Our Tax Dollars and the Future of Data:
FAIR Data Ecosystems and AI-Readiness

Gulf Coast Consortium
26 January 2022

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UNIVERSITY OF CALIFORNIA

Agenda

- SDSC and Research Data Services
- Context – Signposts in the growth of U.S. research data landscape
  - Academic and federal science
- FAIR Principles
  - What they are, and what they are not
  - GO FAIR!
- US efforts for Open Science
- AI Initiatives and Readiness

SDSC & HDSI: Opportunities for Collaborative Research and Learning

SDSC
www.sdsc.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE
datascience.ucsd.edu

SDSC and HDSI are now developing a proposal to form a School of Computation, Information, and Data Science.

San Diego Supercomputer Center

SDSC was founded in 1985 with a $170 million grant from the NSF Supercomputer Centers program. From 1997 to 2004, SDSC extended its leadership in computational science and engineering to form the National Partnership for Advanced Computational Infrastructure (NPACI), teaming with approximately 40 university partners around the country.

Since then, SDSC has continuously hosted national resources that concentrated on specialized needs, from Comet serving the long-tail of science, to Voyager, which is highlighted here.

From large-scale networking and high throughput computing expertise to internet research at CAIDA, SDSC is a leader in computing at scale. New initiatives are aimed the intersection of services for management, analysis, and sharing data at the 10s of Petabytes level. This work supports multidisciplinary programs in academia, industry, and government that address a wide variety of grand challenges from astrophysics and earth sciences to disease research and drug discovery.

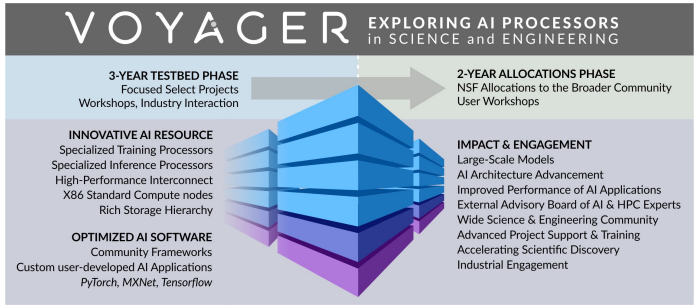
- **CAIDA**
  - Integrated Laboratory for Advanced Network Data Science (ILANDS): Producing data for understanding the changing Internet. (NSF #2120399)
- **TILOS**
  - Pioneering optimization to transform chip design, robotics, networks and other us-inspired domains that are vital to our nation's health, prosperity, and welfare. (NSF #2112665)

The **Research Data Services** (RDS) division at SDSC provides a complete suite of services: research data consulting; active and archival storage (TB to PB scales); virtual and physical compute platforms; data center colocation; enterprise networking. The R+D portfolio includes local and federally-funded projects that address data policy and regulatory management, design and delivery of technical and training resources, and advancing FAIR and Trusted data resources across the academic research ecosystem. We lead research on FAIR digital object infrastructure, general problems in AI reproducibility, and community readiness for data sharing and evolving public access policy. In the Earth Sciences, for example, RDS runs the **EarthCube Coordination Office**, which

- Bridges the divide between geoscientists and cyberinfrastructure builders to promote interdisciplinary earth and space sciences.
- Leads efforts for FAIR and Research Data Management, including:
  - Promoting computational notebooks as scholarship
  - Producing new training and resources for FAIR implementation

The San Diego Supercomputer Center (SDSC) at the University of California, San Diego, which is normally associated with supplying computational power to the scientific community, was one of the earliest organizations to recognize the need to add data to its mission.

*Gordon Bell, 2009**



Voyager is our new high-performance resource for conducting artificial intelligence (AI) research across a wide swath of science and engineering domains. This innovative system will be the first-of-its-kind available in the NSF resource portfolio – its architecture is uniquely optimized for deep learning (DL) operations and AI workloads.

Halıcıoğlu Data Science Institute

HDSI was founded in 2018 as a fully independent academic unit. The institute is based on 3 pillars: academics, research, and industry.

Academics: The institute has its own data science undergraduate degree that gives equal balance to CS, Math and Cognitive Science. Course work is combined with a 2-quarter capstone project for Seniors that includes industry mentorship. The aim is to enable skills and problem-solving ability to tackle the real-world data science challenges of the future.

Research: HDSI has its own faculty (11 hired in the last year), and 200 affiliated faculty across campus representing every department and school. HDSI serves a diverse network of scholars who are all applying data science to solve problems and advance knowledge in their respective research areas.

Industry: Collaborations and industrial partnerships are critical parts of the data science ecosystem. In addition to mentoring and recruiting, powerful research collaborations and knowledge transfer happen when industry and academia, and talent come together. Sponsored research and HDSI Board leadership adds significant value, and there is innovation around dataset production.

