National Institutes of Health
Office of Data Science Strategy

AI Supplements Closeout Meeting

NIH Office of Data Science Strategy

Nov 1 , 2021

Showcase your work and share your experiences with other researchers and the NIH

	Oct 24th	Oct 31st	Nov 1st
AI-Workforce: (NOT-OD-21-079) *Workforce Development at the Interface of Information Sciences, AI/ML, and Biomedical Sciences*	☑		☑
AI-Readiness: (NOT-OD-21-094) *Collaborations to Improve the AI/ML-Readiness of NIH-Supported Data*		☑	☑

https://www.scgcorp.com/odssaicloseout2022/

Contacts and Information

- Event website has up to date agenda
  https://www.scgcorp.com/odssaicloseout2022/

- If you have questions about the **program**, please email:
  Mike Spittel Michael.Spittel@nih.gov

- If you have any **technical or logistics issues**, please email:
  Mark Dennis mdennis@scgcorp.com
  Danielle Johnikin djohnikin@scgcorp.com

- Please be sure to visit the ODSS website to read the Tagline and Abstracts submitted by your colleagues.
  https://datascience.nih.gov/artificial-intelligence/initiatives/Improving-AI-readiness-of-Existing-Data

- Plenary Sessions will be recorded

Request

Please use the chat freely!!

NIH Observers – please add “NIH” prefix to your name in Zoom

- Right click on your video box, or click “…”
- Choose “Rename”
- “NIH – Laura Biven”

Abbreviated Agenda

11/1/2022

https://www.scgcorp.com/odssaicloseout2022/

11:00 a.m. – 11:10 a.m. **Welcome and Introductions**

11:10 a.m. – 12:00 p.m.
Recognizing and Integrating Social Good into the AI Development Lifecycle
Bradley Malin, Accenture Professor of Biomedical Informatics,
Biostatistics, and Computer Science; Vice Chair for Research Affairs
Department of Biomedical Informatics, Vanderbilt University

12:05 p.m. – 1:00 p.m.
Data Science at NIH
Susan Gregurick, Associate Director for Data Science and Director of the Office of Data Science Strategy, NIH

1:10 p.m. – 2:10 p.m. Breakout Discussions
2:15 p.m. – 3:05 p.m. Readouts

3:05 p.m. – 3:30 p.m. **Open Conversation and Closing**

Abbreviated Agenda

11/1/2022

https://www.scgcorp.com/odssaicloseout2022/

11:00 a.m. – 11:10 a.m. **Welcome and Introductions**

11:10 a.m.
Recognizin
Bradley M
Biostatistic
Departmer

12:05 p.m.
Data Scien
Susan Greg ata Science Strategy,
NIH

1:10 p.m. –
2:15 p.m. –

3:05 p.m. – 3:30 p.m. **Open Conversation and Closing**

** ** **

Use the breaks to stretch, get snacks…

** ** **

Recognizing and Integrating Social Good into the AI Development Lifecycle

Bradley Malin
Accenture Professor of Biomedical Informatics, Biostatistics, and Computer Science;
Vice Chair for Research Affairs
Department of Biomedical Informatics, Vanderbilt University

BREAK

12:00 p.m. – 12:05 p.m.

Data Science at NIH

Susan Gregurick
Associate Director for Data Science;
Director of the Office of Data Science Strategy, NIH

National Institutes of Health
Office of Data Science Strategy

AI/Data Science at the NIH: Opportunities to Improve Biomedical Research and Human Health

Susan K. Gregurick, Ph.D.
Associate Director for Data Science
National Institutes of Health

November 1, 2022

Topics for Today

- Background
- Impactful research initiatives built on AI
- Supporting research and infrastructure to address biases in AI
- Creating AI-ready data to address SARS-COV-2
- Where to next?

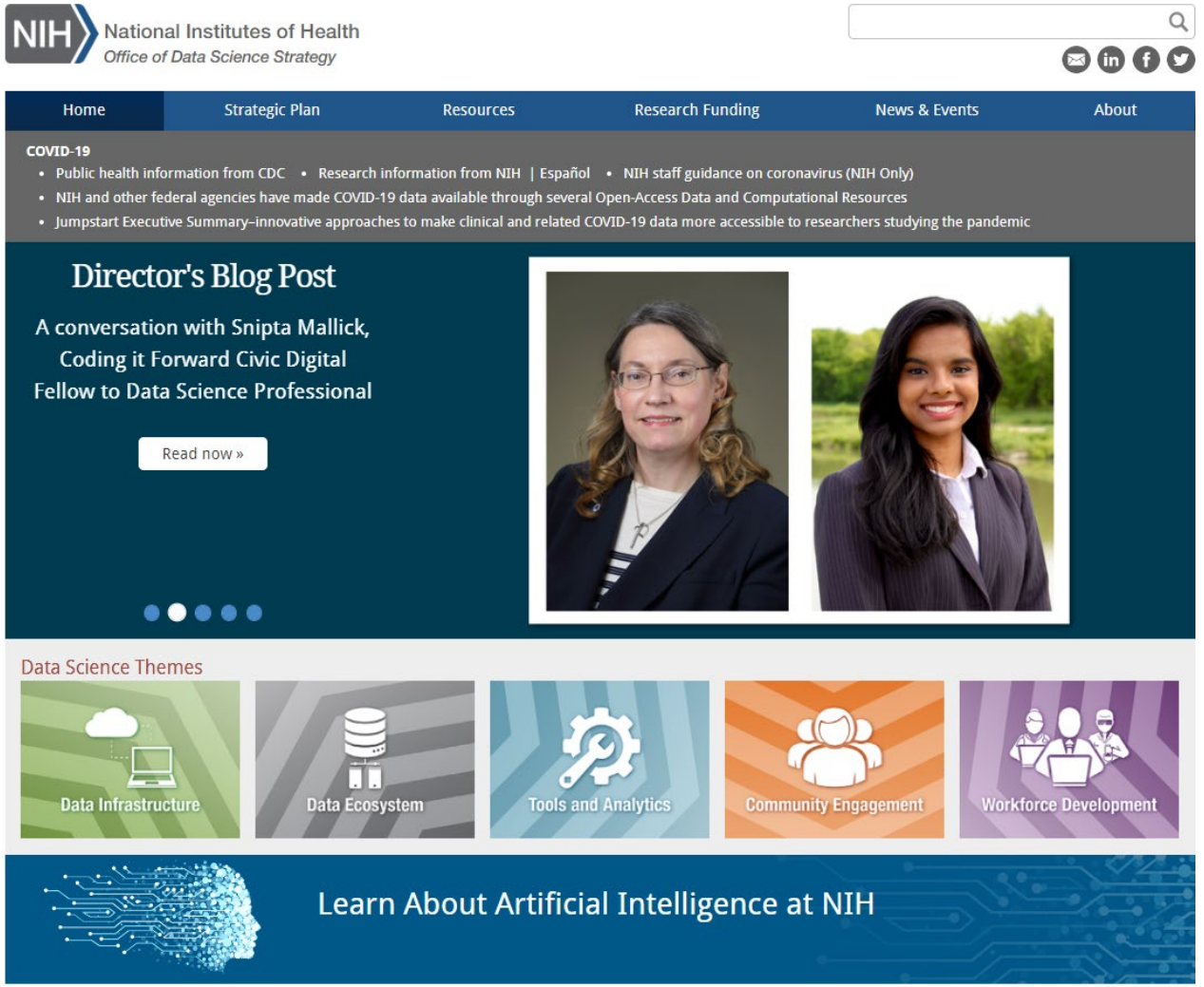
Our vision is built on the Strategic Plan for Data Science

- Support common infrastructure and architecture on which more specialized platforms can be built and interconnected.
- Leverage commercial tools, technologies, services, and expertise; and adopt and adapt tools and technologies from other fields for use in biomedical research.
- Enhance the nation's biomedical data-science research workforce through improved training programs and novel partnerships.
- Enhance data sharing, access, and interoperability such that NIH-supported data resources are FAIR.
- Ensure the security and confidentiality of patient and participant data in accordance with NIH requirements and applicable law.
- Improve the ability to capture, curate, validate, store, and analyze clinical data for biomedical research.
- With community input, develop, promote—and refine as needed—data standards, including standardized data vocabularies and ontologies, applicable to a broad range of fields.

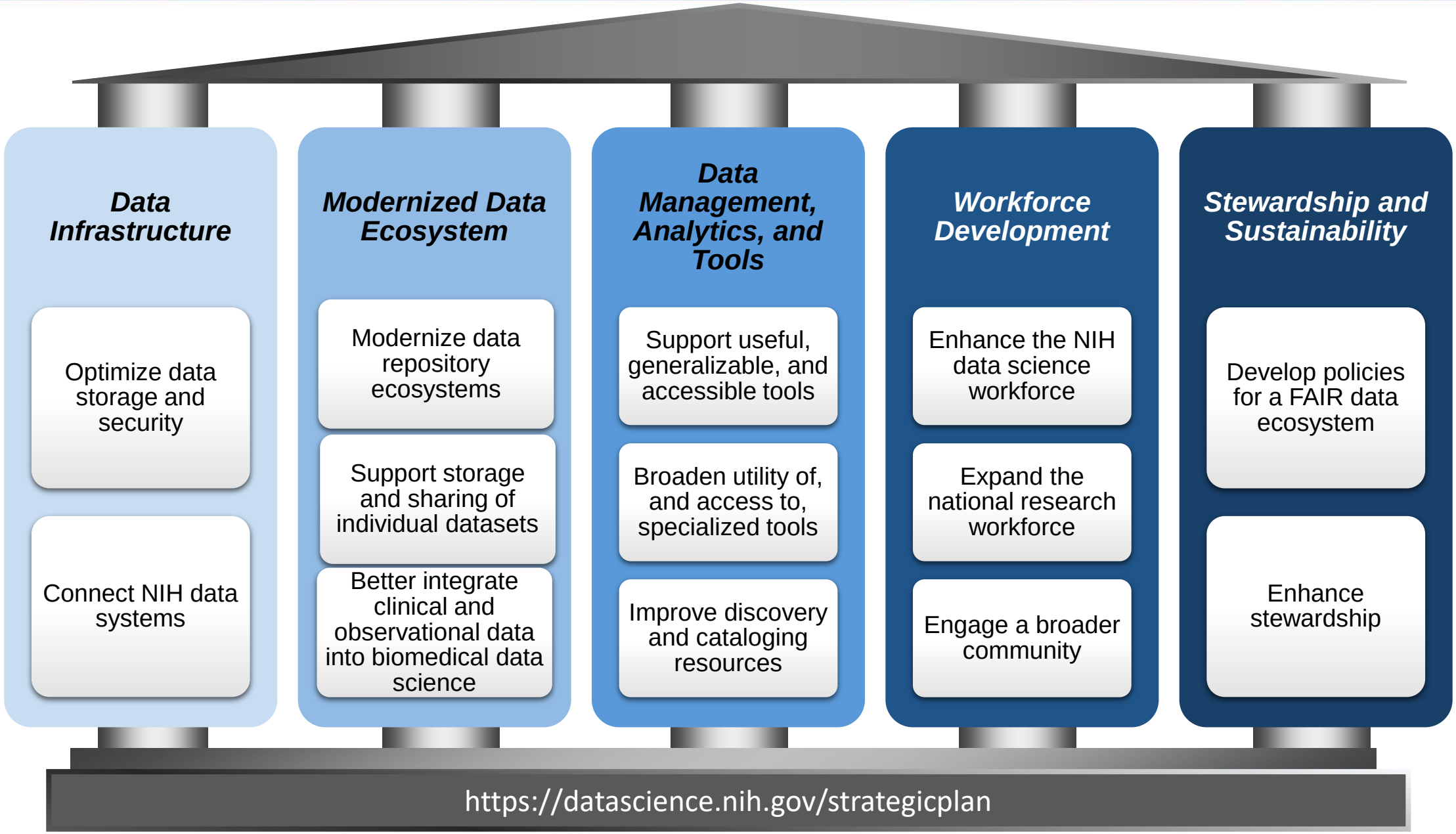
Office of Data Science Strategy

The NIH **Office of Data Science Strategy (ODSS)**, in the Office of the Director

