ScHARe

Think-a-Thons

NIH National Institutes of Health

Computational Data Science Strategies

Getting Ready for a Data Science 101 Course

Deborah Duran, PhD · NIMHD
Luca Calzoni, MD MS PhD Cand. · NIMHD
Kenneth Wilkins, PhD · NIDDK

January 17, 2024

ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

NIH National Institute on Minority Health and Health Disparities + NIH Office of Data Science Strategy + NIH National Institute of Nursing Research

Thank you

NIMHD

Dr. Eliseo Perez-Stable

ODSS

Dr. Susan Gregurick

NIH/OD

Dr. Larry Tabak

NINR

Dr. Shannon Zenk

NINR

Rebecca Hawes
Micheal Steele
John Grason

ORWH

OMH

NIMHD OCPL

Kelli Carrington
Thoko Kachipande
Corinne Baker

BioTeam

STRIDES

Terra

SIDEM

RLA

Broad Institute

CDE Working Group

Deborah Duran
Luca Calzoni
Rebecca Hawes
Micheal Steele
Kelvin Choi
Paula Strassle
Deborah Linares
Crystal Barksdale
Gneisha Dinwiddie
Jennifer Alvidrez
Matthew McAuliffe
Carolina Mendoza-Puccini
Simrann Sidhu
Tu Le

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	□	□	□	□

ScHARe

Overview

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 three 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

ScHARe

nimhd.nih.gov/schare

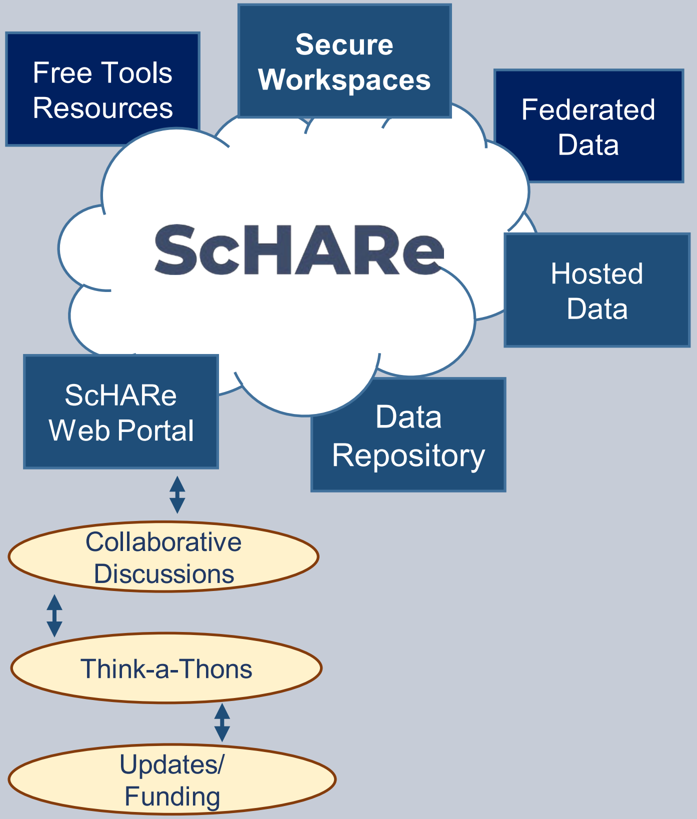
ScHARe Components

ScHARe co-localizes within the cloud:

- **Datasets** (including social determinants of health and social science data) relevant to minority health, health disparities, and health care outcomes research
- **Data repository** to comply with the required hosting, managing, and sharing of data from NIMHD- and NINR-funded research programs
- **Computational capabilities and secure, collaborative workspaces** for students and all career level researchers
- **Tools for collaboratively evaluating and mitigating biases** associated with datasets and algorithms utilized to inform healthcare and policy decisions

Frameworks: Google Platform, Terra, GitHub, NIMHD Web ScHARe Portal

Intramural & Extramural Resource



nimhd.nih.gov/schare

ScHARe Data Ecosystem

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

- **Google Cloud Public Datasets:** publicly accessible, federated, de-identified datasets hosted by Google through the Google Cloud Public Dataset Program

  Example: *American Community Survey (ACS)*

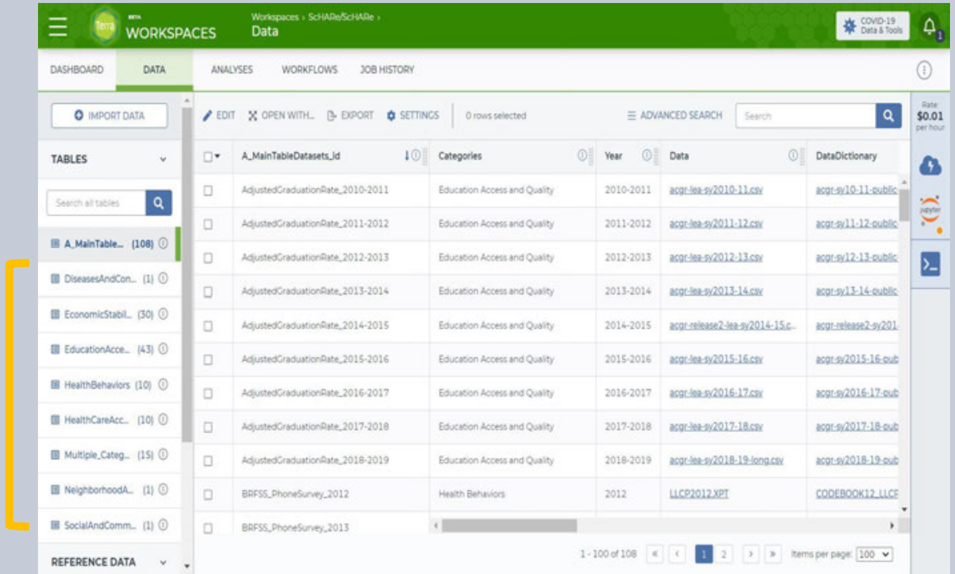
- **ScHARe Hosted Public Datasets:** publicly accessible, de-identified datasets hosted by ScHARe

  Example: *Behavioral Risk Factor Surveillance System (BRFSS)*

- **Funded Datasets on ScHARe:** publicly accessible and controlled-access, funded program/project datasets using Core Common Data Elements shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy

  Examples: *Jackson Heart Study (JHS); Extramural Grant Data; Intramural Project Data*

OVER 240 DATA SETS CENTRALIZED



A_MainTable... (108)
DiseasesAndCon... (1)
EconomicStabil... (30)
EducationAcce... (43)
HealthBehaviors (10)
HealthCareAcc... (10)
Multiple_Categ... (15)
NeighborhoodA... (1)
SocialAndComm... (1)

A_MainTableDatasets_id	Categories	Year	Data	DataDictionary
AdjustedGraduationRate_2010-2011	Education Access and Quality	2010-2011	acgr-lea-sy2010-11.csv	acgr-sy10-11-public
AdjustedGraduationRate_2011-2012	Education Access and Quality	2011-2012	acgr-lea-sy2011-12.csv	acgr-sy11-12-public
AdjustedGraduationRate_2012-2013	Education Access and Quality	2012-2013	acgr-lea-sy2012-13.csv	acgr-sy12-13-public
AdjustedGraduationRate_2013-2014	Education Access and Quality	2013-2014	acgr-lea-sy2013-14.csv	acgr-sy13-14-public
AdjustedGraduationRate_2014-2015	Education Access and Quality	2014-2015	acgr-release2-lea-sy2014-15.c...	acgr-release2-sy201
AdjustedGraduationRate_2015-2016	Education Access and Quality	2015-2016	acgr-lea-sy2015-16.csv	acgr-sy2015-16-pub
AdjustedGraduationRate_2016-2017	Education Access and Quality	2016-2017	acgr-lea-sy2016-17.csv	acgr-sy2016-17-pub
AdjustedGraduationRate_2017-2018	Education Access and Quality	2017-2018	acgr-lea-sy2017-18.csv	acgr-sy2017-18-pub
AdjustedGraduationRate_2018-2019	Education Access and Quality	2018-2019	acgr-lea-sy2018-19-long.csv	acgr-sy2018-19-pub
BRFSS_PhoneSurvey_2012	Health Behaviors	2012	LLCP2012.XPT	CODEBOOK12_LLCF
BRFSS_PhoneSurvey_2013				

1 - 100 of 108 « ‹ 1 2 › » Items per page: 100

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**

Users will be able to **map and link** across datasets

ScHARe Data Ecosystem Structure

FEDERATED
PUBLIC DATA
240+

Hosted by Google
& ScHARe

REPOSITORY

CDE FOCUSED

CDEs enhances Data Interoperability (Aggregation) by using semantic standards and concept codes

Innovative Approach:

CDE Concept Codes Uniform Resource Identifier (URI)

What is a CDE?

1. Name Last
First
2. Date of Birth (mm/dd/yyyy)
5. Social Security Number
dress: Line 1: Street/RFD#, P.O. Box,
learly label Apartment, Company, Sui

A common data element (CDE) is a standardized, precisely defined question that is paired with a set of specific allowable responses, that is then used systematically across different sites, studies, or clinical trials to ensure consistent data collection

ScHARe CDEs Labels

For FUNDED PROJECT DATA – Common Data Elements Centralized for Interoperability and Data Sharing

- 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 (Associated Medications/Treatments)

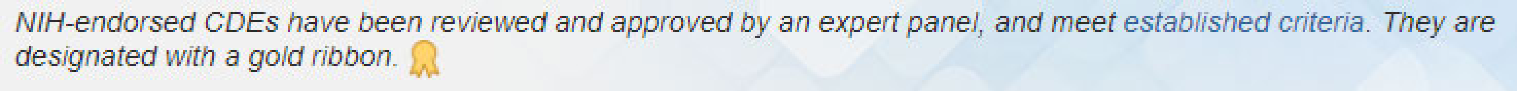
**NIMHD Framework
**Health Disparity Outcomes

NIH Endorsed

(** project level CDE)

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

Cross-walked with PhenX SDoH



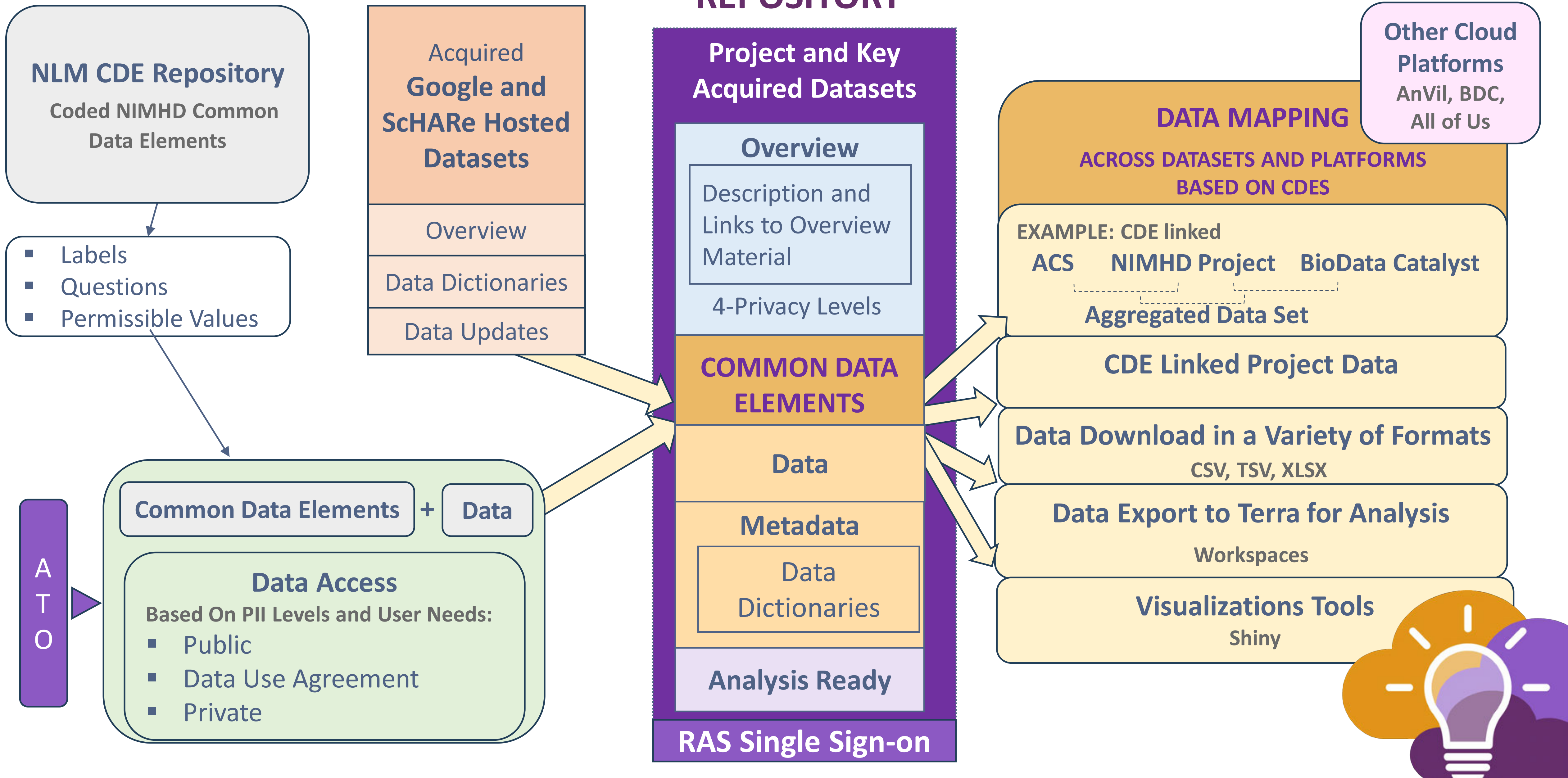
NIH-endorsed CDEs have been reviewed and approved by an expert panel, and meet established criteria. They are designated with a gold ribbon.

