ScHARe

Data Management and Analysis in Python

September 18, 2024

Deborah Duran, PhD · NIMHD
Luca Calzoni, MD MS PhD Cand. · NIMHD
Elif Dede Yildirim, PhD · NIMHD

ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

Outline

5’	Introduction
	▪ Experience poll
	▪ Interest poll
10’	What is ScHARe?
10’	Workshop setup
5’	Why Python?
10’	Recap from August session
5’	Importance of data cleaning
10’	Tools for data cleaning
10’	How data impacts visualizations
10’	Machine Learning primer
1h10’	Examples of Visualizations, Data Cleaning, Machine Learning
5’	Python tutorials and resources
	▪ Evaluation poll

Experience poll

Please check your level of experience with the following:

	None	Some	Proficient	Expert
Python	☐	☐	☐	☐
R	☐	☐	☐	☐
Cloud computing	☐	☐	☐	☐
Terra	☐	☐	☐	☐
Health disparities research	☐	☐	☐	☐
Health outcomes research	☐	☐	☐	☐
Algorithmic bias mitigation	☐	☐	☐	☐

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

What is ScHARe?

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

ScHARe

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 health disparity 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

nimhd.nih.gov/schare

ScHARe

Google Platform
Terra Interface

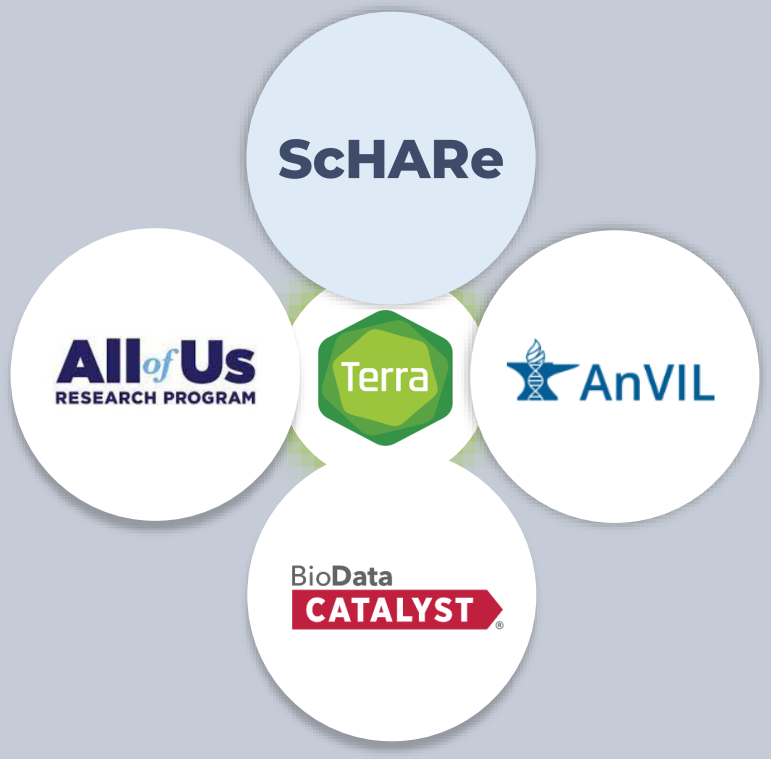
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Copy-and-paste code in Python and R
- Learning Terra on ScHARe prepares you to use other NIH platforms

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 for collaborative research



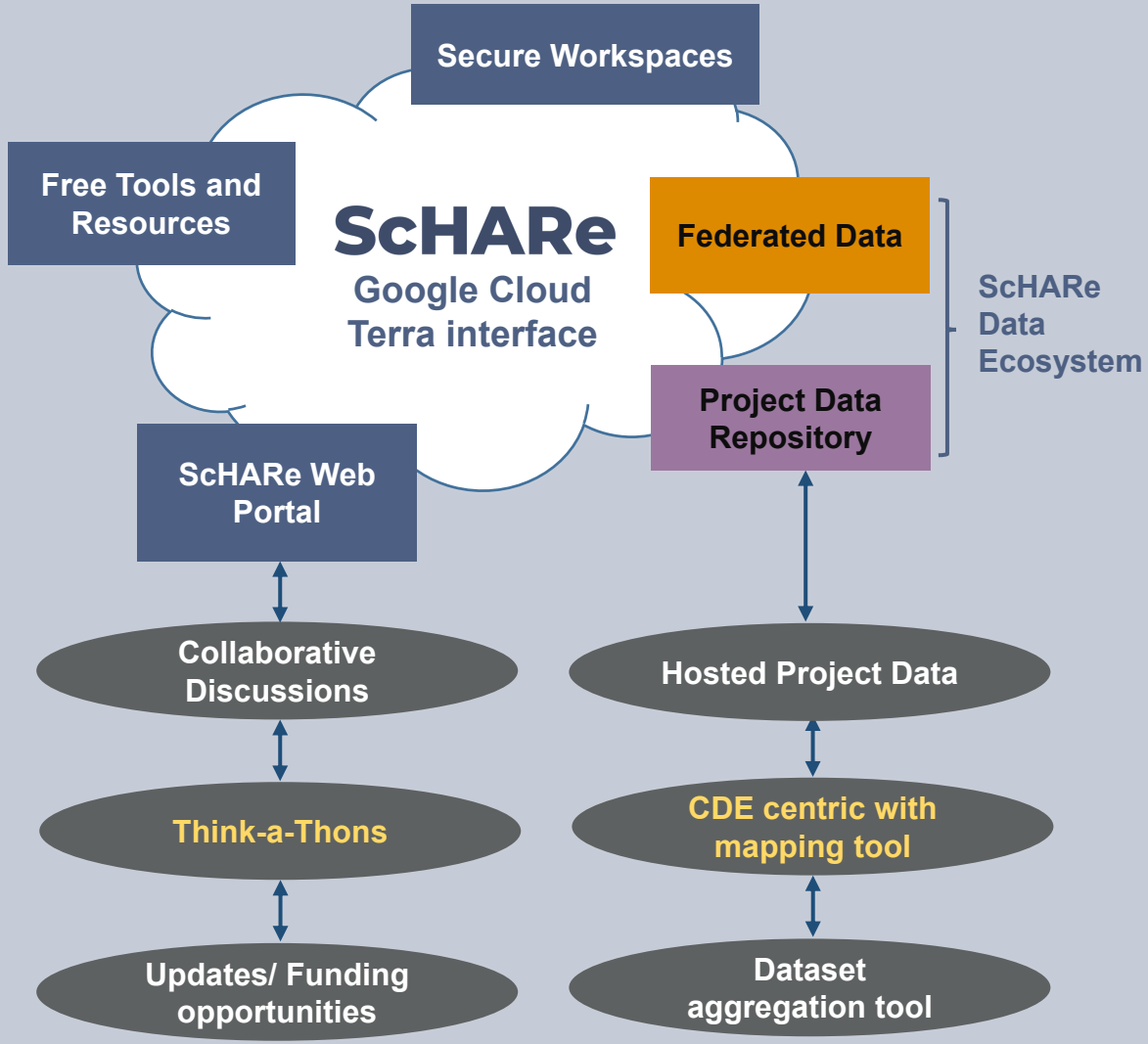
BE A PART OF THE FUTURE OF KNOWLEDGE GENERATION

