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

```
#
# Example file for [illegible]
#
import urllib.request
import json

def printResults([illegible]):
  # Use the json [illegible]
  theJSON = json.[illegible]

  # now we can [illegible]
  if "title" in [illegible]
    print(theJSON[illegible]

  # output the [illegible]
  count = theJSON[illegible]
  print (str(count) [illegible]

  # for each event, [illegible]
  for i in theJSON[illegible]
    print(i["propert[illegible]
  print("[illegible]

  # print the events [illegible]
  for i in the[illegible]
```

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

```
#
# Example file for [illegible]
#
import urllib.request
import json

def printResults([illegible]):
  # Use the json [illegible]
  theJSON = json.[illegible]

  # now we can [illegible]
  if "title" in [illegible]
    print(theJSON[illegible]

  # output the [illegible]
  count = theJSON[illegible]
  print (str(count) [illegible]

  # for each event, [illegible]
  for i in theJSON[illegible]
    print(i["properties"[illegible]
  print("[illegible]

  # print the events [illegible]
  for i in the[illegible]
```

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?

```
#
# Example file for [illegible]
#
import urllib.request
import json

def printResults([illegible]):
  # Use the json [illegible]
  theJSON = json.[illegible]

  # now we can [illegible]
  if "title" in [illegible]
    print(theJSON[illegible]

  # output the [illegible]
  count = theJSON[illegible]
  print (str(count) + [illegible]

  for each event, [illegible]
  for i in theJSON[illegible]
    print(i["properties"][illegible]
  print("[illegible]

  # print the events [illegible]
  for i in the[illegible]
    if [illegible]
```

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

NIH

National Institutes of Health

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

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

The ScHARe CDEs

- **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

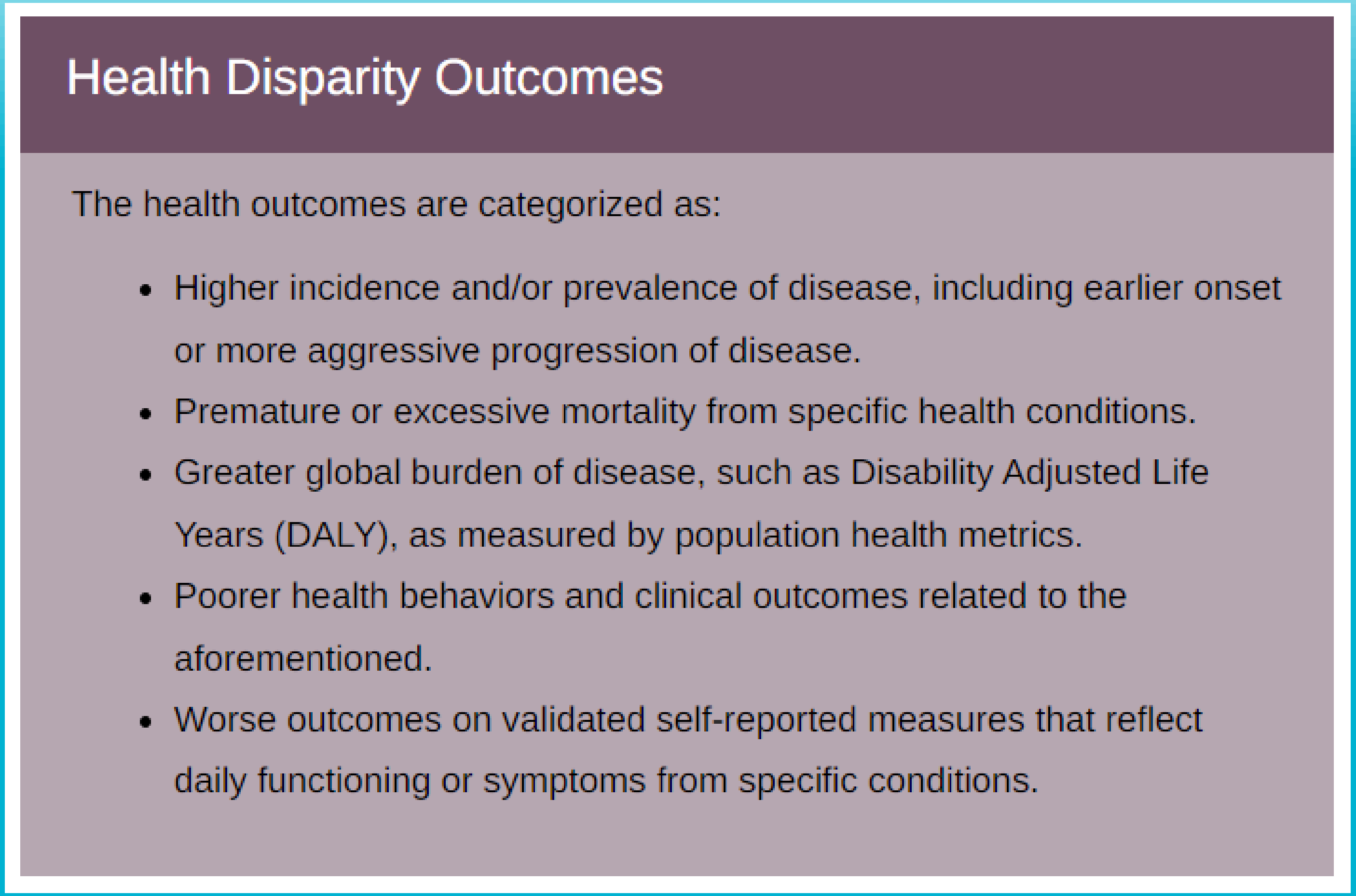
Let’s brainstorm health disparities research ideas

Let’s consider:

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

Health disparities

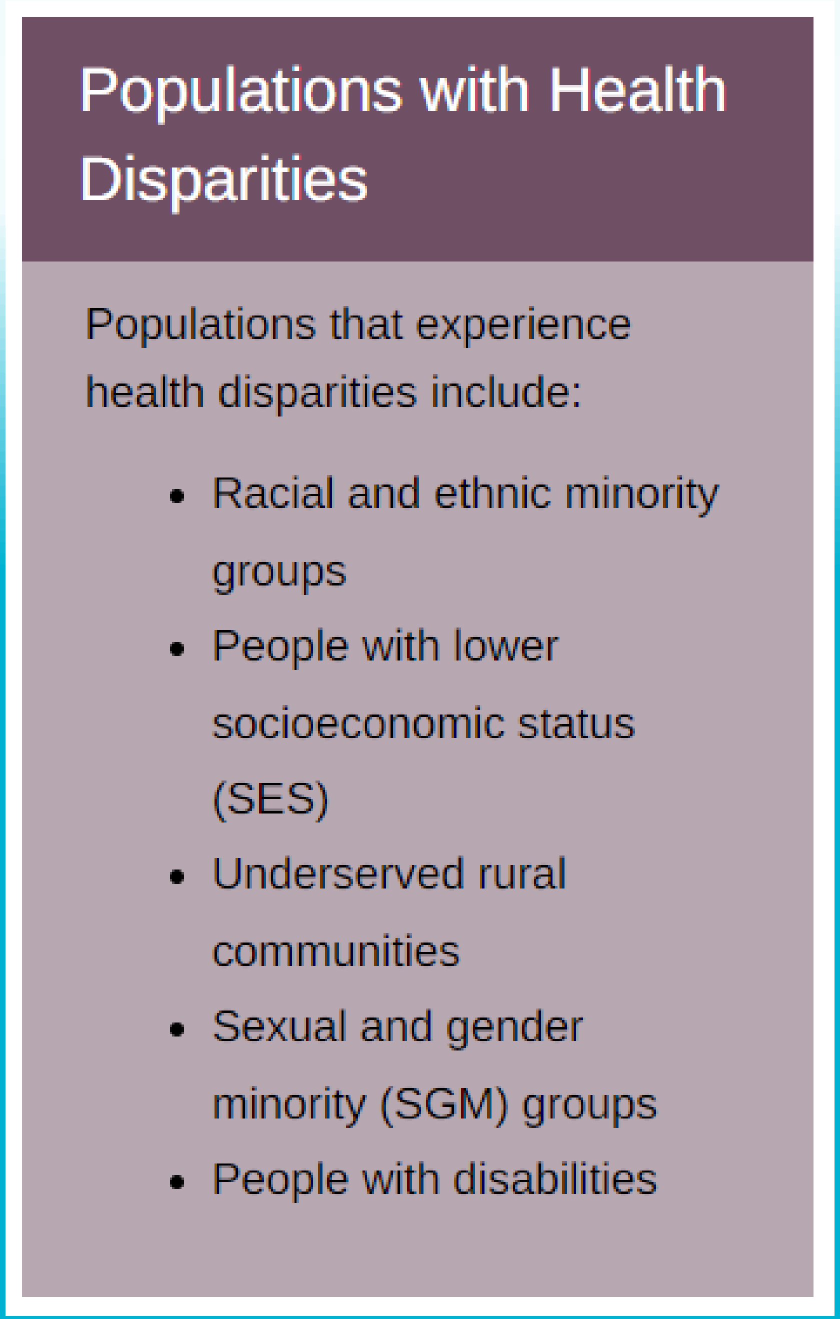
A health disparity is a health difference that adversely affects disadvantaged **populations** in comparison to a reference population, based on one or more **health outcomes**



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:

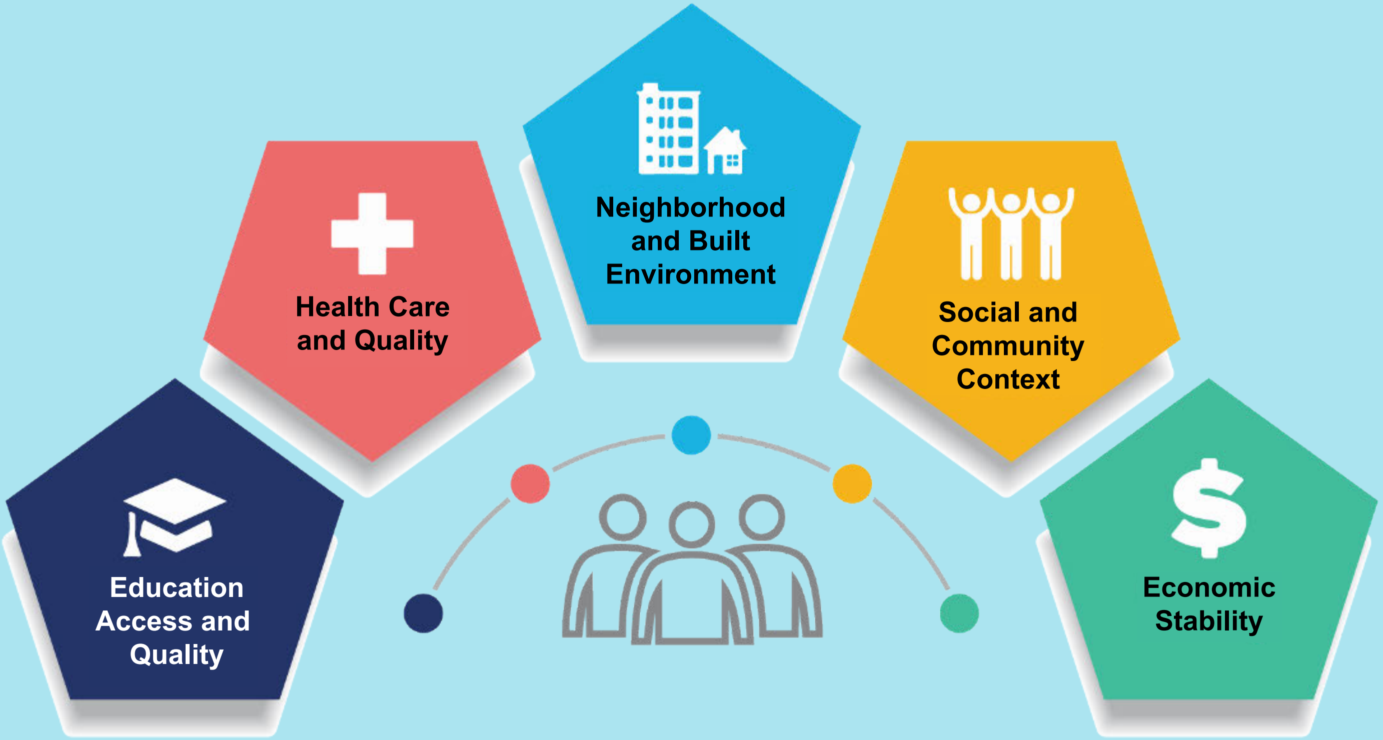
- Racial and ethnic minority groups
- People with lower socioeconomic status (SES)
- Underserved rural communities
- Sexual and gender minority (SGM) groups
- People with disabilities

