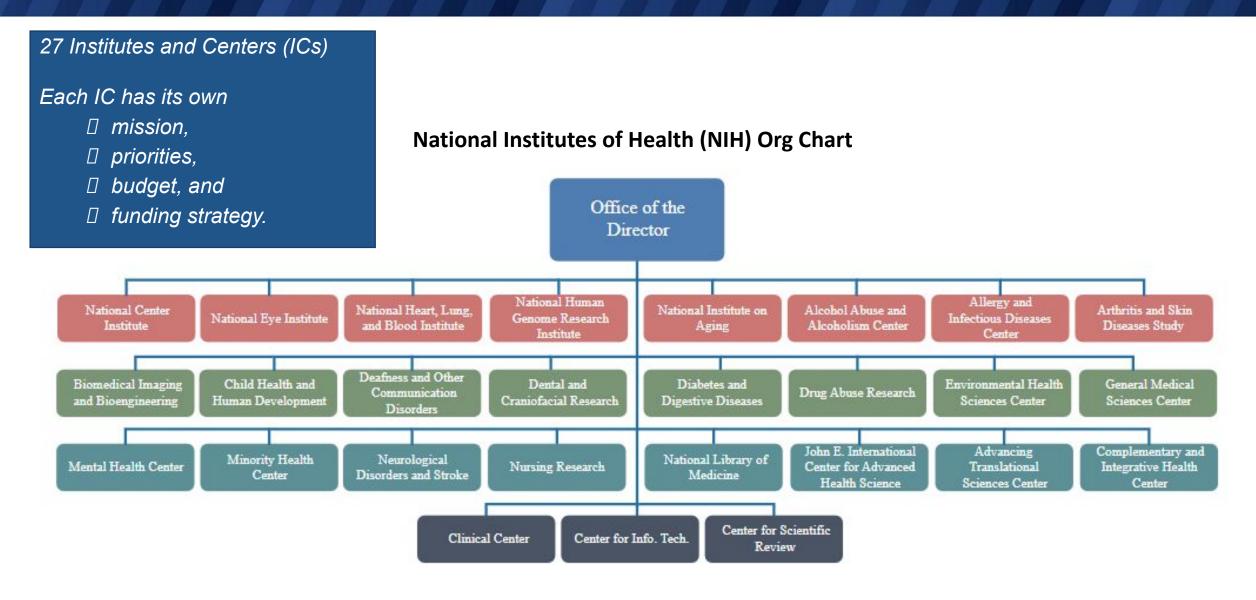


Al at the NIH: Opportunities to Improve Biomedical Research and Human Health

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

June 16, 2022

National Institutes of Health Institutes, Centers, and Offices



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

Support common infrastructure and architecture for specialized platforms.

Leverage industry resourced by adopting and adapting from other fields to use in biomedical research.

Enhance the biomedical data-science research workforce through improved programs and novel partnerships.

Enhance data sharing, access, and interoperability such that NIH-supported data resources are FAIR.

Ensure the security and confidentiality of data in accordance with NIH requirements and applicable law.

With community input, develop, promote, and refine data standards, including standardized data vocabularies and ontologies.

Improve the ability to capture, curate, validate, store, and analyze clinical data for biomedical research.

NIH Strategic Plan for Data Science – Goals & Objectives

Data Infrastructure

Optimize data storage and security

Connect NIH data systems

Modernized Data Ecosystem

Modernize data repository ecosystems

Support storage and sharing of individual datasets

Better integrate clinical and observational data into biomedical data science

Data Management, Analytics, and Tools

Support useful, generalizable, and accessible tools

Broaden utility of, and access to, specialized tools

Improve discovery and cataloging resources

Workforce Development

Enhance the NIH data science workforce

Expand the national research workforce

Engage a broader community

Stewardship and Sustainability

Develop policies for a FAIR data ecosystem

Enhance stewardship

Al and Biomedicine: Visions for the Future



The promise of data science for AI is not in the re-analysis of individual datasets alone, it is in the comparative and correlative relationships among datasets.

Having the ability to infer and model relationships on aggregated sets of disparate data underlies the more difficult components of data interoperability and data for Al.

Al and Biomedicine: Visions for the Future

Future Al promises to also help understand causality. Along the way, "explainable Al" may help machines teach humans.

Advances in semantic artificial intelligence and inference algorithms could create harmonized data models making it possible to integrate biomedical data across federated resources.



Al and Biomedicine: Visions for the Future

One area that remains understudied within the Al realm is its promise for improving the health of racial and ethnic minority populations and reducing health disparities.

These capabilities build on work the that NIH has done toward a modern, FAIR data ecosystem and is an extension of the vision outlined in the Strategic Plan for Data Science.



Creating the largest Al-ready biomedical datasets for research

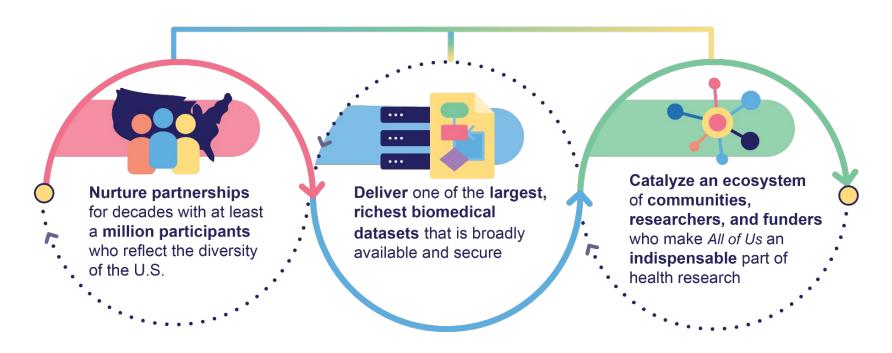
- All of Us
- Nutrition for Precision Health
- Bridge2Al
- AIM-AHEAD
- Medical Imaging and Data Resource Center
- National COVID Cohort Collaborative



