ScHARe

Think-a-Thons

NIH National Institutes of Health

An Introduction to FAIR Data and AI-ready Datasets

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

Rebecca Rodriguez, PhD MS · NIDDK
Anya Dabic · Booz Allen Hamilton
Summer Rankin, PhD · Booz Allen Hamilton
Courtney D. Shelley, PhD · Booz Allen Hamilton

November 15, 2023

ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

ScHARe

NIH National Institute on Minority Health and Health Disparities

+

NIH Office of Data Science Strategy

+

NIH National Institute of Nursing Research

ScHARe

+

Dr. Deborah Duran NIH/NIMHD

Dr. Luca Calzoni NIH/NIMHD

Thank you

NIMHD

Dr. Eliseo
Perez-Stable

ODSS

Dr. Susan
Gregurick

NIH/OD

Dr. Larry
Tabak

NINR

Dr. Shannon
Zenk

NINR
Rebecca Hawes
Micheal Steele
John Grason

ORWH

OMH

NIMHD OCPL
Kelli Carrington
Thoko Kachipande
Corinne Baker

BioTeam

STRIDES

Terra

SIDEM

RLA

Broad Institute

CCDE Working Group

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

Experience poll

Please check your level of experience with the following:

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

ScHARe

Overview

ScHARe

Phase I

Population Science and SDoH datasets
Tutorials and resources
Think-a-Thons

ScHARe

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

ScHARe aims to fill three critical gaps:

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

nimhd.nih.gov/schare

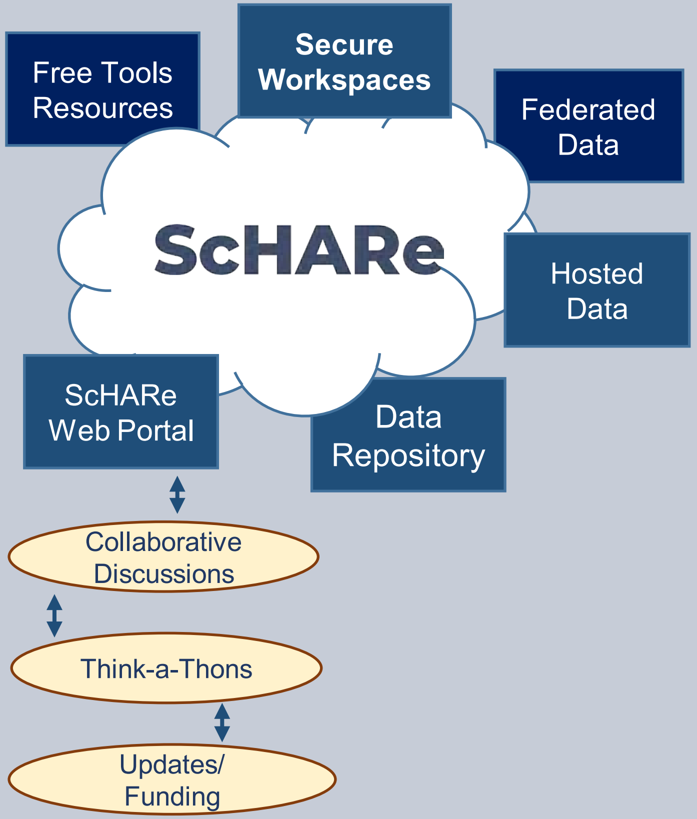
ScHARe Components

ScHARe co-localizes within the cloud:

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

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

Intramural & Extramural Resource



nimhd.nih.gov/schare

ScHARe Data Ecosystem

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

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

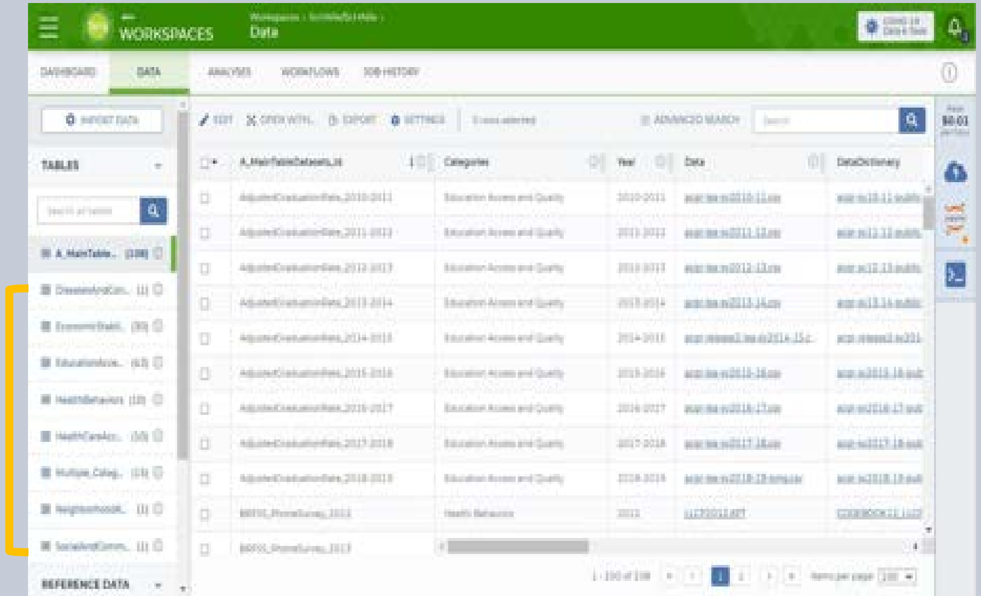
  Example: *American Community Survey (ACS)*

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

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

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

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



A_MainTableDatasets_id	Categories	Year	Data	DataDictionary
AdjustedGraduationRate_2010-2011	Education Access and Quality	2010-2011	[illegible]	[illegible]
AdjustedGraduationRate_2011-2012	Education Access and Quality	2011-2012	[illegible]	[illegible]
AdjustedGraduationRate_2012-2013	Education Access and Quality	2012-2013	[illegible]	[illegible]
AdjustedGraduationRate_2013-2014	Education Access and Quality	2013-2014	[illegible]	[illegible]
AdjustedGraduationRate_2014-2015	Education Access and Quality	2014-2015	[illegible]	[illegible]
AdjustedGraduationRate_2015-2016	Education Access and Quality	2015-2016	[illegible]	[illegible]
AdjustedGraduationRate_2016-2017	Education Access and Quality	2016-2017	[illegible]	[illegible]
AdjustedGraduationRate_2017-2018	Education Access and Quality	2017-2018	[illegible]	[illegible]
AdjustedGraduationRate_2018-2019	Education Access and Quality	2018-2019	[illegible]	[illegible]
BRFSS_PhoneSurvey_2011	Health Behaviors	2011	[illegible]	[illegible]
BRFSS_PhoneSurvey_2013				

1 - 100 of 108

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:

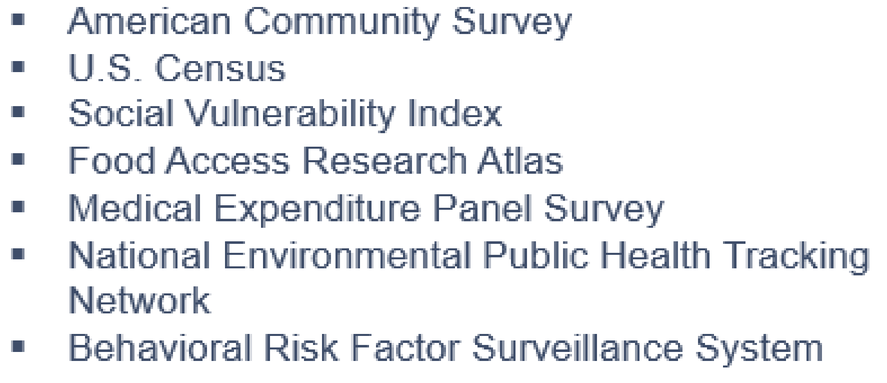
- **Health Behaviors**
- **Diseases and Conditions**

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

Access to Population Science datasets

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

- Google Cloud Public Datasets
- ScHARe Hosted Public Datasets:



- 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

- **Coming Soon:** Repository for Funded Datasets on ScHARe, in compliance with NIH Data Sharing Policy

Cloud computing strategies

ScHARe

- Uses **workflows** in Workflow Description Language (**WDL**), a language easy for humans to read, for batch processing data
- **Python and R**, including most commonly used libraries
- Enables **customization** of computing environments to ensure everyone in your group is using the same software
- **Big Query** and **Tensorflow** access for advanced machine learning
- Enables researchers to create interactive **Jupyter notebooks** (documents that contain live code) and share data, analyses and results with their collaborators in real time
- For novice users, integration with **SAS** is planned

AI bias mitigation strategies

- Widespread use of AI raises a number of ethical, moral, and legal issues – likely not to go away
- Algorithms often are “black boxes”
- **Biases can result from:**
  - **social/cultural context not considered**
  - **design limitations**
  - **data missingness and quality problems**
  - **algorithm development and model training**
  - **Implementation**
- If not rectified, biases may result in decisions that lead to discrimination, unequitable healthcare, and/or health disparities
- **Lack of diverse perspectives:** populations with health disparities are underrepresented in data science
- **Guidelines** and recommendations emerging from HHS, NIST, White House, etc.

ScHARe

Critical thinking can rectify AI biases

ScHARe was created to:

- foster participation of **populations with health disparities in data science**
- promote the collaborative identification of **bias mitigation strategies** across the continuum
- create a **culture of ethical inquiry** and critical thinking whenever AI is utilized
- build **community confidence** in implementation approaches
- focus on **implementation of AI bias** guidelines and recommendations

ScHARe

Phase II

(in process)

Data ecosystem and repository

ScHARe Data Repository

CORE COMMON DATA ELEMENTS

NOVEL CDE FOCUSED REPOSITORY TO FOSTER INTEROPERABILITY

COMPLY WITH DATA SHARING POLICY - HOST PROJECT DATA

DATA ECOSYSTEM

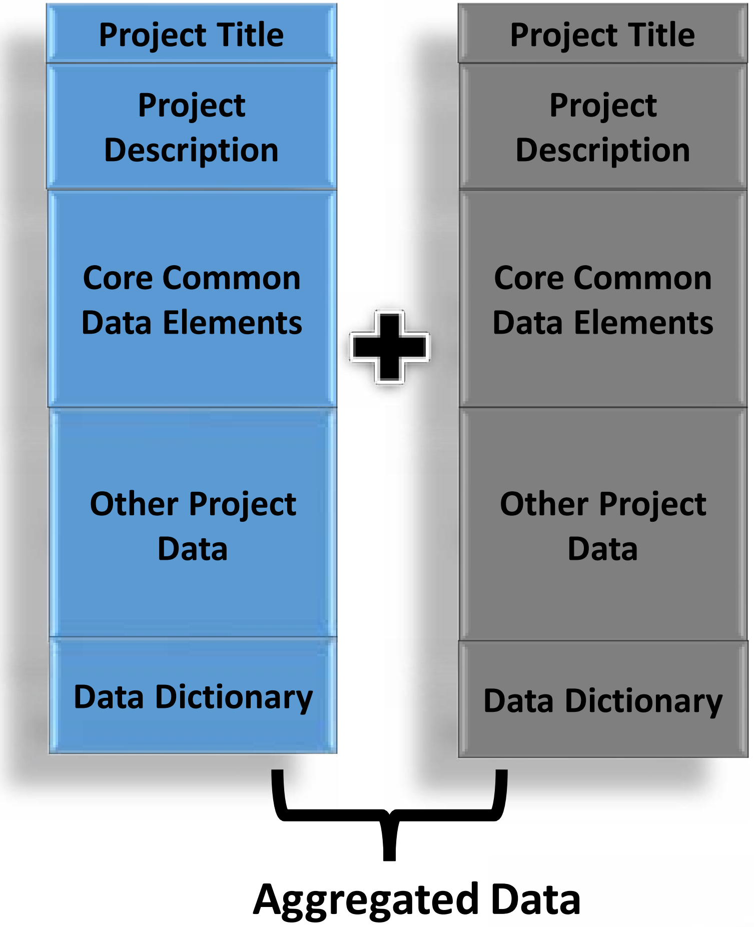
- Map across datasets
- Map across platforms

UPCOMING

ScHARe

Core Common Data Elements
Intramural and Extramural Project Repository

- Complies with **NIH Data Sharing Policy**
- Fosters dataset sharing and interoperability by using or mapping to **Core Common Data Elements**
- Provides resources for **intramural researchers** to work in a secure workspace and host data
- Centralizes **aggregated datasets** for repeat use


Project Title

Project Description

Core Common Data Elements

Other Project Data

Data Dictionary

+

Project Title

Project Description

Core Common Data Elements

Other Project Data

Data Dictionary

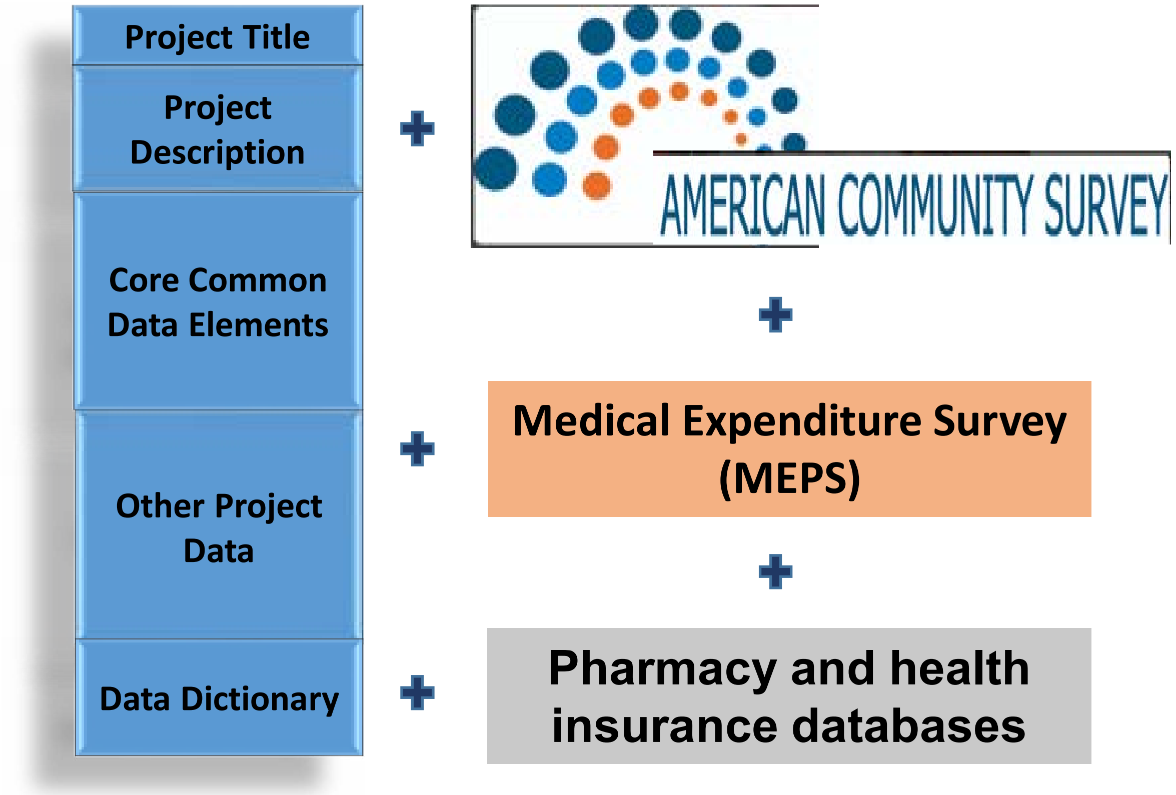
Aggregated Data

smoking
SDoH
geographic
Community
Interventions
Sleep
cancer
smoking
Mental health
geographic
Adherence
Applied Clinical
Mental health
geographic
Health Care Delivery
Decision making
SDoH
Decision making
Diabetes
SDOH
cancer

UPCOMING

ScHARe

Project & federated dataset mapping



Project Title

Project Description

Core Common Data Elements

Other Project Data

Data Dictionary

+ AMERICAN COMMUNITY SURVEY

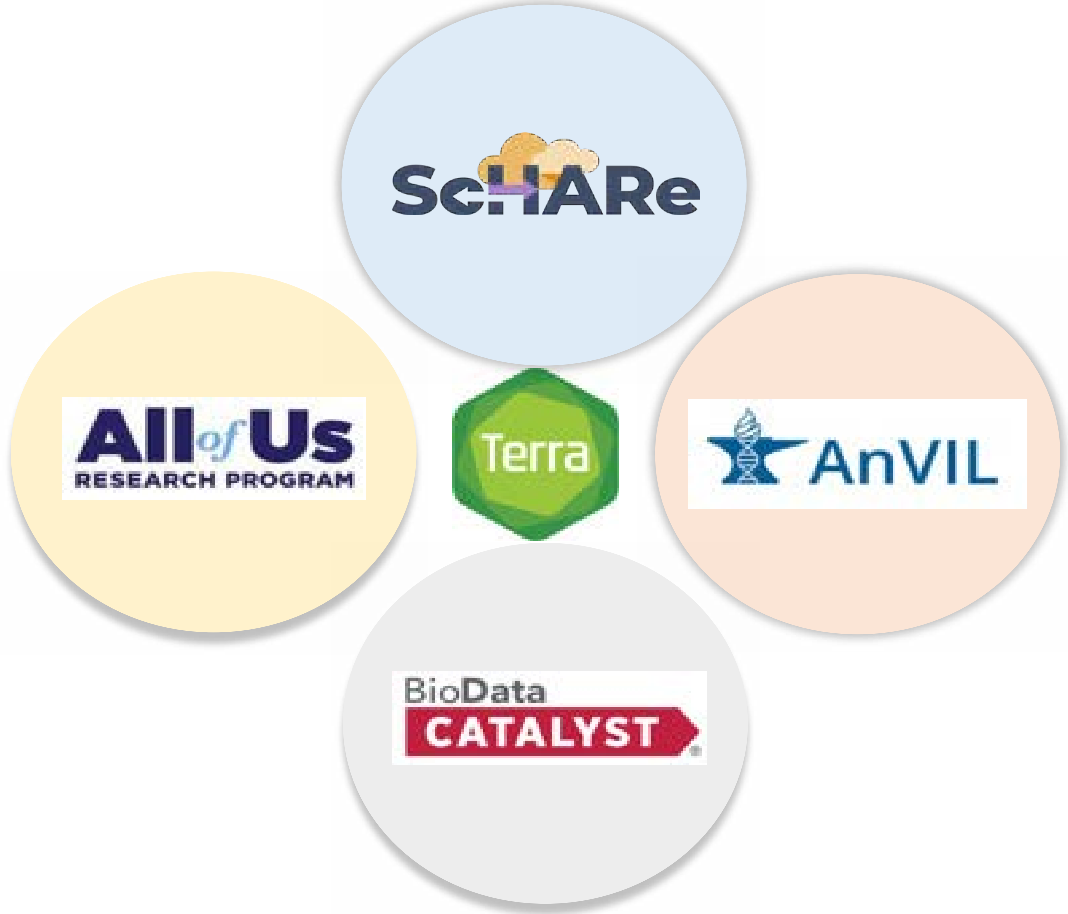
+

+ Medical Expenditure Survey (MEPS)

+

+ Pharmacy and health insurance databases

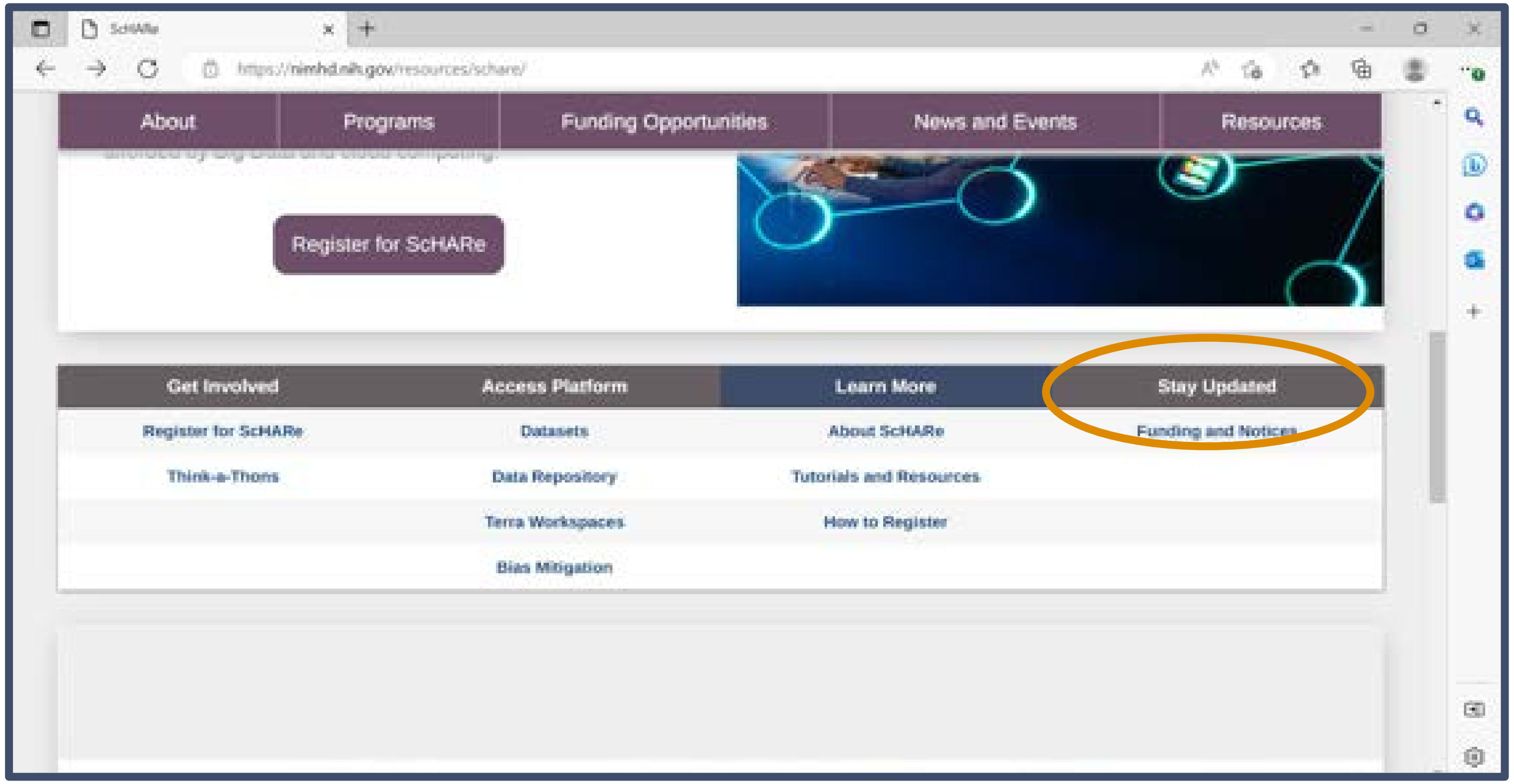
Mapping across cloud platforms



UPCOMING

Two ways to sign up for ScHARe news

Scannable from your screen!



nimhd.nih.gov/schare

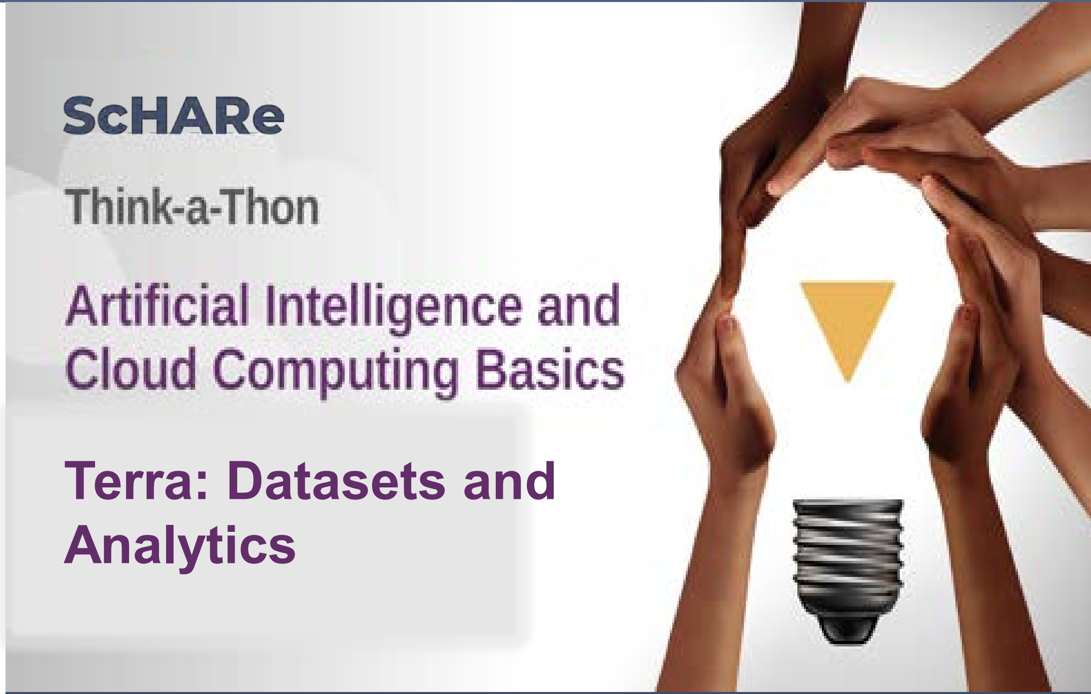
ScHARe Think-a-Thons (TaT)

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

✓ **Instructional**

✓ **Collaboration research teams**

✓ **Bias mitigation**



ScHARe

Think-a-Thon

Artificial Intelligence and Cloud Computing Basics

Terra: Datasets and Analytics

Register:

bit.ly/think-a-thons

Interest poll

I am interested in (check all that apply):

☐ Learning about Health Disparities and Health Outcomes research to apply my data science skills

☐ Conducting my own research using AI/cloud computing and publishing papers

☐ Connecting with new collaborators to conduct research using AI/cloud computing and publish papers

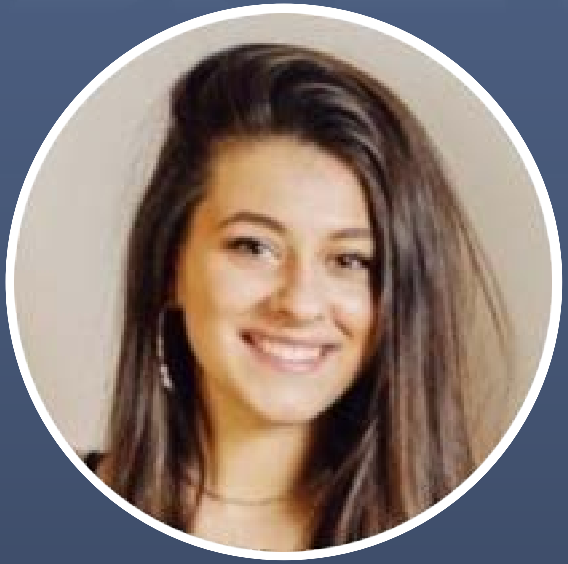
☐ Learning to use AI tools and cloud computing to gain new skills for research using Big Data

☐ Learning cloud computing resources to implement my own cloud

☐ Developing bias mitigation and ethical AI strategies

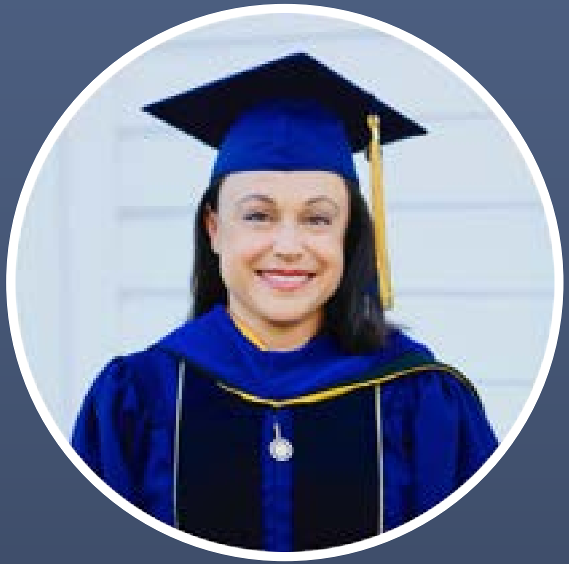
☐ Other

ScHARe Guest experts



Andrijana (Anya) Dabic

Summer Rankin, PhD



Courtney D. Shelley, PhD

Booz Allen Hamilton

NIDDK Central Repository

About Anya

Anya Dabic is a Health Data Scientist at Booz Allen Hamilton that specializes in scientific data management and stewardship.

