ScHARe

Think-a-Thons

NIH National Institutes of Health

Preparing for AI-driven Research on ScHARe - Part 1

A Comprehensive Review and Brainstorming Session

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

February 21, 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	□	□	□	□

Outline

5’ Introduction

- Experience poll

10’ ScHARe overview

- Interest poll

1h10’ Documenting research

- Polls

10’ Selecting the data

25’ Common Data Elements

25’ Making datasets AI-ready

5’ Join our Research Think-a-Thons

- Final poll

Next time

- Selecting computational strategies
- Algorithm testing and implementation
- Publishing research
- Research Think-a-Thons: brainstorming projects

ScHARe

Overview

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

ScHARe

nimhd.nih.gov/schare

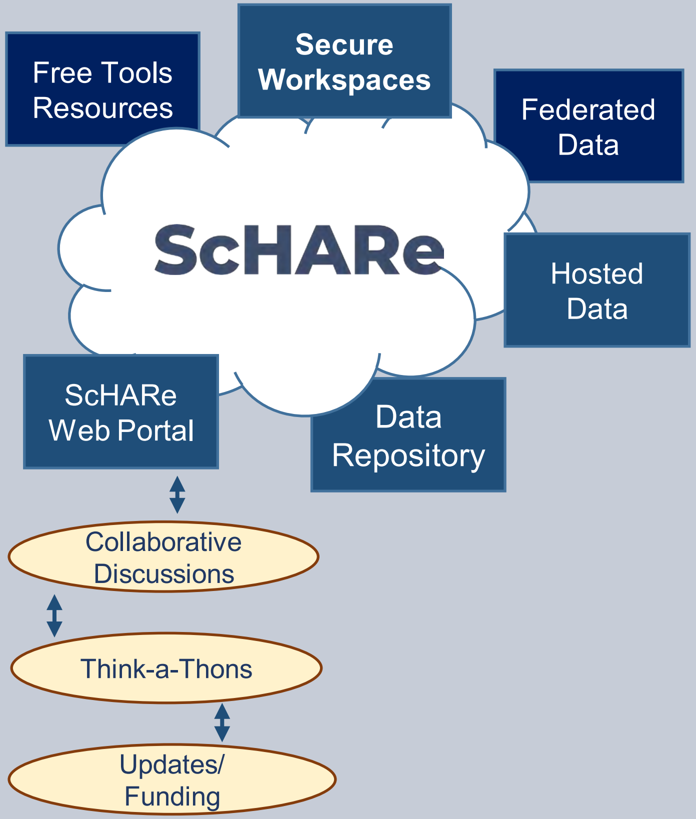
ScHARe Components

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)*

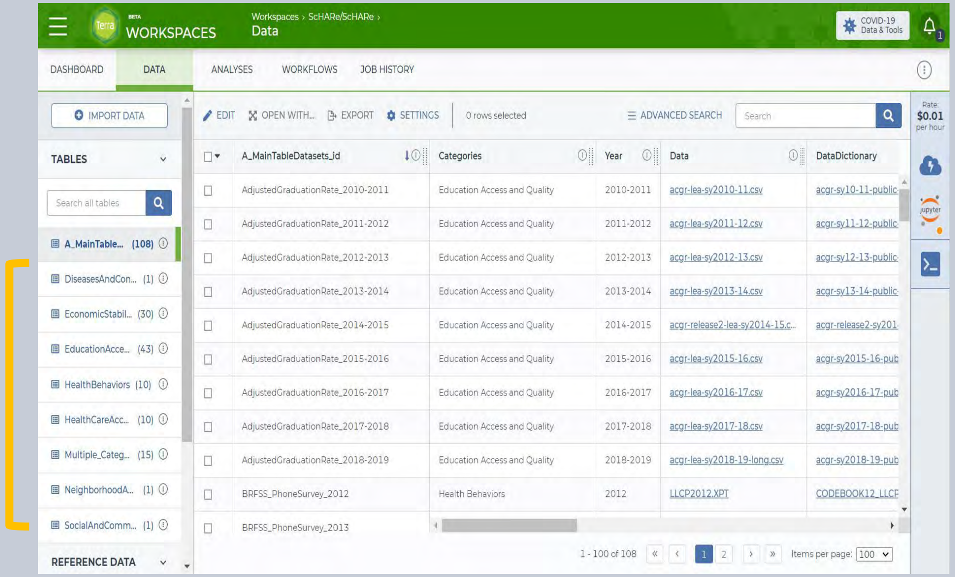
- **ScHARe Hosted Public Datasets:** publicly accessible, de-identified datasets hosted by ScHARe

  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



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_LLCF
BRFSS_PhoneSurvey_2013				

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

CDE FOCUSED

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,
early label Apartment, Company, Sui

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)

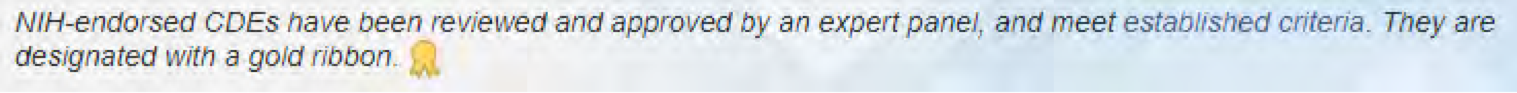
**NIMHD Framework
**Health Disparity Outcomes

NIH Endorsed

(** project level CDE)

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

Cross-walked with PhenX SDoH



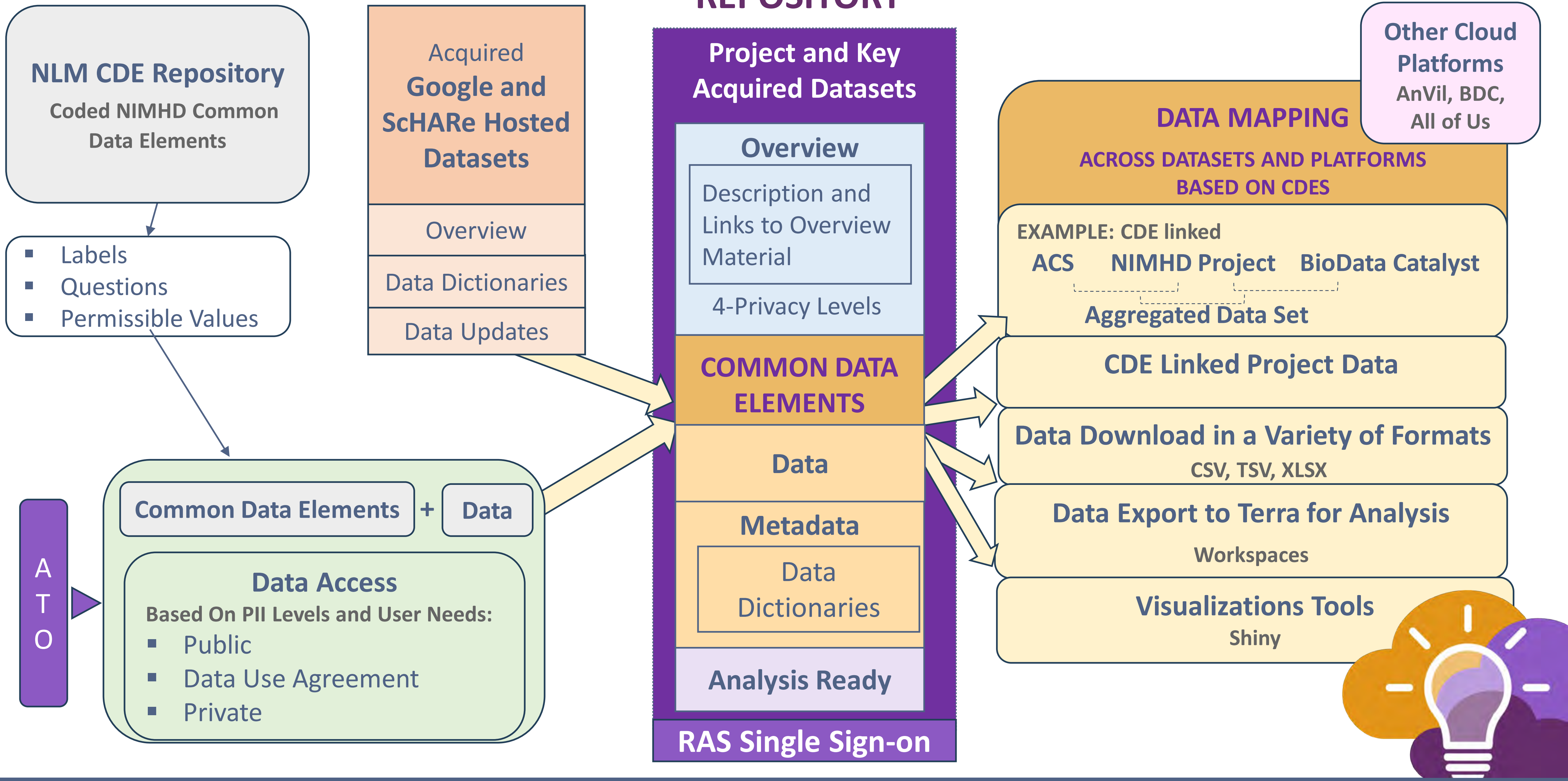
NIH-endorsed CDEs have been reviewed and approved by an expert panel, and meet established criteria. They are designated with a gold ribbon. 🎗

ScHARe REPOSITORY

COMMON DATA ELEMENTS

DATA UPLOAD

DATA MAPPING, DOWNLOAD AND EXPORT



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

Acquired **Google and ScHARe Hosted Datasets**

- Overview
- Data Dictionaries
- Data Updates

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

ACROSS DATASETS AND PLATFORMS BASED ON CDES

Other Cloud Platforms

AnVil, BDC, All of Us

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

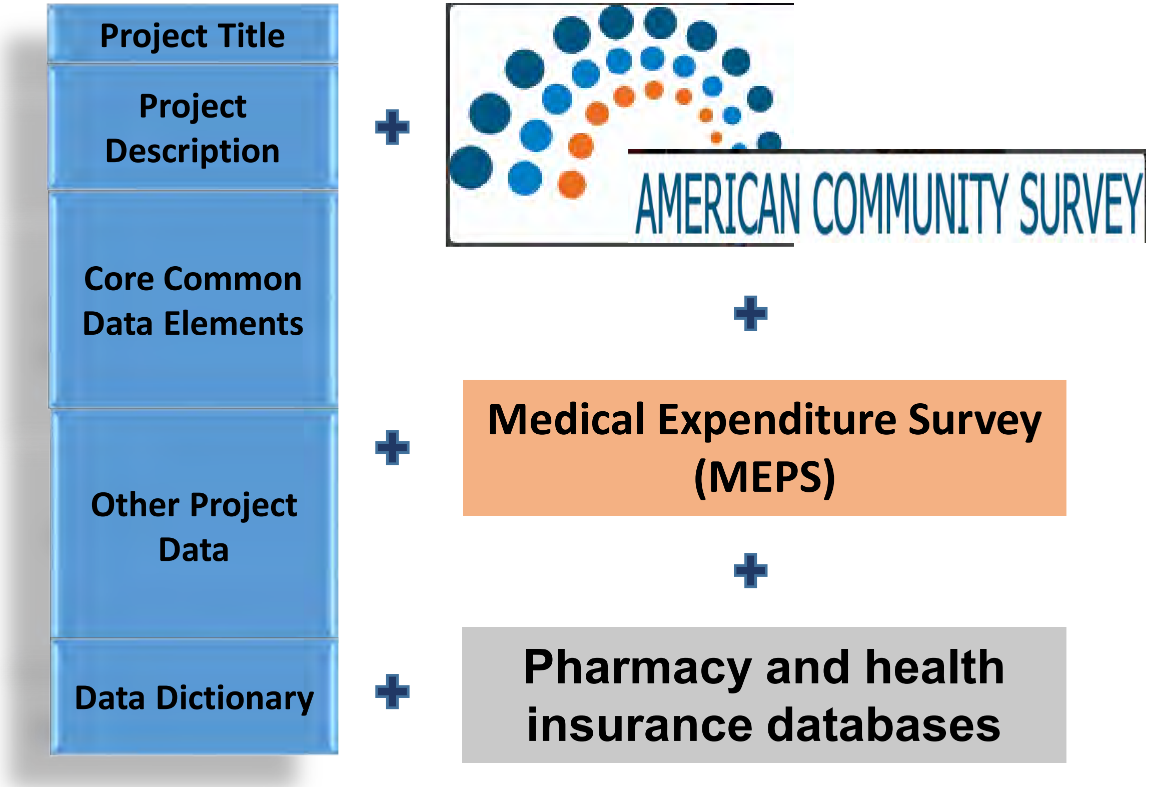
Workspaces

Visualizations Tools

Shiny

ScHARe

Project & federated dataset mapping



Project Title

Project Description

Core Common Data Elements

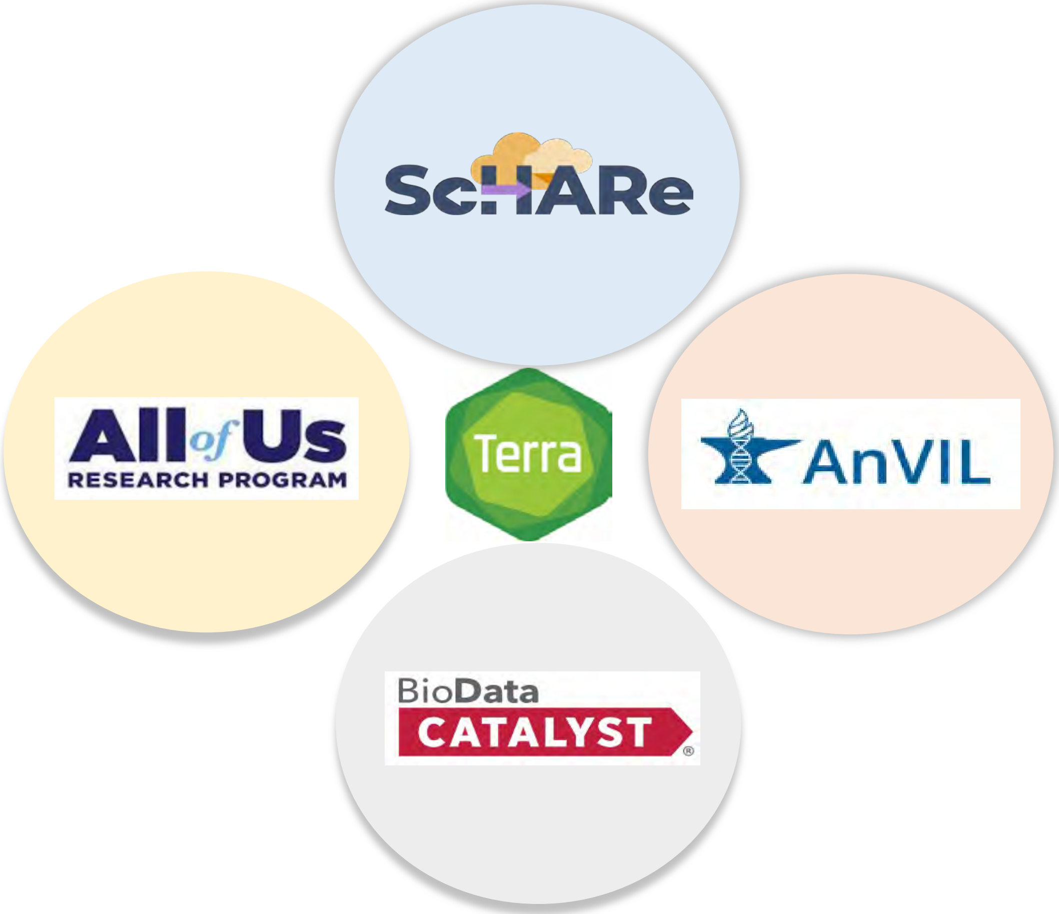
Other Project Data

Data Dictionary

Medical Expenditure Survey (MEPS)

