



ScHARe

Research Think-a-Thons



National Institutes of Health

The logo for ScHARe, featuring the text "ScHARe" in a bold, dark blue sans-serif font, centered within a white circle.

Be a Part of the Future of Knowledge Generation

May 15, 2024

Deborah Duran, PhD • NIMHD

Luca Calzoni, MD MS PhD Cand. • NIMHD



Look deeper with more eyes

“For the first time in history, we have a technology (AI) that is opening our eyes to who we are, is changing us as we speak, and could allow us to play a conscious role in who we want to become.”

Jennifer Aue

IBM Director for AI Transformation
AI professor at the University of Texas

- Diverse perspectives
- Bias mitigation strategies
- Research paradigm shift to Big Data



ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

Outline

- 10'** Introduction
 - Experience poll
 - Interest poll
- 20'** AI and Cloud Computing 101
- 15'** What is ScHARe?
- 15'** Health Disparities, Health Care Delivery, Health Outcomes
- 15'** Python and R
- 40'** Common Data Elements
- 10'** Research Think-a-Thon Expectations
- 10'** NIH Clouds and Resources for ScHARe Collaborations
- 10'** Join a Research Team
- 5'** Training Pipelines
 - Evaluation poll

Experience poll

Please check your level of experience with the following:

	None	Some	Proficient	Expert
Python	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
R	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Cloud computing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Terra	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Health disparities research	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Health outcomes research	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Algorithmic bias mitigation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Interest poll

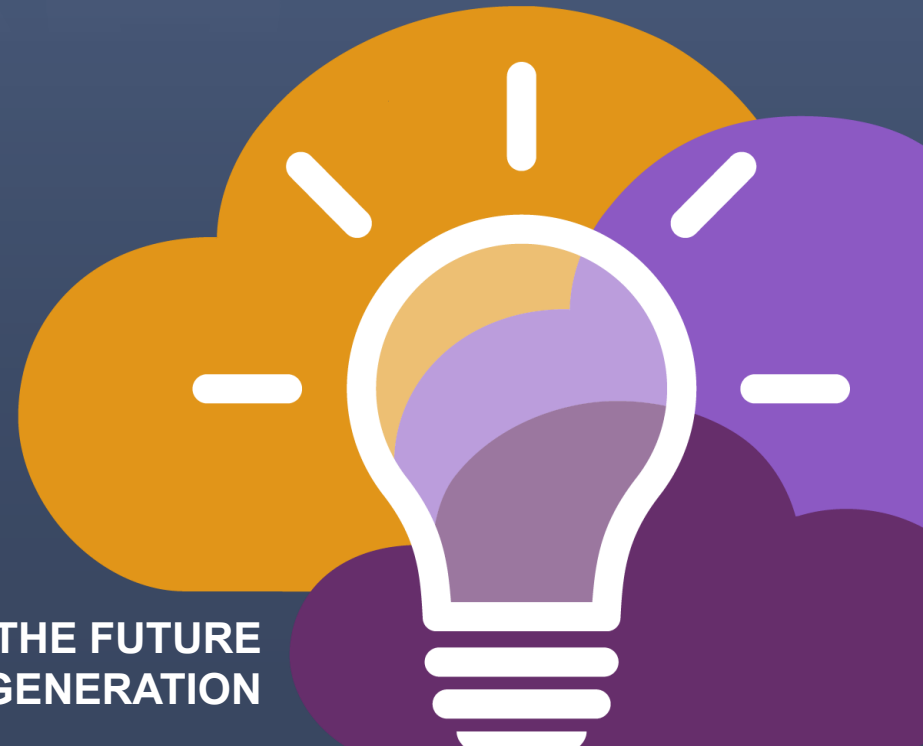
I am interested in (check all that apply):

- ☐ Learning about Health Disparities and Health Outcomes research to apply my data science skills
- ☐ Conducting my own research using AI/cloud computing and publishing papers
- ☐ Connecting with new collaborators to conduct research using AI/cloud computing and publish papers
- ☐ Learning to use AI tools and cloud computing to gain new skills for research using Big Data
- ☐ Learning cloud computing resources to implement my own cloud
- ☐ Developing bias mitigation and ethical AI strategies
- ☐ Other

ScHARe

AI and Cloud Computing 101

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



Big Data

Extremely large data sets that are statistically analyzed to gain detailed insights, often **using AI** and substantial **computer-processing power**.

Datasets are sometimes **linked together (Data Integration)** to see how patterns in one domain affect other areas.

Data Integrity (data quality) is the overarching completeness, accuracy, consistency, accessibility, and security of the data for its intended purpose. This should always be assessed before using a dataset.

FAIR data are data which meet machine-actionability principles of:

- Findability
- Accessibility
- Interoperability
- Reusability



ScHARe

The ScHARe Data Ecosystem will offer access to **250+ datasets**, including:

- American Community Survey
- U.S. Census
- Social Vulnerability Index
- Food Access Research Atlas
- Medical Expenditure Panel Survey
- National Environmental Public Health Tracking Network
- Behavioral Risk Factor Surveillance System

Big Data: structured and unstructured data

Structured data is quantitative data that is organized and easily searchable

Some tools used to work with structured data include:

- OLAP
- MySQL
- PostgreSQL
- Oracle Database



Unstructured data is every other type of data that is not structured.

Some tools used to manage unstructured data include:

- MongoDB
- Hadoop
- Azure



	Structured data	Unstructured data
Main characteristics	Searchable Usually text format Quantitative	Difficult to search Many data formats Qualitative
Storage	Relational databases Data warehouses	Data lakes Non-relational databases Data warehouses NoSQL databases Applications
Used for	Inventory control CRM systems ERP systems	Presentation or word processing software Tools for viewing or editing media
Examples	Dates, phone numbers, bank account numbers, product SKUs	Emails, songs, videos, photos, reports, presentations

Data mining

Techniques that **analyze large amounts of information to gain insights**, spot trends, or uncover patterns.

Data mining helps:

- organizations improve their processes
- researchers identify associations to answer **novel research questions**.

It **involves more use of algorithms** (software-based coding programs - especially machine learning), than traditional statistics.



ScHARe

ScHARe aims to enable a **research paradigm shift** to leverage Big Data and AI tools to develop **more innovative research projects**

Cloud computing

Data storage and processing **used to take place on personal computers or local servers.**

In recent years, **storage and processing have migrated to digital servers** operated by internet platforms.

People can store and process data remotely.

Cloud computing offers **convenience, reliability, and the ability to scale applications** quickly.

Main public cloud service providers:

- **Google**
- Azure
- AWS



ScHARe

Computing environments can be **customized or standardized** (using a custom Docker Image or a startup script) on ScHARe, to make sure everyone in your group is using the **same software in your analyses**

Google Cloud Platform

GCP is a **provider of computing resources** for developing, deploying, and operating applications on the Web.

It provides management tools and modular cloud services, including:

- **computing**
- **data storage**
- **data analytics**
- **machine learning**

GCP is the platform **used for ScHARe**.



ScHARe

Through Google, ScHARe offers:

- **Big query** and **Tensorflow** access for advanced machine learning
- Access to Google Cloud Public Datasets
- **\$300/user in free credits** to cover computing costs

Google email address needed.

Artificial Intelligence (AI)

AI is defined as:

*“machines that respond to stimulation **consistent with traditional responses from humans**, given the human capacity for contemplation, judgment, and intention.”*

This definition emphasizes several qualities that separate AI from traditional computer software:

- **Intentionality**
- **Intelligence**
- **Adaptability**

AI-based computer systems **can learn from data, text, or images and make intentional and intelligent decisions** based on that analysis.



ScHARe

Many AI projects are built using Python.

ScHARe fully supports the **Python libraries** most commonly used for AI tasks.

Machine Learning (ML)

ML is “based on **algorithms that can learn from data** without relying on rules-based programming.”

It represents **a way to classify data/objects without detailed instruction.**

The algorithm learns in the process so that new objects can be identified using the learned info.

Language Learning Models (LLM)

- Computational model notable for its ability to achieve **general-purpose language generation and other natural language processing tasks** such as classification, or learning statistical relationships from text documents.
- Used for copywriting, knowledge base answering, text classification, code generation, text generation.
- Chat GPT acquires knowledge about syntax, semantics and "ontology" inherent in human language corpora, but **also inaccuracies and biases** present in the corpora, including cultural and temporal biases in language and semantics.

Neural networks

Researchers use software to “**perform some task by analyzing training examples**”.

Similar to the neural nodes of a brain, **neural networks learn in layers and build complex concepts** out of simpler ones.

Deep learning and many recent applications of ML use neural networks (e.g., driverless cars, genomics, drug development).

Deep Learning

Deep learning employs **statistics to spot underlying trends, correlations or data patterns** and applies that knowledge to other layers of analysis.

It's a way to “learn by example”.

It requires extensive computing power and labeled data.

Artificial Intelligence Bias

Algorithms are widely used in healthcare- and policy-related decisions. However, many operate as “**black boxes**”, offering little opportunity for testing to identify biases.

Biases can result from:

- **social/cultural context not considered**
- **design limitations**
- **data missingness and quality problems**
- **algorithm development and model training**

If not identified, biased algorithms may result in decisions that lead to discrimination, unequitable healthcare, and/or health disparities.

Trust in AI

Caution Against:

- **Epistemic Trust**, which describes the willingness to accept new information from another person or entity (e.g., synthetic data) as trustworthy, generalizable, and relevant.
- **Synthetic Trust**, a misplaced belief in the model's capabilities and fairness.
 - Synthetic data models could be making predictions based on a narrow set of experiences, which may not be generalizable to the wider population they are meant to serve, which leads to unintended harms and perpetuates health disparities.

Mistrust of AI

- Fear of misuses
- Fear because of harmful impacts of biases
- Lack of underrepresented populations/community trust

Ethical AI

It is crucial that **AI algorithms respect basic human values** and undertake their analysis and decision-making in a trustworthy manner.

Ethical AI builds tools that are faithful to values such as **accountability, privacy, safety, security, and transparency**.

Taken together with explainable AI, it is a way to **deploy AI in ways that further human values**.

Explainable AI (XAI)

One of the complaints about AI is the **lack of transparency** in how it operates. Many developers don't reveal the data used or how various factors are weighted. Outsiders cannot tell how AI reached the decision that it did.

This lack of explainability can lead people to **not trust AI**.

XAI seeks to help **describe either the overall function of AI or the specific way it reaches decisions**, to make AI more understandable and trustworthy.

Synthetic / AI Generated DATA

- Information that is **artificially generated** rather than produced by real-world events.
- Typically created using **algorithms**, synthetic data can be deployed to **validate mathematical models** and to **train machine learning models**
- Generated to meet **specific needs or certain conditions that may not be found in the original, real data**
- Often used for **underrepresented populations** in datasets

Digital Twins

Digital model of an intended or actual real-world physical product, system, or process (a physical twin) that serves as the **effectively indistinguishable digital counterpart** of it for practical purposes, such as simulation, integration, testing, monitoring & maintenance

Digital twin of a person, based on such computer simulations, could help drug developers design, test and monitor, and aid doctors in applying, the **safest and most effective treatments or therapies** that are specific and tailored to our genetics or biochemistry.

**Not the answers to
poor quality or missing data**

Model Autophagy Disorder (MAD)

- Occurs when a **model collapses or “eats itself”** after being repeatedly trained on AI generated data
- In model training the quality (precision) or diversity (recall) of the generated data progressively decrease over successive generations.
- MAD results when there is **not enough fresh data in self-consuming generative models**, leading to a degradation in the quality and diversity of future training loops, as the model forgets the true data distribution over time.

Cloning data

Data that was cloned or synthetically AI generated results in the data that contains the same data flaws and existing patterns as the original data, regardless of data accuracy or representativeness.

Dolly, the cloned sheep, created more Dollys – with no genetic diversity.

**Not the answers to
poor quality or missing data**

AI and cloud computing: benefits and challenges



AI and cloud computing are **revolutionary and beneficial technologies** transforming research and accelerating science progress.



However, they pose various **risks and challenges**.



Benefits

Access to big datasets and large data ecosystems:

- Today, the scientific community confronts a data landscape that more expansive and more varied. The cloud offers access to **vast repositories** of scientific data, and enables **efficient mapping and linking** across data sources



SchARE

The SchARE Data Ecosystem will offer access to **300+ datasets**, including:

- Public Datasets hosted by SchARE and Google
- Funded Datasets on SchARE, in compliance with the **NIH Data Sharing Policy**



Benefits

Deeper insights and better decision making:

- AI in the cloud, linked with **machine learning (ML)** and **data mining** resources, can identify trends in **large datasets** with **quicker and more accurate results, facilitating decision-making** in clinical and policy applications



ScHARe

Terra, standalone or in conjunction with Google Cloud Platform's Vertex AI, **can support your ML-based analyses**

Tutorials will show you how to do large-scale training and model serving



Benefits

Intelligent automation and data management:

- AI can deal with massive amounts of data in a **programmed manner** to analyze them properly without human intervention
- AI can **automate repetitive tasks** and help manage and monitor workflows



ScHARe

Workflows (pipelines) are steps performed by a compute engine for bulk analysis.

ScHARe uses workflows in Workflow Description Language (**WDL**), a language easy for humans to read, for batch processing data.

For novice users, integration with **SAS** is planned.



Benefits

Real-time online collaboration:

- Cloud technology enables **truly collaborative work**, allowing researchers and institutions to break down silos and **connecting people across different disciplines**, multiple functions and from far-away locations.