She has supported various data management and sharing programs at the National Institutes of Health to implement the FAIR (findable, accessible, interoperable, reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability and Technology) principles for digital repositories, including the NIDDK Central Repository, NICHD Data and Specimen Hub, and RADx Data Hub.

Her expertise includes health IT standards and technology, biomedical semantic standards, metadata models and schemas, data governance, and privacy preserving record linkage.

Ms. Dabic completed her B.Sc. in Biomedical Engineering at the University of Virginia.

About Summer

Summer Rankin, PhD is a computational neuroscientist who investigates the boundaries of AI and drives data science solutions for federal government clients.

She has a doctorate in complex systems and brain sciences and works as a senior lead data scientist at Booz Allen Hamilton’s Honolulu Chief Technology Office.

She leads projects that involve a range of machine learning techniques including: deep learning, natural language processing, anomaly detection, and performance measurement.

She serves as an artificial intelligence subject matter expert for Indo-Pacific defense and health projects with recent publications modeling mortality rates in chronic kidney disease (ONC) and adverse event detection from EHRs (FDA).

She holds a PhD in Complex Systems and Brain Sciences and and completed a postdoctoral fellowship with Charles Limb, MD at Johns Hopkins School of Medicine.

She has multiple peer-reviewed publications, public software releases, and conference presentations in the fields of AI, data science and neuroscience.

About Courtney

Courtney D. Shelley, PhD, is a Health Data Scientist at Booz Allen Hamilton, where she focuses on data science education and AI-readiness of health-related data.

She has supported the NIH Office of Data Science Strategy to develop online data science learning resources for pre-college and collegiate audiences, and to assess data science education across US universities to promote collaborative research between biomedical researchers and AI professionals.

Prior to working at Booz Allen Hamilton, Dr. Shelley worked at Los Alamos National Laboratory, where she received the Postdoctoral Distinguished Performance Award for COVID-19 response efforts at local, state, and federal levels, as well as conducted research in suicide prevention with the support of the Department of Veterans Affairs and Million Veteran Program.

She completed her PhD in Epidemiology with a focus on causal inference at University of California, Davis.

NIDDK-CR Data Centric Challenge

The NIDDK Central Repository is conducting a **Data Centric Challenge**

Goals:

1. Seek approaches to enhance the utility of select NIDDK datasets focused on Type 1 Diabetes (T1D) for future secondary research
2. Generate “AI-ready” datasets

Register:

Visit Challenge.gov to learn more and register for the Challenge

by **5PM EST November 30, 2023**

An Introduction to FAIR Data and AI-ready Datasets

Presenting on behalf of the NIDDK Central Repository Program:

Anya Dabic, Booz Allen Hamilton
Summer Rankin, PhD, Booz Allen Hamilton
Courtney D. Shelley, PhD, Booz Allen Hamilton

About the NIDDK Central Repository (NIDDK-CR)

- Established in 2003 to expand the usefulness of extramurally NIDDK-funded multi-center clinical studies' generated resources by providing access to a wider research community
  - Supports receipt and distribution of data and biospecimens in a manner that is ethical, equitable, and efficient
  - Enables investigators not involved with the original work to test new hypotheses without the need to collect new data and biospecimens
  - Promotes FAIR (Findable, Accessible, Interoperable, and Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles – and recently CoreTrustSeal certified
  - NIDDK-CR website is open to all to explore, view available resources, upcoming studies, and register an account which is easy and free

160 study collections,
144 with clinical phenotype data, **94** with samples available

Imaging Data
1394 GB
Compressed data available

Clinical Data
9.7 GB
Compressed data available

Biospecimens
6.2M
>70 different material types not including T1D studies

Website Users
5732
Total registered users and >57K annual visitors

Approved investigators
>800
Approved external and ancillary investigators

Public Releases
>500
Reported from external and ancillary investigators

NIDDK-CR Data Centric Challenge – Overview

Background

- The NIDDK Central Repository is conducting a Data Centric Challenge aimed at augmenting and enhancing existing Repository data for future secondary research, including data-driven discovery by **artificial intelligence (AI)** researchers
- The **NIDDK-CR Data Centric Challenge** will build upon future challenges to develop tools, approaches, models and/or methods to increase data interoperability and usability for artificial intelligence and machine learning applications
- Towards this, NIDDK is seeking innovative approaches to enhance the utility of select NIDDK datasets focused on **Type 1 Diabetes (T1D)**
  - The Environmental Determinants of Diabetes in the Young (TEDDY)
  - Four studies from the Type 1 Diabetes TrialNet

Goals of the Data Centric Challenge

1. Generate an "AI-ready" dataset that can be used for future data challenges
2. Produce methods that can be used to enhance the AI-readiness of NIDDK data

Key Points

- NIDDK-CR will host regular **Office Hours** during the challenge to answer questions and provide participant's the opportunity to continue to gain experience in the AI-research space.
- Visit Challenge.gov to learn more and register for the Challenge by **5PM EST November 30, 2023,** by visiting Challenge.gov today!

NIDDK-CR Data Centric Challenge – Overview

Participation in this challenge is tiered based on the challenge applicants' self-described experience with data science and analytics

All participants will be instructed to 1) prepare a ***single merged dataset*** by aggregating all data files associated with TEDDY or TrialNet, and 2) augment and enhance the single merged dataset to prepare a ***single AI-ready dataset***.

- **Beginner** (TEDDY) – For the beginner-level challenge, the goal for challenge participants will be to *aggregate and harmonize* datasets from the TEDDY study into a single unified and machine-readable dataset, then enhance the aggregated dataset for AI-readiness.
- **Intermediate/Advanced** (TrialNet) – For the intermediate to advanced-level challenge, the goal for challenge participants will be to *aggregate, harmonize, and fuse* four studies within the TrialNet set of studies (TN01, TN16, TN19, and TN20) into a single unified and machine-readable dataset, then enhance the harmonized dataset for AI-readiness.

Data will be made accessible to participants through the **NIDDK-CR Analytics Workbench** which provides computational tools to access and analyze the data

Service Workbench

Service Workbench on AWS (devlce-east-1)

Webinar Topics

1. Overview of AI-Assisted Research
   - Introduction to AI-assisted research as an iterative process
   - Definition of an AI-ready dataset (and other AI concepts)
   - Importance of strong research design and subject matter expertise for transparency
   - Bias in AI
   - FAIR and CARE principals
2. Performing Pre-Model Processing and Data Quality Checks
   - Importing data in an IDE
   - Understanding the data
   - Performing data pre-processing and quality checks
   - Documenting the data
3. **Live Demo:** Data Handling in Jupyter Notebook
4. Conducting AI-Research for Health Data
   - Handling imbalanced datasets
   - Selecting an ML model
   - Real example of an ML model

Overview of AI-Assisted Research

Summer Rankin, PhD, Booz Allen Hamilton

What questions can be answered with AI?

***AI is an outcome**—the ability of machines to perform tasks that typically require human-level intelligence*

	perception	**notification**	**suggestion**	**automation**	**prediction**	**prevention**	**situational awareness**
	Describe and understand surroundings	*Provide alerts, reminders, etc.*	*Build on past preferences and modify over time*	*Follow routine steps to accomplish an objective*	*Forecast the likelihood of future events based on past events*	*Apply cognitive reckoning to identify potential threats*	*Summarize the current, and likely future, environment*
Key Questions Answered	What's happening now?	What do I need to know?	What do you recommend?	What should I do?	What can I expect to happen?	What can/should I avoid?	What do I need to do now?

THE CURRENT ROLE OF AI:

Curator — Recommender — Orchestrator

NOT THE ROLE OF AI:

Critical Thinker — Decision Maker

Slido Quiz

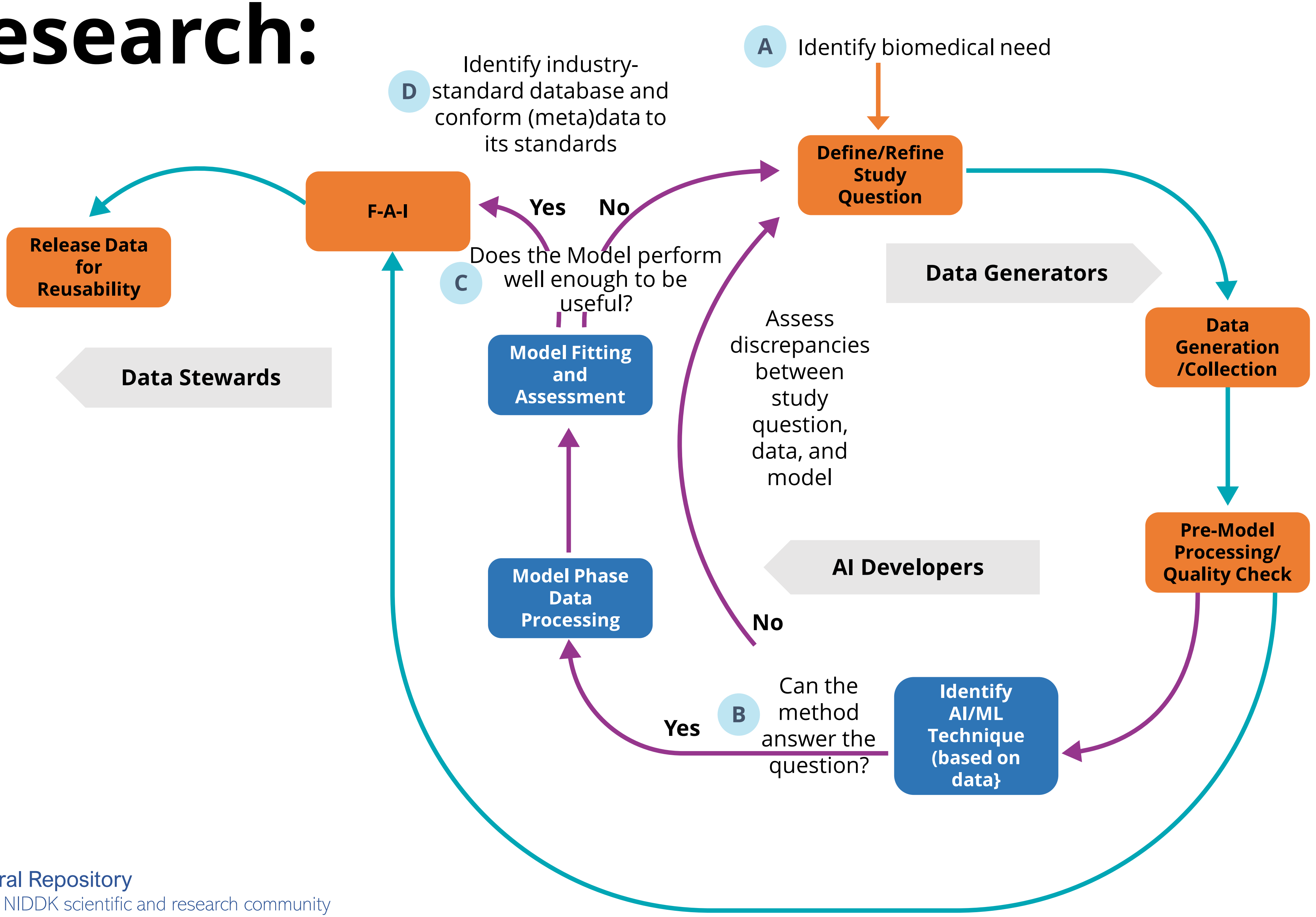
1. What do you think are the important features of an AI-ready dataset?

What is an AI-ready dataset?

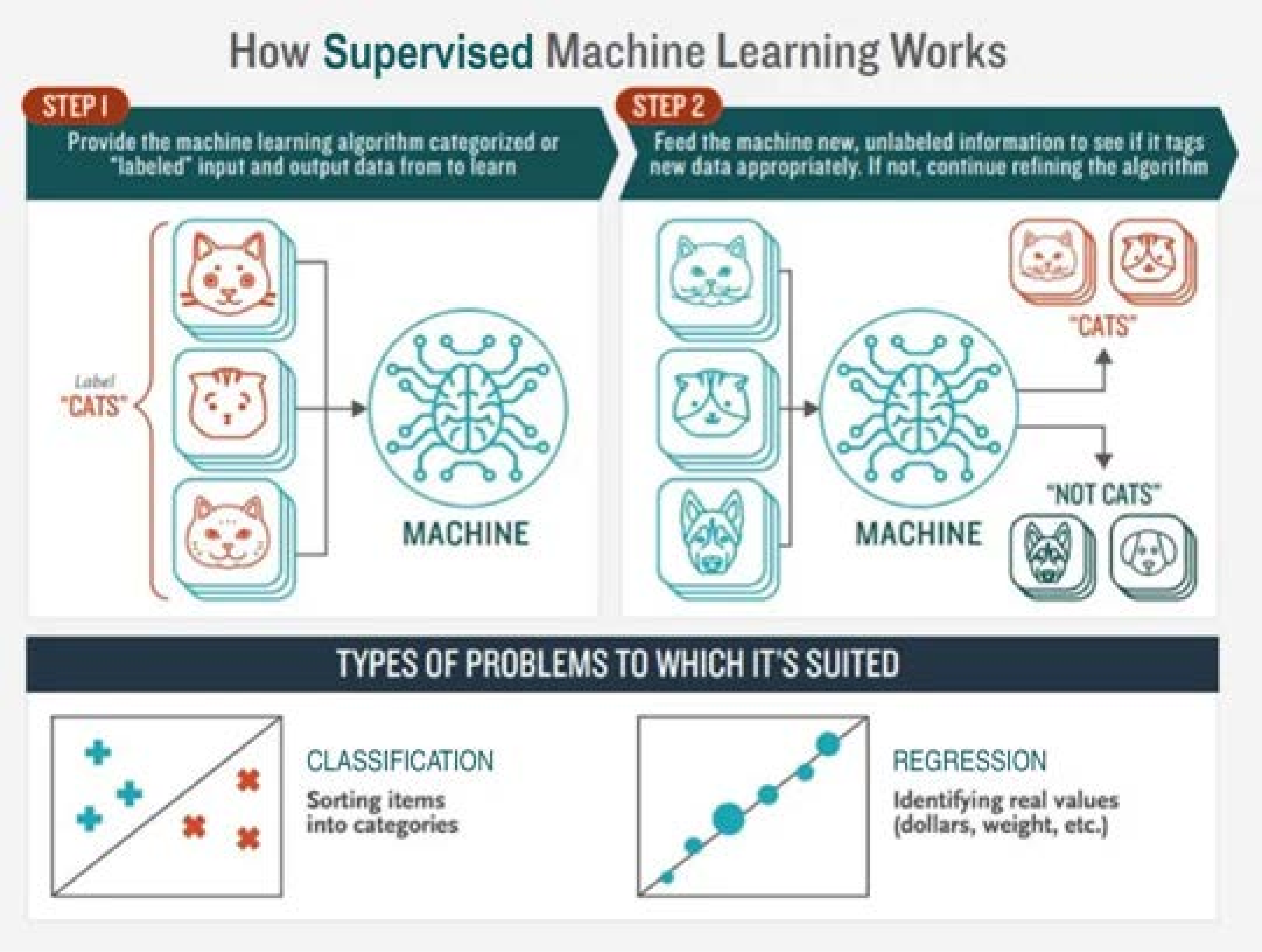
AI-readiness refers to data that are *machine-readable, reliable, accurate, explainable, predictive,* and *accessible for future AI applications*

- An AI-ready dataset consists of:
  - Data that is reflective of the population from which it was drawn
  - Data that is well documented and FAIR (findable, accessible, interoperable, and reusable)
  - Data that is model-agnostic
- AI-readiness will include:
  - ✓ **pre-processing steps** such as addressing errant values,
  - ✓ **handling of missing values,**
  - ✓ **relabeling and recoding** of data elements (aka columns, variables, features, or attributes) and values during harmonization to ensure consistency and standardized formatting
  - ✓ **documentation** of all data handling steps, all variables, and the dataset itself
- When possible,
  - attempt to **retain as much information as possible** by creating new data elements that are transforms of existing elements without deleting or overwriting existing elements.

AI Research:


A
Identify biomedical need
Define/Refine Study Question
Data Generators
Data Generation /Collection
Pre-Model Processing/ Quality Check
Identify AI/ML Technique (based on data}
B
Can the method answer the question?
No
Yes
Assess discrepancies between study question, data, and model
AI Developers
Model Phase Data Processing
Model Fitting and Assessment
C
Does the Model perform well enough to be useful?
Yes
No
D
Identify industry-standard database and conform (meta)data to its standards
F-A-I
Release Data for Reusability
Data Stewards

Supervised Learning



How Supervised Machine Learning Works

STEP 1

Provide the machine learning algorithm categorized or "labeled" input and output data from to learn

STEP 2

Feed the machine new, unlabeled information to see if it tags new data appropriately. If not, continue refining the algorithm

TYPES OF PROBLEMS TO WHICH IT'S SUITED

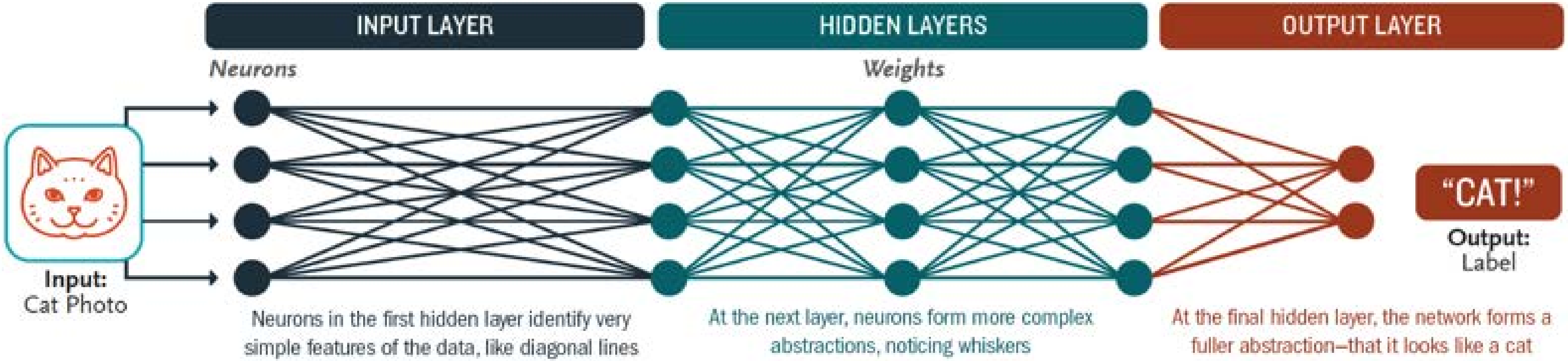
CLASSIFICATION

Sorting items into categories

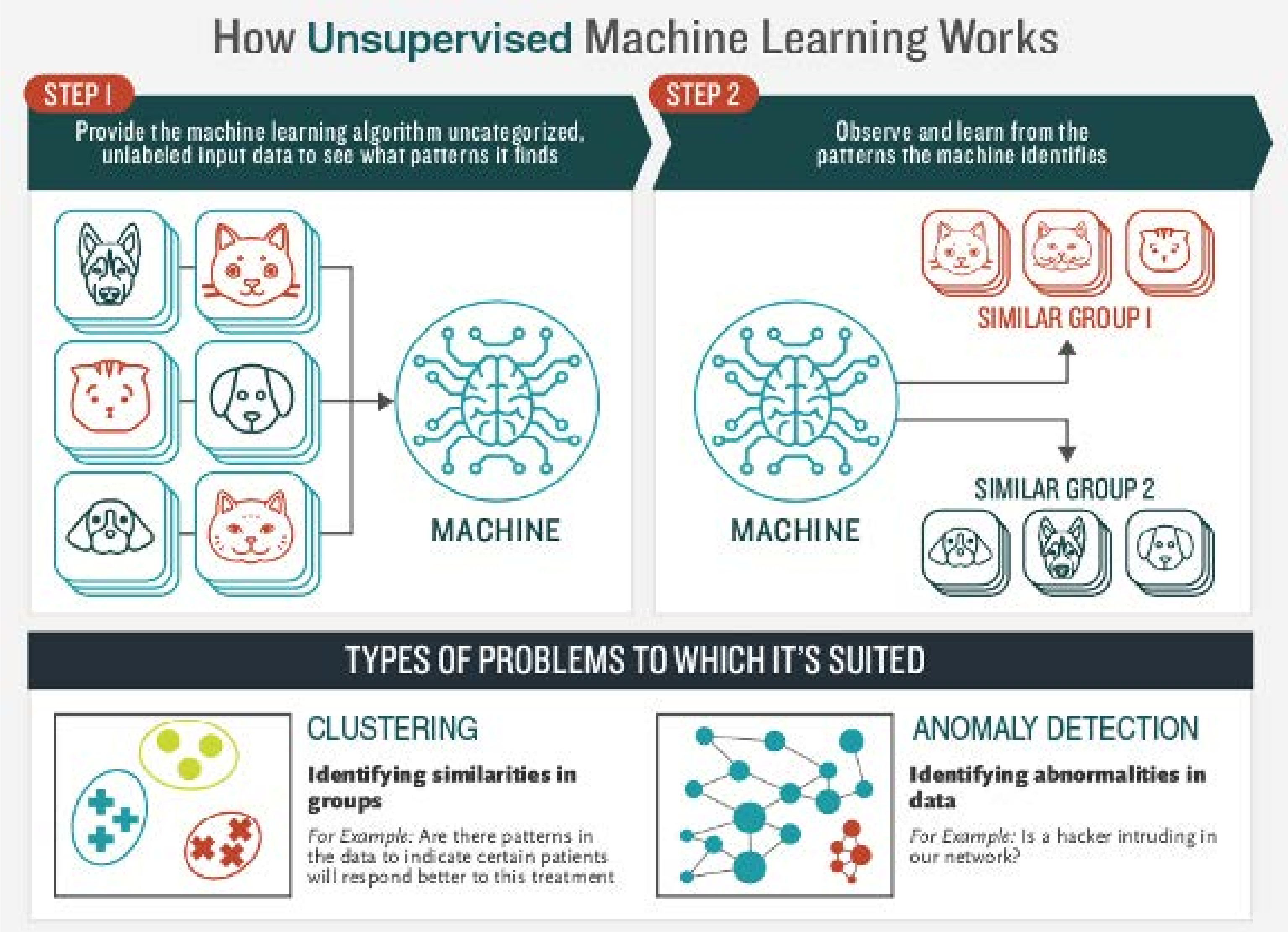
REGRESSION

Identifying real values (dollars, weight, etc.)

Deep Learning


INPUT LAYER
HIDDEN LAYERS
OUTPUT LAYER
Neurons
Weights
Input:
Cat Photo
"CAT!"
Output:
Label
Neurons in the first hidden layer identify very simple features of the data, like diagonal lines
At the next layer, neurons form more complex abstractions, noticing whiskers
At the final hidden layer, the network forms a fuller abstraction–that it looks like a cat

Unsupervised Learning



How Unsupervised Machine Learning Works

STEP 1

Provide the machine learning algorithm uncategorized, unlabeled input data to see what patterns it finds

STEP 2

Observe and learn from the patterns the machine identifies

TYPES OF PROBLEMS TO WHICH IT'S SUITED

CLUSTERING

Identifying similarities in groups

For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment

ANOMALY DETECTION

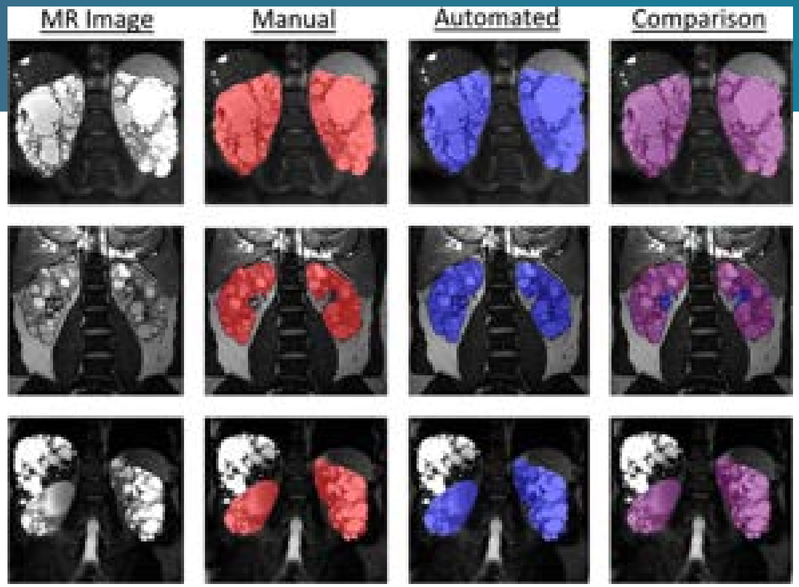
Identifying abnormalities in data

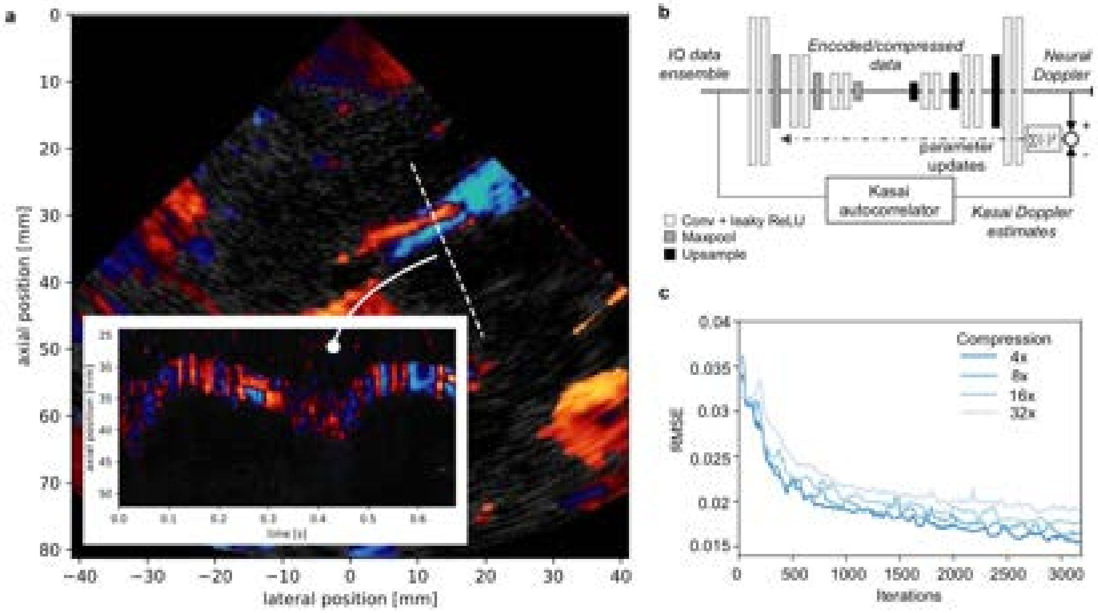
For Example: Is a hacker intruding in our network?

