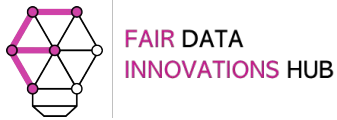
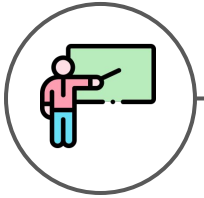
Making Data AI-Ready

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

1989

Born and raised in France

2008-2011

Arts et Métiers
(Mechanical Eng.)

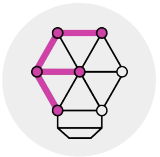
2011-2015

UC Berkeley
(MSc & PhD in ME,
MA in Maths)

2015-Present

California Medical
Innovations Institute
(Nonprofit, San Diego)

2019-Present



Started the FAIR Data Innovations Hub division

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

NIH

The Navigation Fund

Microsoft AI for Good Lab

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

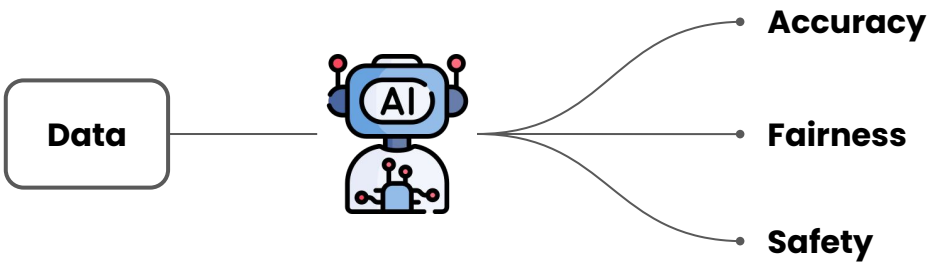
https://fairdataihub.org/team

AI-Ready Data

AI-Ready Data

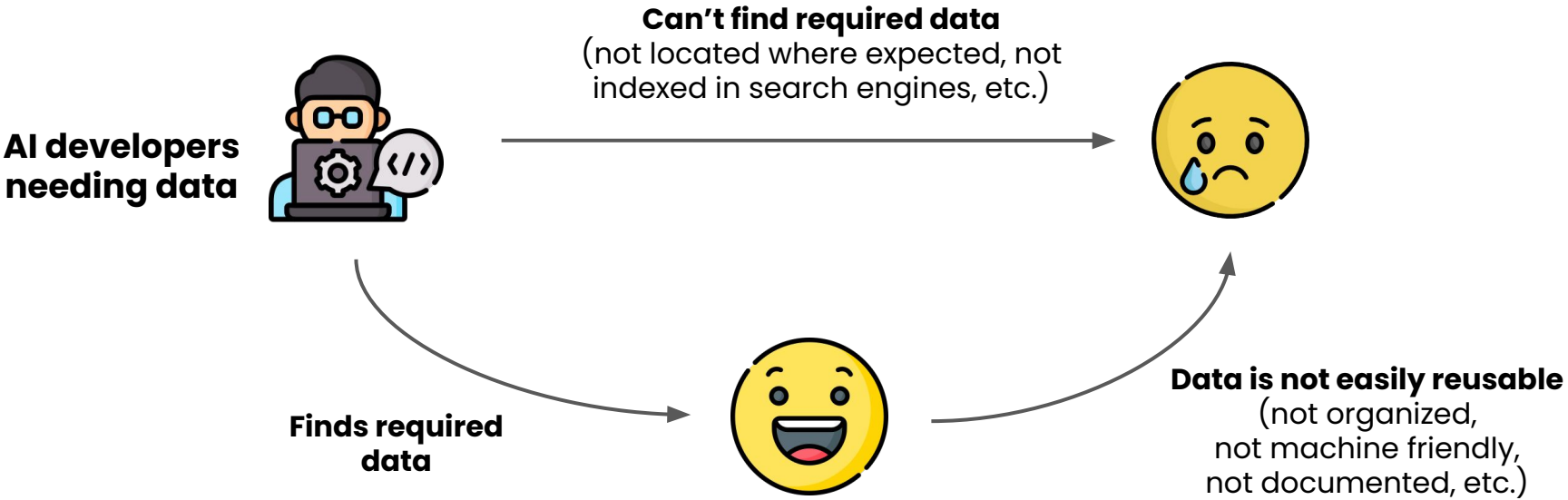
Data is the fuel for AI

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 date 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)”