ScHARe REPOSITORY

COMMON DATA ELEMENTS

DATA UPLOAD

DATA MAPPING, DOWNLOAD AND EXPORT



NLM CDE Repository

Coded NIMHD Common Data Elements

- Labels
- Questions
- Permissible Values

Common Data Elements + Data

Data Access

Based On PII Levels and User Needs:

- Public
- Data Use Agreement
- Private

A T O

Acquired **Google and ScHARe Hosted Datasets**

Overview

Data Dictionaries

Data Updates

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

DATA MAPPING

ACROSS DATASETS AND PLATFORMS BASED ON CDES

Other Cloud Platforms

AnVil, BDC, All of Us

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

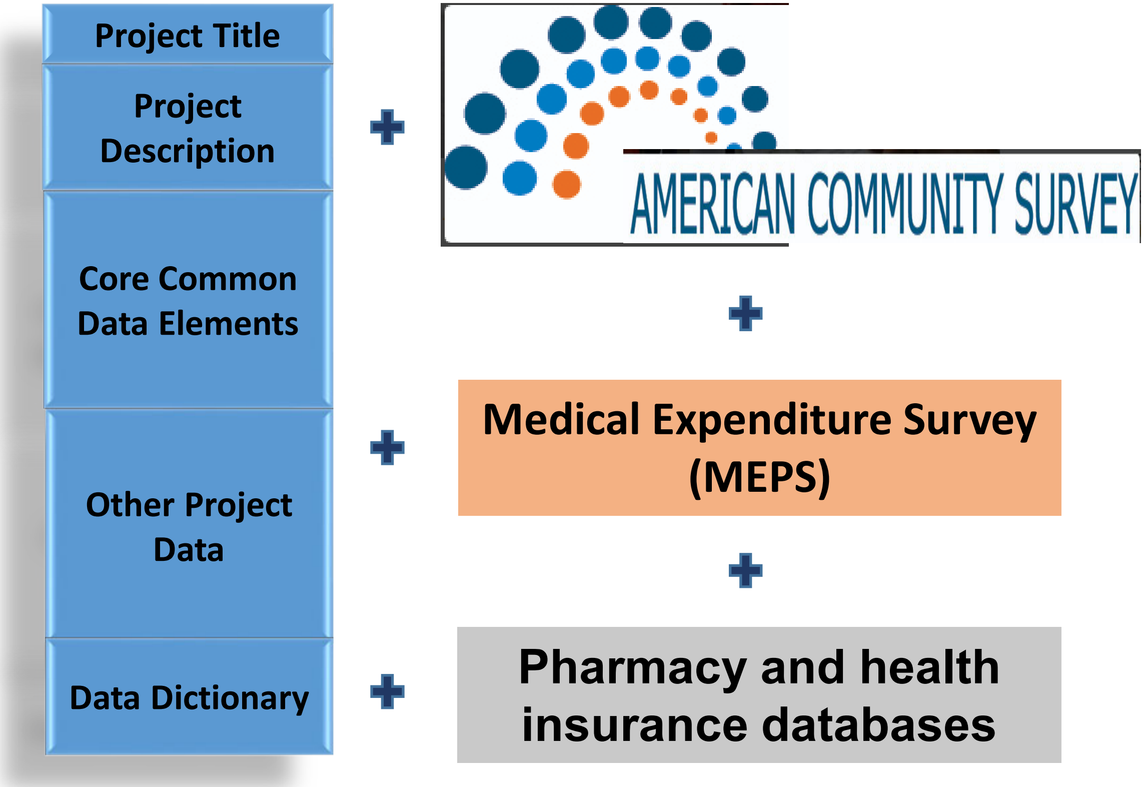
Workspaces

Visualizations Tools

Shiny

ScHARe

Project & federated dataset mapping



Project Title

