

Making Data AI-Ready

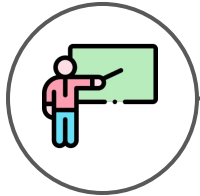
Bhavesh Patel, Ph.D.
Research Professor



About This Presentation



25 min + 5 min questions



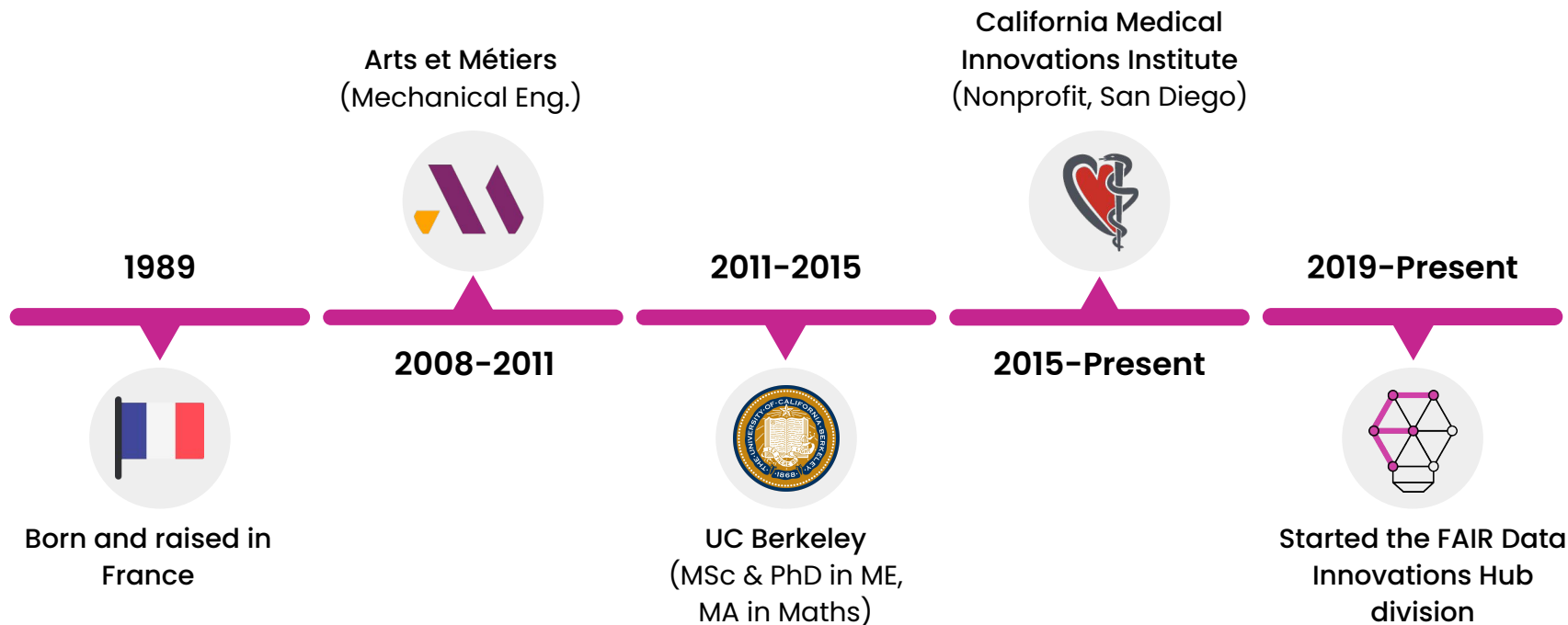
Why and how to make
data ready for AI



Background

Background

About me



Background

What we do at the FAIR Data Innovations Hub

We develop **software** and **guidelines** that help biomedical researchers **prepare and share AI-ready data**

