Opening Science & Scholarship

Michael F. Huerta, Ph.D.
Coordinator of Data Science & Open Science Initiatives
Associate Director for Program Development
National Library of Medicine, NIH

National Academies – September 18, 2017

The National Library of Medicine

The National Library of Medicine

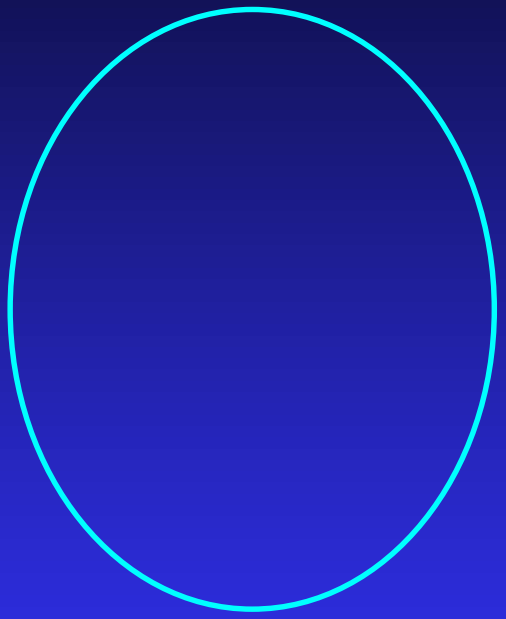
- NIH Institute – Lead, conduct & support research in:
  - Information science
  - Informatics
  - Data science

The National Library of Medicine

- NIH Institute – Lead, conduct & support research in:
  - Information science
  - Informatics
  - Data science
- Biomedical Library – The world's largest
  - Embrace openness – science and scholarship
  - Steward of literature and more
  - Index > 5600 journals in MEDLINE
  - Major data & info resources
    - Sends > 100 terabytes of data to > 4 million users
    - Receives > 10 terabytes of data from > 3,000 users

