Session 4:
Considerations for Data Generated through the HEAL Initiative

NIH National Institutes of Health
Turning Discovery Into Health

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

Susan Gregurick, Ph.D.

Senior Advisor

Office of Data Science Strategy

May 17, 2019

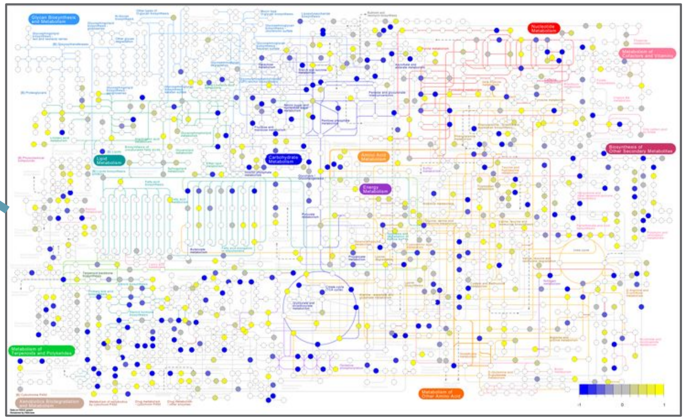
NIH National Institutes of Health
Office of Data Science Strategy

VISION

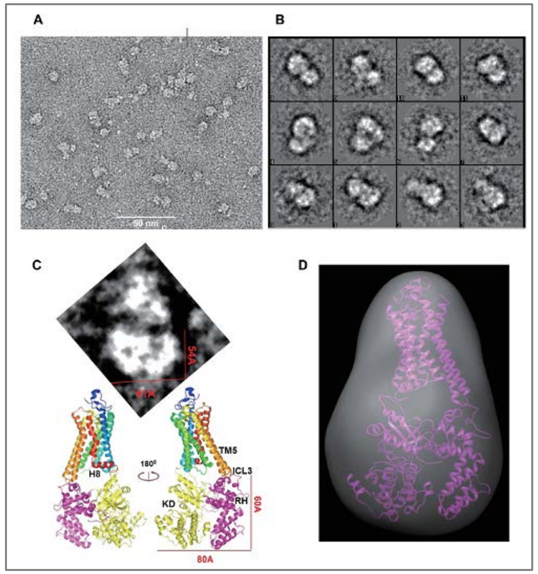
a **modernized, integrated, FAIR** biomedical data ecosystem

IMAGINE…

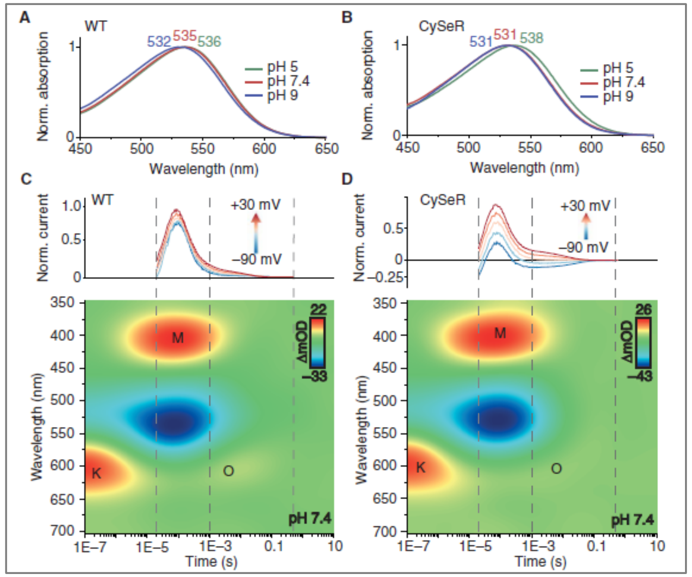
the ability to link electronic health care records with personal data and with clinical and basic research data.