ScHARe Components

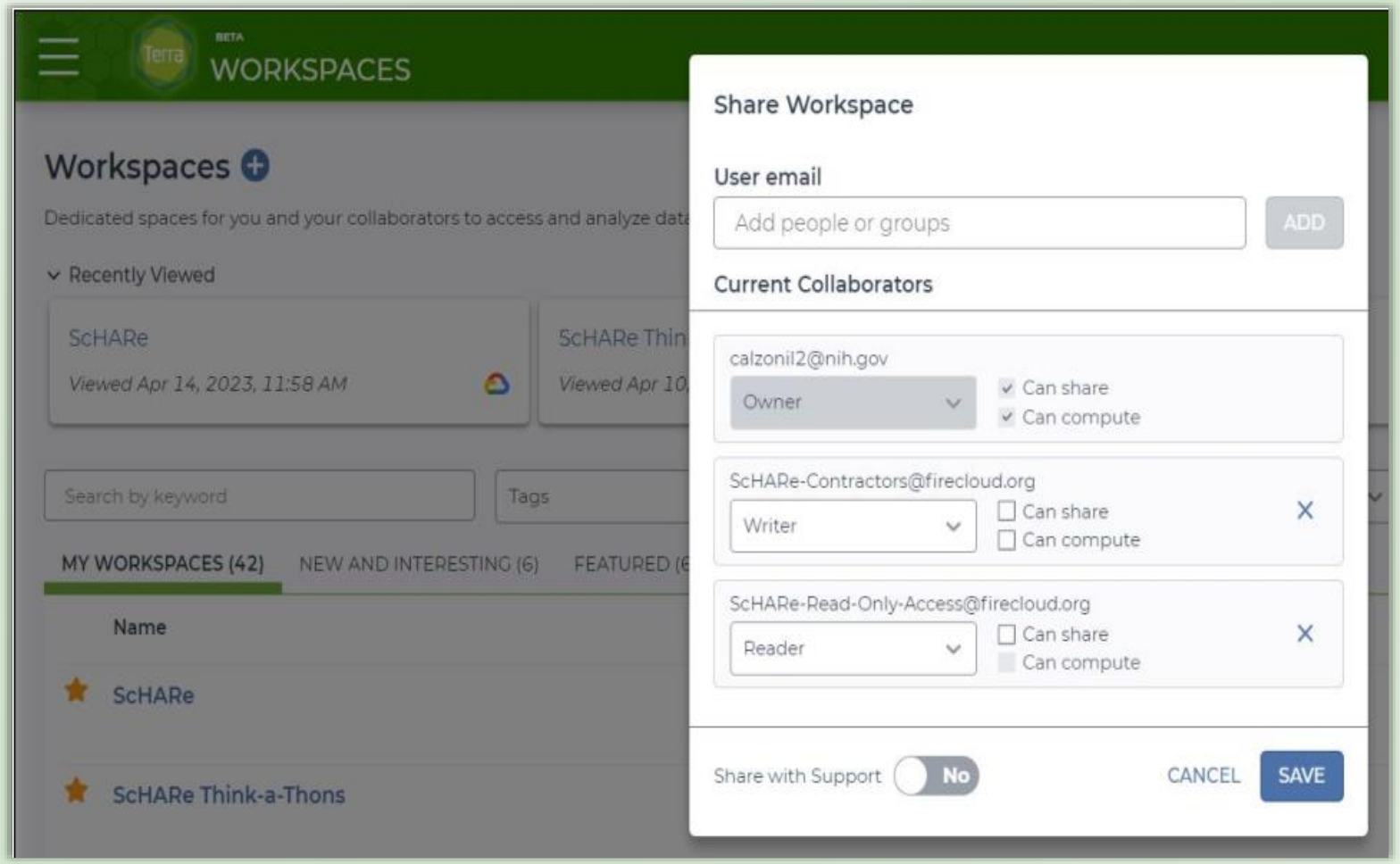
ScHARe co-localizes within the cloud:

1. **Datasets** (including social determinants of health and social science data) relevant to minority health, health disparities, and healthcare outcomes research
2. **CDE-focused data repository** to comply with the required hosting and sharing of data from NIMHD-/NINR-funded programs
3. **User-friendly computational capabilities and secure, collaborative workspaces** for students and all career level researchers
4. **Tools for collaboratively evaluating and mitigating biases** associated with datasets and algorithms utilized to inform healthcare and policy decisions (*upcoming*)

Intramural and Extramural Resource



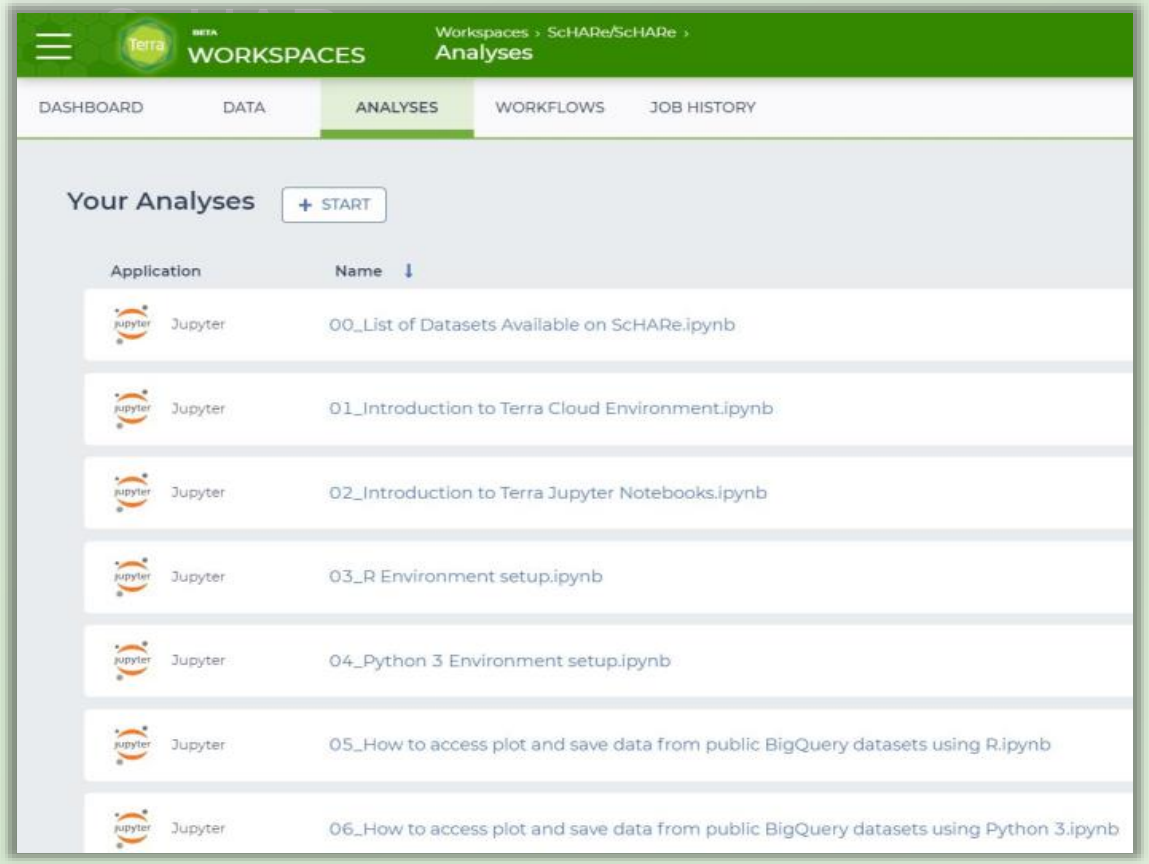
ScHARe Terra interface: secure workspaces



