



ScHARe

Research Think-a-Thons



National Institutes of Health

The logo for ScHARe, featuring the text "ScHARe" in a bold, dark blue, sans-serif font. The letters are contained within a white circle that is set against a dark blue background.

Be a Part of the Future of Knowledge Generation

May 15, 2024

Deborah Duran, PhD • NIMHD

Luca Calzoni, MD MS PhD Cand. • NIMHD



Look deeper with more eyes

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

Jennifer Aue

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

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



ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

Outline

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

Experience poll

Please check your level of experience with the following:

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

Interest poll

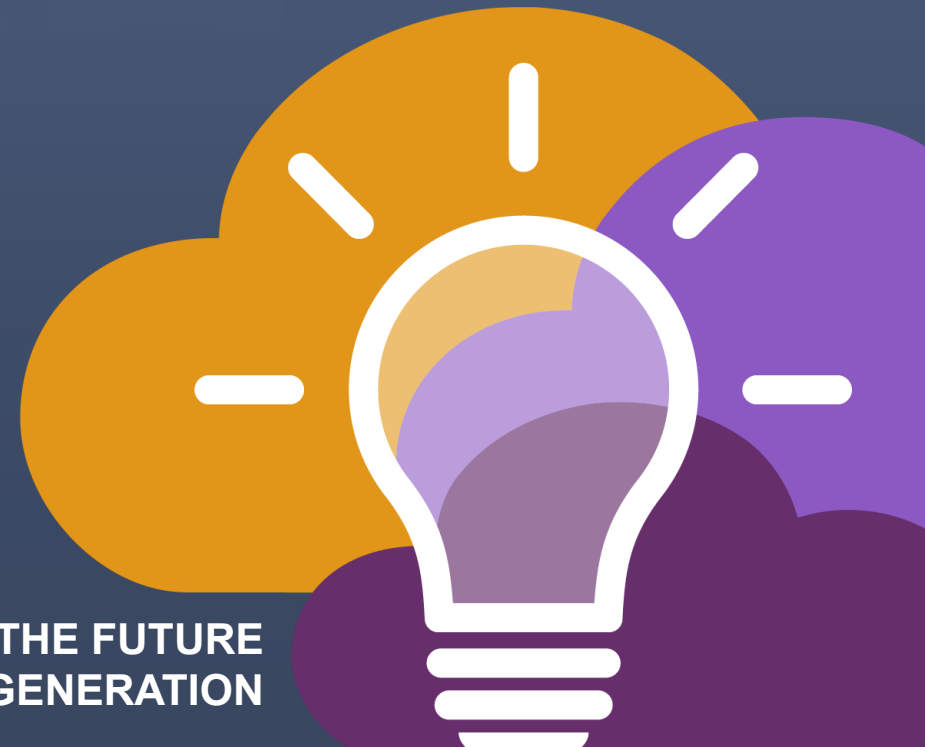
I am interested in (check all that apply):

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

ScHARe

AI and Cloud
Computing 101

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



Big Data

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

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

Data Integrity (data quality) is the overarching completeness, accuracy, consistency, accessibility, and security of the data for its intended purpose.

This should always be assessed before using a dataset.

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

- Findability
- Accessibility
- Interoperability
- Reusability



ScHARe

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

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

Big Data: structured and unstructured data

Structured data is quantitative data that is organized and easily searchable

Some tools used to work with structured data include:

- OLAP
- MySQL
- PostgreSQL
- Oracle Database



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

Some tools used to manage unstructured data include:

- MongoDB
- Hadoop
- Azure



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

Data mining

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

Data mining helps:

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

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



ScHARe

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

Cloud computing

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

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

People can store and process data remotely.

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

Main public cloud service providers:

- **Google**
- Azure
- AWS



ScHARe

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

Google Cloud Platform

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

It provides management tools and modular cloud services, including:

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

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



ScHARe

Through Google, ScHARe offers:

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

Google email address needed.

Artificial Intelligence (AI)

AI is defined as:

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

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

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

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



ScHARe

Many AI projects are built using Python.

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

Machine Learning (ML)

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

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

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

Language Learning Models (LLM)

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

Neural networks

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

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

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

Deep Learning

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

It's a way to “learn by example”.

It requires extensive computing power and labeled data.