Pharmacy and health insurance databases

Mapping across cloud platforms

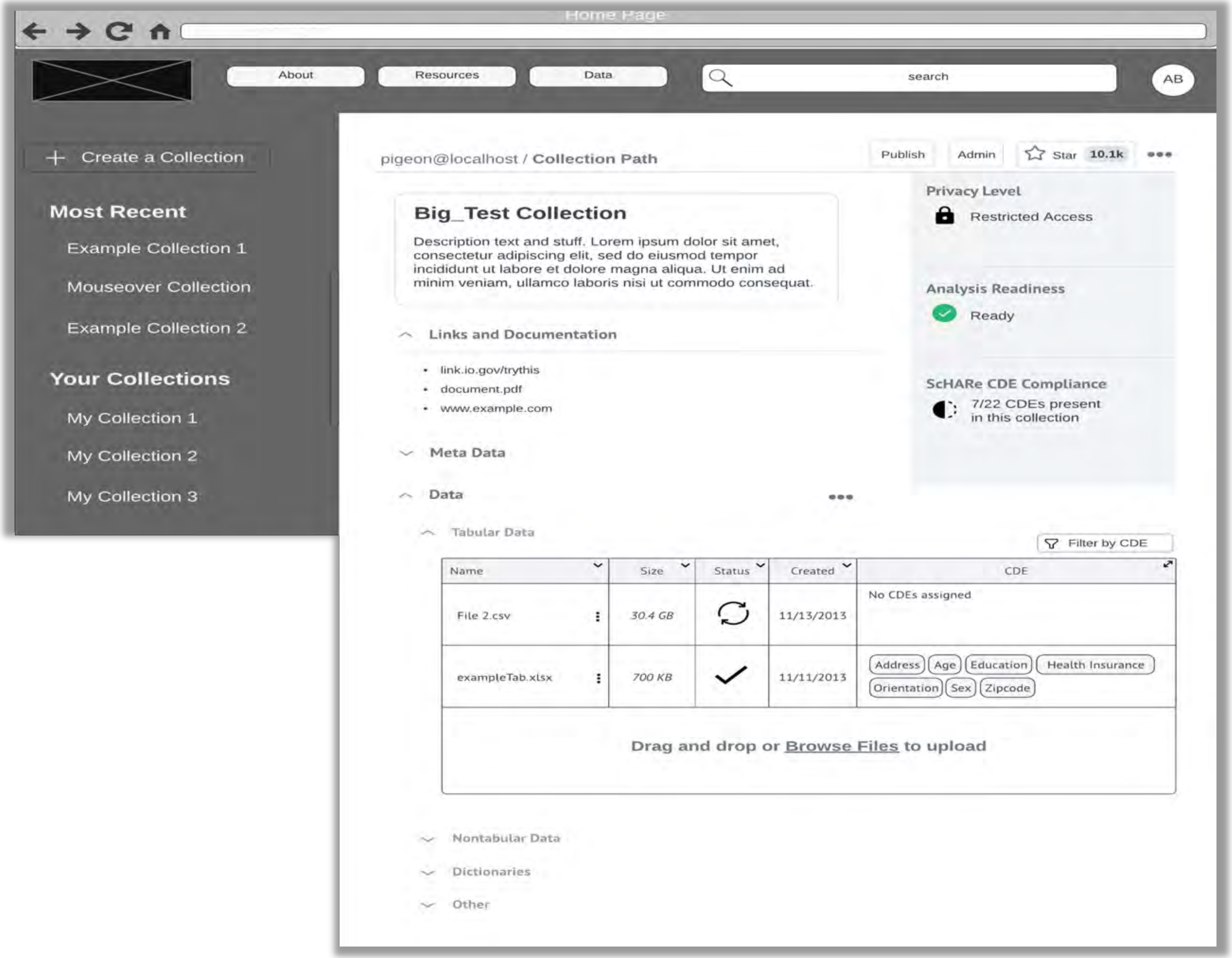


UPCOMING

ScHARe

Repository
CDE Focused
for Data
Interoperability

Coming Soon



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**

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

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

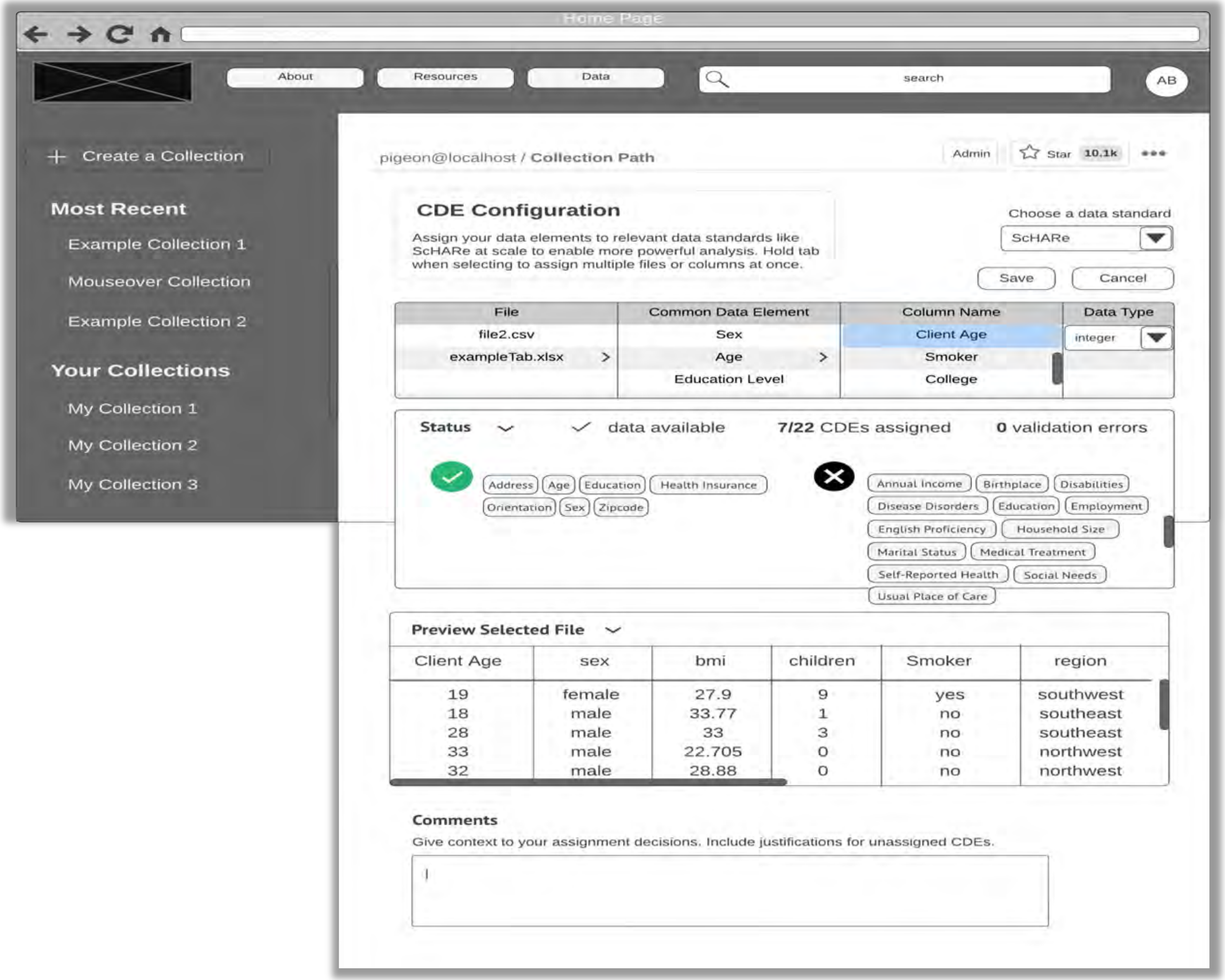
Dictionaries

Other

ScHARe

Repository
CDE Focused for Data Interoperability

Coming Soon



＋ 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**

Admin ☆ Star 10.1k •••

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

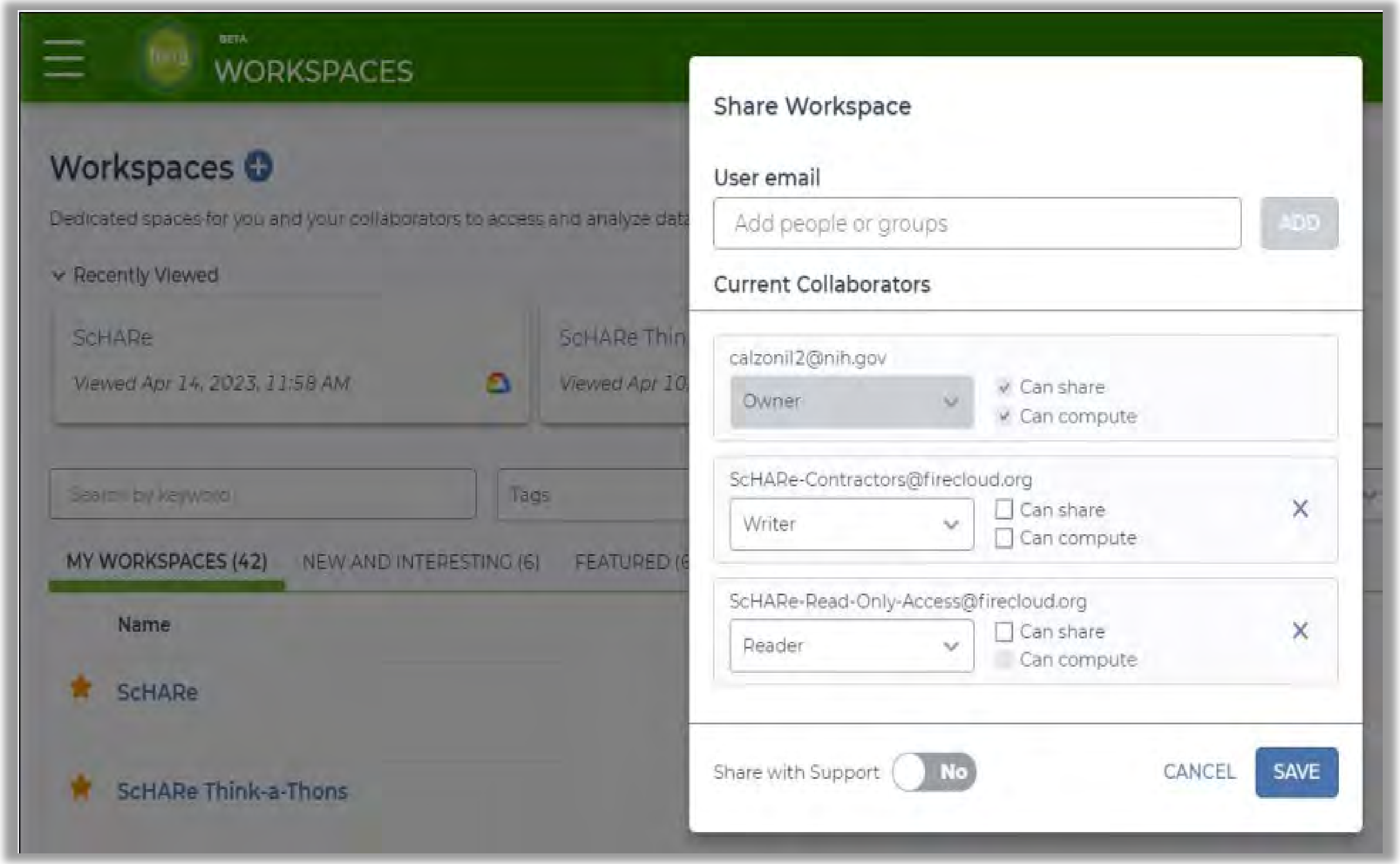
Preview Selected File ⌄

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 dat

Recently Viewed

ScHARe
Viewed Apr 14, 2023, 11:58 AM

ScHARe Thin
Viewed Apr 10,

Search by keyword

Tags

MY WORKSPACES (42) NEW AND INTERESTING (6) FEATURED (6

Name

ScHARe

ScHARe Think-a-Thons

Share Workspace

User email

Add people or groups ADD

Current Collaborators

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

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

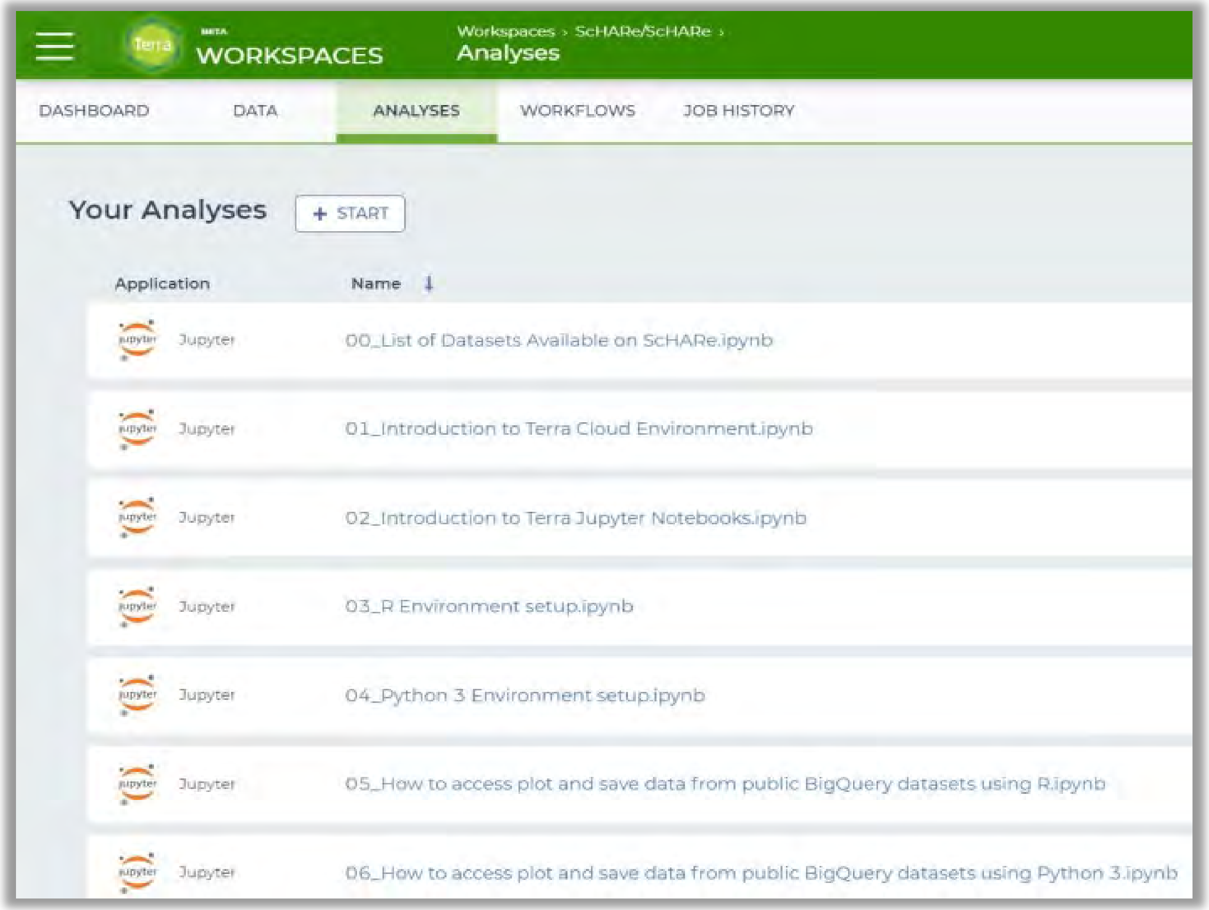
ScHARe-Read-Only-Access@firecloud.org
Reader □ Can share □ Can compute ×

Share with Support No

CANCEL SAVE

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

Notebooks analytics



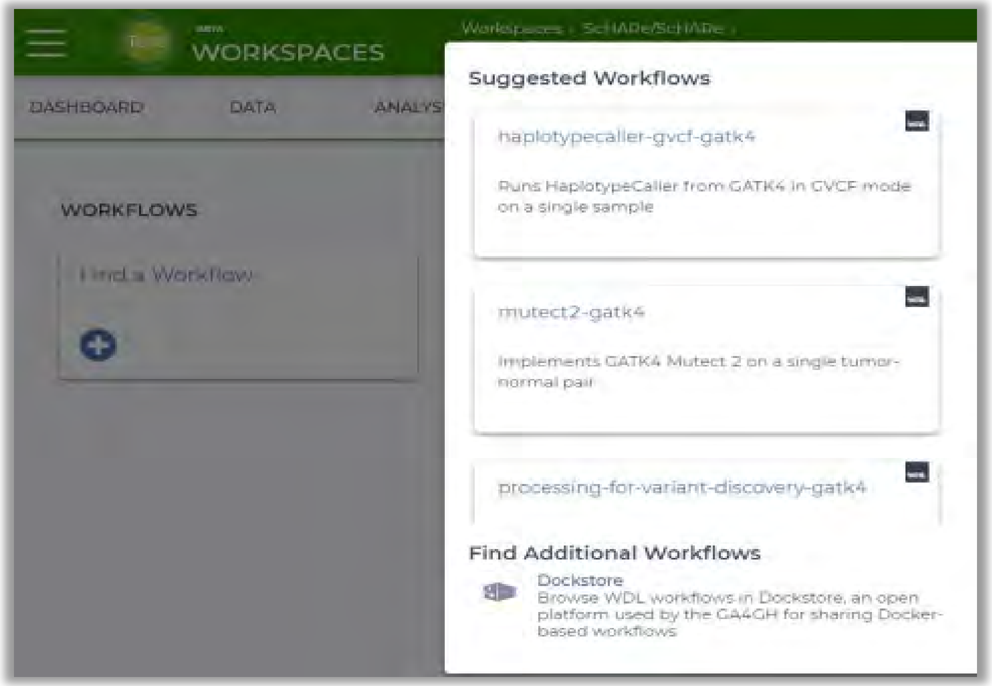
Workspaces › ScHARe/ScHARe ›
Analyses

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

Workflows - Modular codes

- **Copy and paste analytics**



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

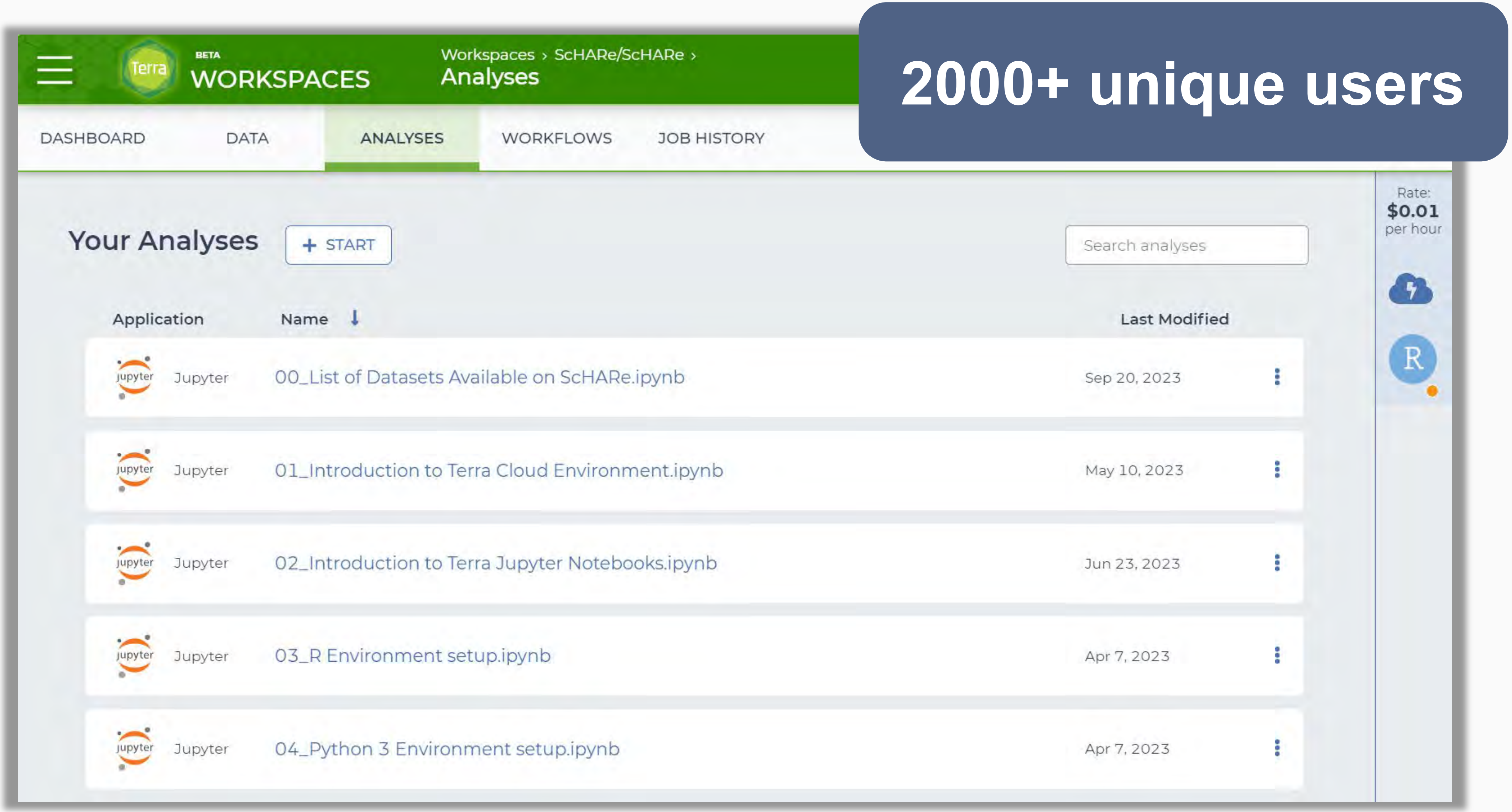
processing-for-variant-discovery-gatk4

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



Your 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

2000+ 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

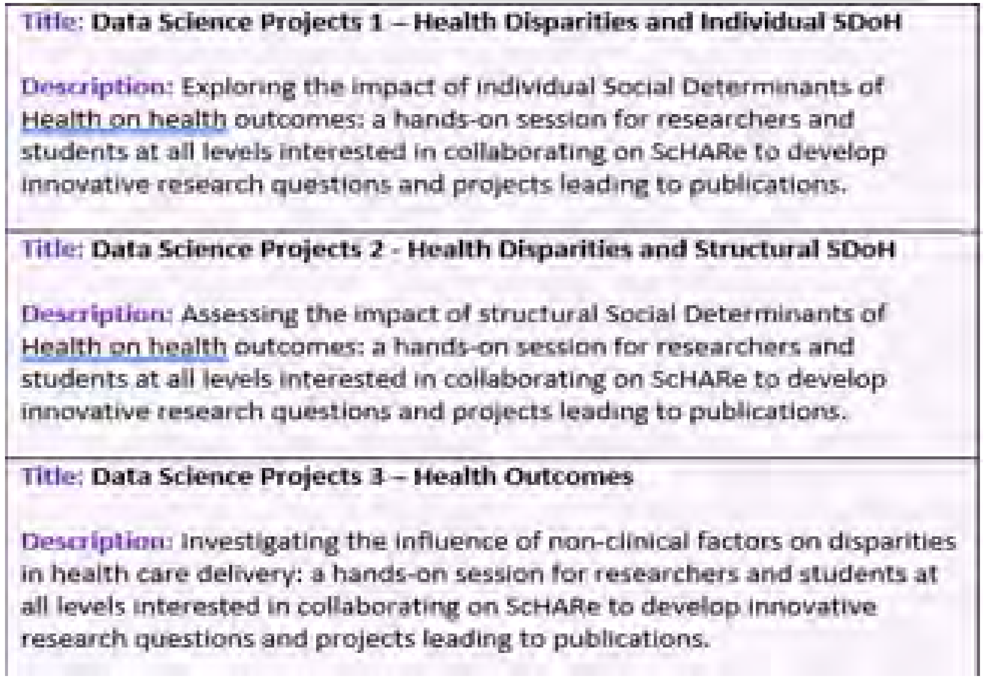
bit.ly/think-a-thons

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

Conducting research projects

Unleashing the power of secondary data

A guide to initiating and conducting research projects

Welcome, researchers!

This comprehensive guide equips you with the knowledge and tools to navigate the exciting world of secondary data analysis research projects.

By leveraging existing datasets, we can:

- unlock valuable insights
- contribute to meaningful advancements across various fields

What is secondary data analysis research?

Secondary data analysis involves utilizing existing datasets collected by other researchers or organizations. This approach offers numerous advantages, including:

- **Cost-effectiveness:** Eliminates the need for expensive data collection efforts.
- **Time-efficiency:** Provides access to historical data and facilitates quicker research completion.
- **Diverse data availability:** Grants access to a wider range of data sources and variables compared to primary data collection.

Importance of research documentation for transparent AI in secondary analysis

Secondary analysis of existing AI research data carries immense potential for advancing the field.

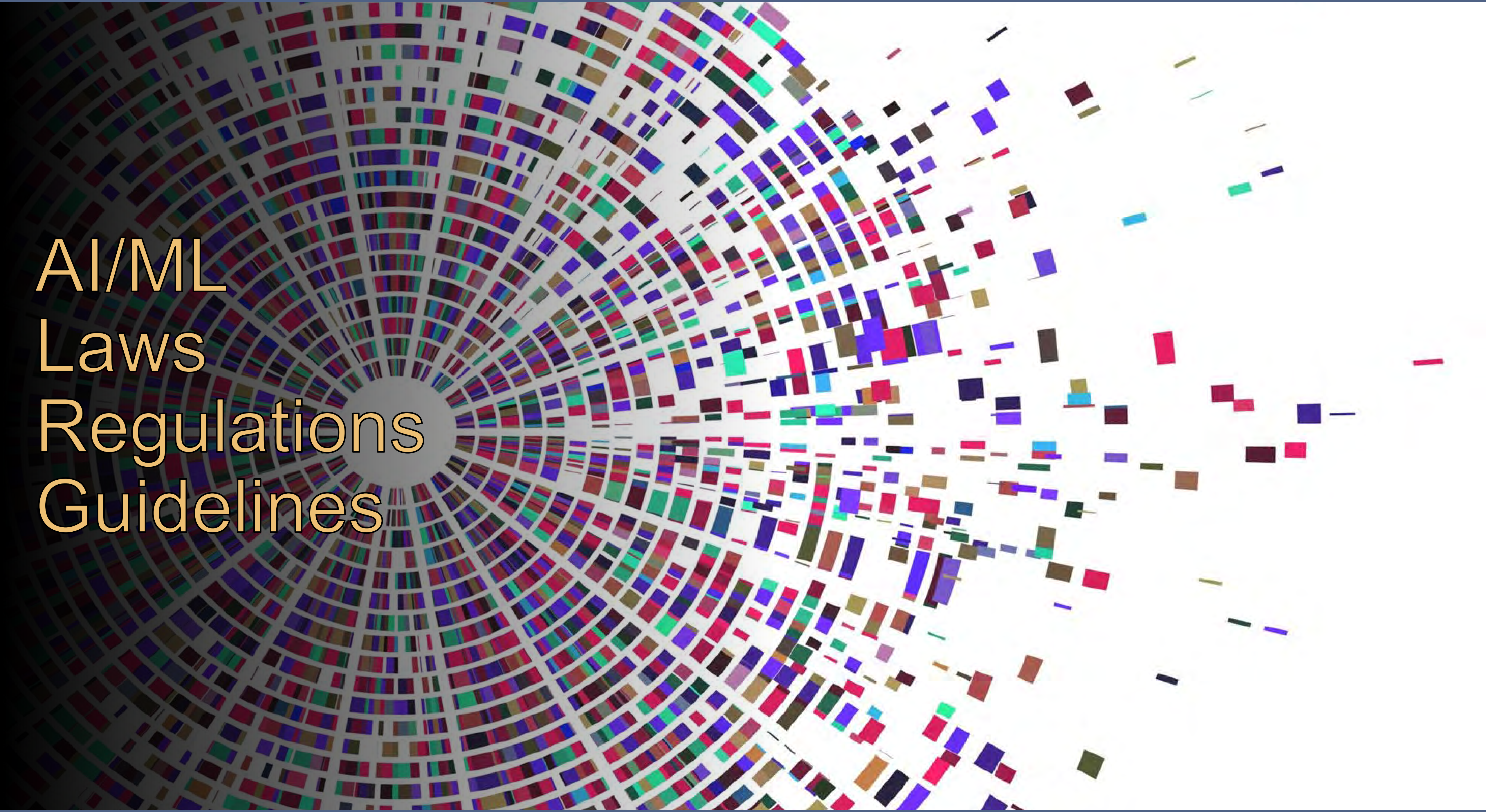
However, this process necessitates meticulous **documentation practices** to ensure transparency and address ethical concerns around data reuse.

