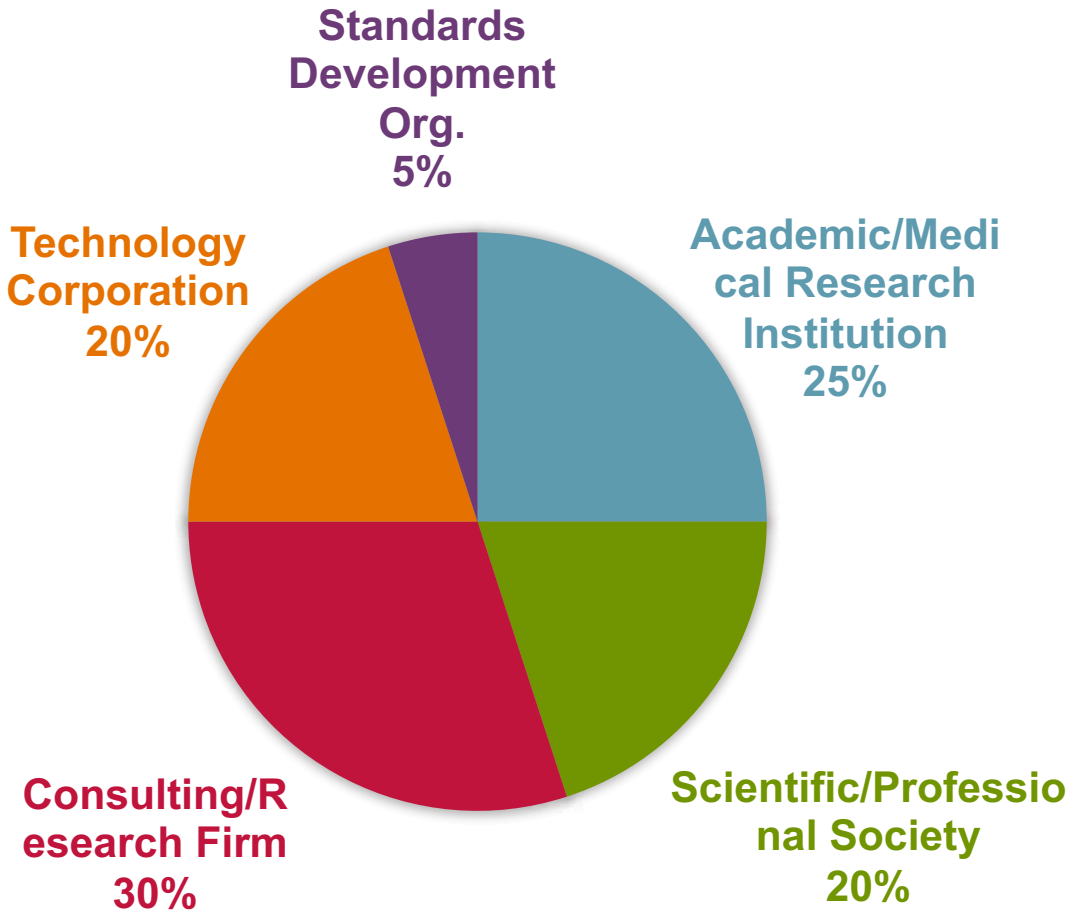
- Two pilot projects
  1. Development and Testing of FHIR Tools for Researchers
  2. Advancing Exchange of Phenotypic Information in Genomic Interpretation through FHIR



https://datascience.nih.gov/news/FHIR-awards-announcement-high-quality-data

Request for Information

- Overall support for NIH's direction
- Noted challenges in using FHIR "as is" for research
- Recommendations regarding
  - Engagement
  - Tool development
  - R&D
  - Related policies



Let’s create a bright future

Coming in 2020…

Coding it Forward

- 19 fellows from 13 universities across 13 institutes, centers, offices
- Masters' and undergraduate
- Majors in information systems, cybersecurity, poly sci, stats, data science, math, chem, public policy, computer science, electrical engineering, quantitative social science, biological science, information science, econ

ON HOLD: Graduate Data Science Summer Program

- 13 Masters' level fellows from 13 universities in labs across 6 institutes and centers
- Majors in math, biostats, data science, health informatics, nursing, bioengineering, bioinformatics, public health

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

National Institutes of Health
Office of Data Science Strategy

Applications due April 30

- 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

- 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

DATA Scholars Potential Projects

Catalyze neuroscience research by moving high-performance brain imaging analysis to the cloud	Unraveling the Alzheimer's Disease Genome using Artificial Intelligence, Machine Learning, and Deep Learning	Supporting Cancer Knowledge Extraction through the Cancer Research Data Commons	Accelerating the Clinical Adoption of Machine Intelligence Applications in Medical Imaging	Harnessing Data Science for Health Discovery and Innovation in Africa	Expanding Theories of Brain Circuits Using Knowledge Integration
NIMH	NIA	NCI	NIBIB	FIC	NINDS

