ScHARe

Think-a-Thons

NIH National Institutes of Health

Preparing for AI-driven Research on ScHARe - Part 2

A Comprehensive Review and Brainstorming Session

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

March 20, 2024

ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

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

National Institute of Dental and Craniofacial Research

U.S. Department of Health and Human Services Office of Minority Health

National Institutes of Health Office of Research on Women's Health

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

NIDCR

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	☐	☐	☐	☐

Outline

5’ Introduction

- Experience poll

15’ ScHARe overview

- Interest poll

35’ Computational strategies

- Polls

20’ Python data science libraries

15’ Testing and monitoring in algorithm development

15’ Open science and reproducible research

45’ Research Think-a-Thons brainstorming

- Final poll

ScHARe

Overview

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

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

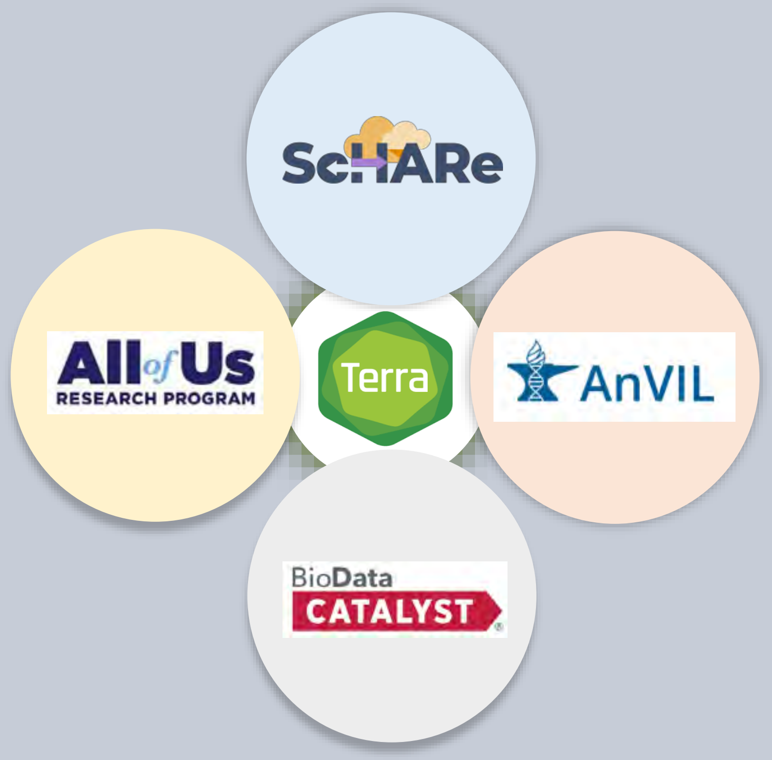
ScHARe

PREPARING FOR AI RESEARCH AND HEALTHCARE USING BIG DATA

Mapping across cloud platforms with Terra Interface

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

BE A PART OF THE FUTURE OF KNOWLEDGE GENERATION

Data Ecosystem structure

Population Science/SDoH

240+
FEDERATED PUBLIC DATASETS

- **Population Science / SDoH / Behavioral**
- Hosted by Google & ScHARe

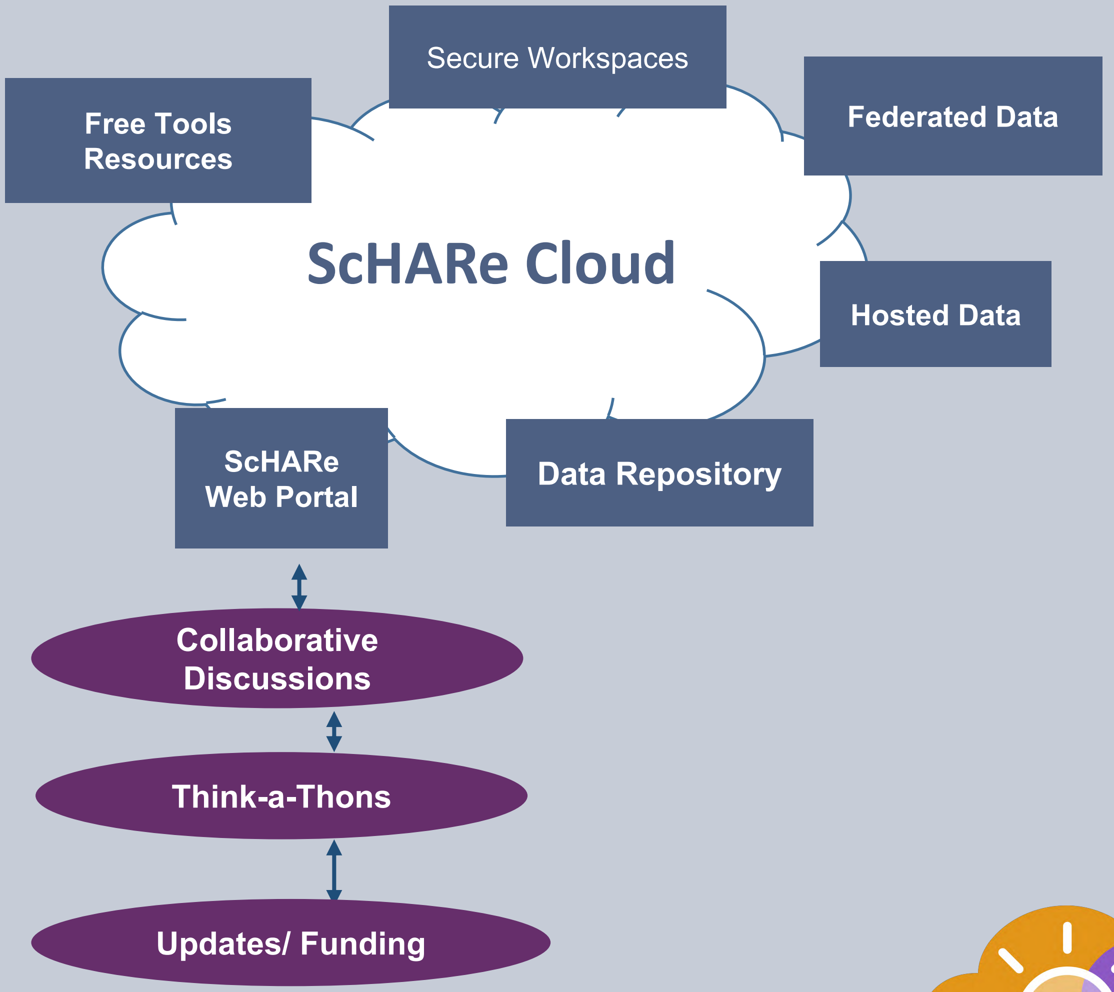
REPOSITORY
CDE
FOCUSED

- **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 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)*

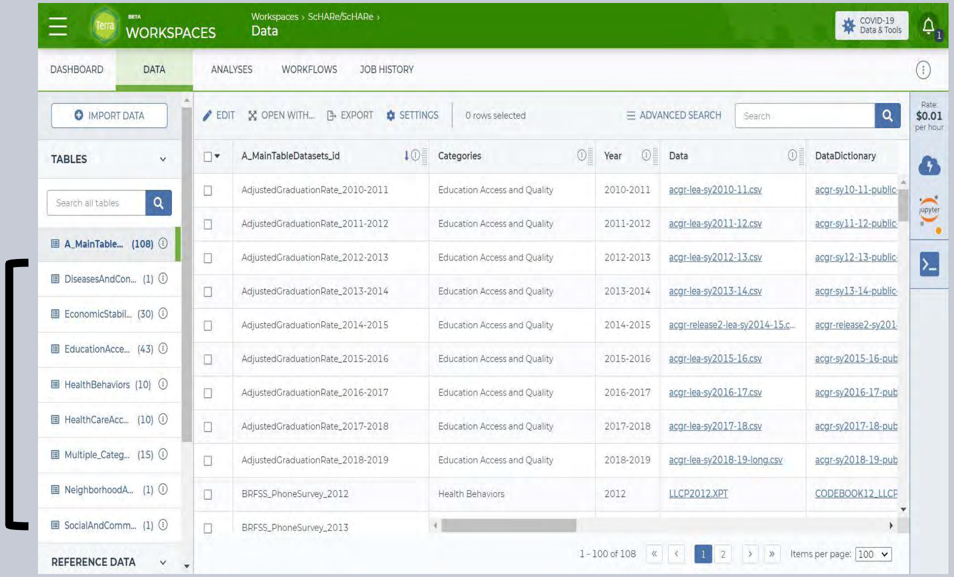
- **ScHARe Hosted Public Datasets:** publicly accessible, de-identified datasets hosted by ScHARe

  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



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 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)

ScHARe Ecosystem: ScHARe hosted datasets

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

200+ 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.

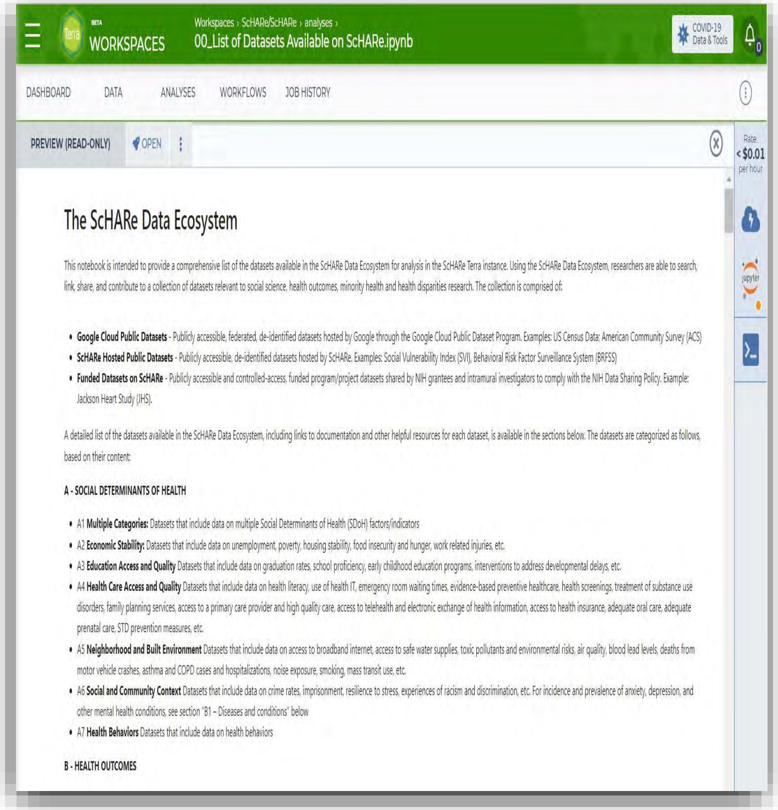
Examples:

- **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

Terra Interface: Datasets and Access to Data

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

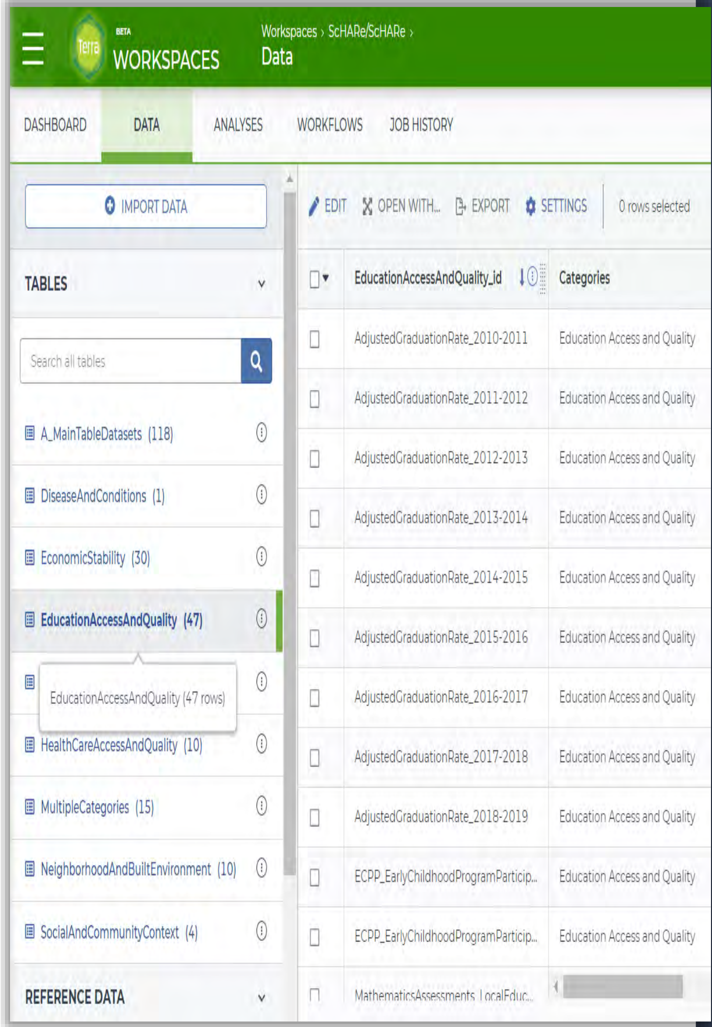
What?