ScHARe

ScHARe enables researchers to create interactive **Jupyter notebooks** (documents that contain live code) and **share data, analyses and results with their collaborators** in real time.



Benefits

Increased security:

- With sensitive data hosted in the cloud, data security is crucial.
- AI-powered network security tools track network traffic and can immediately detect anomalies and block them



ScHARe

ScHARe provides researchers with **secure workspaces** that they can share with their collaborators.

The ScHARe platform is secured according to best practices in information security (the Terra system has been granted Authority to Operate as a **FISMA Moderate** impact system and is **FedRAMP** authorized).



Benefits

Lower costs:

- Restrictive **upfront costs** related to on-site data centers, such as hardware and maintenance, are eliminated
- **Staff costs** are reduced, as AI tools can gain insights from the data with little human intervention



ScHARe

ScHARe leverages **low-cost** and **open-source** components:

- Terra Platform
 - GitHub
 - Open-source tools/libraries
- to keep platform costs at a minimum



Challenges



ScHARe

Lack of knowledge and expertise:

- Research institutions are finding it tough to find and hire the right cloud talent. There is a **shortage of professionals** with the required qualifications, especially among **populations with health disparities**.
- **Many researchers lack the required skills** and knowledge to use AI and cloud computing.

Step-by-step guides, tutorials, and training materials help novice ScHARe users accomplish their research goals and **upskill their careers** by acquiring hands-on AI and cloud computing knowledge



Challenges



ScHARe

Data privacy and security - or misperceptions therein:

- Research institutions use a lot of sensitive information that can be targeted for data breaches by hackers. Hence, they need to create **privacy policies and secure all data** when using AI in the cloud
- **Not all Cloud providers can assure 100% data privacy.** Cloud misconfiguration, data misuse, lack of control tools and poor identity access management can cause privacy leaks

The Terra platform powering ScHARe uses best practices and industry standards, mostly aligned to NIST-800-53 Rev 4 Moderate, to achieve compliance with industry-accepted **security and privacy frameworks.** Future **single sign-on** using RAS.



Challenges

Performance, reliability and availability:

- The performance of cloud computing solutions **depends on the vendors** who offer these services
- If a cloud vendor is affected by reliability and availability issues, so are the organizations using their services



ScHARe

Through Terra, ScHARe partners with a Cloud Service Provider that has real-time **monitoring** policies.

Terra also implements the **NIST Framework** standards in cloud environments.



Challenges

AI bias:

- Widespread use of AI raises a number of **ethical, moral, and legal issues** that are yet to be addressed
- **AI biases** are found in training data, as well as in the algorithm design and implementation phases. They shape healthcare decisions and can result in health disparities.
- **Populations with health disparities are underrepresented** in data science



ScHARe

Critical thinking can identify, if not eliminate, AI biases.

ScHARe was created to:

- foster participation of **populations with health disparities** in data science
- promote the collaborative identification of **bias mitigation strategies**
- create a **culture of ethical inquiry** whenever AI is utilized



We want to hear from you

What **challenges** are **you** experiencing or anticipating in adopting AI/cloud computing?

Cost · Knowledge · Research applications · Other

ScHARe meets challenges of cloud computing adoption

Utility:

- Many centralized social science & SDoH datasets
- Data Sharing requirement compliance
- Secure confidential workspaces
- Workbooks with instructions & code
- Link across data sets & platforms
- SAS

Costs:

- Capitalizes free & low-cost tools
- Google credits
- Download data to personal computer when cloud unnecessary

Collaborations:

- Multi-career level / multi-discipline research & bias mitigation teams
- Dark data use
- Publications
- Upskilling Jr & Sr underrepresented data science & health investigators

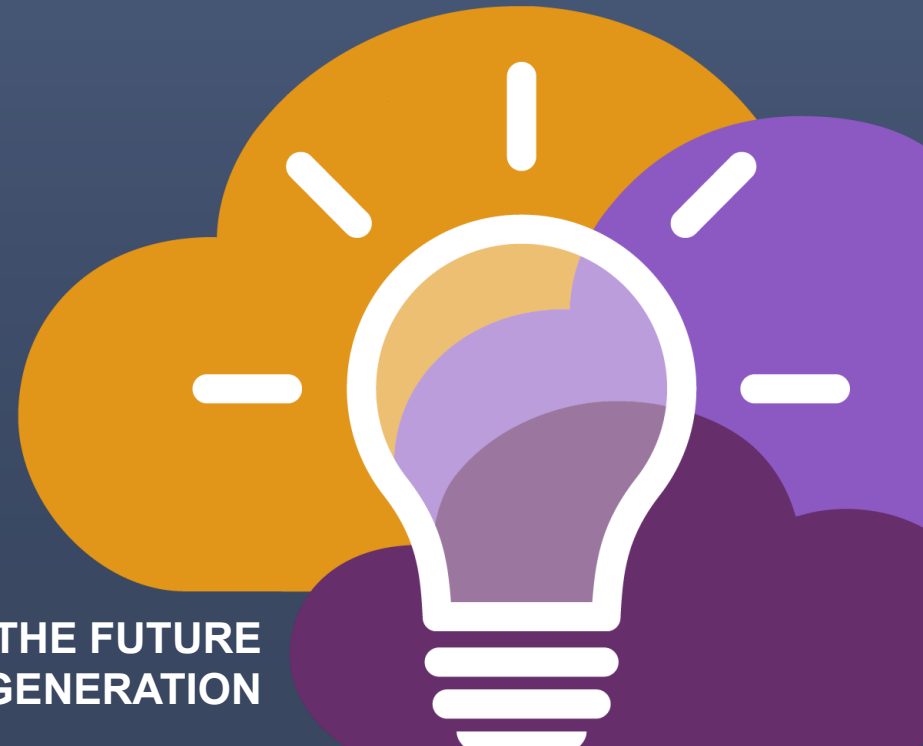
Knowledge:

- Think-a-Thons
- Cloud computing platforms
- Cloud computing resources
- Jargon & Terminology
- Python / R

ScHARe

What is ScHARe?

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



ScHARe is a **cloud-based population science data platform** designed to accelerate research in health disparities, health and healthcare delivery outcomes, and artificial intelligence (AI) bias mitigation strategies

ScHARe aims to fill **four critical gaps**:

- Increase participation of **women & underrepresented populations with health disparities** in data science through data science skills training, cross-discipline mentoring, and multi-career level collaborating on research
- Leverage population science, SDoH, and behavioral Big Data and cloud computing tools to foster a **paradigm shift** in healthy disparity, and health and healthcare delivery outcomes research
- **Advance AI bias mitigation and ethical inquiry** by developing innovative strategies and securing diverse perspectives
- Provide a **data science cloud computing resource** for community colleges and low resource minority serving institutions and organizations

ScHARe



nimhd.nih.gov/schare



ScHARe



Google Platform Terra Interface

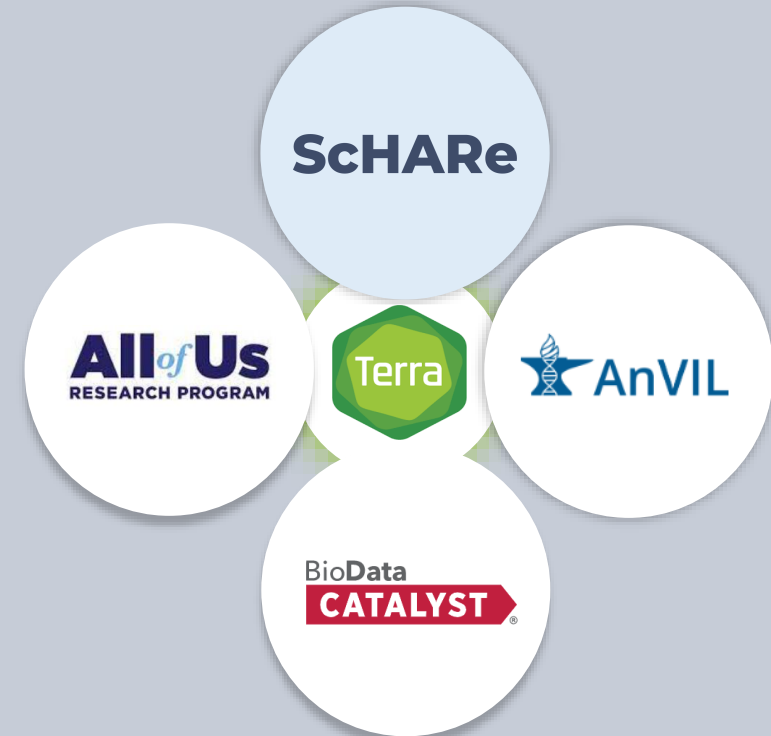
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Cut-and-paste code in Python and R