Benefits of Transparent Documentation:

- **Reproducibility:** Enables other researchers to replicate the analysis, verify findings, and build upon existing research.
- **Accountability:** Allows stakeholders to hold researchers accountable for responsible AI development and mitigate potential bias or ethical concerns.
- **Collaboration:** Fosters collaboration and knowledge sharing within the research community by facilitating clear communication about research methods and findings.

ScHARe

Documenting research



AI/ML Laws Regulations Guidelines

EU and Touring Institute Shining the Way

EU sets global standards with first major AI regulations

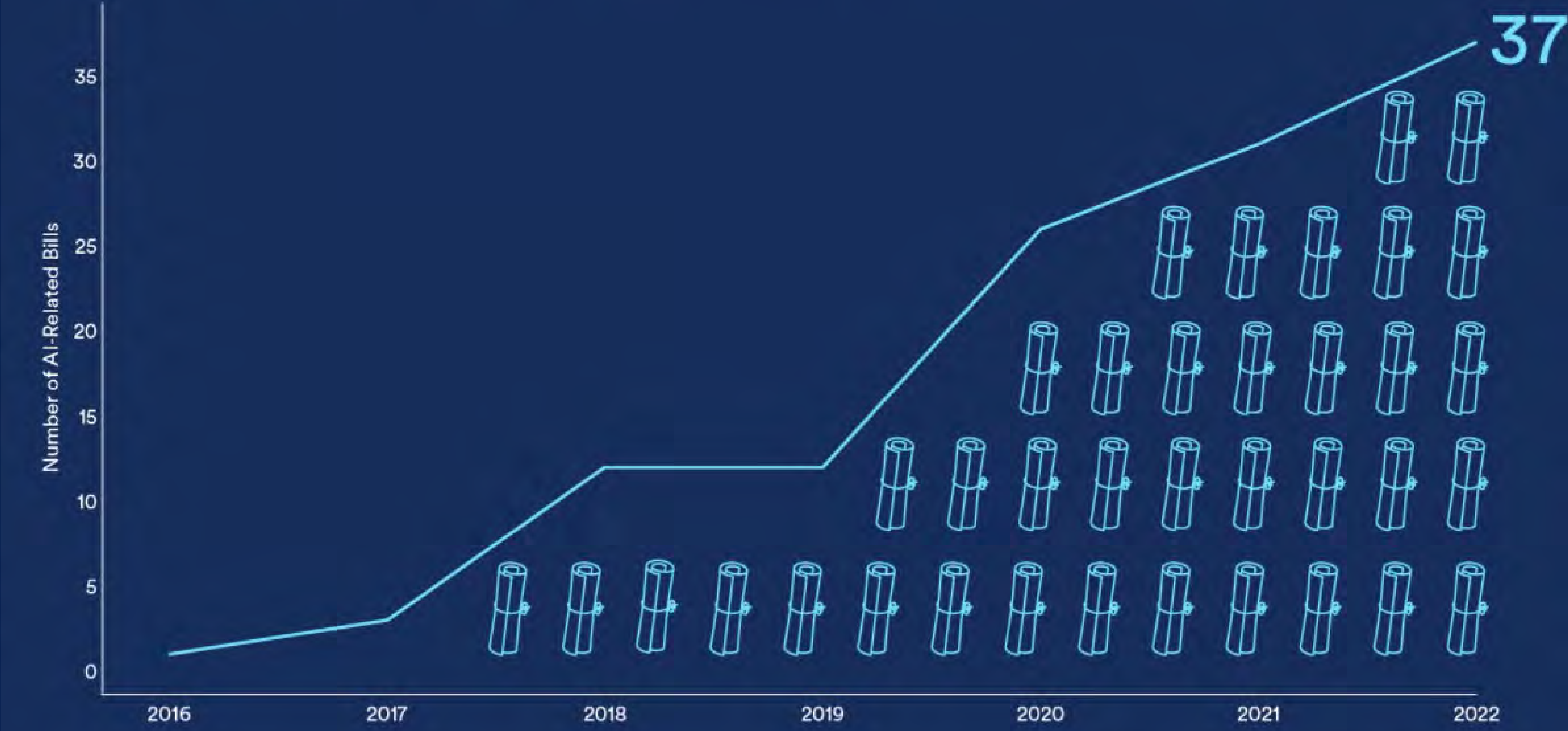
- Europe becomes the first major world power to enact comprehensive AI regulations, covering areas like transparency, use of AI in public spaces, and high-risk systems.
- High-impact models with systemic risks face stricter requirements, including model evaluation, risk mitigation, and incident reporting.
- Governments can use real-time facial recognition in limited cases, excluding cognitive manipulation and social scoring.
- Requires foundation models such as ChatGPT and general purpose AI systems (GPAI) to comply with transparency obligations before they are put on the market. These include drawing up technical documentation, complying with EU copyright law and disseminating detailed summaries about the content used for training.

General Data Protection Regulation (GDPR) - Rights of the data subject

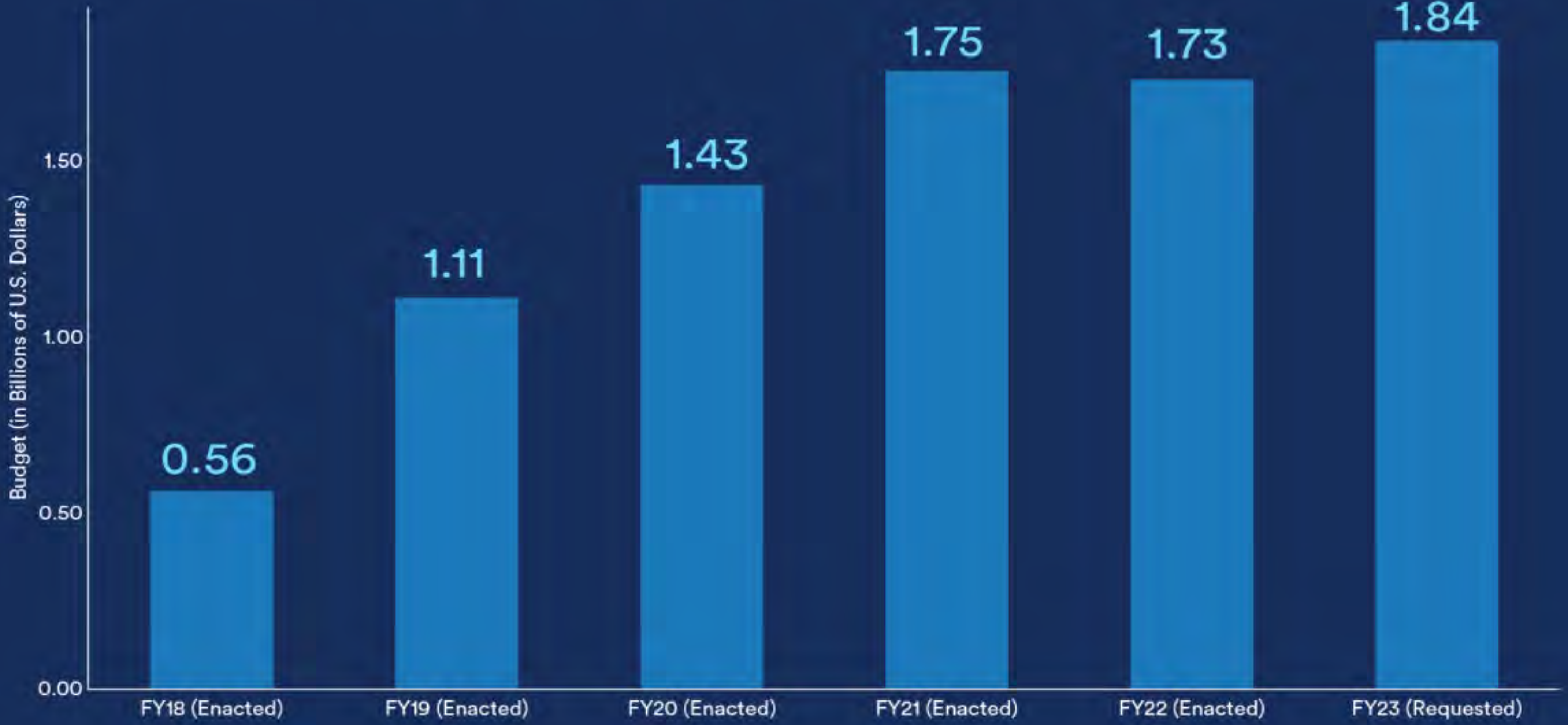
- The right to be informed - what personal data is being processed
- The right to rectification – correct or erase aspects of their personal data
- The right of access – receive data in a structured commonly used and machine-readable format
- The right to be forgotten (erasure) – object to the processing of their personal data
- The right to restrict the processing of your data– restrict processing of their personal data

https://www.weforum.org/agenda/2023/12/europe-landmark-ai-regulation-deal/ https://gdpr-info.eu/chapter-3/

Number of AI-Related Bills Passed into Law Globally



U.S. Federal Budget for AI R&D (Non-defense)



Source: U.S. NITRD Program, 2022 | Chart: 2023 AI Index Report

U.S. lacks a comprehensive federal AI law, existing regulations

EU sets global standards with first major AI regulations

- Europe becomes the first major world power to enact comprehensive AI regulations, covering areas like transparency, use of AI in public spaces, and high-risk systems.
- High-impact models with systemic risks face stricter requirements, including model evaluation, risk mitigation, and incident reporting.
- Governments can use real-time facial recognition in limited cases, excluding cognitive manipulation and social scoring.
- Requires foundation models such as ChatGPT and general purpose AI systems (GPAI) to comply with transparency obligations before they are put on the market. These include drawing up technical documentation, complying with EU copyright law and disseminating detailed summaries about the content used for training.

Federal AI Governance Policy:

- The **White House**, **Congress**, and various federal agencies have been actively shaping AI governance.
- The **Federal Trade Commission**, the **Consumer Financial Protection Bureau**, and the **National Institute of Standards and Technology** have all contributed to AI-related initiatives and policies.
- Notably, existing laws do apply to AI technology, and the focus is on understanding how these laws intersect with AI rather than creating entirely new AI-specific legislation

White House Initiatives for AI Governance

New Standards for AI Safety and Security

- Require that developers of the most powerful AI systems share their safety test results and other critical information with the U.S. government.
- Develop standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy.
- Protect against the risks of using AI to engineer dangerous biological materials
- Protect Americans from AI-enabled fraud and deception by establishing standards and best practices for detecting AI-generated content and authenticating official content.
- Establish an advanced cybersecurity program to develop AI tools to find and fix vulnerabilities in critical software
- Order the development of a National Security Memorandum that directs further actions on AI and security

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence | The White House

Standing Up for Consumers, Patients, and Students

Advance the responsible use of AI in healthcare and the development of affordable and life-saving drugs.

Shape AI's potential to transform education by creating resources to support educators deploying AI-enabled educational tools

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence | The White House

White House Initiatives for AI Governance

Protecting Americans’ Privacy

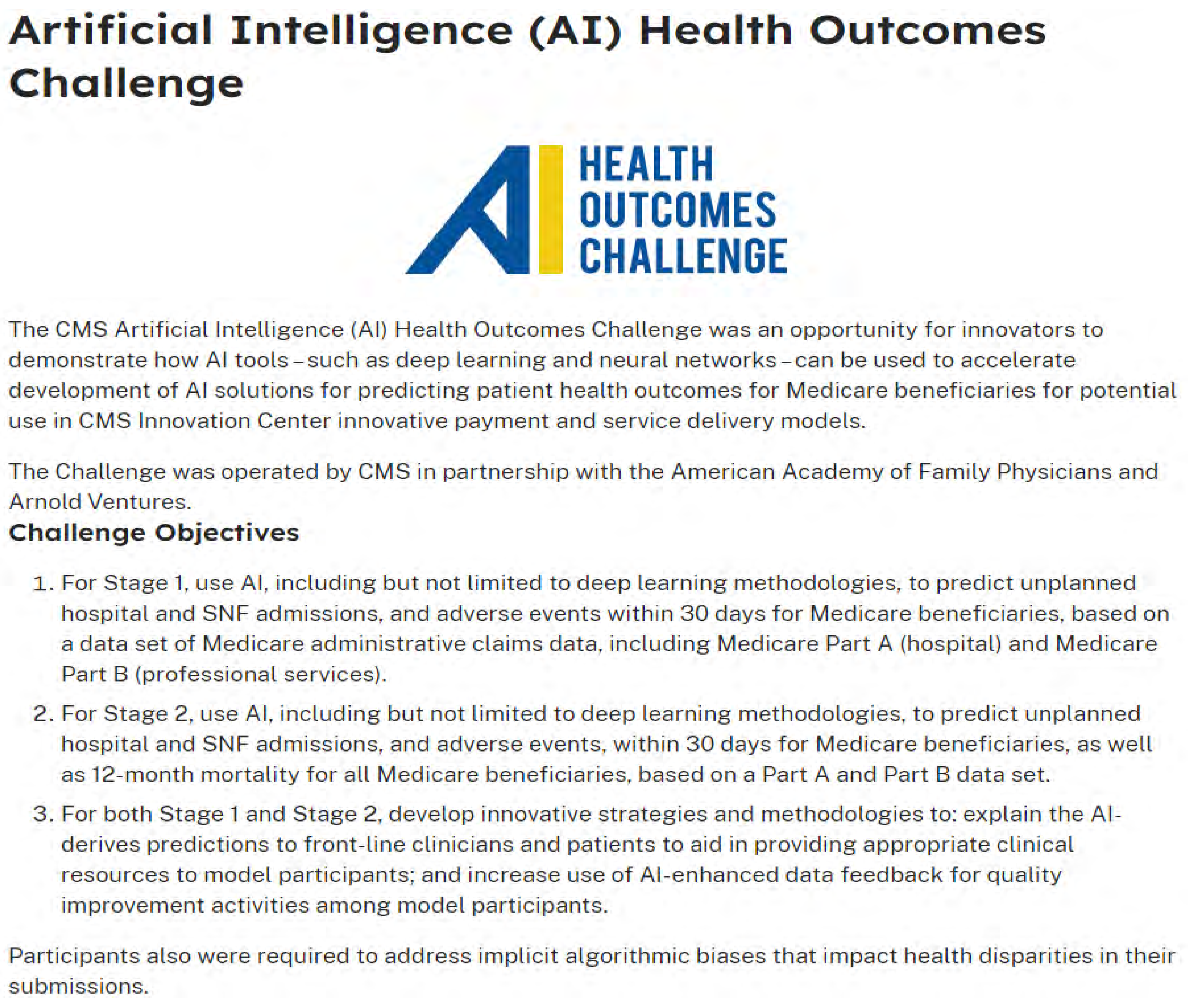
- Protect Americans’ privacy by prioritizing federal support for accelerating the development and use of privacy-preserving techniques
- Strengthen privacy-preserving research and technologies
- Evaluate how agencies collect and use commercially available information
- Develop guidelines for federal agencies to evaluate the effectiveness of privacy-preserving techniques

Advancing Equity and Civil Rights

- Provide clear guidance to landlords, Federal benefits programs, and federal contractors to keep AI algorithms from being used to exacerbate discrimination.
- Address algorithmic discrimination through training, technical assistance, and coordination between the Department of Justice and Federal civil rights offices on best practices for investigating and prosecuting civil rights violations related to AI.
- Ensure fairness throughout the criminal justice system by developing best practices on the use of AI in sentencing, parole and probation, pretrial release and detention, risk assessments, surveillance, crime forecasting and predictive policing, and forensic analysis

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence | The White House

CMS



Artificial Intelligence (AI) Health Outcomes Challenge

The CMS Artificial Intelligence (AI) Health Outcomes Challenge was an opportunity for innovators to demonstrate how AI tools – such as deep learning and neural networks – can be used to accelerate development of AI solutions for predicting patient health outcomes for Medicare beneficiaries for potential use in CMS Innovation Center innovative payment and service delivery models.

The Challenge was operated by CMS in partnership with the American Academy of Family Physicians and Arnold Ventures.

Challenge Objectives

1. For Stage 1, use AI, including but not limited to deep learning methodologies, to predict unplanned hospital and SNF admissions, and adverse events within 30 days for Medicare beneficiaries, based on a data set of Medicare administrative claims data, including Medicare Part A (hospital) and Medicare Part B (professional services).
2. For Stage 2, use AI, including but not limited to deep learning methodologies, to predict unplanned hospital and SNF admissions, and adverse events, within 30 days for Medicare beneficiaries, as well as 12-month mortality for all Medicare beneficiaries, based on a Part A and Part B data set.
3. For both Stage 1 and Stage 2, develop innovative strategies and methodologies to: explain the AI-derives predictions to front-line clinicians and patients to aid in providing appropriate clinical resources to model participants; and increase use of AI-enhanced data feedback for quality improvement activities among model participants.

Participants also were required to address implicit algorithmic biases that impact health disparities in their submissions.

FDA

The **U.S. Food and Drug Administration (FDA)** has been actively addressing the regulation of **Artificial Intelligence (AI) and Machine Learning (ML)** in medical devices. Here are some key actions and guidelines:

1. **AI/ML-Based Software as a Medical Device (SaMD) Action Plan**:
    1. Released in January 2021, this action plan outlines the FDA’s approach to overseeing AI/ML-based medical software. It focuses on improving patient care while ensuring safe and effective software functionality.
    2. Key actions include:
        1. Developing a regulatory framework, including guidance on change control plans for software learning over time.
        2. Supporting good machine learning practices for evaluating and enhancing algorithms.
        3. Prioritizing a patient-centered approach with transparent device communication to users.
        4. Advancing real-world performance monitoring pilots[1].
2. **Draft Guidance on AI/ML in Medical Devices**:
    1. In April 2023, the FDA published groundbreaking draft guidance specifically addressing the use of AI/ML in medical devices. This guidance aims to provide clarity and promote responsible adoption of AI/ML technologies in healthcare[2].
3. **Transparency and Explainability**:
    1. The FDA recognizes the importance of transparency in AI/ML-based technologies. Promoting a patient-centered approach involves ensuring users understand how these technologies work and their potential impact on patient care[3].

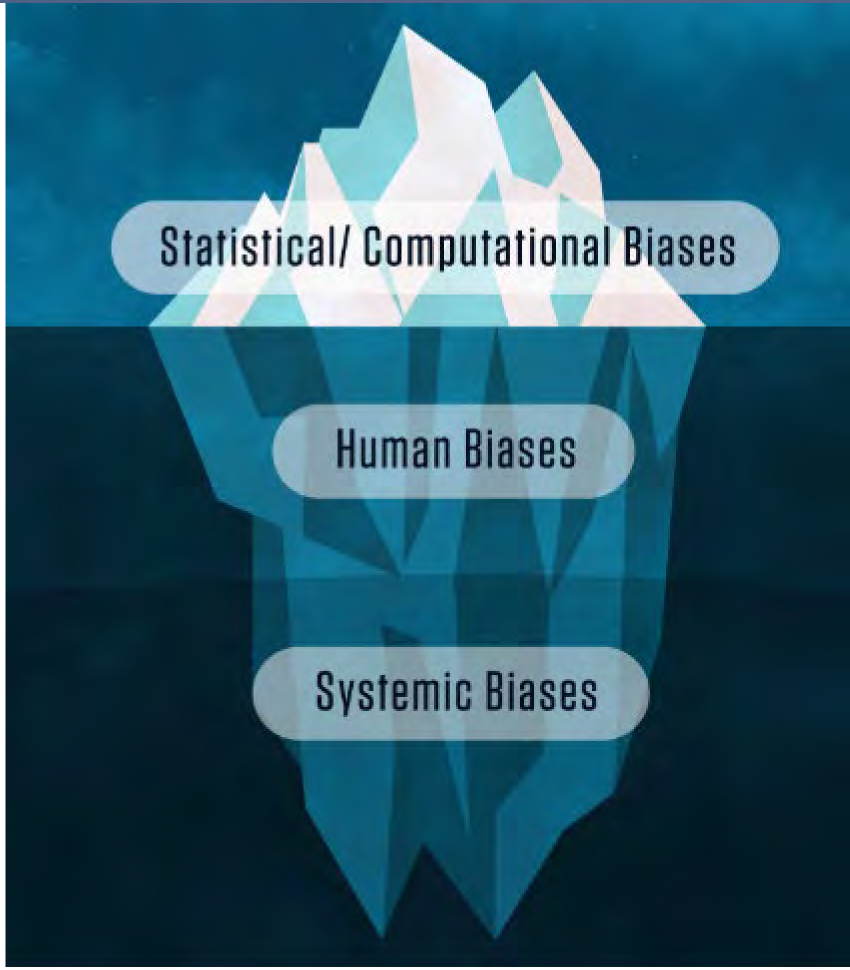
These efforts demonstrate the FDA’s commitment to fostering innovation while safeguarding patient safety in the rapidly evolving field of AI and ML in healthcare.

NIST

Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence--New guidance seeks to cultivate trust in AI technologies and promote AI innovation while mitigating risk.

Discusses how organizations can frame the risks related to AI and outlines the characteristics of trustworthy AI systems.

Describes four specific functions — govern, map, measure and manage — to help organizations address the risks of AI systems in practice.



Towards a Standard for Identifying and Managing Bias in Artificial Intelligence

Bias is neither new nor unique to AI and it is not possible to achieve zero risk of bias in an AI system

Trustworthy and Responsible AI is not just about whether a given AI system is biased, fair or ethical, but whether it does what is claimed.

https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

State-Level Legislation:

While there is no sweeping federal legislation akin to the EU’s Artificial Intelligence Act, several **U.S. states** have passed laws related to AI and ML.