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



Education Access and Quality

Health Care and Quality

Neighborhood and Built Environment

Social and Community Context

Economic Stability

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**

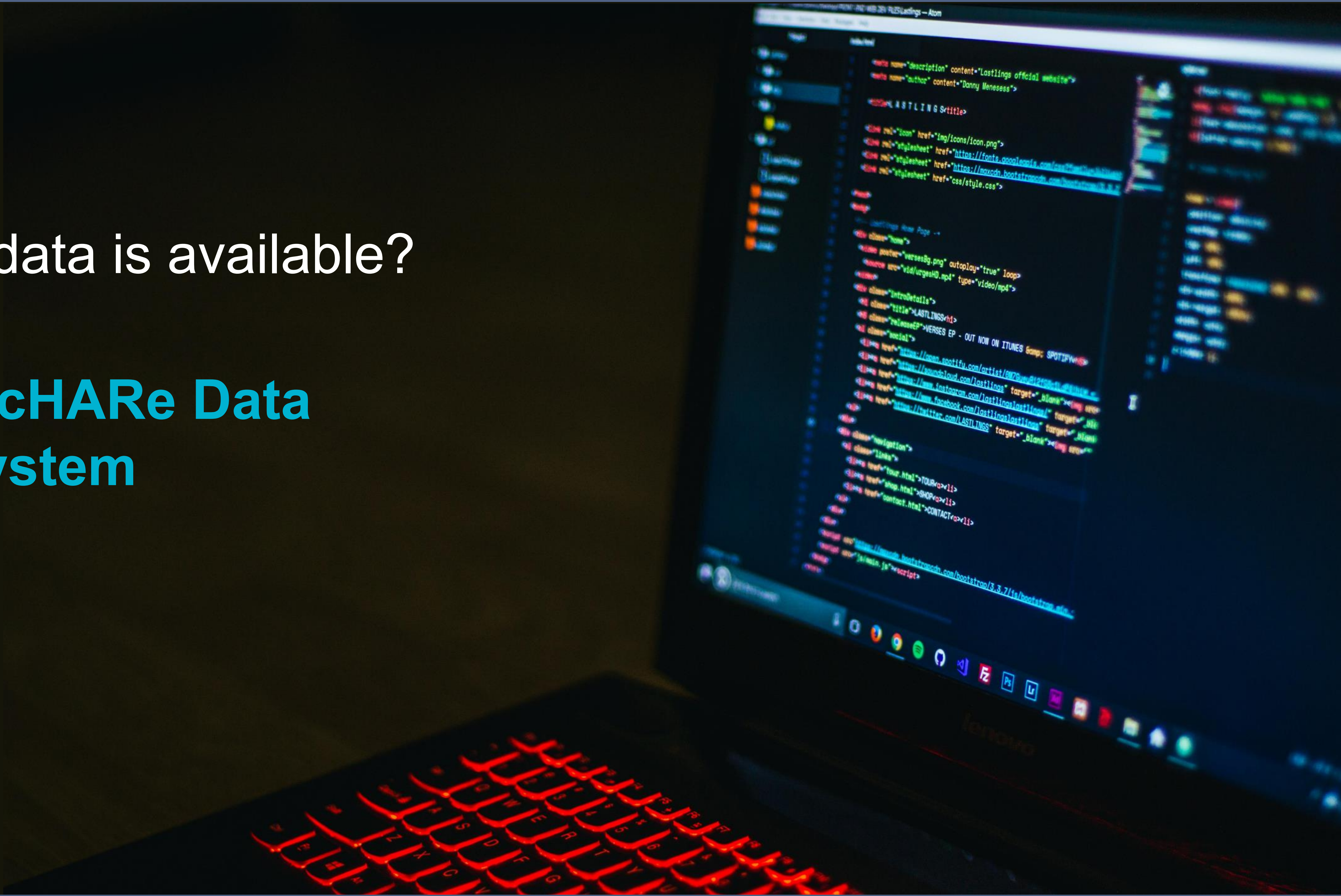
   Lack of access to technology can exacerbate health disparities. How to leverage technology to improve health outcomes for underserved populations while ensuring equitable access?

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?