Artificial Intelligence Bias

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

Biases can result from:

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

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

Trust in AI

Caution Against:

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

Mistrust of AI

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

Ethical AI

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

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

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

Explainable AI (XAI)

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

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

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

Synthetic / AI Generated DATA

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

Digital Twins

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

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

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

Model Autophagy Disorder (MAD)

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

Cloning data

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

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

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

AI and cloud computing: benefits and challenges



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



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



Benefits

Access to big datasets and large data ecosystems:

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



ScHARe

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

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



Benefits

Deeper insights and better decision making:

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



ScHARe

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

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



Benefits

Intelligent automation and data management:

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



ScHARe

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

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

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



Benefits

Real-time online collaboration:

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



ScHARe

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



Benefits

Increased security:

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



ScHARe

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

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



Benefits

Lower costs:

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



ScHARe

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

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



Challenges

Lack of knowledge and expertise:

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



ScHARe

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



Challenges

Data privacy and security - or misperceptions therein:

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



ScHARe

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



Challenges

Performance, reliability and availability:

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



ScHARe

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

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



Challenges

AI bias:

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



ScHARe

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

ScHARe was created to:

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



We want to hear from you

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

Cost · Knowledge · Research applications · Other

ScHARe meets challenges of cloud computing adoption

Utility:

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

Costs:

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

Collaborations:

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

Knowledge:

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

ScHARe

What is ScHARe?

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



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

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

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

ScHARe



nimhd.nih.gov/schare



ScHARe



Google Platform Terra Interface

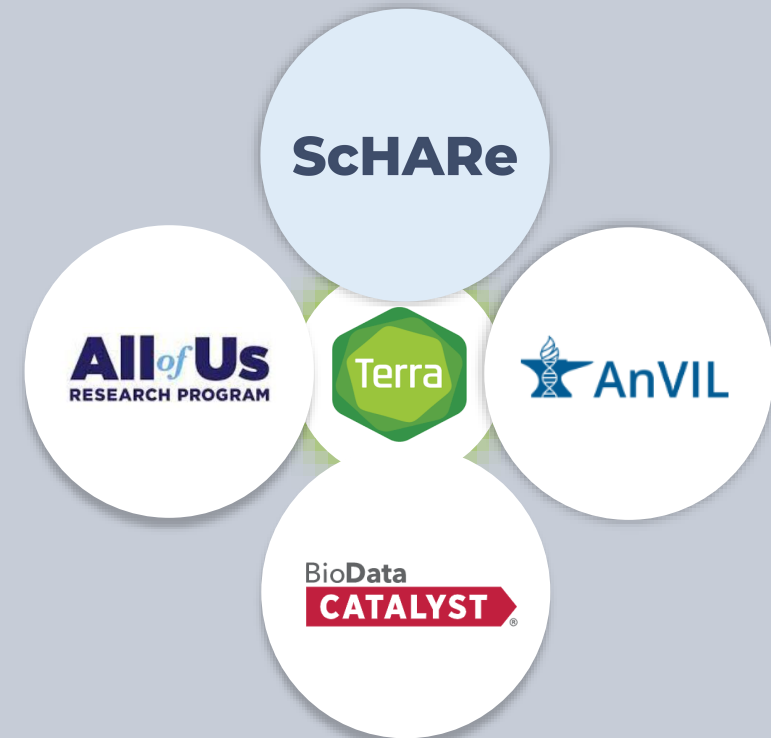
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Cut-and-paste code in Python and R



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

PREPARING FOR AI RESEARCH AND HEALTHCARE USING BIG DATA

Mapping across cloud platforms
with Terra Interface



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



ScHARe

All of Us
RESEARCH PROGRAM



BioData
CATALYST

This creates an extraordinary opportunity for **high-impact collaborations** across platforms

Learning how to use Terra on ScHARe will open up a world of possibilities, giving you access to an interdisciplinary wealth of datasets and resources