AI in Health

Labeled, annotated images

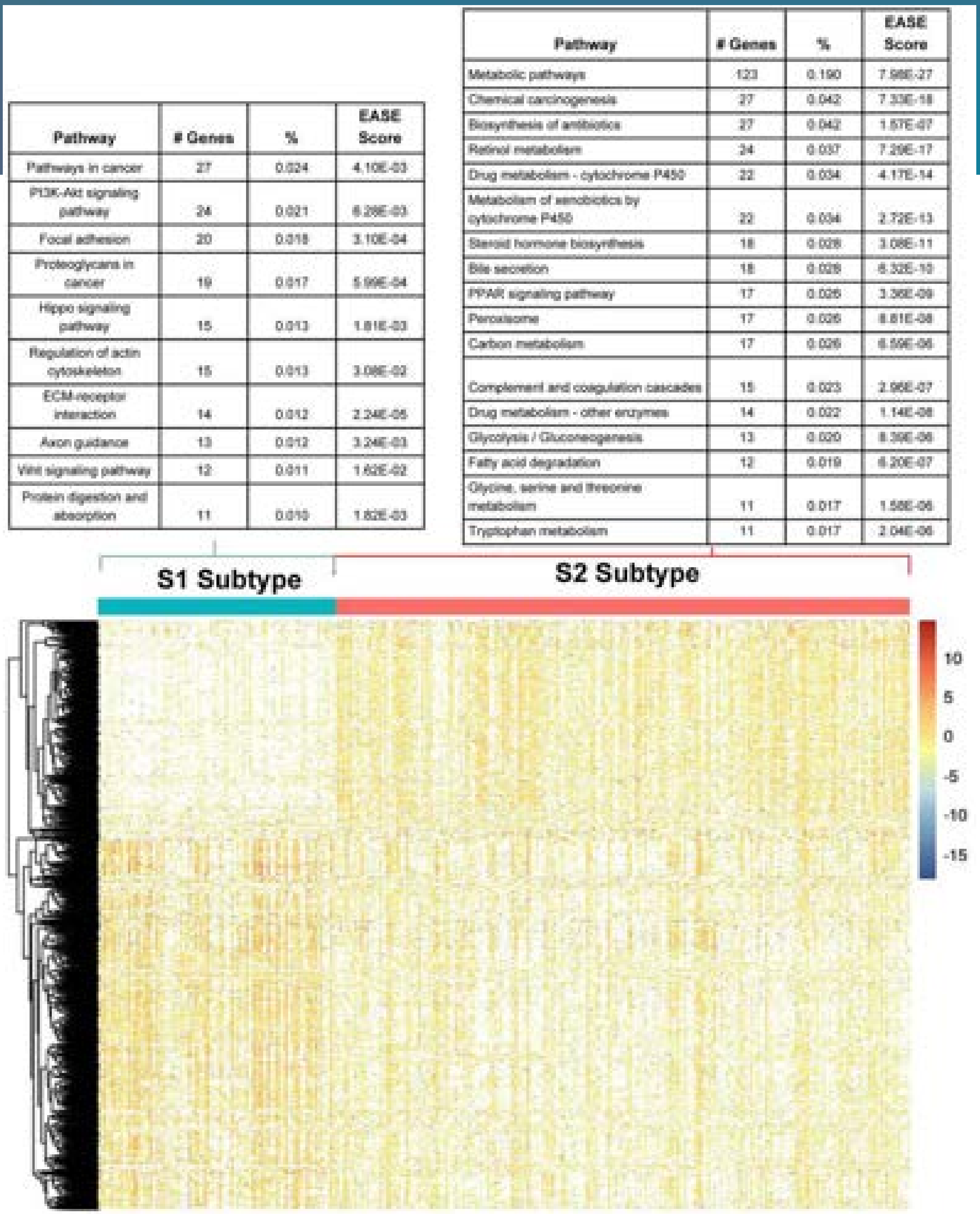
- Feature Extraction - Image segmentation (US, CT, MRI)
- Deep Learning - Learn important low-level and high-level features
  - *Image Augmentation*
  - *Transfer learning*
  - *Architectures for Deep Learning*
    - *Convolutional Neural Nets (CNN)*
    - *Autoencoders (AE)*
    - *Recurrent Neural Networks (RNN)*
    - *Deep Belief Network (DBN)*
  - Voxel-wise classification





AI in Health

- -omic sequence data is treated like a sequence and/or language
- Deep Learning Architectures
  - *Transfer learning from pre-trained models*
  - *Convolutional Neural Nets (CNN) - treat a window of the sequence as an image*
  - *Variational Autoencoders (VAE)*
  - *Recurrent Neural Networks (RNN)*
  - *Long Short-Term Memory (LSTM)*
    - *GENOMIC-ULMFiT – from FAST AI*
  - *Bi-directional Transformer models (BERT)*



Pathway	# Genes	%	EASE Score
Pathways in cancer	27	0.024	4.10E-03
PI3K-Akt signaling pathway	24	0.021	6.28E-03
Focal adhesion	20	0.018	3.10E-04
Proteoglycans in cancer	19	0.017	5.99E-04
Hippo signaling pathway	15	0.013	1.81E-03
Regulation of actin cytoskeleton	15	0.013	3.08E-02
ECM-receptor interaction	14	0.012	2.24E-05
Axon guidance	13	0.012	3.24E-03
Wnt signaling pathway	12	0.011	1.62E-02
Protein digestion and absorption	11	0.010	1.82E-03

Pathway	# Genes	%	EASE Score
Metabolic pathways	123	0.190	7.98E-27
Chemical carcinogenesis	27	0.042	7.33E-18
Biosynthesis of antibiotics	27	0.042	1.57E-07
Retinol metabolism	24	0.037	7.29E-17
Drug metabolism - cytochrome P450	22	0.034	4.17E-14
Metabolism of xenobiotics by cytochrome P450	22	0.034	2.72E-13
Steroid hormone biosynthesis	18	0.028	3.08E-11
Bile secretion	18	0.028	6.32E-10
PPAR signaling pathway	17	0.026	3.36E-09
Peroxisome	17	0.026	8.81E-08
Carbon metabolism	17	0.026	6.59E-06
Complement and coagulation cascades	15	0.023	2.96E-07
Drug metabolism - other enzymes	14	0.022	1.14E-08
Glycolysis / Gluconeogenesis	13	0.020	8.39E-06
Fatty acid degradation	12	0.019	6.20E-07
Glycine, serine and threonine metabolism	11	0.017	1.58E-06
Tryptophan metabolism	11	0.017	2.04E-06

Research Design

- Develop and define a systematic plan to study a scientific problem.
- Identify the type of study (e.g., descriptive, review, experimental), research question, hypothesis, variables, design, data collection, and subsequent statistical analysis plan.
- **Identify the data required to study this question: especially demographic details**
- Types of data that can support outcomes research:
  - Clinical Data – doctors' notes, prescription records, lab images and notes, insurance (claims) data, electronic health record (EHR) data
  - Patient-Sourced Data – sensors, survey measures, social media posts, preferences, wearables data

DATA CONSIDERATIONS

- Domain experts needed to inform data-use assumptions
- **Data source and details need to represent the population of interest**
- All algorithms inherently involve assumptions, some of which are *not* verifiable by the data
- Unmeasured, random variation mitigated by design/replication
- Non-random or systematic variation, more commonly encountered with "found" data (selection/confounding bias)[1]
- The learning 'target' (prediction, estimation) must guide chosen priorities in data considerations

Research Design

Use Case: Predict mortality for chronic kidney disease patients in the first 90 days of dialysis.

- The first 90 days following initiation of chronic dialysis represent a high-risk period for adverse outcomes, including mortality
- While the sudden and unplanned start of dialysis is a known risk factor, other factors leading to poor outcomes during this early period have not been fully delineated
- Tools to identify patients at highest-risk for poor outcomes during this early period are lacking

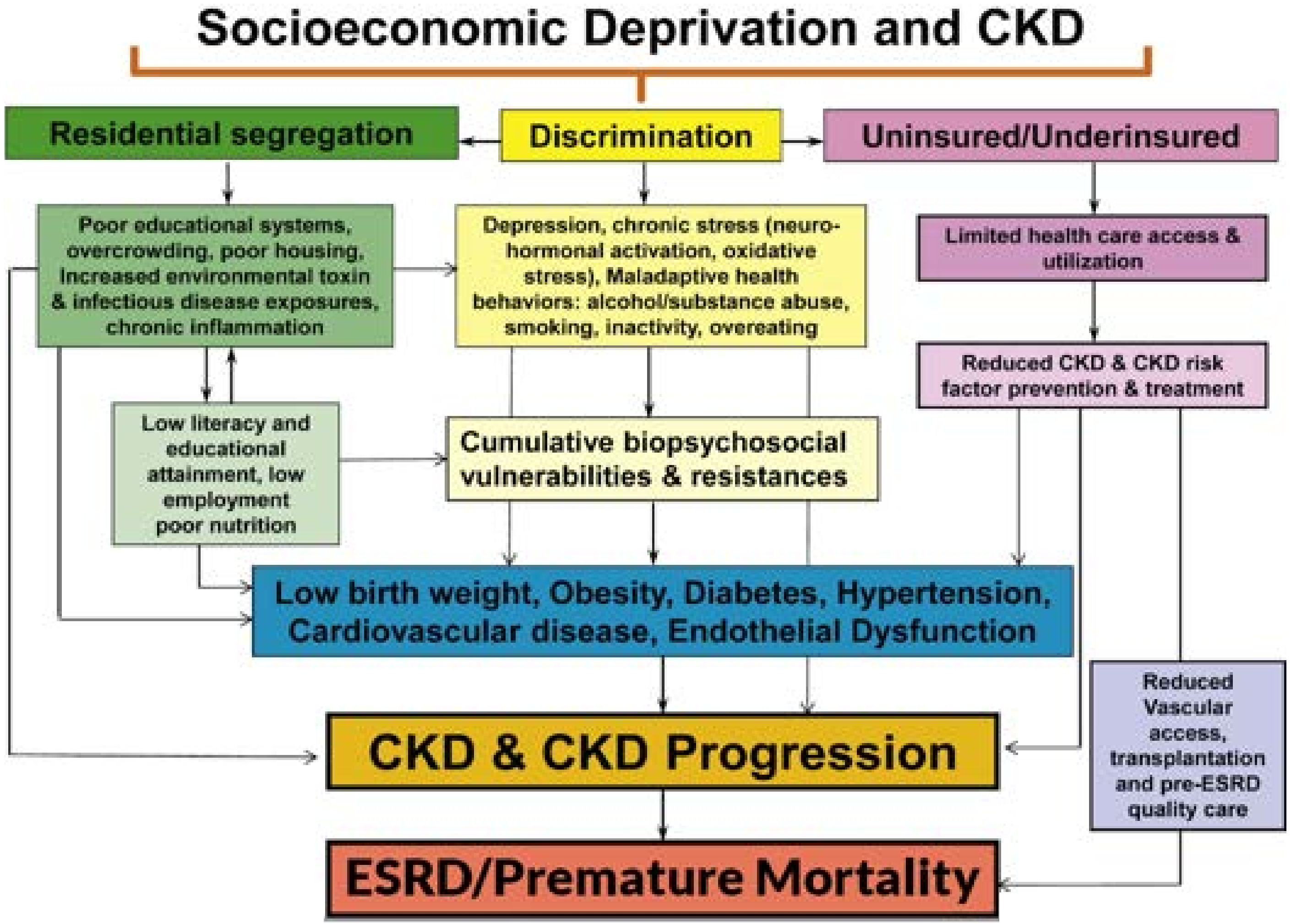
POTENTIAL DATA SOURCES
EHR data from any health system (e.g., VA, Optum)
Health claims data from Medicare/Medicaid and Payers
Vital statistics databases
Disease registries (e.g., USRDS, SEER)

The Office of the National Coordinator for
Health Information Technology

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

Bias (socioeconomic)

- Many of the determinants of chronic kidney disease, such as obesity, diabetes, hypertension, chronic inflammation, neurohormonal activation, and oxidative stress may be related to socioeconomic disparities.
- Factors include substandard living conditions, limited quality health care to the uninsured or underinsured, and limited health literacy.

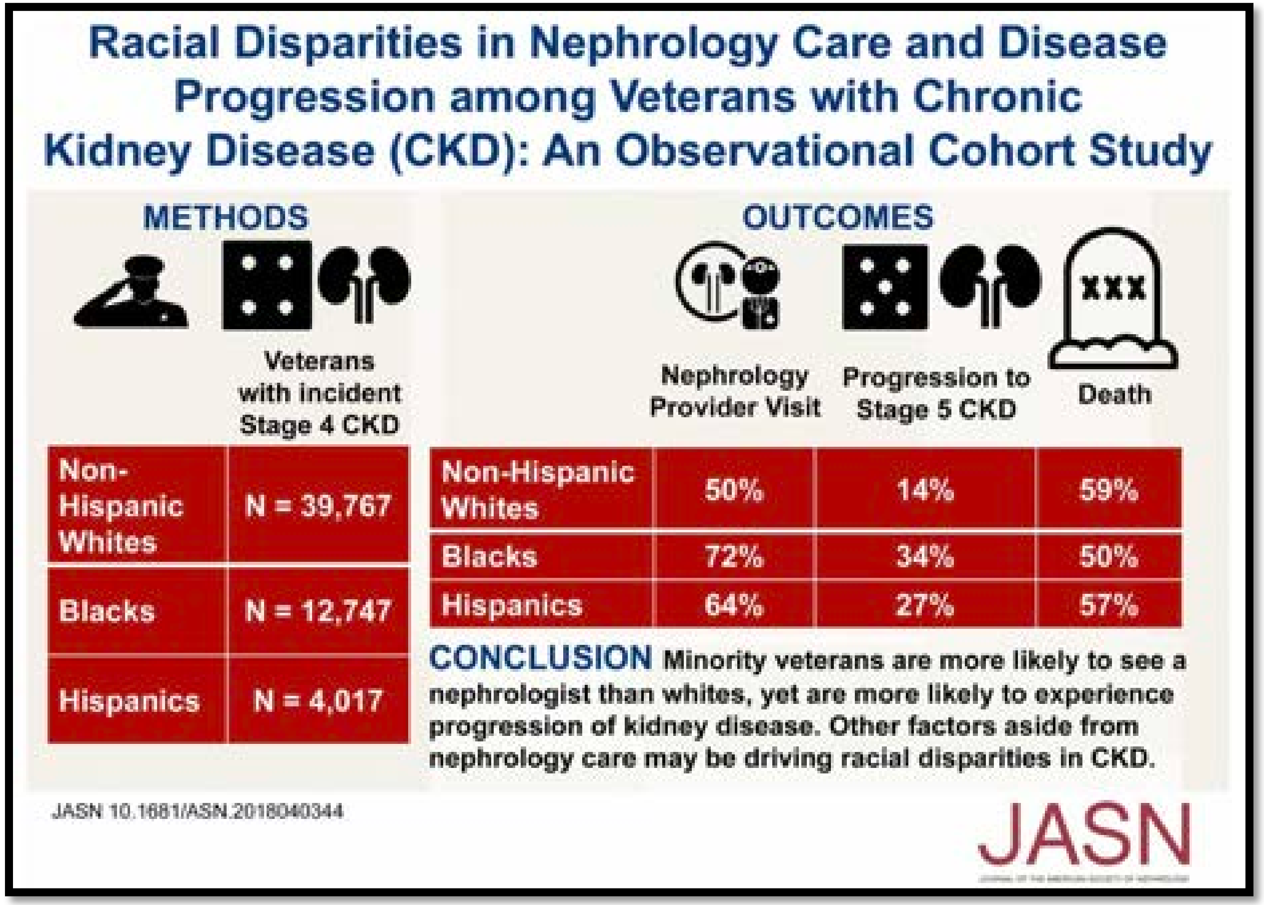


Socioeconomic Deprivation and CKD
Residential segregation
Discrimination
Uninsured/Underinsured
Poor educational systems, overcrowding, poor housing, Increased environmental toxin & infectious disease exposures, chronic inflammation
Depression, chronic stress (neuro-hormonal activation, oxidative stress), Maladaptive health behaviors: alcohol/substance abuse, smoking, inactivity, overeating
Limited health care access & utilization
Reduced CKD & CKD risk factor prevention & treatment
Low literacy and educational attainment, low employment poor nutrition
Cumulative biopsychosocial vulnerabilities & resistances
Low birth weight, Obesity, Diabetes, Hypertension, Cardiovascular disease, Endothelial Dysfunction
Reduced Vascular access, transplantation and pre-ESRD quality care
CKD & CKD Progression
ESRD/Premature Mortality

Source: https://www.sciencedirect.com/science/article/pii/S1548559514001086

Bias (Racial)

- Despite being more likely to receive nephrology consultation, black patients with stage 4 chronic kidney disease (CKD) were 62% more likely to develop end-stage renal disease (ESRD) after adjustment for comorbidities and socioeconomic factors.
- These findings suggest that biologic or environmental factors drive ESRD progression through mechanisms that nephrologists cannot currently treat.



Racial Disparities in Nephrology Care and Disease Progression among Veterans with Chronic Kidney Disease (CKD): An Observational Cohort Study

METHODS

	Veterans with incident Stage 4 CKD
Non-Hispanic Whites	N = 39,767
Blacks	N = 12,747
Hispanics	N = 4,017

OUTCOMES

	Nephrology Provider Visit	Progression to Stage 5 CKD	Death
Non-Hispanic Whites	50%	14%	59%
Blacks	72%	34%	50%
Hispanics	64%	27%	57%

CONCLUSION Minority veterans are more likely to see a nephrologist than whites, yet are more likely to experience progression of kidney disease. Other factors aside from nephrology care may be driving racial disparities in CKD.

JASN 10.1681/ASN.2018040344

JASN

Source: https://jasn.asnjournals.org/content/29/10/2563

Bias in AI

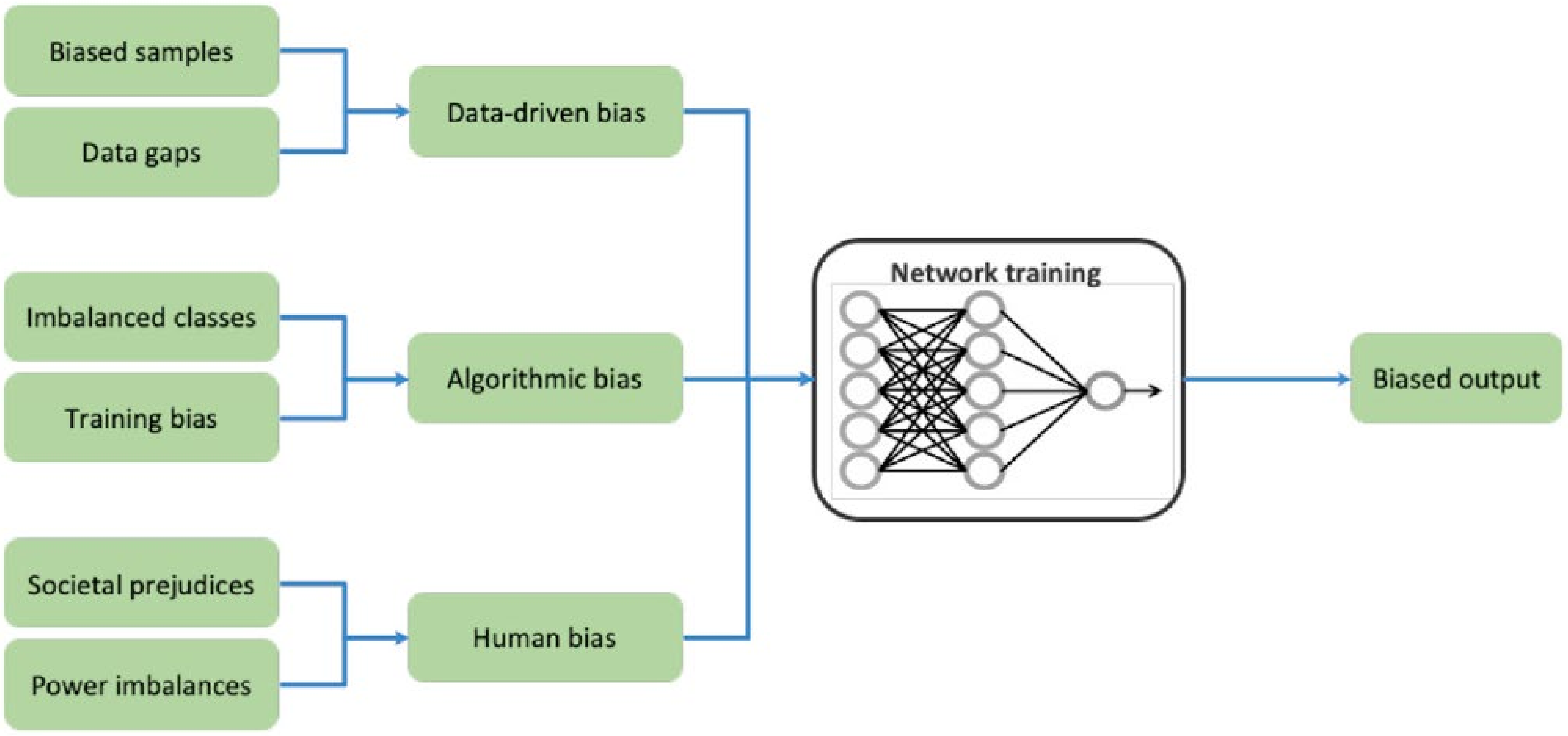
- Advances in AI offer the potential to provide personalized care by taking into account individual differences[1]
- **At the same time, because machine learning algorithms aggregate and assess large volumes of real-world data, AI can reinforce bias in data, potentially reinforcing existing patterns of discrimination**
- Machine learning algorithms may work well for one patient group, but results may not be appropriate for others

SOURCES OF BIAS

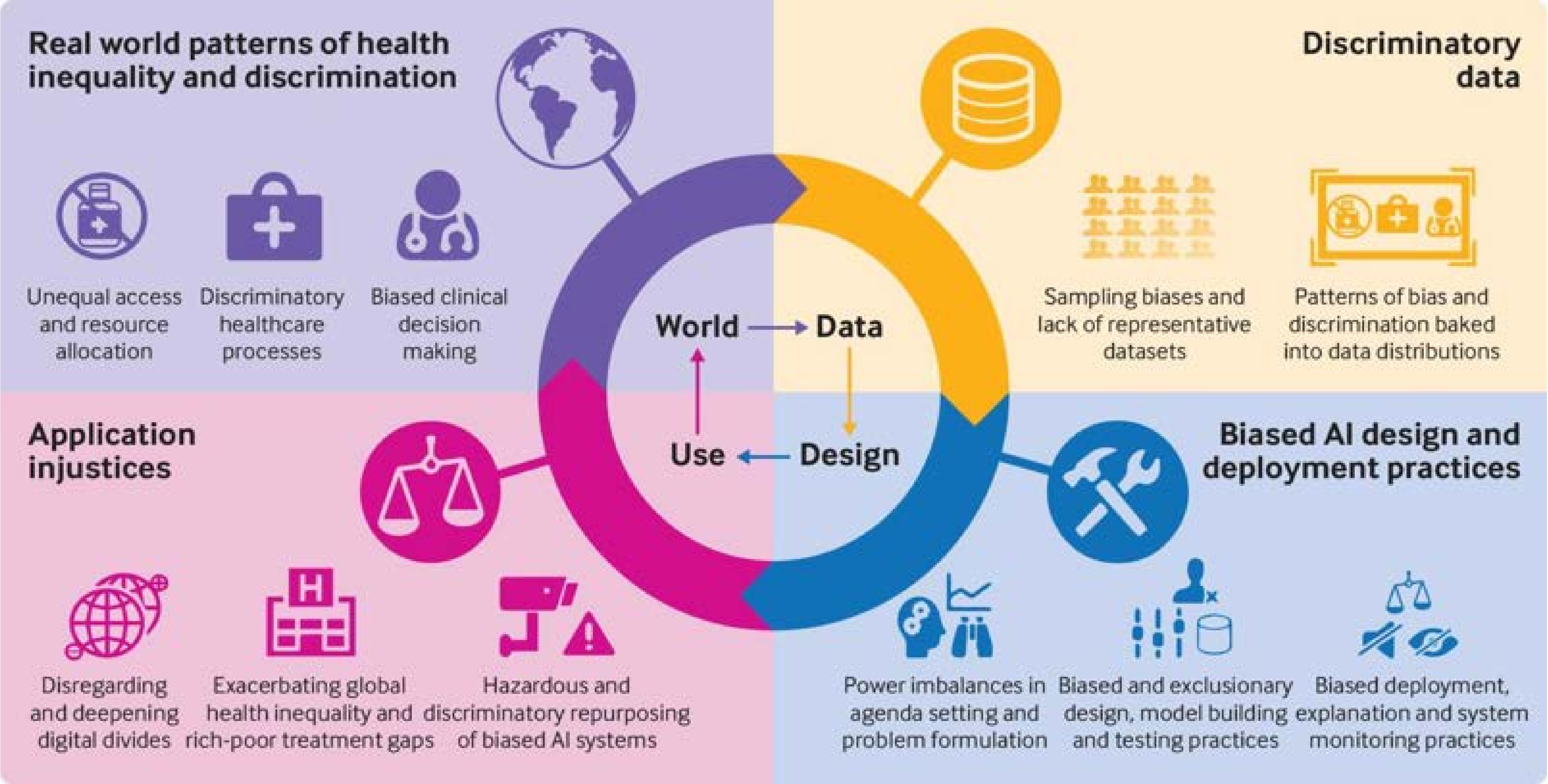
- Missing data – patients without consistent care at a single institution and/or lower health literacy
- Sample size – certain subgroups of patients may not exist in sufficient numbers, leading to uninformative predictions
- Misclassification or measurement error – implicit bias leads to disparities in care, teaching clinics (where patients of low socioeconomic status may be seen) may have less accurate data input[2]

Sources: 1. https://journalofethics.ama-assn.org/article/can-ai-help-reduce-disparities-general-medical-and-mental-health-care/2019-02; 2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/#!po=15.6250

Algorithmic racial bias mechanisms



The big picture


Real world patterns of health inequality and discrimination
Unequal access and resource allocation
Discriminatory healthcare processes
Biased clinical decision making
Discriminatory data
Sampling biases and lack of representative datasets
Patterns of bias and discrimination baked into data distributions
World
Data
Use
Design
Application injustices
Biased AI design and deployment practices
Disregarding and deepening digital divides
Exacerbating global health inequality and rich-poor treatment gaps
Hazardous and discriminatory repurposing of biased AI systems
Power imbalances in agenda setting and problem formulation
Biased and exclusionary design, model building and testing practices
Biased deployment, explanation and system monitoring practices

Example 1: Algorithm favors healthier white patients over sicker black patients

The issue

An algorithm used to predict which patients would benefit from extra medical care **flagged healthier white patients as more at risk than sicker black patients**

- An analysis on 3.7 million patients found that **black patients ranked as equally as in need of extra care** as white patients collectively suffered from 48,772 additional chronic diseases

- The bias was discovered when researchers from a health system in Massachusetts found the **highest scores in their patient population concentrated in the most affluent suburbs of Boston**

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. doi:10.1126/science.aax2342

Example 1: Algorithm favors healthier white patients over sicker black patients

The cause

- **The algorithm used a seemingly race-blind metric**: how much patients would cost the health-care system in the future

- **Cost isn't a race-neutral measure of health-care need**: unequal access to care means that we spend less money caring for black patients than for white patients

The solution

- **Researchers tweaked the algorithm** to make predictions about their future health conditions

- The tweak increased the percentage of black patients receiving additional help from 17.7 to 46.5%

Example 2: Flawed racial adjustments in kidney function estimates

- **Race forms part of the algorithms used to assess kidney function through an eGFR equation** that uses serum creatinine measurement, age, sex, race, body weight

- The inclusion of a **coefficient for black patients** in the eGFR equation was based on small poor-quality studies. The more accurate **CKD-EPI equation** still contains a correction for black patients.

The issue

The CKD-EPI equation modifier **increases eGFR for black individuals by nearly 16%**, altering guideline-based diagnoses and referrals for care

Diao JA, Wu GJ, Taylor HA, et al. Clinical Implications of Removing Race From Estimates of Kidney Function. JAMA. 2021;325(2):184-186. doi:10.1001/jama.2020.22124

Example 2: Flawed racial adjustments in kidney function estimates

The cause

Including adjustment for race in these eGFR equations **ignores the substantial diversity within self-identified black patients and other racial or ethnic minority groups**.

The solution

- Healthcare organizations have started **removing the race-based adjustment from the eGFR equation**, reporting the "White/Other" value for all patients.
- This measure may **increase CKD diagnoses among black adults** and enhance access to specialist care, medical nutrition therapy, kidney disease education, and kidney transplantation.

Example 3: AI-driven dermatology leaves dark-skinned patients behind

- Machine Learning has been used to create **programs capable of distinguishing between images of benign and malignant moles** with accuracy similar to that of board-certified dermatologists.