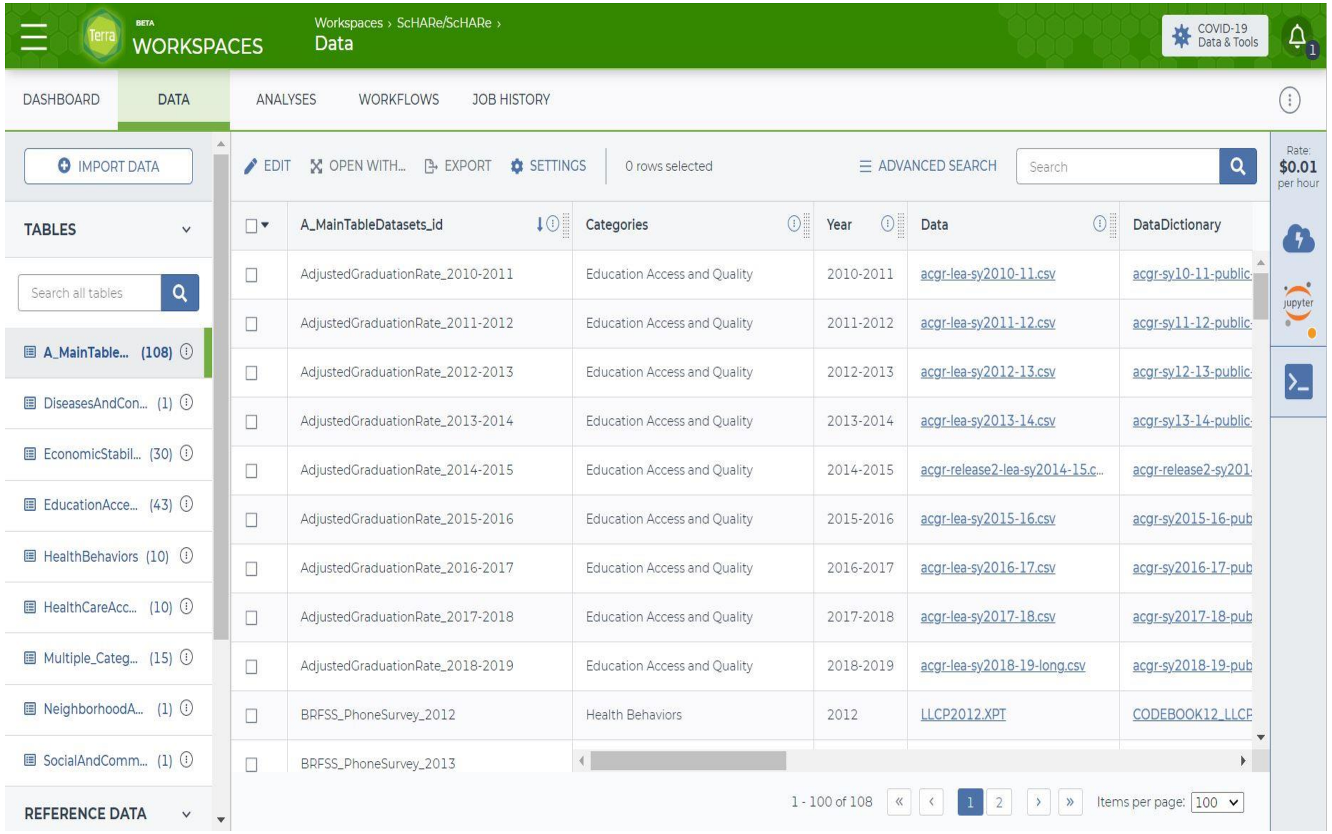
The ScHARe Data Ecosystem



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



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

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

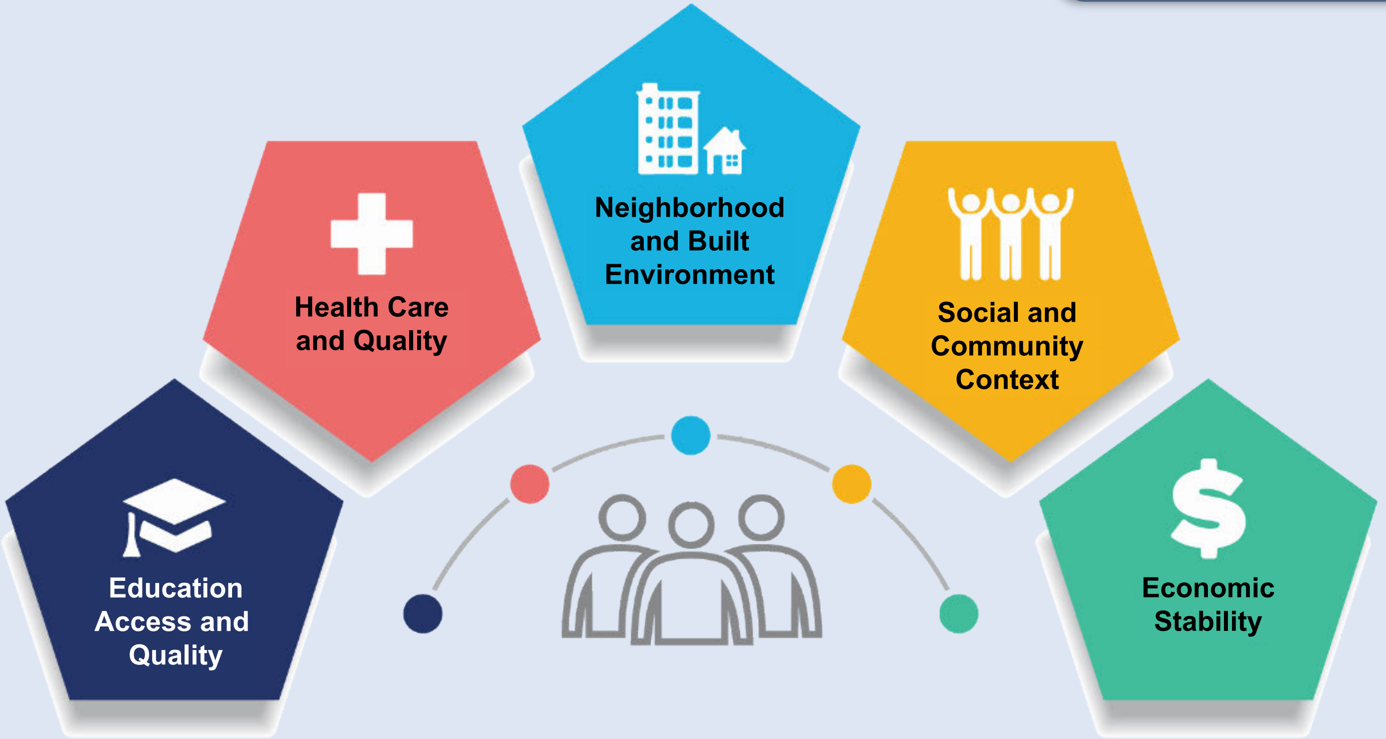
Organized based on the **CDC SDoH categories**, with the addition of *Health Behaviors* and *Diseases and Conditions*:

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**


Education Access and Quality
Health Care and Quality
Neighborhood and Built Environment
Social and Community Context
Economic Stability

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.

