National Institutes of Health
Office of Data Science Strategy

Going Beyond FAIR to Create a Connected Data Ecosystem

Susan Gregurick, Ph.D.

Associate Director for Data Science and
Director, Office of Data Science Strategy

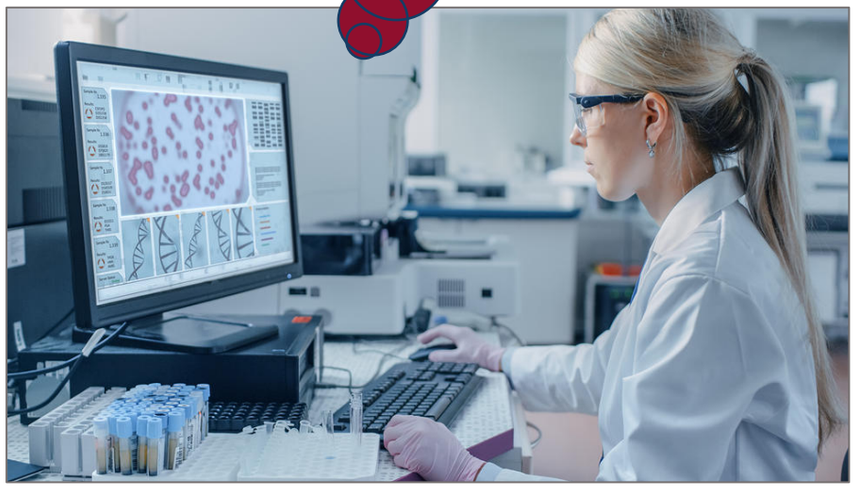
June 4, 2021

Some Compelling Use Cases

I'm writing my RO1 and need to submit a Data Management and Sharing Plan. How do I put the FAIR Principles into practice, where do I start, and what do I do?

Credit: iStock

Studying rare diseases like pediatric cancers are especially challenging because no single source has enough data to allow identification of causative variants on their own and research participant may be duplicated across systems.



National Cancer Institute, Credit: iStock

Data Scientist or AI Engineer? AI has great potential to help scientists make sense of the vast quantities of data being generated by modern instruments.

Credit: Albert Christopher, Medium, Sept 9, 2020

VISION: a modernized, integrated, FAIR biomedical data ecosystem

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

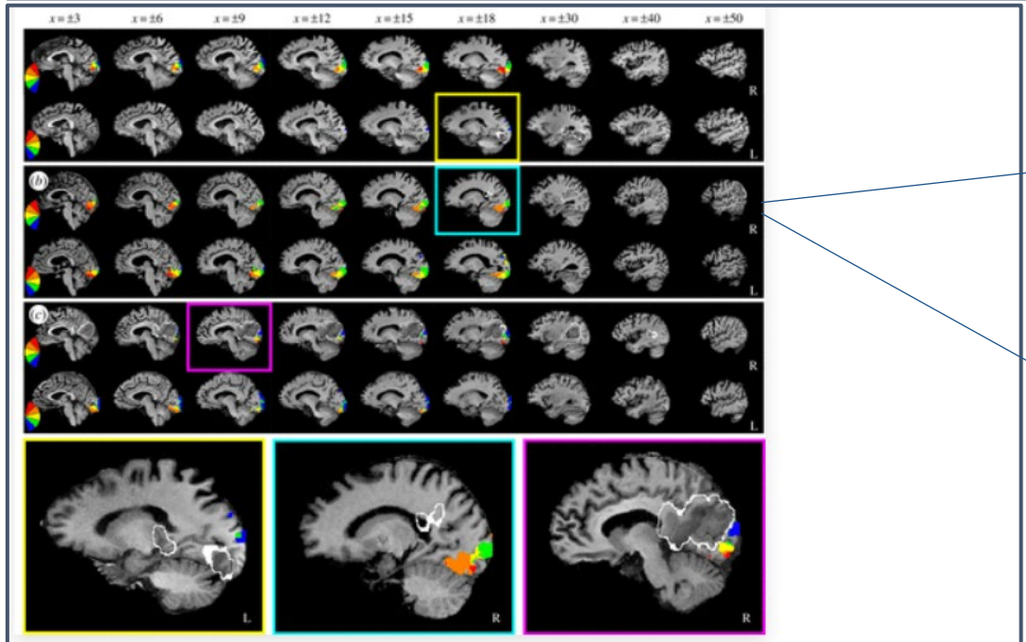


Figure 1. Overview of key measures. (*a*) Example measures from participant 5 collected at the final time point. Winner map of fMRI activity to flickering checkerboard wedges (stimulus example shows random order, lesion outlined from clinical T2 FLAIR or diffusion-weighted image *DWI shown in white;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

Journal articles could link to repository data sets, code, and interactive notebooks

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

Metadata were computable so that a search for similar datasets was possible

There Is Progress

Linking data in repositories embedded in corresponding journal publications or collections:

- Elsevier, Nature Scientific Data, PLOS, Dataverse Crossref, Datacite, researchgate, OCLC CONNECT pilot; related efforts such as bioCADDIE, PubMed, FAIRsharing WG (RDA) etc.
- Many of these efforts are working with a community of repositories for greater discoverability

Recommender (machine readable) Metadata for greater discoverability:

- RDA, FAIRsharing, Crossref, American Mathematical Society, Altmetric, Nature Scientific Data, NIH's NCPI, etc.
- Standards are emerging for minimal metadata for discoverability much like efforts to create common data elements and common data models

On the open-source landscape of *PLOS Computational Biology* *https://doi.org/10.1371/journal.pcbi.1008725*

- An emerging trend is to provide an easily reproducible data & coding environment MyBinder, Google Colab, NeuroLibre, and Code Ocean, interactive figures (e.g., Plotly and Bokeh), widgets (e.g., ipywidget), or dashboards (e.g., Dash and Shiny)

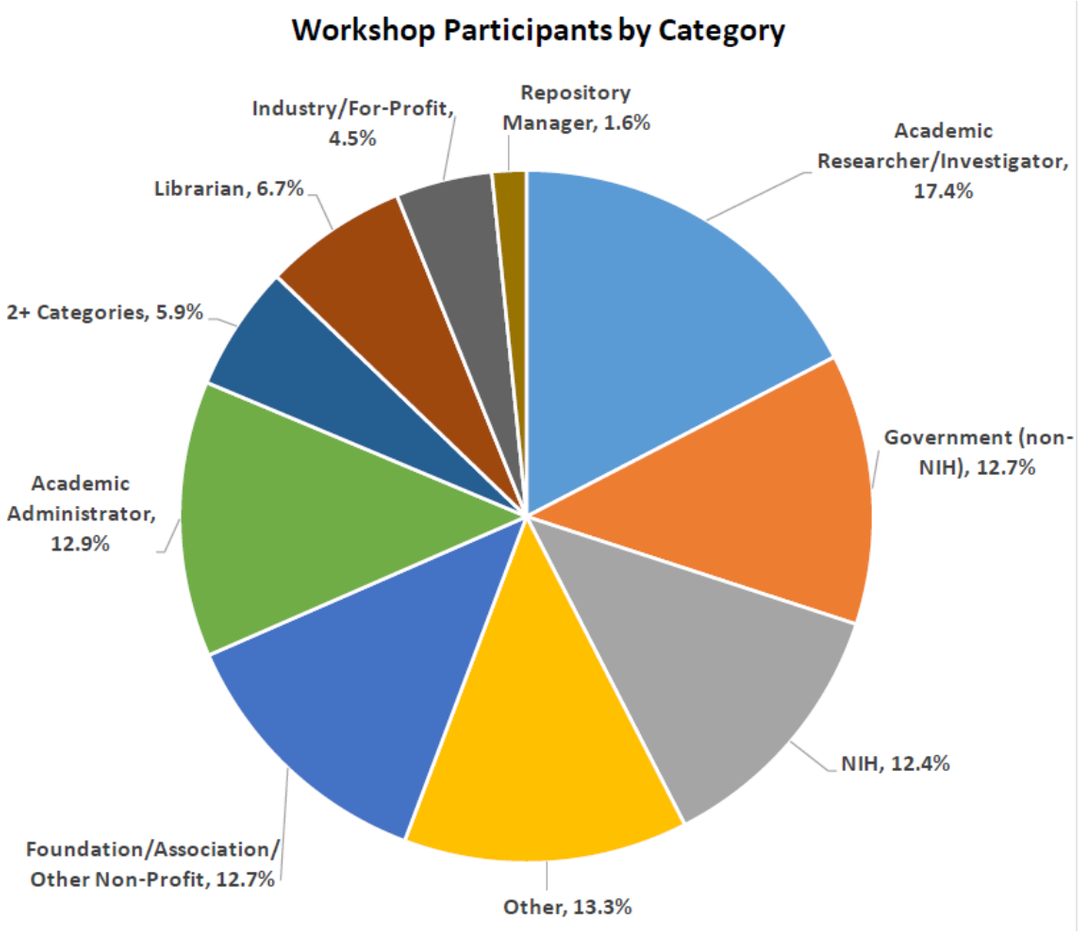
NIH Policy for Data Management and Sharing

- **Submission of Data Management & Sharing Plan for all NIH-funded research** *(how/where/when)*
- **Compliance with the ICO-approved Plan** *(may affect future funding)*
- **Effective January 25, 2023** *(replaces 2003 Data Sharing Policy)*
- **Supplemental info available to assist**
- **Aims to foster data stewardship**

NASEM Workshop on Changing the Culture of Data Management and Sharing, April 28-29

- Workshop aimed to…
  - Identify community training and resource needs
  - Learn about challenges anticipated in implementing policy expectations
  - Understand what the community envisions successful data sharing to look like and how to measure it

