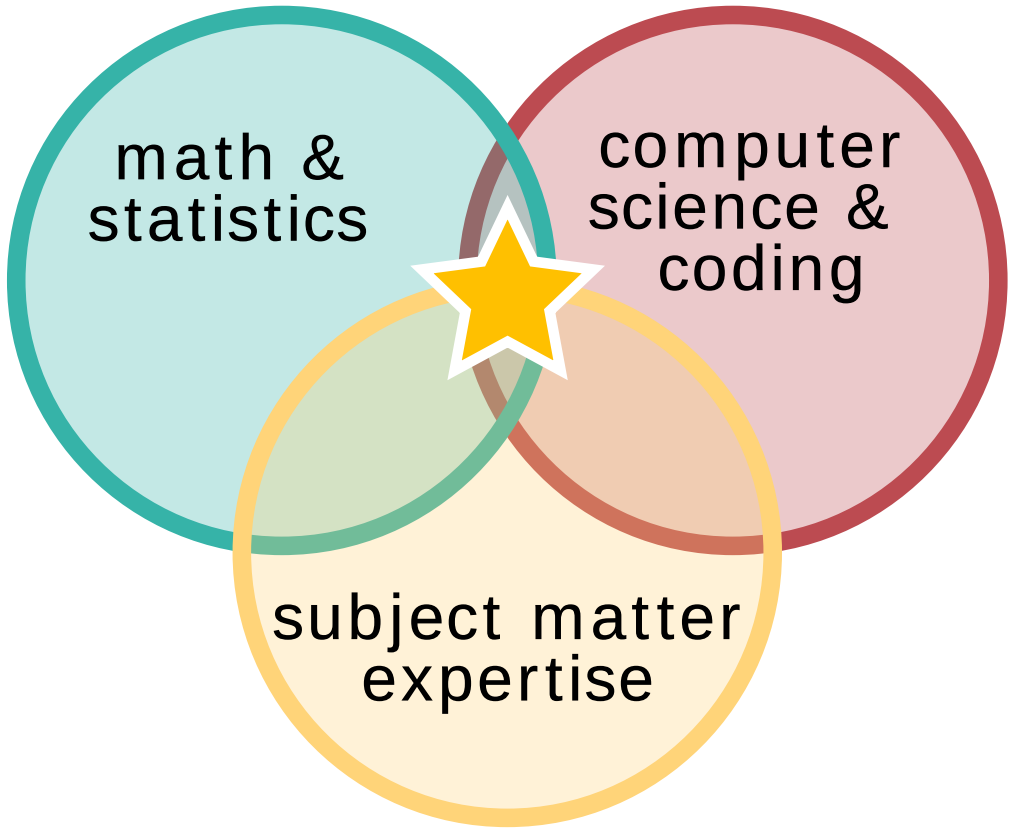
Strategic Approaches to Data Science & Open Science at NLM & NIH

Implementing FAIR Data for People & Machines Symposium
September 11, 2019

Michael F. Huerta, PhD
Director, Office of Strategic Initiatives
Associate Director, National Library of Medicine, NIH

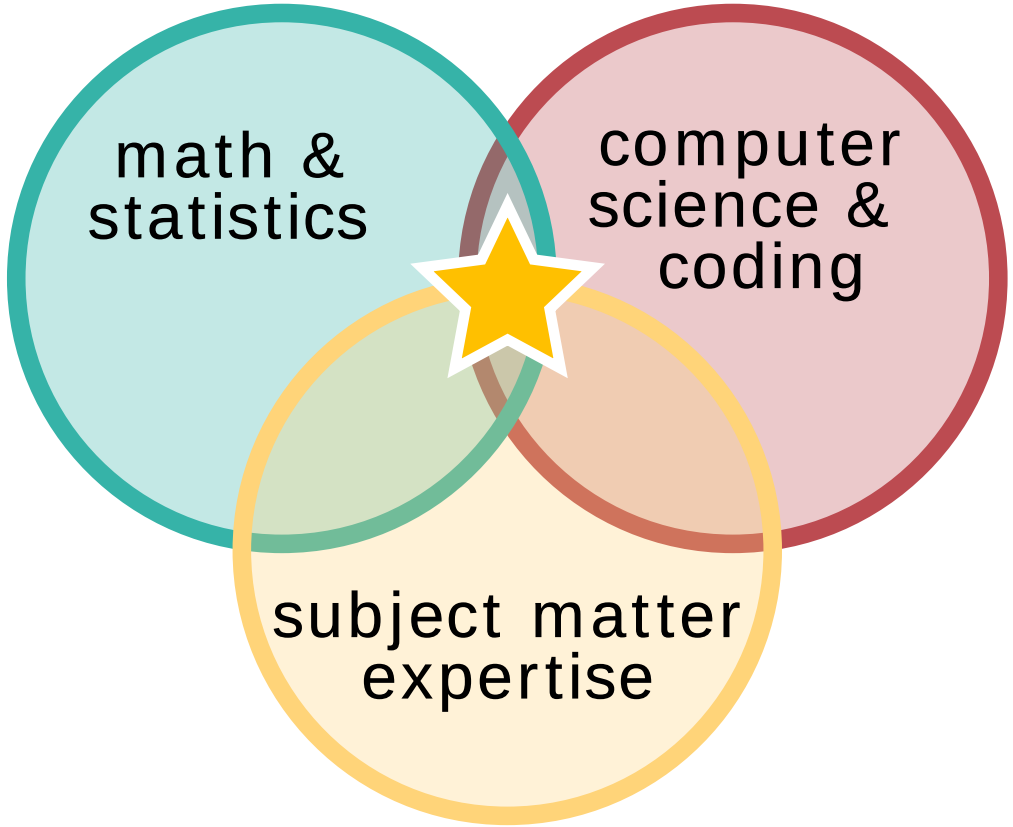
Data Science



Data Science is a **scientific & methodologic approach** to understanding data

(Just as molecular biology is a scientific & methodologic approach to understanding disease)

Data Science



Data Science is a **scientific & methodologic approach** to understanding data

(Just as molecular biology is a scientific & methodologic approach to understanding disease)

New tools for new insights

“New directions in science are launched by new tools much more often than by new concepts.”