- Provides **leadership and coordination** on the strategic plan for data science
- Develops and implement NIH's vision for a **modernized** and **integrated** biomedical data ecosystem
- Enhances a **diverse and talented** data science workforce
- **Builds strategic partnerships** to develop and disseminate advanced technologies and methods

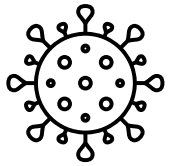


NIH Strategic Plan for Data Science – Goals & Objectives



Accomplishments

200 PB data on 3 Clouds



Identify and share coronavirus sequences collected by the global research community

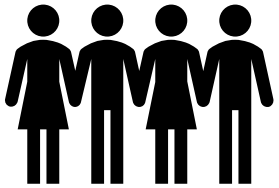
9+ NIH IC data platforms allow for single sign-on of researchers

Researchers can data from the GTEx AnVIL platform & the Cancer Data Commons for combined analysis of LINE-1 expression

Programs to support databases and knowledgebases, includes DataCite

Aligning data resources to support NIH Data Management and Sharing Policy

Programs to enhance data workforce

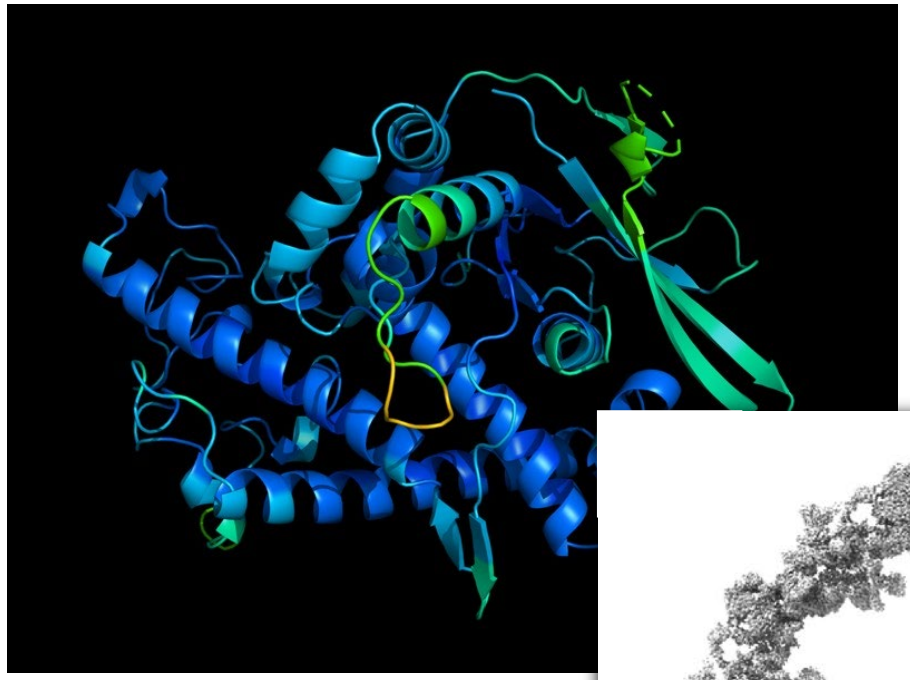


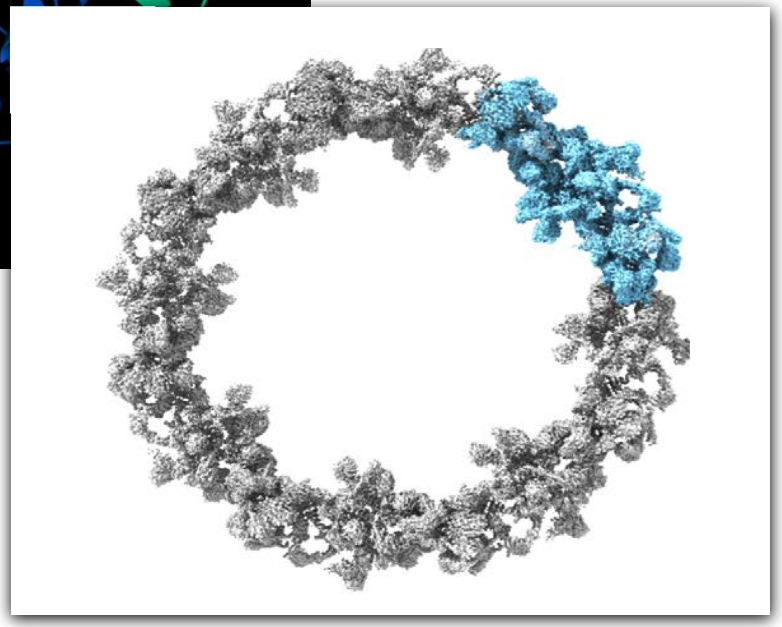
NIH 1–2-year sabbaticals for fellows to work at NIH on challenging data science problems

Promise of AI for Biology

- 2020 - DeepMind's AI makes gigantic leap in solving protein structures

- 2022 - Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold





Nature **588**, 203-204 (2020) *doi:* https://doi.org/10.1038/d41586-020-03348-4
Fontana, et al., *Science* **376,** 1178 (2022) https://doi.org/10.1126/science.abm9326

The Promise of AI for Medicine

Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation

MICHELLE M. CLARK, AMBER HILDRETH, SERGEY BATALOV, YAN DING, SHIMUL CHOWDHURY, KELLY WATKINS, KATARZYNA ELLSWORTH, BRANDON CAMP, CYRIELLE I. KINT, [...] STEPHEN F. KINGSMORE +52 authors Authors Info & Affiliations

SCIENCE TRANSLATIONAL MEDICINE · 24 Apr 2019 · Vol 11, Issue 489 · DOI: 10.1126/scitranslmed.aat6177

AI reduces time and effort for diagnosis of rare genetic disorders in infants in the ICU and can analyze 4.5M variants associated with 13,000 genetic disorders in 5 minutes.

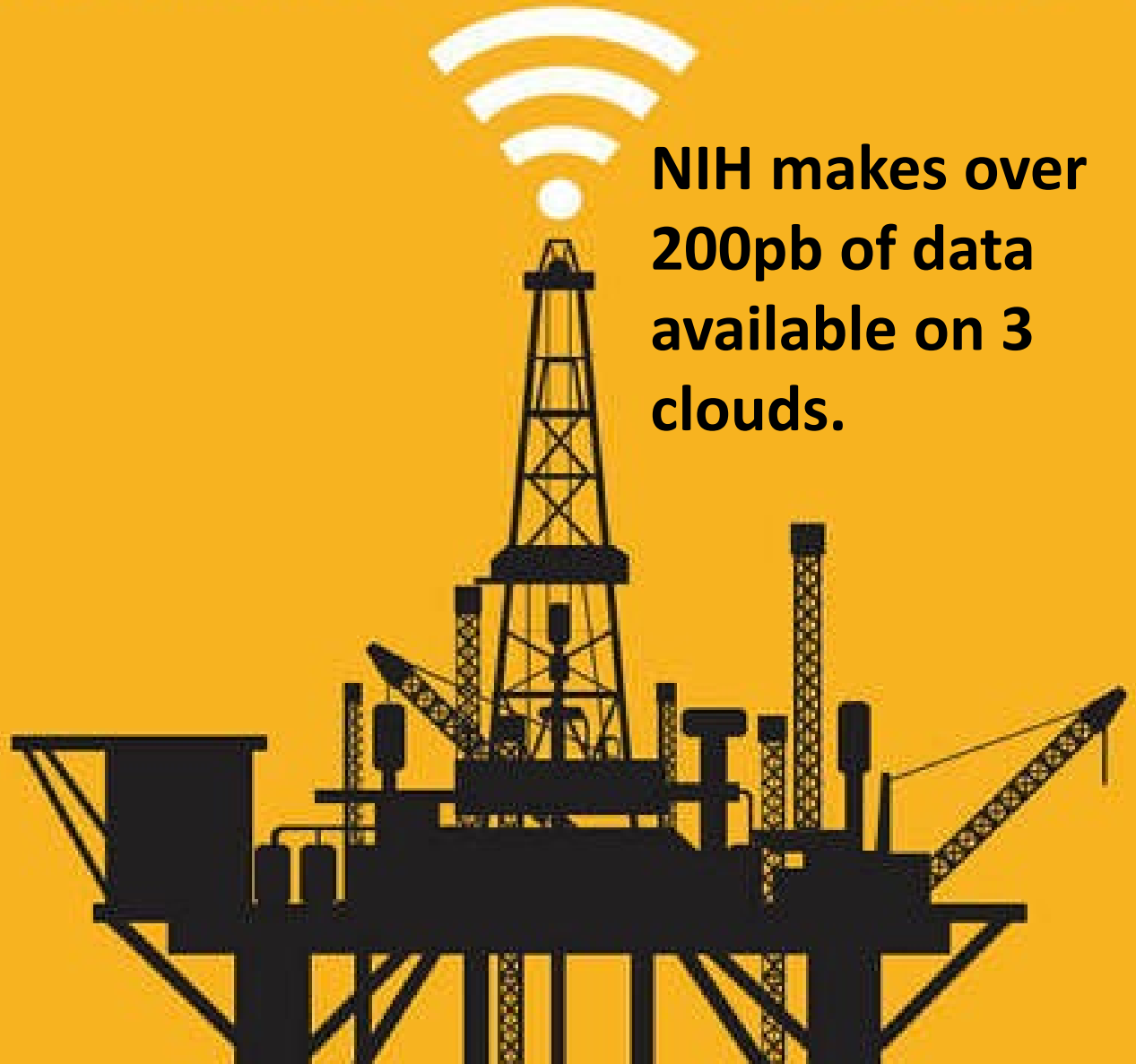
NIH Director's Blog

Whole-Genome Sequencing Plus AI Yields Same-Day Genetic Diagnoses

Dr. Francis Collins

Data is the new oil!

- Genetic Expression and Variation Analysis
- Microbiome Analysis
- Cellular Structure and Functional Analysis
- Neuroscience Analysis
- Genomic and Phenotypic Analysis
- Neuronal Image Analysis



NIH makes over 200pb of data available on 3 clouds.

- Metabolomics Analysis
- Whole Genome Sequence Analysis
- Single-Cell ‘Omics Analysis
- Microscopy Image Analysis
- Cryo-Electron Microscopy Analysis
- Clinical Analytics, new applications of FHIR

Collaborations to Make Data FAIR and AI/ML Ready

NIH supported collaboration, bringing together expertise in biomedicine, data management, and artificial intelligence and machine learning (AI/ML) to make NIH-supported data AI-ready for AI/ML analytics.