ScHARe Ecosystem structure

250+
FEDERATED
PUBLIC
DATASETS

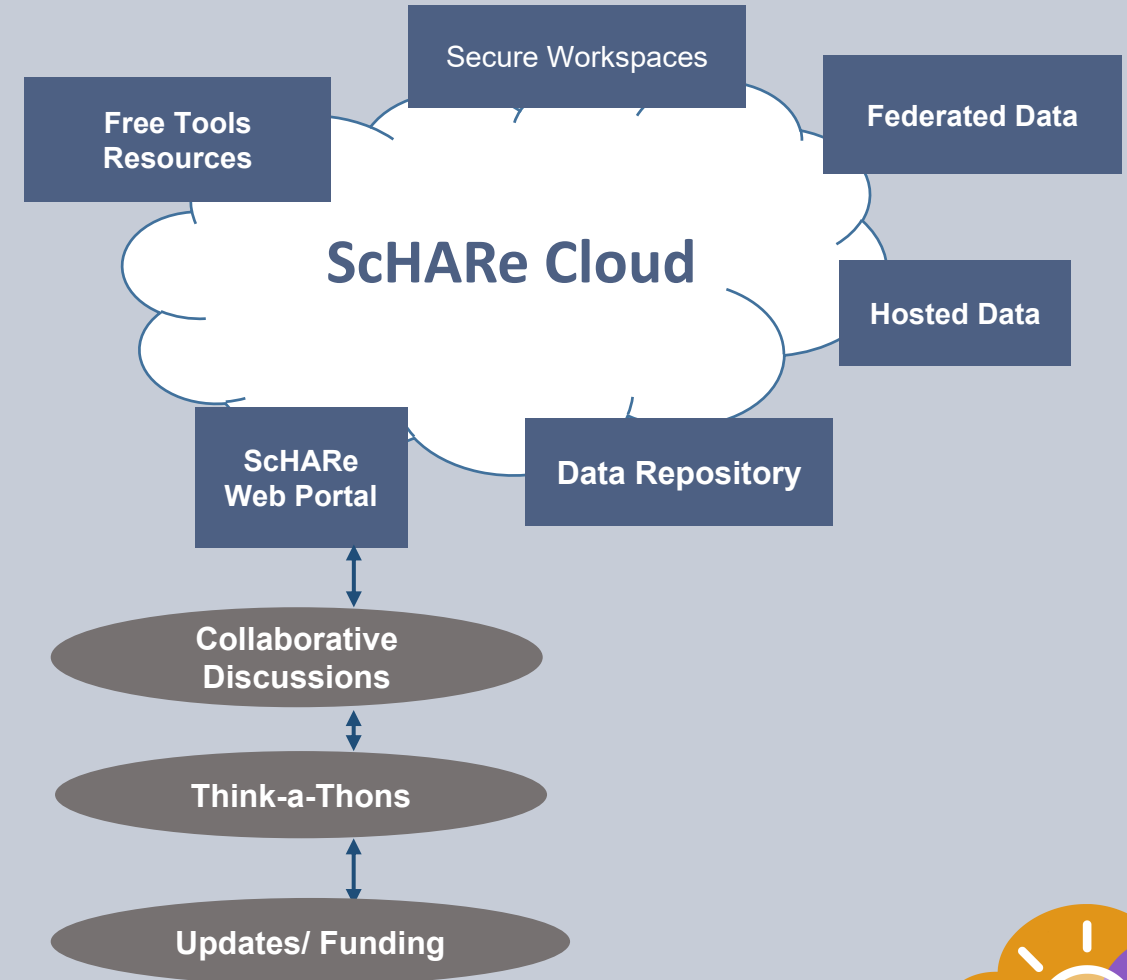
CDE
FOCUSED
REPOSITORY

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

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

COMPONENTS

Intramural and Extramural Resource



ScHARe Ecosystem

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

1

Public datasets

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

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

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

2

Funded datasets

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

e.g.: *Jackson Heart Study (JHS)*
Extramural Grant Data
Intramural Project Data



ScHARe Ecosystem

OVER 250 DATA SETS CENTRALIZED

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

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

with the addition of:

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

The screenshot shows the Terra WORKSPACES Data interface. The top navigation bar includes 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The 'DATA' tab is active, displaying an 'IMPORT DATA' button and a search bar for tables. A list of tables is shown on the left, with 'EconomicStability (62)' highlighted. The main table on the right lists various datasets with their names and sizes in GB.

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



ScHARe Ecosystem: ScHARe hosted datasets

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

250+ datasets

- What are the Social Determinants of Health?

Social determinants of health (SDoH) are the **nonmedical factors that influence health outcomes**

They are the **conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life**



ScHARe Ecosystem: ScHARe hosted datasets

Examples of datasets for each category include:

Education access and quality

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

Examples:

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

ScHARe Ecosystem: ScHARe hosted datasets

Health care access and quality

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

Example:

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

ScHARe Ecosystem: ScHARe hosted datasets

Neighborhood and built environment

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

Examples:

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

ScHARe Ecosystem: ScHARe hosted datasets

Social and community context

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

Example:

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

ScHARe Ecosystem: ScHARe hosted datasets

Economic stability

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

Examples:

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

ScHARe Ecosystem: ScHARe hosted datasets

Health behaviors

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

Examples:

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

ScHARe Ecosystem: ScHARe hosted datasets

Diseases and conditions

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

Examples:

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

ScHARe Ecosystem: Google hosted datasets

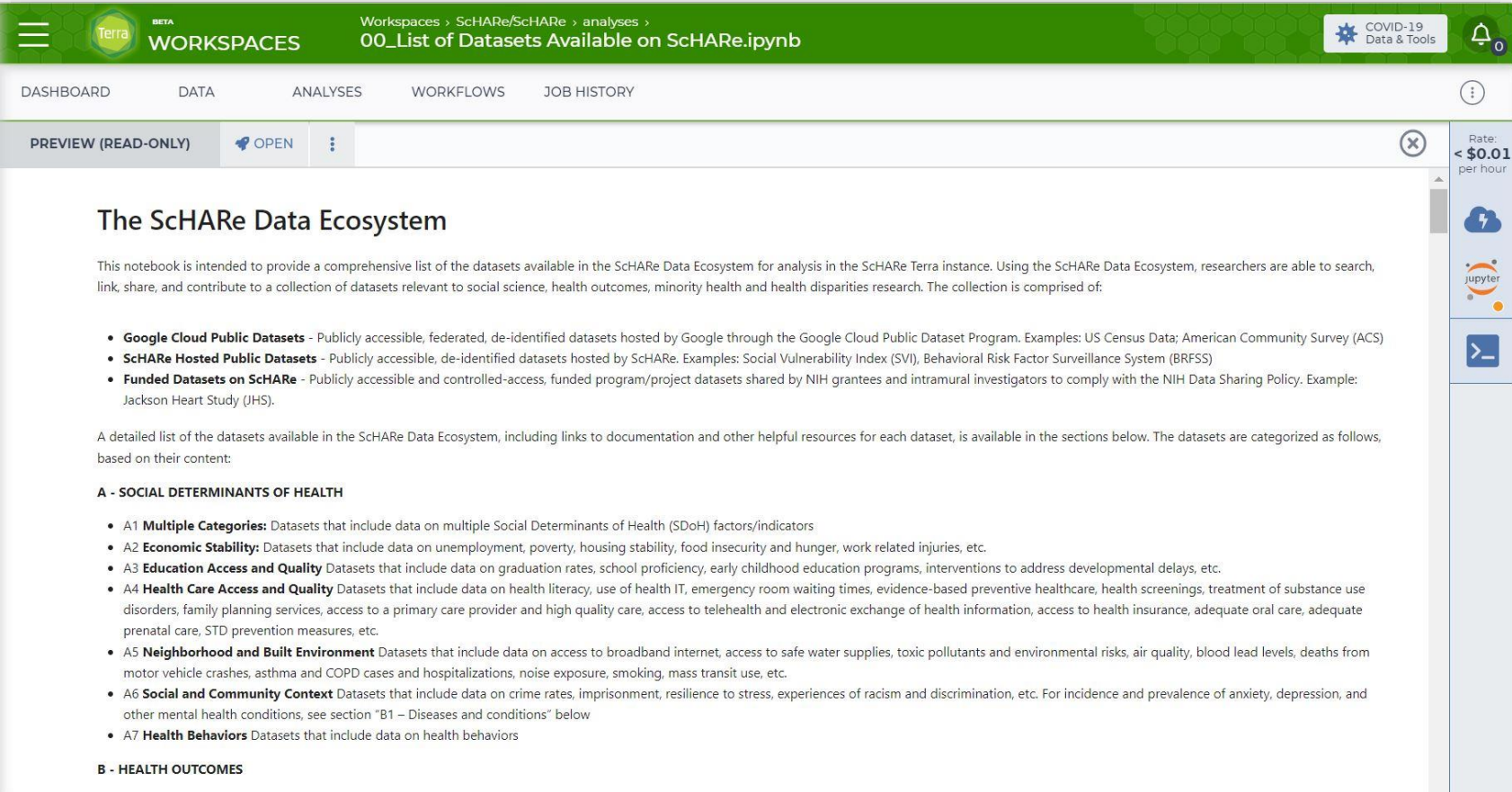
Examples of interesting datasets include:

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

How to check what data is available on ScHARe

1. Analyses tab

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



The screenshot displays the ScHARe workspace interface. At the top, there is a green header with the Terra logo and 'WORKSPACES' text. The breadcrumb navigation shows 'Workspaces > ScHARe/ScHARe > analyses > 00_List of Datasets Available on ScHARe.ipynb'. Below the header, a navigation bar includes 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The 'ANALYSES' tab is active, showing a 'PREVIEW (READ-ONLY)' view of a notebook. The notebook content is titled 'The ScHARe Data Ecosystem' and includes a description of the data ecosystem and a list of dataset categories. The right sidebar shows a 'Rate: < \$0.01 per hour' and various utility icons like a lightning bolt, Jupyter logo, and a right arrow.