Freeman Dyson
Imagined Worlds
(1997)
Harvard University Press

Open Science

Open Science is a **new paradigm**,
a different way of doing science
in which the **products and processes of research** (research objects) are broadly **available & usable**

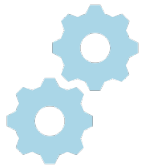
Open Science

Open Science is a **new paradigm**,
a different way of doing science
in which the **products and processes of research** (research objects) are broadly **available & usable**

Best practices for open science abide by **FAIR principles**

Findable

Accessible



Interoperable

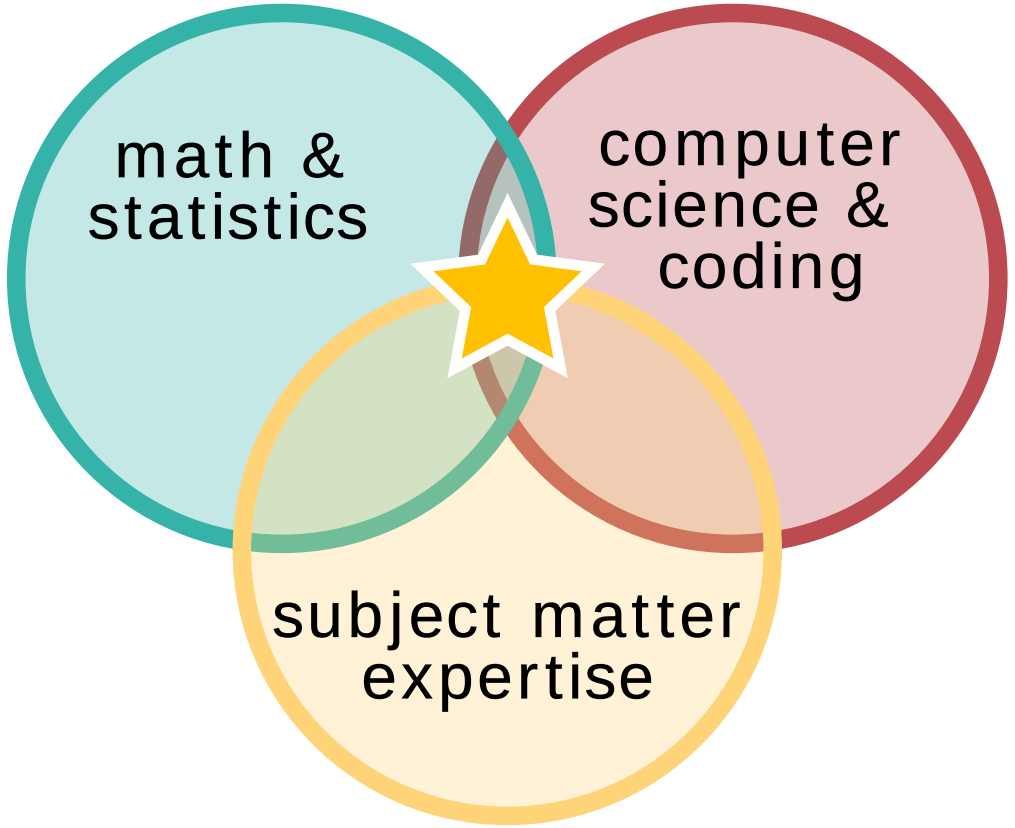
Reusable

"...paradigm changes do cause scientists to see the world of their research engagements differently."

Thomas Kuhn
The Structure of Scientific Revolutions
(1962)
University of Chicago Press