These state laws often address privacy concerns and may extend to AI systems handling specific types of personal data

Data AI Use and Re-Use **Consent** / Privacy

No Laws and Regulations….some principles and guidelines….some emerging laws and regulations

In a putative class action filed on June 28, 2023, in the Northern District of California, and in other similar cases, plaintiffs allege that **OpenAI, Microsoft, and their respective affiliates violated the privacy rights of millions of internet users through the large-scale scraping of their personal data from social media, blog posts, and other websites, and using those data to train machine learning models.**

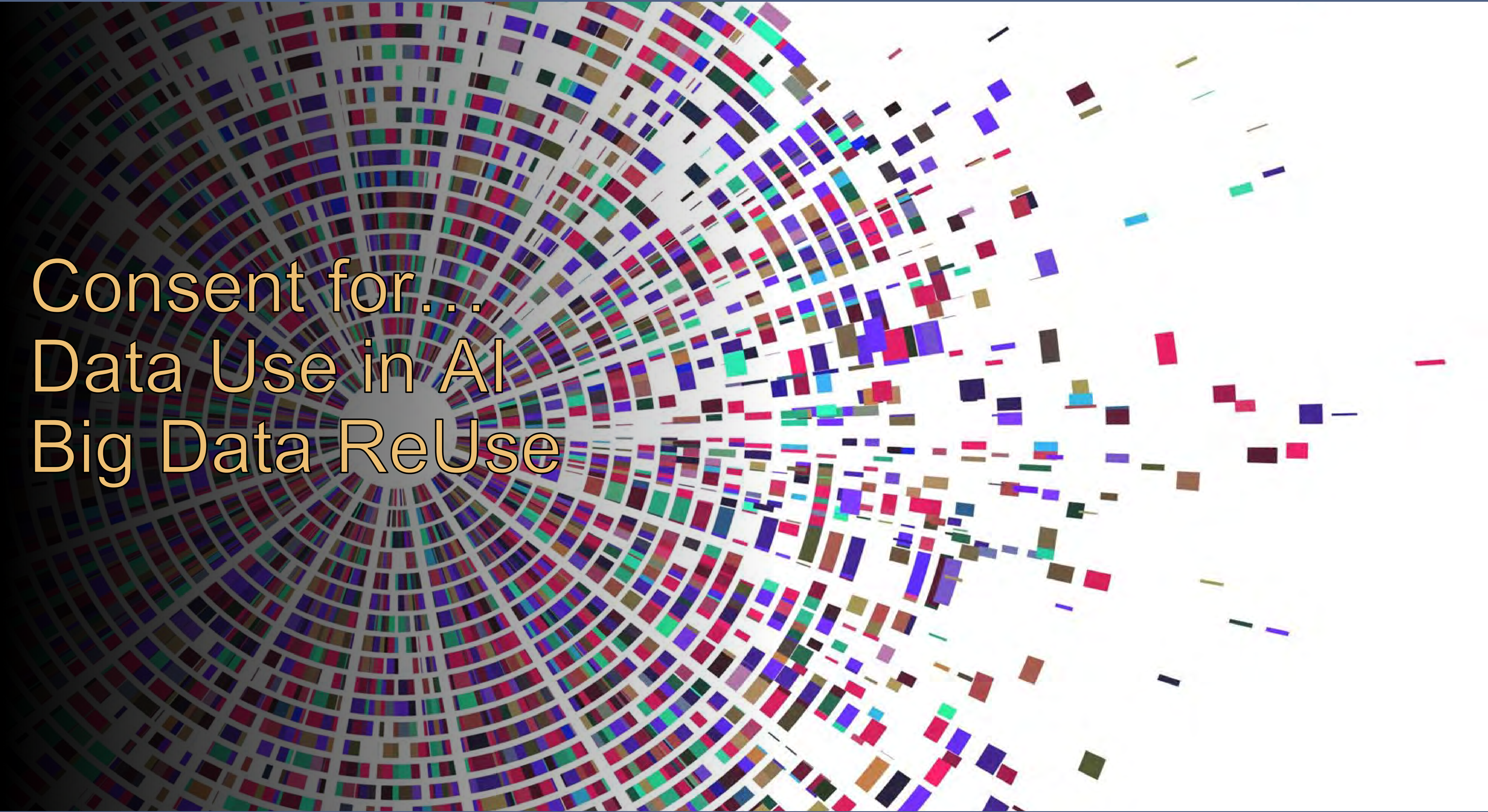
- Violated the Computer Fraud and Abuse Act (CFAA) by intentionally accessing protected computers without authorization and obtaining information through ChatGPT plug-ins integrated across various platforms/ websites
- Failed to adequately disclose that users' data may be used to train machine learning models/generative AI tools.
- Consent is necessary for large-scale scraping of personal data from the internet for use in training AI tools.

In early 2023, stock photo provider Getty Images sued Stability AI, a smaller AI start-up in Delaware federal court, alleging the **illegal use of its photos to train an image-generating bot**.

In Europe, under the **General Data Protection Regulation (GDPR),** organizations must obtain explicit consent from EU citizens before collecting or processing their personal data.

Dark Patterns – a user interface designed or manipulated with the substantial effect of subverting or impairing user autonomy, decision-making, or choice (Emerging Rules: Colorado, California, Federal Trade Commission)

https://www.jdsupra.com/legalnews/data-scraping-privacy-law-and-the-3354981/



Consent for…
Data Use in AI
Big Data ReUse

Data AI Use and Re-Use Consent / **Privacy**

PERSONALLY IDENTIFIABLE INFORMATION (PII)-Anything used to identify someone

- Name
- Social security number (SSN)
- Passport number
- Driver's license number
- Taxpayer identification number
- Patient identification number
- Financial account or credit card number
- Personal address and phone number
- Email

NON-SENSITIVE PII

Accessible from public sources like phonebooks, the Internet, and corporate directories.

- Race
- Zip code
- Gender
- Date of birth
- Place of birth
- Religion

Any information that can be used in combination to identify someone

Emergence of big data has also increased the number of data breaches and cyberattacks by entities who realize the value of this information.

https://toolkit.ncats.nih.gov/glossary/personally-identifiable-information/

Data AI Use and Re-Use Consent / Privacy

Consent Before Big Data and AI Uses

- Current practices not applicable to the evolving applications and innovative research designs of big data research
- Data Use Agreements based on the notion of 'reasonable expectations' for the reuse of data;

Consent for Data AI Use and ReUse (Benefits)

- Understanding the aetiology of diseases requires their study longitudinally
- Big data analytics has ability to make novel predictive inferences across datasets about the interactions of disparate risk factors
- This iterative novelty limits what can be communicated to participants about the purposes for which their data may be used

Key: No overall consensus about…

- how to define optimal participant welfare
- how consent for the reuse of data should be managed

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7862505/

Since big data, machine learning approach is predicated on the value of its capacity to reveal novel findings and causal relationships beyond those that are predictable through conventional means, it follows from this unpredictability that the standard account of prospectively informed consent may be inadequate.

Data AI Use and Re-Use Consent / Privacy

New World of DATAFICATION

Data collected everywhere…social media, phones, buying, healthcare, employment, security cameras, marketing…

Research Privacy Preservation:

- Data can do no harm
  - Insurance discrimination
  - Misuse and abuse of disease/disorder information
  - Identification of individual putting that person at risk for harm
- Users' own tolerance to the use of the data provided varies
- Health literacy varies
- Research might yield findings relevant to the health of the participant and about which they may, or may not, wish to be informed

Current deep learning systems can learn endless correlations between arbitrary bits of information that is not foreseeable, but still go no further; they fail to represent the richness of the world, and lack even any understanding that an external world exists at all.

Data AI Use and Re-Use **Consent** / Privacy

Data Platforms

- Many data repositories and federated data sets are open to the public and identifiable information removed
- In platform-based health data research, reuse and sharing of data by researchers granted access to data with PII is inevitable and necessary.
  - Privacy Protections in Place
  - Data Use Agreements (how many layers of data re-use or re-purposing is this valid?)
- Creation of Synthetic data from existing data

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7862505/

Data AI Use and Re-Use Consent / Privacy

Issue of Consent: Balancing Big Data Benefits and Personal Data Protection

Perspective Matters…..

Data Controller's Perspective:

- Data recycling – using data several times for the same purpose
- Data repurposing – using data for different purposes than for which they were initially collected
- Data recontextualization – using data in another context than in which they were initially collected

Data Subject's Perspective:

- Data sharing or data disclosure – data subjects have the ability to (directly) allow use and reuse of their personal data
- Data portability – data subjects have the ability to use and reuse their personal data across devices and services – potentially gain benefit from reuse of data
- The right to be forgotten – data subjects have the ability to block data use and reuse.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3046774

Data AI Use and Re-Use Consent / Privacy

Three basic approaches to informed consent for data sharing and reuse from platforms.

1. **Most Permissive - Widest sharing and reuse** - Allows data to be used in future research projects by the original research team or others through controlled access to identifiable data and/or release of publicly available de-identified data.
2. **Moderately Permissive - Limited sharing and reuse** - Allows for reuse and public data sharing of deidentified data but does not allow for reuse of identifiable data either by the original research team or others in the future.
3. **Least Permissive - No sharing or reuse** - Only the research team is allowed access, and data use is limited to the specific project under consideration with an expectation that identifiable data will be destroyed at a defined date. This scenario does not allow for reuse of identifiable data in the future or sharing de-identified data.

Data AI Use and Re-Use Consent / Privacy

Issue of Consent: Balancing Big Data Benefits and Personal Data Protection

Assumptions: Forms of data reuse that …..

- Stay close to the awareness and intentions of data subjects should be approached less tight – use same consent
- Are 'at a distance', i.e., in which awareness and transparency may be lacking and data subject's rights may prove more difficult to exercise, more restrictions and additional protection should be considered

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3046774

Data AI Use and Re-Use Consent / Privacy

Blanket Consent: Maximizing the research uses to which data can be put, but the disadvantage of failing to inform the donor of the nature of the research.

Broad Consent: Permission is sought for a range of uses but not assumed for all purposes and is constrained

Dynamic Consent: Consent is sought on a case-by-case basis for reuse of data for each specific purpose rather than for its initial use.

Challenges:

- suitable in instances where data subjects are unwilling or unable to have ongoing engagement with a digital research interface
- some people may refuse to allow their use of their samples for particular uses, and some individuals may be or become uncontactable.

Meta Consent: Individuals would be able to choose how they prefer to provide consent

Challenge: does not circumvent the unknowability of potential future uses that is a function of a predictive analytic machine learning approach

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7862505/

Data AI Use and Re-Use Consent / Privacy

Societal Consent

PUBLIC LEGITIMACY for CONSENTING AI USE and BIG DATA RE-USE

Acknowledge that changing societal attitudes towards a presumption of the reuse of data
in the context of contemporary health care and longitudinal research may be gradual
Shift between what is ‘normatively’ reasonable and what is ‘descriptively’ reasonable.

Data AI Use and Re-Use Consent / Privacy

Premise: Imposition of too many restrictions on the sharing of data for reuse ‘ inhibits data flows necessary to conduct AI research and deliver care in the context of a modern healthcare system’

Committed to the Principle of Collect Once, Use Many times
(Basis of the 2014 memorandum of understanding established between NHS England and the General Pharmaceutical Council)

(1) Optimizing the benefit from data-driven healthcare requires the sharing and reuse of data

(2) Reach of machine learning to uncover otherwise unpredictable associations between, and by extension uses for, health data makes necessary a reassessment of consent in this context.

Trustworthiness of AI: Should not over promise results using machine learning (Cancer vs Alzheimer’s) or the benefits to all diverse populations when not developed for such

Data AI Use and Re-Use **Consent** / Privacy

Open-Source AI/ML

- enables easier debugging
- more flexible approach to building deep learning models

Much of AI data collection is without consent

KEY CONCERNS:

•**Data privacy:** Generative AI tools often use data from the public internet - concerns about the privacy rights of individuals if their data has been used without their consent, potentially leading to violations of data protection laws and regulations.

•**Licensing and open-source issues**: AI tools may use open-source software or data with specific licensing requirements, which can result in legal disputes if these requirements are not met or are violated during the AI development process.

•**Intellectual property:** AI-generated content can closely resemble or even reproduce copyrighted material, raising questions about copyright infringement. The use of copyrighted data in training AI models could lead to legal claims regarding the unauthorized use of such material.

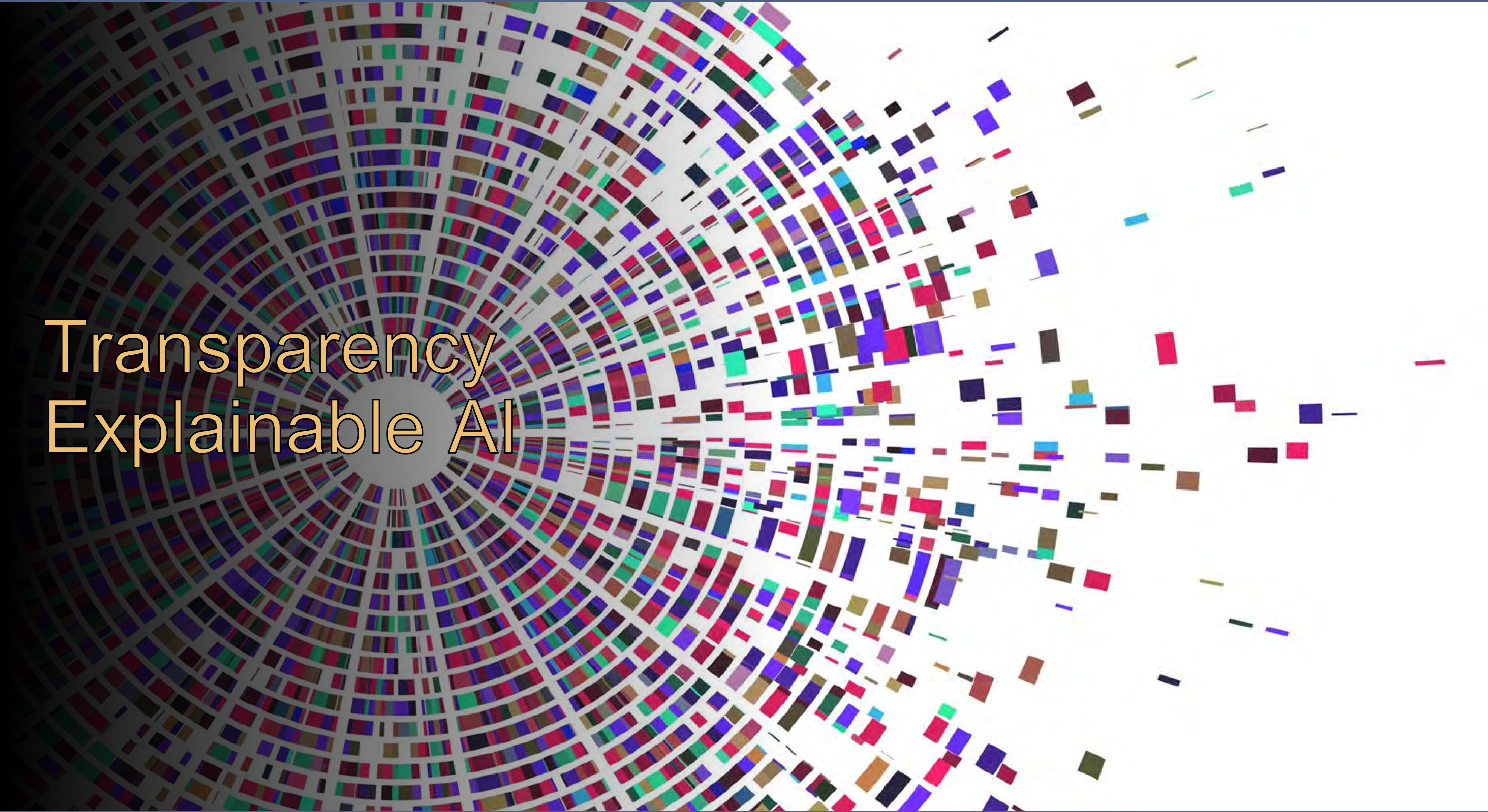
•**Responsibility for Harm**: In open-source collective where code builds upon code, responsibility and accountability are dimensioned for harm done.

Google owner Alphabet (GOOGL.O) launched a new AI model, Gemini, to rival OpenAI.

https://www.jdsupra.com/legalnews/data-scraping-privacy-law-and-the-3354981/

Slido

What are the best ways to engage with the public to explore the tension between the impossibility of providing comprehensive knowledge in advance about the uses to which one’s data might be put, and the probable health gains achievable through those as yet unforeseen uses?



Transparency
Explainable AI

Transparency and Explainability in AI

- Share process
- Reveal the data sets and components of the analytics being used
- Make known the assumptions that went into the development of the algorithm
- Expose how data, model or training issues were handled, like small sample sizes
- Divulge how the models were trained
- Clarify uses and limitations

Tools for Transparency & Explainability – Use and Re-Use of Data & Applications

QUESTIONNAIRE is a set of printed or written questions with a choice of answers, devised for the purposes of a survey or statistical study.

DATA COLLECTION SHEET is document containing a summary of data or other useful information. It doesn't contain choices. Datasheets are similar to Excel sheets.

DATASHEET typically include a variety of information, such as mechanical specifications, configurations, operating conditions, recommended usage guidelines, application notes and performance characteristics. It is essential for the data to be accurate and complete.

DATA MODEL is an abstract model that organizes elements of data and standardizes how they relate to one another and to the properties of real-world entities.

DATA DICTIONARY is a collection of names, definitions, and attributes about data elements that are being used or capture--Centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format

DATA/DATASET CARDS are structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a dataset's lifecycle for responsible AI development.

MODEL CARDS are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type) and intersectional groups (e.g., age and race, or sex) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information.

● DATA FOCUSED

- Data Sheets ✱■◆⬟⬣★✖
- Data Statements ✱■⬣★✖
- Data Nutrition Labels ✱■⬟⬣★
- Data Cards for NLP ✱■⬣★◆
- Dataset Development Lifecycle Documentation Framework ✱■◆⬟⬣★✖
- Data Cards ✱■⬟★✖

● MODELS & METHODS FOCUSED

- Model Cards ✱■◆⬟⬣★✖
- Value Cards ✱■⬟⬣✖
- Method Cards ✱■
- Consumer labels for Models ◆⬟⬣✖

● SYSTEMS FOCUSED

- System Cards ✱■◆⬣★
- FactSheets ✱◆⬣★
- ABOUT ML ✱■◆⬟⬣★✖

SAMPLE OF POTENTIAL AUDIENCES

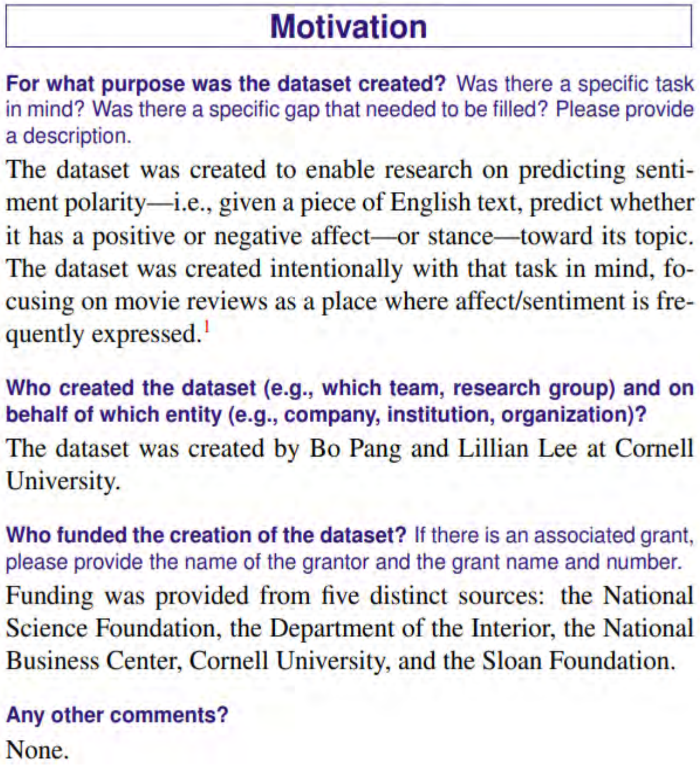
✱ ML Engineers	■ Model Developers/Reviewers	◆ Students	⬟ Policymakers
⬣ Ethicist	★ Data Scientist/Business Analyst	✖ Impacted Individuals	

https://huggingface.co/blog/model-cards

Background

Datasheets are a popularly suggested metadata file…
but there are many variants

Datasheet



Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

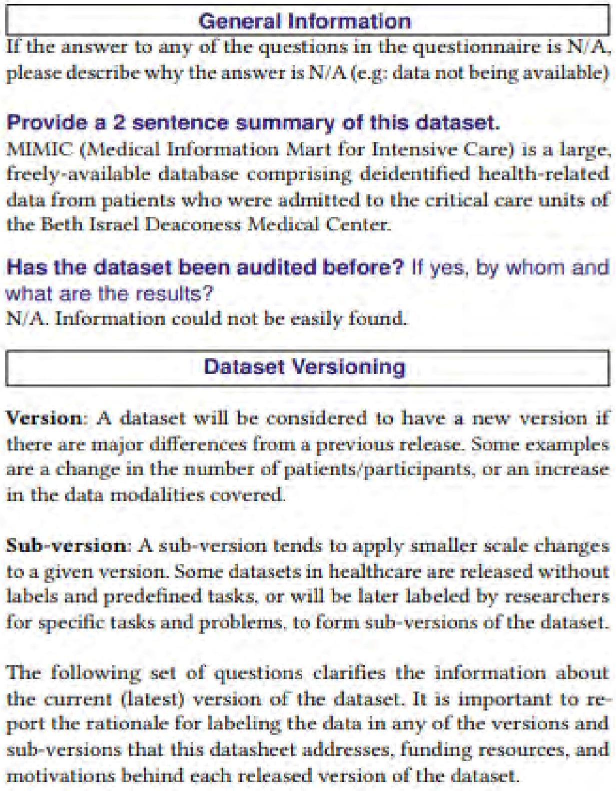
Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Healthsheet



General Information

If the answer to any of the questions in the questionnaire is N/A, please describe why the answer is N/A (e.g: data not being available)

Provide a 2 sentence summary of this dataset.

MIMIC (Medical Information Mart for Intensive Care) is a large, freely-available database comprising deidentified health-related data from patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Center.

Has the dataset been audited before? If yes, by whom and what are the results?

N/A. Information could not be easily found.