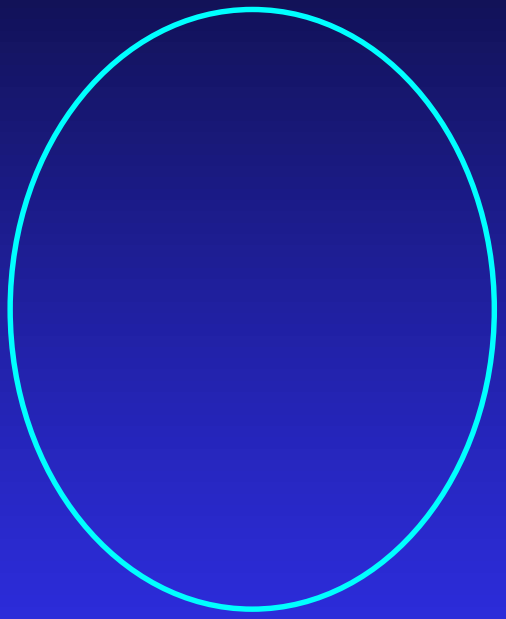
What NLM Does

Biomedical Science



Literature

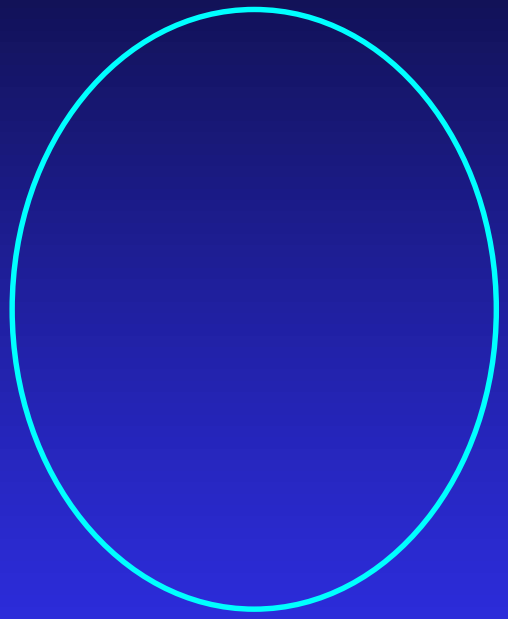
Biomedical Science

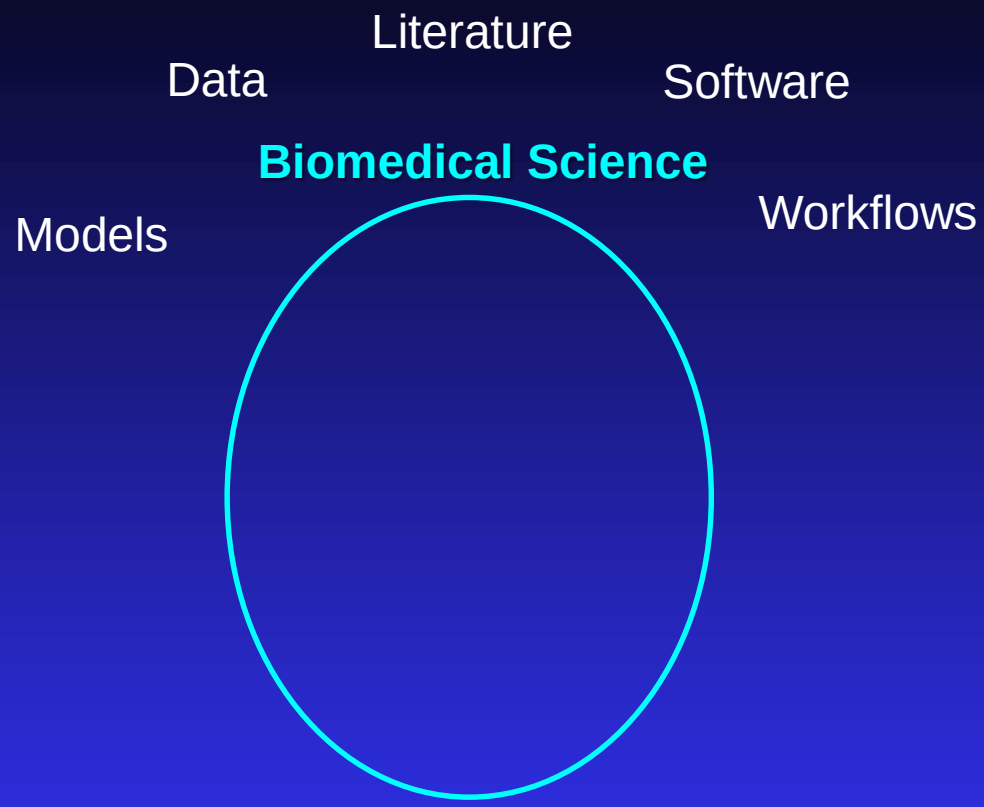


Literature

Data

Biomedical Science




Literature
Data
Software
Biomedical Science
Workflows
Models

Collections of Digital Research Objects

Literature

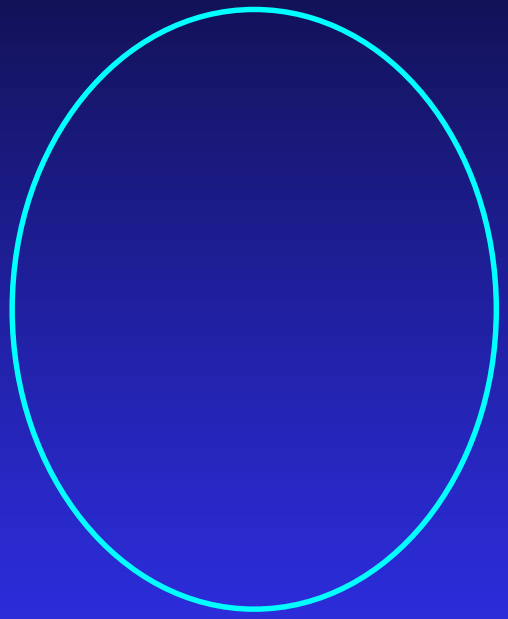
Data

Software

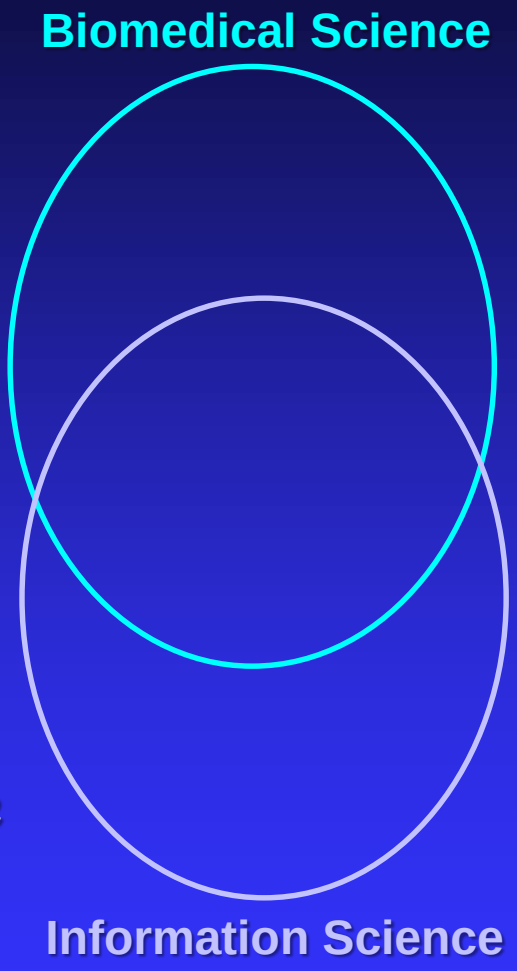
Biomedical Science

Workflows

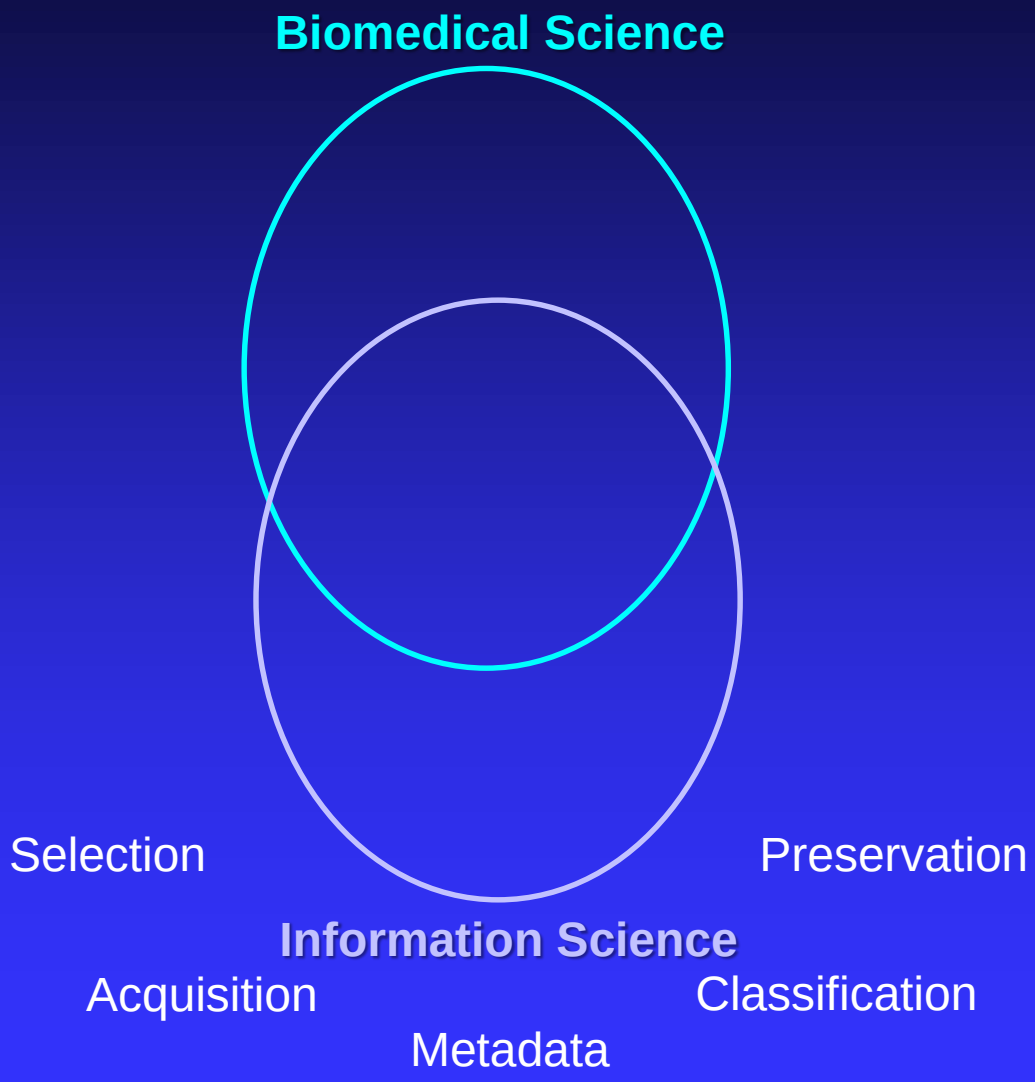
Models



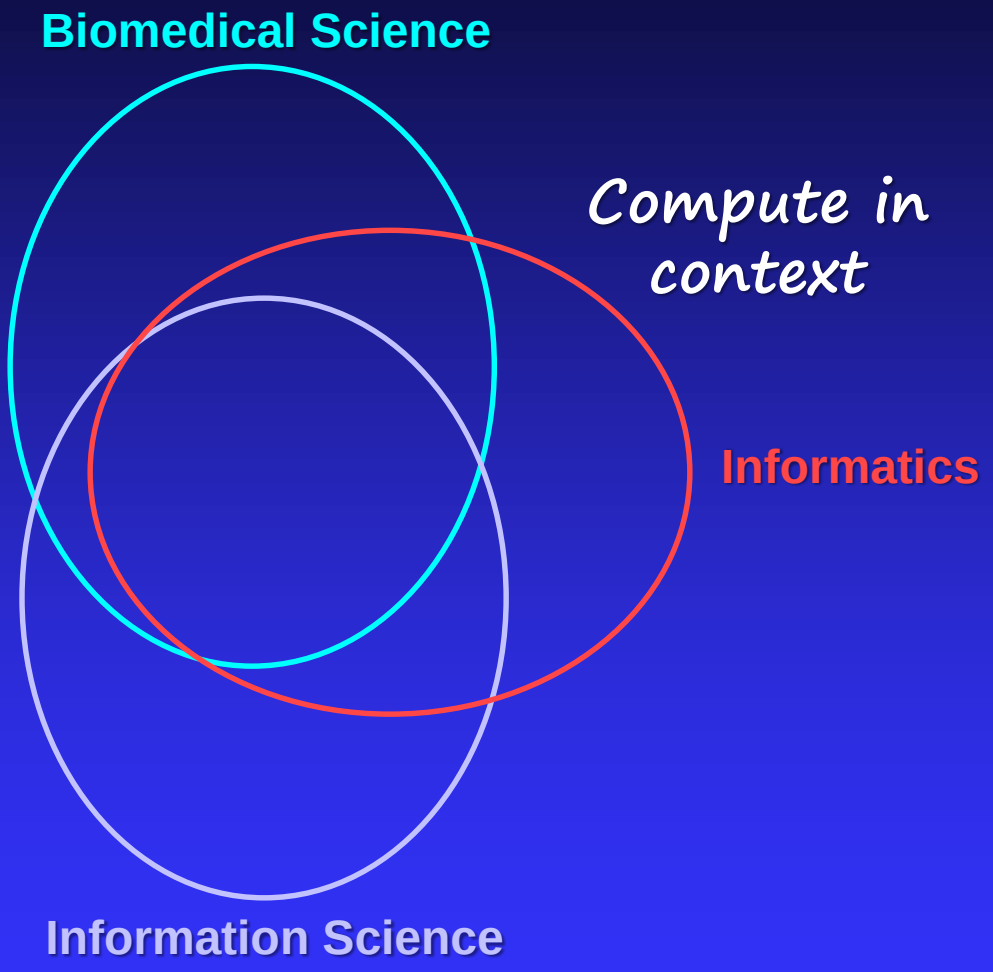
Collections of Digital Research Objects



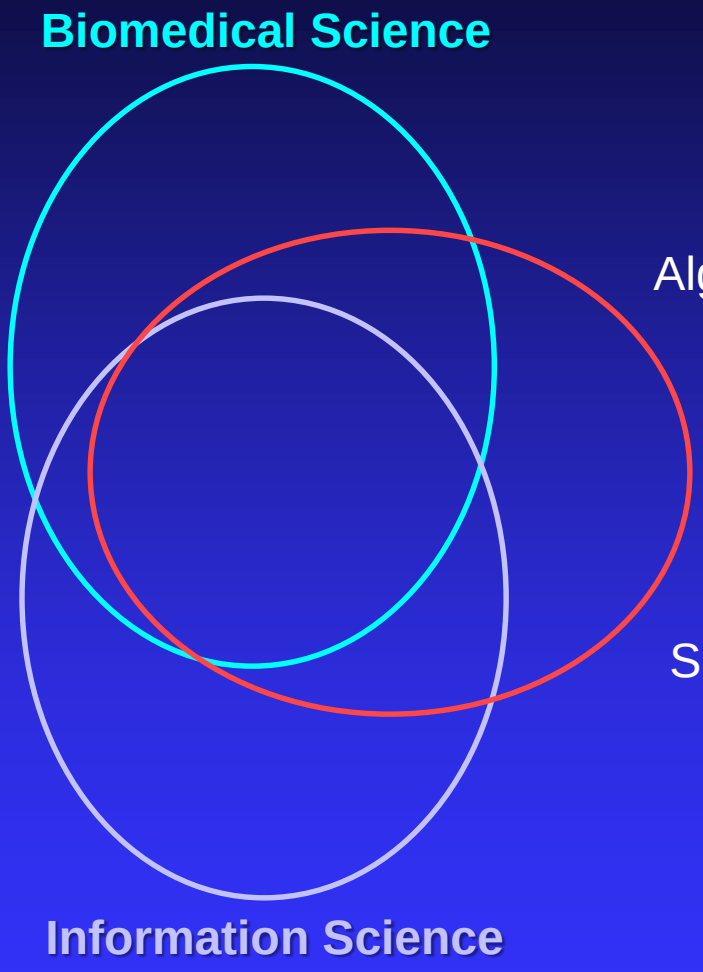
Collections of Digital Research Objects



Collections of Digital Research Objects



Collections of Digital Research Objects



Algorithms

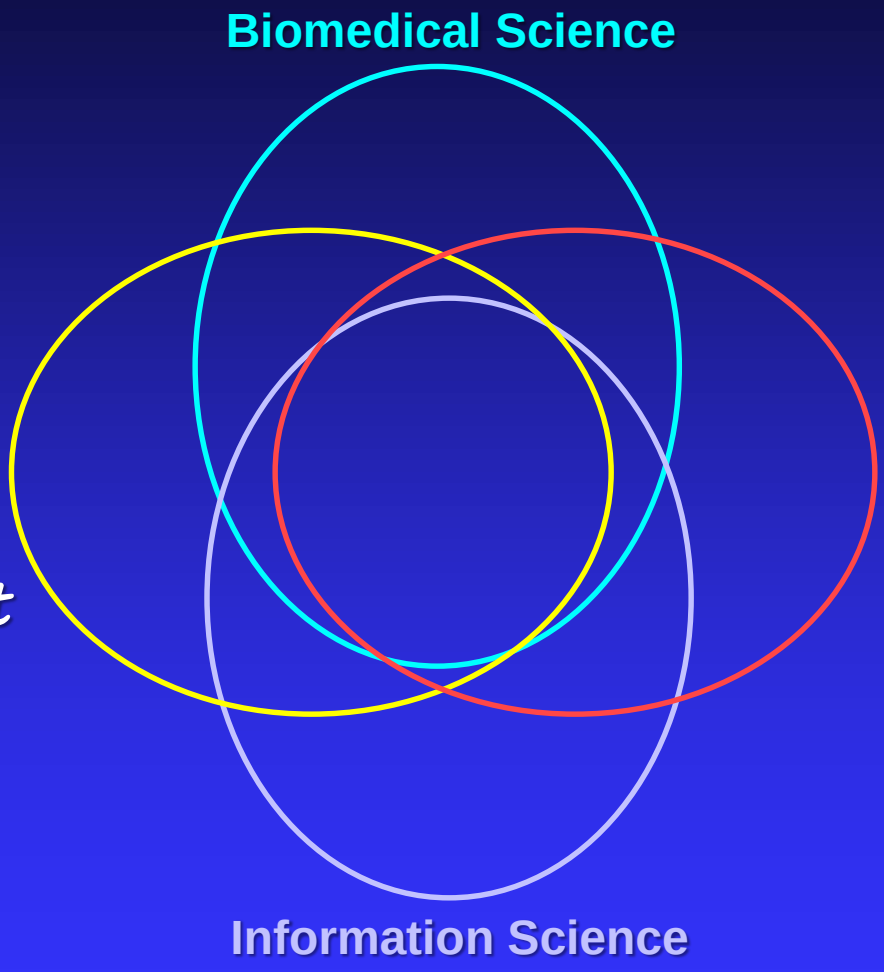
Ontologies

Informatics

Computation

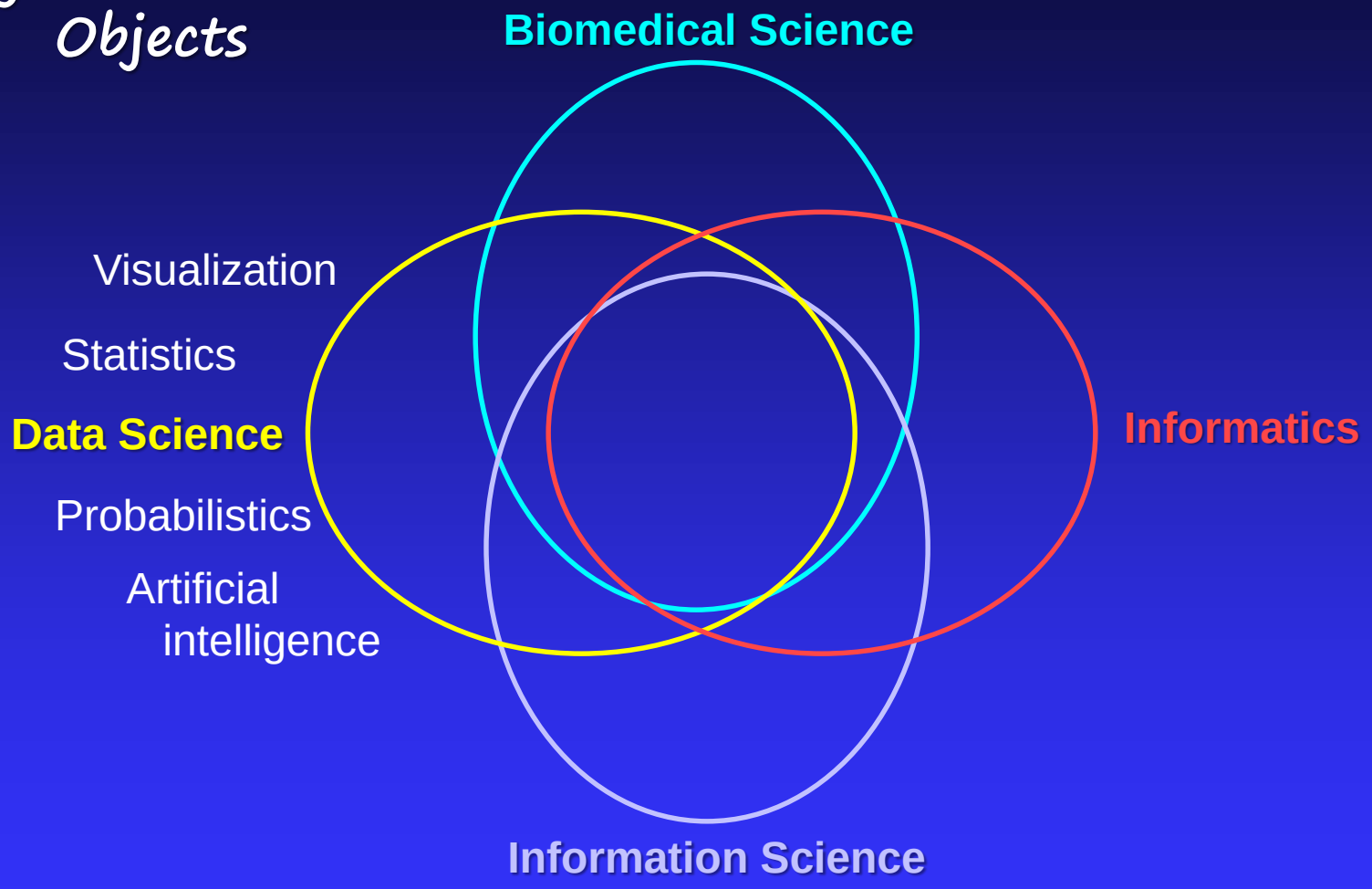
Software tools

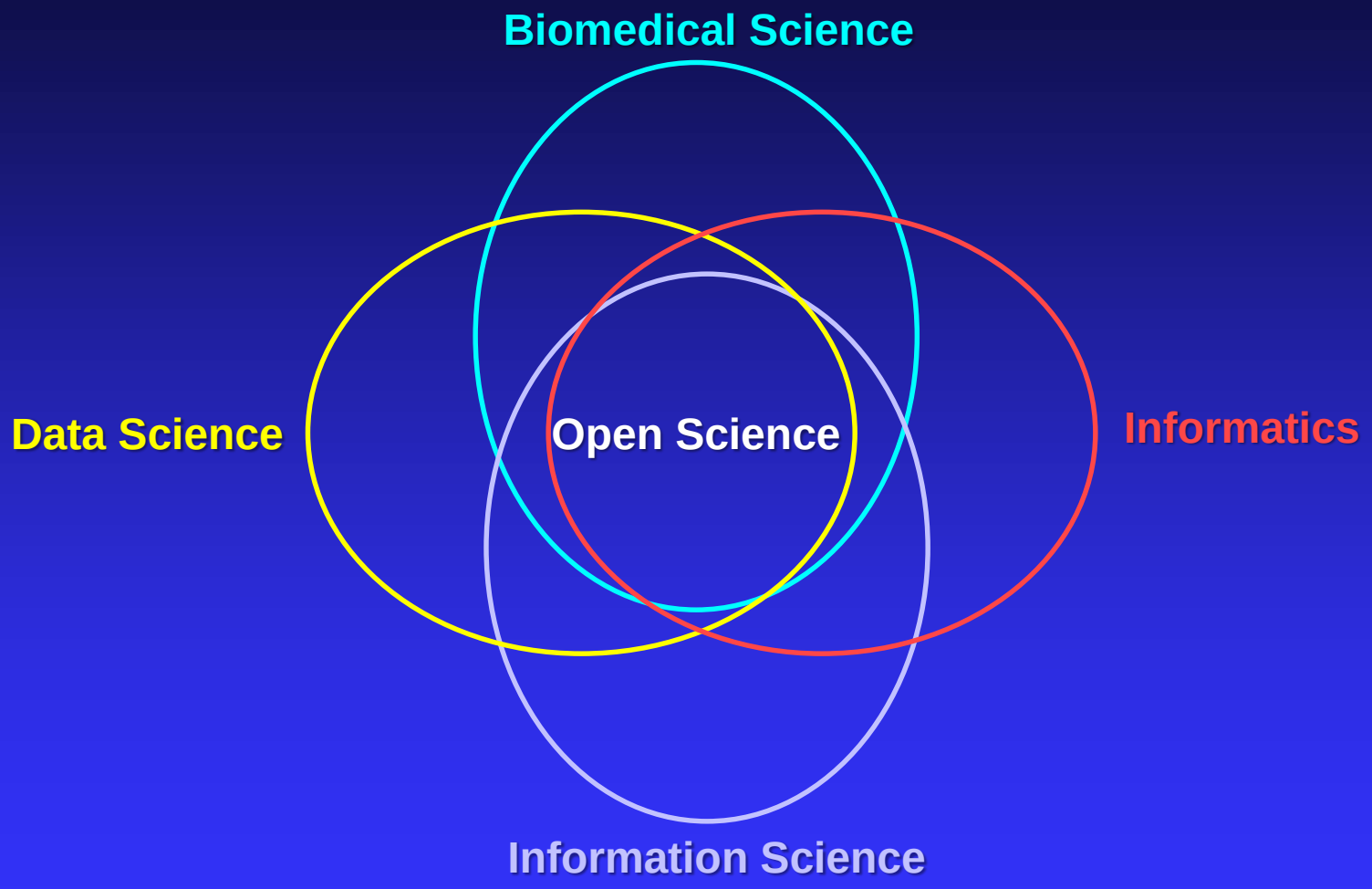
Collections of Digital Research Objects



Extract insight from data

Collections of Digital Research Objects

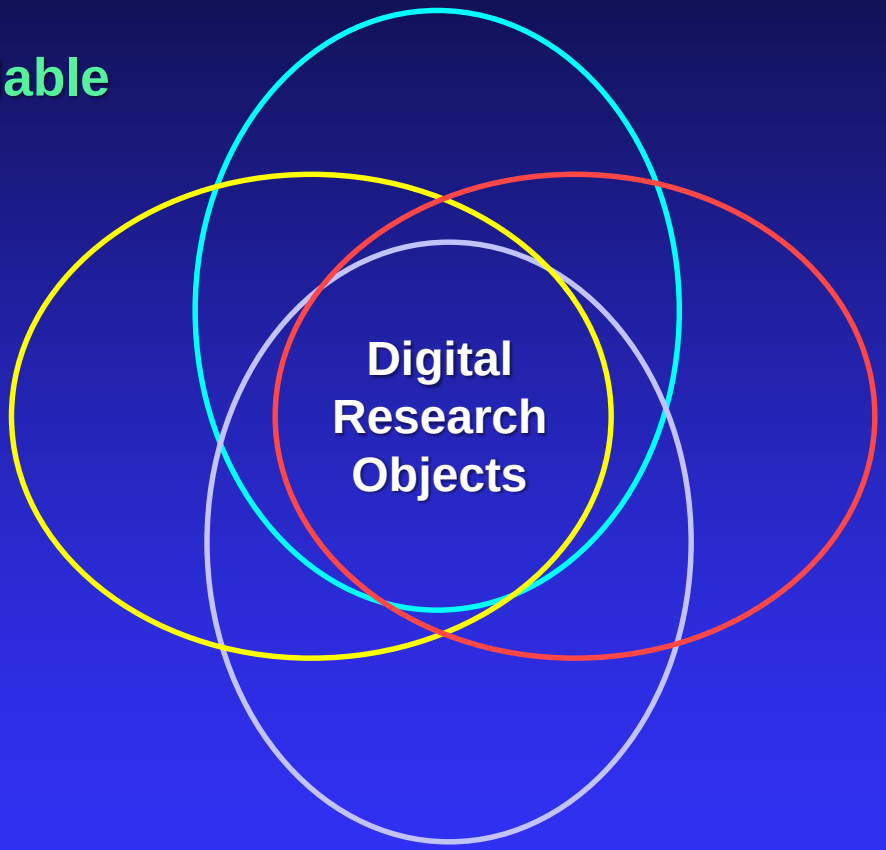



Biomedical Science
Data Science
Open Science
Informatics
Information Science

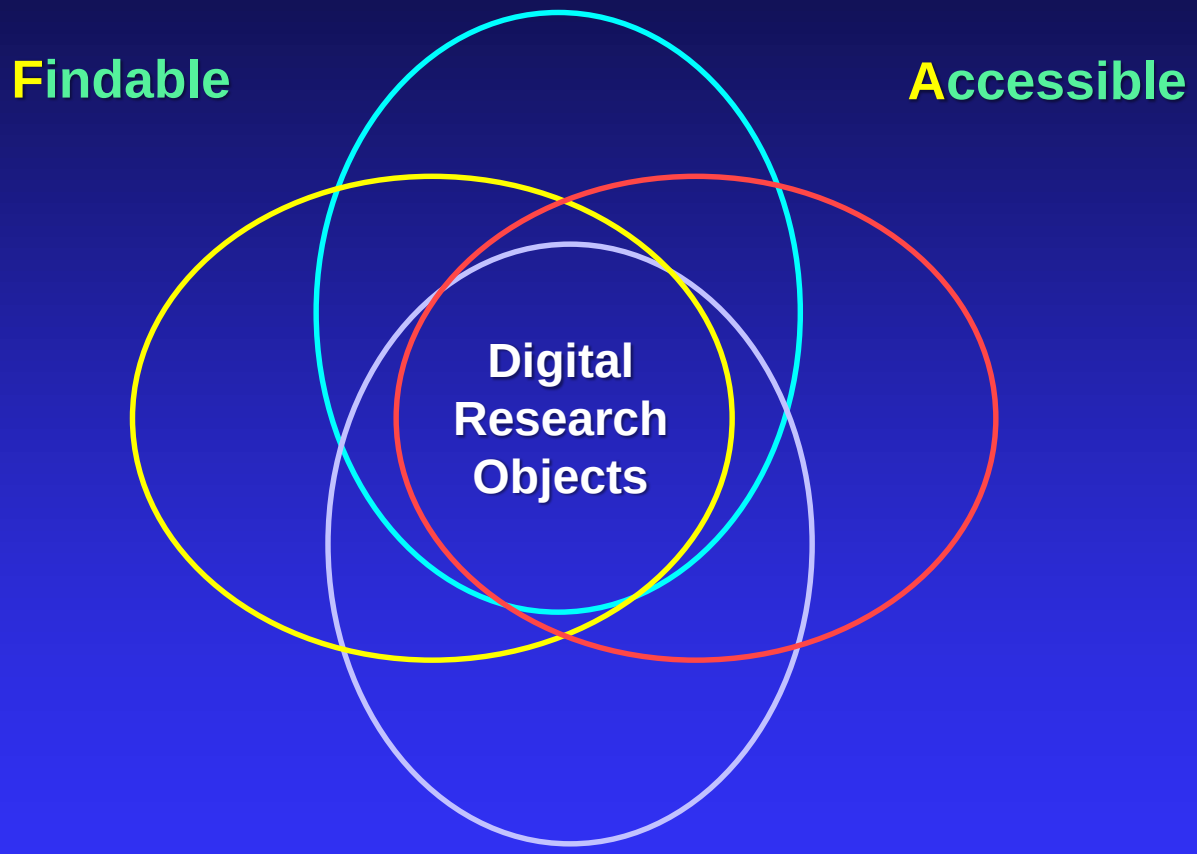

Digital
Research
Objects

F

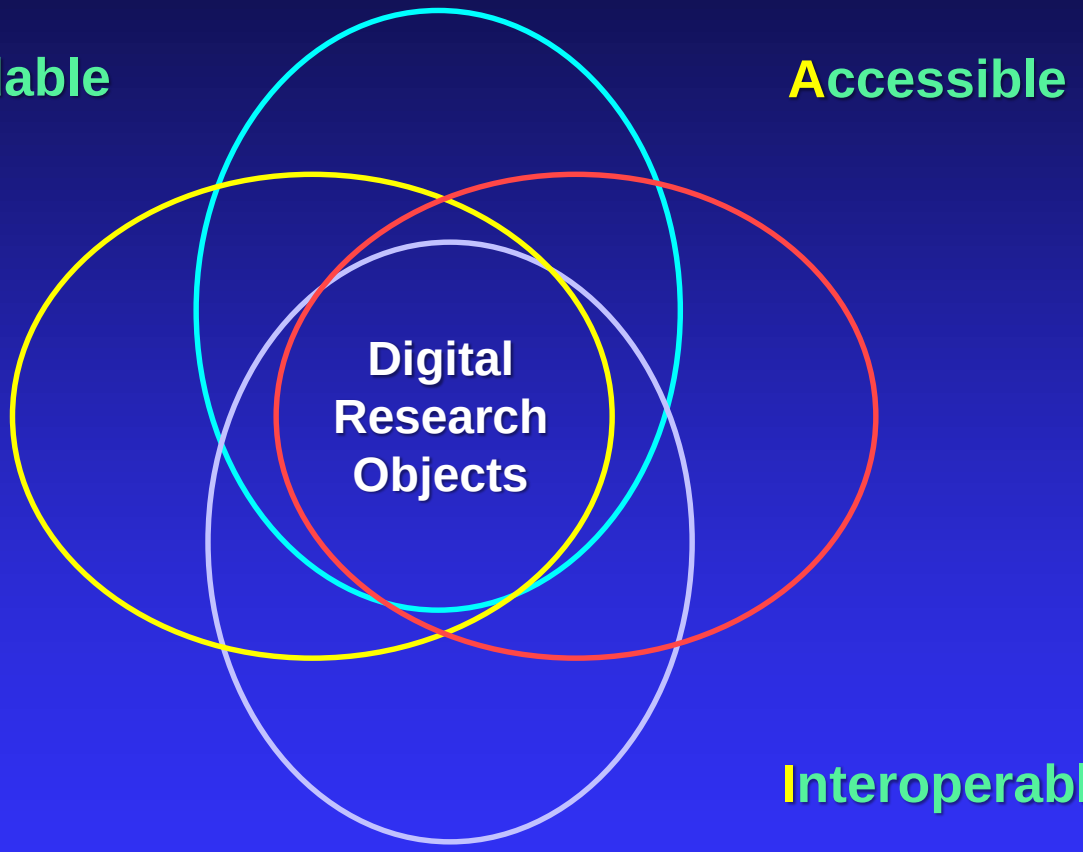