DS + OS = new tools and new paradigm

Data Science



Open Science

Findable

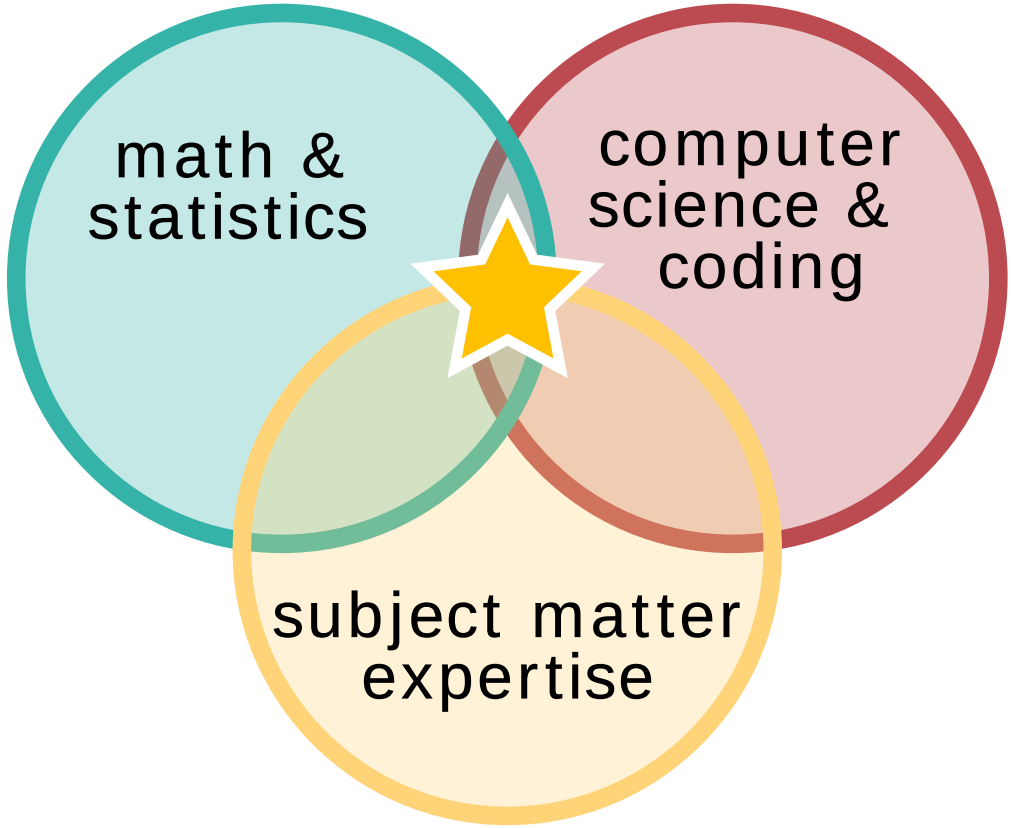
Accessible

Interoperable

Reusable

DS & OS are very powerful when paired

Data Science



Open Science

Findable

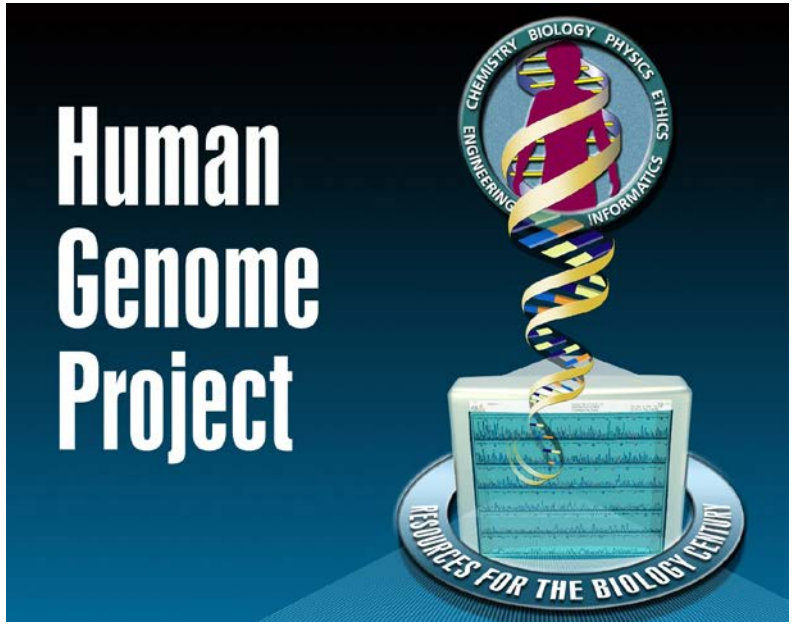
Accessible

Interoperable

Reusable

NIH has paired DS & OS to tackle significant biomedical research questions through several open, large scale, data-centric, digital initiatives

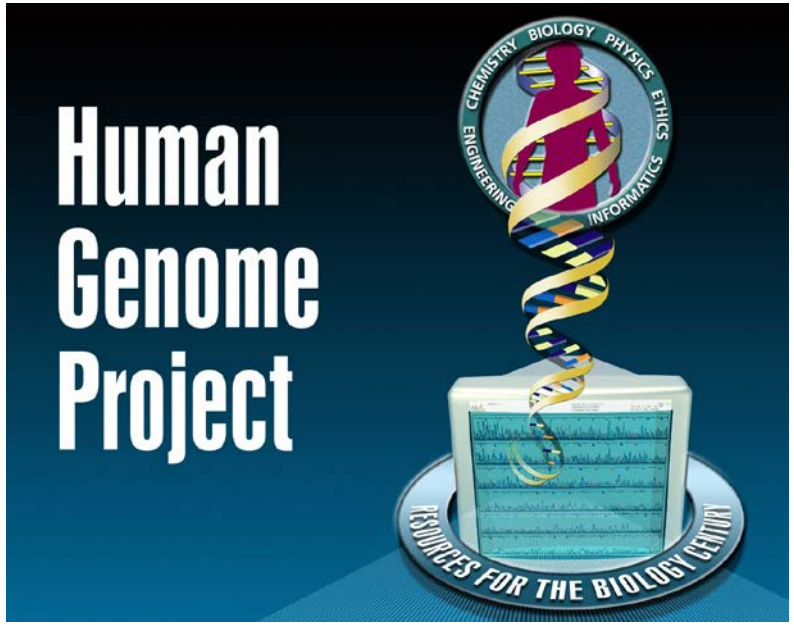
NIH has paired DS & OS to tackle significant biomedical research questions through several open, large scale, data-centric, digital initiatives