https://commonfund.nih.gov/bridge2ai

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

https://commonfund.nih.gov/bridge2ai

AI-READI


AI-READI
AI Ready and Exploratory
Atlas for Diabetes Insights

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

https://aireadi.org

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





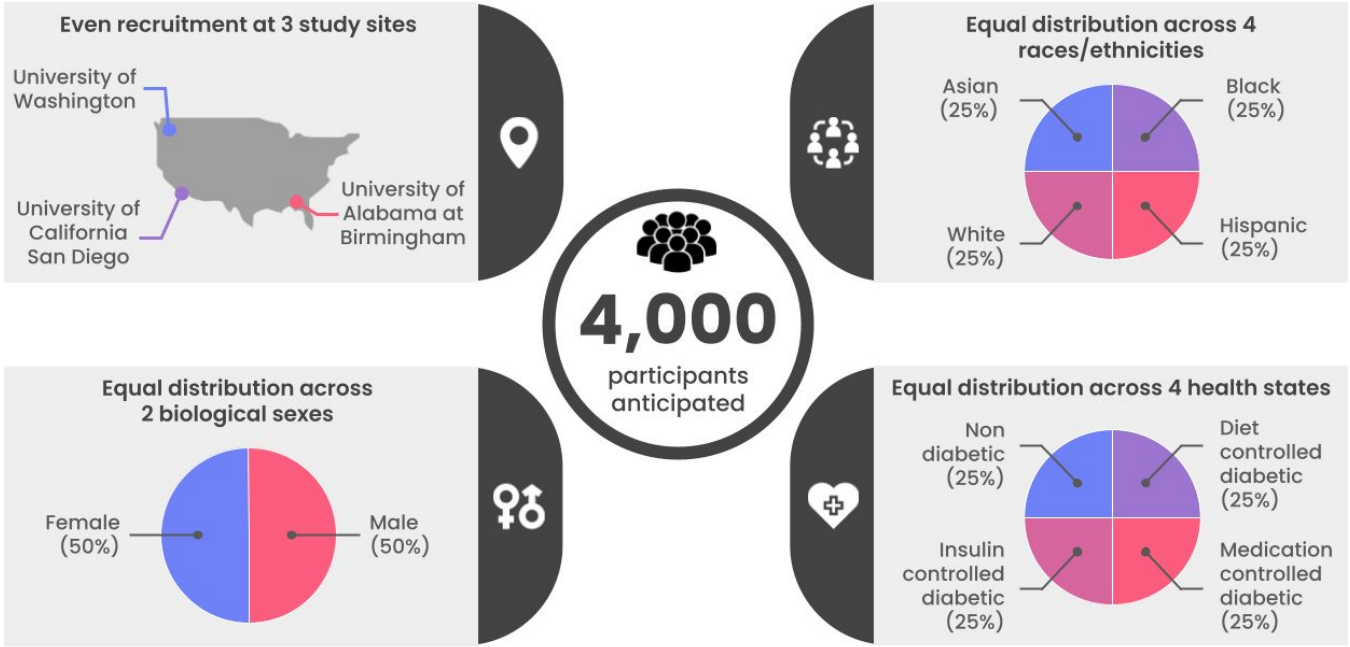
Program and Project Managers