Terra recommends using **Chrome**
Must have a **Gmail** friendly account

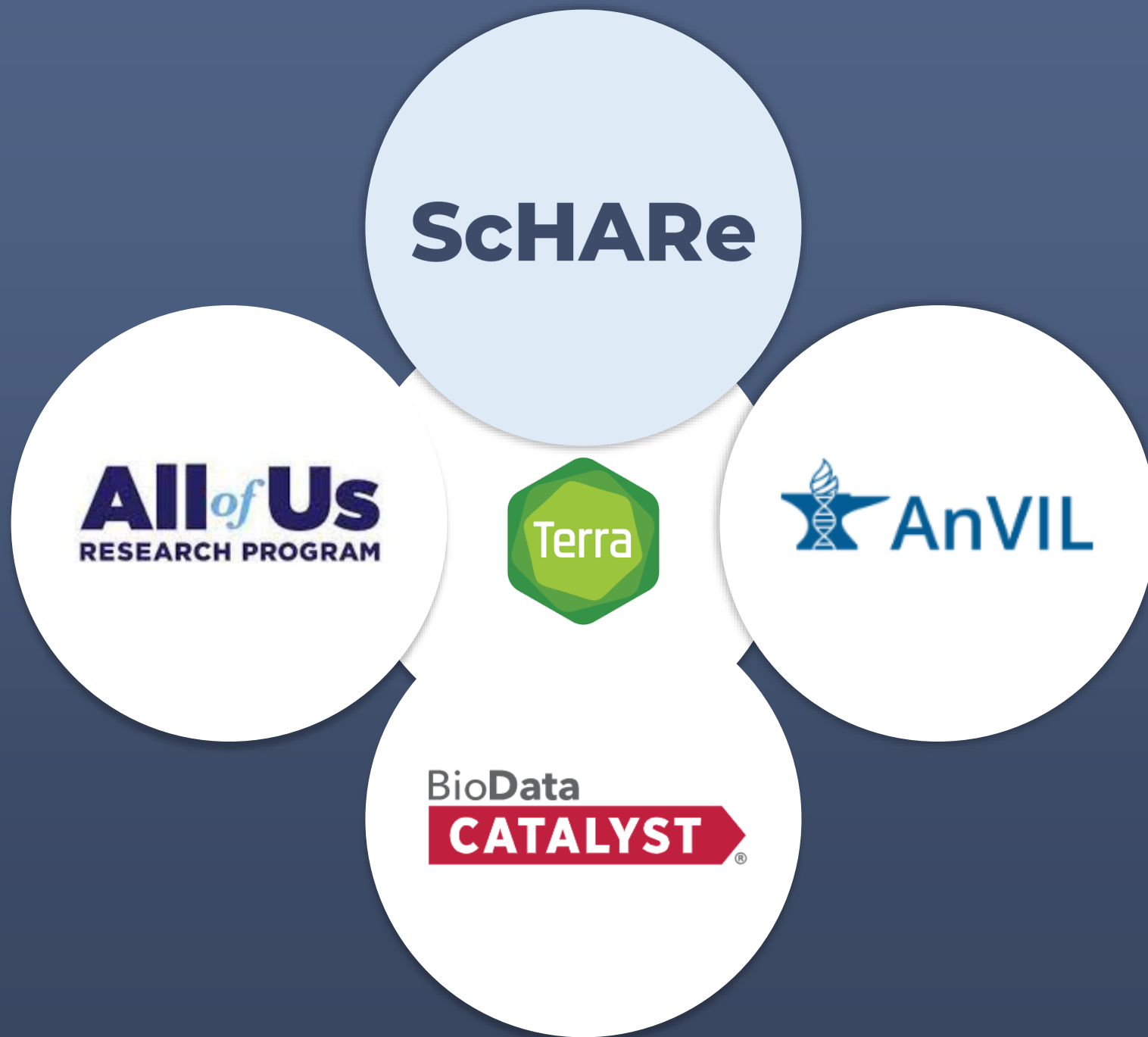
PREPARING FOR AI RESEARCH AND HEALTHCARE USING BIG DATA

Mapping across cloud platforms
with Terra Interface



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION





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 Ecosystem structure

250+
FEDERATED
PUBLIC
DATASETS

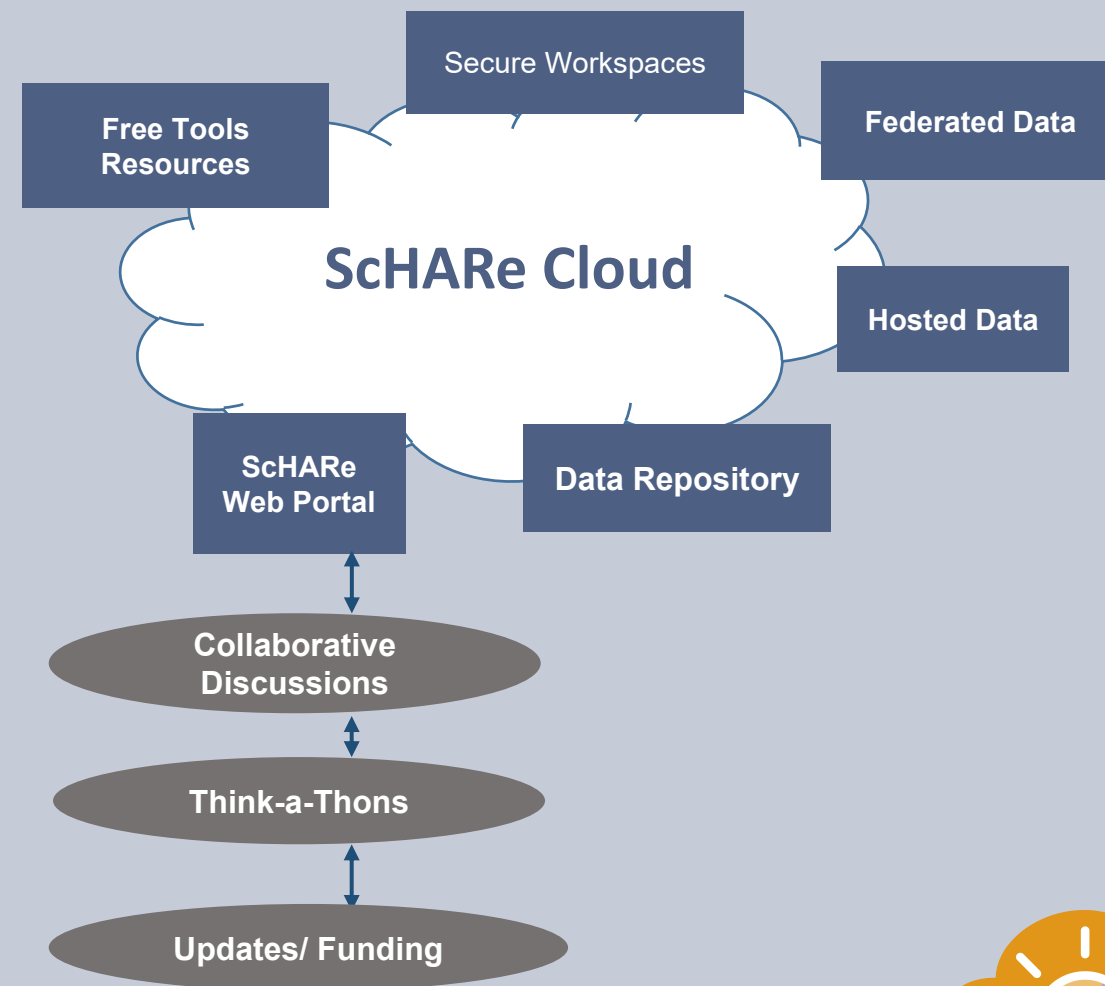
CDE
FOCUSED
REPOSITORY

- **Population Science / SDoH / Behavioral Data**
- Hosted by Google & ScHARe
- **CDEs enhance data interoperability** (aggregation) by using semantic standards and concept codes

Innovative Approach: CDE Concept Codes
Uniform Resource Identifier (**URI**)

COMPONENTS

Intramural and Extramural Resource



ScHARe Ecosystem

Researchers can access, link, analyze, and export a **wealth of SDoH and population science related datasets** within and across platforms relevant to research about health disparities, health care delivery, health outcomes and bias mitigation, including:

1

Public datasets

Publicly accessible, federated, de-identified datasets hosted by ScHARe or hosted by Google through the Google Cloud Public Dataset Program

ScHARe

e.g.: *Behavioral Risk Factor Surveillance System (BRFSS)*

Google

e.g.: *American Community Survey (ACS)*

2

Funded datasets

Publicly accessible and controlled-access, funded program/project datasets using Common Data Elements shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy

e.g.: *Jackson Heart Study (JHS)*

Extramural Grant Data

Intramural Project Data



ScHARe Ecosystem

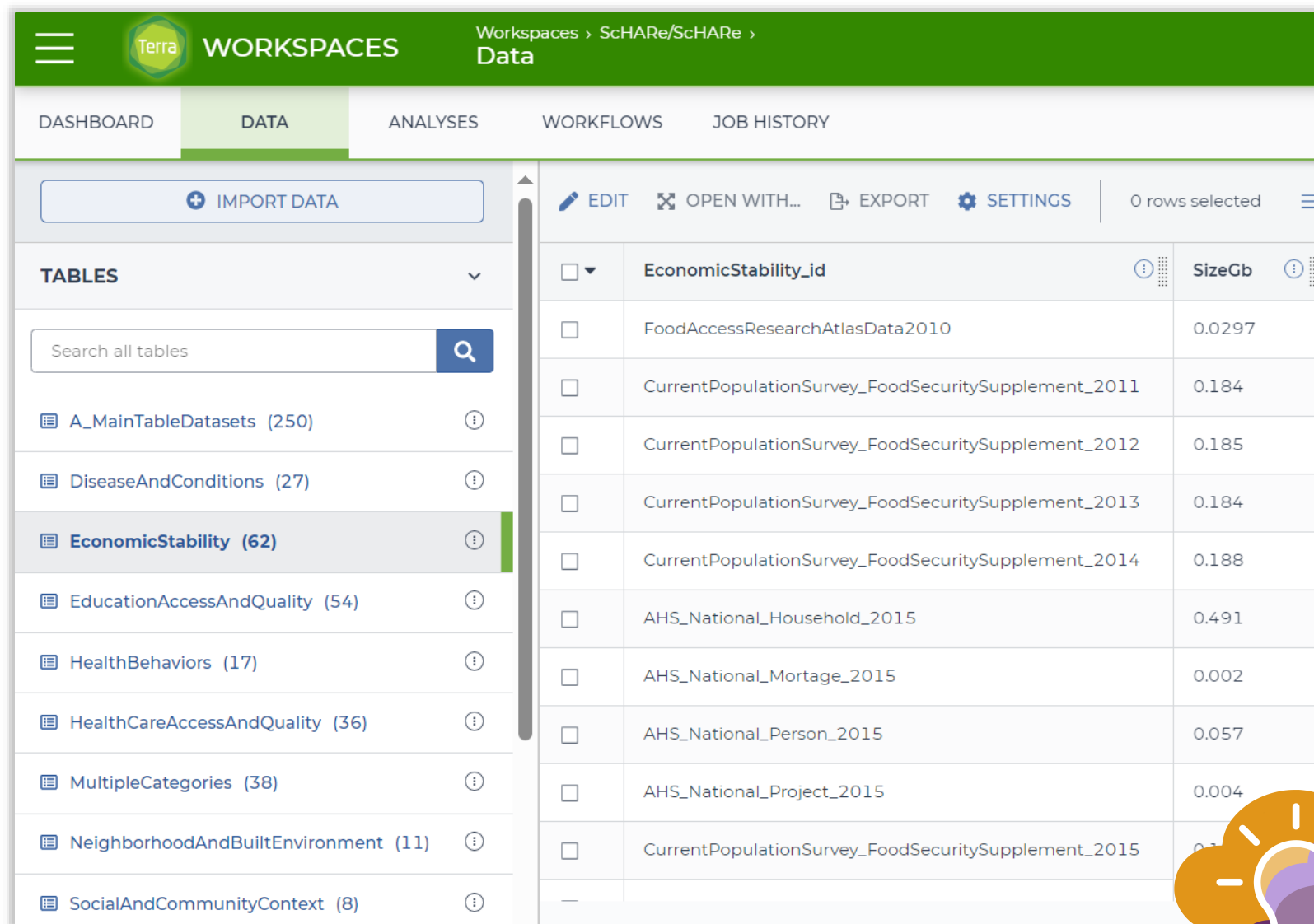
OVER 250 DATA SETS CENTRALIZED

Datasets are categorized by content based on the CDC **Social Determinants of Health** categories:

1. Economic Stability
2. Education Access and Quality
3. Health Care Access and Quality
4. Neighborhood and Built Environment
5. Social and Community Context

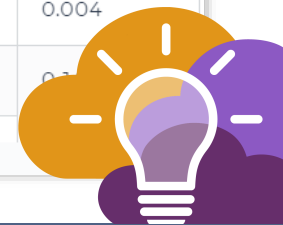
with the addition of:

- **Health Behaviors**
- **Diseases and Conditions**



The screenshot shows the Terra WORKSPACES Data interface. The top navigation bar includes a menu icon, the Terra logo, and the text 'WORKSPACES' and 'Data'. Below this is a sub-navigation bar with tabs: DASHBOARD, DATA (selected), ANALYSES, WORKFLOWS, and JOB HISTORY. The main content area is divided into two panels. The left panel, titled 'TABLES', contains a search bar and a list of datasets. The right panel displays a table of datasets with columns for selection, name, and size.

	EconomicStability_id	SizeGb
<input type="checkbox"/>	FoodAccessResearchAtlasData2010	0.0297
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2011	0.184
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2012	0.185
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2013	0.184
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2014	0.188
<input type="checkbox"/>	AHS_National_Household_2015	0.491
<input type="checkbox"/>	AHS_National_Mortgage_2015	0.002
<input type="checkbox"/>	AHS_National_Person_2015	0.057
<input type="checkbox"/>	AHS_National_Project_2015	0.004
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2015	0.184



ScHARe Ecosystem: ScHARe hosted datasets

Organized based on the **CDC SDoH categories**, with the addition of *Health Behaviors* and *Diseases and Conditions*:

250+ datasets

- What are 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

ScHARe Ecosystem: ScHARe hosted datasets

Examples of datasets for each category include:

Education access and quality

Data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.

Examples:

- **EDFacts Data Files** (U.S. Dept. of Education) - Graduation rates and participation/proficiency assessment
- **NHES - National Household Education Surveys Program** (U.S. Dept. of Education) – Educational activities

ScHARe Ecosystem: ScHARe hosted datasets

Health care access and quality

Data on health literacy, use of health IT, emergency room waiting times, preventive healthcare, health screenings, treatment of substance use disorders, family planning services, access to a primary care provider and high quality care, access to telehealth and electronic exchange of health information, access to health insurance, adequate oral care, adequate prenatal care, STD prevention measures, etc.

Example:

- **MEPS - Medical Expenditure Panel Survey** (AHRQ) - Cost and use of healthcare and health insurance coverage
- **Dartmouth Atlas Data** - Selected Primary Care Access and Quality Measures - Measures of primary care utilization, quality of care for diabetes, mammography, leg amputation and preventable hospitalizations

ScHARe Ecosystem: ScHARe hosted datasets

Neighborhood and built environment

Data on access to broadband internet, access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, deaths from motor vehicle crashes, asthma and COPD cases and hospitalizations, noise exposure, smoking, mass transit use, etc.

Examples:

- **National Environmental Public Health Tracking Network** (CDC) - Environmental indicators and health, exposure, and hazard data
- **LATCH - Local Area Transportation Characteristics for Households** (U.S. Dept. of Transportation) – Local transportation characteristics for households

ScHARe Ecosystem: ScHARe hosted datasets

Social and community context

Data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc.

Example:

- **Hate crime statistics** (FBI) - Data on crimes motivated by bias against race, gender identity, religion, disability, sexual orientation, or ethnicity
- **General Social Survey** (GSS) - Data on a wide range of characteristics, attitudes, and behaviors of Americans.

ScHARe Ecosystem: ScHARe hosted datasets

Economic stability

Data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.

Examples:

- **Current Population Survey (CPS) Annual Social and Economic Supplement** (U.S. Bureau of Labor Statistics) - Labor force statistics: annual work activity, income, health insurance, and health
- **Food Access Research Atlas** (U.S. Dept. of Agriculture) – Food access indicators for low-income and other census tracts

ScHARe Ecosystem: ScHARe hosted datasets

Health behaviors

Data on health-related practices that can directly affect health outcomes.

Examples:

- **BRFSS - Behavioral Risk Factor Surveillance System** (CDC) - State-level data on health-related risk behaviors, chronic health conditions, and use of preventive services
- **YRBSS - Youth Risk Behavior Surveillance System** (CDC) – Health behaviors that contribute to the leading causes of death, disability, and social problems among youth and adults

ScHARe Ecosystem: ScHARe hosted datasets

Diseases and conditions

Data on incidence and prevalence of specific diseases and health conditions.