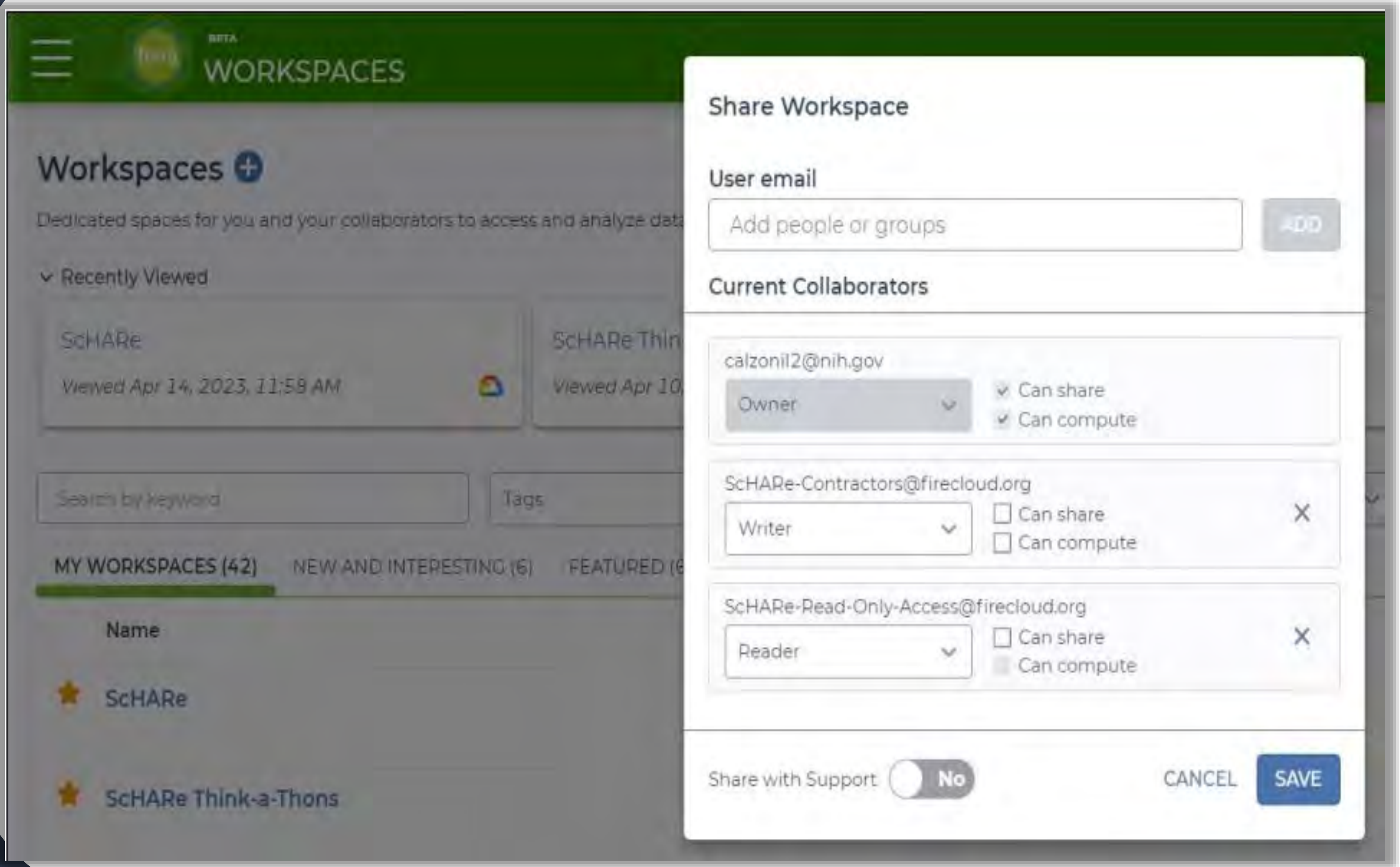
Where?

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

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



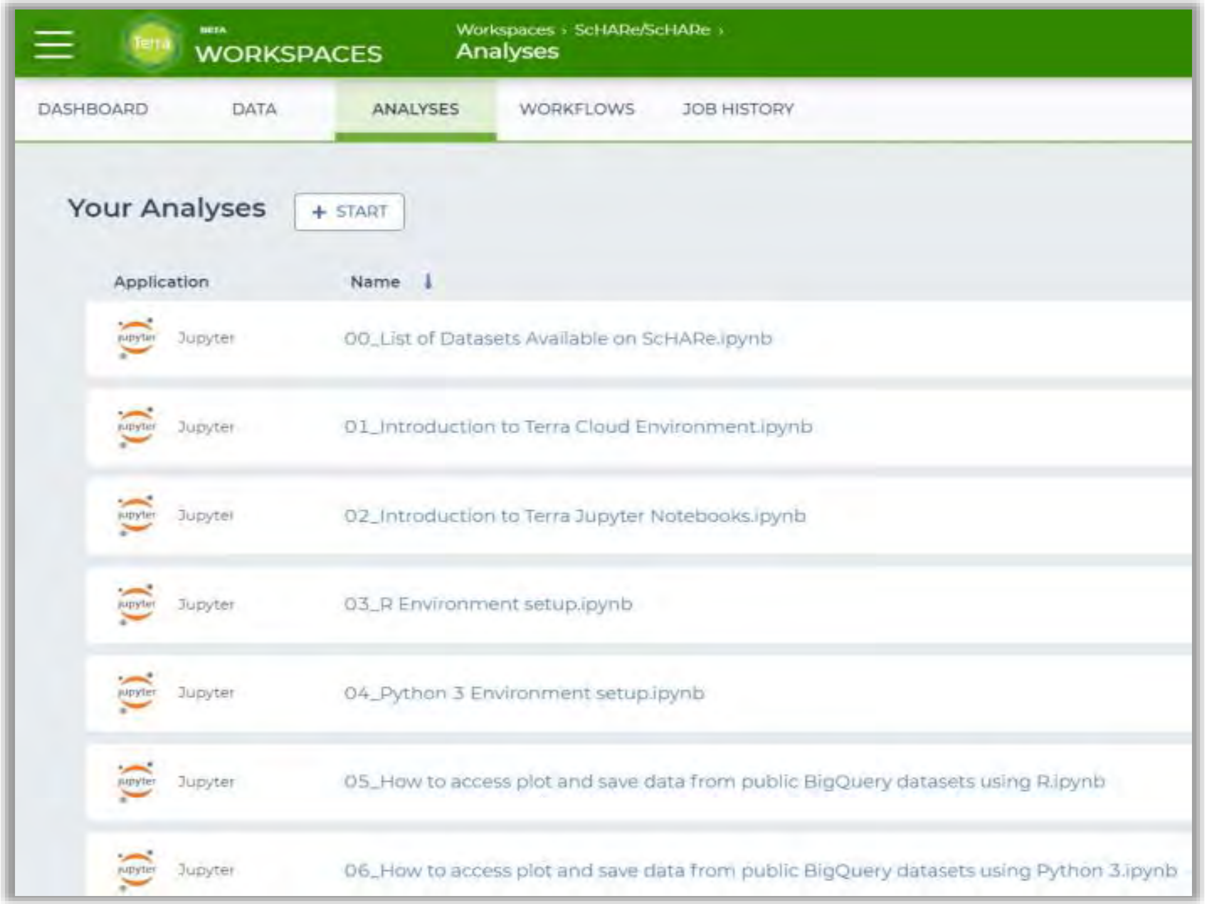
Terra Interface: Secure workspace



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

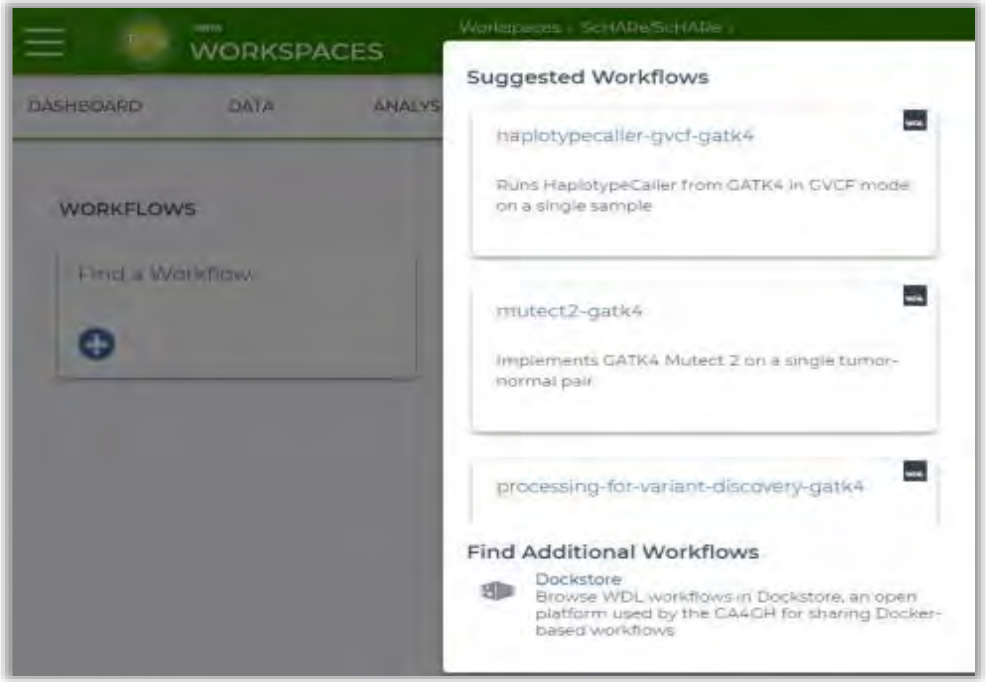
Terra Interface: Notebooks for Analytics & Tutorials

A notebook integrates code and its output into a single document where you can run code, display the output, and also add explanations, formulas, and charts



Workflows Modular codes

Easy to Use--Cut and Paste Analytics



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

ScHARe

PREPARING FOR CLOUD COMPUTING AI – HEALTH DISPARITY AND CARE RESEARCH USING BIG DATA

AGGRAGATING DATA SETS TOOL: Variables & CDEs

CDE Mapping Project & Federated Data

Project Title

Project Description

"Core Common Data Elements"

Other Project Data

Data Dictionary

+

AMERICAN COMMUNITY SURVEY

+

BRFSS

+

Medical Expenditure Survey (MEPS)

*****Assign DOIs to aggregated data sets***

CDE Mapping across Program Projects

NIMHD Grantee Data Proj 1

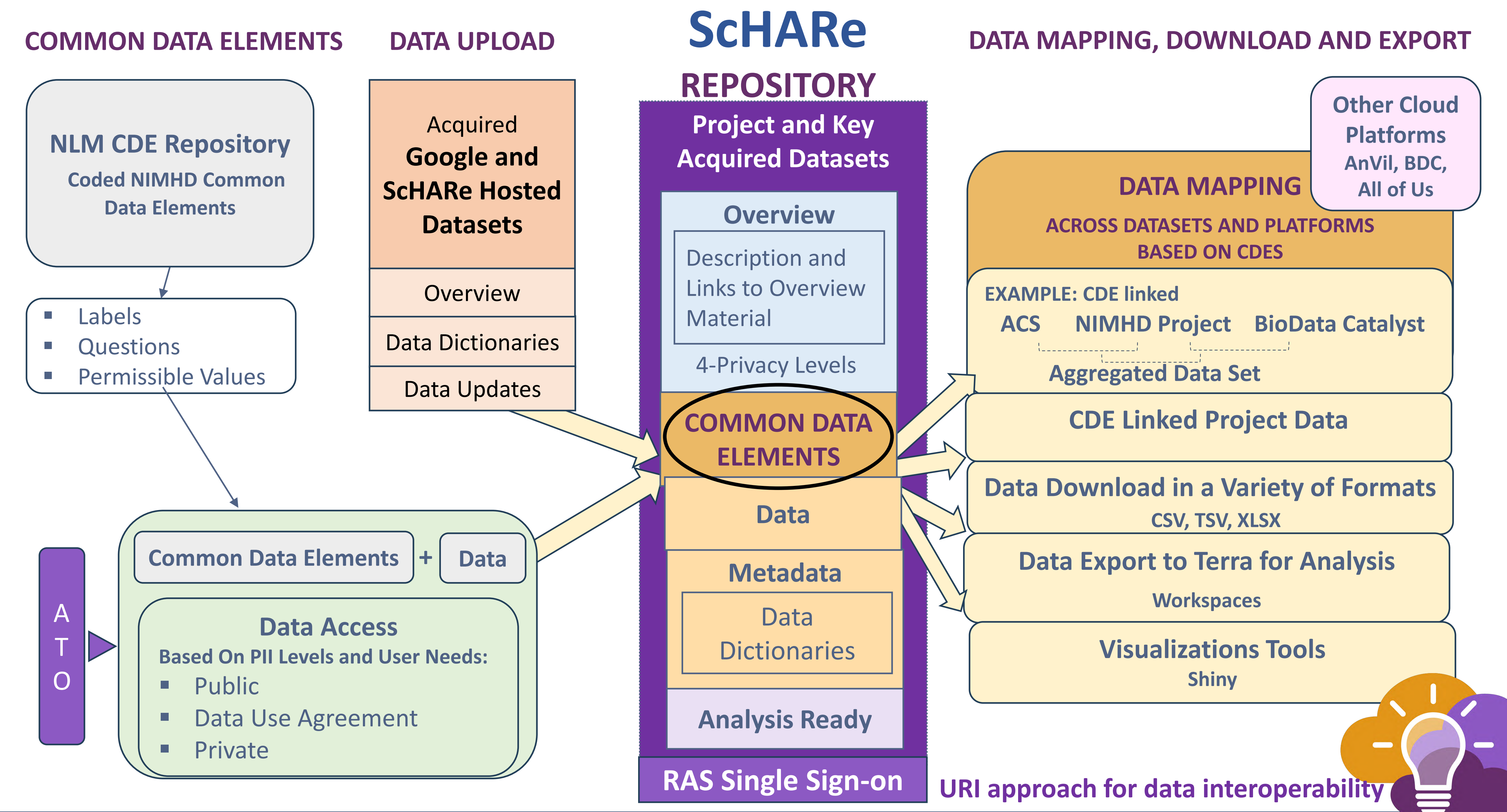
ICO Grantee Data Proj 4

ScHARe CDEs

Contract Data Proj 2

Intramural Data Proj 3

BE A PART OF THE FUTURE OF KNOWLEDGE GENERATION


COMMON DATA ELEMENTS
DATA UPLOAD
ScHARe
REPOSITORY
DATA MAPPING, DOWNLOAD AND EXPORT
NLM CDE Repository
Coded NIMHD Common Data Elements
Labels
Questions
Permissible Values
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
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
Common Data Elements
+
Data
Data Access
Based On PII Levels and User Needs:
Public
Data Use Agreement
Private
A
T
O
URI approach for data interoperability

ScHARe

CDE Adoption

- Making Data Interoperable (URI Approach)
- Concept Code Mapping (Data Harmonization)

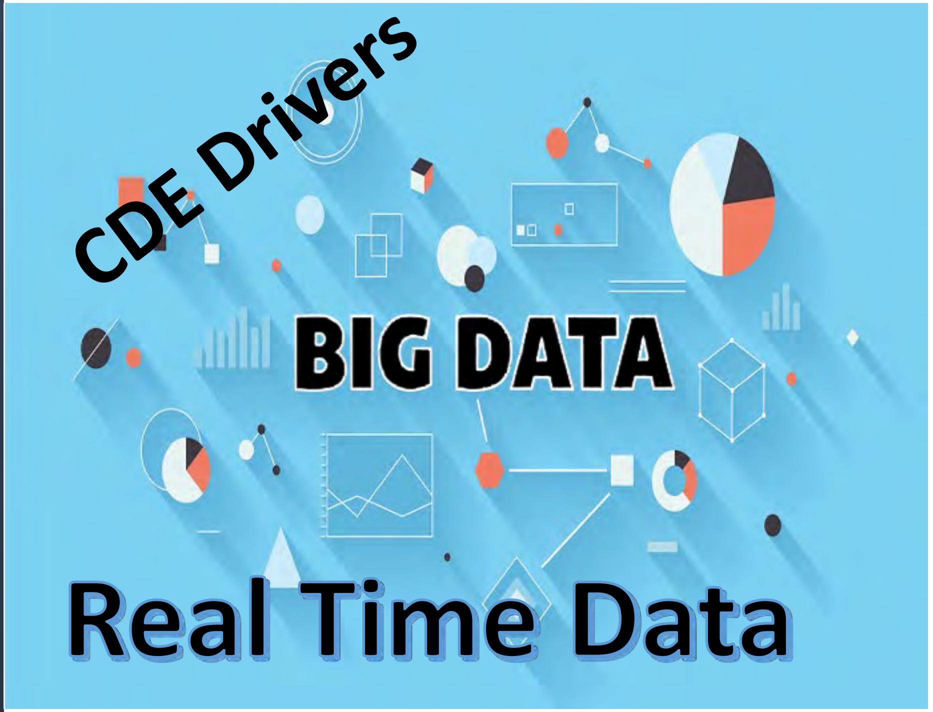
BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