- **~2,000 engaged participants**

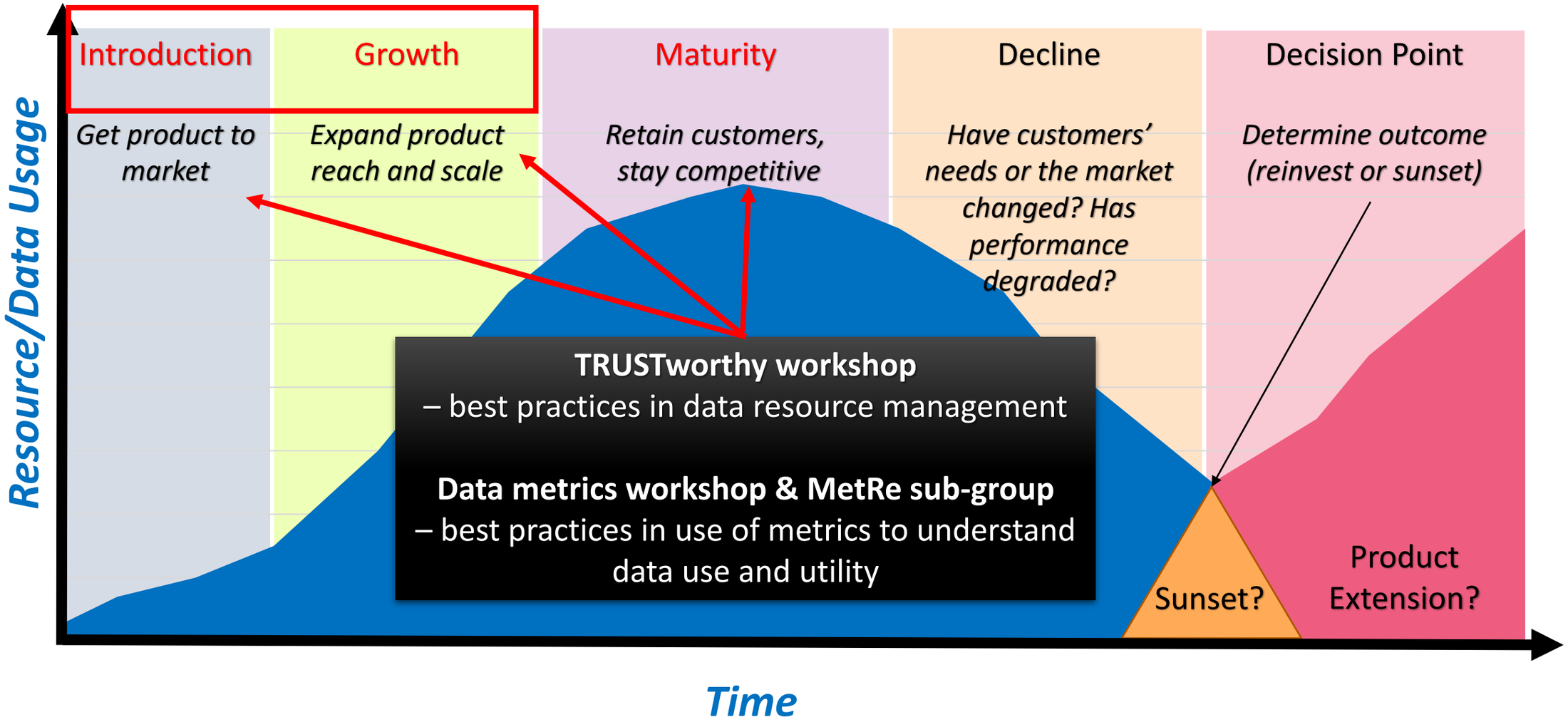


Key Points from the NASEM Workshop

- **Implementation requires a system-wide culture shift**
  - Need for aligned incentives and resources from NIH and all biomedical ecosystem stakeholders (e.g., other funders, institutions, publishers, data repositories, and associations)

- **Impactful data sharing is key to successful policy implementation**
  - Data management practices mindful of secondary data users are necessary for useful data sharing
  - Good data management (e.g., ensuring metadata accompanies data throughout the research lifecycle, QA/QC, and versioning) is the foundation for data sharing
  - Collecting and analyzing data sharing metrics are needed to understand the value of data sharing
  - Data citation adoption promotes data reuse and aids compliance monitoring