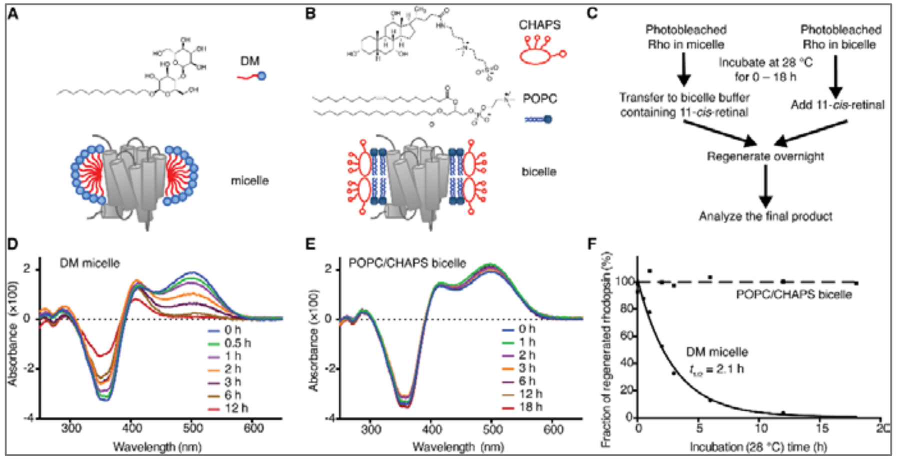
IMAGINE… the ability to quickly obtain access to data, and related information, from published articles.



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



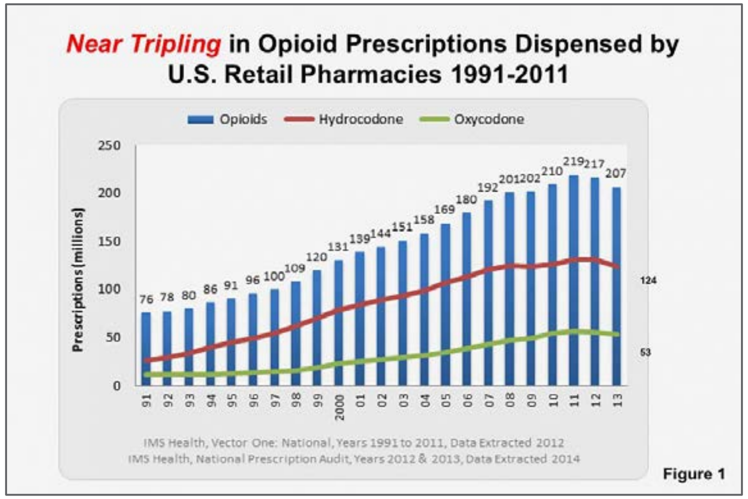
Absorption spectra of purified CsR-WT (A) and CySeR (B) at pH 5 (green), pH 7.4 (red), and pH 9 (blue). R. Fudim, e al, Science Signaling, 2019



