ScHARe

Common Data Elements

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

ScHARe Core CDEs

NIH CDE Repository:
https://cde.nlm.nih.gov/home

NIH Endorsed

- Age
- Birthplace
- Zip Code
- Race and Ethnicity
- Sex
- Gender
- Sexual Orientation
- Marital Status
- Education
- Annual Household Income
- Household Size

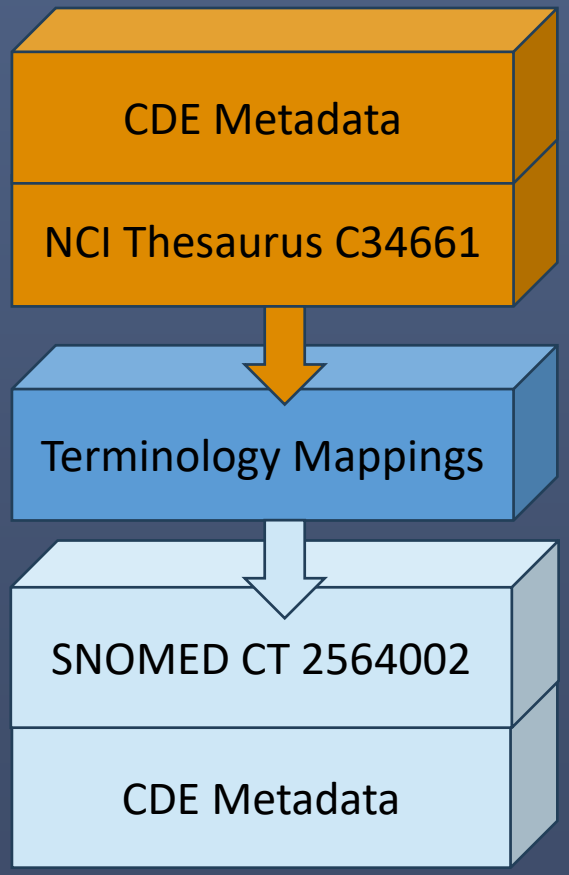
- English Proficiency
- Disabilities
- Health Insurance
- Employment Status
- Usual Place of Health Care
- Financial Security / Social Needs
- Self Reported Health
- Health Conditions (and Associated Medications/Treatments)

- **NIMHD Framework***
- **Health Disparity Outcomes***

* Project Level CDEs

For FUNDED PROJECT DATA – CDEs Centralized for Interoperability and Data Sharing

Importance of Concept Code Mapping and Data Interoperability (Uniform Resource Identifier - URI)



- CDE **unique CONCEPT CODES** represent data semantics
  - Human readable
  - Machine readable format
- **Mapping** enables interoperability even if the same standard terminology was not used in another CDE
- CDE Metadata enables searching for concept codes across CDEs to compare data

ScHARe

Understanding what is a Common Data Element (CDE)

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Precisely Defined-Shared Meaning/Understanding

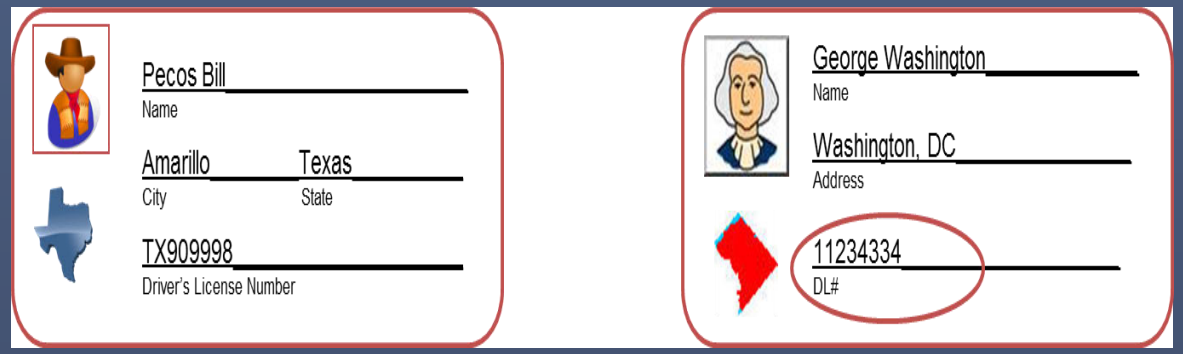
- Context is important in conveying meaning when using CDEs
  - Words have different meanings depending on words around it.
- Some examples:
    - **Agent:** chemical compound or government employee?
    - **Alcohol:** disinfecting or drinking?
    - **Colon:** sentence punctuation or biological organ?
    - **Mole:** animal, blemish, unit of measure, or spy?
    - **Probe:** examination, investigation, or instrument?

The above words are **SEMANTICALLY AMBIGUOUS**.

Words can mean different things in different contexts.

Importance of standardized meaning and concept codes for metadata

- Different formats for different states
  - WHAT we are describing (driver's license) and the
  - HOW – what data we need (name, city, state, driver's license number)



- Label or wording may be different→ but What is being collected *means* the same thing

What is metadata?

WHAT are we referring to? *A person's driver's license*

HOW we collect data, what data do we need?

Name
City and State
Driver's license number

The labels for the data are considered items of ***metadata.***

The data elements are Semantically Equivalent

Standardized Semantics/Concept Codes

- Describes the terms used for a given profession
- Provides consistency and clarity
- Often includes text definitions and synonyms
- Use Standard Terminologies for clear, shared meaning
- Ontologies provide context for a shared meaning

Ensures humans and computers attach the same meaning!

Semantics: Meaning of a word, phrase or sentence

A branch of linguistics and logic concerned with the meaning based on how a is sentence structured, including social and cultural context and relationships between words impact understanding.