- However, the algorithms used by most healthcare organizations are basing most of their knowledge on ISIC, an open-source repository of **skin images from primarily fair-skinned populations.**

Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol. 2018;154(11):1247. doi:10.1001/jamadermatol.2018.2348

The issue

Lesions on patients of color are less likely to be diagnosed. The algorithms provide advancement for the Caucasian population, which already has the highest survival rate.

Example 3: AI-driven dermatology leaves dark-skinned patients behind

The cause

Bias emanates from unrepresentative training data that reflects historical inequalities: decades of clinical research have focused primarily on people with light skin.

The solution

- Researchers are taking measures to ensure a **more equitable demographic participation in clinical trials.**
- ISIC is looking to **expand its archive to include as many skin types as possible,** and has asked dermatologists to contribute photos of lesions on their patients with darker skin.

Bias in AI

POTENTIAL CHALLENGES	RECOMMENDED SOLUTIONS
Data diversity due to limited population representation	• Assess the limitations • Identify the strategy for mitigating a lack of diversity as part of the research design
Overreliance on machine learning solutions	• Ensure interdisciplinary approach and continuous human involvement • Conduct follow-up studies to ensure results are meaningful
Algorithms based on biased data	• Identify the target population and select training and testing sets accordingly • Build and test algorithms in socioeconomically diverse health care systems • Ensure that key variables that are related to race, gender, etc. are being captured and included in algorithms where appropriate • Test algorithms for potential discriminatory behavior throughout processing • Develop feedback loops to monitor and verify output and validity
Non-clinically meaningful algorithms	• Focus on clinically important improvements in relevant outcomes rather than strict performance measures • Impose human values in algorithms at the cost of efficiency

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/#!po=15.6250

Bias in AI

- Preventing algorithms from making biased decisions is challenging and there is often a tradeoff between fairness and accuracy
- Three main strategies for reducing bias:
  - Eliminating sources of unfairness in the data before training a machine learning algorithm
  - Making fairness adjustments as part of the process by which the algorithm is constructed
  - Adjusting performance after an algorithm is applied to make it fairer

WHY IS IT SO DIFFICULT TO ELIMIATE UNFAIRNESS?

- There is a lack of agreement among researchers about which definition of fairness is the most appropriate[1]
- Removing sensitive information from data, such as race, age, and gender, may not result in unbiased outcomes since non-sensitive attributes and outcome variables are often statistically dependent on sensitive information[2,3,4]
- A user's judgment about a model feature may change after learning how the use of the feature impacts decision outcomes[5]

Sources: 1. https://journalofethics.ama-assn.org/article/can-ai-help-reduce-disparities-general-medical-and-mental-health-care/2019-02; 2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/#!po=15.6250

FAIR + CARE

Data that is

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

used for

- the **C**ollective benefit of those from which data was collected
- whose populations maintain **A**uthority to control the data
- data collectors have a **R**esponsibility to interact with minoritized populations respectfully
- the **E**thics of populations from which data was collected are respected

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/; https://www.slideshare.net/carolegoble/building-the-fair-research-commons-a-data-driven-society-of-scientists

Slido Quiz

1. Select a potential source of bias for an electronic health record data set.
   a) Sample size (not enough representation of all subgroups)
   b) Measurement error
   c) Equipment choice
   d) Missing context
   e) All of the above
2. The type of model that can be used if you have a set of labeled data
   a) Unsupervised Learning
   b) Supervised Learning
   c) Independent Learning
   d) Observation Learning
3. An AI-ready dataset does not need to be documented fully because the model will do it automatically.
   a) True
   b) False

Performing Pre-Model Processing and Data Quality Checks

Courtney Shelley, PhD, Booz Allen Hamilton

Importing Data

If working within a Jupyter notebook, save your data file in the same folder as your notebook.
If using R on your own computer,

```
getwd()        # will print out the current directory.
               # Save your data here or ….
setwd("…")     # point to the directory you want to use
list.files()   # shows all files within directory
```

Your data file will have a name and a file extension:

File Type	File Extension	library	function
Comma-Separated Values (i.e., CSV)	.csv	---*	read.csv()
Excel Spreadsheet	.xls or .xlsx	openxlsx	read.xlsx()

Files created for/within R
R objects
	.Rdata, .rda
	.rds
		---	load()
			readRDS()

Files created for/within SAS
Files created for/within STATA
Files created for/within SPSS
	.sas7bdat
	.dta
	.sav
		haven	read_sas()
			read_dta()
			read_spss()

Space-delimited,
Tab-delimited
	.dat	---*	read.table(" ", sep = " ")
			read.table(" ", sep = "\t")

Fixed-width	.txt	---*	Read.fwf()

Importing Data

- Installing packages:
    - Only have to do this ONCE

```
install.packages("package-name")   # Note the double quotations!
```

- To use a package:
    - Will need to do this once every session

```
library(package-name)                 # No quotations here
```

- Putting it all together -

```
install.packages("openxlsx")
library(openxlsx)
df <- read.csv("data.xlsx", header = TRUE)
```

Understanding the Data

- Check you read it in correctly
- Get a feel for its size and complexity

```
>   head(df)    # first five rows of dataset. Does this look as expected?
>   dim(df)     # dimensions of dataset in rows, columns
>   names(df)   # column names
```

Understanding the Data

- Assess Variable Types:
  - Each value is an **element**
  - Data types in R include: character, numeric, integer, complex, and logical
    - Integer (denoted with an L, rarely used in health science)
    - Complex (<*REAL*> + <*IMAGINARY*>I, also rarely used in health science)
    - Numeric are any numbers, including negative and decimal values
    - Logical is TRUE/FALSE
    - Character is “Latino”, “always”, “California”

```
>   class(6)
[1] “numeric”
>   class(TRUE)
[1] “logical”
>   class(“friend”)
[1] “character”
```

Understanding the Data

- R can handle more complexity also, including vectors, matrices, data frames, and lists. These **objects** are composed of elements:
  - A vector is made with the concatenate function, `c()`:

    ```
    >  c(“red”, “yellow”, “green”)
    >  c(1, 2, 87)
    ```

  - A matrix is made of numeric elements:

    ```
    >  matrix(1:25, nrow = 5)
    ```

  - A data frame is made of many vectors of the same length:

    ```
    >  data.frame(color = c(“red”, “yellow”, “green”),
    +             age = c(1, 2, 87))
    ```

Understanding the Data

1. All elements of a vector must be the same type.
2. R uses a process called **coercion** to attempt to make sense of input.

★ THEREFORE: If you create a vector like c("red", 1, TRUE), R will coerce the elements to all be the same type. In this case, it will force 1 and TRUE to also be characters.

```
>  c("red", 1, TRUE)
[1] "red"  "1"    "TRUE"
```

- ***Why do you care?*** Because a typo will coerce a vector to become a different type than you expect.
    - Ex: A typo such as 4. when you meant 4.0 will be read as a character, so that the entire vector will then be coerced to character. Now it won't behave as a numeric vector when you try to analyze it - you won't be able to find min/max or do math.

- ***Always ask yourself - what class do I expect from this variable? If it doesn't look how you expected it to, check for errors!***

Slido Quiz

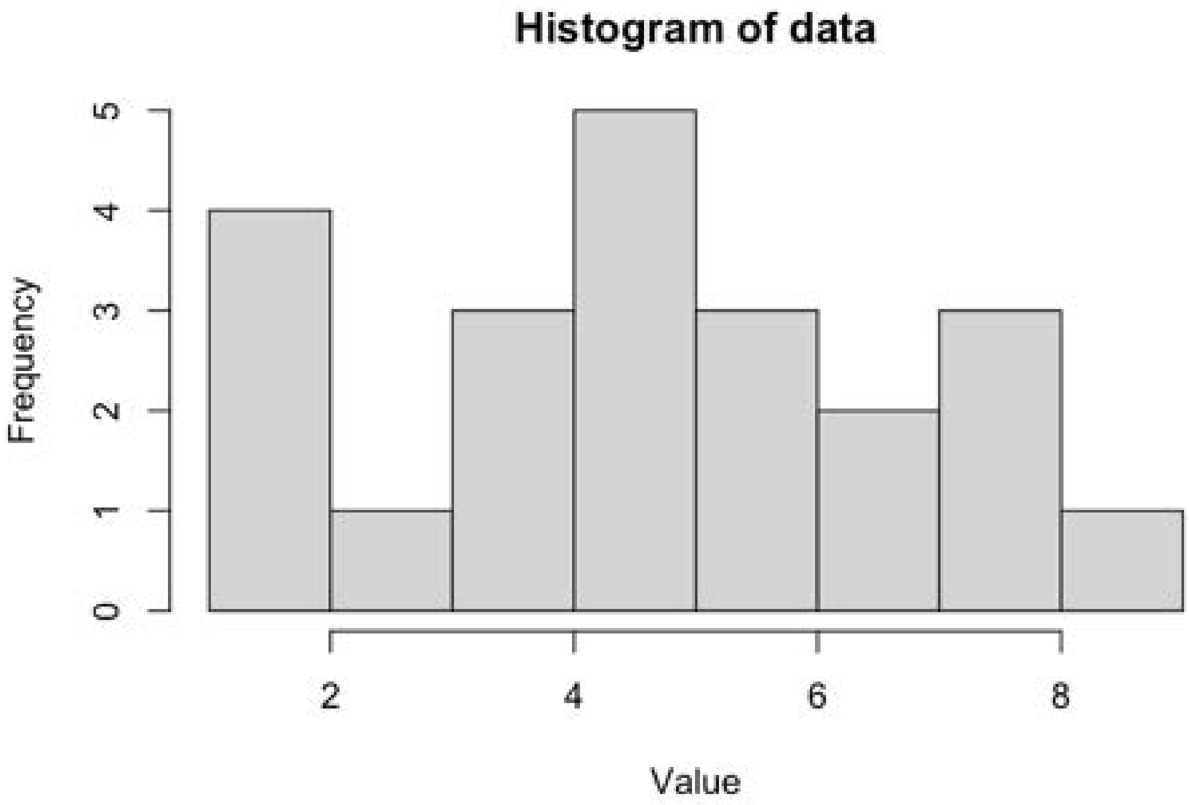
Why do you want to use a programming language like R or Python or SAS, rather than spreadsheet software like Excel for data exploration and analysis?

Visualizing Data

HISTOGRAMS

- Suitable for numeric data with (at least theoretically) continuous values.
- Creates a specified number of bars representing a value range with height equal to the number of observations within that range.

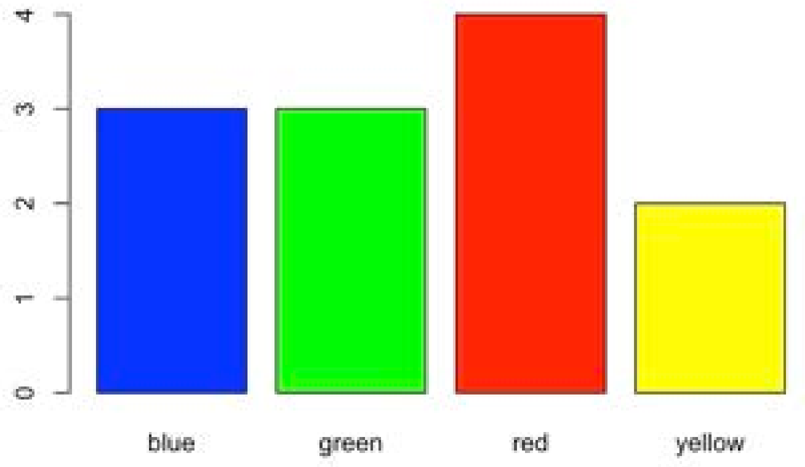
```
data = c(1,1,4,5,8,3,5,7,9,1,7,
             1,4,5,6,8,6,5,4,5,6,8)

hist(data)
```



Visualizing the Data

BARPLOTS

- Suitable for categorical (i.e., count) data.
- Generates a bar for each category with height representing the number of observations within each category.



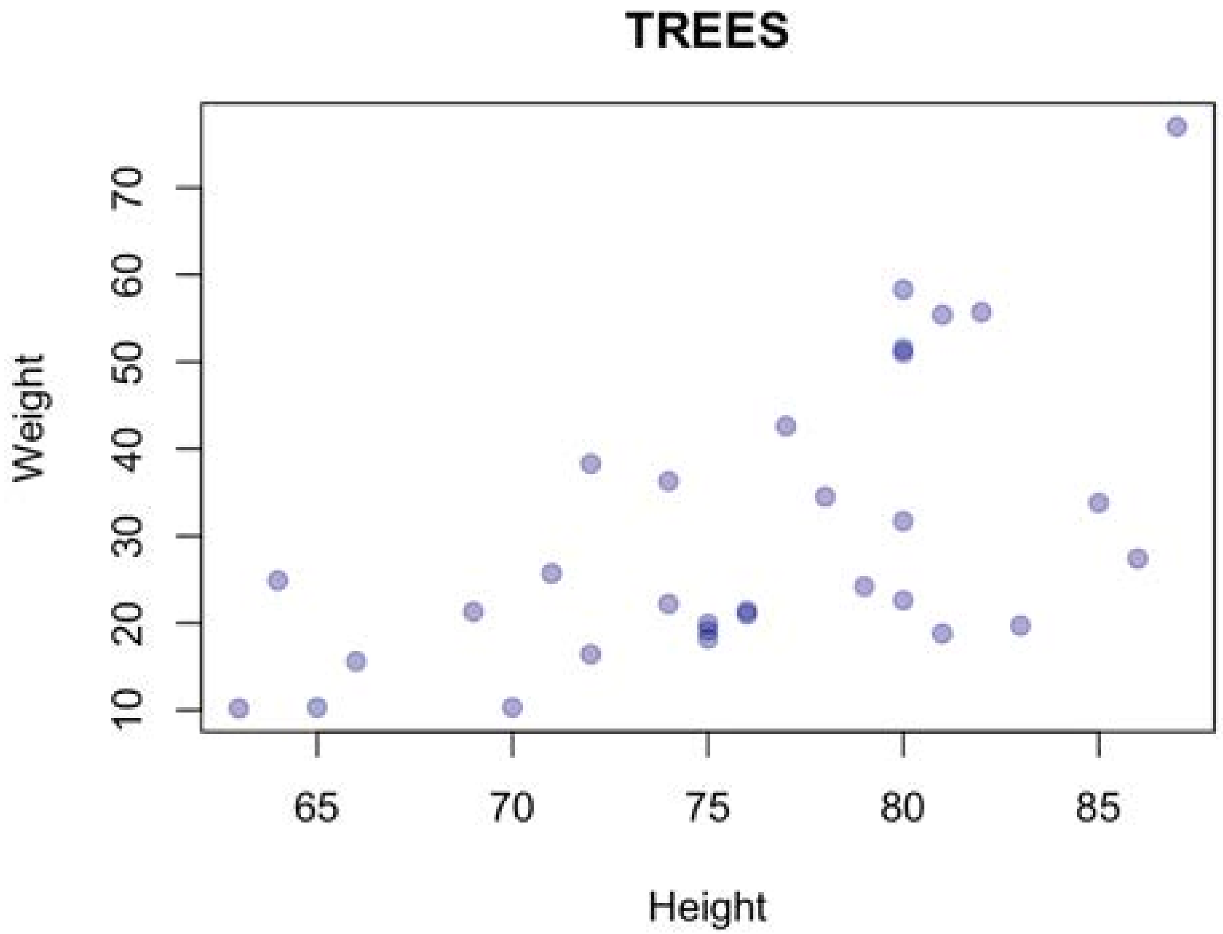
```
data = c("red", "yellow", "green", "red",
         "red", "blue", "blue",
"yellow",             "red", "green",
"green", "blue")
table(data)
data
   blue  green    red yellow
      3      3      4      2

barplot(table(data), col = c("blue",
"green", "red", "yellow"))
```

Visualizing the Data

SCATTERPLOTS

- Suitable for viewing the relationship between two continuous variables
- Most often plotted with the independent variable on the x-axis and the dependent (response) variable on the y-axis

```
plot(Height, Volume, data = TREES)
```



Slido Quiz

I have 20 observations on patients' reported gender, sex assigned at birth, height and weight. The data is in a .xlsx format, so I can open it in Excel.

1. What do you expect the class of gender, sex, height, weight to be?
   a) "category", "category", "number", "number"
   b) "factor", "factor", "integer", "integer"
   c) "character", "character", "numeric", "numeric"
2. Which two variables can I visualize together using a scatterplot?
   a) gender and sex
   b) sex and weight
   c) height and weight

Slido Quiz

3. In R, I ran hist(weight) and received the following error:

```
Error in hist.default(data$weight) : ‘x’ must be numeric
```

What is wrong and what could I do to investigate?

Data Cleaning and Pre-Processing

TYPOS

- A common place we see typos is in self-identified categories.
- Sometimes datasets will use codes rather than the actual labels so that these variables look and behave like numeric count data instead of categorical data:
  - When Race is input as 1 = White, 2 = Black, 3 = Hispanic/Latino, 4 = Other, we can calculate mean(Race), but what does that mean?
  - R can handle **factors** so that categorical names will be treated as labels.
  - **BUT...** if these same values are input by name, we tend to see "white", "White", "WHITE", "hisp", "hispanic", "Latin**a**", which creates all kinds of categories with the same meaning!
- The function `table()` is very helpful because it will work on both numeric data and character data.
- Recoding these isn't too hard – just pick a standard and stick to it! *(we recommend referring to a standard ontology like SNOMED).*

Data Cleaning and Pre-Processing

TYPOS

- Typos may also look like ***clinically implausible values*** or extreme outliers. Assessing `min()` and `max()` for numeric values will often find these.
    - Example: Age of 250 instead of 25 or a height of 5.4 cm that is probably 5.4 ft.
- When handling medical data, you'll likely need to speak with a clinician to understand the clinically plausible values of laboratory measurements, as well as understanding what the laboratory cut-off values may be (such as all readings over 1,000 recorded as 1,000, which will look like a normal curve with its tail cut off).

Data Cleaning and Pre-Processing

MISSINGNESS

- May occur in different ways:
  - **Missing Completely At Random (MCAR)**: the fact that data is missing is independent of the data value itself, such that there is no systematic difference between those with missingness and those without, such as a batch of laboratory samples processed improperly that result in missing laboratory values.
  - **Missing At Random (MAR)**: Missingness is systematically related to ***observed*** data, such as male patients being less likely to answer survey questions about depression. Here the probability of completing the survey is due to being male, not the severity of the depression. This can be seen in medical studies if patients cannot get time off work to attend follow-up visits.
  - **Missing Not At Random (MNAR)**: Similar to MAR, but missingness is systematically related to ***unobserved*** data, such as not answering the survey about depression based on severity of depression. This can be seen in medical studies if patients are too ill to attend follow-up visits

Handling missingness depends on the type of missingness observed!

Data Cleaning and Pre-Processing

- Regrouping categories (such as combining two variables of RACE and ETHNICITY into a single RACE_ETHNICITY variable) can help reduce missingness, though will also reduce information in the data as data handling decisions are made:

RACE	ETHNICITY
White	Hispanic/Latino
Black	Not Hispanic/Latino
Asian	
Other	

RACE_ETHNICITY
NonHispanic_White
Hispanic_White
NonHispanic_Black
Hispanic_Black
NonHispanic_Asian
Hispanic_Asian
NonHispanic_Other
Hispanic_Other

- It's up to you to assess that this regrouping is "better". Statistical tests can assess whether those who did not answer one question were more likely to also not answer the other question, or that those who did not answer these questions did not also systematically differ in other measured ways.

- *But how would you know if they differed in unmeasured ways?*

Data Cleaning and Pre-Processing

MISSINGNESS

- A common approach to handling missingness is to delete those rows or columns that contain missingness. But this will also necessarily reduce sample size or potential explanatory features.
  - When deleting rows (i.e., observations or patients), ensure those with missingness are not systematically different from those with complete information.
  - If feasible, a new variable can be created that indicates missing data such that models can be fitted with this variable that omit the observation from calculations but can show a signal for missingness itself. This is how censoring works in time-to-event analysis.
  - When deleting columns (i.e., variables or features), use a systematic rule such as the column must contain at least 5% filled cells.

Data Cleaning and Pre-Processing

- Combining related columns may help – sort of the reverse of one-hot encoding or indicator (dummy) variables that we'll discuss later.

Do you feel anxious about your diagnosis?	**Do you feel sad about your diagnosis?**	**Do you feel accepting of your diagnosis?**
Yes	***NA***	No
No	Yes	***NA***
No	Yes	***NA***
Yes	No	***NA***
NA	Yes	***NA***
NA	No	***NA***
NA	***NA***	Yes

How do you feel about your diagnosis?
1 = Anxious
2 = Sad
2 = Sad
1 = Anxious
2 = Sad
4 = Other/NA
5 = Accepting

  - Another solution is to create a variable that systematically accounts for missingness such as "Did Patient answer any of Questions 9-12?" Now those who did not answer can be "No" instead of "NA" for four responses.

Data Cleaning and Pre-Processing

- Imputation is a method for predicting values based on surrounding data. If data is MCAR, missing values can be inferred from the complete dataset. If data is MAR, a predictive model fit using only similar participants can predict missing values (in the depression and gender example, predict missing male values from males who did answer). If data is MNAR, imputation is more complex but still possible.

Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017 Mar 15;9:157-166. doi: 10.2147/CLEP.S129785. PMID: 28352203; PMCID: PMC5358992.

Slido Quiz

- In our last quiz example, we encountered an issue with recorded weights. The error caused `hist(weight)` to not run, with an error message that our data was not numeric class. This was probably a typo that caused one value to be character rather than numeric, which coerced the entire vector to character class.

- The data is read into R and stored as an objected called `study`. I viewed `study$weight` and observed the following:

```
[1]  "118.4" "164"   "191.9" "149.2" "156.9" "189.5" "146.2" "135.3" "165.3" "121"   "179.4" "151.2"
[13] "136."  "136.7" "162.1" "164.6" "121.1" "137.8" "149.9" "120.8"
```

What can be done to fix this typo?

Data Cleaning and Pre-Processing

DATES

- Dates are also an element class in R.
- Dates may be recorded in separate columns of DAY, MONTH, and YEAR or may be recorded in single columns but with differing formats, such as “19 AUG 2023”, "2023-11-07", or “2023/11/07”.
- In R, all date types can be handled using `as.Date()`

```
as.Date('1/15/2001',format='%m/%d/%Y')
[1] "2001-01-15"
> as.Date('April 26, 2001',format='%B %d, %Y')
[1] "2001-04-26"
> as.Date('22JUN01',format='%d%b%y')    # %y is system-
specific; use with caution
[1] "2001-06-22"
```

Data Cleaning and Pre-Processing

- If in a three-column form, you will first need to collapse columns of DAY, MONTH, YEAR into a single column

```
dates <- paste(DAY, MONTH, YEAR, sep = "/")
as.Date(dates, "%d/%b/%y")
```

You'll probably have to play around for a while to ensure your code does what you want it to do.

Be sure to check simple cases first and create a new variable rather than altering your data!

Data Cleaning and Pre-Processing

CREATING NEW VARIABLES

- Categorical variables will most likely be treated separately in any statistical analysis you do. If you have a categorical variable of car color, one category (or level or factor) will be treated as the referent category and all other categories will compare to this one. This is achieved using one-hot encoding or indicator variables (i.e., dummy variables). For a categorical variable with n categories, n – 1 new variables will be created and filled with 0/1 to indicate whether or not each observation is in each category.

COLOR
White
Blue
Red
Black

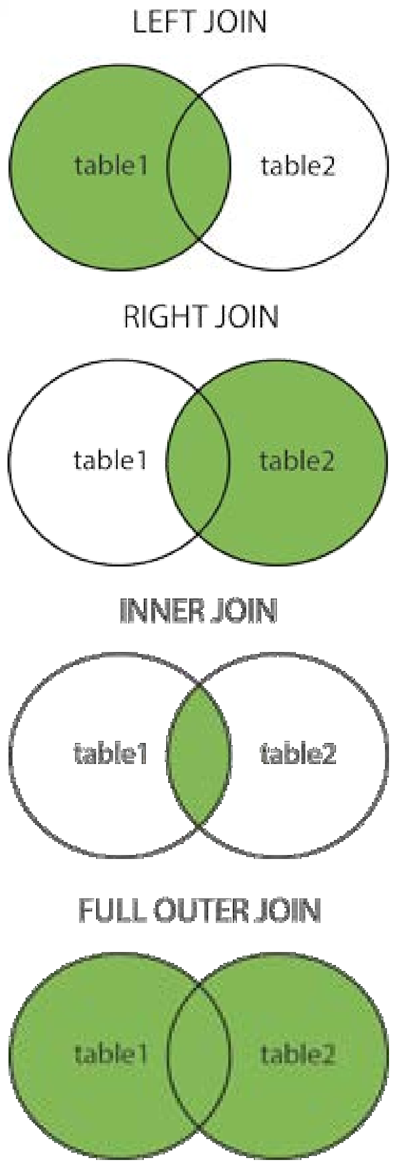
COLOR_BLUE	COLOR_RED	COLOR_BLACK
0	0	0
1	0	0
0	1	0
0	0	1

- Here white is the reference color such that white is inferred by not being blue, red, or black. A model fit with these values will estimate differences from white car color.

Data Handling and Pre-Processing

MERGING FILES

- When working with files from a database, you will likely need to merge separate files on the same patients to achieve a single tabular (i.e., spreadsheet) dataset. Merging can be achieved with several R packages and differing techniques.
- Two easy to use methods for two datasets, df1 and df2:
  - **Base R's** `merge()` function: `merge(df1, df2)`. Arguments within this function can specify which columns to merge if the names differ, and whether you want to keep all rows or only those with a match in both datasets.
  - **dplyr's** `join()` family of functions. The dplyr package uses SQL database syntax.
    - A *left join* means: Include everything on the left (what was the df1 in `merge()`) and all rows that match from the right (df2) data frame. If the join columns have the same name, all you need is `left_join(df1, df2)`. If they don't have the same name, you need a by argument, such as `left_join(x, y, by = c("df1ColName" = "df2ColName"))`.
    - There is also `right_join()`, `inner_join()`, and `full_join()`.



Harmonization and Fusion

HARMONIZATION

- Similar datasets collected over several years or in several locations under the same study design can be combined through ***data harmonization***.
  - This may require renaming columns so they match:
    - `date_diagnosed` and `date_of_diagnosis`, or creation of `date_diagnosed` by combining `day_diagnosed, month_diagnosed`, and `year_diagnosed`
  - When combining studies over several years or locations, create a new column of `YEAR` or `LOCATION` so that the original datasets can be examined separately.
  - ***NOTE: Harmonized datasets should be VERY SIMILAR.***

Harmonization and Fusion

FUSION

- Data fusion is the process of combining multiple datasets to test hypotheses and find patterns that would not be testable in a single available dataset.
- Combining multiple datasets multiplies your potential for introducing bias.
  - One such bias is the **ecological fallacy** – making inferences about individuals based on aggregate data for a group. AI is very good at committing this one by inferring race based on home address and neighborhood characteristics, inferring sexuality based on social medial Likes
  - Another bias is reusing datasets that weren't collected for research purposes. Troublesome datasets will include data collected from cell phones or social media, which have known selection biases of age, urban/rural, and SES.
- Be careful about drawing causal conclusions from fused datasets. **Causal data fusion** is a branch of computational epidemiology with a growing body of theory that should be referenced.