NIH The Common Fund

GTE

NIH HUMAN MICROBIOME PROJECT



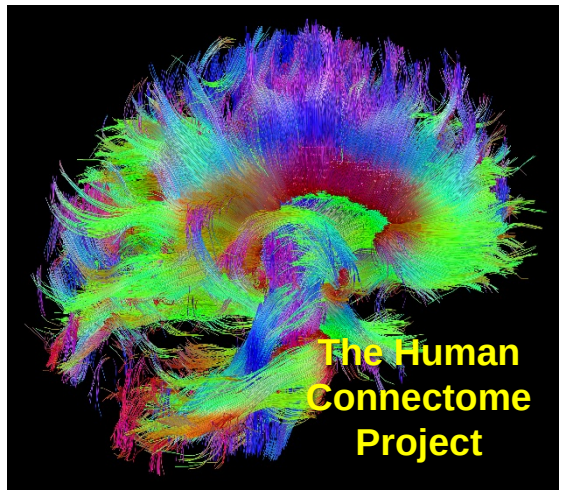
FlyBase

MGI

ZFIN The Zebrafish Model Organism Database

SGD Saccharomyces GENOME DATABASE

WormBase

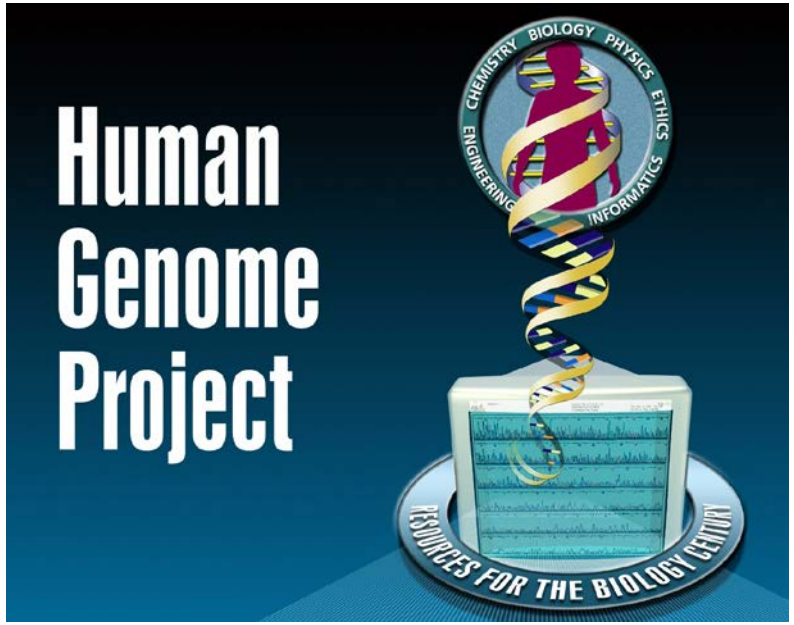


NIH
The Common Fund

GTE

ECHO
Environmental Influences on Child Health Outcomes

HMP
NIH HUMAN MICROBIOME PROJECT



ADNI
Alzheimer's Disease Neuroimaging Initiative

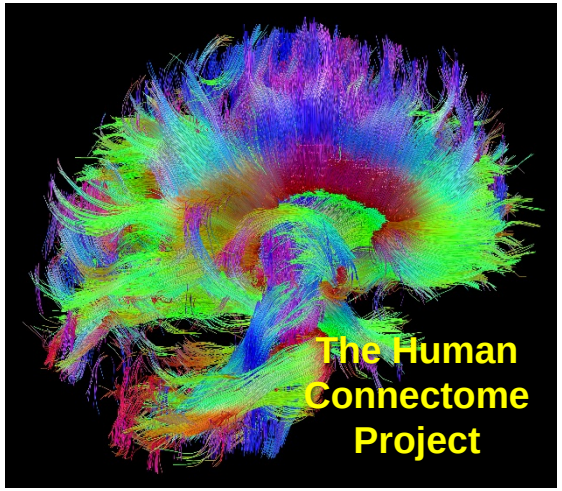
FlyBase

MGI

ZFIN
The Zebrafish Model Organism Database

SGD Saccharomyces GENOME DATABASE

WormBase



NIH
The Common Fund

GTE

ECHO
Environmental Influences on Child Health Outcomes

HMP
NIH HUMAN MICROBIOME PROJECT

CANCER MOONSHOT

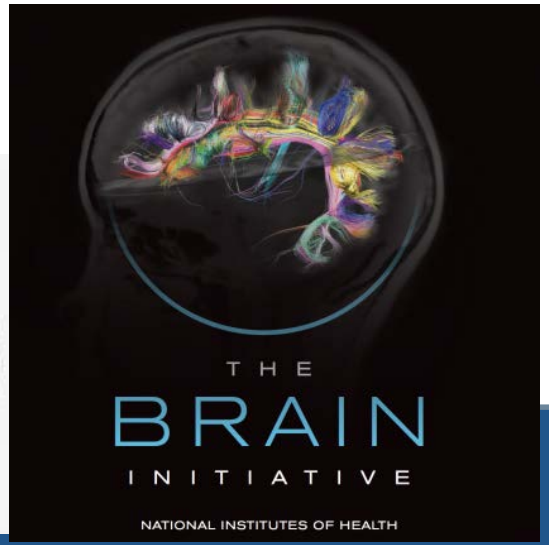
Human Genome Project
CHEMISTRY BIOLOGY PHYSICS ETHICS ENGINEERING INFORMATICS
RESOURCES FOR THE BIOLOGY CENTURY

ADNI
ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

FlyBase
MGI
ZFIN
The Zebrafish Model Organism Database
SGD Saccharomyces GENOME DATABASE
WormBase

Adolescent Brain Cognitive Development
Teen Brains. Today's Science. Brighter Future.

All of Us℠
THE FUTURE OF HEALTH BEGINS WITH YOU
The Precision Medicine Initiative®

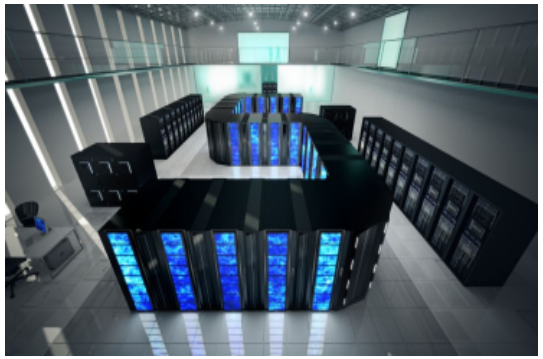


NIH Research to be Even MORE data-centric & open

Societal Expectations

NIH Research to be Even MORE data-centric & open

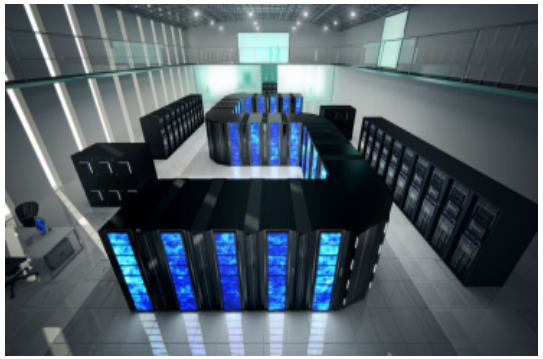
Societal Expectations



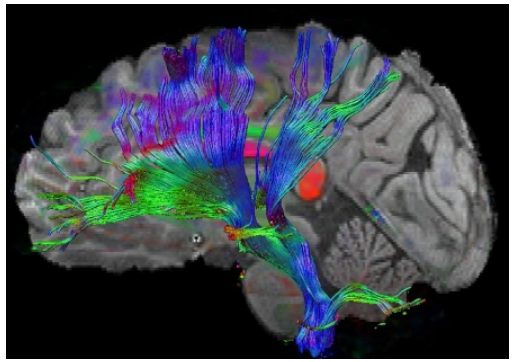
NIH Research to be Even MORE data-centric & open

