ScHARe

Think-a-Thons

NIH National Institutes of Health

Computational Data Science Strategies

Getting Ready for a Data Science 101 Course

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

January 17, 2024

ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

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

Thank you

NIMHD

Dr. Eliseo Perez-Stable

ODSS

Dr. Susan Gregurick

NIH/OD

Dr. Larry Tabak

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

CDE Working Group

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

Experience poll

Please check your level of experience with the following:

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

ScHARe

Overview

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

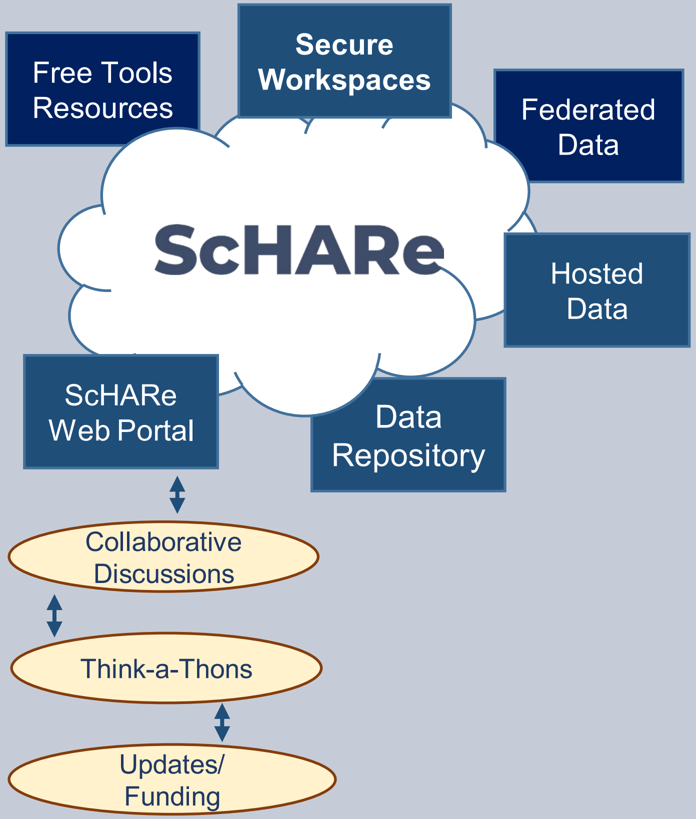
ScHARe Components

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)*

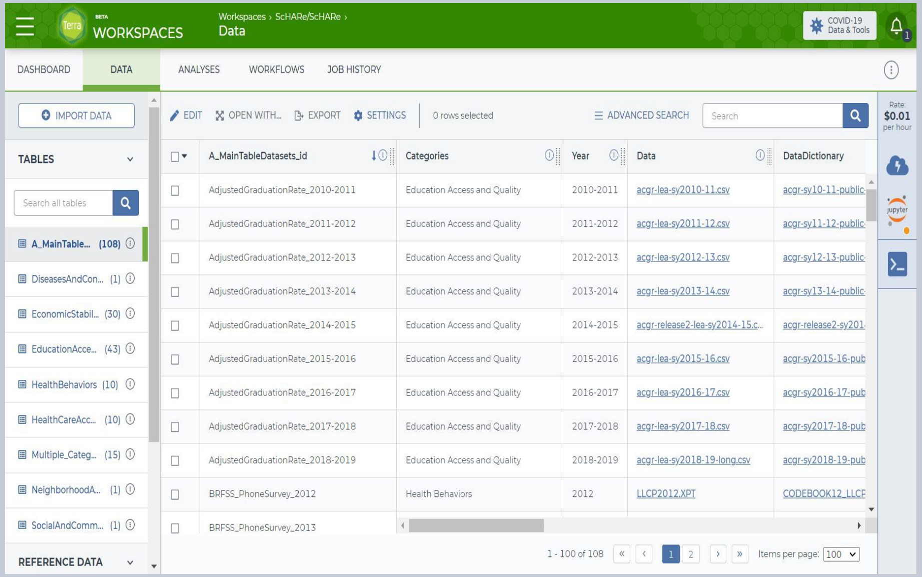
- **ScHARe Hosted Public Datasets:** publicly accessible, de-identified datasets hosted by ScHARe

  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*

OVER 240 DATA SETS CENTRALIZED



0 rows selected

A_MainTable... (108)
DiseasesAndCon... (1)
EconomicStabil... (30)
EducationAcce... (43)
HealthBehaviors (10)
HealthCareAcc... (10)
Multiple_Categ... (15)
NeighborhoodA... (1)
SocialAndComm... (1)

A_MainTableDatasets_id	Categories	Year	Data	DataDictionary
AdjustedGraduationRate_2010-2011	Education Access and Quality	2010-2011	acgr-lea-sy2010-11.csv	acgr-sy10-11-public-
AdjustedGraduationRate_2011-2012	Education Access and Quality	2011-2012	acgr-lea-sy2011-12.csv	acgr-sy11-12-public-
AdjustedGraduationRate_2012-2013	Education Access and Quality	2012-2013	acgr-lea-sy2012-13.csv	acgr-sy12-13-public-
AdjustedGraduationRate_2013-2014	Education Access and Quality	2013-2014	acgr-lea-sy2013-14.csv	acgr-sy13-14-public-
AdjustedGraduationRate_2014-2015	Education Access and Quality	2014-2015	acgr-release2-lea-sy2014-15.c...	acgr-release2-sy201
AdjustedGraduationRate_2015-2016	Education Access and Quality	2015-2016	acgr-lea-sy2015-16.csv	acgr-sy2015-16-pub
AdjustedGraduationRate_2016-2017	Education Access and Quality	2016-2017	acgr-lea-sy2016-17.csv	acgr-sy2016-17-pub
AdjustedGraduationRate_2017-2018	Education Access and Quality	2017-2018	acgr-lea-sy2017-18.csv	acgr-sy2017-18-pub
AdjustedGraduationRate_2018-2019	Education Access and Quality	2018-2019	acgr-lea-sy2018-19-long.csv	acgr-sy2018-19-pub
BRFSS_PhoneSurvey_2012	Health Behaviors	2012	LLCP2012.XPT	CODEBOOK12_LLCP
BRFSS_PhoneSurvey_2013				

1 - 100 of 108 « ‹ 1 2 › » Items per page: 100

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

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

with the addition of:

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

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

ScHARe Data Ecosystem Structure

FEDERATED PUBLIC DATA 240+

Hosted by Google & ScHARe

REPOSITORY

<u>CDE FOCUSED</u>

CDEs enhances Data Interoperability (Aggregation) by using semantic standards and concept codes

Innovative Approach:

CDE Concept Codes Uniform Resource Identifier (URI)

What is a CDE?

1. Name Last
First
2. Date of Birth (mm/dd/yyyy)
5. Social Security Number
dress: Line 1: Street/RFD#, P.O. Box,
learly label Apartment, Company, Su

A common data element (CDE) is a standardized, precisely defined question that is paired with a set of specific allowable responses, that is then used systematically across different sites, studies, or clinical trials to ensure consistent data collection

ScHARe CDEs Labels

For FUNDED PROJECT DATA – Common Data Elements Centralized for Interoperability and Data Sharing

- Age
- Birthplace
- Zip Code
- Race and Ethnicity
- Sex
- Gender
- Sexual Orientation
- Marital Status
- Education
- Annual Household Income
- Household Size

- English Proficiency
- Disabilities
- Health Insurance
- Employment Status
- Usual Place of Health Care
- Financial Security / Social Needs
- Self Reported Health
- Health Conditions (Associated Medications/Treatments)

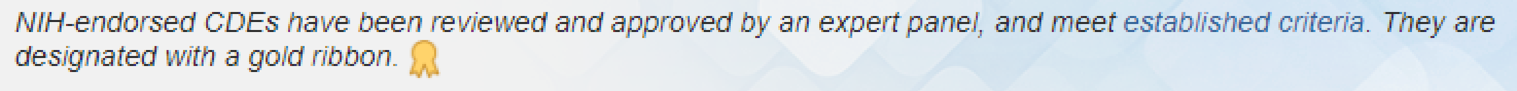
**NIMHD Framework
**Health Disparity Outcomes

NIH Endorsed

(** project level CDE)

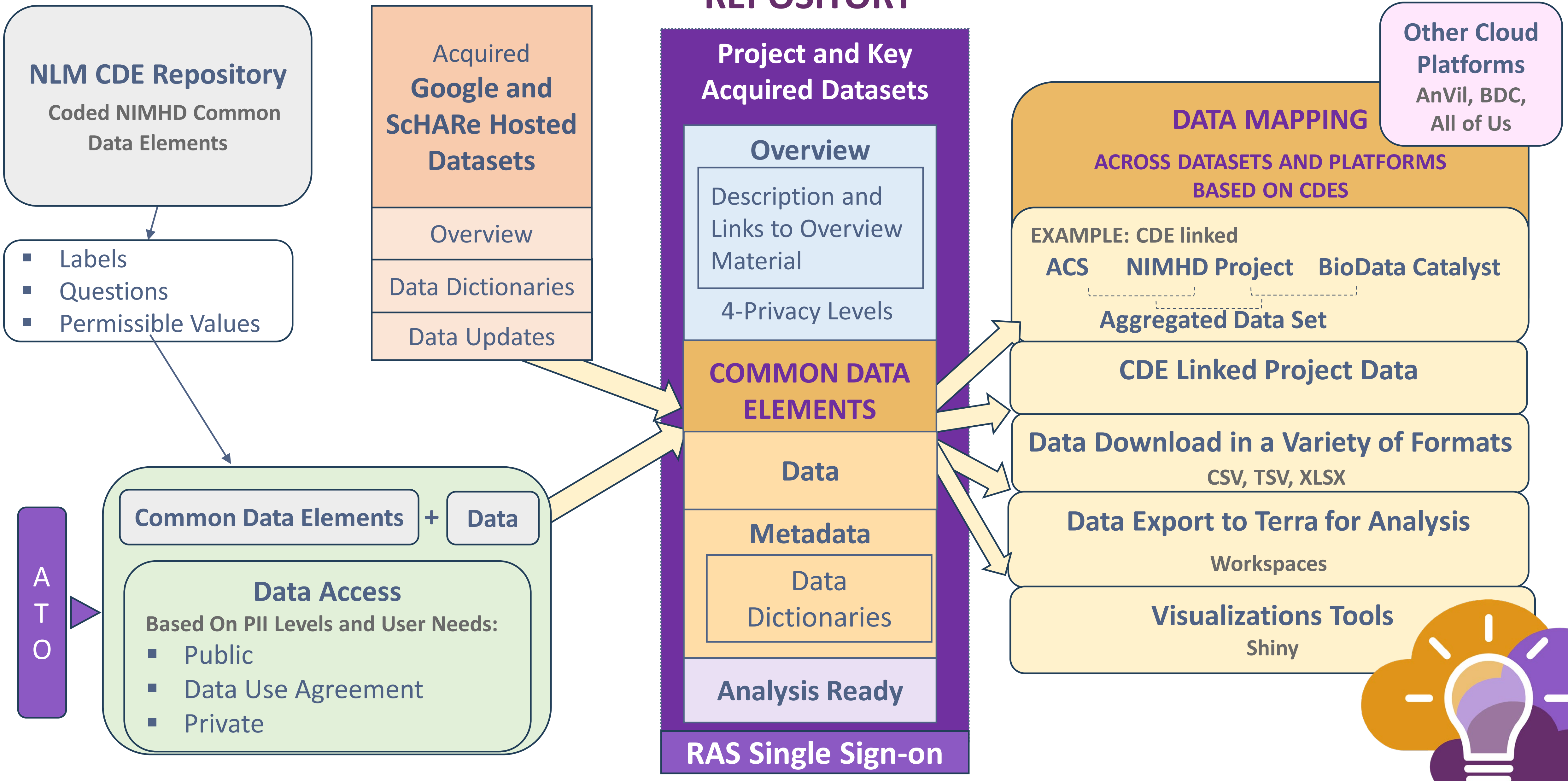
NIH CDE Repository: https://cde.nlm.nih.gov/home

Cross-walked with PhenX SDoH



NIH-endorsed CDEs have been reviewed and approved by an expert panel, and meet established criteria. They are designated with a gold ribbon. 🎗

ScHARe REPOSITORY



COMMON DATA ELEMENTS

NLM CDE Repository

Coded NIMHD Common Data Elements

- Labels
- Questions
- Permissible Values

Common Data Elements + **Data**

Data Access

Based On PII Levels and User Needs:

- Public
- Data Use Agreement
- Private

A T O

DATA UPLOAD

Acquired **Google and ScHARe Hosted Datasets**

Overview

Data Dictionaries

Data Updates

ScHARe REPOSITORY

Project and Key Acquired Datasets

Overview

Description and Links to Overview Material

4-Privacy Levels

COMMON DATA ELEMENTS

Data

Metadata

Data Dictionaries

Analysis Ready

RAS Single Sign-on

DATA MAPPING, DOWNLOAD AND EXPORT

Other Cloud Platforms

AnVil, BDC, All of Us

DATA MAPPING

ACROSS DATASETS AND PLATFORMS BASED ON CDES

EXAMPLE: CDE linked

ACS NIMHD Project BioData Catalyst

Aggregated Data Set

CDE Linked Project Data

Data Download in a Variety of Formats

CSV, TSV, XLSX

Data Export to Terra for Analysis

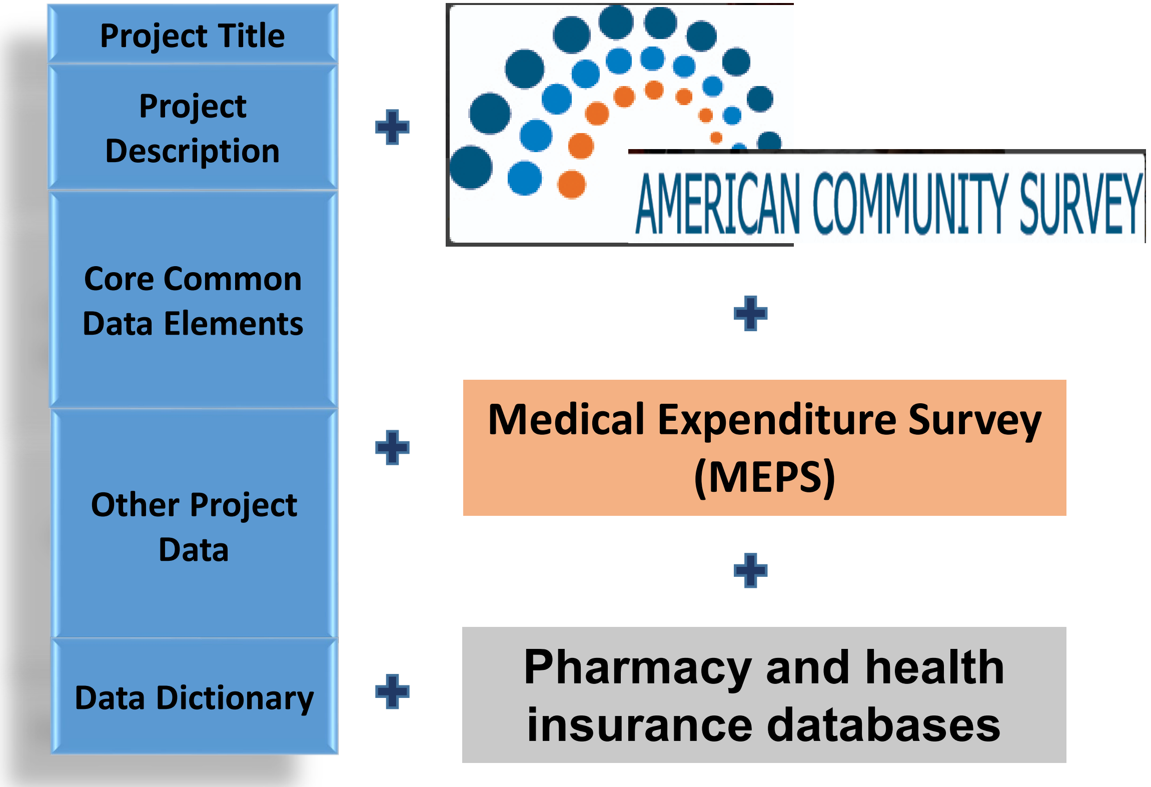
Workspaces

Visualizations Tools

Shiny

ScHARe

Project & federated dataset mapping



Project Title

Project Description

Core Common Data Elements

Other Project Data

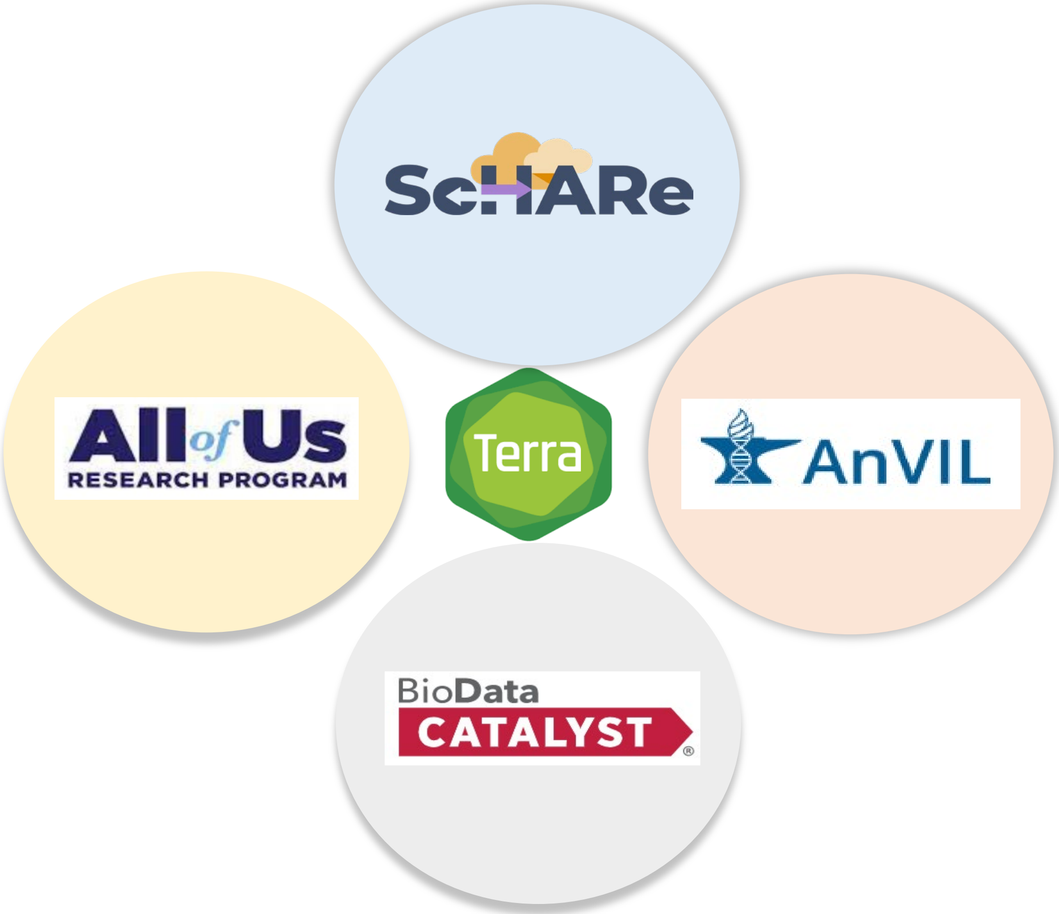
Data Dictionary

+ AMERICAN COMMUNITY SURVEY

+ Medical Expenditure Survey (MEPS)

+ Pharmacy and health insurance databases

Mapping across cloud platforms

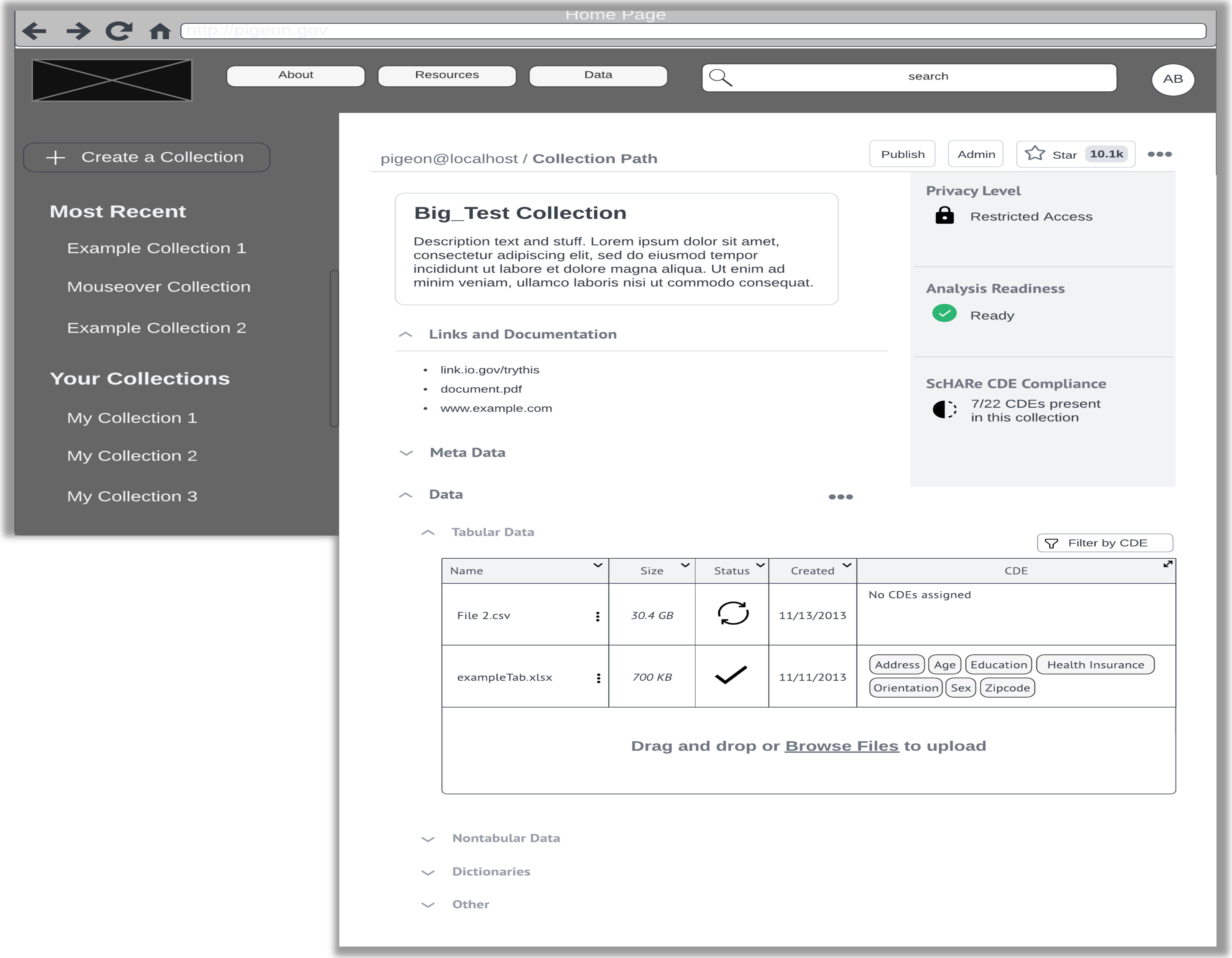


UPCOMING

ScHARe

Repository CDE Focused for Data Interoperability

Coming Soon



Home Page

http://pigeon.gov

About | Resources | Data | search | AB

+ Create a Collection

Most Recent

- Example Collection 1
- Mouseover Collection
- Example Collection 2

Your Collections

- My Collection 1
- My Collection 2
- My Collection 3

pigeon@localhost / **Collection Path**

Publish | Admin | ☆ Star 10.1k | •••

Big_Test Collection

Description text and stuff. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, ullamco laboris nisi ut commodo consequat.

Privacy Level

Restricted Access

Analysis Readiness

✓ Ready

ScHARe CDE Compliance

7/22 CDEs present in this collection

Links and Documentation

- link.io.gov/trythis
- document.pdf
- www.example.com

Meta Data

Data

Tabular Data

Filter by CDE

Name	Size	Status	Created	CDE
File 2.csv	30.4 GB	⟳	11/13/2013	No CDEs assigned
exampleTab.xlsx	700 KB	✓	11/11/2013	Address Age Education Health Insurance Orientation Sex Zipcode

Drag and drop or Browse Files to upload

Nontabular Data

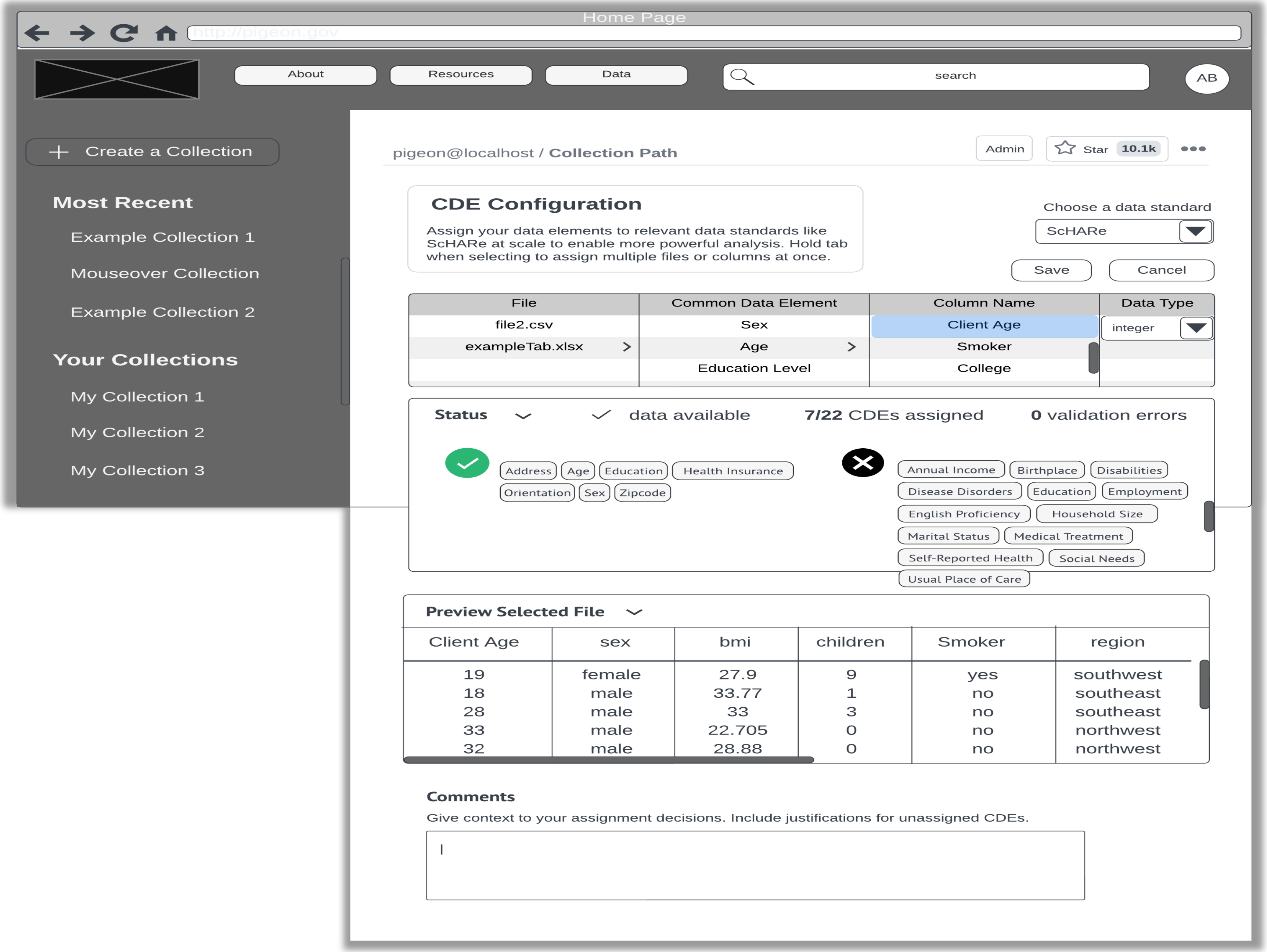
Dictionaries

Other

ScHARe

Repository CDE Focused for Data Interoperability

Coming Soon



pigeon@localhost / **Collection Path**

CDE Configuration

Assign your data elements to relevant data standards like ScHARe at scale to enable more powerful analysis. Hold tab when selecting to assign multiple files or columns at once.

Choose a data standard: ScHARe ▼

Save | Cancel

File	Common Data Element	Column Name	Data Type
file2.csv	Sex	Client Age	integer ▼
exampleTab.xlsx >	Age >	Smoker	
	Education Level	College	

Status ﹀ ✓ data available **7/22** CDEs assigned **0** validation errors

✓ Address | Age | Education | Health Insurance | Orientation | Sex | Zipcode

✕ Annual Income | Birthplace | Disabilities | Disease Disorders | Education | Employment | English Proficiency | Household Size | Marital Status | Medical Treatment | Self-Reported Health | Social Needs | Usual Place of Care

Preview Selected File ﹀

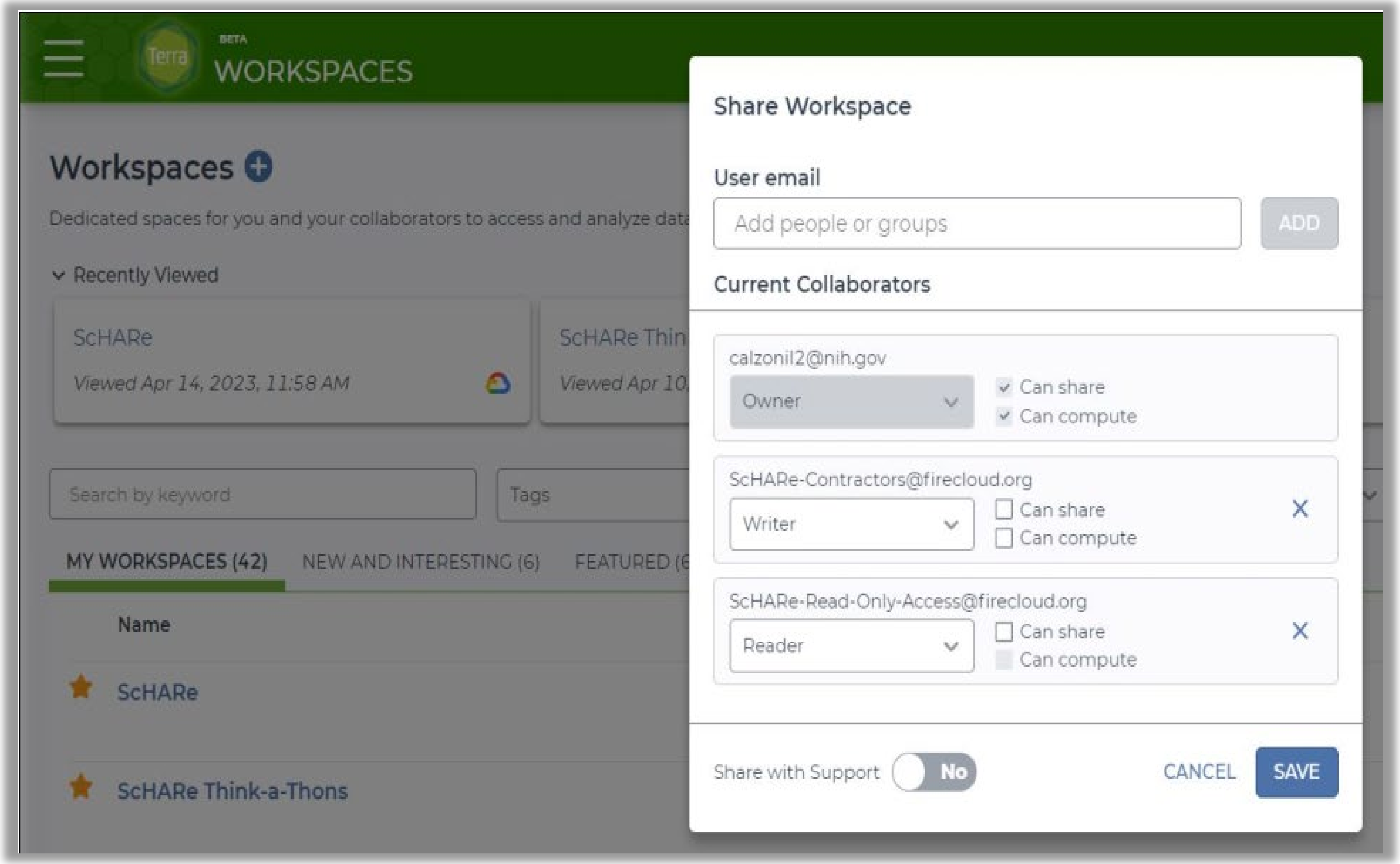
Client Age	sex	bmi	children	Smoker	region
19	female	27.9	9	yes	southwest
18	male	33.77	1	no	southeast
28	male	33	3	no	southeast
33	male	22.705	0	no	northwest
32	male	28.88	0	no	northwest

Comments

Give context to your assignment decisions. Include justifications for unassigned CDEs.

|

Secure workspace



Workspaces

Dedicated spaces for you and your collaborators to access and analyze data

Recently Viewed

ScHARe
Viewed Apr 14, 2023, 11:58 AM

ScHARe Thin
Viewed Apr 10,

ScHARe

ScHARe Think-a-Thons

Share Workspace

User email

Add people or groups

Current Collaborators

calzonil2@nih.gov
Owner
☑ Can share
☑ Can compute

ScHARe-Contractors@firecloud.org
Writer
□ Can share
□ Can compute

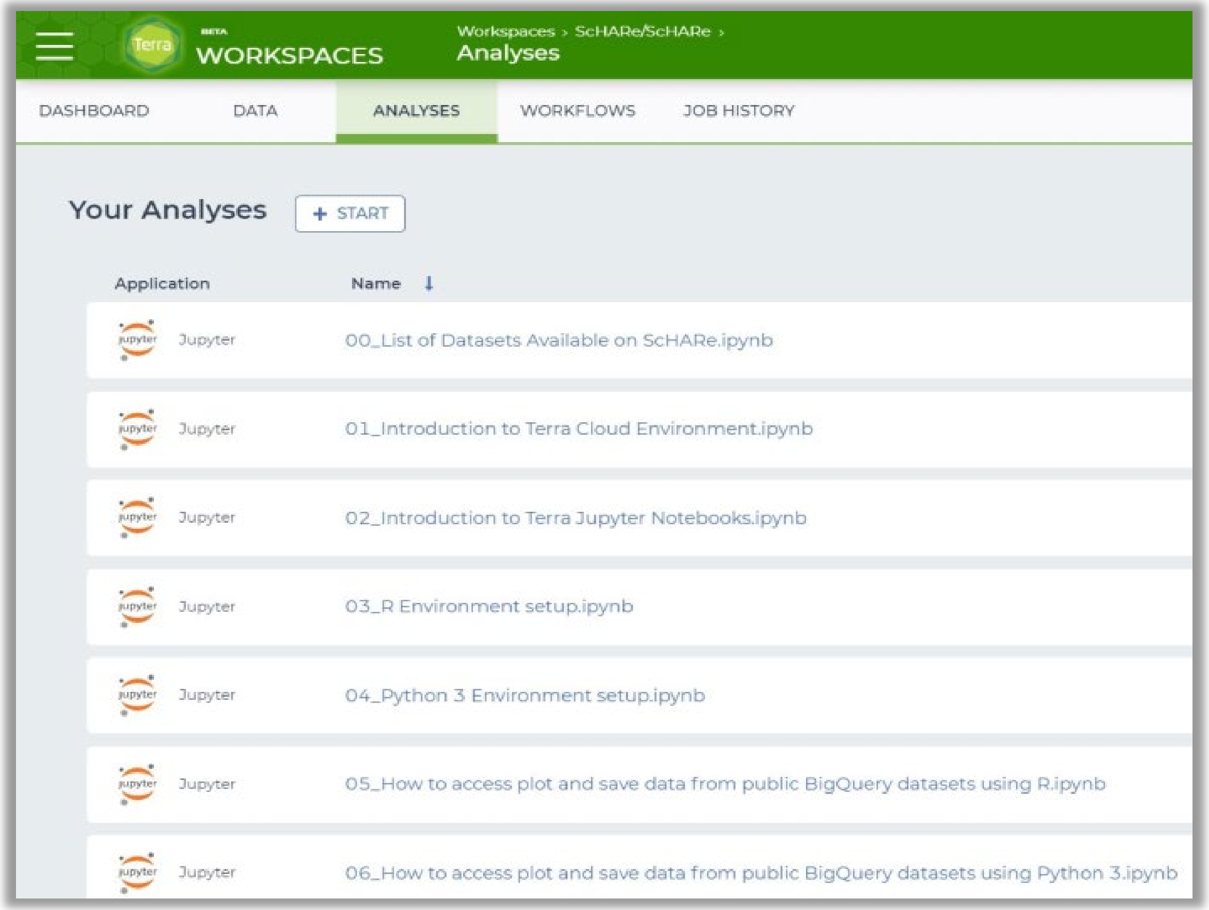
ScHARe-Read-Only-Access@firecloud.org
Reader
□ Can share
□ Can compute

Share with Support: No

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

Notebooks analytics

Workflows - Modular codes



Terra BETA WORKSPACES

Workspaces › ScHARe/ScHARe ›

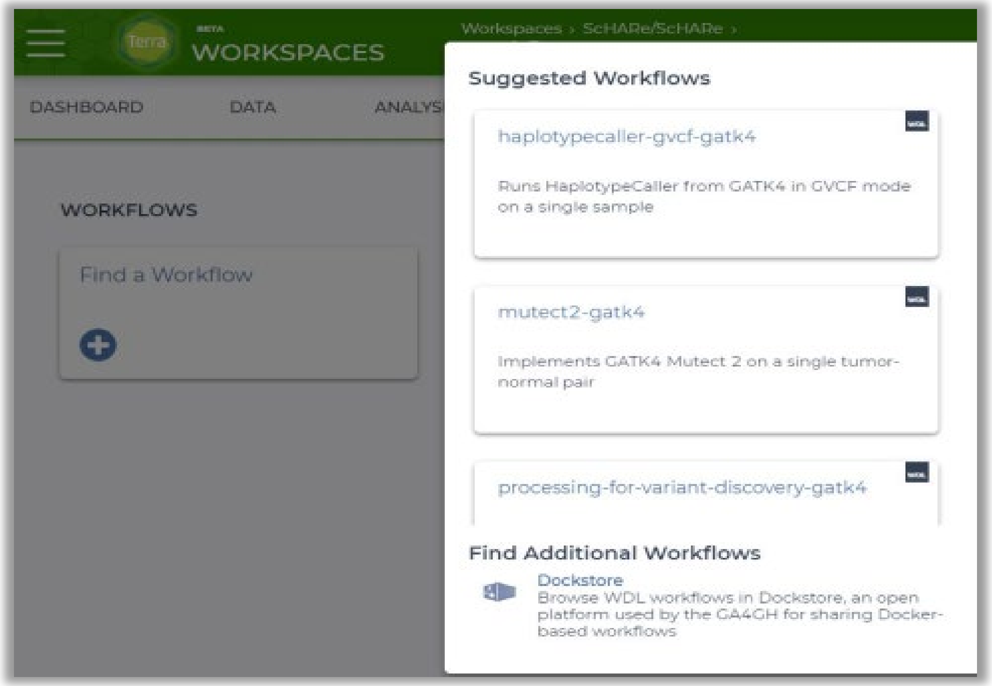
Analyses

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

Your Analyses + START

Application	Name ↓
Jupyter	00_List of Datasets Available on ScHARe.ipynb
Jupyter	01_Introduction to Terra Cloud Environment.ipynb
Jupyter	02_Introduction to Terra Jupyter Notebooks.ipynb
Jupyter	03_R Environment setup.ipynb
Jupyter	04_Python 3 Environment setup.ipynb
Jupyter	05_How to access plot and save data from public BigQuery datasets using R.ipynb
Jupyter	06_How to access plot and save data from public BigQuery datasets using Python 3.ipynb

- **Copy and paste analytics**



Terra BETA WORKSPACES

Workspaces › ScHARe/ScHARe ›

DASHBOARD DATA ANALYSES

WORKFLOWS

Find a Workflow

Suggested Workflows

haplotypecaller-gvcf-gatk4

Runs HaplotypeCaller from GATK4 in GVCF mode on a single sample

mutect2-gatk4

Implements GATK4 Mutect 2 on a single tumor-normal pair

processing-for-variant-discovery-gatk4

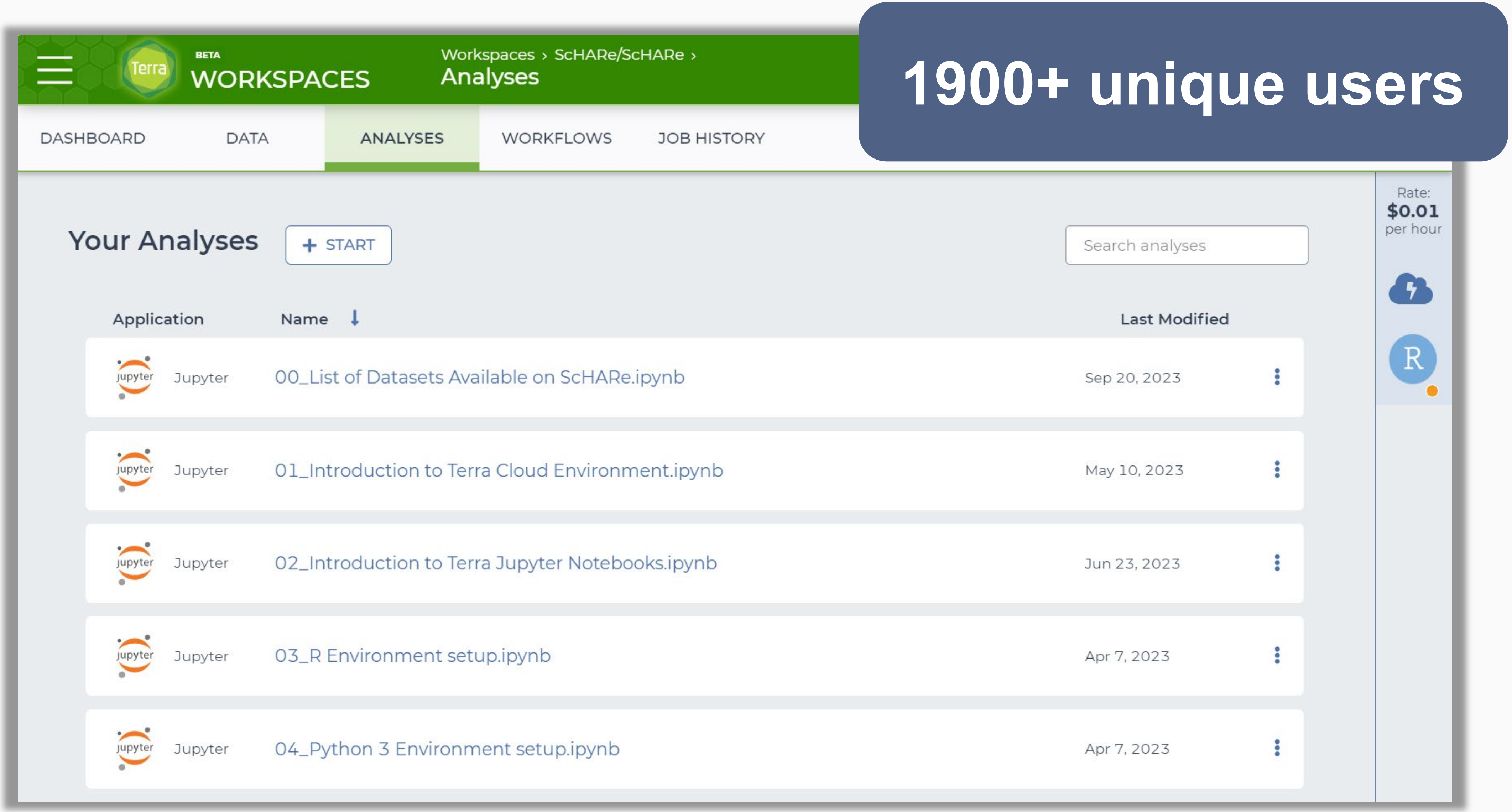
Find Additional Workflows

Dockstore

Browse WDL workflows in Dockstore, an open platform used by the GA4GH for sharing Docker-based workflows

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

ScHARe Registrations



BETA WORKSPACES

Workspaces › ScHARe/ScHARe ›

Analyses

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

Your Analyses + START

Search analyses

Application	Name	Last Modified
Jupyter	00_List of Datasets Available on ScHARe.ipynb	Sep 20, 2023
Jupyter	01_Introduction to Terra Cloud Environment.ipynb	May 10, 2023
Jupyter	02_Introduction to Terra Jupyter Notebooks.ipynb	Jun 23, 2023
Jupyter	03_R Environment setup.ipynb	Apr 7, 2023
Jupyter	04_Python 3 Environment setup.ipynb	Apr 7, 2023

Rate: $0.01 per hour

1900+ unique users

ScHARe

Think-a-Thons

National Institutes of Health

Think-a-Thon Tutorials

February	**Artificial Intelligence and Cloud Computing 101**
March	**ScHARe 1 – Accounts and Workspaces**
April	**ScHARe 2 – Terra Datasets**
May	**ScHARe 3 – Terra Google-hosted Datasets**
	ScHARe for Educators (Community Colleges & Low Resource MSIs)
June	**ScHARe 4 – Terra ScHARe-hosted Datasets**
July	**An Introduction to Python for Data Science – Part 1**
August	**An Introduction to Python for Data Science – Part 2**
	ScHARe for American Indian / Alaska Native Researchers
September	**ScHARe 5: A Review of the ScHARe Platform and Data Ecosystem**
October	**Preparing for AI 1: Common Data Elements and Data Aggregation**
November	**Preparing for AI 2: An Introduction to FAIR Data and AI-ready Datasets**
January	**Preparing for AI 3: Computational Data Science Strategies 101**
	ScHARe for Coders and Programmers to conduct Research (Jan 31)

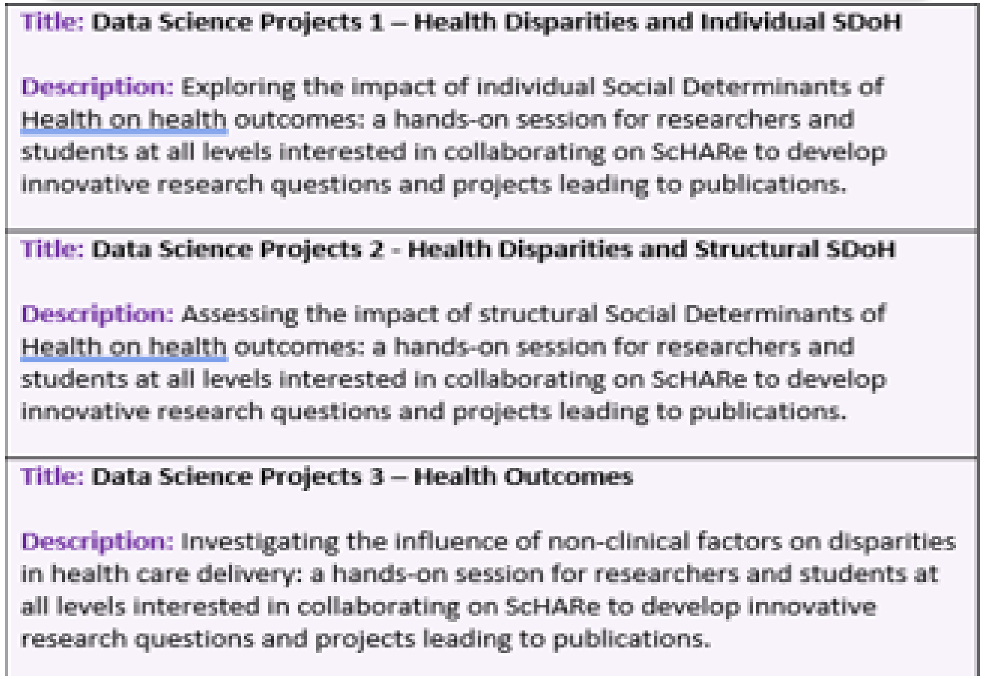
bit.ly/think-a-thons

Upcoming

ScHARe

Think-a-Thons (TaT)

Research Teams



Title: Data Science Projects 1 – Health Disparities and Individual SDoH

Description: Exploring the impact of individual Social Determinants of Health on health outcomes: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.