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

IMAGINE…

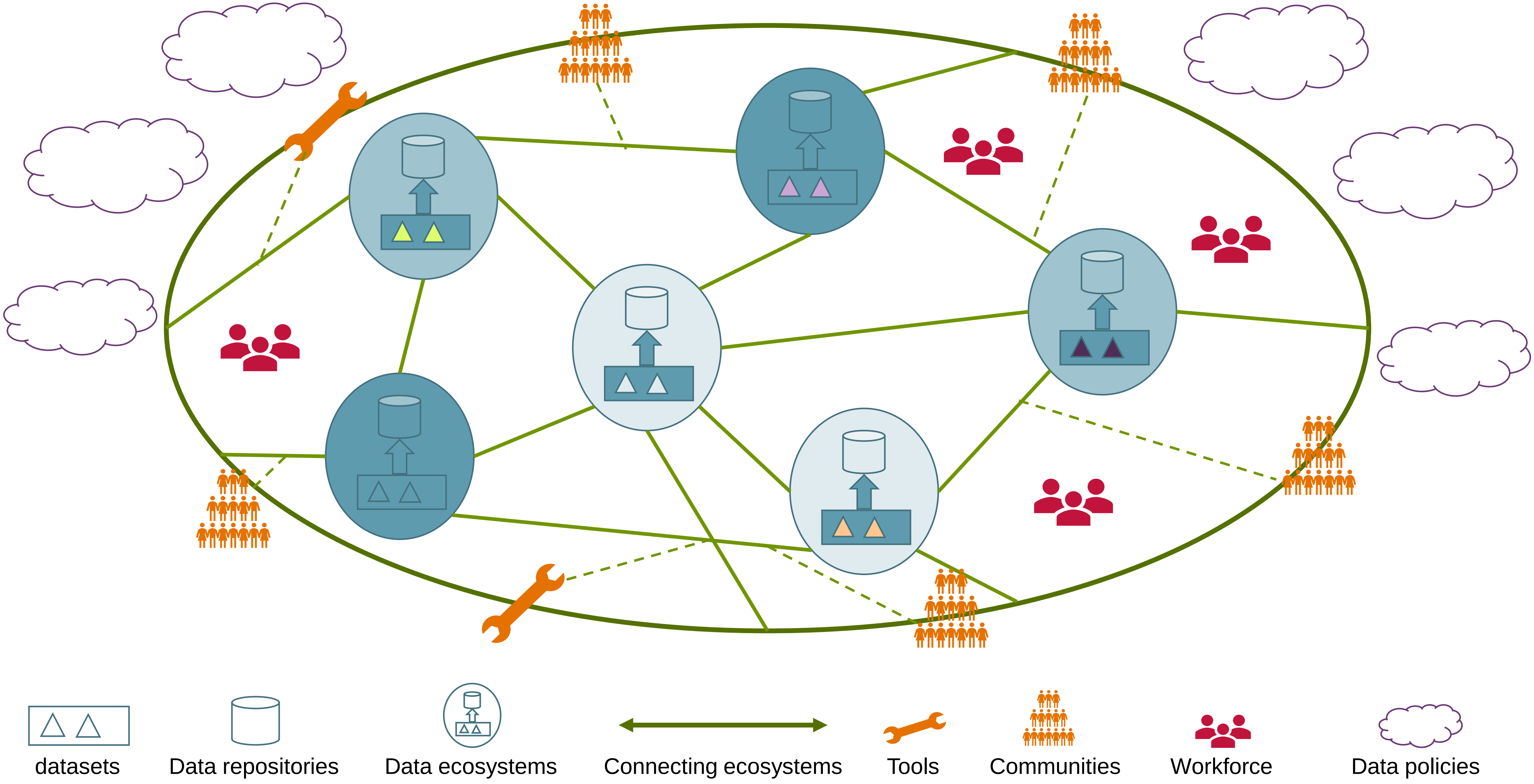
the ability to link data in the HEALing Communities Study with data on opioid prescribing practices and measures of opioid use in other HEAL studies.



Near Tripling **in Opioid Prescriptions Dispensed by U.S. Retail Pharmacies 1991-2011**

IMS Health, Vector One: National, Years 1991 to 2011, Data Extracted 2012
IMS Health, National Prescription Audit, Years 2012 & 2013, Data Extracted 2014

Figure 1



datasets

Data repositories

Data ecosystems

Connecting ecosystems

Tools

Communities

Workforce

Data policies

This is the promise of ***Data Science at NIH***

…and here’s how we will get there.

Recent Progress Toward NIH's Vision for Data Science

datasets Data repositories Data ecosystems

- Link datasets to publications (PubMed)
- Provide FAIR-enabled, open-access options for datasets that underly a publication resulting from NIH funded research
- Supporting data repositories and knowledgebase resources
- Develop criteria for open-access NIH data sharing repositories

↔

Connecting ecosystems

- High-priority datasets moved to cloud service providers (CSPs)
- Single method for sign-on and data access across repositories and CSPs

Tools Communities

- Engaging with a broader community
  - National Science Foundation partnership
  - SBIR/STTR utilization
  - Hackathons, bug bounties, citizen science challenges
  - Software sustainability extension through hardening

