National Institute of Biomedical Imaging and Bioengineering

Creating trustworthy open data for scientific discovery

New York Scientific Data Summit 2024: Addressing Data Challenges in Digital Twins

New York City, New York

September 16, 2024

Grace C.Y. Peng, PhD

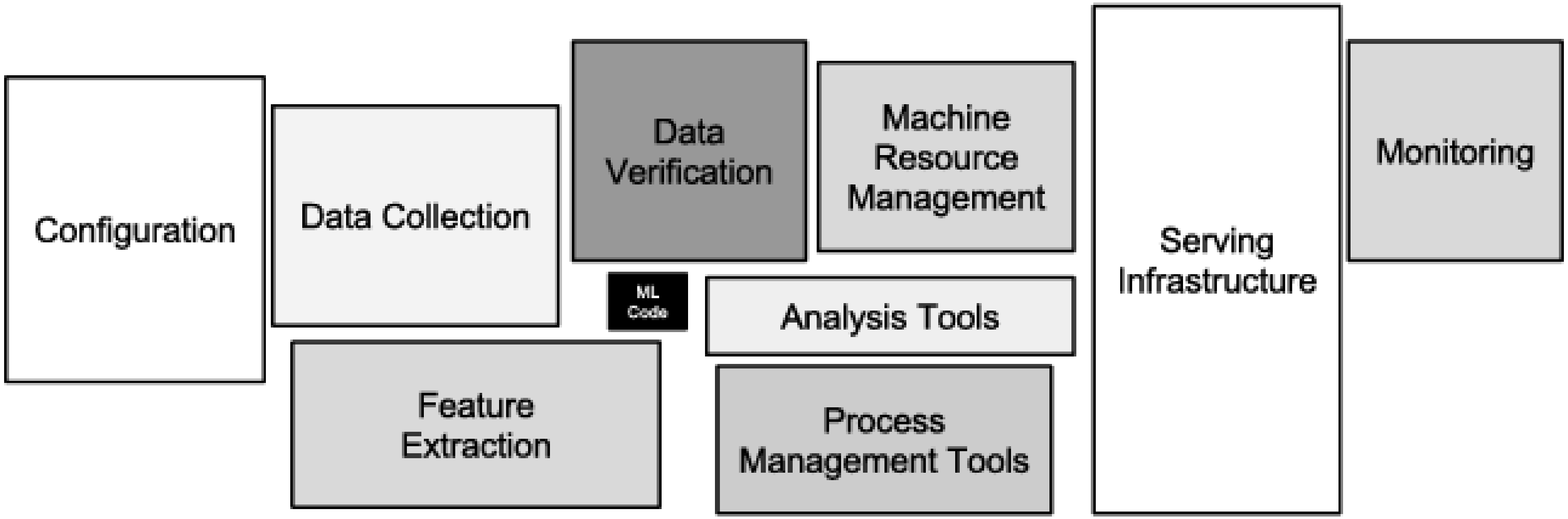


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Scully et al. (2015): Hidden technical debt in Machine learning systems [doi: 10.5555/2969442.2969519]

- **December 6, 2019 ACD AI WG Report**
- https://acd.od.nih.gov/documents/presentations/12132019AI_FinalReport.pdf
- **December 13, 2019 ACD presentation**
- https://acd.od.nih.gov/documents/presentations/12132019AI.pdf

Report of the ACD AI WG

December 6, 2019

TABLE OF CONTENTS

The NIH Bridge2AI Program

Supported by the NIH Common Fund

Bridge2AI Program Management Team

Co-Chairs
Michael Chiang
Eric Green
Helene Langevin
Steve Sherry
Bruce Tromberg

Common Fund Program Leader
Haluk Resat

Common Fund Program Officers
Chris Kinsinger
George Papanicolaou

Working Group Coordinators
James Gao, NEI
Lanay Mudd, NCCIH
Grace Peng, NIBIB
Shurjo Sen, NHGRI

Common Fund Staff
Natalie Vineyard (Comm)
David Dzamashvili (Ops)
Karen Kellton (Prog Mgmt)
Kristina Faulk (Prog Coord)

Awards Management
Kristen Kreuter (DOTM)
Erna Petrich (DOTM)

Federal Working Group (+100 Members)
CC, CIT, FIC, NCATS, NCI, NCCIH, NEI, NHGRI, NIA, NIAID, NICHD, NIBIB, NIDA, NIDDK, NIAMS, NIGMS, NIMHD, NINDS, NLM

Other Federal Agencies:
DARPA, DOE, FDA, NIST, NSF

Bridge to Artificial Intelligence

Vision: to propel biomedical and behavioral research forward by setting the stage for widespread use of artificial intelligence (AI) technologies

Goals:

- Use biomedical and behavioral research grand challenges to generate **flagship datasets**
- **Prepare** AI/ML-friendly data
- Prioritize **ethical** best practices
- Promote **diverse perspectives**