Title: Data Science Projects 2 - Health Disparities and Structural SDoH

Description: Assessing the impact of structural Social Determinants of Health on health outcomes: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.

Title: Data Science Projects 3 – Health Outcomes

Description: Investigating the influence of non-clinical factors on disparities in health care delivery: a hands-on session for researchers and students at all levels interested in collaborating on ScHARe to develop innovative research questions and projects leading to publications.

- **Foster a research paradigm shift to use Big Data**
- **Promote use of Dark Data**

- Multi-career (students to sr. investigators)
- Multi-discipline (data scientist & researchers)
- Feature Datasets with Guest Expert Leads
- Secure experts in topic area, analytics, data sources etc. to provide guidance
- Generate research idea - decide potential design, datasets & analytics
- Select co-leads to coordinate completion outside of TaT
- Publications

Register:

bit.ly/think-a-thons

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 expert

Kenneth J. Wilkins, PhD

NIH/NIDDK

About Ken

Ken is a former mathematics and computer science high school teacher who found his way into biostatistics.

He worked for two decades across sectors in biomedical research, and he is now working with both NIH-employed intramural and NIH-funded extramural researchers in his NIH/NIDDK and trans-NIH roles.

His research interests encompass evolving data methods to better suit researchers’ posed questions given limitations in data and data-interoperability standards.

National Institute of Diabetes and Digestive and Kidney Diseases

ScHARe Think-a-thon Preparing for AI 3: Computational Data Science Strategies 101

Ken Wilkins, PhD
Biostatistics Program, Office of Clinical Research
Data Science Working Group, Office of the Director
National Inst. of Diabetes & Digestive & Kidney Diseases, NIH

National Institute of Diabetes and Digestive and Kidney Diseases

Overview: *a whistlestop tour of a landscape*

- Understanding the Landscape
- Traditional Statistics & Epidemiologic Methods as Baseline
- Artificial Intelligence in Data Science as Broad New Horizon
- Machine Learning Unveiled as a Bridge-building Trailblazer
- Python Libraries for Data Science Computational Strategies
- Ongoing Resources and Decision-Making Tools to use as a Guide
- Q&A and Closing Remarks

National Institute of Diabetes and Digestive and Kidney Diseases

Microsoft 365 Stock Photos; https://media.amtrak.com/wp-content/uploads/2021/05/Amtrak-2021-Corridor-Vision-May27_2021.pdf

Again a ‘whistlestop’ rather than a ‘whirlwind’ tour…

O'REILLY

A Whirlwind Tour of Python

Jake VanderPlas

ScHARe

Science Collaborative for Health disparities and Artificial intelligence bias REduction

Understanding the Landscape

https://www.digitscotland.com/what-is-landscape-surveying-recording/

A. Definitions and Differentiations

1) Preliminaries to get everyone on the same page
2) Context while getting our lay of the land: **health disparities**

B. Decision-Making Framework: early teaser… hard to decide *which* tools without a few things in toolbox

…will use above ‘alarm’ icon to trigger our need to “unpack” some ‘jargon’ terms

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Understanding the Landscape: Preliminaries

- *Consider yourself as a data science practitioner: be* practical *on what to use!*
  - *“data science”: coin termed by a statistician, adopted by computer science/informatics*
  - *Most recently viewed as an ‘interdiscipline’ –interdisciplinary/metadisciplinary nature*
  - *‘practical’ means bringing the most effective tool(s) for the task(s) at hand*
  - *We cover computational strategies ranging from traditional to modern statistics and epidemiologic methods, and where these don’t meet needs: AI & machine learning*
  - We cover working definitions of above, ahead of diving in… but we also bear in mind…

- Context of ScHARe goals of working toward **health disparities** (*primal aim*)
  - *“The aim of the ScHARe program is to increase participation of people from underrepresented populations in data science and cloud computing so that everyone can benefit from the research opportunities afforded by Big Data.”*

Understanding the Landscape: Preliminaries

- *Consider yourself as a data science practitioner: be a* scientist i*n what you do!*
  - *"data science": science as the practice of adding to 'generalizable knowledge'*
  - *Scientists ought to maintain awareness of their 'blind spots': tacit assumptions in data*
  - *Consider how you must check your assumptions… how did data come to be at hand?*
  - *This 'design behind the data' hearken back to 'Research Design' of prior TaT session*
  - We cover working definitions of above, ahead of diving in… but we also bear in mind…
- Context of ScHARe **aims**
  - *Increase participation of women and 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 health disparity, and health and healthcare delivery outcomes research.*
  - *Advance AI bias mitigation and ethical inquiry by developing innovative strategies and securing diverse perspectives.*

Getting our lay of the land: **health disparities**

the lay of the land noun phrase **(US idiom)**

: the arrangement of the different parts in an area of land **:** where things are located in a place - She knew the *lay of the land* from hiking through it daily.

—often used figuratively

It takes time for new employees to get *the lay of the land* in this department.

https://www.merriam-webster.com/dictionary/the%20lay%20of%20the%20land

- Context of ScHARe goals
- **Decreasing Health Disparities – 'dual' problem of mitigating extant biases**

  The primal and dual are two sides of the same coin, with the primal being the original problem and the dual being the derived problem.

- *Mitigating Bias: does it mean the same thing to all parties?*
  - *not necessarily: varied forms of each type of 'bias' ought to be considered*
    - *Bias in perspective/experience (confirmation bias), bias in data available (selection bias), &c.*
  - *Theoretical behavior of data methods: 'bias' if estimates differ from target*
    - *Often referred to as 'statistical bias' – follows from any quantity derived from data being a 'statistic'*
  - *Practical applications to data: inherent imbalances of data's sources →* ***algorithmic bias***
    - *One distinction as written by AI/ML researchers: "In contrast to human bias, algorithmic bias occurs when an AI model, trained on a given data set, produces results that may be completely unintended by the model creators." – Chen, Szolovits, & Ghassemi 2019, AMA Journal of Ethics*

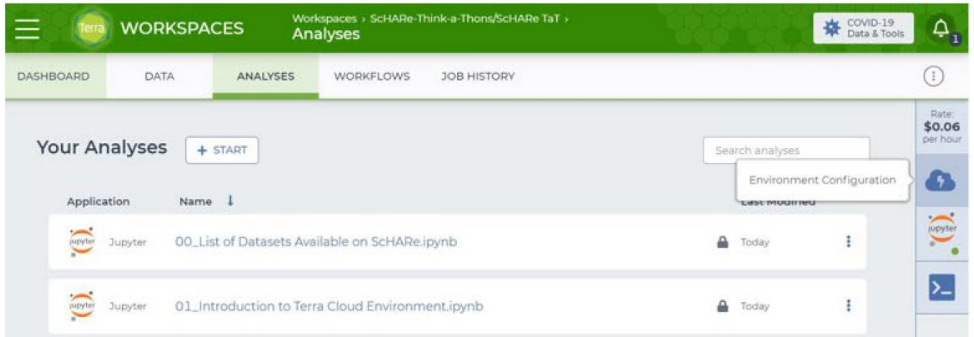
Getting our lay of the land: **health disparities**

- Context of ScHARe goals, while getting our lay of the land: health disparities
- *As a data scientist, you can have* ***agency*** *in some sources of bias*
  - If you lack individual-level features that 'explain' source of bias, use *supplements*
  - *Supplements easier to get with data linkage (e.g., ZIP code for area-level proxies)*
  - ***Ultimately:*** *some features need careful prep, others will be 'missing' (still recognize)*
    - ***Data prep:*** **numeric** *form of features used in algorithms, possible '***weighting***' for missed features*
    - ***Teaser of decision-making framework:*** *can't decide tools to use without actual toolbox… ScHARe@Terra*
    - ***NOTE:*** *today will NOT involve live hands-on work*
      - ❖ ***We have a lot to cover conceptually, prior to coding***
      - ❖ ***Concepts can be reinforced by experiential learning***

Experiencing
Reflecting
Thinking
Acting

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

bit.ly/schare-tat



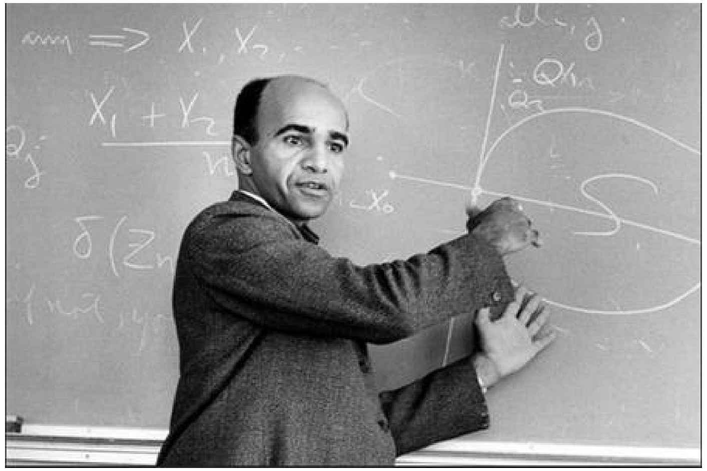
Workspaces › ScHARe-Think-a-Thons/ScHARe TaT › Analyses

Rate: $0.06 per hour

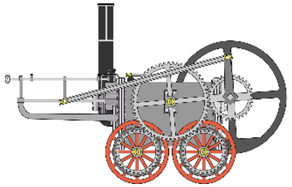
Your Analyses

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

Traditional Statistics & Epidemiologic Methods as a Baseline



A. Simpler, straightforward data summaries



B. More complex modern modeling / exploration:
early forms of machine learning and AI…

National Institute of Diabetes and Digestive and Kidney Diseases

Hypatia of Alexandria (c.335-415BCE) c/o https://www.theutcecho.com/opinion/hypatia-the-first-known-woman-in-stem/article_2f043adc-9dac-11ec-843c-8baacd9ceb64.html; David H. Blackwell (1919-2010) c/o https://ww3.math.ucla.edu/david-harold-blackwell-summer-research-institute/ | also https://en.wikipedia.org/wiki/Locomotive | https://en.wikipedia.org/wiki/Shinkansen

Traditional Statistics & Epidemiologic Methods as a Baseline

- My own take: I'd *not pursued* statistics because of 'stats class':
  - *As HS math teacher, got question: where is math useful?*

Maggie L.Walker
GOVERNOR'S SCHOOL
for Government and International Studies

- *'traditional' statistics class seemed to me like laundry list of 'recipes'*
  - Can be **very dry material** when divorced from its motivating context: using data!
  - Adopt the 'interdisciplinary' view, like John Tukey (coined terms 'bit', 'software')
  - [paraphrase] statisticians (data scientists) get to play in everyone's 'back yard'

- **For you as data scientists: use 'modern' stats (if not AI/ML) methods**
  - Demonstrated to outperform deep learning in tabular structured health data
  - That said, be prepared for *multimodal* data, to *combine* stats with AI/ML

"Multimodal"
Multiple types of data (numeric, image, text) whose information is tied together

Data Methods, Overall: Fundamental Role of **Algorithms**

Machine learning algorithms are the engines of machine learning, meaning it is the algorithms that turn a data set into a model. Which kind of algorithm works best (supervised, unsupervised, classification, regression, etc.) depends on the kind of problem you're solving, the computing resources available, and the nature of the data. Uncovering patterns rather than carrying out a pre-defined task can yield surprising and useful results

How is an AI algorithm made?

At the core level, an AI algorithm takes in training data (labeled or unlabeled, supplied by developers, or acquired by the program itself) and uses that information to learn and grow. Then it completes its tasks, using the training data as a basis.

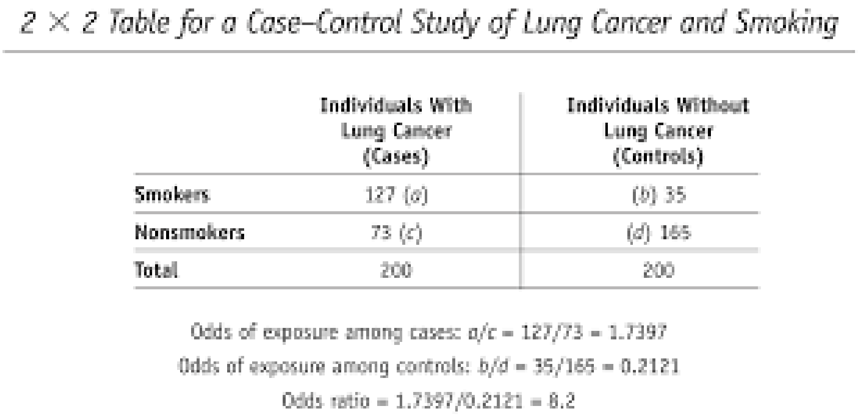
Algorithms: AI algorithms are the core mathematical and computational instructions that enable AI systems to process and analyze data. These algorithms include machine learning, deep learning, reinforcement learning, natural language processing (NLP), and many more.

Traditional Stats & Epi Methods: Simple data summaries

- Easy ‘rule of thumb’ (pun intended):
  - *can you count quantities involved on one hand (or even two)?*
- *If yes, the more ‘traditional’ statistics & epi methods will suffice*
  - Estimates with accompanying quantities that convey uncertainty
  - Many still can be done ‘by hand’…you will learn later to do in 1 line of code
  - Example, important to health disparities, to follow on next slide
- **If not, may need more modern stats/epi methods (if not AI/ML)**
  - Includes methods of regression / statistical learning that have bled into AI/ML
  - These regularly involve special preparation of data to use (later examples)

Traditional Stats & Epi Methods: Simple data summary *examples*

- Epidemiologic simple data summaries:
  - *Typically used in health outcome events to measure association with ‘risk factors’*
  - *Some useful for quantifying disparities, like odds and odds ratios (see 2×2 table @ right)*
  - ***Association ≠ Causation**, bear in mind*
- Continuous measures: mean, median
  - Get a sense of variability around these with standard deviation, interquartile range
  - Can also look at ‘co’-variation, like Pearson’s Correlation Coefficient, estimated by ‘*r*’



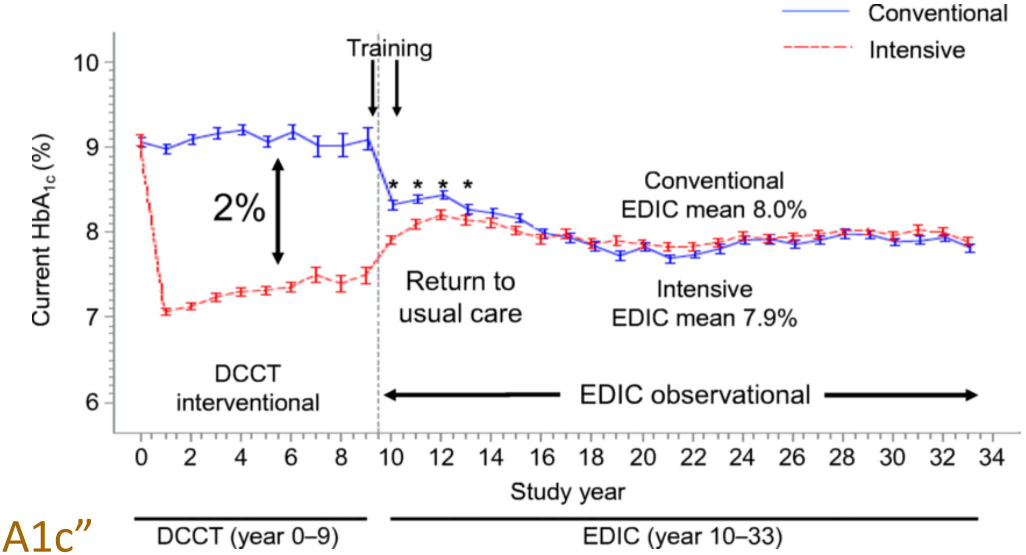
2 × 2 Table for a Case–Control Study of Lung Cancer and Smoking

	Individuals With Lung Cancer (Cases)	Individuals Without Lung Cancer (Controls)
Smokers	127 (*a*)	(*b*) 35
Nonsmokers	73 (*c*)	(*d*) 165
Total	200	200

Odds of exposure among cases: $a/c = 127/73 = 1.7397$

Odds of exposure among controls: $b/d = 35/165 = 0.2121$

Odds ratio = $1.7397/0.2121 = 8.2$



“HbA1c”

Long-term (~3 month) measure of blood sugar: proxy for control of diabetes

http://individual.utoronto.ca/ahmed_3/index_files/NSU/Epi2_3.pdf
http://dx.doi.org/10.1007/s00125-021-05397-4

Traditional Stats & Epi Methods: Simple data summary *pitfalls*

- Easy 'pitfall' with simple data summaries:
  - *Tendency to draw inferences **without** considering influence of variables NOT included, such as socioeconomic advantages*
  - ***Correlation* $\neq$ *Causation**, bear in mind*
  - *Article at right does consider, just not fully*
  - *Discussion by numerous others give caveats*
  - *Lost chance at using regression to 'adjust'*
- **Even with more features or variables used, still is a pitfall**
  - Remains a risk for methods of regression / statistical learning that have bled into AI/ML

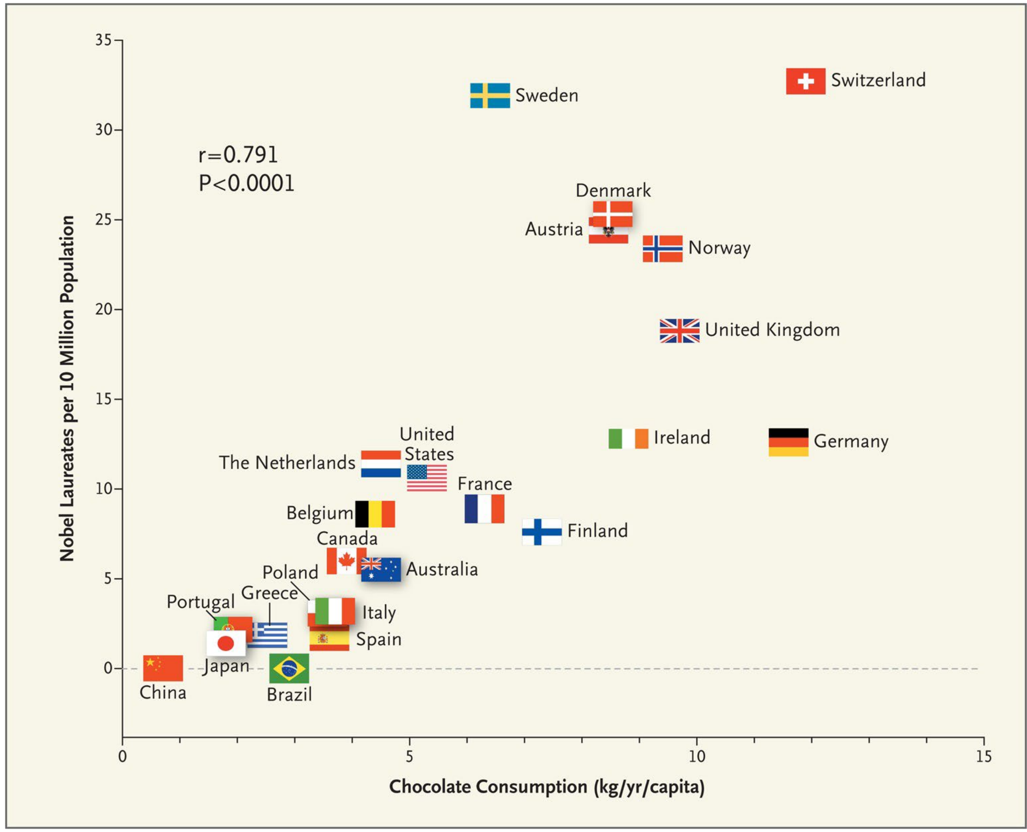


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.
https://www.nejm.org/doi/full/10.1056/NEJMon1211064

Traditional Stats & Epi Methods: Algorithms still in AI/ML (1)

- We now quickly outline a number of algorithms still in use within AI/ML:
  - [*from* *14 popular AI algorithms and their uses post*]
- **1 Linear regression**
- Linear regression, also called least squares regression, is the simplest supervised machine learning algorithm for predicting numeric values. In some cases, linear regression doesn't even require an optimizer, since it is solvable in closed form. Otherwise, it is easily optimized using gradient descent (see below). The assumption of linear regression is that the objective function is linearly correlated with the independent variables. That may or may not be true for your data.
- To the despair of data scientists, business analysts often blithely apply linear regression to prediction problems and then stop, without even producing scatter plots or calculating correlations to see if the underlying assumption is reasonable. Don't fall into that trap. It's not that hard to do your exploratory data analysis and then have the computer try all the reasonable machine learning algorithms to see which ones work the best. By all means, try linear regression, but treat the result as a baseline, not a final answer.
- **2 Gradient descent**
- Optimization methods for machine learning, including neural networks, typically use some form of gradient descent algorithm to drive the back propagation, often with a mechanism to help avoid becoming stuck in local minima, such as optimizing randomly selected mini-batches (stochastic gradient descent) and applying momentum corrections to the gradient. Some optimization algorithms also adapt the learning rates of the model parameters by looking at the gradient history (AdaGrad, RMSProp, and Adam).
- **3 Logistic regression**
- Classification algorithms can find solutions to supervised learning problems that ask for a choice (or determination of probability) between two or more classes. Logistic regression is a method for solving categorical classification problems that uses linear regression inside a sigmoid or logit function, which compresses the values to a range of 0 to 1 and gives you a probability. Like linear regression for numerical prediction, logistic regression is a good first method for categorical prediction, but shouldn't be the last method you try.
- **4 Support vector machines**
- Support vector machines (SVMs) are a kind of parametric classification model, a geometric way of separating and classifying two label classes. In the simplest case of well-separated classes with two variables, an SVM finds the straight line that best separates the two groups of points on a plane. In more complicated cases, the points can be projected into a higher-dimensional space and the SVM finds the plane or hyperplane that best separates the classes. The projection is called a *kernel*, and the process is called the *kernel trick*. After you reverse the projection, the resulting boundary is often nonlinear. When there are more than two classes, SVMs are used on the classes pairwise. When classes overlap, you can add a penalty factor for points that are misclassified; this is called a soft margin.

Traditional Stats & Epi Methods: Algorithms still in AI/ML (2)

- We now quickly outline a number of algorithms still in use within AI/ML:
  - [*from 14 popular AI algorithms and their uses post*]

- **5 Decision tree** Decision trees (DTs) are a non-parametric supervised learning method used for both classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.
- Decision trees are easy to interpret and cheap to deploy, but computationally expensive to train and prone to **overfitting**.
- **6 Random forest** The random forest model produces an *ensemble* of randomized decision trees, and is used for both classification and regression. The aggregated ensemble either combines the votes modally or averages the probabilities from the decision trees. Random forest is a kind of *bagging* ensemble.
- **7 XGBoost** XGBoost (eXtreme Gradient Boosting) is a scalable, end-to-end, tree-boosting system that has produced state-of-the-art results on many machine learning challenges. Bagging and boosting are often mentioned in the same breath. The difference is that instead of generating an ensemble of randomized trees (RDFs), gradient tree boosting starts with a single decision or regression tree, optimizes it, and then builds the next tree from the residuals of the first tree.
- **8 K-means clustering** The k-means clustering problem attempts to divide *n* observations into *k* clusters using the Euclidean distance metric, with the objective of minimizing the variance (sum of squares) within each cluster. It is an unsupervised method of vector quantization, and is useful for feature learning, and for providing a starting point for other algorithms.
- Lloyd’s algorithm (iterative cluster agglomeration with centroid updates) is the most common heuristic used to solve the problem. It is relatively efficient, but doesn’t guarantee global convergence. To improve that, people often run the algorithm multiple times using random initial cluster centroids generated by the Forgy or random partition methods.
- K-means assumes spherical clusters that are separable so that the mean converges towards the cluster center, and also assumes that the ordering of the data points does not matter. The clusters are expected to be of similar size, so that the assignment to the nearest cluster center is the correct assignment.
- **9 Principal component analysis** Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated numeric variables into a set of values of linearly uncorrelated variables called principal components. Karl Pearson invented PCA in 1901. PCA can be accomplished by eigenvalue decomposition of a data covariance (or correlation) matrix, or singular value decomposition (SVD) of a data matrix, usually after a normalization step applied to the initial data.

Traditional Stats & Epi Methods: Assessment Check

- We now engage participants to check our mutual understanding.

Traditional Stats & Epi Methods: More complex models

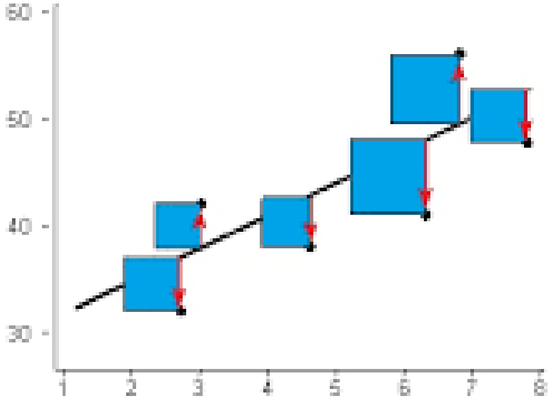
- When you need a **lot more** ‘hands’ on which to *count quantities involved*

Millipeded Photo by Unknown Author is licensed under CC BY; stock photos elsewise

- *Grow number of quantities to track data features, or ‘parameters’*
  - *In these cases, more ‘modern’ statistics & epi methods are needed…*
  - *A fundamental method (to AI/ML also): ‘regression’ often ‘fitted’ using least-squares*

Linear regression, also called least squares regression, is the simplest supervised machine learning algorithm for predicting numeric values. In some cases, linear regression doesn’t even require an optimizer, since it is solvable in closed form. Otherwise, it is easily optimized using gradient descent (see below in later algorithm coverage). The assumption of linear regression is that the objective function is linearly correlated with the independent variables.

We will cover additional fundamental algorithms throughout today’s Think-a-Thon



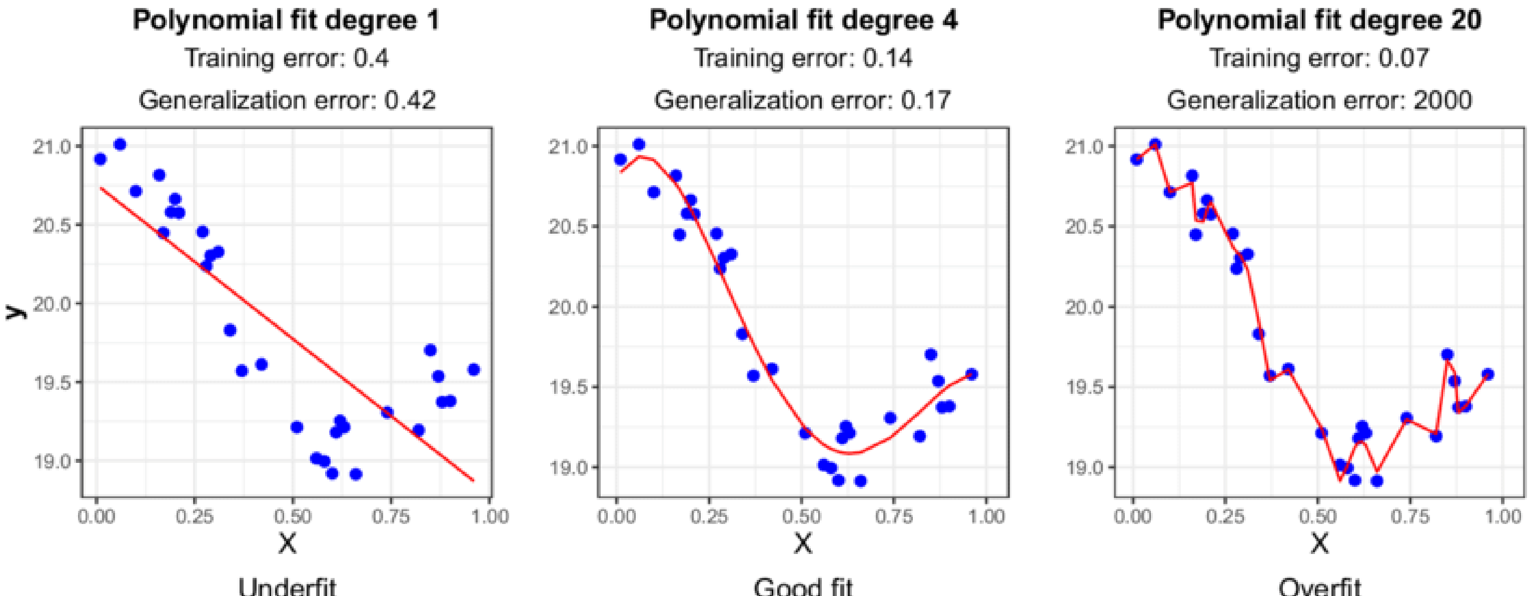
https://www.infoworld.com/article/3695208/14-popular-ai-algorithms-and-their-uses.html

Traditional Stats & Epi Methods: More complex models

- When you need a **lot more** ‘hands’ on which to *count quantities involved*

Millipeded Photo by Unknown Author is licensed under CC BY; stock photos elsewise

- *Grow number of quantities to track data features, or ‘parameters’*
  - *In these cases, more ‘modern’ statistics & epi methods are needed, at risk of ‘overfitting’*



Underfit

Good fit

Overfit

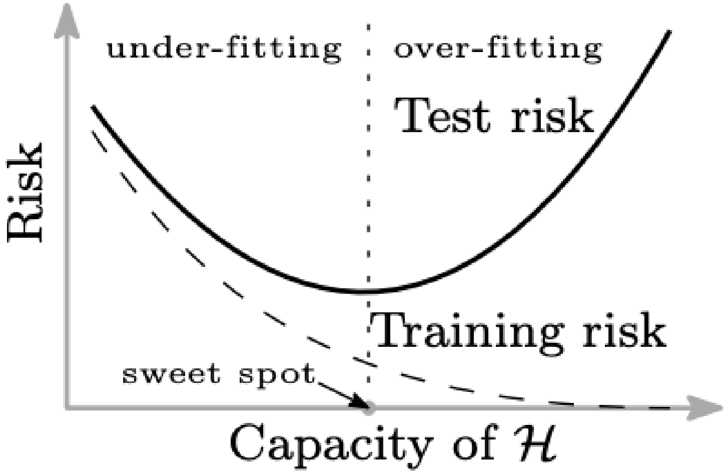
Illustration of the underfitting/overfitting issue on a simple regression case. Data points are shown as blue dots and model fits as red lines. Underfitting occurs with a linear model (left panel), a good fit with a polynomial of degree 4 (center panel), and overfitting with polynomial of degree 20 (right panel). Root mean squared error is chosen as objective function for evaluating the training error and the generalization error, assessed by using 10-fold cross-validation.

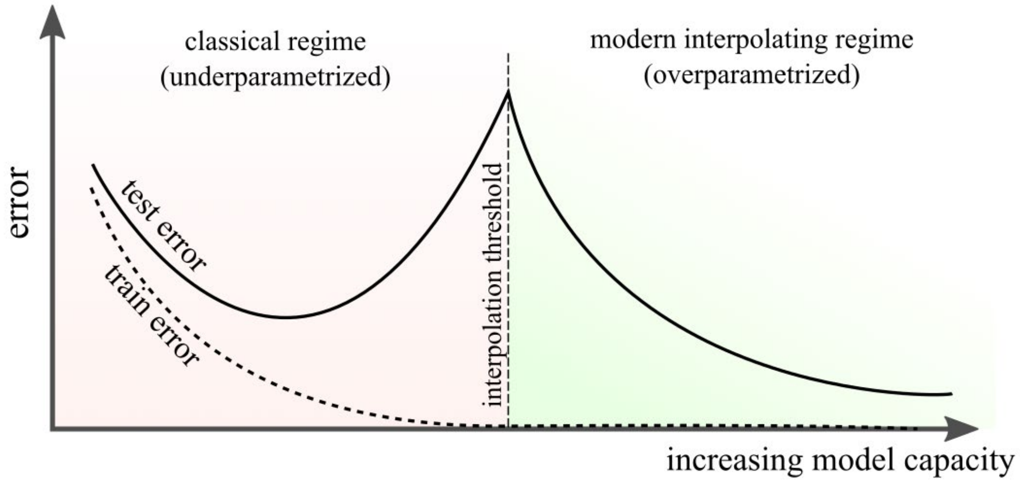
Traditional Stats & Epi Methods: More complex models

- When you need a **lot more** ‘hands’ on which to *count quantities involved*

Millipeded Photo by Unknown Author is licensed under CC BY; stock photos elsewise

- *Grow number of quantities to track data features, or ‘parameters’*
  - *In these cases, more ‘modern’ statistics & epi methods are needed, at risk of ‘overfitting’*
  - *Still in ‘pink’ zone relative to (overparametrized) model architectures*

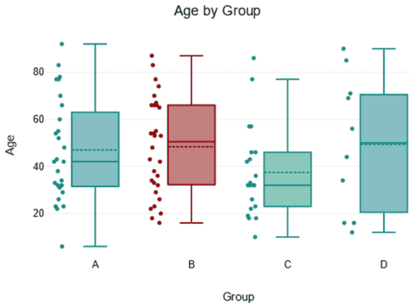


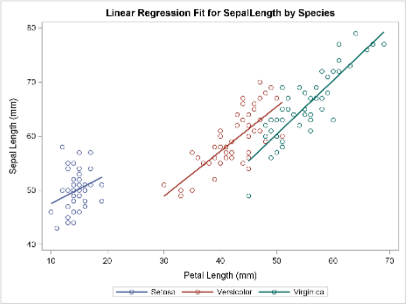


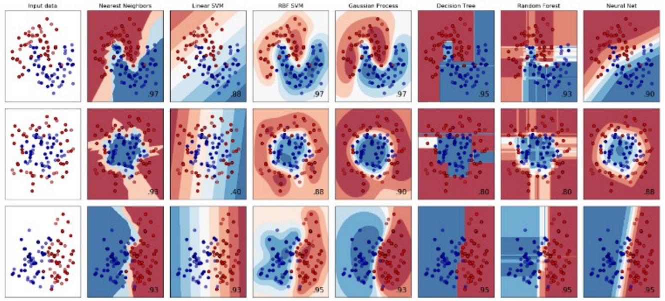
http://dx.doi.org/10.1002/cpt.1796
https://en.wikipedia.org/wiki/Neural_tangent_kernel#/media/File:Double_descent.png

Traditional Stats & Epi Methods: More complex models

- When you need a **lot more** ‘hands’ on which to *count quantities involved*







- *Grow number of quantities to track data features, or ‘parameters’*
  - *In these cases, more ‘modern’ statistics & epi methods (like those ^here) are needed*

- **If not, may need more modern stats/epi methods (if not AI/ML)**
  - Includes methods of regression / statistical learning that have bled into AI/ML
  - Example of special preparation of data to use (later Think-a-thon example)

More Modern Stats & Epi Methods: **assessment check**

- We now engage participants to check our mutual understanding.

Traditional Stats & Epi Methods: Pro’s & Con’s

Per Think-a-thon Planning outline:

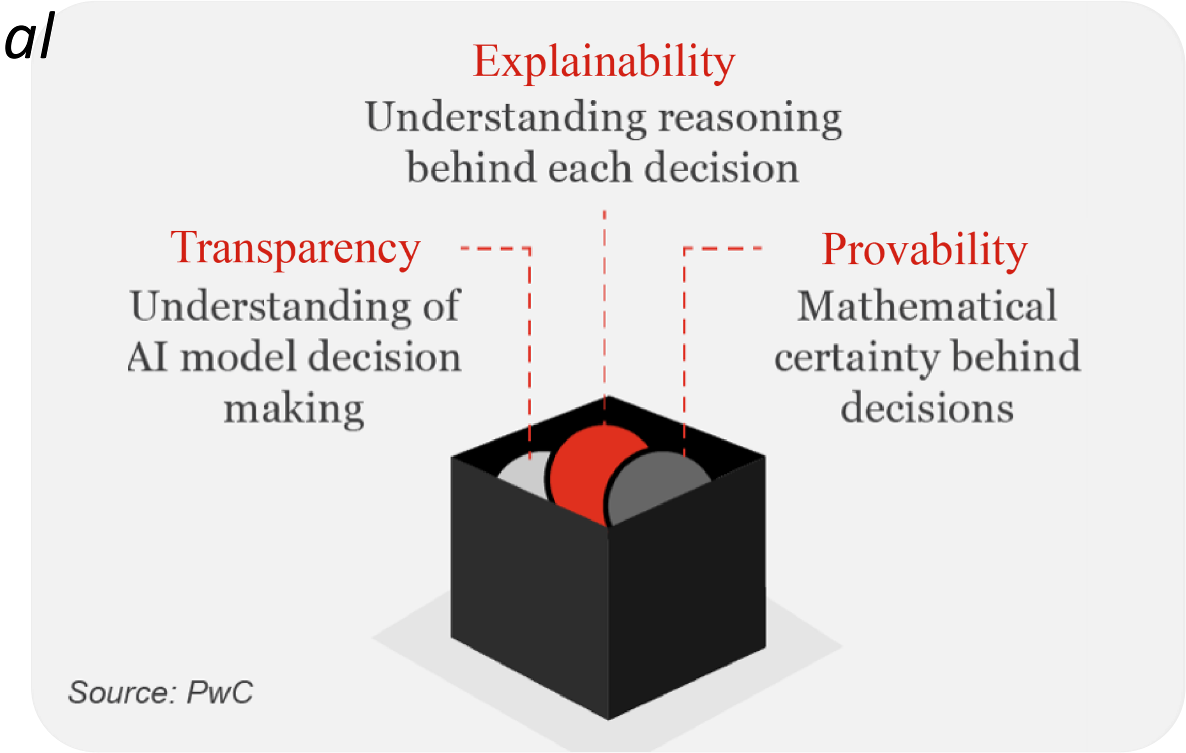
- *Strengths:*
  - *robust,*
  - ***Interpretable (where these shine: covered more by next few slides)**,*
  - *well-established methodology*
  - *assumptions transparently expressed in terms of domain-specific science*
- *Weaknesses:*
  - *limited predictive power when using conventional ‘parametric’ forms,*
  - *assumption-dependent, yet assumptions typically more transparently assessed*
  - *often (over-)focused on hypothesis testing*
-

Traditional Stats & Epi Methods: Pro's & Con's

- Common to *any data science computational strategy*

Setting apart conventional statistical/epidemiologic modeling

- When you need an **interpretability of** *quantities involved*


Explainability
Understanding reasoning behind each decision
Transparency
Understanding of AI model decision making
Provability
Mathematical certainty behind decisions

Source: PwC

JAMA. 2018;320(21):2199-2200. doi:10.1001/jama.2018.17163

- *Distinct from* post hoc *'explainability'*
  - *Often applied after the fact in AI/ML*
  - *'explaining' via repeated 'querying' of models…*

Per JAMA editorial, "Black boxes are unacceptable: A Clinical Decision Support System requires transparency so that users can understand the basis for any advice or recommendations that are offered"

Intrinsic to ANY Data Methods: Pros & Cons

- REMEMBER: for *any data science computational strategy*

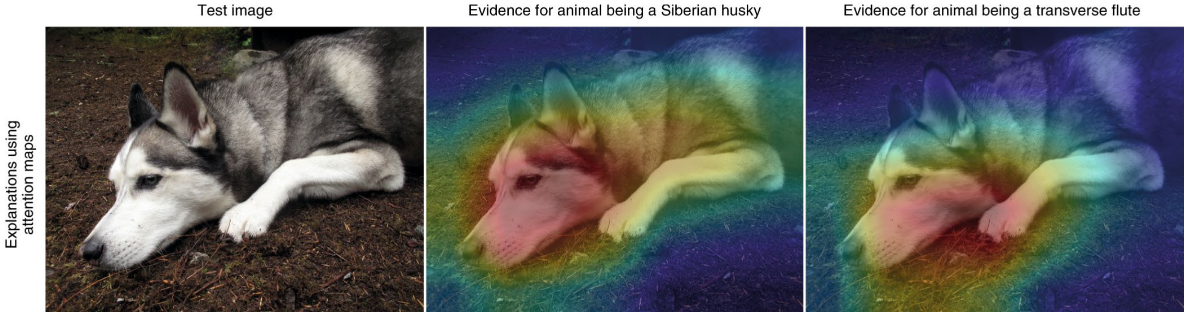
Setting apart conventional modeling

- When you need an **interpretability of** *quantities involved*
- *Distinct from after-the-fact 'explainability'*
  - *Survey of examples / counter-examples here: https://jair.org/index.php/jair/article/view/12228*
- *Assessment check:*
- [sli.do questions]

MEET THE LEADERS # ARTIFICIAL INTELLIGENCE
"Interpretable Machine Learning is critical for fairness and trust."
Cynthia Rudin
Professor of Computer Science

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Fig. 2: Saliency does not explain anything except where the network is looking.