Data Resource Lifecycle: Focus on the Growth and Maturity Phase

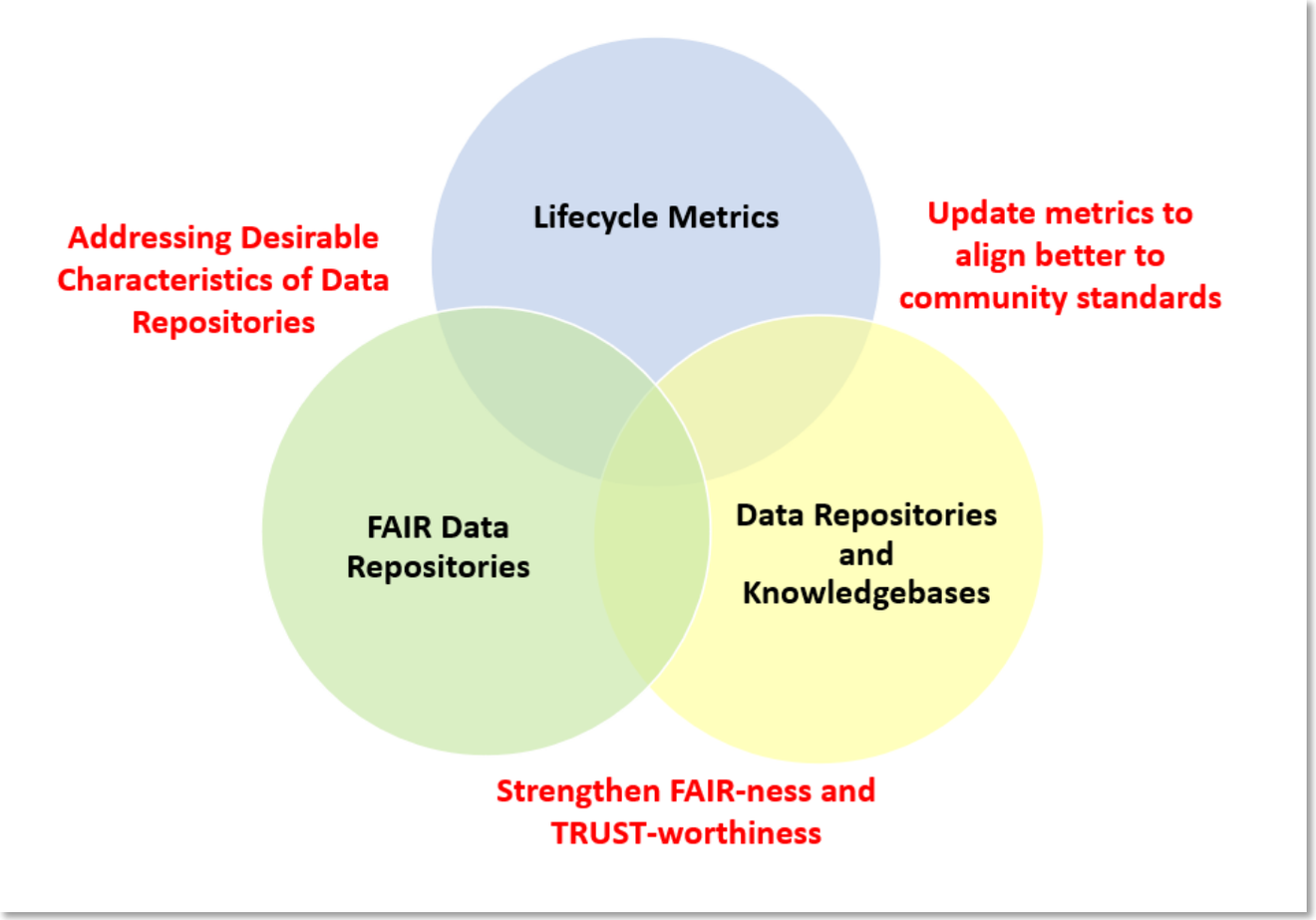


Credit: Kim Pruitt, NCBI

Positioning Repositories for Data Sharing

Challenge: It is critical for repositories to receive support to help them achieve the TRUST and FAIR principles and evaluate usage, utility, and impact via metrics to prepare them for data sharing requirements as per the Final NIH Policy for Data Management and Sharing (NOT-OD-21-073)

Opportunity: Develop an ecosystem that allows for a range of diverse repositories to provide different options for researchers as their data sharing needs evolve, which would leverage existing data and maximize assets



Supplements to Support Existing Repositories

Better enable **data discoverability, interoperability, and reuse**

Goals

- Implement the relevant portions of the Desirable Characteristics for Repositories (i.e., NOT-OD-21-016) with the aim to strengthen adoption of the FAIR principles
- Strengthen the adoption of the TRUST principles, which may include a structured plan to pursue certifications
- Adopt, enhance, or contribute to community-based metrics standards or best practices of metrics to evaluate the usage, utility, and impact of the data resource throughout its lifecycle

NOT-OD-21-089

Desired Characteristics for All Data Repositories

- Assigns unique persistent identifiers
- Plan for long-term sustainability
- Accompanies data with metadata
- Provides mechanisms for curation and quality assurance
- Free and easy access to data
- Broad and measured terms of reuse
- Provides clear use guidance
- Uses documented security and integrity measures
- Ensures confidentiality
- Uses common format
- Tracks provenance
- Provides retention policy

Desired Characteristics for Repositories Preserving and Sharing Human Participant Data