Data policies

- Data management and sharing policy for NIH

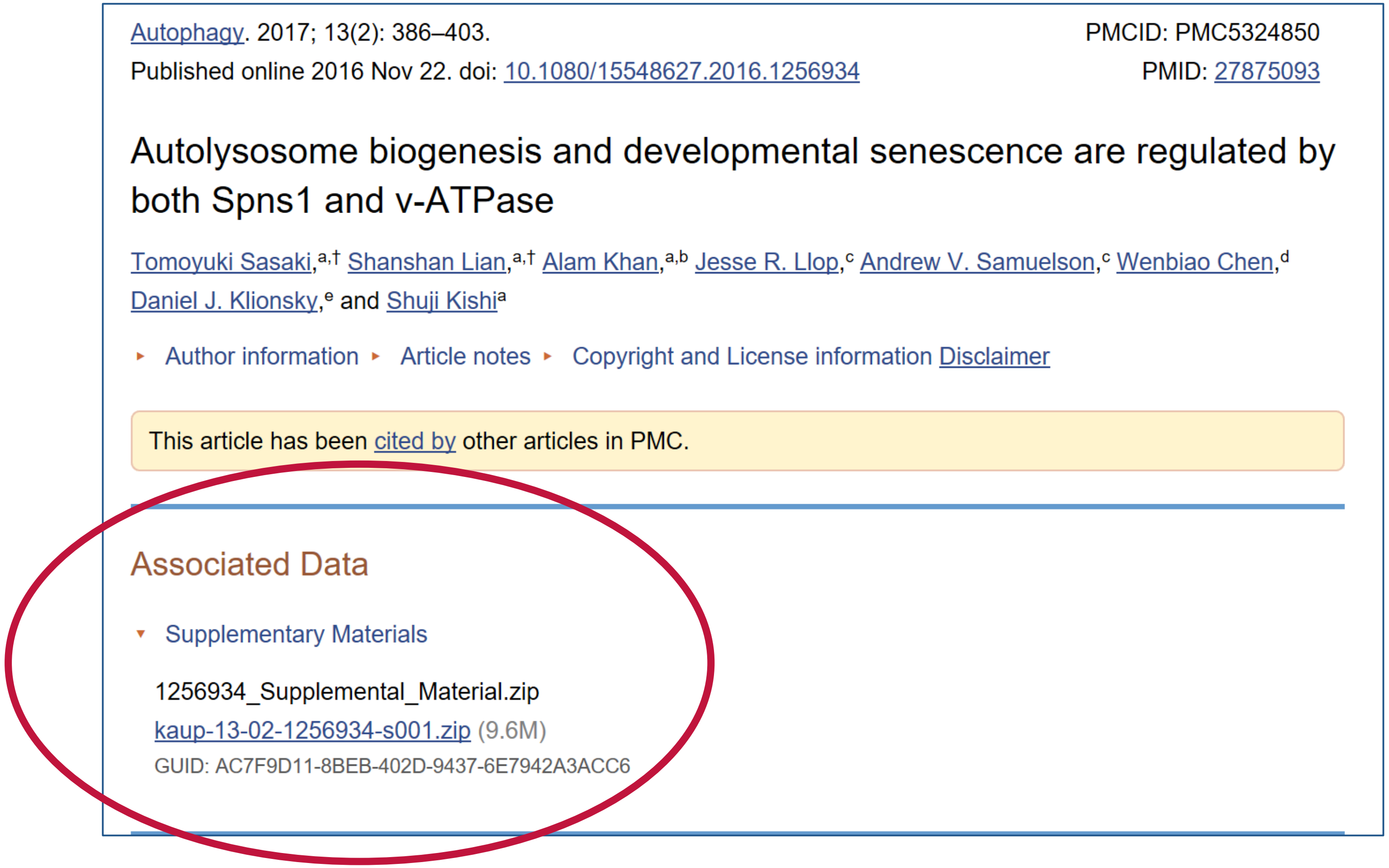
Workforce

- Enhancing biomedical workforce through internships
  - Coding it Forward
  - Graduate Data Science Summer Program
  - NIH Data Science Senior Fellowships

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.

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

Link 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

NIH recommends domain-specific repositories when available.