Examples:

- **U.S. CDI - Chronic Disease Indicators** (CDC) - 124 chronic disease indicators important to public health practice
- **UNOS - United Network of Organ Sharing** (Health Resources and Services Administration) – Organ transplantation: cadaveric and living donor characteristics, survival rates, waiting lists and organ disposition

ScHARe Ecosystem: Google hosted datasets

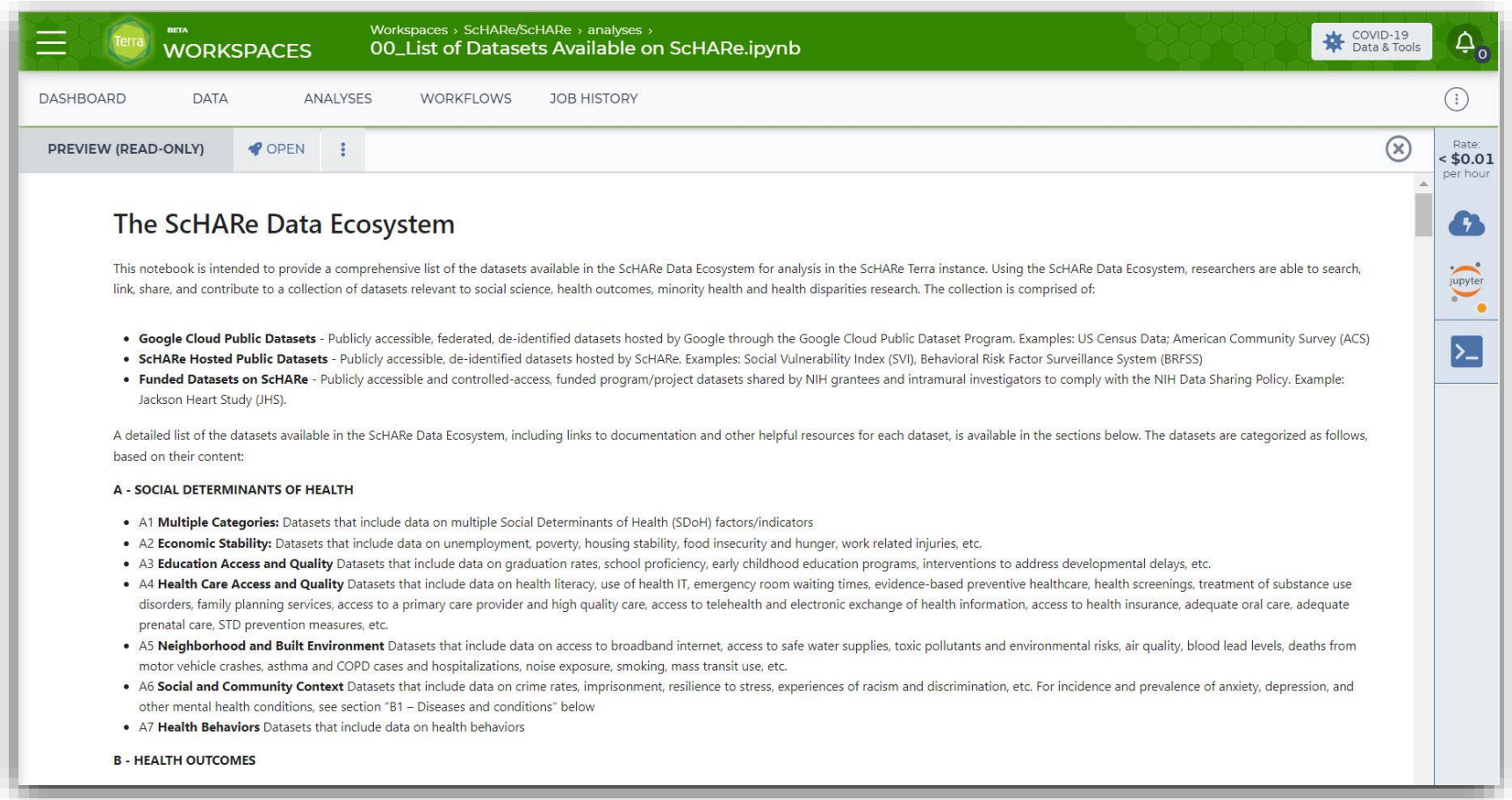
Examples of interesting datasets include:

- **American Community Survey** (U.S. Census Bureau)
- **US Census Data** (U.S. Census Bureau)
- **Area Deprivation Index** (BroadStreet)
- **GDP and Income by County** (Bureau of Economic Analysis)
- **US Inflation and Unemployment** (U.S. Bureau of Labor Statistics)
- **Quarterly Census of Employment and Wages** (U.S. Bureau of Labor Statistics)
- **Point-in-Time Homelessness Count** (U.S. Dept. of Housing and Urban Development)
- **Low Income Housing Tax Credit Program** (U.S. Dept. of Housing and Urban Development)
- **US Residential Real Estate Data** (House Canary)
- **Center for Medicare and Medicaid Services - Dual Enrollment** (U.S. Dept. of Health & Human Services)
- **Medicare** (U.S. Dept. of Health & Human Services)
- **Health Professional Shortage Areas** (U.S. Dept. of Health & Human Services)
- **CDC Births Data Summary** (Centers for Disease Control)
- **COVID-19 Data Repository by CSSE at JHU** (Johns Hopkins University)
- **COVID-19 Mobility Impact** (Geotab)
- **COVID-19 Open Data** (Google BigQuery Public Datasets Program)
- **COVID-19 Vaccination Access** (Google BigQuery Public Datasets Program)

How to check what data is available on ScHARe

1. Analyses tab

In the **Analyses** tab in the ScHARe workspace, the notebook **00_List of Datasets Available on ScHARe** lists all of the datasets available in the ScHARe Datasets collection



The screenshot displays the Terra WORKSPACES interface. The top navigation bar includes a menu icon, the Terra logo, and the text "BETA WORKSPACES". The breadcrumb trail shows "Workspaces > ScHARe/ScHARe > analyses > 00_List of Datasets Available on ScHARe.ipynb". The main navigation tabs are DASHBOARD, DATA, ANALYSES (selected), WORKFLOWS, and JOB HISTORY. Below the tabs, there are buttons for "PREVIEW (READ-ONLY)" and "OPEN". The notebook content is titled "The ScHARe Data Ecosystem" and includes a description of the data ecosystem and a list of dataset categories.

The ScHARe Data Ecosystem

This notebook is intended to provide a comprehensive list of the datasets available in the ScHARe Data Ecosystem for analysis in the ScHARe Terra instance. Using the ScHARe Data Ecosystem, researchers are able to search, link, share, and contribute to a collection of datasets relevant to social science, health outcomes, minority health and health disparities research. The collection is comprised of:

- **Google Cloud Public Datasets** - Publicly accessible, federated, de-identified datasets hosted by Google through the Google Cloud Public Dataset Program. Examples: US Census Data; American Community Survey (ACS)
- **ScHARe Hosted Public Datasets** - Publicly accessible, de-identified datasets hosted by ScHARe. Examples: Social Vulnerability Index (SVI), Behavioral Risk Factor Surveillance System (BRFSS)
- **Funded Datasets on ScHARe** - Publicly accessible and controlled-access, funded program/project datasets shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy. Example: Jackson Heart Study (JHS).

A detailed list of the datasets available in the ScHARe Data Ecosystem, including links to documentation and other helpful resources for each dataset, is available in the sections below. The datasets are categorized as follows, based on their content:

A - SOCIAL DETERMINANTS OF HEALTH

- **A1 Multiple Categories:** Datasets that include data on multiple Social Determinants of Health (SDoH) factors/indicators
- **A2 Economic Stability:** Datasets that include data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.
- **A3 Education Access and Quality** Datasets that include data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.
- **A4 Health Care Access and Quality** Datasets that include data on health literacy, use of health IT, emergency room waiting times, evidence-based preventive healthcare, health screenings, treatment of substance use disorders, family planning services, access to a primary care provider and high quality care, access to telehealth and electronic exchange of health information, access to health insurance, adequate oral care, adequate prenatal care, STD prevention measures, etc.
- **A5 Neighborhood and Built Environment** Datasets that include data on access to broadband internet, access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, deaths from motor vehicle crashes, asthma and COPD cases and hospitalizations, noise exposure, smoking, mass transit use, etc.
- **A6 Social and Community Context** Datasets that include data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc. For incidence and prevalence of anxiety, depression, and other mental health conditions, see section "B1 - Diseases and conditions" below
- **A7 Health Behaviors** Datasets that include data on health behaviors

B - HEALTH OUTCOMES

How to access ScHARe hosted datasets

Data tab

In the **Data** tab in the ScHARe workspace, **data tables help access ScHARe data and keep track of your project data:**

- In the ScHARe workspace, click on the Data tab
- Under Tables, you will see a list of dataset categories
- If you click on a category, you will see a list of relevant datasets
- Scroll to the right to learn more about each dataset

The screenshot shows the Terra WORKSPACES interface, specifically the 'Data' tab. The top navigation bar includes 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The 'DATA' tab is active. On the left, under 'TABLES', there is a search bar and a list of dataset categories. The 'EconomicStability' category is selected, showing 62 datasets. The main panel displays a table of datasets with columns for checkboxes, dataset names, and sizes in GB. The table is currently empty, showing 0 rows selected.

	EconomicStability_id	SizeGb
<input type="checkbox"/>	FoodAccessResearchAtlasData2010	0.0297
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2011	0.184
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2012	0.185
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2013	0.184
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2014	0.188
<input type="checkbox"/>	AHS_National_Household_2015	0.491
<input type="checkbox"/>	AHS_National_Mortgage_2015	0.002
<input type="checkbox"/>	AHS_National_Person_2015	0.057
<input type="checkbox"/>	AHS_National_Project_2015	0.004
<input type="checkbox"/>	CurrentPopulationSurvey_FoodSecuritySupplement_2015	0.185

How to access Google hosted datasets

Big Query

The Google public datasets are **available for access on Terra using BigQuery**

- BigQuery is the Google Cloud storage solution for structured data
- It is easy to use, works with large amounts of data and offers fast data retrieval and analysis
- Our **instructional notebooks in the Analyses tab** provide code and instructions on using Big Query to access Google datasets



Jupyter

06_How to access plot and save data from public BigQuery datasets using Python 3.ipynb

The following Python code will read a BigQuery table into a Pandas dataframe.

From <https://cloud.google.com/community/tutorials/bigquery-ibis>

Ibis is a Python library for doing data analysis. It offers a Pandas-like environment for executing data analysis composable, and familiar replacement for SQL.

```
In [9]: # Connect to the dataset
conn = ibis.bigquery.connect(dataset_id='bigquery-public-data.broadstreet_adi')
```

```
In [10]: # Read table
ADI_table_2 = conn.table('area_deprivation_index_by_census_block_group')
ADI_table_2
```

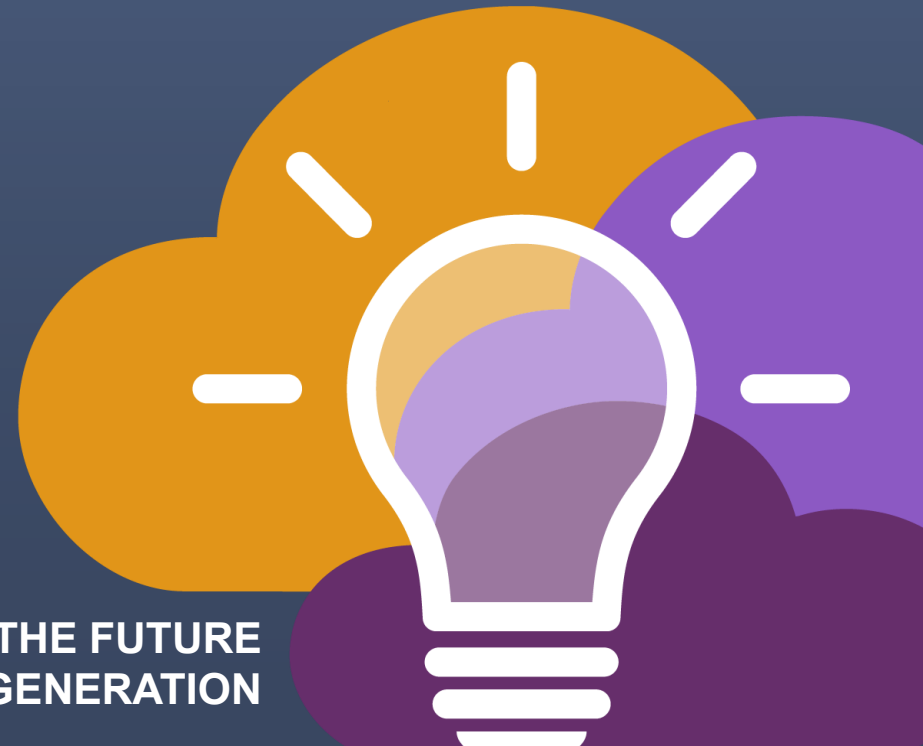
```
Out[10]: BigQueryTable[table]
name: bigquery-public-data.broadstreet_adi.area_deprivation_index_by_census_block_group
schema:
  geo_id : string
  state_fips_code : string
  county_fips_code : string
  block_group_fips_code : string
  description : string
  county_name : string
  state_name : string
  state : string
  year : int64
  area_deprivation_index_percent : float64
```