Data Documentation

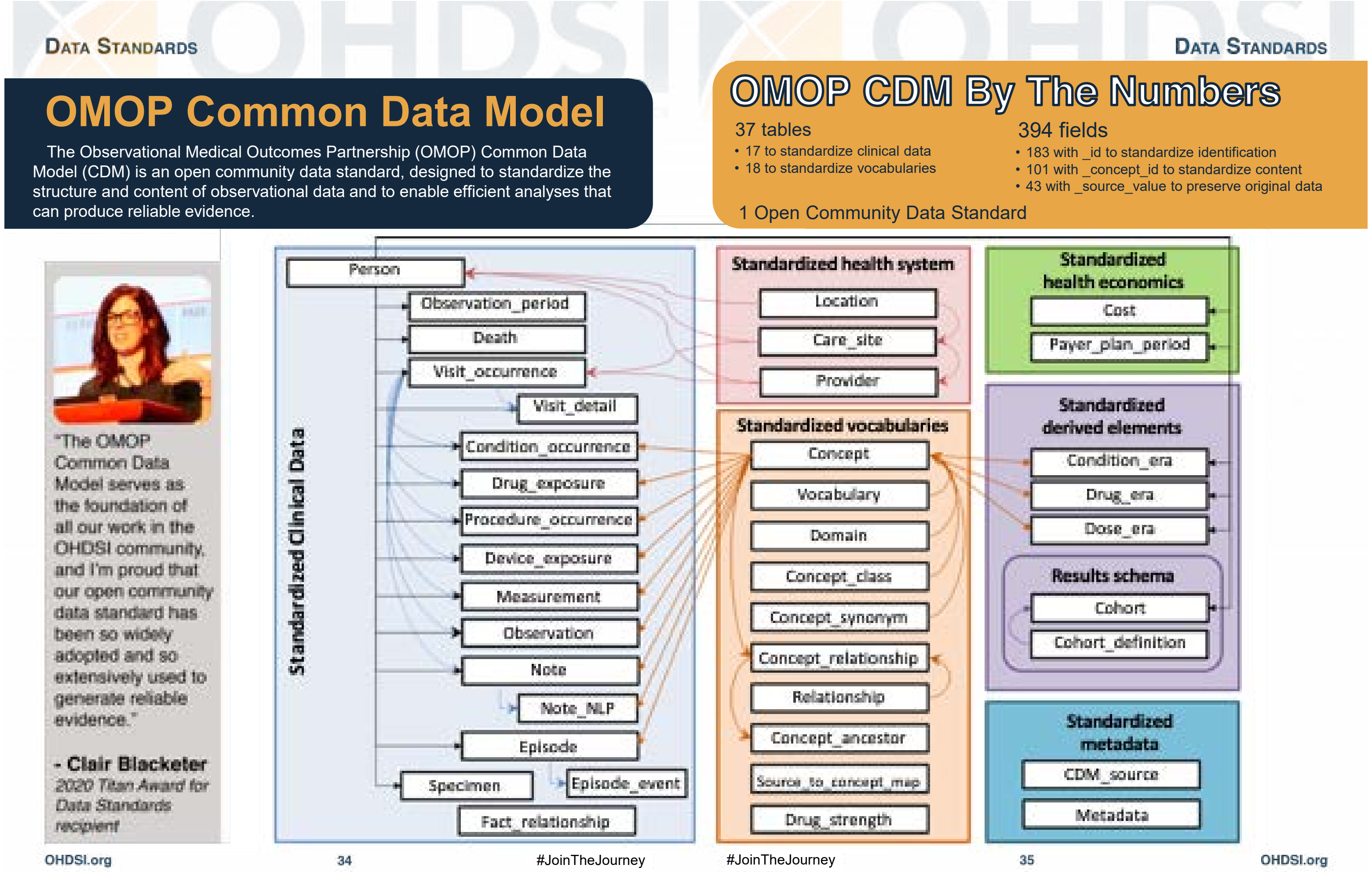
Now that we've done all these steps, we should have a clean and AI-ready dataset that requires accompanying documentation so others can properly use the data.

ELEMENTS OF A DATA DICTIONARY:

1. **Document the dataset itself** – the study or source of the data including details on inclusion/exclusion criteria, source population and target population, sampling schema used (if applicable). The goal is to not need to contact any original data creators for further details but to be able to successfully apply the data to a new application without introducing bias or non-portability
2. **Document the data elements** – again, the goal is for future users to not need to interact with the dataset creator (you!) so sufficient documentation *of each element* means:
   - a thorough description of what was collected, why, and how
   - variable type (e.g., numeric), unit of measurement (e.g., pounds, kilograms), and corresponding code lists (e.g., 0 = "No", 1 = "Yes")
   - summary statistics of each element (*N*, *N* missing, min, max, median, mean, and interquartile range of continuous numeric values or N, categories and *n* for each).
   - missingness notation and codes used for categories if these were used.
   - Along with above, document for each element what pre-processing steps were taken.

Data Models

- A Common Data Model (CDM) is a means to organize *data stored in a database* into a standard structure to facilitate interoperability with other systems.
- CDMs provide a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes) to standardized the data.
- The Observational Medical Outcomes Partnership (OMOP) CDM is used broadly across health domains to standardize the structure and content of observational data.



DATA STANDARDS

OMOP Common Data Model

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is an open community data standard, designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence.

"The OMOP Common Data Model serves as the foundation of all our work in the OHDSI community, and I'm proud that our open community data standard has been so widely adopted and so extensively used to generate reliable evidence."

- Clair Blacketer
2020 Titan Award for Data Standards recipient

DATA STANDARDS

OMOP CDM By The Numbers

37 tables
- 17 to standardize clinical data
- 18 to standardize vocabularies

394 fields
- 183 with _id to standardize identification
- 101 with _concept_id to standardize content
- 43 with _source_value to preserve original data

1 Open Community Data Standard

OHDSI.org 34 #JoinTheJourney #JoinTheJourney 35 OHDSI.org

https://www.ohdsi.org/data-standardization/

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Live Demo: Data Handling in Jupyter Notebook

Courtney Shelley, PhD, Booz Allen Hamilton

ScHARe

Demo Setup

We have registered you for ScHARe

You can choose not to use your account. If you prefer to be removed at any time,

email us at

schare@mail.nih.gov

With your consent, you have been:

- **registered for ScHARe**
- **added to a free temporary billing project** that will allow you to run the event materials with your instructors

➢ You will be active on this billing project for the duration of the Think-a-Thon. If you want to access work-in-progress after this time, you will need to set up your own billing and copy your workspaces to it

In preparation for the Think-a-Thon

We want to make sure that everyone:

1. has provided their Gmail address and has been registered for ScHARe, receiving a registration confirmation email
2. If not, please fill in this form: https://forms.office.com/g/7QybyjDjiw
3. can create and set up their Terra account with our help

The next two slides provide instructions on how to do so
for users who could not attend our Think-a-Thon today

Registering for ScHARe

Complete slides with **step-by-step instructions and screenshots** available at: bit.ly/think-a-thons

Normally, you would have to complete the following steps to register for ScHARe:

1. Visit the ScHARe portal on the NIMHD website:

   nimhd.nih.gov/schare
2. Click on the “Register for ScHARe” button
3. On the registration page, click on the “Register for ScHARe on Terra” button
4. Complete the registration form

Terra recommends using Chrome

The ScHARe team will:

- review and approve your application
- send you an email with additional instructions

➢ Note: you will need a **Gmail account** or another email account (an institutional email, for example) associated with a Google identity. If you do not have it, you can create one here:

bit.ly/3QeUngh

Creating a Terra account

Complete slides with **step-by-step instructions and screenshots** available at: bit.ly/think-a-thons

The email you will receive after ScHARe registration approval will ask you to **complete the following steps:**

1. Access the ScHARe Terra workspace at: bit.ly/access-schare
2. Click on the blue "Log in" button
3. Select "Sign in with Google"
4. Sign into Terra. Your username is the Google email address you provided to request access to ScHARe
5. Click "Next" and enter your Google account password to login
6. You will see a New User Registration page. Insert your name and contact email, then click on "Register"
7. Review and accept the Terra Terms of Service

➢ You will be taken to the ScHARe Terra Workspace: bit.ly/access-schare

Here you can click on the tabs at the top of the page (**Dashboard, Data, Analyses**, etc.) to explore the available resources

Workspaces are the building blocks of Terra - a dedicated space where you and your collaborators can access and organize the same data and tools and run analyses together

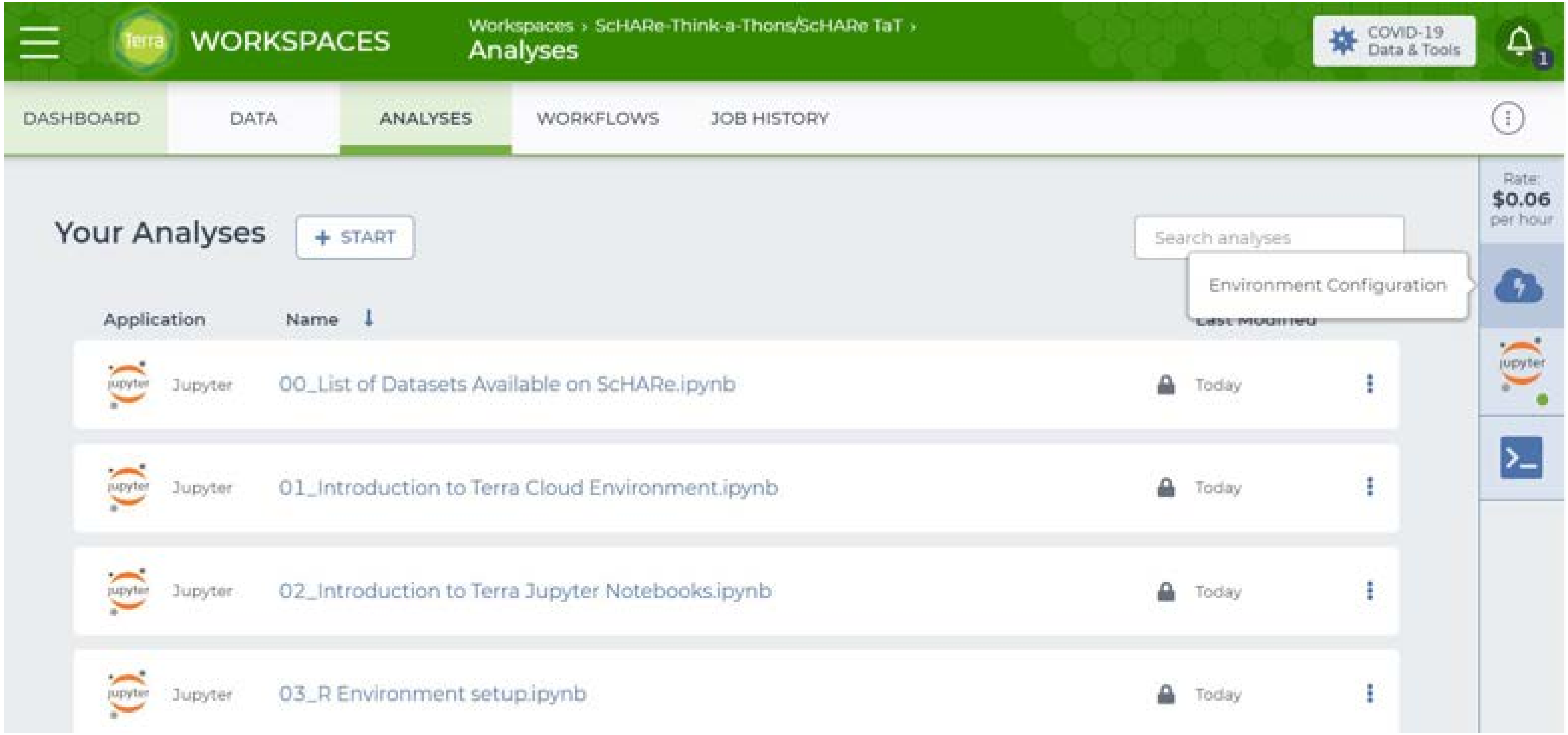
They are like **computational sandboxes** with everything you need to complete your project: data, analysis tools, documentation

Please paste this address in your browser:

bit.ly/schare-tat

If you have already created a Terra account and are logged in, you will see this:

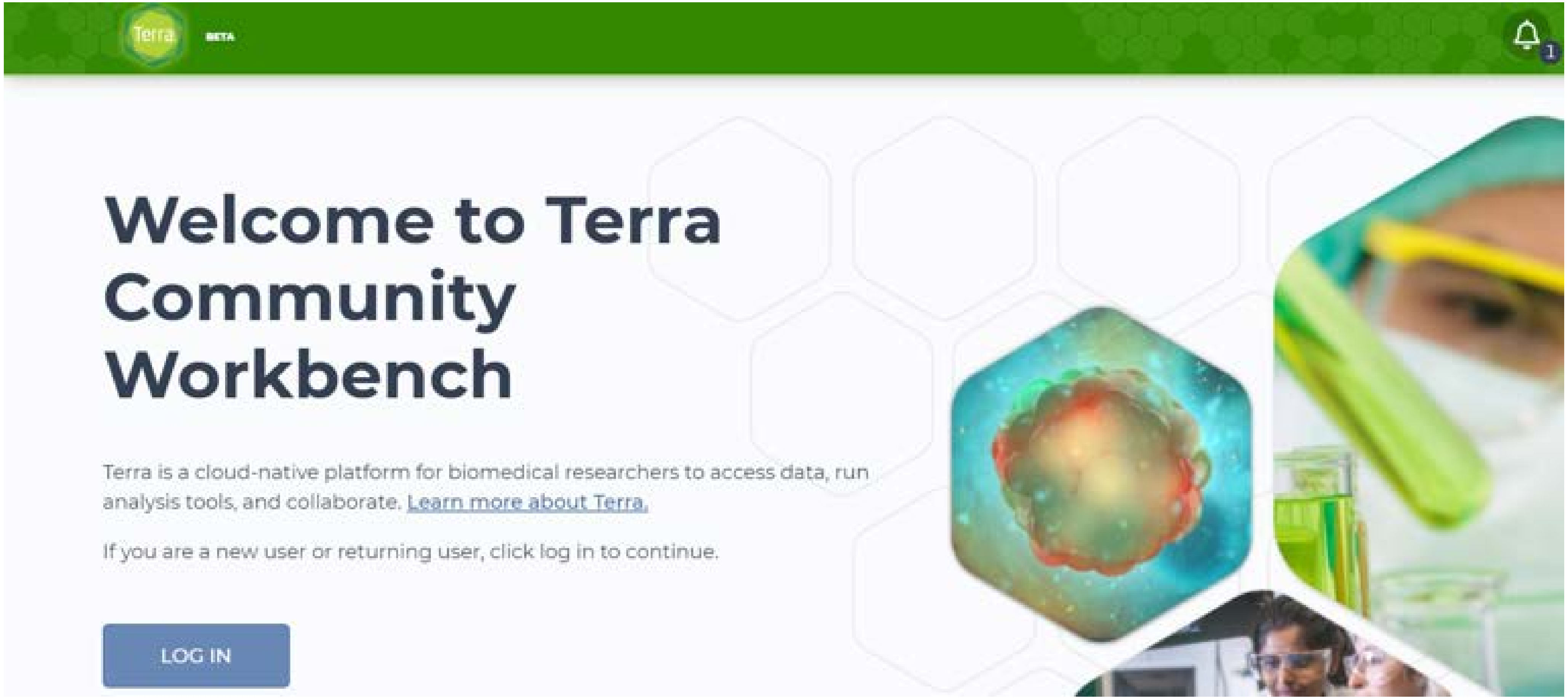
bit.ly/schare-tat



Application	Name	Last Modified
Jupyter	00_List of Datasets Available on ScHARe.ipynb	Today
Jupyter	01_Introduction to Terra Cloud Environment.ipynb	Today
Jupyter	02_Introduction to Terra Jupyter Notebooks.ipynb	Today
Jupyter	03_R Environment setup.ipynb	Today

If you have not logged in, or have not yet created a Terra account, you will see this:

bit.ly/schare-tat



Welcome to Terra Community Workbench

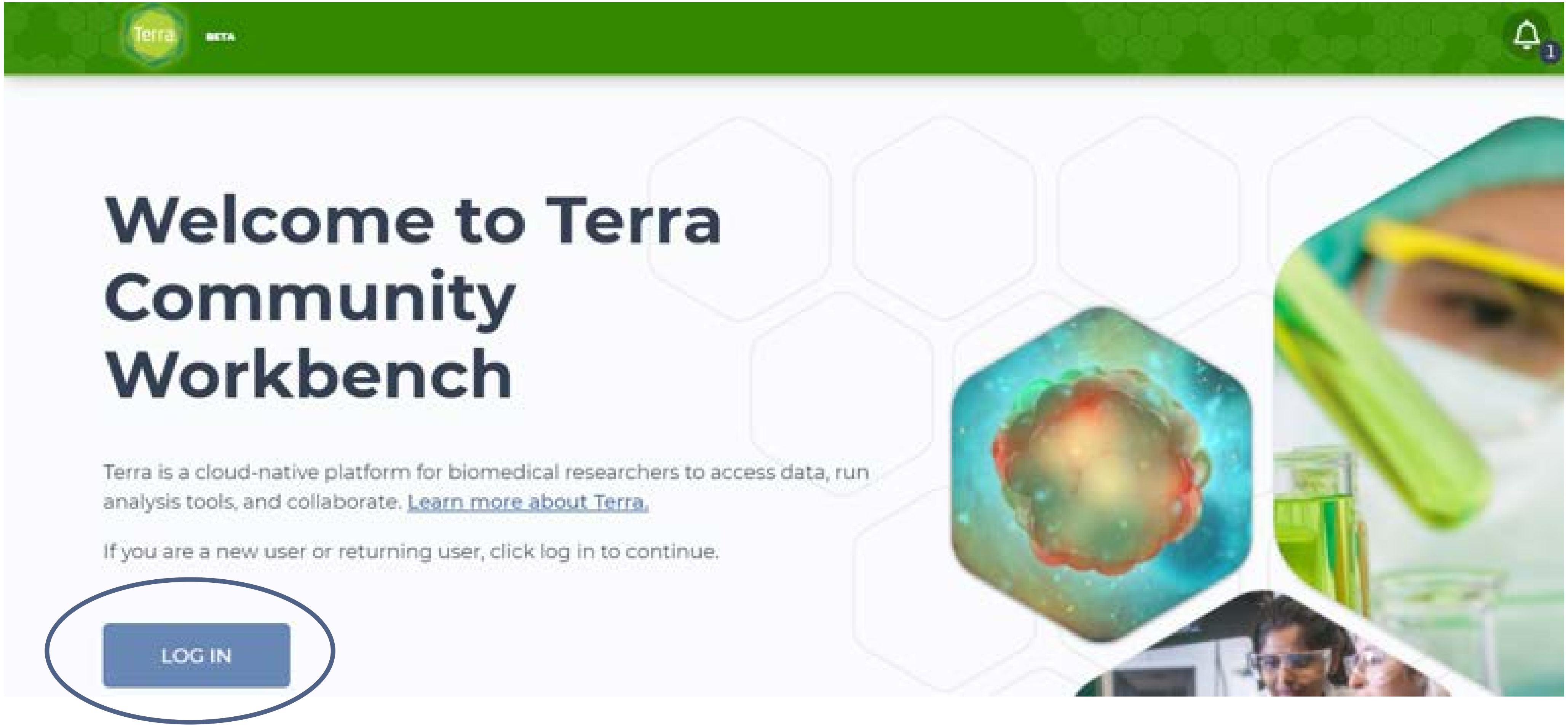
Terra is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate. Learn more about Terra.

If you are a new user or returning user, click log in to continue.

LOG IN

Click on the login button:

bit.ly/schare-tat

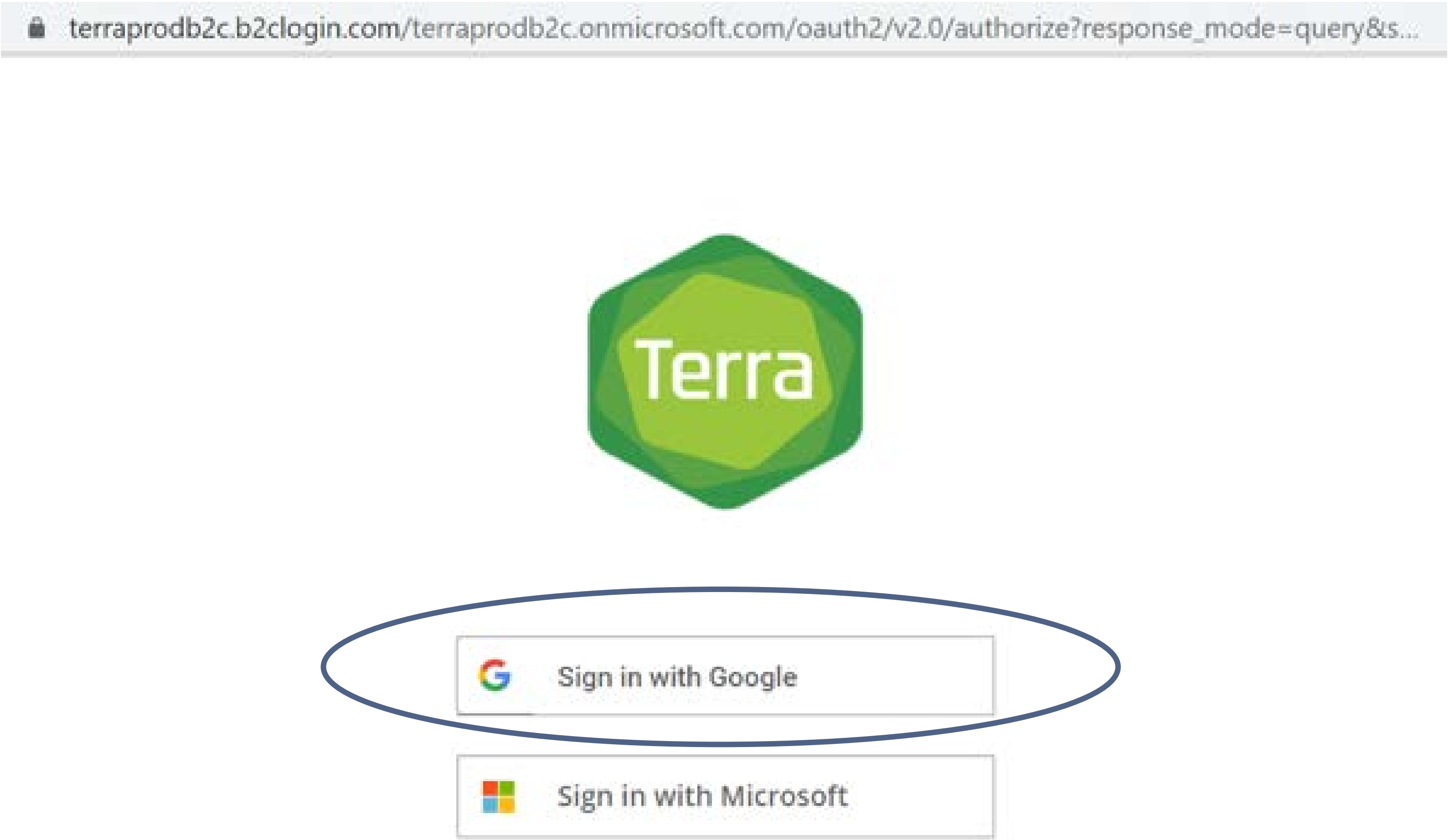


Welcome to Terra Community Workbench

Terra is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate. Learn more about Terra.

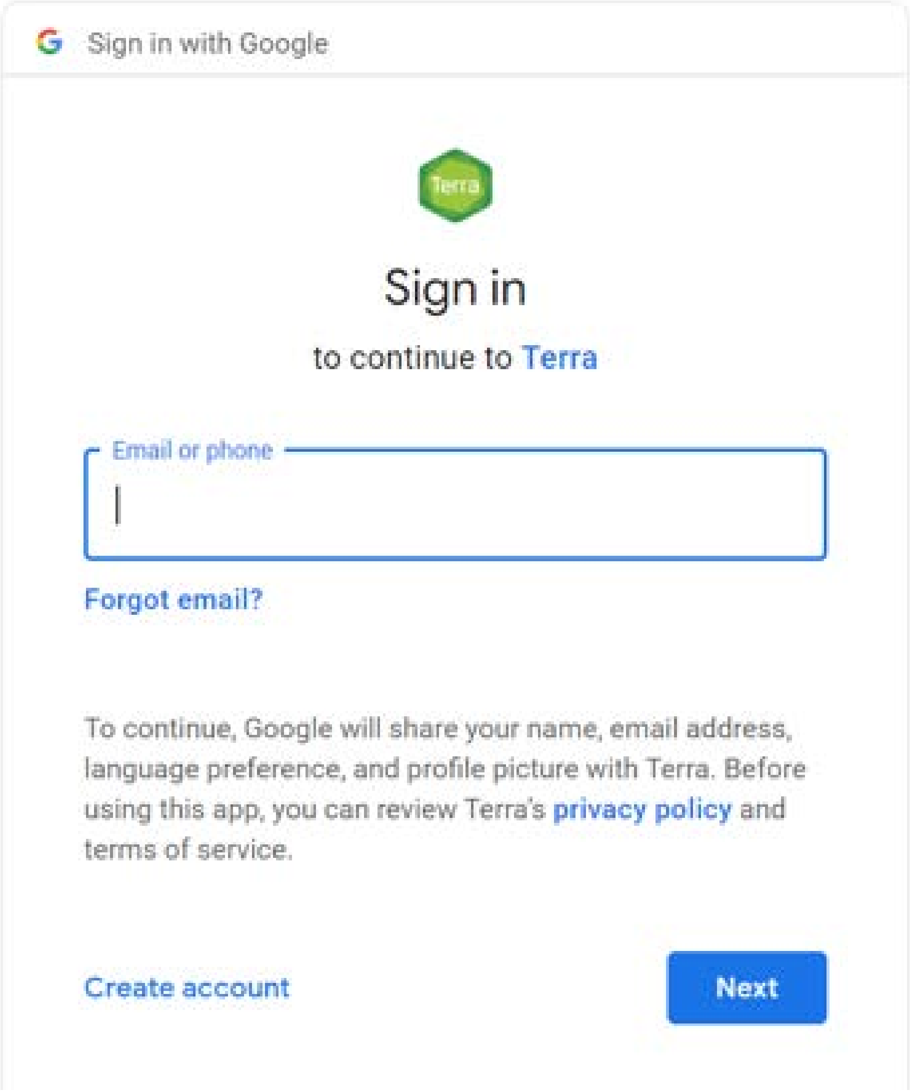
If you are a new user or returning user, click log in to continue.

Use the Gmail address you provided us with to log in:



terraprodb2c.b2clogin.com/terraprodb2c.onmicrosoft.com/oauth2/v2.0/authorize?response_mode=query&s...

Use the Gmail address you provided us with to log in:



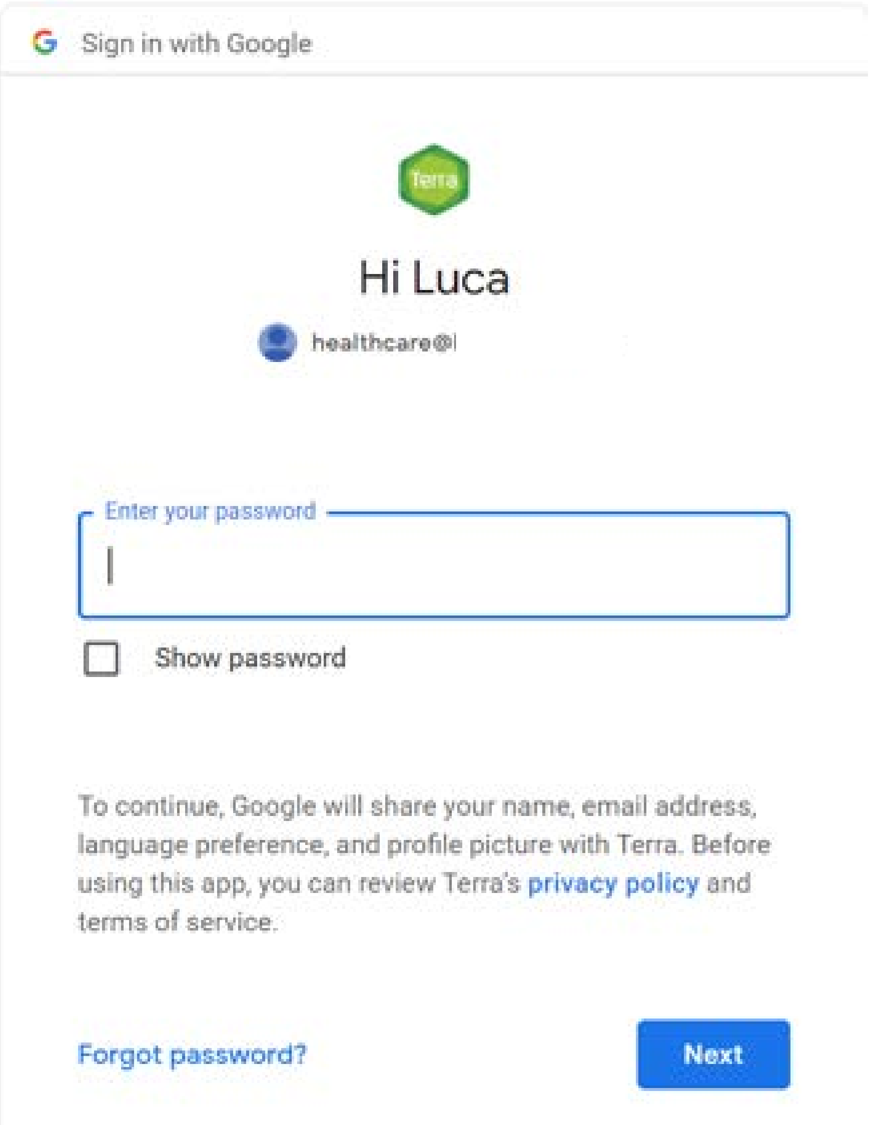
Sign in with Google

Sign in

to continue to Terra

To continue, Google will share your name, email address, language preference, and profile picture with Terra. Before using this app, you can review Terra's privacy policy and terms of service.

