2023 DATA Scholar Program Information Session

Alison Lin, PhD & Bryan Kim, PhD
Training, Workforce Initiatives and Community Engagement (TWICE)
NIH Office of Data Science Strategy

April 19, 2023
Work at NIH: DATA Scholar Program

DAta and Technology Advancement (DATA) National Service Scholar Program

- Attract talent to the NIH to help optimize and accelerate data science in biomedicine and health research
- Encourage transformative approaches that lead to increased efficiency, innovative research, tool development and analytics
- One to two years commitment
- In addition to their own project, scholars participate in workgroups and collaborations, and contribute in many ways to the NIH
• 20 Scholars have been matched to 14 ICOs since program launch in 2020 and 7 have completed the program

FY 2022 Accomplishments Include:

• 7 peer reviewed publications
  Example: Precision Medicine Landscape of Genomic Testing for Patients With Cancer in the National Institutes of Health All of Us Database Using Informatics Approaches by Jay Ronquillo (NCI; cohort 1); JCO Clinical Care Informatics, 2022

• 13 technologies, products (including websites), inventions, patent applications, and shared resources developed
  Example: “Kids First Cloud Credits Program” developed by Anne Deslattes Mays (NICHD; cohort 2) available: https://github.com/kids-first/kf-cloud-credits#readme

• 26 oral presentations at scientific conferences
  Example: "Medical Imaging and Data Resource Center: for a more equitable Medical Imaging AI" presented by Rui Pereira de Sa (NIBIB; cohort 1) at the Radiological Society of North America (RSNA), Nov 2022
Eligibility and Application

Eligibility:

• Must be U.S. citizens, non-citizen nationals, or Permanent Residents
• MD, PhD or other doctoral degrees
• Advanced experience in data science or related fields including:
  ➢ Artificial intelligence, cloud computing, data engineering, data science, database management, project management, software design, supercomputing, and/or bioinformatics

Application:

• **Online form**
  ➢ Applicant information
  ➢ Choices of projects
• Cover letter
  ➢ Motivation – why interested in program
  ➢ Contribution – how experiences can address data challenges at NIH
  ➢ Vision – impact of data science in biomedical research and health sciences
  ➢ Contribution to enhancing diversity in data science
• Resume – accomplishments including:
  ➢ Data science projects, publications and products
  ➢ Expertise in data science skills, tools and technologies
• Contact info for three references
NIH’s Notice of Interest in Diversity

ODSS particularly encourages applications from individuals from groups identified in NIH’s Notice of Interest in Diversity (NOT-OD-20-031) as underrepresented in the biomedical, clinical, behavioral, and social sciences, including women and:

- **Race/Ethnicity**
  - Blacks or African Americans
  - Hispanics or Latinos
  - American Indians or Alaska Natives
  - Native Hawaiians and other Pacific Islanders

- **Disability**
  - Physical or mental impairment that substantially limits one or more major life activities

- **Disadvantaged Background**
  - Homeless
  - Foster care system
  - First generation w/ Bachelor’s degree
  - Federal Pell Grants
  - Special Supplemental Nutrition Program
  - Rural or low income/access areas
DATA Scholar Program: 2023 Projects (11 ICOs)

- Application of Data Science to EHR Data from the All of Us Research Program
- Multi-modal NHLBI Data Integration Solutions: Linking Health and Environmental Data to Improve Patient and Community Health
- Using Large Language Models to Understand Immune Cell Communications
- Integrative Biological and Behavioral Health Disparities Data Scholar
- A Cost-Benefit Analysis of Persistent Identifiers Across the NIH-Funded Biomedical Research Sector
- Analyzing the Nation’s Over-the-Counter Testing Data
- Modernizing the NIDDK Biomedical Data Ecosystem to Enhance Translation of Big Data Science to Clinical Studies and Health Outcomes
- LitCoin Natural Language Processing Coalescence
- Integrating Multi-Dimensional Data to Promote Data-Driven Research in Oral Health and Oral Health Disparities
- Longitudinal Harmonization and Analysis of Alzheimer’s Disease and Related Dementias (AD/ADRD)
- Distributed Genomic Analysis Workflows and Services on NIH Cloud-Based Data Resources
- Enhancement of the FITBIR Data Science Platform Analysis Tools to Advance Traumatic Brain Injury (TBI) Research
- A Computational Framework to Identify Shared Molecular Etiology Among Rare Diseases Towards Drug Discovery
A cost-benefit analysis of persistent identifiers across the NIH-funded biomedical research sector