FY21-FY22: 73 Awards

Most common biomedical focus areas:

Alzheimer's and Parkinson's disease, cardiovascular disease, cancer, and aging

Most common data types:

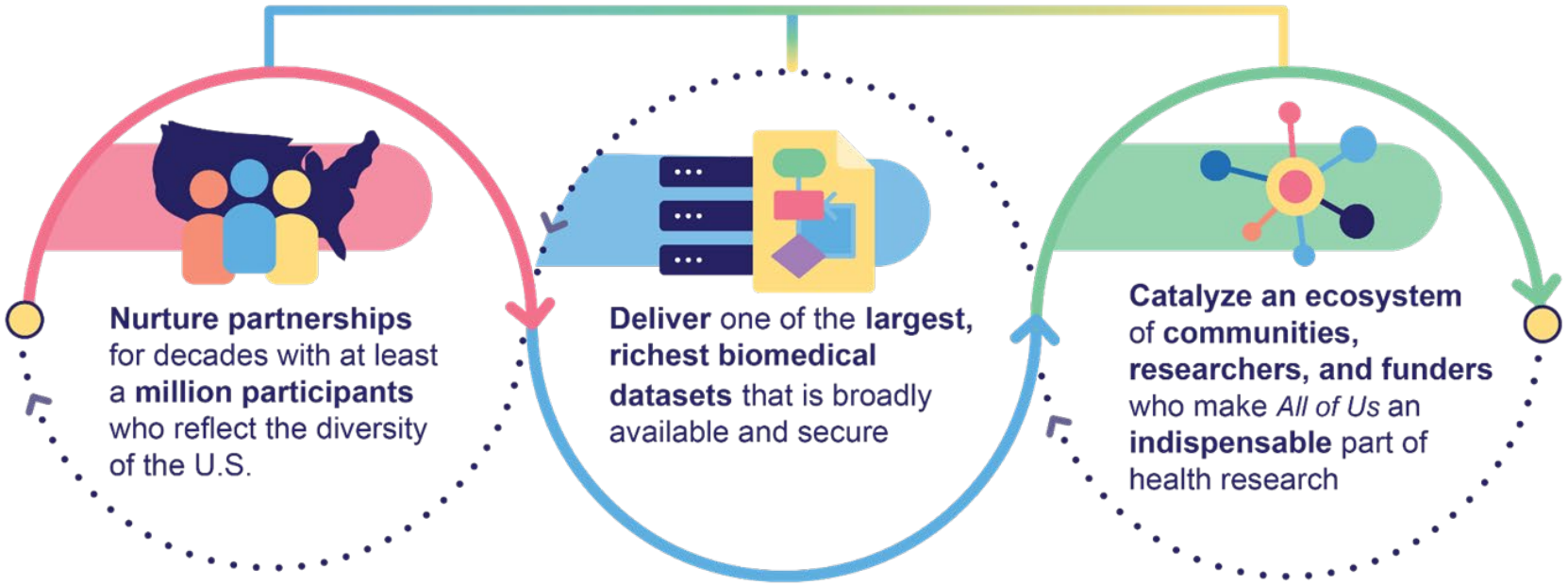
imaging, EHRs, -omics, microbes/pathogens, speech

NHGRI | NIA | NIBIB | NIDA | NIDCD | NIDCR | NIEHS | NIGMS | NIMH | NINDS | NCI | NLM | NIMHD | NIDDK | NICHD | NIAID | NIAMS | NHLBI

All of Us Research Program

Our Mission

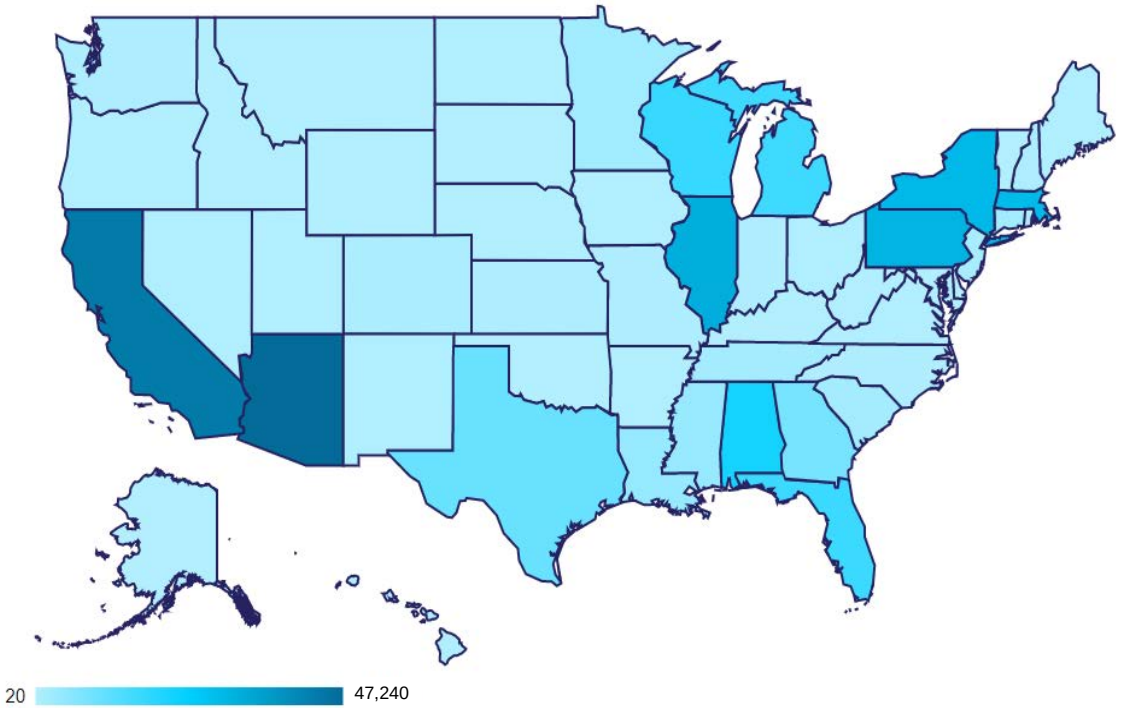
Accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all of us



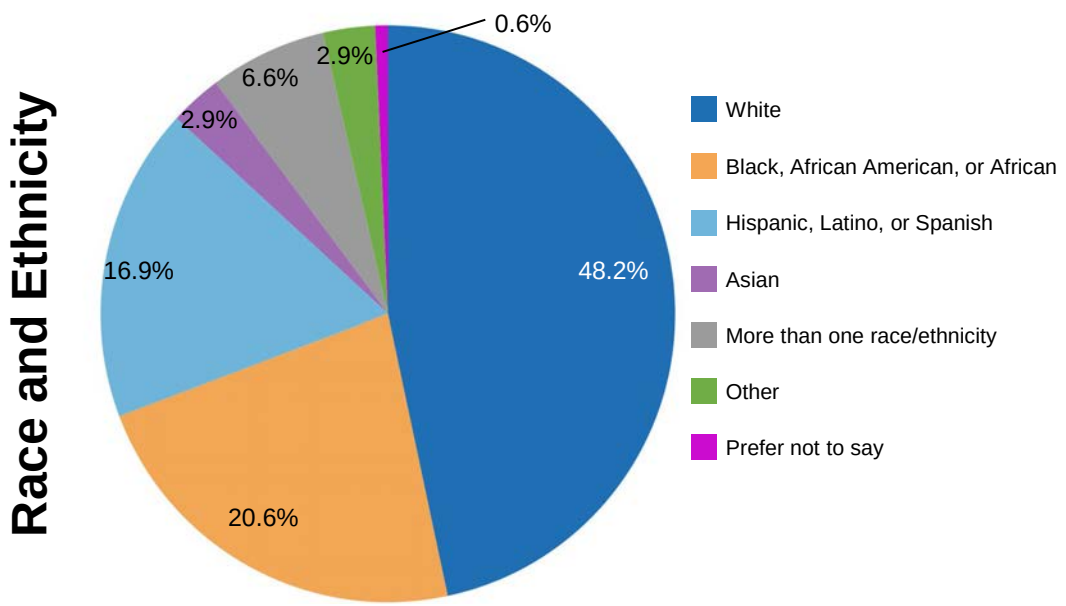
Made possible by a team that maintains a culture built around the program's core values

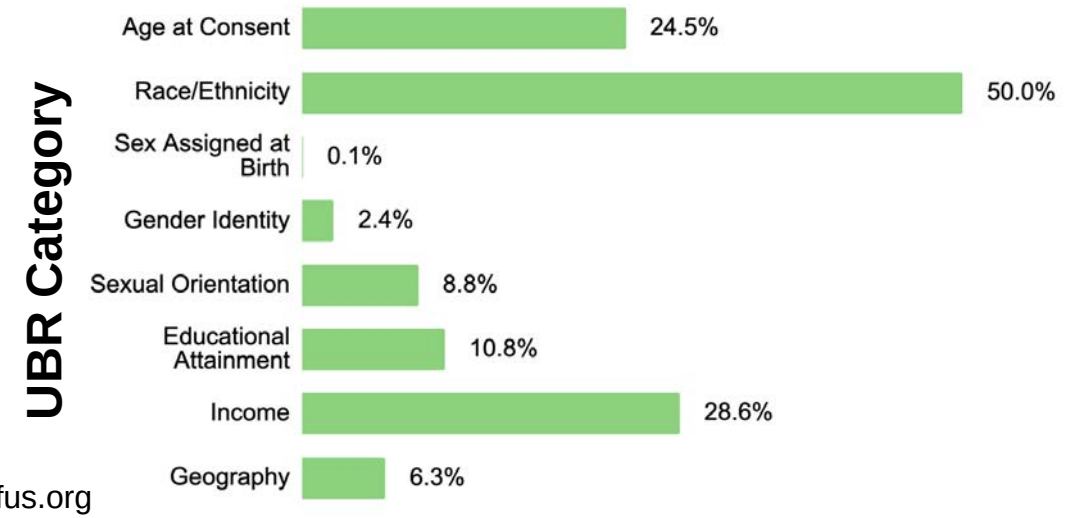
https://allofus.nih.gov/

Status of the *All of Us* Research Program



Over 80% of *All of Us* participants are underrepresented in biomedical research





researchallofus.org

Status of the *All of Us* Research Program

Including the first batch of genomic data in the Controlled Tier

Genomics

Only available via the Controlled Tier

98,600+
Whole Genome Sequences

165,000+
Genotyping Arrays

593,500,000+
Unique Variants
Represented and searchable through public genomic data browser

Genomics Analysis Tools
Hail and PLINK
in addition to R, Python, and Jupyter Notebooks

100M of these occur >2 times in AoU and are not in gnomAD

Genomic Data is Paired with Rich Phenotypic Data

77,000+
Have Whole Genome Sequences + Electronic Health Records + Physical Measurements + Survey Responses

95,000+
Have Whole Genome Sequences + Physical Measurements + Survey Responses

78,200+
Have Whole Genome Sequences + Electronic Health Records

3,500+
Have Whole Genome Sequences + Fitbit Records
Representing >30% of all participant Fitbit records