All of our work is free and open source, supported by various organizations



The
Navigation
Fund



Background

The FAIR Data Innovations Hub Team



Bhavesh Patel
Lead/PI



Sanjay Soundarajan
Software Developer



Christopher Marroquin
Software Developer



Xuebin Dong
Software Developer



Jacob Clark
Software Developer



Dorian Portillo
Software Developer



Aydan Gasimova
Software Developer



Nahid Zeinali
AI Research Scientist

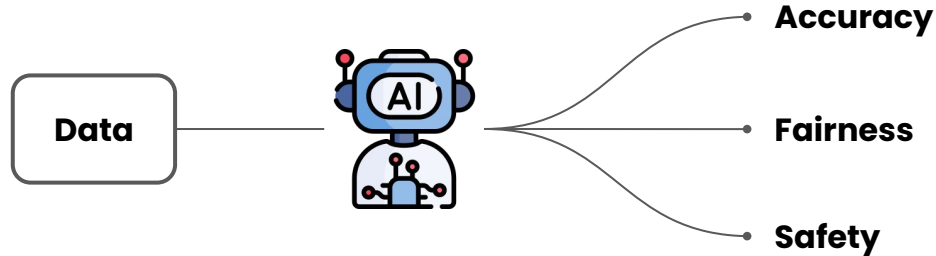


AI-Ready Data

AI-Ready Data

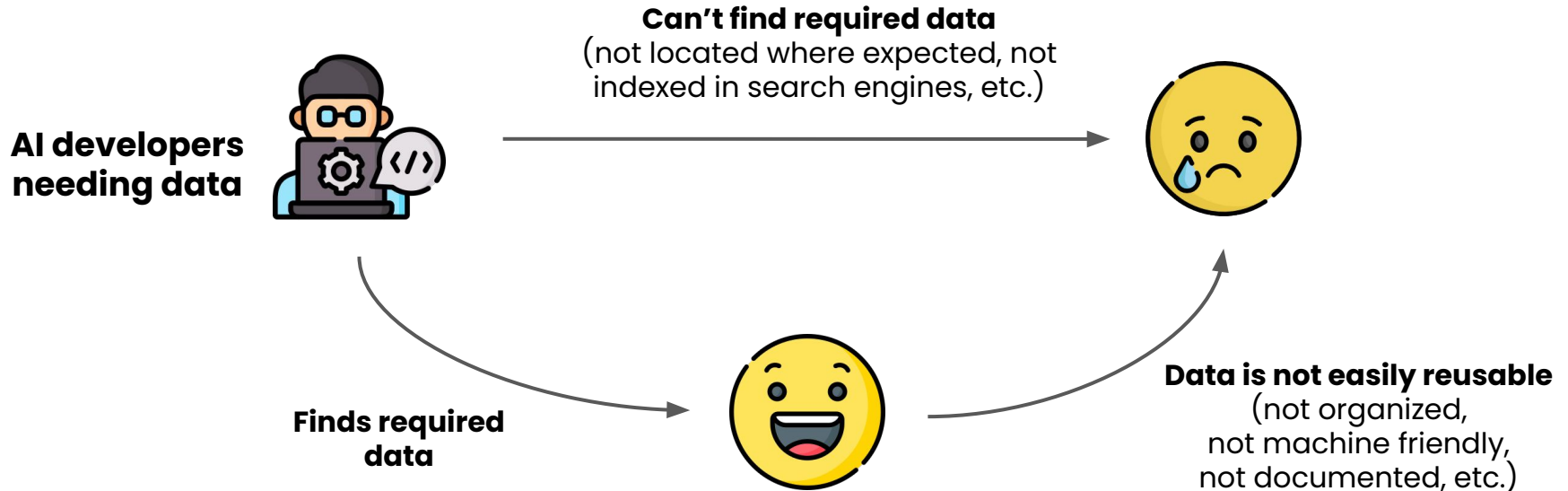
Data is the fuel for AI

Data **enables the development** of AI models
and directly **affects their characteristics**



AI-Ready Data

But often data is not AI-Ready



AI-Ready Data

What if?



**AI developers
needing data**

**Finds required data and
data is easy to reuse**



NIH Bridge2AI



NIH Bridge2AI

About Bridge2AI

Bridge2AI is a new **NIH Common Fund Program** (2022)



Goal: “propel biomedical research forward by setting the stage for widespread adoption of artificial intelligence (AI)”

NIH Bridge2AI

Funding

Four data generation projects

\$130M over 4 years



- 1 Collect and share new human data around a major disease
- 2 Develop a blueprint for preparing and sharing AI-ready data