- Secure workspaces **for self or collaborative research**
- **Assign roles**: review or admin
- **Host own data and code**

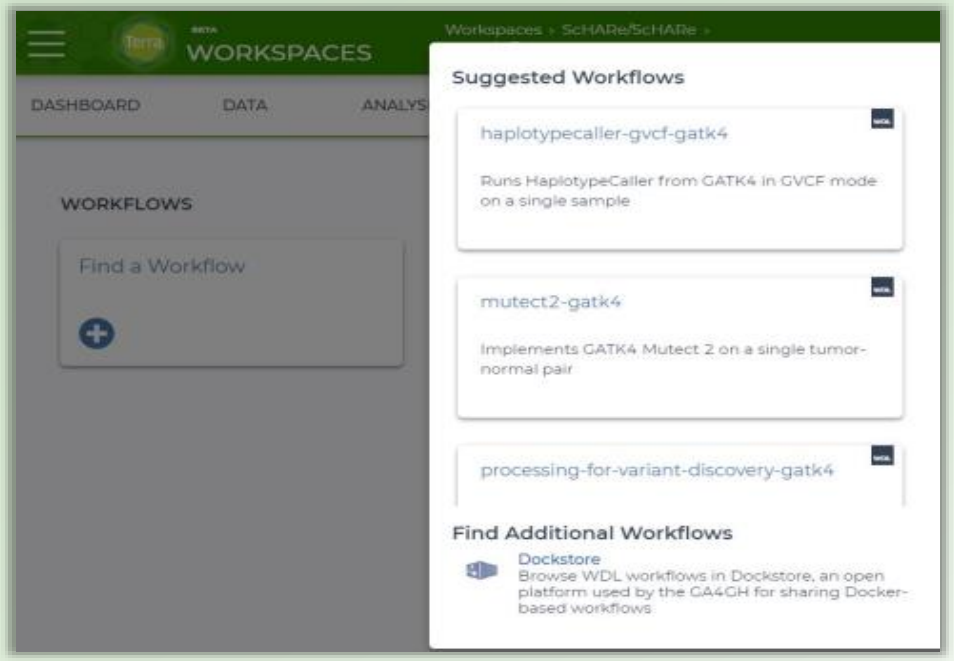
ScHARe Terra interface: analyses

Notebooks for analytics and tutorials



Modular codes

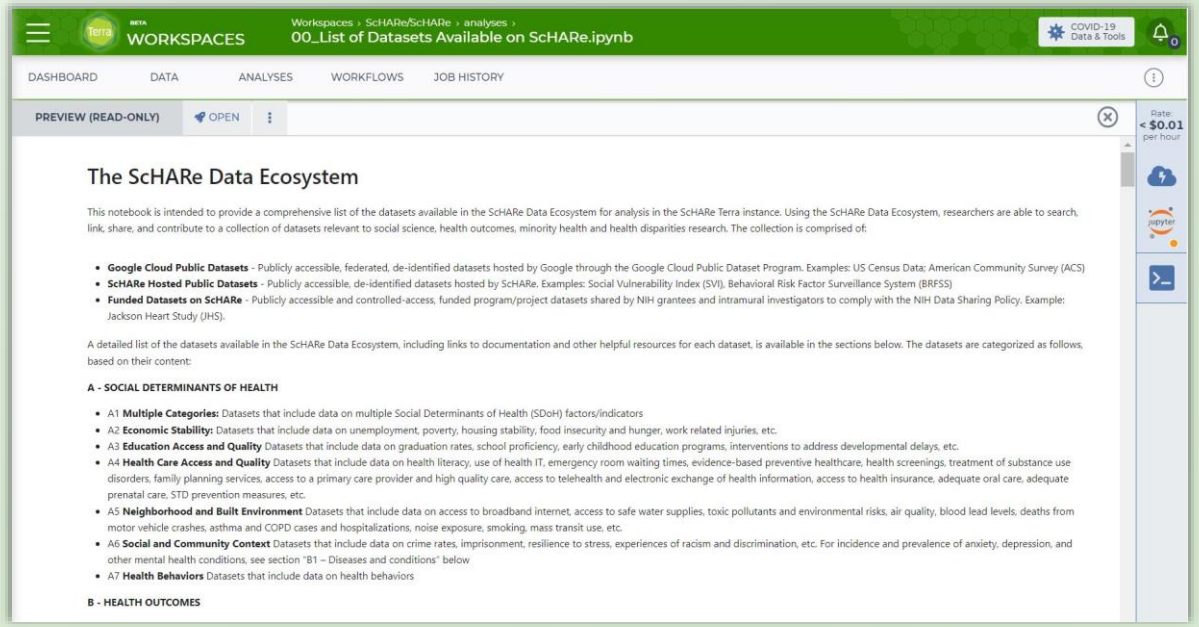
- **Easy-to-use copy-and-paste analytics**



