NIH

February 15th, 2023 Think-a-Thon

Artificial Intelligence and Cloud Computing 101

Science collaborative for Health disparities and Artificial intelligence bias Reduction

SciHARe


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

Thank you

NIMHD

Dr. Eliseo Perez-Stable

ODSS

Dr. Susan Gregurick

NIH/OD

Dr. Larry Tabak

NIMHD OCPL

RLA

BioTeam

STRIDES

Terra

SIDEM

CCDE Working Group

Outline

5’	Introduction ▪ Experience Poll
10’	ScHARe and Think-a-Thons overview ▪ Interest Poll
20’	Artificial Intelligence and cloud computing concepts behind ScHARe
45’	Benefits and challenges in the AI/cloud computing world ▪ Whiteboards
5’	NIH clouds and resources for ScHARe collaborations
15’	Guest: Dr. Lakshmi Matukumalli (NIH/NIGMS): Advancing cloud computing and AI within IDeA states and MSIs
15’	Guest: Dr. Alison Lin (NIH/OD): ODSS training opportunities
5’	Cloud computing in grants

Experience poll

Please check your level of experience with the following:

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

SciAReS

Dr. Deborah Duran NIH/NIMHD

Dr. Luca Calzoni NIH/NIMHD

ScHARe

ScHARe is a cloud-based social science data platform designed to accelerate research in health disparities, health care outcomes, and artificial intelligence (AI) bias mitigation

The platform offers researchers at all career levels and disciplines:

- Access to social determinants of health and other social science **datasets**
- The ability to **collaborate** as they apply AI, machine learning, and other advanced analytical techniques to these datasets in a **secure setting**

nimhd.nih.gov/schare

ScHARe

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

- Foster research collaborations and increase participation of **women and underrepresented populations with health disparities** in data science
- Leverage **research opportunities** afforded by Big Data and cloud computing
- Advance **AI bias mitigation** strategies and **ethical inquiry** by increasing the use of diverse eyes and skills

ScHARe Mission

The ScHARe project goal is to advance health disparities, health care outcomes, and artificial intelligence bias mitigation research by:

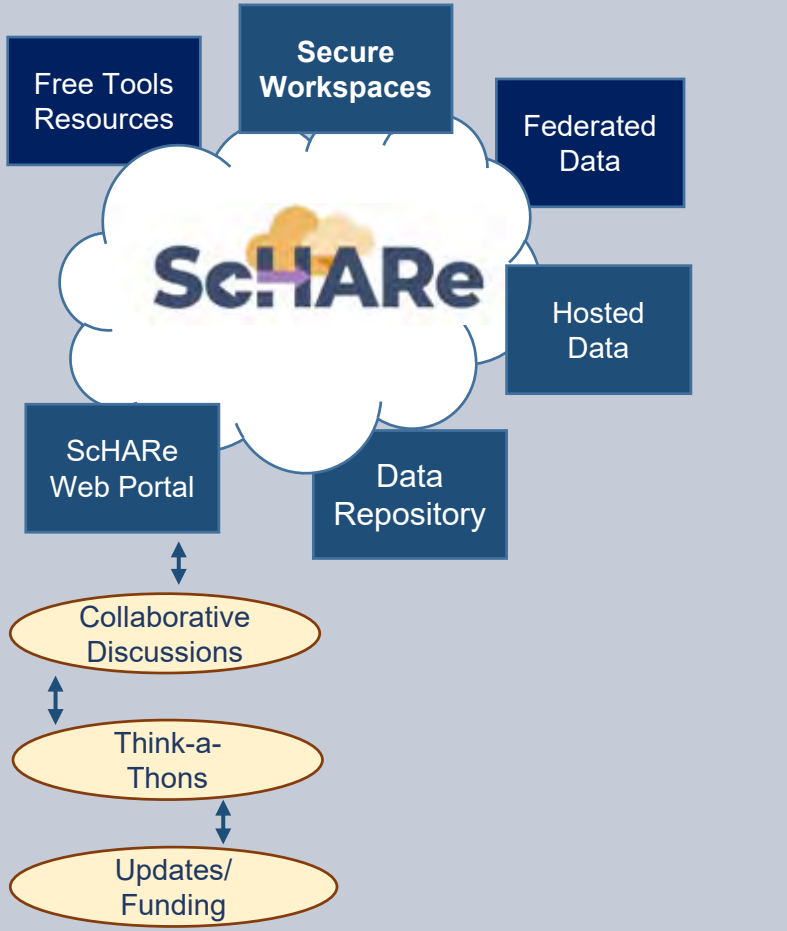
- Providing streamlined and centralized access to relevant **datasets** - including social determinants of health and other social science data
- Developing an **ecosystem** across data sets and across other NIH Terra Platforms
- Encouraging researchers to **leverage Big Data and advanced artificial intelligence analytic tools,** especially in population science and health care delivery
- Offering a **data analytics and research collaboration platform** for biomedical researchers and their collaborators to access the same data and run analyses together in secure online spaces
- Fostering the use of cutting-edge, low-cost data science resources among **researchers from underrepresented populations**, including diverse racial and ethnic groups and women, and from under-resourced Minority Serving Institutions (MSIs)
- Mitigating biases and creating a **culture of ethical inquiry** whenever artificial intelligence tools and algorithms are utilized

ScHARe Components

ScHARe co-localizes within the cloud:

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

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



nimhd.nih.gov/schare

ScHARe Data Ecosystem

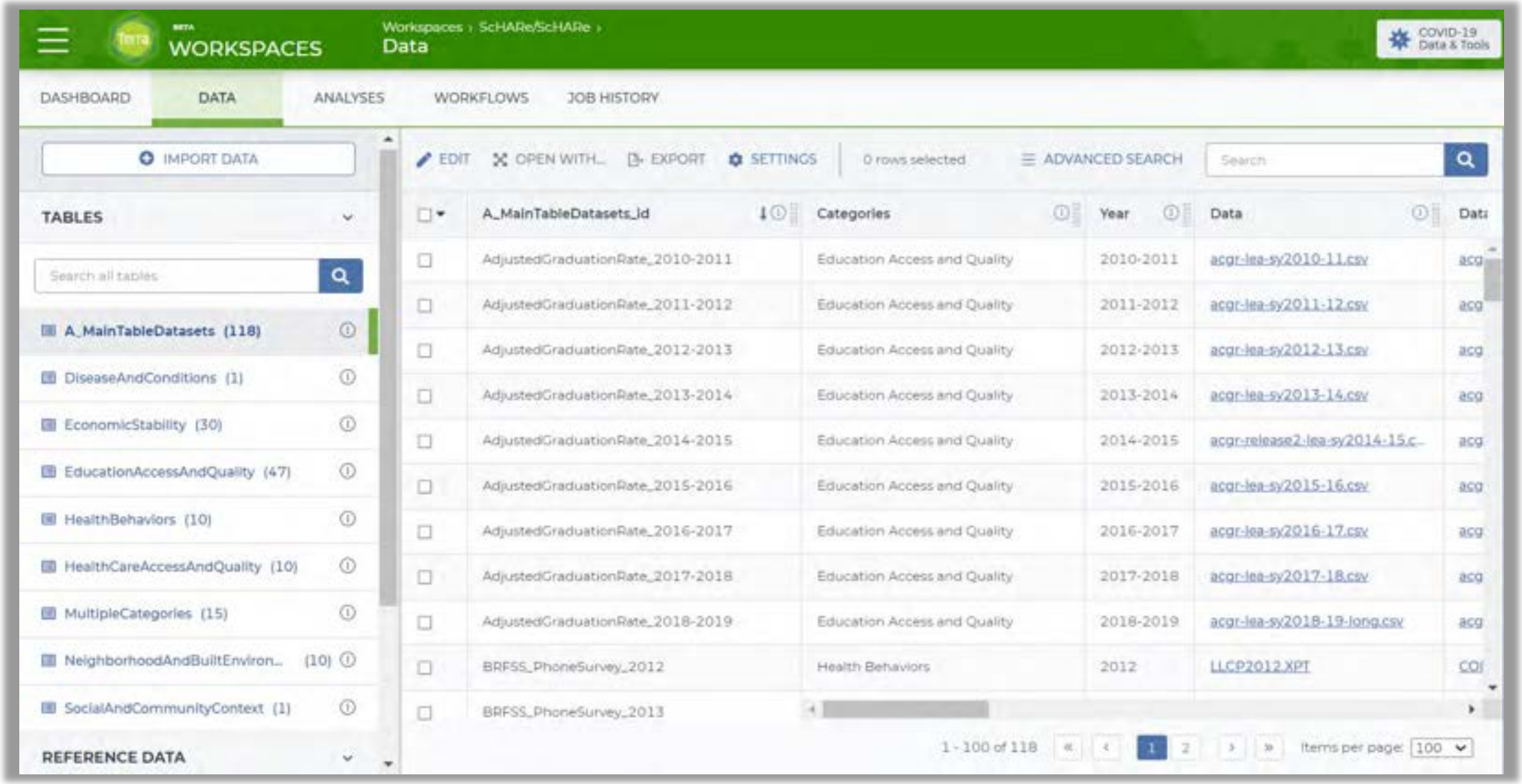
Researchers can access, link, analyze, and export **a wealth of datasets** within and across platforms relevant to research about health disparities, health care outcomes and bias mitigation, including:

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

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

- **Funded Datasets on ScHARe:** publicly accessible and controlled-access, funded program/project datasets 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*

Social science datasets



ScHARe Data Repository

CORE COMMON DATA ELEMENTS

NOVEL CDE FOCUSED REPOSITORY TO FOSTER INTEROPERABILITY

COMPLY WITH DATA SHARING POLICY HOST PROJECT DATA

DATA ECOSYSTEM

- Map across datasets
- Map across platforms

UPCOMING

Collaborations - Upskilling - Mentoring

ScHARe MEETS CHALLENGES OF CLOUD COMPUTING ADOPTION:

Utility:

- Numerous centralized social science & SDoH datasets
- Data sharing requirement compliance
- Secure confidential workspaces
- Workbooks with instructions & code
- Link across datasets and platforms
- SAS

Knowledge:

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

Costs:

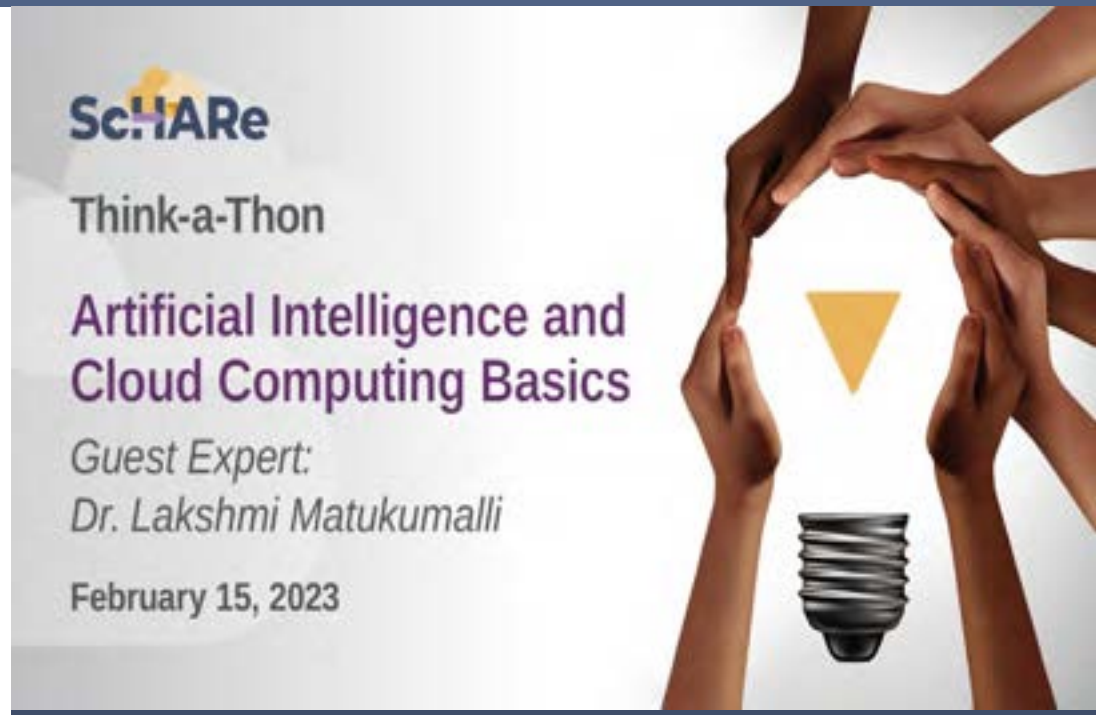
- Capitalizes free and low-cost tools
- Google credits
- Download data to personal computer when cloud unnecessary

Collaborations:

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

ScHARe Think-a-Thons

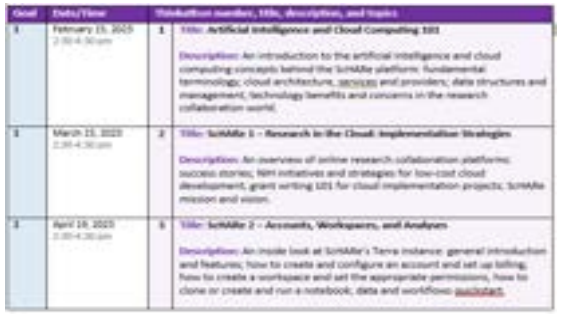
- Monthly sessions (2 hours)
- Instructional/interactive
- Designed for new and experienced users
- Research & analytic teams to:
  - Conduct health disparities, health outcomes, bias mitigation research
  - Analyze/create tools for bias mitigation
- Publications from team collaboration
- Networking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



Register:

bit.ly/think-a-thons

Instructional



Goal	Date/Time	Thinkathon number, title, description, and topics
1	February 15, 2023 2:30-4:30 pm	1 Title: Artificial Intelligence and Cloud Computing 101 Description: An introduction to the artificial intelligence and cloud computing concepts behind the ScHARe platform: fundamental terminology; cloud architecture, services and providers; data structures and management; technology benefits and concerns in the research collaboration world.
1	March 15, 2023 2:30-4:30 pm	2 Title: ScHARe 1 – Research in the Cloud: Implementation Strategies Description: An overview of online research collaboration platforms; success stories; NIH initiatives and strategies for low-cost cloud development; grant writing 101 for cloud implementation projects; ScHARe mission and vision.
1	April 19, 2023 2:30-4:30 pm	3 Title: ScHARe 2 – Accounts, Workspaces, and Analyses Description: An inside look at ScHARe's Terra instance: general introduction and features; how to create and configure an account and set up billing; how to create a workspace and set the appropriate permissions; how to clone or create and run a notebook; data and workflows available

Research teams

2	June 21, 2023 2:30-4:30 pm	5	Title: Data Science Projects 1 – Health Disparities and Individual SDoH Description: Exploring the impact of individual Social Determinants of Health on health outcomes: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.
2	July 19, 2023 2:30-4:30 pm	6	Title: Data Science Projects 2 – Health Disparities and Structural SDoH Description: Assessing the impact of structural Social Determinants of Health on health outcomes: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.
2	August 16, 2023 2:30-4:30 pm	7	Title: Data Science Projects 3 – Health Outcomes Description: Investigating the influence of non-clinical factors on disparities in health care delivery: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.

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

Artificial Intelligence and cloud computing concepts behind

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** to see how patterns in one domain affect other areas.

Data can be **structured** into fixed fields **or unstructured** as free-flowing information.

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

- **F**indability
- **A**ccessibility
- **I**nteroperability
- **R**eusability

ScHARe

The ScHARe Data Ecosystem will offer access to **300+ 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

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 information and process data remotely.

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

Main public cloud service providers:

- **Google**
- Azure
- AWS

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.**

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.

Many AI projects are built using Python.

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

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

Machine learning

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.

Tutorials show you how to scale and manage **model training and serving**, from a Terra notebook or leveraging GCP Vertex AI.

Novice users will find resources to help them learn more about ML applications.

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.

ScHARe

Critical thinking, an ethical inquiry approach to AI, and diverse perspectives are needed to mitigate biases.

Collaboratively **developing bias mitigation tools and strategies** is one of ScHARe's goals.

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

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.

Benefits and challenges in the Artificial Intelligence/cloud computing world

AI and cloud computing

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

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

We want to hear from you

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

Benefits

Access to big datasets and large data ecosystems:

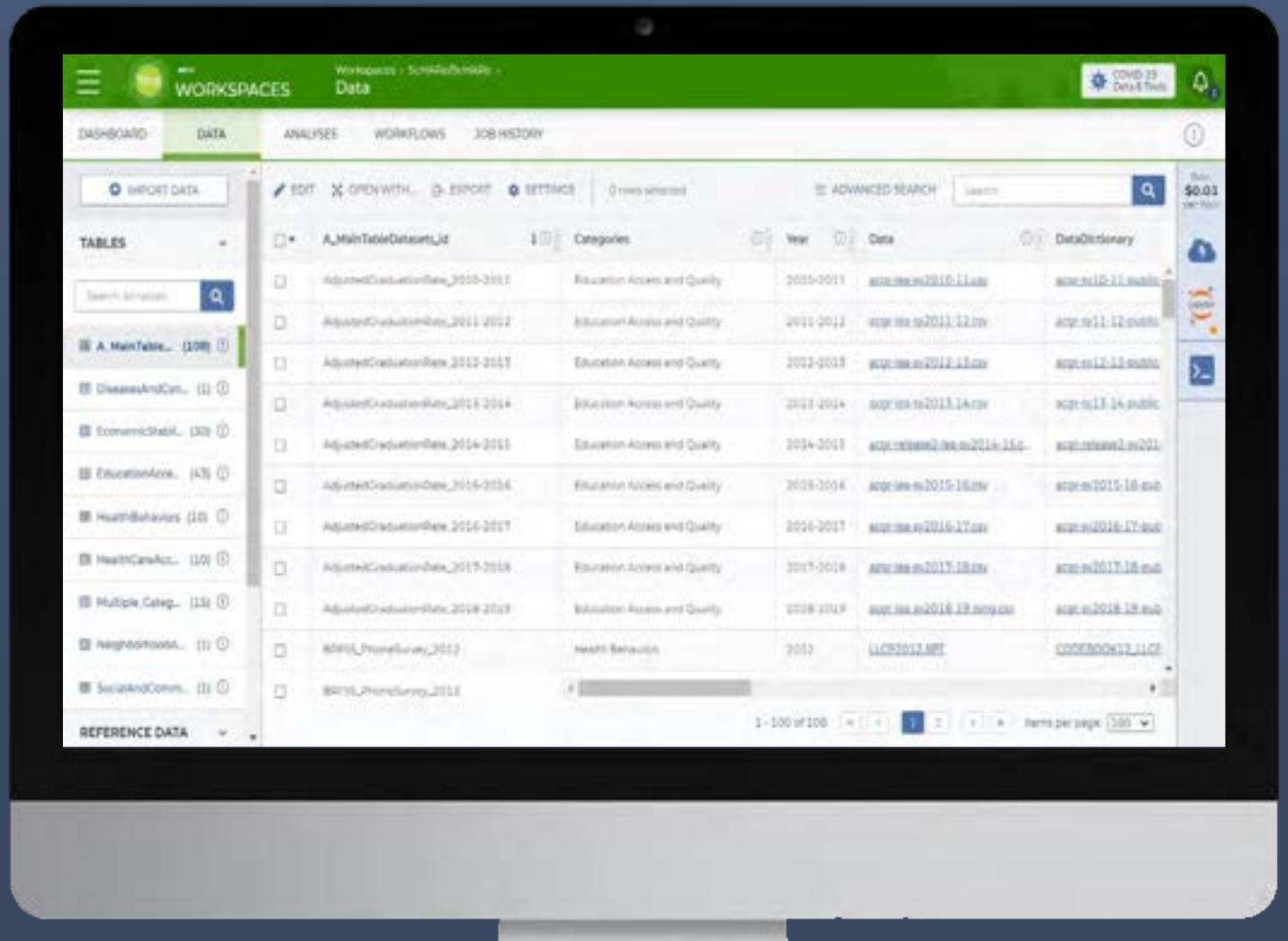
- Today, the scientific community confronts a data landscape that more expansive and more varied. The cloud offers access to **vast repositories** of scientific data, and enables **efficient mapping and linking** across data sources

ScHARe

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

- Google Cloud Public Datasets
- ScHARe Hosted Public Datasets
- Funded Datasets on ScHARe, in compliance with the **NIH Data Sharing Policy**

Benefits



ScHARe

On ScHARe, datasets are categorized by content based on the CDC **Social Determinants of Health categories**:

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

with the addition of:

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

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

Benefits

Deeper insights and better decision making:

- AI in the cloud can help in managing the massive volumes of data available, identifying trends in **large datasets** with **quicker and more accurate results**
- By helping to compare patterns in historical data with current patterns, AI can **facilitate decision-making** in clinical and policy applications

Benefits

Deeper insights and better decision making:

In particular, the Smart Cloud enabled by the relationship between AI and the cloud, linked with **machine learning** (ML) and **data mining** resources is a true revolution in research in terms of:

- productivity
- efficiency
- insights gainable leveraging Big Data

SciHARe

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

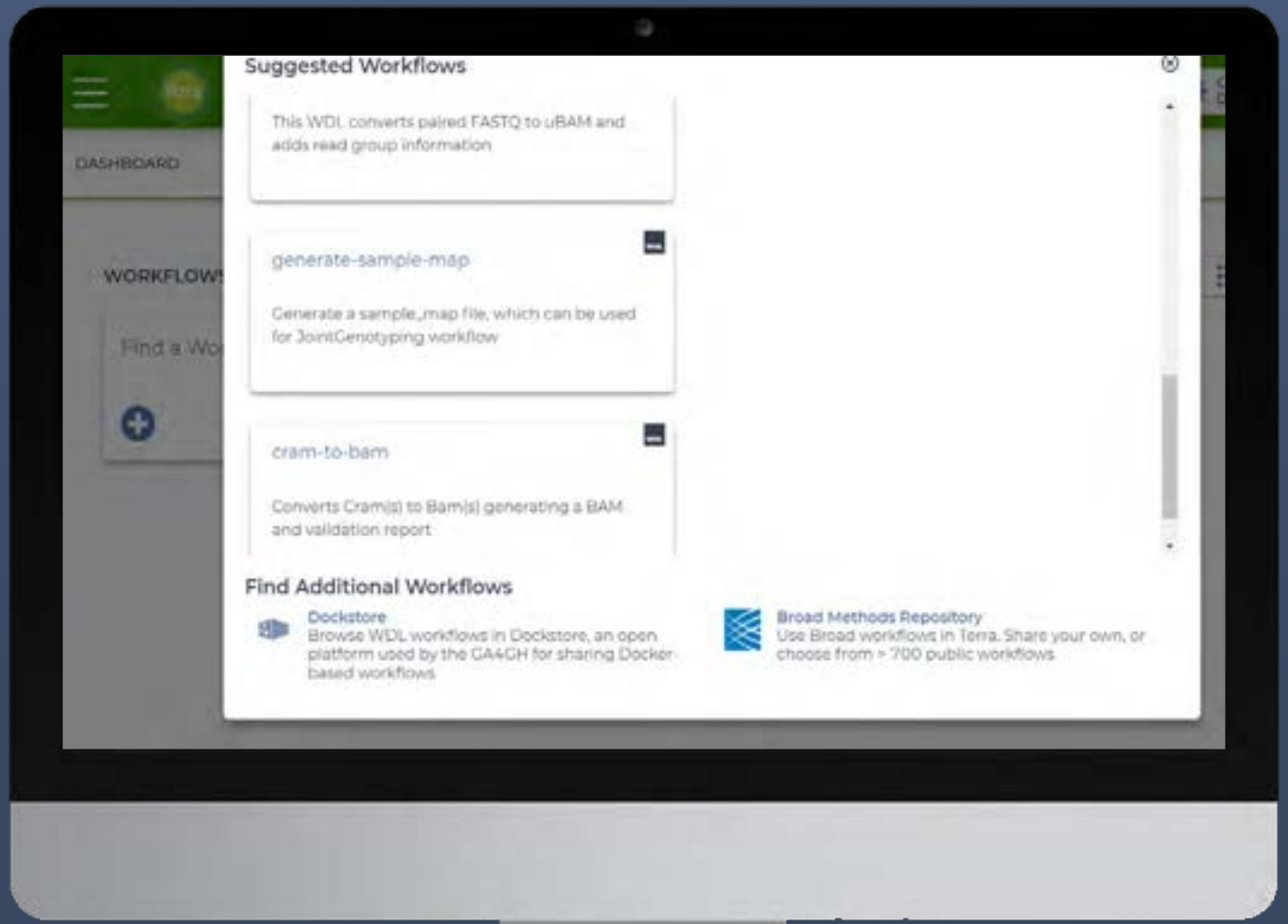
ScHARe

Workflows (pipelines) are a series of 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



ScHARe

Workflows (pipelines) are a series of steps performed by a compute engine - often used for automated, bulk analysis.

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

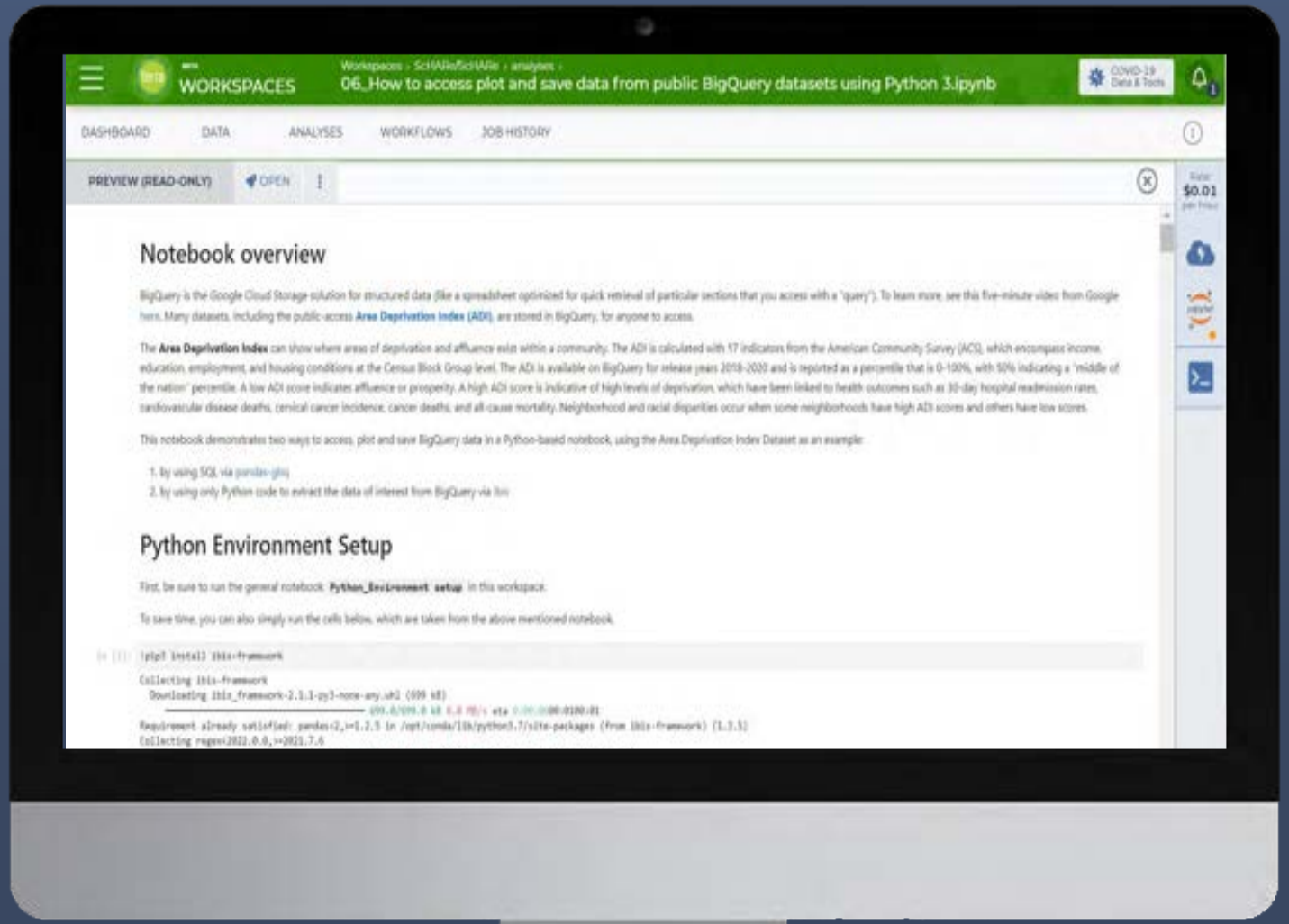
Benefits

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

Benefits



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 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** and critical thinking whenever AI is utilized

We want to hear from you

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

4 whiteboards:

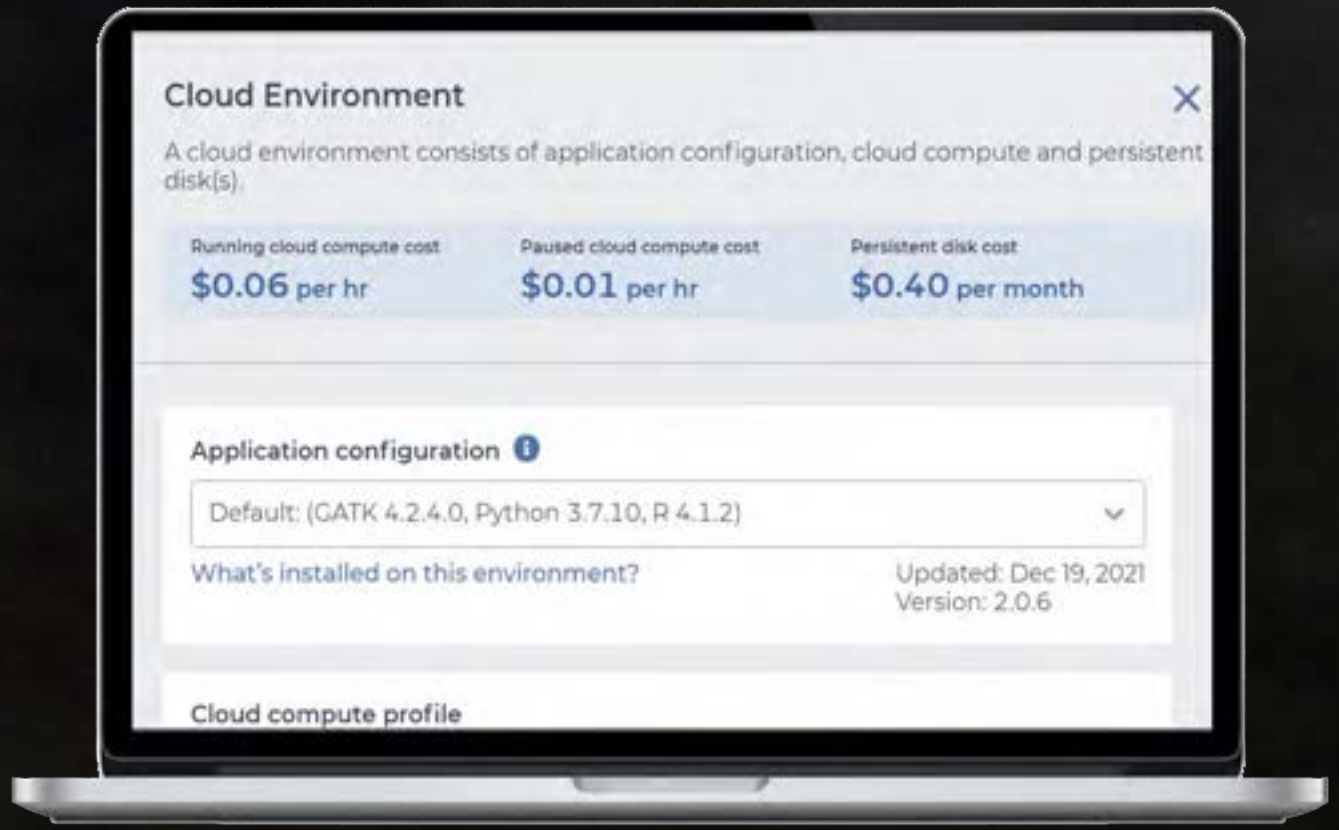
COST

KNOWLEDGE

RESEARCH APPLICATIONS

OTHER

~~COST~~ PROBLEMS SOLUTIONS



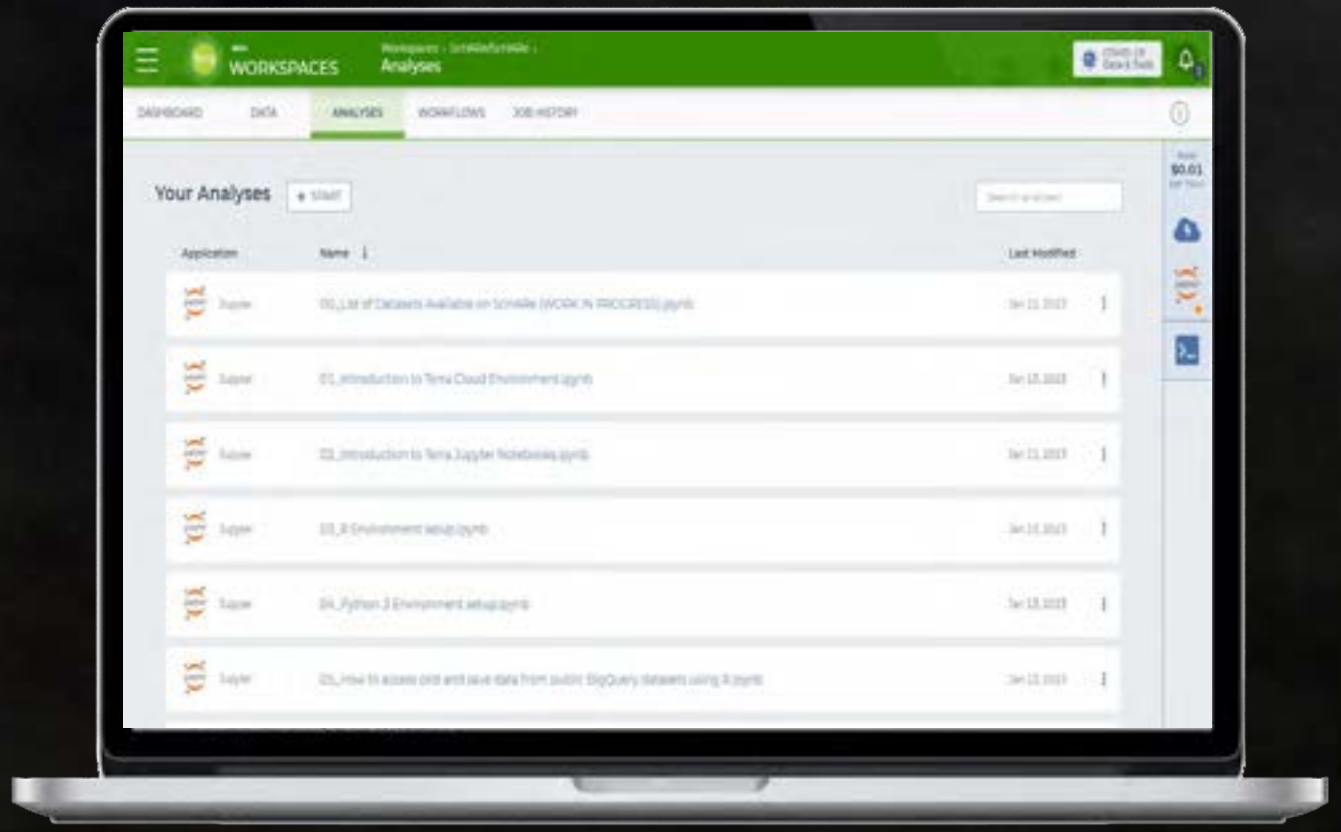
ScHARe leverages **low-cost and open-source** components:

- Terra Platform
- GitHub
- Open-source tools/libraries

to keep platform costs at a minimum and pave the way for the adoption of AI/cloud computing by Minority Serving Institutions

~~KNOWLEDGE~~ ~~PROBLEMS~~

SOLUTIONS



ScHARe provides:

- Step-by-step instructional notebooks for novice users
- An updated list of free tutorials and e-learning resources on its web portal
- Think-a-Thons to share knowledge and form collaborations
- Training opportunities through the NIH STRIDES initiative

app.terra.bio/#workspaces/ScHARe/ScHARe/analyses

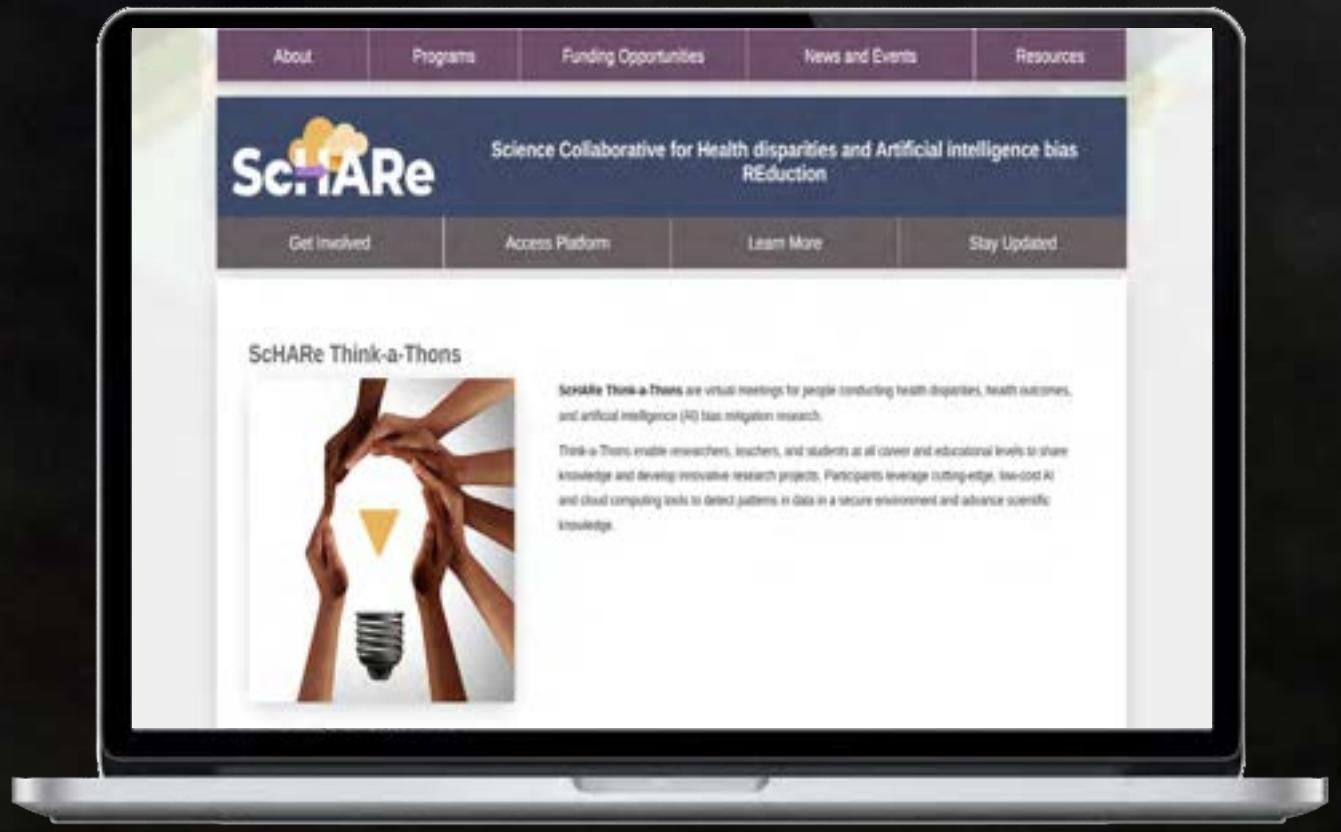
~~KNOWLEDGE~~ ~~PROBLEMS~~ SOLUTIONS

ScHARe provides:

- Step-by-step instructional notebooks for novice users
- An updated list of free tutorials and e-learning resources on its web portal
- Think-a-Thons to share knowledge and form collaborations
- Training opportunities through the NIH STRIDES initiative

nimhd.nih.gov/resources/schare/resources-and-tutorials.html

~~KNOWLEDGE~~ ~~PROBLEMS~~ SOLUTIONS



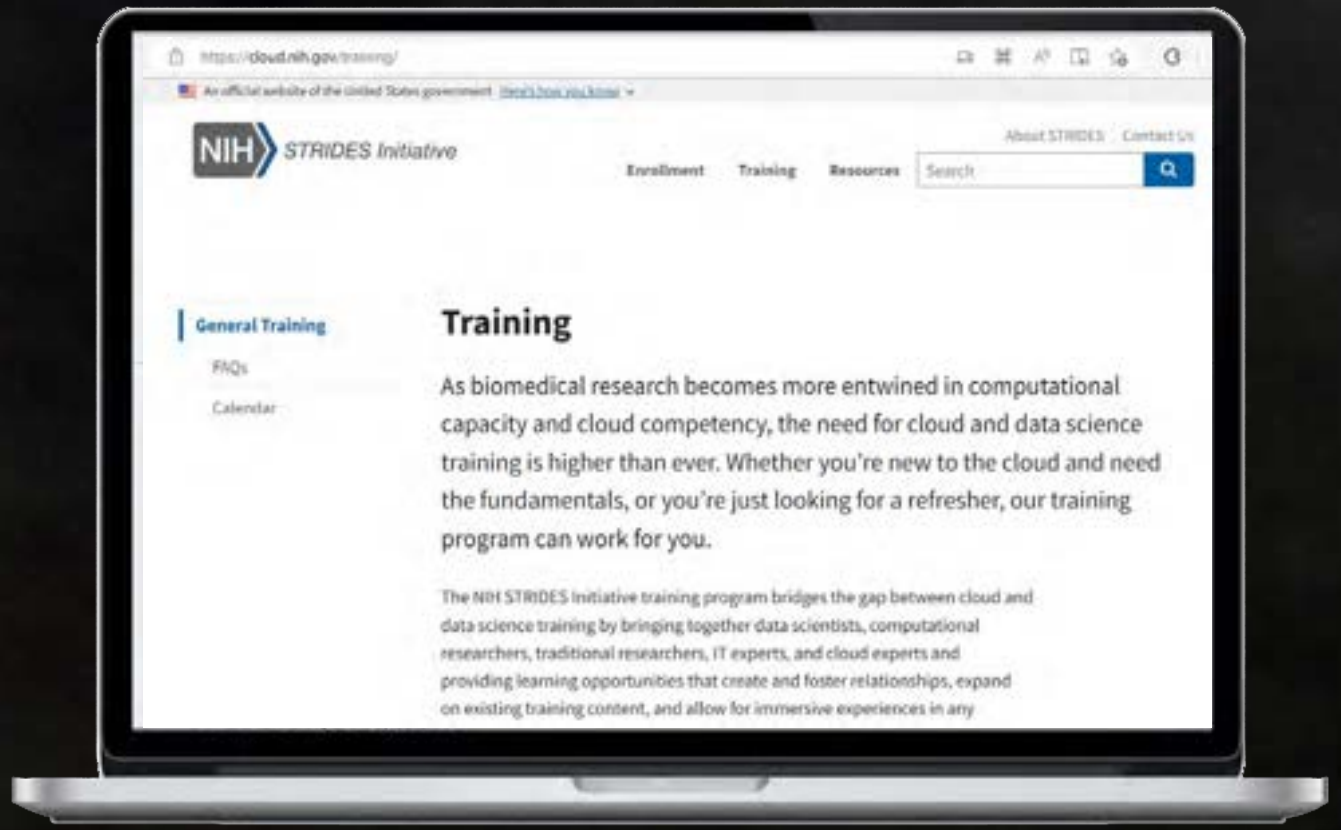
ScHARe provides:

- Step-by-step instructional notebooks for novice users
- An updated list of free tutorials and e-learning resources on its web portal
- Think-a-Thons to share knowledge and form collaborations
- Training opportunities through the NIH STRIDES initiative

bit.ly/think-a-thons

~~KNOWLEDGE~~ ~~PROBLEMS~~

SOLUTIONS



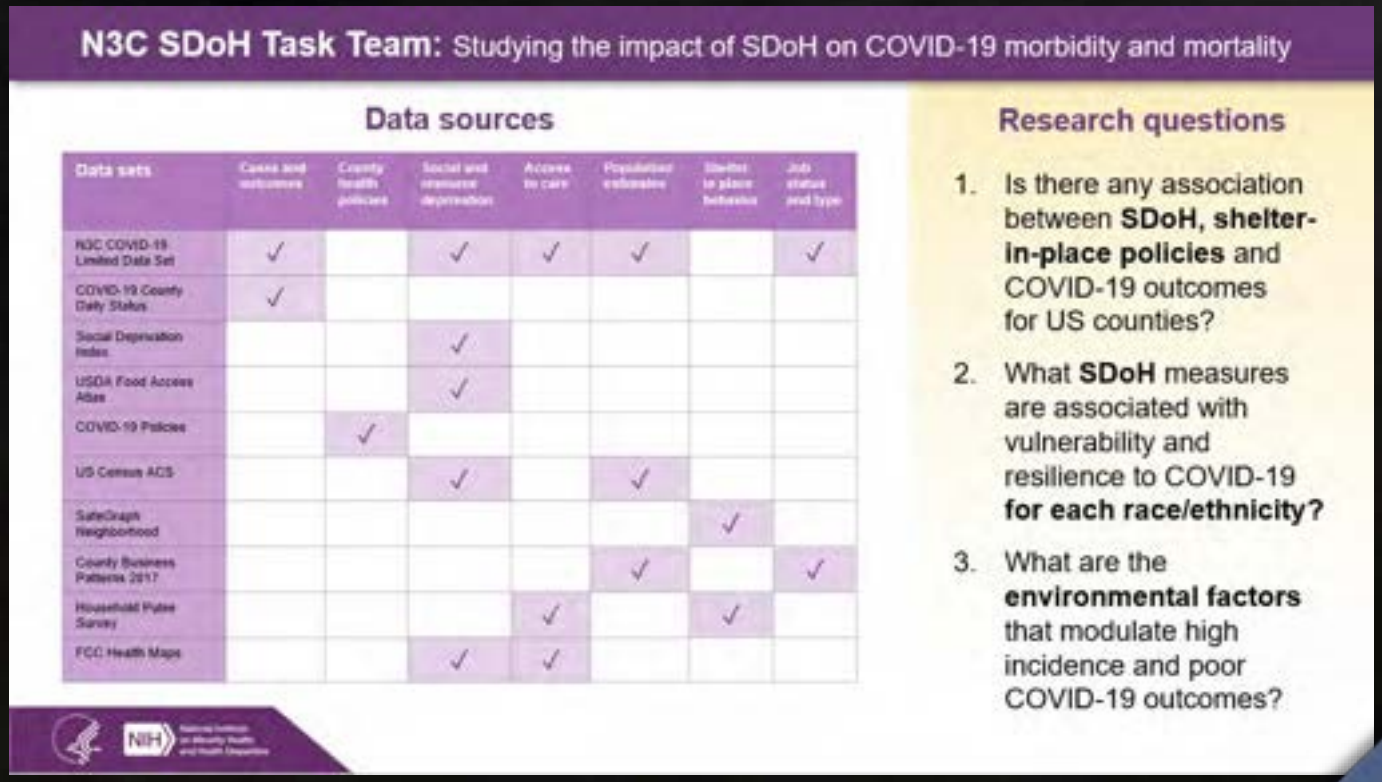
ScHARe provides:

- Step-by-step instructional notebooks for novice users
- An updated list of free tutorials and e-learning resources on its web portal
- Think-a-Thons to share knowledge and form collaborations
- Training opportunities through the NIH STRIDES initiative

cloud.nih.gov/training/

~~APPLICATIONS~~ ~~PROBLEMS~~ SOLUTIONS

An example of research collaboration using the cloud: N3C



ScHARe follows a model similar to **N3C** *(National Covid Cohort Collaborative)*

N3C is a secure cloud portal hosted by NCATS (*National Center for Advancing Translational Sciences*)

N3C Domain Teams of researchers with shared interests collaborate in groups to analyze data in the cloud

APPLICATIONS

~~PROBLEMS~~

SOLUTIONS

Project examples

Exploring the impact of:

1. individual Social Determinants of Health (SDoH) on health outcomes
2. structural SDoH on health outcomes
3. non-clinical factors on disparities in health care delivery

ScHARe promotes **Think-a-Thons** for researchers and students at all levels interested in collaborating to develop innovative research questions and projects leading to publications

NIH clouds and resources for

collaborations

NIH Initiatives

NIH has launched a series of initiatives to:

- harness the power of cloud computing
- provide NIH biomedical researchers access to the most advanced, cost-effective computational infrastructure, tools and services

Examples include:

- **STRIDES** (Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability):
  - NIH partnered with commercial providers to streamline NIH data use leveraging cloud environments
  - Benefits include:
    - Professional services
    - Training
    - Discounts on STRIDES partner services
    - Potential collaborative engagements

NIH STRIDES
Accelerating biomedical research

cloud.nih.gov

Examples include:

- **AIM-AHEAD** (Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity):
  - Establish partnerships to increase the participation of underrepresented researchers in the development of AI/ML models using electronic health record (EHR) data

aim-ahead.net

- **BRIDGE2AI** (Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity):
  - Expand the use of AI in biomedical and behavioral research by generating “flagship” data sets and best practices for ML analysis

bridge2ai.org

Examples include:

- **All of Us**:
  - A historic effort to gather data from 1+ million people in the U.S. to build one of the most diverse health databases in history

allofus.nih.gov

- **AnVIL** (NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space):
  - Unified cloud environment for the analysis of genomic datasets

anvilproject.org

- **BioData Catalyst**:
  - Cloud-based platform for tools, applications, and workflows

biodatacatalyst.nhlbi.nih.gov

Examples include:

- **All of Us**:
  - A historic effort to gather data from 1+ million people in the U.S. to build one of the most diverse health databases in history
- **AnVIL** (NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space):
  - Unified cloud environment for the analysis of genomic datasets
- **BioData Catalyst**:
  - Cloud-based platform for tools, applications, and workflows

Terra powers all four cloud platforms

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

SciARe

Guest expert

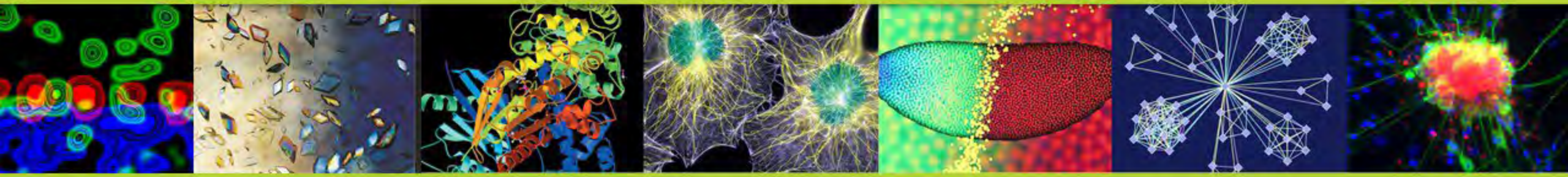
Dr. Lakshmi Matukumalli

NIH/NIGMS

National Institute of General Medical Sciences

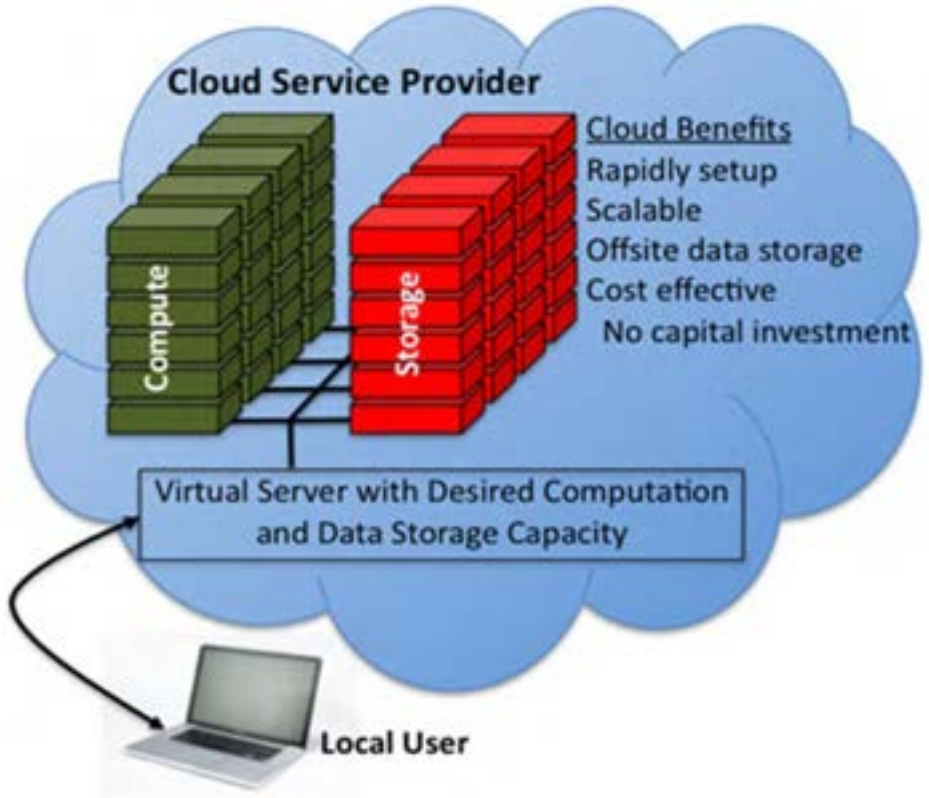
NIGMS Sandbox Cloud Modules

Lakshmi K Matukumalli



NIH Workshop on Broadening Cloud Computing Usage in Biomedical Research (Sept 13 -14, 2021)

- Most RCMIs, Minority serving institutions and network institutions in IDeA states do not have access to high performance computing.



https://datascience.nih.gov/data-ecosystem/nih-workshop-on-broadening-cloud-computing-usage-in-biomedical-research

Challenges

- Need for better understanding of cloud computing (computing costs, storage, data and software resources, IT support, governance, and data security)
- Need for Cloud Computing training, AI/ML, and Data Science

Opportunities

- Scalable computing power and storage, big data, innovation (novel research topic areas and new computing methods - AI/ML)
- Cloud enhances collaborations, data reuse, software sharing, and research reproducibility

Interactive Training on Cloud Fundamentals

48 classes (through 60 six-hour sessions) by Google (27) or AWS (21)
June–December 2021