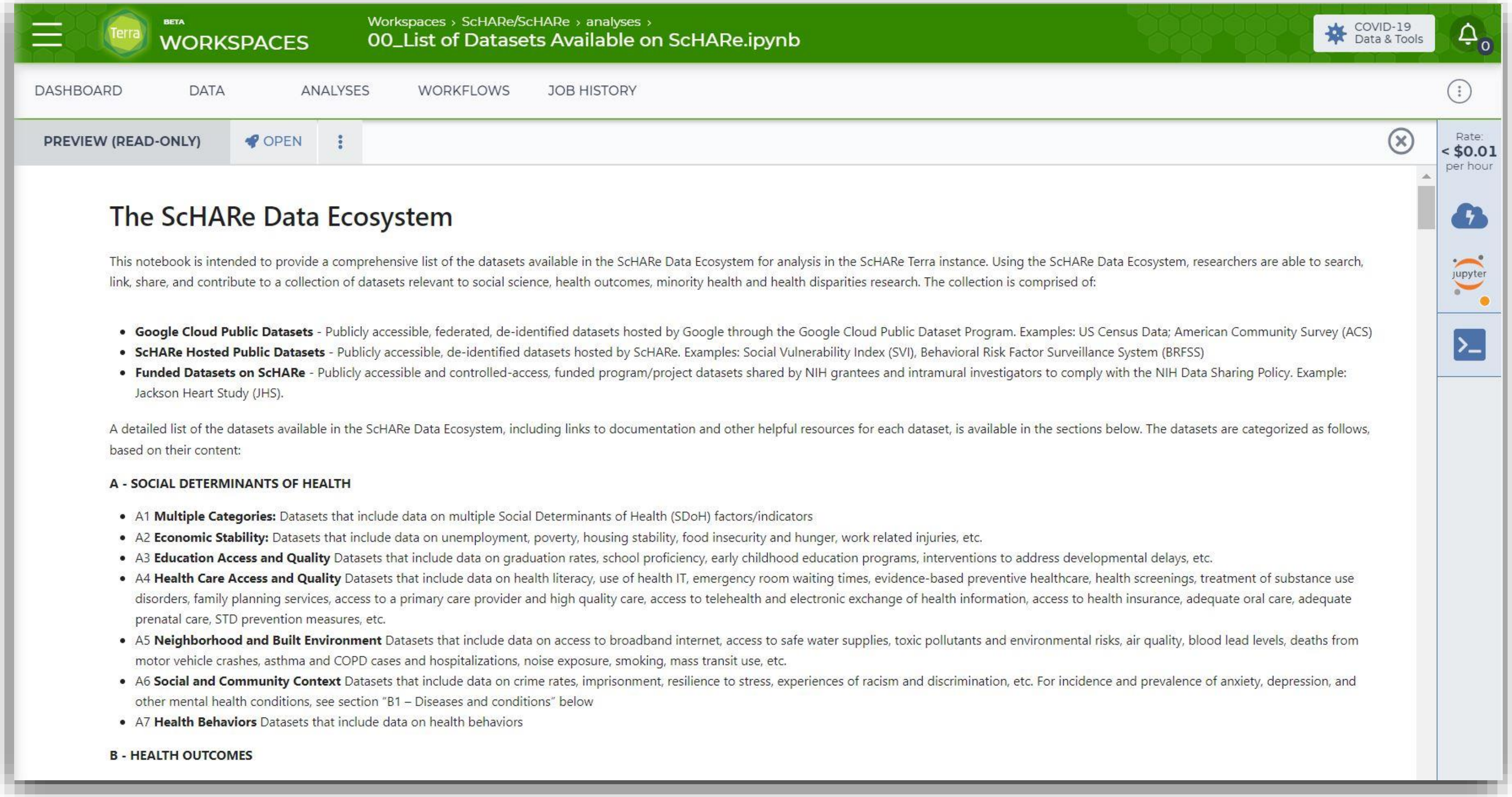
Examples:

- **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



The ScHARe Data Ecosystem

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

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

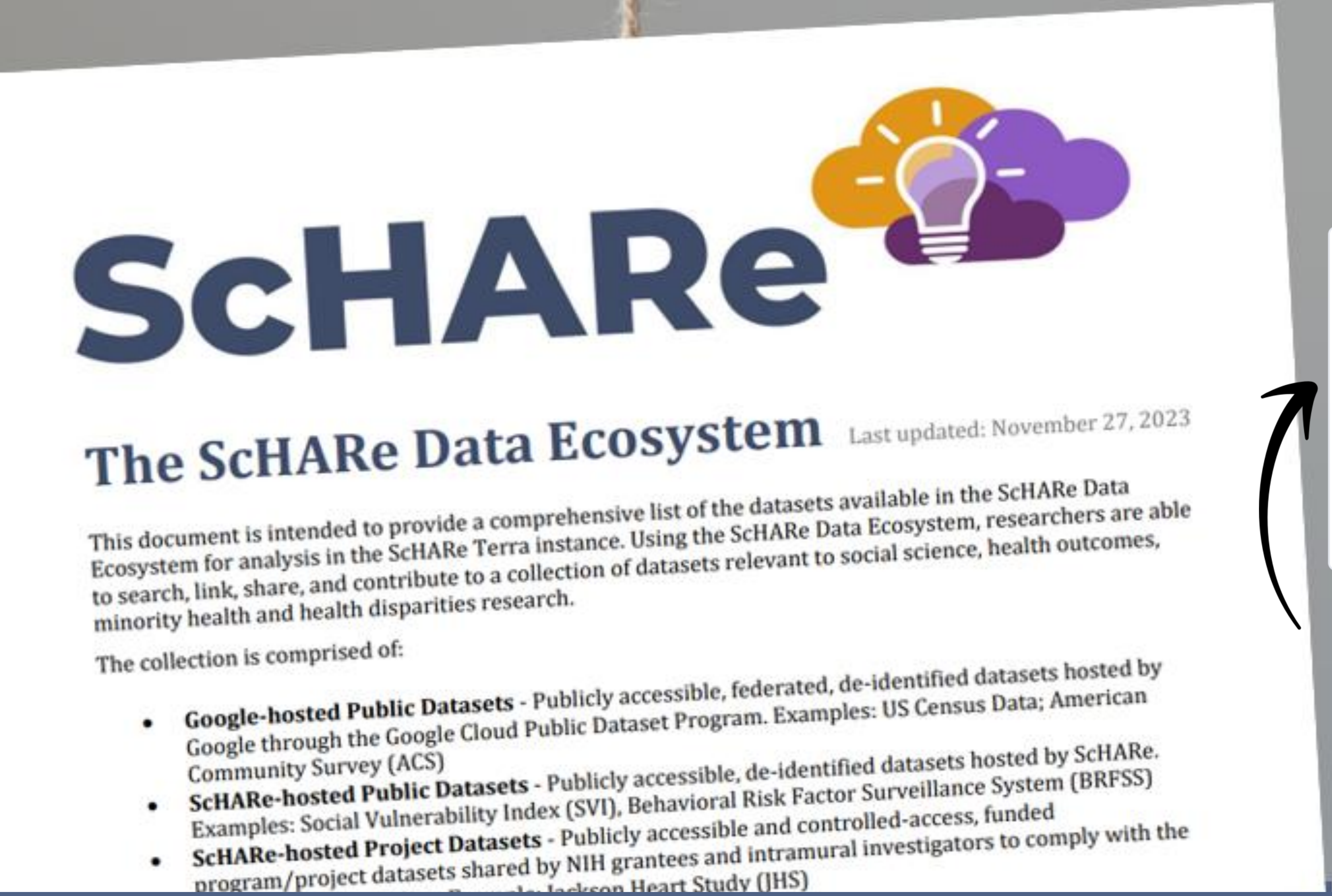
A detailed list of the datasets available in the ScHARe Data Ecosystem, including links to documentation and other helpful resources for each dataset, is available in the sections below. The datasets are categorized as follows, based on their content:

A - SOCIAL DETERMINANTS OF HEALTH

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

B - HEALTH OUTCOMES

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 Jackson Heart Study (JHS)