- Modular codes developed for reuse
- **Adding SAS**

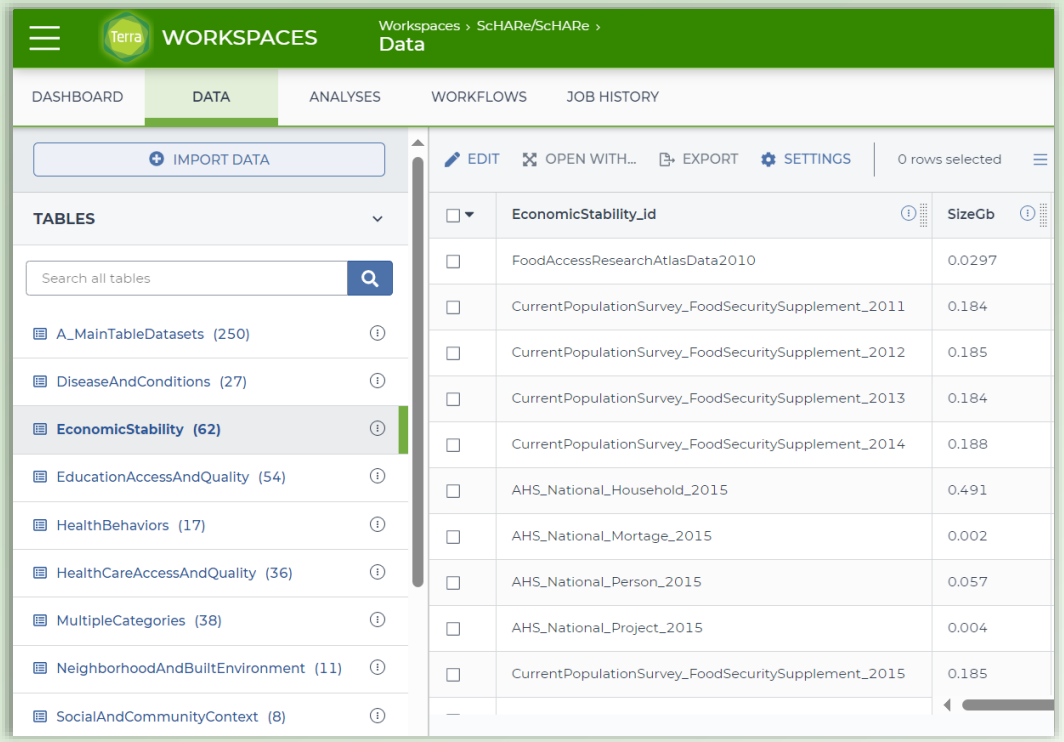
ScHARe Terra interface: access to datasets

What data?



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

Where?



In the **Data tab**, data tables help access data

ScHARe Ecosystem structure

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:

250+
FEDERATED PUBLIC DATASETS

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)*

CDE
FOCUSED REPOSITORY

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

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

ScHARe Ecosystem

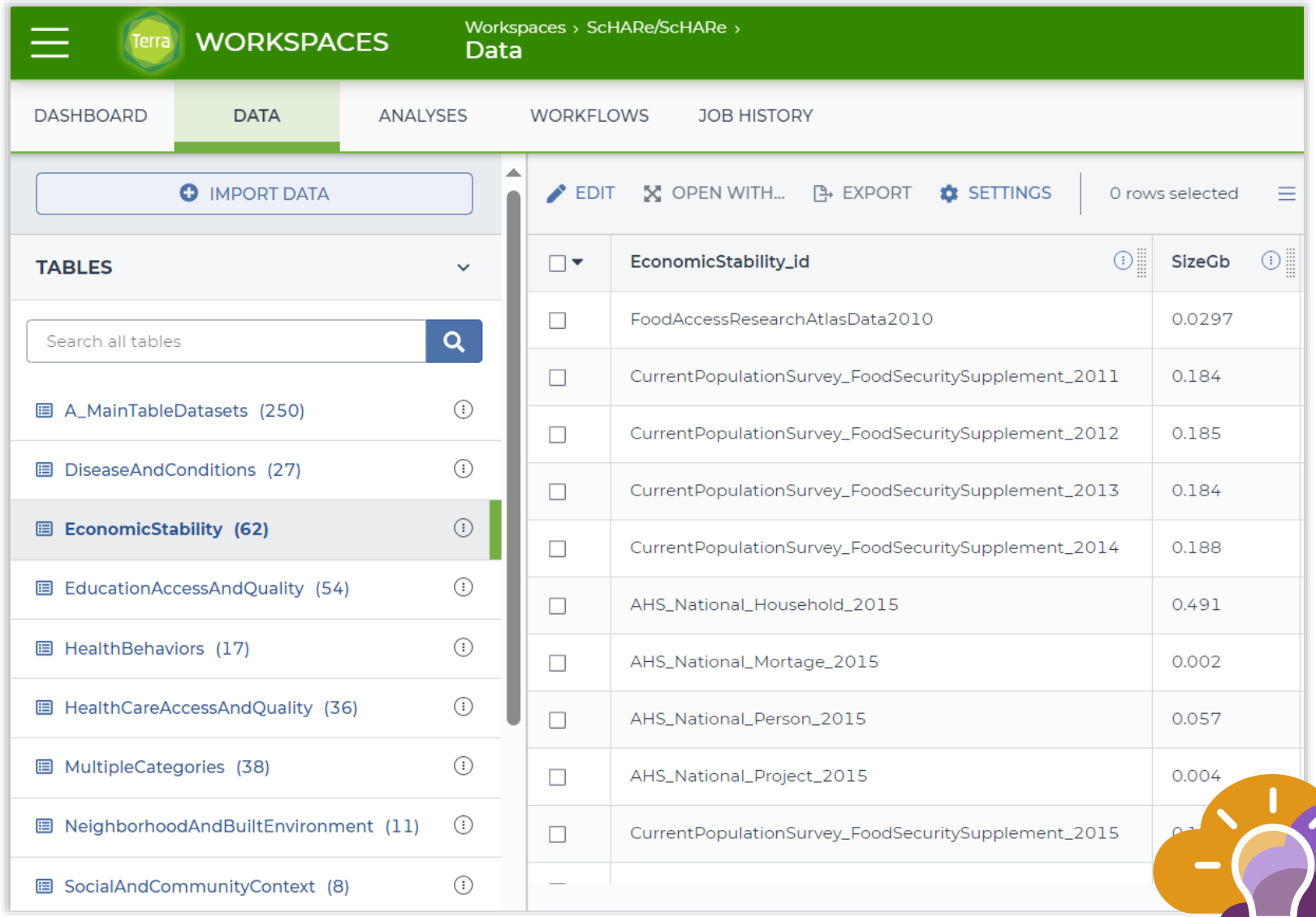
OVER 260 DATA SETS CENTRALIZED

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**



ScHARe Ecosystem: ScHARe hosted datasets

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

260+ 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**

https://www.cdc.gov/about/priorities/social-determinants-of-health-at-cdc.html?CDC_AAref_Val=https://www.cdc.gov/about/sdoh/index.html

ScHARe Ecosystem: ScHARe hosted datasets

Education access and quality

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

Economic stability

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

Health care access and quality

Data on health literacy, use of health IT, preventive healthcare, access to health insurance, etc.

* Health behaviors

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

Neighborhood and built environment

Data on access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, noise exposure, smoking, mass transit use, etc.