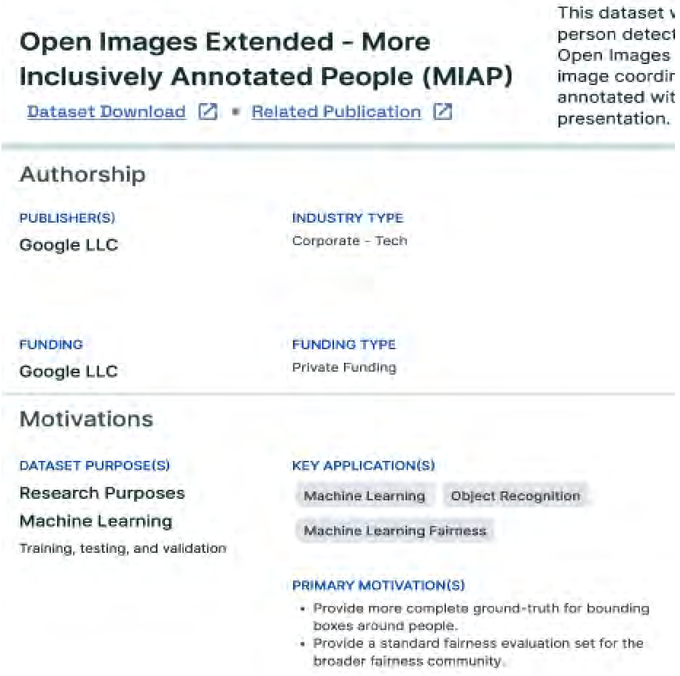
Dataset Versioning

Version: A dataset will be considered to have a new version if there are major differences from a previous release. Some examples are a change in the number of patients/participants, or an increase in the data modalities covered.

Sub-version: A sub-version tends to apply smaller scale changes to a given version. Some datasets in healthcare are released without labels and predefined tasks, or will be later labeled by researchers for specific tasks and problems, to form sub-versions of the dataset.

The following set of questions clarifies the information about the current (latest) version of the dataset. It is important to report the rationale for labeling the data in any of the versions and sub-versions that this datasheet addresses, funding resources, and motivations behind each released version of the dataset.

Data card



Open Images Extended - More Inclusively Annotated People (MIAP)

Dataset Download • Related Publication

This dataset v
person detect
Open Images
image coordir
annotated wit
presentation.

Authorship

PUBLISHER(S)
Google LLC

INDUSTRY TYPE
Corporate - Tech

FUNDING
Google LLC

FUNDING TYPE
Private Funding

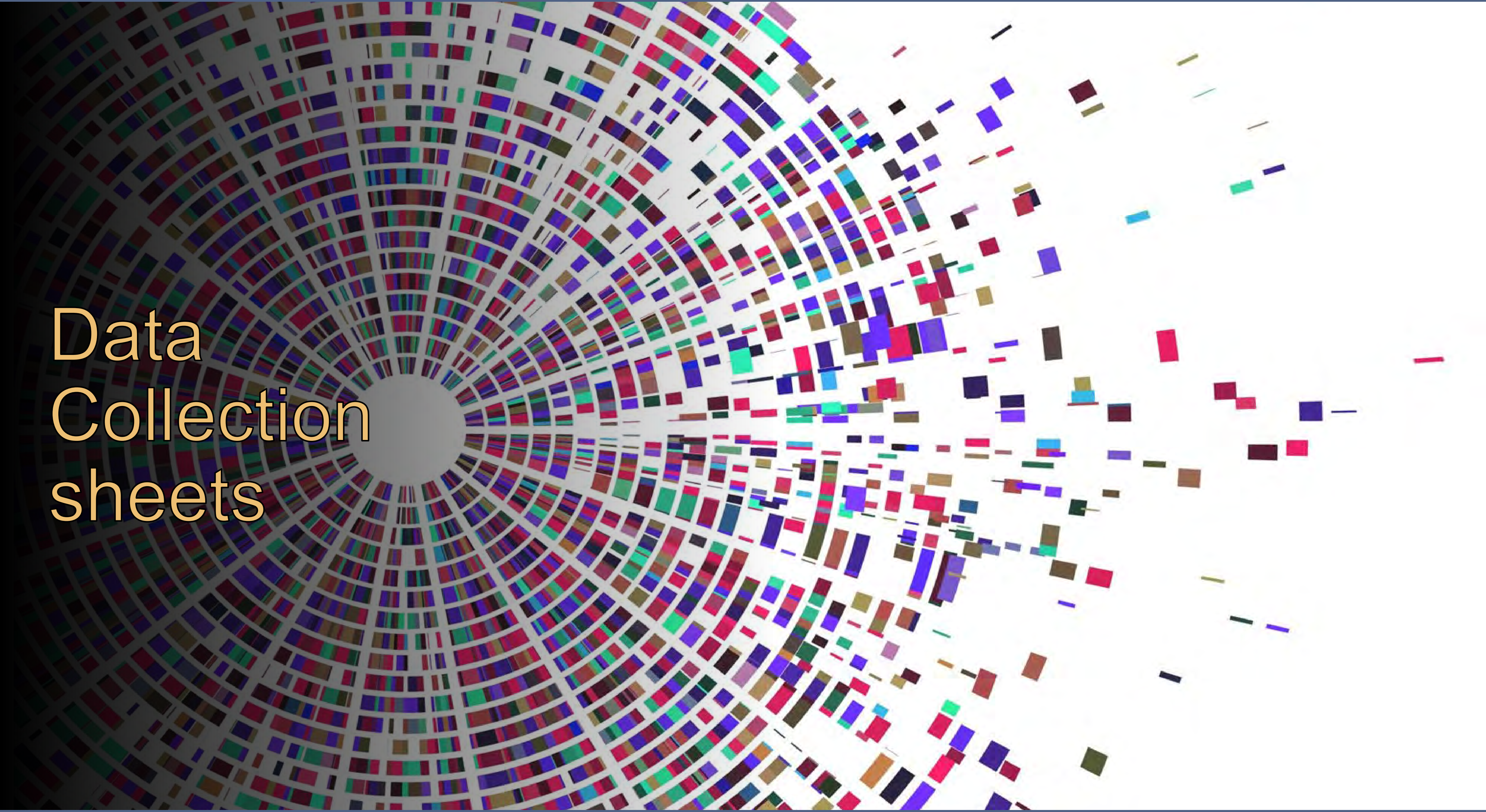
Motivations

DATASET PURPOSE(S)
Research Purposes
Machine Learning
Training, testing, and validation

KEY APPLICATION(S)
Machine Learning
Object Recognition
Machine Learning Fairness

PRIMARY MOTIVATION(S)
- Provide more complete ground-truth for bounding boxes around people.
- Provide a standard fairness evaluation set for the broader fairness community.

Data Collection sheets



Uses of Data Collection Sheets

•**Research**

A data collection sheet is a systematic tool for collecting and analyzing data in research. Quantitative researchers use data collection sheets to track different numerical values in the course of the systematic investigation.

•**It Saves Time**

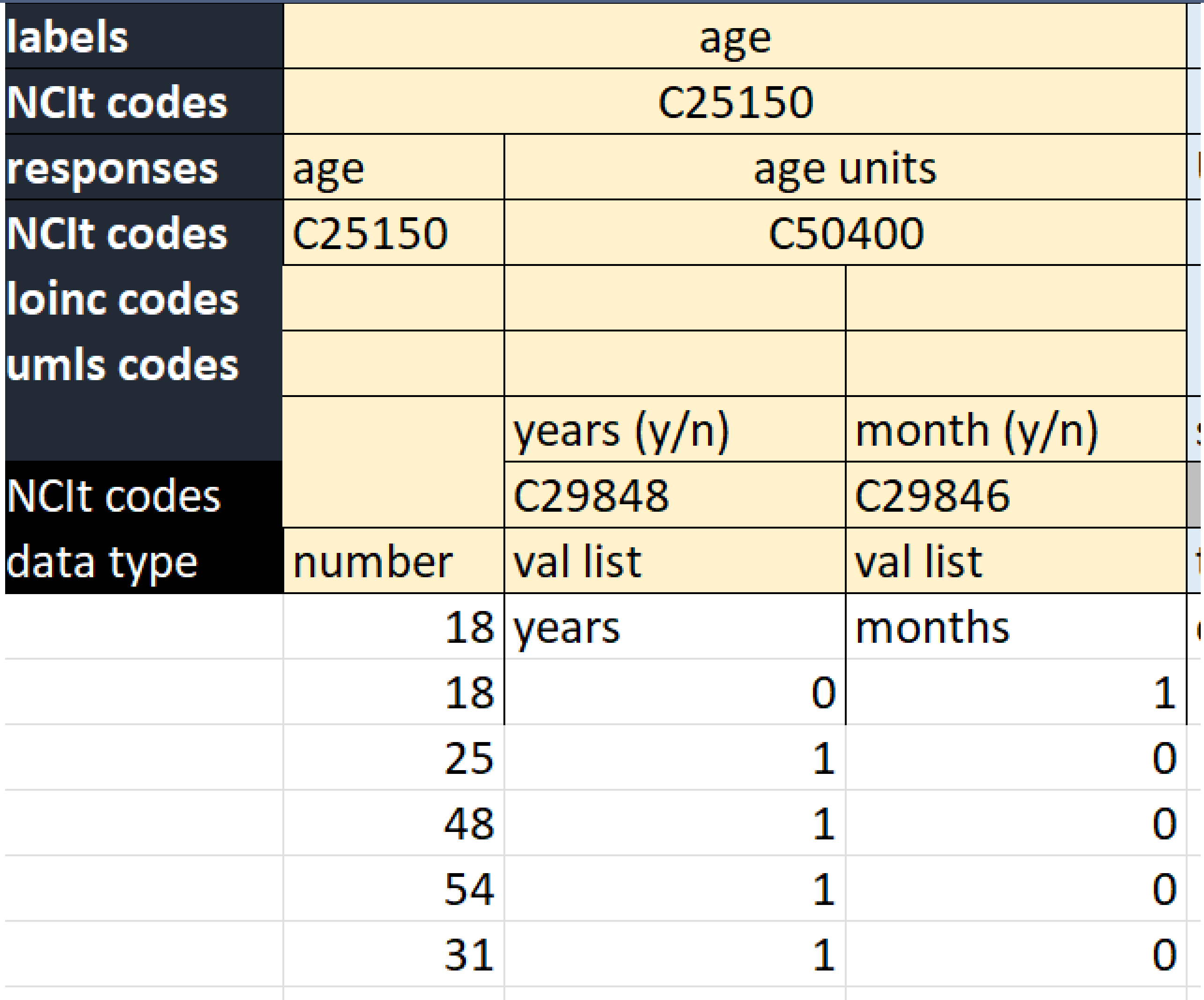
Using a data collection sheet helps you to be more efficient when carrying out a systematic investigation.

Data Categorization

A data collection sheet makes data categorization easy. You can place data variables in categories as you create different columns in your sheet.

•**Research Reporting**

It is a useful tool in research reporting. You can include a copy of your data sheet in your research report to help other parties understand how and why your data was captured.

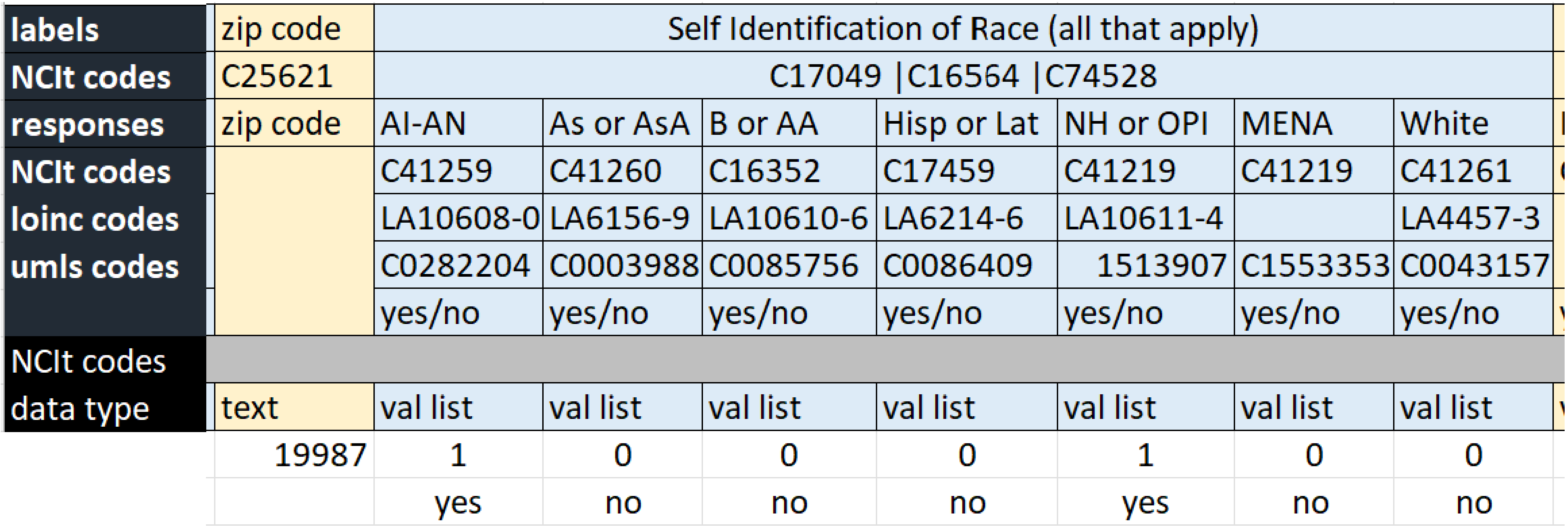


labels	age		
NCIt codes	C25150		
responses	age	age units	
NCIt codes	C25150	C50400	
loinc codes			
umls codes			
		years (y/n)	month (y/n)
NCIt codes		C29848	C29846
data type	number	val list	val list
	18	years	months
	18	0	1
	25	1	0
	48	1	0
	54	1	0
	31	1	0

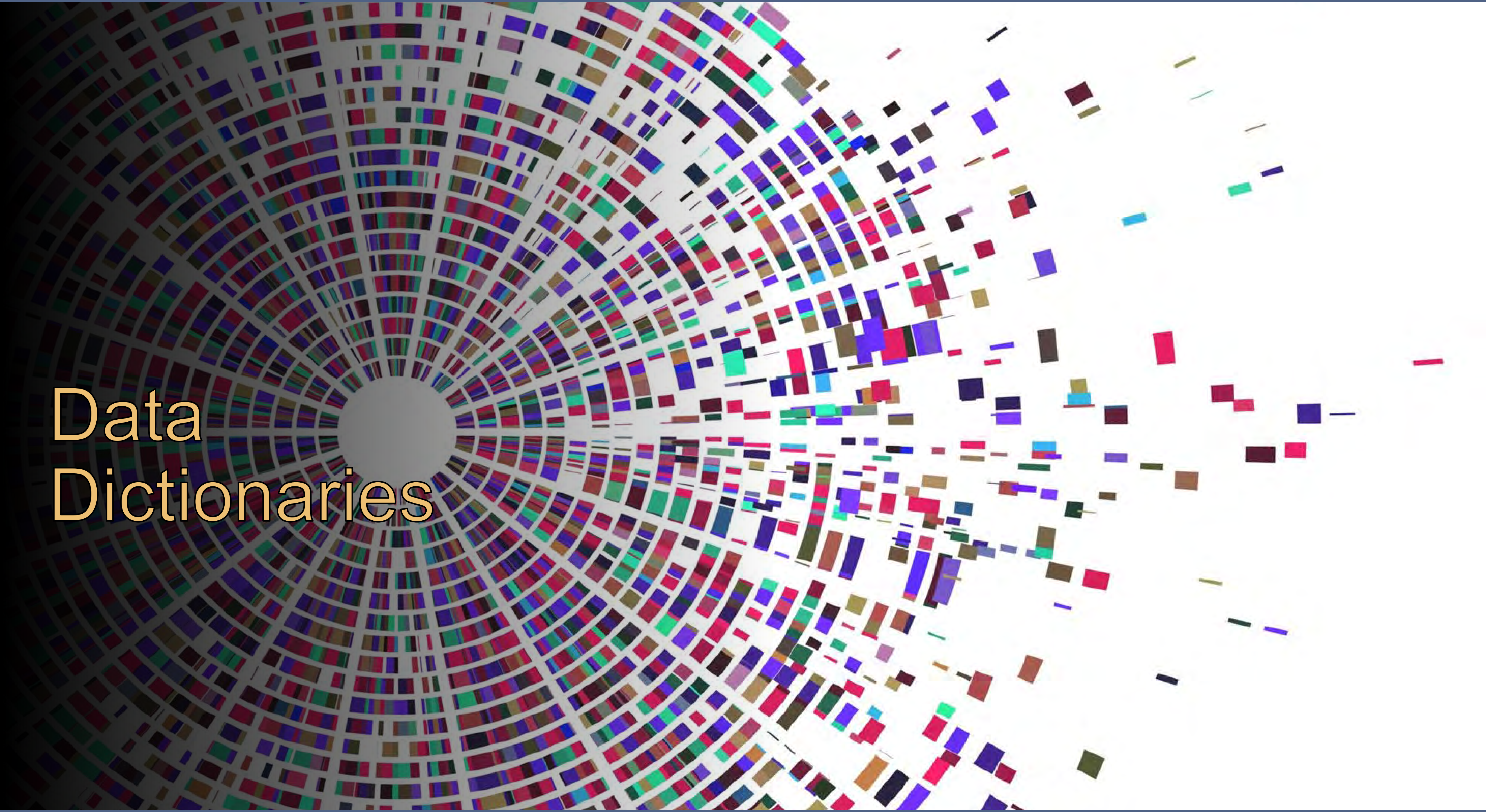
Source to create surveys and data collection sheet: https://www.formpl.us/blog/data-collection-sheet

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://images.template.net/wp-content/uploads/2022/07/Datasheets.pdf

Uses of Data Collection Sheets



labels	zip code	Self Identification of Race (all that apply)						
NCIt codes	C25621	C17049 |C16564 |C74528						
responses	zip code	AI-AN	As or AsA	B or AA	Hisp or Lat	NH or OPI	MENA	White
NCIt codes		C41259	C41260	C16352	C17459	C41219	C41219	C41261
loinc codes		LA10608-0	LA6156-9	LA10610-6	LA6214-6	LA10611-4		LA4457-3
umls codes		C0282204	C0003988	C0085756	C0086409	1513907	C1553353	C0043157
		yes/no	yes/no	yes/no	yes/no	yes/no	yes/no	yes/no
NCIt codes								
data type	text	val list	val list	val list	val list	val list	val list	val list
	19987	1	0	0	0	1	0	0
		yes	no	no	no	yes	no	no



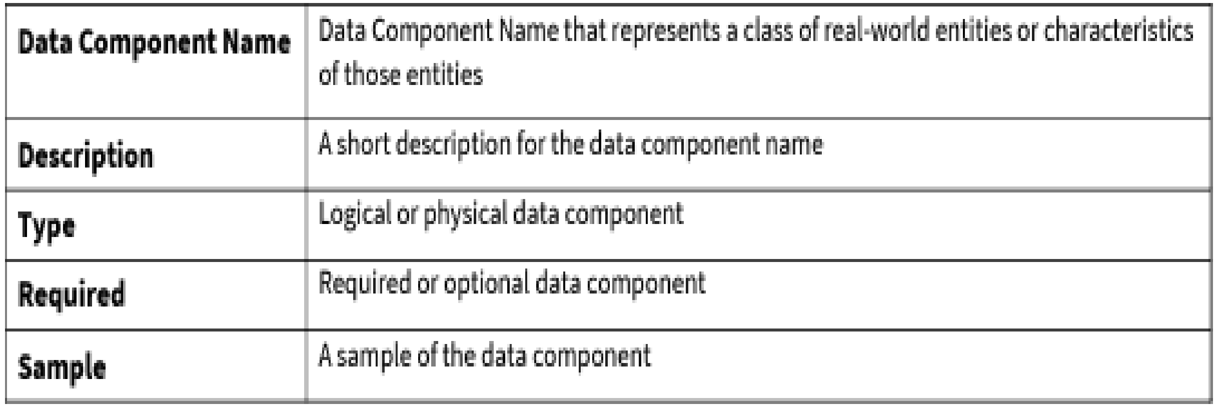
Data Dictionaries

Data Dictionary Description

A data dictionary is a collection of descriptions of data objects or terms, definitions, and properties in a data asset.

- Information about the data
- Essential communications tool for data modeling, curation, governance, and analytics, especially when dealing with datasets that have been collected, compiled, categorized, used, and reused by different internal and external data consumers across the organization.
- provides meaningful descriptions for individually named data objects

A data dictionary consists of several data components, which contains multiple levels: data asset, entity, attribute, and value domain. Each level includes different components, but each component should be defined with the following properties:



Data Component Name	Data Component Name that represents a class of real-world entities or characteristics of those entities
Description	A short description for the data component name
Type	Logical or physical data component
Required	Required or optional data component
Sample	A sample of the data component

The data asset level is composed of one line that contains a data asset profile, which includes the data asset name, description, type, version, and create and last update date.

Data Dictionaries – first step in transparency

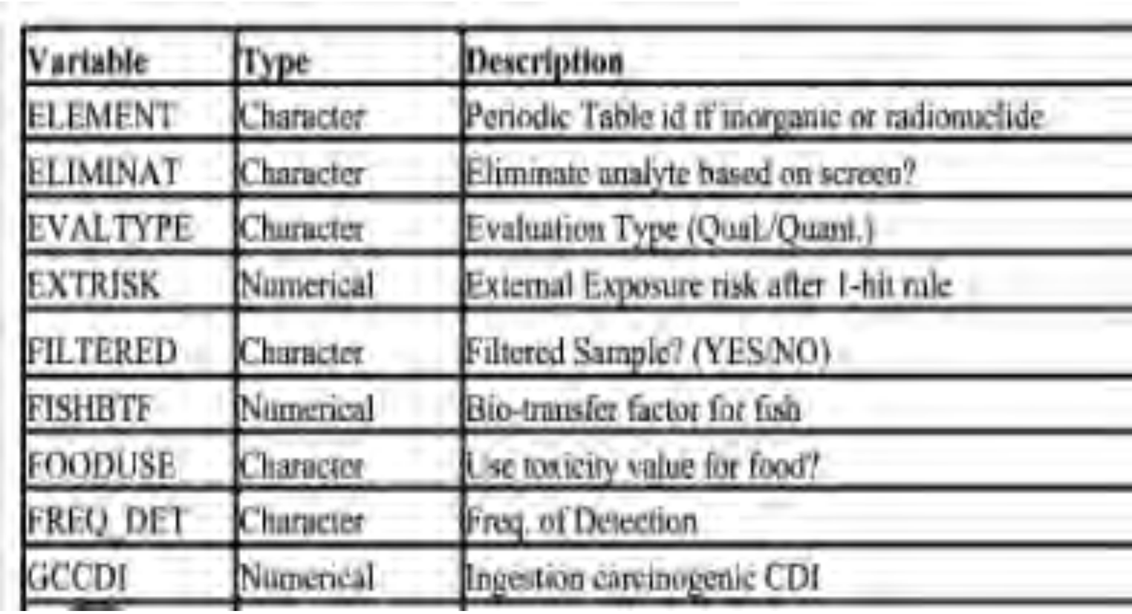
How do you create and maintain a Data Dictionary?