All of Us Research Program Mission

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

Status of the All of Us Research Program



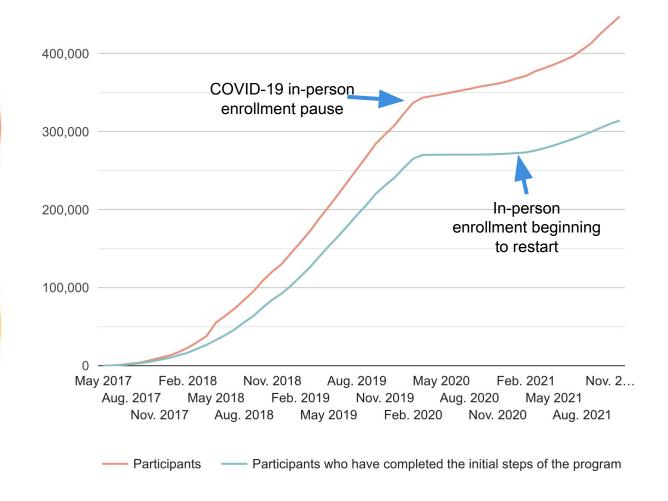
283,000+
Electronic Health
Records

324,000+

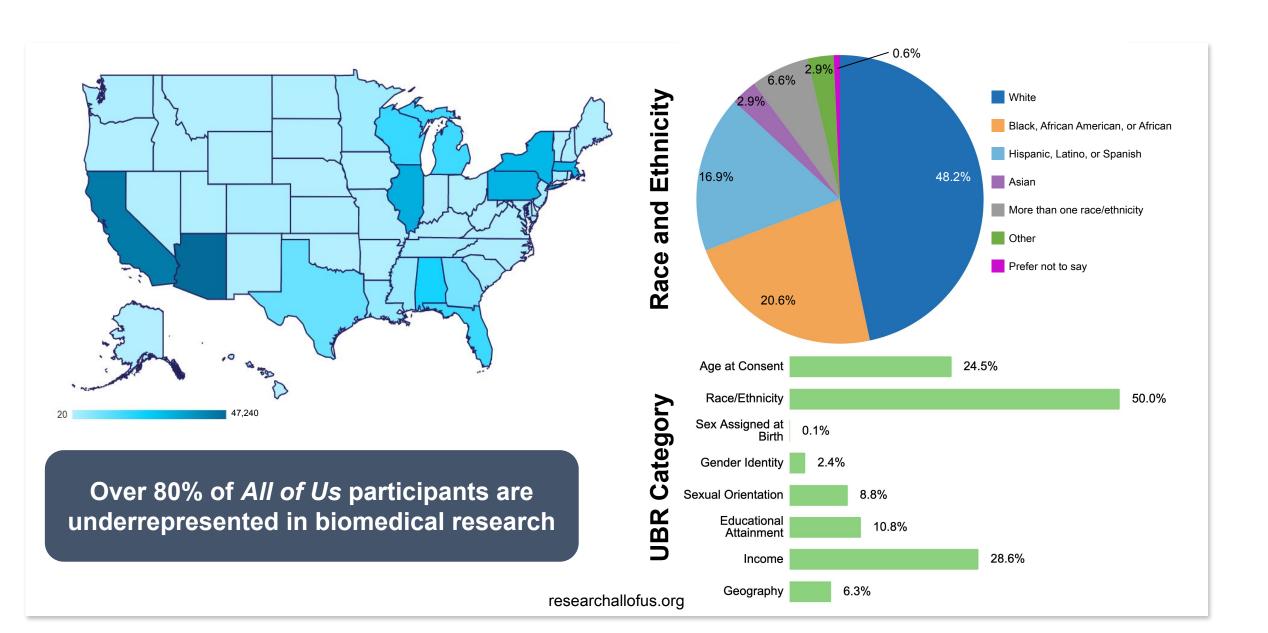
Participants who have completed initial steps of the program

345,000+

Biosamples



Status of the All of Us Research Program



Data on the Researcher Workbench Diverse and Longitudinal



329,000+ Participants



267,600+ Physical Measurements



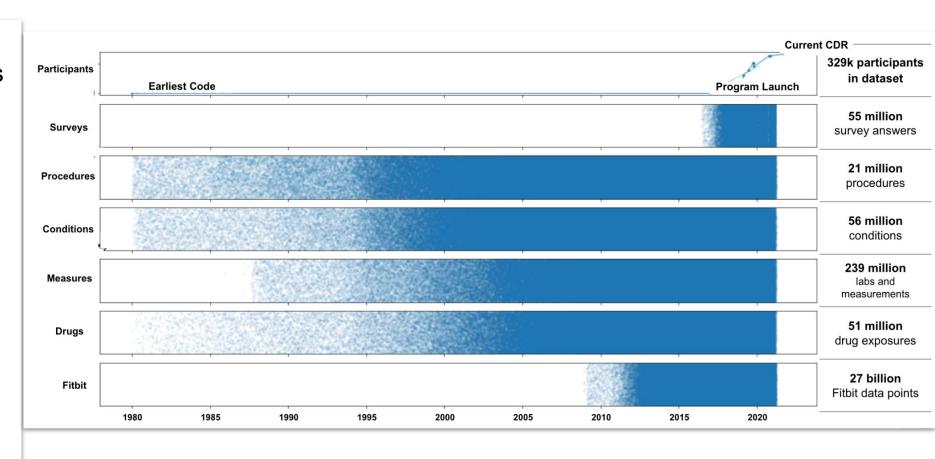
214,200+ EHRs



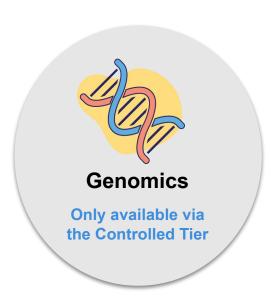
329,000+ Surveys

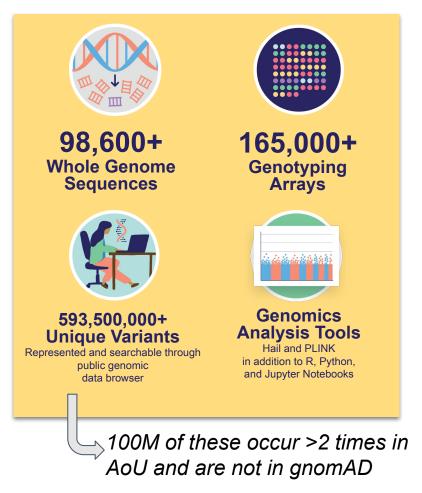


11,600+ Fitbit Records



Including the first batch of genomic data in the Controlled Tier





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

- This new effort, powered by the Allofus program will recruit a diverse pool of 10,000 participants to inform more personalized nutrition recommendations.
- A major challenge in precision nutrition is the inability to combine the 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.