Provide FAIR-enabled, open-access options for datasets that underly a publication resulting from NIH funded research

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)

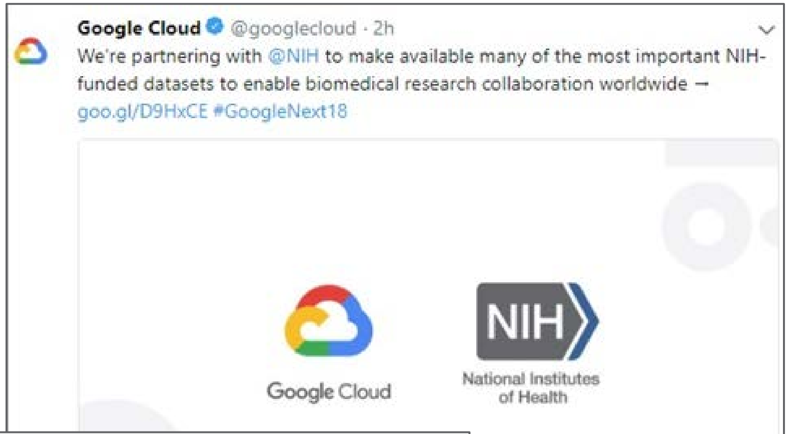
- Characteristics drafted, includes provisions for repositories with human data
- Developed and reviewed in trans-NIH process
- Planned Community Input: Request for Information (RFI)

Develop criteria for open-access
NIH data sharing repositories

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

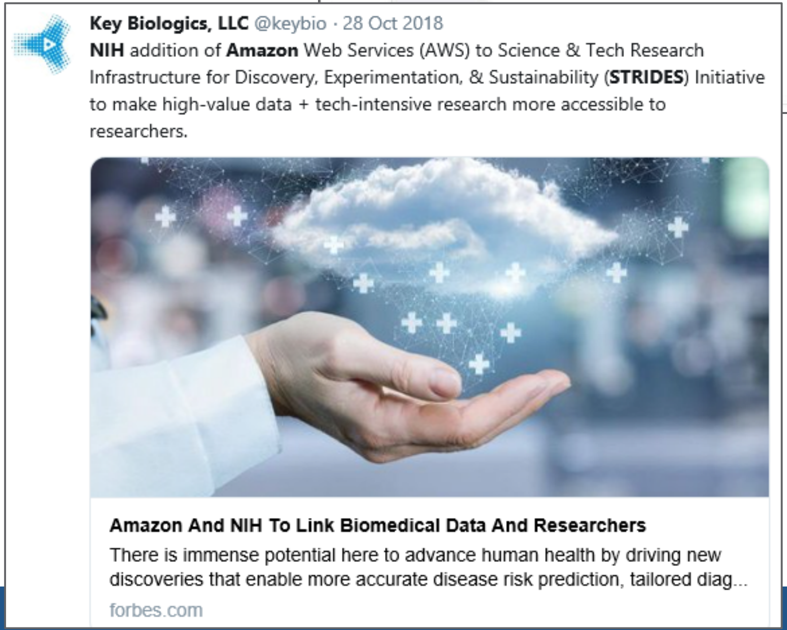
- First **STRIDES** agreement: Google Cloud (July 2018)
- Second **STRIDES** agreement: Amazon Web Services (Oct. 2018)
- Other Transaction mechanism
- Additional partnerships anticipated