Note: Ontologies are often expressed using formal semantics, which provides more precise meaning than other kinds of terminologies.

CDE semantics: (Human Readable) Expression of meaning in a **Standard, Structured** way using terminology concepts.

A Questionnaire to CDEs to Standardize Metadata

These requirements facilitate interoperability:

1. Standard Structure → predictable format
   (Makes CDEs easily used by humans and computers)

2. Standard Terminology → shared meaning
   (Makes the meaning of data clear and more easily reused)

3. Standard Codes → URI approach for machine readability
   (Automates data harmonization by using the codes instead of words)

Data Items that become CDEs are defined Independent of any System, Programming or Cultural language

ScHARe

How Do Common Data Elements Work?

Concept Codes and Mapping

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Survey Questions Become CDEs when…

Each are semantically defined by a standardized coding system for shared meaning

and

In a format that is human and machine readable for ease of reuse

CDEs Defined and Coded – URI Approach

Education

What is the highest level of education you have completed?

Shared Semantics and Concept Code:
An indication of the years of schooling completed in graded public, private, or parochial schools, and in colleges, universities, or professional schools. **C17953**

URI approach in data repository uses codes to harmonize data rather than semantics.

This improves data interoperability.

Answer machine readable format—excel spreadsheet: use of pipes to separate concepts & codes

Permissible Value (PV) Labels	PV Definitions	PV Concept Identifiers
No formal Schooling |	Indicates that a person has never attended an educational program or formal schooling.|	C67122 |
Primary/Grade/Elementary School (approximately grades 1st through 5th) |	Indicates that 5th grade potentially is the highest level of educational achievement.|	C67127 |
Middle School/Lower Secondary Education (approximately grades 6th through 8th) |	Indicates that 8th grade potentially is the highest level of educational achievem...	C67130 |

New OMB Categories

Self-Identification. Please select the racial category or categories with which you most closely identify (select all that apply).

☐ **American Indian or Alaska Native**
[write in] Enter, ie, Navajo Nation, Blackfeet Tribe of the Blackfeet Indian Reservation of Montana, Native Village of Barrow Inupiat Traditional Government, Nome Eskimo Community, Aztec, Maya, etc.

☐ **Asian or Asian American**
Chinese, Asian Indian, Filipino- Vietnamese, Korean- Japanese
[write in] Enter, ie, Pakistani, Hmong, Afghan, etc.

☐ **Black or African American**
African American, Jamaican, Haitian, Nigerian, Ethiopian, Somali
[write in] Enter, ie, Trinidadian and Tobagonian, Ghanaian, Congolese, etc.

☐ **Hispanic or Latino**
Mexican, Puerto Rican, Salvadoran, Cuban, Dominican, Guatemalan
[write in] Enter, ie, Colombian, Honduran, Spaniard, etc.

☐ **Native Hawaiian or Other Pacific Islander**
Native Hawaiian, Samoan, Chamorro, Tongan, Fijian, Marshallese
[write in] Enter, ie, Chuukese, Palauan, Tahitian, etc.

☐ **Middle Eastern or North African**
Lebanese, Iranian, Egyptian, Syrian, Iraqi, Israeli
[write in] Enter, ie, Moroccan, Yemeni, Kurdish, etc.

☐ **White**
English, German, Irish, Italian, Polish, Scottish
[write in] Enter, ie, French, Swedish, Norwegian, etc.

Survey Question to Become a CDE: The Journey

Please select the racial category or categories with which you most closely identify. *(select all that apply)*

- ☐ American Indian or Alaska Native
- ☐ Asian or Asian American
- ☐ Black or African American
- ☐ Hispanic or Latino
- ☐ Native Hawaiian or Other Pacific Islander
- ☐ Middle Eastern or North African (in current reporting tables will be reported as white)
- ☐ White

Start with a survey question in a study to be used by all

Making of a CDE from a Protocol/Question

Use a standardized source to define the concept (main words) and related code. Source: NCI Thesaurus

Race/Ethnicity Self-Identification (what racial category (ies) do you most identify?)

A textual description of a person's race. C17049 |The ethnicity of a person. C16564 | An individual's perspective or subjective interpretation of an event or information. C74528

Answers with pipes for human and machine readability

American Indian or Alaska Native |
Asian or Asian American |
Black of African American |
Hispanic, Latino, or Spanish |
Native Hawaiian or Other Pacific Islander |
Middle Eastern or North African |
White

Making of a CDE from a Protocol/Question

Defined and Coded

- A person having origins in any of the original peoples of North and South America (including Central America) and who maintains tribal affiliation or community attachment. (OMB) C41259 |
- A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent, including for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam. (OMB) C41260 |
- A person having origins in any of the Black racial groups of Africa. Terms such as "Haitian" or "Negro" can be used in addition to "Black or African American". (OMB) C16352 |
- A person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race. The term, "Spanish origin" can be used in addition to "Hispanic or Latino". (OMB) C17459 |
- A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands. (OMB) C41219 |
- Denotes a person having origins in the region of southwest Asia, between the India subcontinent and Europe, including Kuwait, Turkey, Lebanon, Israel, Iraq, Iran, Jordan, Saudi Arabia, lands east of Pakistan or the other countries of the Arabian Peninsula. Also includes people of Jewish ethnicity including Sephardic and Ashkenazic. C77820 :
- Denotes a person whose ancestry is in any of the countries of the northern part of the African continent: Algeria, Egypt, Libya, Morocco, Sudan, Tunisia, and Western Sahara. C126529 |
- A person having origins in any of the original peoples of Europe, the Middle East, or North Africa. (OMB) C41261

CDE Enhances Interoperability

Use codes from any source to harmonize different data sets and remove semantic ambiguity