Findable



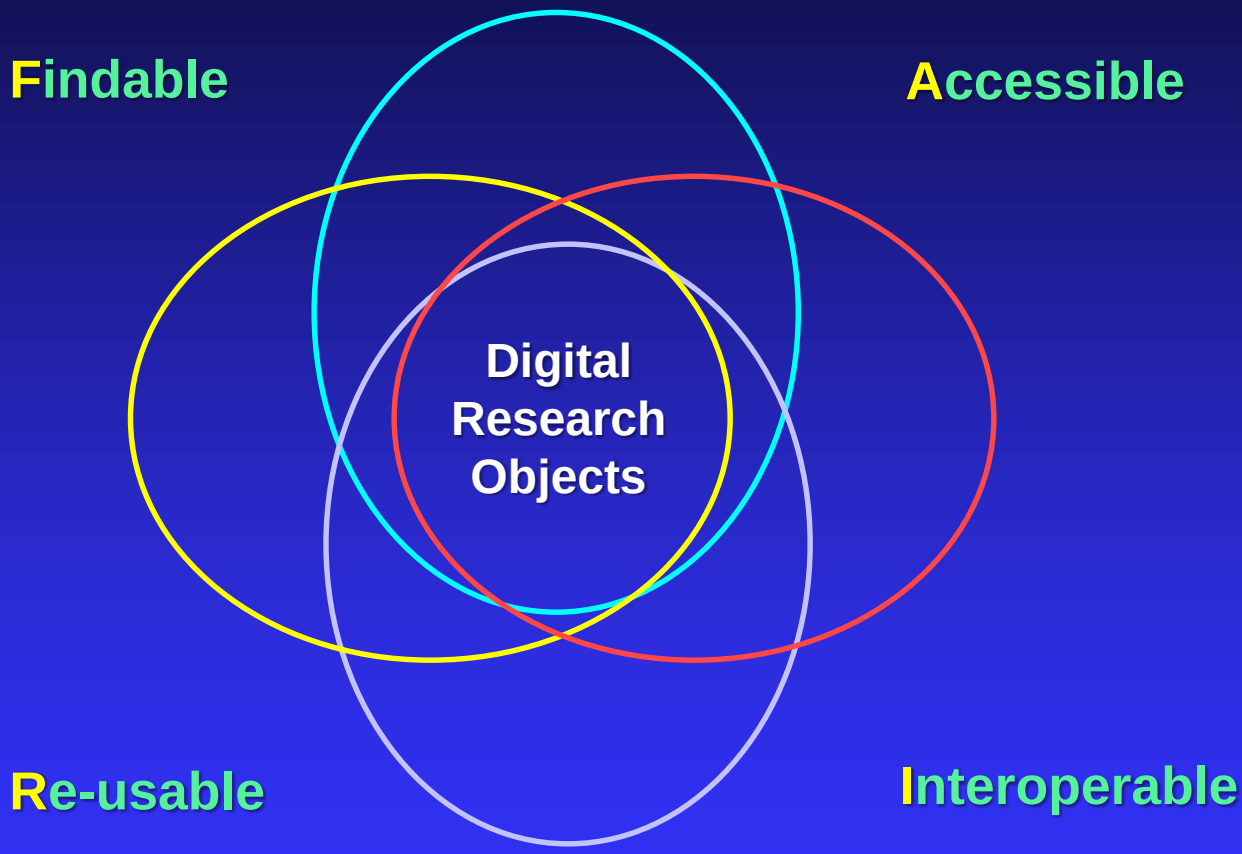
FA

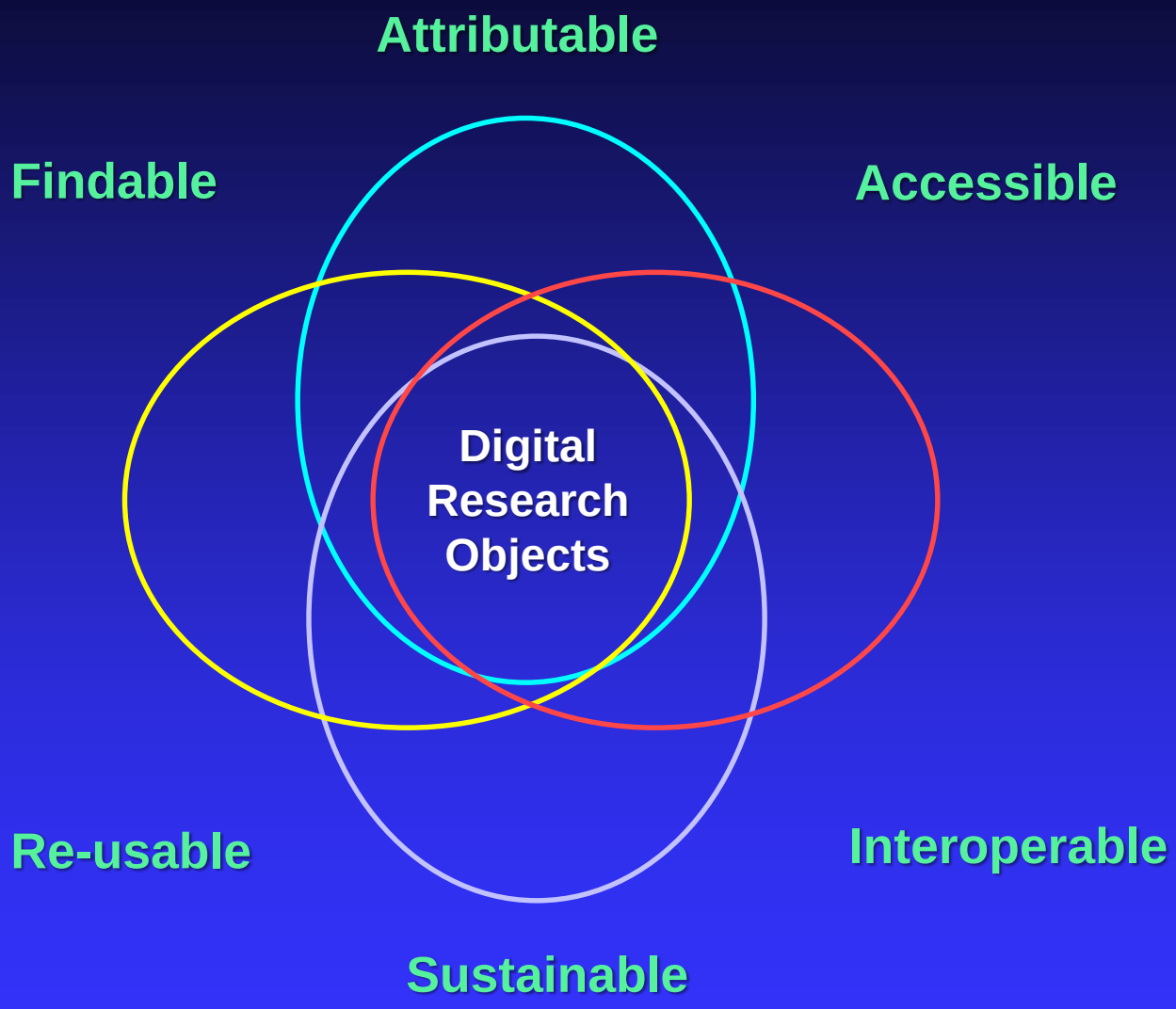


FAI


Findable
Accessible
Digital
Research
Objects
Interoperable

FAIR


Findable
Accessible
Digital
Research
Objects
Re-usable
Interoperable


Attributable
Findable
Accessible
Digital
Research
Objects
Re-usable
Interoperable
Sustainable


GTR
Genetic Testing Registry
ClinVar
Clinically relevant variation
MedGen
Conditions with a genetic component


PubMed


DailyMed

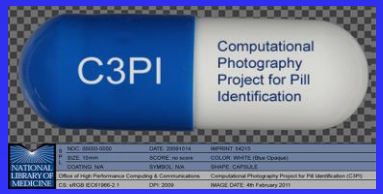

GenBank


PubChem


BioProject


NIH CDE
Repository


CHEMM
CHEMICAL HAZARDS EMERGENCY MEDICAL MANAGEMENT

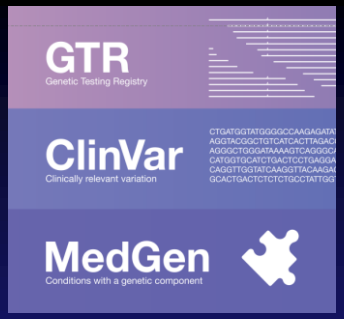

C3PI
Computational
Photography
Project for Pill
Identification


MEDLINE


How do you say glucose?
LOINC
the clinical data exchange lingua franca


REMM


PubMed Central
A free archive of life sciences journals


GTR
Genetic Testing Registry
ClinVar
Clinically relevant variation
MedGen
Conditions with a genetic component


MeSH
MeSH on Demand


PubMed


Daily Med


itk


GenBank


SNOMED CT
The global language of healthcare


NIH
Value Set Authority Center
U.S. National Library of Medicine


PubChem


NCBI
BLAST


OPEN i SM


BioProject

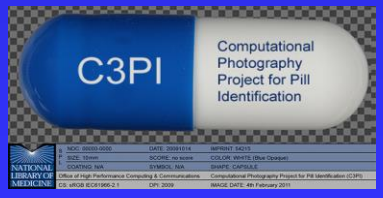

Genetics Home Reference


MTI
NLM Medical Text Indexer
Providing Indexing Assistance Since 2002