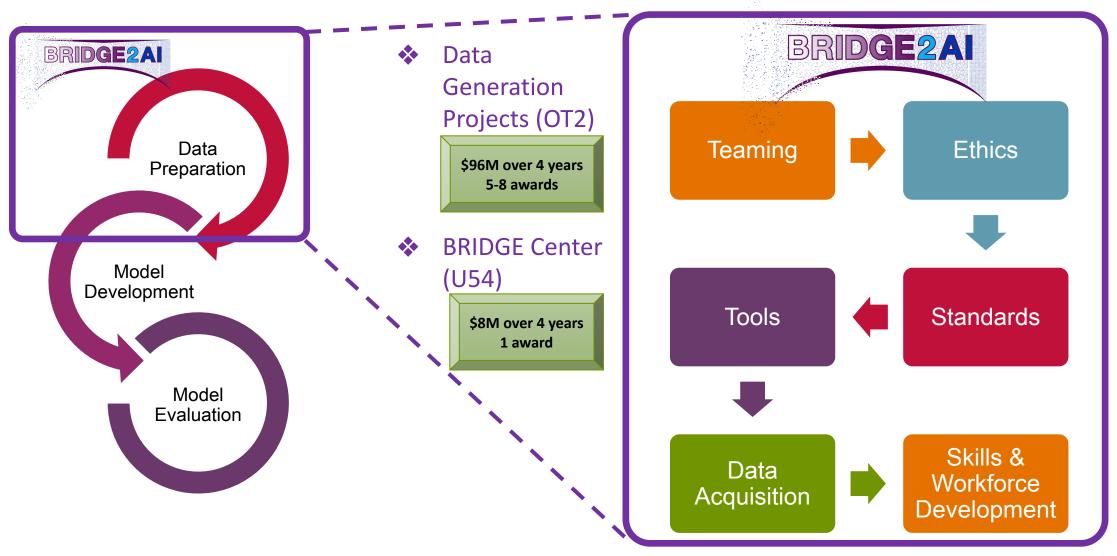
Bridge2Al

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

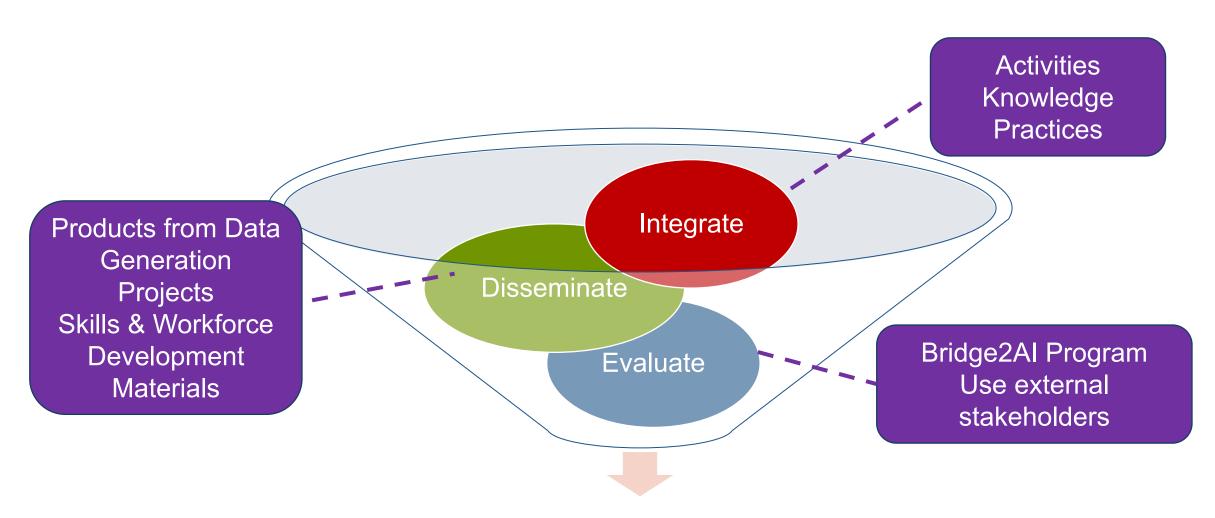


Standardize Data Develop **Automated Tools** Attributes New **Datasets** Create cross-training Disseminate materials for Products & Best Workforce Practices Development

Preparing the Data



BRIDGE Center



Best Practices for Al/ML in Biomedical and Behavioral Research

Who needs to come together?

- Biomedical Scientists
- Behavioral Scientists
- Social Scientists
- Clinicians
- Economists
- Ethicists
- Philosophers
- Anthropologists