Technical Capabilities

Societal Expectations



NIH Research to be Even MORE data-centric & open



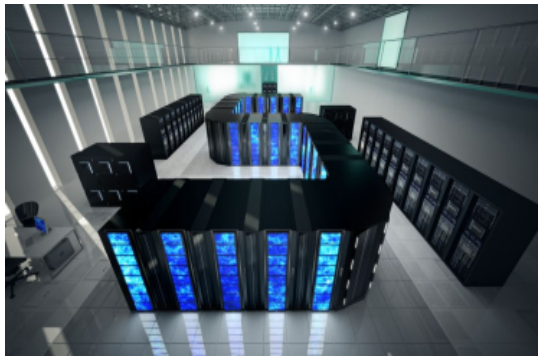
Technical Capabilities

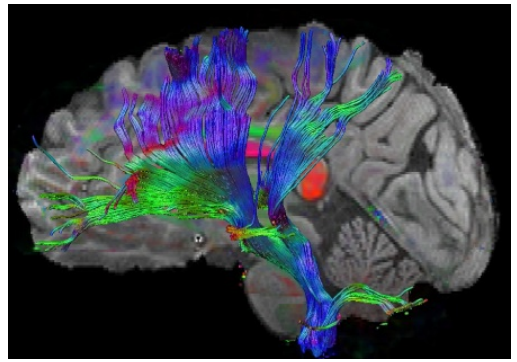
Scientific Opportunities

Societal Expectations

Policy & Legislative Directives

NIH Research to be Even MORE data-centric & open





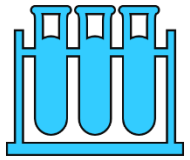
Technical Capabilities

Scientific Opportunities

The ***National Library of Medicine***
lives at the intersection of
data science *&* ***open science***

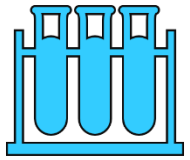
National Library of Medicine

National Library of Medicine



- An Institute of the NIH (1968)
  - Lead, conduct, and support research and training in biomedical:
    - Information science
    - Informatics
    - Data science

National Library of Medicine



- An Institute of the NIH (1968)
  - Lead, conduct, and support research and training in biomedical:
    - Information science
    - Informatics
    - Data science

- The world's largest biomedical library (1836)
  - Create & host major resources, tools, & services for literature, data, standards, & more
    - Send > 115 terabytes of data to > 5 million users daily
    - Receive > 15 terabytes of data from > 3,000 users daily

National Library of Medicine

- An Institute of the NIH (1968)
  - Lead, conduct, and support research and training in biomedical:
    - Information science
    - Informatics
    - Data science
- The world's largest biomedical library (1836)
  - Create & host major resources, tools, & services for literature, data, standards, & more
    - Send > 115 terabytes of data to > 5 million users daily
    - Receive > 15 terabytes of data from > 3,000 users daily
  - Facilitate open science & scholarship by making digital research objects:
    - Findable, Accessible, Interoperable, & Reusable (FAIR)
    - As well as Attributable & Sustainable

Now that the table is set, what's on the menu?

Now that the table is set, what's on the menu?

NLM Strategic Plan

Now that the table is set, what's on the menu?

NLM Strategic Plan

NIH Strategic Plan for Data Science

U.S. National Library of Medicine

A Platform for Biomedical Discovery and Data-Powered Health

Strategic Plan 2017-2027

Innovate, build, & sustain an open digital ecosystem for health info, science, & scholarship

Imp. Goal 1 – Connect & Enhance Resources

- Aligning curation across data & literature
- Universal Dataset Metadata Model Initiative
- Integration of select digital resources
- NLM Labs projects - indexing & linking info
- Blue Ribbon Panel Review of Intramural Research Programs
- Recruitment of Intramural Research Program Scientific Director, three investigators, & staff
- NSF-NLM data science MOU
- Data science research RFI
- Data Science Drivers Workshop & report
- At-scale computational curation FOA
- Research Reproducibility Workshop & report
- Ethnographic study of reproducibility
- Recruitment of 2 DSOS staff in OD
- NASEM study of value assessment for OS
- NASEM roundtable on incentives for OS
- Assessment of IT – 5 aspects & teams
- Assessment of NLM portfolio of offerings
- Assessment of NIH CDE Repository
- Assessment of data center & cloud Use
- Assessment of trans-NLM central functions
- Assessment & comparison of indexing methods
- Assessment of workspace

Imp. Goal 1 – Connect & Enhance Resources

- Aligning curation across data & literature
- Universal Dataset Metadata Model Initiative
- Integration of select digital resources
- NLM Labs projects - indexing & linking info
- Blue Ribbon Panel Review of Intramural Research Programs
- Recruitment of Intramural Research Program Scientific Director, three investigators, & staff
- NSF-NLM data science MOU
- Data science research RFI
- Data Science Drivers Workshop & report
- At-scale computational curation FOA
- Research Reproducibility Workshop & report
- Ethnographic study of reproducibility
- Recruitment of 2 DSOS staff in OD
- NASEM study of value assessment for OS
- NASEM roundtable on incentives for OS
- Assessment of IT – 5 aspects & teams
- Assessment of NLM portfolio of offerings
- Assessment of NIH CDE Repository
- Assessment of data center & cloud Use
- Assessment of trans-NLM central functions
- Assessment & comparison of indexing methods
- Assessment of workspace