AI-READI



AI-READI

About AI-READI

**AI-READI: Artificial Intelligence Ready and
Exploratory Atlas for Diabetes Insights**



Collect a multimodal dataset for studying
Type 2 Diabetes and make it AI-ready

AI-READI Team



Joseph Yracheta
Native BioData



Bhavesh Patel
CALMI



Linda Zangwill
UCSD



Sally Baxter
UCSD



Nick Evans
University of
Massachusetts



Sara Singer
Stanford



Amir Bahmani
Stanford



Samm Hurst
UCSD



Jorge Contreras
University of Utah



Christopher Chute
JHU



Michelle Hribar
OHSU



Cecilia Lee
UW



Aaron Lee
UW



Cynthia Owsley
UAB



Gerald McGwin
UAB



Shannon McWeeney
OHSU

Developers

Interns

Clinical Research Coordinators

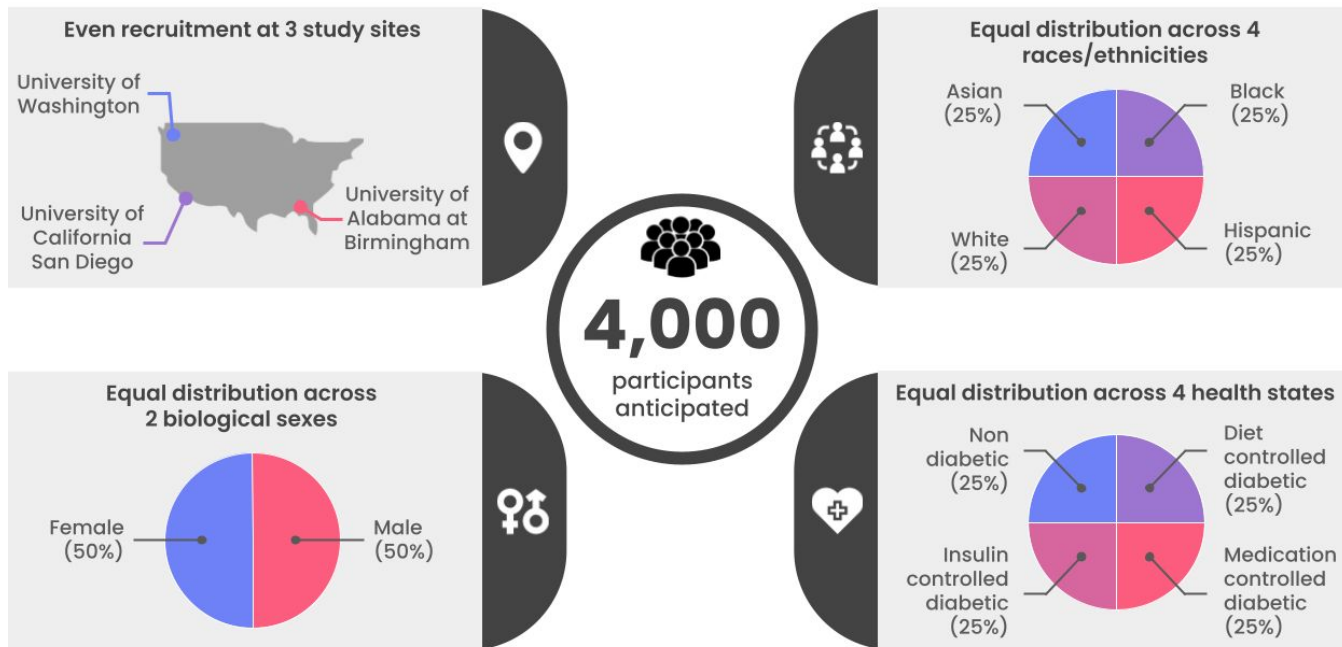