Test image
Evidence for animal being a Siberian husky
Evidence for animal being a transverse flute
Explanations using attention maps

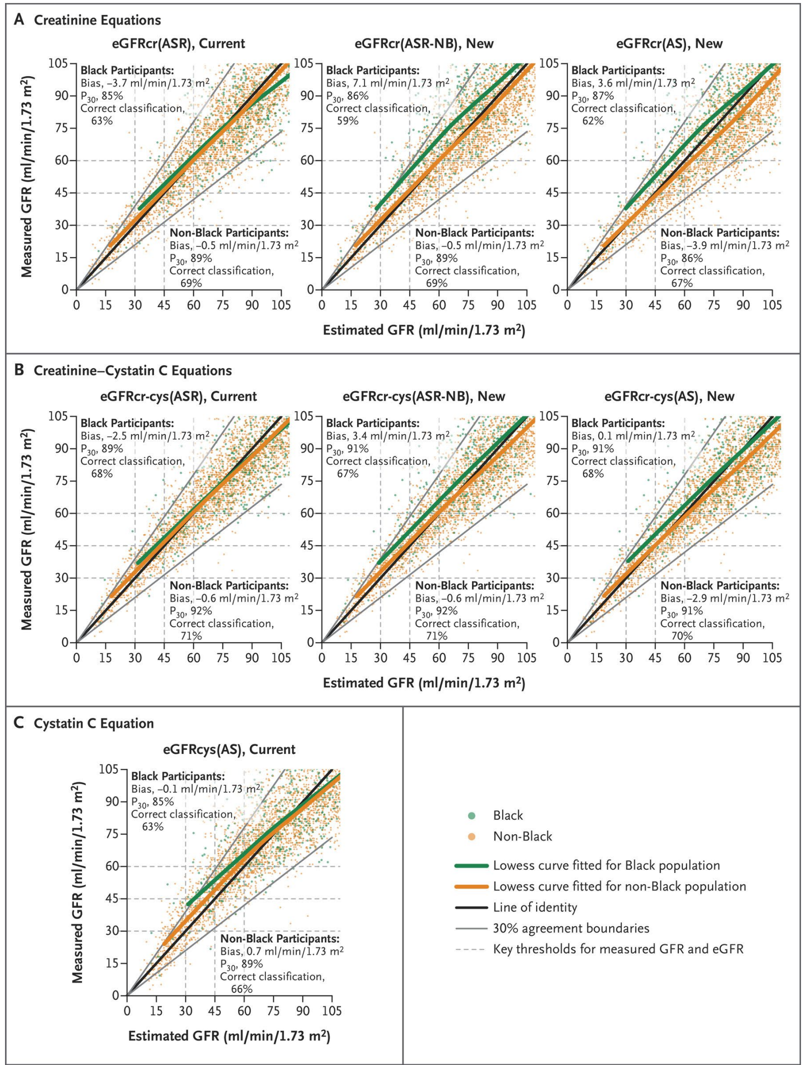
We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University

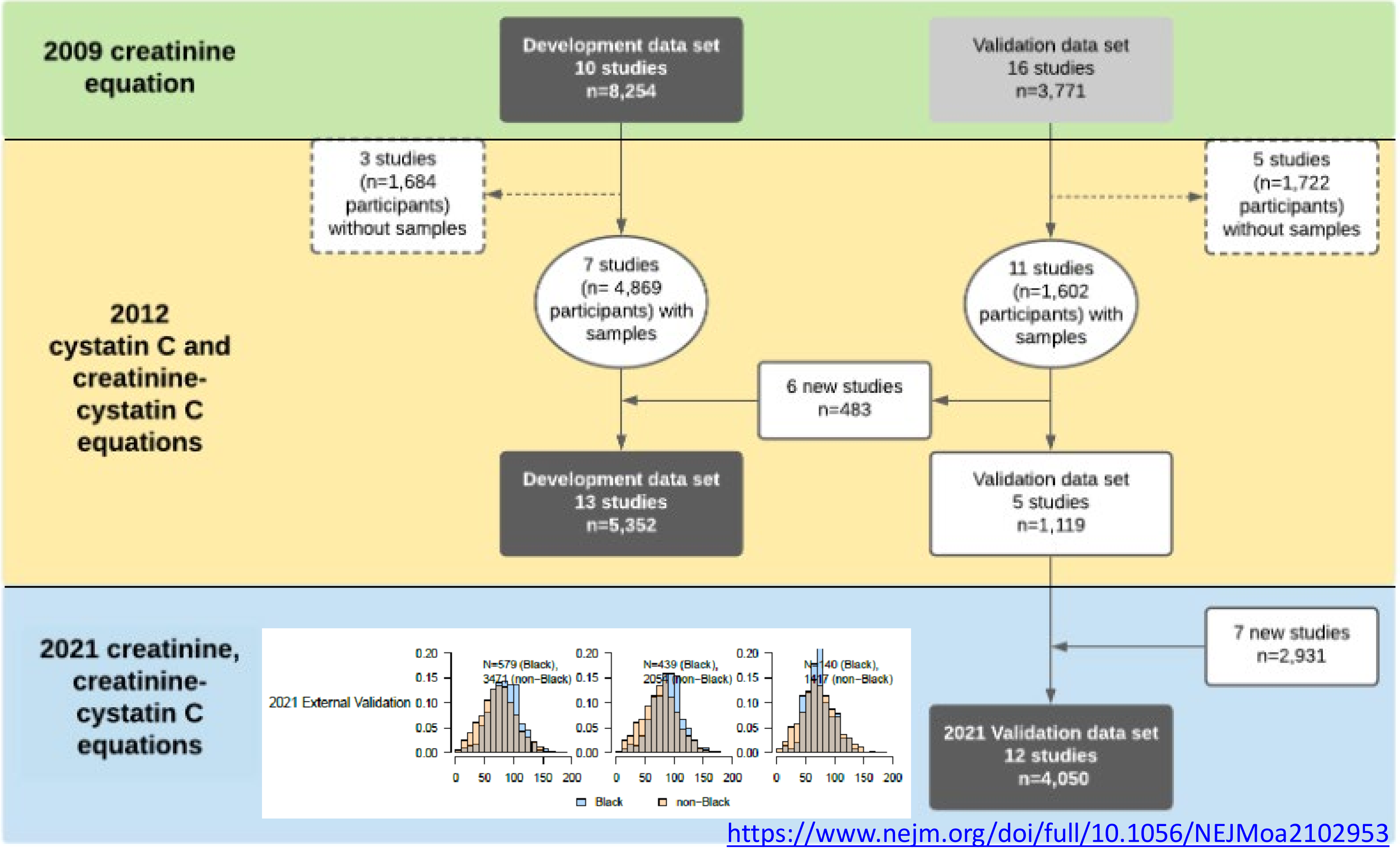
Beyond Traditional Stats & Epi Methods: ***issues remain***
(*counter*-)Example(s) with regard to **health disparity**

- [already be covered in other Think-a-thon slides on CKD-Epi eGFR]

$$\text{eGFR} =$$

$$\mu \times \min\left(\frac{\text{Scr}}{\kappa}, 1\right)^{\alpha_1} \times \max\left(\frac{\text{Scr}}{\kappa}, 1\right)^{\alpha_2} \times \min\left(\frac{\text{Scys}}{0.8}, 1\right)^{\beta_1} \times \max\left(\frac{\text{Scys}}{0.8}, 1\right)^{\beta_2} \times \lambda^{\text{Age}} \times \psi\ [\text{if female}] \times \phi\ [\text{if black}].$$





2009 creatinine equation

Development data set 10 studies n=8,254

Validation data set 16 studies n=3,771

3 studies (n=1,684 participants) without samples

5 studies (n=1,722 participants) without samples

7 studies (n= 4,869 participants) with samples

11 studies (n=1,602 participants) with samples

6 new studies n=483

2012 cystatin C and creatinine-cystatin C equations

Development data set 13 studies n=5,352

Validation data set 5 studies n=1,119

2021 creatinine, creatinine-cystatin C equations

7 new studies n=2,931

2021 Validation data set 12 studies n=4,050

2021 External Validation
N=579 (Black), 3471 (non-Black)
N=439 (Black), 2058 (non-Black)
N=140 (Black), 1417 (non-Black)
0.00 0.05 0.10 0.15 0.20
0 50 100 150 200
Black
non-Black

https://www.nejm.org/doi/full/10.1056/NEJMoa2102953

Beyond Traditional Stats & Epi Methods: Healthcare AI/ML (*counter*-)Example(s) with regard to **health disparities**

- Examples

(Optum algorithm)

Task: Who are the patients requiring more resources for care?

Bias: Black patients assigned the same level of risk by the algorithm are actually sicker than white patients.

Reason: Actual target (cost) is not reflecting true target (needs for health care).

https://www.healthcarefinancenews.com/news/study-finds-racial-bias-optum-algorithm

(Racial/Ethnic Disparities in Suicide prediction)

Task: Prediction of death by Suicide After Mental Health Visits.

Bias: Suicide prediction models disproportionately benefit certain race/ethnic subgroups than the others

13,980,570 mental health visits by 1,433,543 patients from Jan. 2009 to Sep. 2017

Both LASSO and random forests performed better (AUC) for White(0.822/0.812), Hispanic (0.855/0.831) and Asian(0.834/0.882) patients than Black(0.775/0.786) and American Indian/Alaskan Native(0.599/0.642) patients.

Reason: Lack of health record data of minor race/ethnicities for training ML models.

https://pubmed.ncbi.nlm.nih.gov/33909019/ (Coley et. al 2021)

NIH National Institute of Diabetes and Digestive and Kidney Diseases

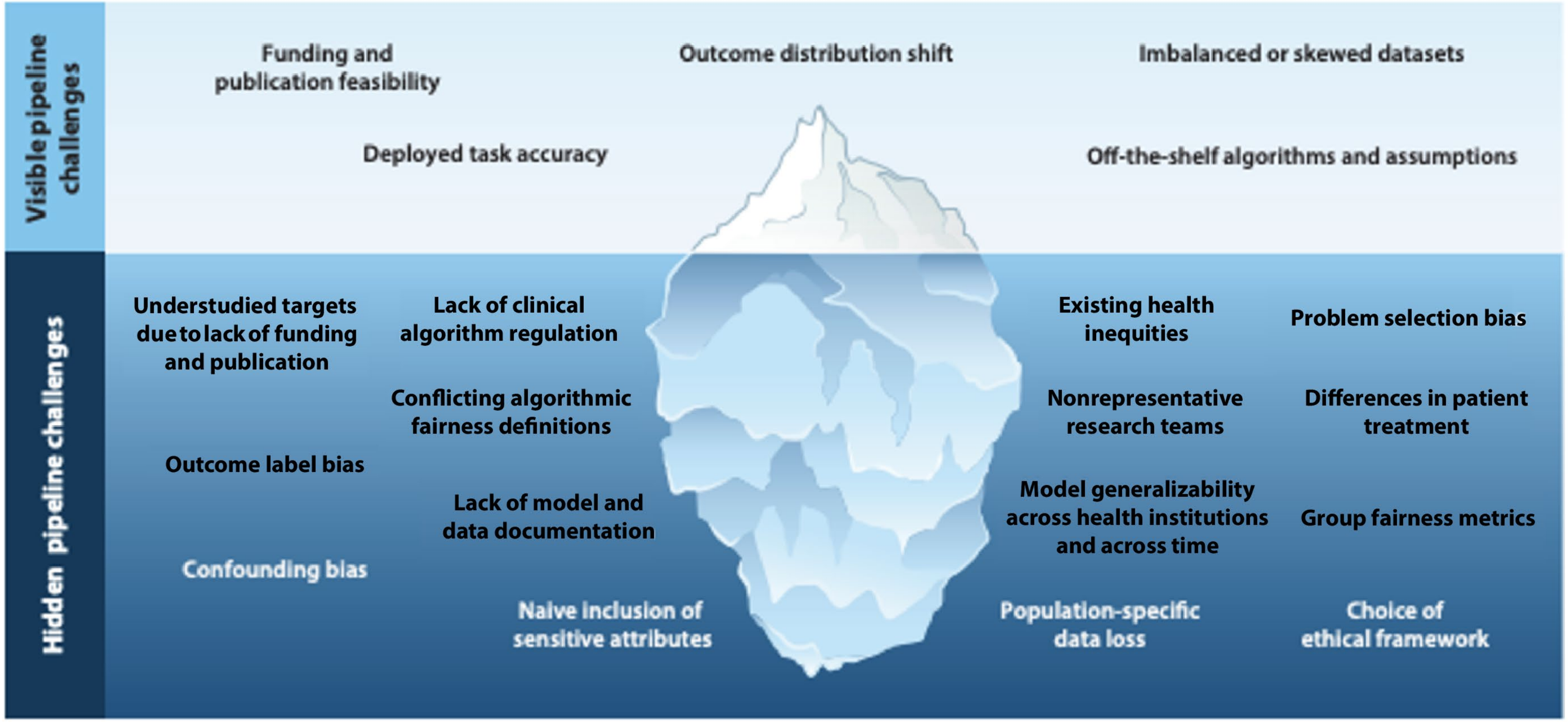
Getting our lay of the land: **health disparity… bias**

"Bias"

Along with 'fairness' will be discussed in next few slides

- 'Fairness' = lack of 'bias'?
  - Not necessarily *due to incompatibility of some fairness/bias measures*
  - *Theorem exists to show this **inherent tradeoff***

The Bias/Fairness Iceberg



Visible pipeline challenges

Funding and publication feasibility

Outcome distribution shift

Imbalanced or skewed datasets

Deployed task accuracy

Off-the-shelf algorithms and assumptions

Hidden pipeline challenges

Understudied targets due to lack of funding and publication

Lack of clinical algorithm regulation

Existing health inequities

Problem selection bias

Conflicting algorithmic fairness definitions

Nonrepresentative research teams

Differences in patient treatment

Outcome label bias

Lack of model and data documentation

Model generalizability across health institutions and across time

Group fairness metrics

Confounding bias

Naive inclusion of sensitive attributes

Population-specific data loss

Choice of ethical framework

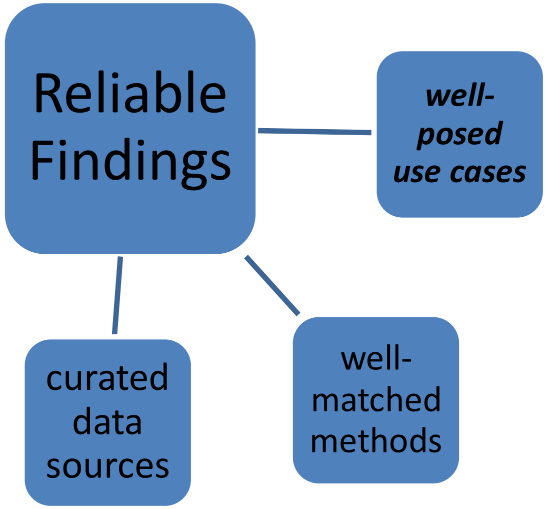
Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Healthcare. Annu Rev Biomed Data Sci. 2021 Jul;4:123-144. doi: 10.1146/annurev-biodatasci-092820-114757. Epub 2021 May 6. PMID: 34396058; PMCID: PMC8362902.

c/o Tony Solomonides

Intrinsic to ANY Data Methods: Pro’s & Con’s

Common to *any data science computational strategy*

- ONLY holds up IF a three-legged stool:
  - well-posed use cases
  - curated data sources, and
  - well-matched methods.



Why a “Three-legged Stool”?

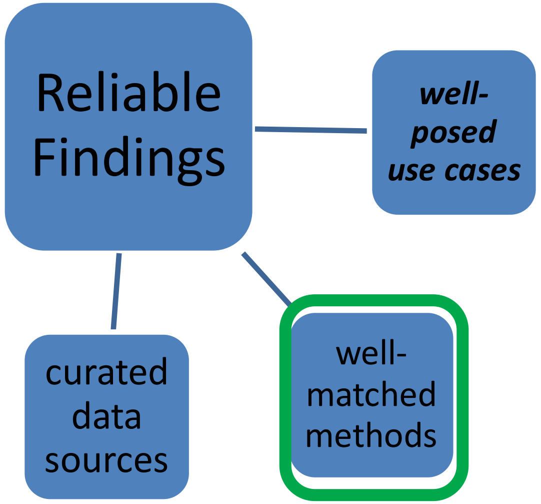
Physics reigns supreme:

- stool couldn’t stay up / support anything
- with only 2 of its 3 legs in place… data scientist needed for all 3

Intrinsic to ANY Data Methods: Pro’s & Con’s

Common to *any data science computational strategy*

- 3-legged stool:
  - **well-posed** use cases
  - **curated** data sources, and
  - **well-matched** methods.



Reliable Findings

well-posed use cases

curated data sources

well-matched methods

Why a “Three-legged Stool”?

Physics reigns supreme:

- stool couldn’t stay up / support anything
- with only 2 of its 3 legs in place, data scientist essential ↗

Intrinsic to ANY Data Methods: Pro’s & Con’s

Common to *any data science computational strategy*

- 3-legged stool:
  - ***well-posed*** use cases (*some are* ***so*** *well-posed, it may function like this stool* ↓ *)*
  - ***curated*** data sources, and
  - ***well-matched*** methods.

- We continue through AI and machine learning use cases
  - Objective is for ScHARe community members to gain intuition
  - We’ll provide some **examples** and ***counter-examples***
  - *Our emphasis today is on grasping concepts via this quick tour*

→

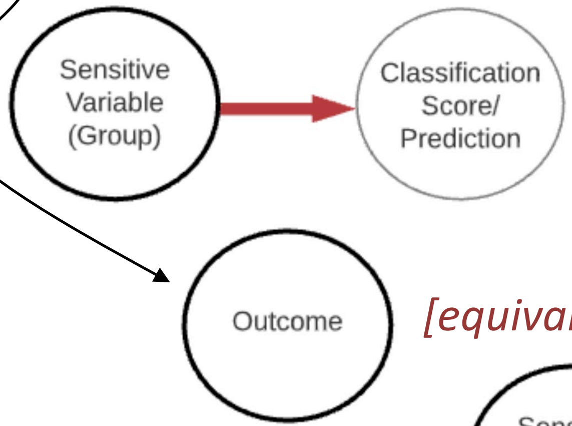
Getting our lay of the land: **health disparities**

Without being concerned about *each jargon term* used (for equity measures) at right, just note how *first* two each presume distinct relationships among variables, as shown at end of curved arrows; ***outcome***?

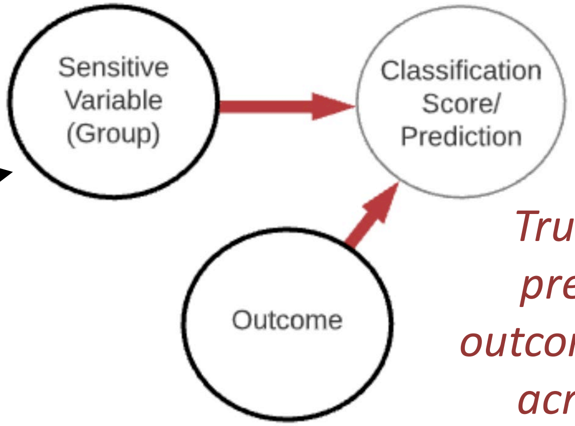
or **Group**?

Demographic parity
Equal opportunity
*Equalized odds
Equal accuracy
Treatment equality
Equalized (dis)incentives
False negative parity

…in the context of a Classification / Prediction task



The outcome is independent of group (the sensitive variable) [equivalent to 'separation' in FairML text]



True-'positive' predictions of outcome is same across groups (distinct values of sensitive variable)

Caton S, Haas C. Fairness in machine learning: A survey. arXiv:2010.04053v1 [cs.LG] 4 Oct 2020

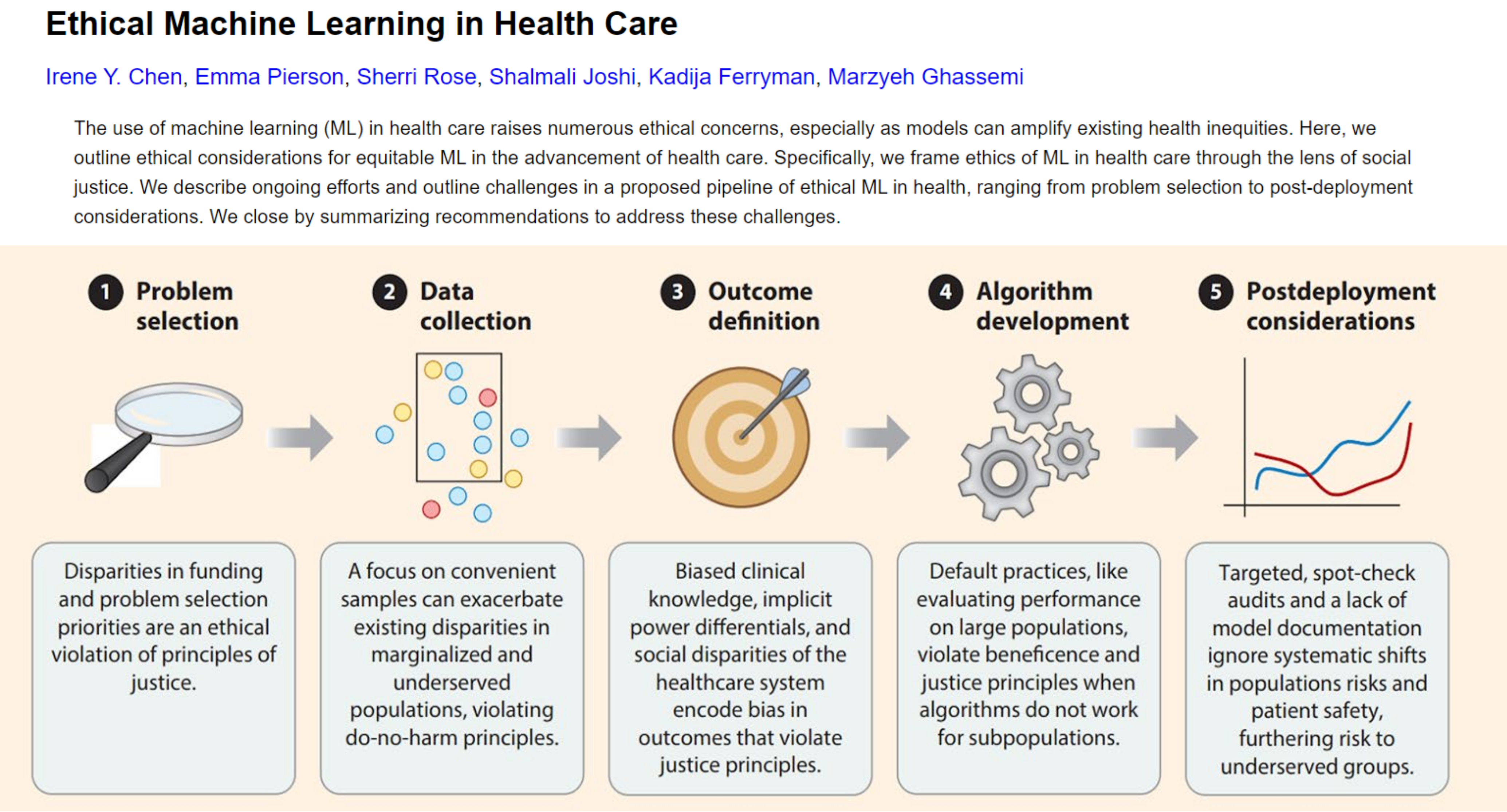
c/o Harold Lehmann

**Odds & Equalized odds will be touched on in later slides; others in later Think-a-thons…*

Getting our lay of the land: **health disparities…** ***lack of*** **fairness**

- ‘Fairness’ = lack of ‘bias’?
  - Not necessarily *due to incompatibility of some fairness/bias measures*
  - *Theorem asserts this mathematically… thus, each use case must prioritize*
- *Bias: does it mean the same thing to all data science practitioners?*
  - *Also not necessarily: ‘statistical bias’ is concept of long-term behavior of estimation… does it approach its target in the long term, is it off (‘biased’)?*
  - *Varied forms of ‘bias’ in medical/epidemiologic evidence (Risk of Bias)*
    - *Many subtypes… ascertainment bias, confounding bias, recall bias, selection bias, etc.*
    - *key one for practicing data scientists & their collaborators: confirmation bias*
    - *Other forms noted in data science circles: gender bias, language bias, political bias, etc.*
  - *‘Bias’ most often considered in data science:*
    - *Lack of ‘fairness’ i.e., differential (if not* ***adverse****) performance for certain subgroups*
    - *Most often unintentionally introduced due to longstanding biases in who’s data we ‘have’*

Getting our lay of the land: *reducing* **health disparities as ethical imperative**



Ethical Machine Learning in Health Care

Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, Marzyeh Ghassemi

The use of machine learning (ML) in health care raises numerous ethical concerns, especially as models can amplify existing health inequities. Here, we outline ethical considerations for equitable ML in the advancement of health care. Specifically, we frame ethics of ML in health care through the lens of social justice. We describe ongoing efforts and outline challenges in a proposed pipeline of ethical ML in health, ranging from problem selection to post-deployment considerations. We close by summarizing recommendations to address these challenges.

1 Problem selection

Disparities in funding and problem selection priorities are an ethical violation of principles of justice.

2 Data collection

A focus on convenient samples can exacerbate existing disparities in marginalized and underserved populations, violating do-no-harm principles.

3 Outcome definition

Biased clinical knowledge, implicit power differentials, and social disparities of the healthcare system encode bias in outcomes that violate justice principles.

4 Algorithm development

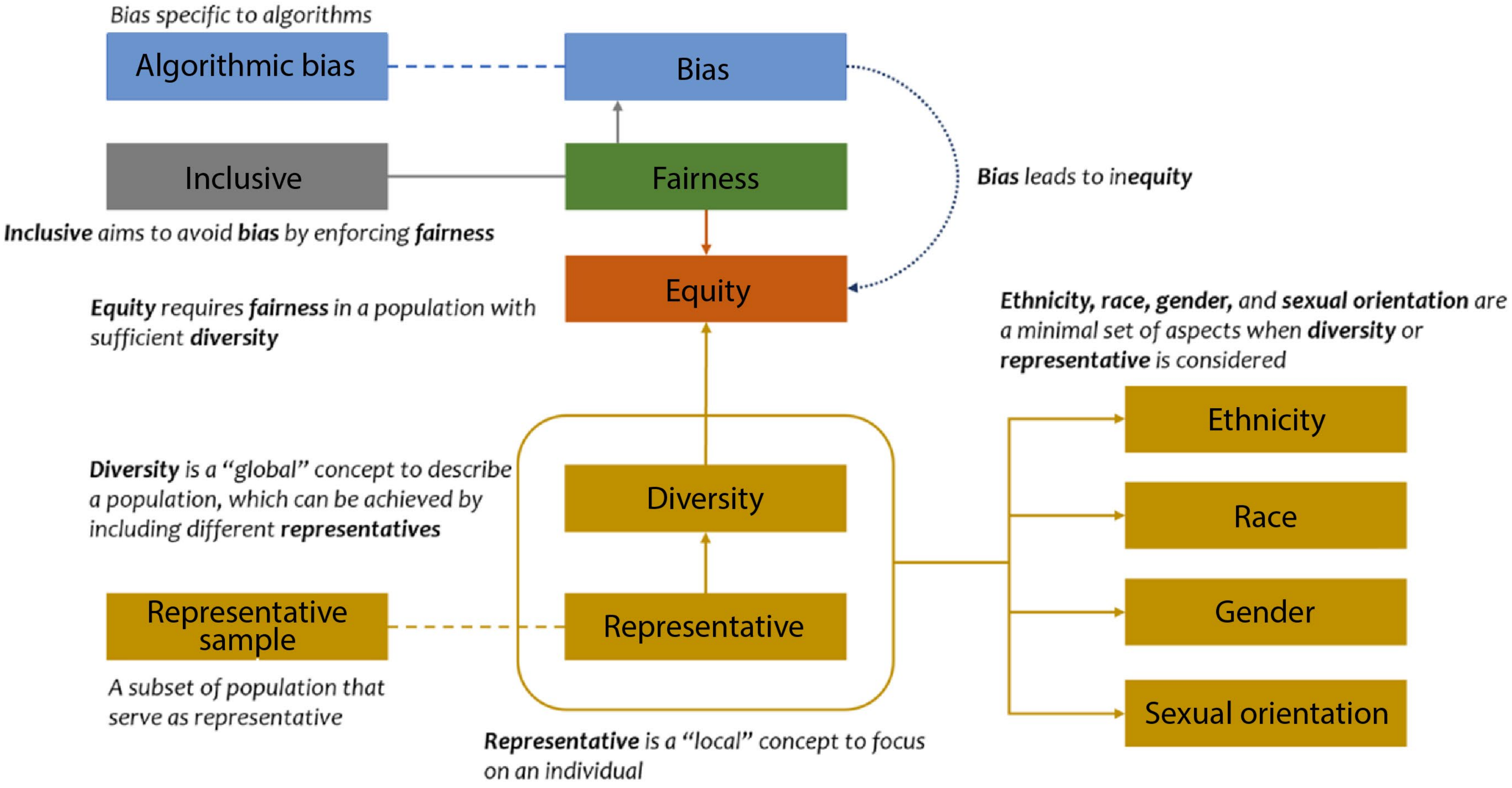
Default practices, like evaluating performance on large populations, violate beneficence and justice principles when algorithms do not work for subpopulations.

5 Postdeployment considerations

Targeted, spot-check audits and a lack of model documentation ignore systematic shifts in populations risks and patient safety, furthering risk to underserved groups.

Getting our lay of the land: **health equity terminologies**

- *Bias: our working use going forward*
  - *AIM-AHEAD*
    - *presented last week by physician member of NIDDK Advisory Council*
    - *Developed by AIM-AHEAD**
  - *ScHARe:*
    - *Looking to align with recent activities within * Artificial Intelligence/Machine Learning Consortium to Advance Health Equity & Researcher Diversity (AIM-AHEAD) Ethics & Equity Workgroup (paper ->)*

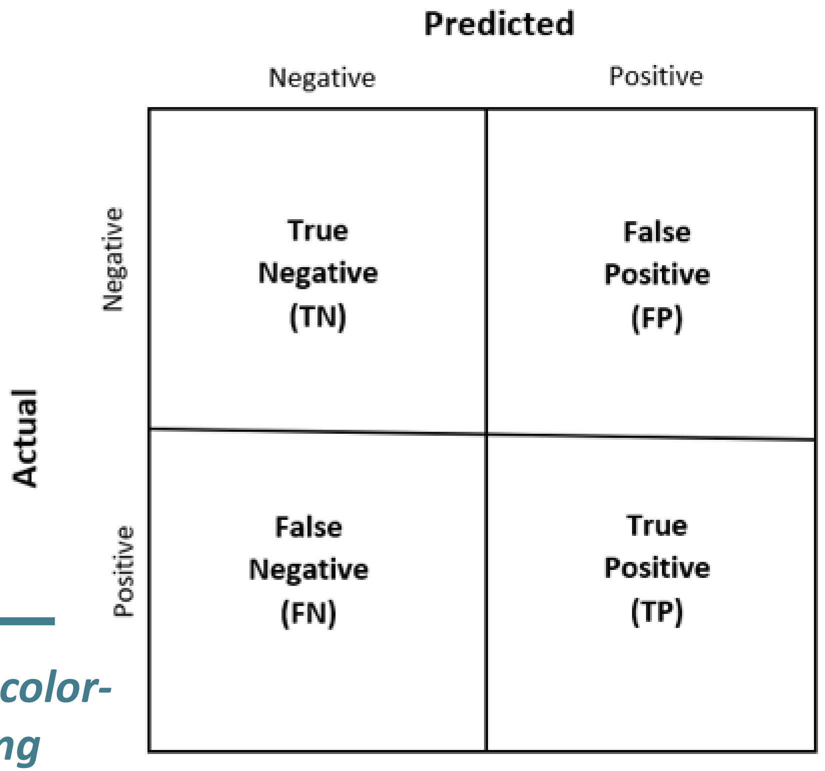


Bias specific to algorithms
Algorithmic bias
Bias
Bias leads to inequity
Inclusive
Fairness
Inclusive aims to avoid bias by enforcing fairness
Equity
Equity requires fairness in a population with sufficient diversity
Ethnicity, race, gender, and sexual orientation are a minimal set of aspects when diversity or representative is considered
Ethnicity
Race
Gender
Sexual orientation
Diversity is a "global" concept to describe a population, which can be achieved by including different representatives
Diversity
Representative sample
Representative
A subset of population that serve as representative
Representative is a "local" concept to focus on an individual

https://ai.jmir.org/2023/1/e52888

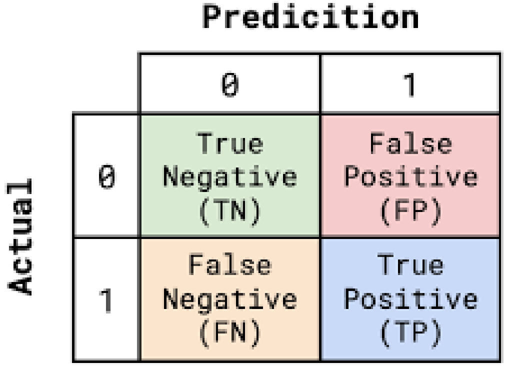
Getting our lay of the land: health disparity terminologies

- *Bias: example application of a fairness measure*
  - *Equalized Odds*
    - *Mentioned above, among many other measures*
    - *Used for binary events (in original paper, now multiclass)*
    - *Also termed 'equality of odds' (of event)*
  - *Used as measure of **group** fairness*
    - *Must know Actual status, v. what's Predicted by method*
    - *From this one can form a 'Confusion Matrix' table, @ right*
    - *As this is a 2 row by 2 column tabulation, 'odds' are natural*

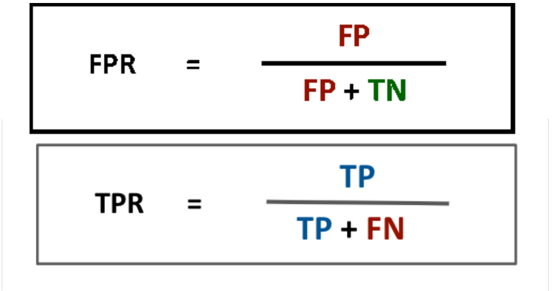


		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Add color-coding

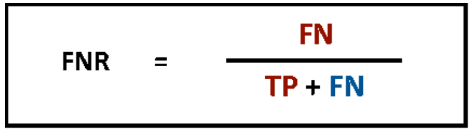


		Predicition	
		0	1
Actual	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)



$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

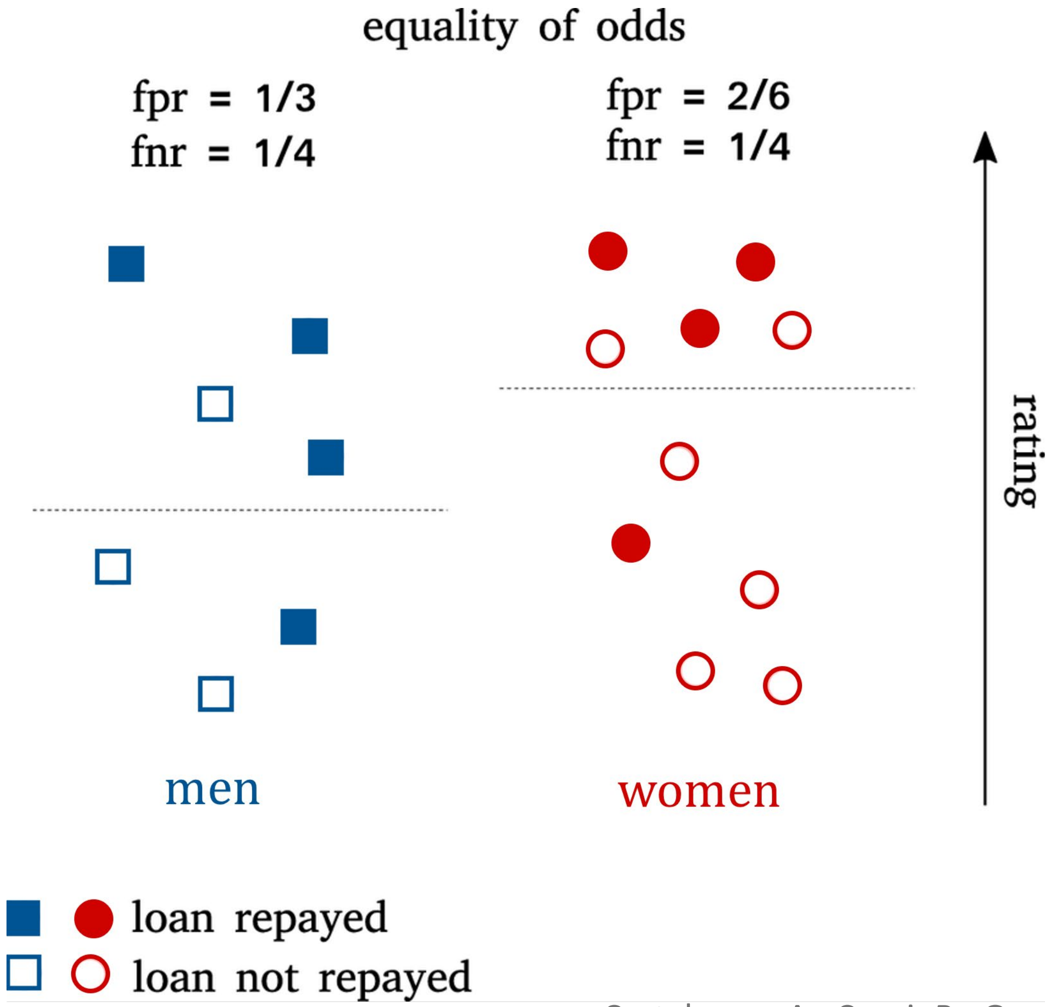


$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}}$$

https://ocw.mit.edu/courses/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/pages/module-three-framework/fairness-criteria/ | https://towardsdatascience.com/analysing-fairness-in-machine-learning-with-python-96a9ab0d0705

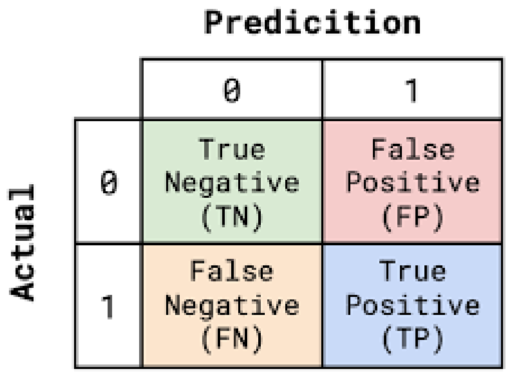
Getting our lay of the land: **health disparity terminologies**

- *Bias:* **example** *application of a fairness measure*
  - *Equalized Odds*
    - *Mentioned above*
    - *Used for binary events, like @ right*
    - *Also termed 'equality of odds' (of event)*
  - ***group** fairness: are FPR & FNR the same across the two groups of men & women?*



Castelnovo, A., Crupi, R., Greco, G. *et al.* A clarification of the nuances in the fairness metrics landscape. *Sci Rep* **12**, 4209 (2022). https://doi.org/10.1038/s41598-022-07939-1

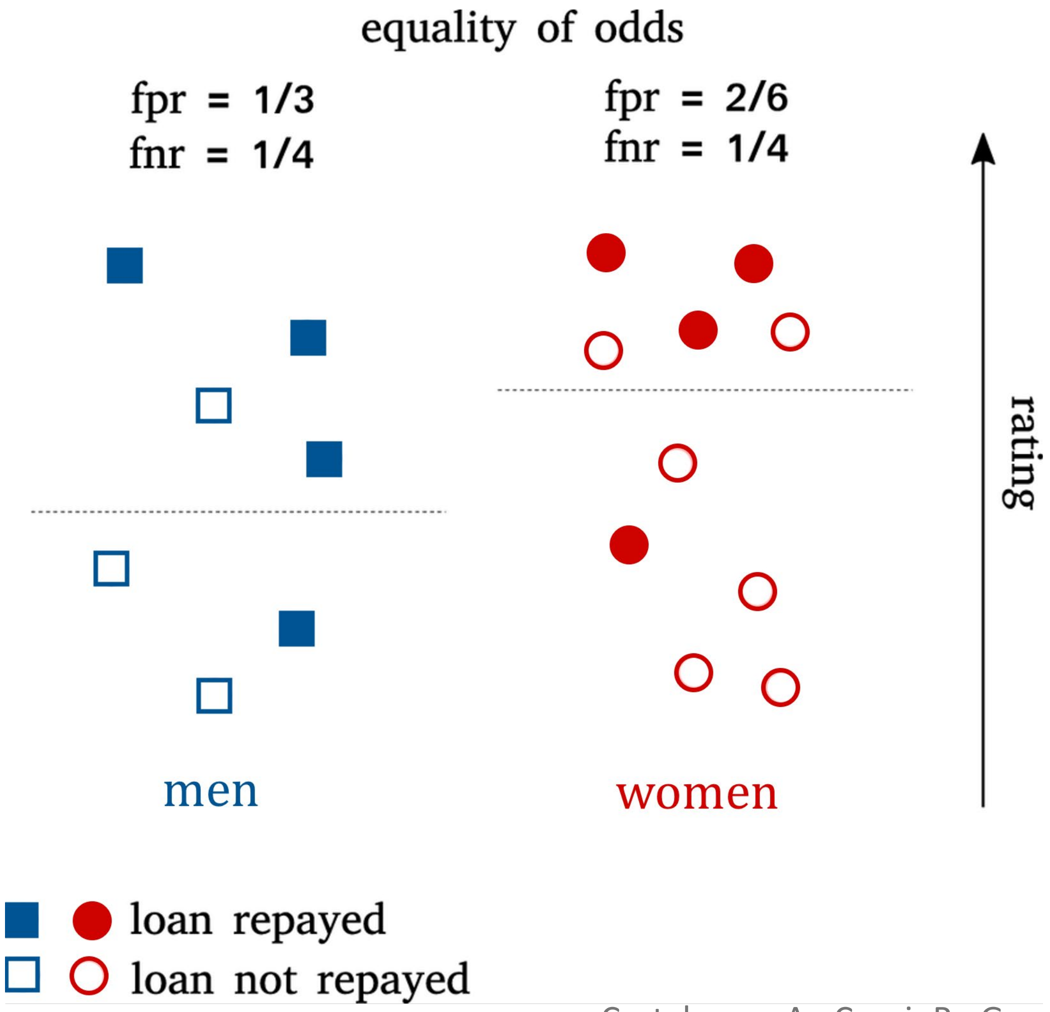
Getting our lay of the land: **health disparity terminologies**

- *Bias:* **example** *application of a fairness measure*
  - 

		Predicition	
		0	1
Actual	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

  - ***group** fairness: are FPR & FNR the same across the two groups of men & women?*

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}}$$



$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Castelnovo, A., Crupi, R., Greco, G. *et al.* A clarification of the nuances in the fairness metrics landscape. *Sci Rep* **12**, 4209 (2022). https://doi.org/10.1038/s41598-022-07939-1

Getting a lay of the land: assessment check

- We now engage participants to check our mutual understanding.