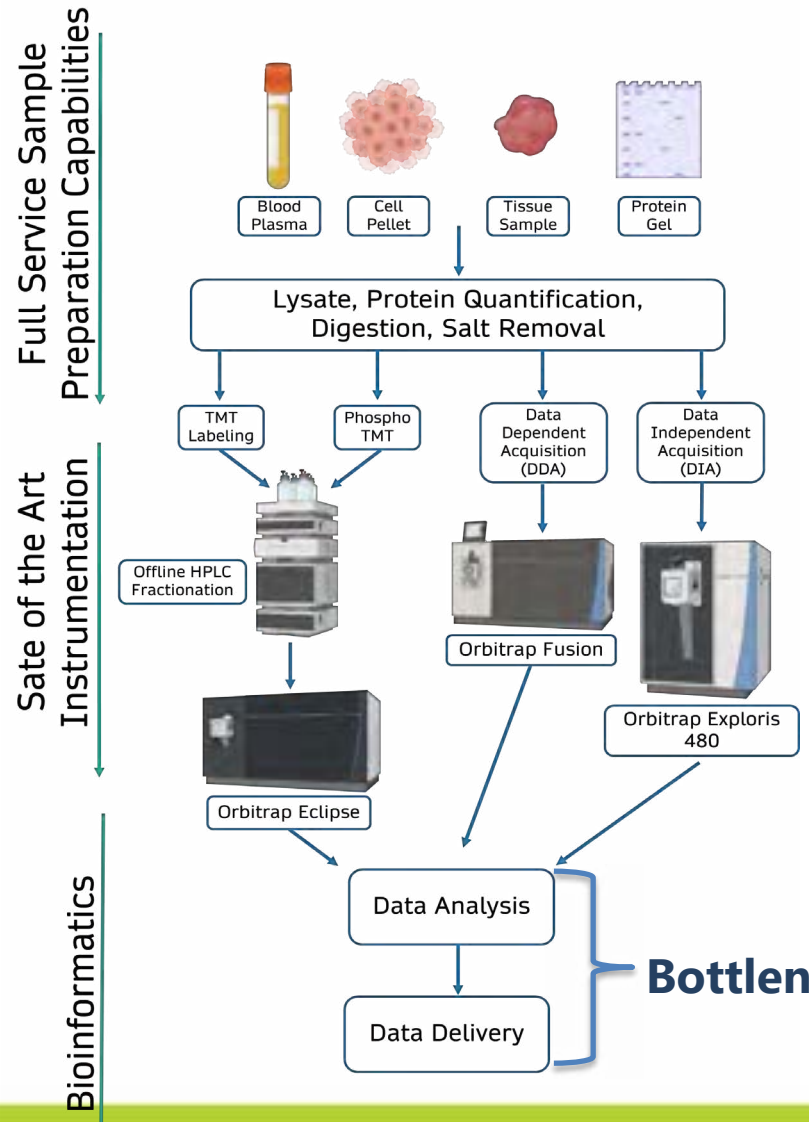
Google Cloud

- Beginner (Cloud Basics) 24 classes; 24 sessions
- Intermediate (App Development) 12 classes; 16 sessions
- Advanced (Architecture) 12 classes; 20 sessions

15 students/class; 720 class seats;
500 participants from ~ 90 Institutions

GM staff enrolling participants;
CIT-STRIDES staff managing logistics;
GCP and AWS providing instruction/training

Bottleneck Halted Growth at IDeA National Resource for Quantitative Proteomics



- Time needed for data processing, delivery, and consultation
- Staff effort/time for data processing, delivery, and consultation
- Cost of computational infrastructure and data storage

Transformative Impact of the Lift-and-Shift Project

- **Goal**: Work with Google and NIH STRIDES to lift-and-shift proteomics pipeline to the cloud for automated data analysis, delivery, consulting, and data storage
- **Outcomes**
    - Led to the establishment of UAMS' own GCP account
    - System launched January 1, 2022
    - Project delivery through signURL link using GCP workflow since launch
    - Multi-plex 5 data searches at a time (previously one at a time)
    - Automated data analysis time 6-times faster
    - Leveraging 5TB of cloud data

"This cloud environment has been absolutely transformative for our resource. I couldn't be more pleased"

Alan Tackett, PhD, PI

NIH Cloud Platforms

NIH National Heart, Lung, and Blood Institute | BioData CATALYST

SevenBridges

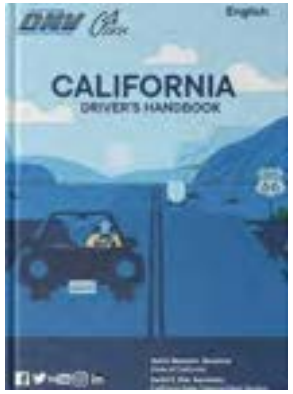
NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

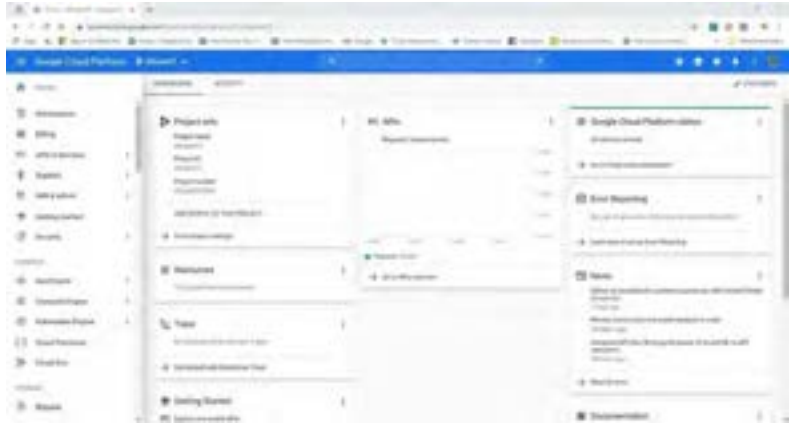
AnVIL

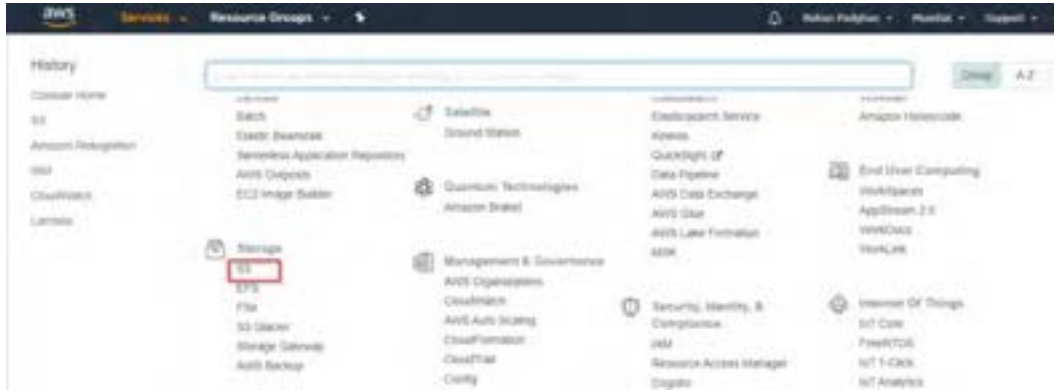
Terra

SciHARe

Cloud Training



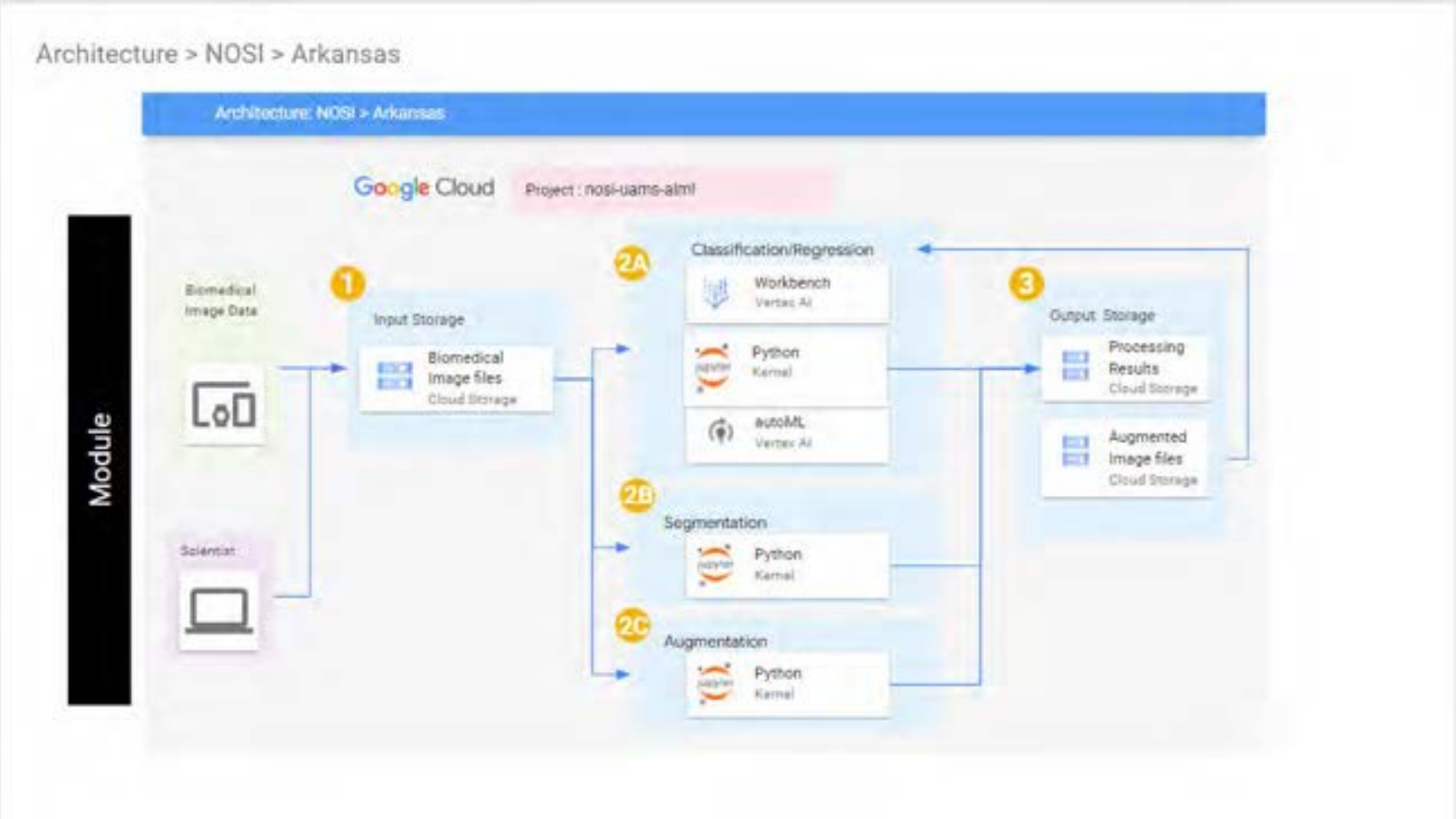




NIGMS Sandbox Modules

- Jan – Dec 2021: RNA-Seq and Proteomics Modules – Pilot
- Jan – May 2022: NOSI proposal review and awards
- June 2022 – April 2023: Researcher Training and Module Building
- Researcher Training:
  - Cloud Computing
  - Developing workflows
  - Jupyter Notebooks and Source repositories
  - Quizzes and Visualizations