Biomedical Literature
MTI MTIFL
MeSH Suggestions


NIH CDE Repository


Rxnorm


CHEMM
CHEMICAL HAZARDS EMERGENCY MEDICAL MANAGEMENT


C3PI
Computational Photography Project for Pill Identification


MEDLINE


How do you say glucose?
LOINC
the clinical data exchange lingua franca


WISER


REMM


MedlinePlus CONNECT
Trusted Health Information for You


PubMed Central
A free archive of life sciences journals
GTR
Genetic Testing Registry
ClinVar
Clinically relevant variation
MedGen
Conditions with a genetic component
MeSH
MeSH on Demand
PubMed
ClinicalTrials.gov
Daily Med
itk
GenBank
SNOMED CT
The global language of healthcare
NIH
Value Set Authority Center
U.S. National Library of Medicine
Unified Medical Language System
PubChem
NCBI
BLAST
NLM APIs
OPEN i SM
BioProject
Genetics Home Reference
TOXNET
Toxicology Data Network
NIH CDE Repository
Rxnorm
MTI
NLM Medical Text Indexer
Providing Indexing Assistance Since 2002
Biomedical Literature
MTI MTIFL
MeSH Suggestions
CHEMM
CHEMICAL HAZARDS EMERGENCY MEDICAL MANAGEMENT
C3PI
Computational Photography Project for Pill Identification
MEDLINE
How do you say glucose?
LOINC
the clinical data exchange lingua franca
WISER
REMM
MedlinePlus CONNECT
Trusted Health Information for You

Data Science at NIH

Data Science at NIH