ScHARe

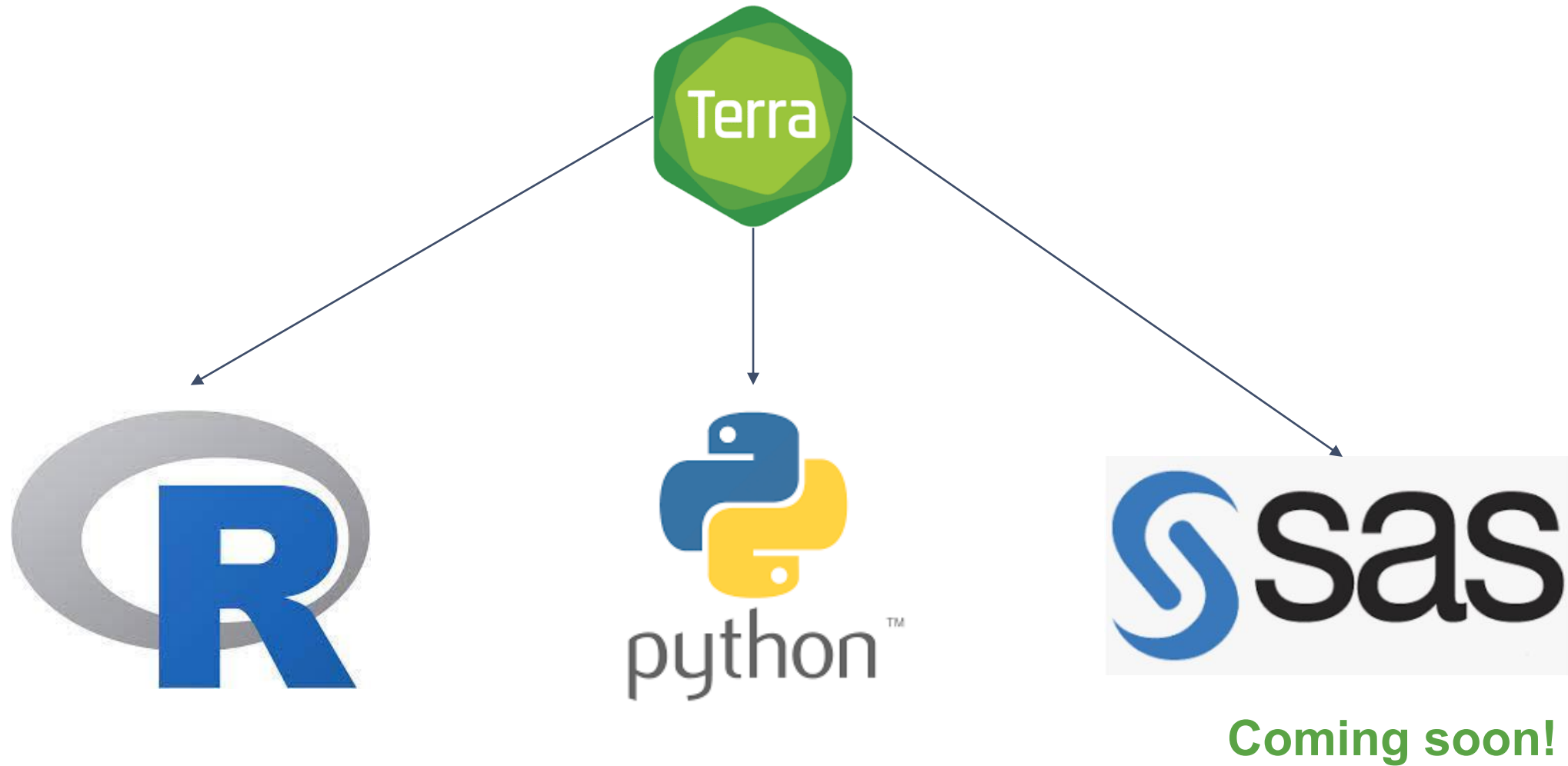
Python and R

The language
of Cloud Computing

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



What programming languages does Terra support?



What is R?

R is a **programming language** for statistical computing and graphics

It is used by data miners, bioinformaticians and statisticians for data analysis

Users have created **packages** to augment its functions

Third-party **graphical user interfaces** are also available, such as Rstudio



What is Python?

Python is a **computer programming language** used in data science to:

- manipulate and analyze data
- create data visualizations
- build machine learning algorithms



Imagine you want to tell your computer what to do, by giving it clear, easy-to-understand commands. That's what Python is like!

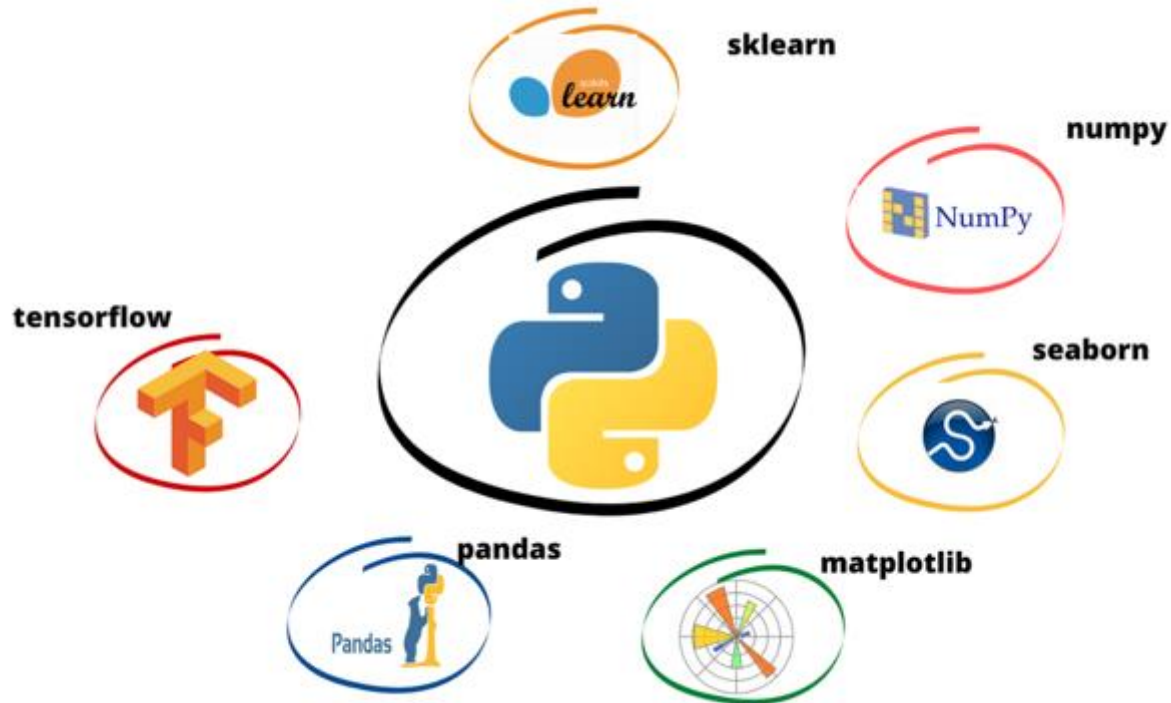
- **Easy to learn:** Python uses words and phrases that are close to everyday English, making it a good choice for beginners
- **Versatile:** You can use Python for many things
- **Free and open-source:** Anyone can use and improve Python for free: there's a large and helpful community to answer your questions
- **Popular:** there are lots of online resources

Sources

www.quanthub.com/python-for-data-science/
[coursera.org](https://www.coursera.org)

Introduction to Python Data Science Libraries

Python offers a **rich ecosystem of libraries** for data science tasks. **Each one serves specific functions** in the data science workflow



What is a Python library?

It's like a **collection of tools or functions** that someone else has **already built and packaged up** for you to use in your own programs

When you're writing a Python program and you need to do something specific, like create visualizations, you can often find a library that **already has the tools you need for that job**

You just need to **"import" the library** into your program, and you can start using its tools right away

Why Python?

According to SlashData:

- there are 8.2 million Python users
- **69%** of machine learning developers and data scientists **use Python (vs. 24% using R)**

Source

stackify.com/learn-python-tutorials/

How to learn Python

How long does it take to learn Python?

It can take **2 to 5 months**, but you can write your first short program in **minutes**

Can you learn Python with no experience?

Python is the **perfect** programming language **for people without any coding experience**, as it has a simple syntax and is very accessible to beginners

Links to additional **free learning resources** will be provided

Python resources

You can take advantage of the dozens of “**Python for data science**” online tutorials for beginners and advanced programmers listed here:

- [Stackify - 30+ Tutorials to Learn Python](#)
- [FreeCodeCamp - Code Class for Beginners](#)
- [Harvard – Free Python Course](#)
- [Coursera – Free and Paid Python Courses](#)
- [LearnPython – Free Interactive Python Tutorials](#)
- [BestColleges – 10 Places to Learn Python for Free](#)



Python resources

Stackify

30+ tutorials to learn Python

Top 30 Python Tutorials

In this article, we will introduce you to some of the best **Python tutorials**. These tutorials are suited for both beginners and advanced programmers. With the help of these tutorials, you can learn and polish your coding skills in Python.

1. [Udemy](#)
2. [Learn Python the Hard Way](#)
3. [Codecademy](#)
4. [Python.org](#)
5. [Invent with Python](#)
6. [Pythonspot](#)
7. [AfterHoursProgramming.com](#)
8. [Coursera](#)
9. [Tutorials Point](#)
10. [Codementor](#)
11. [Google's Python Class eBook](#)
12. [Dive Into Python 3](#)
13. [NewCircle Python Fundamentals Training](#)
14. [Studytonight](#)
15. [Python Tutor](#)
16. [Crash into Python](#)
17. [Real Python](#)
18. [Full Stack Python](#)
19. [Python for Beginners](#)
20. [Python Course](#)
21. [The Hitchhiker's Guide to Python!](#)
22. [Python Guru](#)
23. [Python for You and Me](#)
24. [PythonLearn](#)
25. [Learning to Python](#)
26. [Interactive Python](#)
27. [PythonChallenge.com](#)
28. [IntelliPaat](#)
29. [Sololearn](#)
30. [W3Schools](#)

Python resources

FreeCodeCamp

Code class for beginners

A screenshot of the freeCodeCamp website. The header is dark blue with the freeCodeCamp logo and a flame icon. Below the header is a blue navigation bar with the text "Learn to code — free 3,000-hour curriculum". The main content area is white and features two sections. The first section is titled "Python Tutorial for Beginners (Learn Python in 5 Hours)" in bold black text. Below the title is a paragraph of text: "In [this TechWorld with Nana YouTube course](#), you will learn about strings, variables, OOP, functional programming and more. You will also build a couple of projects including a countdown app and a project focused on API requests to Gitlab." The second section is titled "Scientific Computing with Python" in bold black text. Below the title is a paragraph of text: "In [this freeCodeCamp certification course](#), you will learn about loops, lists, dictionaries, networking, web services and more."

Python resources

Harvard

Free Python course

Catalog > Computer Science Courses > HarvardX's Computer Science for Web Programming



Harvard University: CS50's Introduction to Computer Science

An introduction to the intellectual enterprises of computer science and the art of programming.



12 weeks

6–18 hours per week



Self-paced

Progress at your own speed

There is one session available:

4,974,616 already enrolled! After a course session ends, it will be [archived](#) .

Starts Jul 19

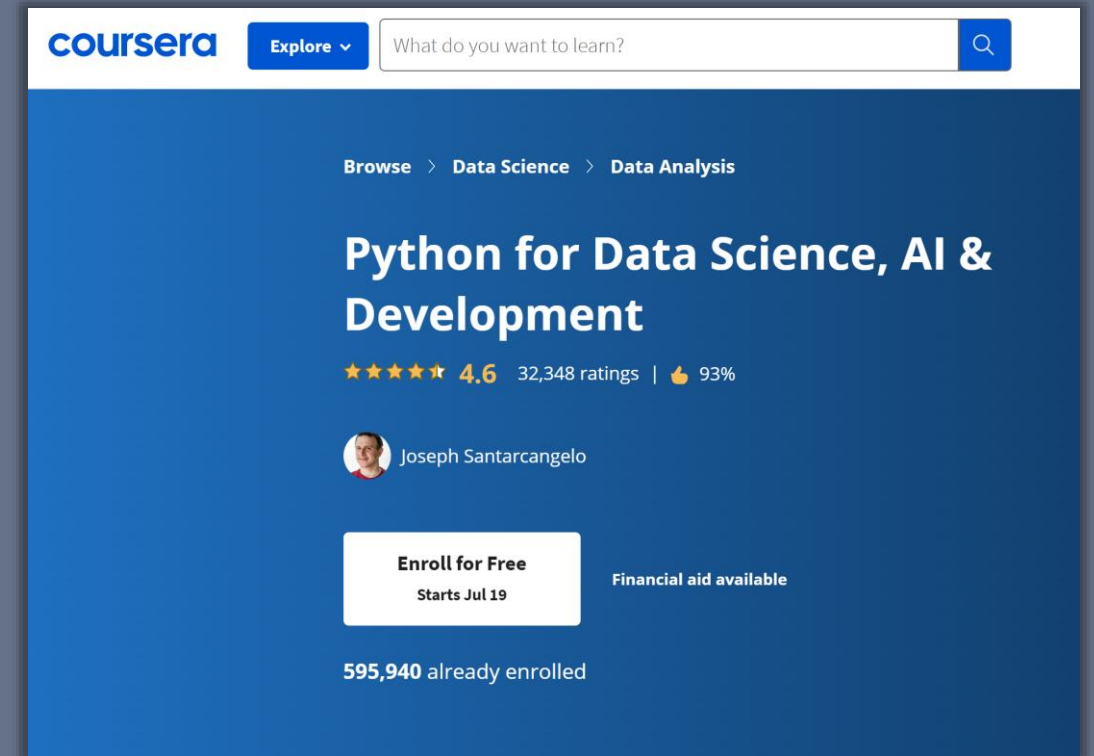
Ends Dec 31

Enroll

Python resources

Coursera

Free and paid Python courses



Python resources

LearnPython

Free interactive Python tutorials

Learn the Basics

- [Hello, World!](#)
- [Variables and Types](#)
- [Lists](#)
- [Basic Operators](#)
- [String Formatting](#)
- [Basic String Operations](#)
- [Conditions](#)
- [Loops](#)
- [Functions](#)
- [Classes and Objects](#)
- [Dictionaries](#)
- [Modules and Packages](#)

Data Science Tutorials

- [Numpy Arrays](#)
- [Pandas Basics](#)

Advanced Tutorials

- [Generators](#)
- [List Comprehensions](#)
- [Lambda functions](#)
- [Multiple Function Arguments](#)
- [Regular Expressions](#)
- [Exception Handling](#)
- [Sets](#)
- [Serialization](#)
- [Partial functions](#)
- [Code Introspection](#)
- [Closures](#)
- [Decorators](#)
- [Map, Filter, Reduce](#)

Python resources

BestColleges

10 places to learn Python for free

[Bootcamp Types](#) ▾ [Reviews](#) ▾ [Resources](#) ▾ [About](#) ▾ [BestColleges.com](#)

Top 10 Free Python Courses

Google's Python Class

Students with some programming language experience can learn Python with Google's intensive two-day course. While there are no official prerequisites, students need a basic understanding of programming language concepts, such as if statements.