Code Mapping

	NCIT	Loinc	UMLS CUI
American Indian or Alaska Native	C41259	LA10608-0	C0282204
Asian or Asian American	C41260	LA6156-9	C0003988
Black of African American	C16352	LA10610-6	C0085756
Hispanic, Latino, or Spanish	C17459	LA6214-6	C0086409
Native Hawaiian or Other Pacific Islander	C41219	LA10611-4	C1513907
Middle Eastern or North African	C43866	Mena no loinc	C1553353
White	C41261	LA4457-3	C0043157

Data Harmonization: matched CDEs from different projects

Core CDE Reported this way	Income (Project 1) Collected this way	Income (Project 2) Collected this way
Less than $10,000	Less than $10,000	Less than $10,000
$10,000-$24,999	$10,000-$24,999	$10,000-$24,999
$25,000-$34,999	$25,000-$34,999	$25,000-$34,999
$35,000-$49,999	$35,000-$49,999	$35,000-$49,999
$50,000-$74,999	$50,000-$74,999	$50,000-$74,999
$75,000-$99,999	$75,000-$99,999	$75,000-$99,999
$100,000-$149,999	$100,000-$149,999	$100,000-$149,999
$150,000-$199,999	$150,000-$199,999	$150,000-$199,999
$200,000 or more	$200,000 or more	$200,000 or more

Data Harmonization: Mappable CDEs

Mapped using algorithms

Income (Project 1)

Collected this way

Less than $24,999|
$25,000-$49,999 |
$50,000-$74,999 |
$75,000-$99,999 |
$100,000-$199,999|
$200,000 or more

Core CDE

Reported this way

Less than $10,000 |
$10,000-$24,999 |
$25,000-$34,999 |
$35,000-$49,999 |
$50,000-$74,999 |
$75,000-$99,999 |
$100,000-$149,999 |
$150,000-$199,999 |
$200,000 or more

Income (Project 2)

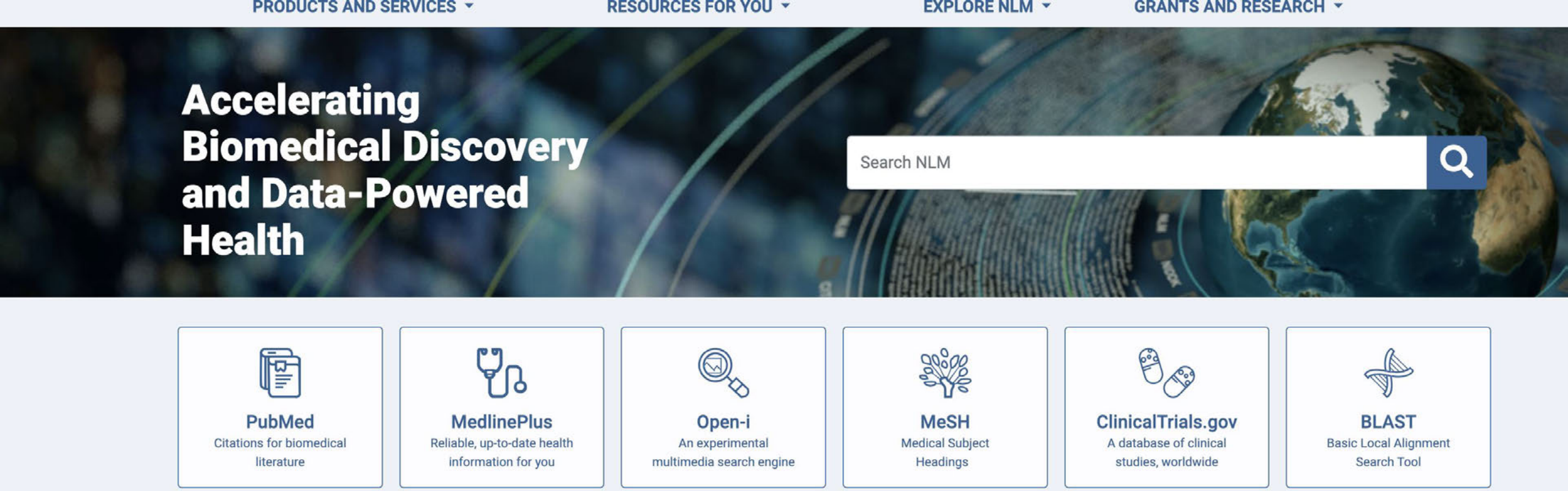
Collected this way

Less than $10,000 |
$10,000-$19,999 |
$20,000-$29,999 |
$30,000-$39,999 |
$40,000-$49,999 |
$50,000-$59,999 |
$60,000-$69,999 |
$70,000-$79,999 |
$80,000-$89,999 |
$90,000-$99,999 |
.....
$200,000 or more