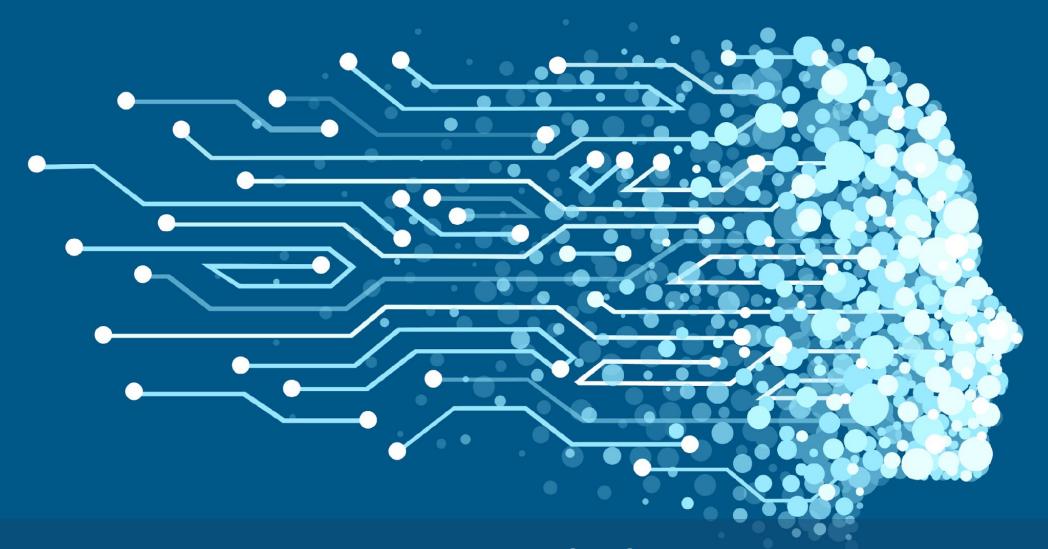


- Diverse career stages
- Diverse social and cultural backgrounds
- Academic, industrial, and technical backgrounds
- Diverse communities and institutions

- AI/ML Experts
- Statisticians
- Computer Scientists
- Data Scientists
- Mathematicians
- Bio-informaticists
- Engineers
- Team Science Experts...

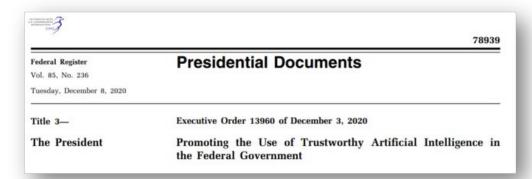
What will Bridge2Al Produce?

- Novel, complete, trustworthy datasets ethically sourced, following <u>FAIR</u> 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



Actioning BioMed-ML (AI) to address health disparities

Ethics is at the forefront of government wide initiatives to accelerate AI/ML



https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government



https://www.hhs.gov/sites/default/files/final-hhs-ai-strategy.pdf

IMMEDIATE RELEASE

DOD Adopts Ethical Principles for Artificial Intelligence

FEB. 24, 2020

https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

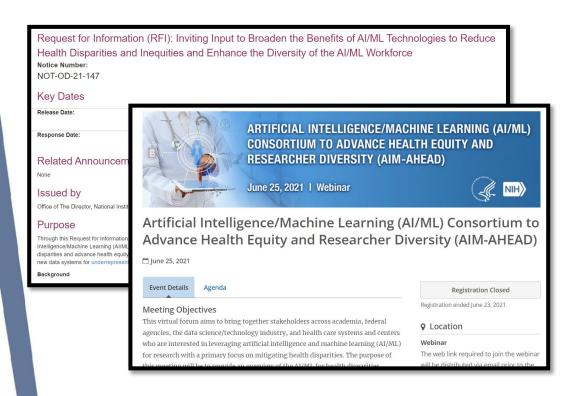


https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community



https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/

Community Input Shaped the Initial Phase

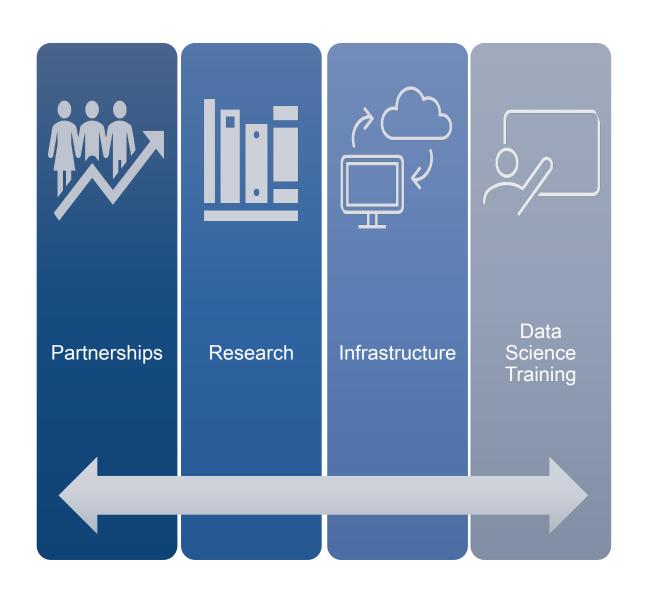


There is a wide variety of interests, needs, and resources across communities.

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 health disparity populations.
- 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

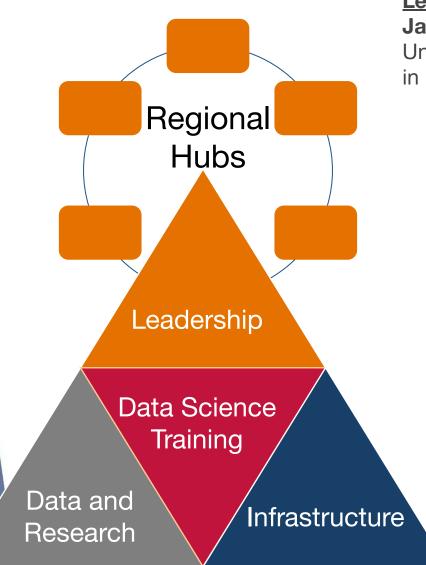
AIM-AHEAD GOALS



Goals

- •Increase participation and representation of researchers and communities currently underrepresented in the development of 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 records (EHR) and extending to other diverse data

The AIM-AHEAD Coordinating Center



Leadership Core

Jamboor K. Vishwanatha, Ph.D.

University of North Texas Health Science Center in Fort Worth

Regional Hubs

Toufeeq Ahmed, Ph.D.

Vanderbilt University Medical Center

Bettina Beech, Dr.P.H.

University of Houston

Harlan P. Jones, Ph.D.

University of North Texas Health Science Center in Fort Worth

Spero Manson, Ph.D.

University of Colorado-Anschutz Medical Center in Aurora

Keith Norris, M.D., Ph.D.

University of California, Los Angeles

Anil Shanker, Ph.D.