Project Description

Core Common Data Elements

Other Project Data

Data Dictionary

+

AMERICAN COMMUNITY SURVEY

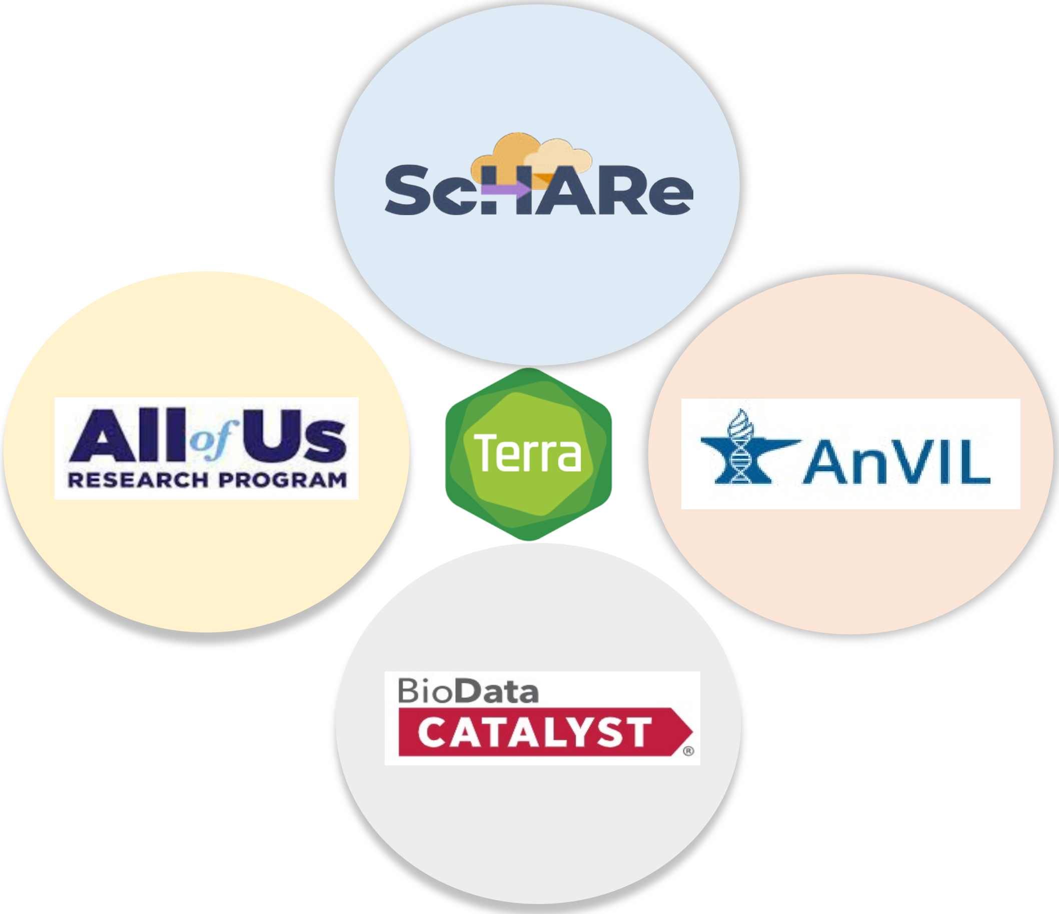
+

Medical Expenditure Survey (MEPS)

+

Pharmacy and health insurance databases

Mapping across cloud platforms

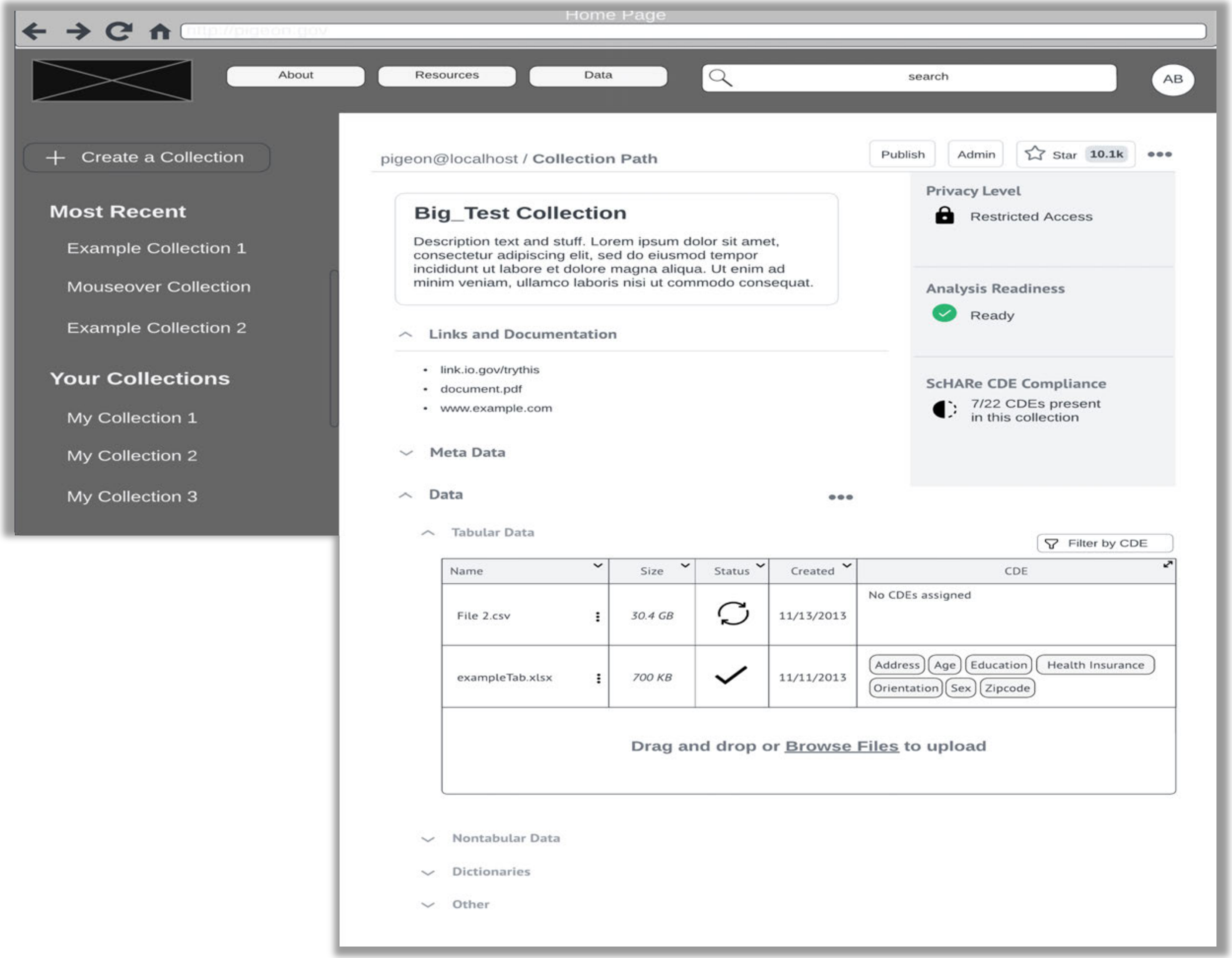


UPCOMING

ScHARe

Repository CDE Focused for Data Interoperability

Coming Soon



+ Create a Collection

Most Recent

- Example Collection 1
- Mouseover Collection
- Example Collection 2

Your Collections

- My Collection 1
- My Collection 2
- My Collection 3

pigeon@localhost / **Collection Path**

Big_Test Collection

Description text and stuff. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, ullamco laboris nisi ut commodo consequat.