Learners initially explore strings and lists using lecture videos and written materials. A coding exercise follows each section, and the exercises become increasingly complex.

This Python course gives students hands-on practice with complete programs, working with text files, processes, and HTTP connections.

Microsoft's Introduction to Python Course

Students can learn Python online and build a simple input/output program with Microsoft's introductory Python course. There are no prerequisites for this short, eight-unit, 16-minute class.

This online Python course is part of Microsoft's Python learning paths. It prepares students with the concepts and basic skills to pursue more advanced learning.

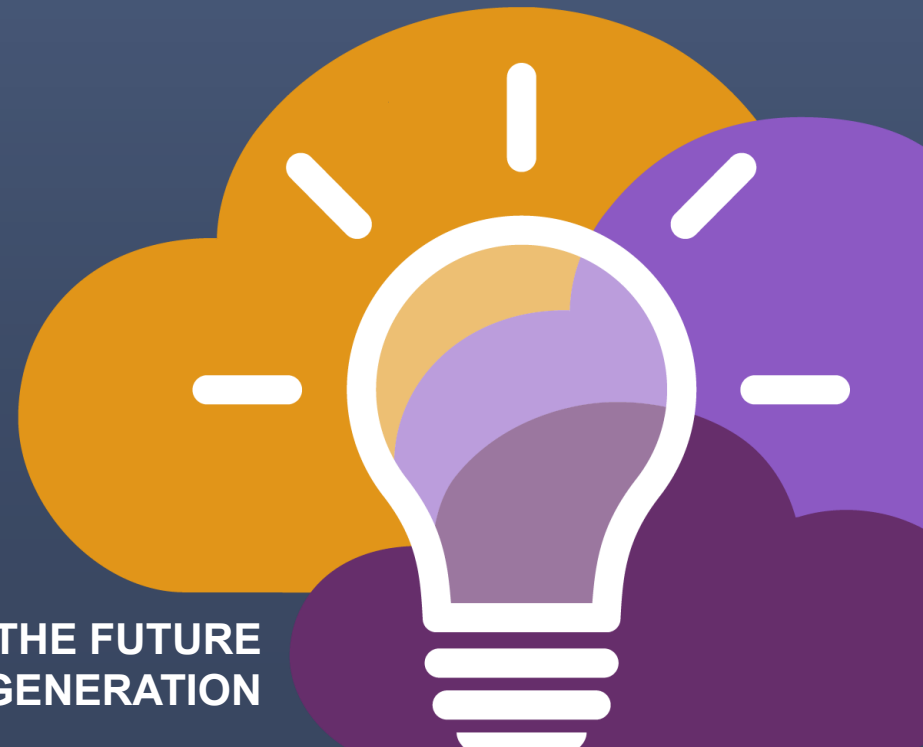
Students explore Python code, where to run Python apps, learn how to declare variables, and use the Python interpreter. They also learn how to access free resources.

ScHARe

Health Disparities,
Health Care Delivery,
Health Outcomes

The Language of Research

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



Minority Health & Health Disparities Definitions and Populations

Examples: hypertension, diabetes mellitus, asthma, cancer, cardiovascular disease, and obesity

Health Disparities

Health differences that adversely affect defined populations, based on one or more health outcomes

Priority Populations:
Minorities / Rural / Low SES / Sexual Gender Minority (SGM) / Disabled

Health Differences across Populations

Minority Health

Distinctive health characteristics and attributes of a racial and/or ethnic group who is socially disadvantaged and/or subject to potential discriminatory acts

Population:
OMB Racial/Ethnic Categories

Health Differences within Populations

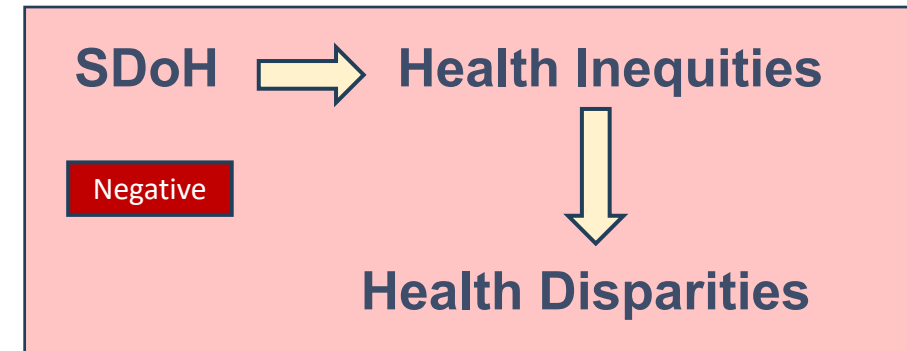
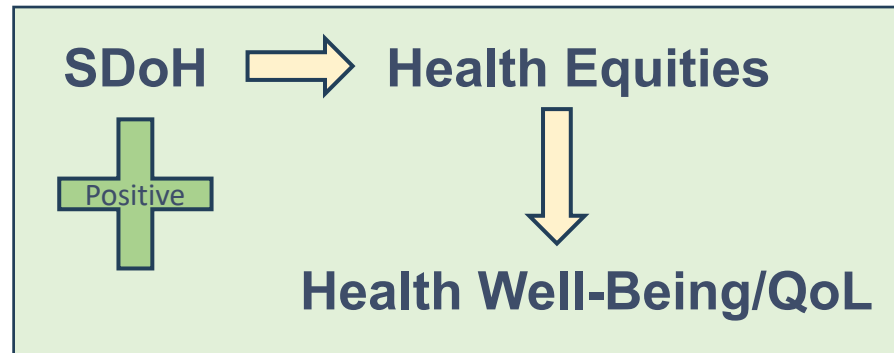
Health Disparity Outcomes

The health outcomes are categorized as:

- Higher incidence and/or prevalence of disease, including earlier onset or more aggressive progression of disease.
- Premature or excessive mortality from specific health conditions.
- Greater global burden of disease, such as Disability Adjusted Life Years (DALY), as measured by population health metrics.
- Poorer health behaviors and clinical outcomes related to the aforementioned.
- Worse outcomes on validated self-reported measures that reflect daily functioning or symptoms from specific conditions.

SDoH: Neutral Measures Impacting Health Outcomes

Individual and Structural SDoH impact Chronic/Infectious Disease Onset & Management



Research Areas:

1. **SDoH Mechanistic pathways** – what factors impact QoL/disease management across the life course?
2. **Interactions with other determinants**, such as biology, behaviors, psychological factors
3. **Structure** refers to “political, social, cultural, historical, and economic forces that influence individual behavior and thus create predictable patterns based on social location”
4. **Intersectionality of SDoH**: Combination of individual factors and the intersecting systems of oppression that perpetuate discrimination and disadvantage based on factors such as race, class, sex, and gender identity

SDoH Impact Health Equity that Determines QoL


“**Health equity** is the principle underlying the continual **process** of assuring that all individuals or populations have optimal opportunities to attain the best health possible. Applying the principle of health equity requires that barriers to promoting good health are removed and resources are allocated among populations and/or communities proportional to their need(s).”

NIMHD 2024

Assuring sustainable health equity often involves changes in laws, policies, processes, norms, values, resource allocation, and power structures (both intentional and unintentional) that affect access to healthcare, employment, education, wealth, public safety, housing, safe green spaces, and other social determinants of health.

Applying a health equity lens in science requires an intentional effort to ensure that **research** is designed explicitly to promote fairness, opportunity, quality, and social justice in access, interventions or treatments, and outcomes.

Health Care Delivery



Advancing health equity into the future.

NINR's mission is to lead nursing research to solve pressing health challenges and inform practice and policy – optimizing health and advancing health equity into the future.

Health Outcome

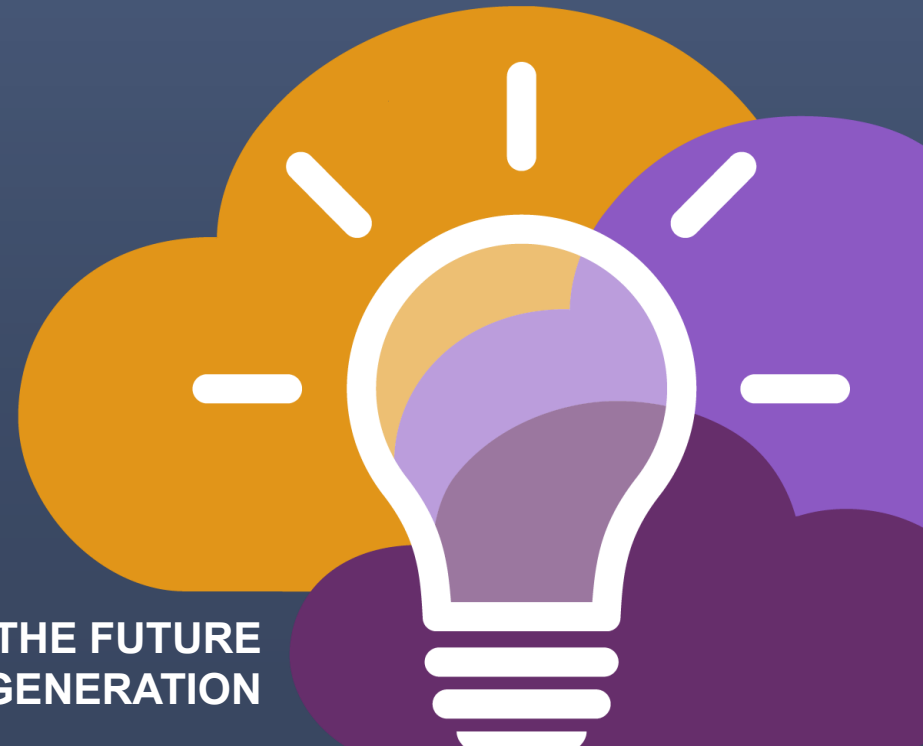
Change in the health of an individual, group of people or population which is attributable to an intervention or series of interventions.

These may be measured clinically (physical examination, laboratory testing, imaging), self-reported, or observed (such as gait or movement fluctuations seen by a healthcare provider or caregiver).

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



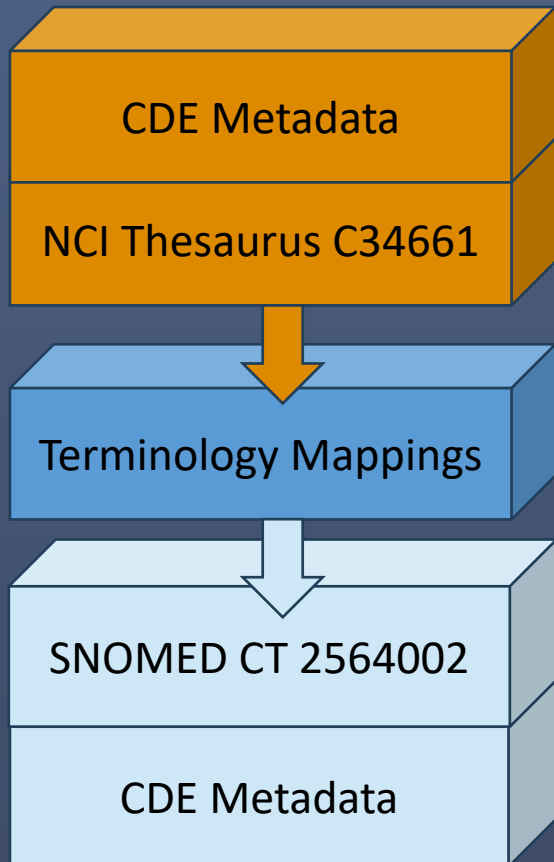
- Age
- Birthplace
- Zip Code
- Race and Ethnicity
- Sex
- Gender
- Sexual Orientation
- Marital Status
- Education
- Annual Household Income
- Household Size

- 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)