ScHARe

Adopted CDEs to:

- Standardize data for people & computers (human and machine readable)
- Enable data sharing across studies (data interoperability)
- Enhance data interpretation & analysis (semantically defined and standardized coded)
- Simplify collaboration
- Reduces project start-up & results time

***BIG DATA AND AI*: REQUIRES NEW APPROACHES FOR COLLECTION, MANAGEMENT, ANALYSIS**



Covid revealed the need to have real time data

ScHARe

"CORE" CDE Development

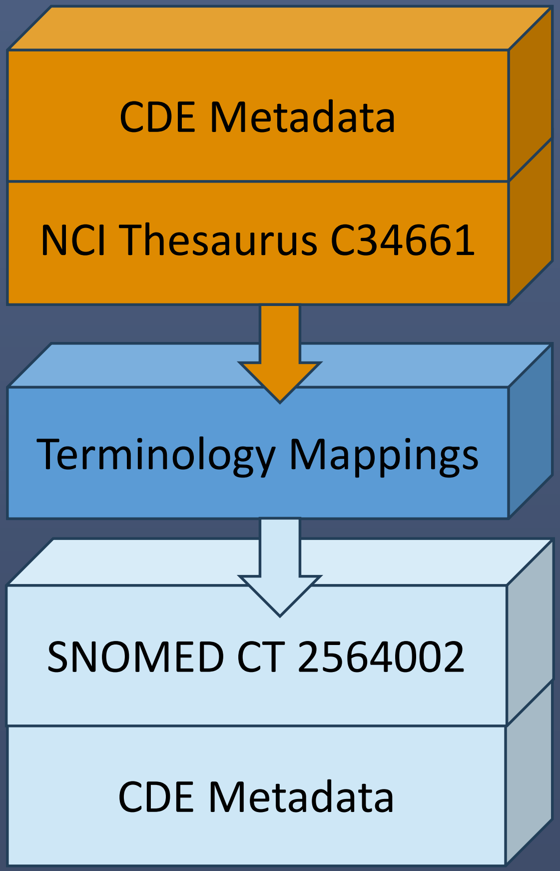
Core Set:

- Few critical questions required from all studies/sites
- Minimal burden
- Allows for questions to be asked in any way, but reported in a standardized format
- Allows for any number of other questions to be collected as collector chooses

Criteria:

- PhenX Toolkit first
- Validated source
- Adaptation of a validated source
- Generate new gap area CDE

Importance of Concept Code *Mapping* and 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

Questions become CDEs When Defined and Coded

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 (words).

Human Readable w Shared Meaning

Codes facilitate Machine Readable

CDES: Words Precisely Defined-Shared Meaning

Words can be **SEMANTICALLY AMBIGUOUS**.

- Context is important in conveying meaning when using CDEs
  - Words have different meanings depending on words around it and context.
- Some examples:
  - **Seizure:** uncontrolled electrical activity between brain cells / spiritual experience?
  - **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?

Words can mean different things in different contexts

Questions become CDEs When Defined and Coded

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 (words).

Human Readable w Shared Meaning

Machine-Readable format—**excel spreadsheet: codes increase interoperability and 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 |

Some Concept Coding Systems
One NOT 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

How a Survey Question Became a CDE

Please select the racial category or <u>categories</u> 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

Survey Questions become CDEs when they are:

- **semantically defined by a standardized coding system for shared meaning**
- **in a format that is human and machine readable for ease of reuse**

Making of a CDE from a Protocol/Question

Need a standardized defined concept and related code.
Source: NCI Thesaurus

Race/Ethnicity Self-Identification

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

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

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

Making of a CDE from a Protocol/Question

- 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

Making of a CDE from a Protocol/Question

Need a standardized defined concept and related code. Source: NCI Thesaurus

Code Mapping

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

Matched CDE

Income (Project 1)

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

Reported this way

Income (Project 2)

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

Collected this way

Mappable CDE

Income (Project 1)

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

Reported this way

Income (Project 2)

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

Collected this way

Mapped using algorithms

ScHARe

CDEs in Survey Format

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

1. Age

What is the person's age? (collapse data over 89 yrs old / 2 yrs and under, report in months-does not exclude asking full birthdate)

__________ ☐ years ☐ months

Project 5 Covid-19 Age https://cde.nlm.nih.gov/cde/search?q=PROJECT%205&nihEndorsed=true

2. Birthplace

Where were you born?

☐ In the United States, including U.S. Territories (Puerto Rico, Guam, U.S. Virgin Islands, American Samoa and Northern Mariana Islands) (**Select from Drop Down-not doable on word doc**)

☐ Outside the United States (**Select from Drop Down-ISSO categories-not doable on word doc**)

PhenX – Birthplace https://www.phenxtoolkit.org/protocols/view/10201 *ADAPTED-Territoires with US; instead of seperate*

Source for PhenX : American Community Survey (ACS), 2008

Potential new approach!

3. ZIP code (caveat collapse zip codes w less than 10)

What is your current postal ZIP code? ____________________

Project 5 Covid-19 Address Postal Code https://cde.nlm.nih.giov/deView?tinyId=w BHatIMoA

4. Self-Identification (This question's intent is to get at bare minimum of identification, which will be determined by the changes proposed by OMB. Study can collect details of Race and Ethnicity as preferred. This does not supplant other required R/E reporting. Awaiting OMB.)

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

ScHARe working group preference based on potential classifications in 2030 census https://www.npr.org/2021/09/30/1037352177/2020-census-results-by-race-some-other-latino-ethnicity-hispanic#:~:text=And%20under%20that%20combined%20question%2C%20the%20list%20of,federal%20agencies%20collect%20data%20on%20race%20and%20ethnicity.

Potential new approach!

5. Sex

What was your sex assigned at birth, on your original birth certificate?

- ☐ Female
- ☐ Male
- ☐ Intersex
- ☐ None of these describe me
- ☐ Prefer not to answer

PhenX Protocol - Biological Sex Assigned at Birth https://www.phenxtoolkit.org/protocols/view/11601

All of Us Research Program, Participant Provided Information (PPI), 2018

National Academies Sciences, Engineering, Medicine report: Measuring Sex, Gender Identity, and Sexual Orientation
https://www.nationalacademies.org/our-work/measuring-sex-gender-identity-and-sexual-orientation-for-the-national-institutes-of-health
and All of Us
chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://allofus.nih.gov/sites/default/files/aou_ppi_basics_version.pdf

Potential new approach!

6. Gender

What is your current gender? [Select only one]

- ☐ Man
- ☐ Woman
- ☐ Non-Binary
- ☐ Transgender
- ☐ None of these describe me-I would like to consider additional options

  Are any of these a closer description to your gender identity?

  [ ] Trans man/Transgender Man/FTM
  [ ] Trans woman/Transgender Woman/MTF
  [ ] Genderqueer
  [ ] Genderfluid
  [ ] Gender variant
  [ ] Questioning or unsure of your gender identity
  [ ] None of these describe me, and I want to specify ____________________
- ☐ Prefer not to answer

PhenX Protocol - Gender Identity https://www.phenxtoolkit.org/protocols/view/11801

All of Us Research Program, Participant Provided Information (PPI), 2018

National Academies Sciences, Engineering, Medicine report: Measuring Sex, Gender Identity, and Sexual Orientation
https://www.nationalacademies.org/our-work/measuring-sex-gender-identity-and-sexual-orientation-for-the-national-institutes-of-health
Adapted: Non Binary added

Potential new approach!

7. Sexual orientation

Which of the following best represents how you think of yourself? [Select only one]

☐ Lesbian
☐ Gay
☐ Straight, that is, not gay or lesbian, etc.
☐ Bisexual

If none of the above represents you, are any of these a closer description of how you think of yourself (drop down)

[ ] Queer
[ ] Polysexual, omnisexual, sapiosexual or pansexual
[ ] Asexual
[ ] Two-spirit
[ ] Have not figured out or are in the process of figuring out your sexuality
[ ] Mostly straight, but sometimes attracted to people of your own sex
[ ] Do not think of yourself as having sexuality
[ ] Do not use labels to identity yourself
[ ] Don't know the answer
[ ] No, I mean something else (optional free text) ____________________

☐ Prefer not to answer

Phen X Sexual Orientation Protocol. https://www.phenxtoolkit.org/protocols/view/11701?origin=subcollection

All of Us Research Program Participant Provided Information (PPI) Version: December 17, 2018

National Academies Sciences, Engineering, Medicine report: Measuring Sex, Gender Identity, and Sexual Orientation https://www.nationalacademies.org/our-work/measuring-sex-gender-identity-and-sexual-orientation-for-the-national-institutes-of-health

8. Marital status

What is your current marital status?

- ☐ Married
- ☐ Living as married or living with a romantic partner
- ☐ Divorced
- ☐ Widowed
- ☐ Separated
- ☐ Single, never been married-not living with romantic partner
- ☐ Prefer not to answer

Hints 5 Cycle 4 (2020) https://hints.cancer.gov/view-questions-topics/question-details.aspx?qid=593 (BRFSS Questionnaire (2001), Section 13: Demographics modified)
All of Us
chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://allofus.nih.gov/sites/default/files/aou_ppi_basics_version.pdf

Unique – potentially mappable

9. Education

What is the highest level of education you have completed?

- ☐ No formal schooling
- ☐ Primary/Grade/Elementary School (approximately grades 1st through 5th)
- ☐ Middle School/Lower Secondary Education (approximately grades 6th through 8th)
- ☐ Secondary/High School/Upper Secondary (grades 9th through 11th) without a high school diploma
- ☐ General Educational Diploma (GED)
- ☐ Secondary/High School/Upper Secondary (grades 9th through 12th) with a high school diploma
- ☐ Occupational/Technical/Vocational Programs/Short Cycle Tertiary Education - Associate's Degree (approximately 2 years)
- ☐ College/University/Bachelor's Degree/Equivalent Tertiary Education (approximately 3-5 years)
- ☐ Graduate/post-graduate degree/professional degree/ (JD, PhD, MD, EdD, Eng, Master's Degree, etc.)

International Standard Classification of Education (ISCED)
https://datatopics.worldbank.org/education/wRsc/classification#:~:text=The%20International%20Standard%20Classification%20of,revised%20in%201997%20and%202011
and
USA standards of Education https://nces.ed.gov/programs/digest/d01/fig1.asp

Potentially mappable

10. Annual household income range

What is your annual household income from all sources within family, not including roommates?

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

All of Us - Basic Information Survey https://allofus.nih.gov/sites/default/files/aou_ppi_basics_version.pdf

BRFSS = Behavioral Risk Factor Surveillance System (CDC)

11. Household family size

Approximately how many individuals (adult and children) does your household family income support?

☐ ____________________

Project 5 Covid-19 Shared Living Space Number of Individuals https://cde.nlm.nih.gov/cde/search?q=PROJECT%205&nihEndorsed=true

12. English proficiency

We are interested in your own opinion of how well you speak English. Would you say you speak English:

- ☐ Very well
- ☐ Well
- ☐ Not well
- ☐ Not at all
- ☐ Refused
- ☐ Don't Know

PhenX Toolkit - English Proficiency https://www.phenxtoolkit.org/protocols/view/270201

Regents of the University of California. (2019). CHIS 2018 Adult Questionnaire, question number "QA18_G8" is represented in this protocol as question 1. Retrieved from http://healthpolicy.ucla.edu/chis/design/Pages/questionnairesEnglish.aspx

13. Disabilities

Do you have a disability or have serious difficulty with any of the following? Select all that apply.

- ☐ Deafness or difficulty hearing
- ☐ Blindness or difficulty seeing
- ☐ Difficulty concentrating, remembering, and deciding
- ☐ Difficulty walking or climbing stairs
- ☐ Difficulty dressing or bathing
- ☐ Difficulty doing errands alone
- ☐ Not disabled

CDC Standard Disability Questions https://www.cdc.gov/ncbddd/disabilityandhealth/datasets.html (format adapted)

14. Health insurance

Are you currently covered by any of the following types of health insurance or health coverage plans?

- ☐ Insurance through a current or former employer or union (of yours or another family member's). This would include COBRA coverage.
- ☐ Insurance purchased directly from an insurance company (by you or another family member). This would include coverage purchased through an exchange or marketplace
- ☐ Medicare, for people 65 and older, or people with certain disabilities.
- ☐ Medicaid, Medical Assistance (MA), the Children's Health Insurance Program (CHIP), or any kind of state or government-sponsored assistance plan based on income or a disability.
- ☐ TRICARE or other military health care, including VA health care.
- ☐ Indian Health Service
- ☐ Any other type of health insurance. coverage or health coverage plan
- ☐ Uninsured

PhenX Health Insurance Coverage https://www.phenxtoolkit.org/protocols/view/11502

15. Employment status

We would like to know about what you do: are you working now, looking for work, retired, keeping house, a student, or what?

- ☐ Working now or paid sick leave/parental leave/family leave/administrative leave
- ☐ Only temporarily laid off, or unpaid sick leave/parental leave/family leave/administrative leave
- ☐ Looking for work, unemployed
- ☐ Retired
- ☐ Disabled, permanently or temporarily
- ☐ Raising children full-time, full-time caregiver, or keeping house
- ☐ Student
- ☐ Other/specify: ______________________________

PhenX - Current Employment Status https://www.phenxtoolkit.org/protocols/view/11301 (Adapted-used parental instead of maternal, and family leave added with paid/unpaid)

16. Usual place of health care

Is there a place that you USUALLY go to when you are sick or need advice about your health? Select all that apply.

- ☐ A doctor's office or community health center, including Indian Health Service, or hospital-based clinics
- ☐ Walk-in clinic, urgent care center, or retail clinic in a pharmacy or grocery store
- ☐ Emergency room
- ☐ A VA Medical Center or VA outpatient clinic
- ☐ Some other place
- ☐ Does not go to one place most often
- ☐ Don't know

PhenX Protocol Access to Health Services Ques #5 https://www.phenxtoolkit.org/protocols/view/270101 (adapted with hospital-based clinics)

Project 5 Covid-19 Usual Place of Health Care Type https://cde.nlm.nih.gov/cde/search?q=PROJECT%205&nihEndorsed=true\ (adapted with hospital-based clinics)

17. Economic Stability – Social Needs

In the past year, have you or any family members you live with been unable to get any of the following when it was really needed? Select all that apply.