DATA

Diverse

FAIR

AI-ready

ETHICS

Accurate

Reliable

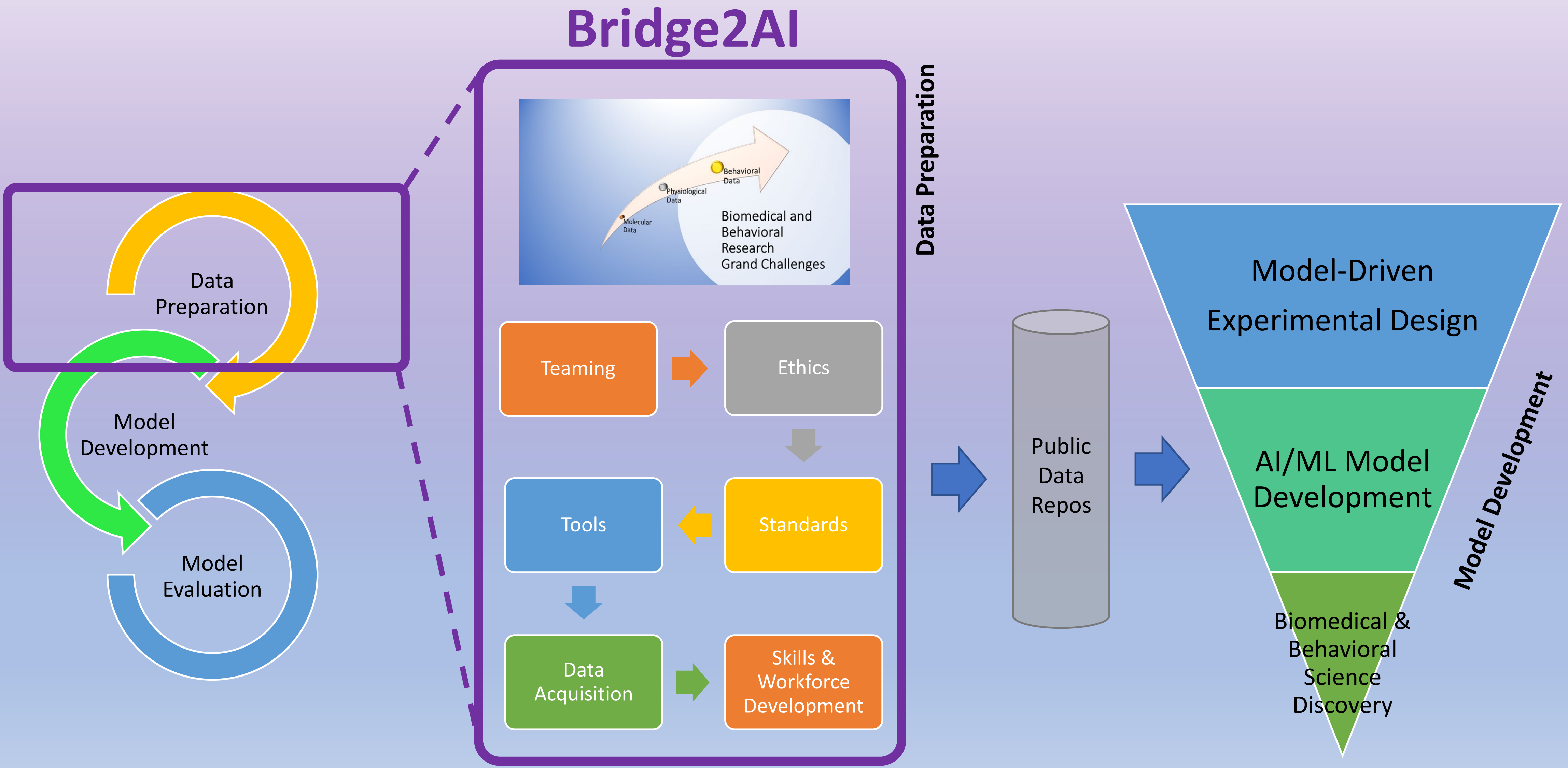
Ethically-sourced

PEOPLE

Diverse teams & research cohorts

Training

Scientific Discovery Pipeline



Grand Challenges -- Data Generation Projects

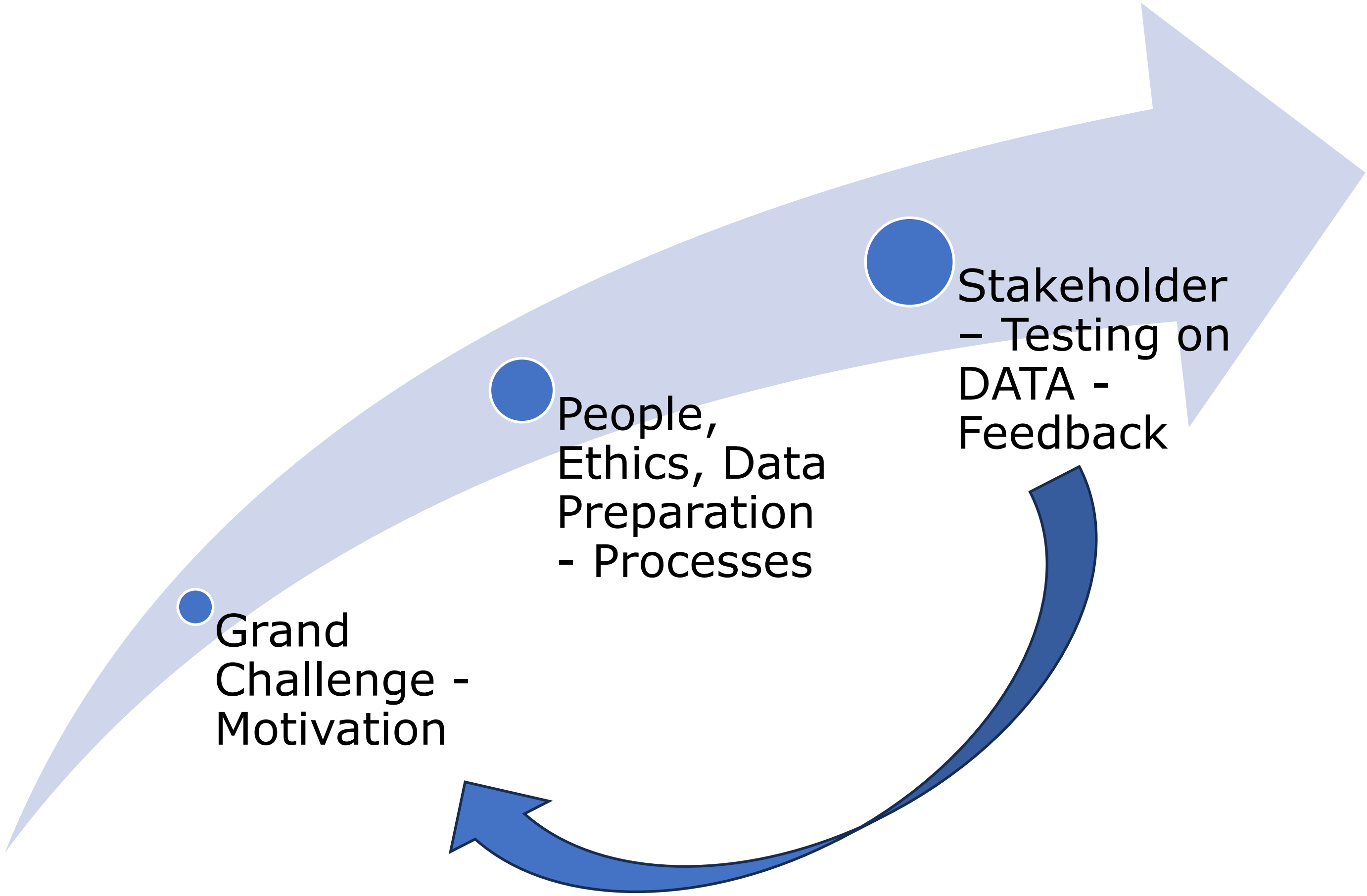
Clinical Care - Using imaging, clinical, and other data collected in an **ICU setting** for diagnosis and risk prediction

Precision Public Health - Using **voice as a biomarker** for human health, revealing how genomic variation, human development, behavioral, and environmental factors affect individual and population health

Salutogenesis (Return to Health) - Uncovering the details of how human health is restored after disease, using **type 2 diabetes** as a model

Functional Genomics - Mapping spatiotemporal architecture of human cells to interpret cell structure/function in health and disease