- Ensures fidelity to consent
- Compliant with data use restrictions
- Ensures participant privacy
- Plan for breach
- Controls and audits download
- Addresses violations of terms-of-use
- Uses request review process

From NOT-OD-21-016 Desirable Characteristics for Data Repositories

Improving Discoverability of Existing Resources and Improving Data Within These Resources

https://findwise.com/blog/data-that-really-saves-lives-and-possibly-your-organisation/

(*used with permission from Ingrid Dillo*)

The TRUST Principles

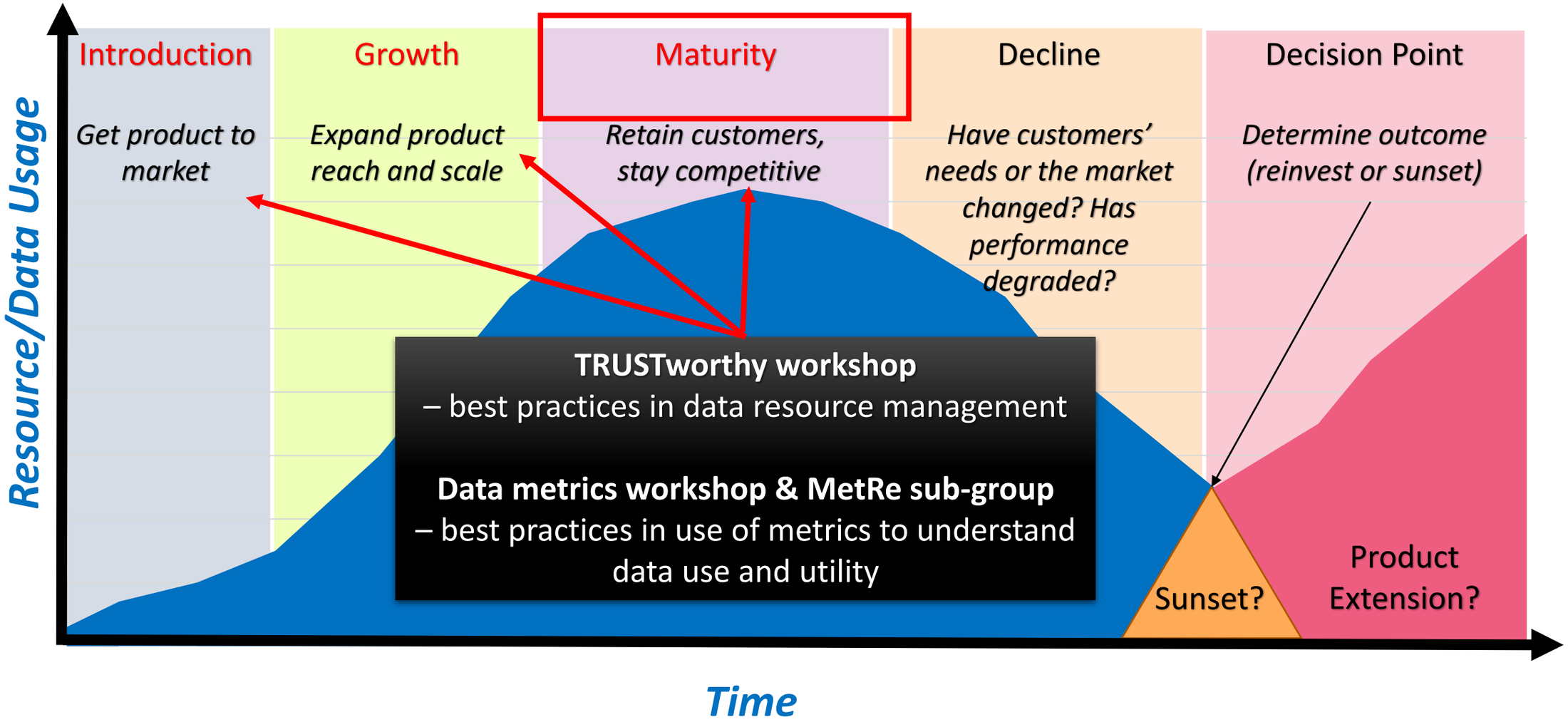
TRUST

Transparency · Responsibility · User focus · Sustainability · Technology

Principle	Guidance for Repositories
Transparency	To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.
Responsibility	To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.
User Focus	To ensure that the data management norms and expectations of target user communities are met.
Sustainability	To sustain services and preserve data holdings for the long-term.
Technology	To provide infrastructure and capabilities to support secure, persistent, and reliable services.