- ☐ Childcare
- ☐ Clothing
- ☐ Food
- ☐ Housing
- ☐ Internet/Broadband
- ☐ Phone (e.g., mobile or landline)
- ☐ Transportation (e.g., private or public)
- ☐ Utilities (e.g., gas, electric, propane, natural gas, etc.)
- ☐ Medicine or any health care (medical, dental, mental health, vision)
- ☐ Other/specify: ______________________

Source: Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE) tool (Adapted-internet, housing, transportation added to question #14) Housing and transportation is included in survey. https://prapare.org/wp-content/uploads/2023/01/PRAPARE-English.pdf

U.S. Census Bureau, 2015 and 2016 American Community Survey – Internet/Broadband https://www.census.gov/content/dam/Census/library/publications/2018/acs/ACS-39.pdf

High-level SDoH assessment

18. Self-reported health

Would you say your health in general is excellent, very good, good, fair, or poor?

- ☐ Excellent
- ☐ Very good
- ☐ Good
- ☐ Fair
- ☐ Poor

Patient-Reported Outcomes Measurement Information (PROMIS)

https://www.healthmeasures.net/index.php?option=com_instruments&task=downloadComponentFile&file=PROMIS%20Scale%20v1.2%20-%20Global%20Health%20Physical%202a%2009062016.pdf

19. Health conditions and medications or other Treatments

Has a health care provider told you that you have any one or more of the following conditions? Select all that apply currently. Check the second box if you are taking medications or receiving some other treatment for the condition.

☐ ☐ Cancer

☐ ☐ Coronary heart disease

☐ ☐ Heart failure

☐ ☐ High blood pressure/hypertension

☐ ☐ Stroke

☐ ☐ Thrombotic disorders

☐ ☐ High cholesterol

☐ ☐ Diabetes (type I)

☐ ☐ Diabetes (type II)

☐ ☐ Obesity

☐ ☐ Hepatitis

☐ ☐ Other chronic liver disease

Addresses co-morbidities

☐ ☐ Asthma

☐ ☐ Other chronic respiratory disease (e.g., COPD, emphysema)

☐ ☐ Chronic kidney disease

☐ ☐ Psychological and/or psychiatric disease or disorder (e.g., anxiety, depression, bipolar disorder)

☐ ☐ Alzheimer's disease

☐ ☐ Dementia

☐ ☐ Epilepsy

☐ ☐ Multiple sclerosis

☐ ☐ Other chronic neurological condition (e.g., Parkinson's disease, migraine)

☐ ☐ Immunodepression

☐ ☐ HIV/AIDS

☐ ☐ Autoimmune condition (e.g., rheumatoid arthritis, systemic lupus erythematosus, vasculitis)

☐ ☐ Chronic musculoskeletal condition (e.g., back pain, osteoarthritis, osteoporosis)

☐ ☐ Sickle cell disease

☐ ☐ Sleep disorder (e.g., insomnia, sleep apnea, narcolepsy)

☐ ☐ Solid organ transplant

☐ ☐ Smoking

☐ ☐ Other substance use disorder (e.g., drugs and/or alcohol dependence)

☐ ☐ Long Covid (also known as long-haul COVID, long-term effects of COVID, chronic COVID, post-acute COVID-19, and PASC - post-acute sequelae of SARS-CoV-2)

☐ ☐ Chronic fatigue

☐ ☐ Dental diseases and conditions (e.g., caries, periodontal disease, oral and pharyngeal cancer)

☐ ☐ Eye diseases and conditions (e.g., cataract, glaucoma, amblyopia, myopia and other refractive errors, age-related macular degeneration, diabetic retinopathy, ocular trauma, uveitis, keratoconus)

☐ ☐ Other chronic disease/specify:

☐ None of the above

Project 5 Covid-19 Comorbidity or Underlying conditions https://cde.nlm.nih.gov/cde/search?q=PROJECT%205&nihEndorsed=true\ (Adapted for Medications-Added Chronic musculoskeletal conditions, High Cholesterol, Sleep Disorders and Stroke)

Project level

20. Minority Health and Health disparities research content area

Which of the following content areas of research is this study addressing, if any?
Select all that apply.

- ☐ Minority health study focused on a one race or ethnic population and not addressing a health disparity.
- ☐ Health Disparity Outcome (select the focus area)
  - ☐ 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/or clinical outcomes using established measures
  - ☐ Worse outcomes on validated self-reported measures that reflect daily functioning or symptoms from specific conditions
- ☐ Other Health Outcomes / Healthcare Delivery

Duran D, Perez-Stable, E. Novel Approaches to Advance Minority Health and Health Disparities; Am J Public Health. 2019, Jan;109(S1):S8-S10. doi:10.2105/AJPH. 2019.304952. PMID: 30699026; PMCID:PMC6356133. *ADAPTED with Other health outcomes delivery/care*

21. NIMHD Framework

What NIMHD Research framework levels and domains of influence is your study targeting? (Select all that apply)

<u>Levels of Influence</u>

☐ Individual

☐ Interpersonal

☐ Community

☐ Societal

<u>Domains of Influence</u>

☐ Biological

☐ Behavioral

☐ Physical/Built Environments

☐ Sociocultural Environment

☐ Health Care Systems and Clinical Care

NIMHD Research Framework. https://www.nimhd.nih.gov/about/overview/research-framework/nimhd-framework.html

ScHARe

What are Think-a-Thons?

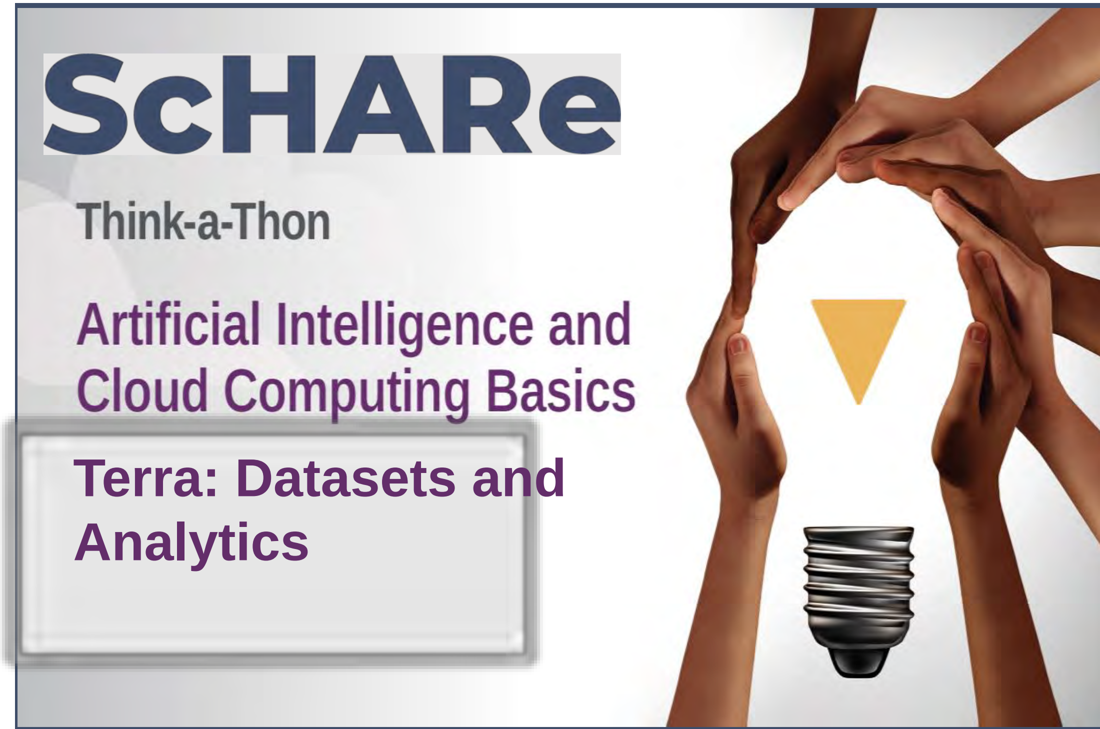
BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Think-a-Thons (TaT)

- Monthly sessions (2 1/2 hours)
- Instructional/interactive
- Designed for new and experienced users
- Research & analytic teams to:
  - Conduct health disparities, health outcomes, bias mitigation research
  - Analyze/create tools for bias mitigation
- Publications from research team collaboration
- Networking
- Mentoring and coaching
- Focus:

Types:

✓Instructional / Tutorial
✓Collaborative Research Teams
✓Bias mitigation



Register:

bit.ly/think-a-thons

Think-a-Thon Instructional Tutorials

Web: 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	**Preparing for AI 4: Overview Prep for AI Summary with Transparency, Privacy, Ethics**

ScHARe for Educators (Community Colleges & Low Resource MSIs)
ScHARe for American Indian / Alaska Native Researchers
ScHARe for Non-Biomedical Researcher Coders and Programmers to conduct Research

The monthly **ScHARe Think-a-Thons** scheduled or archived below are designed so participants reach one of these goals (as noted with each session):

- **Goal 1:** Achieve a **better understanding of both the fields and the terminology** used to describe the AI/cloud computing infrastructure, components and processes.
- **Goal 2: Develop research questions and projects relevant to AI and cloud computing** that leverage the cutting-edge technology and data/computing resources now available to health disparities researchers (including the ones at their disposal on the ScHARe platform).

Upcoming Think-a-Thons | Past Think-a-Thons | See FAQs

ScHARe

Think - a - Thons

Think-a-Thon Schedule

Think-a-Thons are held on the third Wednesday of each month. Accommodations information | Think-a-Thon recordings

Date	Time	Topic	Register
March 20, 2024	2:00 – 4:30 p.m. ET	**Preparing for AI-driven Research on ScHARe: A Comprehensive Review and Brainstorming Session – Part 2** Toward Goal 2: Prepares participants for ScHARe research collaborations by covering: • Choosing computational strategies (AI, ML, statistics) • An overview of Python data science libraries • The significance of testing and monitoring in algorithm development • The role of open science in ensuring reproducible and transparent AI-based research For researchers and students at all levels who want to collaborate on ScHARe to develop innovative and publishable research projects	Register *Registration closes at 12:00 p.m. ET on the day of the event.*

PAST Think-a-Thons Posted

November 15, 2023	2.5 hours	View video: Preparing for AI 2: An Introduction to FAIR Data and AI-ready Datasets Toward Goal 1: How to prepare an AI-ready dataset using gold standard data management principles, including: • Making datasets findable, accessible, interoperable, and reusable (FAIR) • Using transparent data documentation to foster data re-use • Ensuring that selected data addresses expected outcomes and drives meaningful AI insights • Handling missing data through strategies, proxies, and synthetic data	View slides (PDF, 4 MB)

https://www.nimhd.nih.gov/resources/schare/think-a-thons.html

Upcoming

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**
- **Generational Career & Discipline Exchange**

- 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

Generational Career & Discipline Exchange

Research Think - a - Thons

Expectation of the Research Project.

- The launch of the project will occur during the Think-a-Thon.
  - Pre-Assigned Co-Leads: Data Science Expert and a Health Disparity/Health Care Delivery Expert
  - There will be 4 sessions: 2 python, 1 R and 1 Statistic defined research collaborative
  - Volunteers who want to participate in health disparity/health care delivery research will select one of the 4 sessions based upon the analytics expected to be used
  - In the breakouts, the group will decide the research topic and which data sets will be used.

- The co-leads will assign tasks to the participants for the next **three months** to complete the project in preparation for publication. There will be 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

https://www.nimhd.nih.gov/resources/schare/think-a-thons.html

ScHARe

Research Think - a - Thons

<u>During Think-a-Thon</u>

- ScHARe Terra Workspace (Data Co-lead is primary to create and to monitor workspace collaborators
- Research Topic (Science co-lead will guide the discussion to the consensus topic)
- Likely data sets to be used for topic

<u>Project Expectations:</u>

- Literature review
- Data set assessment for AI readiness (i.e. complete variables needed for project, fair representation of populations, missing variables, incompleteness of variables, data gaps, etc)
- Data Dictionary
- Data Sheet and Data Set Facts
- Design description to ensure that the outcome expected is probable.
- Decision on analytics and training to be used (complete a methodology description, including a model card)
- Test results for biases (document the types of biases encountered and how each addressed)
- Draft Publication

https://www.nimhd.nih.gov/resources/schare/think-a-thons.html

ScHARe

Research Think - a - Thons

Experience Conducting Ethical AI

TRANSPARENCY:

- Def:
  - Public Perception & understanding of how AI works
  - Comprehend the algorithmic views and decisions taken based on them
- Technical Documentation for duplication / re-use
- Tools:
  - Data Dictionary
  - Health Sheet (Data Sheet)
- Model Cards (capabilities & purpose of algorithms are openly and clearly communicated to relevant stakeholders)
- Documentation of methodologies
- Doesn't disclose intellectual property

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

https://www.nimhd.nih.gov/resources/schare/think-a-thons.html

Think-a-Thons Training/Mentoring Pipeline

NLM
OIC Experts
Fellows

Think-a-Thons

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

N3C
Aim AHEAD
BioData Catalyst
All of US
AnVil
HEAL

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

Goal: "Upskilling"

- ✓ Data science specialist 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 scientist with no or little research experience.
- ✓ Resource & Tool for Community Colleges and Low Resource MSIs and Organizations

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

Computational Strategies

Data Science Computational Strategies

Choosing Between Traditional Statistics and AI/ML

- Welcome to the **second part** of our workshop on conducting research projects

- Today’s **overview** will cover selecting computational strategies in data science. We'll explore the decision-making process involved in choosing between traditional statistics and Artificial Intelligence (AI)/Machine Learning (ML)

- We will help you understand the **fundamental differences** between these approaches and their respective advantages and disadvantages, and point you to helpful **Python libraries** for each strategy

Decision-Making Process

Choosing the **right approach to analyzing data** is critical for achieving research objectives effectively

The decision-making process in selecting computational strategies involves several **key steps**:

1. We must clearly **define the research problem** or question we aim to address
2. Next, we must consider the **nature of the data we have**, our **research goals**, and the **resources available** to us

Based on these factors, we then choose the most appropriate computational strategy

Traditional Statistics vs. AI and ML

Traditional Statistics and modern computational techniques such as Artificial Intelligence (AI) and Machine Learning (ML) offer **distinct approaches to data analysis:**

- **Traditional Statistics** focuses on **hypothesis testing, inference, and the application of parametric and non-parametric methods.** It emphasizes **interpretability, reliance on assumptions, and limitations in handling complex datasets**

- In contrast, **AI and ML** prioritize **pattern recognition, prediction, and the development of predictive models.** They emphasize **scalability, complexity management, and the ability to process large volumes of data efficiently**. However, AI and ML models often **trade interpretability for increased predictive power**, leading to challenges in understanding their decision-making processes

Analyses Enabled by AI and Big Data

- Artificial Intelligence (AI) and Big Data enable a **wide range of advanced analyses that go beyond traditional statistical methods**, including:
  - **predictive analytics**
  - **natural language processing**
  - **image recognition**
  - **anomaly detection** (identifying unusual patterns or data points that deviate significantly from the expected norm)

- AI and Big Data empower researchers to extract **valuable insights from vast and complex datasets**, leading to more accurate predictions, enhanced decision-making and problem-solving capabilities

- Examples of AI and Big Data analyses include predictive modeling for **disease outbreak prediction** and **sentiment analysis** of social media data for public health monitoring