Meharry Medical College in Nashville, Tennessee **Herman Taylor, M.D.**

Morehouse School of Medicine in Atlanta, Georgia Roland J. Thorpe, Jr., Ph.D.

Johns Hopkins University in Baltimore, Maryland

Data Science Training Core

Legand L. Burge, Ph.D.

Howard University in Washington, D.C.

Infrastructure Core

Alex J. Carlisle, Ph.D.

National Alliance Against
Disparities in Patient Health in

Woodbridge, Virginia

Paul Avillach, M.D., Ph.D.

Harvard Medical School in

Boston, Massachusetts

Bradley A. Malin, Ph.D.

Vanderbilt University Medical Center in Nashville, Tennessee

Data and Research Core

Jon Puro, M.P.A.

OCHIN in Portland, Oregon

https://aim-ahead.net/

AIM-AHEAD PARTNERSHIP MAP



AIM-AHEAD Research Fellowship

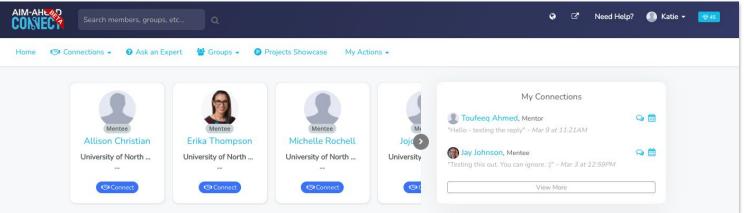
Purpose: To engage a group of 20 early career researchers to actively participate in biomedical research that involves the use of AI/ML methodologies.

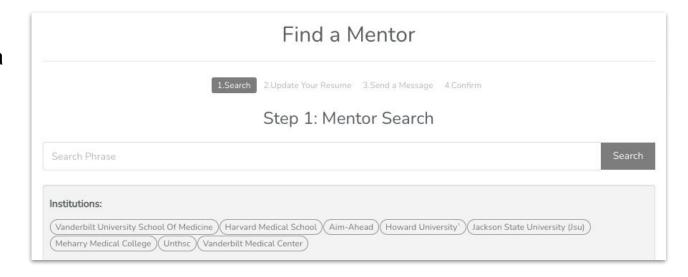
- AIM-AHEAD Research Fellowship start date: September 1st, 2022
- Limited access to AIM-AHEAD data, (OCHIN's ADVANCE EHR) infrastructure (Centralized cloud computing environment), and training will be made available starting in the Fall of 2022.
- Fellows will be matched with an AIM-AHEAD and/or an Institutional mentor(s)
- Stipend for each Fellow
- FAQ and help desk available during the application period and throughout the fellowship
- As the AIM-AHEAD Consortium develops, more data, infrastructure and training will be made available to the fellows.

AIM-AHEAD CONSECT

What is AIM-AHEAD Connect?

- AIM-AHEAD Connect is a mentoring, networking, and professional development platform to connect the AIM-AHEAD Consortium and community.
- Once someone joins AIM-AHEAD Connect as a Mentee (learner) or Mentor (expert), they will have access to all of the resources, such as:
 - Mentoring
 - Groups
 - Projects
 - Publication Sync
 - CV/Resume Builder
 - And more!
- Some Mentors may quality to mentor AIM-AHEAD Research Fellows



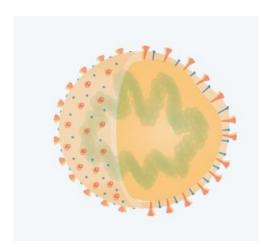


AIM-AHEAD

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

- "...we need to build on programs like the new NIH AIM-AHEAD (or at least ensure their funding continues), to not only make sure diversity is covered in biomedical data sets, but diversity is promoted and enhanced among the data scientists themselves."
- -- Atul Butte, MD, PhD, Priscilla Chan, and Mark Zuckerberg Distinguished Professor, University of California, San Francisco, Director, Bakar Computational Health Sciences Institute and Chief Data Scientist, University of California Health
- "I never anticipated the appetite for this initiative in [the American Indian, Alaskan Native, and Hispanic] communities... There is a thirst for this."
- -- Spero Manson (Pembina Chippewa), Distinguished Professor of Public Health and Psychiatry, Director for the Centers for American Indian and Alaska Native Health, Associate Dean of Research at the Colorado School of Public Health at the University of Colorado Denver's Anschutz Medical Center

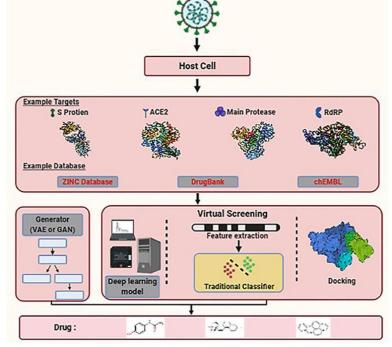
Accelerating us of Artificial Intelligence to address COVID-19



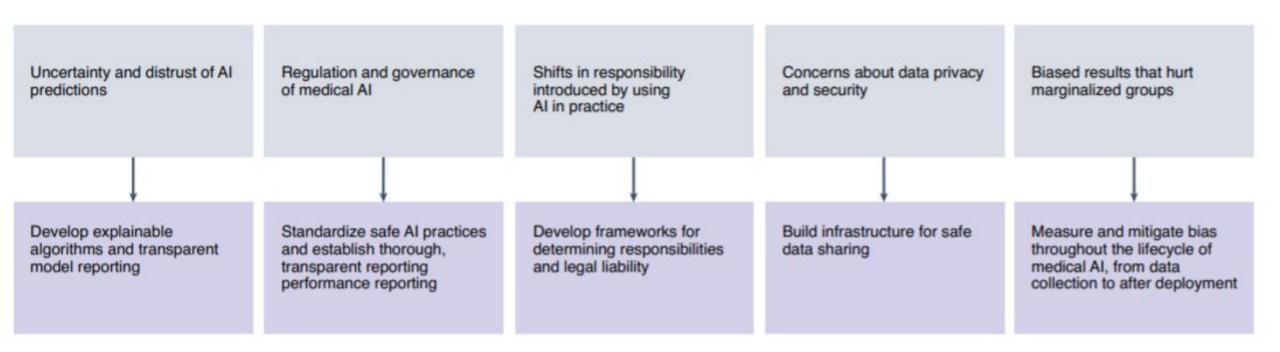
LitCovid is a curated literature hub for tracking up-to-date scientific information about the 2019 novel Coronavirus.