WORKSPACES BETA
Workspaces > ScHARe/ScHARe > analyses > 00_List of Datasets Available on ScHARe.ipynb
COVID-19 Data & Tools

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

PREVIEW (READ-ONLY) OPEN

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



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 program/project requirements. Examples: Jackson Heart Study (JHS)

2. ScHARe Datasets PDF list



Scan me

bit.ly/ScHARe-datasets

How to access ScHARe hosted datasets

Data tab

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

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

The screenshot shows the Terra WORKSPACES interface. The top navigation bar includes 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The 'DATA' tab is active, showing an 'IMPORT DATA' button and a 'TABLES' section with a search bar and a list of dataset categories. The 'EconomicStability (62)' category is selected, displaying a list of datasets with columns for 'EconomicStability_id' and 'SizeGb'.

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

How to access Google hosted datasets

Big Query

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

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



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

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

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

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

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

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

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

ScHARe

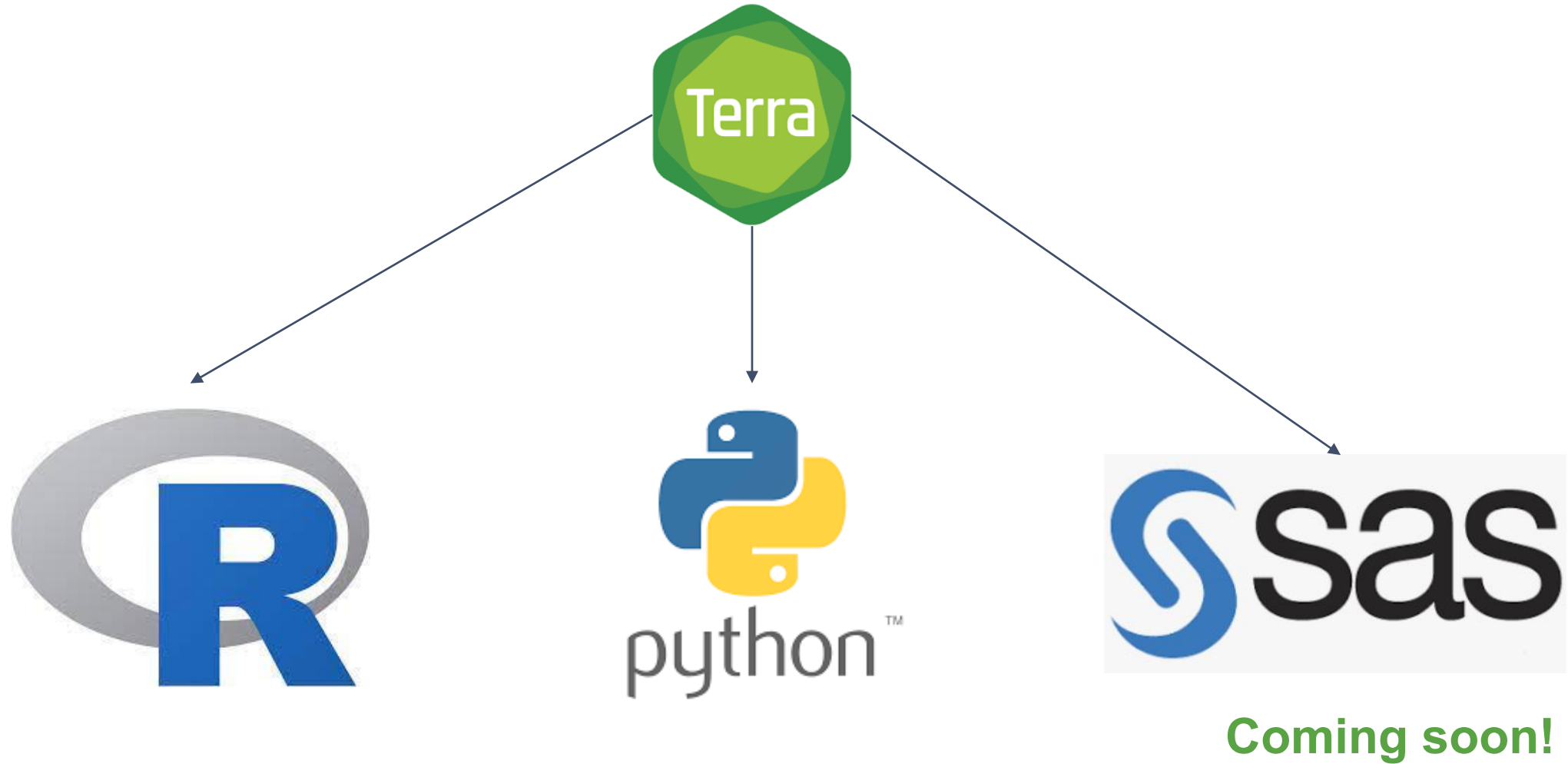
Python and R

The language
of Cloud Computing

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



What programming languages does Terra support?