From Vision to Deliverables



Flagship Data Generation

Criteria for ML-Friendly Data

Ethical Principles

Consent Standards

Data Access Standards

Curricula

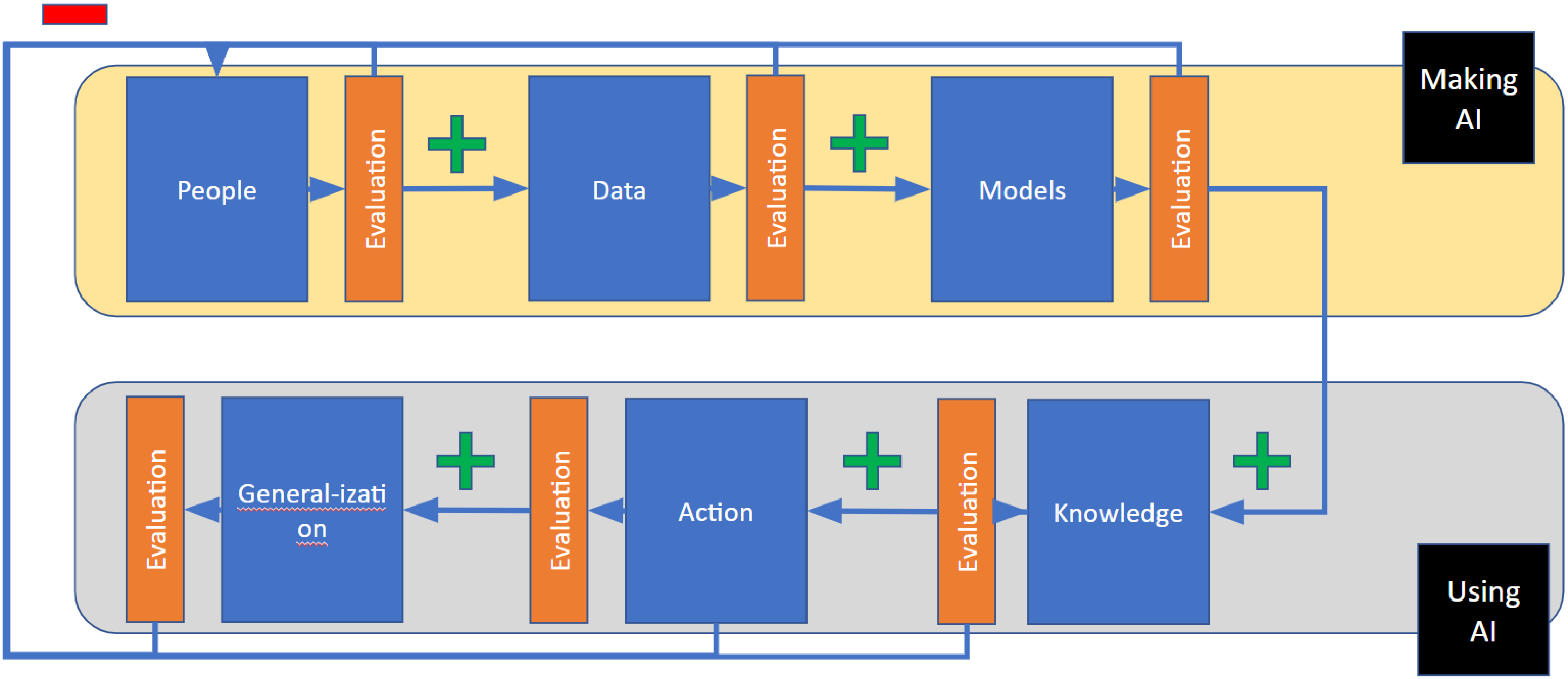
Data Sheets

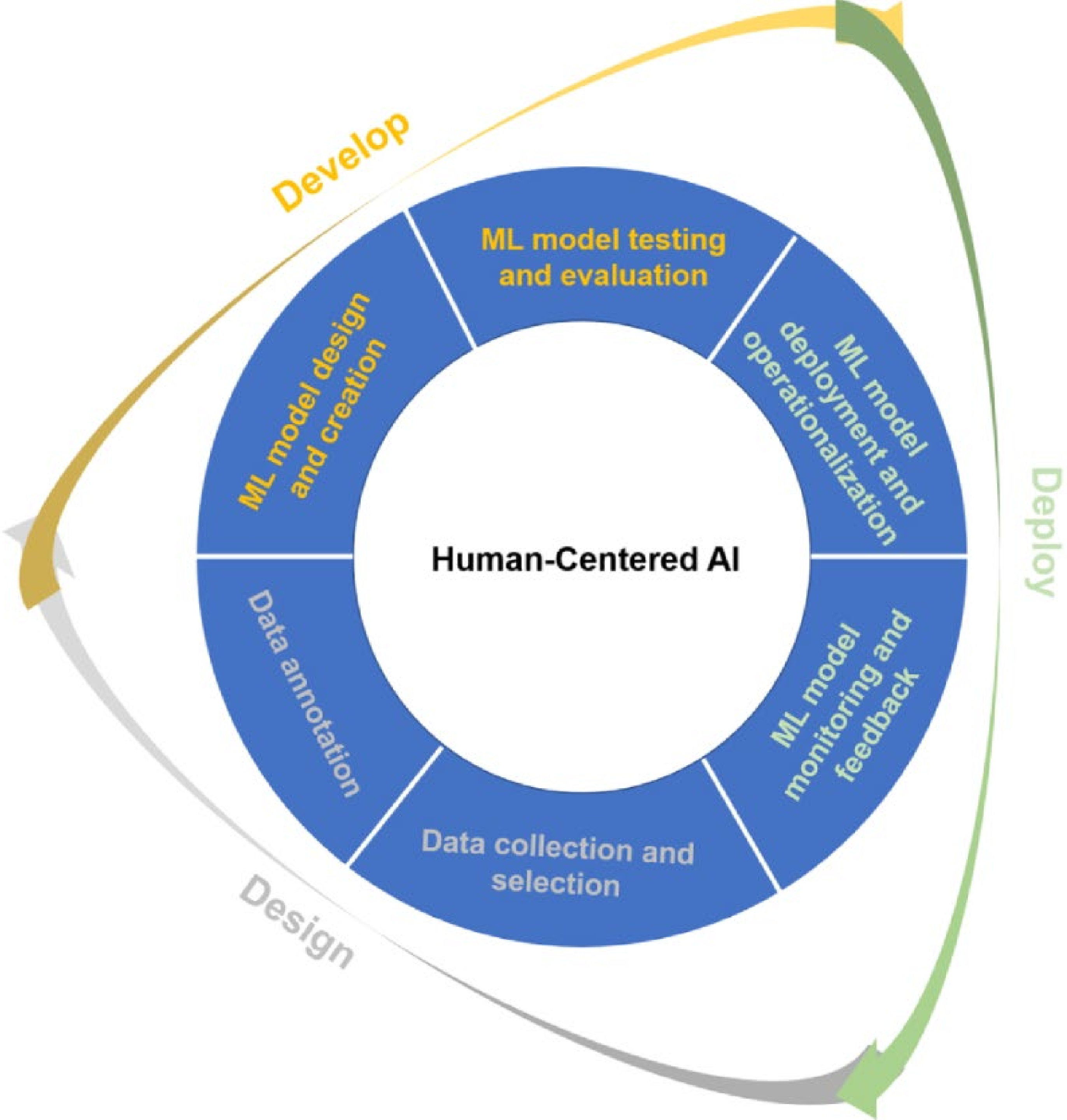
Model Cards

Bridge2AI

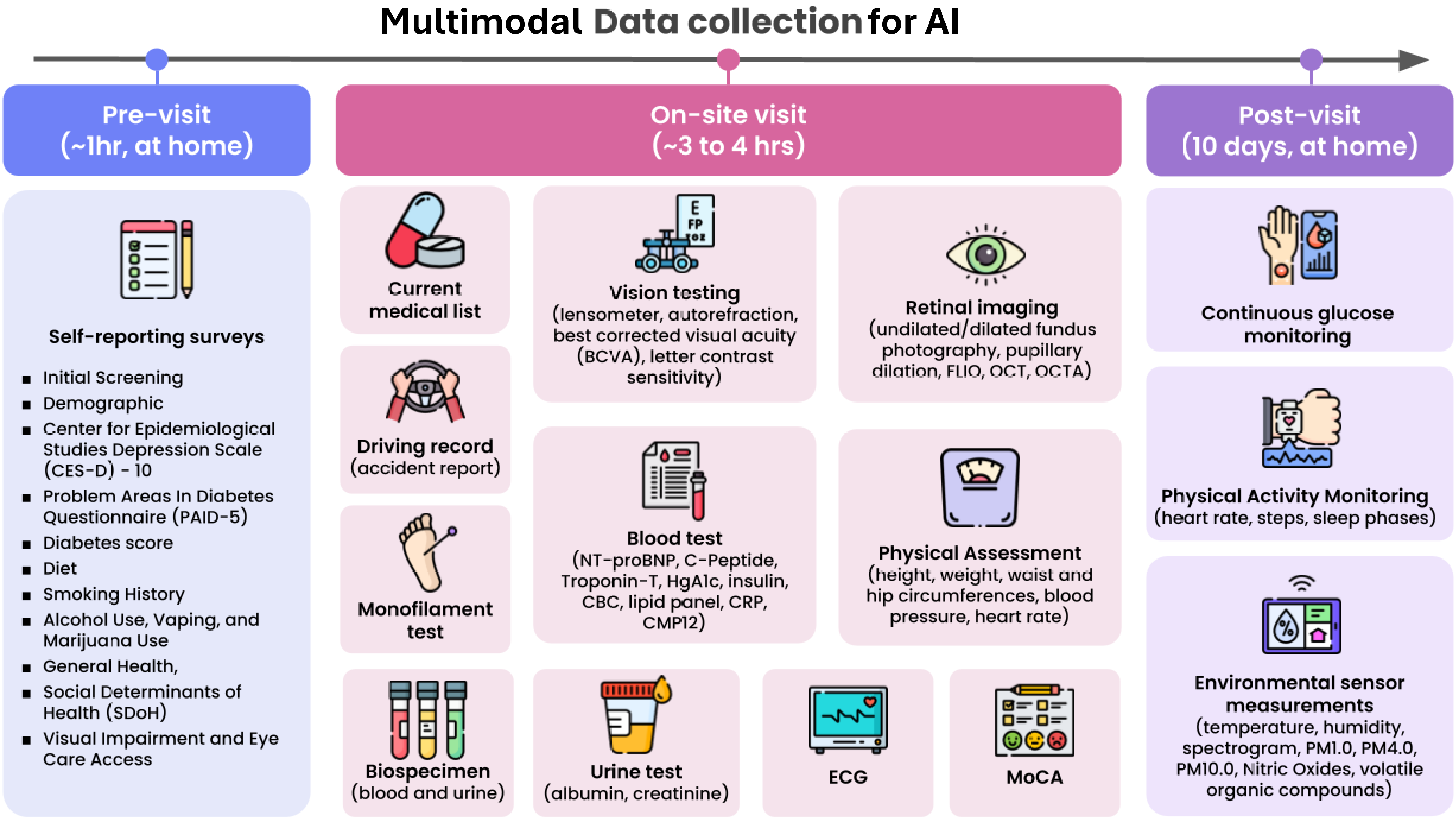
Generating ethically sourced data and best practices

Ethics Must be Embedded from the Outset





Chen, Clayton, Novak, Anders, Malin. Human-Centered Design to Address Biases in Artificial Intelligence. JMIR. 2022.



FLIO = Fluorescence Lifetime Imaging, OCT = Optical Coherence Tomography, OCTA = Optical Coherence Tomography Angiography,
ECG = Electrocardiogram, MoCA = Montreal Cognitive Assessment, PM1.0, 4.0, and 10.0 = Particulate matter less than 1, 4, and 10 microns, respectively

Ethics beyond compliance

BRIDGE2AI

Consent example:

- By signing this consent, you agree that all the medical data that is collected, apart from your direct HIPAA will be released in a public repository.
- Although low, there is a risk that someone will attempt to re-identify you through the data release and it there is a residual risk that development of new technologies will allow people to re-identify you in the future
- Companies who download your data are not allowed to sell it but may use your data to develop models for commercial intent