Advisory Committee to the NIH Director Recommended:

Data Science at NIH

Advisory Committee to the NIH Director Recommended:

- **Data Science** - *NLM should be the intellectual and programmatic epicenter for data science at NIH* and stimulate its advancement throughout biomedical research and application.

Data Science at NIH

Advisory Committee to the NIH Director Recommended:

- **Data Science** - *NLM should be the intellectual and programmatic epicenter for data science at NIH* and stimulate its advancement throughout biomedical research and application.
- **Open Science** - *NLM should lead efforts to support and catalyze open science*, data sharing, and research reproducibility, striving to promote the concept that biomedical information and its transparent analysis are public goods.

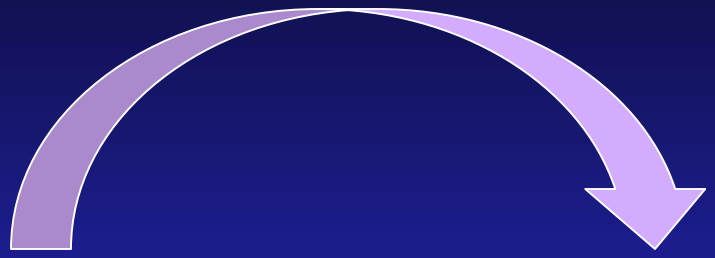
Data Science at NIH

Advisory Committee to the NIH Director Recommended:

- **Data Science** - *NLM should be the intellectual and programmatic epicenter for data science at NIH* and stimulate its advancement throughout biomedical research and application.
- **Open Science** - *NLM should lead efforts to support and catalyze open science*, data sharing, and research reproducibility, striving to promote the concept that biomedical information and its transparent analysis are public goods.