The White House partnered with AI research institutions to mine scientific literature to better understand Covid-19.

Leveraging AI to understand the structure of the novel coronavirus and to expedite drug discovery



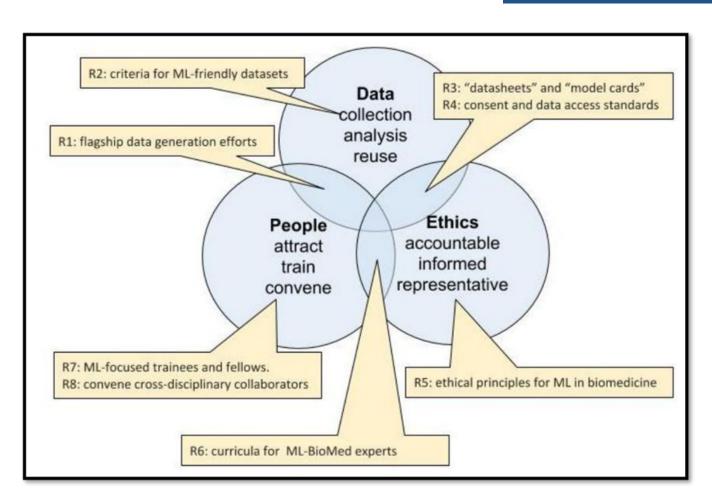
Prioritizing Ethics in Biomed-Al



Rajpurkar, P., Chen, E., Banerjee, O. *et al.* Al in health and medicine. *Nat Med* **28,** 31–38 (2022). https://doi.org/10.1038/s41591-021-01614-0

Biomedical AI: Visions for an ETHICAL Future

NIH ACD AI Working Group Recommendations



 Outlined opportunities to fuse AI/ML with exponential increase in biomedical data

 Ethics was identified as equally important to Data and People, reflecting the primary importance of infusing ethical thinking into AI/ML use in biomedical research

Collaboratively Envisioning Al and Ethics in Biomedical Research

The NIH is bringing together a diverse cross-section of scientists, social scientists, ethicists, advocates, legal scholars, communicators, and artists interested in the social implications of technology to

- o Forge new collaborations among these cross-disciplinary groups
- o Identify important areas of consideration at the intersection of artificial intelligence (AI) and machine learning (ML), biomedicine, and ethics.
- o Generate creative strategies to solve ethical dilemmas in biomedical AI/ML

Collaboratively Envisioning Al and Ethics in Biomedical Research

Micro Lab #1

Dec 15th, 2021, 2-4pm ET

Who are the relevant stakeholders?

Micro Lab #2

Jan 12th, 2022, 2-4pm ET

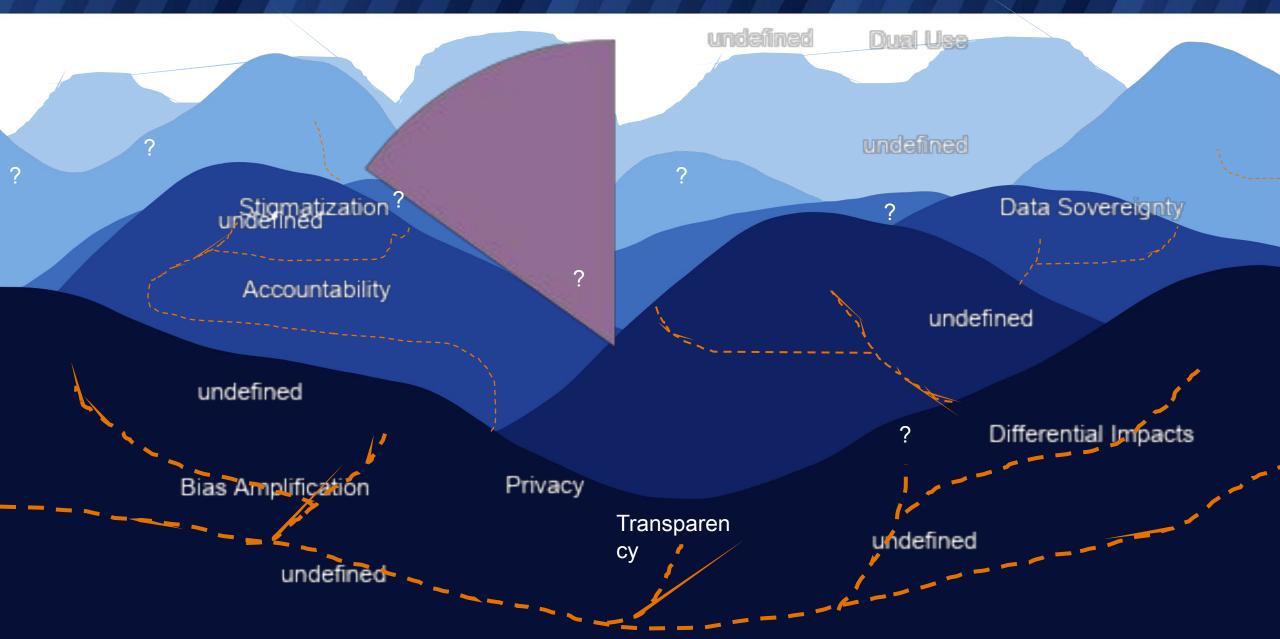
What are the key opportunities, challenges, and themes?

Micro Lab #3

Jan 26th, 2022, 2-4pm ET

Organizing and understanding opportunity

What are the key opportunities, challenges, and themes?



Collaborations to Advance Ethical Use of Al/ML

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

ODSS will support collaborations that bring together expertise in ethics, biomedicine, data collection, and AI/ML to advance the understanding, tools, metrics, and practices for the ethical development and use of AI/ML in biomedical and behavioral sciences.