Clinical Research Coordinators

https://aireadi.org/team

AI-READI

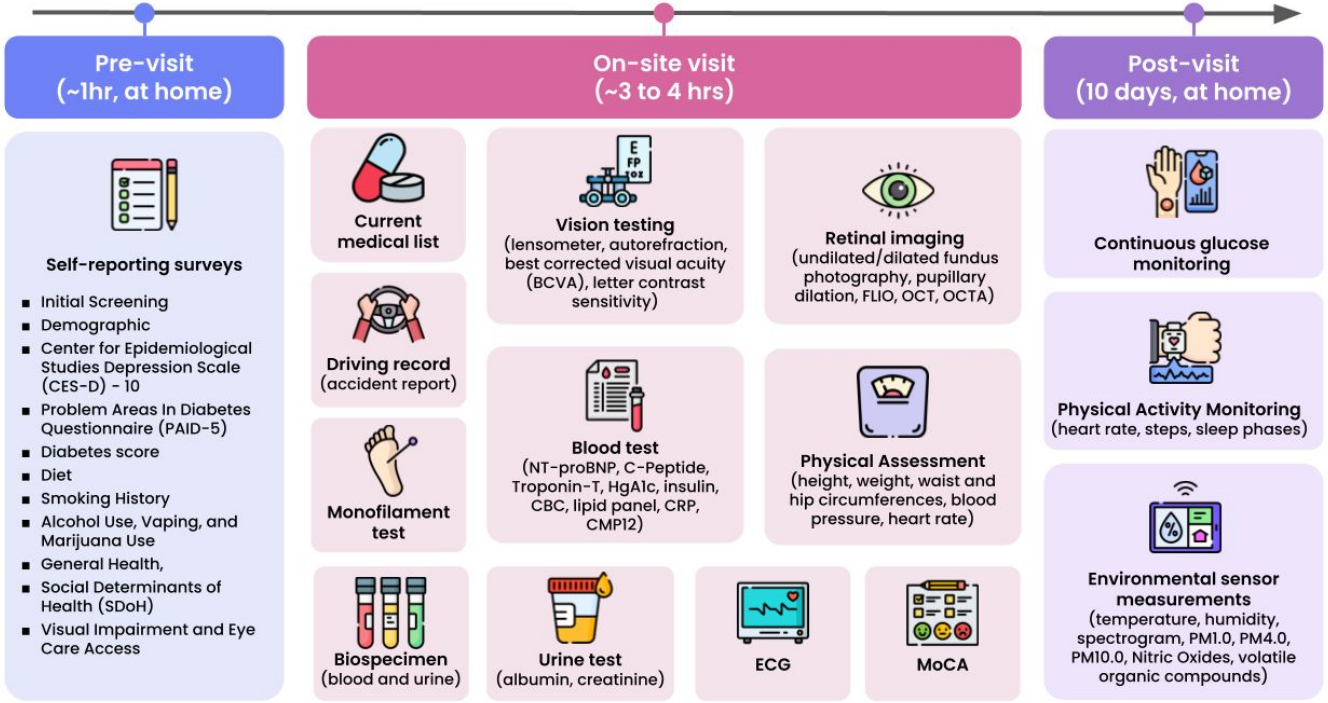
Study design



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

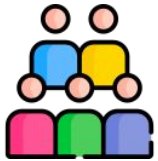
Data collection protocol



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

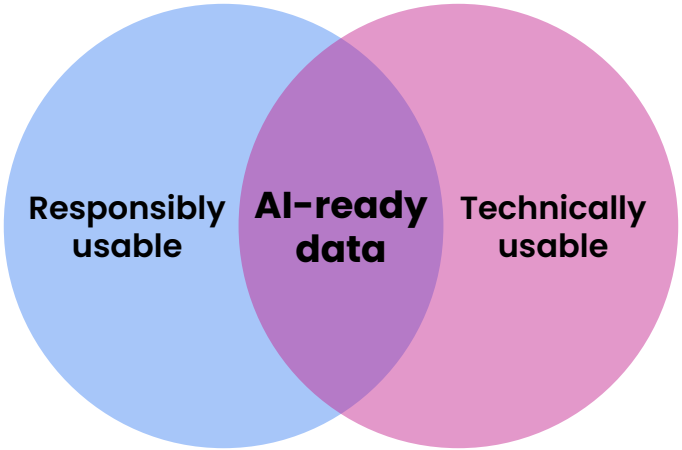
https://fairhub.io/datasets/2

AI-Ready

How to make data AI-ready?