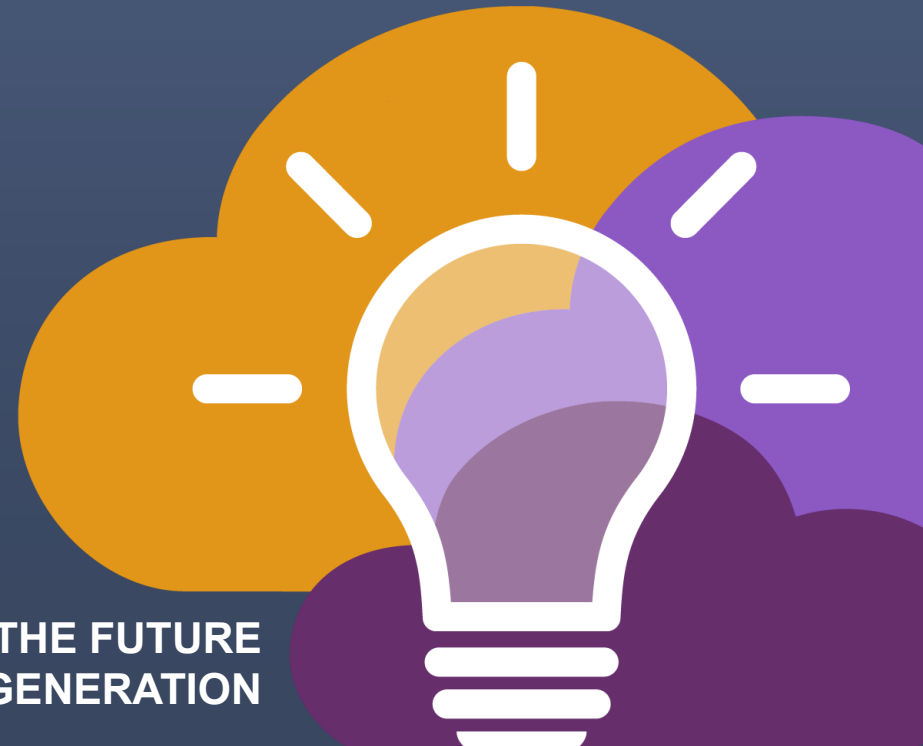


- CDE unique **CONCEPT CODES** represent data semantics
 - Human readable
 - Machine readable format
- **Mapping** enables interoperability even if the same standard terminology was not used in another CDE
- CDE Metadata enables searching for concept codes across CDEs to compare data

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)

	Pecos Bill
	Name
	Amarillo Texas
	City State
	TX909998
	Driver's License Number

	George Washington
	Name
	Washington, DC
	Address
	11234334
	DL#

- 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 sentence is 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 achievement.	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 or 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 or 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

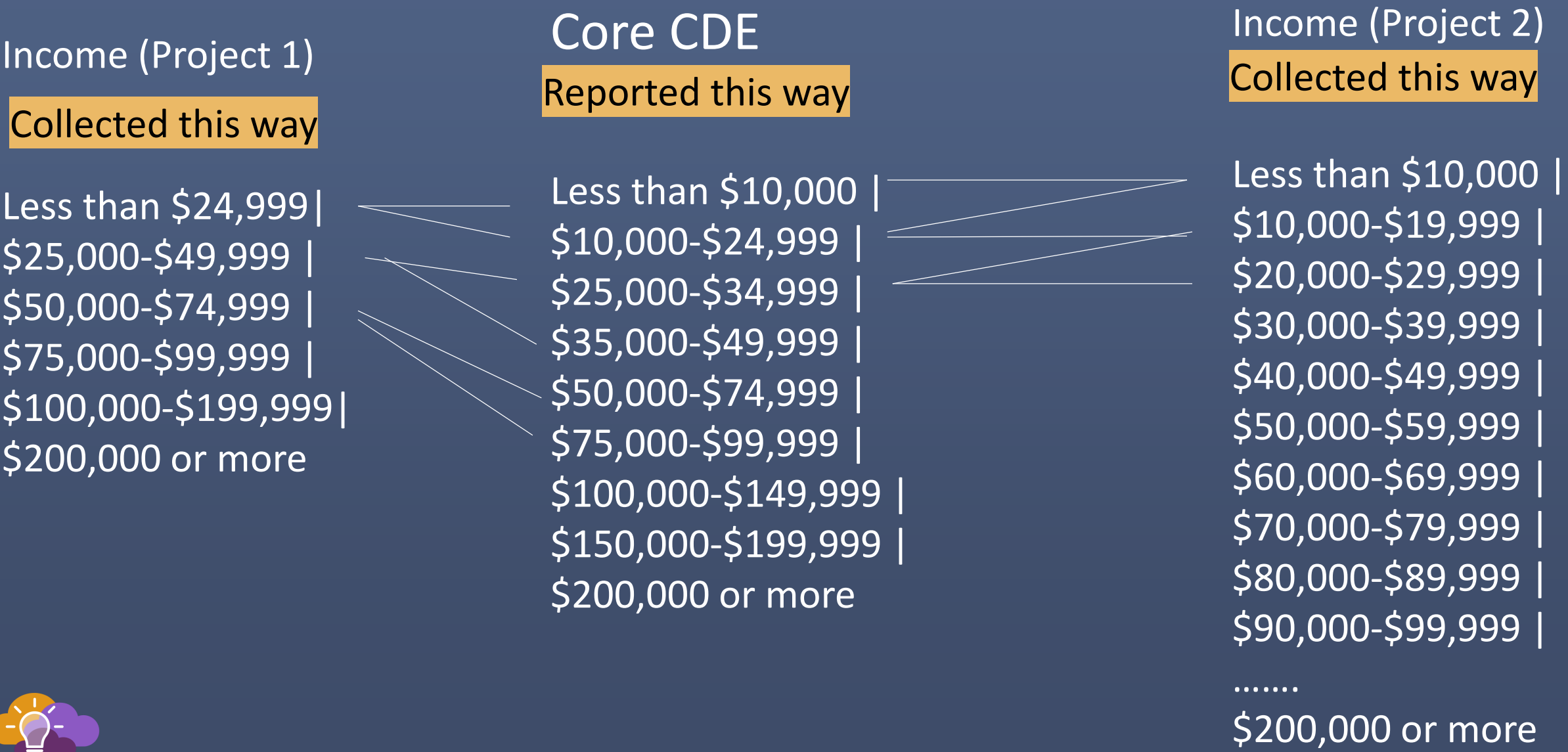
Data Harmonization: matched CDEs from different projects



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

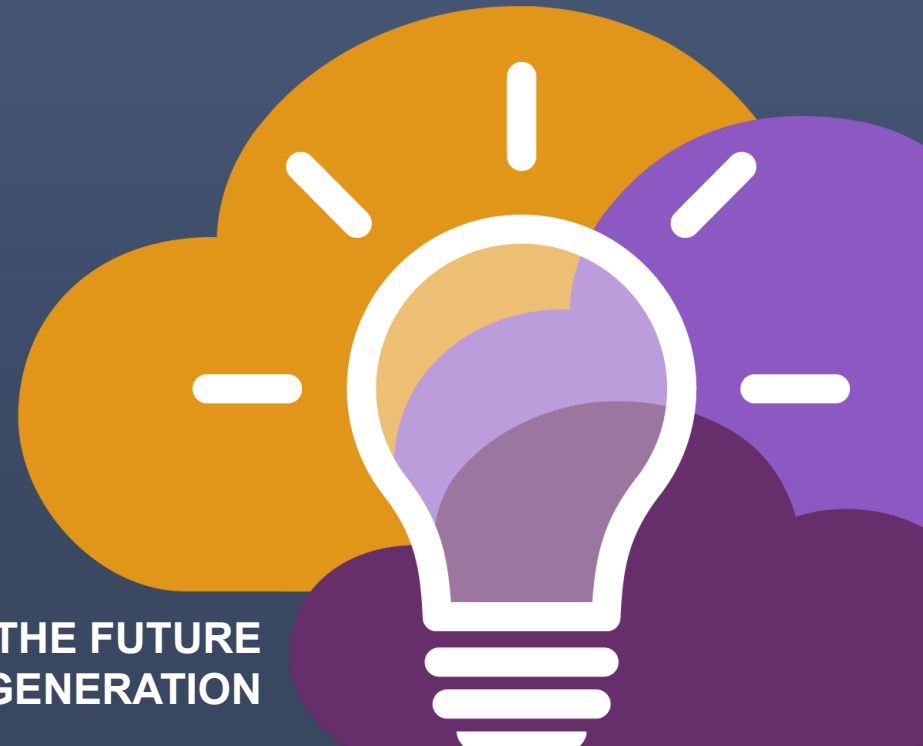
Mapped using algorithms



ScHARe

CDE Terminology
and Concept Coding

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



Accelerating Biomedical Discovery and Data-Powered Health



PubMed

Citations for biomedical
literature



MedlinePlus

Reliable, up-to-date health
information for you



Open-i

An experimental
multimedia search engine



MeSH

Medical Subject
Headings



ClinicalTrials.gov

A database of clinical
studies, worldwide



BLAST

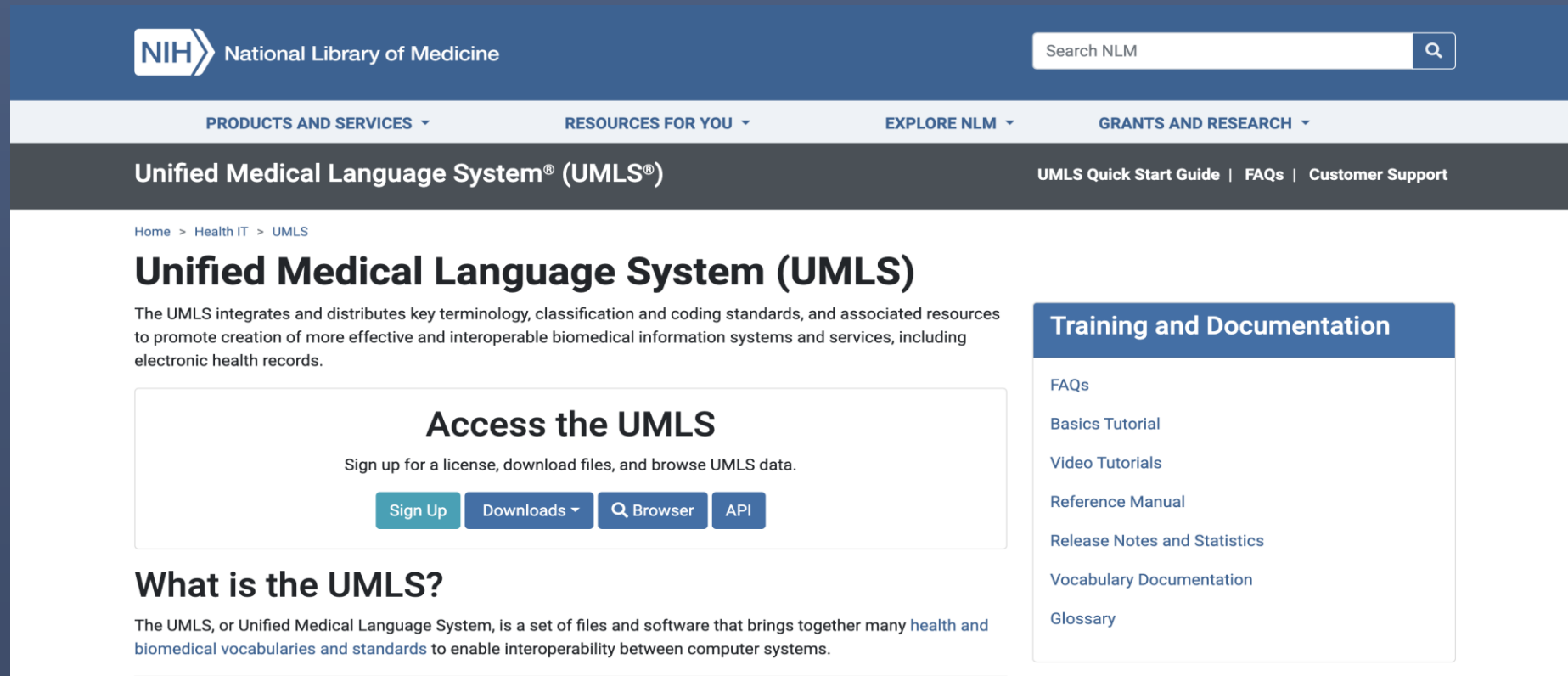
Basic Local Alignment
Search Tool

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



The screenshot shows the official NLM UMLS website. At the top is the NIH National Library of Medicine logo and a search bar. Below this is a navigation bar with links to PRODUCTS AND SERVICES, RESOURCES FOR YOU, EXPLORE NLM, and GRANTS AND RESEARCH. The main heading is 'Unified Medical Language System® (UMLS®)' with links to a Quick Start Guide, FAQs, and Customer Support. A breadcrumb trail indicates the path: Home > Health IT > UMLS. The main content area features a large heading 'Unified Medical Language System (UMLS)' followed by a descriptive paragraph. Below this is a section titled 'Access the UMLS' with a subtext 'Sign up for a license, download files, and browse UMLS data.' and four buttons: 'Sign Up', 'Downloads', 'Browser', and 'API'. To the right is a 'Training and Documentation' sidebar with links to FAQs, Basics Tutorial, Video Tutorials, Reference Manual, Release Notes and Statistics, Vocabulary Documentation, and Glossary. At the bottom left, there is a section titled 'What is the UMLS?' with a brief description of the system.

NIH National Library of Medicine

Search NLM

PRODUCTS AND SERVICES ▾ RESOURCES FOR YOU ▾ EXPLORE NLM ▾ GRANTS AND RESEARCH ▾

Unified Medical Language System® (UMLS®) UMLS Quick Start Guide | FAQs | Customer Support

Home > Health IT > UMLS

Unified Medical Language System (UMLS)

The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records.

Access the UMLS

Sign up for a license, download files, and browse UMLS data.

Sign Up Downloads ▾ Browser API

What is the UMLS?

The UMLS, or Unified Medical Language System, is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.

Training and Documentation

- FAQs
- Basics Tutorial
- Video Tutorials
- Reference Manual
- Release Notes and Statistics
- Vocabulary Documentation
- Glossary

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 HL7 (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



FHIR

Important Concepts, Terms and Definitions

FHIR is the latest interoperability standard based on a RESTful API architecture published by HL7. HL7 has been working for over 25 years in publishing standards for Healthcare data interoperability. The purpose of this article is to get the reader to understand the difference between the earlier versions of HL7 interoperability standards and then present the important concepts that will help you to understand the FHIR standard



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.