Program and Project Managers

<https://aireadi.org/team>

AI-READI

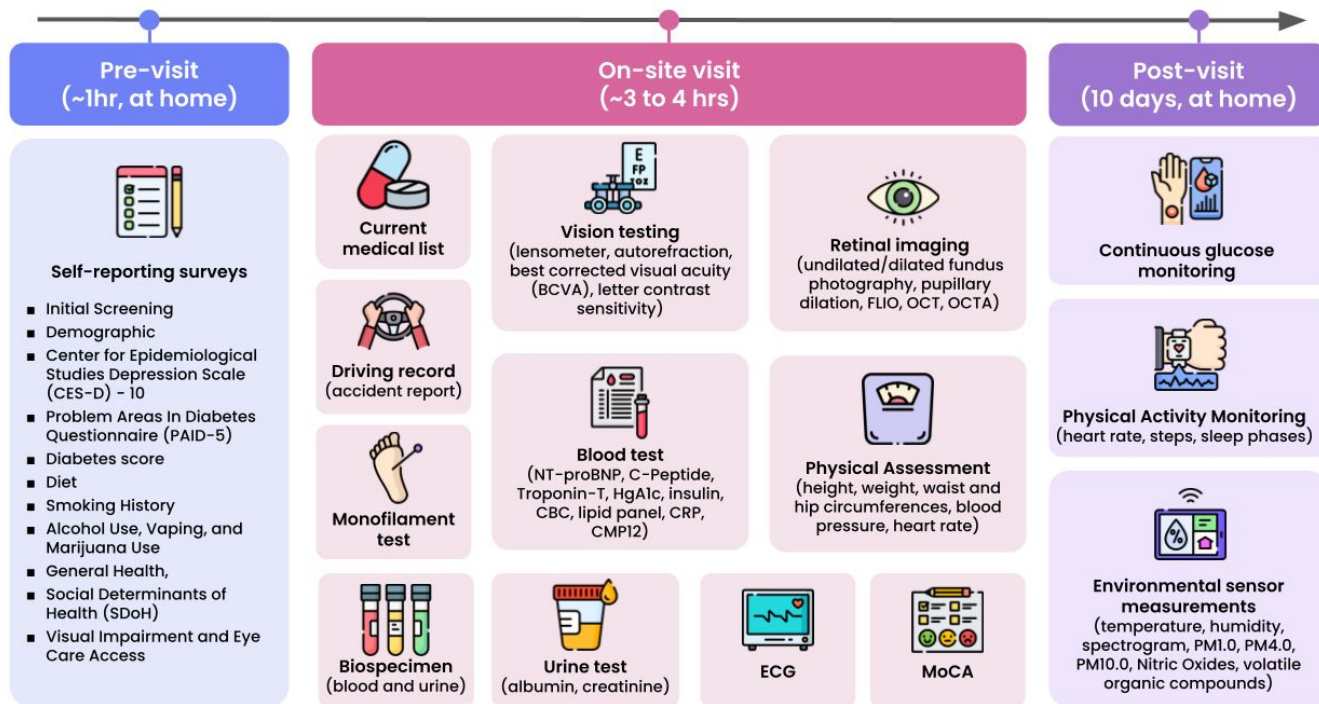
Study design



Note: the study may additionally include a cohort of participants from Native American communities but is contingent on finding a suitable agreement between representatives of Native American communities and the NIH

AI-READI

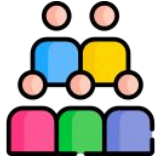
Data collection protocol



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

AI-READI

Current dataset (version 2 released in November 2024)



Data from 1067
participants



165,000+ files



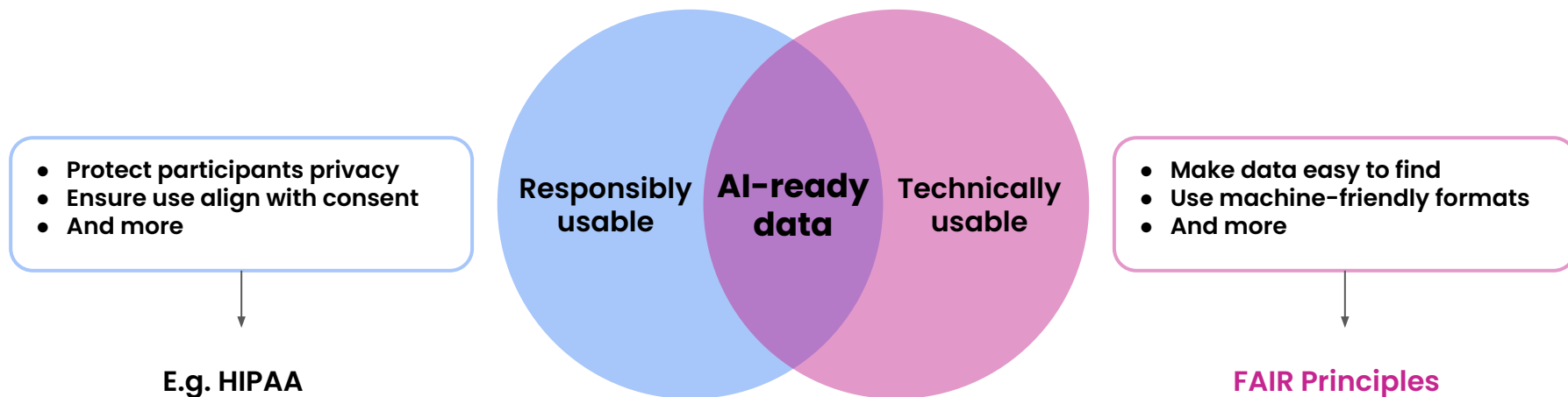
2 TB of data total



600+ downloads

AI-Ready

How to make data AI-ready?