Privacy Level

Restricted Access

Analysis Readiness

Ready

ScHARe CDE Compliance

7/22 CDEs present in this collection

Links and Documentation

- link.io.gov/trythis
- document.pdf
- www.example.com

Meta Data

Data

Tabular Data

Name	Size	Status	Created	CDE
File 2.csv	30.4 GB	⟳	11/13/2013	No CDEs assigned
exampleTab.xlsx	700 KB	✓	11/11/2013	Address Age Education Health Insurance Orientation Sex Zipcode

Drag and drop or Browse Files to upload

Nontabular Data

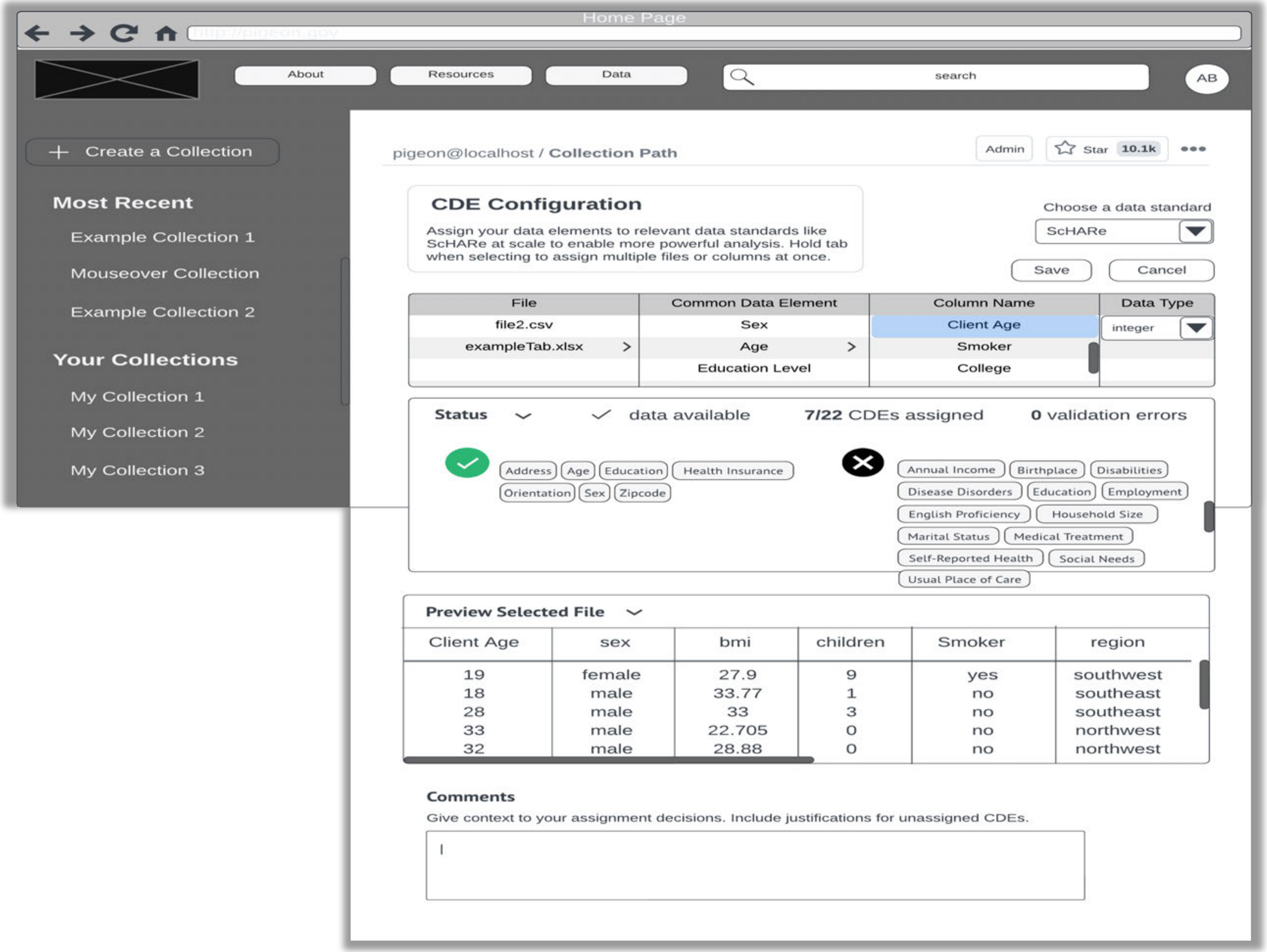
Dictionaries

Other

ScHARe

Repository CDE Focused for Data Interoperability

Coming Soon



pigeon@localhost / **Collection Path**

CDE Configuration

Assign your data elements to relevant data standards like ScHARe at scale to enable more powerful analysis. Hold tab when selecting to assign multiple files or columns at once.

Choose a data standard: ScHARe

File	Common Data Element	Column Name	Data Type
file2.csv	Sex	Client Age	integer
exampleTab.xlsx >	Age >	Smoker	
	Education Level	College	

Status ✓ data available **7/22** CDEs assigned **0** validation errors

✓ Address, Age, Education, Health Insurance, Orientation, Sex, Zipcode

✕ Annual Income, Birthplace, Disabilities, Disease Disorders, Education, Employment, English Proficiency, Household Size, Marital Status, Medical Treatment, Self-Reported Health, Social Needs, Usual Place of Care