Data on Researcher Workbench

Diverse and Longitudinal

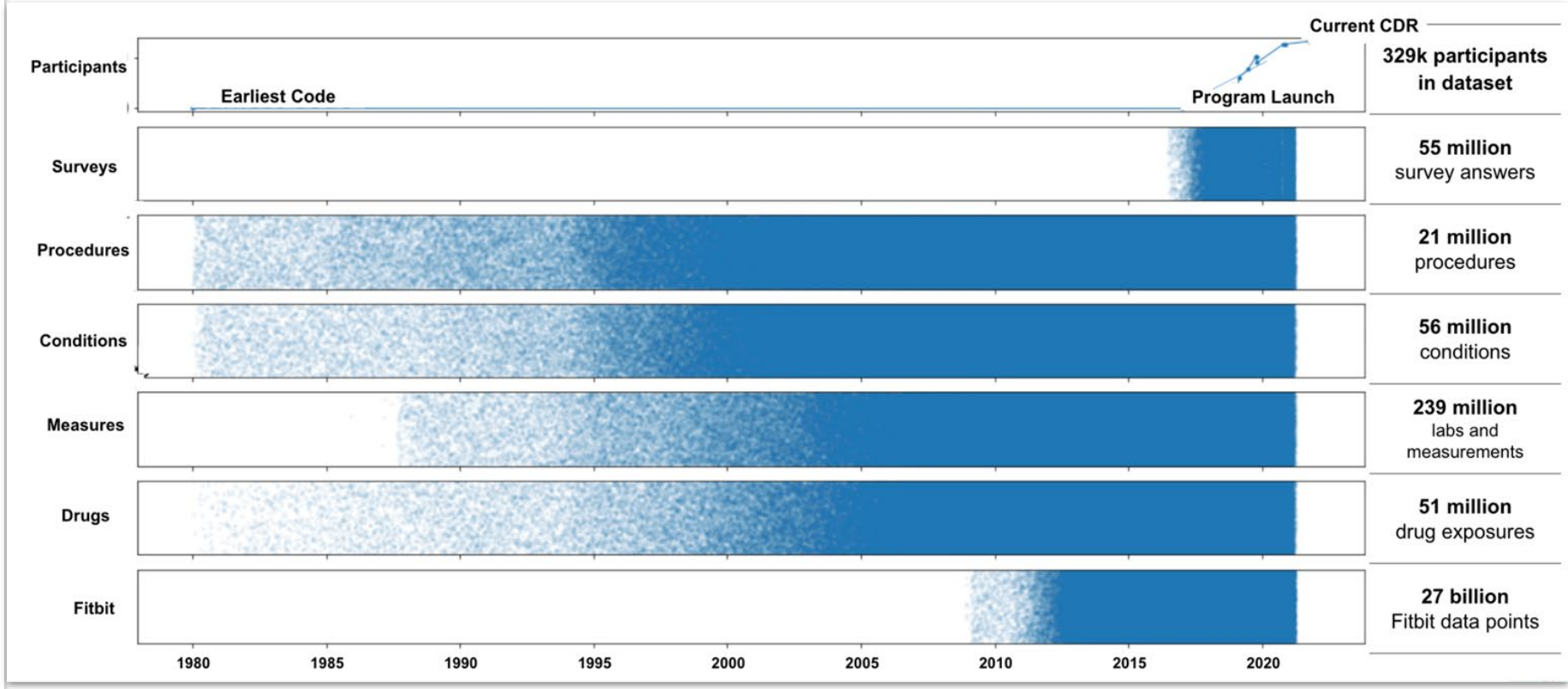
329,000+ Participants

267,600+ Physical Measurements

214,200+ EHRs

329,000+ Surveys

11,600+ Fitbit Records



Nutrition for Precision Health

A major challenge in nutrition is the inability to combine factors that affect how individuals respond to diet into a personalized nutrition regimen including the community of microbes that live in our gut, metabolism, nutritional status, genetics, and the environment.

To address these gaps, this effort will collect new data on multiple potential predictive factors and combine it with existing data in the ***All of Us*** database to develop a more complete picture of how individuals respond to different foods or dietary routines.

These data linkages will be one of the largest, most diverse precision nutrition studies to date.

Bridge2AI Program Goals

- Use biomedical and behavioral research grand challenges to generate **flagship data sets**
- **Prepare** AI/ML-friendly data
- Emphasize **ethical** best practices
- Promote **diverse teams**

BRIDGE2AI

Standardize Data Attributes

Develop Automated Tools

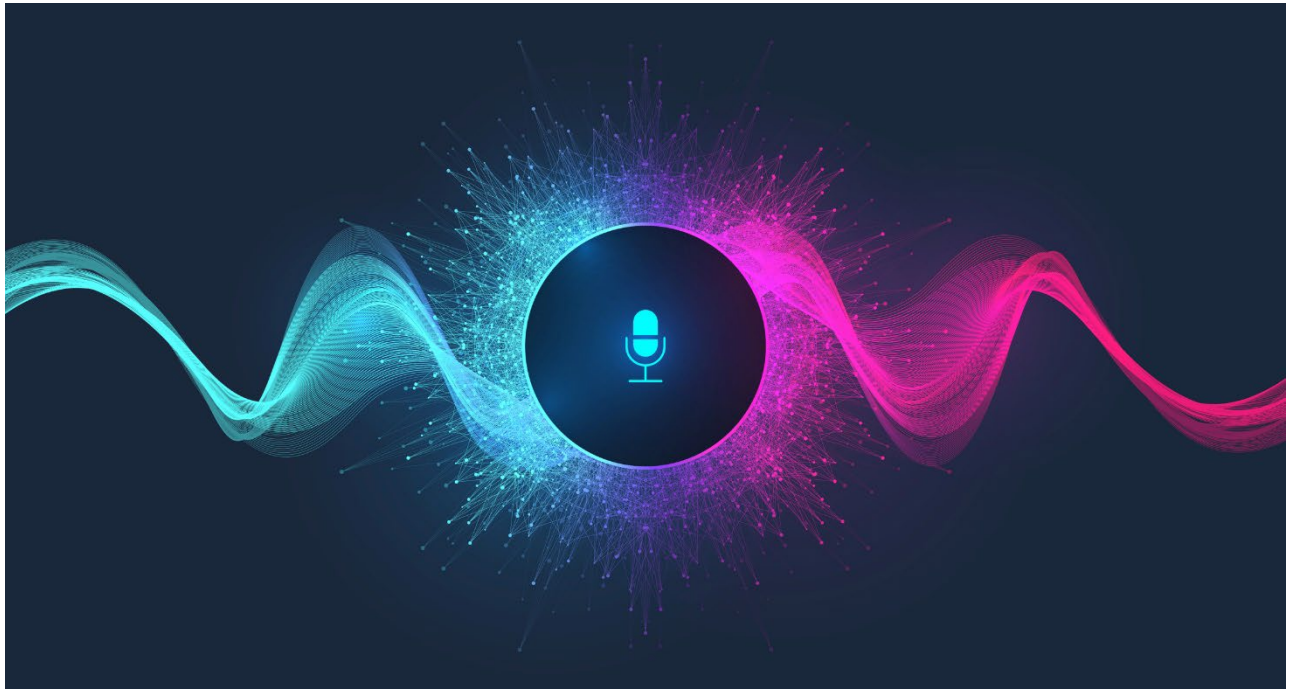
New Datasets

Create cross-training materials for Workforce Development

Disseminate Products & Best Practices

Precision Public Health

- Use voice as a biomarker for respiratory disorders, sleep apnea, mental health, and neurological disorders
- Create a database of bioacoustic waveforms
- Ethics: voice-hacking
- Tools: waveform compression, minimum acoustic quality

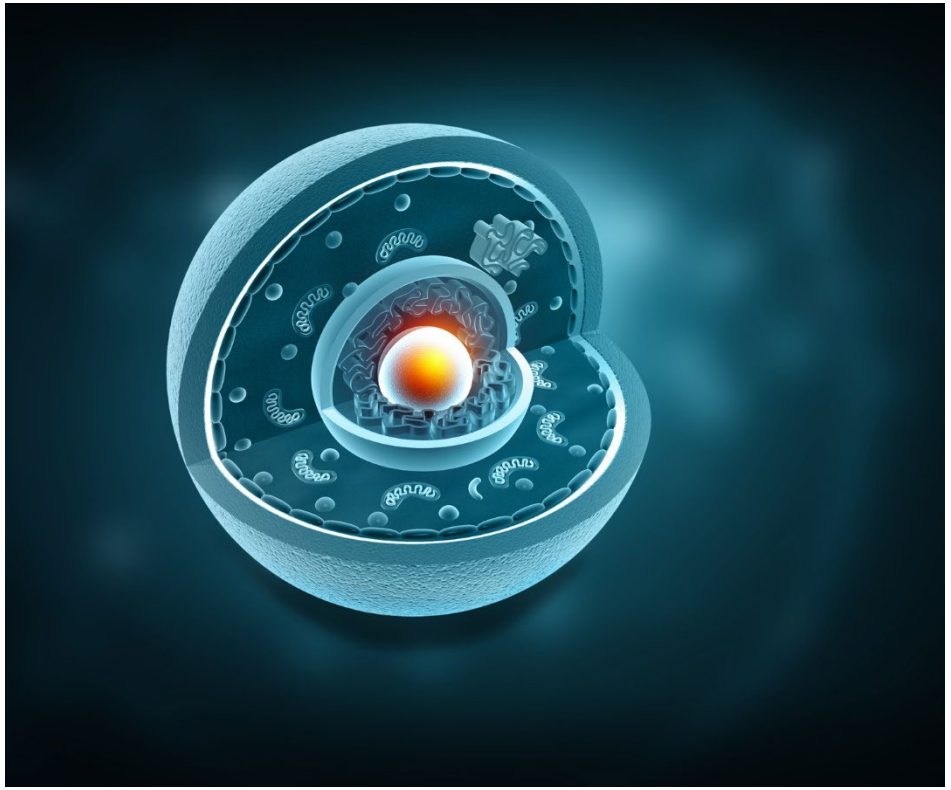


Functional Genomics

- Map the spatiotemporal architecture of human cells and use these maps toward the grand challenge of interpretable genotype-phenotype learning.