- Protect participants privacy
- Ensure use align with consent
- And more

E.g. HIPAA



- Make data easy to find
- Use machine-friendly formats
- And more

FAIR Principles

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)

https://doi.org/10.1038/sdata.2016.18

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

https://doi.org/10.1038/sdata.2016.18

FAIR Principles
Adoption

G20

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

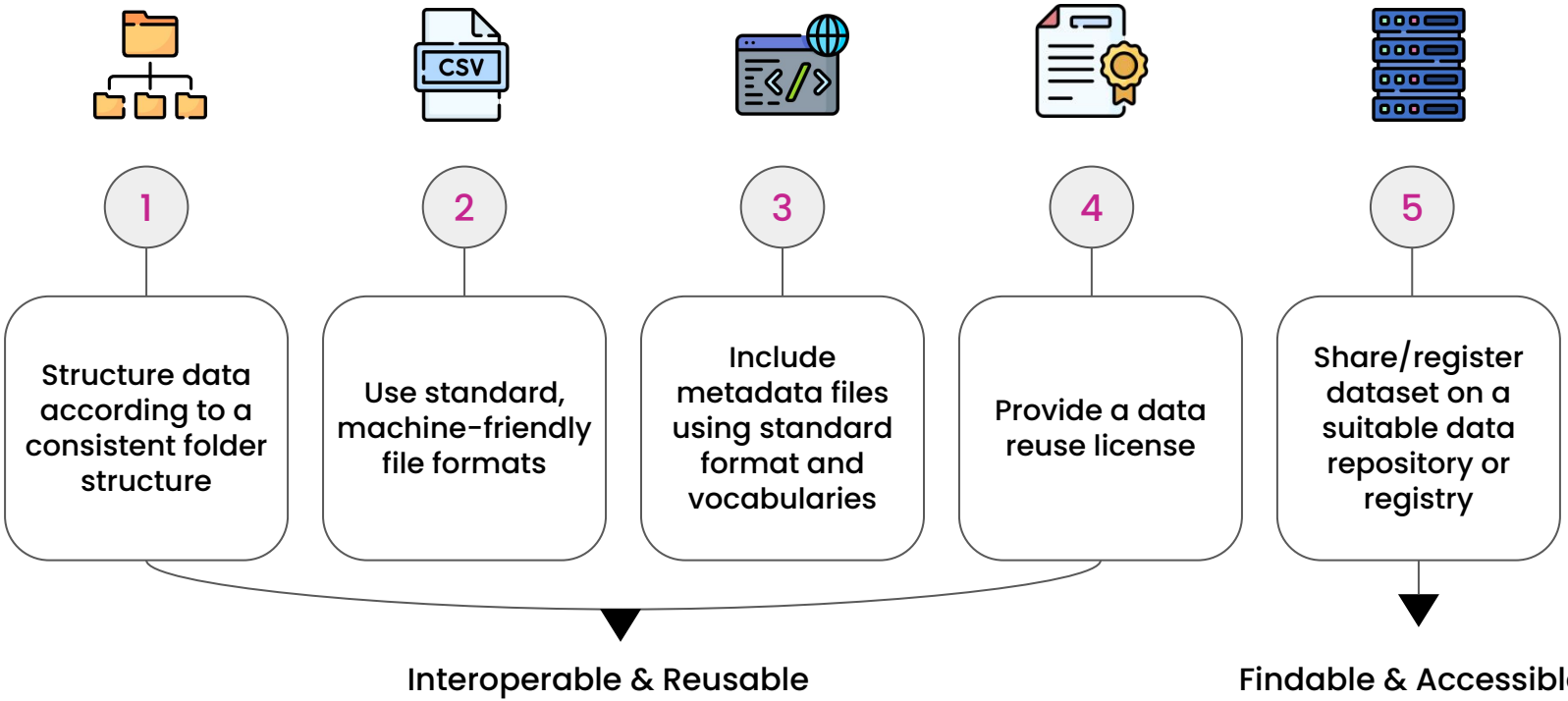
NIH
National Institutes of Health