Input the password associated with your Gmail account:



Sign in with Google

Hi Luca

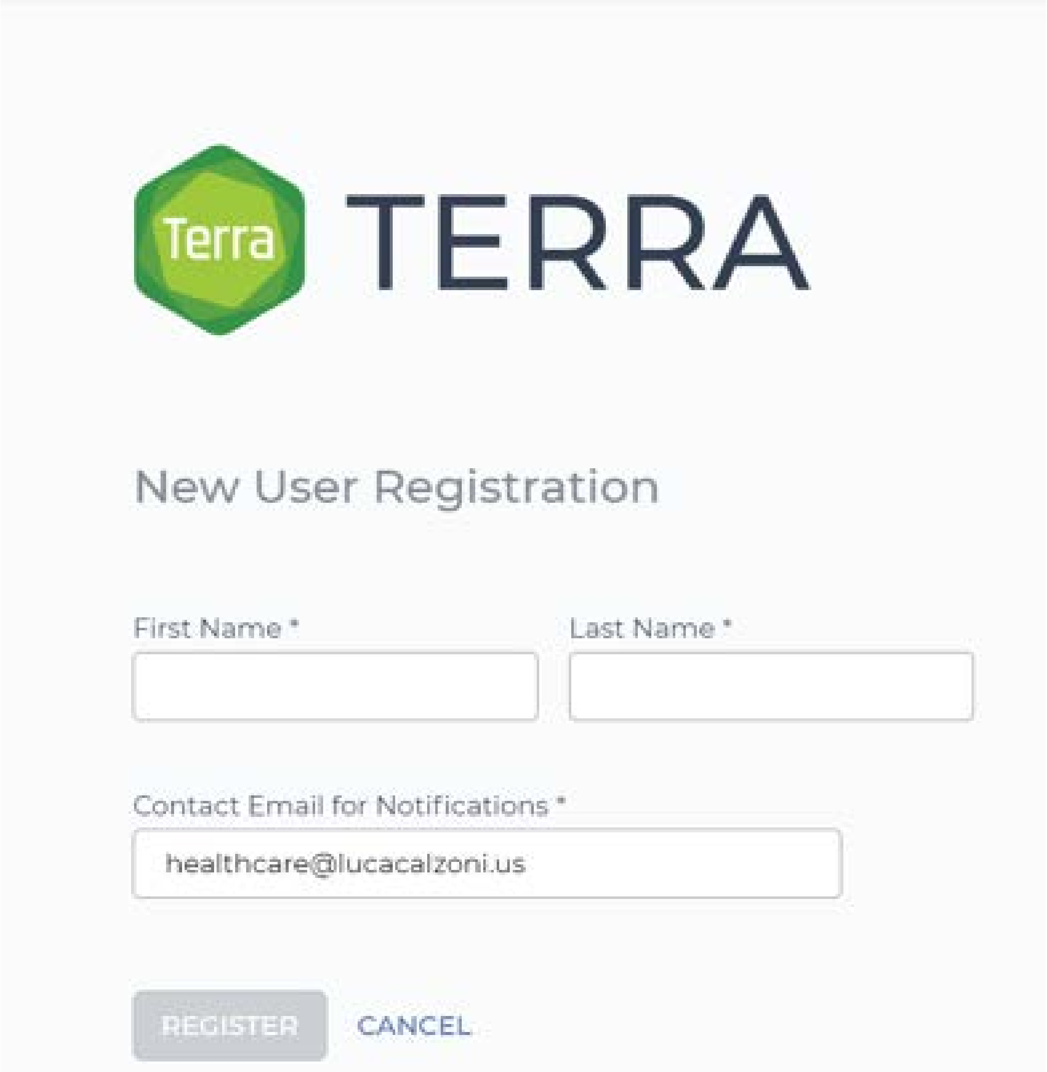
healthcare@l

Enter your password

□ Show password

To continue, Google will share your name, email address, language preference, and profile picture with Terra. Before using this app, you can review Terra's privacy policy and terms of service.

If you are new to Terra, create an account now:



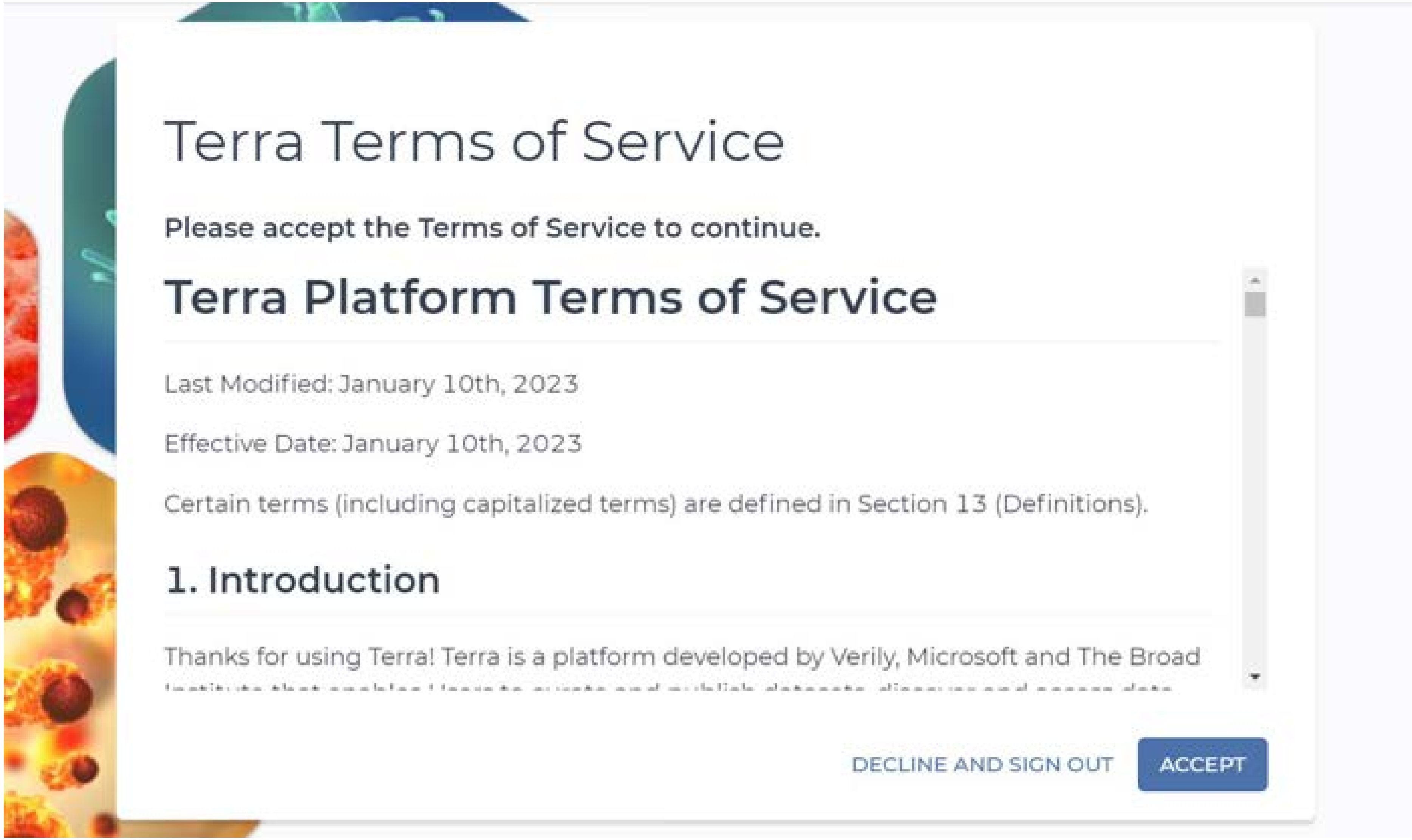
New User Registration

First Name *

Last Name *

Contact Email for Notifications *
healthcare@lucacalzoni.us

Accept the Terra Terms of Service:



Terra Terms of Service

Please accept the Terms of Service to continue.

Terra Platform Terms of Service

Last Modified: January 10th, 2023

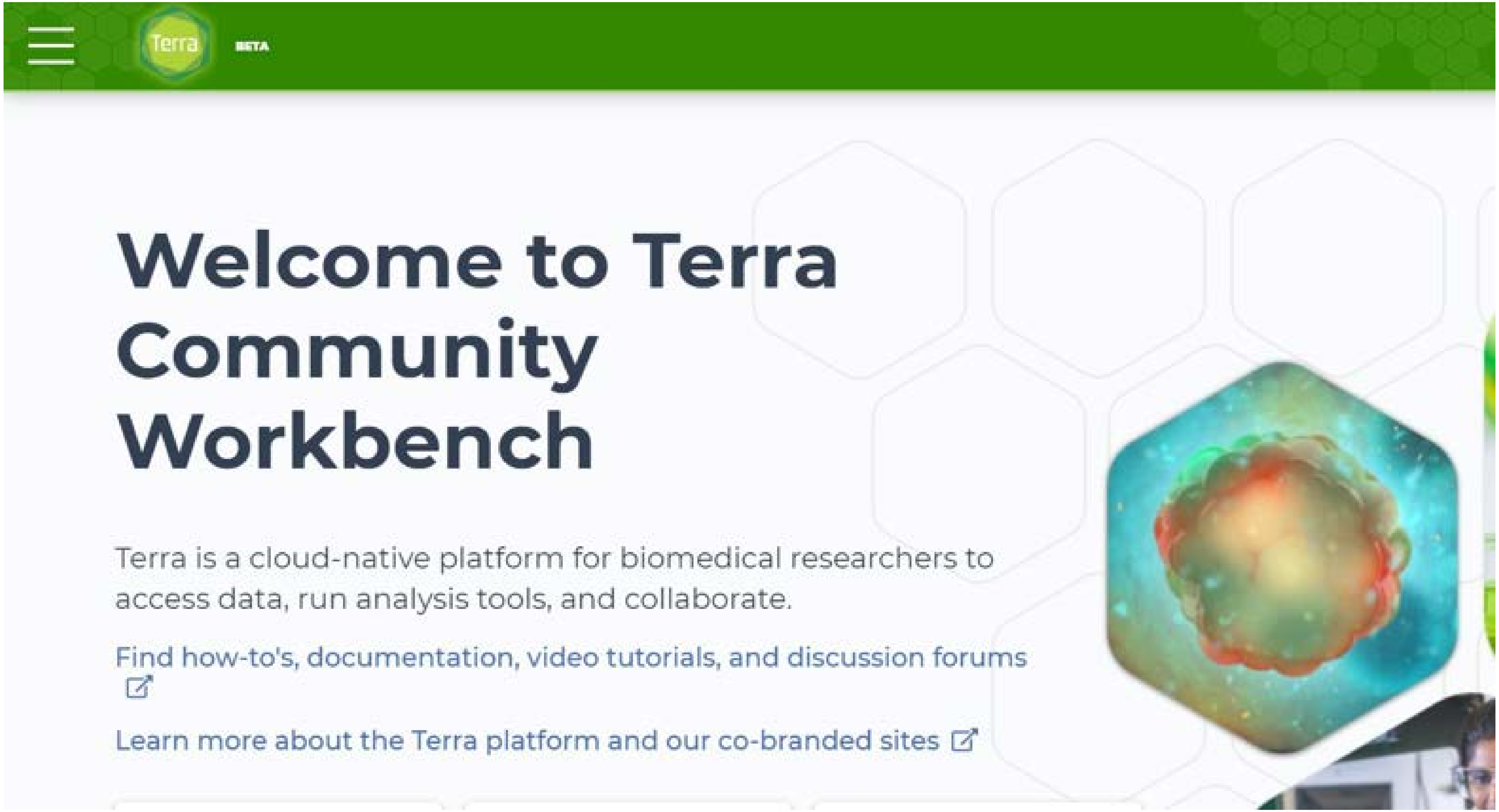
Effective Date: January 10th, 2023

Certain terms (including capitalized terms) are defined in Section 13 (Definitions).

1. Introduction

Thanks for using Terra! Terra is a platform developed by Verily, Microsoft and The Broad

You will see this welcome page:



Welcome to Terra Community Workbench

Terra is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate.

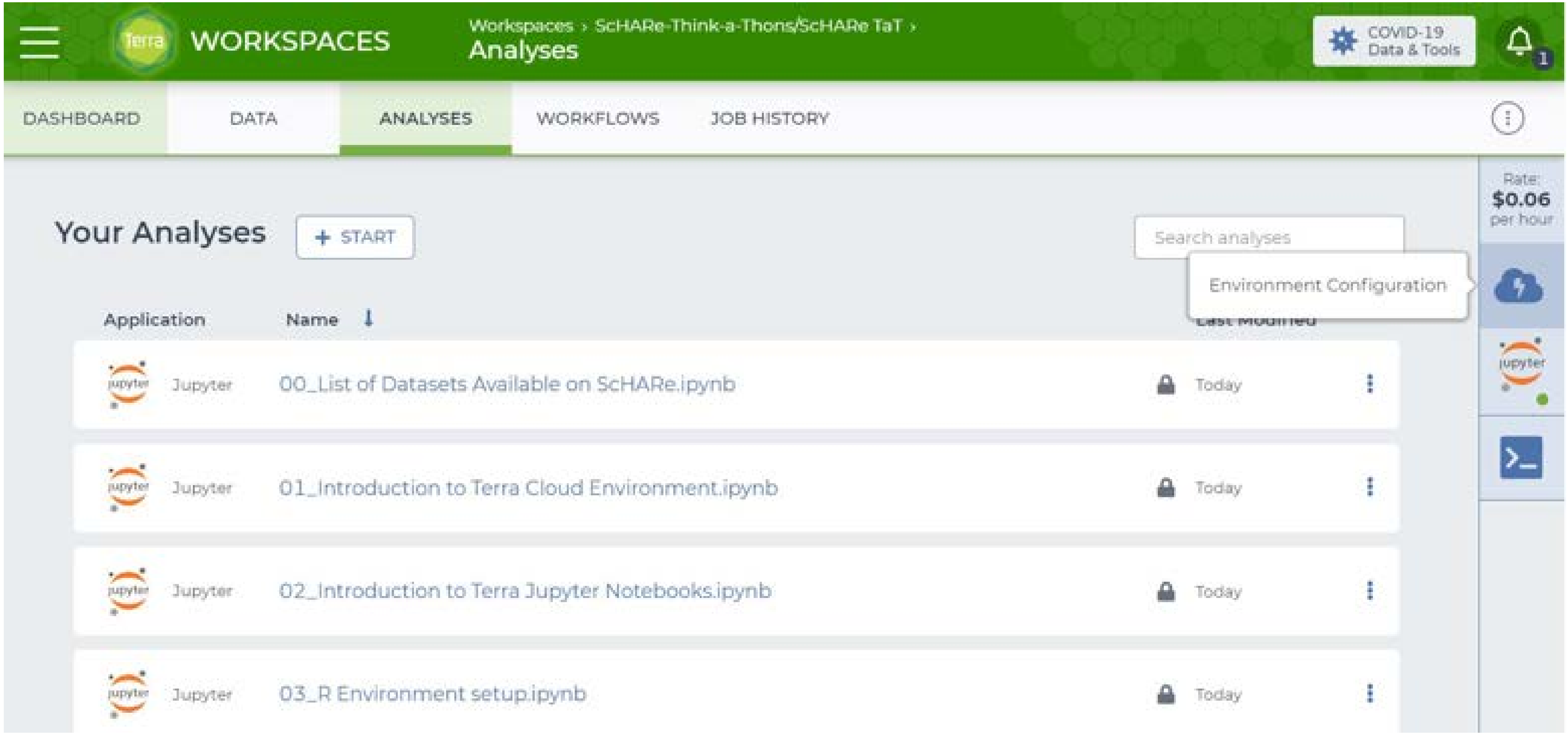
Find how-to's, documentation, video tutorials, and discussion forums

Learn more about the Terra platform and our co-branded sites

Please paste this address in your browser:

bit.ly/schare-tat

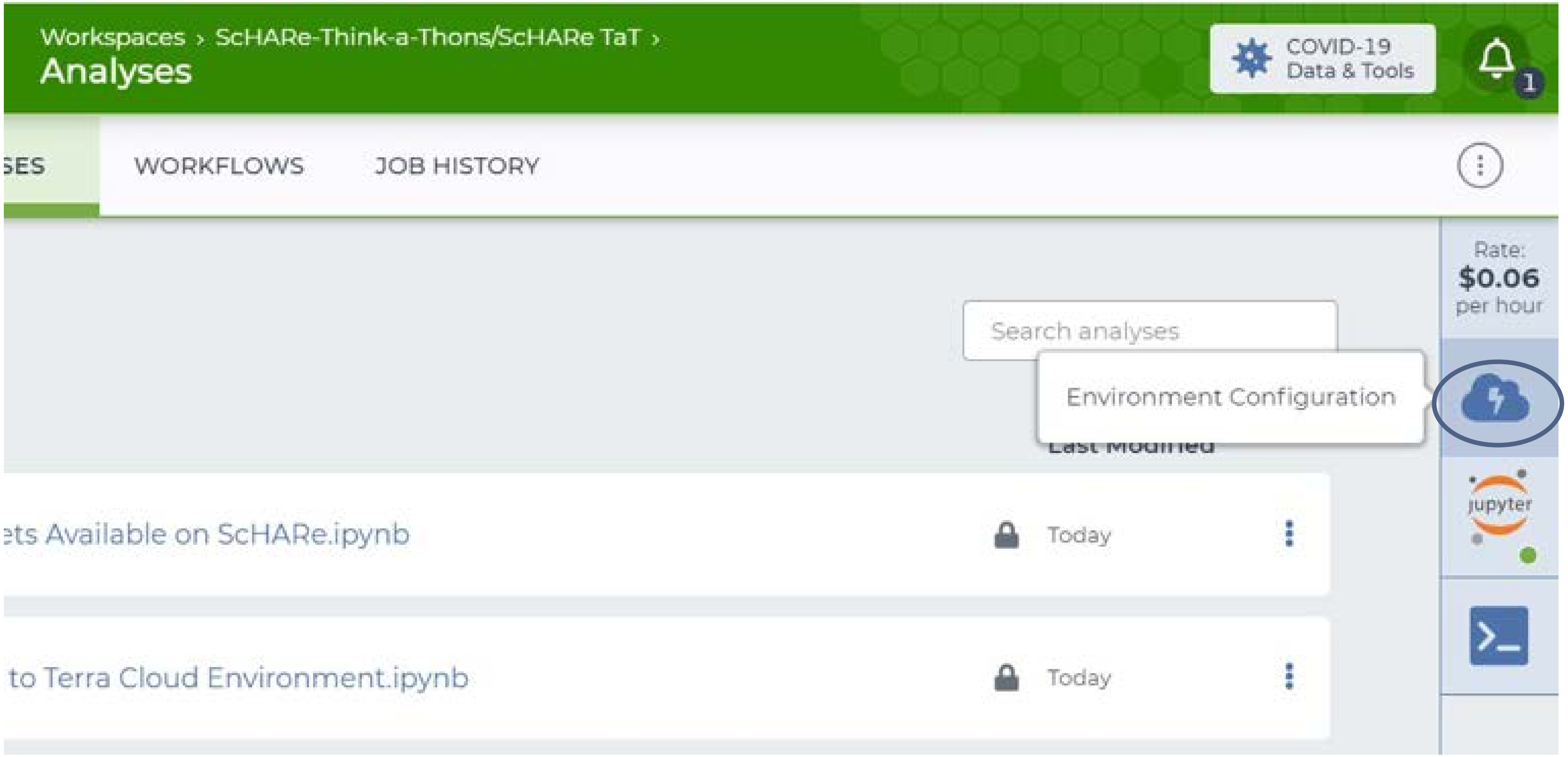
You will see this:



Your Analyses

Application	Name		Last Modified
Jupyter	00_List of Datasets Available on ScHARe.ipynb	🔒	Today
Jupyter	01_Introduction to Terra Cloud Environment.ipynb	🔒	Today
Jupyter	02_Introduction to Terra Jupyter Notebooks.ipynb	🔒	Today
Jupyter	03_R Environment setup.ipynb	🔒	Today

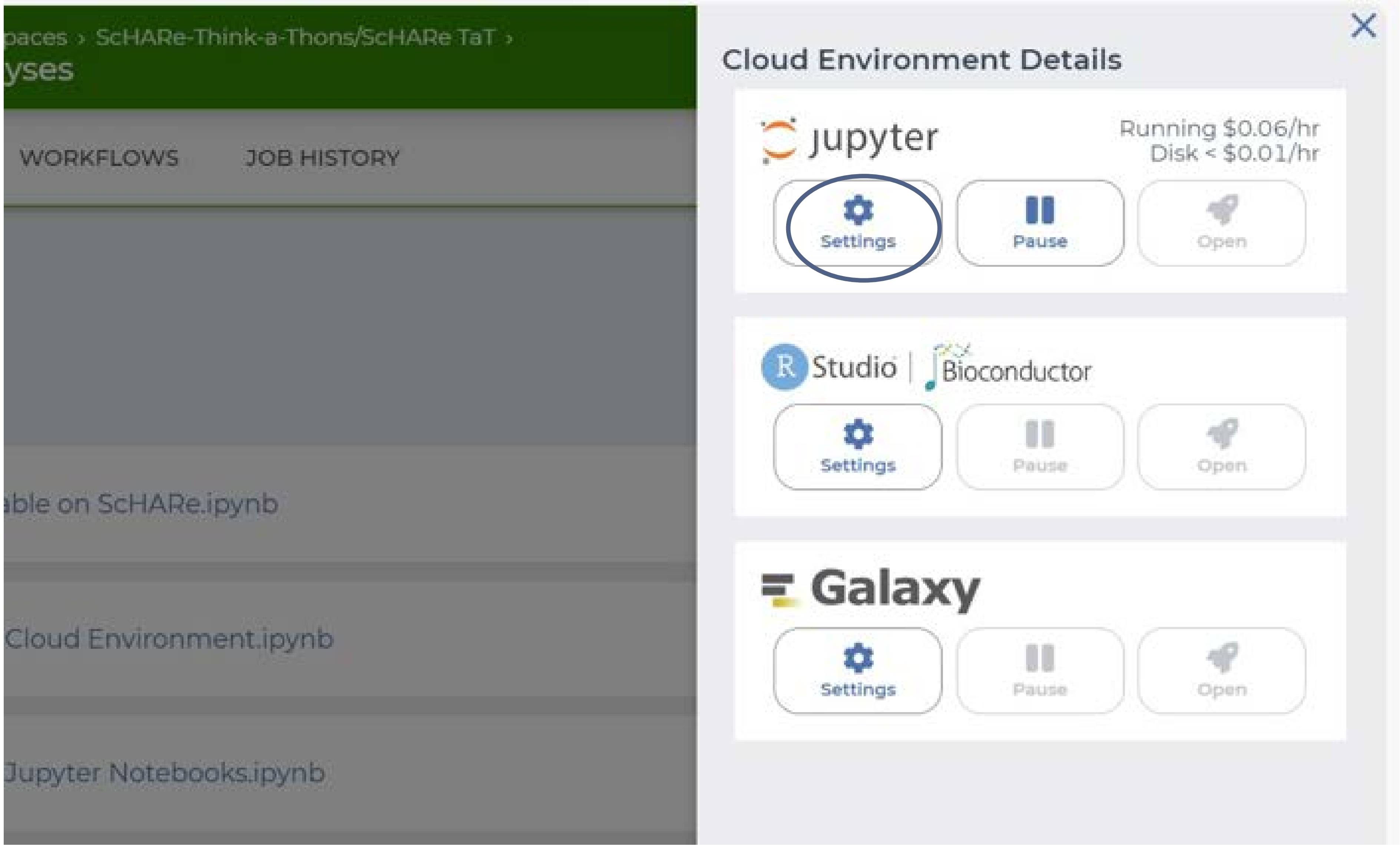
Click on the “Environment configuration” button:



ets Available on ScHARe.ipynb Today

to Terra Cloud Environment.ipynb Today

Click on Jupyter settings:



paces › ScHARe-Think-a-Thons/ScHARe TaT ›
yses

WORKFLOWS JOB HISTORY

able on ScHARe.ipynb

Cloud Environment.ipynb

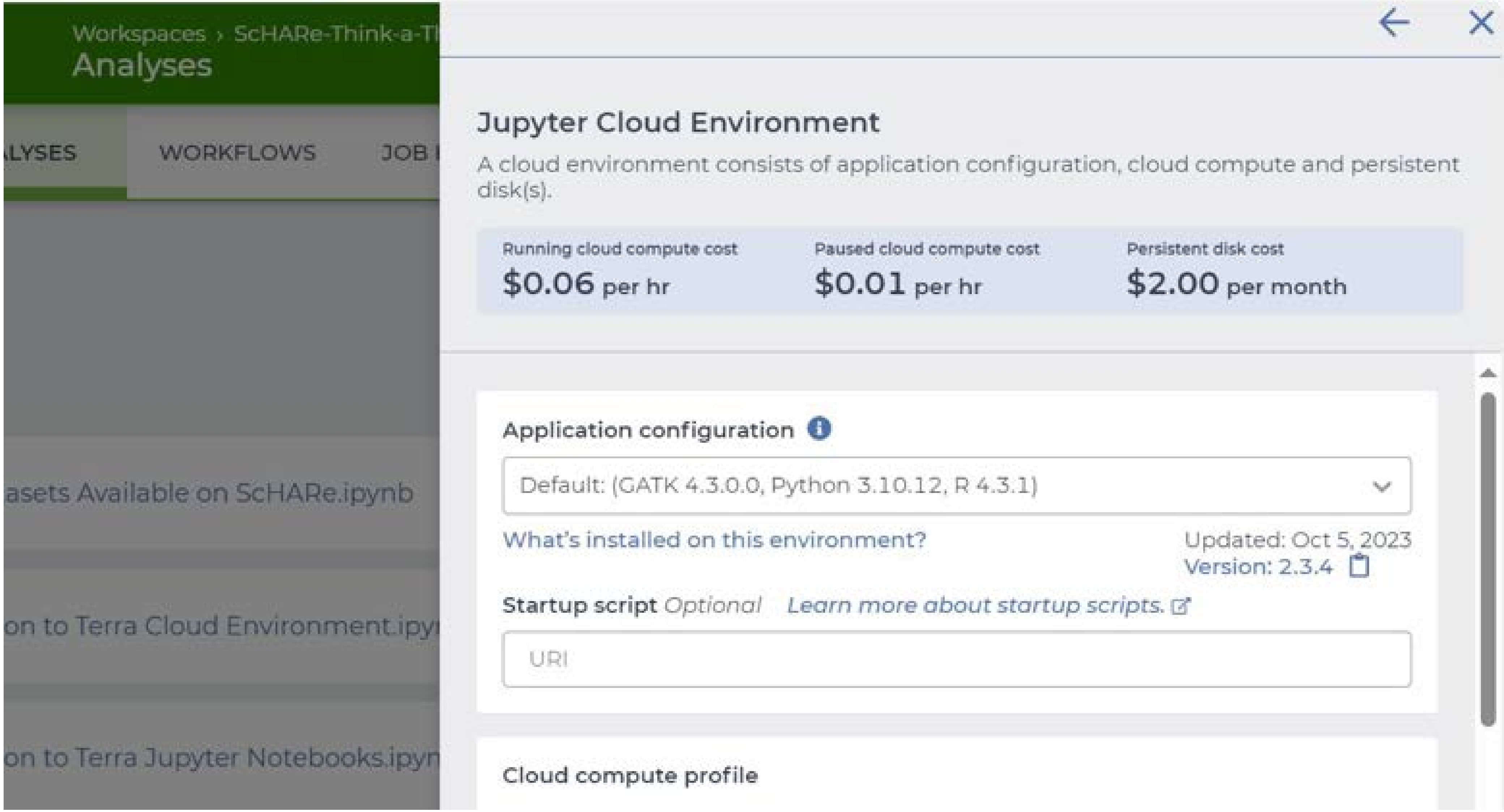
Jupyter Notebooks.ipynb

Cloud Environment Details

jupyter

Running $0.06/hr
Disk < $0.01/hr

Configure the Environment leaving the default values checked:



Workspaces › ScHARe-Think-a-Th
Analyses
LYSES
WORKFLOWS
JOB
asets Available on ScHARe.ipynb
on to Terra Cloud Environment.ipy
on to Terra Jupyter Notebooks.ipyn

Jupyter Cloud Environment

A cloud environment consists of application configuration, cloud compute and persistent disk(s).