Imp. Goal 1 – Innovate, Expand, & Enhance Research

- Aligning curation across data & literature
- Universal Dataset Metadata Model Initiative
- Integration of select digital resources
- NLM Labs projects - indexing & linking info
- Blue Ribbon Panel Review of Intramural Research Programs
- Recruitment of Intramural Research Program Scientific Director, three investigators, & staff
- NSF-NLM data science MOU
- Data science research RFI
- Data Science Drivers Workshop & report
- At-scale computational curation FOA
- Research Reproducibility Workshop & report
- Ethnographic study of reproducibility
- Recruitment of 2 DSOS staff in OD
- NASEM study of value assessment for OS
- NASEM roundtable on incentives for OS
- Assessment of IT – 5 aspects & teams
- Assessment of NLM portfolio of offerings
- Assessment of NIH CDE Repository
- Assessment of data center & cloud Use
- Assessment of trans-NLM central functions
- Assessment & comparison of indexing methods
- Assessment of workspace

Imp. Goal 1 – Innovate, Expand, & Enhance Research

- Aligning curation across data & literature
- Universal Dataset Metadata Model Initiative
- Integration of select digital resources
- NLM Labs projects - indexing & linking info
- Blue Ribbon Panel Review of Intramural Research Programs
- Recruitment of Intramural Research Program Scientific Director, three investigators, & staff
- NSF-NLM data science MOU
- Data science research RFI
- Data Science Drivers Workshop & report
- At-scale computational curation FOA
- Research Reproducibility Workshop & report
- Ethnographic study of reproducibility
- Recruitment of 2 DSOS staff in OD
- NASEM study of value assessment for OS
- NASEM roundtable on incentives for OS
- Assessment of IT – 5 aspects & teams
- Assessment of NLM portfolio of offerings
- Assessment of NIH CDE Repository
- Assessment of data center & cloud Use
- Assessment of trans-NLM central functions
- Assessment & comparison of indexing methods
- Assessment of workspace

Imp. Goal 1 – Advance Open Science

- Aligning curation across data & literature
- Universal Dataset Metadata Model Initiative
- Integration of select digital resources
- NLM Labs projects - indexing & linking info
- Blue Ribbon Panel Review of Intramural Research Programs
- Recruitment of Intramural Research Program Scientific Director, three investigators, & staff
- NSF-NLM data science MOU
- Data science research RFI
- Data Science Drivers Workshop & report
- At-scale computational curation FOA
- Research Reproducibility Workshop & report
- Ethnographic study of reproducibility
- Recruitment of 2 DSOS staff in OD
- NASEM study of value assessment for OS
- NASEM roundtable on incentives for OS
- Assessment of IT – 5 aspects & teams
- Assessment of NLM portfolio of offerings
- Assessment of NIH CDE Repository
- Assessment of data center & cloud Use
- Assessment of trans-NLM central functions
- Assessment & comparison of indexing methods
- Assessment of workspace

Imp. Goal 1 – Advance Open Science

- Aligning curation across data & literature
- Universal Dataset Metadata Model Initiative
- Integration of select digital resources
- NLM Labs projects - indexing & linking info
- Blue Ribbon Panel Review of Intramural Research Programs
- Recruitment of Intramural Research Program Scientific Director, three investigators, & staff
- NSF-NLM data science MOU
- Data science research RFI
- Data Science Drivers Workshop & report
- At-scale computational curation FOA
- Research Reproducibility Workshop & report

- Ethnographic study of reproducibility
- Recruitment of 2 DSOS staff in OD
- NASEM study of value assessment for OS
- NASEM roundtable on incentives for OS
- Assessment of IT – 5 aspects & teams
- Assessment of NLM portfolio of offerings
- Assessment of NIH CDE Repository
- Assessment of data center & cloud Use
- Assessment of trans-NLM central functions
- Assessment & comparison of indexing methods
- Assessment of workspace

Imp. Goal 1 – Assure Sustainable Infrastructures

- Aligning curation across data & literature
- Universal Dataset Metadata Model Initiative
- Integration of select digital resources
- NLM Labs projects - indexing & linking info
- Blue Ribbon Panel Review of Intramural Research Programs
- Recruitment of Intramural Research Program Scientific Director, three investigators, & staff
- NSF-NLM data science MOU
- Data science research RFI
- Data Science Drivers Workshop & report
- At-scale computational curation FOA
- Research Reproducibility Workshop & report
- Ethnographic study of reproducibility
- Recruitment of 2 DSOS staff in OD
- NASEM study of value assessment for OS
- NASEM roundtable on incentives for OS
- Assessment of IT – 5 aspects & teams
- Assessment of NLM portfolio of offerings
- Assessment of NIH CDE Repository
- Assessment of data center & cloud Use
- Assessment of trans-NLM central functions
- Assessment & comparison of indexing methods
- Assessment of staff workspace

Imp. Goal 1 – Assure Sustainable Infrastructures

- Aligning curation across data & literature
- Universal Dataset Metadata Model Initiative
- Integration of select digital resources
- NLM Labs projects - indexing & linking info
- Blue Ribbon Panel Review of Intramural Research Programs
- Recruitment of Intramural Research Program Scientific Director, three investigators, & staff
- NSF-NLM data science MOU
- Data science research RFI
- Data Science Drivers Workshop & report
- At-scale computational curation FOA
- Research Reproducibility Workshop & report
- Ethnographic study of reproducibility
- Recruitment of 2 DSOS staff in OD
- NASEM study of value assessment for OS
- NASEM roundtable on incentives for OS
- Assessment of IT – 5 aspects & teams
- Assessment of NLM portfolio of offerings
- Assessment of NIH CDE Repository
- Assessment of data center & cloud Use
- Assessment of trans-NLM central functions
- Assessment & comparison of indexing methods
- Assessment of staff workspace

Innovate, build, & sustain an open digital ecosystem for health info, science, & scholarship