BRIDGE2AI

		EHR/CLINICAL	SURVEYS	IMAGING	SENSOR-BASED	OMICS	WAVEFORM
PRECISION PUBLIC HEALTH Bridge2AI Voice Cloud environment Microsoft Azure	A database of 10,000 diverse bioacoustic waveforms is being established to establish voice biomarkers in mental health, respiratory, neurological, and other areas.	• Demographics • Diagnosis (ICD) • Severity of disease • Treatment information • Social history (smoking, alcohol)	• 12 validated questionnaires (e.g., MOCA, GAD-7, VHI-10, PANAS, DI, etc.)	• Brain MRI/CTs • Chest/neck CTs • Laryngoscopy		• Whole genome sequencing	• Bioacoustic data tasks of voice and non-voice sounds, shared as waveforms, Mel spectrograms, features
		OMOP	OMOP	Brain imaging: DICOM; laryngoscopy: MP4		CRAM & VCFs with metadata	Waveform database (WFDB); creating new standard for bioacoustics
SALUTOGENESIS AI-READI AI Ready and Equitable Atlas for Diabetes Insights Cloud environment Microsoft Azure	Creating a temporal atlas from 3,000 individuals around pathogenesis and salutogenesis to expand applications of AI in clinical care, focusing on Type 2 diabetes	• Demographics, SDoH • Diet • Social history • Lab tests (blood, urine) • Monofilament test • Physical assessment • Medications • Vision testing	• Multiple validated self-reporting surveys (CES-D, PAID-5, etc.)	• Retinal imaging (undliated/dilated fundus photography, pupillary dilation, FLIO, optical coherence tomography (OCT), OCT angiography)	• Continuous glucose monitoring (CGM) • Physical activity monitoring (heart rate, steps, sleep phases) • Environmental sensors (air quality and particulate measures, temperature)	• Whole genome sequencing	• Electrocardiogram (ECG)
		OMOP, LOINC	OMOP, LOINC	DICOM	CGM, physical activity: open mHealth; Air: Earth Science Data Spec	CRAM & VCFs with metadata	Waveform database (WFDB)
CLINICAL CARE CHoRUS Cloud environment Microsoft Azure	Establishing a set of >100,000 patients from 14 ICU sites across the United States to improve recovery from acute illnesses	• Demographics, SDoH • Clinical notes • Lab tests • Medications • Encounters • Procedures		• All imaging acquired during ICU setting and captured in PACS (MR, CT, US, x-ray)			• Physiological data (ECG; electroencephalogram, EEG)
		OMOP, LOINC		DICOM			Waveform database (WFDB)
FUNCTIONAL GENOMICS CM4AI Cell Maps for AI Cloud environment Google Cloud	Creating a library of large-scale maps of cellular structure, function, and disease contexts using cell lines. 200 genes/proteins are the subject of coordinated experiments in three modalities			• Immunofluorescence imaging data for cell imaging		• Proteomic mass spectrometry • CRISPR perturbation scRNA-Seq Datasets • Cell maps	
				Cell imaging: RO-Crate with JPEG 4-channel (red, green blue, yellow) and metadata		Mass spec: RO-Crate w/TSV & metadata; CRISPR: RO-Crate with h5ad file & metadata; Cell maps: RO-Crate with Cytoscape CX & metadata	

Bridge2AI is supported by NIH U54 HG012510, U54 HG012513, U54 HG012517, OT2 OD032720, OT2 OD032742, OT2 OD032644, OT2 OD032701

Towards Best Practices

What type of Data are you collecting?	Identifier under HIPAA	Non-Identifier under HIPAA	High/low Risk of Re-identification under Common Rule		
Do you need consent and what should it contain?	Consent exempt	Consent	Assent	Blanket Consent, Opt-in/Opt Out, Menu?	
Who will be using your data and how?	Only Pis from Academia	Only academic researchers	Public population	Companies	
What kind of regulatory contracts do you need?	Codes of Conducts	DUA	License		
What are the risk vs benefits of your data being released?	Technology available	Type of similar data online	Public health context	Urgency	
Can technology be used to diminish risk?	Blockchain technology	Watermarking	Others!		

Bridge2AI

Lessons Learned so far

What make Bridge2AI challenging?

Our Goal:
Propel Scientific Discovery

Biomedical Research

- Humans → inferred knowledge
- Heterogeneous, messy data
- Non-standardized processes
- Validation through scientific method → Diversity
- Open culture of sharing

AI/ML

- Machines → explicit knowledge
- “Complete” data
- Standardized algorithms
- Training → Bias
- Closed culture of security

Ethical Challenges → for Open Science

- **Biases:** Issues related to inherent biases of the data
- **Informed Consent:** Going beyond a legal consent form
  - How do we ensure consent given the evolving landscape of AI/ML?
- **Re-identification:** Navigating the risk of re-identification with multi-modal data
- **Unauthorized Use:** How do we prevent unauthorized secondary use?

People Challenges

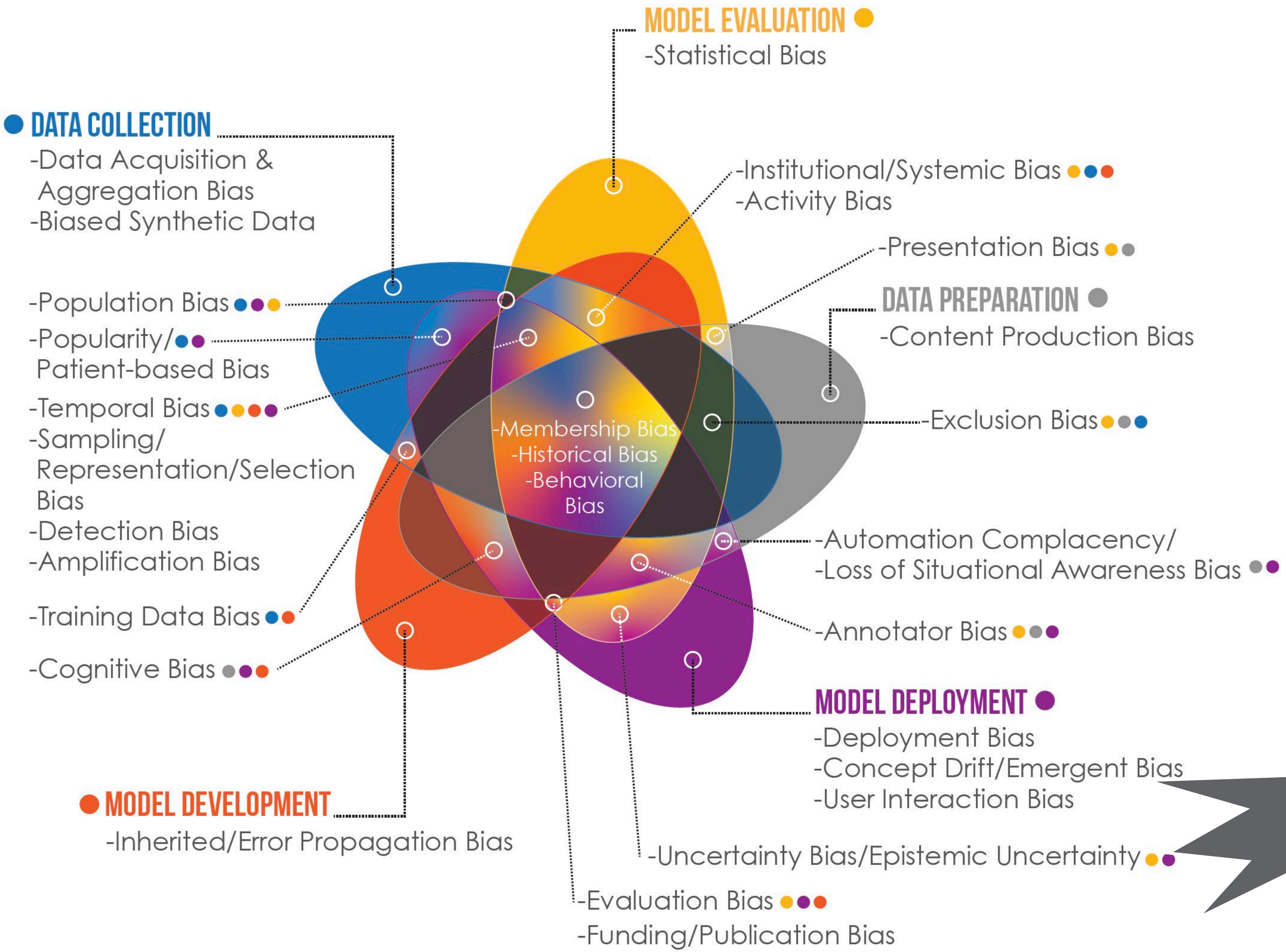
- **Teaming & Collaboration**
  - Multidisciplinary teams
  - Cross-Consortium collaboration
  - Community engagement committees
- **Diverse cohorts for data collection**
  - Consent & privacy
  - Legal issues
  - Sovereignty issues
- **AI/ML Training Needs**
  - Computational science training on the ethical, legal, and social implications
  - New material with use cases
  - Training for non-computational scientists (e.g., clinicians, physician scientists)
  - Hands-on training