Source: Lin et al., 2020. The TRUST Principles for Digital Repositories. Scientific Data https://doi.org/10.1038/s41597-020-0486-7

Data Resource Lifecycle: Focus on the Growth and Maturity Phase



Credit: Kim Pruitt, NCBI

Optimized Funding for NIH Data Repositories and Knowledgebases

Funding Opportunities

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

Scientific Impact

Community Engagement

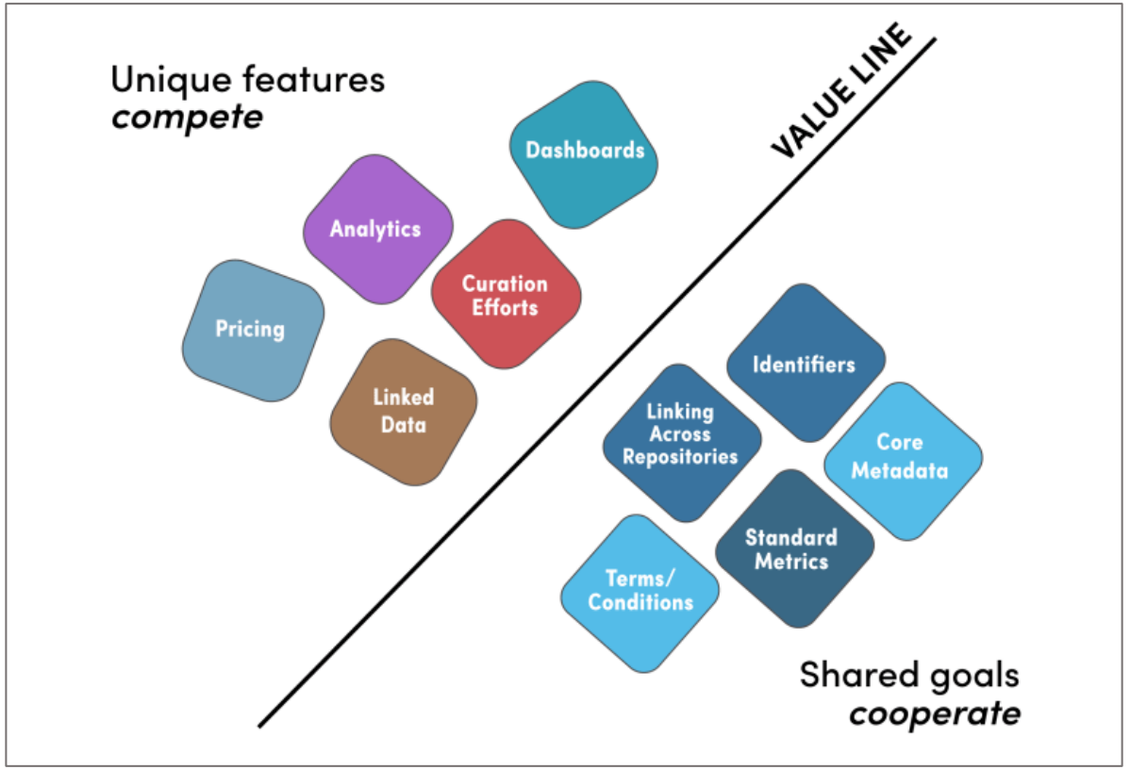
Quality of Data and Services and Efficiency of Operations

Governance

Coopetition: Creating an Ecosystem

The **value line** defines when members:

- ***Compete* on specific/unique features**
  - Dashboards
  - Visualization
  - Analytics
  - Linked data
  - Etc.
- ***Collaborate* on common/shared goals**
  - Core metadata
  - Identifiers
  - Authentication
  - Etc.



Community Workshop on the Role of Generalists Repositories, Feb 11-12, 2020
Co-Chairs *Maryann Martone and Shelley Stall*

National Institutes of Health
Office of Data Science Strategy

Data Sharing and Reuse Seminar Series

Diversifying the AD Target Pipeline: Data and Tool Sharing to Catalyze the Study of Understudied Drug Targets