Most data modeling tools and database management systems (DBMS) have built-in, active data dictionaries the capable of generating and maintaining data dictionaries.

Data stewards - generate data dictionaries – include in grant application proposals

Best Practices:

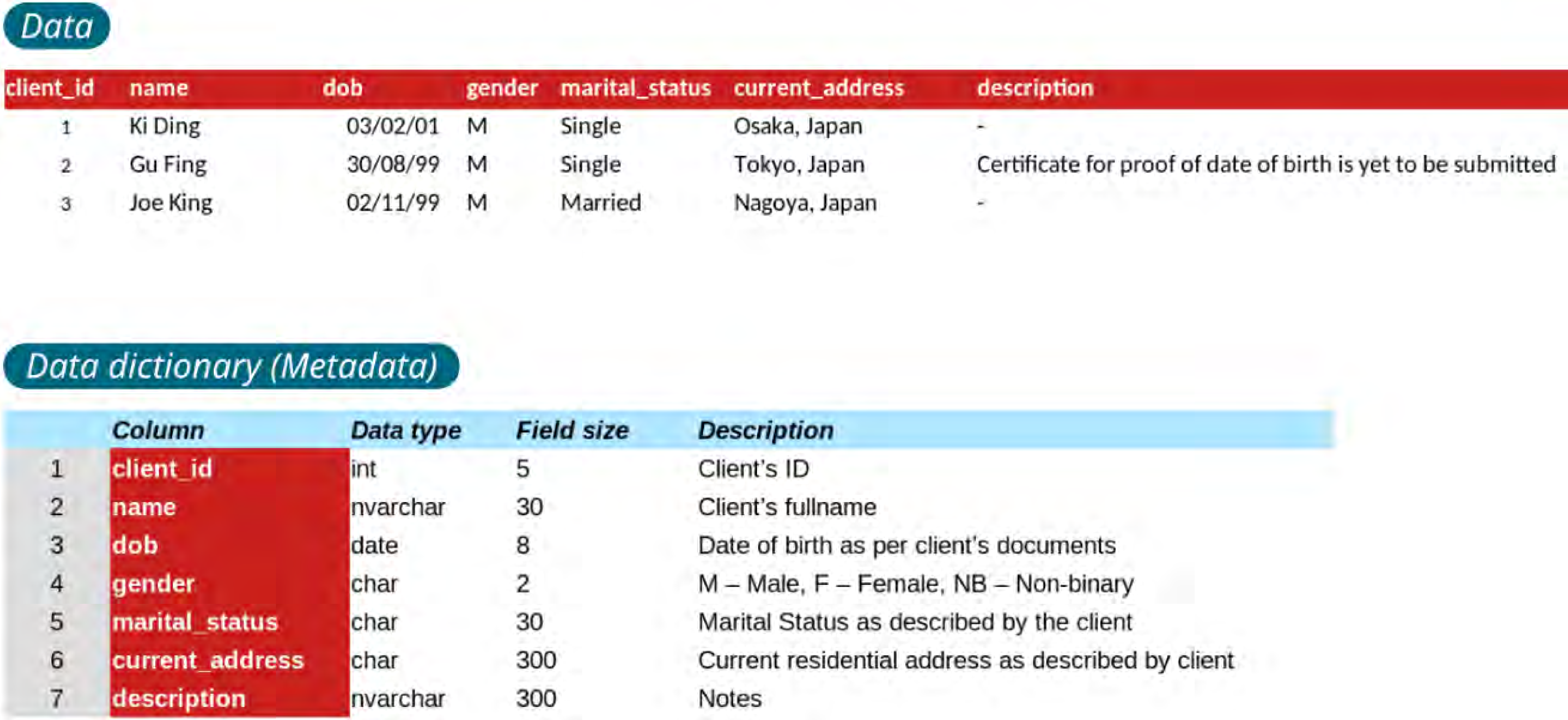
- Start building a data dictionary during the gathering business requirements phase.
- The data dictionary is a living document that must be regularly maintained.
- If utilizing erwin, then Consumers should build the data dictionary from their data model using Report Designer.



Variable	Type	Description
ELEMENT	Character	Periodic Table id if inorganic or radionuclide
ELIMINAT	Character	Eliminate analyte based on screen?
EVALTYPE	Character	Evaluation Type (Qual./Quant.)
EXTRISK	Numerical	External Exposure risk after 1-hit rule
FILTERED	Character	Filtered Sample? (YES/NO)
FISHBTF	Numerical	Bio-transfer factor for fish
FOODUSE	Character	Use toxicity value for food?
FREQ_DET	Character	Freq. of Detection
GCCDI	Numerical	Ingestion carcinogenic CDI

Ques: Name / Variable: Name

Column name	Definition	Data type	Required
Name	This column refers to the first name of customers	String	Yes



Data

client_id	name	dob	gender	marital_status	current_address	description
1	Ki Ding	03/02/01	M	Single	Osaka, Japan	-
2	Gu Fing	30/08/99	M	Single	Tokyo, Japan	Certificate for proof of date of birth is yet to be submitted
3	Joe King	02/11/99	M	Married	Nagoya, Japan	-

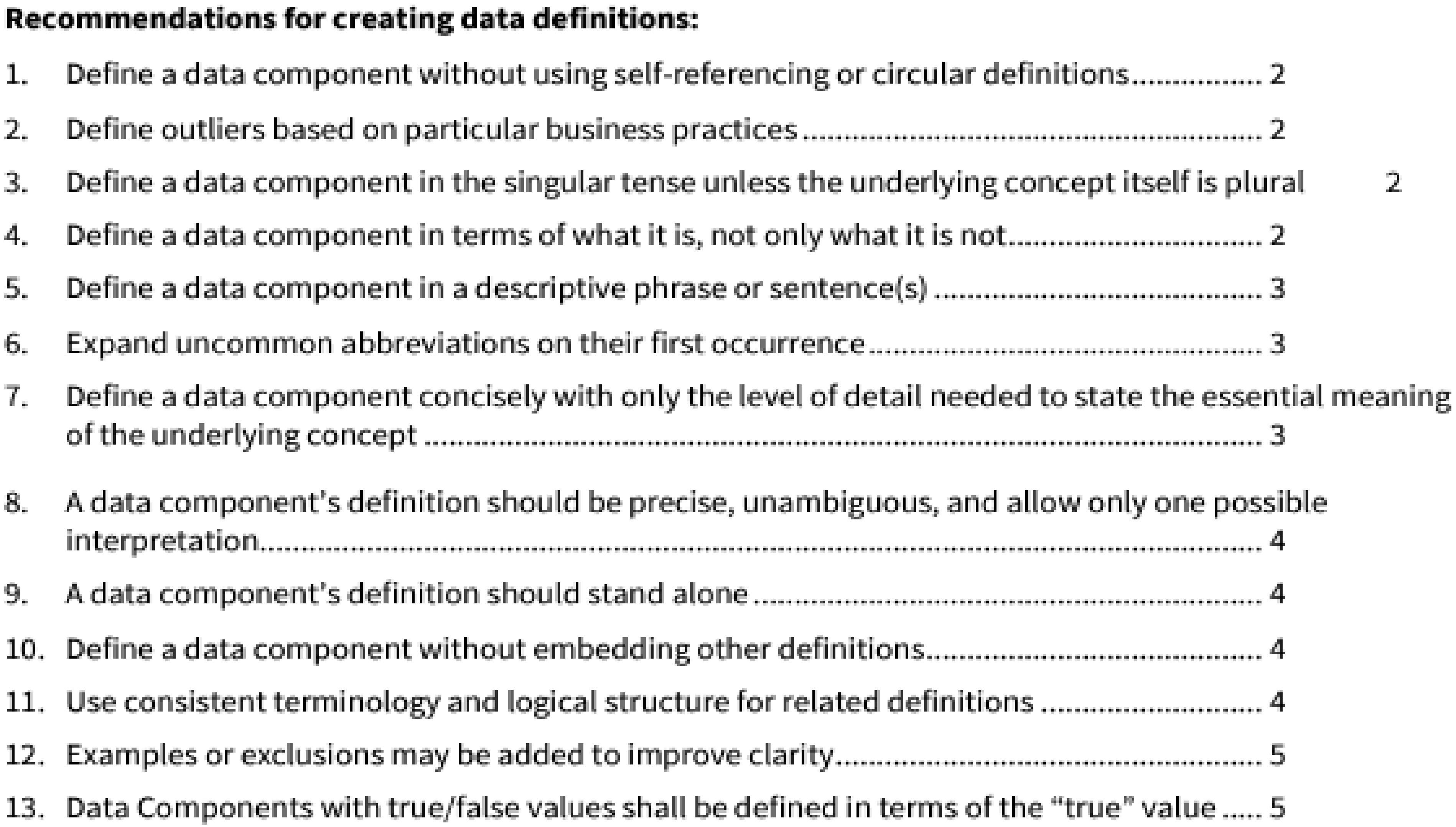
Data dictionary (Metadata)

	Column	*Data type*	*Field size*	*Description*
1	client_id	int	5	Client's ID
2	name	nvarchar	30	Client's fullname
3	dob	date	8	Date of birth as per client's documents
4	gender	char	2	M – Male, F – Female, NB – Non-binary
5	marital_status	char	30	Marital Status as described by the client
6	current_address	char	300	Current residential address as described by client
7	description	nvarchar	300	Notes

https://en.wikipedia.org/wiki/Data_dictionary

Here is a non-exhaustive list of typical items found in a data dictionary for columns or fields:

- Entity or form name or their ID (EntityID or FormID). The group this field belongs to.
- Field name, such as RDBMS field name
- Displayed field title. May default to field name if blank.
- Field type (string, integer, date, etc.)
- Measures such as min and max values, display width, or number of decimal places. Different field types may interpret this differently. An alternative is to have different attributes depending on field type.
- Field display order or tab order
- Coordinates on screen (if a positional or grid-based UI)
- Default value
- Prompt type, such as drop-down list, combo-box, check-boxes, range, etc.
- Is-required (Boolean) - If 'true', the value can not be blank, null, or only white-spaces
- Is-read-only (Boolean)
- Reference table name, if a foreign key. Can be used for validation or selection lists.
- Various event handlers or references to. Example: "on-click", "on-validate", etc. See event-driven programming.
- Format code, such as a regular expression or COBOL-style "PIC" statements
- Description or synopsis
- Database index characteristics or specification

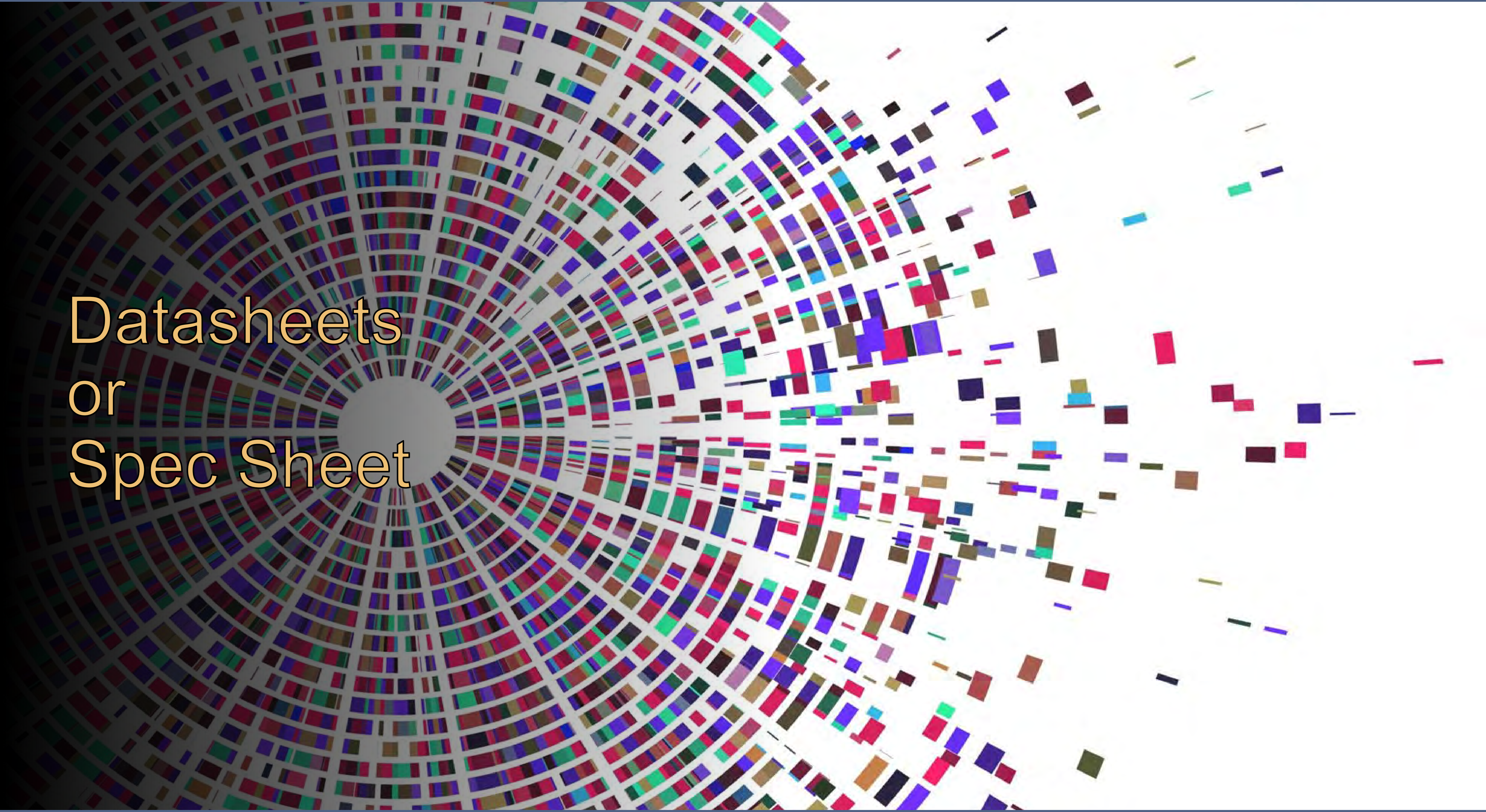


Recommendations for creating data definitions:

1. Define a data component without using self-referencing or circular definitions................ 2
2. Define outliers based on particular business practices 2
3. Define a data component in the singular tense unless the underlying concept itself is plural 2
4. Define a data component in terms of what it is, not only what it is not................................ 2
5. Define a data component in a descriptive phrase or sentence(s) 3
6. Expand uncommon abbreviations on their first occurrence 3
7. Define a data component concisely with only the level of detail needed to state the essential meaning of the underlying concept 3
8. A data component's definition should be precise, unambiguous, and allow only one possible interpretation.. 4
9. A data component's definition should stand alone 4
10. Define a data component without embedding other definitions.. 4
11. Use consistent terminology and logical structure for related definitions 4
12. Examples or exclusions may be added to improve clarity.. 5
13. Data Components with true/false values shall be defined in terms of the “true” value 5

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.cms.gov/files/document/data-definition-quick-reference-guide.pdf

Datasheets
or
Spec Sheet



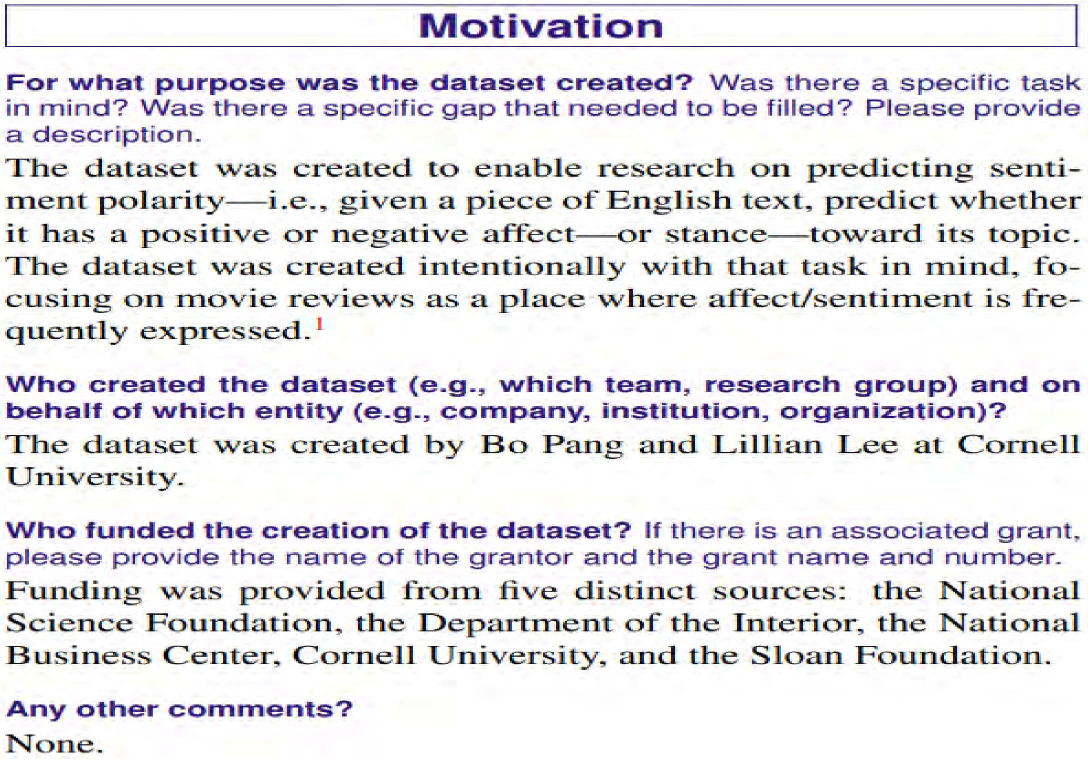
Datasheet

What is it?

How it is structured?

A datasheet is a document consisting of a series of questions/answers that is intended to document motivation, composition, collection process, recommended uses, etc. for a dataset.

A datasheet is a document, printed or electronic that provides details/characteristics about a product- summarizes performance



Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

7 sections and 56 questions

1. “**Motivation**”: Reasons for creating the dataset, funding source, etc. 4 questions
2. “**Composition**” : Describe the content of the dataset, de-identification level, etc. 16 question
3. “**Collection Process**” : Describe the data collection process. 12 questions
4. “**Preprocessing/cleaning/labeling**” : Describe data processing. 4 questions
5. “**Uses**” : Specify tasks for which the dataset should and should not be used. 6 questions
6. **“Distribution”**: Describe the dataset distribution/sharing process. 7 questions
7. “**Maintenance”:** Communicate plan for maintaining the dataset. 7 questions

Data sheet template

Relentlessly Focus on the Essential Information – Set up a template
Most datasheets are short. Use concise statements to describe your product. What is it? What does it do? How can it be used?

Compose Headlines and Sub-Heads to Summarize Your Main Points
Readers will scan these first, so make them succinct and be sure they encapsulate your main content points. (use bullet points)

Summary-Include a Product Definition on the First Page. Include a brief (two sentences or less) product definition at the top of the first page, including how it solves your audience's high-level problem.

Specifications Describe what a particular part needs and can do. Explain every part of your product or service in this section.

Images/Graphs Can be models, diagrams, pictures – anything that helps the reader

CLEAR AND CONCISE COPY THAT QUICKLY DELIVERS THE KEY POINTS AND THEN UTILIZE A FORMAT THAT HELPS THE READER FOCUS ON WHAT MATTERS TO THEM.

https://www.template.net/editable/datasheets

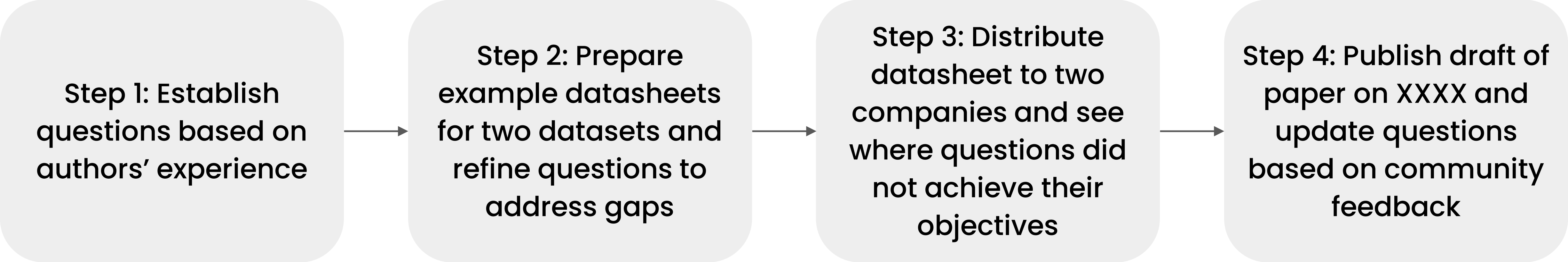
https://www.pragmaticinstitute.com/resources/articles/product/how-to-write-a-kick-butt-product-datasheet/

Datasheet
Timeline

Start: ?	It is not specified when this effort was started
March 2018	Preprint v1 published on XXXX https://doi.org/10.48550/arXiv.1803.09010
2018-2021	Several revised versions published on XXXX
December 2021	Final version published in Communications of the ACM https://doi.org/10.1145/3458723

Datasheet

How was it developed?