FAIR Principles

FAIR Principles

Origin

How to make all research outcomes, including data,
optimally reusable by humans and machines?



**Findable, Accessible, Interoperable, and Reusable (FAIR)
Principles (2016)**

FAIR Principles

15 principles to optimize data reuse for humans and machines

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

FAIR Principles

Adoption

The G20 logo consists of the letters "G20" in a bold, black, sans-serif font, centered within a light blue circle. This circle is itself centered within a larger, light gray rounded square.

Leaders at the 2016 G20 meeting released a joint press release expressing their intention to support implementation of FAIR principles in publicly funded research



"Turning FAIR into Reality" report (2018)

Research data not being FAIR cost the EU economy at least €10bn/year



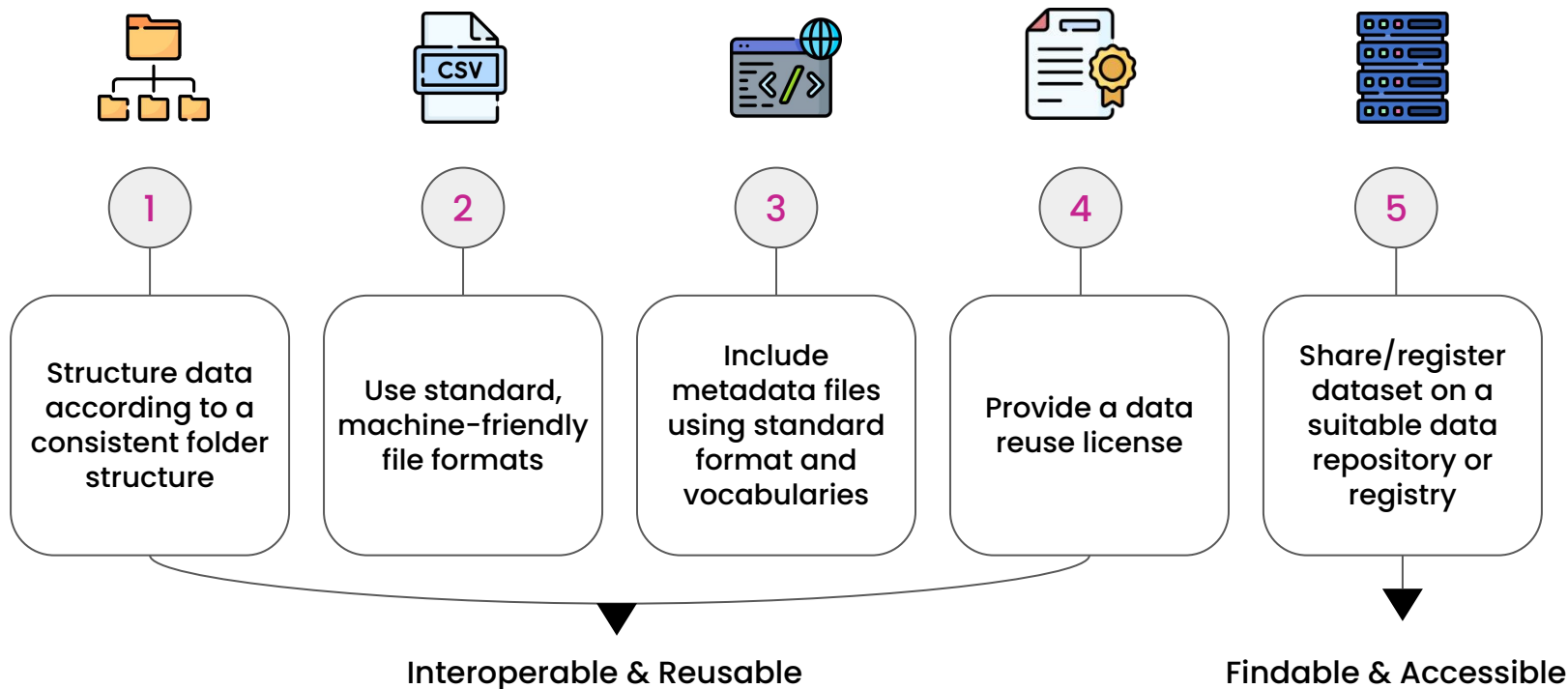
National Institutes of Health

New data sharing policy (January 2023) requires all grant proposals to include a Data Management Plan that describes how data will be made FAIR