- Understand the potential benefits of persistent identifiers (PIDs) to strengthen infrastructure supporting metadata re-use, automation of data systems, and improved quality of disambiguated data elements
- Estimate the cost savings based on expected improved productivity of researchers and research support staff, against the costs incurred to integrate PIDs into funder and research organization data systems
- Make recommendations for an NIH-wide PID strategy that is aligned with existing recommendations made by the Subcommittee on Open Science, Office of Science & Technology Policy
A computational framework to identify shared molecular etiology among rare diseases towards drug discovery

- Develop a computational framework with integration of biomedical, genomics, and clinical data
- Identify shared molecular etiology among rare diseases to help inform drug repurposing and basket drug trial design
Analyzing the Nation’s Over-the-Counter Testing Data

• Perform a full analysis of the government’s OTC testing dataset, ultimately leading to a set of recommendations, opportunities, and limitations regarding use of self-test data for public health surveillance.

• Perform QA/QC on the data, assessing the degree to which OTC data can replace or augment the role of laboratory/POC testing data, characterizing human behaviors around self-test reporting, answering health equity questions around self-testing, and building predictive models.
Application of Data Science to EHR Data from the All of Us Research Program

• Develop standardized analytic and AI/ML tools to expand the analysis, use, and integration of electronic health record (EHR) data within the All of Us Researcher Workbench

• Provide training and technical assistance through sample code and documentation in using these data and tools among consortium staff and users of the Researcher Workbench and, by doing so, help democratize data analysis and usage to accelerate discovery using All of Us data
Distributed Genomic Analysis Workflows and Services on NIH Cloud-Based Data Resources

- Develop and implement interoperable, privacy-preserving and secure genomic analysis workflows and services that utilize the combination of controlled-access datasets from the NHGRI AnVIL and other NIH cloud-based data resources.

- Develop best practices learned from this effort and share them with other cloud-based resources that contribute to the NIH FAIR data ecosystem.
Enhancement of the FITBIR Data Science Platform Analysis Tools to Advance Traumatic Brain Injury (TBI) Research

• Increase transparency, scientific quality, and collaboration through improved access to Traumatic Brain Injury (TBI) data

• Enhance the system to better support AI ready data through the continued refinement of Common Data Elements (CDEs) by assigning Unified Medical Language System (UMLS) codes to each data element and its permissible values (i.e., value sets)

• Develop requirements and support the development to extend the Query GUI interface to efficiently use UMLS codes to develop cohorts
Integrating Multi-Dimensional Data to Promote Data-Driven Research in Oral Health and Oral Health Disparities

- Develop a framework and platform for mining, harmonization and integration, and analysis of research and clinical data, leading to insights about oral health and disparities in oral health.
- Demonstrate the value of data science approaches in the exploration and analysis of integrated health records. Methods and AI/ML-ready datasets generated from the project will be useful for biomedical research, health disparities research, and future methods development.
Integrating Multilevel Data to Understand Mechanisms of Health Disparities Data Scholar

- Develop a health disparities scan tool using data science approaches that can be used to mine extant research literature and large-scale data sets that will help determine gaps and new areas for research on common pathways and mechanisms leading to disease(s) and health advantages among populations who experience health disparities.

- Serve as a technical advisor to help guide and inform the community on the use of AI and data science techniques in health disparities research such as multi-omics, natural language processing, deep learning, artificial intelligence, and/or other ML approaches to help determine new lines of inquiry.
LitCoin Natural Language Processing Coalescence

• Combine and coalesce the successful submissions from the LitCoin Natural Language Processing Challenge into a single, powerful system for generating knowledge assertions from biomedical text

• Create a tool that can be used in several upcoming projects within the LitCoin program, including the generation of a foundational knowledge graph representing a huge corpus of knowledge in the Helping to End Addiction Long-term (HEAL) data ecosystem
Longitudinal Harmonization and Analysis of Alzheimer’s Disease and Related Dementias (AD/ADRD)