Advancing Interoperability for Environmental Health	Broadening the Impact of Data Resources for Heart, Lung, Blood and Sleep Research	Innovative Solutions for Data Harmonization, Mobile Analytics, and End-User Support	Interoperability of NIH Cloud-Based Platforms for Genomics Research	Architecting search across petabyte-scale genomic sequence
NIEHS	NHLBI	OD OSC	NHGRI	NLM

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	

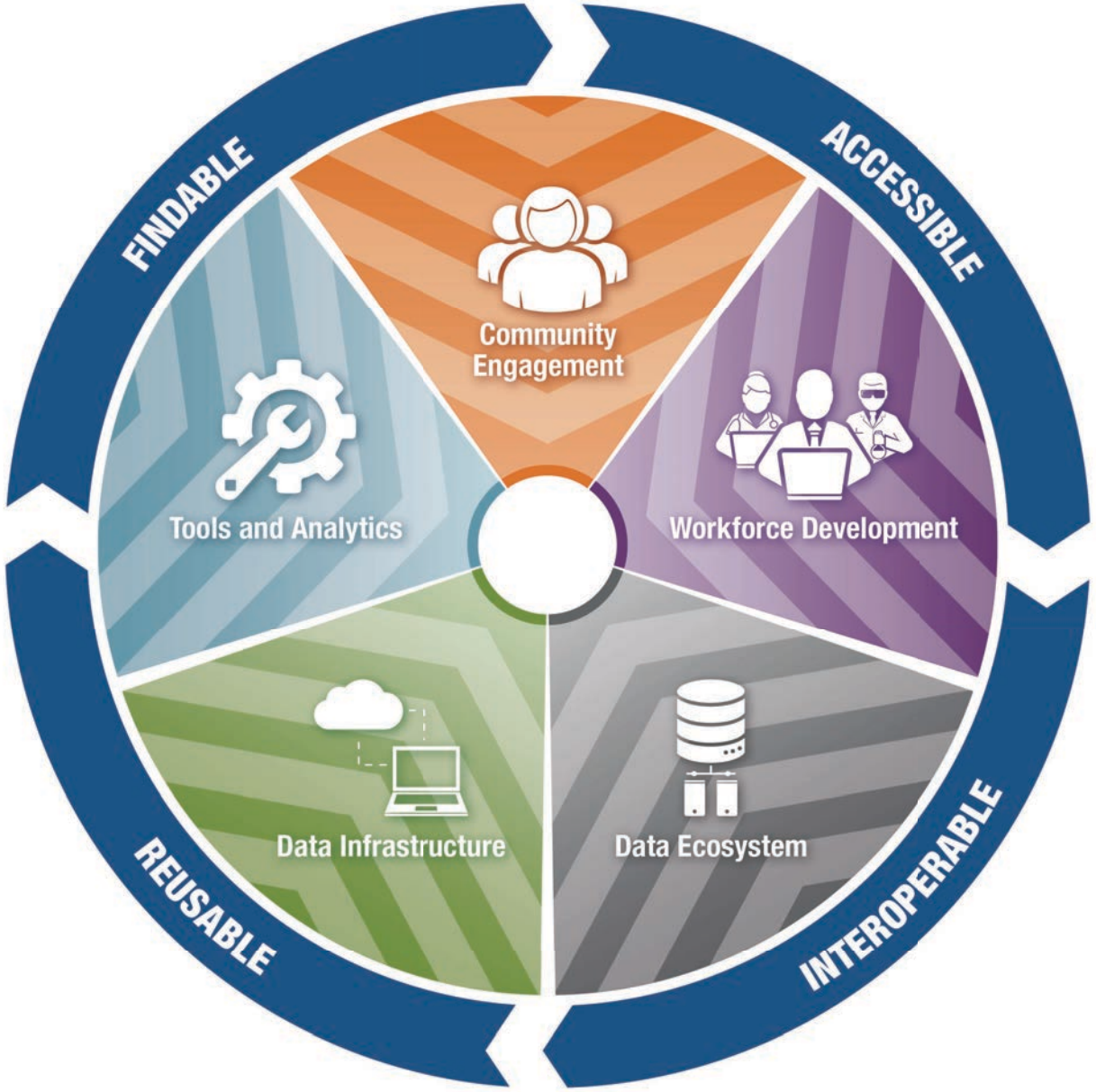
VISION

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

Office of Data Science Strategy

www.datascience.nih.gov

A modernized, integrated, FAIR biomedical data ecosystem



NIH staff who deserve all the credit

- **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, Allen Dearry
- **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, Denise Warzel, Kerry Goetz, Ken Wiley, Alison Cernick, Kenith Wilkins, Carolina Mendoza-Puccini, Matt McAuliffe, 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, Lisa Federer, Maryam Zaringhalam, and Jennie Larkin
- **Coding-it-Forward Fellows Summer Program & DATA Scholars Program:** Jess Mazerik, Wynn Meyer
- **Graduate Data Science Summer Program**: Sharon Milgram and Phil Ryan

Stay Connected

@NIHDataScience

/NIH.DataScience

www.datascience.nih.gov

National Institutes of Health
Office of Data Science Strategy