ScHARe

Science Collaborative for Health disparities and Artificial intelligence bias REduction

Artificial Intelligence in Data Science as a Broad New Horizon

A. AI Fundamentals

B. Computational Strategies: forms of AI that may not be conventionally referred to as machine learning… e.g., Generative AI & other forms of Deep Learning (DL)

Generative AI

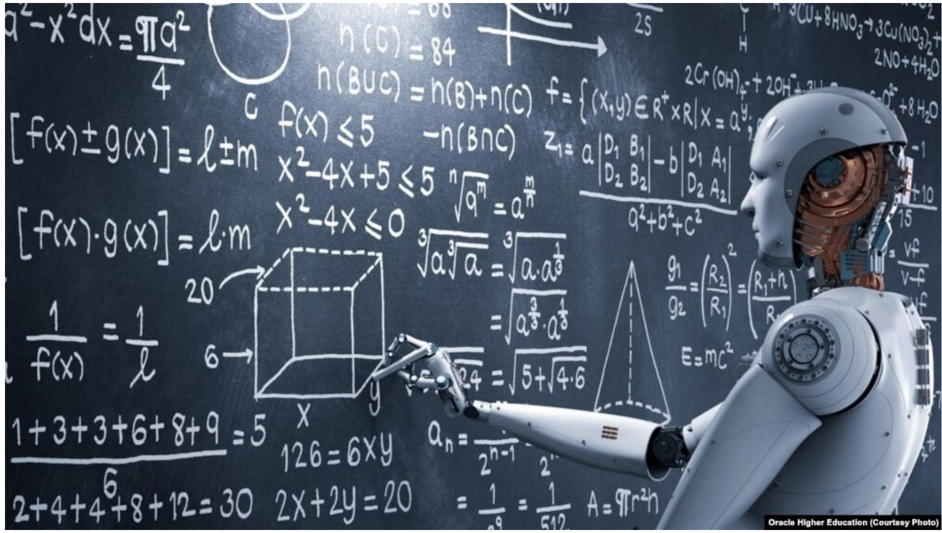
Generative AI is a subset of DL models that generates content like text, images, or code based on provided input. Trained on vast data sets, these models detect patterns and create outputs without explicit instruction, using a mix of supervised and unsupervised learning.

National Institute of Diabetes and Digestive and Kidney Diseases

Artificial Intelligence Fundamentals: Definitions

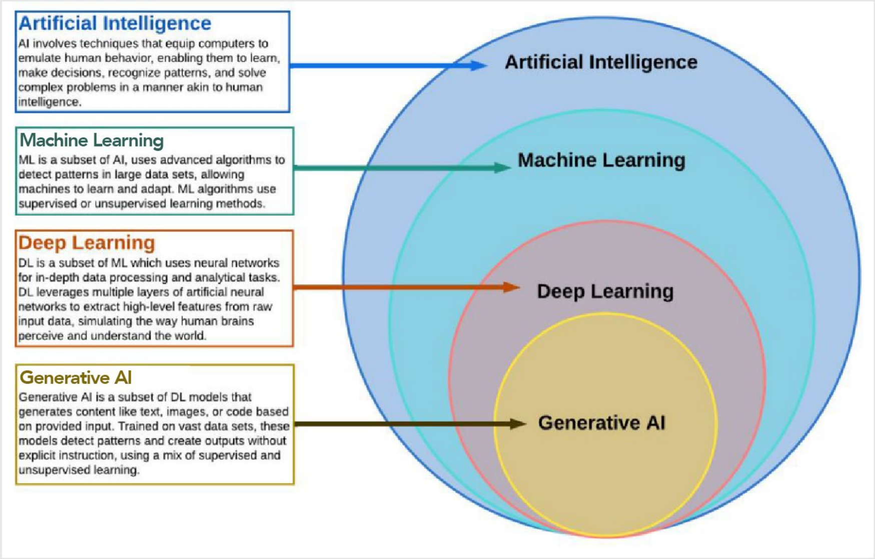
- Definitions (& *distinctions* with specific subset of 'machine learning')
  - ***OURS:*** NIH Strategic Plan for Data Science (2018-2023*):
    - Artificial Intelligence: "the power of a machine to copy intelligent human behavior"
    - Machine Learning: "field of computer science that gives computers the ability to learn without being explicitly programmed by humans"

*NOTE: NIH Strategic Plan for Data Science **2023-2028** (in revision, open for public comment)



Oracle Higher Education (Courtesy Photo)

Oracle Higher Education



Artificial Intelligence
AI involves techniques that equip computers to emulate human behavior, enabling them to learn, make decisions, recognize patterns, and solve complex problems in a manner akin to human intelligence.

Machine Learning
ML is a subset of AI, uses advanced algorithms to detect patterns in large data sets, allowing machines to learn and adapt. ML algorithms use supervised or unsupervised learning methods.

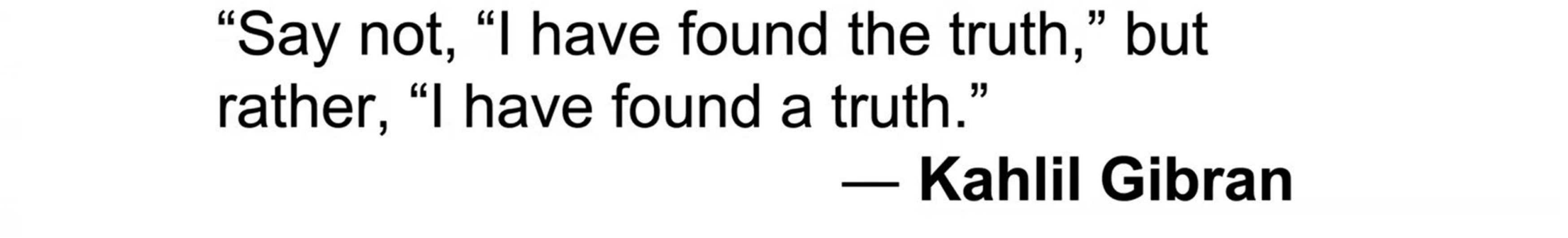
Deep Learning
DL is a subset of ML which uses neural networks for in-depth data processing and analytical tasks. DL leverages multiple layers of artificial neural networks to extract high-level features from raw input data, simulating the way human brains perceive and understand the world.

Generative AI
Generative AI is a subset of DL models that generates content like text, images, or code based on provided input. Trained on vast data sets, these models detect patterns and create outputs without explicit instruction, using a mix of supervised and unsupervised learning.

https://doi.org/10.3390/su151813484

Artificial Intelligence Fundamentals

- *Despite all the potential that AI has, and compelling performance shown… remain humble: per quote selected by an AIM-AHEAD leader*



"Say not, "I have found the truth," but rather, "I have found a truth."

— **Kahlil Gibran**

[sli.do questions]

Artificial Intelligence Computational Strategies

1. Natural Language Processing (NLP) for Text Mining:

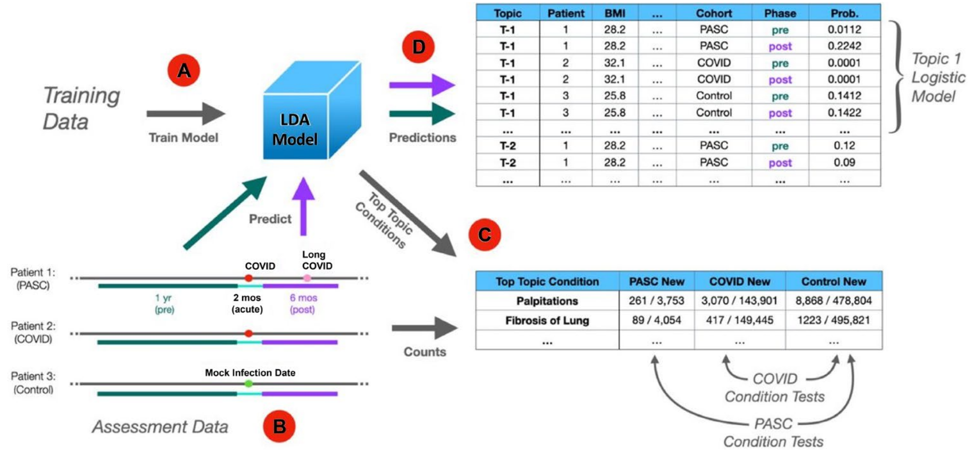
- a. Strategy: Extracting meaningful insights from large volumes of unstructured text data, such as medical literature, clinical notes, or patient narratives.
- b. Applications: Analyzing patient experiences, identifying disparities in healthcare narratives.
- c. Python Libraries: NLTK, SpaCy, gensim.
- d. Large Language Models: GPT, Llama

Artificial Intelligence Computational Strategies

1. Natural Language Processing (NLP) for Text Mining:

- b. Application **examples**:
  - Analyzing patient experiences
  - identifying disparities in healthcare narratives
  - *Classifying diagnostic coding of comorbidities*.

Finding Long-COVID: Temporal Topic Modeling of Electronic Health Records from the N3C and RECOVER Programs

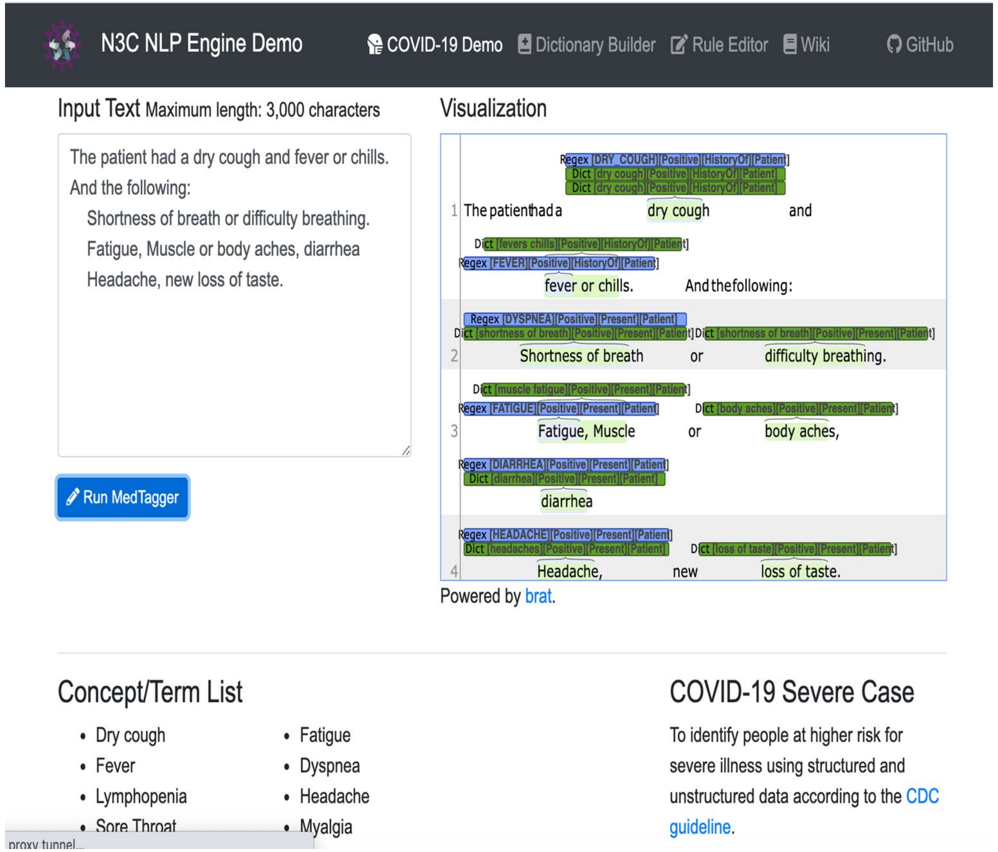


Topic	Patient	BMI	...	Cohort	Phase	Prob.
T-1	1	28.2	...	PASC	pre	0.0112
T-1	1	28.2	...	PASC	post	0.2242
T-1	2	32.1	...	COVID	pre	0.0001
T-1	2	32.1	...	COVID	post	0.0001
T-1	3	25.8	...	Control	pre	0.1412
T-1	3	25.8	...	Control	post	0.1422
...	...	...	...	...	...	...
T-2	1	28.2	...	PASC	pre	0.12
T-2	1	28.2	...	PASC	post	0.09
...	...	...	...	...	...	...

Top Topic Condition	PASC New	COVID New	Control New
Palpitations	261 / 3,753	3,070 / 143,901	8,868 / 478,804
Fibrosis of Lung	89 / 4,054	417 / 149,445	1223 / 495,821
...	...	...	...

https://www.medrxiv.org/content/10.1101/2023.09.11.23295259v1.full-text

c/o Hongfang Liu: N3C NLP Engine …in production


N3C NLP Engine Demo
COVID-19 Demo
Dictionary Builder
Rule Editor
Wiki
GitHub
Run MedTagger
Regex [DRY_COUGH][Positive][HistoryOf][Patient]
Dict [dry cough][Positive][HistoryOf][Patient]
Dict [dry cough][Positive][HistoryOf][Patient]
Dict [fevers chills][Positive][HistoryOf][Patient]
Regex [FEVER][Positive][HistoryOf][Patient]
Regex [DYSPNEA][Positive][Present][Patient]
Dict [shortness of breath][Positive][Present][Patient]
Dict [shortness of breath][Positive][Present][Patient]
Dict [muscle fatigue][Positive][Present][Patient]
Regex [FATIGUE][Positive][Present][Patient]
Dict [body aches][Positive][Present][Patient]
Regex [DIARRHEA][Positive][Present][Patient]
Dict [diarrhea][Positive][Present][Patient]
Regex [HEADACHE][Positive][Present][Patient]
Dict [headaches][Positive][Present][Patient]
Dict [loss of taste][Positive][Present][Patient]
proxy tunnel...

Input Text Maximum length: 3,000 characters

The patient had a dry cough and fever or chills.
And the following:
Shortness of breath or difficulty breathing.
Fatigue, Muscle or body aches, diarrhea
Headache, new loss of taste.

Visualization

1 The patienthad a dry cough and fever or chills. And thefollowing:
2 Shortness of breath or difficulty breathing.
3 Fatigue, Muscle or body aches, diarrhea
4 Headache, new loss of taste.

Powered by brat.

Concept/Term List

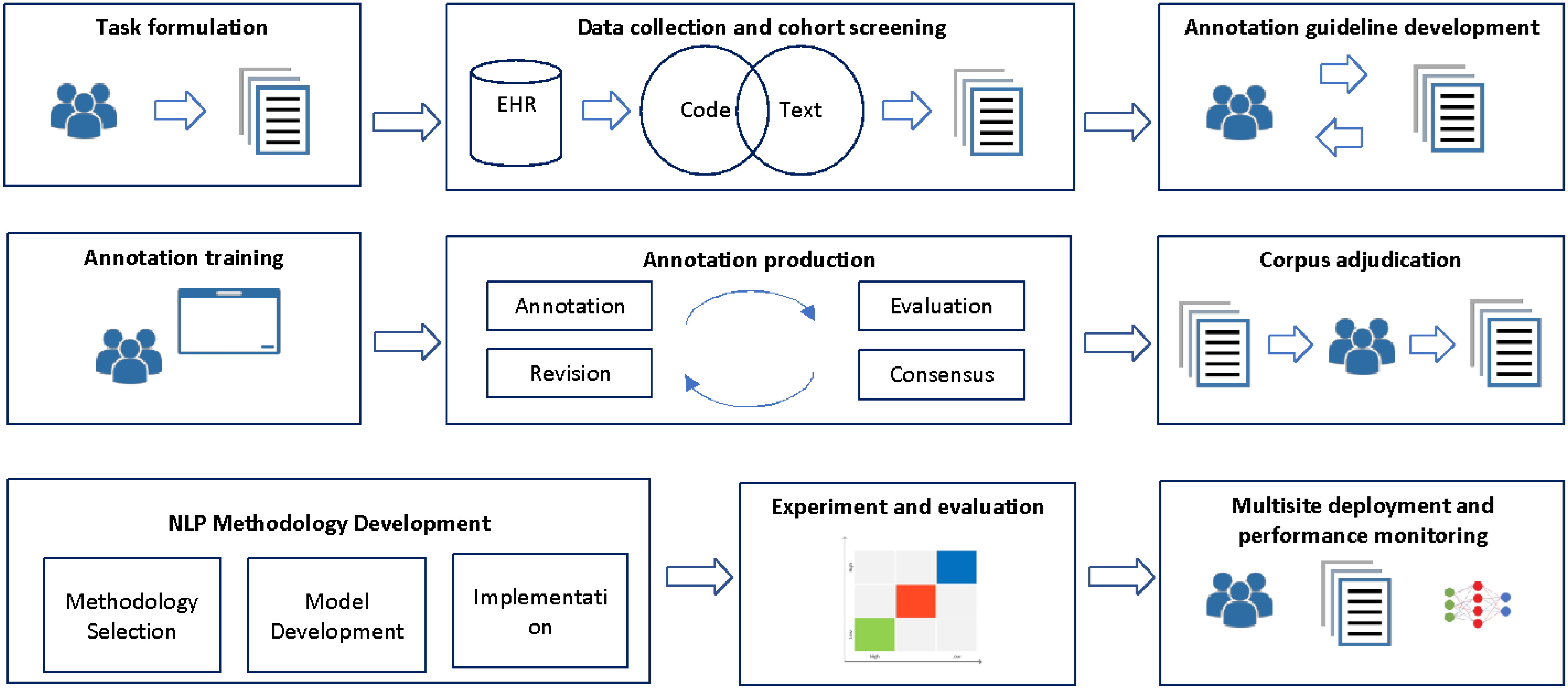
- Dry cough
- Fever
- Lymphopenia
- Sore Throat
- Fatigue
- Dyspnea
- Headache
- Myalgia

COVID-19 Severe Case

To identify people at higher risk for severe illness using structured and unstructured data according to the CDC guideline.

A TRUST Process for NLP Model Development
Text Retrieval and Use towards Scientific rigor and Transparency

c/o Hongfang Liu



Task formulation

Data collection and cohort screening

EHR

Code

Text

Annotation guideline development

Annotation training

Annotation production

Annotation

Revision

Evaluation

Consensus

Corpus adjudication

NLP Methodology Development

Methodology Selection

Model Development

Implementati on

Experiment and evaluation

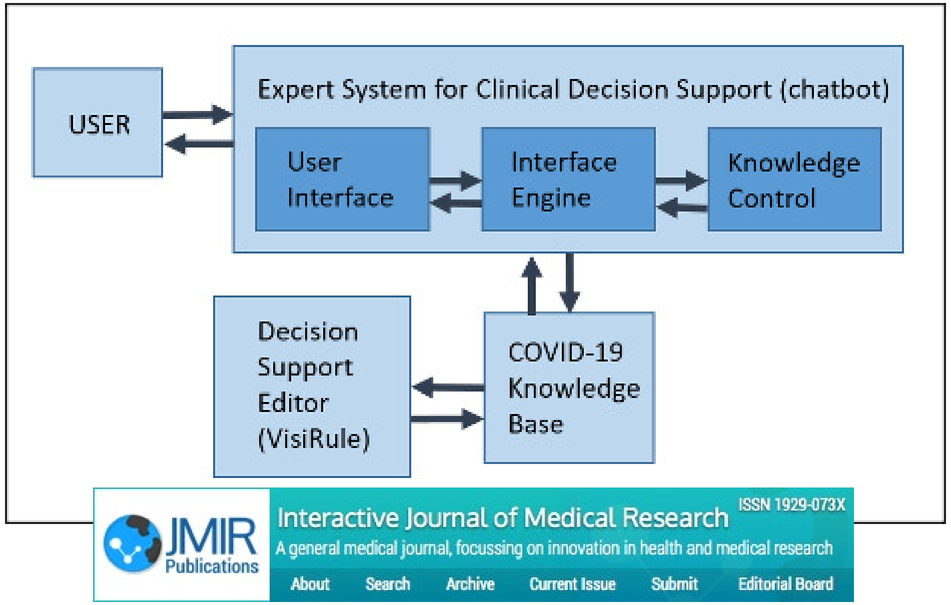
Multisite deployment and performance monitoring

Artificial Intelligence Computational Strategies

2. Expert Systems for Decision Support:

- a. Strategy: Building rule-based systems that emulate human expertise to assist healthcare professionals in decision-making.
- b. Applications: Diagnosis support, treatment planning. **Example @ right**
- c. Python Libraries: mainly ‘rules engines’ like [Experta](), c.2018 [PyKnow](), c.2010 [Pyke]()…

Using Decision Trees as an Expert System for Clinical Decision Support for COVID-19


USER
Expert System for Clinical Decision Support (chatbot)
User Interface
Interface Engine
Knowledge Control
Decision Support Editor (VisiRule)
COVID-19 Knowledge Base

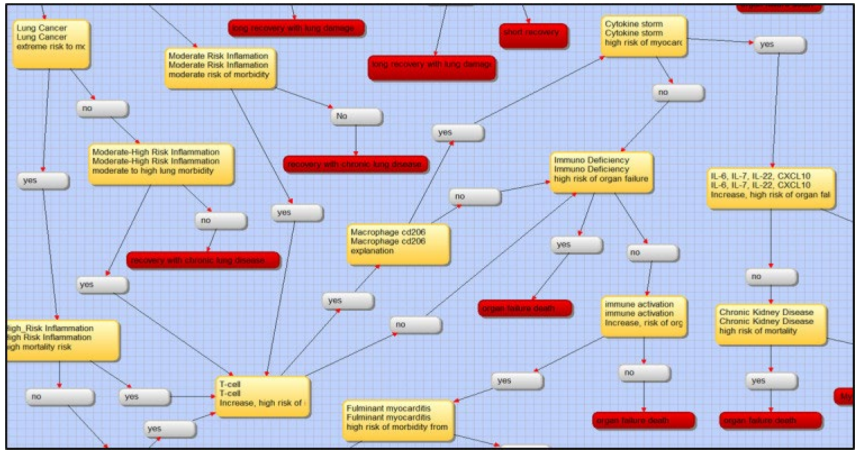
JMIR Publications

Interactive Journal of Medical Research

ISSN 1929-073X

A general medical journal, focussing on innovation in health and medical research

About Search Archive Current Issue Submit Editorial Board

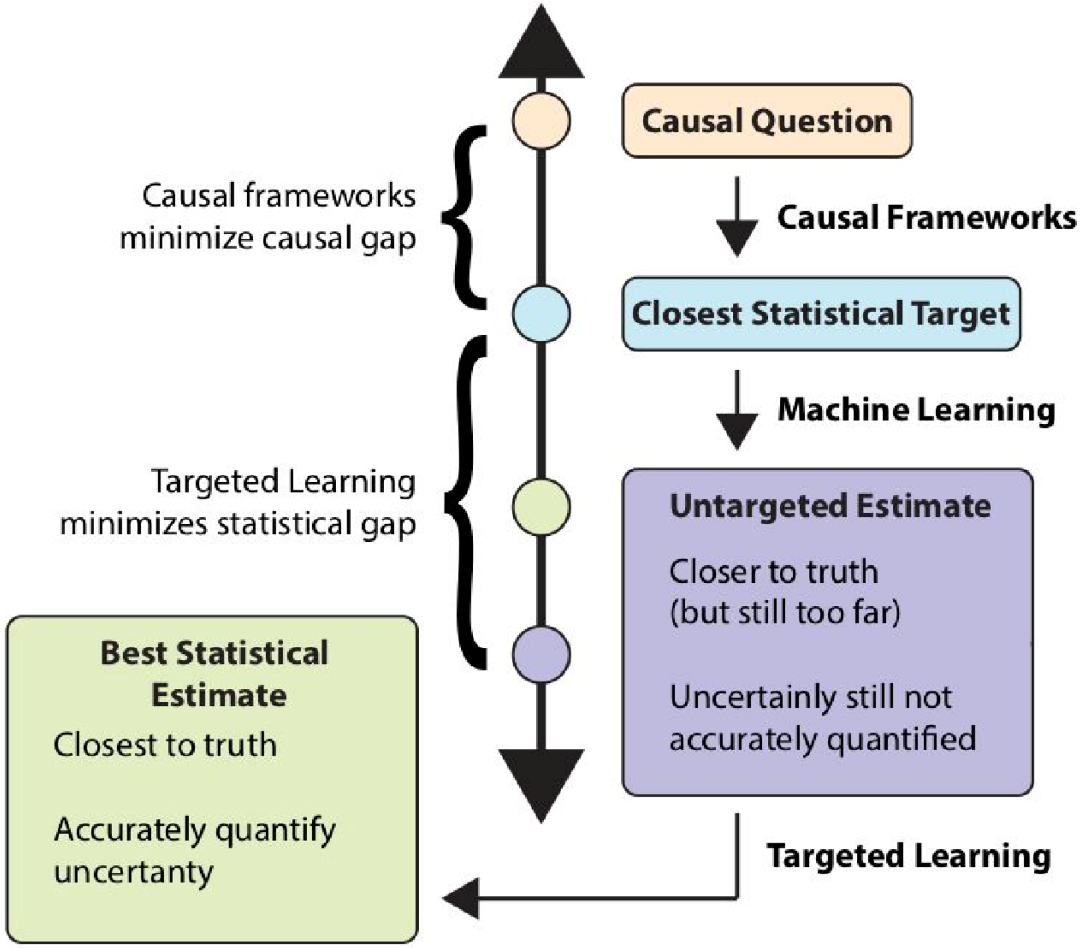


Visualized with graphical artificial intelligence software VisiRule

Artificial Intelligence Computational Strategies

3. Causal Inference Modeling using Machine Learning Algorithms:

- a. Strategy: Inferring causal relationships between variables in healthcare data to understand the impact of interventions or factors on health outcomes.
- b. Applications: Studying the effect of interventions on healthcare disparities, unclear if adequate portion of Big Tech investment.
- c. Python Libraries: CausalImpact, DoWhy, CausalLib (TMLE example doc), zEpid (TMLE doc), causal-curve, mossspider.



Causal Question

Causal Frameworks

Closest Statistical Target

Machine Learning

Untargeted Estimate

Closer to truth (but still too far)

Uncertainly still not accurately quantified

Targeted Learning

Best Statistical Estimate

Closest to truth

Accurately quantify uncertanty

Causal frameworks minimize causal gap

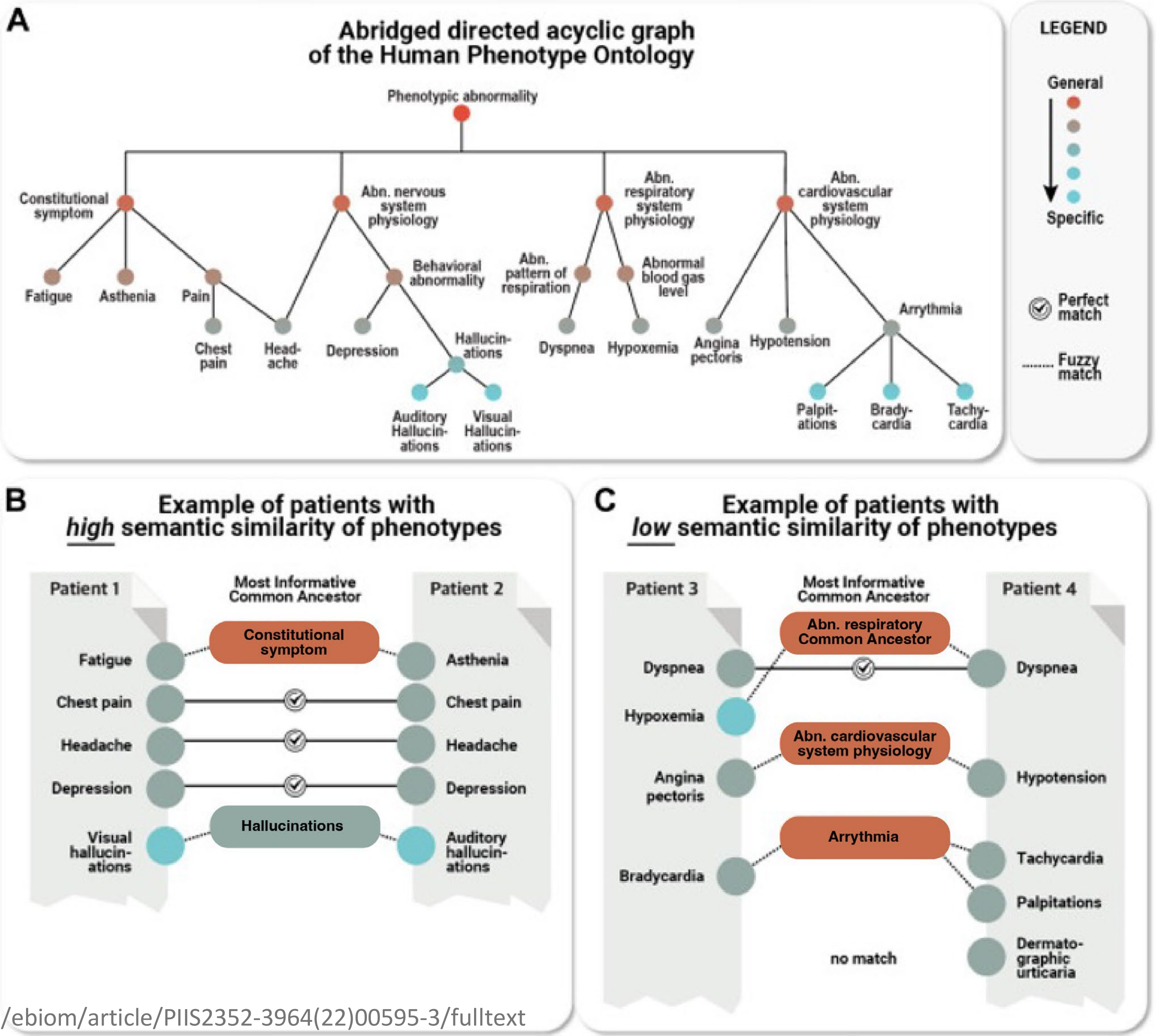
Targeted Learning minimizes statistical gap

Coyle, Jeremy R., Nima S. Hejazi, Ivana Malenica, Rachael V. Phillips, Benjamin F. Arnold, Andrew N. Mertens, Jade Benjamin-Chung, Weixin Cai, Sonali Dayal, John M. Colford, Alan E. Hubbard and Mark J. van der Laan. "Targeting Learning: Robust Statistics for Reproducible Research." *arXiv: Methodology* (2020): n. pag.

Artificial Intelligence Computational Strategies

4. Ontology and Knowledge Graphs:

- a. Strategy: Organizing and representing medical knowledge in structured formats to facilitate semantic understanding.
- b. Applications: Linking disparate healthcare data sources, enhancing interoperability.
- c. Python Libraries: [RDFlib](), [Owlready2](), [OntoGPT]()

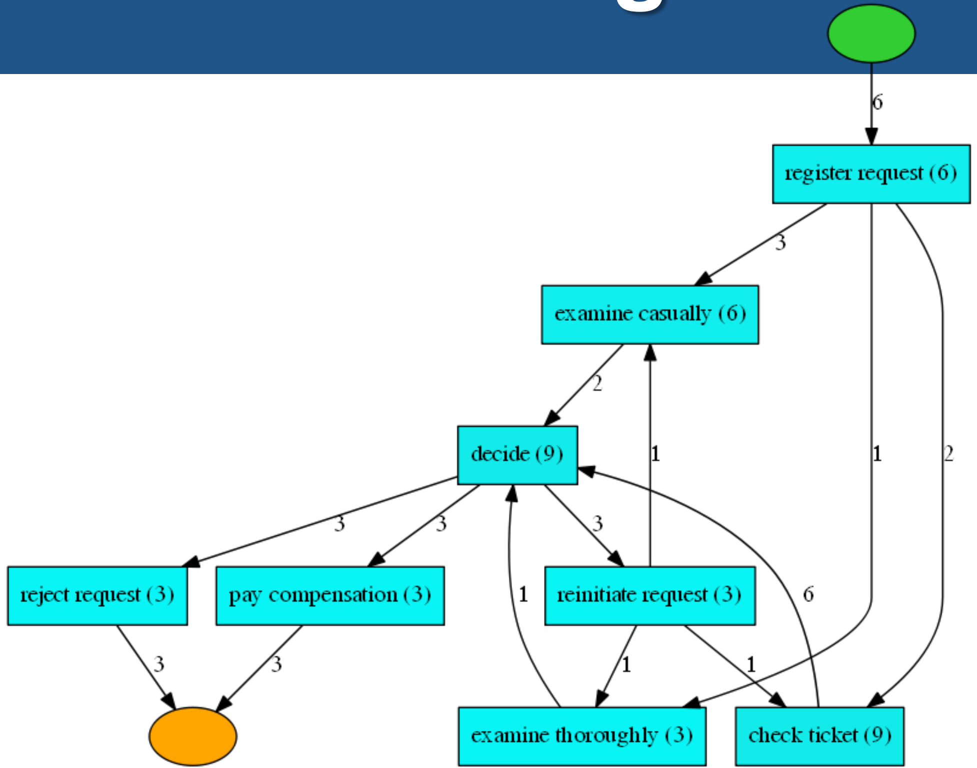


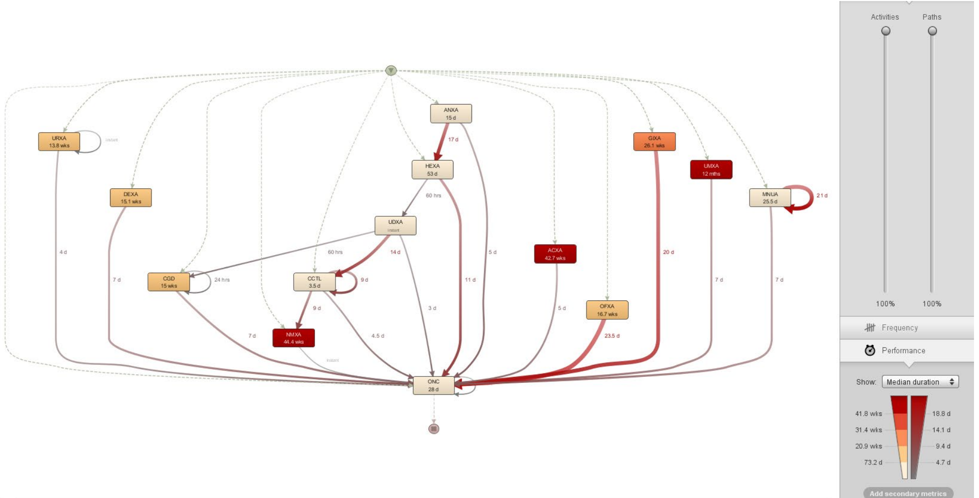
https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(22)00595-3/fulltext

Artificial Intelligence Computational Strategies

5. Process Mining:

- a. Strategy: Analyzing healthcare processes to understand workflow, identify bottlenecks, and optimize resource allocation.
- b. Applications: Improving efficiency in healthcare delivery.
- c. Python Libraries: pm4py, ProM.



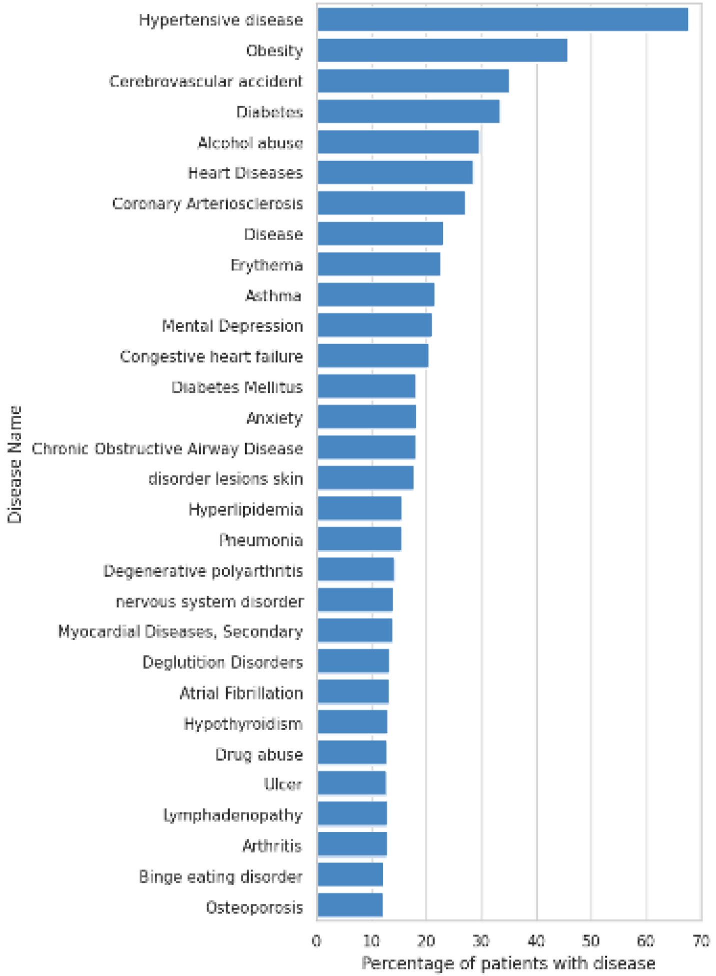


process mining workbench

Artificial Intelligence Computational Strategies

6. Automated Coding and Classification:

- a. Strategy: Developing systems that automate the coding and classification of medical records for standardized reporting and analysis.
- b. Applications: Streamlining data coding processes, ensuring consistency.
- c. Python Libraries: MedCAT, PyCaret.



https://colab.research.google.com/github/CogStack/MedCATtutorials/blob/main/notebooks/introductory/Part_3_2_Extracting_Diseases_from_Electronic_Health_Records.ipynb#scrollTo=TupbSS6OVfgM

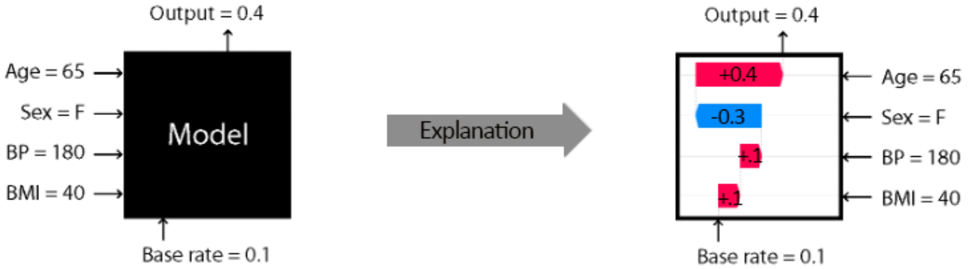
Artificial Intelligence Computational Strategies

7. Decision Support Systems with ***Explainability***:

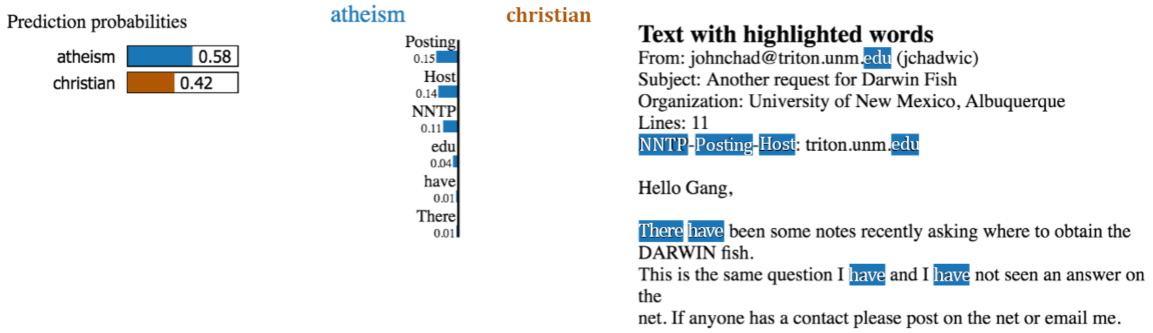
- a. Strategy: Creating AI systems that not only provide recommendations but also explain the reasoning behind the suggestions.
- b. Applications: Enhancing transparency and trust in decision support. **examples**
- c. Python Libraries: SHAP, Lime (Local Interpretable Model-Agnostic Explanations).

Examples quantify and visually show how specific features 'weigh in' on results…

SHAP



Local Interpretable Model-Agnostic Explanations



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the
DARWIN fish.
This is the same question I have and I have not seen an answer on
the
net. If anyone has a contact please post on the net or email me.

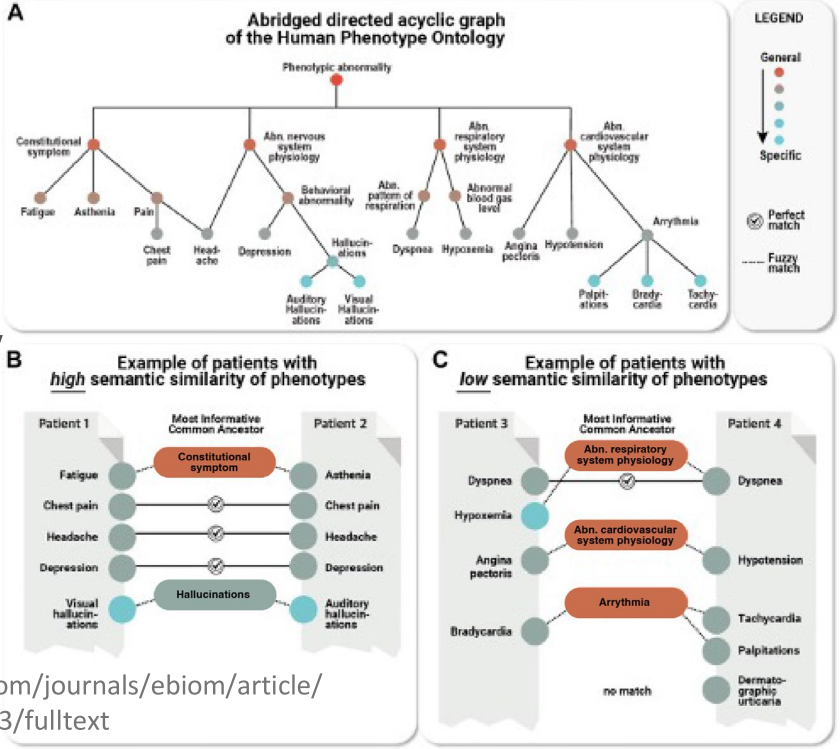
Artificial Intelligence Computational Strategies

Example: semantic-similarity-elicited long COVID types

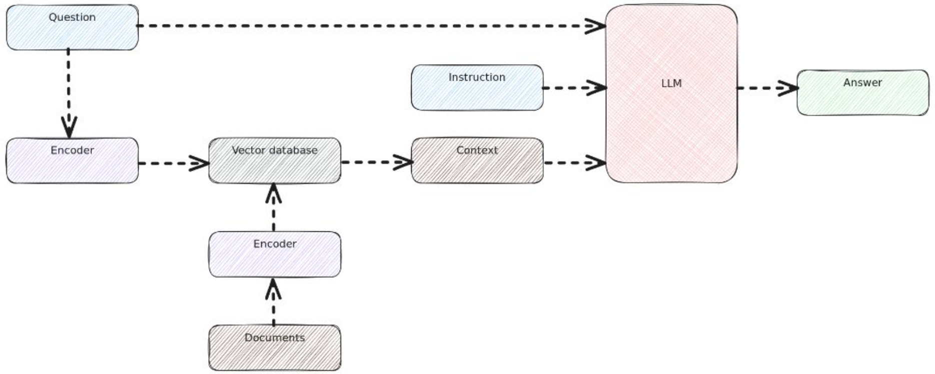
8. Semantic Analysis for Data Integration:

- a. Strategy: Applying semantic techniques to integrate heterogeneous healthcare data from various sources.
- b. Applications: Facilitating cross-domain data integration, enhancing data interoperability.
- c. Python Libraries: RDFlib, Owlready2, OntoGPT
- d. other examples recently emerging:
  - OntoGPT-related SPIRES - Semantic similarity
  - Retrieval Augmented Generation
    - within Large Language Model Prompts (diagram @ right)

Calculating patient semantic similarity based on HPO phenotypes. A) HPO terms are arranged in a directed acyclic graph with specific terms -- excerpt of the entire ontology (15,247 terms) is shown. B) Example showing a pair of patients with relatively high phenotypic similarity



https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(22)00595-3/fulltext



Paper on race-bias in this space is, unfortunately, behind a paywall (even for NIH): https://ieeexplore.ieee.org/document/9669617

Artificial Intelligence Computational Strategies

Strategies Employed in Use Cases

AI Computational Strategies

- We now engage participants to check our mutual understanding.