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



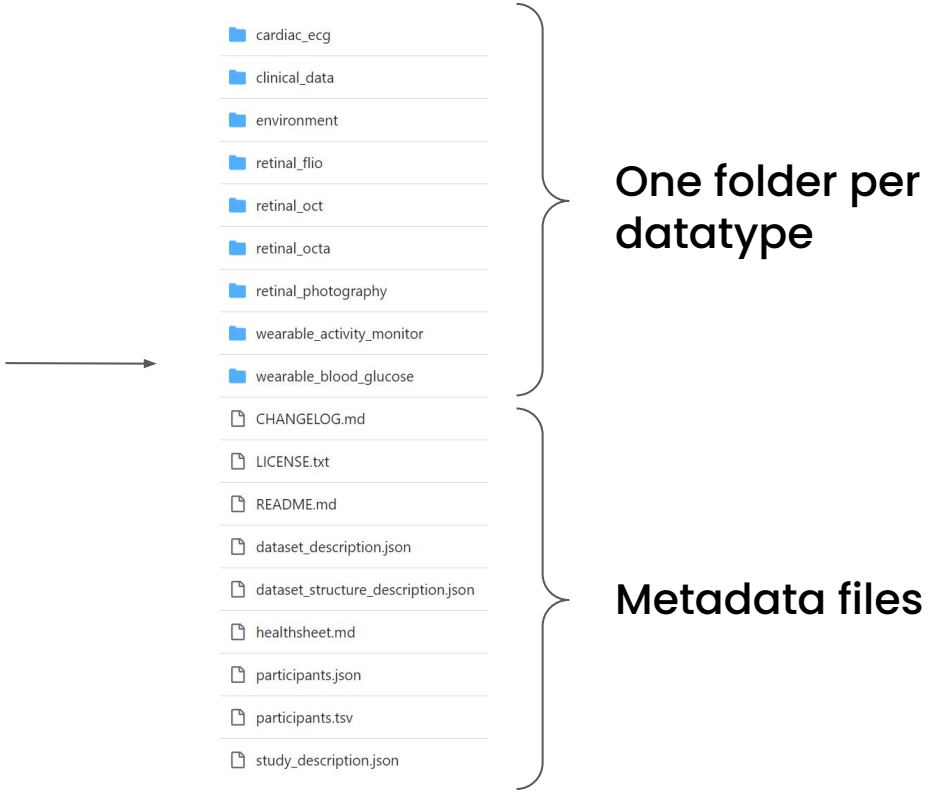
FAIR AI-READI Dataset

FAIR AI-READI Dataset

FAIR: 1. Consistent folder structure

Clinical Dataset Structure (CDS)

cds-specification.readthedocs.io



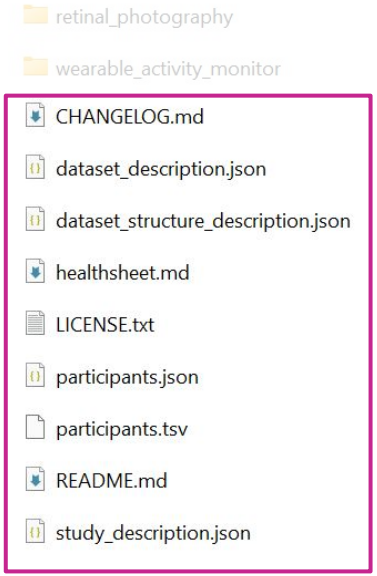
FAIR AI-READI Dataset

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



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

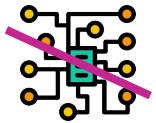
Provided in both human and machine-friendly formats