- Increase transparency, scientific quality and collaboration through improved public access to the aging and AD/ADRD cohort’s descriptive information
- Assist the research community in identifying and accessing population-based resources for research in aging and AD/ADRD
- Develop a pilot cohort database of four longitudinal studies, in the Biomedical Research Informatics Computing System (BRICS - https://brics.cit.nih.gov/) data science platform
Modernizing the NIDDK Biomedical Data Ecosystem to Enhance Translation of Big Data Science to Clinical Studies and Health Outcomes

- Design and develop a centralized metadata and meta-standard catalogue of NIDDK data science resources, including those for clinical, behavioral, social, and basic biomedical research, with a dynamic, searchable web interface
- Design and develop a platform to assist users adopting common data elements and metadata standards for both clinical and observational data types
- Design and develop a data science model to connect and harmonize clinical, observational, and basic science data types
Multi-modal NHLBI Data Integration Solutions: Linking Health and Environmental Data to Improve Patient and Community Health

• Assess the data sources and geo-spatial factors that are important for the NHLBI mission, create an inventory of NHLBI’s datasets with available longitudinal geocoded data (e.g., residential, occupational, etc.), and develop strategies for integrating environmental (both natural and manmade) data into the BDC ecosystem with NHLBI datasets

• Develop analytical procedures and tools, including ML, computing algorithms and statistical risk/disease prediction models, to study linked temporal-spatial relationships between NHLBI disease biomarker and health outcome data and climate and environmental data
Using Large Language Models to Understand Immune Cell Communications

- Apply large language models to understand how messaging works in the immune system and explore the possibility of applying advanced AI-based language models to understand immune cell communication.
- Develop understanding of how the immune system responds to foreign pathogens and allergens and protect its cells.
• **Is there a U.S. citizenship requirement?**
  Applicants must be U.S. citizens, non-citizen nationals, or Permanent Residents
  *Non-citizen national*: A person born in an outlying possession of the U.S. (e.g., American Samoa or Swain’s Island) on or after the date the U.S. acquired the possession, or a person whose parents are U.S. non-citizen nationals.

• **What are the degree requirements to apply as a DATA Scholar?**
  Applicants must have an M.D., Ph.D. or other doctoral degree and have advanced experience in data science or related fields.

• **Can I be involved in an active NIH grant as a DATA Scholar?**
  With concurrence of the matched ICO supervisor and ODSS, you maybe involved in an NIH grant during a DATA Scholar appointment but cannot be the contact PI. Details need to be worked out with the ICO Ethics Office.
• **Do you require full time commitment?**
  Full time commitment is preferred. Part-time commitment (with at least 60% time effort for the DATA Scholar appointment) is possible with concurrence of matched ICO and ODSS.

• **What is the term of a DATA Scholar appointment?**
  DATA Scholar appointments are for a duration of one to two years.

• **What if I want to stay longer? Can I apply to be a DATA Scholar again?**
  To stay longer at NIH, you must apply for open NIH positions. DATA Scholar appointments cannot be extended beyond 2 years. An individual is only allowed one-time participation as a DATA Scholar.
• **Do I work with more than one ICO during the DATA Scholar period?**
  You will be matched with one main IC during your DATA Scholar period. You might have opportunity to work with other ICOs through collaborations and workgroups.

• **Is there travel support for the scholar such as attending conferences?**
  Yes, ODSS will support $4,000 per year for the DATA Scholar for approved project related travel. The matched ICO may provide additional support.

• **Can a DATA Scholar work remotely or need to be in-person?**
  If selected, location of work will be determined by the supervising ICO with concurrence of ODSS. Applicants should be prepared to relocate to one of the NIH institutes, centers, or the Office of the Director, located in Bethesda and Rockville, MD, or Research Triangle Park, NC.
Important Dates

- **Website:** [https://datascience.nih.gov/data-scholars-2023](https://datascience.nih.gov/data-scholars-2023)
- **Application Deadline:** May 15, 2023, 11.59 pm EST
- **Anticipated Interview Time:** May – June
- **Anticipated Final Selection:** July, 2023
- **Anticipated Cohort Start Date:** September, 2023

**Contact:** Dr. Bryan Kim ([ds-workforce@nih.gov](mailto:ds-workforce@nih.gov))