ScHARe

CDE Terminology and Concept Coding

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

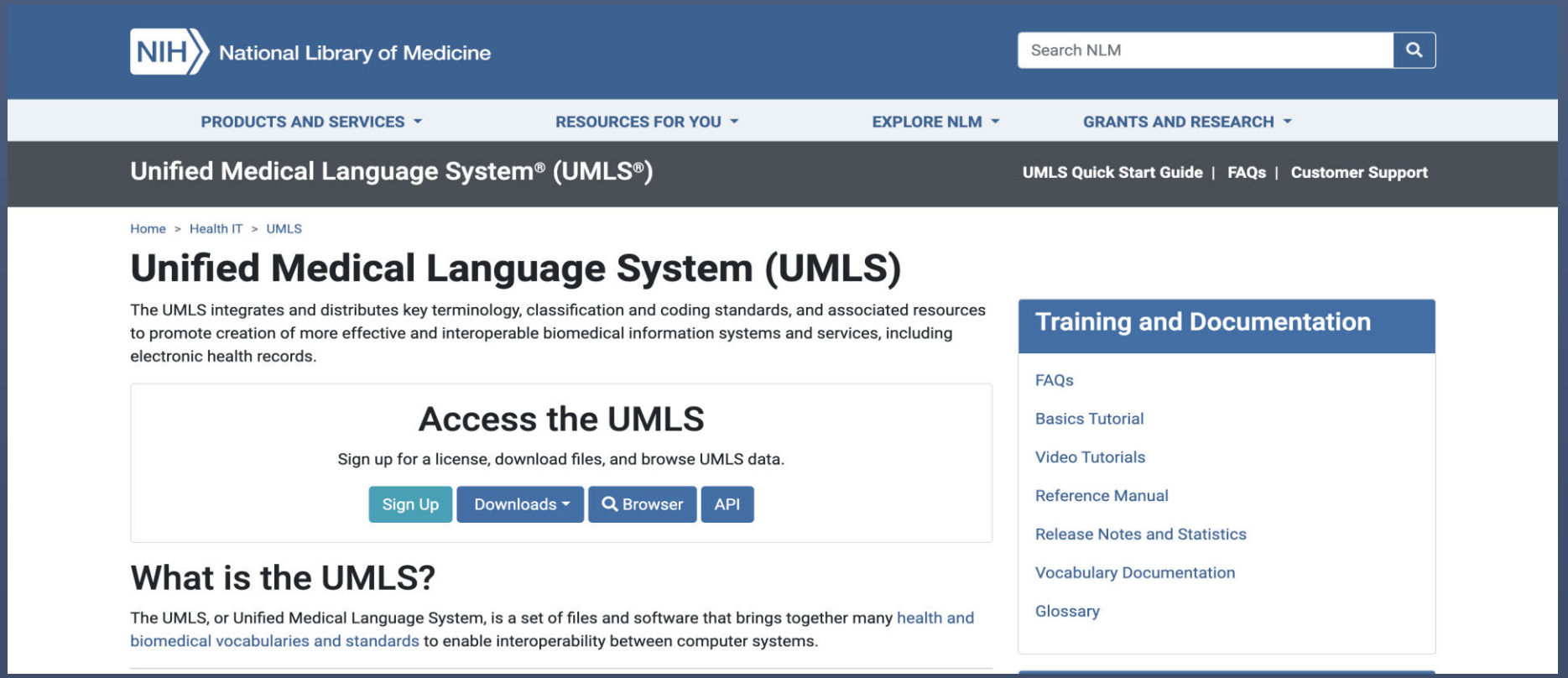


National Library of Medicine (NLM)

- "World's largest biomedical Library"
  - Public resource
- Familiar resources including PubMed, MeSH AND….

Unified Medical Language System (UMLS)

- Files and software that integrate and distribute health and biomedical terminologies and standards
- Database containing cross-terminology mappings (185 terminologies)
- **Assign Concept Unique Identifiers (CUIs) e.g. C0018681**



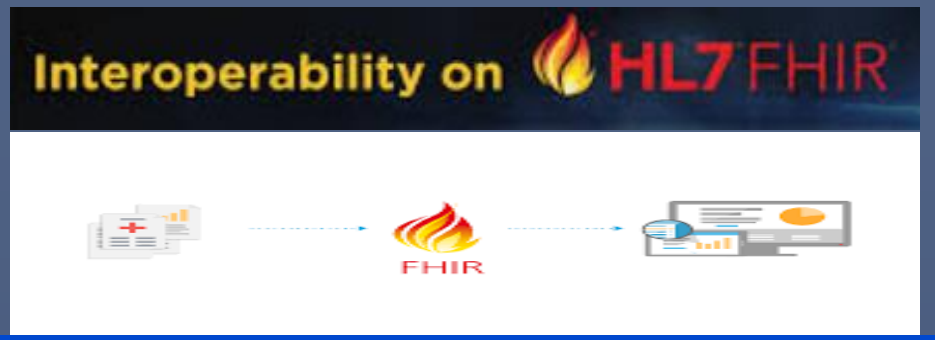
NLM UMLS: https://www.nlm.nih.gov/research/umls/index.html

LOINC (Logical Observation Identifiers Names and Codes)

- Establishes international standard for identifying health measurements, observations, and documents
- Provides a set of universal names and ID codes for identifying laboratory and clinical test results
- Facilitate the exchange and pooling of results for clinical care, outcomes management, and research