- 3 complementary mapping approaches:
  - proteomic mass spectrometry,
  - cellular imaging,
  - genetic perturbation via CRISPR/Cas9 –

- Create a library of large-scale maps of cellular structure/function and disease contexts using cell lines



Salutogenesis

- Reconstruct a temporal atlas of pathogenesis and salutogenesis to expand AI/ML applications in clinical care
- Utilize type 2 diabetes as a model for understanding disease progression
- Native American partnership

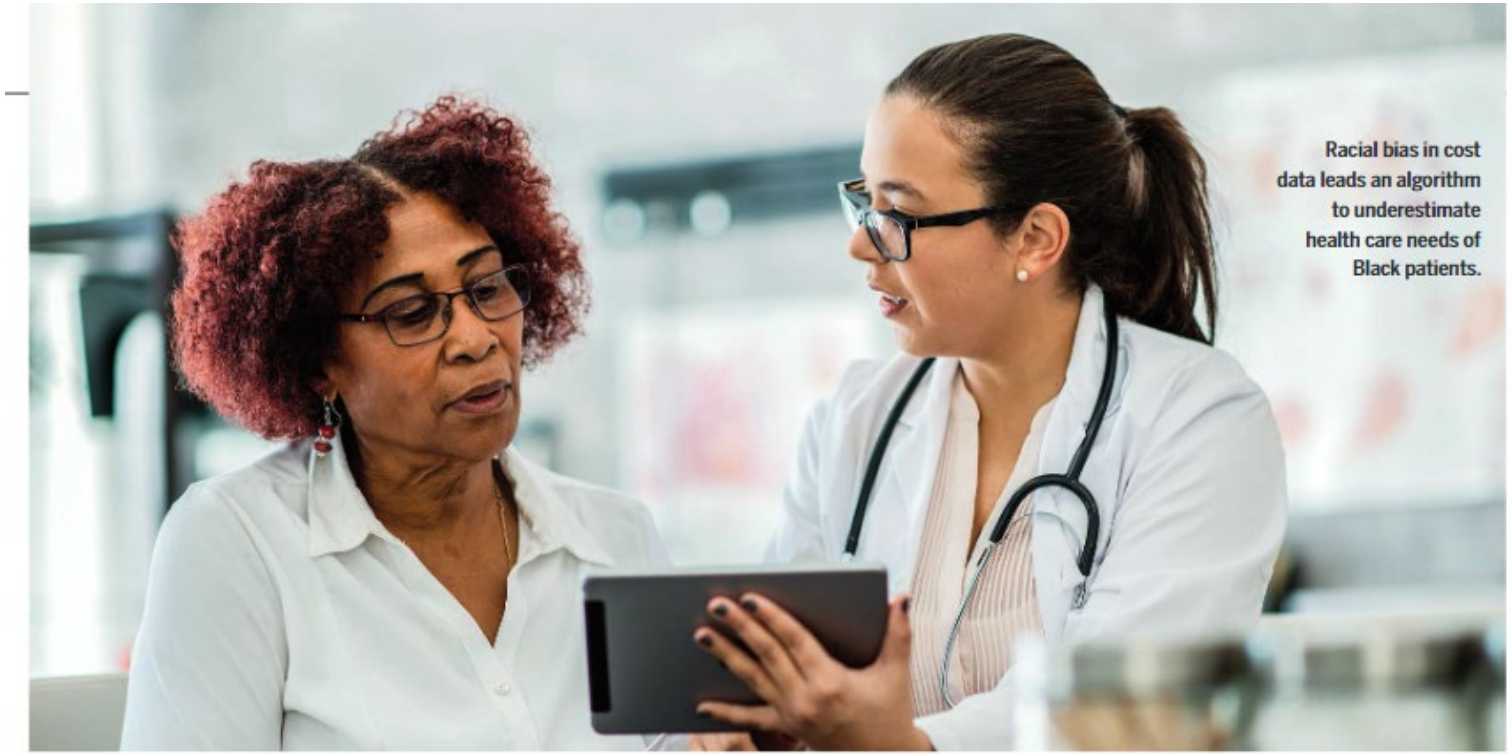
Expanding AI/ML in Clinical Care

- Create a dataset of >100,000 patients from 14 ICU sites to improve recovery from acute illnesses through AI
- Phenotyping from clinical notes via natural language processing
- Develop a model to predict adverse events from ICU-monitoring data

What will Bridge2AI Produce?

- Novel, complete, trustworthy datasets - ethically sourced, following FAIR principles, motivated by biomedical and behavioral grand challenges
- Tools to accelerate the creation of data sets for AI/ML analysis (intelligent annotators, metadata-filling instruments)
- Community evaluation of datasets -- **culture change** to embrace data preparation -- for AI/ML analysis
- Interdisciplinary AI/ML-Biomedical and Behavioral Research Community

Racial Bias



Racial bias in cost data leads an algorithm to underestimate health care needs of Black patients.

SOCIAL SCIENCE

Assessing risk, automating racism

A health care algorithm reflects underlying racial bias in society

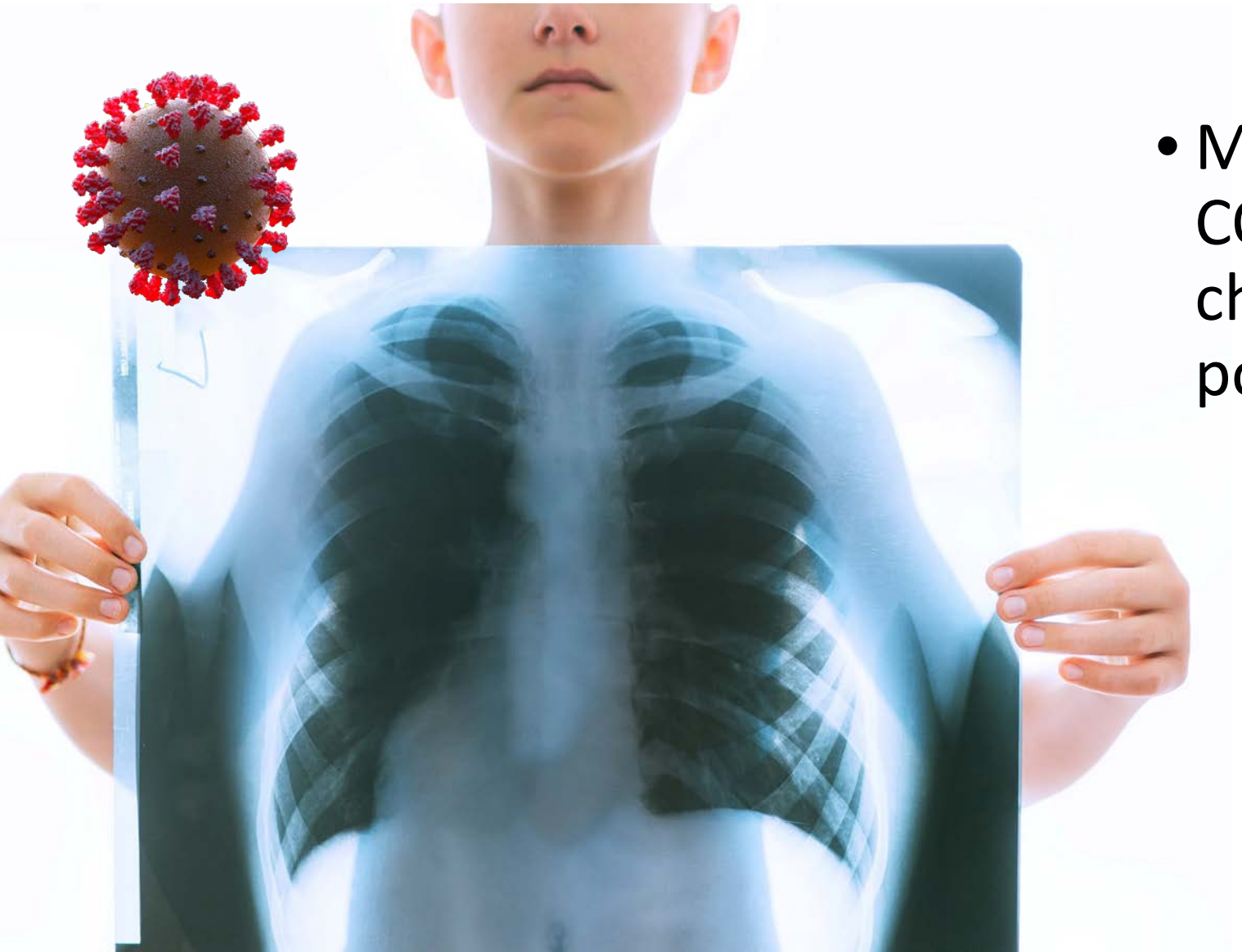
By **Ruha Benjamin**

era, the intention to deepen racial inequities was more explicit, today coded ineq-

beyond the algorithm developers by constructing a more fine-grained measure of

Credit: https://www.science.org/doi/10.1126/science.aaz3873

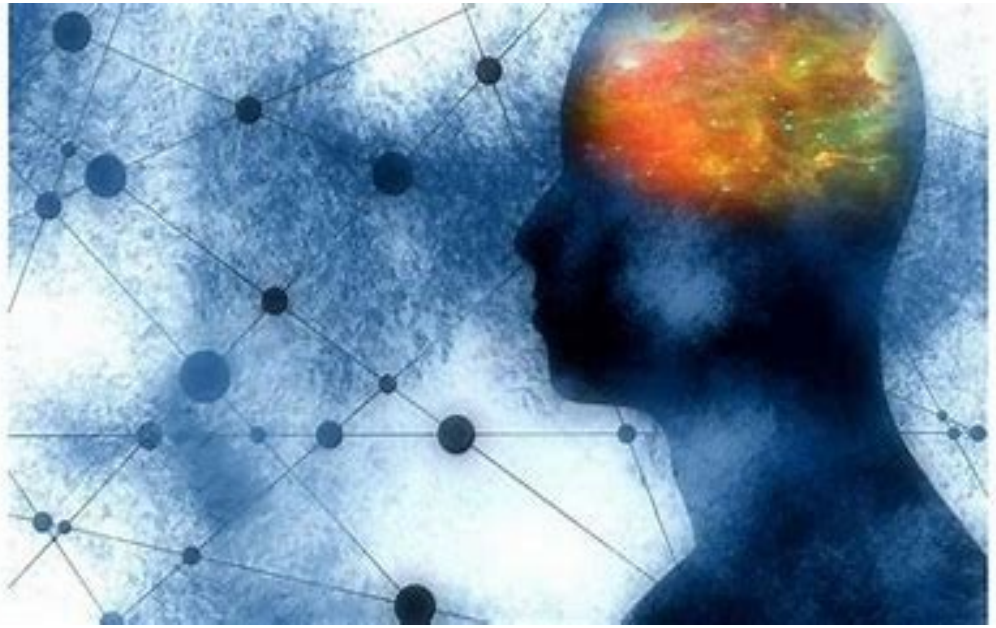
Age Disparities



- Model trained to detect COVID-19 using adult chest X-rays gave false positives in children

Bias in the Application of ML for Mental Health Diagnosis/Treatment

- Researchers at IBM used machine learning to **predict** diagnosis or treatment for **postpartum depression**
- Diagnosis and use of mental health services did not align with **known rates of PPD** incidence



Embi, 2021

Challenges for Widespread Use of AI/ML

- Biomedical studies and the datasets they yield may **lack diverse representation**, leading to an **inadequate understanding** of health disparities and inequities
- If these datasets are used to train the algorithms that make up AI/ML approaches, the results obtained will be flawed
- Many under-represented communities have the potential to **contribute data from diverse participants, and perform cutting-edge studies** but may **lack financial, infrastructural, and training support**
- Integration of Electronic Health Record (EHR) data with other **data** types (e.g., Social Determinants of Health (SDOH), genomic, and imaging) are **needed to understand health disparities and inequities**

Actioning BioMed-ML (AI) to address health disparities

Collaborations to Advance Ethical Use of AI/ML

Advancing the Ethical Development and Use of AI/ML in Biomedical and Behavioral Sciences

NIH will support collaborations that bring together expertise in ethics, biomedicine, data collection, and AI/ML

These collaborations are intended to generate **new understanding, practices, tools, techniques, metrics, or resources that will aid *others*** in making ethical decisions throughout the development and use of AI/ML, including the collection and generation of data as well as the reuse of data and models by others.