Optimize user experience with, and use of, NLM digital resources

Imp. Goal 2 – Optimize User Experience

- Assessment of tools to evaluate resources
- Assessment of comparative web metrics
- Audit & assessment of outreach efforts
- User experience/development initiatives

- Reorganized and enhanced outreach, engagement, & training on resource use
- Personal health libraries FOA
- Health information resources to reduce health disparities FOA
- National Network of Libraries of Medicine outreach around data science & open science

Imp. Goal 2 – Optimize User Experience

- Assessment of tools to evaluate resources
- Assessment of comparative web metrics
- Audit & assessment of outreach efforts
- User experience/development initiatives

- Reorganized and enhanced outreach, engagement, & training on resource use
- Personal health libraries FOA
- Health information resources to reduce health disparities FOA
- National Network of Libraries of Medicine outreach around data science & open science

Imp. Goal 2 – Optimize Use of NLM Resources

- Assessment of tools to evaluate resources
- Assessment of comparative web metrics
- Audit & assessment of outreach efforts
- User experience/development initiatives

- Reorganized and enhanced outreach, engagement, & training on resource use
- Personal health libraries FOA
- Health information resources to reduce health disparities FOA
- National Network of Libraries of Medicine outreach around data science & open science

Imp. Goal 2 – Optimize Use of NLM Resources

- Assessment of tools to evaluate resources
- Assessment of comparative web metrics
- Audit & assessment of outreach efforts
- User experience/development initiatives

- Reorganized and enhanced outreach, engagement, & training on resource use
- Personal health libraries FOA
- Health information resources to reduce health disparities FOA
- National Network of Libraries of Medicine outreach around data science & open science

Innovate, build, & sustain an open digital ecosystem for health info, science, & scholarship

Optimize user experience with, and use of, NLM digital resources

Assure a data-savvy biomedical workforce and a data-ready public

Imp. Goal 3 – Data Science Research Training

- Expand extramural predoctoral & postdoctoral data science research training
- Research reproducibility training of intramural scientists
- Data science core skills analysis & report
- Comprehensive analysis & report on state of data science training in biomedicine
- Data science skills for librarian workforce workshop & report
- Data science assessment and training of all NLM staff
- Data science & open science needs assessment of NIH extramural staff
- Establish trans-NIH coordination on NIH staff training in DSOS
- Increase partnerships with I-Schools
- Increase partnerships with minority serving institutions
- Data Science Drivers Workshop & report
- Expand summer research training of high school and undergrad students

Imp. Goal 3 – Data Science Research Training

- Expand extramural predoctoral & postdoctoral data science research training
- Research reproducibility training of intramural scientists
- Data science core skills analysis & report
- Comprehensive analysis & report on state of data science training in biomedicine
- Data science skills for librarian workforce workshop & report
- Data science assessment and training of all NLM staff
- Data science & open science needs assessment of NIH extramural staff
- Establish trans-NIH coordination on NIH staff training in DSOS
- Increase partnerships with I-Schools
- Increase partnerships with minority serving institutions
- Data Science Drivers Workshop & report
- Expand summer research training of high school and undergrad students

Imp. Goal 3 – Data-Savvy Workforce

- Expand extramural predoctoral & postdoctoral data science research training
- Research reproducibility training of intramural scientists
- Data science core skills analysis & report
- Comprehensive analysis & report on state of data science training in biomedicine
- Data science skills for librarian workforce workshop & report
- Data science assessment and training of all NLM staff
- Data science & open science needs assessment of NIH extramural staff
- Establish trans-NIH coordination on NIH staff training in DSOS
- Increase partnerships with I-Schools
- Increase partnerships with minority serving institutions
- Data Science Drivers Workshop & report
- Expand summer research training of high school and undergrad students

Imp. Goal 3 – Data-Savvy Workforce

- Expand extramural predoctoral & postdoctoral data science research training
- Research reproducibility training of intramural scientists
- Data science core skills analysis & report
- Comprehensive analysis & report on state of data science training in biomedicine
- Data science skills for librarian workforce workshop & report
- Data science assessment and training of all NLM staff
- Data science & open science needs assessment of NIH extramural staff
- Establish trans-NIH coordination on NIH staff training in DSOS
- Increase partnerships with I-Schools
- Increase partnerships with minority serving institutions
- Data Science Drivers Workshop & report
- Expand summer research training of high school and undergrad students

Imp. Goal 3 – Data-Ready Public

- Expand extramural predoctoral & postdoctoral data science research training
- Research reproducibility training of intramural scientists
- Data science core skills analysis & report
- Comprehensive analysis & report on state of data science training in biomedicine
- Data science skills for librarian workforce workshop & report
- Data science assessment and training of all NLM staff
- Data science & open science needs assessment of NIH extramural staff
- Establish trans-NIH coordination on NIH staff training in DSOS
- Increase partnerships with I-Schools
- Increase partnerships with minority serving institutions
- Data Science Drivers Workshop & report
- Expand summer research training of high school and undergrad students

Imp. Goal 3 – Data-Ready Public

- Expand extramural predoctoral & postdoctoral data science research training
- Research reproducibility training of intramural scientists
- Data science core skills analysis & report
- Comprehensive analysis & report on state of data science training in biomedicine
- Data science skills for librarian workforce workshop & report
- Data science assessment and training of all NLM staff
- Data science & open science needs assessment of NIH extramural staff
- Establish trans-NIH coordination on NIH staff training in DSOS
- Increase partnerships with I-Schools
- Increase partnerships with minority serving institutions
- Data Science Drivers Workshop & report
- Expand summer research training of high school and undergrad students

NIH Strategic Plan for Data Science – ODSS & SDC

NIH Strategic Plan for Data Science – ODSS & SDC

Support Highly Efficient and Effective Data **Infrastructure** for Biomedical Research

Promote the Modernization of the Research Data Resources **Ecosystem**

Support the Development and Dissemination of Advanced Management, Analytics, and Visualization **Tools**

Enhance **Workforce** Development for Biomedical Data Science

Enact Appropriate Policies to Promote Stewardship and **Sustainability**

Key Objective: FAIR data sharing for NIH research

Key Objective: FAIR data sharing for NIH research

- Provide FAIR-enabled, open access for all datasets that underlie publications resulting from NIH-funded research

Key Objective: FAIR data sharing for NIH research

- Provide FAIR-enabled, open access for all datasets that underlie publications resulting from NIH-funded research
- Identify key characteristics of repositories housing those data
  - Initial key characteristics identified by BMIC Subgroup
  - Government-wide RFI for public comment on characteristics currently being drafted

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific repositories as a first choice

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

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific repositories as a first choice

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

Options of scaled implementation for sharing datasets

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific 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 GB**

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.