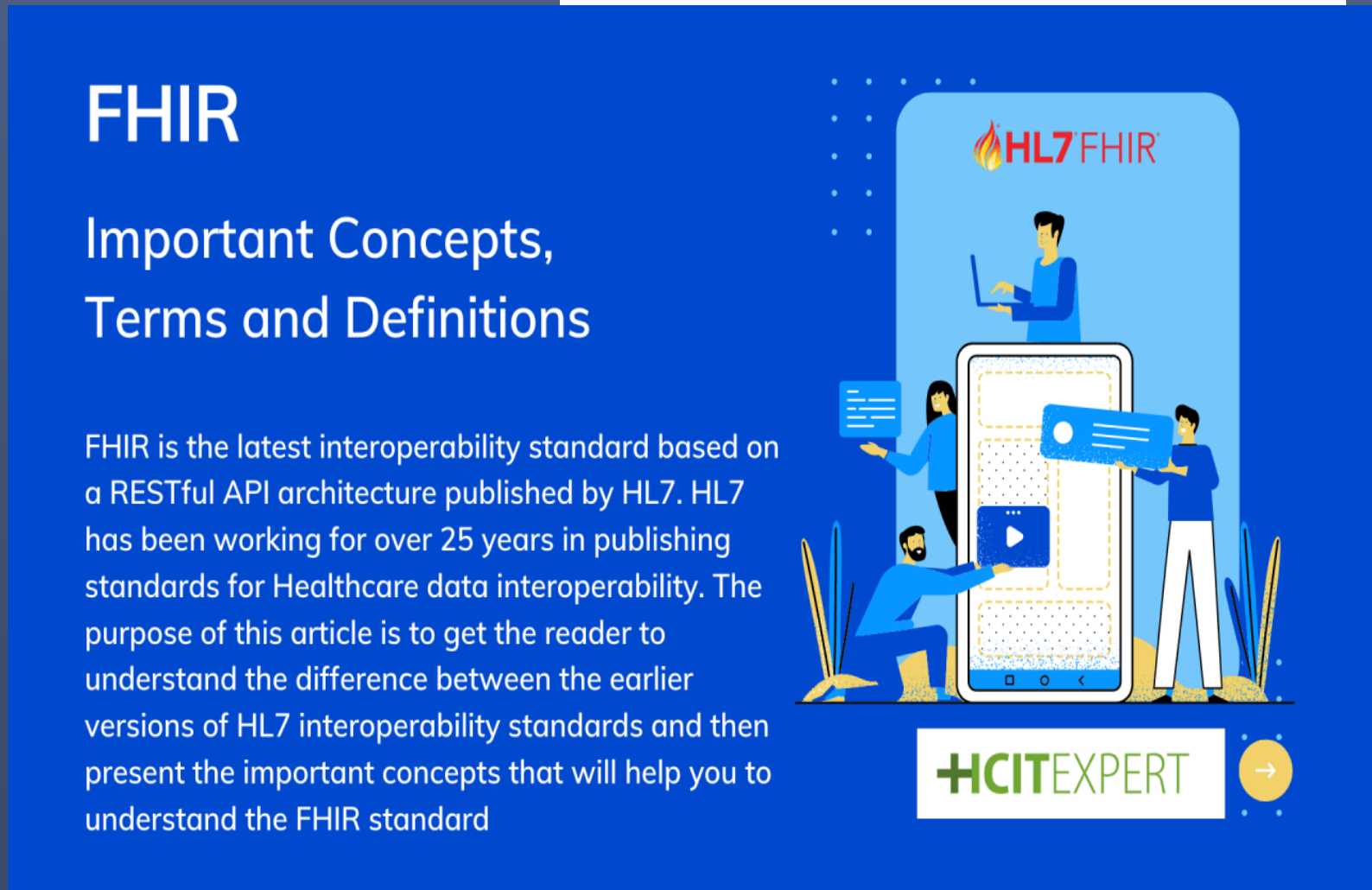
https://loinc.org/

FHIR HI7 (Fast Healthcare Interoperability Resource)



FHIR and HL7 are standards for exchanging electronic healthcare data, with FHIR offering enhanced security measures, mobile device compatibility, and simpler implementation. Can be leveraged to create innovative mobile health apps for patients and providers

HL7 is widely used for data management in hospital systems



NCI Enterprise Vocabulary Services (EVS)

- Terminology Services for NCI and other NIH Institutes/Centers (ICs)
  - Develop new concepts → unique identifiers and definitions
  - Record concept relationships → **scientific evidence**
  - Facilitates standardization across NCI and the larger biomedical community
- Two terminology products:
  - NCI Metathesaurus (NCIm)
  - NCI Thesaurus (NCIt)

Use of common terminologies are a key component of CDE Metadata.

Publicly Available

Standard Semantic & Concept Codes

- Unique concepts → unambiguous meaning
- Unique concept code → “C-code” (E.g. C16977)
- Linked to other coding systems via the UMLS CUI
- Additional concept relationships, definitions and code sources
- Concept Codes provide critical linkage to a specific meaning

Easily compared by computers to identify equivalent meaning regardless of the “words”

No Coding System is Better Than the Other

General use….

LOINC	Laboratory and Clinical Research
ULMS (CUI)	Biomedical
FHIR	Electronic Health Records
*NCIt	Cancer

*ScHARe used NCIt because it has several population concepts

ScHARe

NIH Clouds and Resources for ScHARe Collaborations

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

NIH Initiatives

NIH has launched a series of initiatives to:

- harness the power of cloud computing
- provide NIH biomedical researchers access to the most advanced, cost-effective computational infrastructure, tools and services

Examples include:

- **STRIDES** (Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability):
  - NIH partnered with commercial providers to streamline NIH data use leveraging cloud environments
  - Benefits include:
    - Professional services
    - Training
    - Discounts on STRIDES partner services
    - Potential collaborative engagements

NIH STRIDES
Accelerating biomedical research

cloud.nih.gov

Examples include:

- **AIM-AHEAD** (Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity):
  - Establish partnerships to increase the participation of underrepresented researchers in the development of AI/ML models using electronic health record (EHR) data

aim-ahead.net

- **BRIDGE2AI** (Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity):
  - Expand the use of AI in biomedical and behavioral research by generating “flagship” data sets and best practices for ML analysis

bridge2ai.org

Examples include:

- **All of Us**:
  - A historic effort to gather data from 1+ million people in the U.S. to build one of the most diverse health databases in history

allofus.nih.gov

- **AnVIL** (NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space):
  - Unified cloud environment for the analysis of genomic datasets

anvilproject.org

- **BioData Catalyst**:
  - Cloud-based platform for tools, applications, and workflows