AIM-AHEAD

Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity

Partnerships

Research

Infrastructure

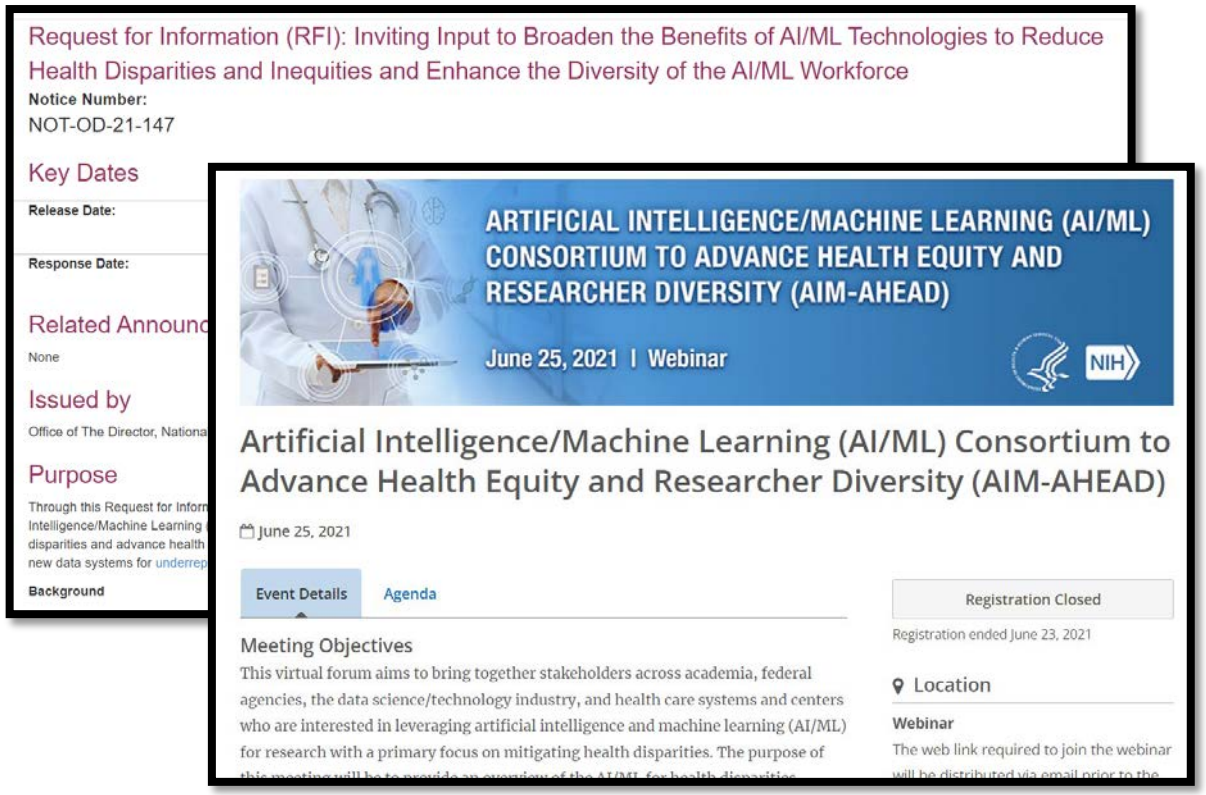
Training

Goals:

- enhance the **participation** and **representation** of researchers and communities currently underrepresented in the development of artificial intelligence and machine learning (AI/ML) models
- address health disparities and inequities using AI/ML
- improve the capabilities of this emerging technology, beginning with the use of electronic health record (EHR) and extending to other diverse data

https://aim-ahead.net/
https://datascience.nih.gov/artificial-intelligence/aim-ahead

Community Input Shaped the Initial Phase



AIM-AHEAD will:

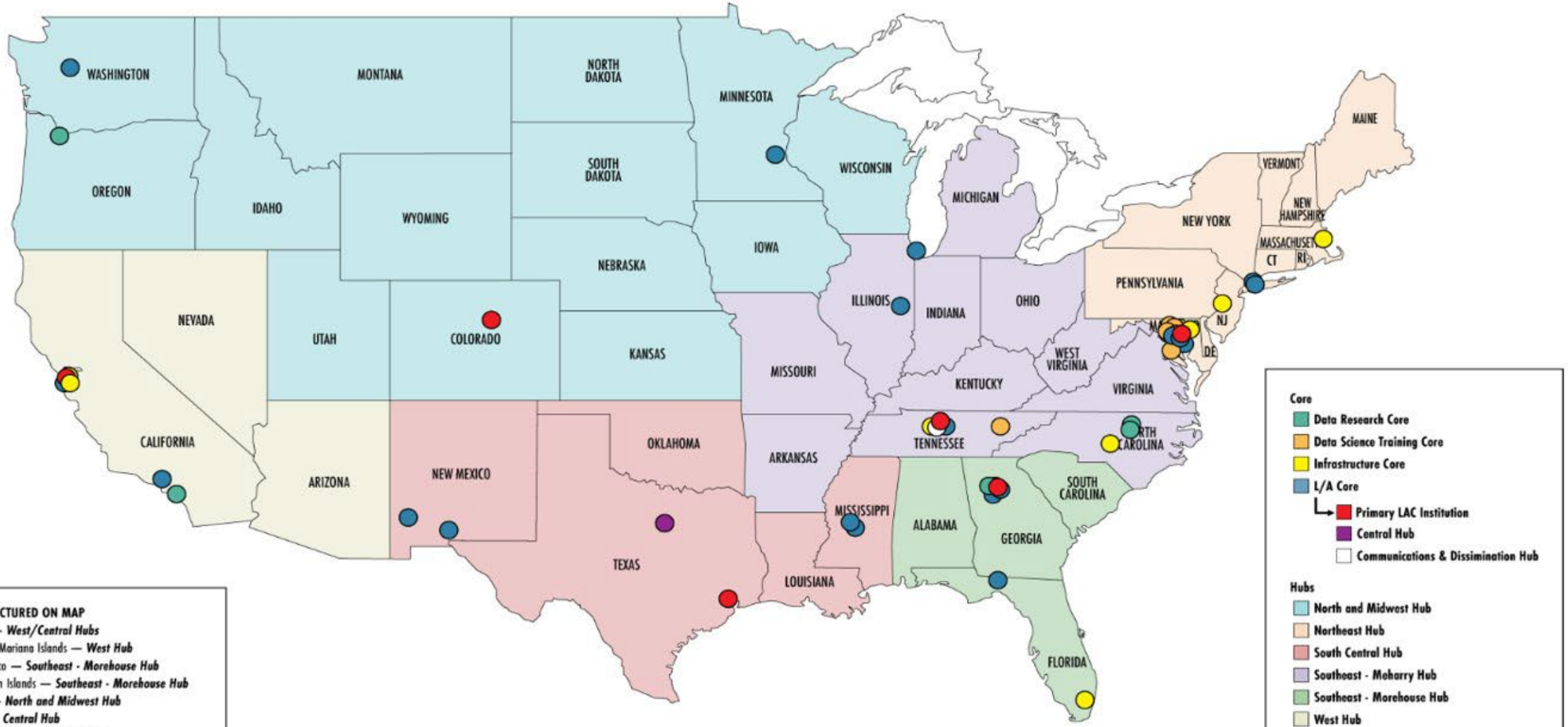
- **Develop** a consortium of organizations and institutions that
  - wish to develop capabilities in AI/ML
  - wish to build a more inclusive basis for AI/ML
  - have a core mission to serve populations experiencing health disparities.
- Begin with a two-year planning, assessment, and capacity building phase
- Establish a coordinating center with the essential expertise in AI/ML and health disparities research, data science training, and data and computing infrastructure

There are a wide variety of interests, needs, and resources across communities.

AIM-AHEAD Cores

CORE	PRIMARY ROLE
Leadership/Admin Core	Leadership of the AIM-AHEAD Coordinating Center (A-CC); recruitment and coordination of consortium members; funding and support for activities of the A-CC; stakeholder engagement and outreach to enhance the diversity; and establishing relationships with populations affected by health disparities
Data Science Training Core	Identify training needs and gaps to address AI/ML and health disparities research, training in AI / ML and health disparities research as well as related competencies; and collaboration with the other Cores
Data and Research Core	Identify AI/ML and health disparities research use cases to drive the design of the data and computing infrastructure; execute pilot AI/ML projects with consortium members; prepare and support the use of existing or purpose-built data resources; and collaboration with the other Cores
Infrastructure Core	Engage with the largest; most diverse array of MSIs; design, develop, and implement appropriate data infrastructure; collaboration with the other Cores, and development of pilot data and analysis environments to advance Coordinating Center aims.

AIM-AHEAD PARTNERSHIP MAP

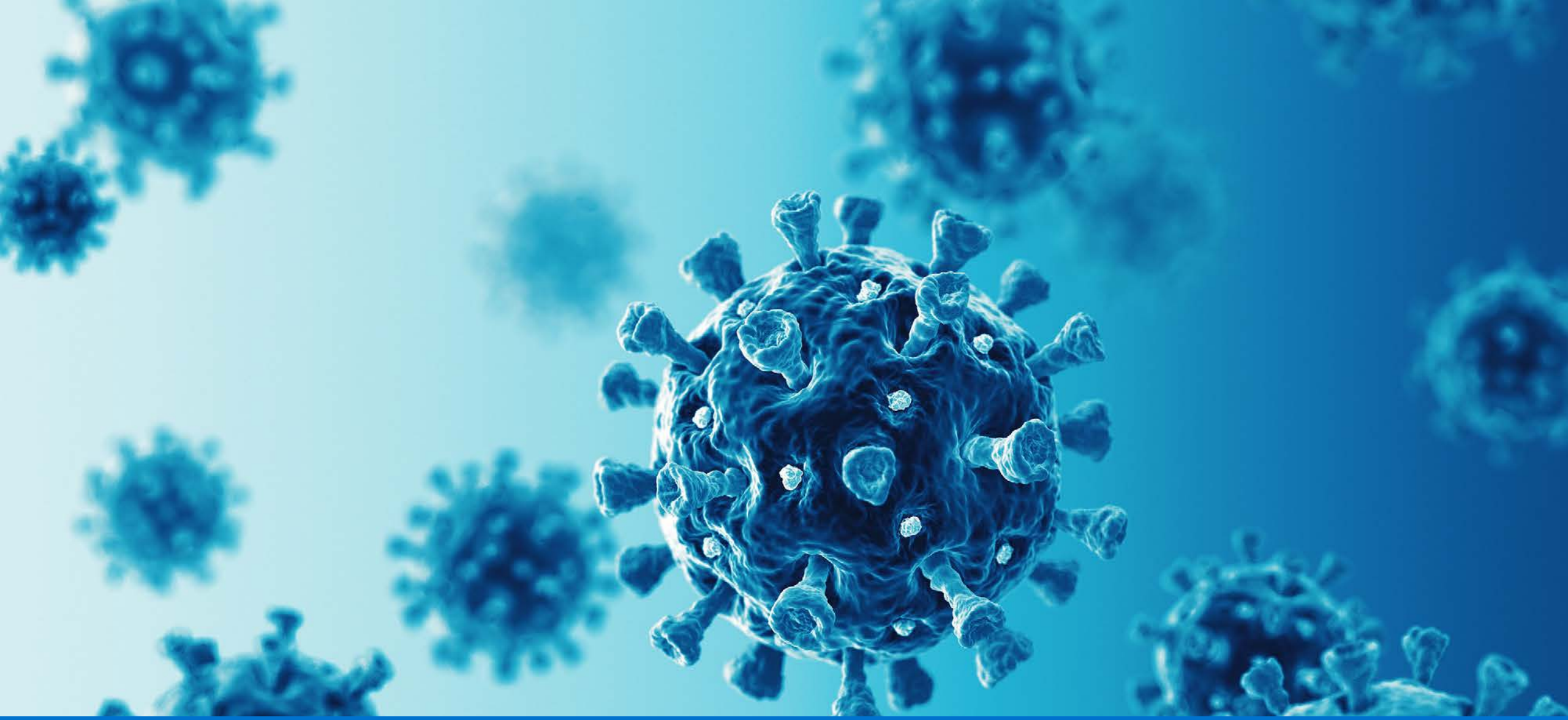


AIM-AHEAD Priority Areas

Priority Area	Description/Examples
Develop a diverse, equitable, and inclusive AI/ML workforce.	Train the workforce in AI/ML using existing and/or synthetic datasets through the lens of health equity.
Increase knowledge, awareness and national-scale community engagement/empowerment in AI/ML.	Develop a culturally-sensitive Community Health Worker (CHW) training in AI/ML.
Use AI/ML to address disparities and minority health in behavioral health, cardiometabolic health, and cancer.	Utilize existing data sets, then co-design data sets from new and/or existing data.
Build community capacity and infrastructure in AI/ML to address community-centric health disparities and minority health.	Conduct a needs assessment to prioritize health disparity areas/populations and invest in relevant and appropriate tools/resources/ expertise.