* Diseases and conditions

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

Social and community context

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

** Not Social Determinants of Health*

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 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

ScHARe

The ScHARe Data Ecosystem

This document 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-hosted 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)
- **ScHARe-hosted Project Datasets** - Publicly accessible and controlled-access, funded program/project datasets shared by NIH grantees and intramural investigators to comply with the [illegible] Jackson Heart Study (JHS)

ScHARe
Datasets
PDF list

Scan me

bit.ly/ScHARe-datasets

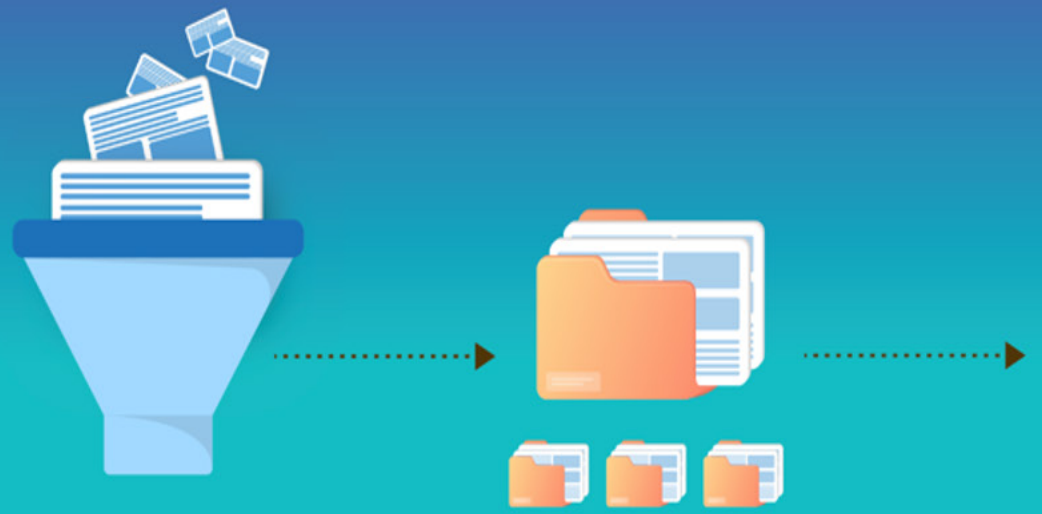
CDE benefits:

- **Faster start-up for project**
- **Better data aggregation across projects**
- **Shared meaning**
- **Concept-focused to allow questions/answers variations**
- **Coding enables an URI approach for better data interoperability**

A **Common Data Element** (CDE) is a standardized, precisely defined question, paired with a set of allowable responses, used systematically across different sites, studies, or clinical trials to ensure consistent data collection

Because Researchers use CDEs…

they can more quickly share data and get results faster, which ultimately can help make a **meaningful difference to our nation's health.**



For more information about how CDEs accelerate research discoveries, visit: cde.nlm.nih.gov/resources

ScHARe PhenX Toolkit Core CDEs

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

ScHARe has developed **Common Data Elements** to ensure consistent data collection across studies, facilitate interoperability, and link data from different sources

NIH CDE Repository:
cde.nlm.nih.gov/home
PhenX Toolkit:
www.nimhd.nih.gov/resources/phenx/

COMMON DATA ELEMENTS

NLM CDE Repository

Coded NIMHD Common Data Elements

- Labels
- Questions
- Permissible Values

DATA UPLOAD

Acquired **Google and ScHARe Hosted Datasets**

- Overview
- Data Dictionaries
- Data Updates

ScHARe REPOSITORY

Project and Key Acquired Datasets

Overview

Description and Links to Overview Material

4-Privacy Levels

COMMON DATA ELEMENTS

Data

Metadata

Data Dictionaries

Analysis Ready

RAS Single Sign-on

ATO

Common Data Elements + **Data**

Data Access

Based On PII Levels and User Needs:

- Public
- Data Use Agreement
- Private

DATA MAPPING, DOWNLOAD AND EXPORT

Other Cloud Platforms
AnVil, BDC, All of Us

DATA MAPPING

ACROSS DATASETS AND PLATFORMS BASED ON CDES

EXAMPLE: CDE linked

ACS **NIMHD Project** **BioData Catalyst**

Aggregated Data Set

CDE Linked Project Data

Data Download in a Variety of Formats

CSV, TSV, XLSX

Data Export to Terra for Analysis

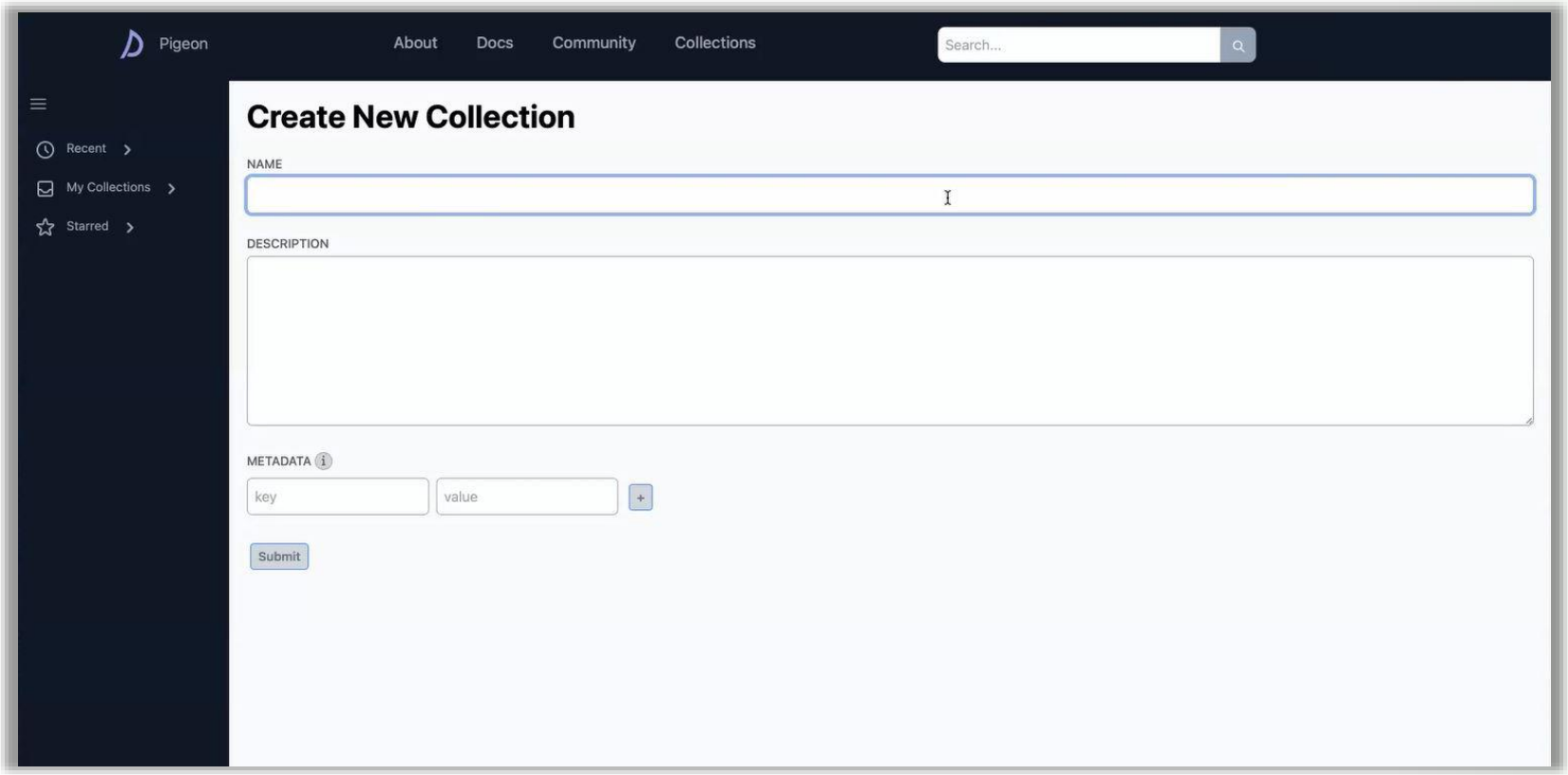
Workspaces

Visualizations Tools

Shiny

ScHARe Repository

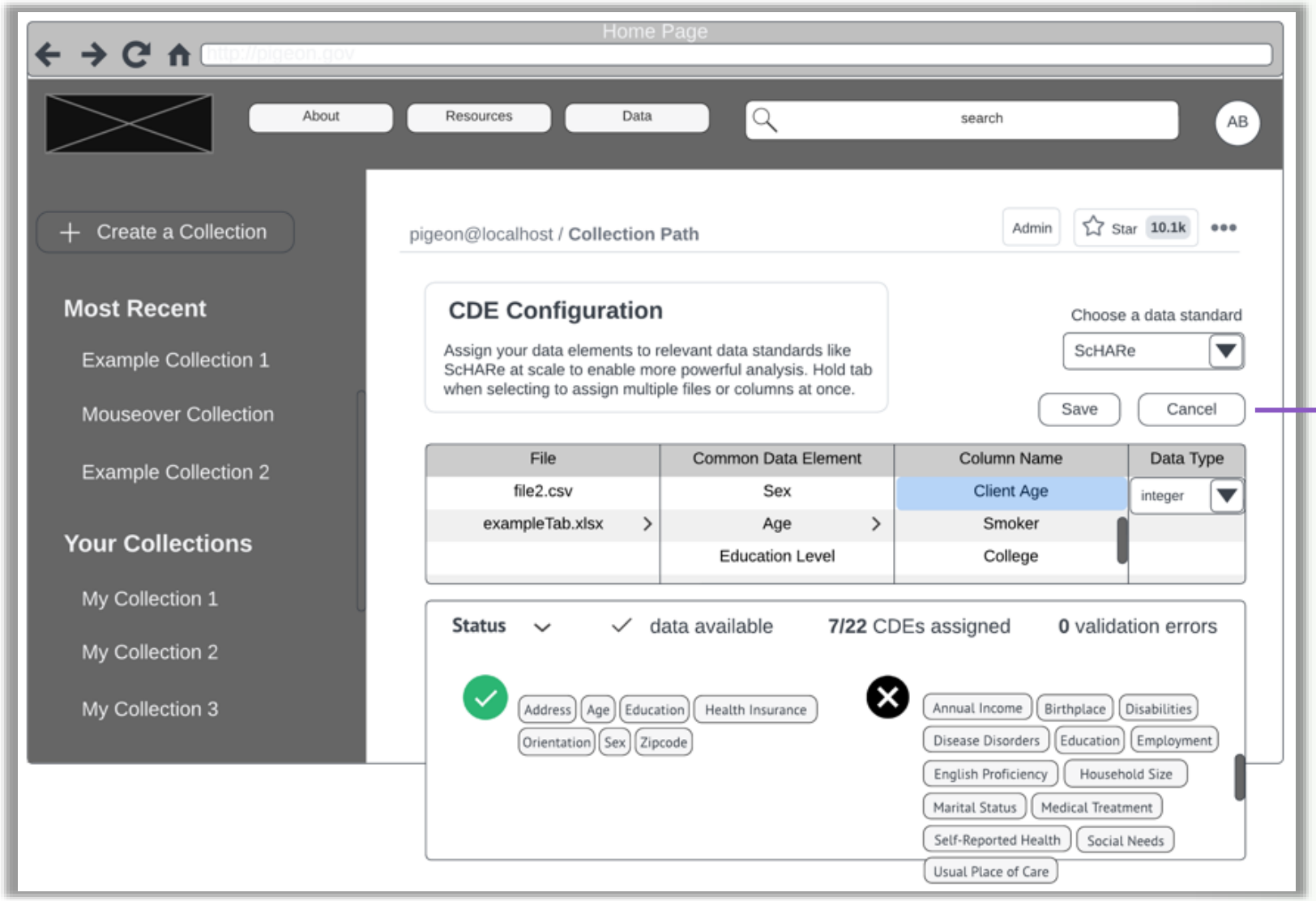
PUBLICLY AVAILABLE FALL 2024



- Host your project data in a **safe space** with privacy levels, secure workspaces, collaboration platform
- **CDE centric**
- **Focus**: Social Science, SDoH, Health Disparities, Health Outcomes Research
- Comply with **NIH Data Management and Data Sharing Policy**
- **Link your data** with others and federated data