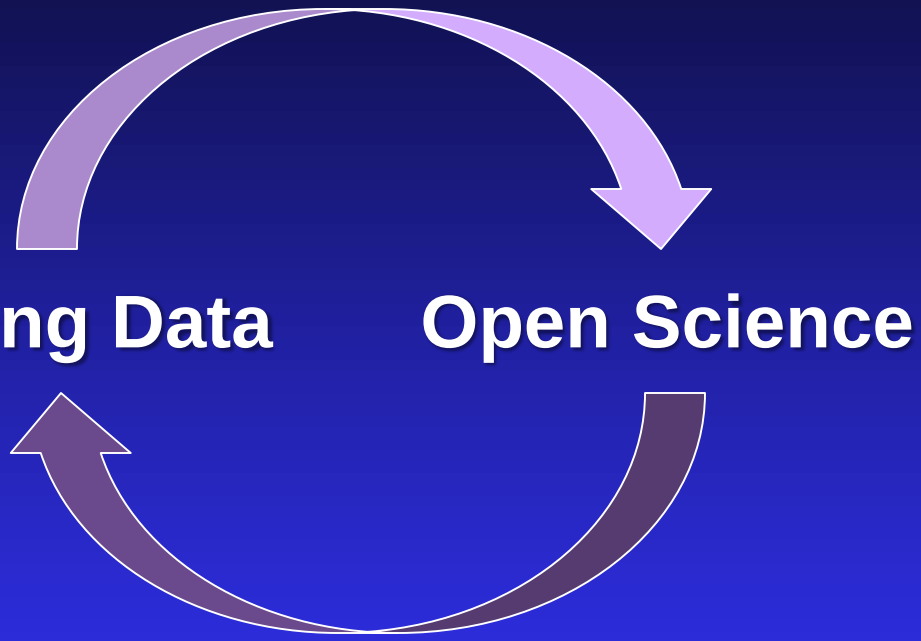
Recommendations Accepted


Data
Science



Sharing Data

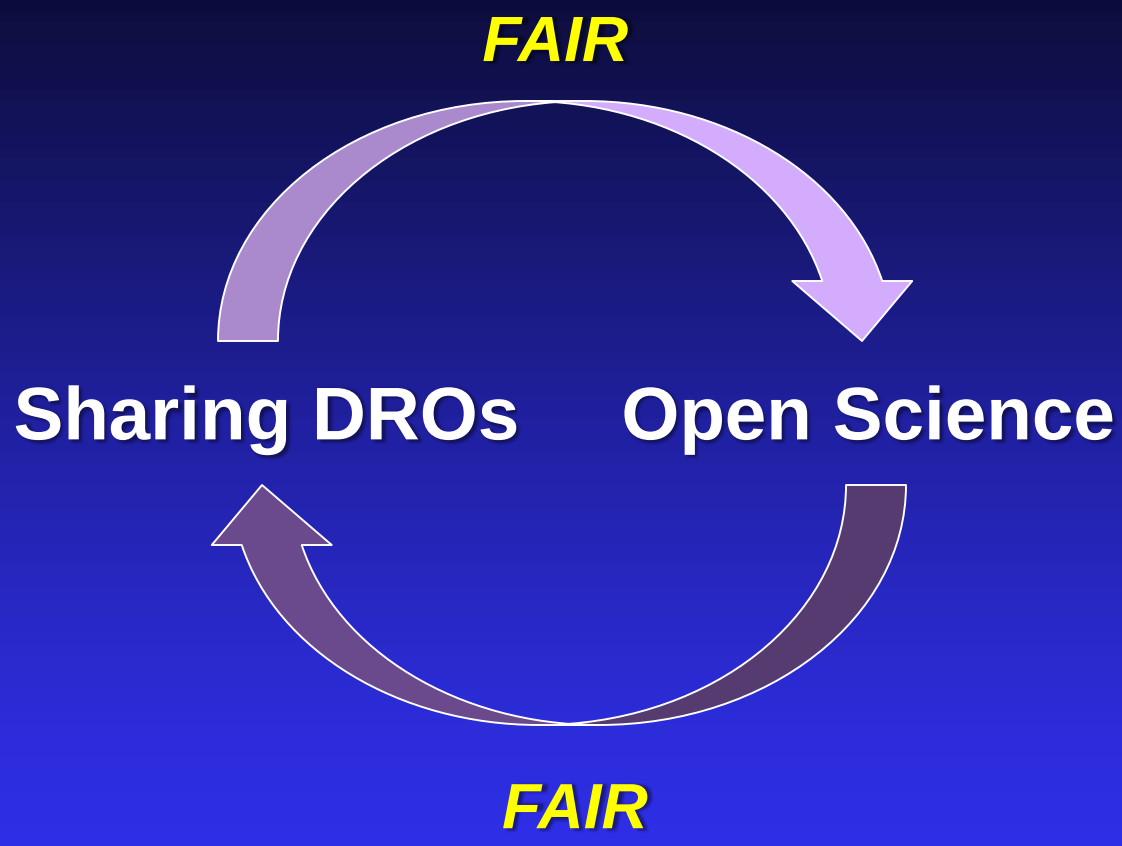
Open Science

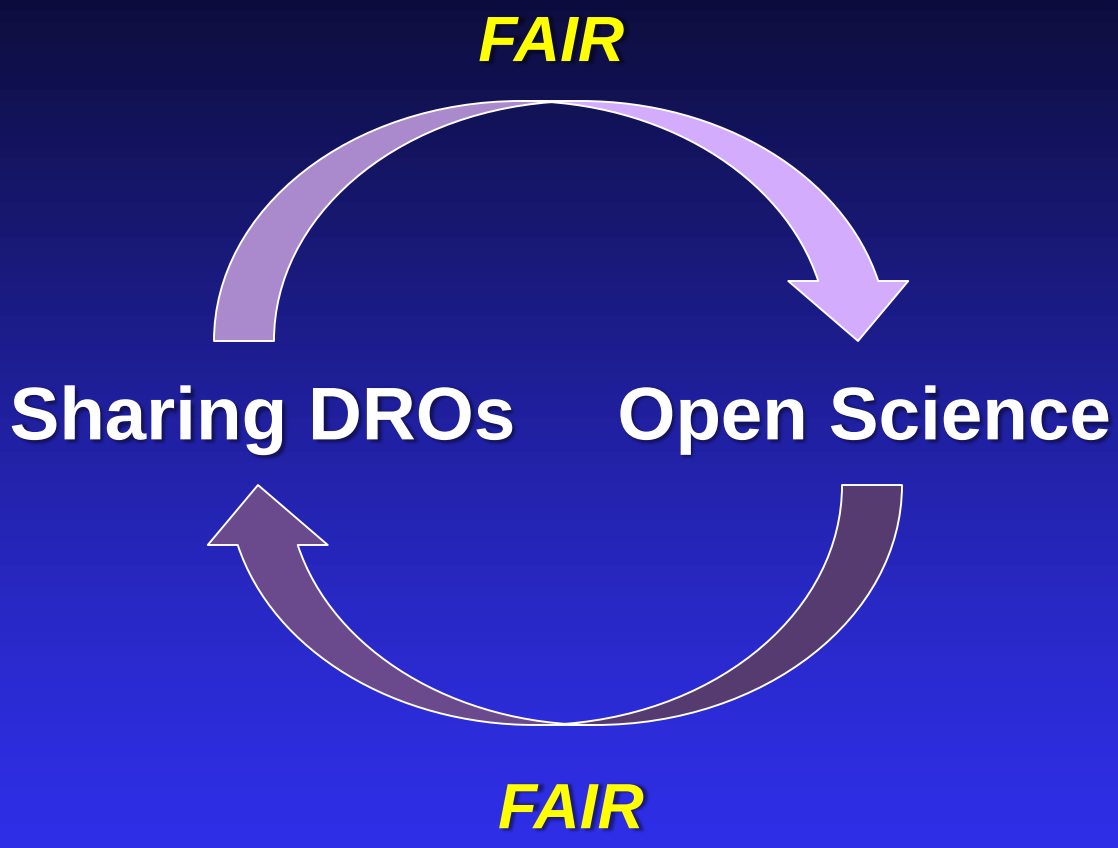


Data-centric & open paradigm

*Digital Research Objects

- Data
- Software
- Publications
- Workflows, etc.

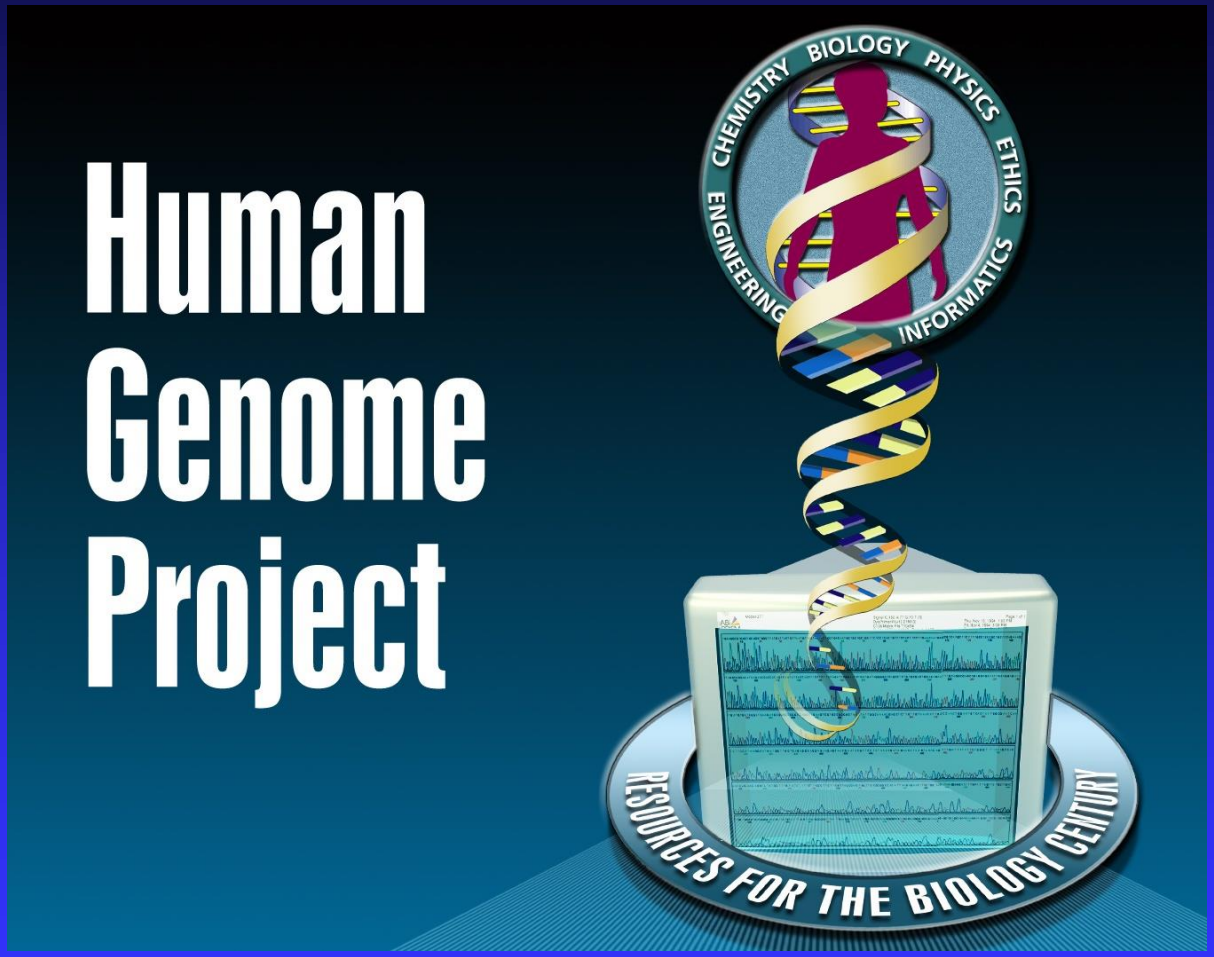

FAIR
Sharing DROs
Open Science
FAIR

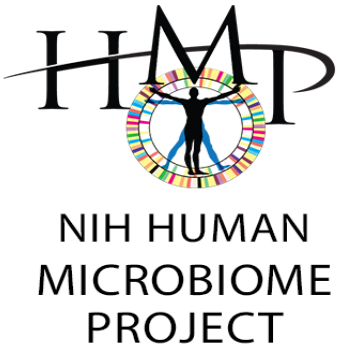


Data-centric & open paradigms
have proven successful

NIH Support for Data-Centric & Open Science

Large Scale Data




HMP
NIH HUMAN
MICROBIOME
PROJECT

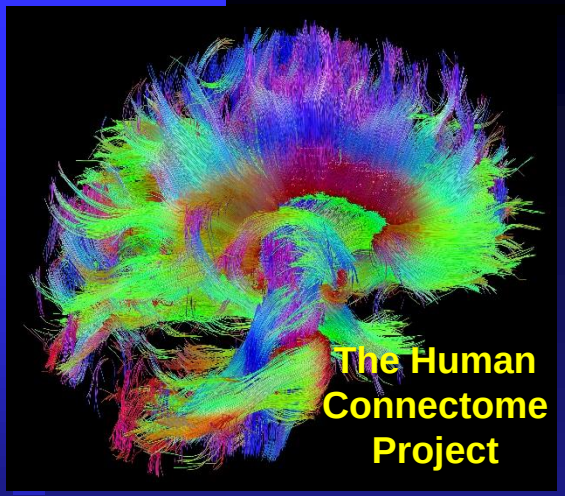

FlyBase


MGI


ZFIN
The Zebrafish Model Organism Database


SGD
Saccharomyces
GENOME DATABASE


WormBase



NIH
The Common Fund

GTE

HMP
NIH HUMAN MICROBIOME PROJECT

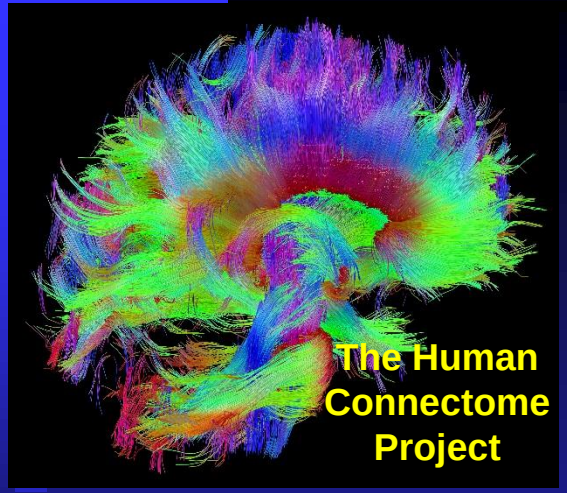
FlyBase

MGI

ZFIN
The Zebrafish Model Organism Database

SGD Saccharomyces GENOME DATABASE

WormBase



NIH
The Common Fund

GTE

HMP
NIH HUMAN MICROBIOME PROJECT

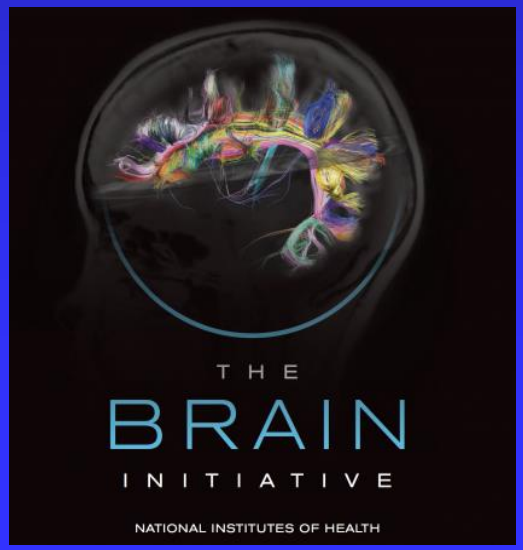
All of Us℠
THE FUTURE OF HEALTH BEGINS WITH YOU
The Precision Medicine Initiative®