What is R?

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

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

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

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



What is Python?

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

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



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

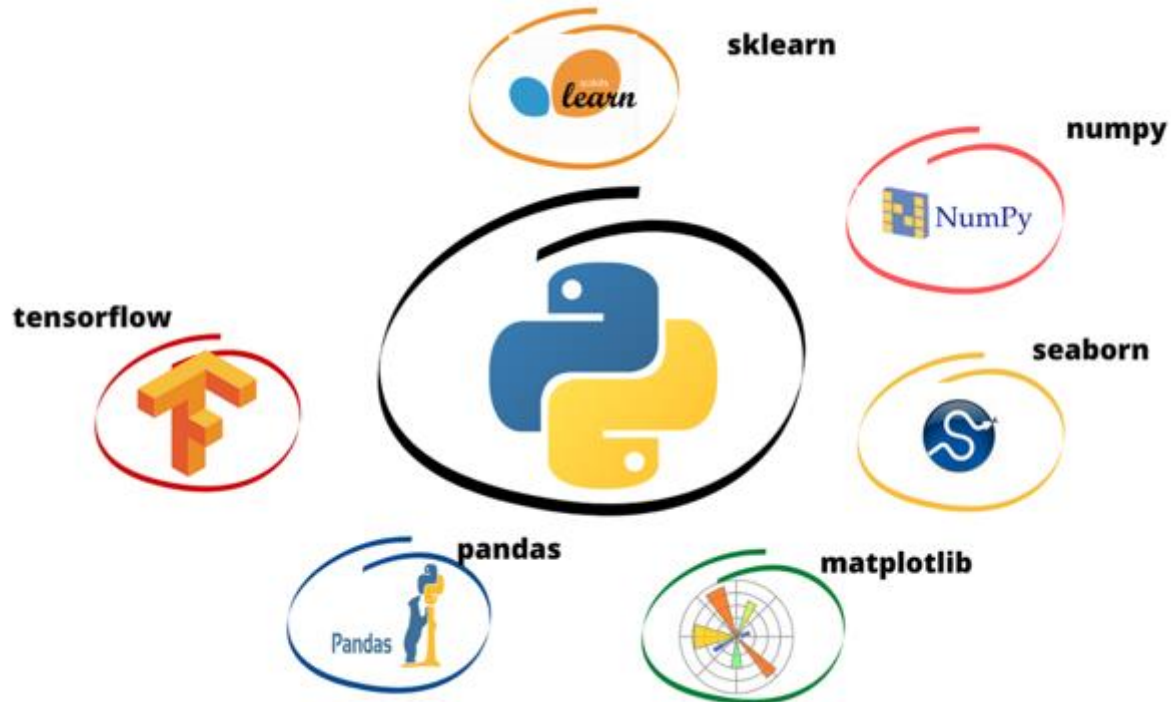
- **Easy to learn:** Python uses words and phrases that are close to everyday English, making it a good choice for beginners
- **Versatile:** You can use Python for many things
- **Free and open-source:** Anyone can use and improve Python for free: there's a large and helpful community to answer your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

Sources

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

Introduction to Python Data Science Libraries

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



What is a Python library?

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

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

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

Why Python?

According to [SlashData](#):

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

Source

stackify.com/learn-python-tutorials/

How to learn Python

How long does it take to learn Python?

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

Can you learn Python with no experience?

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

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

Python resources

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

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



Python resources

Stackify

30+ tutorials to learn Python

Top 30 Python Tutorials

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

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

Python resources

FreeCodeCamp

Code class for beginners

A screenshot of a webpage from freeCodeCamp. The page has a dark blue header with the freeCodeCamp logo and a tagline "Learn to code – free 3,000-hour curriculum". The main content area is white and features two sections. The first section is titled "Python Tutorial for Beginners (Learn Python in 5 Hours)" and includes a paragraph describing a course by TechWorld with Nana, covering topics like strings, variables, OOP, and functional programming, along with project examples. The second section is titled "Scientific Computing with Python" and includes a paragraph describing a certification course covering loops, lists, dictionaries, networking, and web services.

freeCodeCamp (🔥)

Learn to code – [free 3,000-hour curriculum](#)

Python Tutorial for Beginners (Learn Python in 5 Hours)

In [this TechWorld with Nana YouTube course](#), you will learn about strings, variables, OOP, functional programming and more. You will also build a couple of projects including a countdown app and a project focused on API requests to Gitlab.

Scientific Computing with Python

In [this freeCodeCamp certification course](#), you will learn about loops, lists, dictionaries, networking, web services and more.