Artificial Intelligence: Fundamental Algorithms

Note: we provide link to asynchronous hands-on after ML portion…

We now quickly outline remaining number of algorithms primarily in use within AI/ML:

[*from 14 popular AI algorithms and their uses post*]

Popular deep learning algorithms

There are a number of very successful and widely adopted deep learning paradigms, the most recent being the transformer architecture behind today's generative AI models.

10 Convolutional neural networks

Convolutional neural networks (CNNs) are a type of deep neural network often used for machine vision. They have the desirable property of being position-independent. The understandable summary of a convolution layer when applied to images is that it slides over the image spatially, computing dot products; each unit in the layer shares one set of weights. A *convnet* typically uses multiple convolution layers, interspersed with activation functions. CNNs can also have pooling and fully connected layers, although there is a trend toward getting rid of these types of layers.

11 Recurrent neural networks

While convolutional neural networks do a good job of analyzing images, they don't really have a mechanism that accounts for time series and sequences, as they are strictly feed-forward networks. Recurrent neural networks (RNNs), another kind of deep neural network, explicitly include feedback loops, which effectively gives them some memory and dynamic temporal behavior and allows them to handle sequences, such as speech. That doesn't mean that CNNs are useless for natural language processing; it does mean that RNNs can model time-based information that escapes CNNs. And it doesn't mean that RNNs can *only* process sequences. RNNs and their derivatives have a variety of application areas, including language translation, speech recognition and synthesis, robot control, time series prediction and anomaly detection, and handwriting recognition. While in theory an ordinary RNN can carry information over an indefinite number of steps, in practice it generally can't go many steps without losing the context. One of the causes of the problem is that the gradient of the network tends to vanish over many steps, which interferes with the ability of a gradient-based optimizer such as stochastic gradient descent (SGD) to converge.

Artificial Intelligence: Fundamental Algorithms

- *Note: we are including these passages only to expose you to terms…*

12 Long short-term memory Long short-term memory networks (LSTMs) were explicitly designed to avoid the vanishing gradient problem and allow for long-term dependencies. The design of an LSTM adds some complexity compared to the cell design of an RNN, but works much better for long sequences. In LSTMs, the network is capable of forgetting (gating) previous information as well as remembering it, in both cases by altering weights. This effectively gives an LSTM both long-term and short-term memory, and solves the vanishing gradient problem. LSTMs can deal with sequences of hundreds of past inputs.

13 Transformers Transformers are neural networks that solely use *attention* mechanisms, dispensing with recurrence and convolutions entirely. Transformers were invented at Google. Attention units (and transformers) are part of Google's BERT (Bidirectional Encoder Representations from Transformers) algorithm and OpenAI's GPT-2 algorithm (transformer model with unsupervised pre-training) for natural language processing. Transformers continue to be integral to the neural architecture of the latest large language models, such as ChatGPT/Bing Chat (based on GPT-3.5 or GPT-4) and Bard (based on LaMDA, which stands for Language Model for Dialogue Applications). Attention units are not terribly sensitive to how close two words in a sentence appear, unlike RNNs; that makes them good at tasks that RNNs don't do well, such as identifying antecedents of pronouns that may be separated from the referent pronouns by several sentences. Attention units are good at looking at a context larger than just the last few words preceding the current word.

14 Q-learning Q-learning is a model-free, value-based, off-policy algorithm for reinforcement learning that will find the best series of actions based on the current state. The "Q" stands for quality. Quality represents how valuable the action is in maximizing future rewards. Q-learning is essentially learning by experience. Q-learning is often combined with deep neural networks. It's used with convolutional neural networks trained to extract features from video frames, for example for teaching a computer to play video games or for learning robotic control. AlphaGo and AlphaZero are famous successful game-playing programs from Google DeepMind that were trained with reinforcement learning combined with deep neural networks. As we've seen, there are many kinds of machine learning problems, and many algorithms for each kind of problem. These range in complexity from linear regression for numeric prediction to convolutional neural networks for image processing, transformer-based models for generative AI, and reinforcement learning for game-playing and robotics.

Artificial Intelligence: Pros & Cons

Per Think-a-thon Planning outline:

- *Strengths:*
  - *Flexible to multiple data modalities and – with ENOUGH data – quite robust,*
  - *Some aspects are ‘explainable’ through additional ‘extra’ steps*
- *Weaknesses:*
  - *NOT interpretable,*
  - *assumption-dense, yet assumptions typically NOT transparently assessed*
  - *often very dependent upon the tacit decisions made by those applying AI*

ScHARe

AI Bias

AI bias

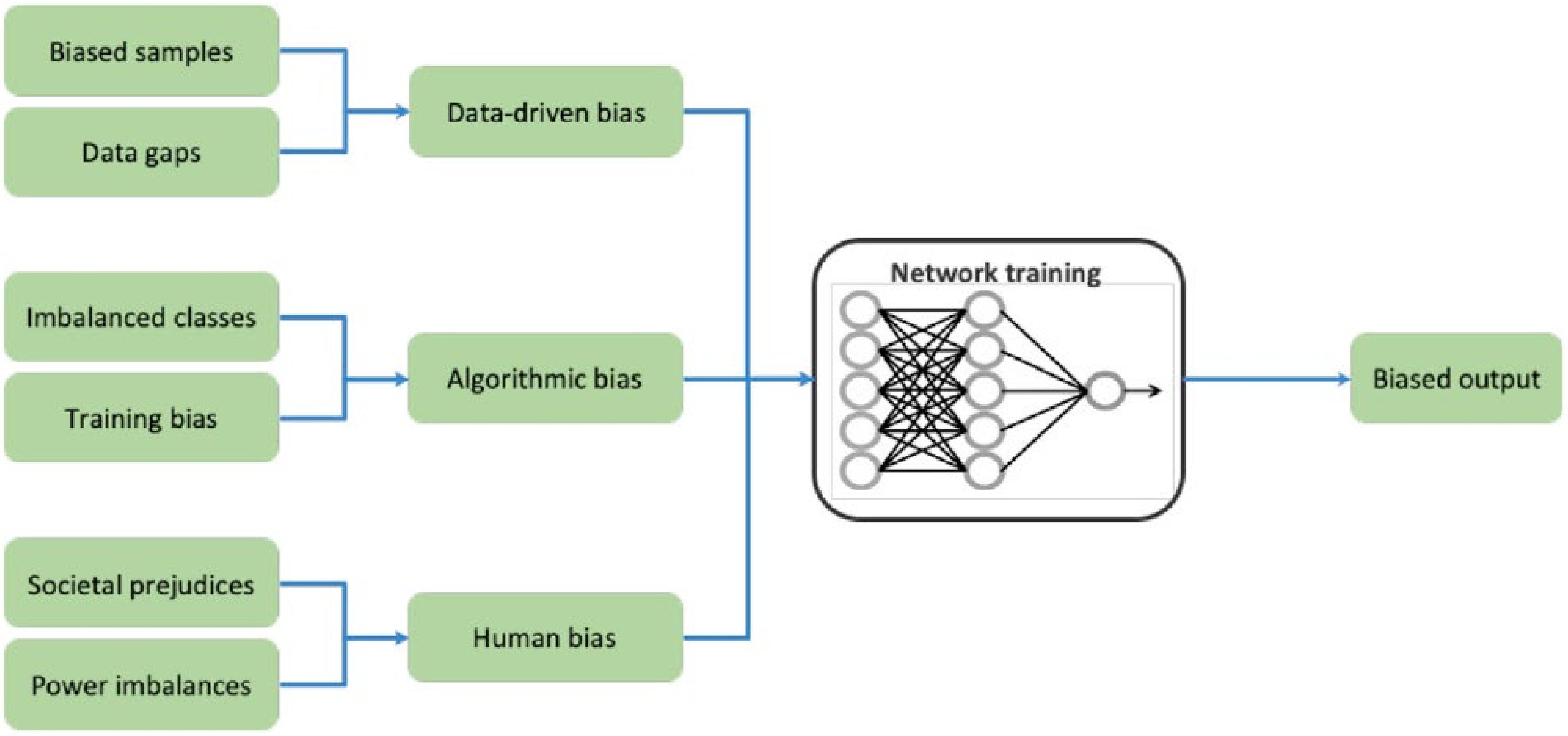
- Algorithms are using Big Data to **influence decisions affecting people's health.**

- **Training data** that specifies what the correct outputs are for some people/objects **is used to learn a model** which is then applied to other people/objects to make predictions about the correct outputs for them

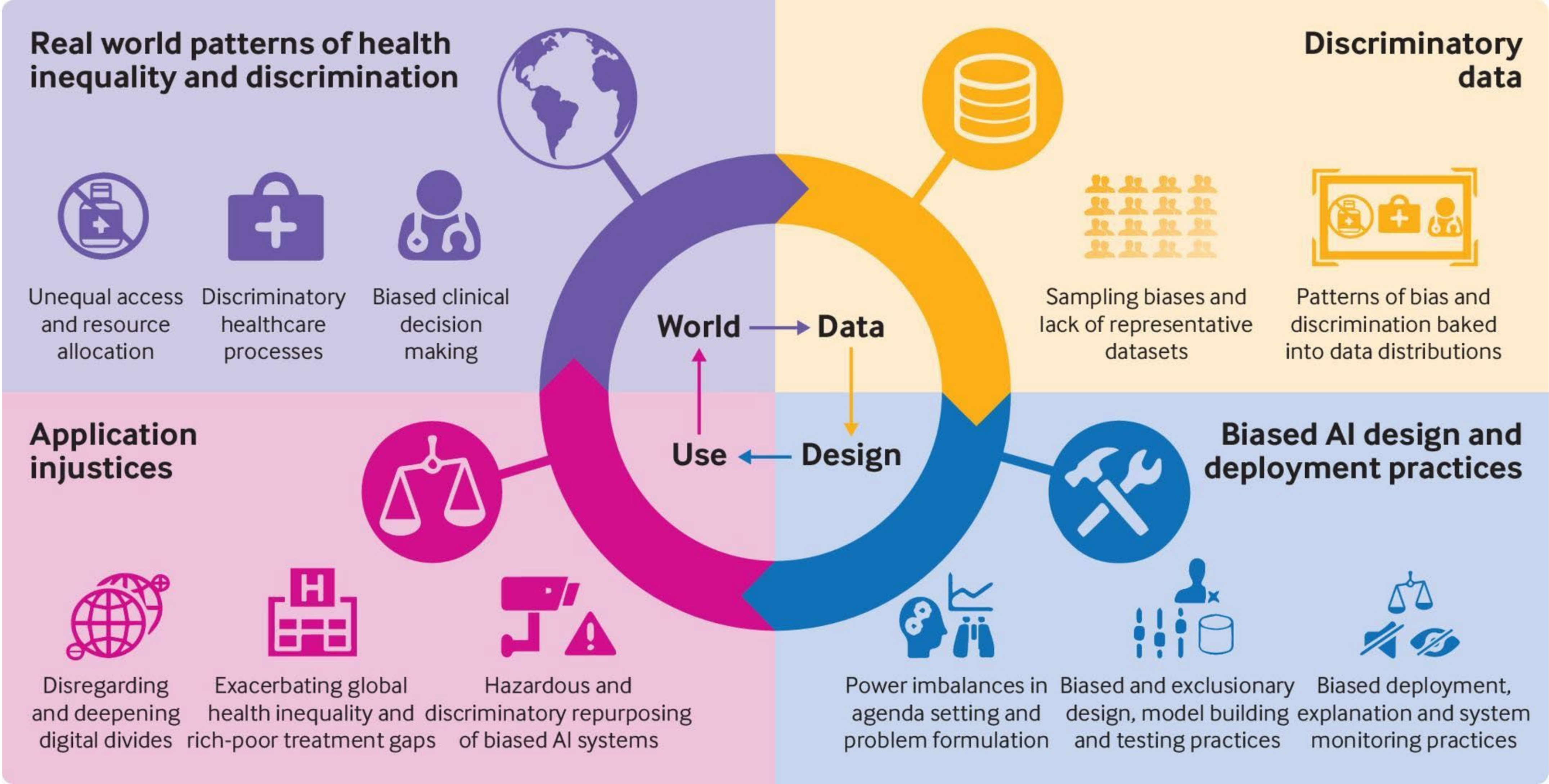
- Algorithms run the **risk of replicating and amplifying human biases** affecting protected groups, leading to outcomes systematically less favorable to them

- **Bias can originate from unrepresentative/incomplete training data** that reflects historical inequalities, or manifest at various points in the algorithm development process

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

Application injustices

Disregarding and deepening digital divides

Exacerbating global health inequality and rich-poor treatment gaps

Hazardous and discriminatory repurposing of biased AI systems

Biased AI design and deployment practices

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.

Testing for biases in datasets and algorithms

- Testing for biases in datasets and algorithmic models is **crucial for ensuring fairness and reliability** in data science.

- Here are general strategies and **techniques for testing biases**, categorized into datasets and algorithmic models.

Testing for biases in datasets

1. **Exploratory Data Analysis (EDA):**

   - **Explanation:** EDA involves visualizing and summarizing the main characteristics of the dataset using histograms, box plots, and summary statistics. The goal is to understand the data distribution
   - **Importance:** EDA helps identify outliers, imbalances, and biases
   - **Example:** If EDA reveals a dataset on job applicants is heavily skewed towards a specific gender, it might indicate a bias in the sampling process
   - **Python Libraries:** Pandas, Matplotlib, Seaborn

Testing for biases in datasets

2. **Demographic Analysis (DA):**

   - **Explanation:** Break down the dataset based on demographic attributes (e.g., age, gender, ethnicity) and analyze the distribution within each group
   - **Importance:** DA can identify imbalances/over-representations in specific groups
   - **Example:** In a healthcare dataset, if one demographic group is over-represented, it may lead to biased predictions
   - **Python Libraries:** Pandas, Matplotlib, Seaborn

Testing for biases in datasets

3. **Data Stratification:**

   - **Explanation:** Divide the dataset into subgroups based on relevant features and analyze each subgroup independently
   - **Importance:** This helps detect biases that may exist disproportionately in specific subgroups
   - **Example:** In a credit scoring dataset, stratifying by income levels can reveal biases in credit approval rates
   - **Python Libraries:** Pandas

Testing for biases in datasets

4. **Bias Detection Tools:**

   - **Explanation:** Use tools like IBM's AI Fairness 360 or Google's What-If Tool that offer automated metrics for assessing bias in datasets and models
   - **Importance:** Automated tools efficiently identify subtle biases and provide quantitative measures, facilitating a systematic approach to bias detection
   - **Examples:**
     - AI Fairness 360 provides a set of algorithms to evaluate fairness across various demographic groups
     - Google's What-If Tool allows interactive exploration of model predictions and visualization of outcomes across different subsets of data
   - **Tools:** AI Fairness 360, What-If Tool

Fixing biases in datasets

Several techniques can be employed to address bias in datasets:

- **Oversampling** involves increasing the representation of underrepresented groups in the dataset, ensuring a more balanced distribution
- **Undersampling** reduces overrepresented groups
- **Using synthetic data** generation introduces artificially generated data points to mitigate imbalances
- **Reweighting** or adjusting the importance of specific instances during model training helps address bias
- Regularly **updating and expanding datasets** with diverse, representative samples further contribute to minimizing bias

Testing for biases in algorithms

1. **Performance Metrics Disaggregation:**

   - **Explanation:** Evaluate model performance metrics (e.g., accuracy, precision) separately for different subgroups defined by sensitive attributes
   - **Importance:** Disparities in performance metrics across groups may indicate bias
   - **Example:** Testing a healthcare algorithm disaggregating accuracy by racial groups reveals slightly lower accuracy for Black patients. Fixes: root cause analysis and algorithm adjustments
   - **Python Libraries:** Scikit-learn

Testing for biases in algorithms

2. **Confusion Matrix Analysis:**

   - **Explanation:** Analyze the confusion matrix (a table that summarizes the performance of a classification algorithm by comparing predicted and actual values) for different subgroups to identify disparities in model predictions, particularly for false positives and false negatives
   - **Importance:** Disparities in errors can pinpoint areas where bias may exist
   - **Example:** Analyzing a medical diagnosis algorithm using a confusion matrix to evaluate the model's effectiveness in making medical diagnoses. Differences in false positives between genders might indicate bias. Fix: adjusting decision thresholds, retraining with balanced data, consulting domain experts
   - **Python Libraries:** Scikit-learn

Testing for biases in algorithms

3. **Fairness Indicators:**

    - **Explanation:** Integrate fairness indicators (measures that assess whether a model's predictions treat different groups equitably) into the model evaluation process to identify bias
    - **Importance:** Fairness indicators provide a structured approach to measure bias
    - **Example:** Using Google's TensorFlow Fairness Indicators to compare prediction accuracies of a healthcare decision support algorithm across different racial groups. Fixes: retraining the algorithm with balanced data, adjusting decision thresholds
    - **Python Libraries:** TensorFlow Fairness Indicators

Testing for biases in algorithms

4. **Sensitivity Analysis:**

   - **Explanation:** Assess how changes in input features impact model predictions. This involves tweaking one feature at a time and observing the model's response
   - **Importance:** It helps identify features that disproportionately influence the model, potentially leading to biases
   - **Example:** In a healthcare decision support algorithm predicting diabetes risk, assessing how variations in input variables (e.g., age, BMI) impact predictions for different racial groups. The analysis reveals that the algorithm disproportionately relies on a single variable affecting certain groups. Fixes: recalibrating the model to minimize the influence of that variable, retraining with a more diverse dataset
   - **Python Libraries:** Scikit-learn

Testing for biases in algorithms

5. **Counterfactual Analysis:**

    - **Explanation:** Counterfactual analysis involves exploring hypothetical scenarios by determining the minimal changes needed in input features to alter a model's prediction
    - **Importance:** It helps understand the model's decision boundaries and can highlight biases
    - **Example:** In a credit approval algorithm, if a loan application from a certain racial group is denied, the analysis involves identifying the minimal changes needed in the application features (income, credit score) for approval, shedding light on potential biases. Fixes: adjusting the decision thresholds, mitigating the impact of sensitive features, or retraining the model
    - **Python Libraries:** Alibi Counterfactual

Machine Learning Unveiled as a Bridge-building Trailblazer (*really a set of bridging paths falling under the AI* umbrella!*)

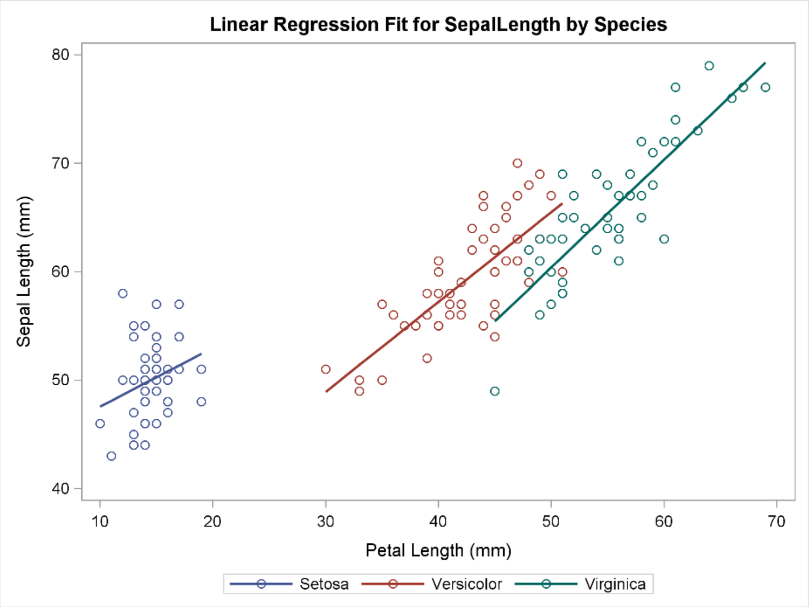
A. ML Essentials: roots in data analysis methods

B. Computational Strategies: varied forms of 'learning' and applying 'learning' algorithms…

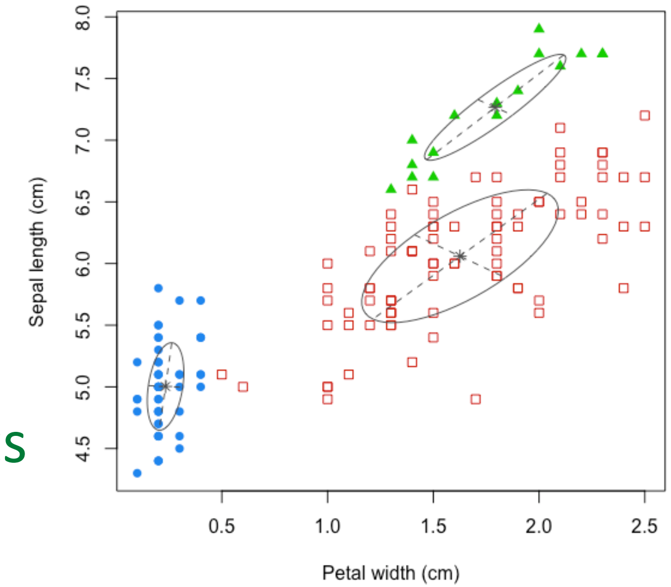
* Recall *cynical* definition offered at recent NIH meeting: if it *actually* works in practice somehow, it's 'machine learning' otherwise it may just be termed 'artificial intelligence' that still has more to learn…

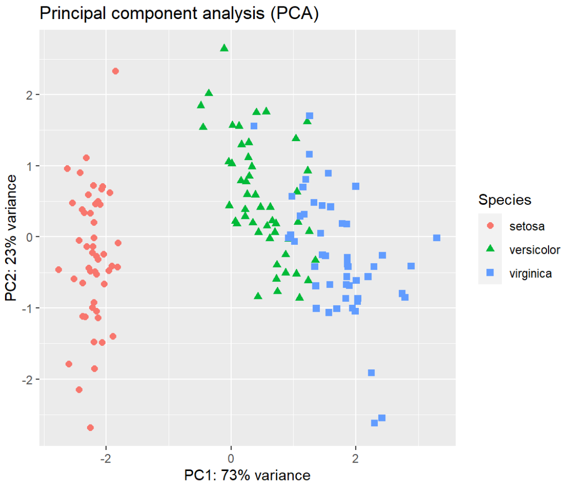
ML Essentials: roots in data analysis methods

- Data analysis methods to ‘learn’ how to predict patterns in data
  - Classic iris flower regression example



- Data analysis methods to ‘learn’ *novel* patterns in data: clustering & ‘mixture modeling’
  - Discover ‘clusters’ by length measures
  - Data reduction by principal components

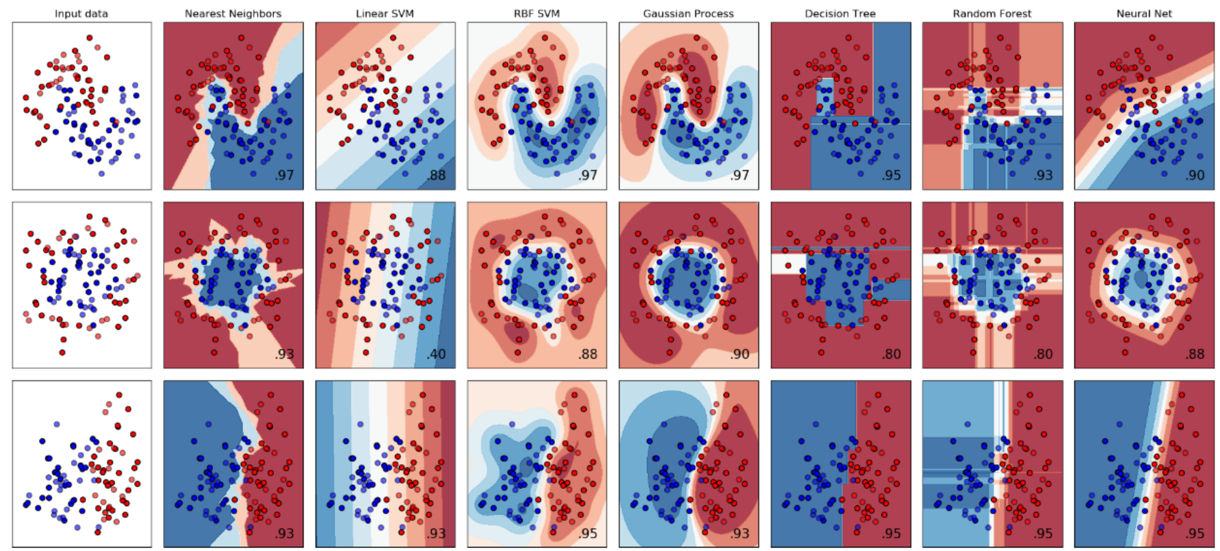




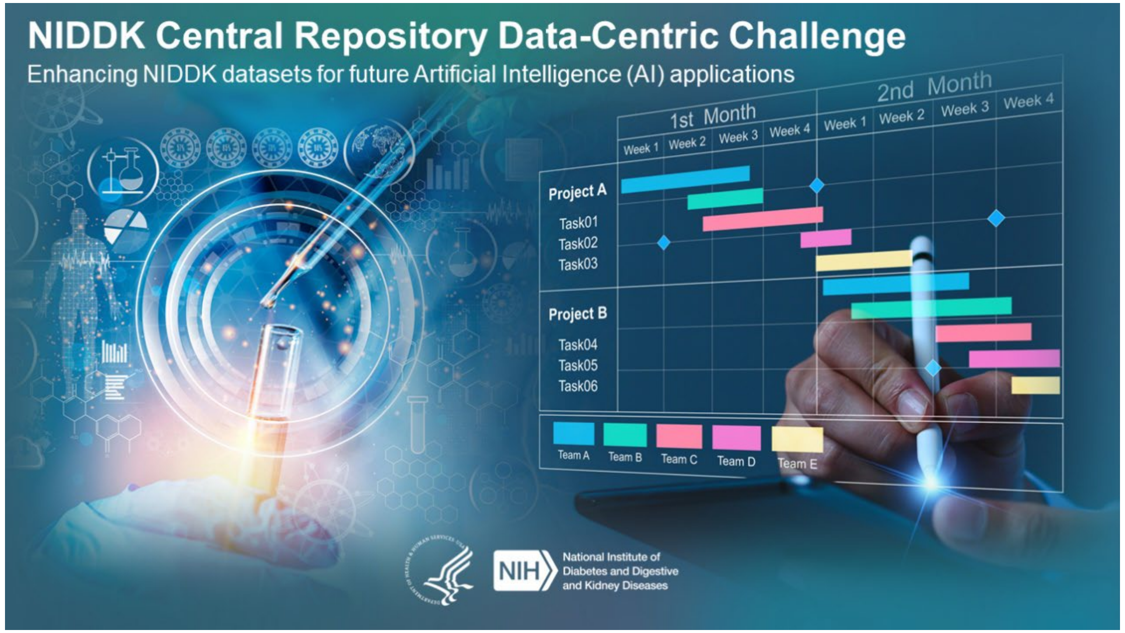
https://exeter-data-analytics.github.io/MachineLearning/_main_files/figure-html/unnamed-chunk-72-1.png | https://gexijin.github.io/learnR/book_files/figure-html/13-5-1.png

ML Essentials: roots in data analysis methods

- Data analysis methods to ‘learn’ how to predict patterns in data
- Data analysis methods to ‘learn’ *novel* patterns in data: clustering



- Relates to UN-supervised v. semi-supervised v. Supervised learning
  - Hearken back to prior ScHARe Think-a-thon
  - Underway: PHASE 2 of NIDDK CR Data-Centric Challenge (till Jan 22, 2024)

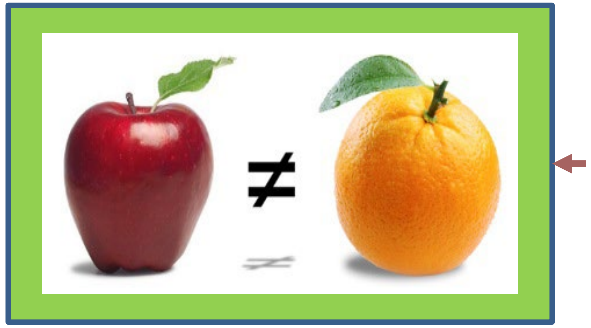

NIDDK Central Repository Data-Centric Challenge
Enhancing NIDDK datasets for future Artificial Intelligence (AI) applications
1st Month
2nd Month
Week 1
Week 2
Week 3
Week 4
Week 1
Week 2
Week 3
Week 4
Project A
Task01
Task02
Task03
Project B
Task04
Task05
Task06
Team A
Team B
Team C
Team D
Team E
NIH
National Institute of Diabetes and Digestive and Kidney Diseases

ML Essentials: supervised v. semi-supervised v. unsupervised learning

‘Machine Learning’ as a tool for Data Science (thus, for health equity research)

- *Does one term cover all approaches?* Types of ML, matching use cases & data
- *e.g. (**extent of ‘supervision’**; goals of analysis)*

- *What does “extent of ‘supervision’” mean in this context?*



Supervised Learning

Supervision here: *each instance is given exactly 1 ‘label’ to distinguish*

https://www.gettyimages.ca/detail/news-photo/republician-presidential-hopeful-and-texas-governor-george-news-photo/51543471?adppopup=true

ML Essentials: supervised v. semi-supervised v. unsupervised learning

'Machine Learning' as a tool for Data Science (thus, for health equity research)

- *Does one term cover all approaches?* Types of ML, matching use cases & data
- *e.g. (**extent of 'supervision'**; goals of analysis)*

- *What does "extent of 'supervision'" mean in this context?*

Semi-supervised Learning

RED color shows up on the **RIGHT**

No picture lacks a 'mirror' image

Supervision here: *Only some instances given a 'label' to distinguish 'labeling' pattern overall…*

https://www.gettyimages.ca/detail/news-photo/republican-presidential-hopeful-and-texas-governor-george-news-photo/51543471?adppopup=true

ML Essentials: supervised v. semi-supervised v. unsupervised learning

'Machine Learning' as a tool for Data Science (thus, for health equity research)

- *Does one term cover all approaches?* Types of ML, matching use cases & data
- *e.g. (**extent of 'supervision'**; goals of analysis)*

- *What does "extent of 'supervision'" mean in this context?*

Four-eyed Fox
Four-eyed Wolf
Red kidneys & butterflies around a 2 person drum circle

Unsupervised Learning

Supervision here: *Only intrinsic parts of instances used to 'label' them, elicit any pattern overall…*

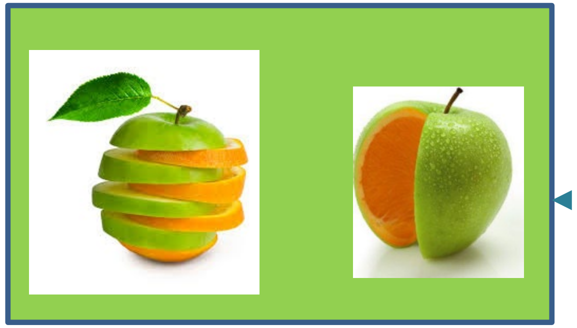
https://www.gettyimages.ca/detail/news-photo/republician-presidential-hopeful-and-texas-governor-george-news-photo/51543471?adppopup=true

ML Essentials: supervised v. semi-supervised v. unsupervised learning

‘Machine Learning’ as a tool for Data Science (thus, for health equity research)

- *Does one term cover all approaches?* Types of ML, matching use cases & data
- *e.g. (**extent of ‘supervision’**; goals of analysis)*

- *What does “extent of ‘supervision’” mean in this context?*



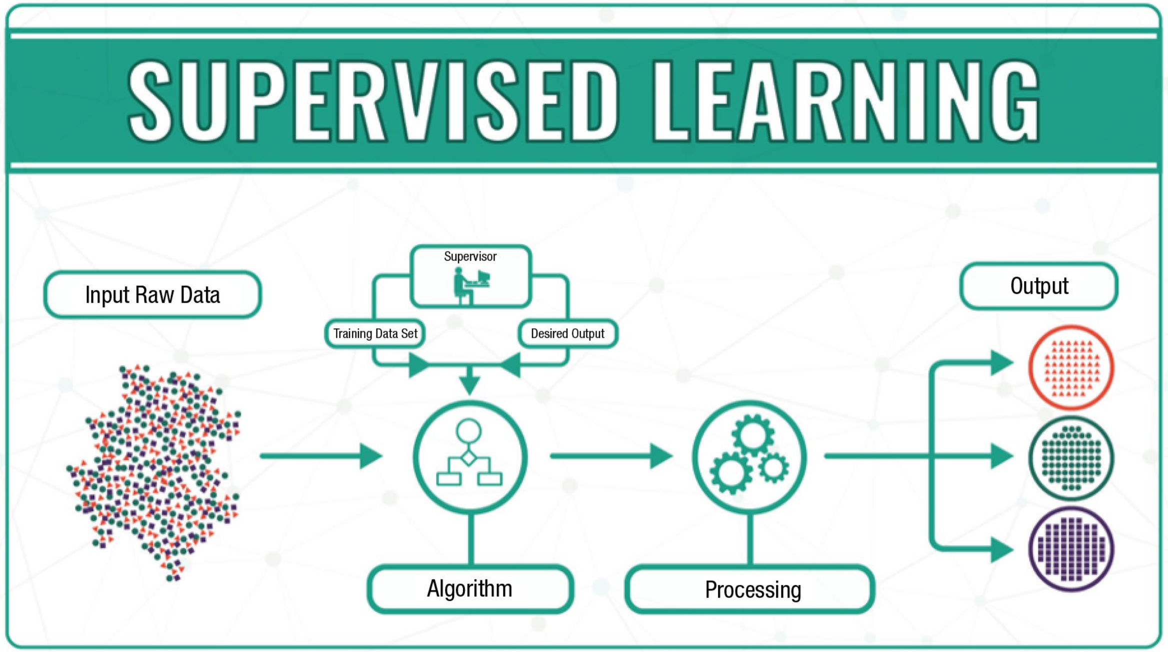
Unsupervised Learning

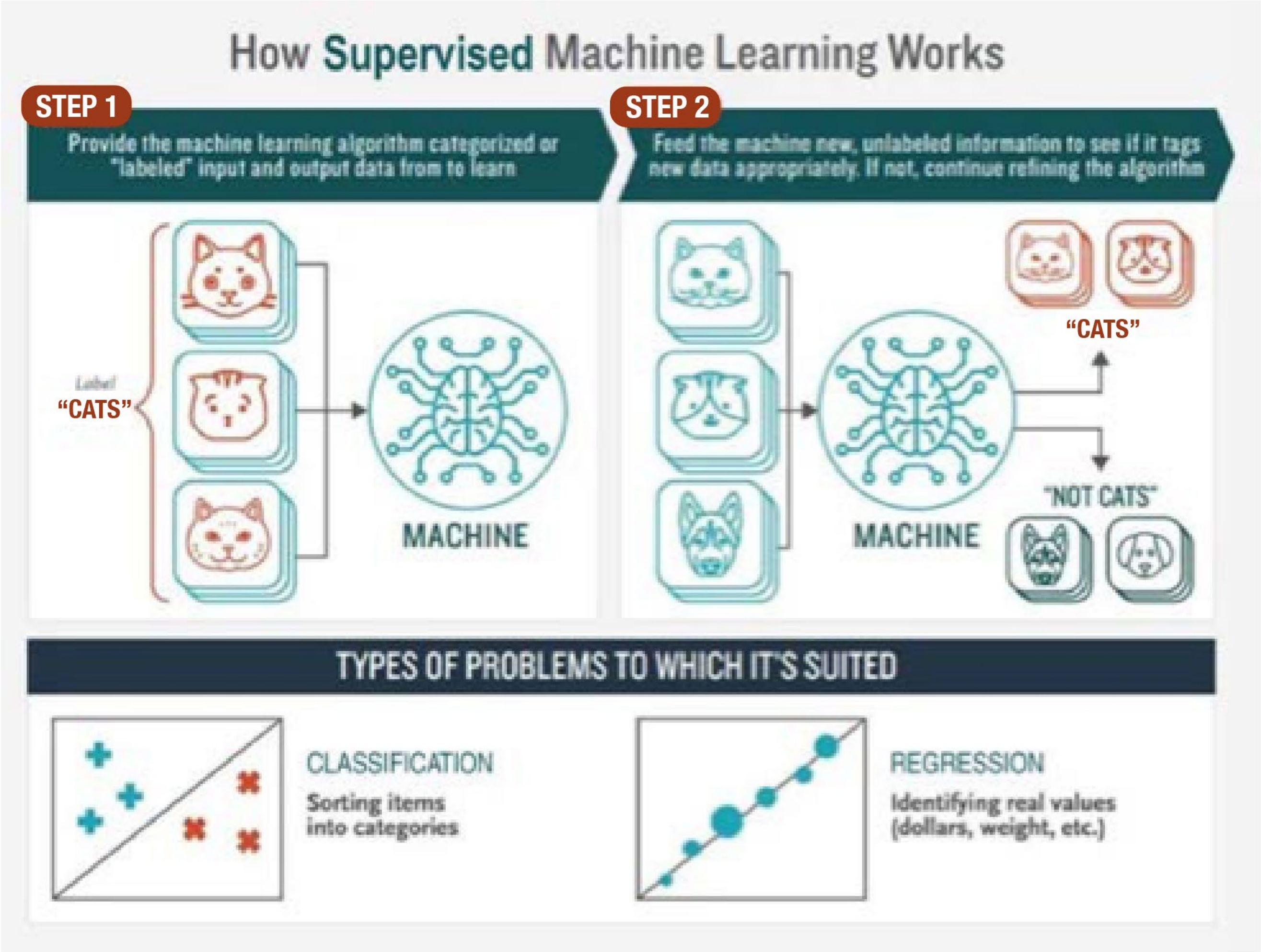
<u>Supervision here</u>: *LACK of such can elicit patterns NOT typically within human intuition*

https://www.gettyimages.ca/detail/news-photo/republican-presidential-hopeful-and-texas-governor-george-news-photo/51543471?adppopup=true

ML Essentials: supervised v. semi-supervised v. unsupervised learning

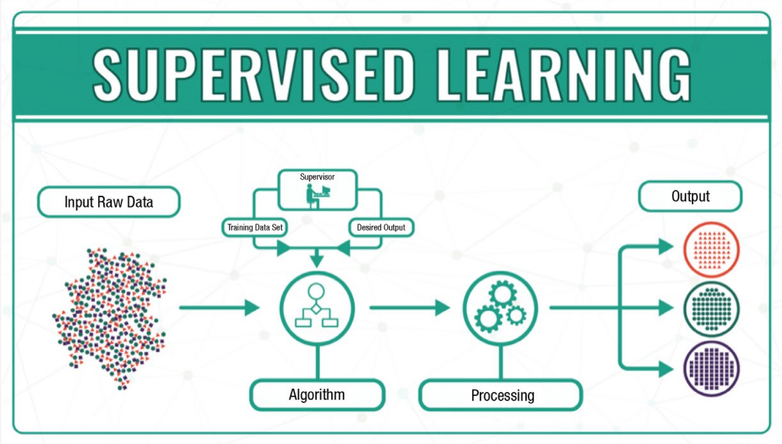
- From Booz Allen Team for CKD




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
Label
“CATS”
MACHINE
“CATS”
"NOT CATS"
MACHINE
TYPES OF PROBLEMS TO WHICH IT'S SUITED
CLASSIFICATION
Sorting items into categories
REGRESSION
Identifying real values (dollars, weight, etc.)

ML Essentials: supervised v. semi-supervised v. unsupervised learning

- From non-NIH-funded team



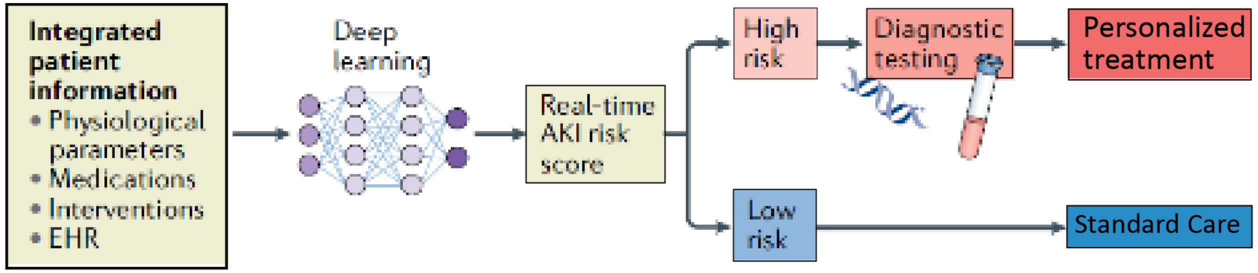
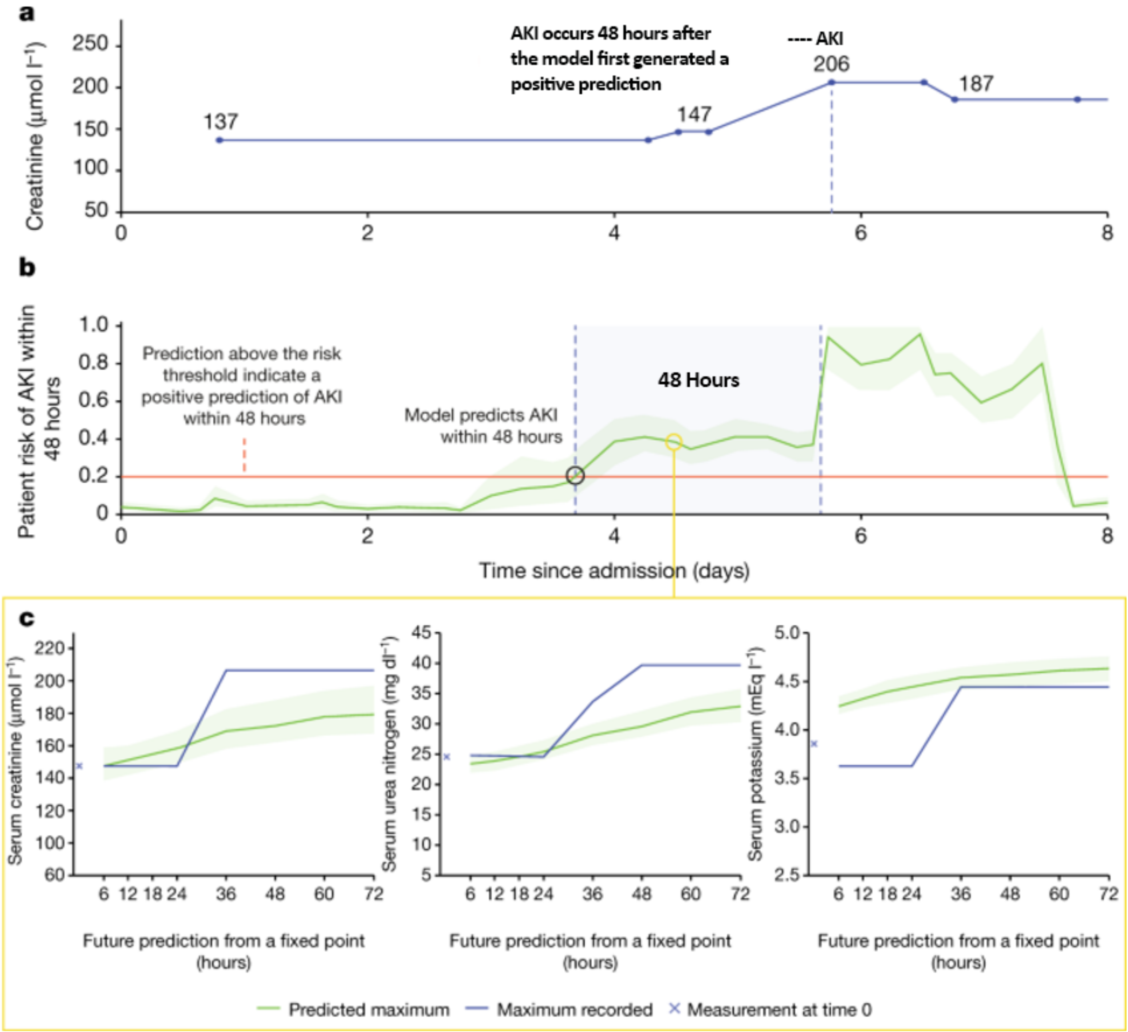


Fig. 1 | **Implementation of deep learning algorithms to identify patients at high risk of AKI.** Deep learning algorithms developed to support clinical decisions in real time should be based on integrated patient information, including electronic health records (EHRs) with detailed medical history (including ongoing problems and procedures), physiological parameters (such as vital signs and laboratory results) and medication details. Acute kidney injury (AKI) risk scores derived from such an algorithm would stratify patients and inform clinical decisions, including the use of additional diagnostics to enable personalized treatment.