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

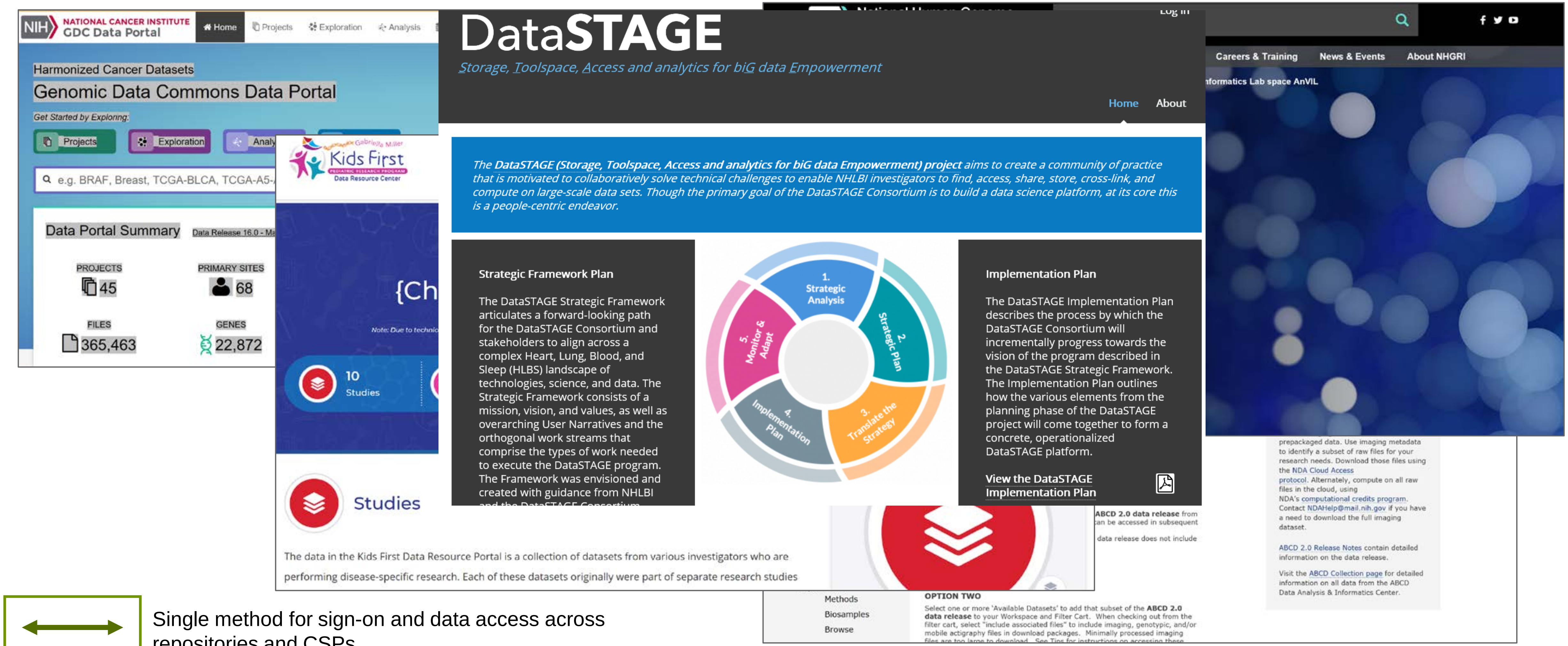
↔ Move/Access to high priority data sets in cloud service providers

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!**

↔ Move/Access to high priority data sets in cloud service providers

NIH's Data Environments are Rich, but Siloed



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Data Portal Summary

Data Release 16.0 - Ma

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

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

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

ABCD 2.0 data release from ... can be accessed in subsequent ... data release does not include

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.