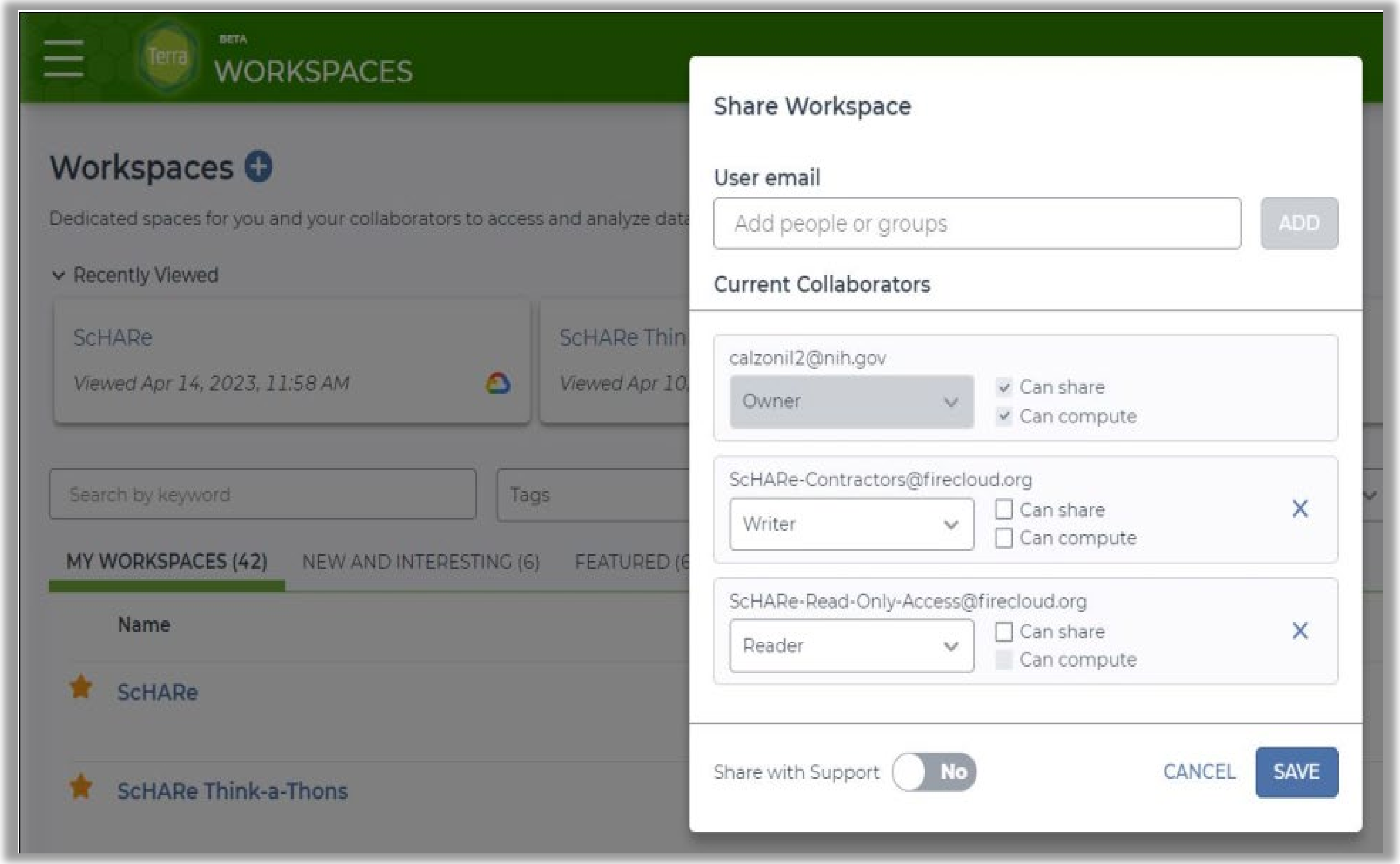
Preview Selected File

Client Age	sex	bmi	children	Smoker	region
19	female	27.9	9	yes	southwest
18	male	33.77	1	no	southeast
28	male	33	3	no	southeast
33	male	22.705	0	no	northwest
32	male	28.88	0	no	northwest

Comments

Give context to your assignment decisions. Include justifications for unassigned CDEs.

Secure workspace



Workspaces

Dedicated spaces for you and your collaborators to access and analyze dat

Recently Viewed

ScHARe
Viewed Apr 14, 2023, 11:58 AM

ScHARe Thin
Viewed Apr 10,

ScHARe

ScHARe Think-a-Thons

Share Workspace

User email

Add people or groups

Current Collaborators

calzonil2@nih.gov
Owner
☑ Can share
☑ Can compute

ScHARe-Contractors@firecloud.org
Writer
□ Can share
□ Can compute

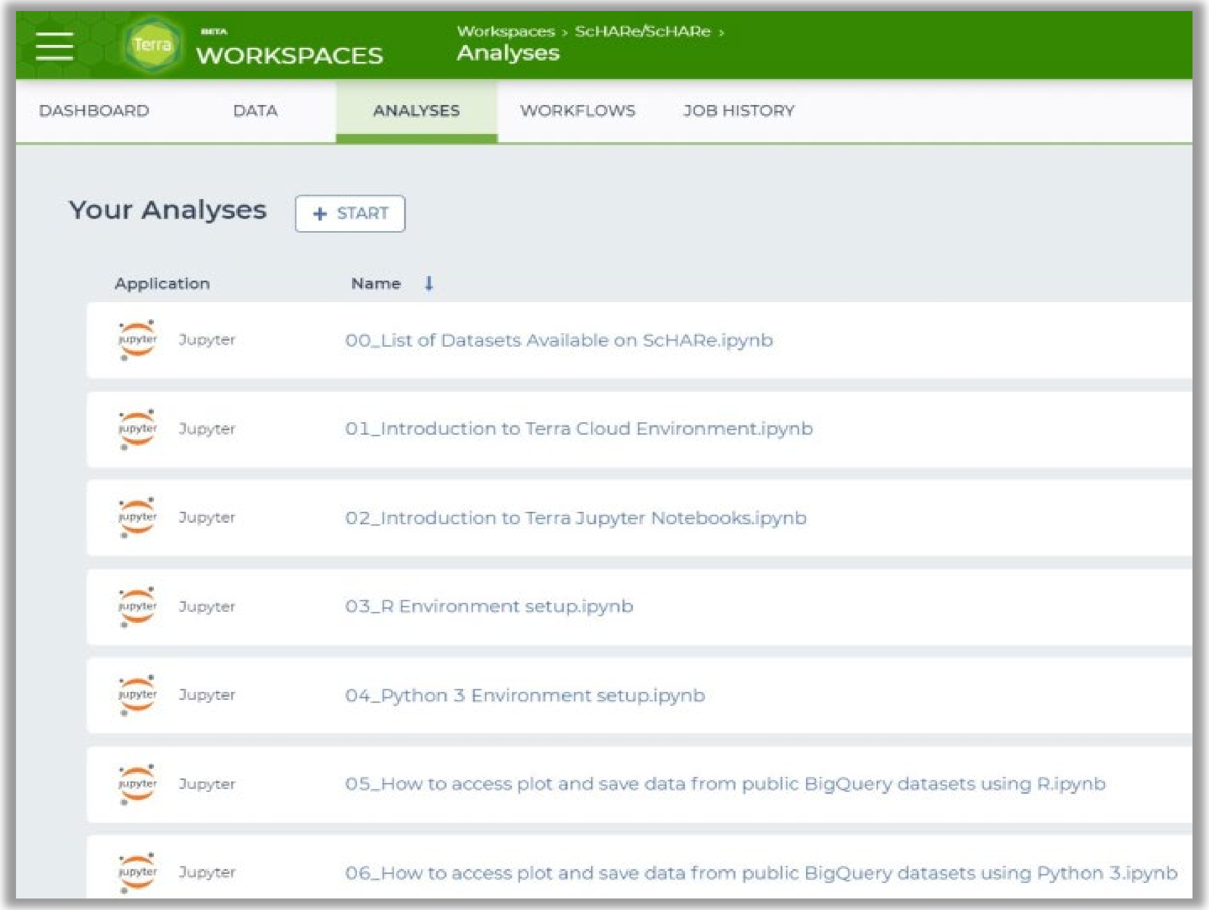
ScHARe-Read-Only-Access@firecloud.org
Reader
□ Can share
□ Can compute