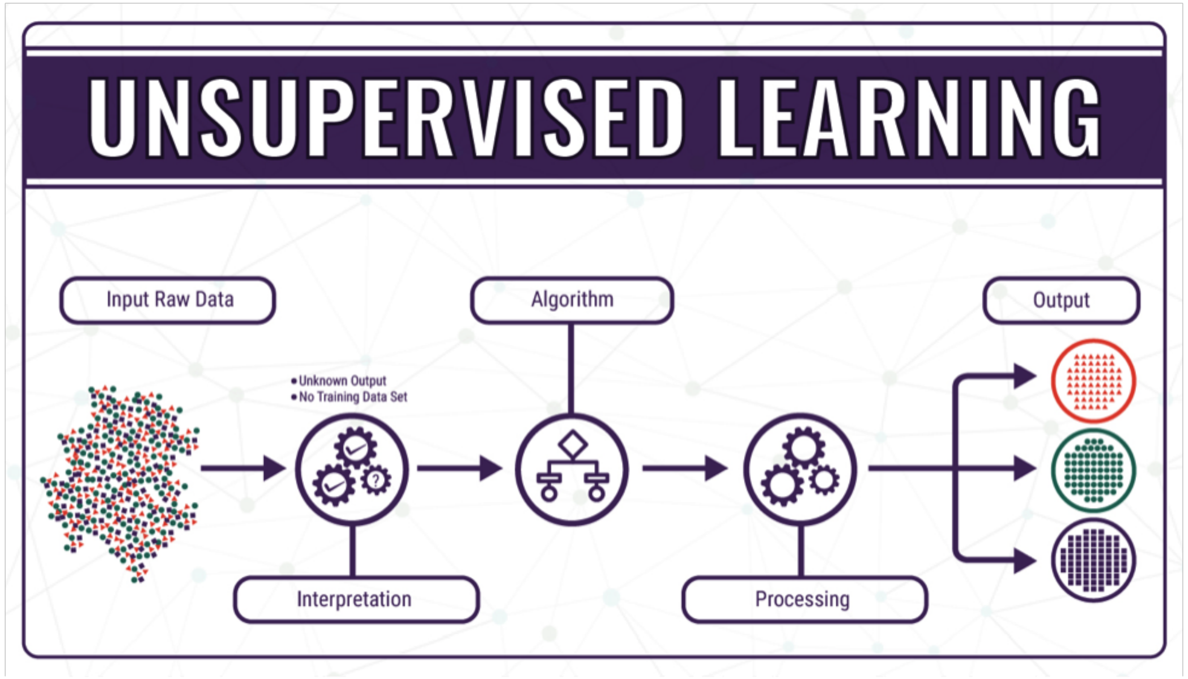


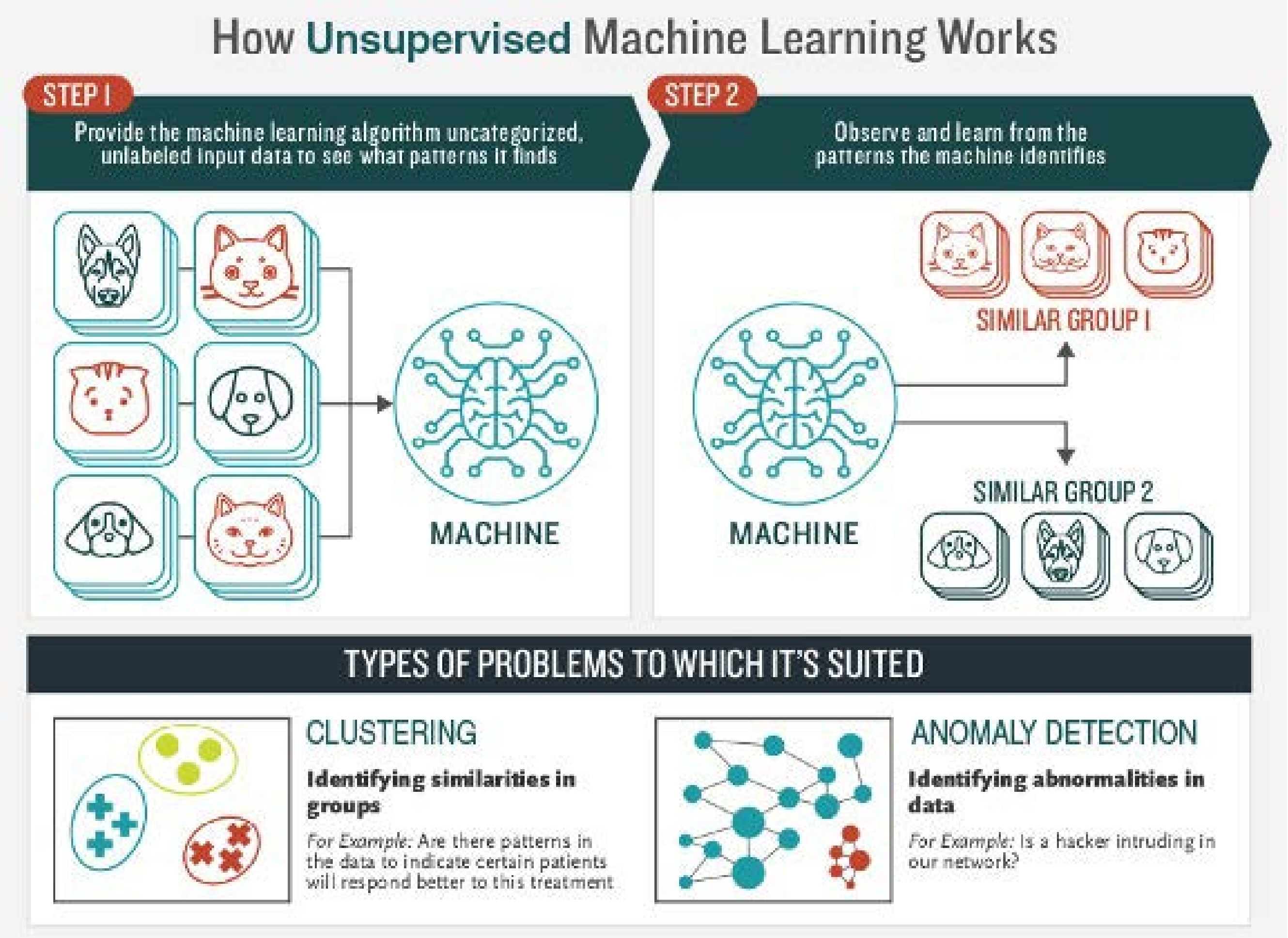
Figures 1 from editorial on and paper of DeepMind's AKI approach in Tomašev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).

"We make use of several open-source libraries to conduct our experiments: the machine learning framework TensorFlow (https://github.com/tensorflow/tensorflow) along with the TensorFlow library Sonnet (https://github.com/deepmind/sonnet)"

ML Essentials: supervised v. semi-supervised v. unsupervised learning

- From Booz Allen Team for CKD





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

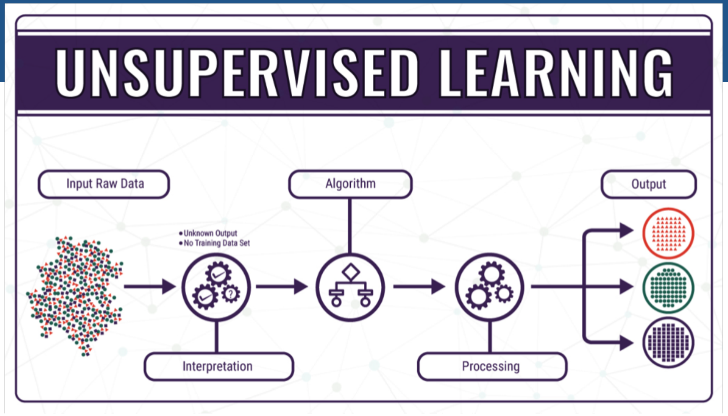
For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment

ANOMALY DETECTION

Identifying abnormalities in data

For Example: Is a hacker intruding in our network?

ML Essentials: supervised v. semi-supervised v. unsupervised learning



- From NIDDK-funded team →
- From other NIH-funded team ↓
  - Mammograms
  - Role of density
  - Blend: un+sup

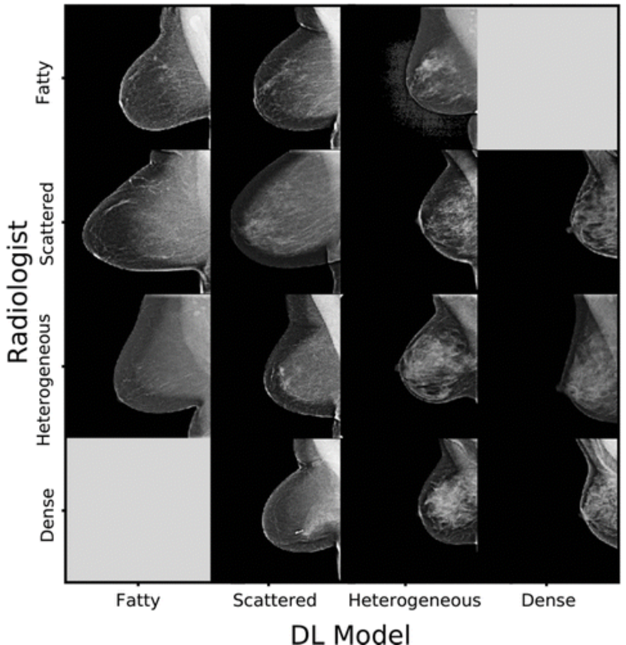
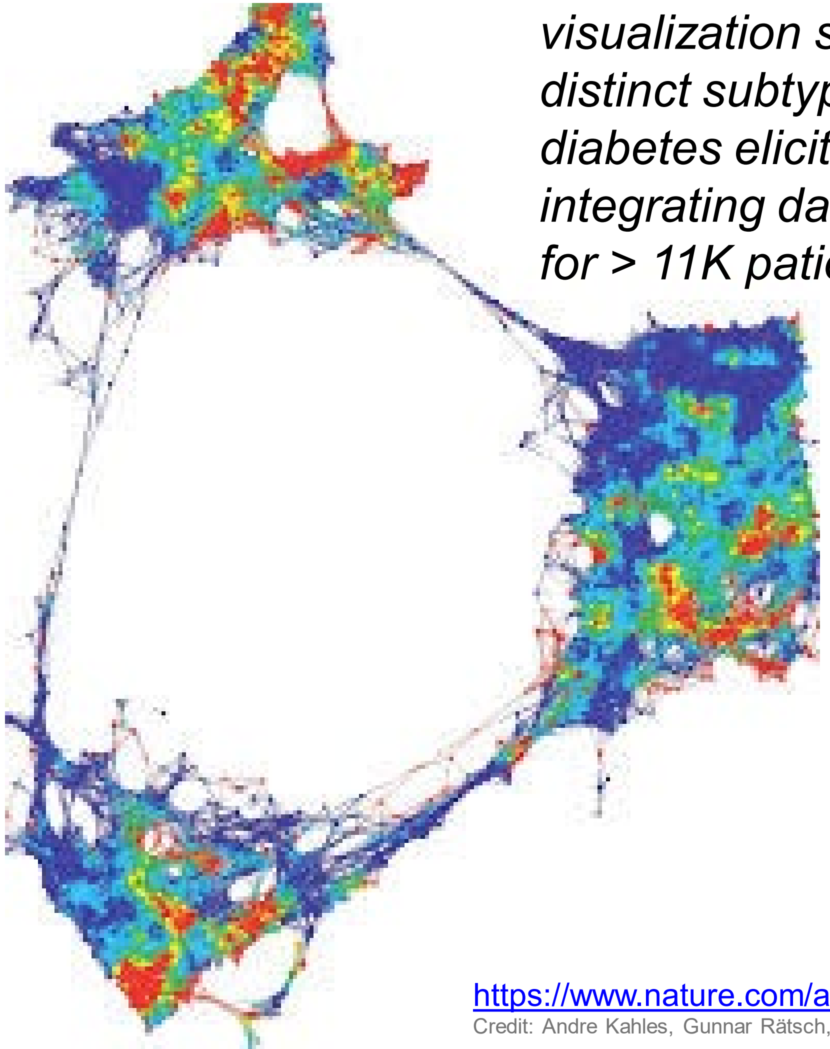


Figure 1d: Test set assessment. Comparison of the original interpreting radiologist assessment with the deep learning (DL) model assessment for **(a)** binary and **(c)** four-way mammographic breast density classification. **(b, d)** Corresponding examples of mammograms with concordant and discordant assessments by the radiologist and with the DL model.

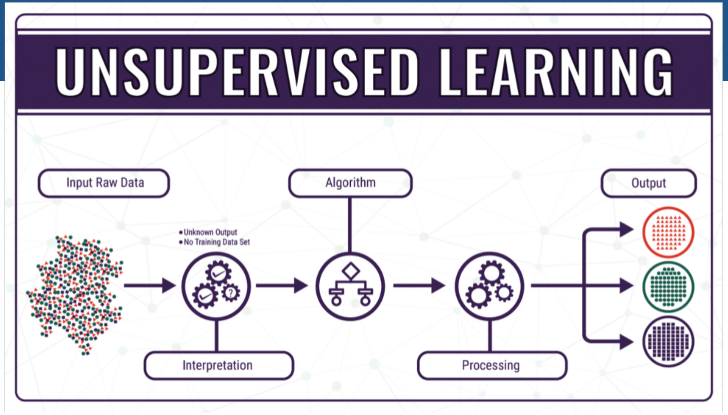


From NIDDK report: *network visualization showing 3 distinct subtypes of Type 2 diabetes elicited after integrating data from EHRs for > 11K patients*

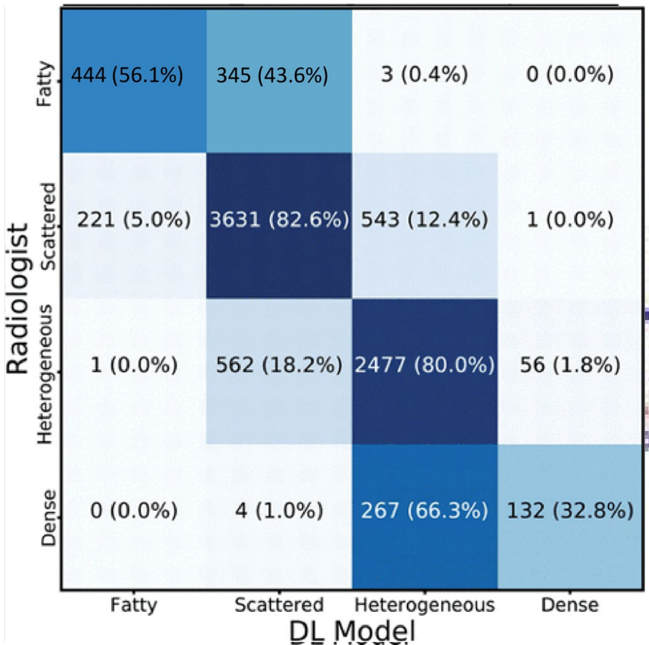
https://www.nature.com/articles/d42473-019-00035-5

Credit: Andre Kahles, Gunnar Rätsch, Chris Sander

ML Essentials: supervised v. semi-supervised v. unsupervised learning

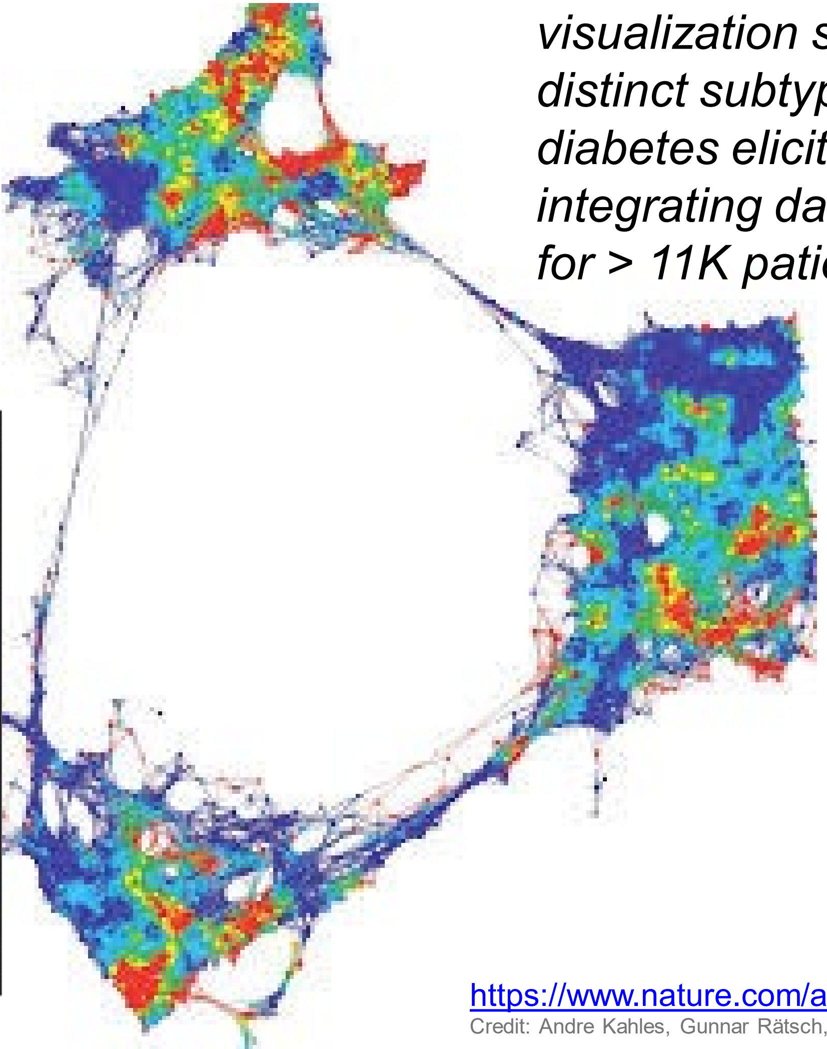


- From NIDDK-funded team →
- From other NIH-funded team ↓
  - Mammograms
  - Role of density
  - Blend: un+sup



Radiologist \ DL Model	Fatty	Scattered	Heterogeneous	Dense
Fatty	444 (56.1%)	345 (43.6%)	3 (0.4%)	0 (0.0%)
Scattered	221 (5.0%)	3631 (82.6%)	543 (12.4%)	1 (0.0%)
Heterogeneous	1 (0.0%)	562 (18.2%)	2477 (80.0%)	56 (1.8%)
Dense	0 (0.0%)	4 (1.0%)	267 (66.3%)	132 (32.8%)

Figure 1d: Test set assessment. Comparison of the original interpreting radiologist assessment with the deep learning (DL) model assessment for **(a)** binary and **(c)** four-way mammographic breast density classification. **(b, d)** Corresponding examples of mammograms with concordant and discordant assessments by the radiologist and with the DL model.



From NIDDK report: *network visualization showing 3 distinct subtypes of Type 2 diabetes elicited after integrating data from EHRs for > 11K patients*

https://www.nature.com/articles/d42473-019-00035-5

Credit: Andre Kahles, Gunnar Rätsch, Chris Sander

Machine Learning Computational Strategies

- We now engage participants to check our mutual understanding

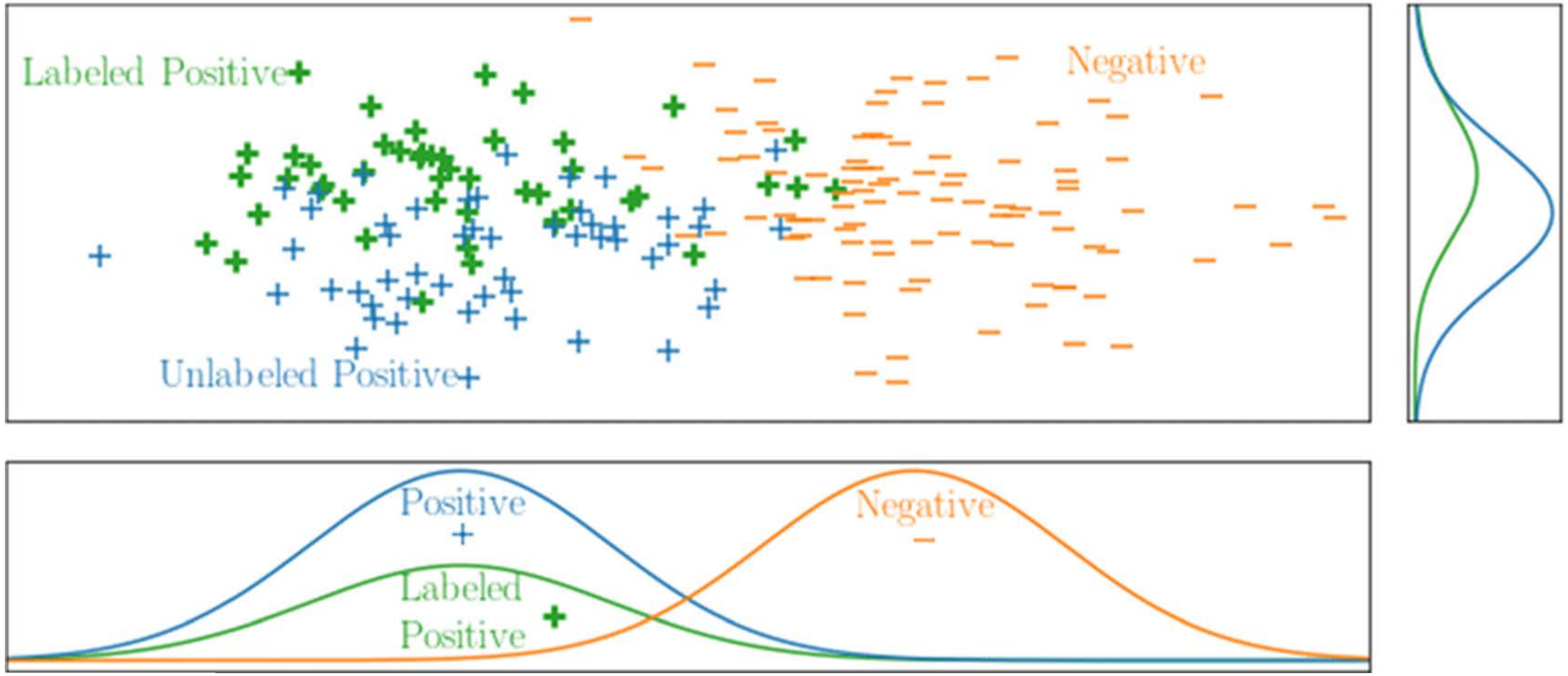
ML Essentials: supervised v. semi-supervised v. unsupervised learning

- **Semi-supervised**: a mix between supervised and unsupervised learning

  Classic **examples**

  – Positive & unlabeled
    - Only **green** instances labeled
    - Algorithm adapts iteratively

  – Role of ‘learning’ objective
    - Entropy v. other criteria



Bekker, J., Davis, J. Learning from positive and unlabeled data: a survey. *Mach Learn* **109**, 719–760 (2020). https://doi.org/10.1007/s10994-020-05877-5

- Survey of DL use cases

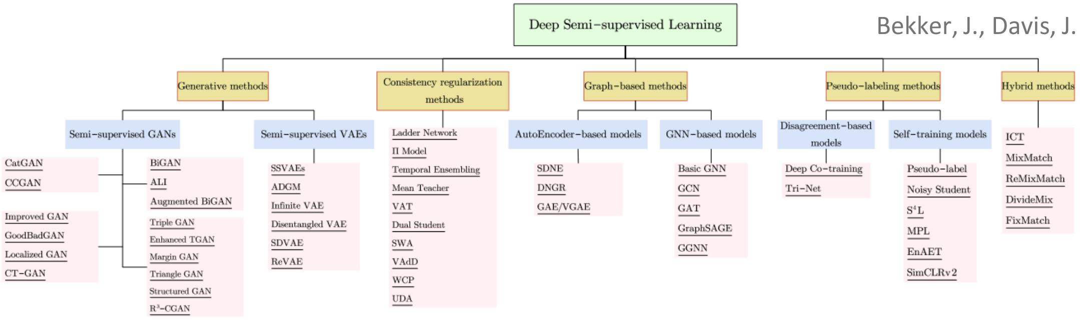


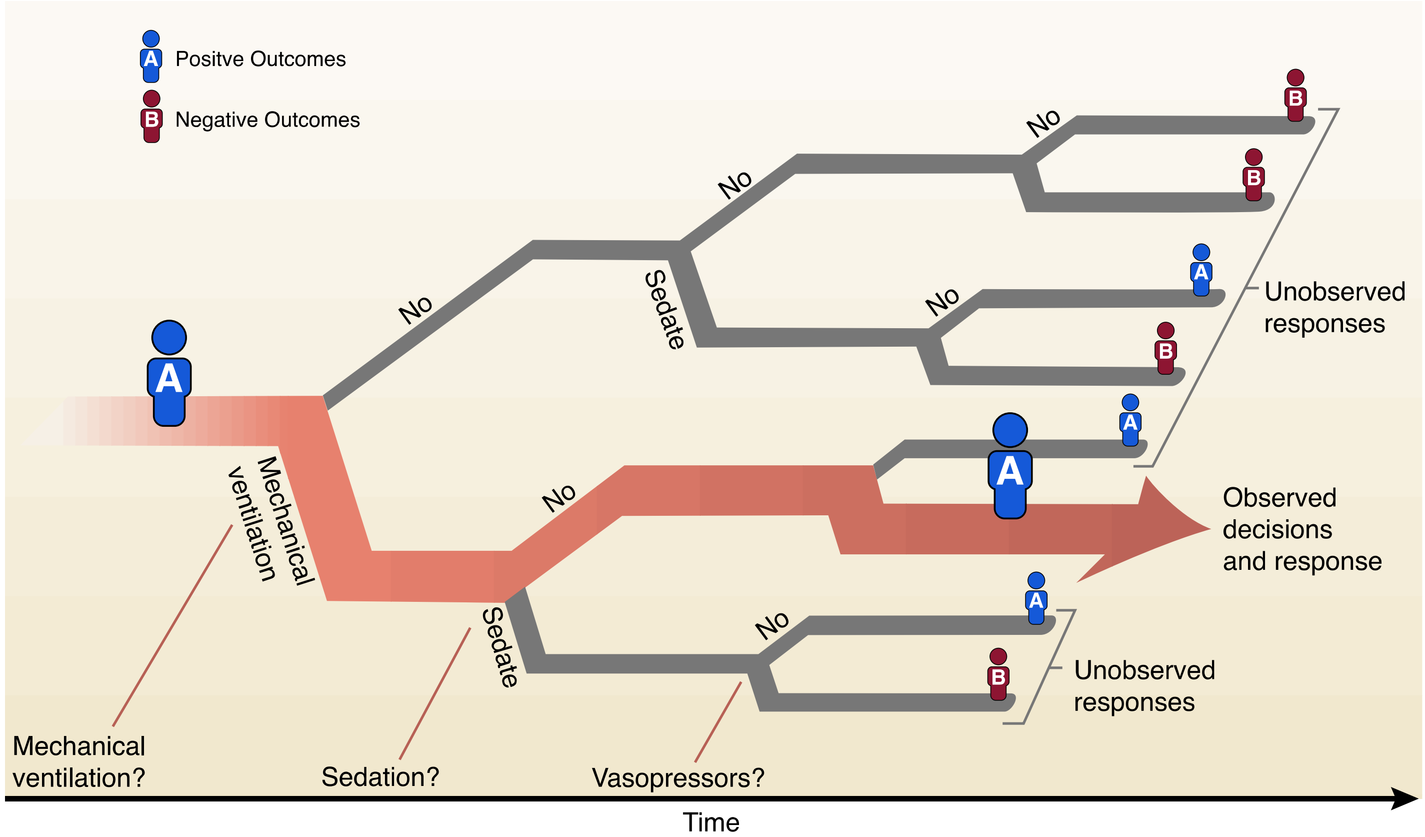
Fig. 1. The taxonomy of major deep semi-supervised learning methods based on loss function and model design.

ML Essentials: supervised v. semi-supervised v. unsupervised learning

- From Semi-supervised to ***Reinforcement Learning*** (frequently uses *Q-learning)*
  - *Particularly useful over time*
  - *suited to decision sequences*
  - *Caveats in health settings,*
    - *Nature editorial poses challenges*
    - *Example at right: intensive care*

Reinforcement learning is a type of machine learning that focuses on training AI agents to make a sequence of decisions to maximize a cumulative reward. It's used in gaming, robotics, and autonomous systems.

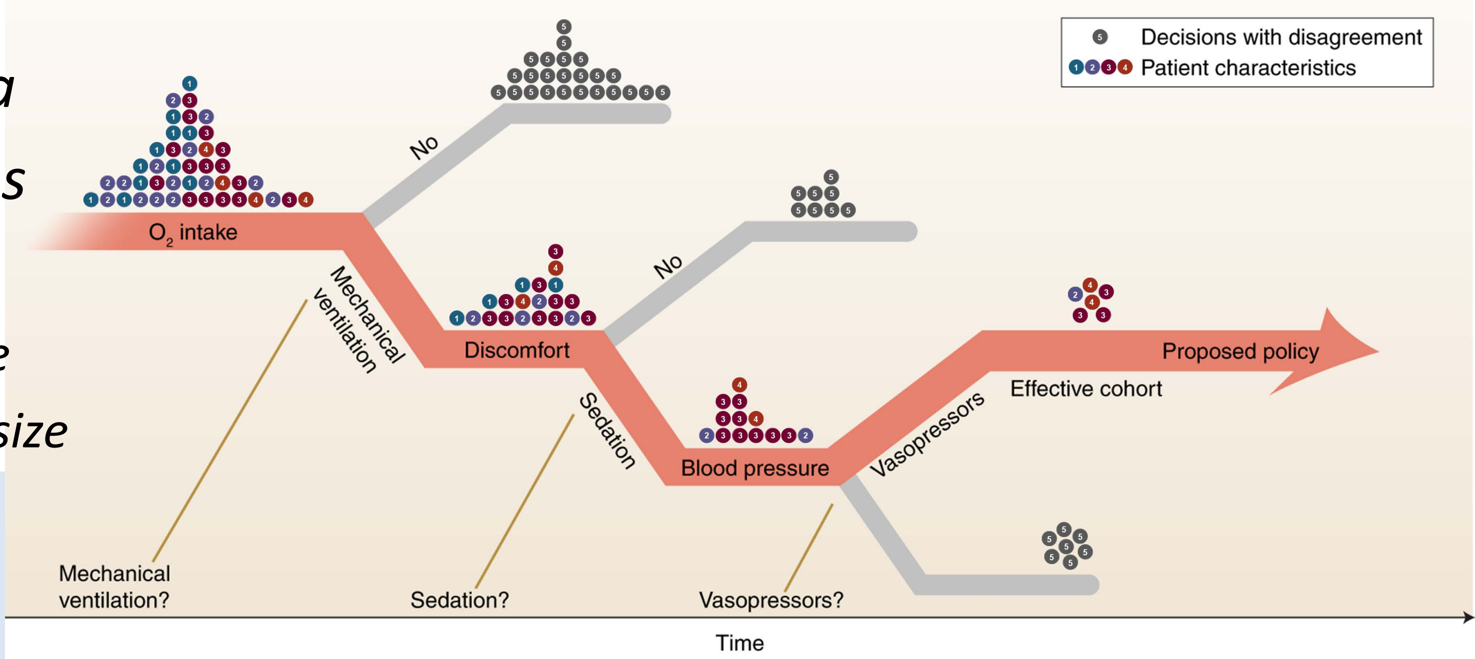
To perform sequential decision making, such as for sepsis management, treatment-effect estimation must be solved at a grand scale—every possible combination of interventions could be considered to find an optimal treatment policy. The diagram shows the scale of such a problem with only three distinct decisions. **Blue** and **red** people denote positive and negative outcomes, respectively.



ML Essentials: supervised v. semi-supervised v. unsupervised learning

- From Semi-supervised to ***Reinforcement Learning*** (frequently uses *Q-learning)*
  - *Particularly reliant on BIG data*
  - *Need cases along all sequences*
  - *Caveats in health settings,*
    - *Nature editorial shows challenge*
    - *Figure @right: effective sample size*

Reinforcement learning is a type of machine learning that focuses on training AI agents to make a sequence of decisions to maximize a cumulative reward. It's used in gaming, robotics, and autonomous systems.

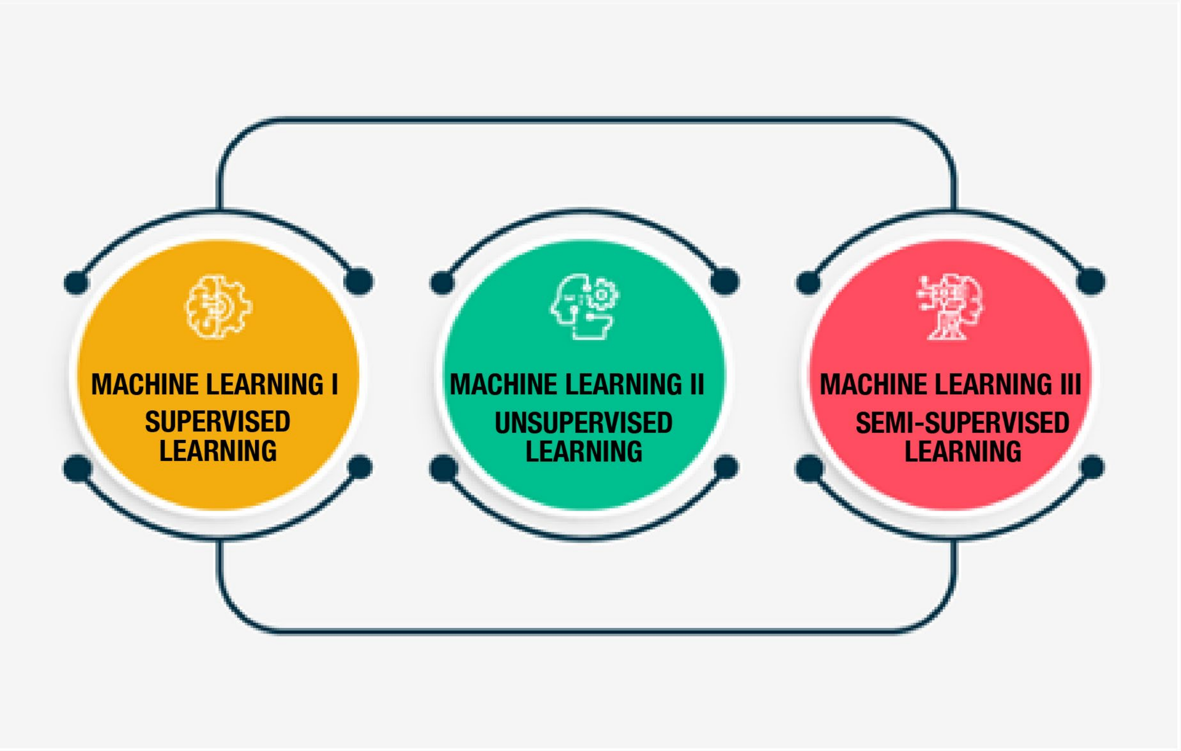


Each dot represents a single patient at each stage of treatment, and its color (gradation from **blue**↔**red**) indicates the patient's characteristics. The more decisions that are performed in sequence, the likelier it is that a new policy disagrees with the one that was learned from. **Gray** decision points indicate disagreement. Use of only samples for which the old policy agrees with the new results in a small effective sample size and a biased cohort, as illustrated by the difference in color distribution in the original and final cohort.

Gottesman, O., Johansson, F., Komorowski, M. *et al.* Guidelines for reinforcement learning in healthcare. *Nat Med* **25**, 16–18 (2019). https://doi.org/10.1038/s41591-018-0310-5

Machine Learning Essentials: concept check

- We now engage participants to check our mutual understanding:

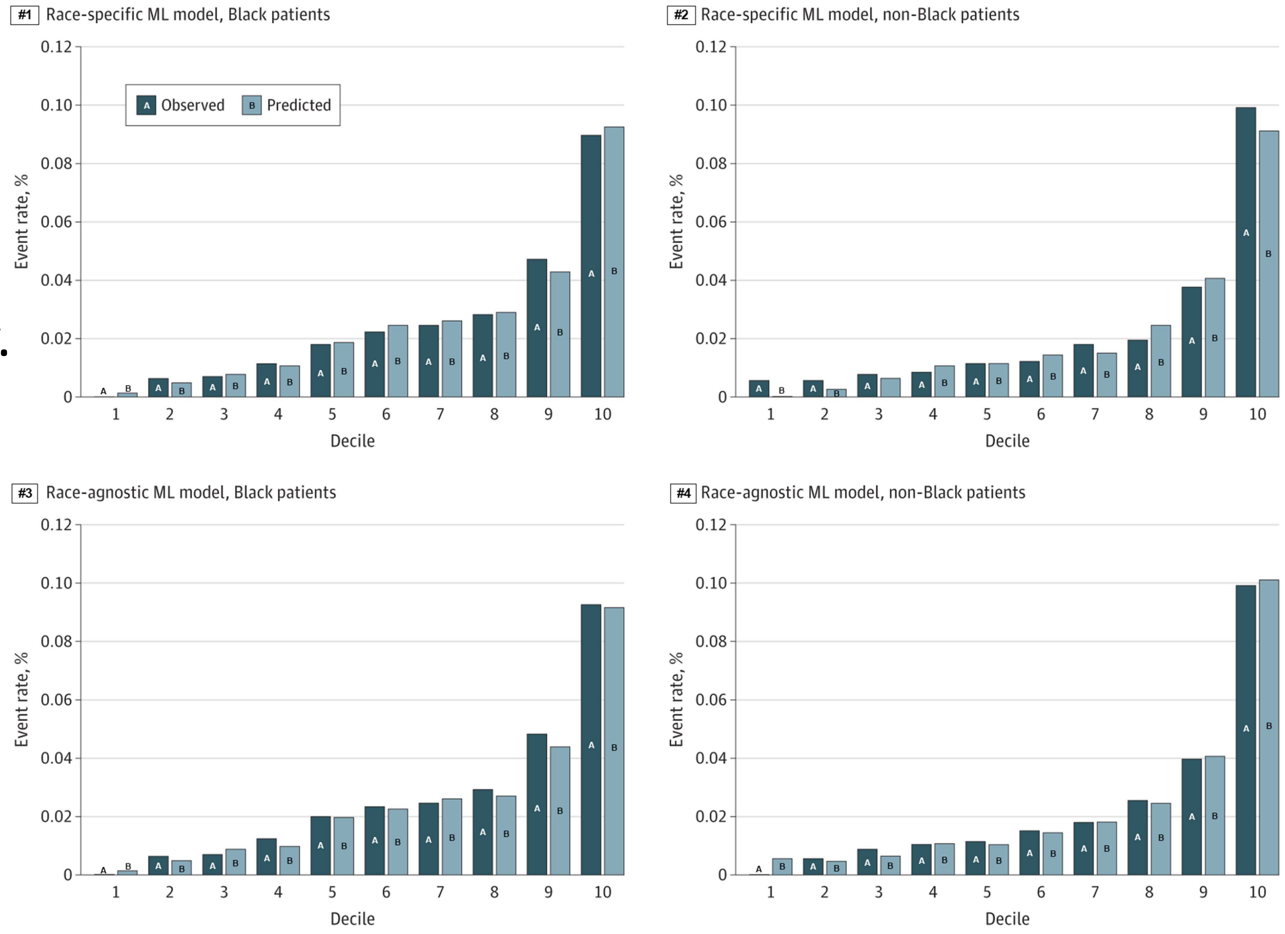


[recall sli.do questions re: supervised v. unsupervised v. semisupervised]

Machine Learning Computational Strategies

- We now provide detailed explanations and use cases for ML strategies, which can improve upon traditional/modern stats / epi data methods.
  - **Example:** See differences in race-specific v. race-agnostic for machine learning predicted in-hospital mortality…
  - either improved on logistic regression

- Detailed Examples of ML computational strategies used in healthcare disparities research (the list of examples to follow is not exhaustive)


#1 Race-specific ML model, Black patients

#2 Race-specific ML model, non-Black patients

#3 Race-agnostic ML model, Black patients

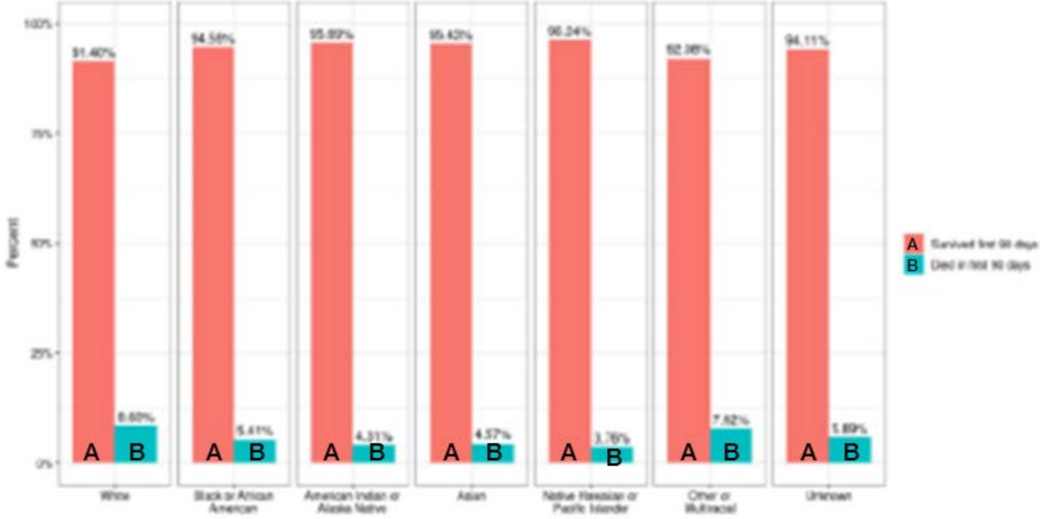
#4 Race-agnostic ML model, non-Black patients

https://jamanetwork.com/journals/jamacardiology/fullarticle/2793728

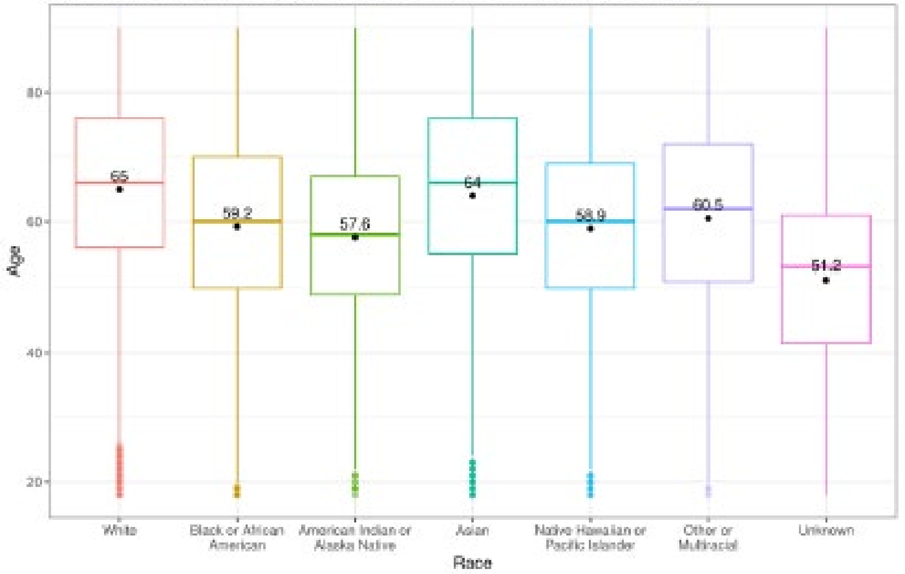
Machine Learning Computational Strategies

1. Predictive Modeling for Patient Outcomes:

- a. Strategy: Using machine learning algorithms to predict patient outcomes.
- b. Application: Identifying high-risk populations for specific diseases [**examples**].
- c. Python Libraries: Scikit-learn, TensorFlow, PyTorch.