OPTION TWO

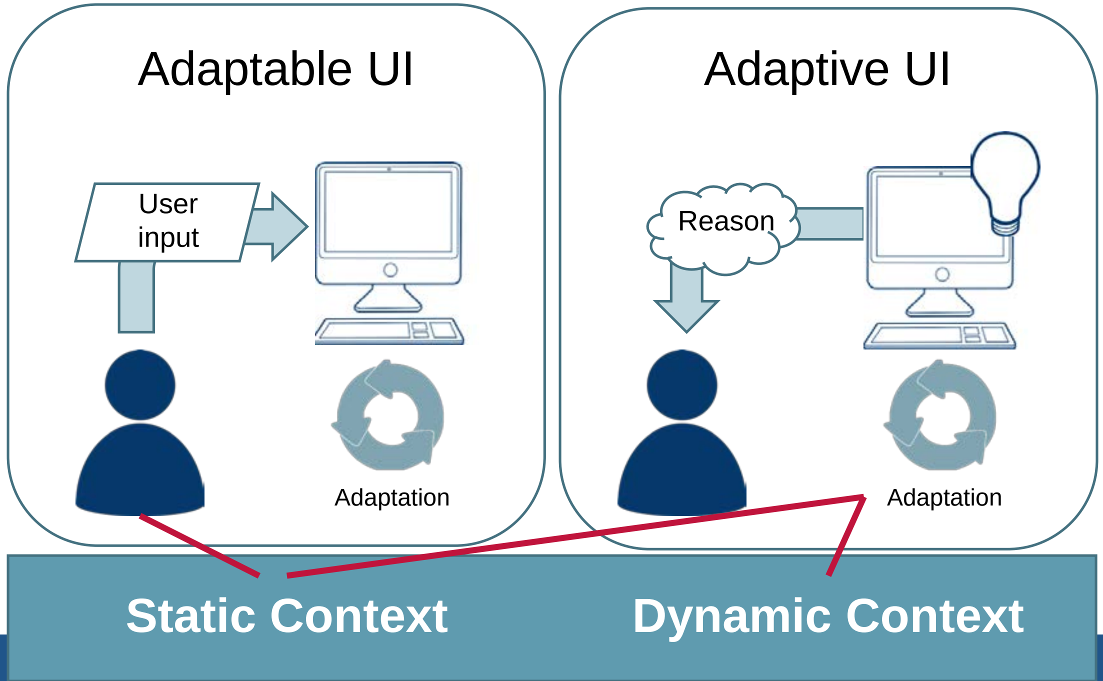
Select one or more 'Available Datasets' to add that subset of the **ABCD 2.0 data release** to your Workspace and Filter Cart. When checking out from the filter cart, select "include associated files" to include imaging, genotypic, and/or mobile actigraphy files in download packages. Minimally processed 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



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

Principles for Data Sharing and Open Access in HEAL Research

Rebecca Baker, Ph.D.

Director, HEAL Initiative

Office of the Director, NIH

May 17, 2019

NIH National Institutes of Health

Considerations for HEAL Data

HHS has declared the national opioid crisis a public health emergency

Many HEAL projects are funded through cooperative agreements

Plans for a central data repository for HEAL

HEAL should leverage ongoing data science innovations at NIH

Maximizing the Utility of HEAL Research Data

- **Goal: Simple and FAIR data through HEAL**
  - Publications and underlying research data should be made available
    - Any file format
    - Assign an institutionally (NIH) branded DOI
    - Central HEAL or other data repository
  - Documented with customizable, discipline-specific metadata
    - Enabling research across different HEAL projects
  - Discoverable content across major search engines and frameworks

Data Sharing Policy Landscape at NIH

- Projects with budgets > 500K direct costs must submit a plan for data sharing in their applications

- Special considerations for certain types of data and projects, e.g. genomic data, Cancer Moonshot

- Publications resulting from NIH-funded research must be deposited into PubMed Central no later than one year after publication

Plan for Open Access to HEAL-Funded Publications

Incorporate into terms and conditions of certain awards:

Rapid deposition of electronic copies of publications in PubMed Central with proper tagging of metadata.

Publications will be published under the Creative Commons Attribution 4.0 Generic License (CC BY 4.0) or an equivalent.

Publications will be made publicly available immediately with no embargo period.

Underlying primary data for the publications will be made broadly available through an appropriate data repository such as the HEAL central data repository.