Running cloud compute cost $0.06 per hr

Paused cloud compute cost $0.01 per hr

Persistent disk cost $2.00 per month

Application configuration

Default: (GATK 4.3.0.0, Python 3.10.12, R 4.3.1)

What's installed on this environment?

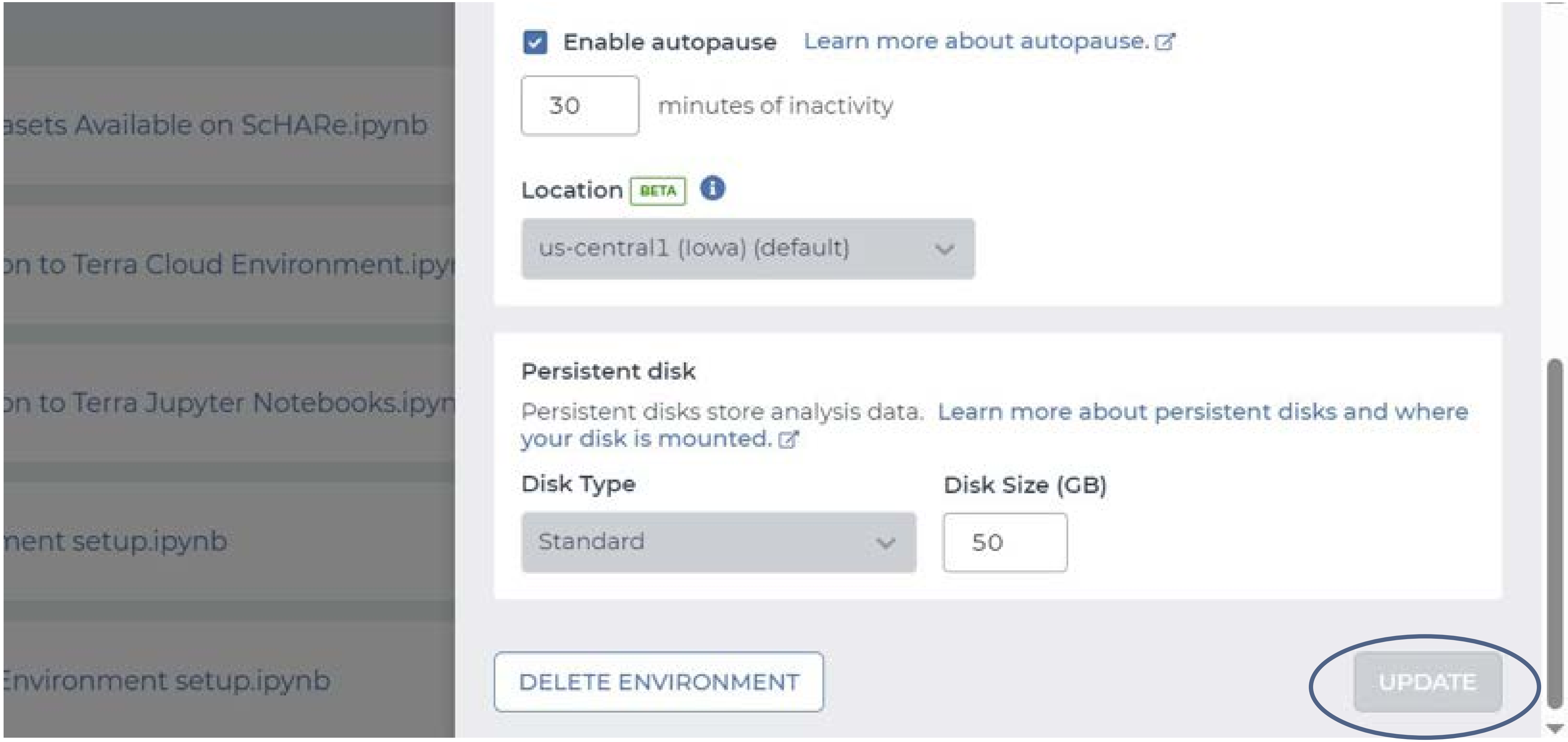
Updated: Oct 5, 2023
Version: 2.3.4

Startup script *Optional* *Learn more about startup scripts.*

URI

Cloud compute profile

Click on “Create” (if first time) or “Update” at the bottom:



☑ Enable autopause Learn more about autopause.

30 minutes of inactivity

Location BETA

us-central1 (Iowa) (default)

Persistent disk

Persistent disks store analysis data. Learn more about persistent disks and where your disk is mounted.

Disk Type: Standard

Disk Size (GB): 50

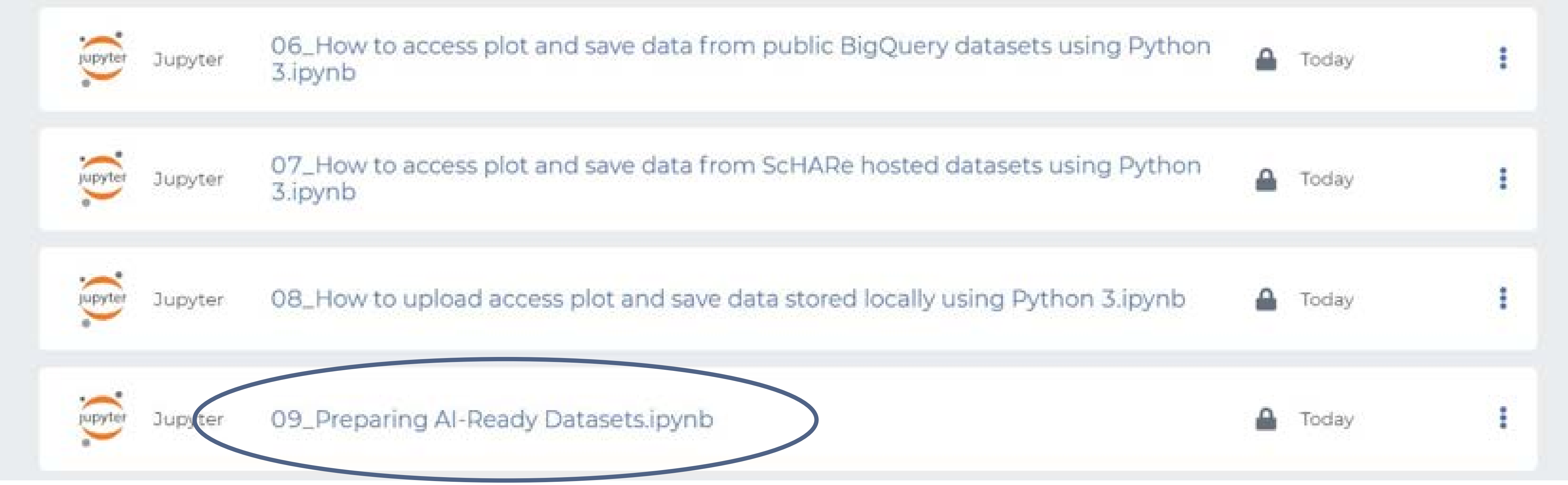
asets Available on ScHARe.ipynb
on to Terra Cloud Environment.ipy
on to Terra Jupyter Notebooks.ipyn
nent setup.ipynb
Environment setup.ipynb
DELETE ENVIRONMENT
UPDATE

Congratulations!

Your virtual machine is being created.

bit.ly/schare-tat

Scroll down to Notebooks #09 and click on the appropriate version (based on the first letter your last name) to open it:



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

Jupyter 07_How to access plot and save data from ScHARe hosted datasets using Python 3.ipynb Today

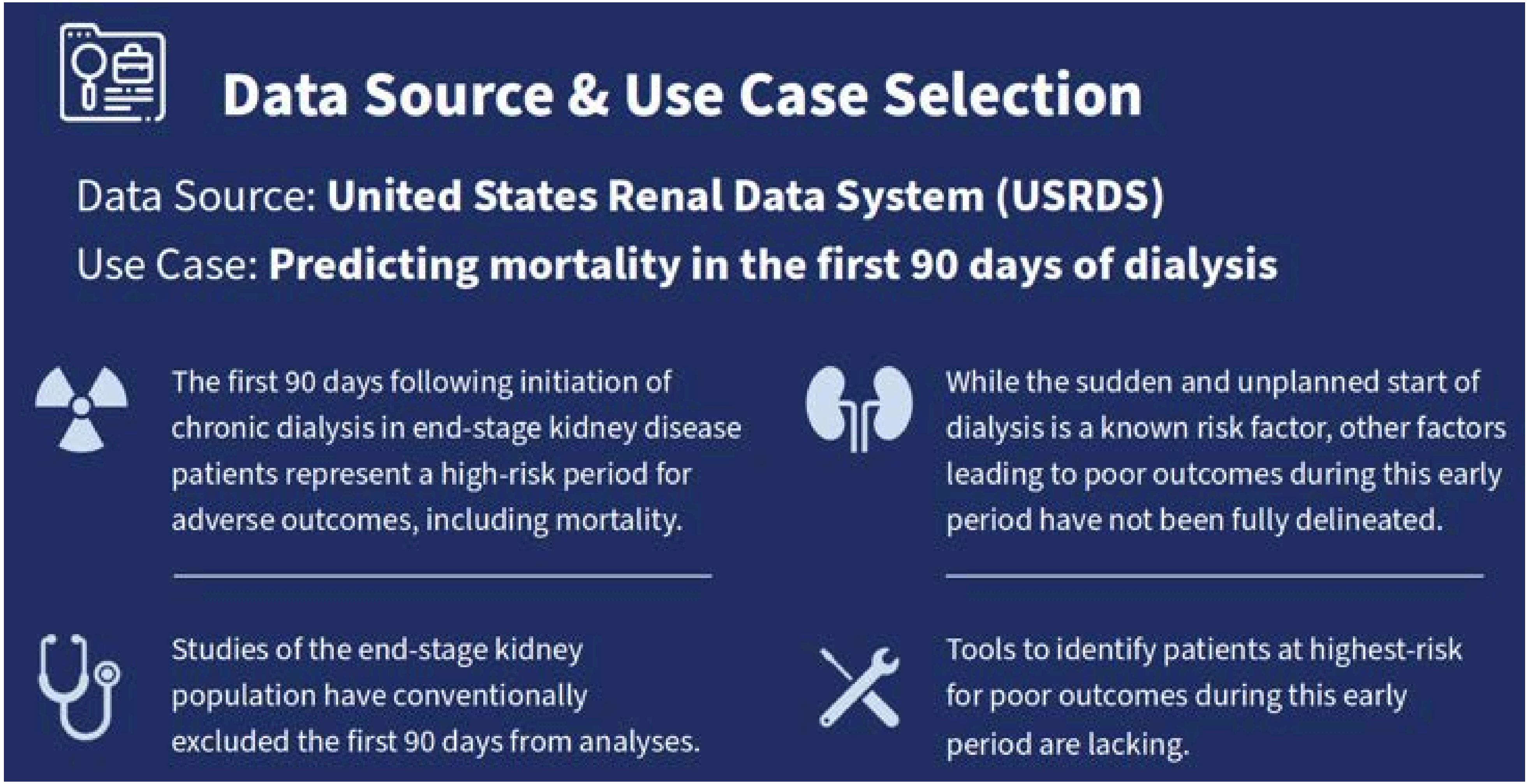
Jupyter 08_How to upload access plot and save data stored locally using Python 3.ipynb Today

Jupyter 09_Preparing AI-Ready Datasets.ipynb Today

Slido Quiz

- Properly handling self-reported demographic data is an emerging field of interest. What are your thoughts?
- Some points to consider:
  - *Free response is the most accurate, but how do you analyze this?*
  - *Offering many categories can lead to "small n", where few observations are recorded in some categories. Is it then okay to combine categories?*
  - *Is "race" or "sex" or "gender" just a proxy for something you even mean within your population? Are there variables you should be recording instead that are more related to exposure or outcome?*

CKD Example - Research Design



Data Source & Use Case Selection

Data Source: **United States Renal Data System (USRDS)**

Use Case: **Predicting mortality in the first 90 days of dialysis**

The first 90 days following initiation of chronic dialysis in end-stage kidney disease patients represent a high-risk period for adverse outcomes, including mortality.

While the sudden and unplanned start of dialysis is a known risk factor, other factors leading to poor outcomes during this early period have not been fully delineated.

Studies of the end-stage kidney population have conventionally excluded the first 90 days from analyses.

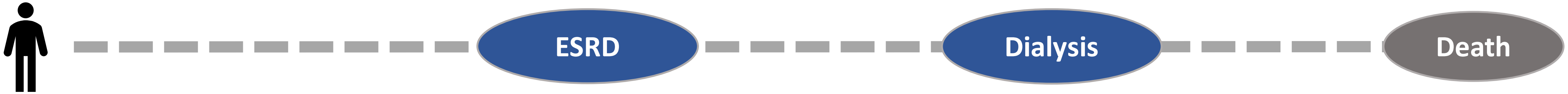
Tools to identify patients at highest-risk for poor outcomes during this early period are lacking.

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – USRDS Data Mapping to Use Case

CKD Patient

Selected use case: *Predicting mortality in the first 90 days of dialysis*



1. **CMS Pre-ESRD Claims Datasets**
   - Parts A and B claims prior to ESRD diagnosis
   - Used to build features, such as prior nephrology care

2. **ESRD Medical Evidence Report (MEDEVID) (CMS 2728)/ PATIENTS Dataset**
   - Form is completed when a patient is diagnosed as ESRD and receives their first chronic dialysis treatment(s) or transplant
   - Used to build features such as patient demographics, comorbid conditions, primary cause of renal failure, and laboratory values

2A. PATIENTS Dataset
- Provides basic demographic and ESRD-related data
- Used to obtain dialysis start date and modality
- Used in conjunction with MEDEVID to build demographic features such as age, sex, race, etc.

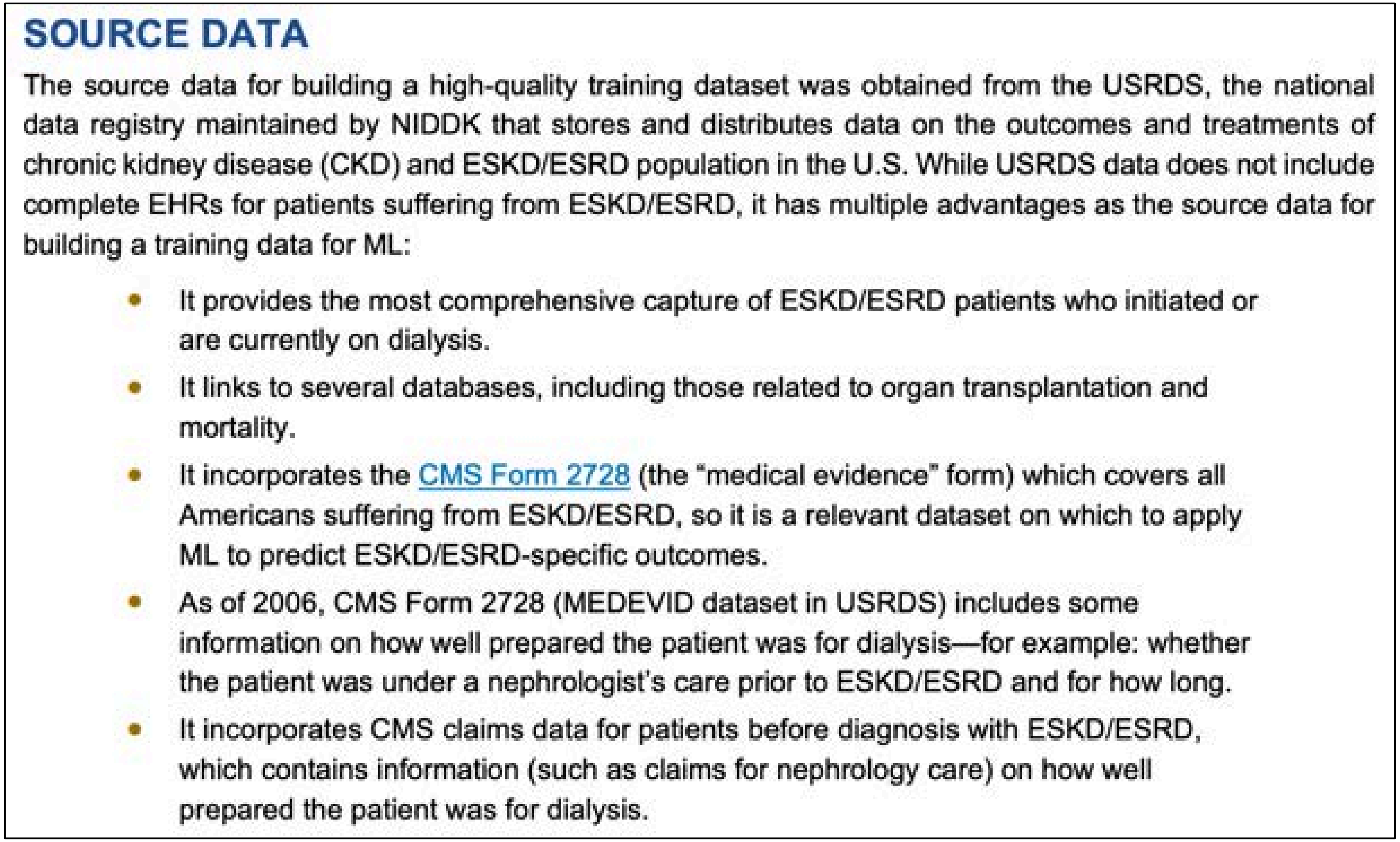
2B. Transplant Dataset (TX)
- Provides information on kidney transplants such as list date/data on eligibility pre-dialysis
- Used to build features such as transplant waitlist status

3. **PATIENTS Dataset/ DEATH Dataset (CMS ESRD Death Notification Form 2726)**
   - Used to determine if a patient died in the first 90 days after dialysis start

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – Data Documentation

Documentation of source for dataset(s)



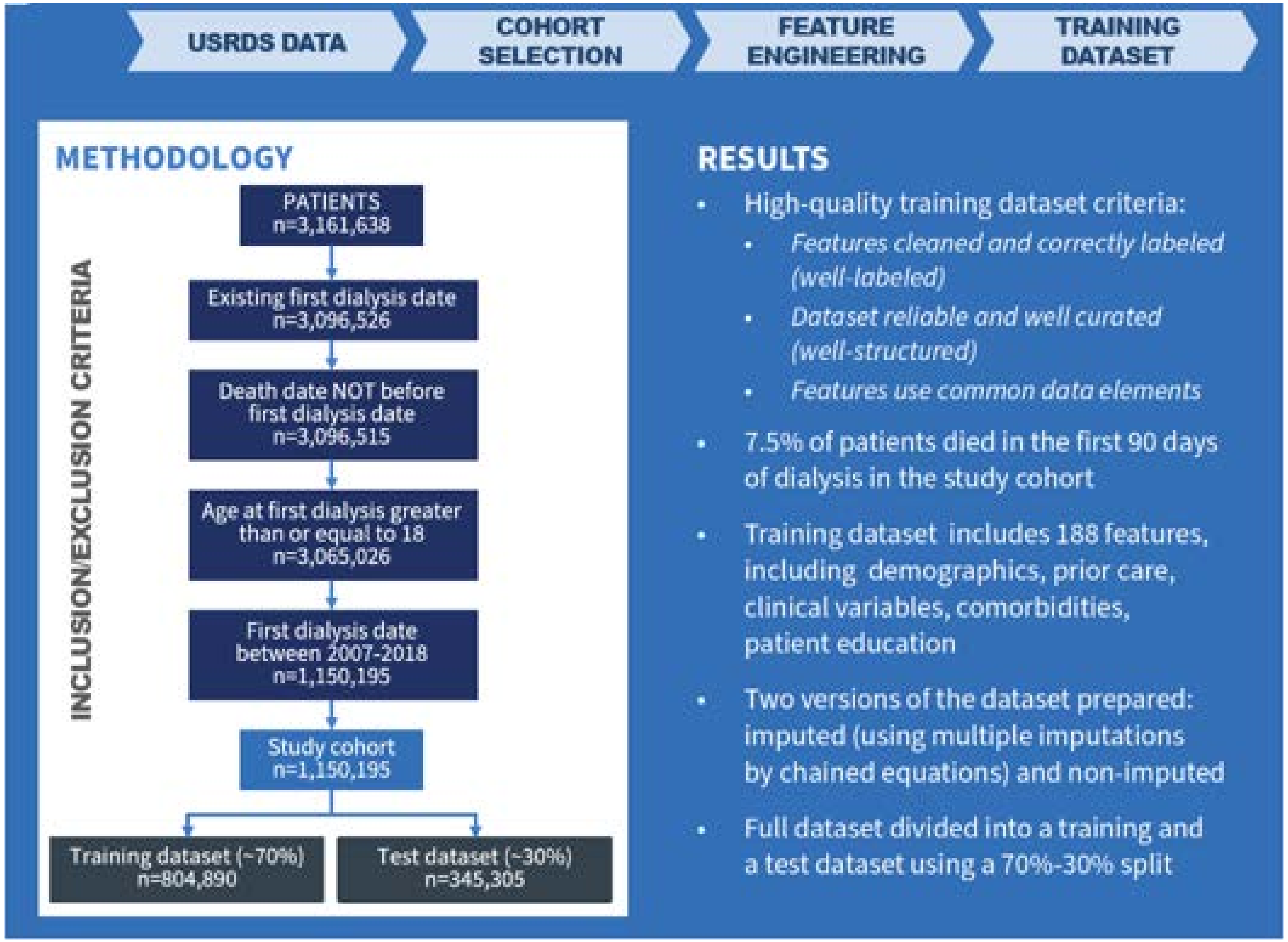
SOURCE DATA

The source data for building a high-quality training dataset was obtained from the USRDS, the national data registry maintained by NIDDK that stores and distributes data on the outcomes and treatments of chronic kidney disease (CKD) and ESKD/ESRD population in the U.S. While USRDS data does not include complete EHRs for patients suffering from ESKD/ESRD, it has multiple advantages as the source data for building a training data for ML:

- It provides the most comprehensive capture of ESKD/ESRD patients who initiated or are currently on dialysis.
- It links to several databases, including those related to organ transplantation and mortality.
- It incorporates the CMS Form 2728 (the “medical evidence” form) which covers all Americans suffering from ESKD/ESRD, so it is a relevant dataset on which to apply ML to predict ESKD/ESRD-specific outcomes.
- As of 2006, CMS Form 2728 (MEDEVID dataset in USRDS) includes some information on how well prepared the patient was for dialysis—for example: whether the patient was under a nephrologist’s care prior to ESKD/ESRD and for how long.
- It incorporates CMS claims data for patients before diagnosis with ESKD/ESRD, which contains information (such as claims for nephrology care) on how well prepared the patient was for dialysis.