New partnership provides unique data sets from National Lab

In collaboration with the Lawrence Livermore National Laboratory (LLNL) Data Science Institute, the UCSD Library and our Data Science Librarian, HDSI is hosting and curating scientific datasets from LLNL. Six unique data sets are now publicly available for research and learning via the Lawrence Livermore National Laboratory Open Data Initiative, including:

- "Cars Overhead with Context (COWC)" (annotated images)
- "The JAG inertial confinement fusion simulation data set for multi-modal scientific deep learning" (pre-trained ML models)

https://doi.org/10.6075/J0HD7T2Q

EarthCube
Transforming Geosciences Research

TILOS
Institute for Learning-Enabled Optimization at Scale

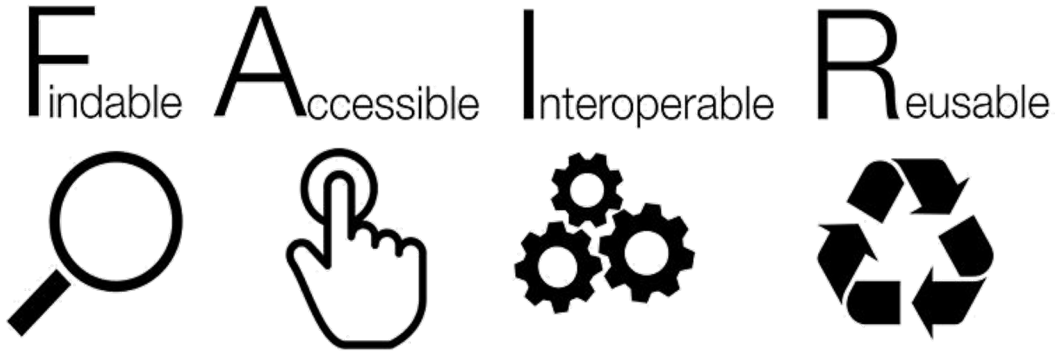
caida
Center for Applied Internet Data Analysis

* Hey, T., Tansley, S., and Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research.

UNIVERSITY OF CALIFORNIA

FAIR is…

A set of principles that describe the attributes data need to have to enable and enhance reuse, by humans and machines.



FAIR stands for Findable, Accessible, Interoperable, and Reusable

Findable

- (Meta)data are assigned a globally unique and persistent identifier
- Data are described with rich metadata
- Metadata clearly and explicitly include in the identifier of the data it describes
- (Meta)data are registered or indexed in a searchable resource

Accessible

- (Meta)data are retrievable by their identifier using a standardized protocol
- The protocol is open, free and universal
- The protocol allows for authentication and authorization, as needed
- Metadata are accessible, even when the data are no longer available

Interoperable

- (Meta)data use a formal, accessible, shared and broadly applicable language
- (Meta)data use vocabularies that follow FAIR principles
- (Meta)data include qualified references to other (meta)data

Reusable

- (Meta)data are richly described with a plurality of accurate and relevant attributes
- (Meta)data are released with a clear and accessible data usage licence
- (Meta)data are associated with a detailed provenance
- (Meta)data meet domain-relevant community standards

https://www.ccdc.cam.ac.uk/researchandconsultancy/fair-data-principles/

FAIR Aims:

The Principles are:

- An aspiration, a journey
- Ambiguous
- A spectrum
- Domain respectful / specific
- Implementable with todays protocols and standards
- A small part of indicators
- A framework for prompting organizational change
- Work in progress

The Principles are not:

- A standard
- Strict
- One size fits all
- One domain
- Inventing new protocols
- Technology specific
- Anything to do with quality
- Synonymous with open
- An architecture
- "Tablets of stone"

Pillars

- GFISCO (GO FAIR International Support & Coordination Office)
- National Offices
  - France, Brazil, Germany, U.S.
- Mobilization around Pillars

Advancing FAIR in the US

Our Aim is to connect FAIR stakeholders and foster a community where FAIR approaches can be shared, discussed, and advanced collaboratively.

Read More

News

25 JAN 2022

GO FAIR USA's Christine Kirkpatrick Elected Secretary General for CODATA

Christine Kirkpatrick, GO FAIR US Office Head and Division Director of Research Data Services at the San Diego Supercomputer Center (SDSC) at UC San Diego was nominated by the U.S. National Academies Committee on Data to serve as Secretary General for the International Science Council's (ISC) Committee on Data (CODATA). The elections took place in November and Kirkpatrick was elected to the position, which lasts from 2022-2026.

13 DEC 2021

EarthCube FAIR How-to Series — Data Identifiers : Anchoring Your Data in Times of Change

Earthcube, as part of their FAIR "How-to Series," has developed a new training resource to educate researchers about how obtaining and utilizing identifiers for their datasets can make their materials more FAIR. Check out "Data Identifiers: Anchoring Your Data in Times of

Workshop : Advancing FAIR and GO FAIR in the U.S.

- Facilitate community of practice
- Improve understanding of FAIR technologies
  - how to teach this to others
- Preparation to support FAIR data management and policies in multiple contexts
- Collaboration with GFISCO & the South Big Data Innovation Hub

Workshop Funding - NSF Award #1937953

Upcoming workshop: FAIR for US

Goals:

1. Assemble a set of diverse stakeholders that represent experts in data and research, U.S. funders and Mission Agencies, and industry where appropriate.