Lara Mangravite, Ph.D.
President, Sage Bionetworks

Learn more and register @ bit.ly/NIHDataSeminars

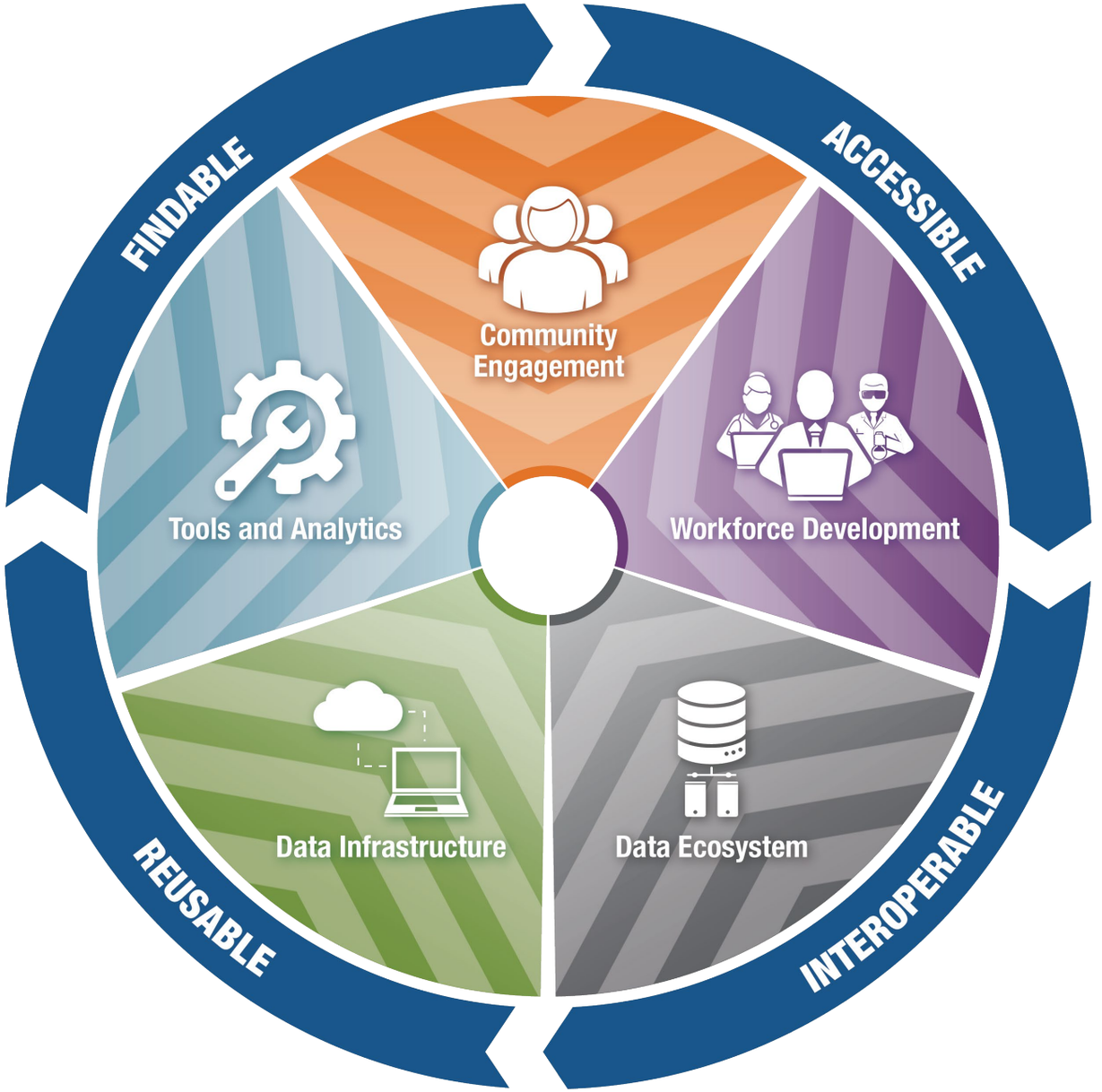
#NIHData

June 11, 2021
12:00 p.m. ET

Office of Data Science Strategy

www.datascience.nih.gov

A modernized, integrated, FAIR biomedical data ecosystem



Building a Data Science Workforce

Coding it Forward

CIVIC DIGITAL FELLOWSHIP

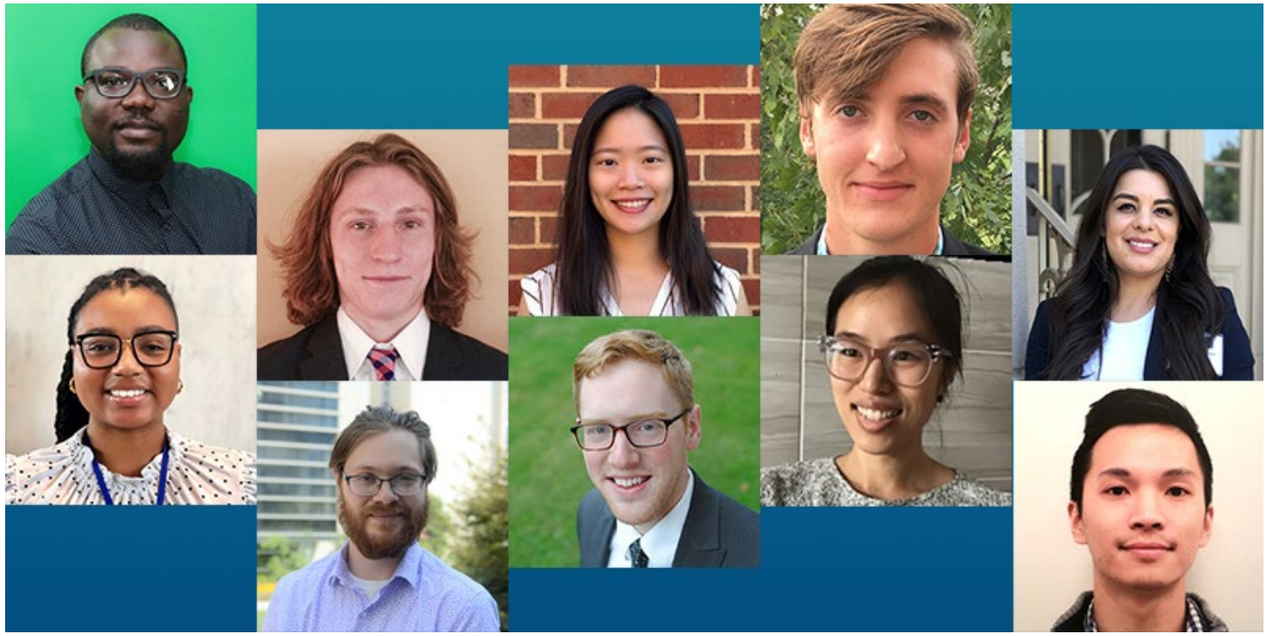
Student-led non-profit places tech-savvy students in federal agencies