FAIR Principles paper has close to 18k citations

FAIR Principles

How to practically make data FAIR





FAIR AI-READI Dataset

FAIR AI-READI Dataset

FAIR: 1. Consistent folder structure

Clinical Dataset Structure (CDS)

cds-specification.readthedocs.io



- cardiac_ecg
- clinical_data
- environment
- retinal_flio
- retinal_oct
- retinal_octa
- retinal_photography
- wearable_activity_monitor
- wearable_blood_glucose
- CHANGELOG.md
- LICENSE.txt
- README.md
- dataset_description.json
- dataset_structure_description.json
- healthsheet.md
- participants.json
- participants.tsv
- study_description.json

One folder per
datatype

Metadata files

FAIR AI-READI Dataset

FAIR: 2. Standard data format

Data category	Data format followed
Clinical data	Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)
Retinal imaging data	Digital Imaging and Communications in Medicine (DICOM)
ECG data (standard 12-lead)	WaveForm DataBase (WFDB)
Wearables data (physical activity monitoring, continuous glucose monitoring data)	Open mHealth
Environmental sensor data	Earth Science Data Systems (ESDS) format

FAIR AI-READI Dataset

FAIR: 3. Extensive metadata

retinal_photography

wearable_activity_monitor

- CHANGELOG.md
- dataset_description.json
- dataset_structure_description.json
- healthsheet.md
- LICENSE.txt
- participants.json
- participants.tsv
- README.md
- study_description.json

Broad metadata: study background, dataset provenance, dataset structure, license terms, participants information

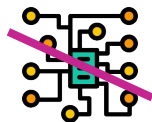
Provided in both human and machine-friendly formats

FAIR AI-READI Dataset

FAIR: 4. Shared under a new usage license



Prohibits re-identification of study participants or harm



Prohibits distribution of the data and of models that may “copy” or “memorize” the data



Allows commercial use

FAIR AI-READI Dataset

FAIR: 5. Shared on FAIRhub

The screenshot shows the FAIRhub website interface for the 'Flagship Dataset of Type 2 Diabetes from the AI-READI Project'. The page includes a navigation bar with links to 'Find Datasets', 'Share datasets', 'About', and 'Contact'. Below the title, there are buttons for 'Access this dataset' and 'View the dataset documentation'. A circular logo for the AI-READI Consortium is visible. The main content area features a tabbed interface with 'About' selected, showing an 'Info' section and an 'Overview of the study'. The 'Overview' text describes the project's goal to create a foundational dataset for T2DM research. On the right, a 'Usage statistics' box displays 6809 views, 0 citations, and 258 access approvals. Below this, a box shows '2.01 TB' and '165,051 Files'. A 'License' section indicates 'Health Data License'. A 'Keywords' section lists terms like 'Diabetes mellitus', 'Machine Learning', 'Artificial Intelligence', 'Electrocardiography', 'Continuous Glucose Monitoring', 'Retinal imaging', and 'Eye exam'.

FAIRhub

Find Datasets Share datasets About Contact

Flagship Dataset of Type 2 Diabetes from the AI-READI Project

AI-READI Consortium

[Access this dataset](#) [View the dataset documentation](#)

[About](#) [Healthsheet](#) [Study Dashboard](#) [Study Metadata](#) [Dataset Metadata](#) [Dataset Structure Preview](#) [Dataset Quality Dashboard](#) [Dataset Uses](#)

Info
This page provides an overview of the dataset and associated study.

Overview of the study

The Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READI) project seeks to create a flagship ethically-sourced dataset to enable future generations of artificial intelligence/machine learning (AI/ML) research to provide critical insights into type 2 diabetes mellitus (T2DM), including salutogenic pathways to return to health. The ability to understand and affect the course of complex, multi-organ diseases such as T2DM has been limited by a lack of well-designed, high quality, large, and inclusive multimodal datasets. The AI-READI team of investigators will aim to collect a cross-sectional dataset of 4,000 people and longitudinal data from 10% of the study cohort across the US. The study cohort will be balanced for self-reported race/ethnicity, gender, and diabetes disease stage. Data collection will be specifically designed to permit downstream pseudo-time manifold analysis, an approach used to predict disease trajectories by collecting and learning from complex, multimodal data from participants with differing disease severity (normal to insulin-dependent T2DM). The long-term objective for this project is to develop a foundational dataset in T2DM, agnostic to existing classification criteria or biases, which can be used to reconstruct a temporal atlas of T2DM development and reversal towards health (i.e., salutogenesis). Data will be optimized for downstream AI/ML research and made publicly available. This project will also create a roadmap for ethical and equitable research that focuses on the diversity of the research participants and the workforce involved at all stages of the research process (study design and data collection, analysis, synthesis, and dissemination).