Lessons Learned

- **Program vision & goals:** Promote repeatedly and continuously and consistently
- **Governance:** Create iterative governance structure to adapt to the changing needs
- **Iterative AI model build and evaluation:** As data and best practices are being created
- **Synchronized stakeholders:** Partner with each team from the outset, equitably
- **Sustainability plan:** For data storage, access, distribution, sovereignty from the outset

Other NIH Programs

Supporting trustworthy data for open science



MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER

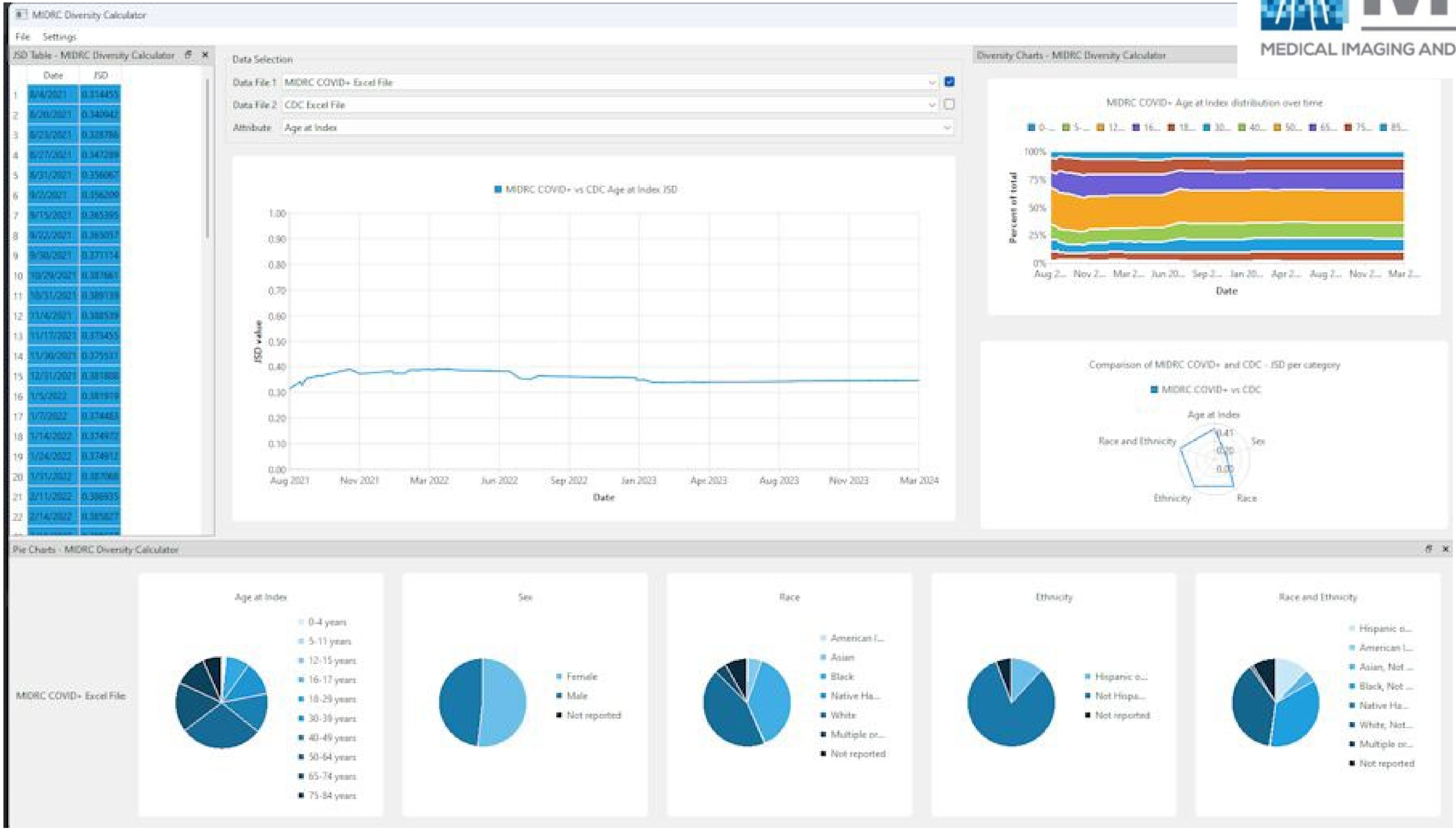
Major Bias Sources

- Data Collection
- Data Preparation
- Model Development
- Model Evaluation
- Model Deployment

Bias Calculator

Bias Awareness Tool:
https://www.midrc.org

Diversity Calculator



MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.

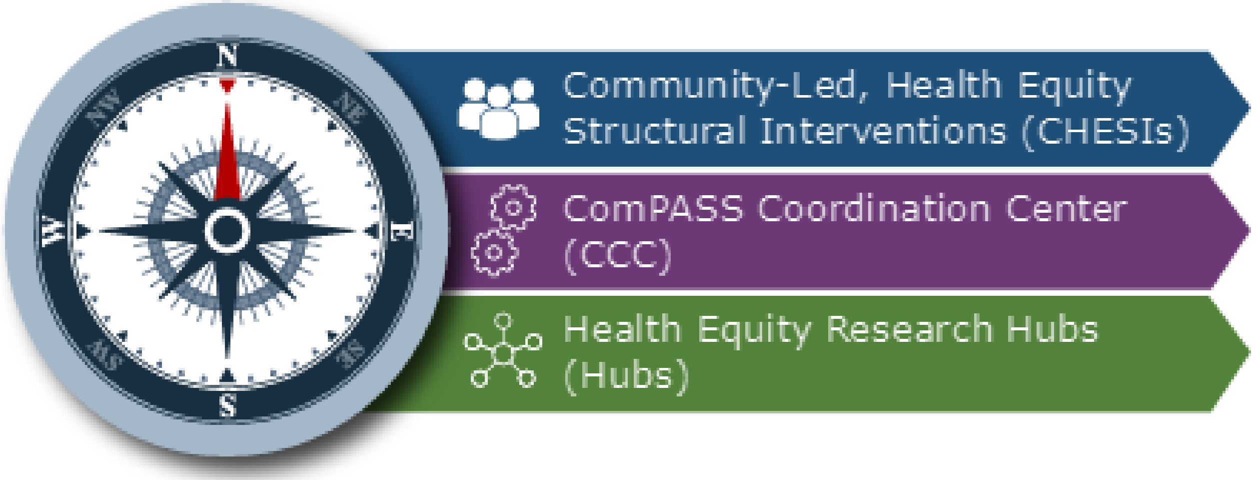
Community Partnerships to Advance Science for Society (ComPASS)