Conclusion and Recommendations

- It's important to consider the **nature of the research problem**, the **characteristics of the data,** and the **desired outcomes** when choosing a computational strategy

- **Recommendations:**
  1. Assess the **research objectives and data characteristics** before selecting a strategy
  2. Leverage the **strengths of each approach** to maximize the insights gained from data analysis
  3. Stay informed about **emerging technologies and methodologies** in data science to adapt to evolving research needs

- By making **informed decisions** about computational strategies, researchers can enhance the quality and impact of their research outcomes

ScHARe

Traditional Statistics

Overview

- **Strengths:** robust, interpretable, well-established methodologies
- **Weaknesses:** limited predictive power, assumption-dependent, often focused on hypothesis testing
- **Data types & use cases:** numerical data, identifying trends, correlations, causal relationships

- **Popular Python libraries:** NumPy, SciPy, Pandas

1. Descriptive Statistics

- **Strategy:** Summarizing and describing key features of healthcare data, such as mean, median, standard deviation, and percentiles
- **Applications:** Understanding the central tendency and variability in healthcare variables
- **Python Libraries:** NumPy, pandas

2. Inferential Statistics

- **Strategy:** Making predictions or inferences about a population based on a sample from that population

- **Applications:** Drawing conclusions about healthcare disparities from a subset of relevant data

- **Python Libraries:** SciPy, statsmodels

3. Hypothesis Testing

- **Strategy:** Evaluating statistical significance to determine whether observed differences are likely to be real or due to chance
- **Applications:** Testing hypotheses about healthcare interventions or disparities
- **Python Libraries:** SciPy, statsmodels

4. Analysis of Variance (ANOVA)

- **Strategy:** Assessing the statistical significance of differences among group means in healthcare data

- **Applications:** Comparing means across multiple categories to identify significant differences

- **Python Libraries:** SciPy, statsmodels

5. Chi-Square Test

- **Strategy:** Assessing the association between categorical variables in healthcare datasets
- **Applications:** Examining relationships between demographic factors and health outcomes
- **Python Libraries:** SciPy, pandas

6. Regression Analysis

- **Strategy:** Modeling the relationship between dependent and independent variables in healthcare data

- **Applications:** Predicting health outcomes based on various factors, identifying disparities

- **Python Libraries:** Statsmodels, scikit-learn

7. Survival Analysis

- **Strategy:** Analyzing time-to-event data, such as the time until a patient experiences a particular health event

- **Applications:** Studying disparities in disease progression or survival rates

- **Python Libraries:** Lifelines, statsmodels

8. Correlation Analysis

- **Strategy:** Examining the strength and direction of relationships between two continuous variables in healthcare datasets

- **Applications:** Assessing associations between risk factors and health outcomes

- **Python Libraries:** NumPy, pandas

9. Logistic Regression:

- **Strategy:** Modeling the probability of a binary outcome in healthcare data
- **Applications:** Analyzing factors influencing the likelihood of specific health events
- **Python Libraries:** Statsmodels, scikit-learn

10. Bayesian Statistics

- **Strategy:** Updating beliefs about parameters based on new evidence in a probabilistic framework

- **Applications:** Incorporating prior knowledge into healthcare disparities research

- **Python Libraries:** PyMC3, Stan

11. Time Series Analysis

- **Strategy:** Analyzing temporal patterns and trends in healthcare data
- **Applications:** Studying disparities over time in health outcomes or interventions
- **Python Libraries:** Statsmodels, Pandas

ScHARe

Artificial Intelligence and Machine Learning

Main AI Computational Strategies

- Artificial Intelligence (AI) encompasses various computational strategies aimed at **mimicking human intelligence**

- These strategies are implemented using **different algorithms and techniques**

- **Python,** being a versatile language, offers numerous libraries for implementing AI strategies effectively

Machine Learning (ML)

- Machine learning involves the development of algorithms that **enable computers to learn from and make predictions or decisions based on data.** ML allows computers to improve at a specific task without explicit programming, by learning from data

- Examples:
  - Linear Regression
  - Decision Trees
  - Random Forest

- Commonly Used Python Libraries:
  - scikit-learn
  - TensorFlow
  - Keras
  - PyTorch

Deep Learning

- Deep learning is a subset of machine learning that **is inspired by the structure and function of the brain.** It uses **artificial neural networks** comprising multiple layers to **learn complex patterns from data**

- Examples:
  - Convolutional Neural Networks (CNNs) for image recognition
  - Recurrent Neural Networks (RNNs) for sequence data
  - Generative Adversarial Networks (GANs) for generating synthetic data

- Commonly Used Python Libraries:
  - TensorFlow
  - Keras
  - PyTorch

Natural Language Processing (NLP)

- NLP involves the interaction between computers and humans using natural language. It focuses on giving **computers the ability to understand and manipulate human language**

- Examples:
  - Sentiment Analysis
  - Named Entity Recognition (NER) (it categorizes specific elements within text)
  - Machine Translation

- Commonly Used Python Libraries:
  - NLTK (Natural Language Toolkit)
  - spaCy
  - Transformers

Reinforcement Learning

- Reinforcement learning focuses on training agents to make sequential decisions by interacting with an environment

- Examples:
  - Game playing (e.g., AlphaGo)
  - Robotics control
  - Recommendation systems

- Commonly Used Python Libraries:
  - OpenAI Gym
  - TensorFlow Agents
  - Stable Baselines

Evolutionary Algorithms

- Evolutionary algorithms are inspired by biological evolution and involve **optimization techniques** based on natural selection and genetic variation. Specifically, they **mimic natural selection** to solve problems by iteratively **refining populations of candidate solutions**

- Examples:
  - Genetic Algorithms
  - Genetic Programming
  - Evolutionary Strategies

- Commonly Used Python Libraries:
  - DEAP (Distributed Evolutionary Algorithms in Python)
  - PyGMO (Python Parallel Global Multiobjective Optimizer)

Quiz 1

Scenario: You are a public health researcher investigating the factors contributing to higher rates of heart disease among a specific minority population in your community. You have a dataset containing information about thousands of individuals, including demographics, socioeconomic factors, health history, and access to healthcare.

Question: Which approach would be most suitable for analyzing this data to understand the disparities in heart disease rates?

a) **Traditional Statistics:** Calculate average income levels and compare them to heart disease prevalence across different zip codes within the community.

b) **Machine Learning:** Develop a machine learning model to predict the likelihood of developing heart disease for individuals based on their data.

c) **Both Traditional Statistics and Machine Learning:** Use traditional statistics to explore initial relationships and then build a machine learning model to identify complex patterns contributing to the disparities.

Quiz 1

Answer: (C) Both Traditional Statistics and Machine Learning

Explanation:

- Traditional statistics can reveal basic trends, like correlations between income and heart disease prevalence across zip codes. This can provide initial clues about potential disparities.
- Machine learning can be powerful in health disparities research. It can analyze complex interactions between various factors (e.g., income, access to healthcare, environmental factors) and their combined influence on heart disease risk within the specific population.
- By combining traditional statistics for initial exploration with machine learning for in-depth analysis, you gain a comprehensive understanding of the factors contributing to the observed health disparities.

ScHARe

Python Libraries

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

python™

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 to help you learn

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

Introduction to Python Data Science Libraries

- Python offers a **rich ecosystem of libraries** for data science tasks
- In this section, we'll introduce some of the most commonly used Python libraries in data science
- **Each library 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

Overview of Python Data Science Libraries

Python data science libraries are essential for **data manipulation, analysis, and visualization** tasks.

NumPy: The Foundation for Numerical Computing

NumPy

Overview

A fundamental package for scientific computing, providing support for large, multi-dimensional arrays (ordered collections of items) and matrices

Characteristics

- Provides efficient **multidimensional arrays** for data storage and manipulation
- Enables **mathematical operations** on large datasets
- Lays the **groundwork for data analysis** with other libraries

Example application

Calculating statistical measures such as mean, median, and standard deviation of health indicators (e.g., life expectancy) across various demographic groups

SciPy: Extending Computing Capabilities

Overview

An open-source library that builds on NumPy and provides additional functionality for mathematical and scientific computing

Characteristics

- Offers **advanced algorithms** for scientific computing beyond NumPy
- Includes **tools for optimization**, integration, and signal processing
- **Complements NumPy** for diverse scientific computing tasks

Example application

Conducting hypothesis testing to evaluate the effectiveness of interventions aimed at reducing health disparities, such as comparing pre- and post-intervention health indicators

Pandas: Wrangling Data Like a Pro

pandas

Overview

A powerful library for data manipulation and analysis, offering data structures and functions for manipulating structured data

Characteristics

- Offers powerful data structures like **DataFrames** for handling tabular data
- Enables **data cleaning, manipulation, and exploration** with ease
- **Integrates** seamlessly with other data science libraries

Example application

Exploring correlations between socio-economic factors (e.g., income, education level) and health outcomes (e.g., mortality rates)

Matplotlib: Visualizing Insights

Overview

A comprehensive library for creating static, animated, and interactive visualizations in Python, offering a wide range of plotting functions

Characteristics

- Creates a wide variety of static, animated, and interactive **visualizations**
- Enables **customization** for clear and compelling data storytelling
- Integrates with other libraries for comprehensive **data exploration**

Example application

Creating visualizations such as bar charts or pie charts to illustrate disparities in healthcare access among different ethnic or socio-economic groups

Seaborn: Building on Matplotlib for Stats

seaborn

Overview

A statistical data visualization library based on Matplotlib, providing a high-level interface for creating informative and attractive visualizations

Characteristics

- Offers a high-level interface built upon Matplotlib for statistical graphics
- Creates aesthetically pleasing and informative visualizations
- Ideal for exploring relationships and distributions within your data

Example application

Creating box plots or violin plots to compare distributions of health indicators (e.g., blood pressure levels) among different population segments

Scikit-learn: Machine Learning Made Accessible

Overview

A machine learning library that offers simple and efficient tools for data mining and data analysis, including classification, regression, clustering, and dimensionality reduction

Characteristics

- Provides a comprehensive library for various **machine learning** algorithms
- Enables tasks like **classification, regression, and clustering**
- Facilitates **model building, evaluation, and deployment**

Example application

Implementing machine learning algorithms to classify patients into different risk categories based on socio-economic factors and predict healthcare outcomes (e.g., hospital readmissions)

Statsmodels: Diving Deeper into Statistical Analysis

statsmodels

Overview

A library for estimating statistical models and conducting statistical tests, providing a wide range of statistical techniques

Characteristics

- Provides a collection of tools for **statistical modeling and econometrics**
- Enables robust **hypothesis testing**, estimation, and model selection
- Ideal for **in-depth statistical analysis** of health disparities data

Example application

Exploring correlations between socio-economic factors (e.g., income, education level) and health outcomes (e.g., mortality rates)

TensorFlow: Building Powerful Deep Learning Models

TensorFlow

Overview

An open-source machine learning framework developed by Google, widely used for building and training deep learning models

Characteristics

- Open-source framework for numerical computation **and large-scale machine learning**
- Particularly adept at **deep learning**, a powerful subset of machine learning
- Enables building and training **complex models** for tasks like natural language processing

Example application

Training convolutional neural networks (CNNs) to analyze medical images (e.g., X-rays, MRIs) and detect signs of disease or abnormalities associated with health disparities

PyTorch: A Powerful Deep Learning Framework

PyTorch

Overview

Provides support for distributed training across multiple GPUs and devices, enabling researchers to train large-scale machine learning models efficiently

Characteristics

- Well-suited for **rapid prototyping and experimentation**
- **User-friendly** approach that lowers the barrier to entry for deep learning
- PyTorch models can be efficiently deployed in production environments

Example application

Adapting pre-trained language models for healthcare-specific NLP tasks, such as extracting information about social determinants of health from unstructured text data

Quiz 2

Which Python library is commonly used for data manipulation and analysis, offering data structures and functions for working with structured data?

a) NumPy

b) Pandas

c) SciPy

d) Statsmodels

Quiz 3

Which Python library is known for its statistical data visualization capabilities and is based on Matplotlib?

a) NumPy

b) Pandas

c) Seaborn

d) TensorFlow

Example Application with Code (NumPy)

python Copy code

```python
import numpy as np

# Sample health outcome data for different demographic groups
health_outcomes = np.array([
    [120, 80, 100],  # Group 1
    [90, 110, 95],   # Group 2
    [100, 95, 105]   # Group 3
])

# Calculate mean, median, and standard deviation
mean_outcomes = np.mean(health_outcomes, axis=1)
median_outcomes = np.median(health_outcomes, axis=1)
std_outcomes = np.std(health_outcomes, axis=1)

print("Mean outcomes:", mean_outcomes)
print("Median outcomes:", median_outcomes)
print("Standard deviation of outcomes:", std_outcomes)
```

NumPy simplifies numerical computations and array operations in Python

Example: Calculating **summary statistics** for health outcome data and detecting variations across demographic groups

Libraries in notebooks

A **Jupyter Notebook** is an interactive analysis tool that includes:

- **code cells** for manipulating and visualizing data in real time (Terra notebooks support **Python or R**)
- **documentation** to make it easier to share and reproduce your analysis

In past Think-a-Thons, we:

- covered the basics of **creating your first notebook**
- **explored the instructional notebooks** available in the ScHARe workspace

If you are not familiar with **programming**, the code in our notebooks is very easy to understand and reuse, and our tutorials will help you understand how notebooks work.

Why use notebooks?

A notebook integrates code and its output into a single document where you can run code, display the output, and also add explanations, formulas, and charts

Using notebooks:

- **is now a major part of the data science workflow** at research institutions across the globe
- can make your work **more transparent, understandable, repeatable, and shareable**
- will **speed up your workflow** and make it easier to communicate and share your results

ScHARe notebooks

Take a look at what a notebook can do by checking out the instructional notebooks that **ScHARe offers to help novice users** learn how to use the workspace and its resources

A list of the available notebooks is provided on the right.

List of ScHARe instructional notebooks

- **00_List of Datasets Available on ScHARe**: a list of the datasets available in the ScHARe Datasets collection.
- **01_Introduction to Terra Cloud Environment**: an introduction to the Terra platform and cloud environment.
- **02_Introduction to Terra Jupyter Notebooks**: an introduction to Jupyter Notebooks on the Terra platform.
- **03_R Environment setup**: instructions on how to setup your cloud environment for R-based notebooks.
- **04_Python 3 Environment setup**: instructions on how to setup your cloud environment for Python 3-based notebooks.
- **05_How to access plot and save data from public BigQuery datasets using R**: instructions on how to access, plot, and save data from datasets available on the cloud through the Google Cloud Public Datasets Program, using R.
- **06_How to access plot and save data from public BigQuery datasets using Python 3**: instructions on how to access, plot, and save data from datasets available on the cloud through the Google Cloud Public Datasets Program, using Python 3.
- **07_How to access plot and save data from ScHARe hosted datasets using Python 3:** instructions on how to access, plot, and save data from datasets hosted by ScHARe in this workspace.
- **08_How to upload access plot and save data stored locally using Python 3:** instructions on how to import to Terra, access, plot, and save data from datasets stored locally on your computer.