Share with Support No

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

Notebooks analytics

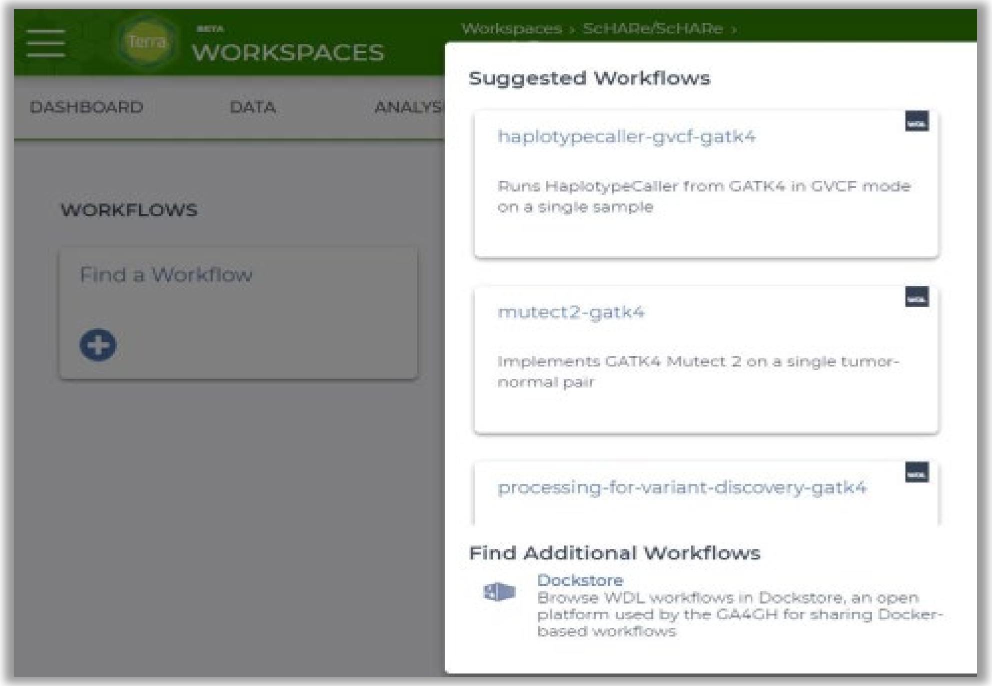
Workflows - Modular codes



Your Analyses

Application	Name ↓
Jupyter	00_List of Datasets Available on ScHARe.ipynb
Jupyter	01_Introduction to Terra Cloud Environment.ipynb
Jupyter	02_Introduction to Terra Jupyter Notebooks.ipynb
Jupyter	03_R Environment setup.ipynb
Jupyter	04_Python 3 Environment setup.ipynb
Jupyter	05_How to access plot and save data from public BigQuery datasets using R.ipynb
Jupyter	06_How to access plot and save data from public BigQuery datasets using Python 3.ipynb