CANCER MOONSHOT

ECHO
Environmental Influences on Child Health Outcomes

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

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



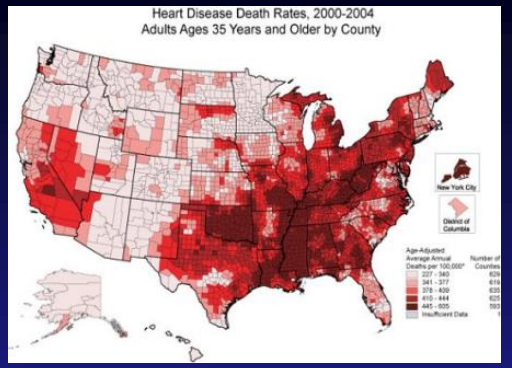
Data Centric & Open Science

- Requires soft & hard infrastructure:
  - Clear & heeded policies of funders, publishers, etc.
  - Widely-used data-related standards (incl metadata)
  - Data repositories, platforms & tools
  - Appropriate incentive structure

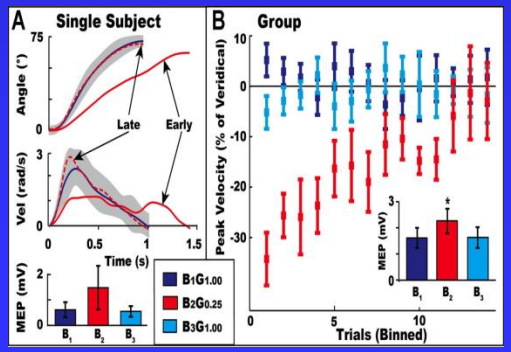
Data Centric & Open Science

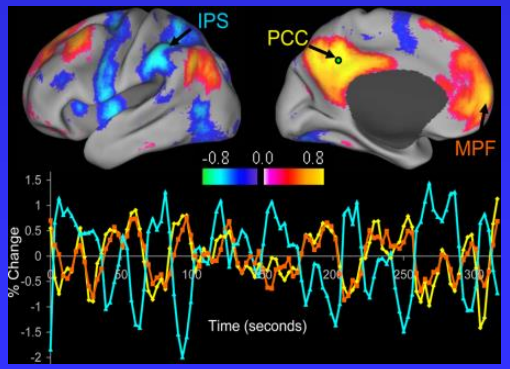
- Requires soft & hard infrastructure:
  - Clear & heeded policies of funders, publishers, etc.
  - Widely-used data-related standards (incl metadata)
  - Data repositories, platforms & tools
  - Appropriate incentive structure

- When implemented with FAIR principles
  - Forms basis of digital ecosystem – transformational
    - Accelerating pace of discovery
    - Changing the nature of discovery



Most domains of biomedical research are **neither data-centric nor open**





For these domains, the major public products of research are **scientific papers** that describe the authors' **conclusions about** the data...

…but the underlying **data are never seen**.

…but the underlying **data are never seen**.

Much less shared

This is about to change

Societal expectations

Societal expectations

Policy directives

Societal expectations

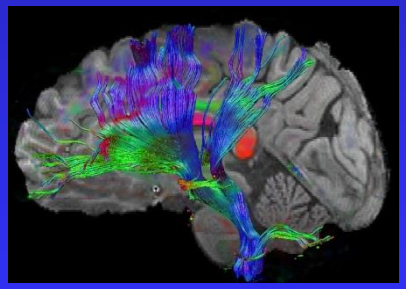
Policy directives

Technical capabilities

Societal expectations

Policy directives

Technical capabilities



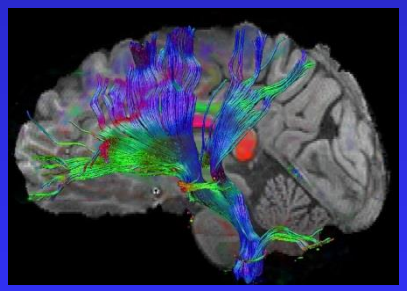
Scientific opportunities

Societal expectations

Policy directives

More Sharing & More Open Across All Domains

Technical capabilities



Scientific opportunities

DataScience@NIH

Build on NIH-Wide Opportunities

DataScience@NIH

Build on NIH-Wide Opportunities

- Findable – PubMed
  - Finding literature
  - Finding data via PubMed Central data deposit, Link Out, etc.

DataScience@NIH

Build on NIH-Wide Opportunities

- Findable – PubMed
  - Finding literature
  - Finding data via PubMed Central data deposit (10/17)
- Accessible - Holdren Memo to increase access
  - NIH plan for publications – PubMed Central
  - NIH plan for data – Peer reviewed DMP for all research
  - Many repositories open for data deposit and withdrawal

DataScience@NIH

Build on NIH-Wide Opportunities

- Findable – PubMed
  - Finding literature
  - Finding data via PubMed Central data deposit, Link Out, etc.
- Accessible - Holdren Memo to increase access
  - NIH plan for publications – PubMed Central
  - NIH plan for data – Peer reviewed DMP for all research
  - Many repositories open for data deposit and withdrawal
- Interoperable - Standards
  - NLM – UMLS, SNOMED-CT, LOINC, RxNorm, etc.
  - Repository & Initiative-related standards across NIH
  - NIH Clinical Common Data Element Task Force

DataScience@NIH

Build on NIH-Wide Opportunities

- Findable – PubMed
  - Finding literature
  - Finding data via PubMed Central data deposit, Link Out, etc.
- Accessible - Holdren Memo to increase access
  - NIH plan for publications – PubMed Central
  - NIH plan for data – Peer reviewed DMP for all research
  - Many repositories open for data deposit and withdrawal
- Interoperable - Standards
  - NLM – UMLS, SNOMED-CT, LOINC, RxNorm, etc.
  - Repository & Initiative-related standards across NIH
  - NIH Clinical Common Data Element Task Force
- Re-usable – Linking systems of DROs
  - PubMed – publication & data citations
  - NIH data repositories - data
  - NIH administrative systems (info about grants, DMPs, PIs, etc.)
  - NIH Data Commons Cloud – Shared space for compliant DROs, tools, compute, etc.

NIH
Big Data to Knowledge(BD2K)

Pivot to the Future

Pivot to the Future

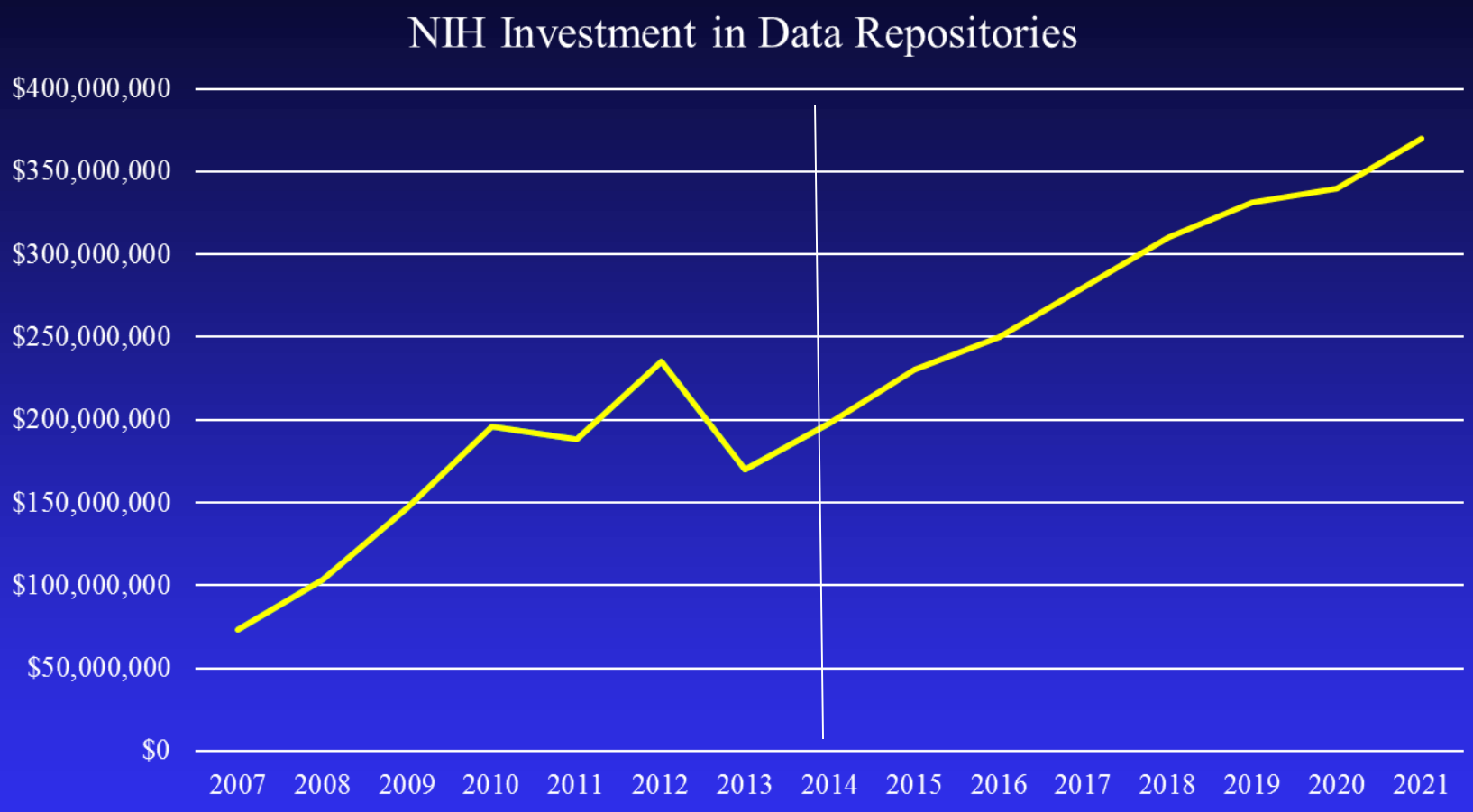
Strategic Engagement Across & Beyond NIH