Usage statistics

6809 Views 0 Cited by 258 Access approved

All versions Current version

2.01 TB 165,051 Files

License

Health Data License

Keywords

Diabetes mellitus Machine Learning Artificial Intelligence Electrocardiography Continuous Glucose Monitoring Retinal imaging Eye exam

DOI

Accessible metadata

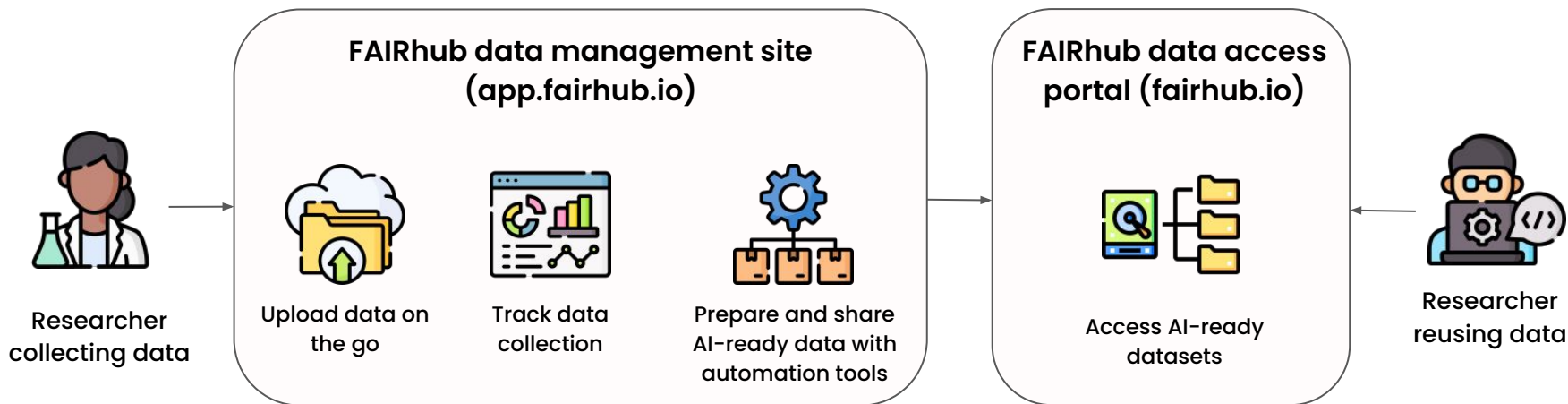
Indexed metadata

Clear access process

<https://doi.org/10.60775/fairhub.2>

FAIR AI-READI Dataset

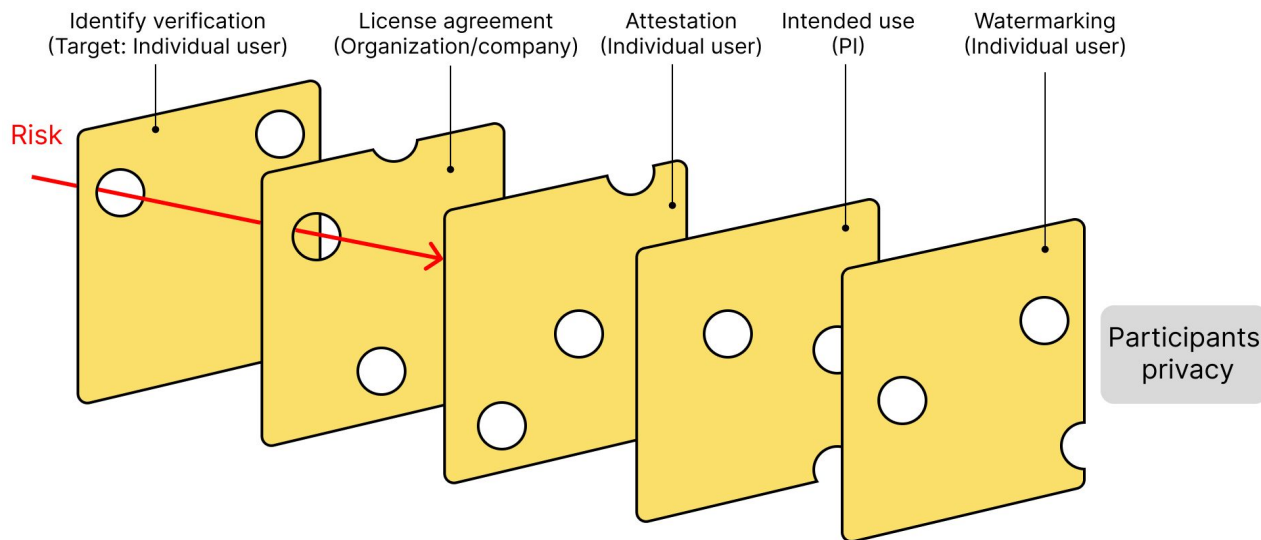
FAIRhub



FAIR AI-READI Dataset

New open data access model

Swiss-cheese model of open data sharing: Multiple steps each designed to help minimize risks to participants privacy and prevent misuse





Closing Remarks

Closing Remarks

Making data FAIR is not always easy...

> 1,000 standards

> 1,600 databases

> 100 policies

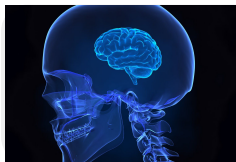