Scan me

bit.ly/ScHARe-datasets

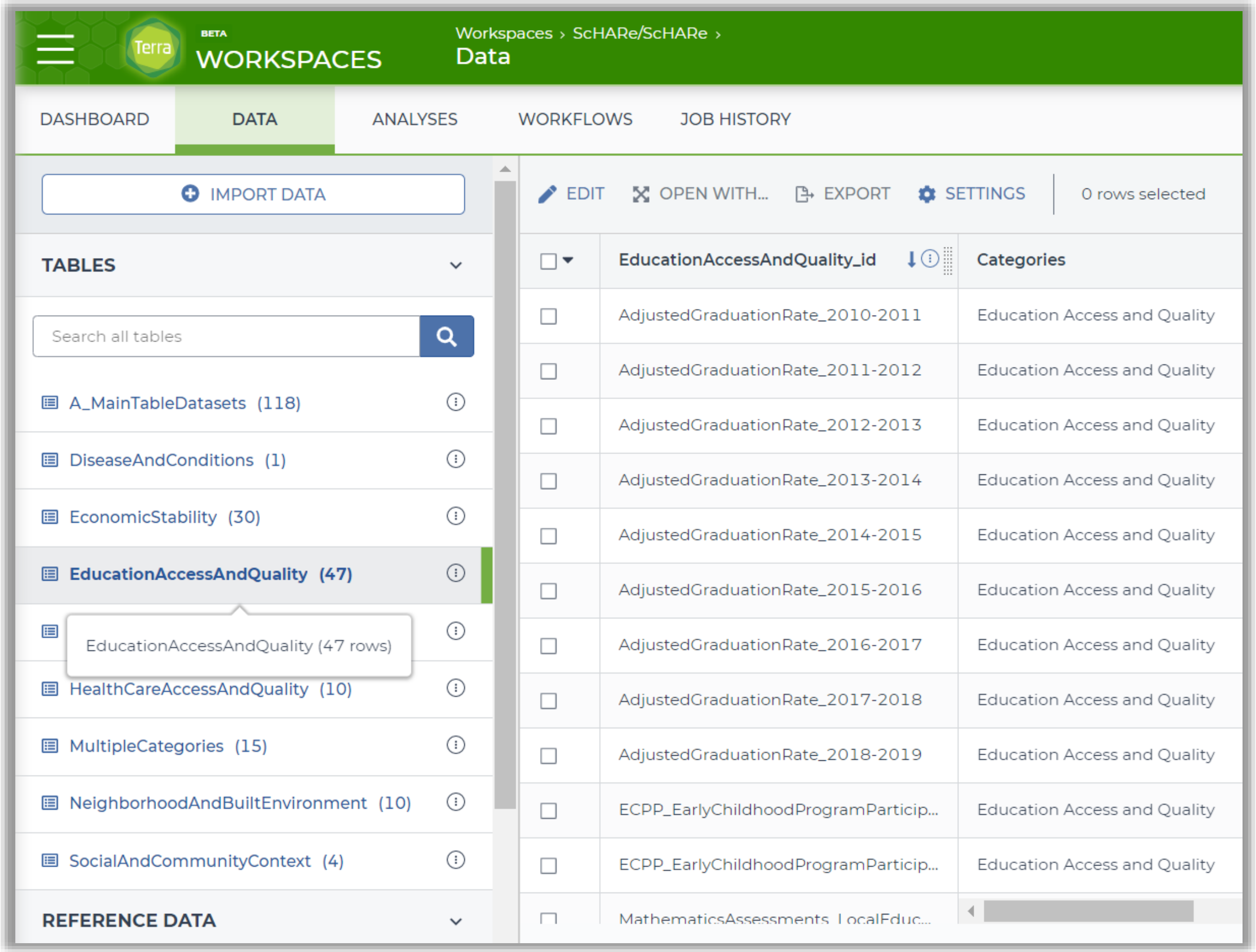
ScHARe
demo

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



- A_MainTableDatasets (118)
- DiseaseAndConditions (1)
- EconomicStability (30)
- EducationAccessAndQuality (47)
- EducationAccessAndQuality (47 rows)
- HealthCareAccessAndQuality (10)
- MultipleCategories (15)
- NeighborhoodAndBuiltEnvironment (10)
- SocialAndCommunityContext (4)

	EducationAccessAndQuality_id	Categories
□	AdjustedGraduationRate_2010-2011	Education Access and Quality
□	AdjustedGraduationRate_2011-2012	Education Access and Quality
□	AdjustedGraduationRate_2012-2013	Education Access and Quality
□	AdjustedGraduationRate_2013-2014	Education Access and Quality
□	AdjustedGraduationRate_2014-2015	Education Access and Quality
□	AdjustedGraduationRate_2015-2016	Education Access and Quality
□	AdjustedGraduationRate_2016-2017	Education Access and Quality
□	AdjustedGraduationRate_2017-2018	Education Access and Quality
□	AdjustedGraduationRate_2018-2019	Education Access and Quality
□	ECPP_EarlyChildhoodProgramParticip...	Education Access and Quality
□	ECPP_EarlyChildhoodProgramParticip...	Education Access and Quality
□	MathematicsAssessments_LocalEduc...	

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

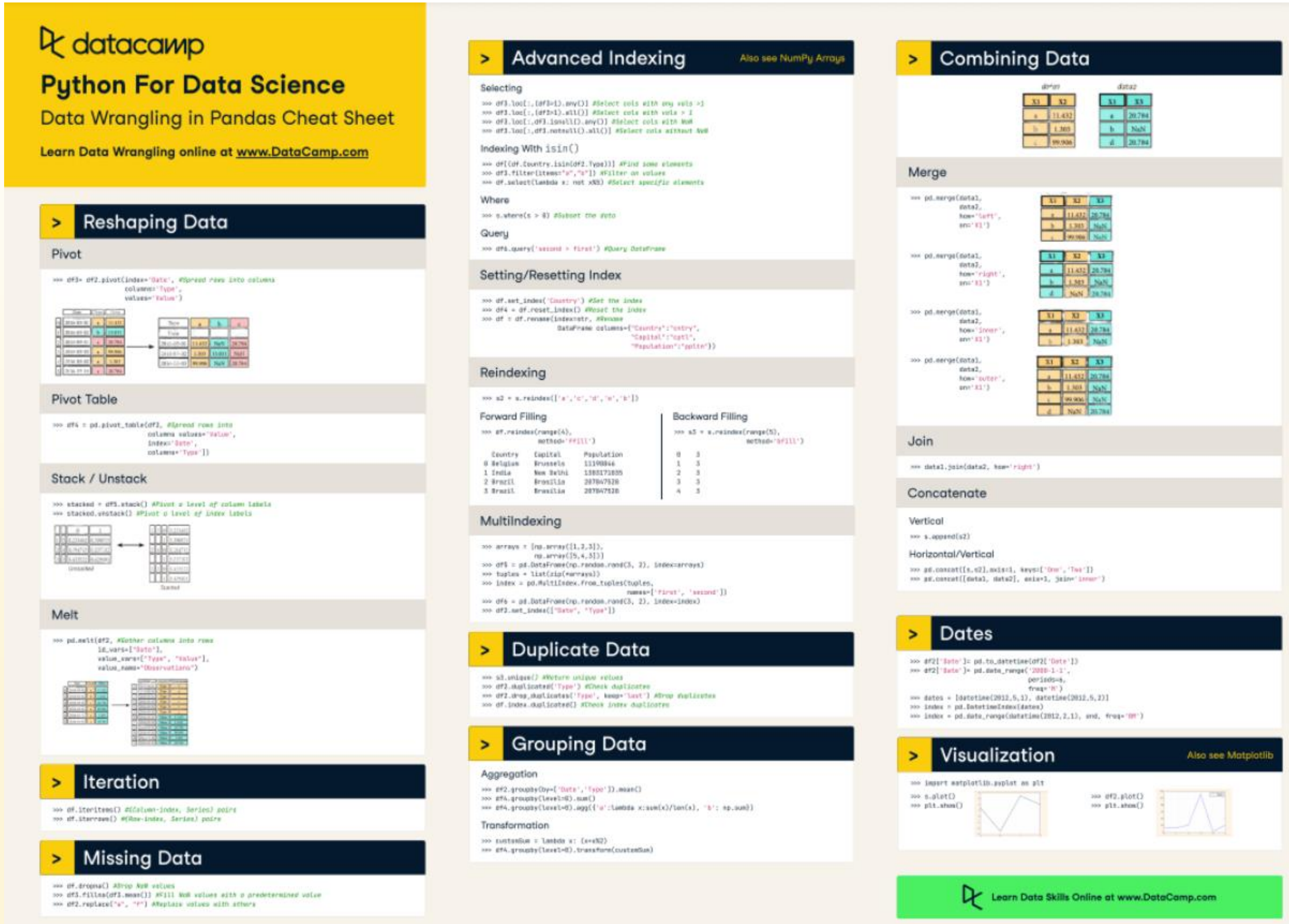
- Terra-specific support
- Terra-specific documentation