Module Building – Cloud-based Workflows



Functional Module Development completed in Dec 2022

Cloud Learning Module – Final Product

- Introductory video with an interesting biological story
- Intuitive Readme instructions;
- Interactive quizzes, graphs, visualizations
- Workflows, Jupyter notebooks Test and training Datasets
- Practicum – test dataset and solutions

NIGMS Sandbox

Metagenomics Analysis of Biofilm-Microbiome

Proteomics Data Analysis

Consensus Pathway Analysis

ATAC-seq and Single Cell ATAC-seq Analysis

Data Science For Biology

Fundamentals of Bioinformatics in Terminal Environment

Biomedical Imaging Analysis Using AI/ML

DNA Methylation Sequencing Data Analysis

Biomarker Discovery

Transcriptome Assembly, Refinement, and Analysis

Integrating Multi-omics Datasets

RNAseq Analysis

NIH CLOUD LAB - Overview

NIH Cloud Lab provides a cloud workspace for research, training, and developing new applications **(for limited time/budget to authorized users)**

Benefits

Reduce barriers to entry by providing researchers an easy route to access cloud

Facilitate technical development through prototyping of new architectures and evaluation of new software/hardware combinations

Connect training with accessible cloud environment

Full Access to the Cloud Console

- Deploy a full range of resources
- CPU or GPU VMs
- Managed Jupyter notebooks
- Advanced AI/ML capabilities
- Bioinformatic workflow managers
- Access to compute clusters
- High-speed networking
- Support from Cloud Team & CSPS
- On-demand training

Sandbox Modules

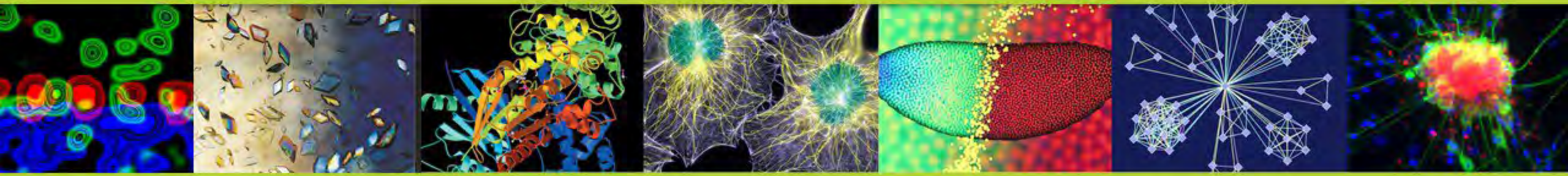
Cloud Lab Environments

- Amazon Web Services
- Google Cloud Services
- Azure

Users

- Intramural
- Extramural

Questions ?



SciHARe

Guest expert

Dr. Alison Lin

NIH/OD

National Institutes of Health
Office of Data Science Strategy

Building a Strong and Diverse Data Science Community

Alison Lin, PhD

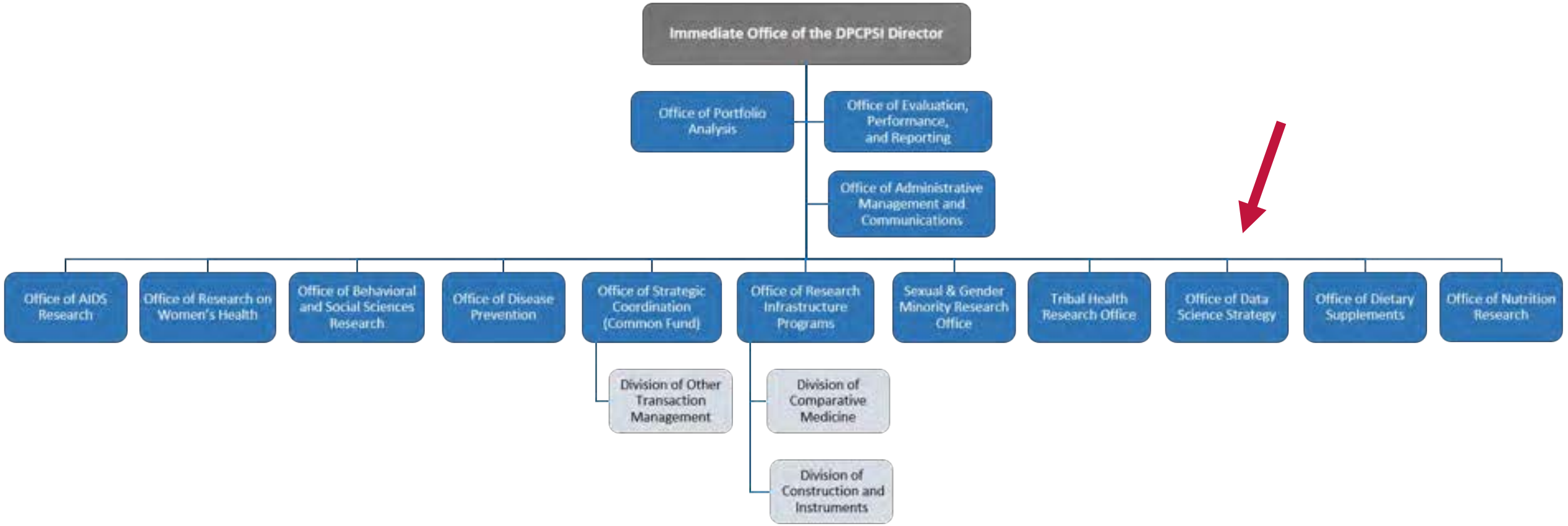
Lead, Training, Workforce Initiatives and Community Engagement (TWICE)

NIH Office of Data Science Strategy

February 15, 2023

Where is NIH Office of Data Science Strategy (ODSS)?

ODSS is within the NIH Office of the Director –
Division of Program Coordination, Planning and Strategic Initiatives (DPCPSI)



Mission and Goals of ODSS

The NIH Office of Data Science Strategy:

- Provides leadership and coordination on the strategic plan for data science
- Develops and implement NIH's vision for a modernized and integrated biomedical data ecosystem
- **Build a strong and diverse data science community**
- Builds strategic partnerships to develop and disseminate advanced technologies and methods

Health Research Needs a Strong Data Science Community

In health research:

- Data science literate
  - Not intimidated by data science
  - Can read and understand reported outcomes resulting from data science approaches
  - Know where to find relevant resources
- Data science savvy – data science literate and
  - Will actively use data sciences approaches in research projects
  - Can initiate and/or participate in collaborations with data scientists
- Data scientist
  - Has skills and expertise in bioinformatics, artificial intelligence, clinical informatics, cloud computing, statistics, computational science, software design and programming, bioinformatics, visualization, machine learning, predictive analytics, supercomputing, modeling and simulation, digital health, data sharing and access, data management, and/or other data science areas
  - Can communicate what they learn and creatively display the information
  - Can formulate implications and implement follow up studies

"Data scientists want to build things, not just give advice … make discoveries while swimming in data …"

Harvard Business Review

Diversity Mitigates Health Disparities and Benefits Research Efforts

Improve access to health care for underserved patients

Increase racial/ethnic minority patient choice and satisfaction

Improve quality of education

- Cultural competency
- Improved learning outcomes

Diverse teams out-perform homogenous teams

Diversity expands range of questions

Facilitates translation of findings to diverse communities

Helps to recruit and retain diverse students and scientists

Training, Workforce Initiatives and Community Engagement (TWICE)

- Supports extramural and intramural training efforts in data science

- Supports development of extramural data science workforce within the extramural biomedical research community
- Facilitates recruitment and retention of data science talents at NIH

- Enhances data science capacity in the extramural community for all populations
- Provides training tools and resources to engage all communities in growing data science knowledge and skills

TWICE Supports Extramural Training Efforts

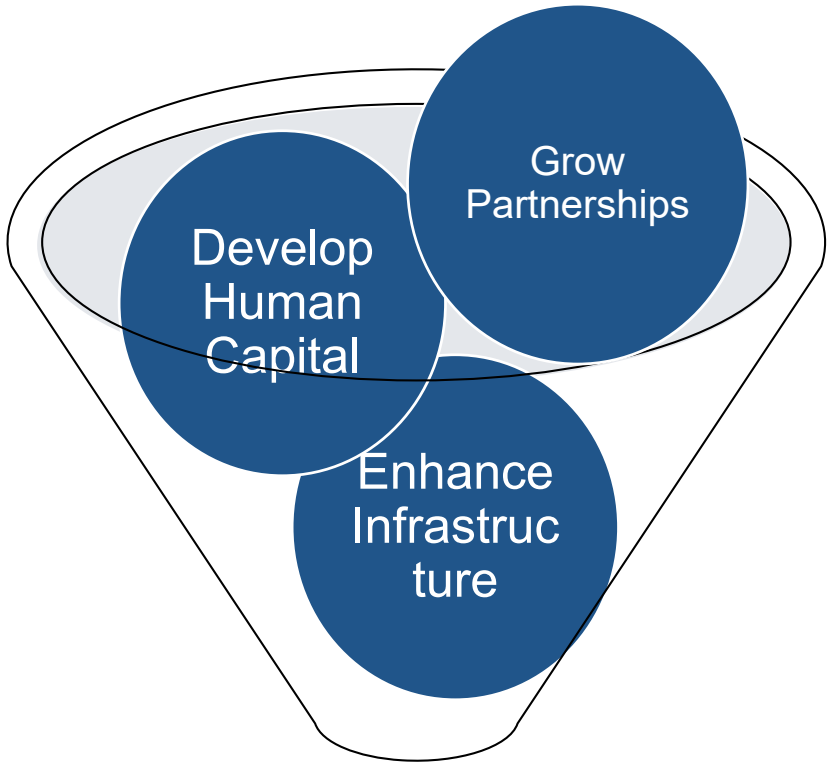
TWICE works to discover, sustain and grow diverse talent

TWICE Training Efforts

- **Provide funding support**
- **Provide support in data skills training and professional development**
- **Promote inclusion of essential data science training elements (such as ethics and cybersecurity)**

- **Div. Sup.** — PA-21-071 Diversity Supplements
- **F31** — F31 Diversity: PA-21-052
- **K** — Diversity-focused K awards
- **T32** — Strong Recruitment Plan to Enhance Diversity
- **R25** — Pre-college research education programs with strong approach to recruit participants from diverse backgrounds

Promote Capacity Building in Low-Resource Institutions



Capacity Building

U54 U24

- Supplements to enhance data science capacity at low resource institutions, NOSI coming soon!
- Collaborate with NIMHD, NIGMS and NCI

S06

- Native American Research Centers for Health (NARCH), data science-relevant project
- Data repository, data training, faculty recruitment, etc.