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 Al/ML, including the collection and generation of data as well as the reuse of data and models by others. Research products developed under this NOSI will be shared and made broadly reusable.

Collaborations to Make Data FAIR and AI/ML Ready

AI/ML is inclusive of machine learning (ML), deep learning (DL), and neural networks (NN). Making data AI/ML-ready is not simply formulaic. It requires engagement with and feedback from AI/ML applications.

NIH supported **collaborations** to bring together expertise in biomedicine, data management, and artificial intelligence and machine learning (Al/ML) to make NIH-supported data Al-ready for Al/ML analytics to:

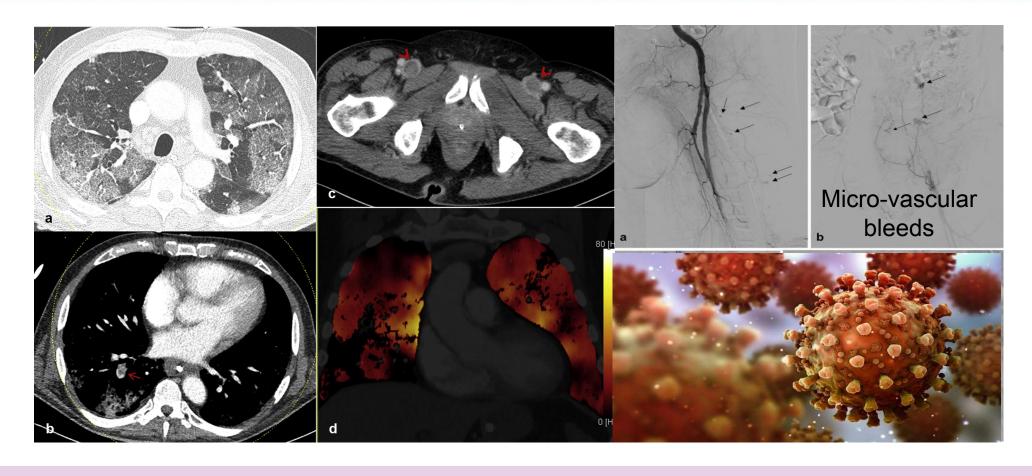
- Imputing missing metadata; improve data preprocessing; adoption of ontologies or other standards Discovering and identifying imbalances in the data, biases in data labels or metadata, or other attributes of the dataset that would help researchers make better, more ethical decisions when using the data for AI/ML.
- Addressing specific challenges related to federated data and federated learning.
- Developing and sharing documentation, e.g. datasheets, that document the provenance, motivation, composition, collection process, recommended uses, and other relevant information for AI/ML re-users of the data
- Demonstrating the usability of the data in Al/ML applications through, for example, Al/ML applications, hackathons, or citizen science challenges and competitions.





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

JANUARY 2020 SARS 2 COVID-19



A rapidly progressive viral infection causing pulmonary (ARDS) & cardiovascular havoc resulting from <u>leaky blood vessels & inflammation</u>, often associated with death or prolonged recovery!

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/

Medical Imaging & Data Resource Center EXPECTED IMMEDIATE IMPACT

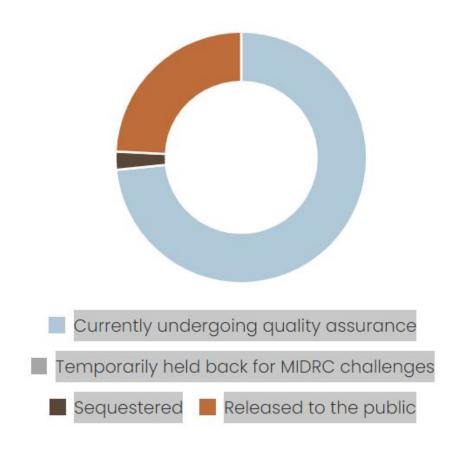
Help corroborate, refine & advance ongoing efforts* in:

- ✓ 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

MIDRC serves as a linked-data commons that coordinates access to data and harmonizes data management activities at three critical stages:

- (1) intake, including curation, de-identification, abstraction, and quality assessment
 - (2) annotation and labelling of images and other data using semi-automated approaches
 - (3) distributed access and query methods.



Total number of imaging exams ingested into MIDRC to date N=85,397

https://www.midrc.org/

National COVID Cohort Collaborative (N3C)



N3C Enclave Data: Current Stats (04/04/2022)

COVID-19 Positive Patients

4.9m

Total Patients

13m

Sites

72

Rows of Data

14.4b

Procedures

663m

Lab Results

6.9b

Drug Exposures

2.2b

Visits

711m

Observations

1.3b

N3C Data

Institutions Contributing Data (92) Geographics Distribution 48/50 States Representative of US population Health Centers: Community, Academic, FQHCs Community Volunteers (>3500) Visits Inpatient, Outpatient ED Data Set 1/1/2018 to Present

N3C Community

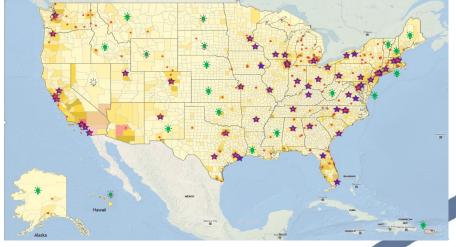
Institutions Using N3C (296)

Investigators (>2400)

Research Projects (336)

Domain Teams (35)

	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 Enclave Data: Current Stats (04/04/2022)

COVID-19 Positive Patients

4.9m

Total Patients

13m

Sites

72

Rows of Data

14.4b

Procedures

663m

Lab Results

6.9b

Drug Exposures

2.2b

Visits

711m

Observations

1.3b

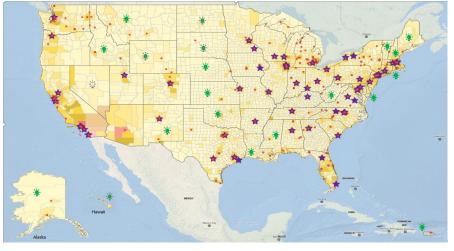
N3C Data

Institutions Contributing Data (92)
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 (>2400)
Research Projects (336)
Community Volunteers (>3500)
Domain Teams (35)

	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



Using AI to Advance Understanding of Long COVID Syndrome

The COVID-19 pandemic continues to present considerable public health challenges in the United States and around the globe.