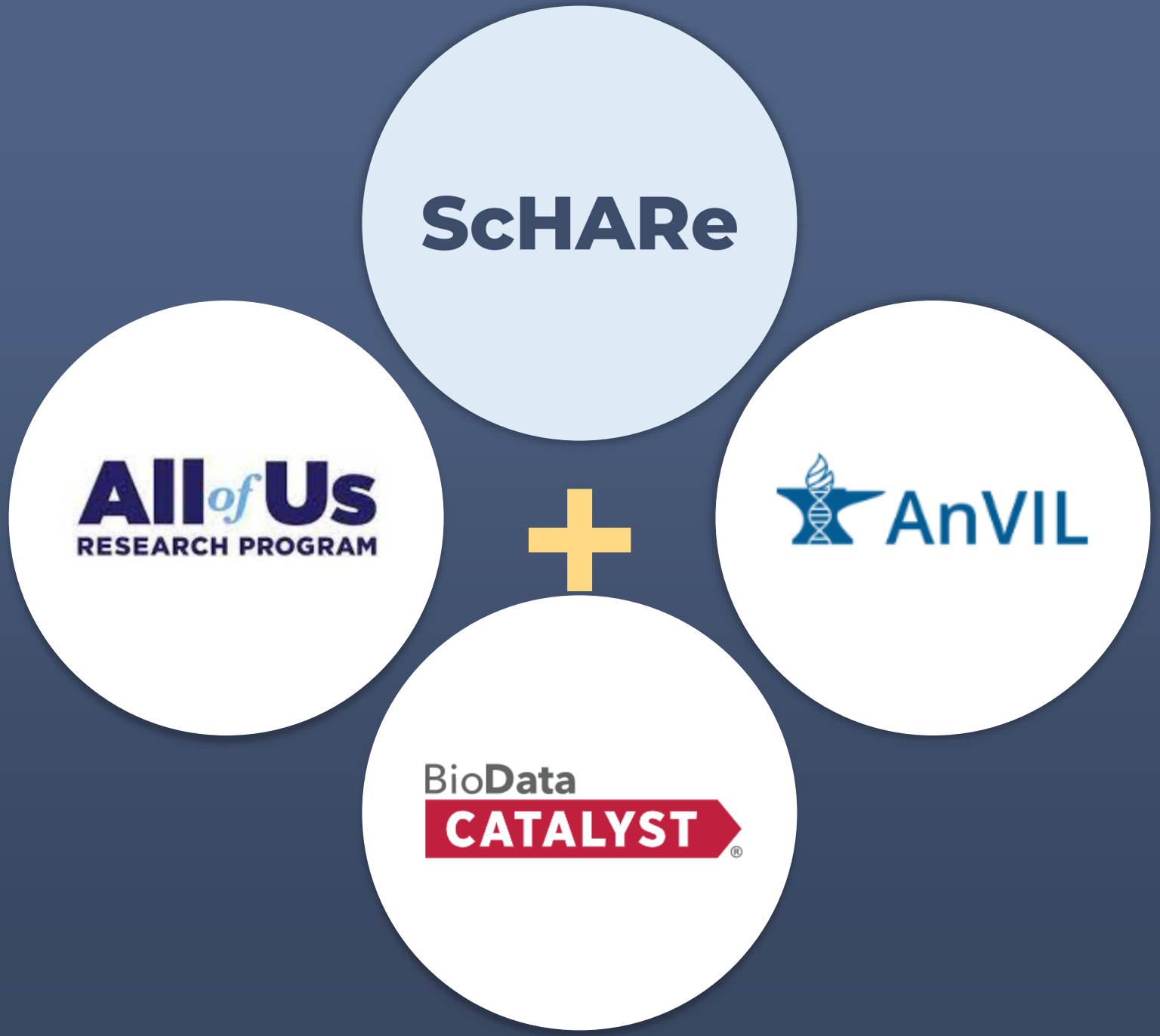
biodatacatalyst.nhlbi.nih.gov

Examples include:

- **All of Us**:
  - A historic effort to gather data from 1+ million people in the U.S. to build one of the most diverse health databases in history