U24

- Tribal Epidemiology Center, funding for data science-relevant project (NIMHD)
- Education Hub to Enhance Diversity (NHGRI)

P20 T32

- Supplements to the NIGMS' Institutional Development Award (IDeA) Networks of Biomedical Research Excellence (INBRE) awards to build cloud-based learning modules

P20

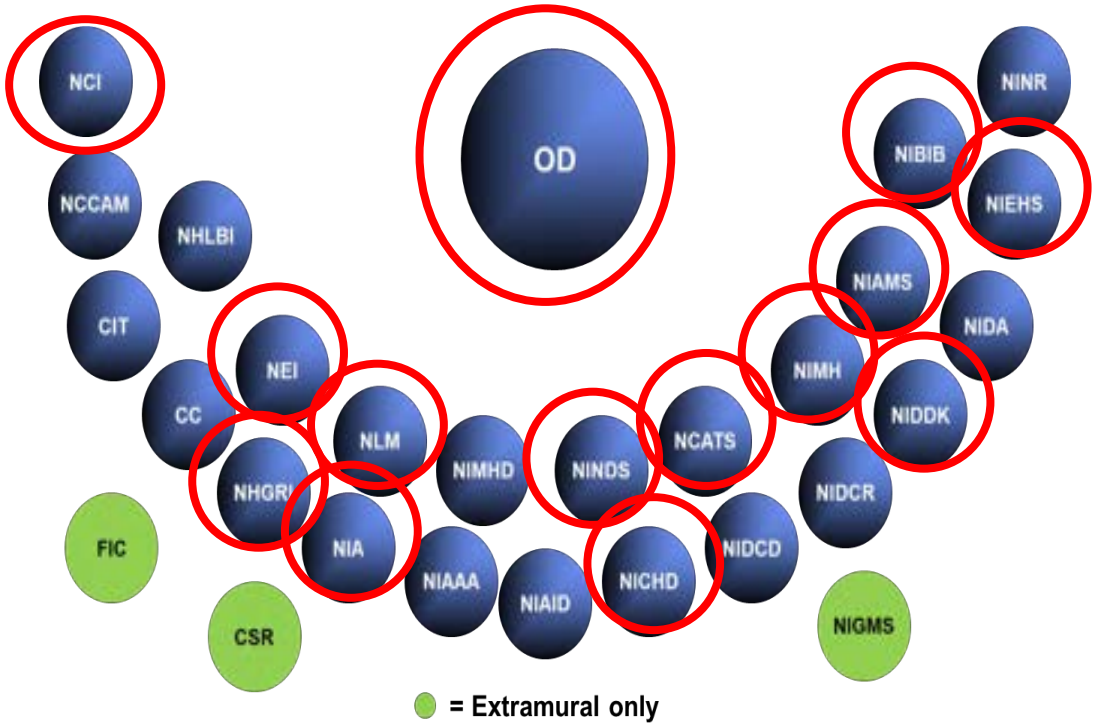
- NIGMS Centers of Biomedical Research Excellence (COBRE) Phase 1 with data science focus

Work at NIH: DATA Scholar Program

DAta and Technology Advancement (DATA) National Service Scholar Program

- Attract talent to the NIH to help optimize and accelerate data science in biomedicine and health research
- Encourage transformative approaches that lead to increased efficiency, innovative research, tool development and analytics
- One to two years commitment
- In addition to their own project, scholars participate in workgroups and collaborations, and contribute in many ways to the NIH
- Applications will open soon for cohort 4 recruitment

DATA Scholar Program: Accomplishments



- 20 Scholars have been matched to 14 ICOs since program launch in 2020 and 7 have completed the program

FY 2022 Accomplishments Include:

- 7 peer reviewed publications
  *Example: Precision Medicine Landscape of Genomic Testing for Patients With Cancer in the National Institutes of Health All of Us Database Using Informatics Approaches by **Jay Ronquillo (NCI; cohort 1)**; JCO Clinical Care Informatics, 2022*
- 13 technologies, products (including websites), inventions, patent applications, and shared resources developed
  *Example: "Kids First Cloud Credits Program" developed by **Anne Deslattes Mays (NICHD; cohort 2)** available: https://github.com/kids-first/kf-cloud-credits#readme*
- 26 oral presentations at scientific conferences
  *Example: "Medical Imaging and Data Resource Center: for a more equitable Medical Imaging AI" presented by **Rui Pereira de Sa (NIBIB; cohort 1)** at the Radiological Society of North America (RSNA), Nov 2022*

NIH's Notice of Interest in Diversity

ODSS particularly encourages applications from individuals from groups identified in NIH's Notice of Interest in Diversity (NOT-OD-20-031) as underrepresented in the biomedical, clinical, behavioral, and social sciences, including **women** and:

Race/Ethnicity

- Blacks or African Americans
- Hispanics or Latinos
- American Indians or Alaska Natives
- Native Hawaiians and other Pacific Islanders

Disability

- Physical or mental impairment that substantially limits one or more major life activities

Disadvantaged Background

- Homeless
- Foster care system
- First generation w/ Bachelor's degree
- Federal Pell Grants
- Special Supplemental Nutrition Program
- Rural or low income/access areas

DATA Scholar Program: Eligibility and Application

Eligibility:

- Must be U.S. citizens, non-citizen nationals, or Permanent Residents
- MD, PhD or other doctoral degrees
- Advanced experience in data science or related fields including:
  - Artificial intelligence, cloud computing, data engineering, data science, database management, project management, software design, supercomputing, and/or bioinformatics

Application:

- Online form
  - Applicant information
  - Choices of projects
- Cover letter
  - Motivation – why interested in program
  - Contribution – how experiences can address data challenges at NIH
  - Vision – impact of data science in biomedical research and health sciences
  - Contribution to enhancing diversity in data science
- Resume – accomplishments including:
  - Data science projects, publications and products
  - Expertise in data science skills, tools and technologies
- Contact info for three references

Coming Soon: NIH FHIR® Training for Extramural Communities

- **Intended Audience:** FHIR® is an interoperability standard that facilitates clinical and healthcare data exchange. This training, which consisted of one 3-hour webinar and two 4-hour hands-on workshops in Spring 2022, was designed for *NIH research scientists and program officers* with some or limited prior experience and knowledge of FHIR®.
- **Course Goal:** To equip participants to better leverage FHIR® in their own research by deepening participants' knowledge of the FHIR® ecosystem and potential applications.

- **Learning Objectives:** Following this course series, you will be able to:
  - Discuss the background and context of the standard;
  - Explain the formatting and components of the FHIR® architecture;
  - Identify the tools, libraries, and resources available for use;
  - Demonstrate how to extract health data using FHIR® standards for the purpose of research;
  - Demonstrate how to use extracted data with other application programming interfaces (APIs).
- We plan to expand this training to extramural communities in 2023

Contact: ***Dr. Hsinyi (Steve) Tsang****; https://datascience.nih.gov/fhir-initiatives/researchers-training*

TWICE Team and Contact

Dr. Alison Lin
Lead

Dr. Raphael Isokpehi
Program Director

Dr. Bryan Kim
Program Director

Nicholas Andrade
Training Specialist

Contact

ds-workforce@nih.gov

ODSS webpage: https://datascience.nih.gov/

Data science funding opportunities: https://datascience.nih.gov/nih-grants-and-funding-opportunities

Data science job opportunities at NIH (federal): https://datascience.nih.gov/jobs


SciHARe

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**

Include in funding applications

Budget for:

- **Staff:**
  - Python expert
  - Data wrangler/manager
- **Data sharing cost,** including CDE use
- **Cloud computing cost:**
  - Analytic time
  - Resource utilization time
  - Data Sharing Access

NIH Data Management and Sharing (DMS) policy:
https://sharing.nih.gov/data-management-and-sharing-policy

FY23 ODSS NOSIs

Notice Number	NOSI Title	Status	Shared Interest Due Date	(Tentative) Publication Date	Submission Due Date
NOT-OD-23-044	Notice of Special Interest (NOSI): Support for existing data repositories to align with FAIR and TRUST principles and evaluate usage, utility, and impact	Published	NA	01/05/2023	03/01/2023
TEMP-21037	Notice of Special Interest (NOSI): Administrative Supplements to Enhance Software Tools for OpenScience	Preliminary Policy Review	01/18/2023	01/31/2023	03/31/2023
TEMP-21039	Notice of Special Interest (NOSI): Administrative Supplements to Support the Exploration of Cloud in NIH-supported Research	Preliminary Policy Review	01/20/2023	02/01/2023	04/07/2023
TEMP-21255	Notice of Special Interest (NOSI): Administrative Supplements to Support Collaborations to Improve the AI/ML-Readiness of NIH-Supported Data	Shared Interest	02/01/2023	02/14/2023	04/18/2023

Use SciHARe

Next Think-a-Thons:

bit.ly/think-a-thons

Register for ScHARe:

bit.ly/join-schare

schare@mail.nih.gov

Thank you

References

1. Tech Blogger, “Benefits and drawbacks of AI in cloud computing”, https://contenteratechspace.com/benefits-and-drawbacks-of-ai-in-cloud-computing/ (accessed January 3, 2023).
2. Kenyon T, “The advantages and disadvantages of AI in cloud computing”, AI Magazine, https://aimagazine.com/ai-strategy/advantages-and-disadvantages-ai-cloud-computing (accessed February 2023).
3. Anand B, “Top 15 Cloud Computing Challenges [with Solution]”, https://www.knowledgehut.com/blog/cloud-computing/cloud-computing-challenges (accessed February 2023).
4. Shukla Shubhendu and Jaiswal Vijay, “Applicability of Artificial Intelligence in Different Fields of Life,” International Journal of Scientific Engineering and Research, vol. 1, no. 1 (September 2013), pp. 28–35.

References

5. Euromoney Learning, “What is Blockchain?”, (accessed January 3, 2023).
6. Jason Brownlee, “A Gentle Introduction to Computer Vision,” Machine Learning Mastery, July 5, 2019.
7. Math Works, “What Is Deep Learning?”, (accessed January 3, 2023).
8. Jason Brownlee, “A Gentle Introduction to Generative Adversarial Networks (GANs),” Machine Learning Mastery, July 19, 2019.
9. John R. Allen and Amir Husain, “Hyperwar and Shifts in Global Power in the AI Century,” in Amir Husain and others, Hyperwar: Conflict and Competition in the AI Century (Austin, TX: SparkCognition Press, 2018), p. 15.
10. Dorian Pyle and Cristina San Jose, “An Executive’s Guide to Machine Learning,” McKinsey Quarterly, June, 2015.

References

11. John Herrman and Kellen Browning “Soon, the Metaverse. Unless It’s Here Now”, New York Times, July 11, 2021.
12. Larry Hardesty, “Explained: Neural Networks,” MIT News, April 14, 2017.
13. Cade Metz, “In Quantum Computing Race, Yale Professors Battle Tech Giants,” New York Times, November 14, 2017, p. B3.
14. Quoted in Tom Wheeler, From Gutenberg to Google: The History of Our Future (Brookings, 2019), p. 226. Also see Ray Kurzweil, The Singularity Is Near: Where Humans Transcend Biology (London: Penguin Books, 2006).
15. Jack Karsten and Darrell M. West, “China’s Social Credit System Spreads to More Daily Transactions,” TechTank (blog), Brookings, June 18, 2018.

References

16. Matthew Hutson, “AI Glossary: Artificial Intelligence, in So Many Words,” Science, July 7, 2017.
17. Wikipedia contributors, "Google Cloud Platform," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Google_Cloud_Platform&oldid=1131253580 (accessed January 3, 2023).
18. Wikipedia contributors, "FAIR data," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/FAIR_data (accessed January 3, 2023).