To advance the science of health disparities and health equity research, the National Institutes of Health (NIH) Common Fund launched the ComPASS Program.

The goals of ComPASS are to:

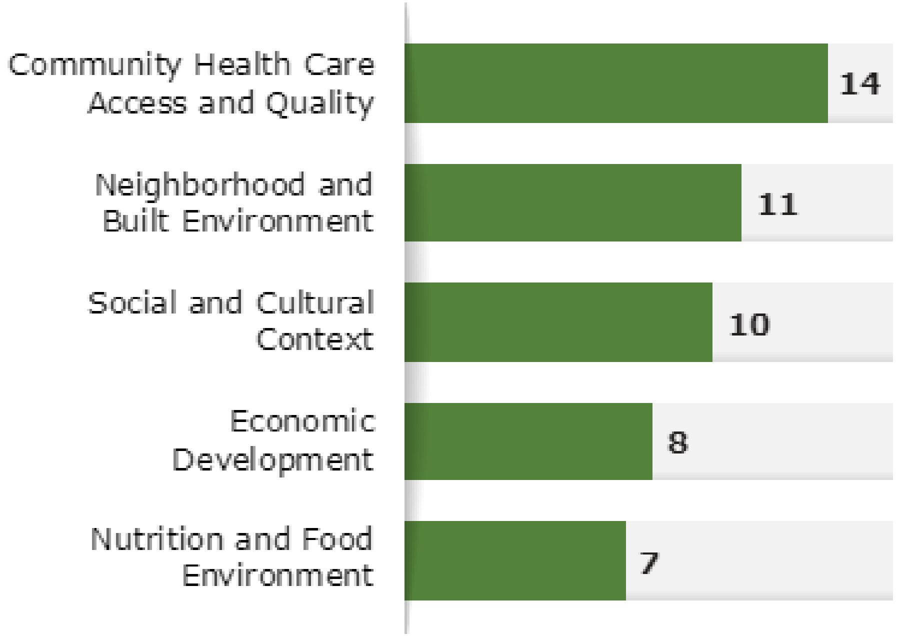
1. Study ways to reduce health disparities by addressing underlying structural factors within communities.
2. Develop a new research model for NIH where the projects are led by community organizations in collaboration with research partners.

ComPASS has three initiatives:

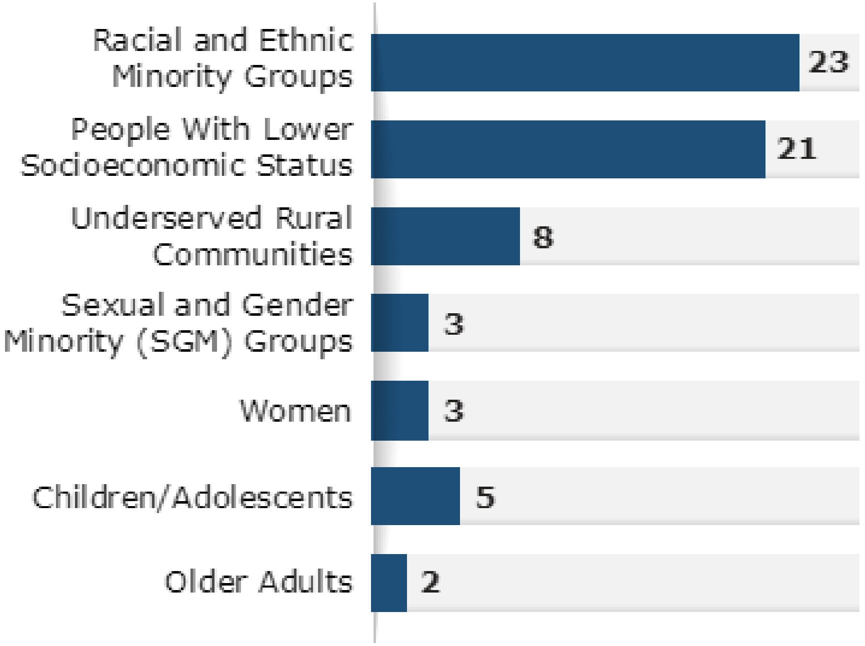


The 25 CHESI Structural Factors and Participant Populations

Social Determinants of Health and Structural Factors of the Projects



Populations That Experience Health Disparities and Other Participant Populations*



*Note that CHESI projects that focus on more than one social determinants of health and/or population experiencing health disparities are counted more than once.