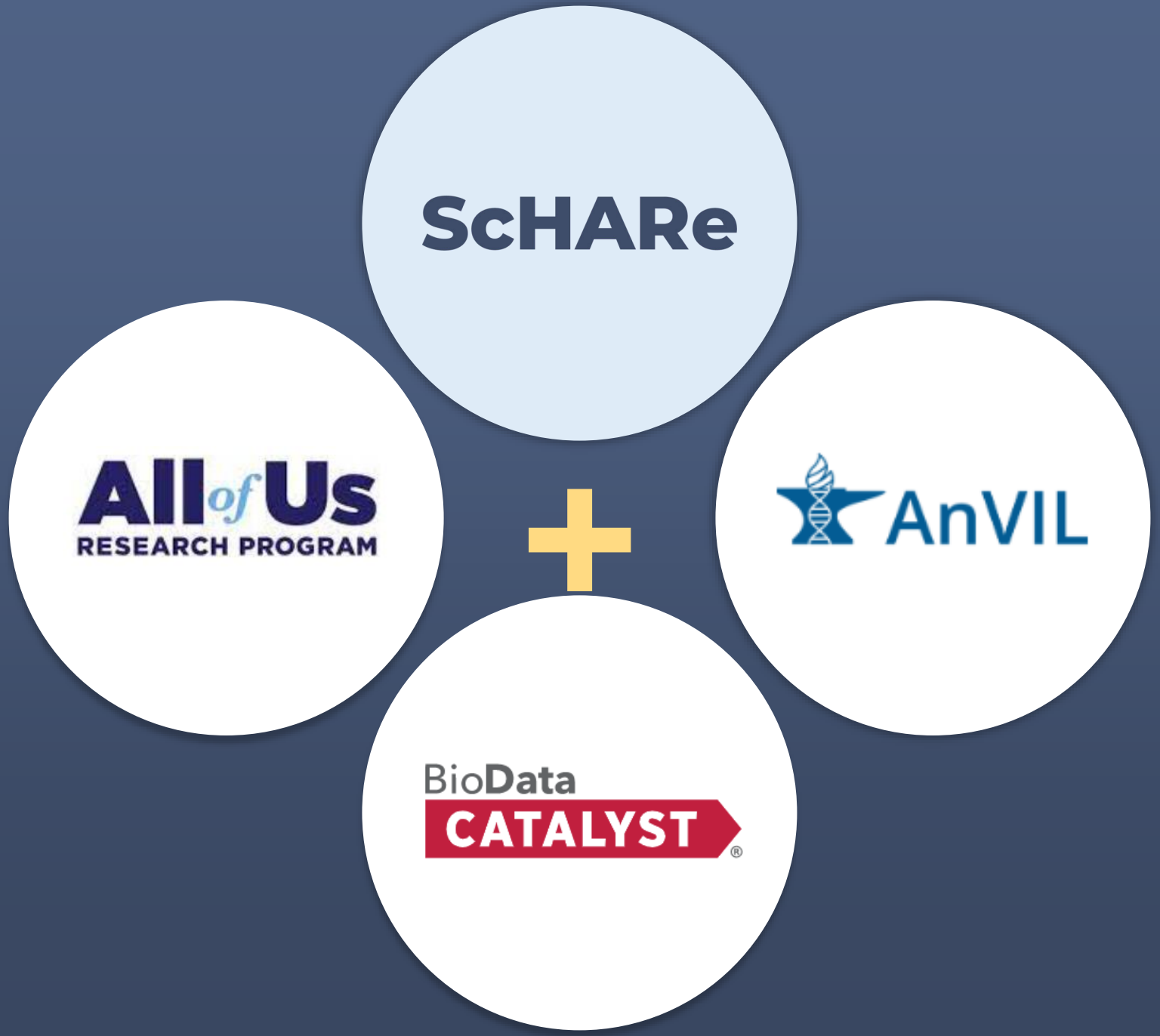
- **AnVIL** (NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space):
  - Unified cloud environment for the analysis of genomic datasets

- **BioData Catalyst**:
  - Cloud-based platform for tools, applications, and workflows



Terra powers all four cloud platforms

This creates an extraordinary opportunity for high-impact collaborations across platforms



Learning how to use Terra on ScHARe will open up a world of possibilities, giving you access to an interdisciplinary wealth of datasets and resources

ScHARe

Interested in a ScHARe Think-a-Thon Research Team?

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

ScHARe

Research Think-a-Thons

- Novice **training webinars** for data science, cloud computing & research using Big Data
- **Target:** underrepresented populations, women, racial/ethnic, sexual gener minorities, rural and poor populations

Generational Career & Discipline Exchange

Think-a-Thon Tutorials

bit.ly/think-a-thons

February	**Artificial Intelligence and Cloud Computing 101**
March	**ScHARe 1 – Accounts and Workspaces**
April	**ScHARe 2 – Terra Datasets**
May	**ScHARe 3 – Terra Google-hosted Datasets**
June	**ScHARe 4 – Terra ScHARe-hosted Datasets**
July	**An Introduction to Python for Data Science – Part 1**
August	**An Introduction to Python for Data Science – Part 2**
September	**ScHARe 5: A Review of the ScHARe Platform and Data Ecosystem**
October	**Preparing for AI 1: Common Data Elements and Data Aggregation**
November	**Preparing for AI 2: An Introduction to FAIR Data and AI-ready Datasets**
January	**Preparing for AI 3: Computational Data Science Strategies 101**
February/March	**Preparing for AI 4: Overview Prep for AI Summary with Transparency, Privacy, Ethics**
April	**Research Teams – SDoH and Health Disparities**
	ScHARe for Educators (Community Colleges & Low Resource MSIs)
	ScHARe for American Indian / Alaska Native Researchers
	ScHARe for Coders and Programmers to conduct Research

Think-a-Thon Research Teams

4 inaugural Research Teams to experience:

- Learn ethical AI strategies
- Utilize AI transparency practices
- Collaborate in a multigenerational, multidiscipline approach

Title: Data Science Projects 1 – Health Disparities and Individual SDoH

Description: Exploring the impact of individual Social Determinants of Health on health outcomes: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.

Title: Data Science Projects 2 - Health Disparities and Structural SDoH

Description: Assessing the impact of structural Social Determinants of Health on health outcomes: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.

Title: Data Science Projects 3 – Health Outcomes

Description: Investigating the influence of non-clinical factors on disparities in health care delivery: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.

April 2024

- **Foster a research paradigm shift to use Big Data**
- **Promote use of Dark Data**

- Multi-career (students to sr. investigators)
- Multi-discipline (data scientist & researchers)
- Feature Datasets with Guest Expert Leads
- Secure experts in topic area, analytics, data sources etc. to provide guidance
- Generate research idea - decide potential design, datasets & analytics
- Select co-leads to coordinate completion outside of TaT
- Publications

Register:

bit.ly/think-a-thons

Current Research Teams: Meetings and time commitment

3-4 months to complete the project in preparation for publication

The co-leads will **assign tasks** to the participants

Meetings other than Think-a-Thons to:

- review progress of tasks
- help/teach others what each participant is contributing
- assessing what else needs to be completed

Focus on the Social Determinants of Health

Social determinants of health (SDoH) are the **nonmedical factors that influence health outcomes**

They are the **conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life**

www.cdc.gov/about/sdoh/index.html

If certain communities have less access to education, jobs, fresh food or healthcare, they might face **more challenges in staying healthy** or may not have the same **opportunities to make healthy choices**