ScHARe

Python External Resources

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

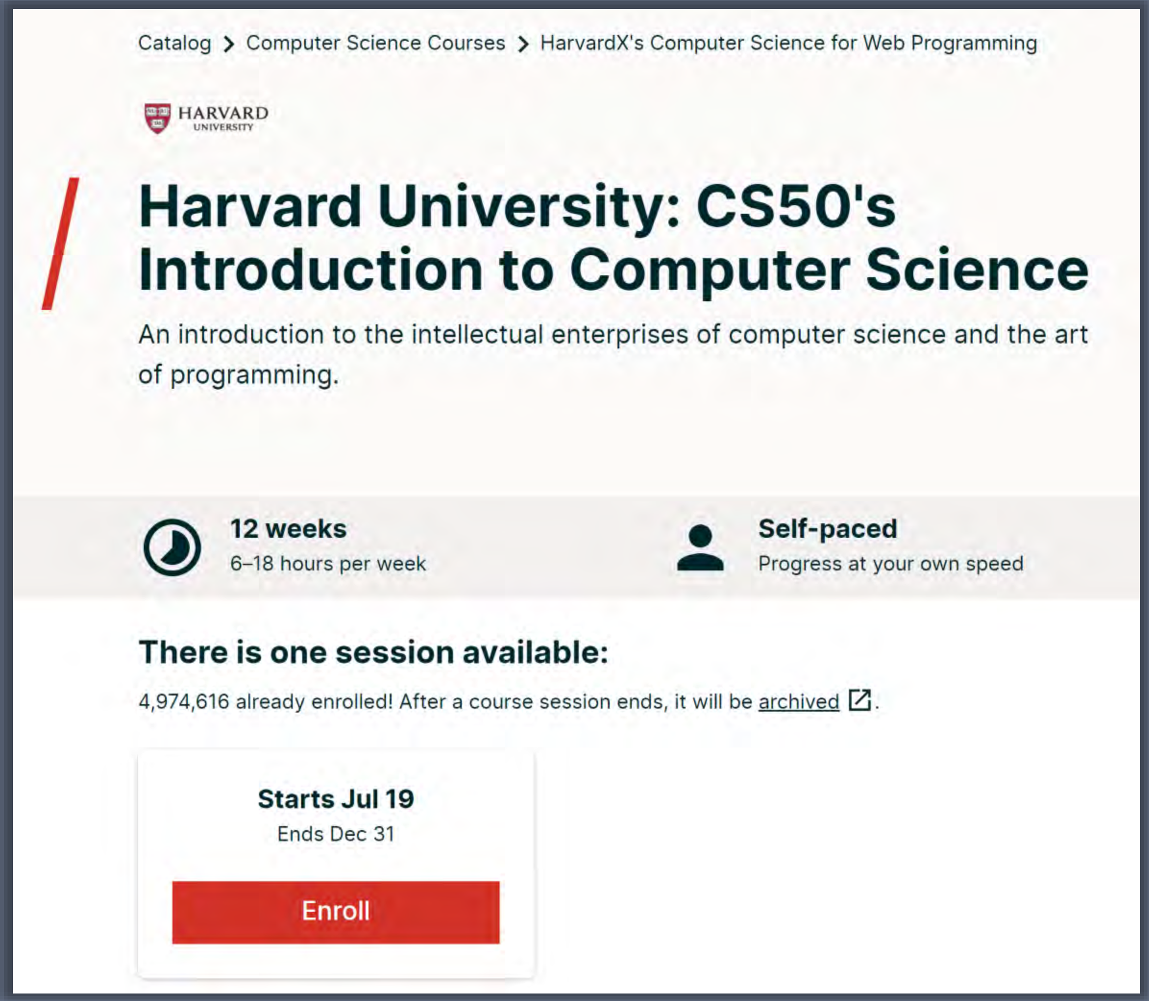
[FreeCodeCamp]()

Code class for beginners

Python resources

[Harvard]

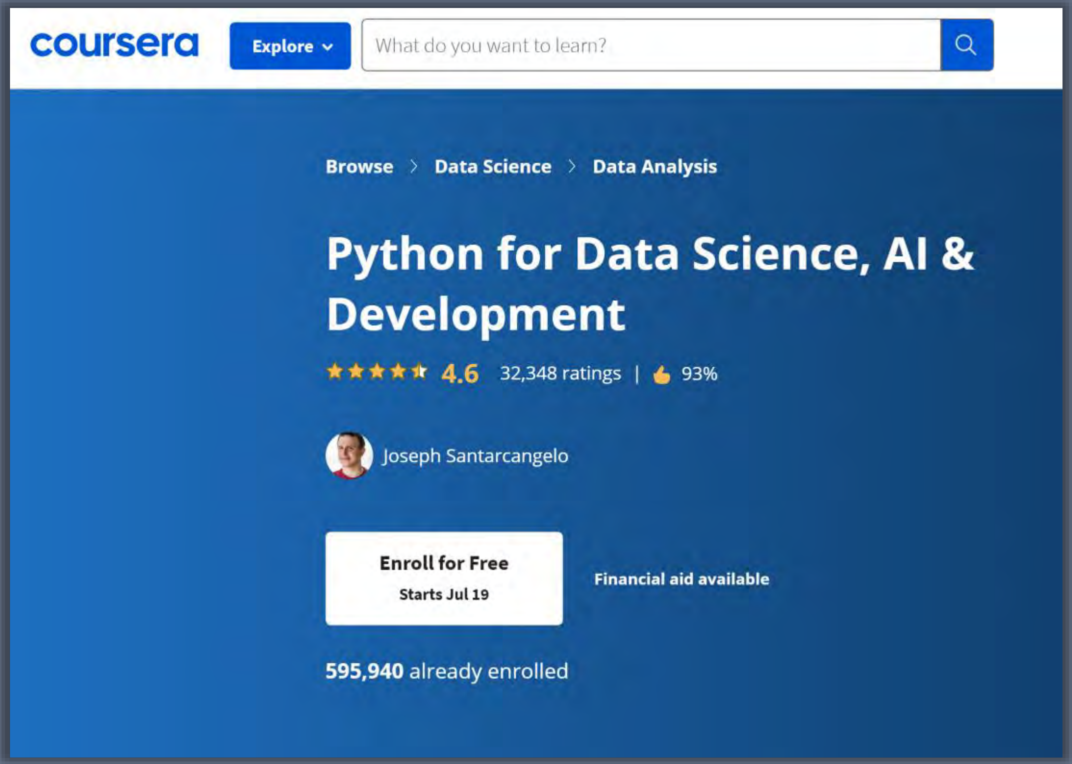
Free Python course



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

ScHARe

Algorithm Testing and Monitoring

What are Algorithms?

- An **algorithm** is a finite set of well-defined instructions designed to perform a specific task or solve a particular problem, often expressed in a logical sequence in a step-by-step process that can be executed by a computer

- Algorithms enable efficient and accurate **decision-making** and problem-solving across various domains, including healthcare policy decisions

Here's how it works in healthcare

- **Lots of data:** Hospitals collect tons of data about patients (diagnoses, treatments, and meds)
- **Data analysis:** Algorithms sift through this massive amount of data and identify patterns
- **Informing decisions:** Healthcare policymakers might use these patterns to decide how to allocate resources, like which treatments are most effective or where to offer more services

Sounds helpful, right? But there's a catch:

Algorithmic Bias

Algorithmic Bias

- Algorithmic bias refers to systematic and unfair outcomes arising from algorithms used for decision-making

- Algorithms trained on biased data or with flawed design can perpetuate or amplify existing societal biases in healthcare

Where can bias creep in during the development and implementation of algorithms?

1. **Data Acquisition and Selection:**

- **Sampling Bias:** if the data used to train the algorithm doesn't represent the entire target population
- **Historical Bias:** if historical healthcare data reflects past discrimination

2. **Feature Engineering and Model Design:**

- **Choosing the wrong features** can lead the algorithm to make unfair decisions based on irrelevant factors.
- **Model design:** inherent limitations

3. **Model Training and Evaluation:**

- **Training data quality:** inaccurate or incomplete datas
- **Evaluation metrics:** focusing solely on overall accuracy might mask disparate impacts on different populations. We need fairness metrics that assess how the algorithm performs across different subgroups (e.g., race, ethnicity, socioeconomic status)

4. **Implementation and Monitoring:**

- **Limited transparency:** If the decision-making process of the algorithm is a "black box," it's hard to identify and mitigate biases
- **Unintended consequences:** Even well-intentioned algorithms can lead to unintended consequences if not continuously monitored for potential biases emerging in real-world use

Importance of Algorithm Testing

1. Crucial during the design phase of a research project
2. Ensures the reliability and validity of algorithms before implementation
3. Enhances the accuracy and effectiveness of algorithms in real-world applications

Importance of Algorithm Monitoring

1. **Evolving Data and Real-World Use:** The data an algorithm encounters in real-world use might differ from the training data

2. **Unforeseen Consequences:** Even well-designed algorithms can have unintended consequences

3. **Shifts in Societal Biases:** Societal biases are constantly evolving. Monitoring helps ensure the algorithm doesn't become biased due to changes in the social landscape

4. **Building Trust and Transparency:** Regular monitoring demonstrates a commitment to fairness and helps build trust in the algorithms used for healthcare decisions

Avoiding Perpetuating Bad AI

Strategies to mitigate bias in datasets:

1. **Identify potential sources of bias:** Analyze data collection methods, sampling procedures, and variable selection for potential biases
2. **Utilize bias mitigation techniques:** Apply techniques like data balancing, weighting, or fairness-aware algorithms to mitigate bias in the data
3. **Promote transparency and responsible AI practices:** Document the limitations of the data and potential biases to ensure responsible use of AI models trained on the dataset.

Legal and Regulatory Frameworks

Legal and regulatory frameworks govern the use of algorithms include:

1. **Anti-Discrimination Laws:** Prohibiting discrimination based on protected characteristics such as race, gender, or age
2. **Privacy Regulations:** Safeguarding individuals' privacy rights and regulating the collection and use of personal data
3. **Ethical Guidelines:** Providing guidelines for ethical algorithm development and deployment, issued by professional organizations or government agencies

Compliance is essential

Ethical Considerations and Responsible AI

Principles of responsible AI include:

1. **Fairness:** Ensuring algorithms produce unbiased outcomes across different demographic groups
2. **Transparency:** Making algorithms transparent and understandable to stakeholders
3. **Accountability:** Holding developers and users accountable for the impact of algorithms
4. **Privacy:** Protecting individuals' privacy rights and sensitive information

Adhering to ethical principles is essential for building trust and mitigating potential harms associated with AI

Quiz 4

To mitigate bias in algorithms used for real-world applications, it's important to:

a) Only use the algorithm on datasets with perfectly balanced representation

b) Continuously monitor the algorithm's performance across different demographics and adjust as needed

c) Focus solely on optimizing the accuracy of the algorithm during development

d) Limit the complexity of the algorithm to ensure easy interpretability

Quiz 5

Societal biases can potentially be reflected in algorithm output because:

a) Algorithms are inherently malicious and designed to discriminate
b) Algorithms are completely objective and not influenced by external factors
c) Algorithms learn from data, which can contain societal biases
d) Algorithms are programmed by biased human creators

ScHARe

Open Science and Reproducibility

Introduction to Open Science

Open Science is a paradigm shift in research practices aimed at fostering **transparency, collaboration, and accessibility**

It promotes the sharing of:

- research data
- methodologies
- findings

to accelerate scientific progress and innovation

Open Science Principles

1. **Open Access:** Making research findings freely available online, often through open access journals or repositories

2. **Open Data:** Sharing the raw data used in research studies

3. **Open Methodology:** Making the research methods and protocols used in a study openly available

4. **Open Source:** Using and sharing open-source software for data analysis and other research tasks

5. **Open Peer Review:** Making the peer review process more transparent, allowing reviewers' identities or comments to be disclosed to some extent

6. **Reproducibility:** Conducting research in a way that allows others to reproduce the findings

7. **Collaboration:** Encouraging researchers to work together and share their findings openly

8. **Public Engagement:** Communicating scientific findings to the public in a clear, understandable way

NIH and ScHARe Embrace Open Science

Promoting Open Science is crucial for:

- advancing **knowledge** discovery
- improving research **reproducibility**
- promoting public **trust** in science

The new NIH Data Sharing Policy

The National Institutes of Health (NIH) implemented a new Data Management and Sharing (DMS) Policy in January 2023

Goal: promote transparency and responsible data management in scientific research

The new NIH Data Sharing Policy

Who is affected:

- Researchers applying for NIH funding (grants, contracts)
- Intramural NIH researchers (conducting research within the NIH itself)

- **Core principle:** Maximize the appropriate sharing of scientific data

The new NIH Data Sharing Policy

Benefits of data sharing:

- Enables verification and **reproduction** of findings
- Fosters **collaboration** among researchers
- Accelerates scientific **progress**
- Increases the **impact** of funded research

The new NIH Data Sharing Policy

NIH National Institutes of Health

What data needs to be shared?

- Scientific data generated from the funded/conducted research, with exceptions for data with privacy risks, commercialization potential, or security concerns

How is data shared?

- Researchers must submit a *Data Management and Sharing Plan* outlining how they will handle data, ensure its quality and security, and deposit it in a suitable public repository

The **ScHARe repository** is designed to meet the data sharing requirements of the NIH data sharing policy

The ScHARe Repository for Data Management:

- serves as a **centralized platform** for storing, managing, and sharing research data related to health disparities
- adheres to **FAIR principles** (Findable, Accessible, Interoperable, Reusable) to ensure data discoverability and usability
- supports **Open Science** initiatives and promotes collaborative research

The ScHARe CDEs

The ScHARe repository focuses on **Common Data Elements** (CDEs), standardizing data and metadata to facilitate interoperability and data reuse

- **Common Data Elements** (CDEs) are standardized, precisely defined data points used consistently across different research studies
- They act as building blocks for collecting and sharing data in a comparable and interoperable manner

NIH CDE Repository

Benefits of CDEs

- **Standardization:** Ensures consistency in data collection, formatting, and reporting across studies
- **Interoperability:** Facilitates data integration and comparison between different datasets and projects
- **Efficiency:** Streamlines data management processes, reducing redundancy and errors in data handling
- **Collaboration:** Promotes collaboration and data sharing among researchers
- **Quality:** Enhances data quality and reliability by adopting standardized data collection and reporting practices

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

Quiz 6

The core principle of the NIH Data Management and Sharing (DMS) Policy emphasizes:

a) Restricting access to all scientific data generated by NIH-funded research
b) Maximizing the appropriate sharing of scientific data while considering ethical and legal limitations
c) Encouraging the publication of research findings in open-access journals only
d) Prioritizing data privacy over all other considerations

Quiz 7

Which of the following is a primary benefit of using common data elements (e.g., standardized variable names, units of measure)?

a) Improves data security and privacy
b) Simplifies data analysis and comparison across studies
c) Reduces the overall size of data storage requirements
d) Enhances the visual appeal of data presentations

ScHARe

Brainstorming for Research Projects

Let’s brainstorm health disparities research ideas

Let’s consider:

- **innovative approaches and methodologies, such as AI**
- **datasets publicly available on ScHARe**

Health disparities

A health disparity is a health difference that adversely affects disadvantaged **populations** in comparison to a reference population, based on one or more **health outcomes**

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.