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – Data



USRDS DATA → COHORT SELECTION → FEATURE ENGINEERING → TRAINING DATASET

METHODOLOGY

INCLUSION/EXCLUSION CRITERIA

PATIENTS n=3,161,638

Existing first dialysis date n=3,096,526

Death date NOT before first dialysis date n=3,096,515

Age at first dialysis greater than or equal to 18 n=3,065,026

First dialysis date between 2007-2018 n=1,150,195

Study cohort n=1,150,195

Training dataset (~70%) n=804,890

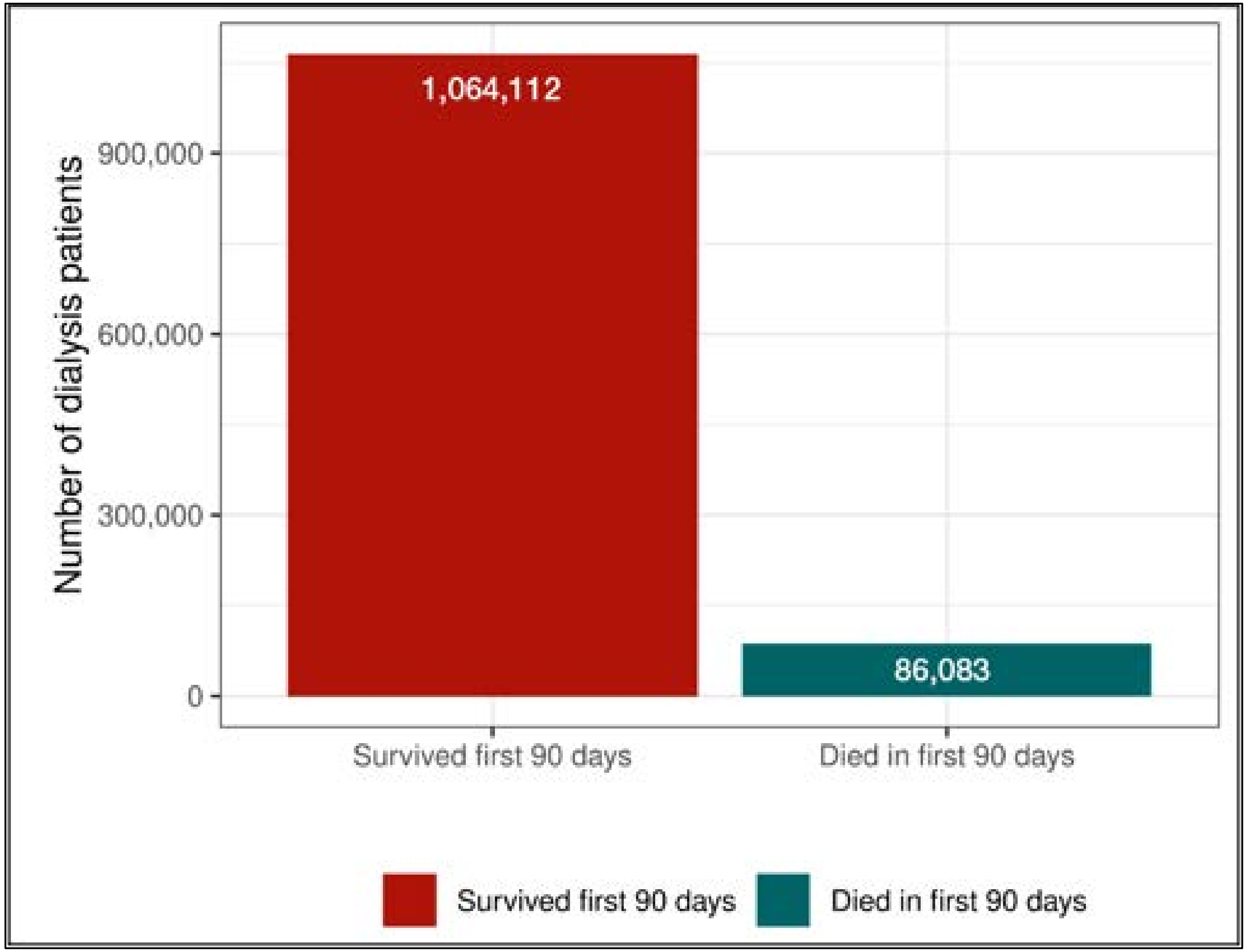
Test dataset (~30%) n=345,305

RESULTS

- High-quality training dataset criteria:
  - *Features cleaned and correctly labeled (well-labeled)*
  - *Dataset reliable and well curated (well-structured)*
  - *Features use common data elements*
- 7.5% of patients died in the first 90 days of dialysis in the study cohort
- Training dataset includes 188 features, including demographics, prior care, clinical variables, comorbidities, patient education
- Two versions of the dataset prepared: imputed (using multiple imputations by chained equations) and non-imputed
- Full dataset divided into a training and a test dataset using a 70%-30% split

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

Imbalanced Data



https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

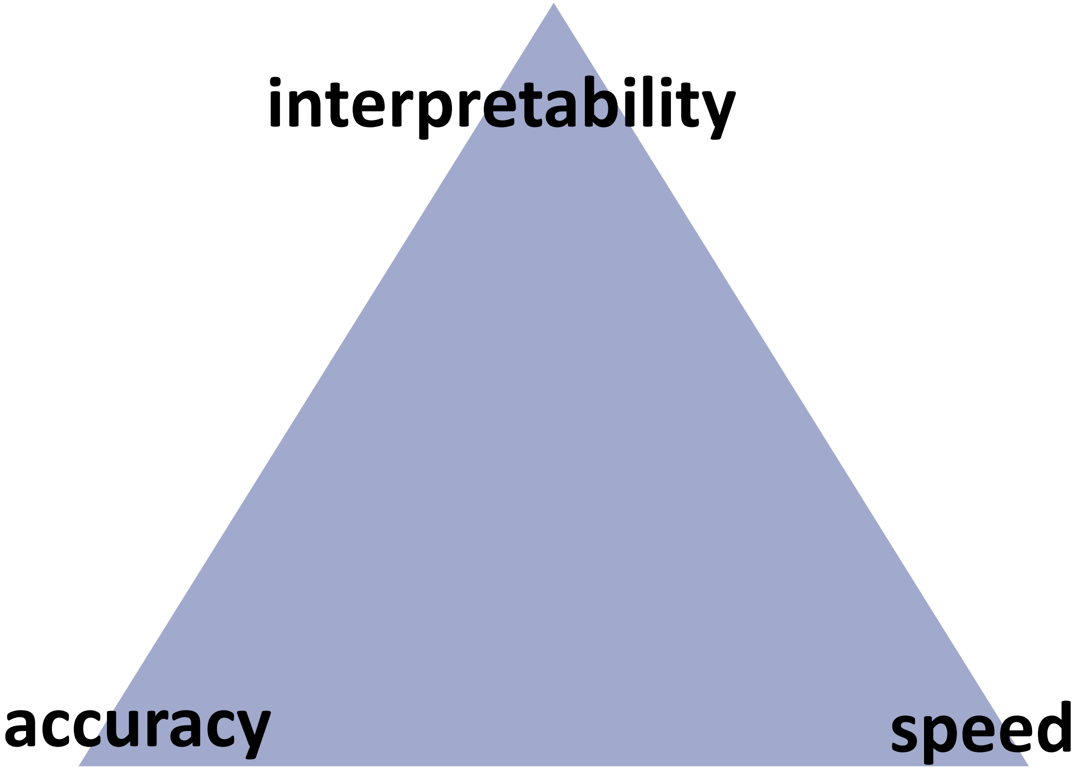
Model Selection

Labeled Data

- Supervised learning

Unlabeled Data

- Unsupervised Learning
- Dimensionality Reduction

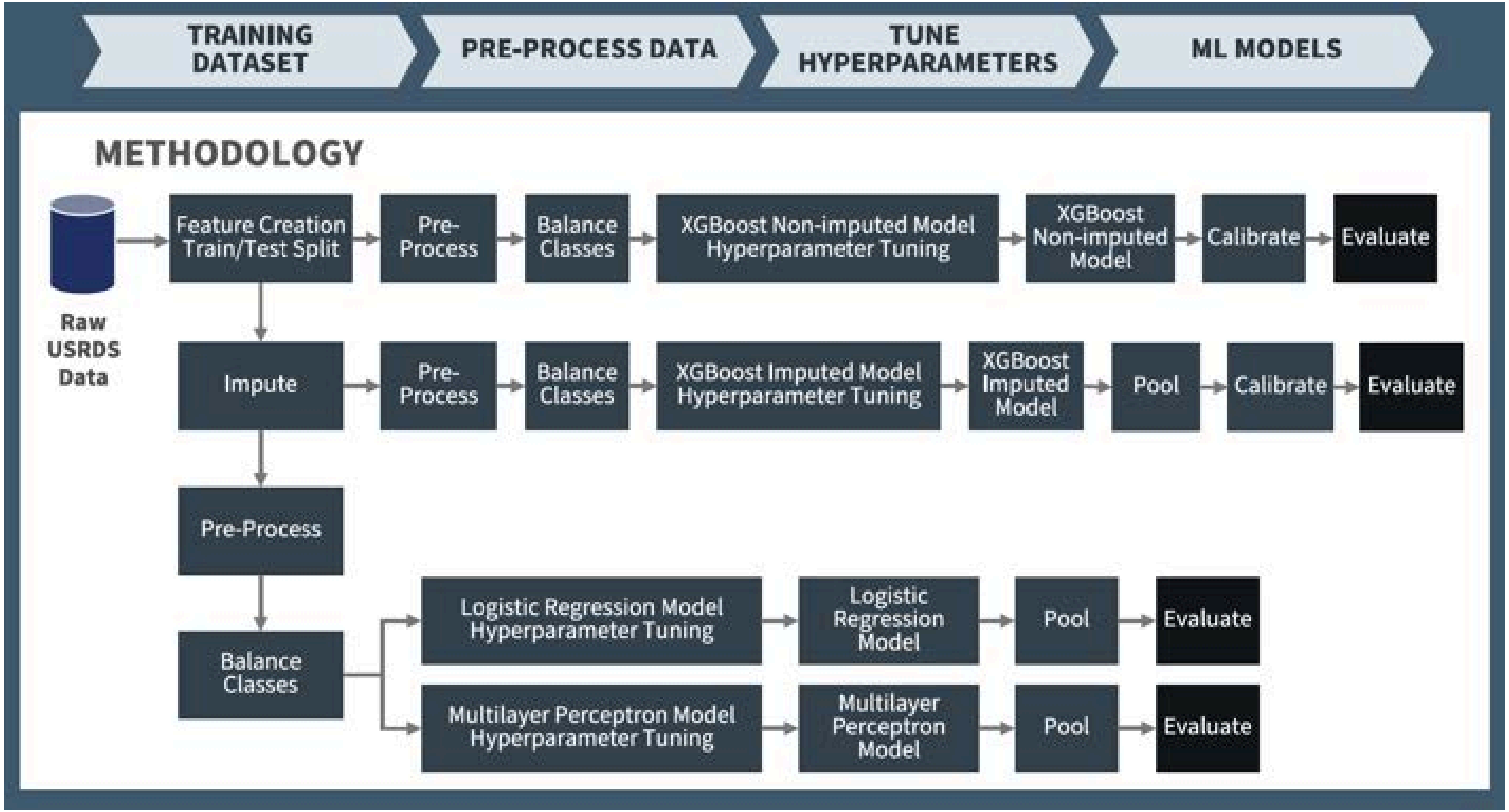


interpretability

accuracy

speed

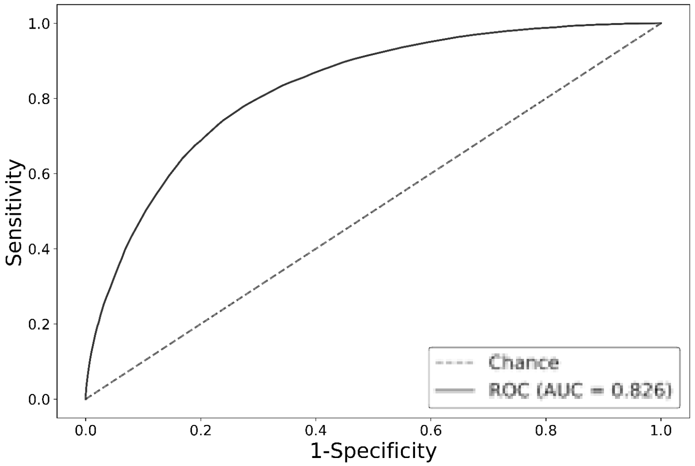
CKD Example – Model Selection


TRAINING DATASET
PRE-PROCESS DATA
TUNE HYPERPARAMETERS
ML MODELS
METHODOLOGY
Raw USRDS Data
Feature Creation Train/Test Split
Pre-Process
Balance Classes
XGBoost Non-imputed Model Hyperparameter Tuning
XGBoost Non-imputed Model
Calibrate
Evaluate
Impute
Pre-Process
Balance Classes
XGBoost Imputed Model Hyperparameter Tuning
XGBoost Imputed Model
Pool
Calibrate
Evaluate
Pre-Process
Balance Classes
Logistic Regression Model Hyperparameter Tuning
Logistic Regression Model
Pool
Evaluate
Multilayer Perceptron Model Hyperparameter Tuning
Multilayer Perceptron Model
Pool
Evaluate

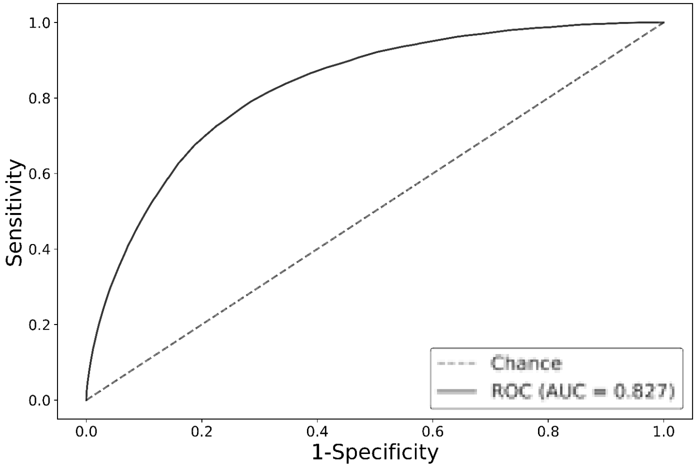
https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – Model Results

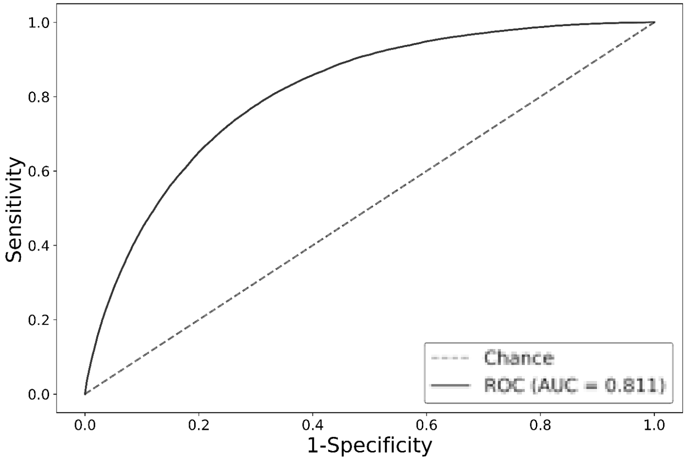
XGBoost Non-imputed



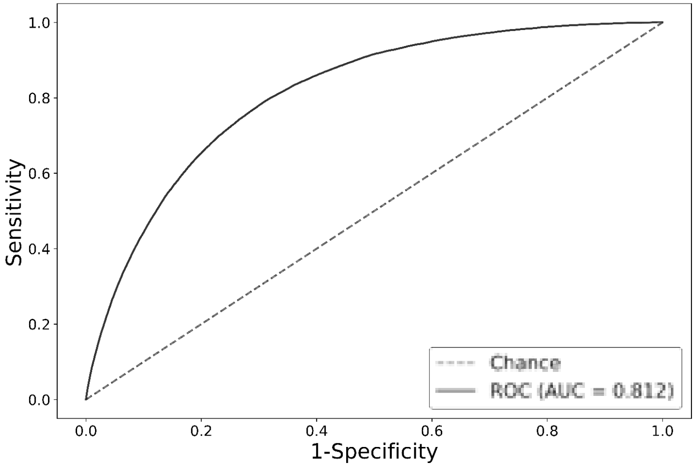
XGBoost Imputed



Logistic Regression



Multilayer Perceptron



https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – Model Interpretability

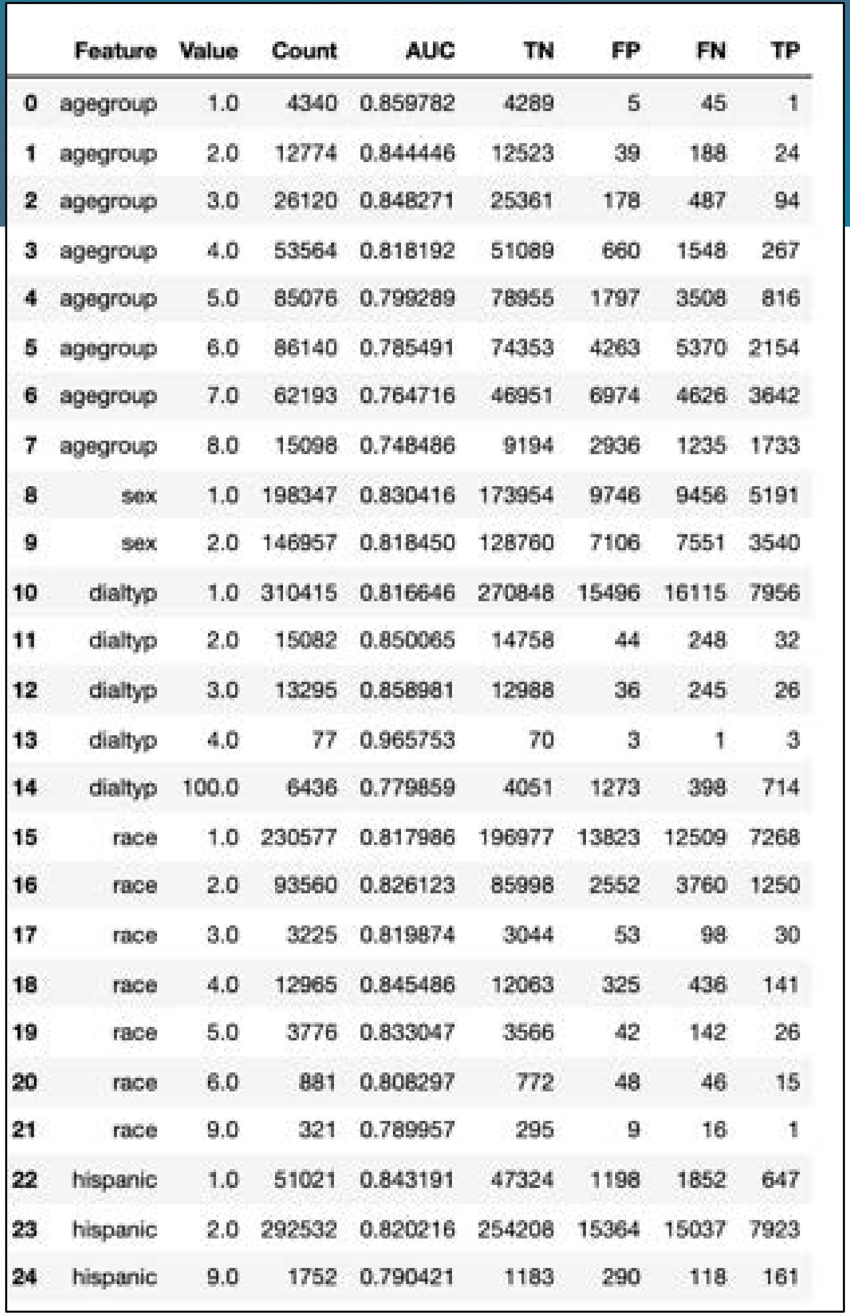
	Feature	Explanation
1.	**Age**	• Older age is associated with worse survival
2.	**Inpatient stays**	• Longer inpatient stays is more common in older and sicker patients and has been associated with early mortality
3.	**Received erythropoietin (EPO)**	• EPO hormone is produced by kidneys when it senses low oxygen levels in the blood; EPO triggers bone marrow to produce more red blood cells which raises blood oxygen • Patients on EPO typically have advanced CKD at the time of dialysis and are under the care of a nephrologist • Patients with kidney failure produces less EPO; therefore, are given EPO
4.	**Albumin**	• Albumin reflects the patient's overall health status (including nutrition and inflammation) • Risk of death is increased by poor serum albumin levels reflecting inadequate nutrition
5.	**Arteriovenous Fistula (AVF)**	• The presence of a maturing AVF indicates prior nephrology care • Hemodialysis through AVF access is associated with reduced mortality

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example – Fairness Assessment

- ML models can perform differently for different categories of patients, so the non-imputed XGBoost model was assessed for fairness, or how well the model performs for each category of interest (demographics—sex, race, and age—as well as initial dialysis modality). Age were binned into the following categories based on clinician input and an example in literature: 18-25, 26-35, 36-45, 46-55, 56-65, 66-75, 76-85, 86+. The USRDS predefined categories for race, sex, and dialysis modality were used for the fairness assessment.

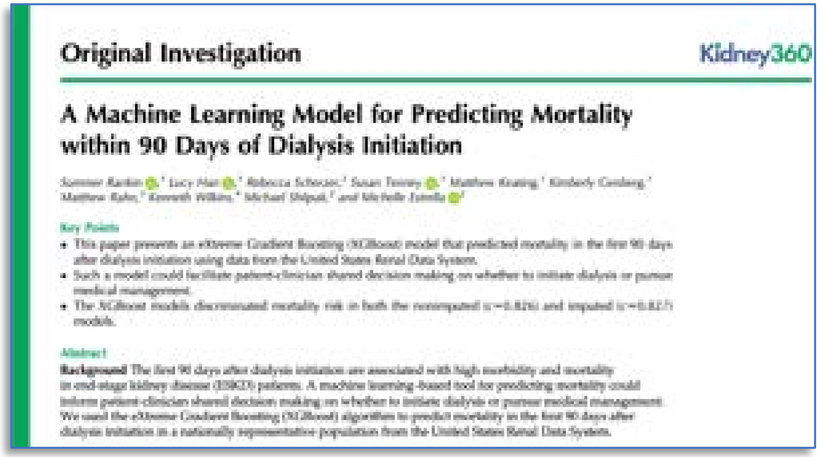
- Performing the fairness assessment on the categories of interest gives additional insight into how the model performs by different patient categories of interest (by demographics, etc.). Future researchers should perform fairness assessments to better evaluate model performance, especially for models that may be deployed in a clinical setting. Other methods of assessing fairness include evaluating true positives, sensitivity, positive predictive value, etc. at various threshold across the different groups of interest, which would allow selection of a threshold that balances model performance across the groups of interest.



	Feature	Value	Count	AUC	TN	FP	FN	TP
0	agegroup	1.0	4340	0.859782	4289	5	45	1
1	agegroup	2.0	12774	0.844446	12523	39	188	24
2	agegroup	3.0	26120	0.848271	25361	178	487	94
3	agegroup	4.0	53564	0.818192	51089	660	1548	267
4	agegroup	5.0	85076	0.799289	78955	1797	3508	816
5	agegroup	6.0	86140	0.785491	74353	4263	5370	2154
6	agegroup	7.0	62193	0.764716	46951	6974	4626	3642
7	agegroup	8.0	15098	0.748486	9194	2936	1235	1733
8	sex	1.0	198347	0.830416	173954	9746	9456	5191
9	sex	2.0	146957	0.818450	128760	7106	7551	3540
10	dialtyp	1.0	310415	0.816646	270848	15496	16115	7956
11	dialtyp	2.0	15082	0.850065	14758	44	248	32
12	dialtyp	3.0	13295	0.858981	12988	36	245	26
13	dialtyp	4.0	77	0.965753	70	3	1	3
14	dialtyp	100.0	6436	0.779859	4051	1273	398	714
15	race	1.0	230577	0.817986	196977	13823	12509	7268
16	race	2.0	93560	0.826123	85998	2552	3760	1250
17	race	3.0	3225	0.819874	3044	53	98	30
18	race	4.0	12965	0.845486	12063	325	436	141
19	race	5.0	3776	0.833047	3566	42	142	26
20	race	6.0	881	0.808297	772	48	46	15
21	race	9.0	321	0.789957	295	9	16	1
22	hispanic	1.0	51021	0.843191	47324	1198	1852	647
23	hispanic	2.0	292532	0.820216	254208	15364	15037	7923
24	hispanic	9.0	1752	0.790421	1183	290	118	161

https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning

CKD Example - Project Resources



Original Investigation

Kidney360

A Machine Learning Model for Predicting Mortality within 90 Days of Dialysis Initiation

Sommer Rankin, Lucy Han, Rebecca Scherzer, Susan Tenney, Matthew Keating, Kimberly Genberg, Matthew Rahn, Kenneth Wilkins, Michael Shlipak, and Michelle Estrella

Key Points

- This paper presents an eXtreme Gradient Boosting (XGBoost) model that predicted mortality in the first 90 days after dialysis initiation using data from the United States Renal Data System.
- Such a model could facilitate patient-clinician shared decision making on whether to initiate dialysis or pursue medical management.
- The XGBoost models discriminated mortality risk in both the nonimputed (c=0.826) and imputed (c=0.827) models.

Abstract

Background The first 90 days after dialysis initiation are associated with high morbidity and mortality in end-stage kidney disease (ESKD) patients. A machine learning-based tool for predicting mortality could inform patient-clinician shared decision making on whether to initiate dialysis or pursue medical management. We used the eXtreme Gradient Boosting (XGBoost) algorithm to predict mortality in the first 90 days after dialysis initiation in a nationally representative population from the United States Renal Data System.

- Main:
  - https://www.healthit.gov/topic/scientific-initiatives/pcor/machine-learning
- Blog Post:
  - https://www.healthit.gov/buzz-blog/health-it/the-application-of-machine-learning-to-address-kidney-disease
- Peer-reviewed publication
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9528387/
- Infographic
  - https://www.healthit.gov/sites/default/files/page/2021-09/ONC%20Training%20Data%20Project_Infographic-FINAL.pdf
- Code Repository:
  - https://github.com/onc-healthit/2021PCOR-ML-AI

Training Data for Machine Learning to Enhance PCOR Data Infrastructure

Project Webinar

September 15, 2021

Slido Quiz

1. Select the techniques that can be used to handle imbalanced data.
   a) Tiprapping
   b) Bootstrapping
   c) None. A model cannot be fit to imbalanced data
   d) Oversampling

2. Select the reason that interpretability in AI models is important for health domain.
   a) The weights can be compared to benchmarks
   b) The importance of the features can be analyzed
   c) A peer-reviewed paper can be published
   d) Trick question, it isn't important

ScHARe

Resources

ScHARe resources

Support made available to users:

ScHARe-specific

- ScHARe documentation
- Email support

Platform-specific

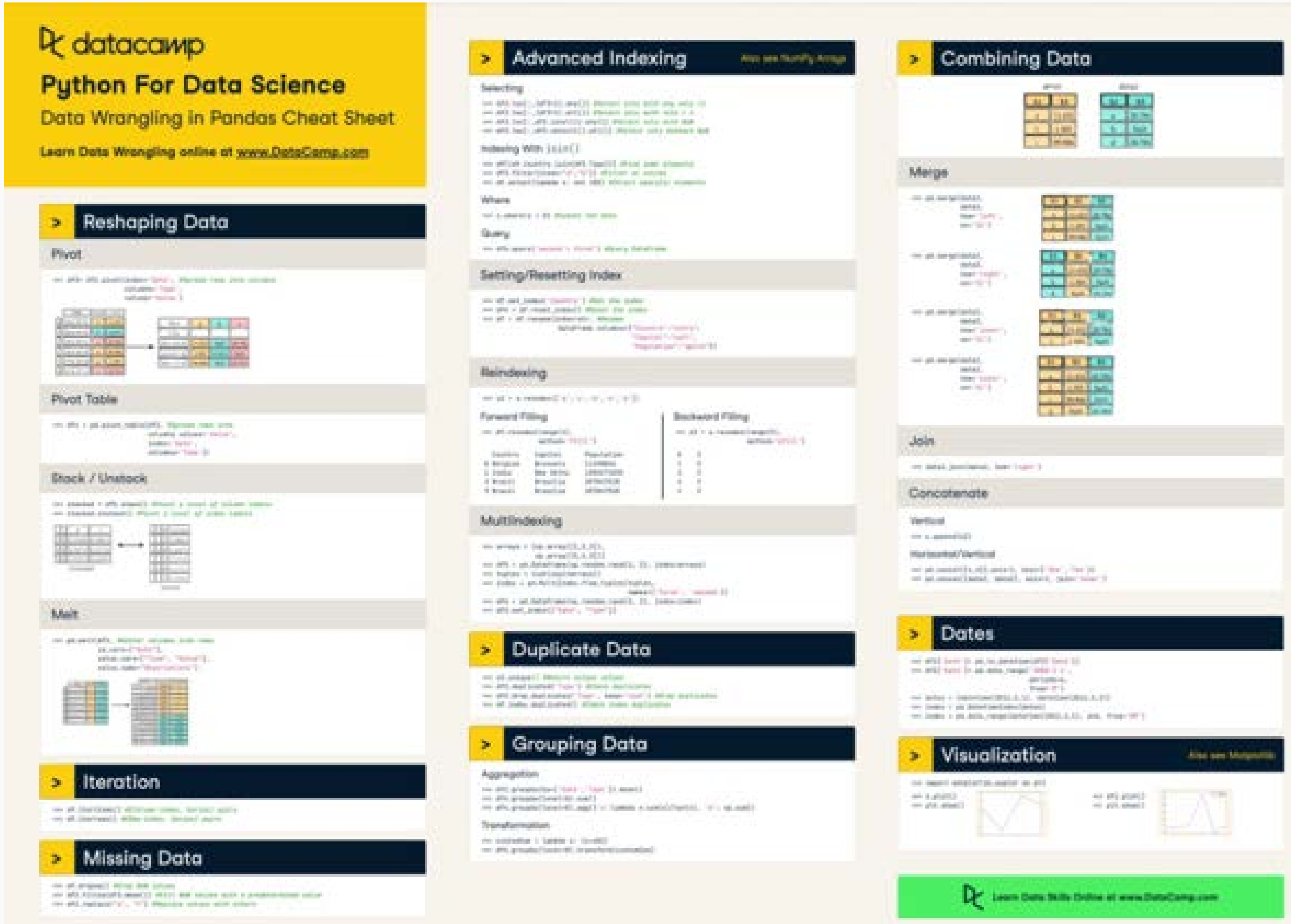
- Terra-specific support
- Terra-specific documentation

ScHARe resources

Training opportunities made available to users:

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

ScHARe resources: cheatsheets



datacamp

Python For Data Science

Data Wrangling in Pandas Cheat Sheet

Learn Data Wrangling online at www.DataCamp.com

Reshaping Data

Pivot

Pivot Table

Stack / Unstack

Melt

Iteration

Missing Data

Advanced Indexing

Also see NumPy Arrays

Selecting

Indexing With isin()

Where

Query

Setting/Resetting Index

Reindexing

Forward Filling

Backward Filling

MultiIndexing

Duplicate Data

Grouping Data

Aggregation

Transformation

Combining Data

Merge

Join

Concatenate

Vertical

Horizontal/Vertical

Dates

Visualization

Also see Matplotlib

Learn Data Skills Online at www.DataCamp.com

Credits: datacamp.com

Terra resources

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

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

ScHARe

Thank you

Think-a-Thon poll

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

□ Very useful

□ Useful

□ Somewhat useful

□ Not at all useful

Think-a-Thon poll

2. **Rate the pace of the instruction for yourself:**

□ Too fast

□ Adequate for me

□ Too slow

Think-a-Thon poll

3. **How likely will you participate in the next Think-a-Thon?**

□ Very interested, will definitely attend

□ Interested, likely will attend

□ Interested, but not available

□ Not interested in attending any others

ScHARe

Next Think-a-Thons:

bit.ly/think-a-thons

Register for ScHARe:

bit.ly/join-schare

✉ schare@mail.nih.gov