- **Copy and paste analytics**



Suggested Workflows

haplotypecaller-gvcf-gatk4

Runs HaplotypeCaller from GATK4 in GVCF mode on a single sample

mutect2-gatk4

Implements GATK4 Mutect 2 on a single tumor-normal pair

processing-for-variant-discovery-gatk4

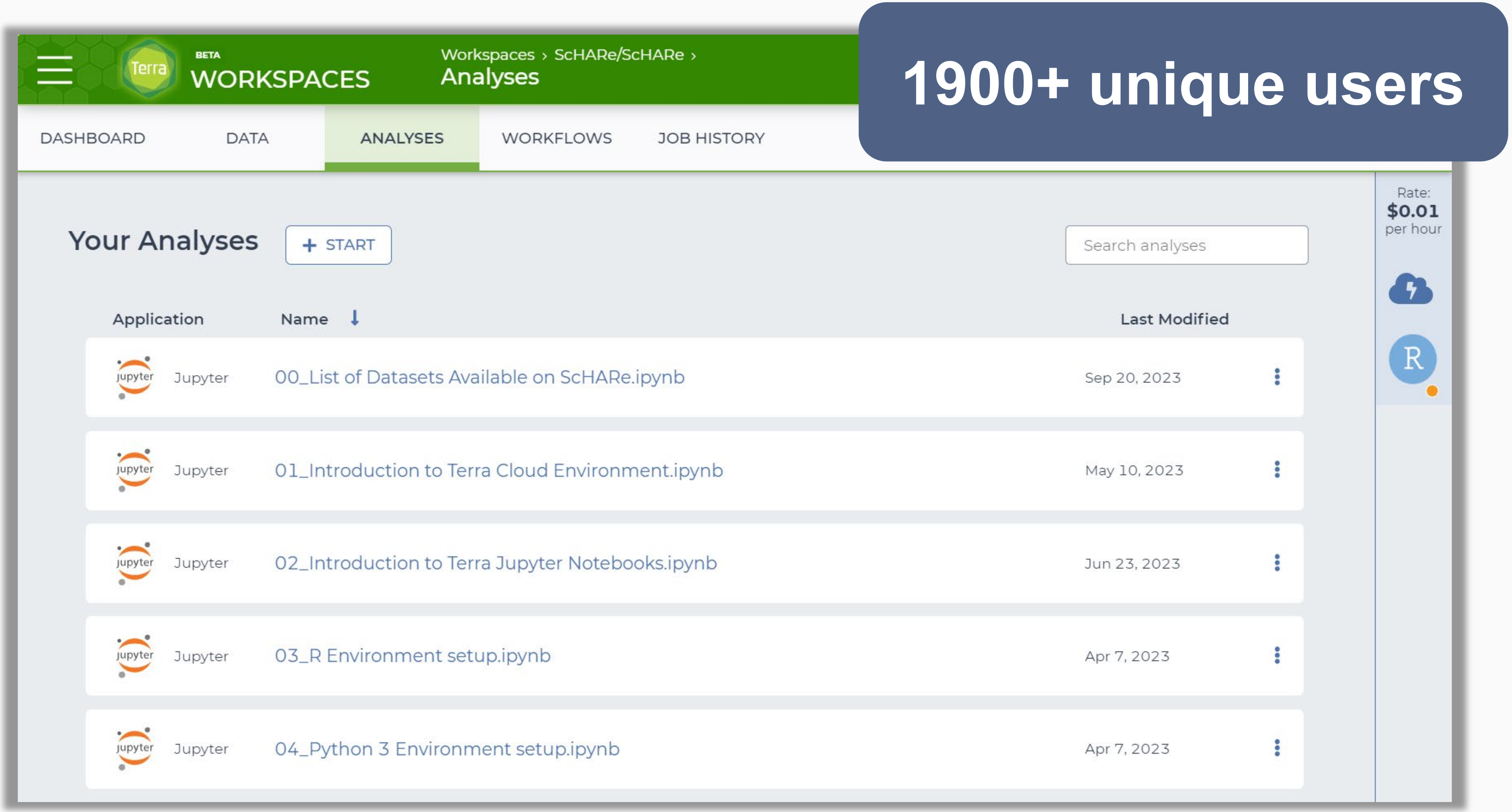
Find Additional Workflows

Dockstore

Browse WDL workflows in Dockstore, an open platform used by the GA4GH for sharing Docker-based workflows

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

ScHARe Registrations



Your Analyses

Application	Name ↓	Last Modified
Jupyter	00_List of Datasets Available on ScHARe.ipynb	Sep 20, 2023
Jupyter	01_Introduction to Terra Cloud Environment.ipynb	May 10, 2023
Jupyter	02_Introduction to Terra Jupyter Notebooks.ipynb	Jun 23, 2023
Jupyter	03_R Environment setup.ipynb	Apr 7, 2023
Jupyter	04_Python 3 Environment setup.ipynb	Apr 7, 2023

1900+ unique users

ScHARe

Think-a-Thons

NIH National Institutes of Health

Think-a-Thon Tutorials

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**
	ScHARe for Educators (Community Colleges & Low Resource MSIs)
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**
	ScHARe for American Indian / Alaska Native Researchers
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**
	ScHARe for Coders and Programmers to conduct Research (Jan 31)

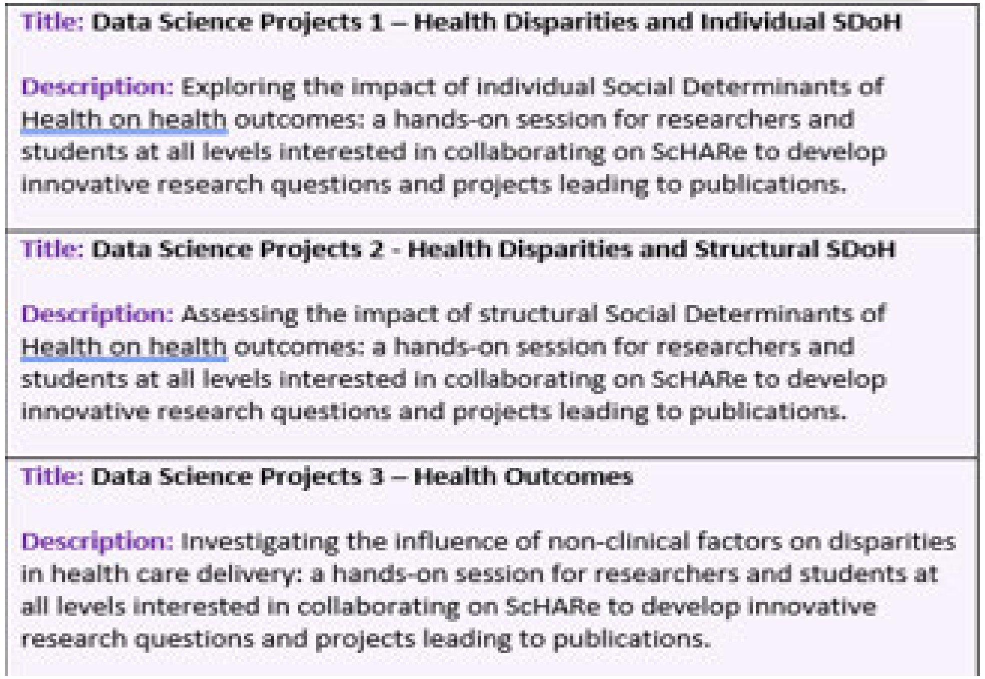
bit.ly/think-a-thons

Upcoming

ScHARe

Think-a-Thons (TaT)

Research Teams



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

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

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

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

Title: Data Science Projects 3 – Health Outcomes

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

- **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

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 Guest expert

Kenneth J. Wilkins, PhD

NIH/NIDDK

About Ken

Ken is a former mathematics and computer science high school teacher who found his way into biostatistics.

He worked for two decades across sectors in biomedical research, and he is now working with both NIH-employed intramural and NIH-funded extramural researchers in his NIH/NIDDK and trans-NIH roles.

His research interests encompass evolving data methods to better suit researchers’ posed questions given limitations in data and data-interoperability standards.