Step 1: Establish questions based on authors’ experience

Step 2: Prepare example datasheets for two datasets and refine questions to address gaps

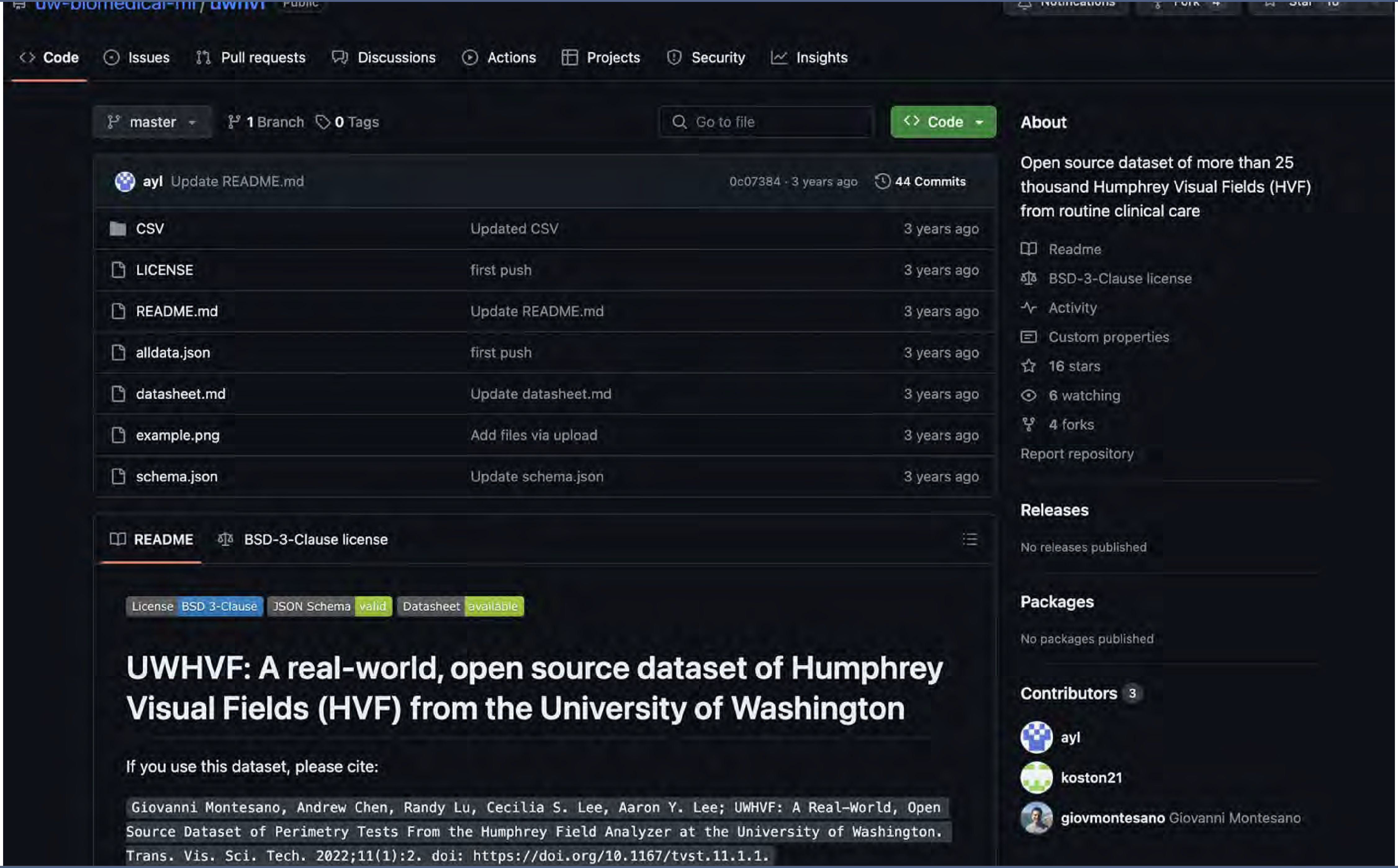
Step 3: Distribute datasheet to two companies and see where questions did not achieve their objectives

Step 4: Publish draft of paper on XXXX and update questions based on community feedback

Datasheet

Are there templates/tools available to create it?

- Note that the paper mentions: "We emphasize that the process of creating a datasheet is not intended to be automated. Although automated documentation processes are convenient, they run counter to our objective of encouraging dataset creators to carefully reflect on the process of creating, distributing, and maintaining a dataset."
- We could not find any tool that helps preparing a datasheet.
- Templates are available in different formats:
  - Markdown: https://github.com/fau-masters-collected-works-cgarbin/datasheet-for-dataset-template
  - Markdown: https://github.com/JRMeyer/markdown-datasheet-for-datasets/blob/master/DATASHEET.md
  - JSON: https://github.com/JRMeyer/json-datasheet-for-datasets/blob/main/DATASHEET.json
  - LaTex: https://github.com/AudreyBeard/Datasheets-for-Datasets-Template
  - Latex: https://www.overleaf.com/latex/templates/datasheet-for-dataset-template/jgqyyzyprxth



ayl Update README.md 0c07384 · 3 years ago 44 Commits

CSV	Updated CSV	3 years ago
LICENSE	first push	3 years ago
README.md	Update README.md	3 years ago
alldata.json	first push	3 years ago
datasheet.md	Update datasheet.md	3 years ago
example.png	Add files via upload	3 years ago
schema.json	Update schema.json	3 years ago

UWHVF: A real-world, open source dataset of Humphrey Visual Fields (HVF) from the University of Washington

If you use this dataset, please cite:

```
Giovanni Montesano, Andrew Chen, Randy Lu, Cecilia S. Lee, Aaron Y. Lee; UWHVF: A Real-World, Open
Source Dataset of Perimetry Tests From the Humphrey Field Analyzer at the University of Washington.
Trans. Vis. Sci. Tech. 2022;11(1):2. doi: https://doi.org/10.1167/tvst.11.1.1.
```

About

Open source dataset of more than 25 thousand Humphrey Visual Fields (HVF) from routine clinical care

Releases

No releases published

Packages

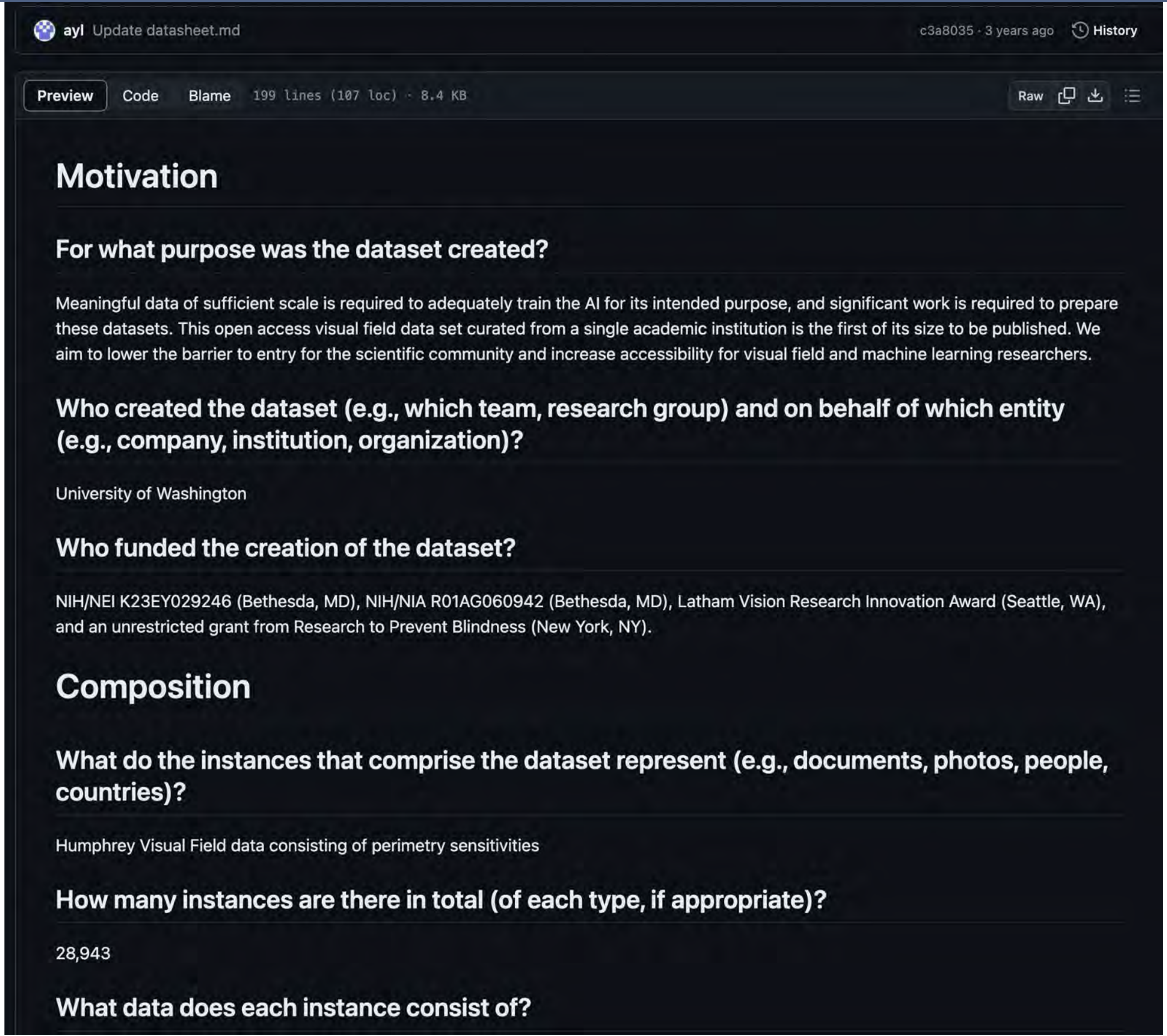
No packages published

Contributors 3

ayl

koston21

giovmontesano Giovanni Montesano



Motivation

For what purpose was the dataset created?

Meaningful data of sufficient scale is required to adequately train the AI for its intended purpose, and significant work is required to prepare these datasets. This open access visual field data set curated from a single academic institution is the first of its size to be published. We aim to lower the barrier to entry for the scientific community and increase accessibility for visual field and machine learning researchers.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

University of Washington

Who funded the creation of the dataset?

NIH/NEI K23EY029246 (Bethesda, MD), NIH/NIA R01AG060942 (Bethesda, MD), Latham Vision Research Innovation Award (Seattle, WA), and an unrestricted grant from Research to Prevent Blindness (New York, NY).

Composition

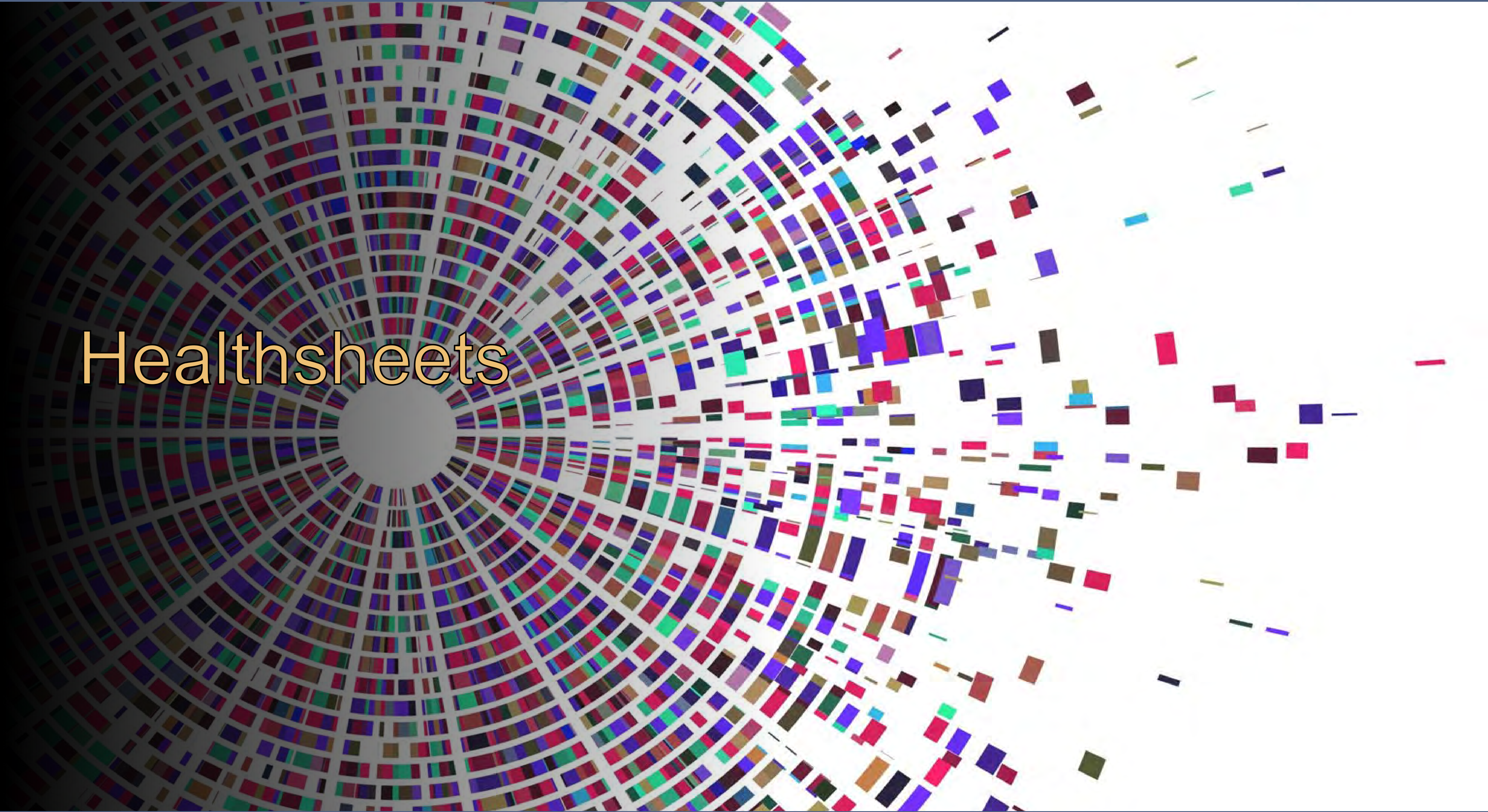
What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

Humphrey Visual Field data consisting of perimetry sensitivities

How many instances are there in total (of each type, if appropriate)?

28,943

What data does each instance consist of?



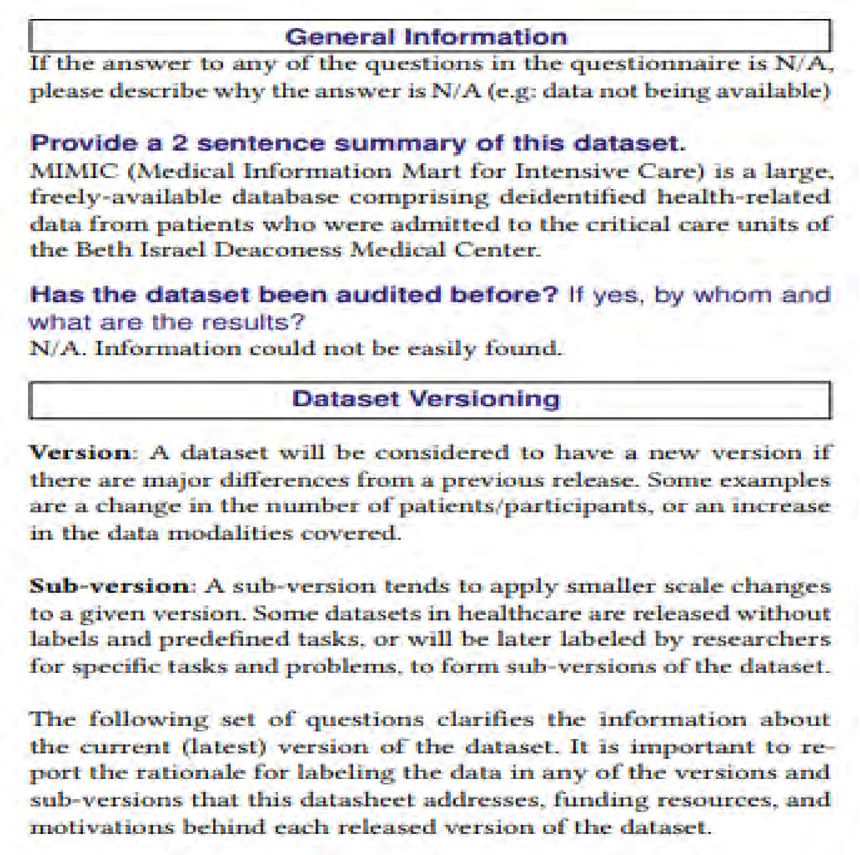
Healthsheets

Healthsheet

What is it?

How it is structured?

Healthsheet is a contextualized adaptation of the original datasheet questionnaire for health specific applications.



General Information

If the answer to any of the questions in the questionnaire is N/A, please describe why the answer is N/A (e.g: data not being available)

Provide a 2 sentence summary of this dataset.
MIMIC (Medical Information Mart for Intensive Care) is a large, freely-available database comprising deidentified health-related data from patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Center.

Has the dataset been audited before? If yes, by whom and what are the results?
N/A. Information could not be easily found.

Dataset Versioning

Version: A dataset will be considered to have a new version if there are major differences from a previous release. Some examples are a change in the number of patients/participants, or an increase in the data modalities covered.

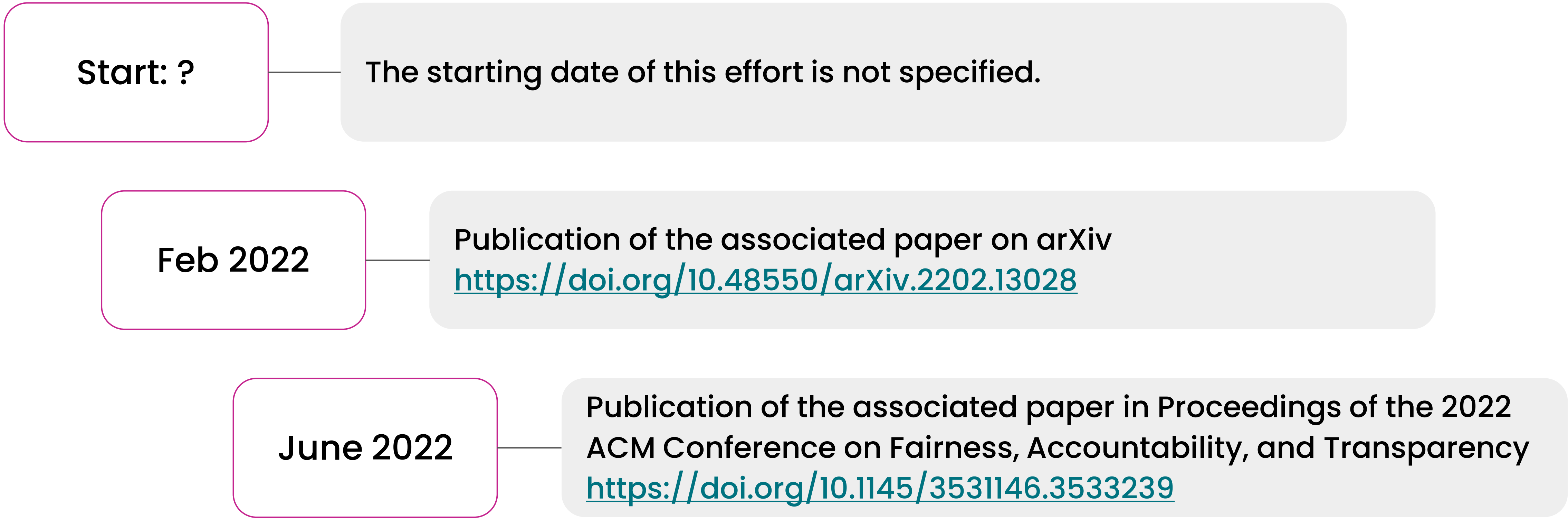
Sub-version: A sub-version tends to apply smaller scale changes to a given version. Some datasets in healthcare are released without labels and predefined tasks, or will be later labeled by researchers for specific tasks and problems, to form sub-versions of the dataset.

The following set of questions clarifies the information about the current (latest) version of the dataset. It is important to report the rationale for labeling the data in any of the versions and sub-versions that this datasheet addresses, funding resources, and motivations behind each released version of the dataset.

- General Information
- Dataset versioning
- Motivation
- Data composition
- Collection and use of demographic information
- Pre-processing, de-identification
- Labeling and subjectivity of labeling
- Collection process
- Uses
- Data distribution
- Maintenance

Healthsheet

Timeline



Start: ?

The starting date of this effort is not specified.

Feb 2022

Publication of the associated paper on arXiv
https://doi.org/10.48550/arXiv.2202.13028

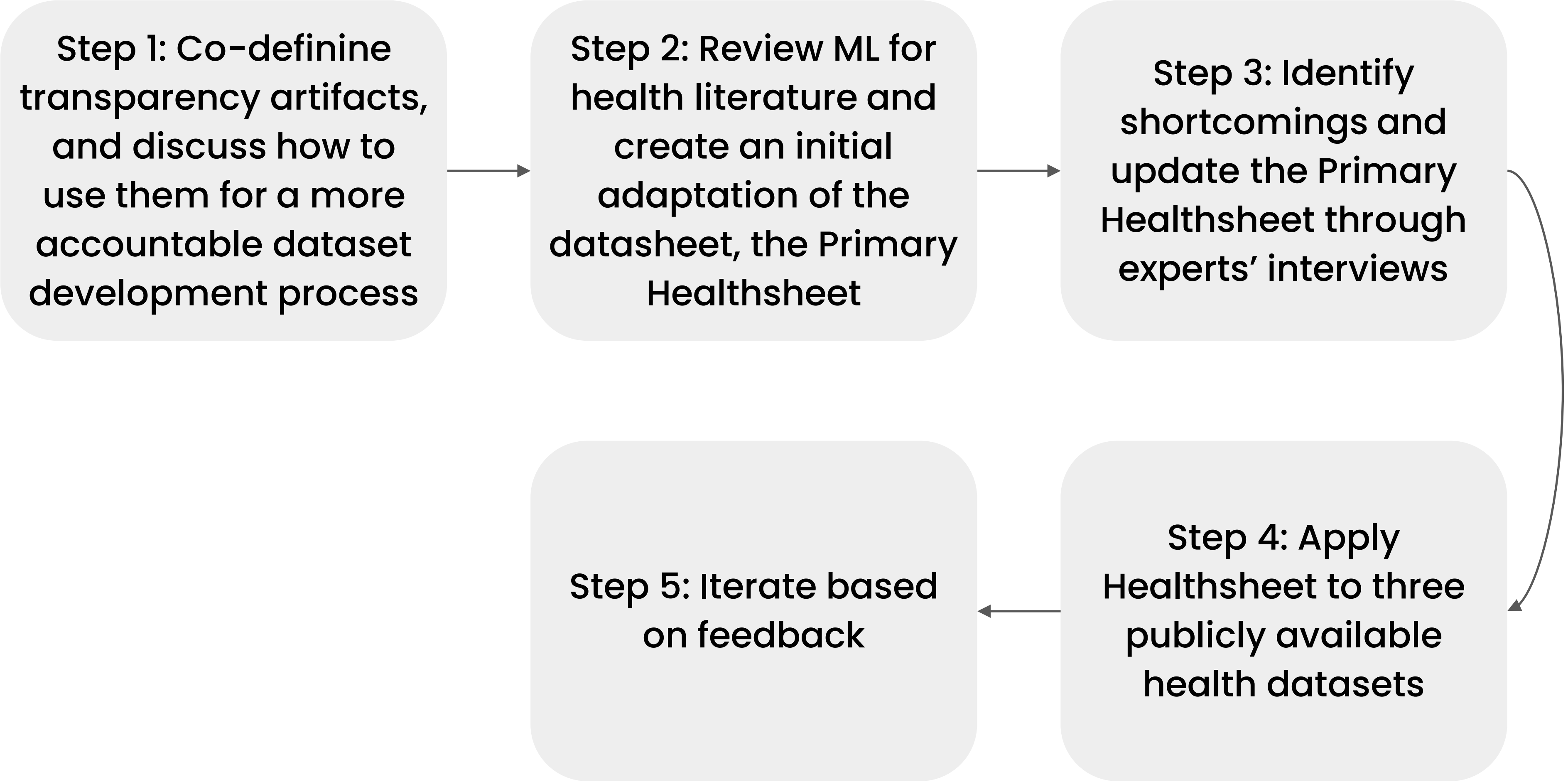
June 2022

Publication of the associated paper in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency
https://doi.org/10.1145/3531146.3533239

Example: Open Dataset of Flat-mounted Images for the Oxygen-induced Retinopathy Mouse Model: https://doi.org/10.6084/m9.figshare.23690973.v3

Healthsheet

How was it developed?