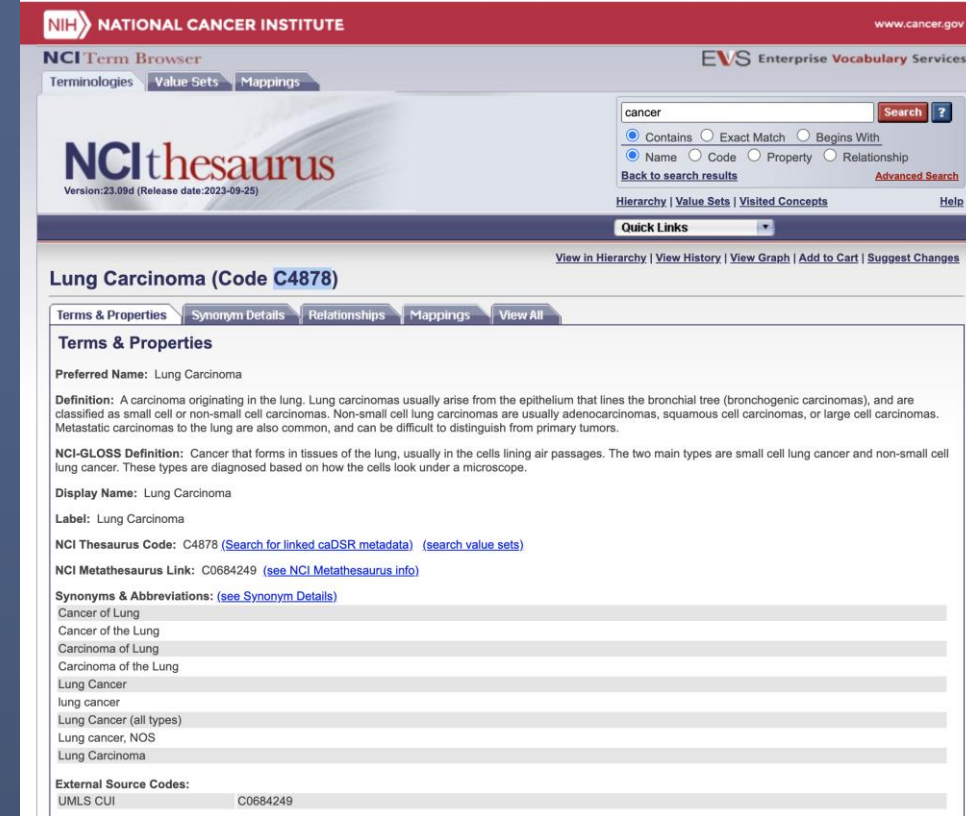
Publicly Available



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”



The screenshot displays the NCI Thesaurus web interface. At the top, the header includes the NIH logo, "NATIONAL CANCER INSTITUTE", and the URL "www.cancer.gov". Below this is the "NCI Term Browser" section with tabs for "Terminologies", "Value Sets", and "Mappings". A search bar contains the word "cancer", and a "Search" button is next to it. Below the search bar, there are radio buttons for "Contains", "Exact Match", and "Begins With", and checkboxes for "Name", "Code", "Property", and "Relationship". A "Back to search results" link and an "Advanced Search" link are also present. The main content area is titled "Lung Carcinoma (Code C4878)" and includes tabs for "Terms & Properties", "Synonym Details", "Relationships", "Mappings", and "View All". The "Terms & Properties" tab is selected, showing the "Preferred Name: Lung Carcinoma", a "Definition" paragraph, a "NCI-GLOSS Definition" paragraph, a "Display Name: Lung Carcinoma", a "Label: Lung Carcinoma", and "NCI Thesaurus Code: C4878" with links to "Search for linked caDSR metadata" and "search value sets". It also shows the "NCI Metathesaurus Link: C0684249" with a link to "see NCI Metathesaurus info". A section for "Synonyms & Abbreviations" lists various terms like "Cancer of Lung", "Carcinoma of Lung", "Lung Cancer", "lung cancer", "Lung Cancer (all types)", "Lung cancer, NOS", and "Lung Carcinoma". At the bottom, "External Source Codes" are listed, including "UMLS CUI: C0684249".



No Coding System is Better Than the Other

General use....

LOINC

Laboratory and Clinical Research

ULMS (CUI)

Biomedical

FHIR

Electronic Health Records

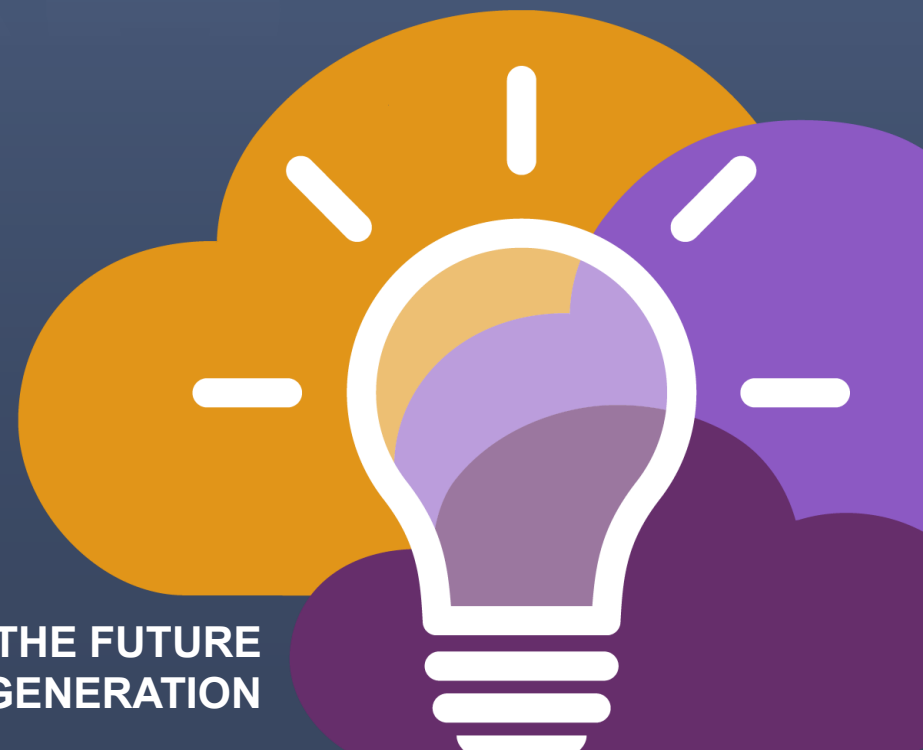
*NCIt

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
- **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



aim-ahead.net



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

ScHARe

All of Us
RESEARCH PROGRAM



 **AnVIL**

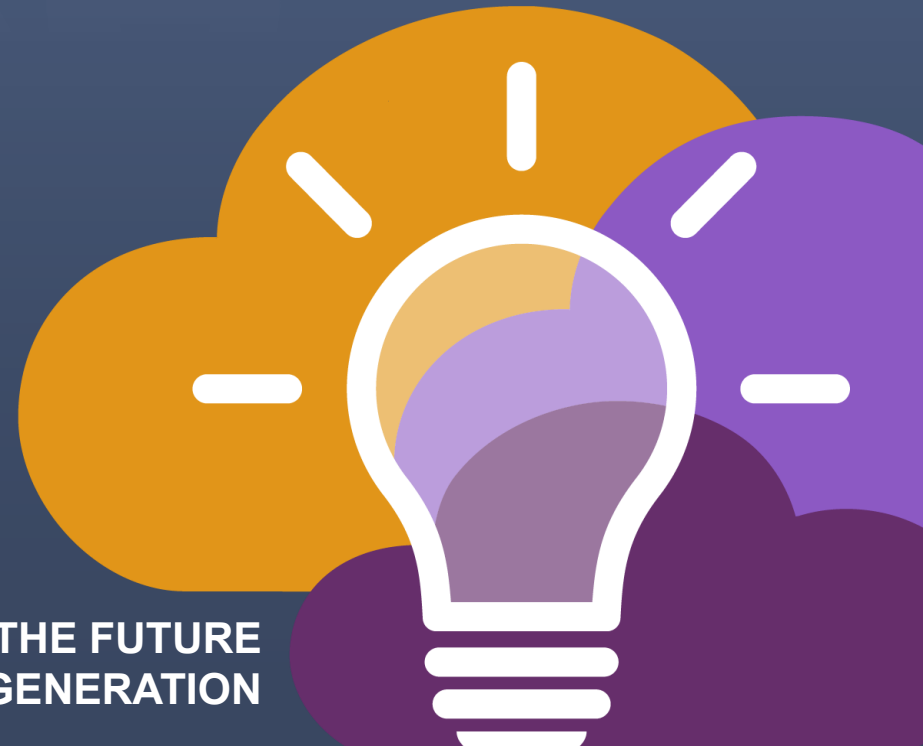
BioData
CATALYST®

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 gender 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
	<i>ScHARe for Educators (Community Colleges & Low Resource MSIs)</i>
	<i>ScHARe for American Indian / Alaska Native Researchers</i>
	<i>ScHARe for Coders and Programmers to conduct Research</i>



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

<p>Title: Data Science Projects 1 – Health Disparities and Individual SDoH</p> <p>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.</p>
<p>Title: Data Science Projects 2 - Health Disparities and Structural SDoH</p> <p>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.</p>
<p>Title: Data Science Projects 3 – Health Outcomes</p> <p>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.</p>

**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

Findable: *providing metadata, documentation, and clear identifiers*

Accessible: *wide audience*

Interoperable: *standardized formats and APIs enable seamless integration*

Reusable: *clear documentation, licensing, reduce redundancy*

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

A vertical image on the left side of the slide shows a hand reaching for a door handle. The hand is dark, and the door handle is a simple, dark, cylindrical shape. Below the handle, a set of keys is hanging from a ring. The background of the image is a light, neutral color.

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.



[forms.office.com/
g/jtYNfEdRUG](https://forms.office.com/g/jtYNfEdRUG)

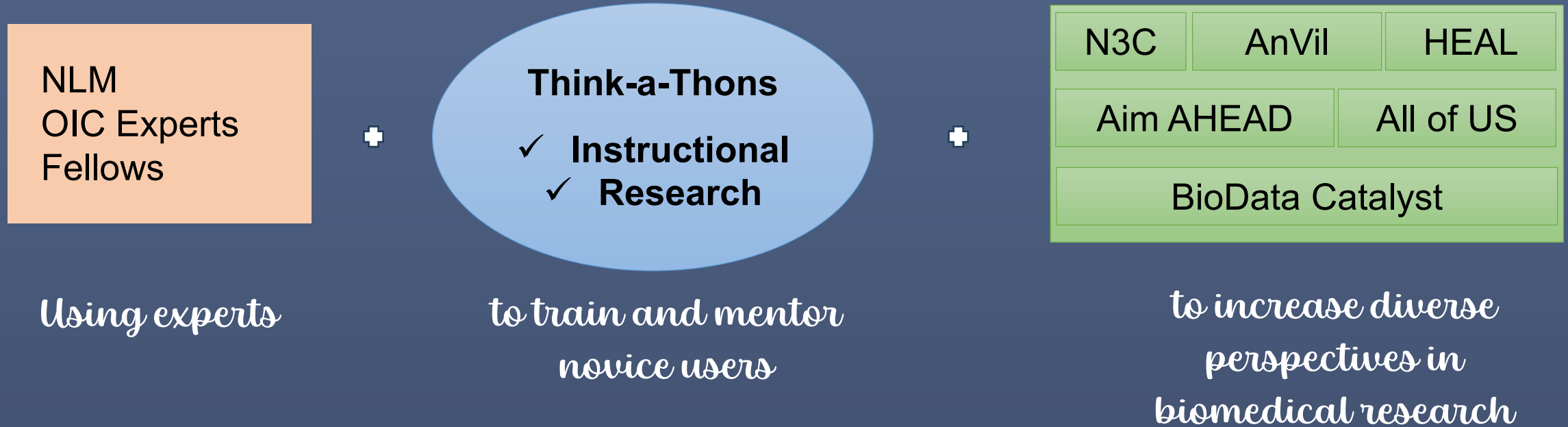
SCHARe

Training Pipelines

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



Think-a-Thons Training/Mentoring Pipeline



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



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

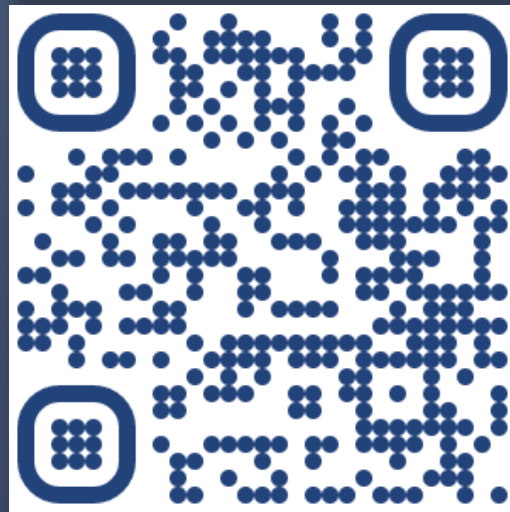
ScHARe

Next Think-a-Thons:



bit.ly/think-a-thons

Register for ScHARe:



bit.ly/join-schare

 schare@mail.nih.gov