https://cds-specification.readthedocs.io

FAIR AI-READI Dataset

FAIR: 4. Shared under a new usage license

Prohibits re-identification of study participants or harm



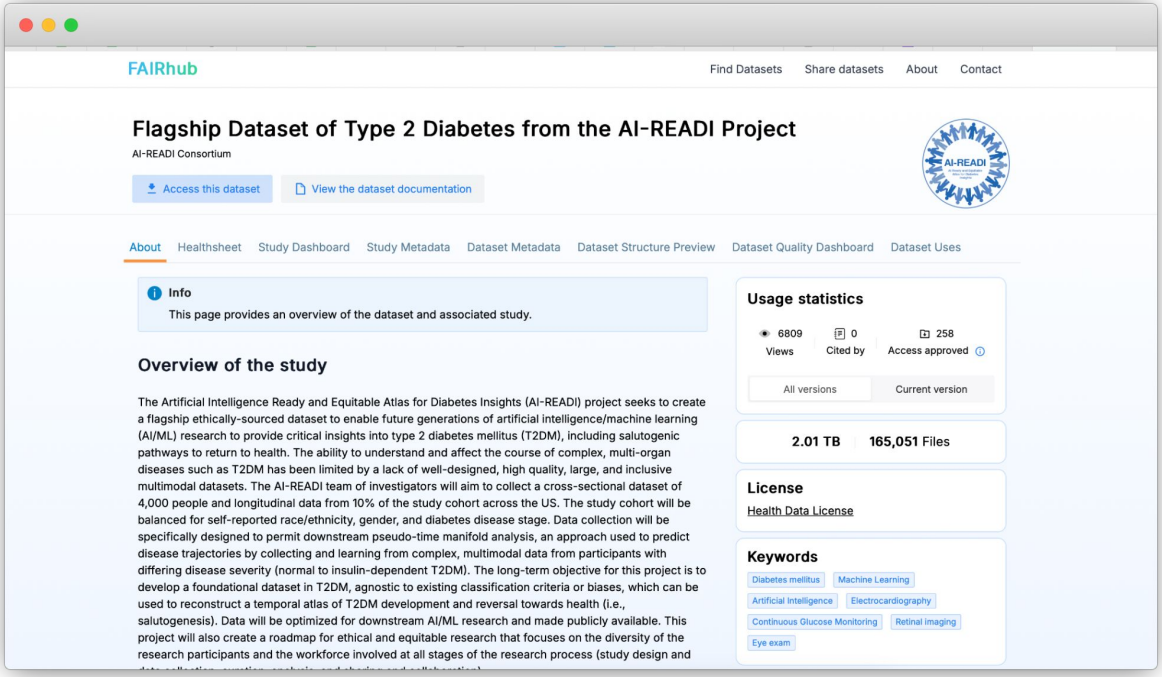
Prohibits distribution of the data and of models that may “copy” or “memorize” the data