Experience conducting ethical AI

Transparency

Public perception and understanding of how AI works

- Technical documentation for duplication/re-use
- Tools:
  - Data dictionary
  - Health sheet (Data sheet)
  - Model cards (capabilities and purpose of algorithms are openly and clearly communicated to relevant stakeholders)

Fairness

***F**indable:* *providing metadata, documentation, and clear identifiers*
***A**ccessible:* *wide audience*
***I**nteroperable:* *standardized formats and APIs enable seamless integration*
***R**eusable:* *clear documentation, licensing, reduce redundancy*

- Metadata and data should be easy to find for both humans and computers
- Ensure that data represents relevant populations

Please join a Research Team

- **Team 1:** Gerido – Moore — R
- **Team 2:** Algarin – Hamner — Python
- **Team 3:** Kohan Ghadr – Zanwar — Python
- **Team 4:** Higginbotham – Yildirim — Statistics
- **Team 5:** Vidal - West — TBD

Team 1

R

Gerido – Moore

Examining multi-level factors associated with sex, gender, and sexuality-related cancer disparities

- Using geospatial and network analyses to characterize disparities in cancer **screening, prevention, diagnosis, and outcomes**
- Characterizing disparities in uptake of cancer prevention and screening behaviors across **intersectional identities**
- Characterizing **digital environmental factors** associated with cancer health disparities

Team 2

Python

Algarin – Hamner

Leveraging ScHARe data to evaluate health inequities at the community level

- County Health Rankings National Findings Report: Data to support community-led efforts to improve health equity
  Source: University of Wisconsin Population Health Institute

- U.S. Chronic Disease Indicators (CDI): 124 chronic disease indicators important to public health practice
  Source: Centers for Disease Control

Team 3

Python

Kohan Ghadr – Zanwar

Disparities in dental visits, oral health and costs among racial/ethnic populations

- Examine **oral health** (e.g. loss of teeth) disparities in access to dental visits and costs among racial/ethnic populations
- Investigate what **social and economic determinants of health** influence oral health disparities among racial/ethnic populations

Team 4

Statistics

Higginbotham – Yildirim

Identify geographical disparities in healthcare access by measuring the distances to various healthcare facilities, including hospitals with specialized care, health clinics, emergency departments, and ICUs

- Explore how different demographics (age, race, income level) face barriers to healthcare access to identify specific **vulnerable groups** that might be disproportionately affected by longer distances to necessary healthcare services
- Examine the **impact of distance** on healthcare service utilization and how access to primary care facilities influences the management of **chronic conditions and rates of preventable hospitalizations**

Team 5

Statistics

Vidal – West

Structural SDoH influence on stressors that foster the early onset of breast cancer and cardiovascular disease in women

- **Structural SDoH** - community, neighborhood policies and practices (environmental justice)
- Impacts on **family and individual stressors**
- Focus on breast cancer and cardiovascular disease in women
- Structural drivers foster the early onset of disease/disorders

ScHARe

join a team

Please fill out this form to indicate if you would like to participate in a research team

Note

The teams have limited capacity.
You may get reassigned to a different team.

ScHARe

Training Pipelines

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Think-a-Thons Training/Mentoring Pipeline

NLM
OIC Experts
Fellows

Think-a-Thons

- ✓ **Instructional**
- ✓ **Research**

N3C	AnVil	HEAL
Aim AHEAD		All of US
BioData Catalyst		

Using experts *to train and mentor novice users* *to increase diverse perspectives in biomedical research*

Goal: "Upskilling"

- ✓ Data science specialists into health disparities and health outcomes research
- ✓ Health Disparity/Outcomes researchers into using big data and cloud computing

Target Audience:

- ✓ Underrepresented populations (women, race/ethnic) users not trained in data science
- ✓ Data scientists with no or little research experience
- ✓ Resource & Tool for Community Colleges and Low Resource MSIs and Organizations

AIM AHEAD

Key Areas:

- Partnerships
- Research
- Infrastructure
- Data Science Training

Programs:

- Leadership Fellowship
- Research Fellowship
- Program for AI Readiness (PAIR)
- AIM-AHEAD Training Practicum (PRIME)
- Professional Development Program
- Federated Network

www.aim-ahead.net

Join AIM-AHEAD Connect

- AIM-AHEAD's community, networking, mentoring, and career development platform
- Virtual space to engage with the entire AIM-AHEAD Consortium and build community!
- Custom tools available to the AIM-AHEAD Coordinating Center:
  - Connect with experts, learners, stakeholders, etc.
  - Mentoring, Q&A, video calls, groups, funding & jobs board, etc.
  - SignUp: Event registration & information solicitation
  - Surveys: Request feedback on various activities
  - HelpDesk: Respond to topic-specific questions
  - Programs: Collaborative space, exclusive content, and mentor matching

Scan QR Code to Join AIM-AHEAD Connect

Program Information Updates

Year 3 Program Webpage

https://www.aim-ahead.net/call-for-proposals-year-3/

ScHARe

Thank you

Evaluation poll

1. **Rate how useful this session was:**

- ☐ Very useful
- ☐ Useful
- ☐ Somewhat useful
- ☐ Not at all useful

Evaluation poll

2. Rate the pace of the instruction for yourself:

- ☐ Too fast
- ☐ Adequate for me
- ☐ Too slow

Evaluation poll

3. How likely will you participate in the next Think-a-Thon?

- ☐ Very interested, will definitely attend
- ☐ Interested, likely will attend
- ☐ Interested, but not available
- ☐ Not interested in attending any others

ScHARe

Next Think-a-Thons:

bit.ly/think-a-thons

Register for ScHARe:

bit.ly/join-schare

schare@mail.nih.gov