2. Gather input
    a. Seek consensus on where the US wants to be in 3-5 years in respect to FAIR;
    b. US participation in the coordination of Global FAIR initiatives; and,
    c. Community-based positions on what's needed to clarify investment strategies for FAIR (who and what).

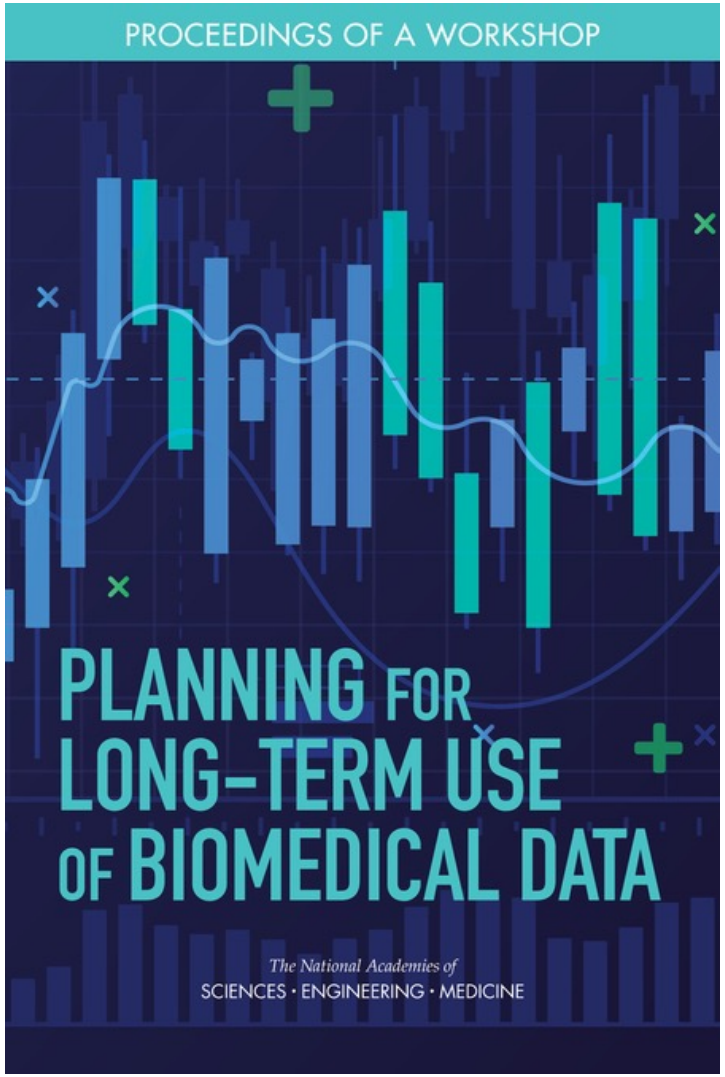
3. Surface divergent experiences, observations, and value assessments (perspectives) related to FAIR and institutional investment.