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



datacamp

Python For Data Science

Data Wrangling in Pandas Cheat Sheet

Learn Data Wrangling online at www.DataCamp.com

> Reshaping Data

Pivot

```
>>> df3= df2.pivot(index='Date', #Spread rows into columns
                   columns='Type',
                   values='Value')
```

Pivot Table

```
>>> df4 = pd.pivot_table(df2, #Spread rows into
                         columns values='Value',
                         index='Date',
                         columns='Type'])
```

Stack / Unstack

```
>>> stacked = df5.stack() #Pivot a level of column labels
>>> stacked.unstack() #Pivot a level of index labels
```

Unstacked

Stacked

Melt

```
>>> pd.melt(df2, #Gather columns into rows
            id_vars=["Date"],
            value_vars=["Type", "Value"],
            value_name="Observations")
```

> Iteration

```
>>> df.iteritems() #(Column-index, Series) pairs
>>> df.iterrows() #(Row-index, Series) pairs
```

> Missing Data

```
>>> df.dropna() #Drop NaN values
>>> df3.fillna(df3.mean()) #Fill NaN values with a predetermined value
>>> df2.replace("a", "f") #Replace values with others
```

> Advanced Indexing

Also see NumPy Arrays

Selecting

```
>>> df3.loc[:,(df3>1).any()] #Select cols with any vals >1
>>> df3.loc[:,(df3>1).all()] #Select cols with vals > 1
>>> df3.loc[:,df3.isnull().any()] #Select cols with NaN
>>> df3.loc[:,df3.notnull().all()] #Select cols without NaN
```

Indexing With isin()

```
>>> df[(df.Country.isin(df2.Type))] #Find same elements
>>> df3.filter(items="a","b"]) #Filter on values
>>> df.select(lambda x: not x%5) #Select specific elements
```

Where

```
>>> s.where(s > 0) #Subset the data
```

Query

```
>>> df6.query('second > first') #Query DataFrame
```

Setting/Resetting Index

```
>>> df.set_index('Country') #Set the index
>>> df4 = df.reset_index() #Reset the index
>>> df = df.rename(index=str, #Rename
                   DataFrame columns={"Country":"cntry",
                                      "Capital":"cptl",
                                      "Population":"ppltn"})
```

Reindexing

```
>>> s2 = s.reindex(['a','c','d','e','b'])
```

Forward Filling

```
>>> df.reindex(range(4),
               method='ffill')

  Country    Capital     Population
0 Belgium    Brussels    11190846
1 India      New Delhi   1303171035
2 Brazil     Brasília    207847528
3 Brazil     Brasília    207847528
```

Backward Filling

```
>>> s3 = s.reindex(range(5),
                   method='bfill')

0    3
1    3
2    3
3    3
4    3
```

MultiIndexing

```
>>> arrays = [np.array([1,2,3]),
              np.array([5,4,3])]
>>> df5 = pd.DataFrame(np.random.rand(3, 2), index=arrays)
>>> tuples = list(zip(*arrays))
>>> index = pd.MultiIndex.from_tuples(tuples,
                                      names=['first', 'second'])
>>> df6 = pd.DataFrame(np.random.rand(3, 2), index=index)
>>> df2.set_index(["Date", "Type"])
```

> Duplicate Data

```
>>> s3.unique() #Return unique values
>>> df2.duplicated('Type') #Check duplicates
>>> df2.drop_duplicates('Type', keep='last') #Drop duplicates
>>> df.index.duplicated() #Check index duplicates
```

> Grouping Data

Aggregation

```
>>> df2.groupby(by=['Date','Type']).mean()
>>> df4.groupby(level=0).sum()
>>> df4.groupby(level=0).agg({'a':lambda x:sum(x)/len(x), 'b': np.sum})
```

Transformation

```
>>> customSum = lambda x: (x+x%2)
>>> df4.groupby(level=0).transform(customSum)
```

> Combining Data

data1

X1	X2
a	11.432
b	1.303
c	99.906

data2

X1	X3
a	20.784
b	NaN
d	20.784

Merge

```
>>> pd.merge(data1,
             data2,
             how='left',
             on='X1')
```

X1	X2	X3
a	11.432	20.784
b	1.303	NaN
c	99.906	NaN

```
>>> pd.merge(data1,
             data2,
             how='right',
             on='X1')
```

X1	X2	X3
a	11.432	20.784
b	1.303	NaN
d	NaN	20.784

```
>>> pd.merge(data1,
             data2,
             how='inner',
             on='X1')
```

X1	X2	X3
a	11.432	20.784
b	1.303	NaN

```
>>> pd.merge(data1,
             data2,
             how='outer',
             on='X1')
```

X1	X2	X3
a	11.432	20.784
b	1.303	NaN
c	99.906	NaN
d	NaN	20.784

Join

```
>>> data1.join(data2, how='right')
```

Concatenate

Vertical

```
>>> s.append(s2)
```

Horizontal/Vertical

```
>>> pd.concat([s,s2],axis=1, keys=['One','Two'])
>>> pd.concat([data1, data2], axis=1, join='inner')
```

> Dates

```
>>> df2['Date']= pd.to_datetime(df2['Date'])
>>> df2['Date']= pd.date_range('2000-1-1',
                               periods=6,
                               freq='M')
>>> dates = [datetime(2012,5,1), datetime(2012,5,2)]
>>> index = pd.DatetimeIndex(dates)
>>> index = pd.date_range(datetime(2012,2,1), end, freq='BM')
```

> Visualization

Also see Matplotlib

```
>>> import matplotlib.pyplot as plt
>>> s.plot()
>>> plt.show()
```

```
>>> df2.plot()
>>> plt.show()
```

Learn Data Skills Online at www.DataCamp.com

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