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific 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 GB**

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* GB**

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.

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific 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 GB**

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* GB**

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 **PBs**

STRIDES Cloud Partners

- Store and manage large scale, high priority NIH datasets (Partnership with STRIDES)
- Assign Unique Identifiers, implement authentication, authorization & access control

STRIDES

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

STRIDES

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

- Agreements with
  - Google Cloud
  - Amazon Web Services
  - Additional partnerships anticipated
- Other Transaction Authority used

STRIDES

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

- Agreements with
  - Google Cloud
  - Amazon Web Services
  - Additional partnerships anticipated
- Other Transaction Authority used

- Benefits to NIH-supported Investigators
  - Discounted **rates** on cloud services
  - Access to engineering, consulting, and other professional **services** from cloud service provider partners
  - Access to cloud **training** programs (standard & custom)

Examples of datasets moving to STRIDES-clouds

- NHLBI Framingham Heart Study
- All of Us Research Program
- NCI Genomic Data Commons
- NLM Sequence Read Archive
- 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 Data

Examples of datasets moving to STRIDES-clouds

- NHLBI Framingham Heart Study
- All of Us Research Program
- NCI Genomic Data Commons
- NLM Sequence Read Archive
- 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 Data

Implementation of NIH SP for DS (so far)

Data Infrastructure
Enabling FAIR

- Establishing **partnerships** with commercial clouds
- Moving High value **datasets to clouds**
- **Training & outreach** on clouds across ICs

Modernize the Data Ecosystem
Enabling data sharing across ICs and grantees

- Conducting **inventory & gap analysis** of current large scale data resources
- **FHIR initiatives** for sharing phenotypic data & tool development
- Implementing NIH-wide system for user **authentication & authorization**

Data Management, Tools and Analytics

- Linking **datasets to publications** in PubMed Central
- Identifying key **characteristics of data repositories**
- NLM **FOA on Curation at Scale**
- NIH **FOAs to support Trans-NIH Databases & Knowledgebases** coming soon!

Implementation of NIH SP for DS (so far)

- Coding-it-forward; 10 **undergraduate fellows** placed in ICs for summer 2019
- 13 **Masters fellows** placed in intramural laboratories for summer 2019
- Quantitative and computational skills language for **training FOAs** developed & being disseminated

- **RFI** on key provisions for data management and sharing policy
- **Draft policy** target date is summer 2019
- **Final policy** by ~ December 2019

Needed to Realize the Promise of DS/OS

Needed to Realize the Promise of DS/OS

- **Incentives** – Establish & align incentives to promote open science practices (e.g., sharing data, adopting standards, using appropriate repositories)
  - Strategically align incentives across entire ecosystem to maximize impact
  - Likely best done domain-by-domain

Needed to Realize the Promise of DS/OS

- **Incentives** – Establish & align incentives to promote open science practices (e.g., sharing data, adopting standards, using appropriate repositories)
  - Strategically align incentives across entire ecosystem to maximize impact
  - Likely best done domain-by-domain

- **At-Scale Curation & Provenance** – Rapid increase in number of DROs and the need to find, associate, and monitor their versions is outstripping the ability to apply consistent, useful metadata to them. Move from applying metadata to having DROs imply their metadata
  - Move from search to learning, and from learning to awareness
  - Draw from artificial intelligence, blockchain, etc.

Needed to Realize the Promise of DS/OS

- **Incentives** – Establish & align incentives to promote open science practices (e.g., sharing data, adopting standards, using appropriate repositories)
  - Strategically align incentives across entire ecosystem to maximize impact
  - Likely best done domain-by-domain

- **At-Scale Curation & Provenance** – Rapid increase in number of DROs and the need to find, associate, and monitor their versions is outstripping the ability to apply consistent, useful metadata to them. Move from applying metadata to having DROs imply their metadata
  - Move from search to learning, and from learning to awareness
  - Draw from artificial intelligence, blockchain, etc.

- **Sustainability** – Assure ROI by assessing the value of particular investments in the ecosystem (e.g., in infrastructure, data acquisition, preservation, policy changes, etc.)
  - Rigorous cost vs benefit analyses
  - Metrics & models

NLM Office of Strategic Initiatives
Data Science & Open Science Team

Lisa Federer, PhD, MLIS
Data Science & Open Science Librarian

Teresa Zayas-Caban, PhD
Coordinator, NIH FHIR Acceleration
Chief Scientist, ONC, DHHS

Rebecca Goodwin, JD
Policy Analyst & Open Science Specialist

Maryam Zaringhalam, PhD
Data Science & Open Science Specialist

Tony Chu, PhD, MLIS
Information Scientist

NLM Office of Strategic Initiatives

Data Science & Open Science Team

Lisa Federer, PhD, MLIS
Data Science & Open Science Librarian

Teresa Zayas-Caban, PhD
Coordinator, NIH FHIR Acceleration
Chief Scientist, ONC, DHHS

Rebecca Goodwin, JD
Policy Analyst & Open Science Specialist

Maryam Zaringhalam, PhD
Data Science & Open Science Specialist

Tony Chu, PhD, MLIS
Information Scientist

THANKS!