Allows commercial use

https://doi.org/10.5281/zenodo.10642459

FAIR AI-READI Dataset

FAIR: 5. Shared on FAIRhub

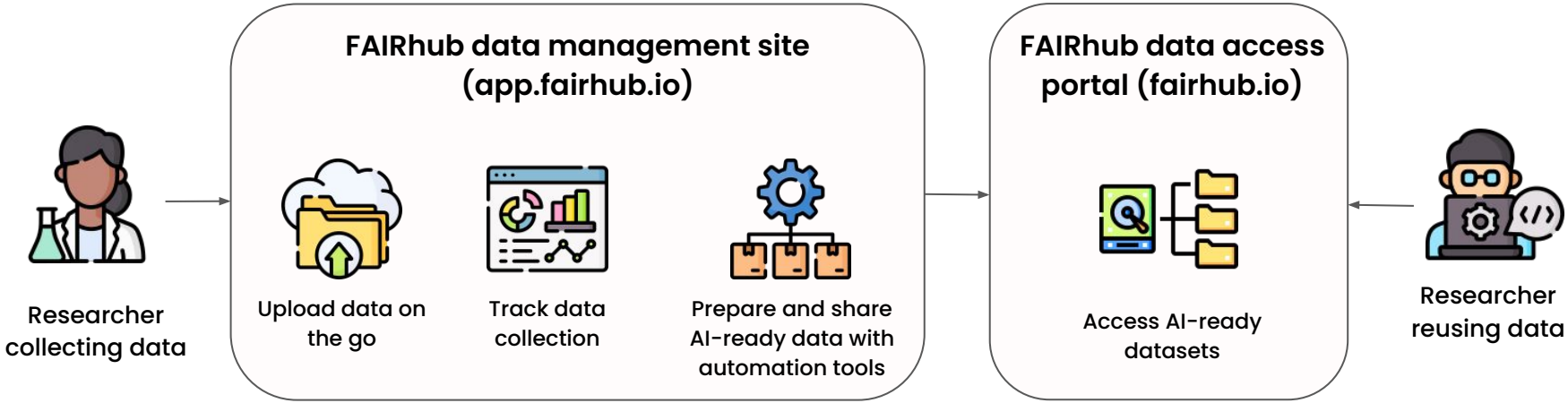


- DOI
- Accessible metadata
- Indexed metadata
- Clear access process

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

FAIR AI-READI Dataset

FAIRhub

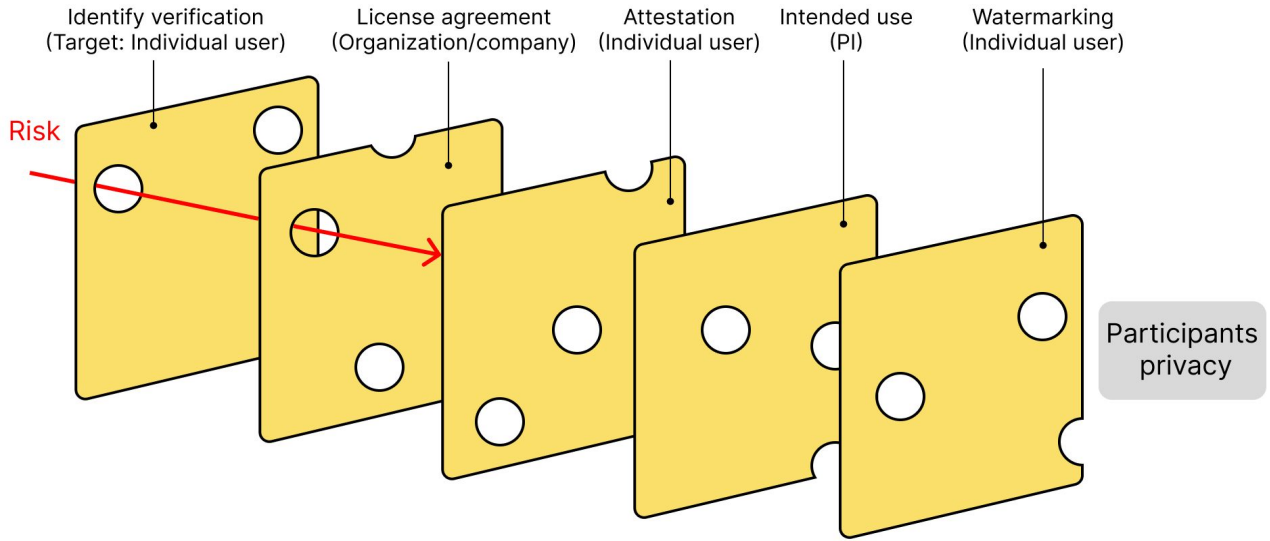


https://doi.org/10.60775/fairhub.2

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)