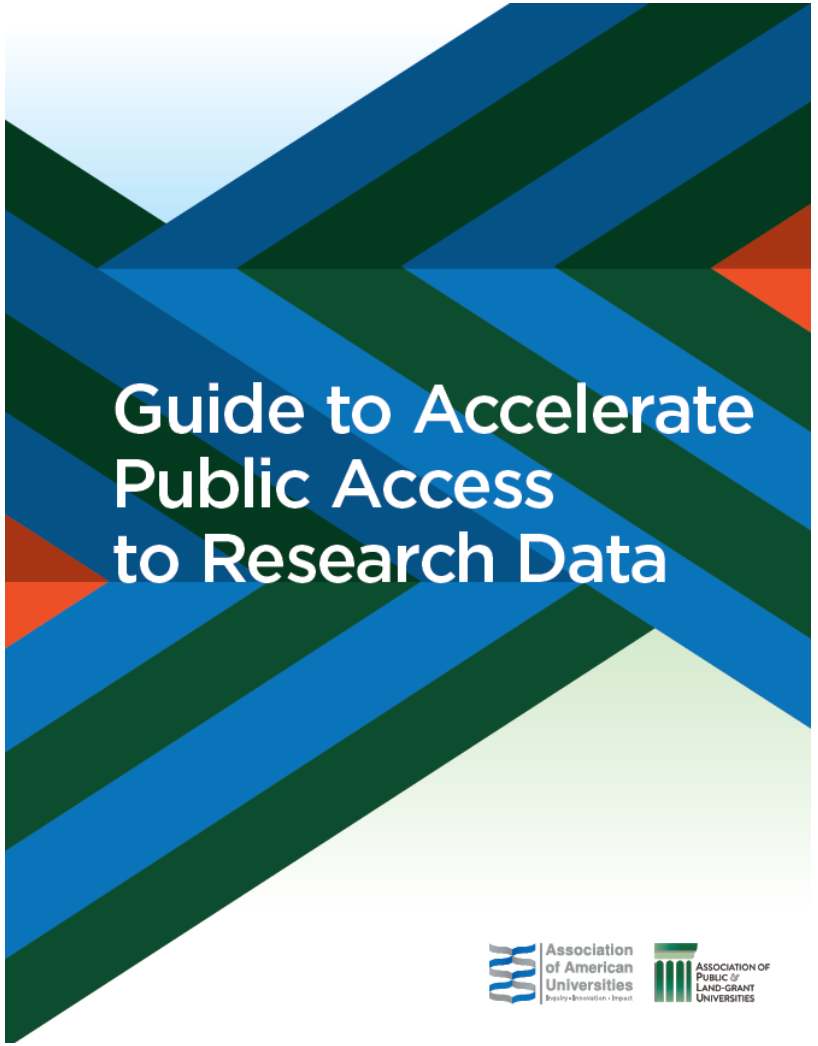
4. Scoping Report

Workshop Funding - NSF Award # 2138314

FAIR, Open Science, and AI Initiatives

US Data Policy, Open Science, and FAIR



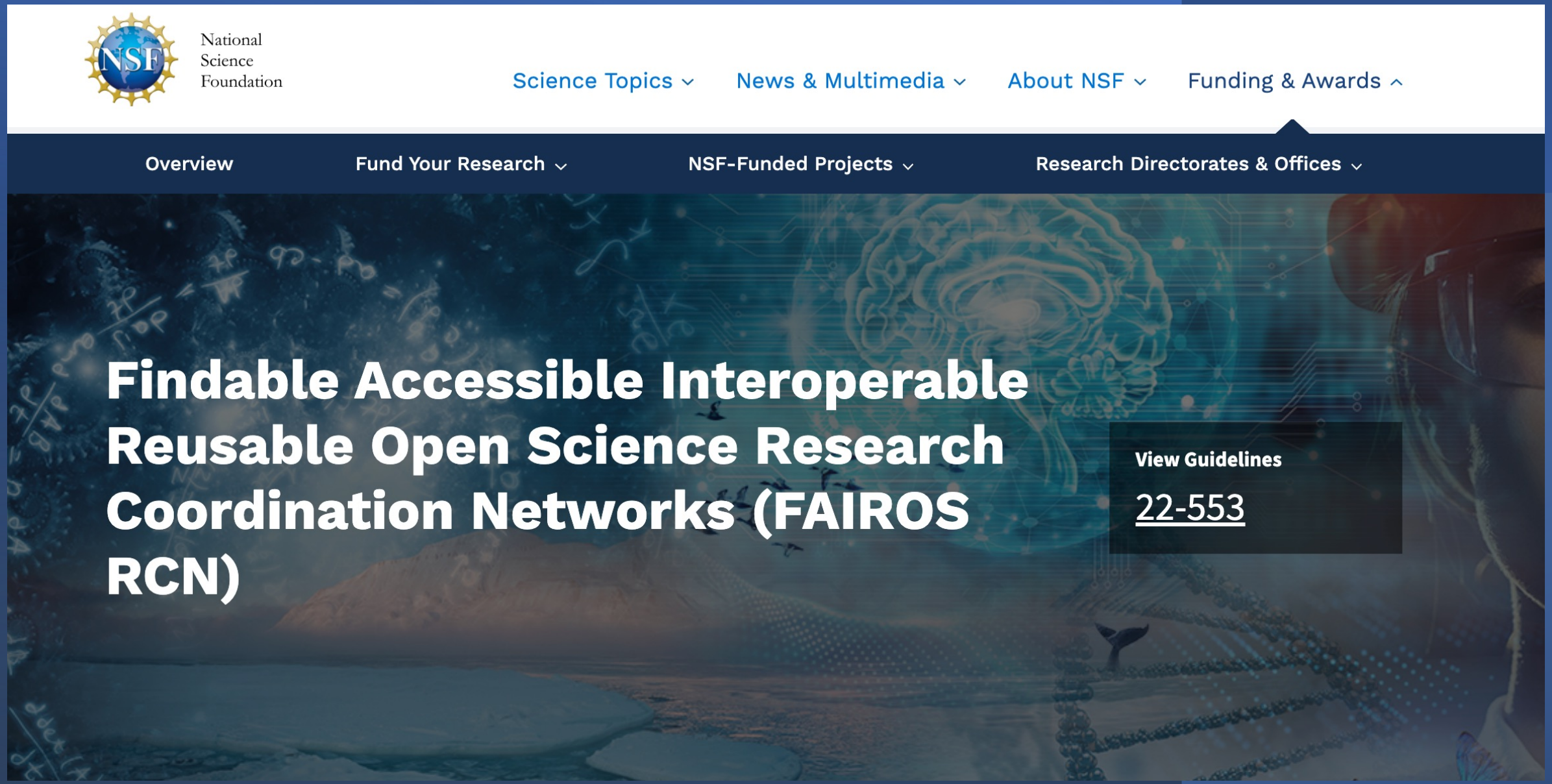


NSF Public Access Initiative

https://beta.nsf.gov/public-access

Components:

- Administration of ~70 awards annually, including
  - NSF National Big Data Hubs
  - other awards which catalyze advances in Public Access to Data (~$6M in 2021)
- Operation and development of NSF Public Access Repository (PAR)
- Coordination of Collaborative Internal (Intra-Agency) Relationships
- Coordination with External (Inter-Agency) Public Access Planning and Activities



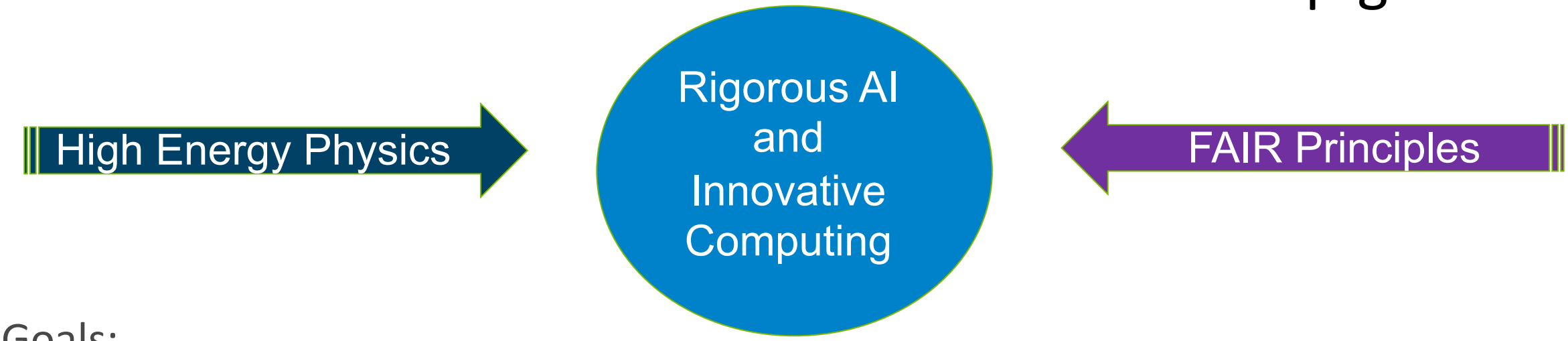
https://beta.nsf.gov/funding/opportunities/findable-accessible-interoperable-reusable-open-science-research-coordination

Recent DOE awards – FAIR and HPC

Award Number	Title	Institution	PI
DE-SC0021258	FAIR Framework for Physics-Inspired Artificial Intelligence in High Energy Physics	Board of Trustees of the University of Illinois, Champaign, IL	Neubauer, Mark
DE-SC0021418	FAIR Surrogate Benchmarks Supporting AI and Simulation Research	The Trustees of Indiana University, Bloomington, IN	Fox, Geoffrey
DE-SC0021352	FAIR Surrogate Benchmarks Supporting AI and Simulation Research (SBI)	Rutgers, The State University of New Jersey, Piscataway, NJ	Jha, Shantenu
DE-SC0021358	FAIR Data and Interpretable AI Framework for Architectured Metamaterials	Duke University, Durham, NC	Brinson, L. Catherine
DE-SC0021396	FAIR Framework for Physics-Inspired AI in High Energy Physics	The Regents of the University of California - UCSD, La Jolla, CA	Duarte, Javier
DE-SC0021395	FAIR Framework for Physics-Inspired AI in High Energy Physics	Regents of the University of Minnesota, Minneapolis, MN	Rusack, Roger
DE-SC0021293	HPC-FAIR: A Framework Managing Data and AI Models for Analyzing and Optimizing Scientific Applications	North Carolina State University, Raleigh, NC	Liu, Xu
DE-SC0021253	FAIR Data and Interpretable AI Framework for Architectured Metamaterials	California Institute of Technology, Pasadena, CA	Daraio, Chiara
DE-SC0021419	FAIR Surrogate Benchmarks Supporting AI and Simulation Research	The University of Tennessee, Knoxville, TN	Dongarra, Jack
DE-SC0021225	FAIR Framework for Physics-Inspired Artificial Intelligence in High Energy Physics	Massachusetts Institute of Technology, Cambridge, MA	Harris, Philip
DE-SC0021514	Developing a FAIR Digital Ecosystem for DOE Materials Modeling	Exabyte Inc., , CA	Bazhirov, Timur