("Bio" related results on FAIRsharing.org)

Closing Remarks

... but you can get started today with simple steps!

Kids



Share/register dataset on a suitable repository or registry

[NIH list of data repositories](#)

Adults



Provide as much metadata as possible

For example: authors, funding source, keywords, devices used, etc.

Legends

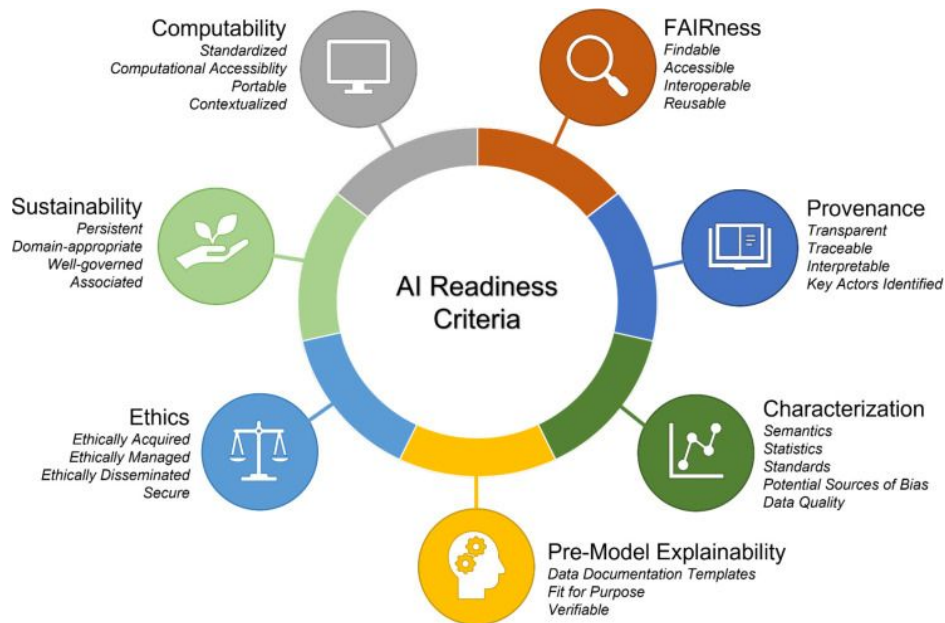


Structure (meta)data following standards

[fairsharing.org](#)

Closing Remarks

Read the Bridge2AI recommendations for AI-ready data

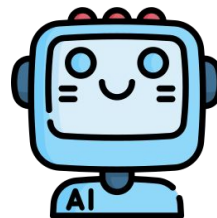
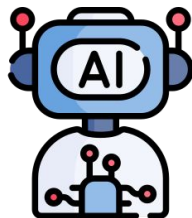


Closing Remarks

Keep an eye on our work!

Vision: **Let AI make data ready for AI**

I got your
data ready



Thanks
buddy!

Thank You!



bpatel@calmi2.org



fairdataihub.org

*Find these slides and
all resources here*



bit.ly/aiready-data