Hosted 40 students in 2020; 16 for a 10-week summer fellowship and 24 for special fall fellowship

Hosting 16 students for a 10-week summer program in 2021

https://www.codingitforward.com/

Graduate Data Science Summer Program



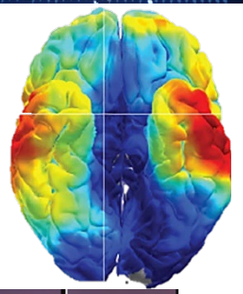
- Led by Office of Intramural Training and Education as part of Summer Internship Program
- Hosting **14** master's-level students for 10-week summer program
  - First cohort in 2019 had 10 students
- Pilot driven by discussion with local universities consortium
- All interns placed in intramural research labs

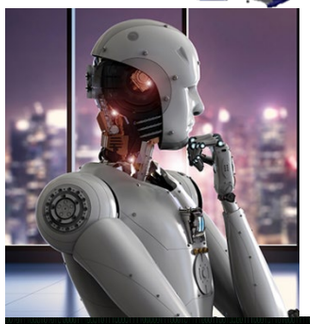
https://www.training.nih.gov/data_science_summer

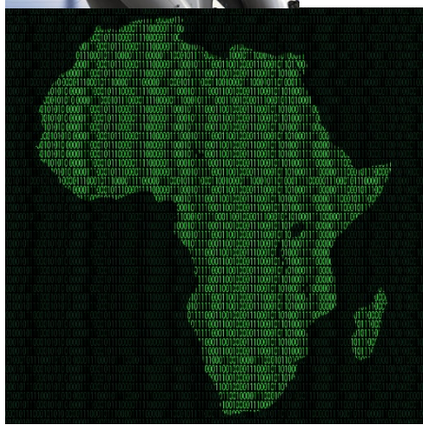
DATA Scholars at NIH

7 Scholars are currently onboard working on:

- Catalyzing neuroscience research (NIMH/NIDA)
- Unraveling the Alzheimer's Disease Genome (NIA)
- Supporting cancer knowledge extraction (NCI)
- Accelerating the clinical adoption of machine intelligence applications in medical imaging (NIBIB)
- Harnessing data science for health discovery and innovation in Africa (FIC)
- Expanding theories of brain circuits (NINDS)
- Integrating NIH cloud-based platforms for genomics research (NHGRI)







Data Scholars for 2021

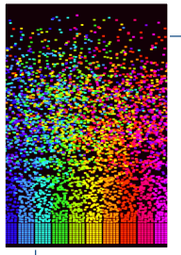
Architecting Search Across Petabyte-Scale Genomic Sequence (NLM)

Amplifying and Sustaining the Impact of Childhood Cancer, Structural Birth Defect, and Down Syndrome Data (NICHD)

Big Data Integration to Better Health for All of Us (All of Us)

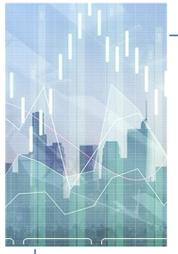
Bringing Hope to Untreatable Rare Diseases through Data Science (NCATS)

Harnessing Geospatial Data for Environmental Public Health Protection (NIEHS)



Implementing Machine Learning Algorithms to Improve Cancer Surveillance Data (NCI)

Innovative Solutions for Data Harmonization, Mobile Analytics, and End-User Support (OSC)



Mitigating Decision Biases of Marginalized Populations with Algorithm Assessments (NIMHD)

Shifting the Paradigm of Heart Failure Research with Advanced Analytics (NHLBI)

https://datascience.nih.gov/data-scholars-2021

Data Sharing and Reuse Seminar Series

- **Goal**: spotlight NIH-funded researchers who have taken existing data and found clever ways to reuse the data or generate new findings
- Second Friday of the month at 12 ET
- Speakers have included:
  - Russ Poldrack, Ph.D., Stanford University
  - Alisa Manning, Ph.D., Massachusetts General Hospital
  - Yang Chai, D.D.S., Ph.D., University of Southern California
  - Deanne Taylor, Ph.D., The Children's Hospital of Philadelphia Research Institute

http://bit.ly/NIHDataSeminars