FAIR4HEP

fair4hep.github.io

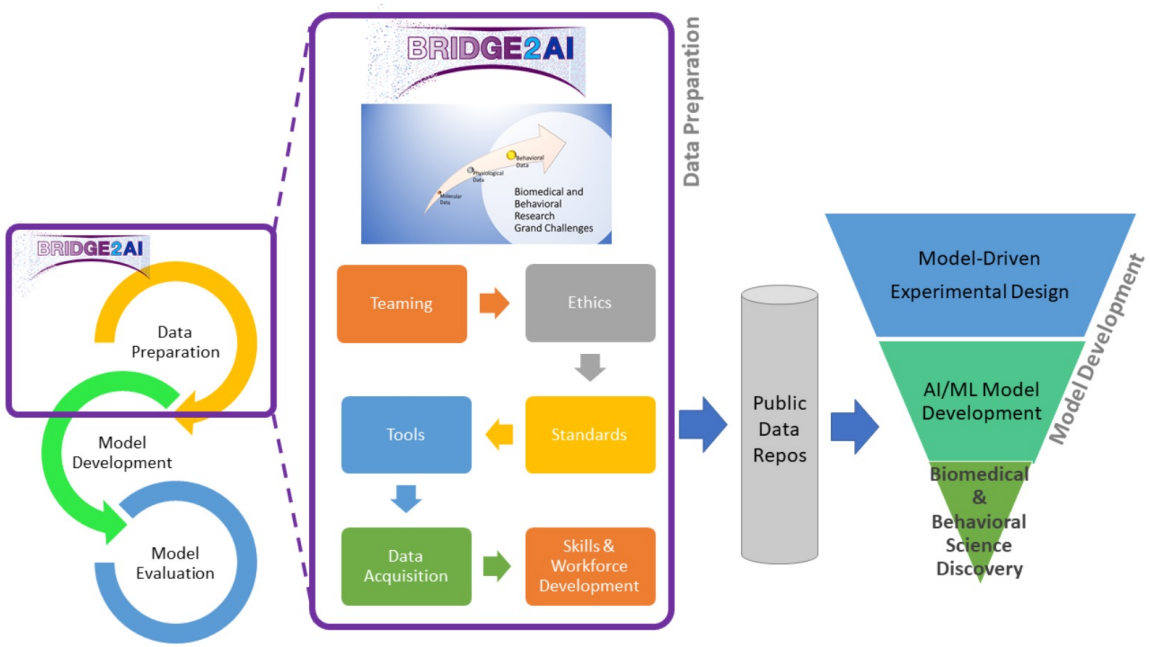


Goals:

- Create FAIR AI datasets and models
- Automate reusability of scientific datasets by humans and machines with little human intervention
- Unveil new connections between data and models
- Accelerate the creation of domain-informed, rigorous, interpretable, reusable and accelerated AI models that adhere to FAIR principles