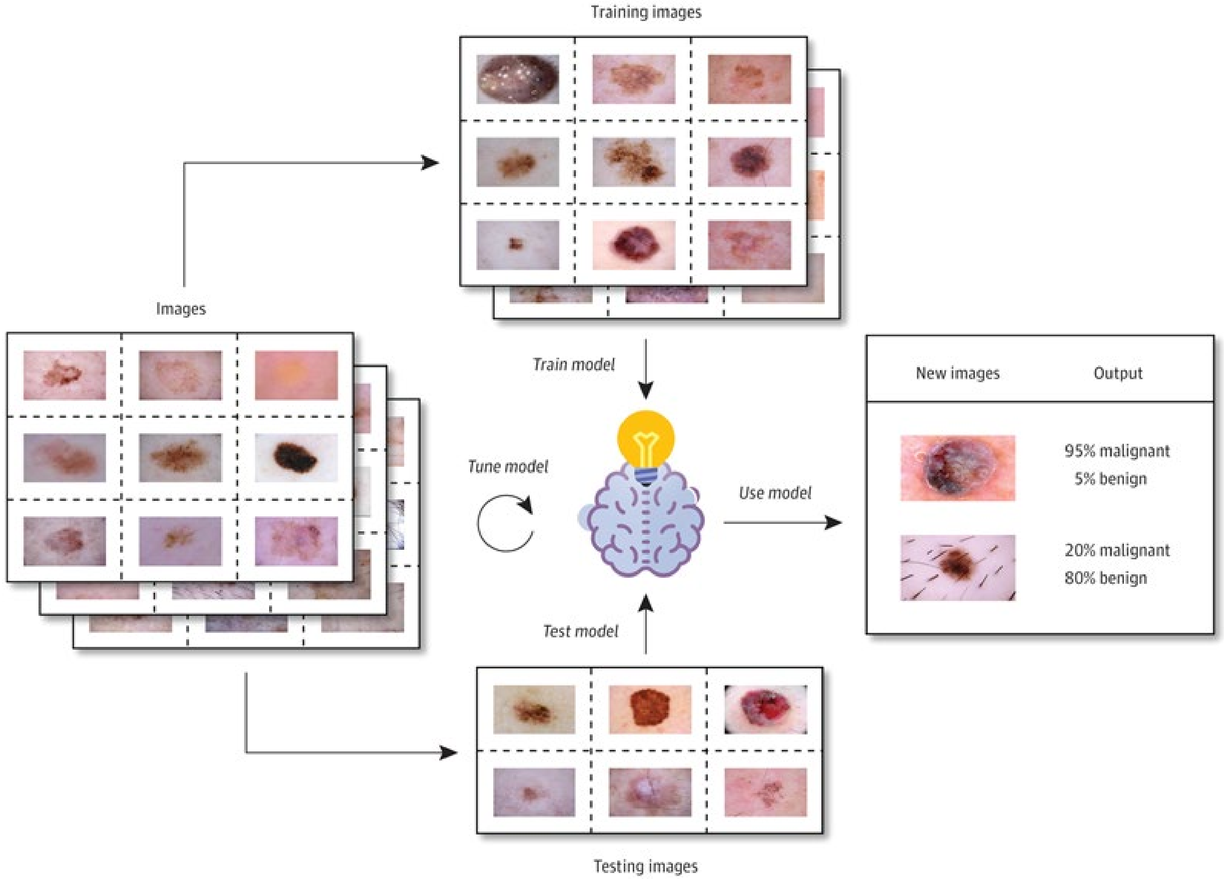
Box plot showing distribution of age at first dialysis treatment for each race



Machine Learning Computational Strategies

2. Image Analysis for Diagnostics:

- a. Strategy: Applying computer vision and deep learning.
- b. Application: Improving diagnostic accuracy from medical images [breast density **example above**; melanoma w/o regard to skin color ***counter*-example** @right]
- c. Python Libraries: TensorFlow, PyTorch, OpenCV.



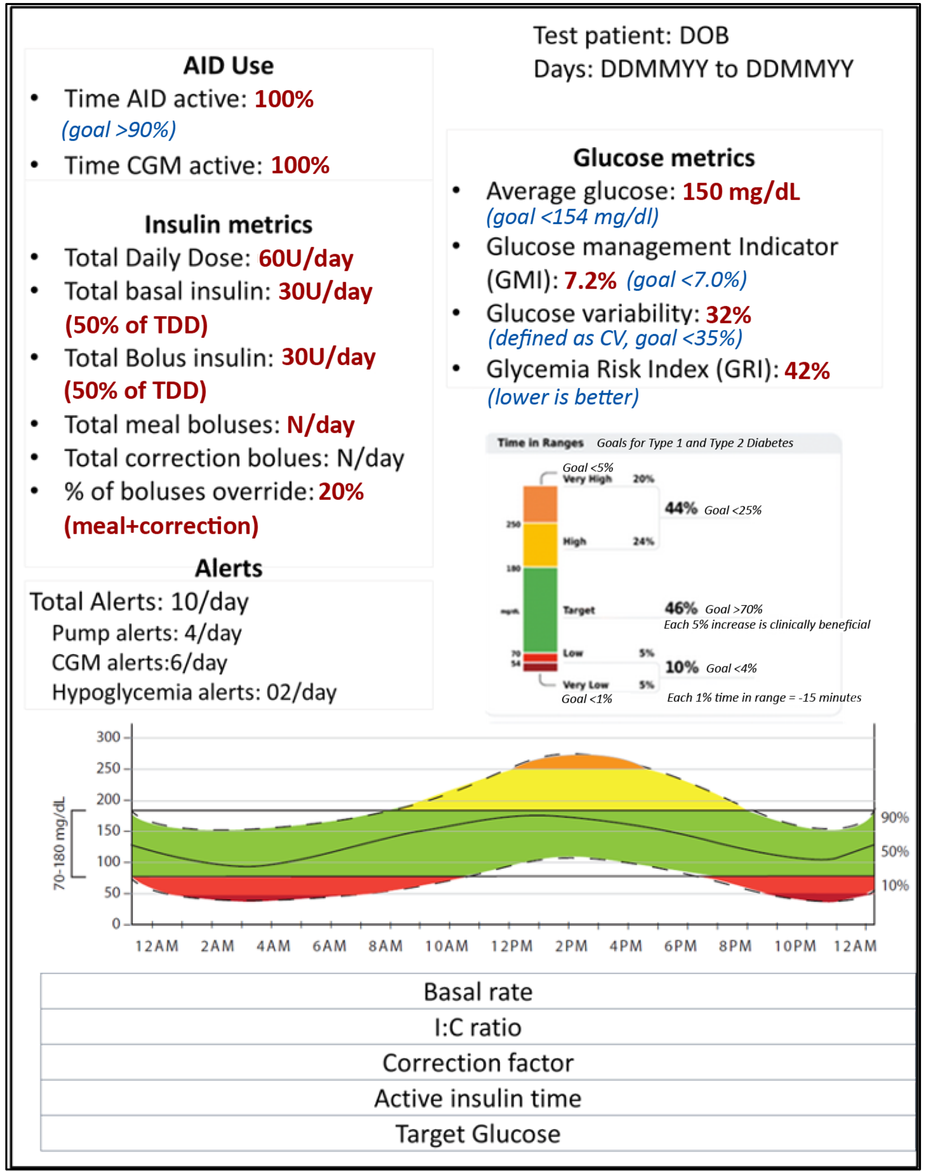
Images are collected of pigmented lesions and split into a larger training image set and a smaller testing image set. The machine learning algorithm (center) uses the training images to learn how to correctly categorize pigmented lesions based on their visual features. The model is then tested with the testing images set to determine model accuracy. The algorithm model is fine-tuned with more training and testing images. Once the machine learning algorithm is developed, it can be used on new images. The output gives an estimate of the likelihood of a given result.

https://jamanetwork.com/journals/jamadermatology/fullarticle/2688587

Machine Learning Computational Strategies

4. Remote Patient Monitoring:

- a. Strategy: Using AI to analyze data from wearable devices.
- b. Application: Monitoring patient health in real-time **examples** [e.g., continuous glucose monitoring, or CGM for Active Insulin Dosing, AID]
- c. Python Libraries: TensorFlow, scikit-learn.



Test patient: DOB
Days: DDMMYY to DDMMYY

AID Use

- Time AID active: **100%** *(goal >90%)*
- Time CGM active: **100%**

Insulin metrics

- Total Daily Dose: **60U/day**
- Total basal insulin: **30U/day (50% of TDD)**
- Total Bolus insulin: **30U/day (50% of TDD)**
- Total meal boluses: **N/day**
- Total correction bolues: N/day
- % of boluses override: **20% (meal+correction)**

Alerts

Total Alerts: 10/day
Pump alerts: 4/day
CGM alerts:6/day
Hypoglycemia alerts: 02/day

Glucose metrics

- Average glucose: **150 mg/dL** *(goal <154 mg/dl)*
- Glucose management Indicator (GMI): **7.2%** *(goal <7.0%)*
- Glucose variability: **32%** *(defined as CV, goal <35%)*
- Glycemia Risk Index (GRI): **42%** *(lower is better)*

Basal rate
I:C ratio
Correction factor
Active insulin time
Target Glucose

Machine Learning Computational Strategies

5. Population Health Management:

- a. Strategy: Employing machine learning algorithms for population-level health data.
- b. Application: Identifying disparities in health outcomes.
- c. Python Libraries: Scikit-learn, TensorFlow, PyTorch.

A counter-example: ***mentioned in last Think-a-thon***

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**

Example: **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%

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

Machine Learning Computational Strategies

6. Social Determinants of Health (SDOH) Analysis:

- a. Strategy: Integrating AI to analyze social, economic, and environmental factors.
- b. Application: Understanding the impact of social determinants on healthcare disparities. **example**
- c. Python Libraries: Scikit-learn, pandas, NumPy.

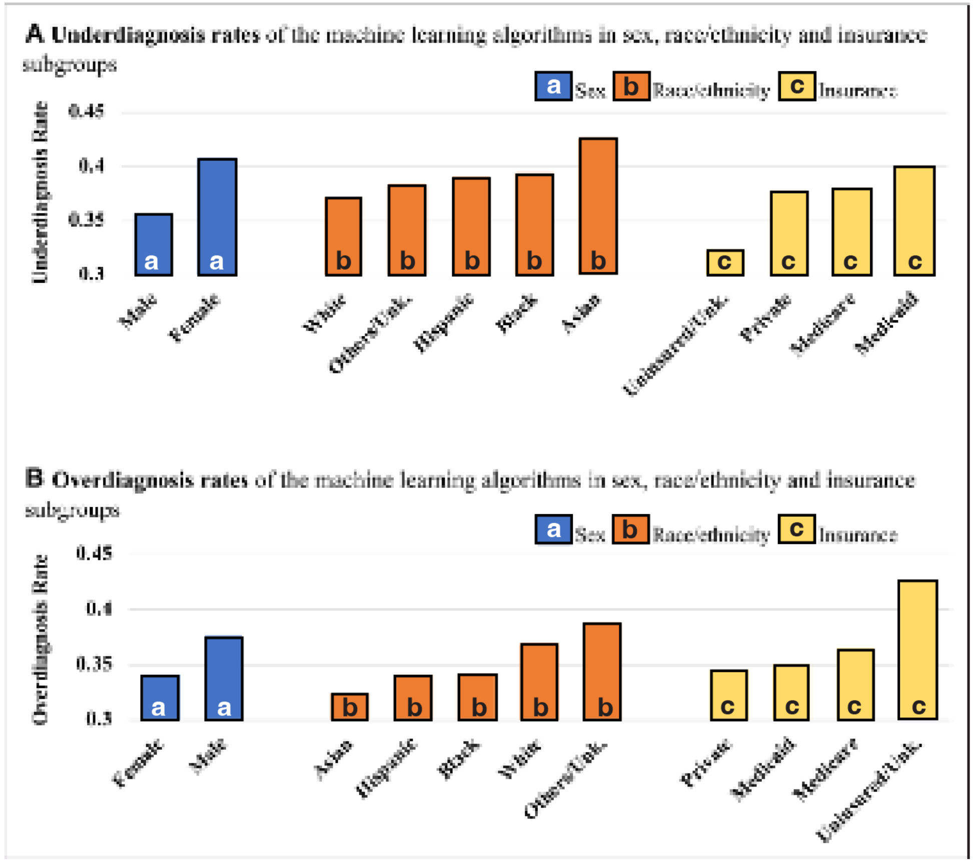


Figure. **Underdiagnosis (false negative rate) and overdiagnosis (false positive rate) rates in each sex, ethnoracial, and insurance subgroup, when using random forest classifier to predict the composite heart failure outcome.**
The model achieves the highest performance and fairness scores. Unk indicates unknown.

b. Application example by Luo's team: Social Deprivation Index (SDI) & Area Deprivation Index (ADI) at both state and national levels) can *somewhat* mitigate the Figure-noted heart failure risk disparities

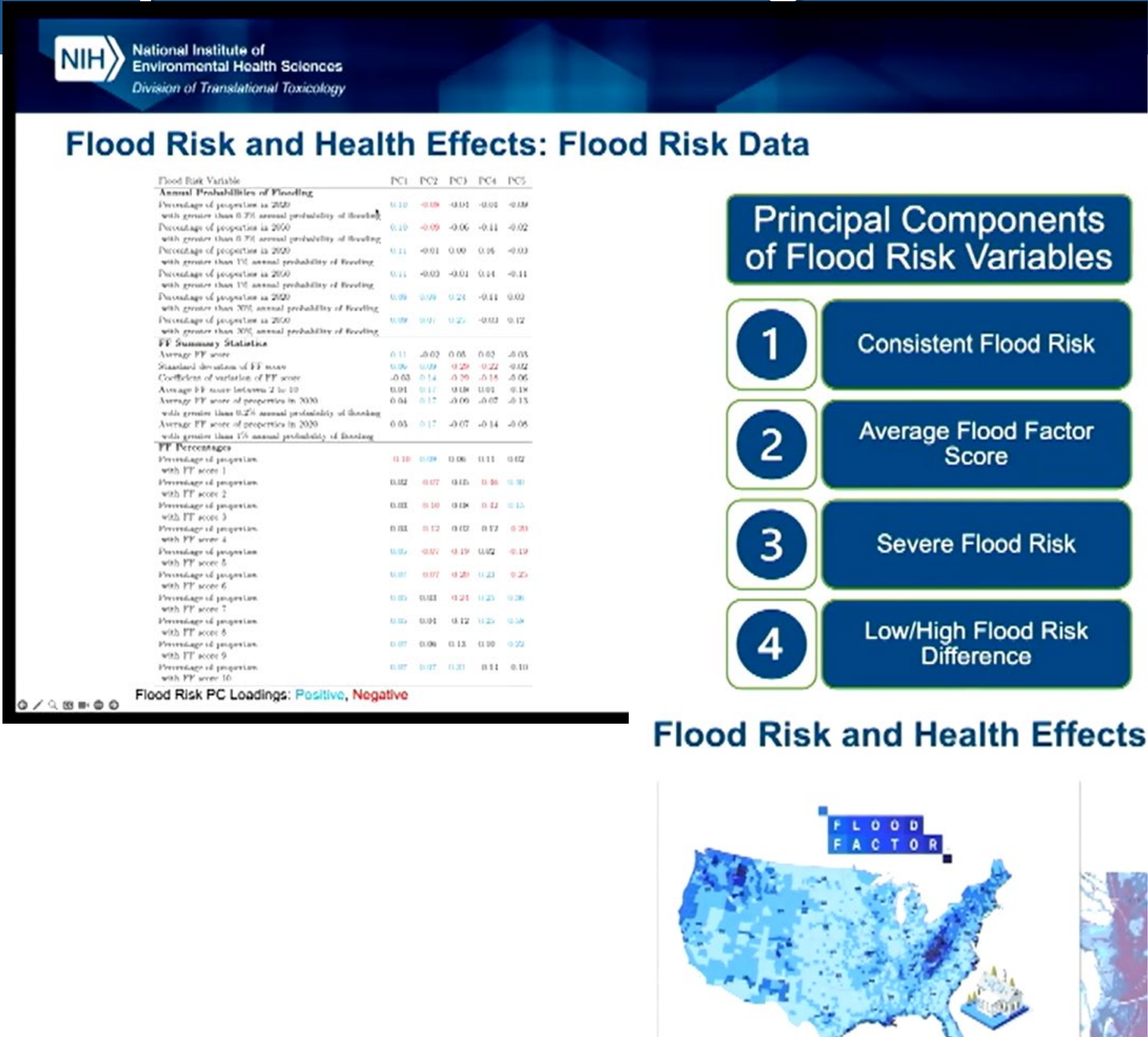
https://news.feinberg.northwestern.edu/2022/12/15/investigating-disparities-in-machine-learning-algorithms/

Machine Learning Computational Strategies

6. Social Determinants of Health (SDOH) Analysis:

- b. Application: Understanding the impact of social determinants on healthcare disparities… can be *less often considered sources for SDOH,* if the use case points to a need
- b. **Example**: stark climate-change related vulnerabilities, like flooding

b. Application example by NIEHS/NIMHD PI Messier's SET group: used First Street Foundation's Flood measures panel at granular area levels -- can *somewhat* mitigate the noted__ risk disparities



NIH National Institute of Environmental Health Sciences
Division of Translational Toxicology

Flood Risk and Health Effects: Flood Risk Data

Flood Risk Variable	PC1	PC2	PC3	PC4	PC5
Annual Probabilities of Flooding					
Percentage of properties in 2020 with greater than 0.2% annual probability of flooding	0.10	-0.09	-0.01	-0.01	-0.10
Percentage of properties in 2050 with greater than 0.2% annual probability of flooding	0.10	-0.09	-0.06	-0.11	-0.02
Percentage of properties in 2020 with greater than 1% annual probability of flooding	0.11	-0.01	0.09	0.16	-0.03
Percentage of properties in 2050 with greater than 1% annual probability of flooding	0.11	-0.03	-0.01	0.14	-0.11
Percentage of properties in 2020 with greater than 20% annual probability of flooding	0.09	0.09	0.24	-0.11	0.02
Percentage of properties in 2050 with greater than 20% annual probability of flooding	0.09	0.07	0.23	-0.03	0.12
FF Summary Statistics					
Average FF score	0.11	-0.02	0.08	0.02	-0.03
Standard deviation of FF score	0.06	0.09	-0.29	-0.22	-0.02
Coefficient of variation of FF score	-0.03	0.14	-0.29	-0.15	-0.06
Average FF score between 2 to 10	0.01	0.17	0.08	0.01	-0.18
Average FF score of properties in 2020 with greater than 0.2% annual probability of flooding	0.04	0.17	-0.09	-0.07	-0.13
Average FF score of properties in 2020 with greater than 1% annual probability of flooding	0.03	0.17	-0.07	-0.14	-0.08
FF Percentages					
Percentage of properties with FF score 1	-0.10	0.09	0.06	0.11	0.02
Percentage of properties with FF score 2	0.02	-0.07	0.05	-0.16	0.30
Percentage of properties with FF score 3	0.03	-0.10	0.08	-0.12	0.13
Percentage of properties with FF score 4	0.03	-0.12	0.02	0.17	-0.20
Percentage of properties with FF score 5	0.05	-0.07	-0.19	0.02	-0.19
Percentage of properties with FF score 6	0.07	-0.07	-0.20	0.23	-0.25
Percentage of properties with FF score 7	0.05	0.03	-0.24	0.25	0.28
Percentage of properties with FF score 8	0.05	0.04	-0.12	0.25	0.58
Percentage of properties with FF score 9	0.07	0.06	0.13	0.10	0.22
Percentage of properties with FF score 10	0.07	0.07	0.30	0.11	0.10

Flood Risk PC Loadings: Positive, Negative

Principal Components of Flood Risk Variables

1 Consistent Flood Risk

2 Average Flood Factor Score

3 Severe Flood Risk

4 Low/High Flood Risk Difference

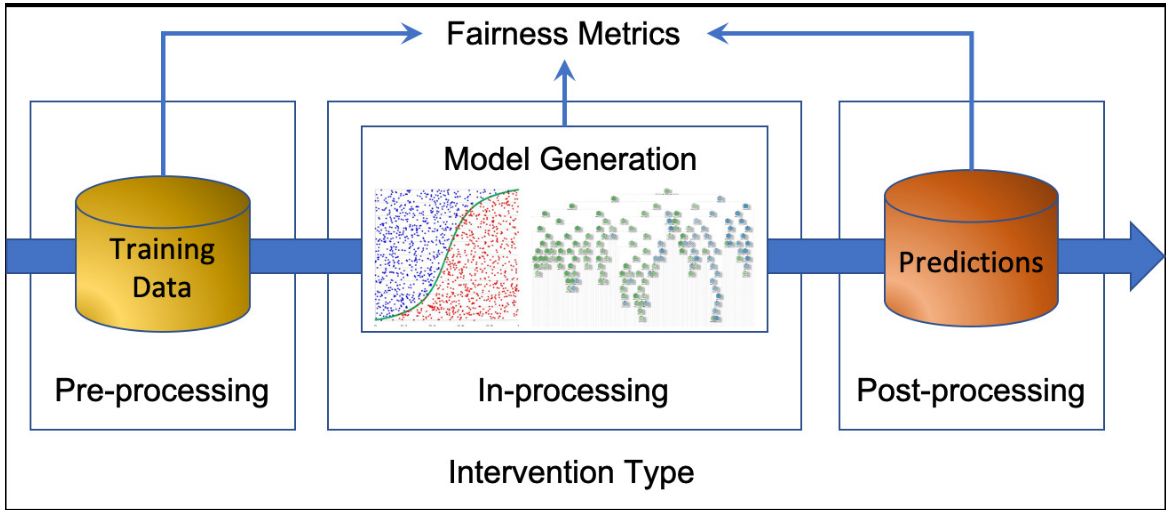
Flood Risk and Health Effects

https://videocast.nih.gov/watch=53935

Machine Learning Computational Strategies

7. Ethical AI for Bias Mitigation:

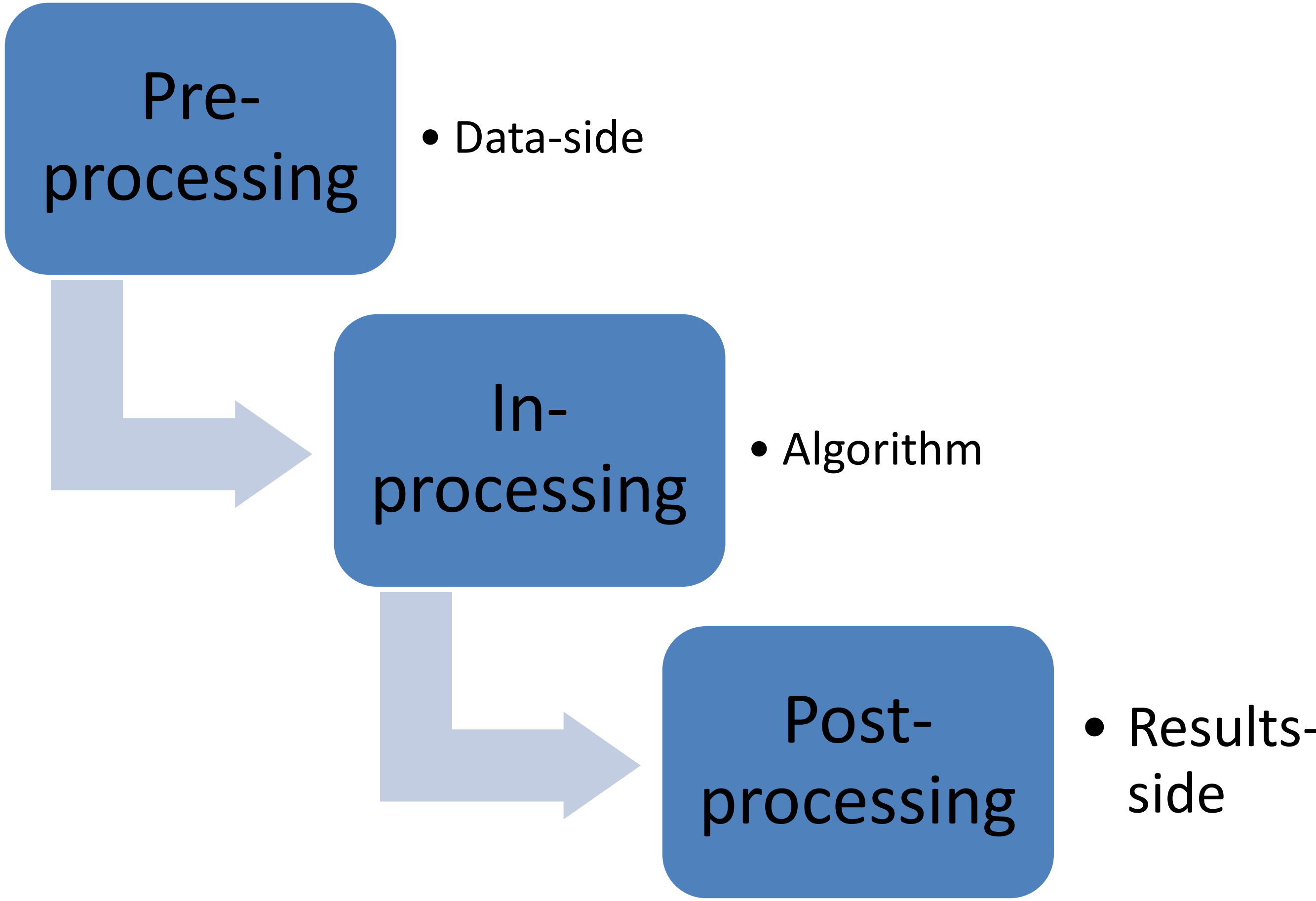
- a. Strategy: Implementing fairness-aware and explainable AI models.
- b. Application: Ensuring AI systems do not perpetuate biases.
- c. Python Libraries: AIF360, Fairness Indicators (Caton & Haas review), AI Fairness 360
  - NB: includes a scikit-learn compatible Application-Programmer Interface (API)!



Machine Learning Computational Strategies

7. Ethical AI for Bias Mitigation:

- b. Application: Ensuring AI systems do **not** perpetuate biases… may be *most tractable* by applying [Caton&Haas framework]
  - Pre-processing
  - *IN-processing*
  - Post-processing: helpful capacity to apply to *any* data science workflow



Pre-processing

- Data-side

In-processing

- Algorithm

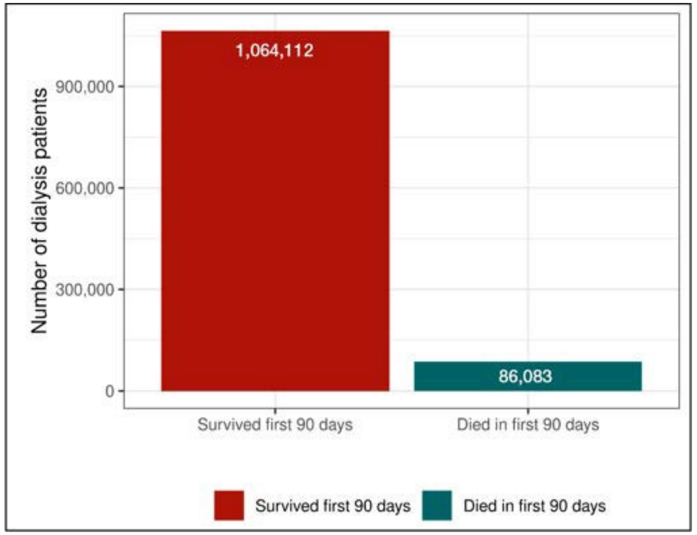
Post-processing

- Results-side

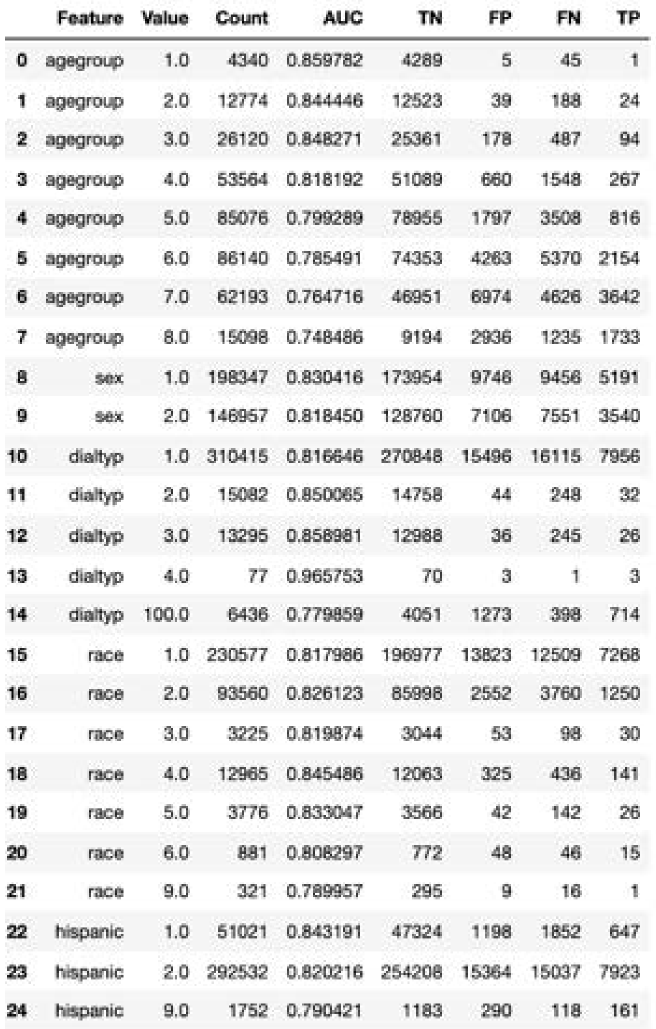
Machine Learning Computational Strategies

7. Ethical AI for Bias Mitigation:

- b. Application: **example** of applying Caton&Haas framework
  - **Post-processing**: helpful capacity to apply to *any* data science workflow



From prior ScHARE Think-a-thon slides (not covered):
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

Machine Learning Computational Strategies

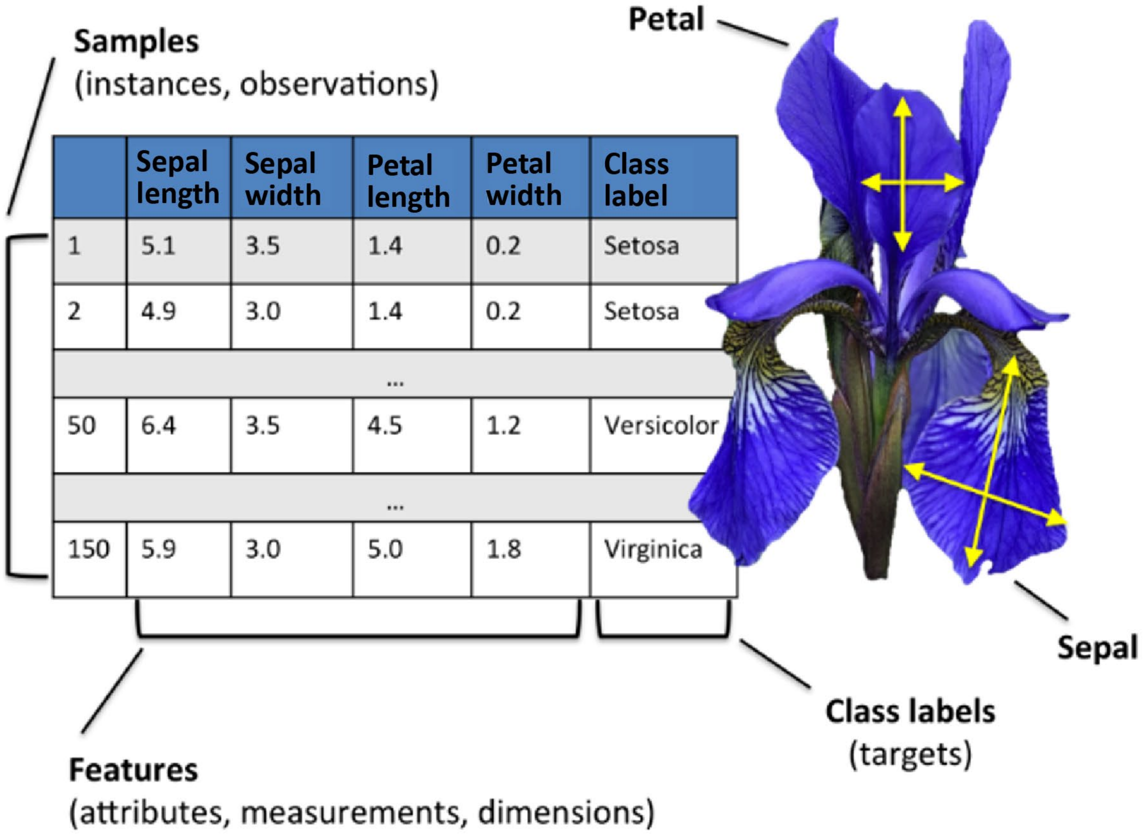
Concept check [slido]

Machine Learning Computational Strategies

Practical hands-on

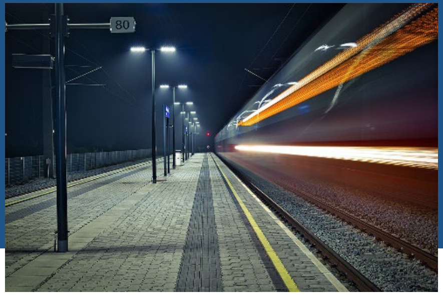
(on your own, using *ScHARe@Terra*)

- Instances of iris flowers …do their petal/sepal length/width vary naturally?
  - Vary by species… exploratory plots confirm [try scikit learn vignette]



Samples
(instances, observations)
Petal
Sepal
Features
(attributes, measurements, dimensions)
Class labels
(targets)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
…					
50	6.4	3.5	4.5	1.2	Versicolor
…					
150	5.9	3.0	5.0	1.8	Virginica



ScHARe

Science Collaborative for Health disparities
and Artificial intelligence bias REduction

Python Libraries & other Software Resources for Data Science Computational Strategies

A. Python’s Pre-eminence in Data Science

python SOFTWARE FOUNDATION

B. Inventory (non-exhaustive) of Examples:

i. Python

ii. Complementary software suites…

R (methods NOT YET in Python)

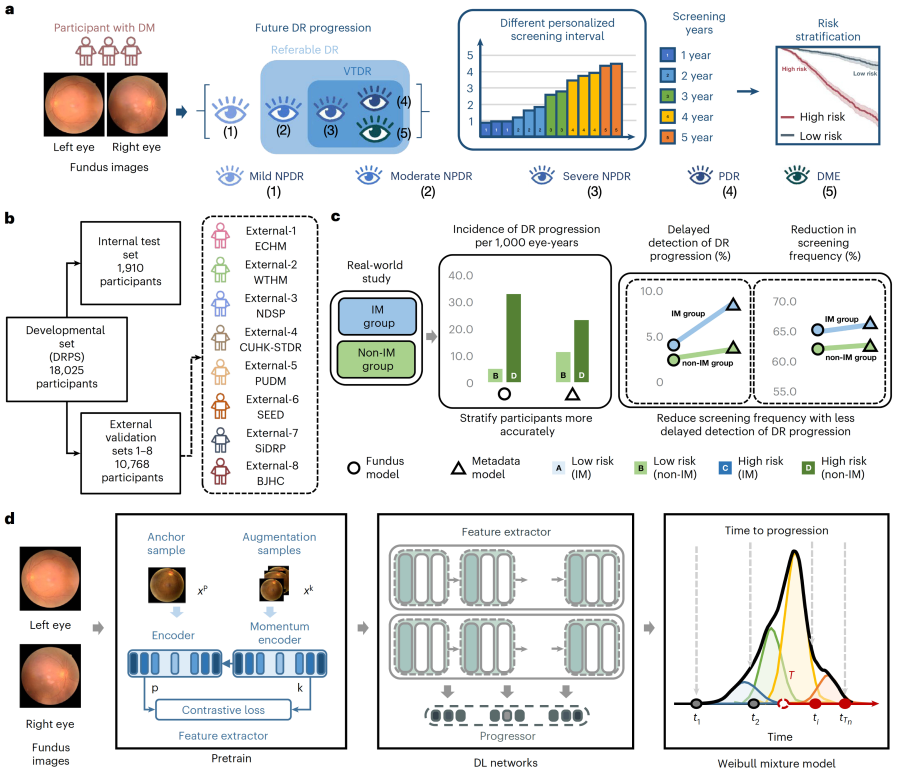
Commercial Software (specialized methods)

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

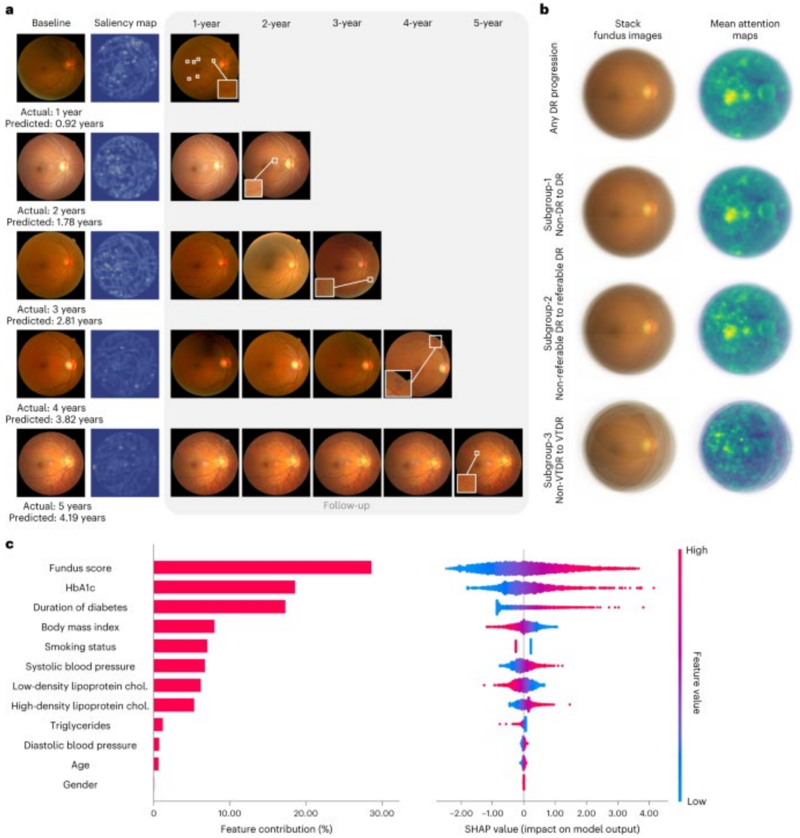
NIH National Institute of Diabetes and Digestive and Kidney Diseases

Python’s Pre-eminence in Data Science

- Jan2024 **example** with details on all software used, number **Python-based**:



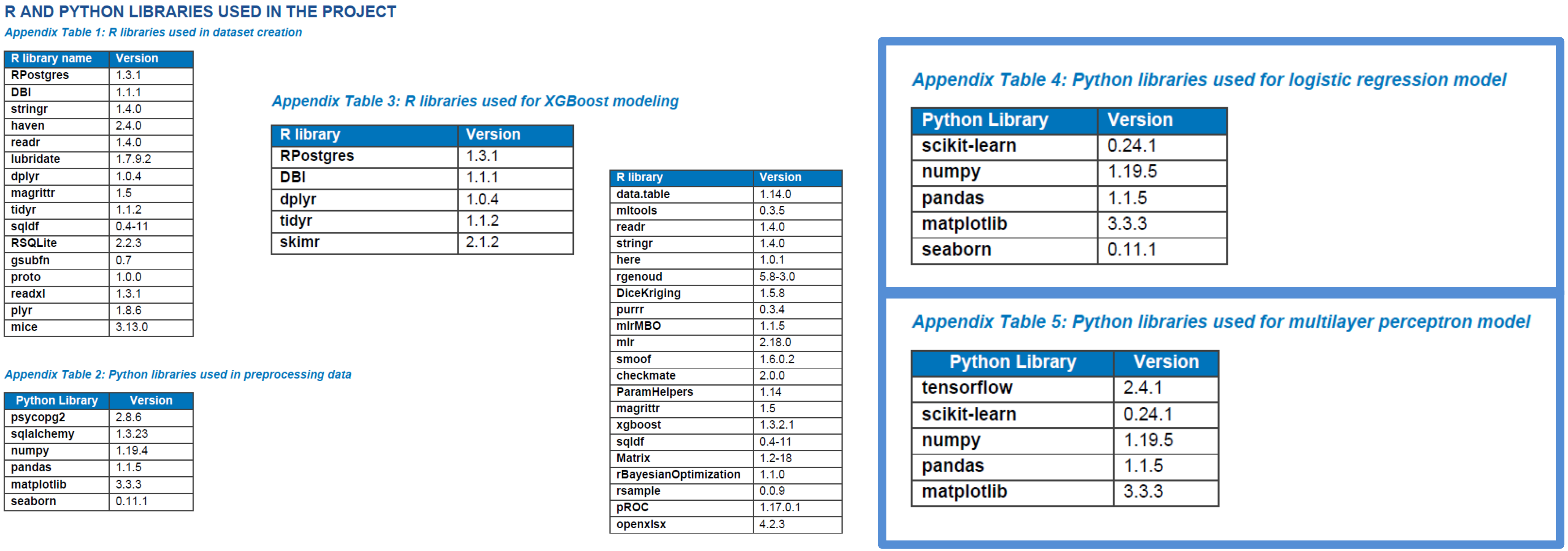
we used the SHAP Python package to illustrate the importance of clinical features as well as the fundus score (that is, predicted time-to-event by fundus model) involved in the combined model. SHAP stands for Shapley Additive exPlanations



The code used in the current study for developing the algorithm is provided at https://github.com/drpredict/DeepDR_Plus. Python version 3.9.0 was used for all statistical analvses in this study. The following third-party Python packages were used: Pytorch version 2.0.1 was used to build the DL models; Scikit-earn version 1.3.0 was used for calculating AUC. NumPy version 1.25.2 used for calculating C-index and Brier score. Lifelines version 0.27.7 was used for survival analysis.