Pivot to the Future

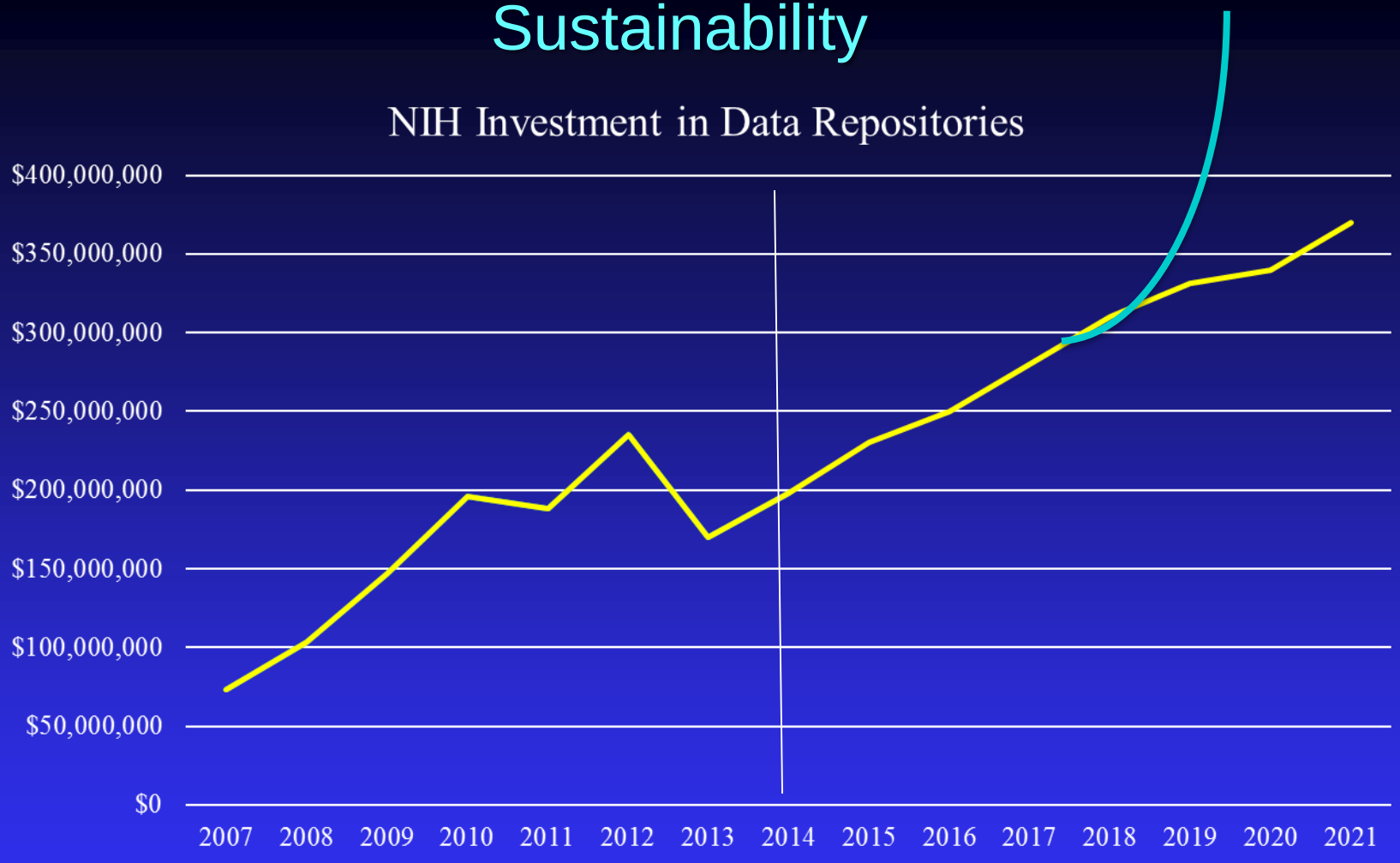
Strategic Engagement Across & Beyond NIH

- Sustainability solutions – urgent to address

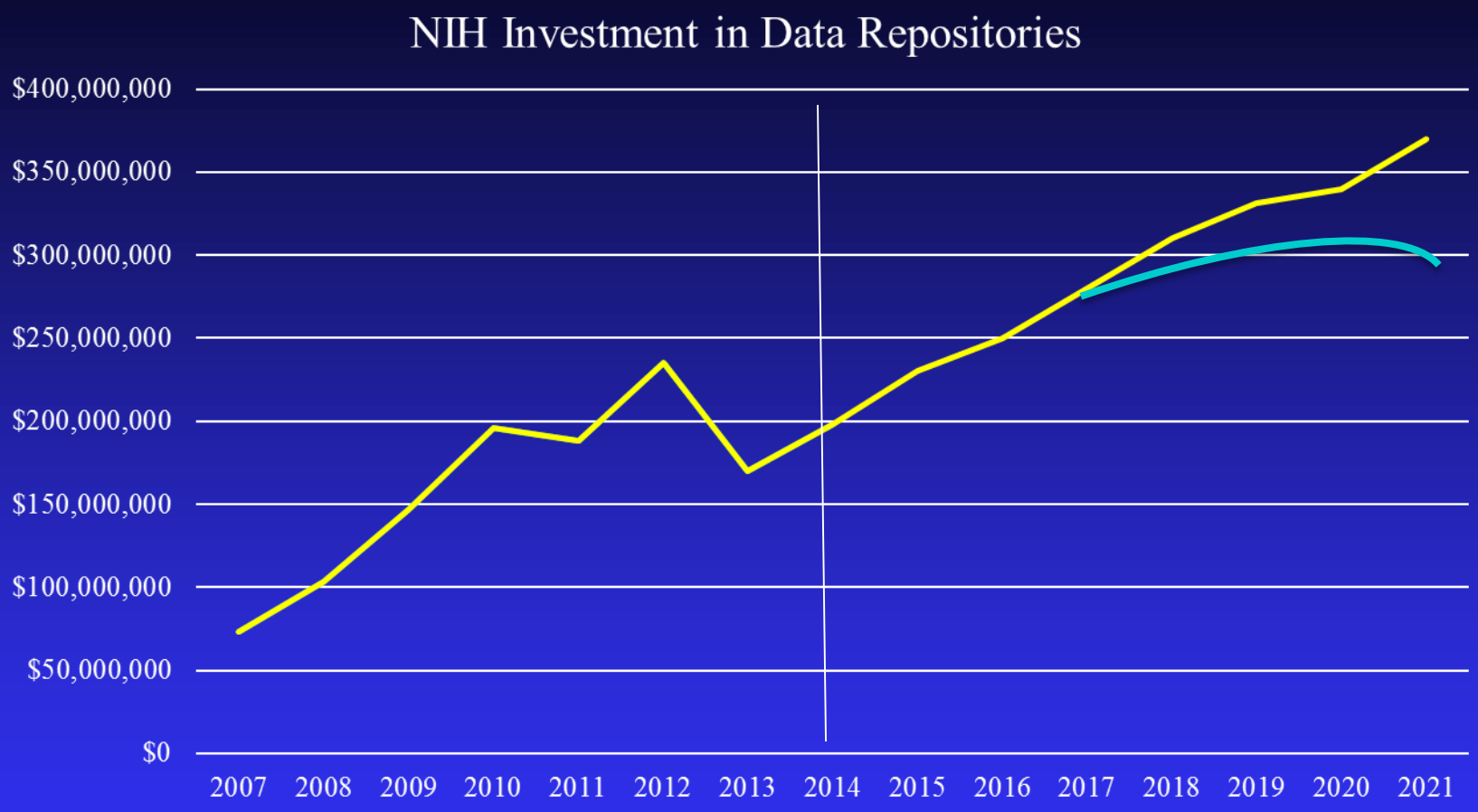
Sustainability



Sustainability



Sustainability



Strategic approach may bend the cost-curve

Pivot to the Future

Strategic Engagement Across & Beyond NIH

- Sustainability solutions

Pivot to the Future

Strategic Engagement Across & Beyond NIH

- Sustainability solutions
  - Enterprise-wide approaches (balance w IC needs)
    - Solve common problems once
    - Lessons learned & best practices
    - Converge on common:
      - Data-related standards
      - Architectures
      - Acquisitions
      - Operational approaches

Pivot to the Future

Strategic Engagement Across & Beyond NIH

- Sustainability solutions
  - Enterprise-wide approaches (balance w IC needs)
    - Solve common problems once
    - Lessons learned & best practices
    - Converge on common:
      - Data-related standards
      - Architectures
      - Acquisitions
      - Operational approaches
  - Evidence-based value assessment for investment in policy changes, infrastructure, data acquisition, preservation, etc.
    - Cost vs benefit analyses
    - Develop and use evidence base & models

Pivot to the Future

Strategic Engagement Across & Beyond NIH

Pivot to the Future

Strategic Engagement Across & Beyond NIH

- Grow a talented workforce intra- & extramural
  - Data science experts
  - Train across bio & data science
  - NIH staff – research, technical, program, review & policy

Pivot to the Future

Strategic Engagement Across & Beyond NIH

- Grow a talented workforce intra- & extramural
  - Data science experts
  - Train across bio & data science
  - NIH staff – research, technical, program, review & policy
- Promote open science & citizen science
  - Evidence-based changes in policies & practices
  - Tools to empower research participants, patients, & citizens

Pivot to the Future

Strategic Engagement Across & Beyond NIH

- Grow a talented workforce intra- & extramural
  - Data science experts
  - Train across bio & data science
  - NIH staff – research, technical, program, review & policy
- Promote open science & citizen science
  - Evidence-based changes in policies & practices
  - Tools to empower research participants, patients, & citizens
- Continue research & innovation in data science
  - Artificial intelligence, analytics, statistics, probabilistics, etc.
  - At-scale curation (metadata, provenance, etc.)

Pivot to the Future

Strategic Engagement Across & Beyond NIH

- Grow a talented workforce intra- & extramural
  - Data science experts
  - Train across bio & data science
  - NIH staff – research, technical, program, review & policy
- Promote open science & citizen science
  - Evidence-based changes in policies & practices
  - Tools to empower research participants, patients, & citizens
- Continue research & innovation in data science
  - Artificial intelligence, statistics, probabilistics, analytics, etc.
  - At-scale curation (metadata, provenance, etc.)
- Strategic incentive structure for data-centric & open paradigm
  - Will require incentives to change behavior of people & orgs
  - Strategically align incentives across ecosystem to maximize impact
  - Likely best done domain-by-domain