NIH Initiatives

Bridge2ai



AIM-AHEAD

AIM AHEAD

Administrative Leadership | **Research**

Infrastructure | **Training**

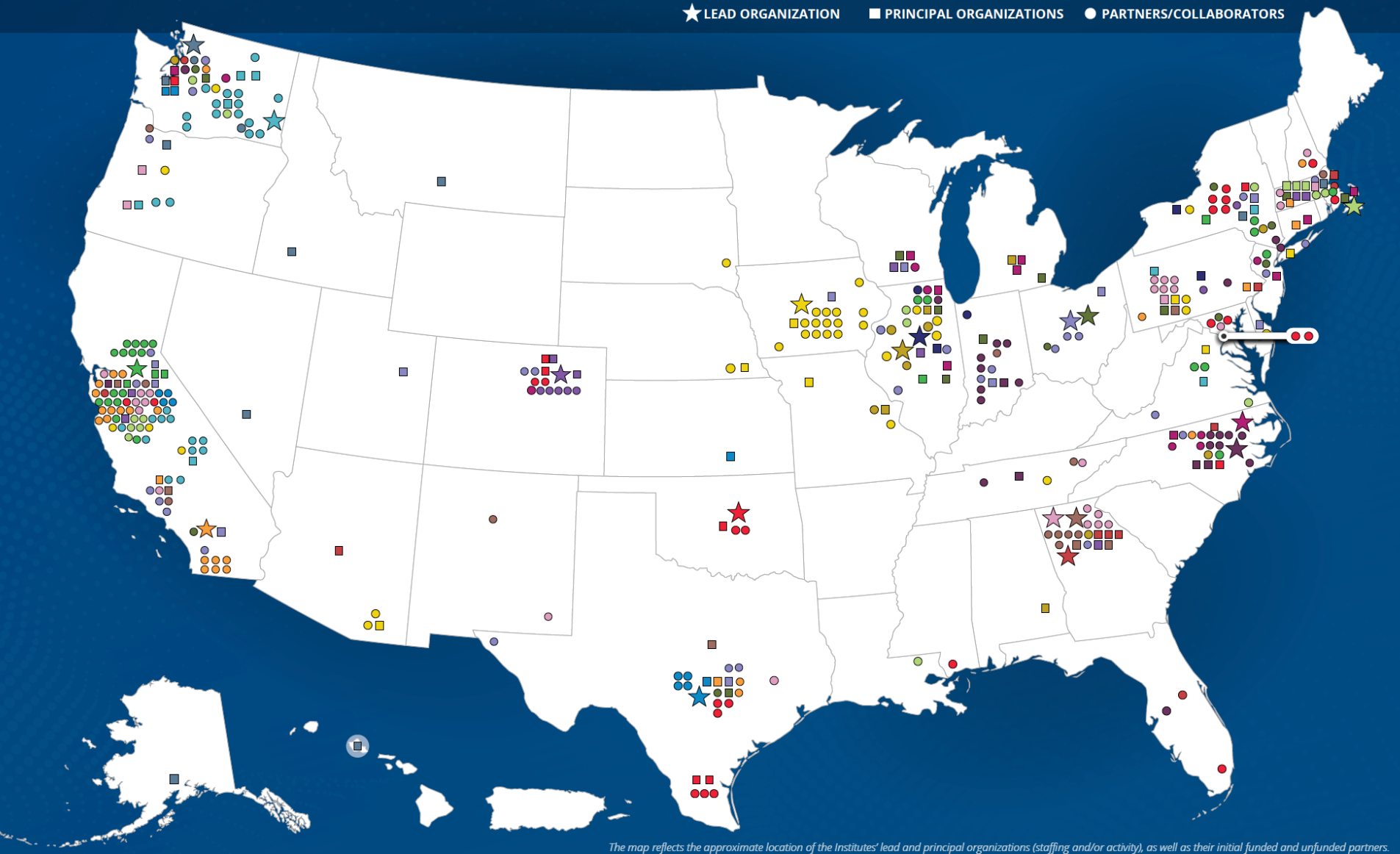
NSF-LED NATIONAL AI RESEARCH INSTITUTES

*The U.S. National Science Foundation (NSF) announced a **$220** million investment in eleven new Artificial Intelligence (AI) Research Institutes, building on the first round of seven AI Institutes totaling **$140** million funded last year. **(The default map view below shows all awards combined).***

This is an Interactive PDF and is best viewed using Adobe Acrobat. **Hover cursor** over dates below or **circles to the right** to display more information. If you have issues with these features you can download a sandard PDF available **here**.

2020 *Awards* **2021** *Awards*

★ LEAD ORGANIZATION ■ PRINCIPAL ORGANIZATIONS ● PARTNERS/COLLABORATORS

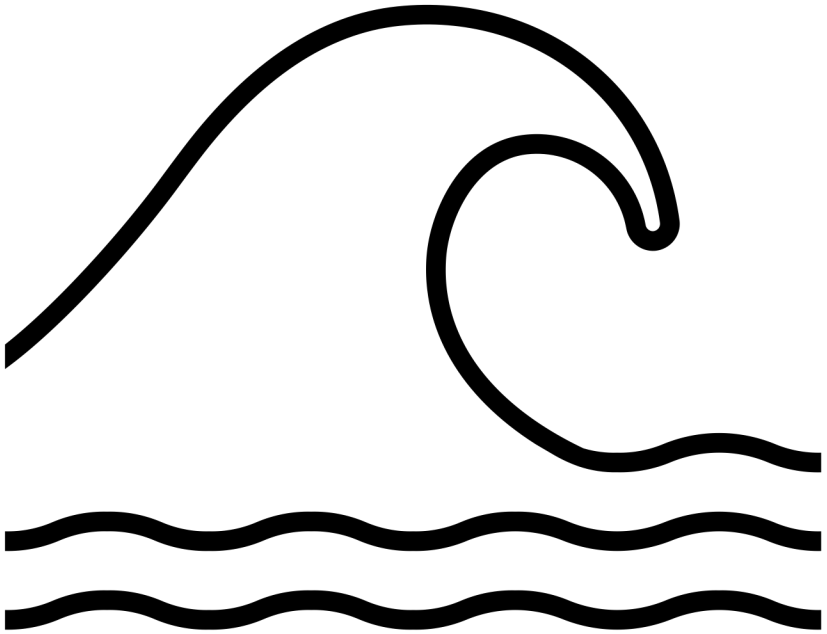


The map reflects the approximate location of the Institutes' lead and principal organizations (staffing and/or activity), as well as their initial funded and unfunded partners.
***Note:** Partners and collaborators related to an Institute may be represented with a single plot due to space limitations.*

AWARDS

- NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography
- NSF AI Institute for Foundations of Machine Learning
- USDA-NIFA AI Institute for Next Generation Food Systems
- USDA-NIFA AI Institute for Future Agricultural Resilience, Management, and Sustainability (AIFARMS)
- NSF AI Institute for Student-AI Teaming
- Molecule Maker Lab Institute (MMLI): NSF AI Institute for Molecular Discovery, Synthetic, and Manufacturing
- NSF AI Institute for Artificial Intelligence and Fundamental Interactions
- NSF AI Institute for Collaborative Assistance and Responsive Interaction for Networked Groups (AI-CARING)
- NSF AI Institute for Learning-enabled Optimization at Scale (TILOS)
- NSF AI Institute for Optimization
- NSF AI Institute for Intelligent Cyberinfrastructure with Computational Learning in the Environment (ICICLE)
- NSF AI Institute for Future Edge Networks and Distributed Intelligence (AI-EDGE)
- NSF AI Institute for Edge Computing Leveraging Next Generation Networks (Athena)
- NSF AI Institute for Dynamic Systems
- NSF AI Institute for Engaged Learning
- NSF AI Institute for Adult Learning and Online Education (ALOE)
- USDA-NIFA AI Institute: Agricultural AI for Transforming Workforce and Decision Support (AgAID)
- USDA-NIFA AI Institute: AI Institute for Resilient Agriculture (AIIRA)

AI and the academic research ecosystem



How do we conceptualize and prepare for what's coming?