Python resources

Harvard

Free Python course

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



Harvard University: CS50's Introduction to Computer Science

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



12 weeks

6–18 hours per week



Self-paced

Progress at your own speed

There is one session available:

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

Starts Jul 19

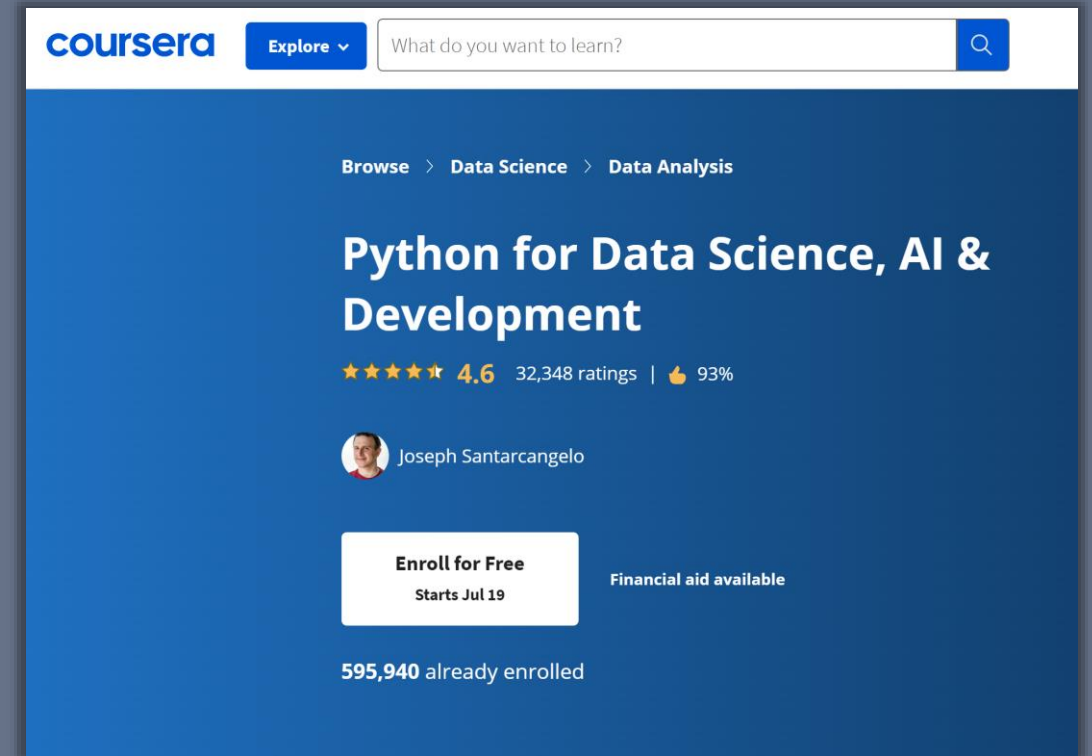
Ends Dec 31

Enroll

Python resources

Coursera

Free and paid Python courses



The screenshot displays the Coursera website interface. At the top, the Coursera logo is on the left, followed by an 'Explore' dropdown menu and a search bar containing the text 'What do you want to learn?'. Below the navigation bar, the breadcrumb trail reads 'Browse > Data Science > Data Analysis'. The main heading for the course is 'Python for Data Science, AI & Development'. Below the title, the course has a rating of 4.6 stars based on 32,348 ratings, with a 93% approval rate. The instructor's name, Joseph Santarcangelo, is listed next to his profile picture. A prominent white button says 'Enroll for Free' with 'Starts Jul 19' underneath. To the right of this button, it says 'Financial aid available'. At the bottom of the course card, it states '595,940 already enrolled'.