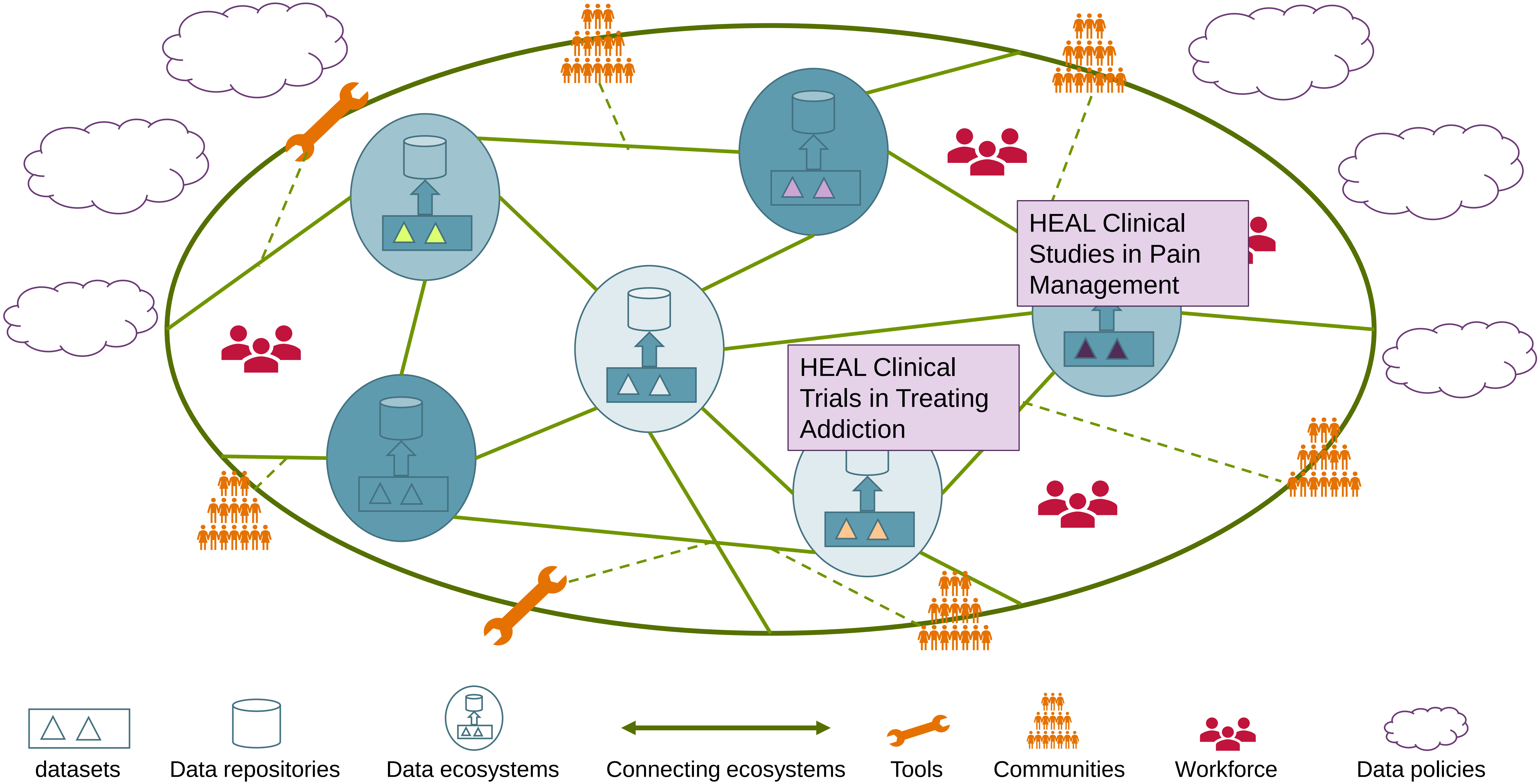
To the extent feasible, underlying primary data will be shared simultaneously with the publication and made immediately accessible.

Implementation Plans for HEAL Data Sharing Policy

- Some awards will need to wait until FY2020

- Broad and responsible sharing of data that protects and maintains privacy and confidentiality

- Investigators required to plan for protecting and maintaining privacy rights of participants and confidentiality



HEAL Clinical Studies in Pain Management

HEAL Clinical Trials in Treating Addiction

datasets

Data repositories

Data ecosystems

Connecting ecosystems

Tools

Communities

Workforce

Data policies

Leveraging NIH Data Science Opportunities for HEAL

HEAL Central Data Repository	→	Characteristics for NIH-supported data repositories
Storage of HEAL data	→	STRIDES program
HEAL data ***not*** in the HEAL Central Data Repository	→	Figshare
“Protected” HEAL data	→	single sign-on system

NIH • Helping to End Addiction Long-term