AI-readiness in academia

- "…preparedness of organizations to implement change involving applications and technology related to AI." (AlSheibani et. al., 2018)
  - readiness to deploy and manage new technologies,
  - implementation flexibility, and
  - support users in applications of AI

- Organizational conditions (coordination, procurement, cybersecurity)
  - steady state of assessment
  - "organizational chassis" – managing a constant state of change

- Institutional conditions
  - regulatory and fiduciary responsibilities
  - Public Access
  - procurement
  - competition for faculty and students

- Workforce capabilities

AI and the research ecosystem

- Empirical study:
  - Need to scale research in this space
    - Across domains both for biomed domain, and observe and describe the emergent landscape
  - Observe the emergent landscape
    - describe, conceptualize and compare

- How is AI expanding across the academic research enterprise?
  - How will AI research change the structure of research services, or the ways that research services function?
  - If so, what does this mean? If not, should it?
  - How do academic constituencies understand AI and its roles?
    - this will have implications for services, policy implementation, resource allocation, data governance

AI-readiness in Health & Biomed

- Practice Challenges
  - More complex planning and management of data, materials, code, etc.
  - Coming changes in data sharing and publishing requirements
- Data Challenges
  - Multi-source data sets
- Technical / technology challenges
  - Costs to produce and *maintain* high quality, richly labeled data
  - New AI approaches will help with some of this
- Good medicine for everyone
  - Studies designed with broad, diverse representation in mind
    - Resulting data sets and their applications can improve what we can know about health disparities
- Fiduciary and other responsibilities – reshaping the “who owns it” ecosystem
  - Institutional Review Boards

What does all this mean for data sharing *today* and thinking about AI?

Basics to consider when sharing your data

- Describing the data
  - Can you use recognized standards (e.g. MESH terms; .csv format;)
  - FAIR Sharing - https://fairsharing.org/standards/
- Place a copy of the metadata record in a repository
  - Include links to the data set(s)
  - Include re-use parameters and license information
- Acquire Persistent Identifiers for the data (and related materials)
- Where will the data “live?”
  - Near term *and* long term

https://osp.od.nih.gov/scientific-sharing/nih-data-management-and-sharing-activities-related-to-public-access-and-open-science/

OKN Innovation Sprint

- NSF, with White House OSTP, is inviting interested parties from government, industry, academia, non-profits, and other communities to participate in an Innovation Sprint to designing an Open Knowledge Network
  - An opportunity for a couple of representatives from this community
- Envisioning An Inclusive, Open, Public-serving OKN
- Using The Innovation Sprint To Define The Scope
- Timetable:
  - **Sept. – Dec. 2021**: Build enthusiasm and shape th
  - **Feb. 2022:** Virtual Kick-Off Workshop – select sp
  - **Feb. – May 2022:** Concurrent Innovation Sprin based on use cases identified by agencies
  - **June 8 & 9, 2022:** Proto-OKN DESIGN Worksho
  - **Sept. 2022**: Launch an activity to develop the pi

Emerging themes:

- **Use Cases:**
  - Equity
  - Natural Disasters / Climate Change
  - Public Health / Biomedicine
  - Power Grid / Energy
- **Cross-cutting topics:**
  - Provenance
  - Information Quality
  - Using large datasets
  - Real-time data

Lara Campbell
Convergence Accelerator OKN
Prog. Director, NSF
lcampbel@nsf.gov

Tess DeBlanc-Knowles
OSTP Liaison, Tech Policy Advisor, NSF
tdeblanc@nsf.gov
Tess.K.DeBlanc-Knowles2@ostp.eop.gov

Wo Chang
Digital Data Advisor, NIST
wchang@nist.gov

References and Resources

Alsheibani, S.; Cheung, Y.; and Messom, C. (2018), Artificial Intelligence Adoption: AI-readiness at Firm-Level. PACIS 2018 Proceedings. 37. https://aisel.aisnet.org/pacis2018/37/

Demchenko, Y., Grosso, P., De Laat, C. and Membrey, P. (2013). Addressing big data issues in scientific data infrastructure. In, 2013 International Conference on Collaboration Technologies and Systems (CTS) (pp. 48-55). IEEE.

Jöhnk, J., Weißert, M. & Wyrtki, K. (2021). Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors. Business & Information Systems Engineering, **63,** 5–20. https://doi.org/10.1007/s12599-020-00676-7

National Academies of Sciences, Engineering, and Medicine. 2020. Planning for Long-Term Use of Biomedical Data: Proceedings of a Workshop. Washington, DC: The National Academies Press. https://doi.org/10.17226/25707.

Open Science by Design - https://www.nap.edu/catalog/25116/open-science-by-design-realizing-a-vision-for-21st-century

Rajpurkar, P., Chen, E., Banerjee, O. et al. (2022). AI in health and medicine. Nat Med. https://doi.org/10.1038/s41591-021-01614-0

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

SDSC

Thank you

mcragin@sdsc.edu

This work is supported in part by the National Science Foundation grants: # 1916481 (West BDI Hub); 1937953 (Advancing FAIR); 2032705 (EAGER)