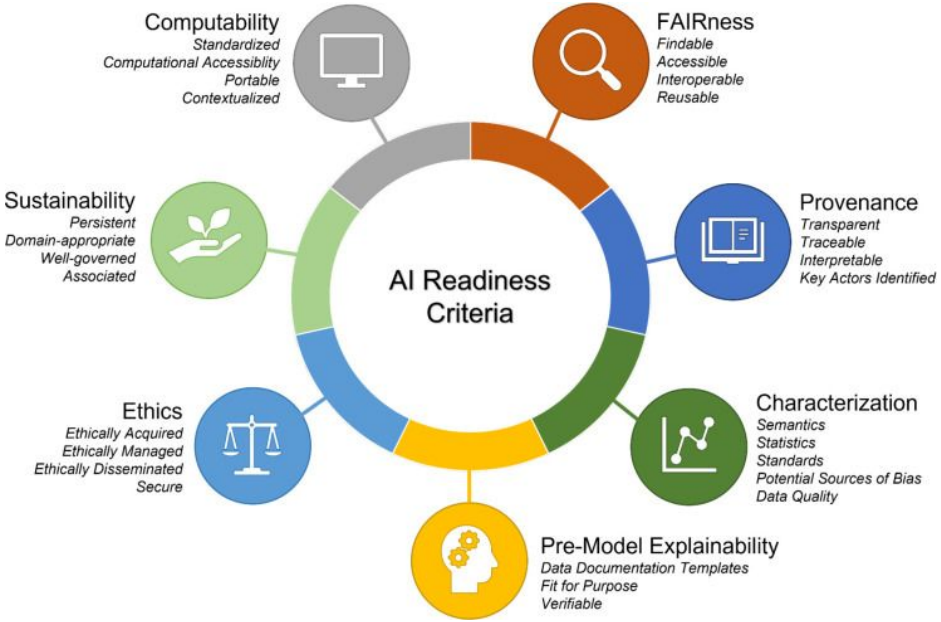
Closing Remarks

… 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

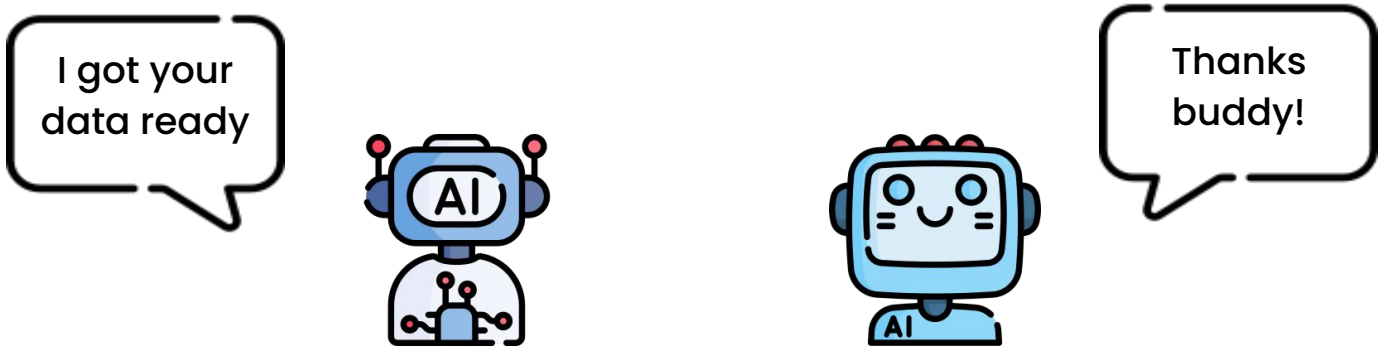


https://doi.org/10.1101/2024.10.23.619844

Closing Remarks

Keep an eye on our work!

Vision: Let AI make data ready for AI



https://fairdataihub.org

Thank You!

bpatel@calmi2.org

fairdataihub.org

Find these slides and all resources here

bit.ly/aiready-data