Python's Pre-eminence in Data Science

- ***Counter* example** with details on all software used, a number **Python-based**:
  - ML-in-Patient-Centered-Outcomes Research Supervised Learning Task of mortality within 90-days of dialysis initiation, among patients diagnosed with end-stage disease



R AND PYTHON LIBRARIES USED IN THE PROJECT

Appendix Table 1: R libraries used in dataset creation

R library name	Version
RPostgres	1.3.1
DBI	1.1.1
stringr	1.4.0
haven	2.4.0
readr	1.4.0
lubridate	1.7.9.2
dplyr	1.0.4
magrittr	1.5
tidyr	1.1.2
sqldf	0.4-11
RSQLite	2.2.3
gsubfn	0.7
proto	1.0.0
readxl	1.3.1
plyr	1.8.6
mice	3.13.0

Appendix Table 2: Python libraries used in preprocessing data

Python Library	Version
psycopg2	2.8.6
sqlalchemy	1.3.23
numpy	1.19.4
pandas	1.1.5
matplotlib	3.3.3
seaborn	0.11.1

Appendix Table 3: R libraries used for XGBoost modeling

R library	Version
RPostgres	1.3.1
DBI	1.1.1
dplyr	1.0.4
tidyr	1.1.2
skimr	2.1.2

R library	Version
data.table	1.14.0
mltools	0.3.5
readr	1.4.0
stringr	1.4.0
here	1.0.1
rgenoud	5.8-3.0
DiceKriging	1.5.8
purrr	0.3.4
mlrMBO	1.1.5
mlr	2.18.0
smoof	1.6.0.2
checkmate	2.0.0
ParamHelpers	1.14
magrittr	1.5
xgboost	1.3.2.1
sqldf	0.4-11
Matrix	1.2-18
rBayesianOptimization	1.1.0
rsample	0.0.9
pROC	1.17.0.1
openxlsx	4.2.3

Appendix Table 4: Python libraries used for logistic regression model

Python Library	Version
scikit-learn	0.24.1
numpy	1.19.5
pandas	1.1.5
matplotlib	3.3.3
seaborn	0.11.1

Appendix Table 5: Python libraries used for multilayer perceptron model

Python Library	Version
tensorflow	2.4.1
scikit-learn	0.24.1
numpy	1.19.5
pandas	1.1.5
matplotlib	3.3.3

Inventory (non-exhaustive) of Complements to Python

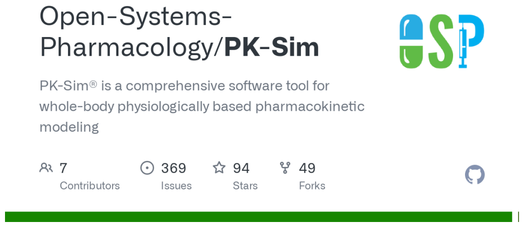
- Complementary software suites…

R / Julia / Stan (methods NOT FULLY in Python)

R

Stan

julia



Open-Systems-Pharmacology/**PK-Sim**

PK-Sim® is a comprehensive software tool for whole-body physiologically based pharmacokinetic modeling

7 Contributors 369 Issues 94 Stars 49 Forks

- Commercial Software (specialized methods)

Cytel

StatXact

LogXact

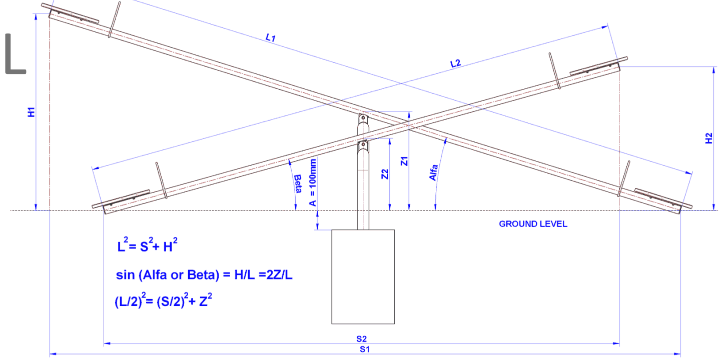
sas®

STATA®

Resources and Decision-Making Tools

A. infographics: decision support for participants

- Which use case features ‘tilt’ a data scientist toward AI/ML



B. links to online repositories for further exploration: participants can please check back with each new Think-a-Thon session…

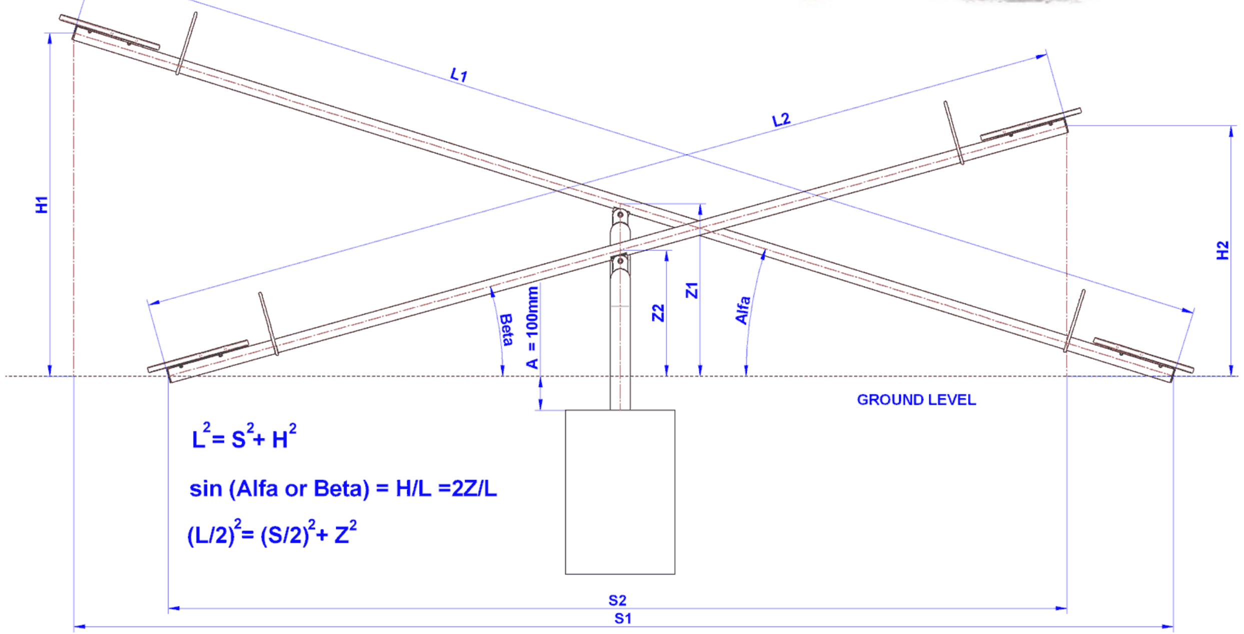
National Institute of Diabetes and Digestive and Kidney Diseases

https://www.craftsmanspace.com/sites/default/files/free-plans-articles/seesaw_playground_equipment_calculation.gif

Resources and Decision-Making Tools: infographic guidelines

- Data Science remains a (inherently *interdisciplinary*) profession that over-demand in the face of under-supply for the very reason that the only consistent guiding answer to question of *what to do is*: "**It depends**"

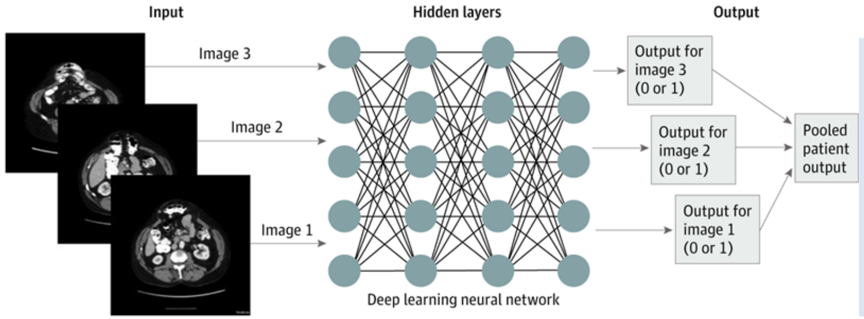
- Will propose in an online resource over coming Think-a-thons (with each TaT topic), what '***tilts***' choices in favor of one data methods over another…



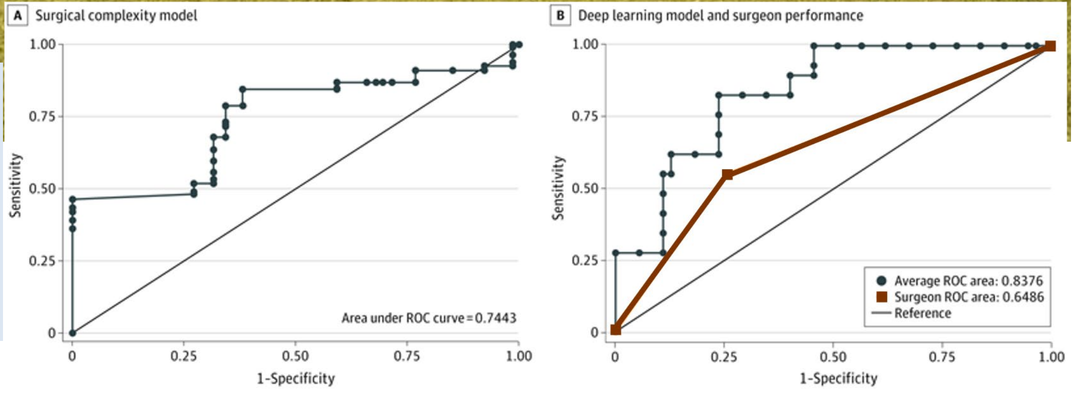
https://www.craftsmanspace.com/sites/default/files/free-plans-articles/seesaw_playground_equipment_calculation.gif

Resources and Decision-Making Tools: infographic guidelines

- Some use cases involve data that are so difficult to ‘structure’ that preference tilts *naturally* toward AI/ML
  - *Akin to kids looking to tilt see-saw @ right*
- Examples: images, sound-signals’ series & other multi-modal data fusion items
  - DL classifier to triage abdominal surgery


Input
Hidden layers
Output
Image 3
Image 2
Image 1
Output for image 3 (0 or 1)
Output for image 2 (0 or 1)
Output for image 1 (0 or 1)
Pooled patient output
Deep learning neural network

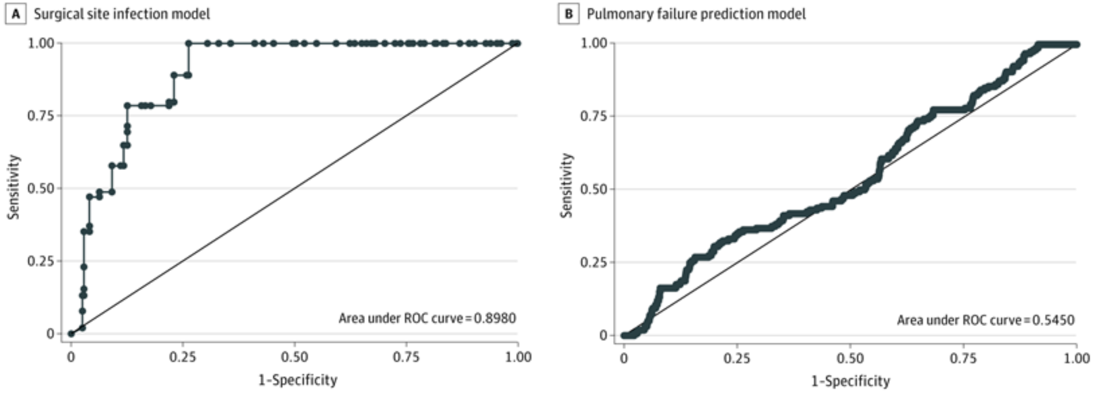
A, Surgical complexity model performance compared with a reference receiver operating characteristic curve (ROC) of 0.5 is depicted. Model performance vs reference value: $P < .001$. B, Deep learning model performance (blue line) and surgeon performance (orange line). The ROC is 0.19 greater for the deep learning model vs surgeon ($P < .001$).



http://www.atmo.arizona.edu/students/courselinks/spring08/atmo336s1/courses/fall18/atmo170a1s1/lecture_notes/mass_weight_density_pressure/publicworkshop_seesaw.jpg
https://jamanetwork.com/journals/jamasurgery/fullarticle/2781744 |

Resources and Decision-Making Tools: infographic guidelines

- Same use case above involved outcome so difficult to 'predict' that even deep learning AI/ML couldn't tilt process to improve over chance (50:50 coin-toss as diagonal reference area under ROC curve)
  - *Again, like kid hopes to tilt see-saw @right*
- Counter-example: expert-based decision-support system is needed
  - DL unhelpful to detect pulmonary failure


A Surgical site infection model
1.00
0.75
0.50
0.25
0
Sensitivity
0
0.25
0.50
0.75
1.00
1-Specificity
Area under ROC curve=0.8980
B Pulmonary failure prediction model
1.00
0.75
0.50
0.25
0
Sensitivity
0
0.25
0.50
0.75
1.00
1-Specificity
Area under ROC curve=0.5450

A, Surgical site infection model performance compared with a reference receiver operating characteristic curve (ROC) of 0.5 is depicted. Model performance vs reference value: *P* < .001. B, Pulmonary failure prediction model compared with a reference ROC of 0.5 is depicted. Model performance vs reference value: *P* = .03.

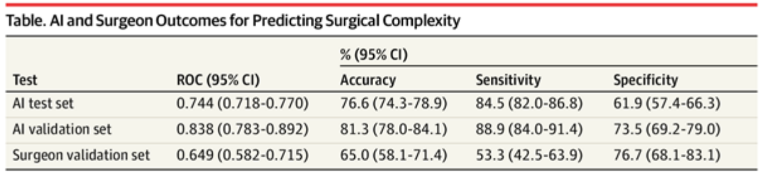
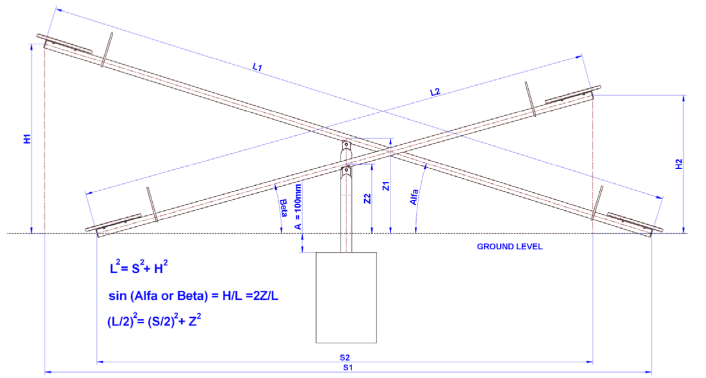


Table. AI and Surgeon Outcomes for Predicting Surgical Complexity

		% (95% CI)		
Test	ROC (95% CI)	Accuracy	Sensitivity	Specificity
AI test set	0.744 (0.718-0.770)	76.6 (74.3-78.9)	84.5 (82.0-86.8)	61.9 (57.4-66.3)
AI validation set	0.838 (0.783-0.892)	81.3 (78.0-84.1)	88.9 (84.0-91.4)	73.5 (69.2-79.0)
Surgeon validation set	0.649 (0.582-0.715)	65.0 (58.1-71.4)	53.3 (42.5-63.9)	76.7 (68.1-83.1)

http://www.atmo.arizona.edu/students/courselinks/spring08/atmo336s1/courses/fall18/atmo170a1s1/lecture_notes/mass_weight_density_pressure/publicworkshop_seesaw.jpg
https://jamanetwork.com/journals/jamasurgery/fullarticle/2781744 |

Resources and Decision-Making Tools: assessment check

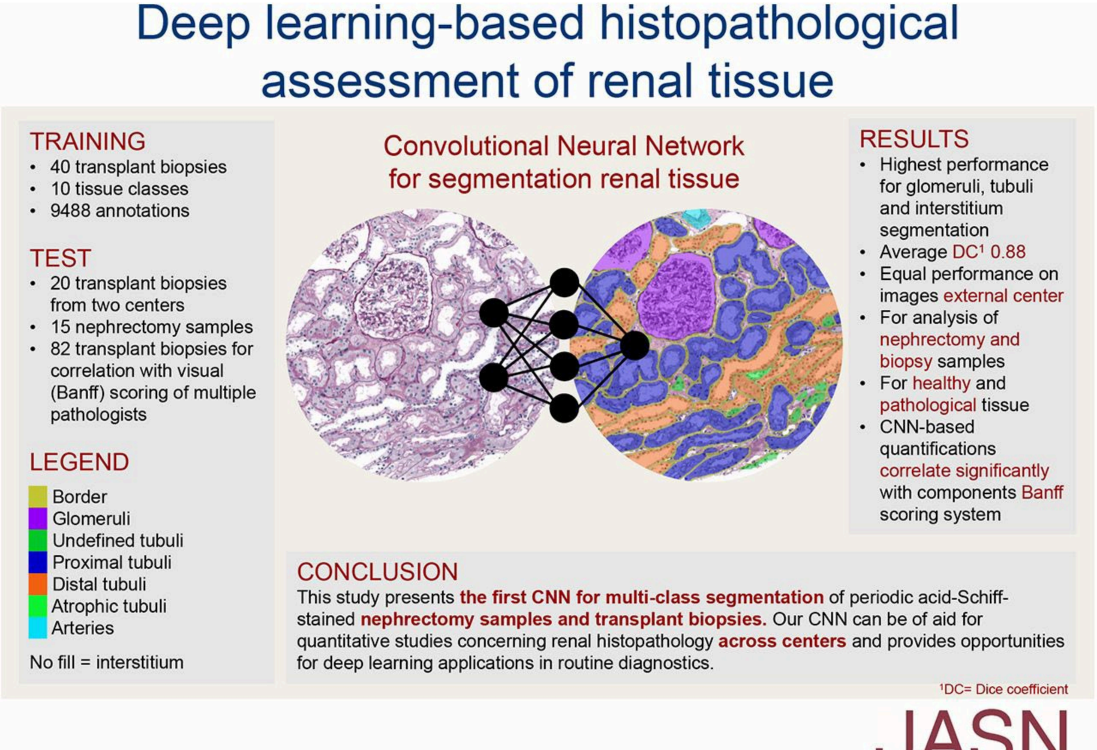


https://www.craftsmanspace.com/sites/default/files/free-plans-articles/seesaw_playground_equipment_calculation.gif

- Remember the only consistent guiding answer to question of *what to do is*: "**It depends**" – on what?

- *Modality* of data is thus one clear factor that '***tilts***' choices in favor of one data methods over another…

- Consider the task of 'segmenting' histopathologic images of kidney biopsies… what works **best?**



Deep learning-based histopathological assessment of renal tissue

TRAINING

- 40 transplant biopsies
- 10 tissue classes
- 9488 annotations

TEST

- 20 transplant biopsies from two centers
- 15 nephrectomy samples
- 82 transplant biopsies for correlation with visual (Banff) scoring of multiple pathologists

LEGEND

🟨	Border
🟪	Glomeruli
🟩	Undefined tubuli
🟦	Proximal tubuli
🟧	Distal tubuli
🟩	Atrophic tubuli
🟦	Arteries

No fill = interstitium

Convolutional Neural Network for segmentation renal tissue

RESULTS

- Highest performance for glomeruli, tubuli and interstitium segmentation
- Average DC[1] 0.88
- Equal performance on images external center
- For analysis of nephrectomy and biopsy samples
- For healthy and pathological tissue
- CNN-based quantifications correlate significantly with components Banff scoring system

CONCLUSION

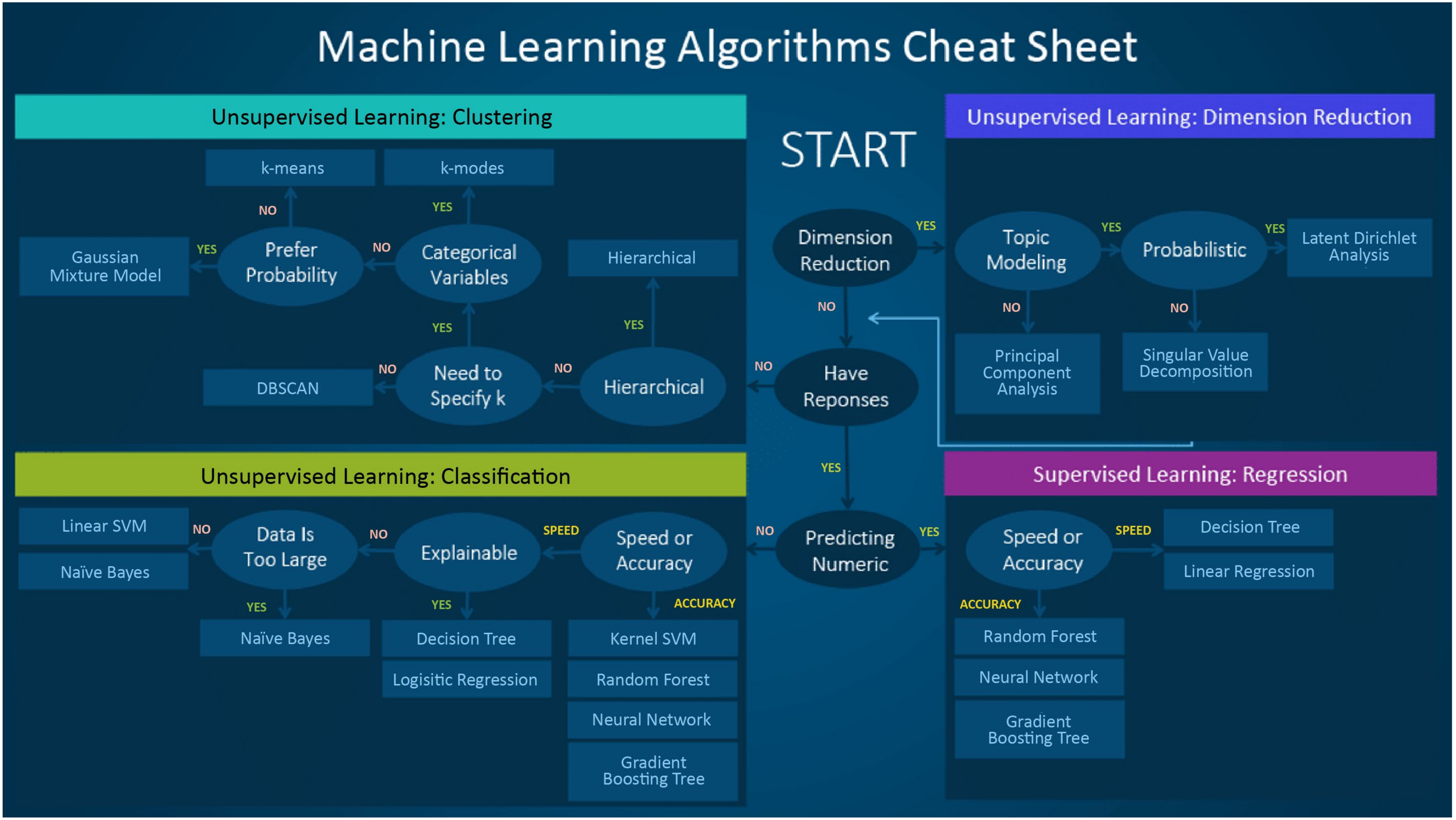
This study presents **the first CNN for multi-class segmentation** of periodic acid-Schiff-stained **nephrectomy samples and transplant biopsies.** Our CNN can be of aid for quantitative studies concerning renal histopathology **across centers** and provides opportunities for deep learning applications in routine diagnostics.

[1] DC= Dice coefficient

JASN
JOURNAL OF THE AMERICAN SOCIETY OF NEPHROLOGY

https://jasn.asnjournals.org/content/30/10/1968.abstract for more details: not only were Convolutional Neural Networks useful, but also statistical models using clinical scoring data

Resources and Decision-Making Tools: infographic guidelines



How can we navigate these different types of machine learning, to decide what's well-matched to our *use cases and data*?

Semi-supervised Learning, Reinforcement Learning evolved recently, so less amenable to any decision flow, like above cheat-sheet

https://medium.com/@dr.thomas.keil/which-algorithm-to-use-for-what-65d187ecc8d5

Resources and Decision-Making Tools: repositories for further exploration

https://www.pinterest.com/pin/data-science-stack-exchange--34340015886292381/

Python for data science:

- https://www.coursera.org/learn/python-for-applied-data-science-ai - This 4 module introduction to Python will kickstart your learning of Python for data science, as well as programming in general. This beginner-friendly Python course will take you from zero to programming in Python in a matter of hours.

- https://www.coursera.org/learn/data-analysis-with-python - Learn how to analyze data using Python. This course will take you from the basics of Python to exploring many different types of data. You will learn how to prepare data for analysis, perform simple statistical analysis, create meaningful data visualizations, predict future trends from data, and more!

- https://www.coursera.org/learn/python-for-data-visualization - The main goal of this Data Visualization with Python course is to teach you how to take data that at first glance has little meaning and present that data in a form that makes sense to people. Various techniques have been developed for presenting data visually but in this course, we will be using several data visualization libraries in Python, namely Matplotlib, Seaborn, and Folium.

- https://www.coursera.org/learn/machine-learning-with-python - This course dives into the basics of machine learning using an approachable, and well-known programming language, Python. You will learn about the purpose of Machine Learning and where it applies to the real world. You will also get a general overview of Machine Learning topics such as supervised vs unsupervised learning, model evaluation, and Machine Learning algorithms.

- https://jakevdp.github.io/WhirlwindTourOfPython/ - A fast-paced introduction to essential features of the Python language, aimed at researchers and developers who are already familiar with programming in another language. The material is particularly designed for those who wish to use Python for data science and/or scientific programming

- http://www.pythonchallenge.com/index.php - is a game in which each level can be solved by a bit of programming. You will be able to solve most riddles in any programming language, but some of them will require Python.

- http://data8.org/ - This is the UC Berkeley Foundations of Data Science course which combines three perspectives: inferential thinking, computational thinking, and real-world relevance. The course teaches critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of real-world datasets, including economic data, document collections, geographical data, and social networks. It delves into social issues surrounding data analysis such as privacy and design. Python based.

- https://automatetheboringstuff.com/ - You'll learn how to use Python to write programs that do in minutes what would take you hours to do by hand-no prior programming experience required. Once you've mastered the basics of programming, you'll create Python programs that effortlessly perform useful and impressive feats of automation.

- http://www.practicepython.org/ - There are over 30 beginner Python exercises just waiting to be solved. Each exercise comes with a small discussion of a topic and a link to a solution. New exercise are posted monthly.

https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks - This page is a curated collection of Jupyter/IPython notebooks that include interesting visual or technical content on a wide variety of programming and scientific computing topics such as image processing, NLP, and machine learning

Resources and Decision-Making Tools: repositories for further exploration

Broader resource materials:

- https://learngitbranching.js.org/ - An interactive way to learn git.

- **https://missing.csail.mit.edu/ - Classes teach you all about advanced topics within CS, from operating systems to machine learning, but there's one critical subject that's rarely covered, and is instead left to students to figure out on their own: proficiency with their tools. Learn how to master the command-line, use a powerful text editor, use fancy features of version control systems, and much more!**
- https://runestone.academy/runestone/books/published/thinkcspy/index.html- The goal of this book is to teach you to think like a computer scientist. This way of thinking combines some of the best features of mathematics, engineering, and natural science. Like mathematicians, computer scientists use formal languages to denote ideas (specifically computations).

- https://github.com/jmoon018/PacVim - PacVim is a fun game that teaches you vim commands. Vim is often called a "programmer's editor". It's not just for programmers, though. Vim is perfect for all kinds of text editing, from composing email to editing configuration files.

- **https://github.com/fabsta/interesting_notebooks - Collection of useful Kaggle notebooks**

- https://www.coursera.org/specializations/introduction-computer-science-programming - This specialization covers topics ranging from basic computing principles to the mathematical foundations required for computer science. You will learn fundamental concepts of how computers work, which can be applied to any software or computer system. You will also gain the practical skillset needed to write interactive, graphical programs at an introductory level.

- https://www.coursera.org/learn/software-processes - In this course, you will get an overview of how software teams work? What processes they use? What are some of the industry standard methodologies? What are pros and cons of each? You will learn enough to have meaningful conversation around software development processes.

- https://www.coursera.org/specializations/software-design-architecture - In the Software Design and Architecture Specialization, you will learn how to apply design principles, patterns, and architectures to create reusable and flexible software applications and systems. You will learn how to express and document the design and architecture of a software system using a visual notation

R for data science:

- https://rstudio.cloud/learn/primers Learn data science basics using these R cloud interactive tutorials. Topics include everything from data tidying to building interactive apps.

- https://r4ds.had.co.nz/ - This is an online book that will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it, visualize it and model it. In this book, you will find a practicum of skills for data science.

- https://github.com/rfordatascience/tidytuesday - Join the R4DS Online Learning Community in the weekly #TidyTuesday event! Every week we post a raw dataset, a chart or article related to that dataset, and ask you to explore the data. The goal of TidyTuesday is to apply your R skills, get feedback, explore other's work, and connect with the greater #RStats community!

- https://datacarpentry.org/semester-biology/nav/getting-started/ - This website hosts introductory material for teaching biologists how to interact with data including: data structure, database management systems, and programming for data manipulation, analysis, and visualization. Most of the modules use R.

- https://www.coursera.org/specializations/statistics - Master Statistics with R in this coursera mooc. Statistical mastery of data analysis including inference, modeling, and Bayesian approaches.

- https://www.coursera.org/specializations/jhu-data-science - This 10 course data science specialization covers the concepts and tools you'll need throughout the entire data science pipeline, from asking the right kinds of questions to making inferences and publishing results using R.

- https://www.coursera.org/specializations/genomic-data-science - This specialization covers the concepts and tools to understand, analyze, and interpret data from next generation sequencing experiments. It teaches the most common tools used in genomic data science including how to use the command line, Python, R, Bioconductor, and Galaxy.

- https://leanpub.com/universities/set/jhu/cloud-based-data-science - Cloud Based Data Science (CBDS) is a free online educational to help anyone who can read, write, and use a computer to move into data science. It is a sequence of 11 MOOCs offered by faculty members in the Johns Hopkins Department of Biostatistics, Bloomberg School of Public Health.

- https://leanpub.com/rprogramming - This book brings the fundamentals of R programming to you, using the same material developed as part of the industry-leading Johns Hopkins Data Science Specialization. The skills taught in this book will lay the foundation for you to begin your journey learning data science.

- https://swirlstats.com/ - swirl teaches you R programming and data science interactively, at your own pace, and right in the R console.

- https://exercism.io/tracks/r/ offers programming puzzles to solve against a provided set of test cases. Mimicking the workflow of test-driven development (TDD), Exercism emphasizes iteration and refactoring. After solving a puzzle, solutions can be discussed with a mentor and peers' solutions can be reviewed.

- https://dreamrs.github.io/esquisse/index.html- The purpose of this add-in is to let you explore your data quickly to extract the information they hold. The interactive plots also come with the code used to generate them, so it can be a useful way to learn data visualization with ggplots.

- https://github.com/calligross/ggthemeassist- this will help you with ggplot visualization themes. You can modify the attributes of the graph in real time and this package will modify your code for the graph output.

- https://happygitwithr.com/ - This tutorial will help you install Git and get it working smoothly with GitHub, in the shell and in RStudio, develop a few key workflows that cover your most common tasks and integrate Git and GitHub into your daily work with R and RMarkdown.

ScHARe

Science Collaborative for Health disparities
and Artificial intelligence bias REduction

Q&A and Closing Remarks

A. Remarks on context, recap of main points

- ScHARe staff will cover

B. Avenues for follow-up, & further exploration: office hours tomorrow, other NIH resources, curated decision support…

Join ZoomGov Meeting / Single-click Direct link
https://nih.zoomgov.com/j/16186685057?pwd=RXhkZkJ6QVQ2UTJadEV2bHJ5ay9mZz09
Meeting ID: 161 8668 5057
Passcode: 008707
One tap mobile+16692545252,,16186685057#,,,,*008707#
US (San Jose)+16469641167,,16186685057#,,,,*008707#
US (US Spanish Line)Dial by your location
+1 669 254 5252 US (San Jose)
+1 646 964 1167 US (US Spanish Line)
+1 646 828 7666 US (New York) | +1 551 285 1373 US (New Jersey)
+1 669 216 1590 US (San Jose) | +1 415 449 4000 US (US Spanish Line)
Meeting ID: 161 8668 5057
Passcode: 008707
Find your local number: https://nih.zoomgov.com/u/aca1qfBfaVJoin by
SIP16186685057.008707@sip.zoomgov.com
Join by H.323161.199.138.10 (US West)
161.199.136.10 (US East)

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Q&A and Closing Remarks

NIH National Institute of Diabetes and Digestive and Kidney Diseases

NIH National Institute of Diabetes and Digestive and Kidney Diseases

office hours tomorrow, Thursday, January 18th: 11am EST – 12:30pm EST

NIH

National Institute of
Diabetes and Digestive
and Kidney Diseases

NIH

National Institute of
Diabetes and Digestive
and Kidney Diseases

Data Science computational strategies glossary

- Glossary (for internal ref)
- Here's an overview of the critical elements that make up the anatomy of AI:

- Data: Data are the lifeblood of AI. It includes structured and unstructured information, such as text, images, audio, etc. AI systems rely on large datasets for training and learning.

- Algorithms: AI algorithms are the core mathematical and computational instructions that enable AI systems to process and analyze data. These algorithms include machine learning, deep learning, reinforcement learning, natural language processing (NLP), and many more.

- Machine Learning: Machine learning is a subset of AI that focuses on developing algorithms that allow computers to learn and make predictions or decisions without being explicitly programmed. Standard techniques include supervised learning, unsupervised learning, and reinforcement learning.

- Deep Learning: Deep learning is a subset of machine learning that uses neural networks with multiple layers (deep neural networks) to process data. It is particularly effective for tasks like image and speech recognition.

- Neural Networks: Neural networks are inspired by the structure and function of the human brain. They consist of interconnected artificial neurons that process and transfer information. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are standard in deep learning.

- Natural Language Processing (NLP): NLP is a subfield of AI that focuses on the interaction between computers and human language. It enables tasks like language translation, sentiment analysis, and chatbots.

- Computer Vision: Computer vision is the field of AI that enables machines to interpret and understand visual information from the world, such as images and videos. It's used in applications like image recognition, facial recognition, and object detection.

- Speech Recognition: This technology enables machines to understand and transcribe spoken language. It's used in voice assistants and voice command systems.

- Reinforcement Learning: Reinforcement learning is a type of machine learning that focuses on training AI agents to make a sequence of decisions to maximize a cumulative reward. It's used in gaming, robotics, and autonomous systems.

- Big Data: AI often relies on large datasets for training and analysis. Big data technologies and tools, including distributed computing and storage, play a significant role in the AI ecosystem.

- Training Data: AI models require training data to learn patterns and make predictions. The quality and quantity of training data are critical factors in AI performance.

- Hardware: AI workloads can be computationally intensive. Specialized hardware, such as Graphics Processing Units (GPUs) and TPUs (Tensor Processing Units), are often used to accelerate AI training and inference.

- Cloud Computing: Many AI applications are deployed on cloud platforms, which offer scalability and accessibility to AI resources and services.

- Ethics and Bias Mitigation: As AI systems are trained on data, there is a growing emphasis on addressing bias and ethical considerations in AI development and usage.

- Robotic Process Automation (RPA): In AI, RPA automates rule-based tasks in business processes, often involving software bots.

- Decision-Making: AI systems are designed to make decisions or recommendations based on the patterns they've learned from data.

- User Interface: AI often interacts with users through chatbots, voice assistants, and recommendation systems.

- Regulation and Compliance: As AI technologies become more prevalent, there's a growing focus on regulations and compliance related to AI, particularly in areas like data privacy and security.

Data Science Computational Strategies

- AI anatomy notes ->
- Add’l slides, if needed

The anatomy of AI is diverse, incorporating various technologies, techniques, and considerations to enable machines to exhibit intelligent behavior and perform a wide range of tasks. It's a rapidly evolving field with applications across industries.

The anatomy of Artificial Intelligence (AI) can be divided into the following three main components:

1. *Hardware: AI systems need powerful hardware to process large amounts of data and perform complex calculations. This hardware can include CPUs, GPUs, and TPUs.*

2. *Software: AI systems need software to implement AI algorithms and to interact with the real world. This software can include machine learning frameworks, deep learning libraries, and natural language processing tools.*

3. *Data: AI systems need data to learn from. This data can come from various sources, such as sensors, databases, and the Internet.*

ScHARe

Resources

ScHARe resources

Support made available to users:

ScHARe-specific

- ScHARe documentation
- Email support

Platform-specific

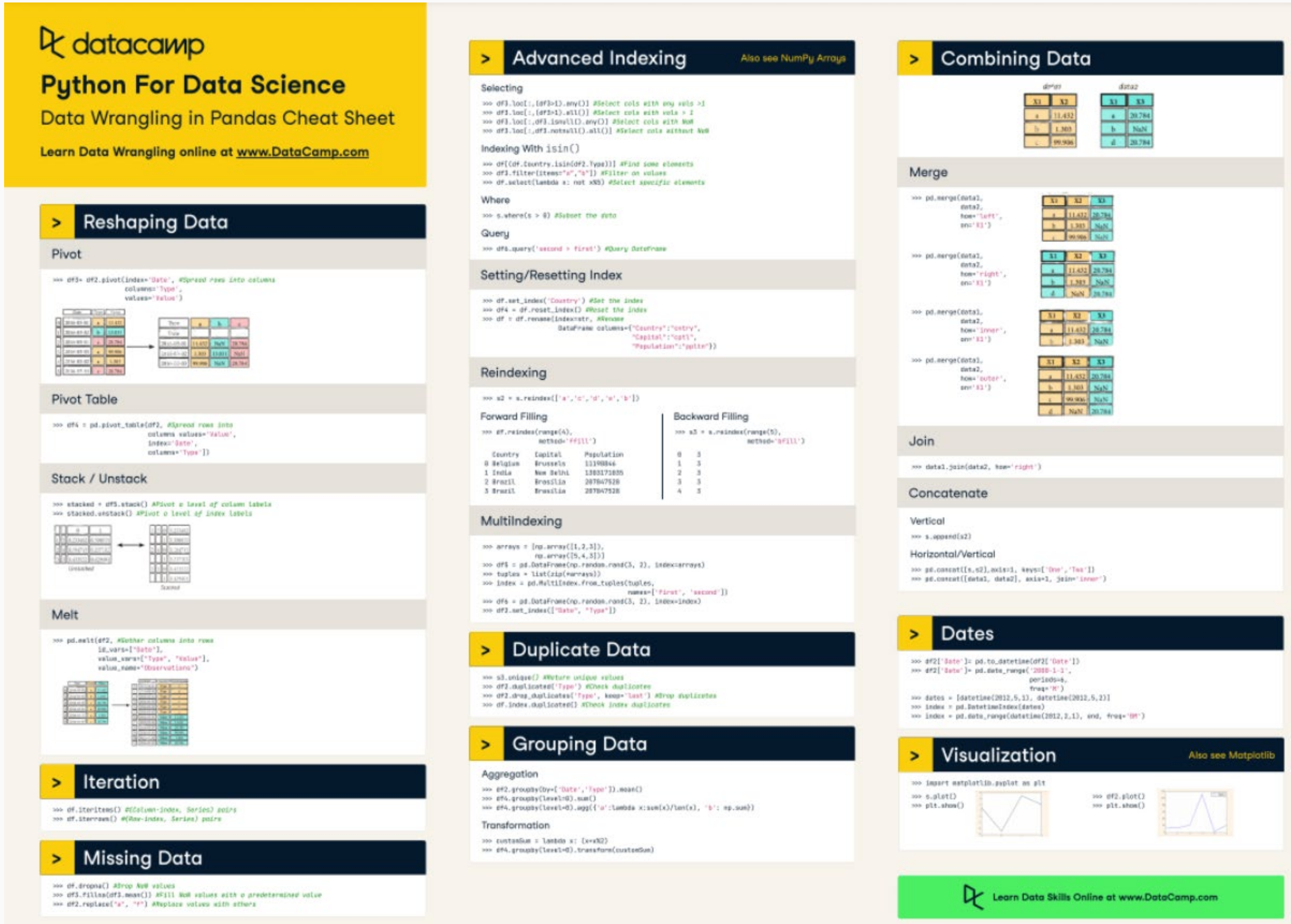
- Terra-specific support
- Terra-specific documentation

ScHARe resources

Training opportunities made available to users:

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

ScHARe resources: cheatsheets



datacamp

Python For Data Science

Data Wrangling in Pandas Cheat Sheet

Learn Data Wrangling online at www.DataCamp.com

> Reshaping Data

Pivot

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

Pivot Table

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

Stack / Unstack

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

Melt

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

> Iteration

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

> Missing Data

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

> Advanced Indexing

Also see NumPy Arrays

Selecting

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

Indexing With isin()

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

Where

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

Query

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

Setting/Resetting Index

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

Reindexing

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

Forward Filling

```
>>> df.reindex(range(4),
               method='ffill')
  Country   Capital    Population
0 Belgium   Brussels   11190846
1 India     New Delhi  1303171035
2 Brazil    Brasilia   207847528
3 Brazil    Brasilia   207847528
```

Backward Filling

```
>>> s3 = s.reindex(range(5),
                   method='bfill')
0    3
1    3
2    3
3    3
4    3
```

MultiIndexing

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

> Duplicate Data

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

> Grouping Data

Aggregation

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

Transformation

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

> Combining Data

data1

X1	X2
a	11.432
b	1.303
c	99.906

data2

X1	X3
a	20.784
b	NaN
d	20.784

Merge

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

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

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

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

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

X1	X2	X3
a	11.432	20.784
b	1.303	NaN

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

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

Join

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

Concatenate

Vertical

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

Horizontal/Vertical

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

> Dates

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

> Visualization

Also see Matplotlib

```
>>> import matplotlib.pyplot as plt
>>> s.plot()
>>> plt.show()
>>> df2.plot()
>>> plt.show()
```

Learn Data Skills Online at www.DataCamp.com

Credits: datacamp.com

Terra resources

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

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

ScHARe

Thank you

Think-a-Thon poll

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

□ Very useful

□ Useful

□ Somewhat useful

□ Not at all useful

Think-a-Thon poll

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

□ Too fast

□ Adequate for me

□ Too slow

Think-a-Thon poll

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

□ Very interested, will definitely attend

□ Interested, likely will attend

□ Interested, but not available

□ Not interested in attending any others

ScHARe

Next Think-a-Thons:

bit.ly/think-a-thons

Register for ScHARe:

bit.ly/join-schare

✉ schare@mail.nih.gov