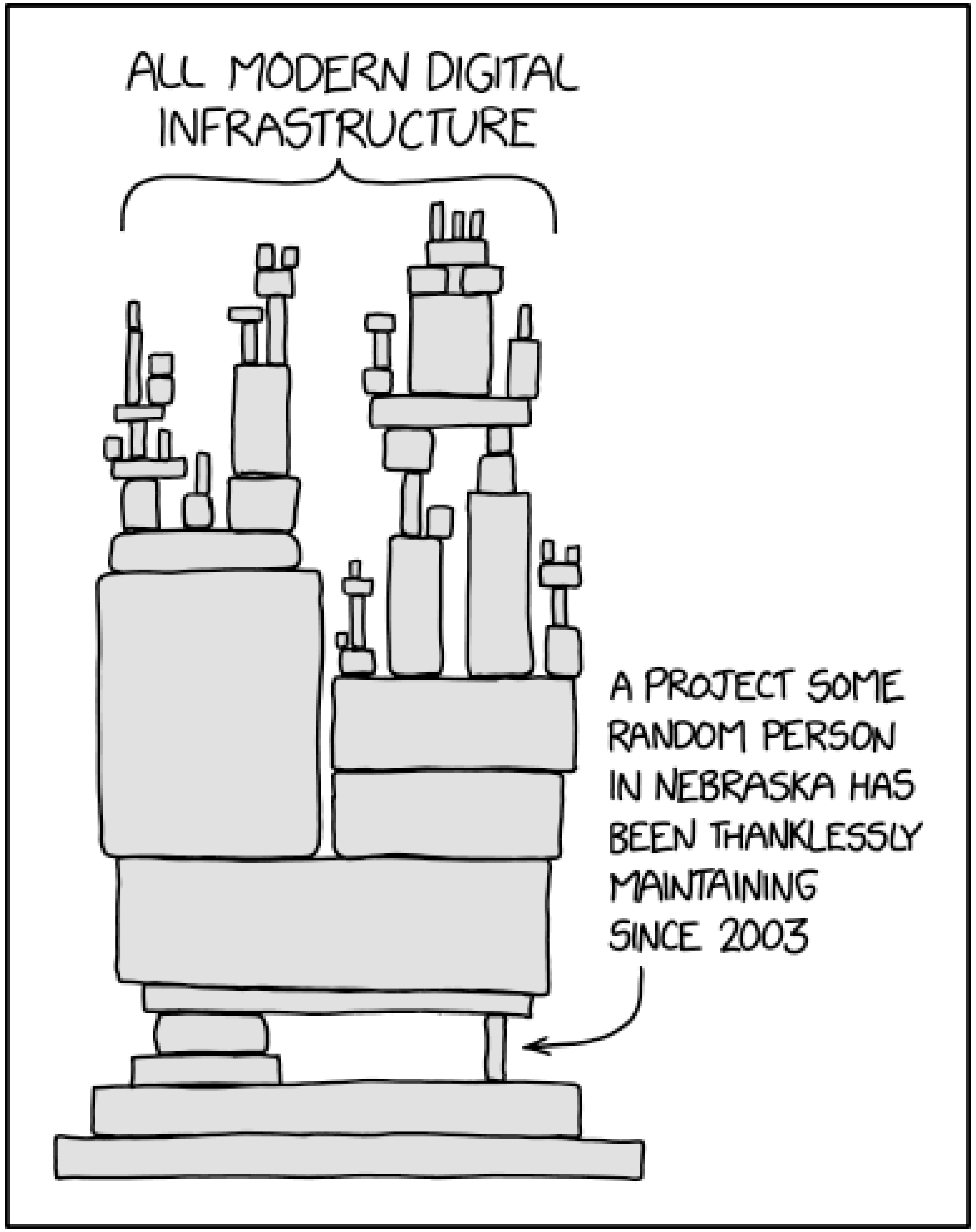
Connect With Us!

For more information, visit the NIH Common Fund ComPASS website at commonfund.nih.gov/compass.

Learn more by viewing the ComPASS Video Overview.

To receive ComPASS program announcements and information about funding opportunities, join the ComPASS listserv.

Trustworthy open data
→ requires understanding dependencies!
https://xkcd.com/2347/



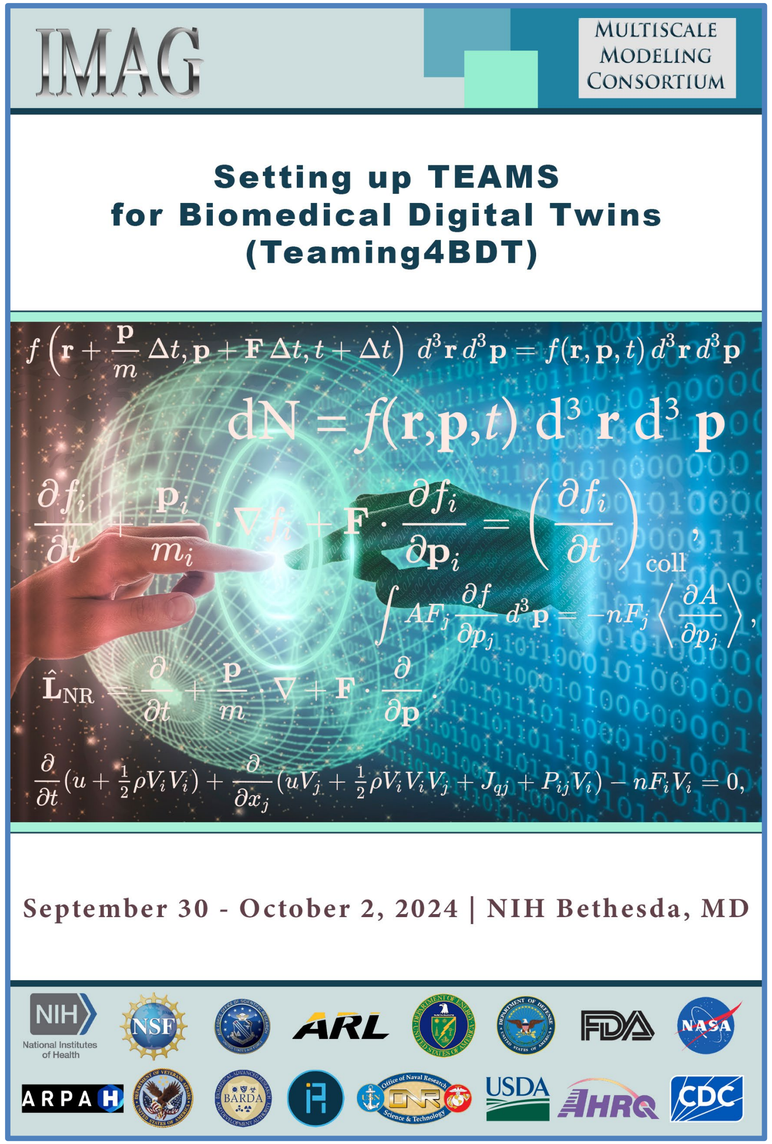
2024 IMAG MSM Consortium Meeting

Setting up TEAMS for Biomedical Digital Twins (Teaming4BDT)

- **September 30 - October 2, 2024**
- Bethesda, Maryland
- Register on the [IMAG WIKI]
- In-person and online – open to all!

Special thanks to NSF for providing Travel Awards

Special thanks to the Society for Mathematical Biology for providing refreshments



Day 1 - Defining Biomedical Digital Twins (BDT)

- Goal 1: To understand the NASEM Digital Twin components
- Goal 2: To identify unique features for digital twins in the biomedical domain (BDT)

Create requirements template for BDT

Day 2 - Approaches to address BDT challenges

- Goal 1: To understand the challenges unique to developing BDT
- Goal 2: To discuss needs with experts and compile BDT component resources

Create review template for BDT

Day 3 - Operationalizing Team Science for BDT

- Goal 1: To form BDT idea teams guided by team science approaches
- Goal 2: To present and review realizable, fit for purpose BDT ideas

Utilize consensus requirements and review templates developed in Day 1 and Day 2

Organized and hosted by the Interagency Modeling and Analysis Group (IMAG) and the Multiscale Modeling (MSM) Consortium