AIM-AHEAD Accomplishments

- 22 Research Fellowships awarded in 2022, engaging early-career researchers from under-represented populations in biomedical research that involves the use of AI/ML methodologies on Electronic Health Record Data.
- 25 Leadership Fellowships awarded in 2022, preparing a diverse leaders to champion the use of AI/ML in addressing persistent health disparities
- 22 **Pilot research projects** awarded in 2022, to test new paradigms of research, data analysis, and the new ways for underrepresented groups to derive value from their own health data.
- AIM-AHEAD connect platform launched as virtual hub for research at the intersection of AI/ML and health equity.
- Numerous webinars and symposia, including AI for Health Equity (AIEHS 2022)



Creating AI-ready data to address SARS-COV-2

Medical Imaging and Data Resource Center

Rapid Response to COVID-19 Pandemic

July 2020, NIBIB received a responsive document from a consortium of leading medical imaging organizations.

https://www.midrc.org/

EXPECTED IMMEDIATE IMPACT

Help corroborate, refine & advance ongoing efforts* in:

✓Detection, triaging, and differential diagnosis of COVID-19 patients.

✓Prognostic information, including prediction and monitoring of response, for use in patient management.

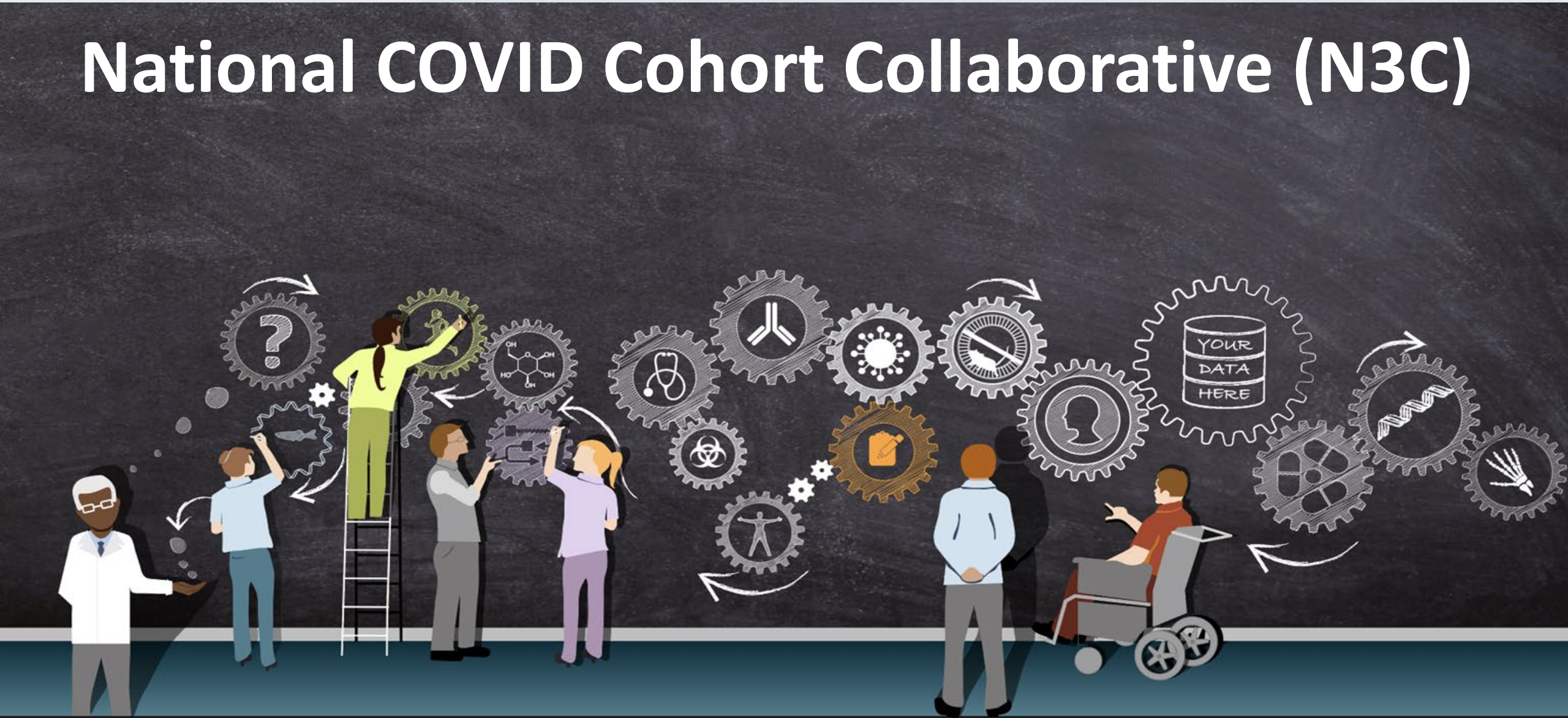
✓Surveillance of & early detection of COVID-19 resurgence.

*** Harmon SH et al. NATURE COMMUNICATIONS | (2020) 11:4080 |** https://doi.org/10.1038/s41467-020-17971-2
Artificial Intelligence for the Detection of COVID-19 Pneumonia on Chest CT using Multinational Datasets

NCATS

COLLABORATE. INNOVATE. ACCELERATE.

National COVID Cohort Collaborative (N3C)

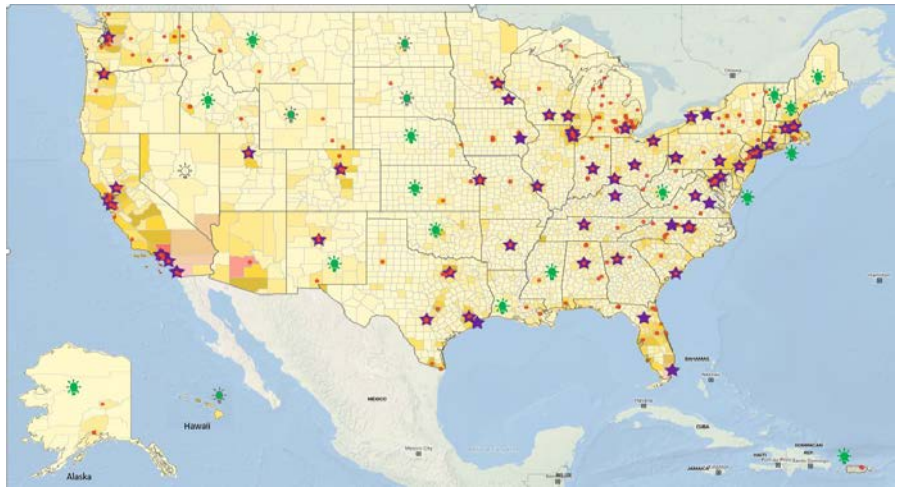


N3C Enclave Data: Current Stats (09/29/2022)

COVID-19 Positive Patients	Total Patients	Sites	Rows of Data
6.5m	16m	76	19.2b

Procedures	Lab Results	Drug Exposures	Visits	Observations
879m	9.1b	2.2b	942m	1.7b

	COVID (N = 4,942,399)	NON-COVID (N = 8,023,979)	OVERALL (N = 12,966,378)
Gender			
Male	2,226,566	3,556,683	5,783,249
Female	2,707,599	4,462,865	7,170,464
Unknown	8,234	4,431	12,665
Age			
0 - 17	661,007	1,244,337	1,905,344
18 - 29	870,584	1,122,119	1,992,703
30 - 49	1,497,864	2,137,611	3,635,475
50 - 64	1,040,583	1,675,633	2,716,216
65+	807,917	1,720,378	2,528,295
Unknown	64,444	123,901	188,345
Race			
White	3,257,810	5,188,191	8,446,001
Other	42,904	98,137	141,041
Black or African American	645,397	1,179,148	1,824,545
Asian	101,850	265,358	367,208
Pacific Islander	7,485	13,905	21,390
Unknown	886,953	1,279,240	2,166,193
Ethnicity			
Not Hispanic or Latino	3,620,865	6,177,402	9,798,267
Hispanic or Latino	637,331	933,659	1,570,990
Unknown	684,203	912,918	1,597,121



N3C Data

Institutions Contributing Data (101)
Geographics Distribution 48/50 States
Representative of US population
Health Centers: Community, Academic, FQHCs
Visits Inpatient, Outpatient ED

N3C Community

Institutions Using N3C (>296)
Investigators (>2900)
Research Projects (390)
Community Volunteers (>3500)
Domain Teams (34)

https://covid.cd2h.org/dashboard/

Using AI to Advance Understanding of Long COVID Syndrome

The COVID-19 pandemic continues to present considerable public health challenges around the world.

One of the most puzzling questions is why many people who get over an initial and often relatively mild COVID-19 illness later develop new and potentially debilitating symptoms.

In a groundbreaking study, NIH-supported researchers relied on machine learning to sift through vast amounts of electronic health care data in the N3C to look for patterns.

They developed three machine learning models: one to identify potential long COVID patients across the whole dataset, and two others that focused separately on people who had or hadn't been hospitalized.

All three models proved effective for identifying people with potential long-COVID. **Each of the models had an 85 percent or better discrimination threshold.**

Lancet Digit Health. 2022 May 16:S2589-7500(22)00048-6.