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



- In a groundbreaking study, NIH-supported researchers relied on machine learning to sifts 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.

Partnerships in AI/ML National Cancer Institute (NCI) and the Department of Energy (DOE)

<u>Joint Design of Advanced Computing Solutions for</u> <u>Cancer</u> (JDACS4C) program providing cutting-edge computational models, algorithms, data sets, software, and other capabilities to the broader research community

- Identify promising new treatment options using advanced computation to rapidly develop, test, and validate predictive preclinical models for precision oncology
- Deepen understanding of cancer biology using molecular, functional, and structural data from the <u>NCI RAS</u>
 <u>Initiative</u> through improved dynamic computer simulations and predictive models
- Transform cancer surveillance by applying advanced computational capabilities to population-based cancer data to understand the impact of new diagnostics, treatments, and patient factors
- Software, data set, models public pre-release.
 https://datascience.cancer.gov/collaborations/joint-design-advanced-computing

National Cancer Institute & Department of Energy Collaborations Joint Design of Advanced Computing **CANcer Distributed Learning** Solutions for Cancer Pilots **Environment (CANDLE)** Cellular Level Pilot for Predictive Modeling for Pre-clinical Screening Accelerating methods to identify promising new treatments DOE exascale computing project Information gathered from the pilots will: Molecular Level Pilot for RAS Z Structure and Dynamics in Cellular Provide insight into Membranes scalable machine learning tools Deepening understanding of cancer 2 Provide analytics to reduce time to solution Inform the design of Population Level Pilot for Population future computing Information Integration, Analysis solutions and Modeling Understanding the impact of new diagnostics, treatments and patient factors in cancer outcomes Address critical needs in computing, data transfer, and data management in cancer research.

NIH NATIONAL CANCER INSTITUTE

Scarching for new antibiotics in the human

- Antibiotic resistance is a significant public health problem, with bacterial infections becoming increasingly difficult to treat. In 2019, approx 35,000 people in the U.S. alone died from antibiotic-resistant infections.
- ☐ To identify peptide antibiotics, the team used artificial intelligence to screen the entire human proteome—the set of all proteins in the human body.
- ☐ The team selected 55 top candidates and synthesized them in the lab for further testing. In cultures of eight common infectious bacteria, more than 60% of the peptides showed some ability to kill at least one of the microbes.

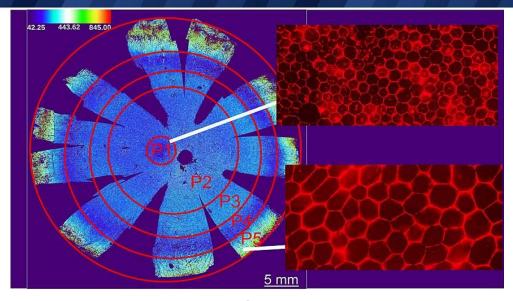


Research team led by Dr. Cesar de la Fuente-Nunez from the University of Pennsylvania funded in part by NIH's National Institute of General Medical Sciences (NIGMS).

Results were published on November 4, 2021, in *Nature Biomedical Engineering*

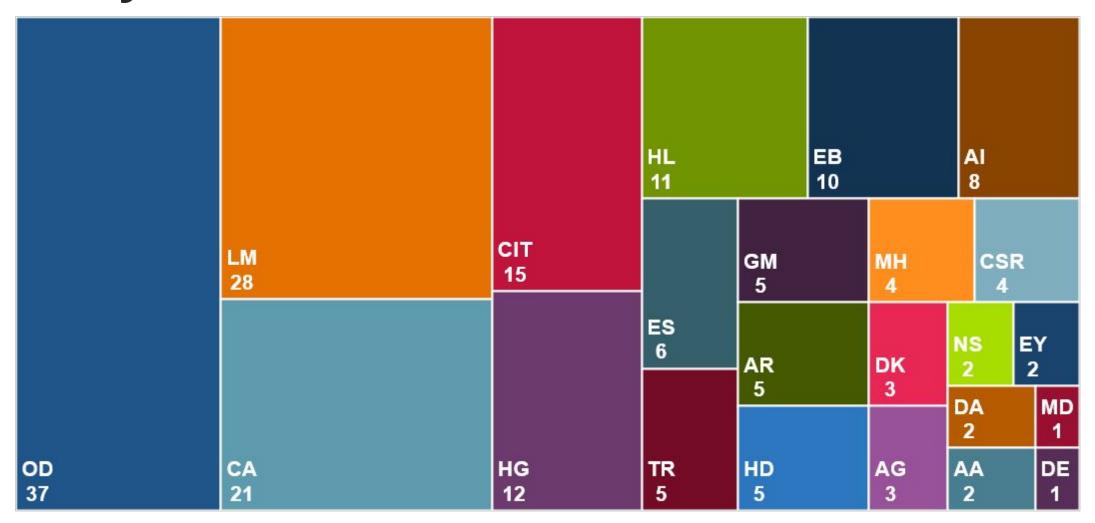
Retinal cell map could advance precise therapies for blinding diseases

- Last month scientists from the National Eye Institute (NEI), led by Dr. Kapil Bharti, discovered five subpopulations of retinal pigment epithelium (RPE)—a layer of tissue that nourishes and supports the retina's light-sensing photoreceptors
- Using artificial intelligence, the researchers analyzed images of RPE at single-cell resolution to create a reference map that locates each subpopulation within the eye
- By analyzing changes to these cell's morphometry researcher believe this may signal visual degeneration
- The goal is to develop noninvasive imaging technologies, such as adaptive optics, coupled with AI, for use in ophthalmology
- A report on the research was published in Proceedings of the National Academy of Sciences



Five subpopulations of RPE (P1-P5) were identified based on cell area, aspect ratio, hexagonality and number of neighbors. Foveal RPE (P1) are tightly packed hexagons. Peripheral RPE (P5) are spread out. *Davide Ortolan, Ph.D.*

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