ScHARe Repository

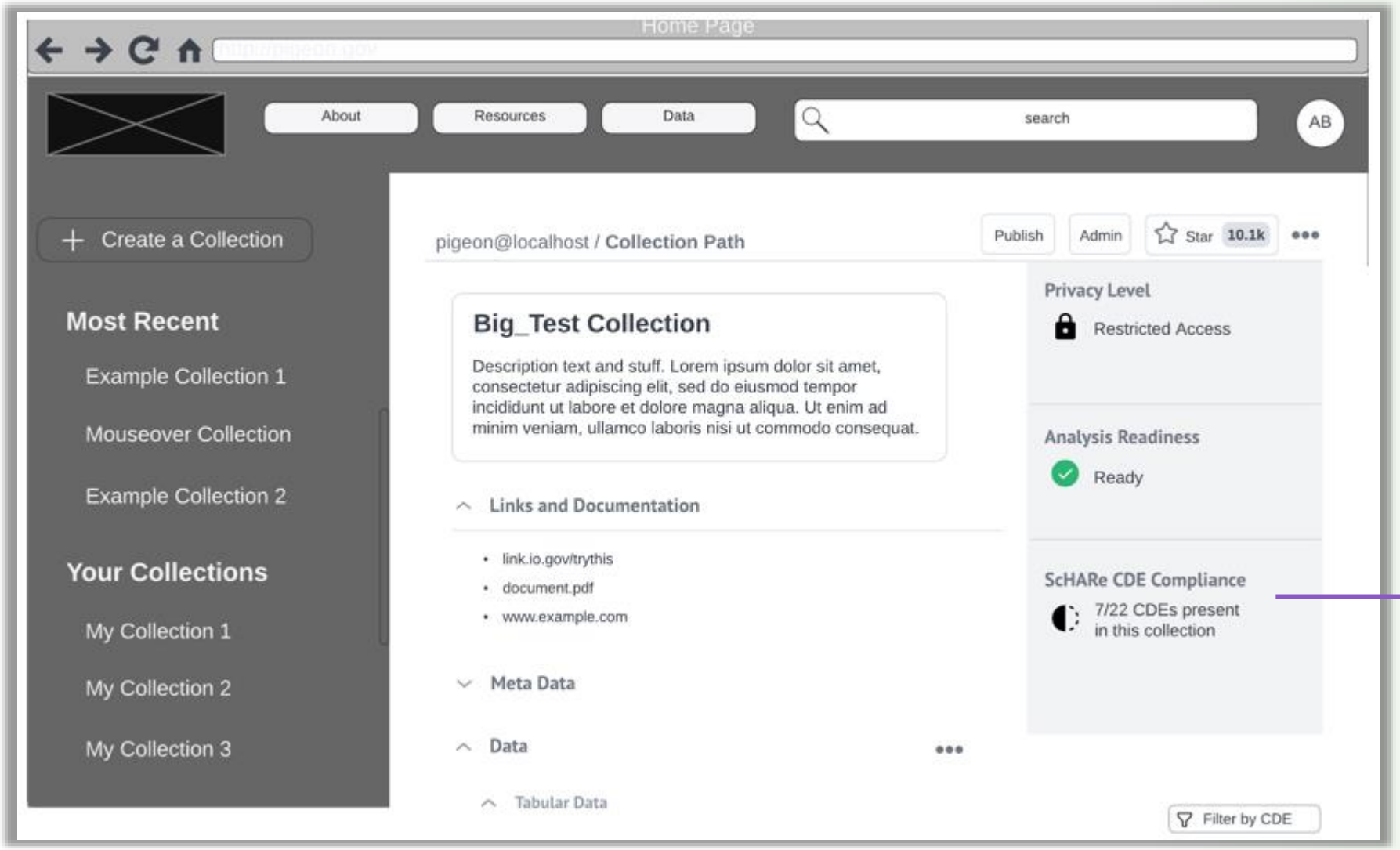
PUBLICLY AVAILABLE FALL 2024



Map project CDEs or variables to ScHARe-PhenX CDEs

PUBLICLY AVAILABLE FALL 2024

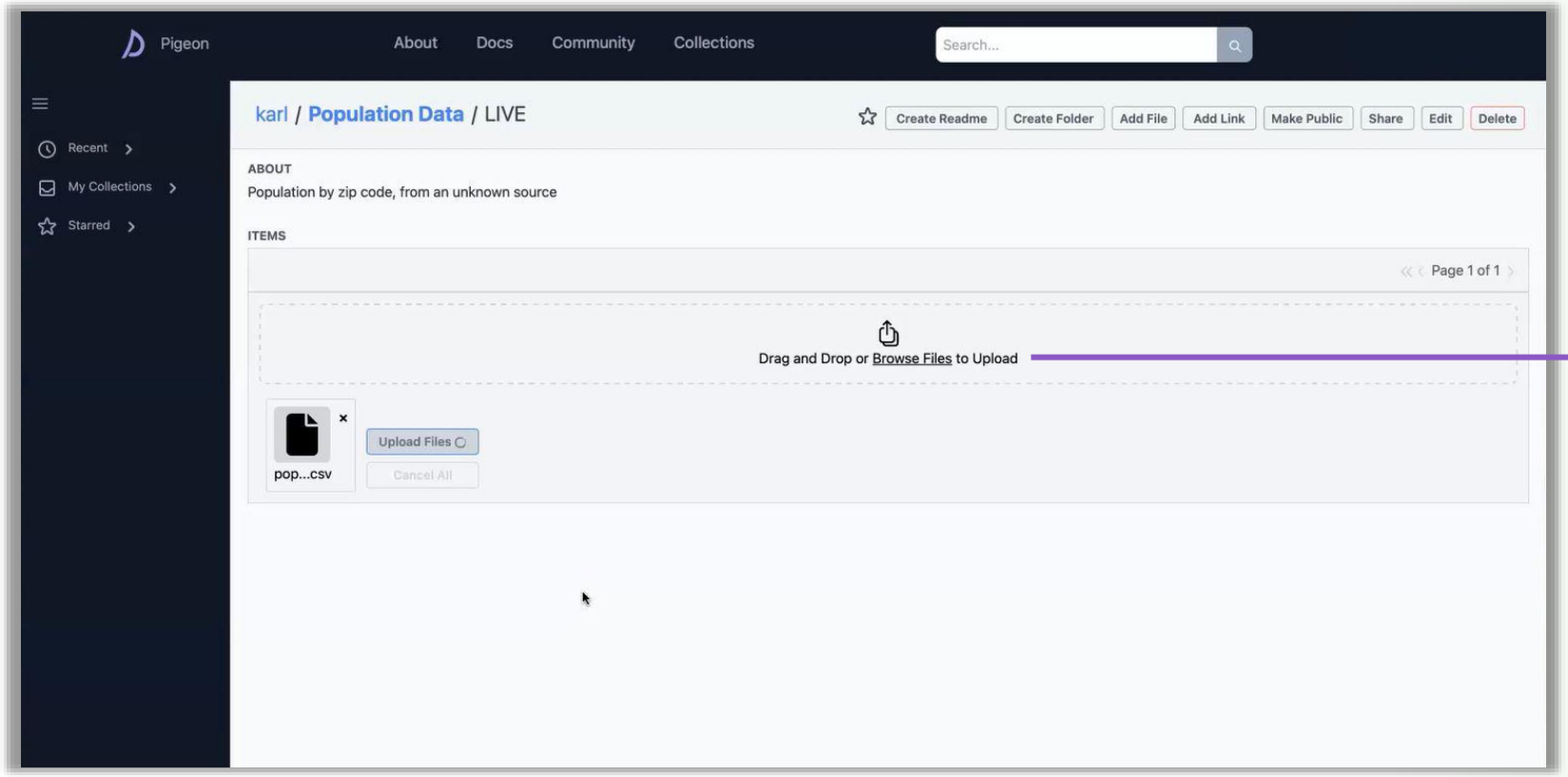
ScHARe Repository



Shows number of project CDEs that match or can map to ScHARe-PhenX CDEs

ScHARe Repository

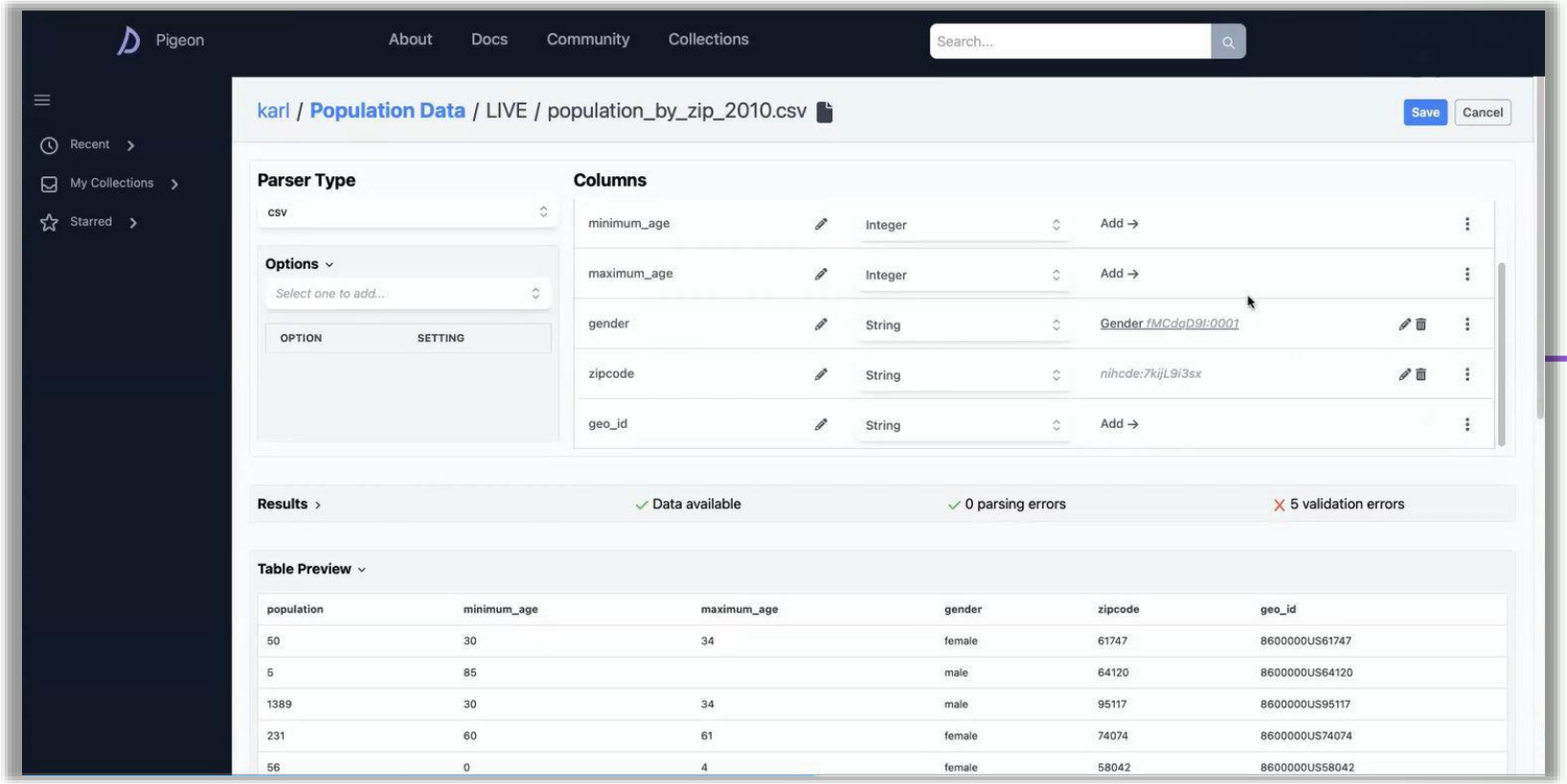
PUBLICLY AVAILABLE FALL 2024



Aggregate datasets with drag-and-drop features

ScHARe Repository

PUBLICLY AVAILABLE FALL 2024



View aggregated dataset

ScHARe

Research Think-a-Thons

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

Generational career & discipline exchange

Think-a-Thons

3rd Wednesday of every month 2 pm

Goals:

- Upskill underrepresented populations in data science and cloud computing
- Foster a research paradigm shift to use Big Data in health disparities/health outcomes research
- Promote use of Dark Data

1. TUTORIAL AND TARGETED THINK-A-THONS

Launched April 2024

- Monthly sessions (2 1/2 hours)
- Instructional/interactive
- Designed for new/experienced users
- Networking
- Mentoring and coaching
- Topics include:
  - Data Science 101
  - Terra
  - Social Determinants of Health analytics
  - Common Data Elements
  - AI readiness
  - Ethical and transparent AI
  - Bias mitigation

2. RESEARCH THINK-A-THONS

- Multi-career (students to senior investigators)