Training support to create a diverse AI-ready workforce

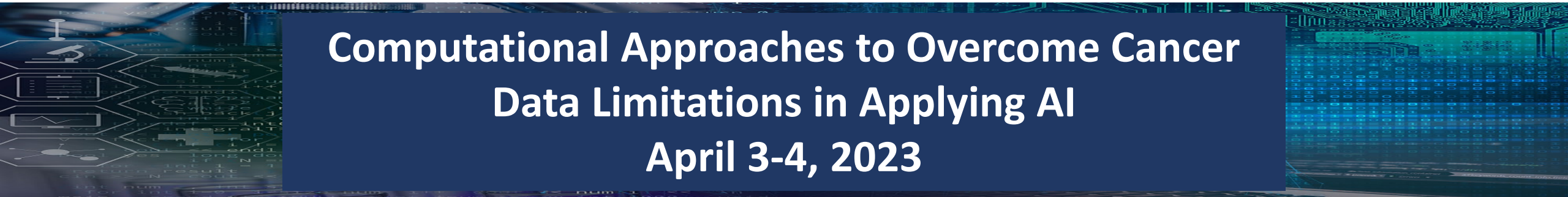
- Direct support of underrepresented groups in the AI workforce through programs like **AIM-AHEAD** and **Bridge2AI**
- Development of quantitative and computational skills in training programs aimed at underrepresented groups, like **NARCH** and the **SEPA diversity program**
- Creating collaborative communities (such as **NCI Cancer AI Accelerator**)
- Incorporating AI/ML education activities (conferences, workshops, webinars)

Other Equity Efforts in the AI/ML Workforce

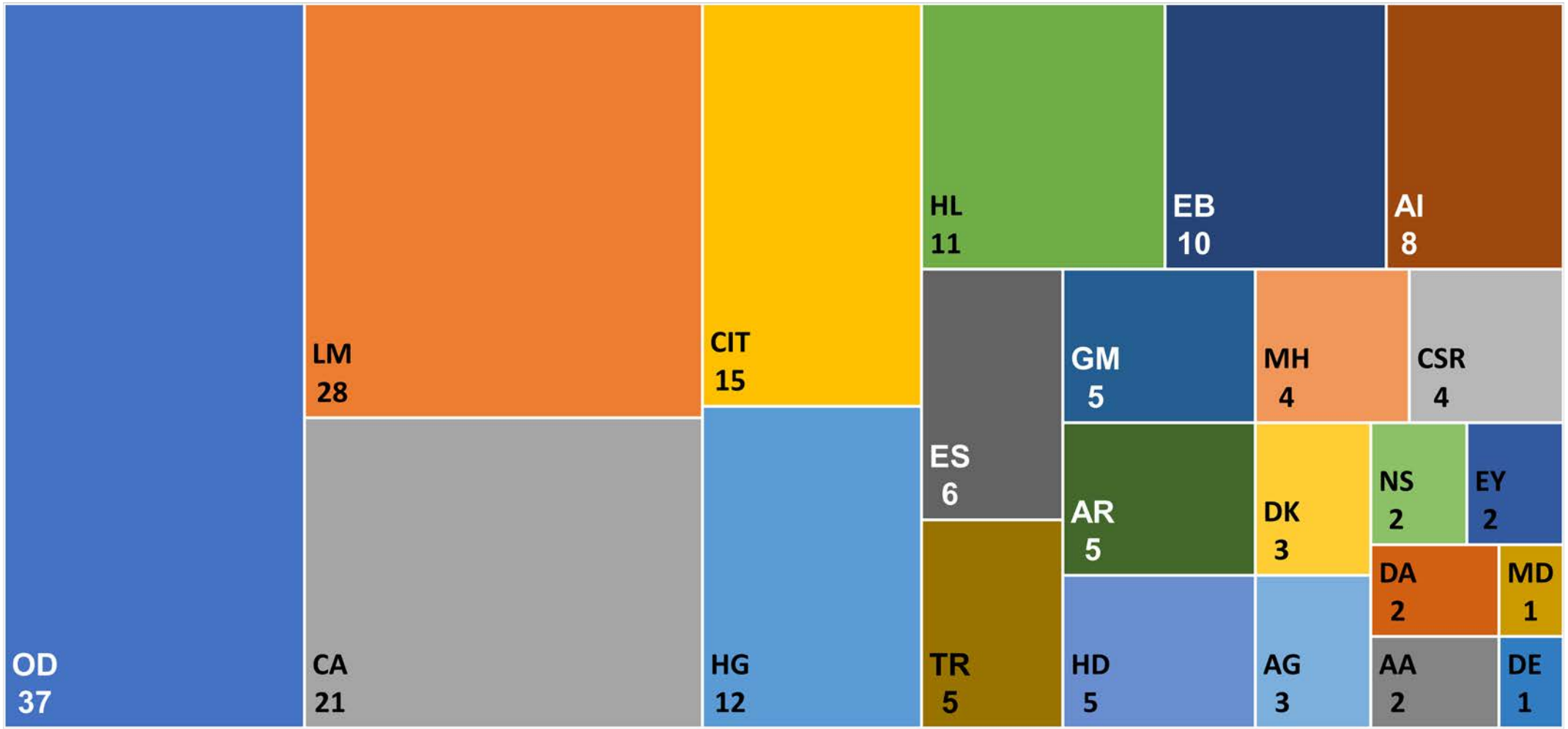
- Incorporating AI/ML educational activities, e.g. “Demystifying Machine Learning and Best Data Practices Workshop Series for Underrepresented STEM Undergraduate and MS Researchers Bound for PhD Training Programs”
- Supporting computational skills development as part of research capacity building, e.g. applying AI/ML to population health within AI/AN communities within the Native American Research Centers for Health (NARCH) program
- Pre-K-12 data science projects for students and teachers through the SEPA diversity program
- The Annual Conference to Increase Diversity in Mathematical Modeling and Public Health hosted by the NIGMS-supported Models of Infectious Disease Agent Study coordinating center
- Research Education Program to Support Short-term Research Experiences in Bioinformatics and Data Science and Enhance Diversity

NCI Cancer AI Accelerator: Convening communities to advance AI in cancer to address common challenges through diverse perspectives

- Brings together a diversity of ideas, approaches and participants
- Fosters discussion and adoption of promising practices in AI ethics, collection and use of data from diverse populations, and AI-relevant standards
- Nimble and adaptive organization that responds to community needs
- Focuses on community identified challenges and opportunities
- Provides agile mechanisms to support sand-pit style meetings, prize challenges, and pilot projects



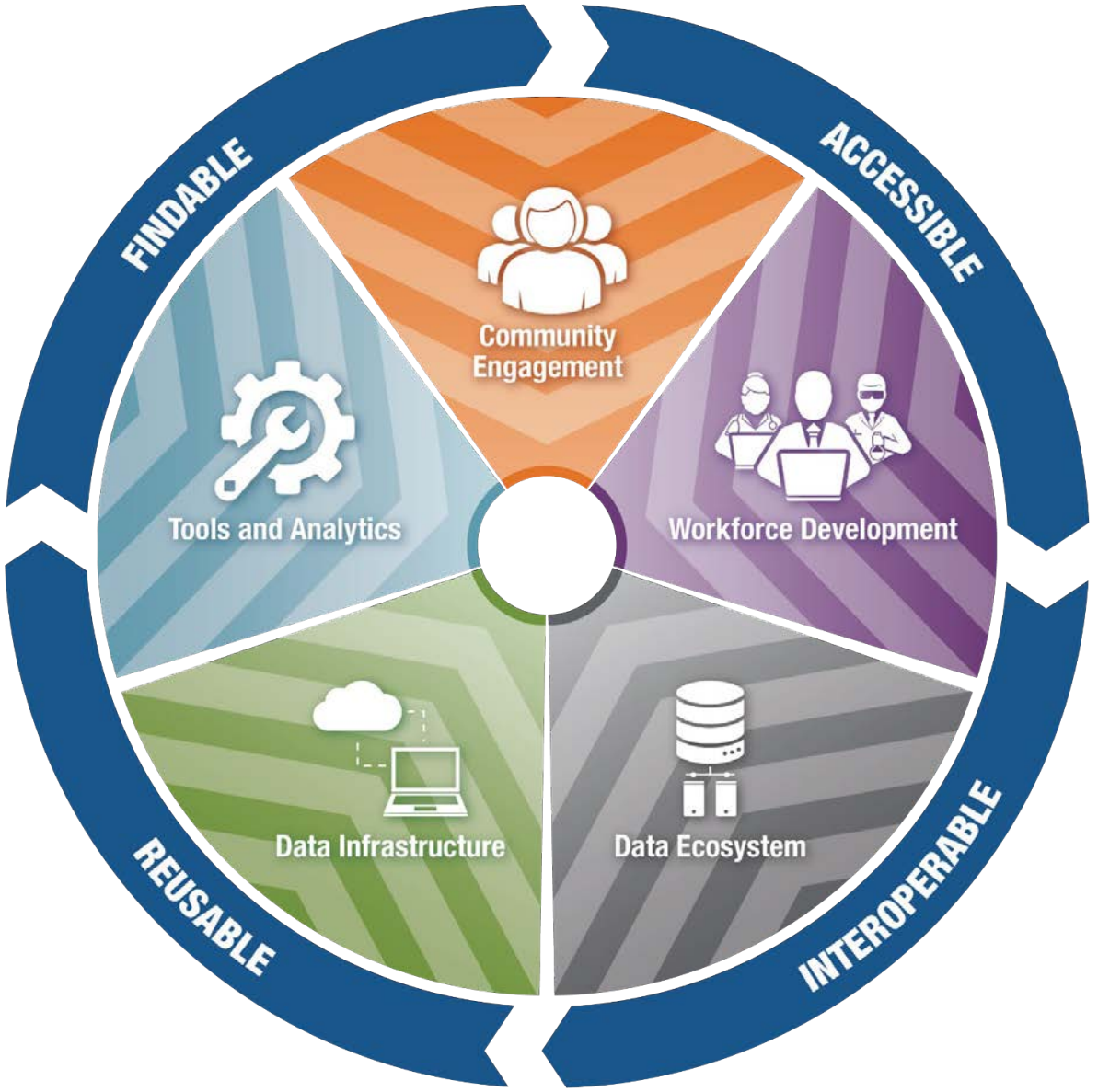
Credits – More than 200 NIH Staff from 23 ICOs Catalyze Data Science Across NIH



Office of Data Science Strategy

www.datascience.nih.gov

A modernized, integrated, FAIR biomedical data ecosystem



BREAK

1:00 p.m. – 1:10 p.m.

Instruction for Breakout Discussions

Breakout Discussion Instructions

- **You have a total of 60 mins – until 2:30pm ET**
- **Take just a few minutes for brief, "one-breath" introductions.**
- **Designate one member of the group to give a ~5 min read out of key themes from this discussion in the next session**
- **Capture your thoughts in the google doc as much as possible. This document will stay open and can be referenced during the read-out**

Breakout Discussion Instructions

- **Links to google docs:** https://docs.google.com/document/d/1Z4Dr3zEtCV4MKNeehHEdw5O87YpSXhiqDLkCX2wDl0A/edit?usp=sharing
- **When you get to your breakout room, open the google doc that matches your breakout room number**

Discussion Questions

- What does it mean for data to be AI Ready?
  - How do you create AI ready data? (what steps need to be taken?)
  - How can you tell when data are AI-ready? (what are the attributes of AI-ready data?)
- What does the ethical development of AI mean?
- This is a dynamic space: For example, technologies and AI/ML applications are evolving rapidly; new concepts of community and stakeholder engagement are emerging; many open questions remain regarding best practices for upholding privacy and ethics goals. How can NIH keep pace? How can NIH researchers help innovate?

BREAK

2:10 p.m. – 2:15 p.m.

Readout from Breakouts

2:15 p.m. – 3:05 p.m. Readout from Breakouts

- Recap: One volunteer from each group, share most interesting points of discussion from your breakout in ~5 minutes

Open Conversation

Open Conversation

- Please join the SLIDO platform:
  - https://app.sli.do/event/aCTaYatrrbnrJABv1Ffq7e
    Code: #2567081

Thank you!!!

Thank you!

Michael Spittel, Ph.D.
Health Scientist Administrator
Office of Data Science Strategy

Mark Dennis
Deputy Director, Conference Services at The Scientific Consulting Group, Inc.

NIH-Wide working groups in AI-Workforce and AI-Readiness

https://www.scgcorp.com/odssaicloseout2022/

Thank you!