Python resources

LearnPython

Free interactive Python tutorials

Learn the Basics

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

Data Science Tutorials

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

Advanced Tutorials

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

Python resources

BestColleges

10 places to learn Python for free



Bootcamp Types ▾ Reviews ▾ Resources ▾ About ▾ BestColleges.com

Top 10 Free Python Courses

Google's Python Class

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

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

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

Microsoft's Introduction to Python Course

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

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

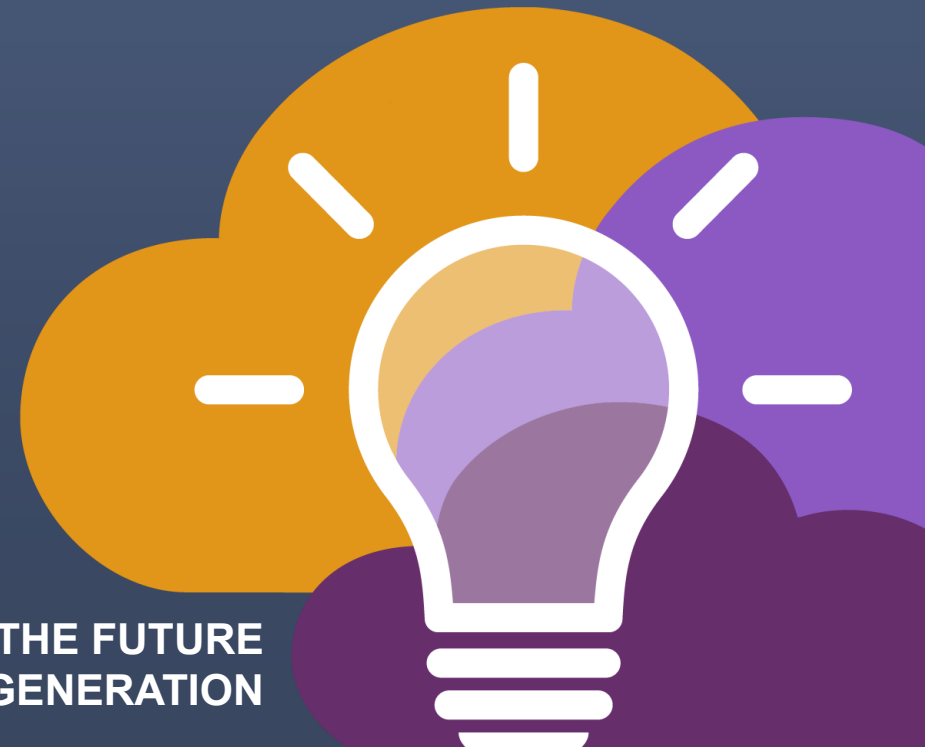
Students explore Python code, where to run Python apps, learn how to declare variables, and use the Python interpreter. They also learn how to access free resources.

SCHARE

Health Disparities,
Health Care Delivery,
Health Outcomes

The Language of Research

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION



Minority Health & Health Disparities Definitions and Populations

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

Health Disparities

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

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

Health Differences across Populations

Minority Health

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

Population:
OMB Racial/Ethnic Categories

Health Differences within Populations

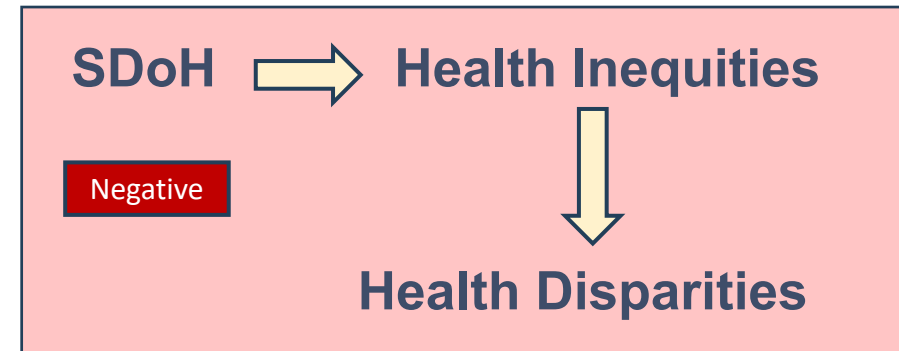
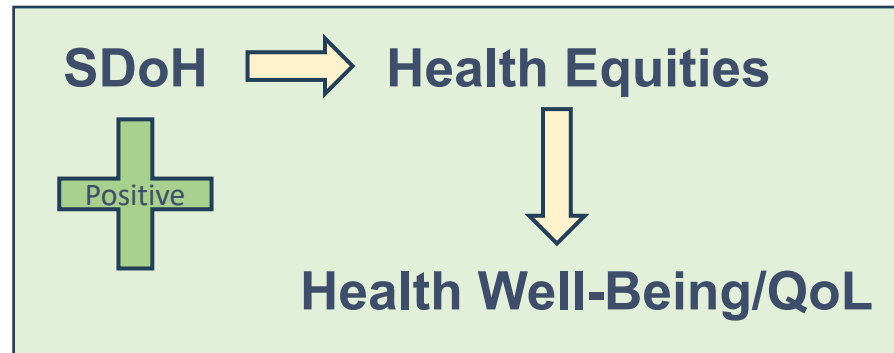
Health Disparity Outcomes

The health outcomes are categorized as:

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

SDoH: Neutral Measures Impacting Health Outcomes

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



Research Areas:

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

SDoH Impact Health Equity that Determines QoL

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

NIMHD 2024

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

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

Health Care Delivery



Advancing health equity into the future.

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

Health Outcome

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

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