- Multi-discipline (data scientists and researchers)
- Featured datasets with guest experts leads
- Guest experts in topic areas, analytics, data sources etc. to provide guidance
- Generate research idea - decide design, datasets and analytics
- Learn Ethical AI
- Publications

Register:

bit.ly/think-a-thons

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**
May	**Be a Part of the Future of Knowledge Generation 1: AI/Cloud Computing Basics and CDEs**
July	**Be a Part of the Future of Knowledge Generation 2: AI-Ready Datasets and Computations**

SPECIAL EVENTS

- ScHARe for **Educators** (Community Colleges and low-resource MSIs)
- ScHARe for **American Indian/ Alaska Native Researchers**
- ScHARe for **Coders and Programmers** to conduct research

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*

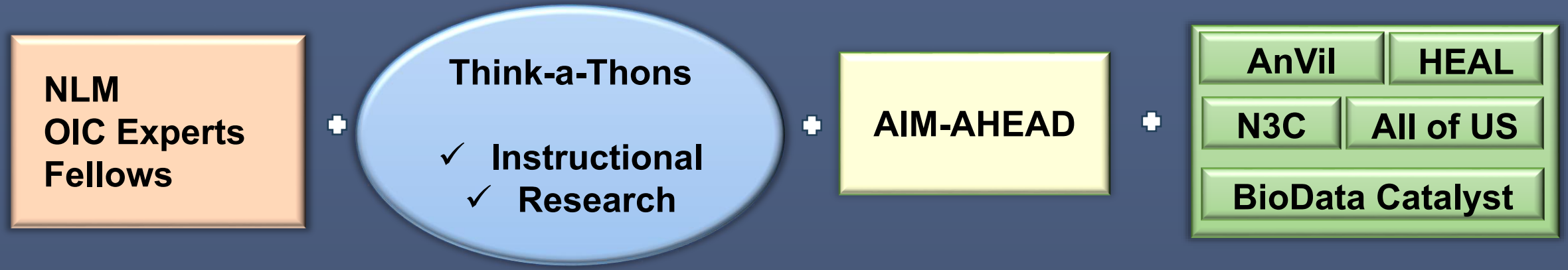
- Metadata and data should be easy to find for both humans and computers
- Ensure that data represents relevant populations

ScHARe

Training pipeline

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Think-a-Thons training/mentoring pipeline



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

Goal: "Upskilling"

- ✓ Data science specialists into health disparities and health outcomes research
- ✓ Health disparities/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 and tool for Community Colleges and low-resource MSIs and organizations