Populations with Health Disparities

Populations that experience health disparities include:

- Racial and ethnic minority groups
- People with lower socioeconomic status (SES)
- Underserved rural communities
- Sexual and gender minority (SGM) groups
- People with disabilities

Inequities can lead to health disparities

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 good 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**

How do these **nonmedical factors interact with each other and biology** to influence health?

Artificial Intelligence
may have the answer

Identifying research gaps

Areas with limited research in the current health disparity research landscape:

1. **Social Determinants of Health (SDoH) Interactions**

   A gap exists in understanding the complex interactions between SDoH factors and how they contribute to health disparities across different populations

2. **Precision Disparities**

   The rise of personalized medicine using genomics raises concerns about potential disparities in access and benefits. How do genetic and social factors intertwine to create “precision health disparities”?

3. **Intersectionality and Health**
   Traditional research focuses on single demographic factors (e.g., race, gender). How do multiple social identities (e.g., Black woman, LGBTQ+, immigrant) intersect and influence health outcomes?

4. **Role of Implicit Bias in Healthcare Systems**
   How does implicit bias in healthcare delivery affect treatment recommendations and patient experiences? What interventions can mitigate its effects?

5. **Digital Divide and Disparities**

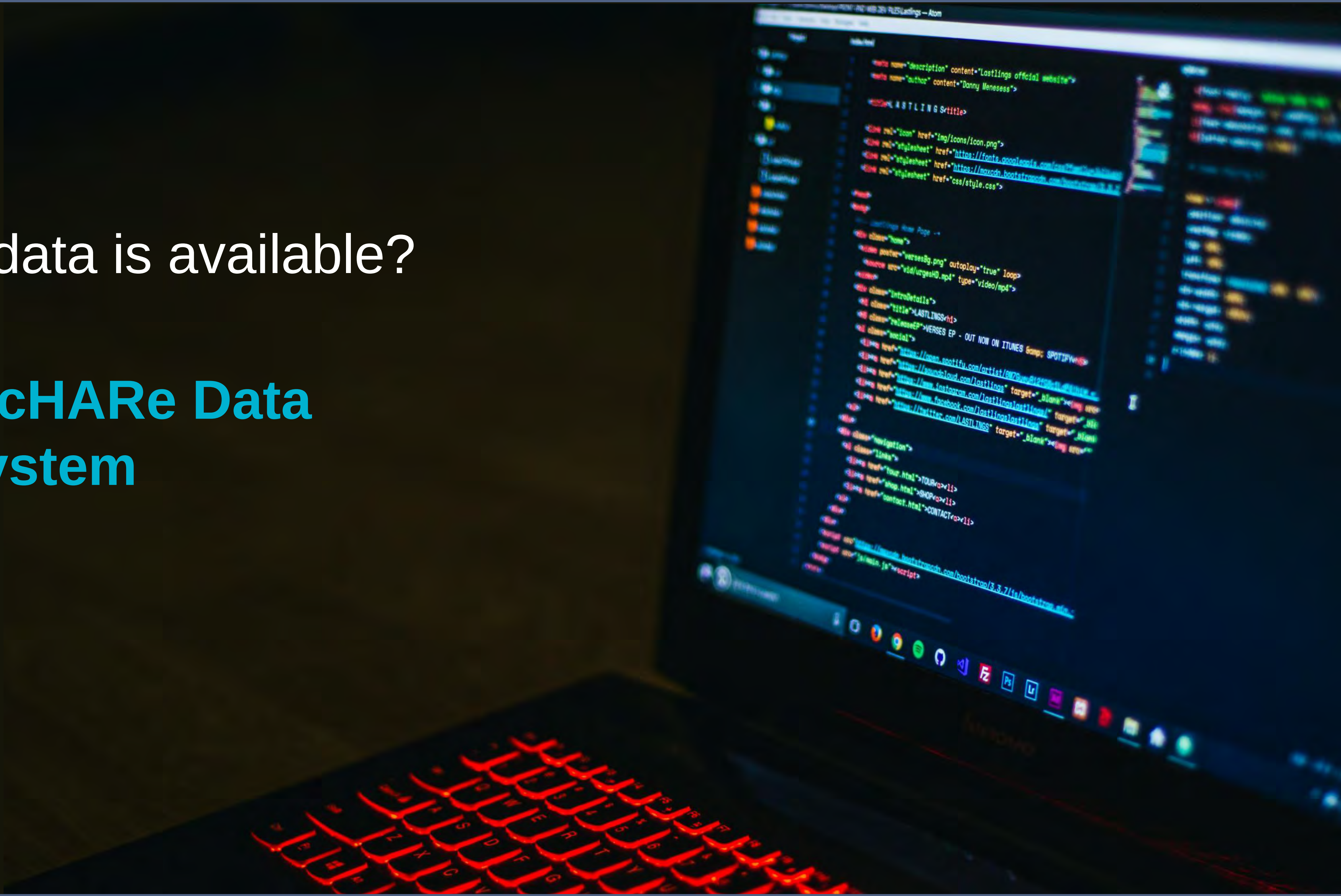
   Lack of access to technology can exacerbate health disparities. How to leverage technology to improve health outcomes for underserved populations while ensuring equitable access?

6. **Environmental Exposures and Disparities**
   Communities of color and low-income populations are often disproportionately exposed to environmental hazards. What are the long-term health effects of these exposures?

7. **Longitudinal Studies on Disparities**
   Many studies are cross-sectional. Longitudinal studies that track individuals over time are crucial for understanding the progression of disparities

What data is available?

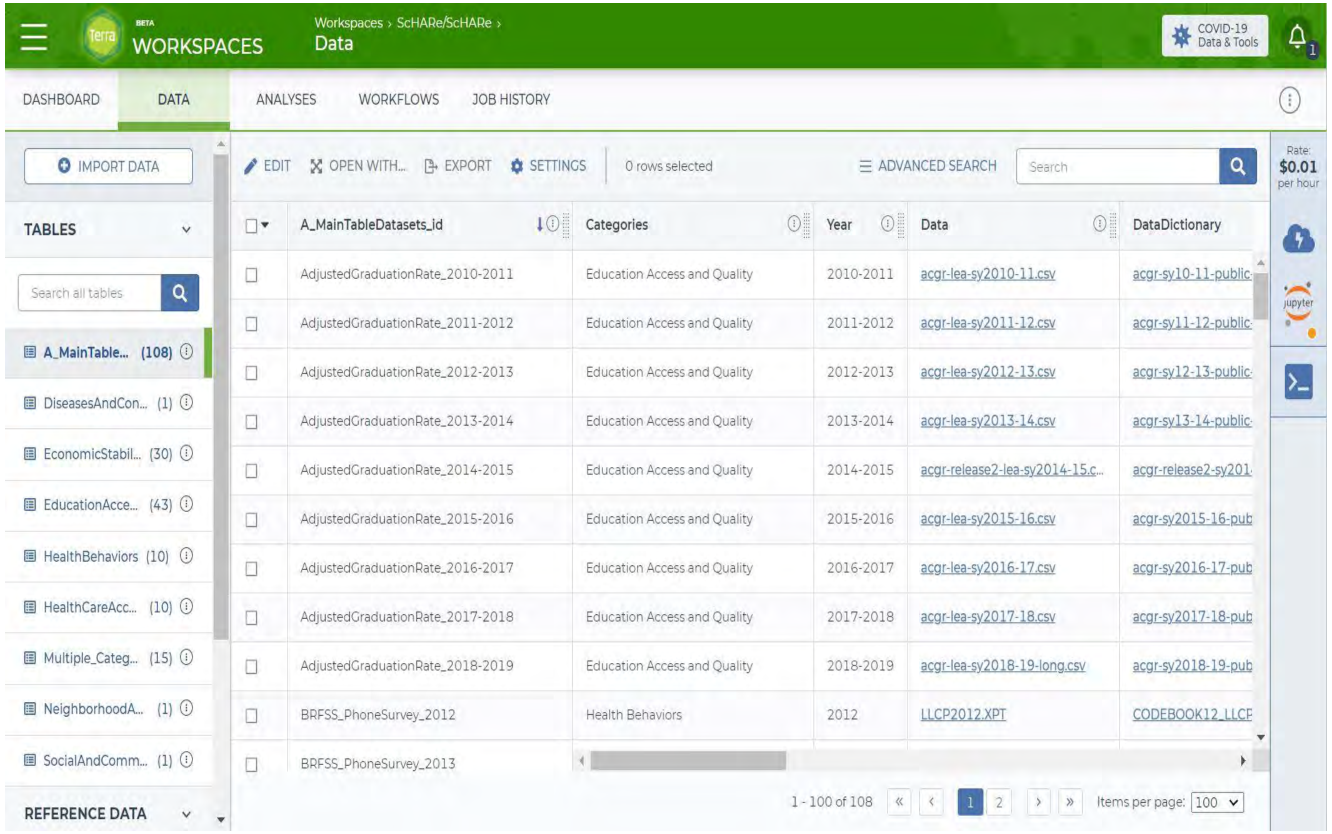
The ScHARe Data Ecosystem



SDoH-related Datasets Available on ScHARe: A Valuable Resource

ScHARe provides a valuable platform for researchers seeking **SDoH-related data**

Explore the available datasets to identify potential resources that align with your research interests in social determinants of health and their impact on various health outcomes



ScHARe Ecosystem

The ScHARe Data Ecosystem is comprised of:

1. **Google Hosted Public Datasets:** publicly accessible, federated, de-identified datasets hosted by Google through the Google Cloud Public Dataset Program
   Example: *American Community Survey (ACS)*

2. **ScHARe Hosted Public Datasets:** publicly accessible, de-identified datasets hosted by ScHARe
   Example: *Behavioral Risk Factor Surveillance System (BRFSS)*

3. **ScHARe Hosted Project Datasets:** 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*

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)

ScHARe Ecosystem: ScHARe hosted datasets

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

240+ 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.

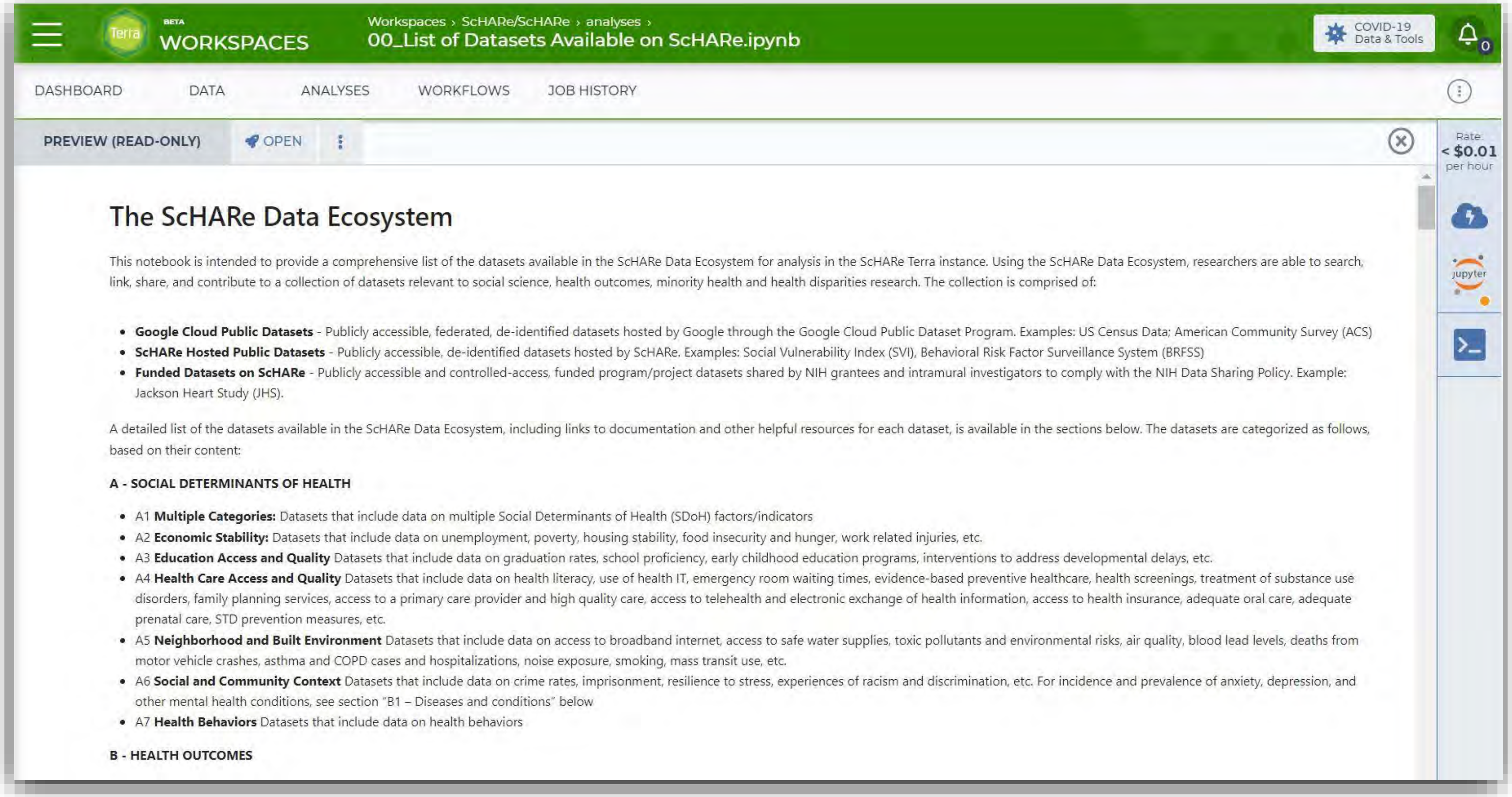
Examples:

- **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

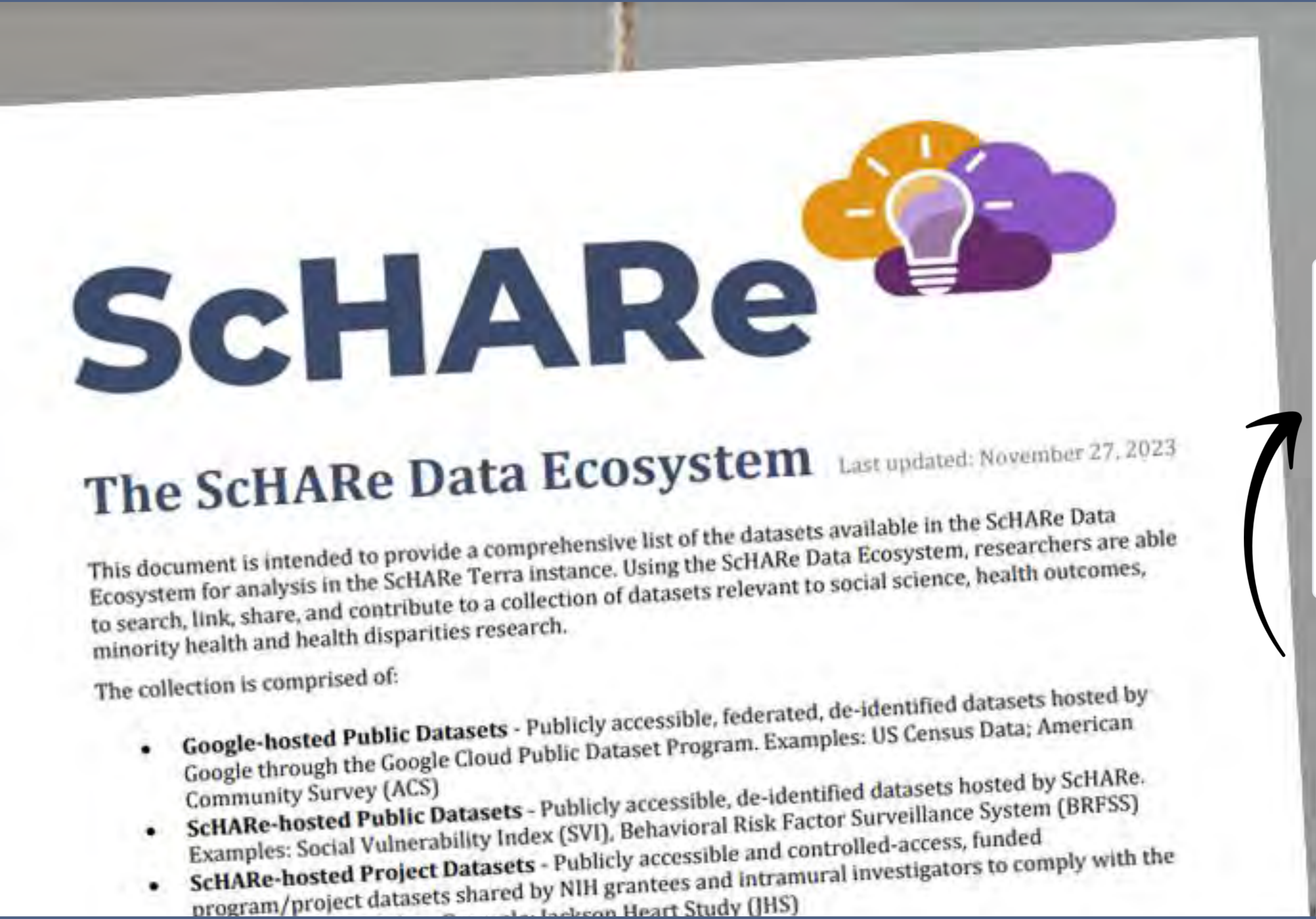
How to check what data is available on ScHARe

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



ScHARe datasets


ScHARe

The ScHARe Data Ecosystem

Last updated: November 27, 2023

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 ... Examples: Jackson Heart Study (JHS)

Scan me

bit.ly/ScHARe-datasets

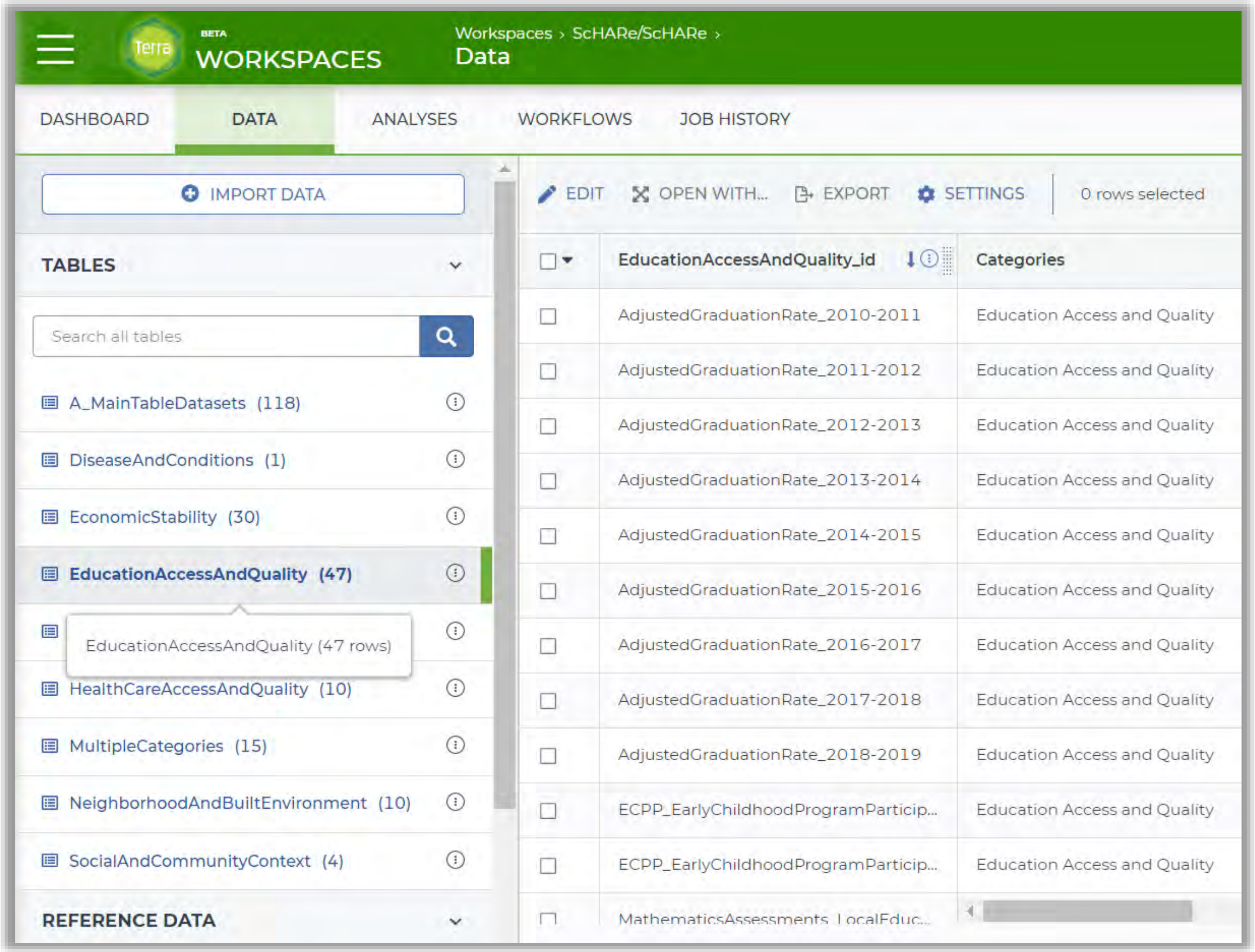
ScHARe demo

How to access available data on ScHARe

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



ScHARe

Potential Projects

Potential projects leveraging AI and Big Data

These examples showcase the intersection of different **datasets** and the application of diverse **AI tools** to gain insights into the social determinants of health and their impact on health outcomes and disparities in minority populations

#1: Geospatial Analysis of Environmental Factors

- **Objective:** Explore the impact of environmental conditions on health outcomes, especially in minority communities

- **Methodology:**
  - Combine environmental datasets (air quality, pollution levels) with health records using geospatial analytics
  - AI models can reveal spatial patterns, helping identify areas with higher health risks in minority populations

This information can inform policies addressing environmental justice and public health

#1: Geospatial Analysis of Environmental Factors

- **Datasets:**
  - Environmental Protection Agency (EPA) Air Quality Data: Provides information on air pollutants and air quality indices
  - Health and Nutrition Examination Survey (NHANES): Includes health data, including respiratory health indicators

- **AI Tools:**
  - Geospatial Analytics Tools: Geographic Information System (GIS) platforms like ArcGIS or QGIS to map environmental data and health outcomes
  - Machine Learning for Spatial Analysis: Algorithms for spatial regression or clustering to identify areas with higher health risks

#2: Education and Health Disparities Analysis

- **Objective:** Examine the link between educational disparities and health outcomes in minority communities

- **Methodology:**
  - Merge educational attainment data with health records, applying AI techniques to discern patterns
  - Explore how educational opportunities influence health behaviors, preventive care, and overall well-being

This interdisciplinary research can inform education and public health policies aimed at reducing health disparities

#2: Education and Health Disparities Analysis

- **Datasets:**
  - National Center for Education Statistics (NCES) Educational Attainment Data: Contains data on educational attainment by demographics
  - Behavioral Risk Factor Surveillance System (BRFSS): Includes self-reported health data and behaviors
- **AI Tools:**
  - Predictive Modeling: Utilize algorithms like logistic regression or neural networks to predict health outcomes based on educational disparities
  - Causal Inference Techniques: Apply methods such as propensity score matching to isolate the impact of education on health

#3: Causal links between chronic stress associated with social adversity and health disparities

Health is adversely affected by social disadvantage:

1. **Neighborhoods influence health through their physical and geographic characteristics**:
   - air and water quality
   - lead paint exposure
   - proximity to health promoting features (e.g.: hospitals, healthy food stores)
   - proximity to health suppressing features (e.g.: toxic factories, fast food)
   - access to green space, etc.

2. **Chronic stress of social disadvantage, socioeconomic inequality, and racial discrimination can influence health** through a variety of biological pathways:
   1. neuroendocrine
   2. developmental
   3. immunologic
   4. vascular

Objective: Examine the role of epigenetic modifications as a causal link between chronic stress associated with social adversity and health disparities, and impact of mitigating factors

Research Projects Brainstorming

What research projects do you believe it would be worthwhile to pursue?

ScHARe

Resources

ScHARe resources

Support made available to users:

ScHARe-specific

- ScHARe documentation
- Email support

Platform-specific

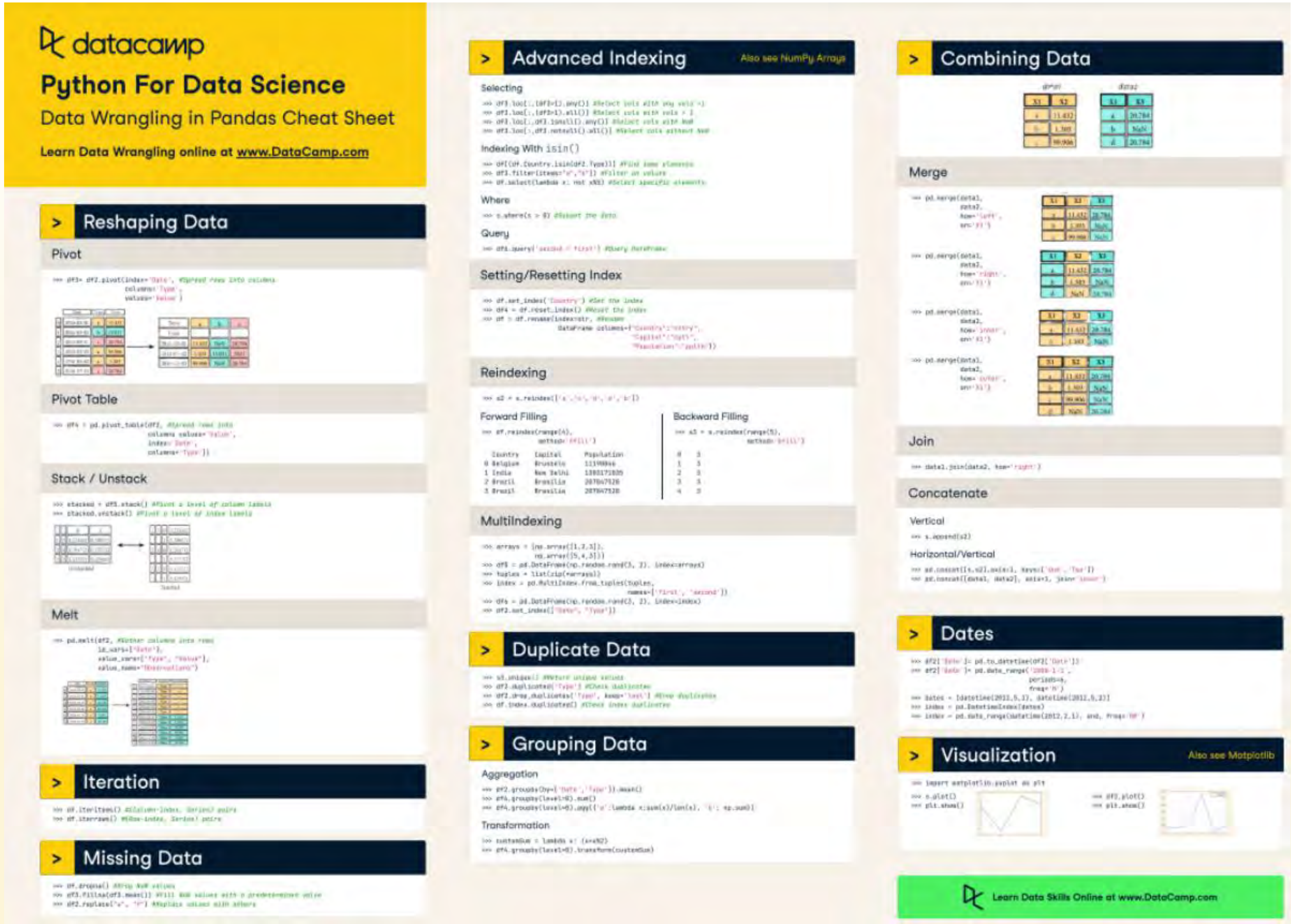
- Terra-specific support
- Terra-specific documentation

ScHARe resources

Training opportunities made available to users:

- Monthly **Think-a-Thons**
- **Instructional materials** and slides made available online on NIMHD website
- **YouTube videos**
- **Links to relevant online resources** and training on NIMHD website
- **Pilot credits** for testing ScHARe for research needs
- **Instructional Notebooks** in ScHARe Workspace with instructions for:
  - Exploring the data ecosystem
  - Setting your workspace up for use
  - Accessing and interacting with the categories of data accessible through ScHARe

ScHARe resources: cheatsheets



Credits: datacamp.com

Terra resources

If you are new to Terra, we recommend exploring the following resources:

- Overview Articles: Review high-level docs that outline what you can do in Terra, how to set up an account and account billing, and how to access, manage, and analyze data in the cloud
- Video Guides: Watch live demos of the Terra platform's useful features
- Terra Courses: Learn about Terra with free modules on the Leanpub online learning platform
- Data Tables QuickStart Tutorial: Learn what data tables are and how to create, modify, and use them in analyses
- Notebooks QuickStart Tutorial**:** Learn how to access and visualize data using a notebook
- Machine Learning Advanced Tutorial: Learn how Terra can support machine learning-based analysis

ScHARe

Thank you

Think-a-Thon poll

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

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

Think-a-Thon poll

2. Rate the pace of the instruction for yourself:

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

Think-a-Thon 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