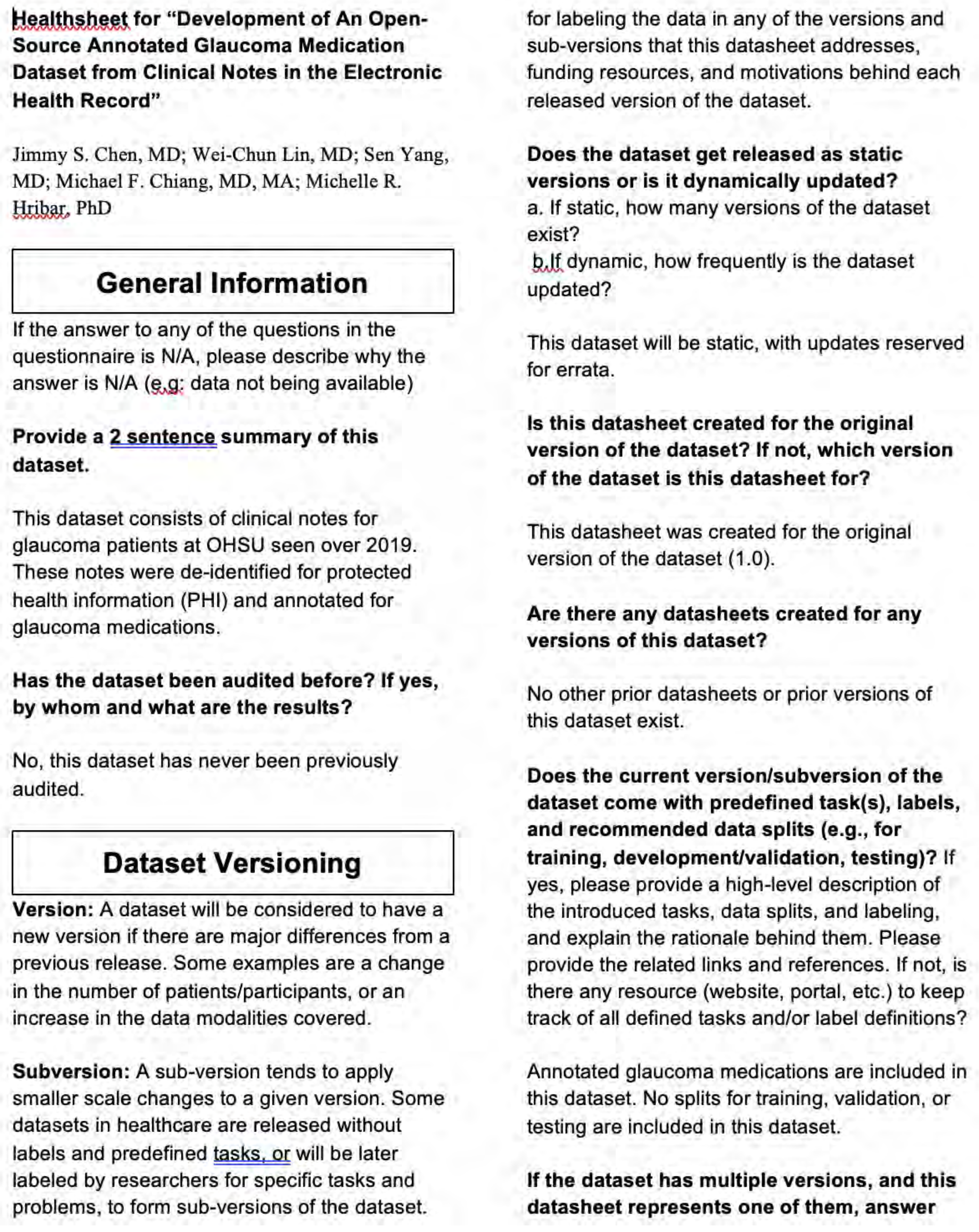
Step 1: Co-definine transparency artifacts, and discuss how to use them for a more accountable dataset development process

Step 2: Review ML for health literature and create an initial adaptation of the datasheet, the Primary Healthsheet

Step 3: Identify shortcomings and update the Primary Healthsheet through experts' interviews

Step 4: Apply Healthsheet to three publicly available health datasets

Step 5: Iterate based on feedback



Healthsheet for “Development of An Open-Source Annotated Glaucoma Medication Dataset from Clinical Notes in the Electronic Health Record”

Jimmy S. Chen, MD; Wei-Chun Lin, MD; Sen Yang, MD; Michael F. Chiang, MD, MA; Michelle R. Hribar, PhD

General Information

If the answer to any of the questions in the questionnaire is N/A, please describe why the answer is N/A (e.g: data not being available)

Provide a 2 sentence summary of this dataset.

This dataset consists of clinical notes for glaucoma patients at OHSU seen over 2019. These notes were de-identified for protected health information (PHI) and annotated for glaucoma medications.

Has the dataset been audited before? If yes, by whom and what are the results?

No, this dataset has never been previously audited.

Dataset Versioning

Version: A dataset will be considered to have a new version if there are major differences from a previous release. Some examples are a change in the number of patients/participants, or an increase in the data modalities covered.

Subversion: A sub-version tends to apply smaller scale changes to a given version. Some datasets in healthcare are released without labels and predefined tasks, or will be later labeled by researchers for specific tasks and problems, to form sub-versions of the dataset. for labeling the data in any of the versions and sub-versions that this datasheet addresses, funding resources, and motivations behind each released version of the dataset.

Does the dataset get released as static versions or is it dynamically updated?
a. If static, how many versions of the dataset exist?
b.If dynamic, how frequently is the dataset updated?

This dataset will be static, with updates reserved for errata.

Is this datasheet created for the original version of the dataset? If not, which version of the dataset is this datasheet for?

This datasheet was created for the original version of the dataset (1.0).

Are there any datasheets created for any versions of this dataset?

No other prior datasheets or prior versions of this dataset exist.

Does the current version/subversion of the dataset come with predefined task(s), labels, and recommended data splits (e.g., for training, development/validation, testing)? If yes, please provide a high-level description of the introduced tasks, data splits, and labeling, and explain the rationale behind them. Please provide the related links and references. If not, is there any resource (website, portal, etc.) to keep track of all defined tasks and/or label definitions?

Annotated glaucoma medications are included in this dataset. No splits for training, validation, or testing are included in this dataset.

If the dataset has multiple versions, and this datasheet represents one of them, answer

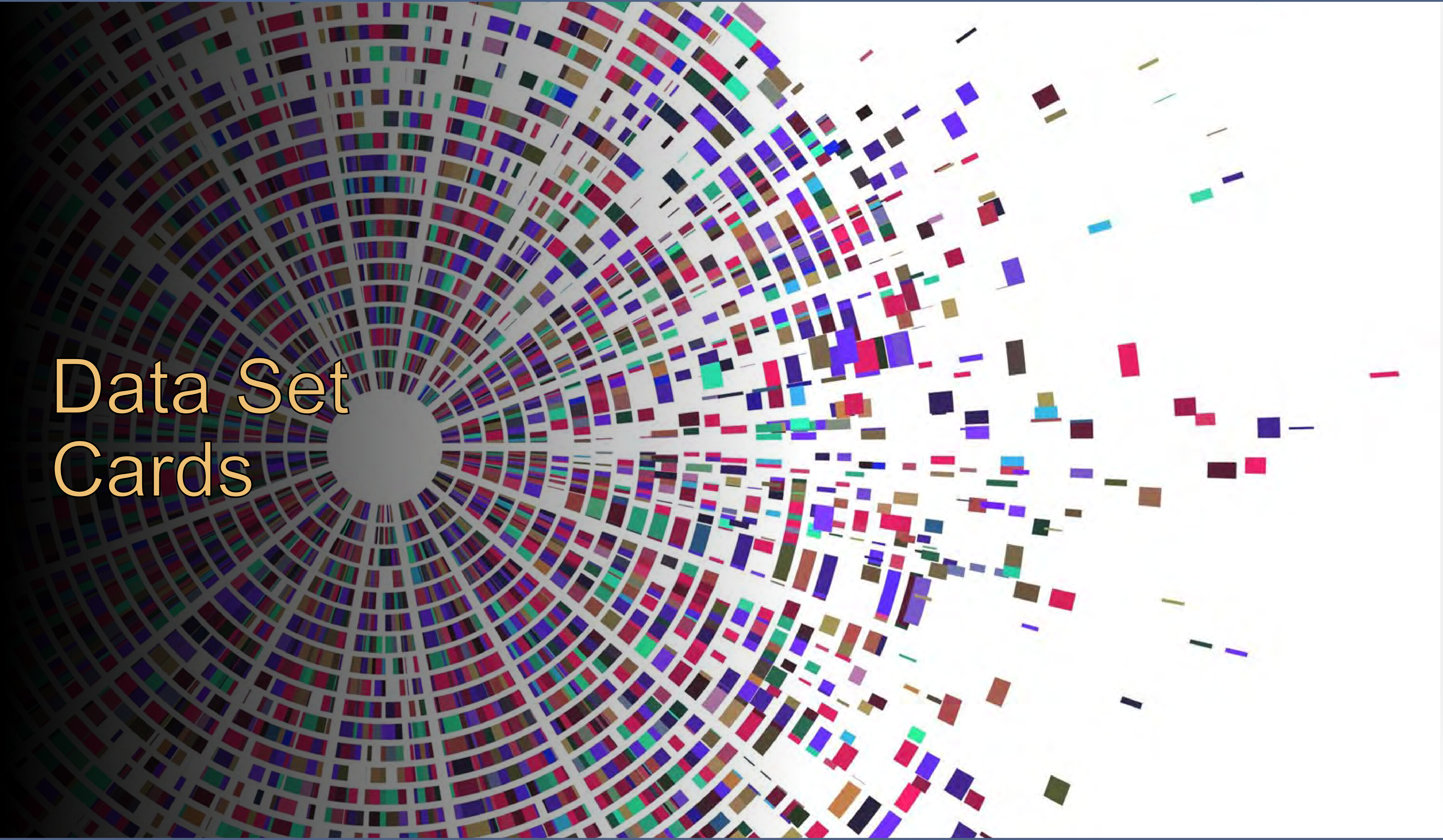
Healthsheet

How does it differ from datasheet?

	Healthsheet	**Datasheet**
Purpose and Focus	Tailored for healthcare datasets	Primarily designed for machine learning datasets
Context and industry	Targeted at healthcare industry	Applicable across various industries using ML
Elements and sections	- Dataset versioning- Accessibility- Demographic info- Racism/social conditions	- Motivation- Composition- Collection process- Fairness considerations
Use cases	Clinical research, healthcare applications.	Machine learning research, model development.
Interdisciplinary collaboration	Collaboration with healthcare professionals, ethicists.	Collaboration between data scientists and domain experts.
Depth of information	Detailed information on demographic factors, accessibility.	In-depth insights into dataset creation, biases.
Application scope	Clinical research, healthcare analytics.	General machine learning applications.

Slido

When you see a data set, what questions about the data pop into your head?



Data Set Cards

Data used in AI need Data Cards

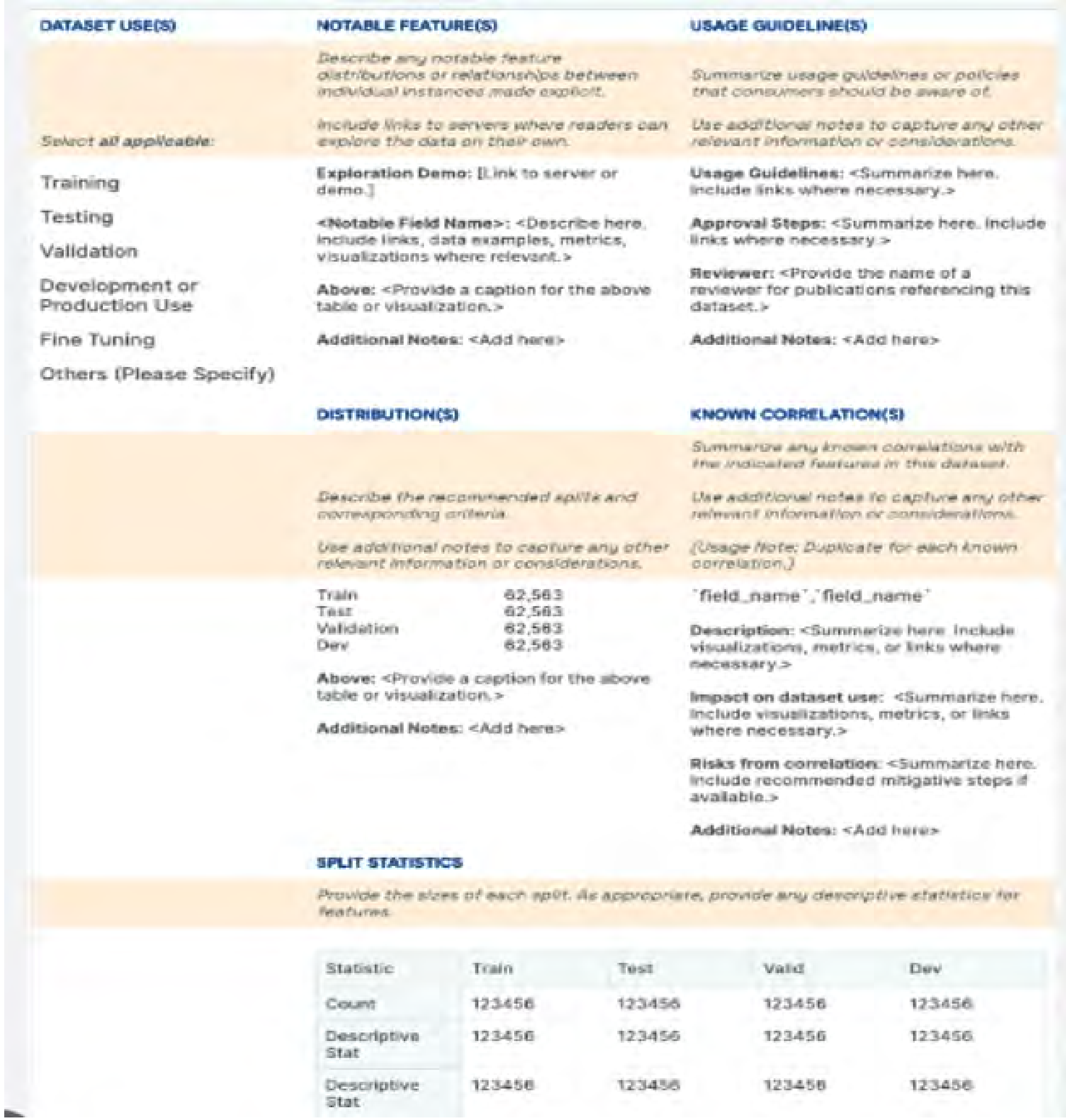
AI datasets typically present in rows and columns, with each row containing an observation. This observation can be in the form of text, an image, or a video.

It is not enough for your dataset to contain a large amount of well-structured data, unless these data have been labeled in the required way.

To construct AI dataset (and before doing data transformation):

1. Collect the raw data.
2. Identify feature and label sources.
3. Select a sampling strategy.
4. Split the data.

https://arxiv.org/abs/2204.01075



DATASET USE(S)

*Select **all applicable**:*

Training

Testing

Validation

Development or Production Use

Fine Tuning

Others (Please Specify)

NOTABLE FEATURE(S)

Describe any notable feature distributions or relationships between individual instances made explicit.

Include links to servers where readers can explore the data on their own.

Exploration Demo: [Link to server or demo.]

<Notable Field Name>: <Describe here. Include links, data examples, metrics, visualizations where relevant.>

Above: <Provide a caption for the above table or visualization.>

Additional Notes: <Add here>

USAGE GUIDELINE(S)

Summarize usage guidelines or policies that consumers should be aware of.

Use additional notes to capture any other relevant information or considerations.

Usage Guidelines: <Summarize here. Include links where necessary.>

Approval Steps: <Summarize here. Include links where necessary.>

Reviewer: <Provide the name of a reviewer for publications referencing this dataset.>

Additional Notes: <Add here>

DISTRIBUTION(S)

Describe the recommended splits and corresponding criteria.

Use additional notes to capture any other relevant information or considerations.

Train	62,563
Test	62,563
Validation	62,563
Dev	62,563

Above: <Provide a caption for the above table or visualization.>

Additional Notes: <Add here>

KNOWN CORRELATION(S)

Summarize any known correlations with the indicated features in this dataset.

Use additional notes to capture any other relevant information or considerations.

(Usage Note: Duplicate for each known correlation.)

`field_name`, `field_name`

Description: <Summarize here. Include visualizations, metrics, or links where necessary.>

Impact on dataset use: <Summarize here. Include visualizations, metrics, or links where necessary.>

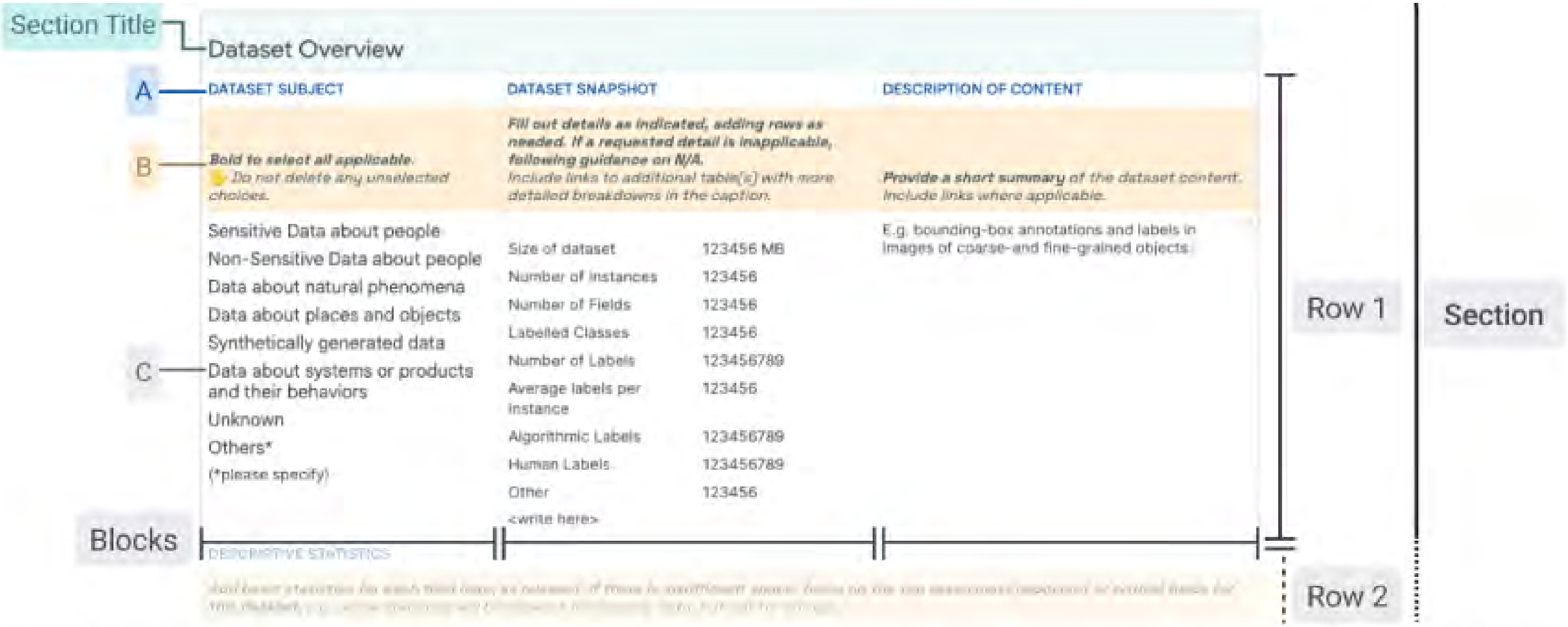
Risks from correlation: <Summarize here. Include recommended mitigative steps if available.>

Additional Notes: <Add here>

SPLIT STATISTICS

Provide the sizes of each split. As appropriate, provide any descriptive statistics for features.

Statistic	Train	Test	Valid	Dev
Count	123456	123456	123456	123456
Descriptive Stat	123456	123456	123456	123456
Descriptive Stat	123456	123456	123456	123456



Dataset Overview

DATASET SUBJECT

Bold to select all applicable.
Do not delete any unselected choices.

Sensitive Data about people
Non-Sensitive Data about people
Data about natural phenomena
Data about places and objects
Synthetically generated data
Data about systems or products and their behaviors
Unknown
Others*
(*please specify)

DATASET SNAPSHOT

Fill out details as indicated, adding rows as needed. If a requested detail is inapplicable, following guidance on N/A.
Include links to additional table(s) with more detailed breakdowns in the caption.

Size of dataset	123456 MB
Number of instances	123456
Number of Fields	123456
Labeled Classes	123456
Number of Labels	123456789
Average labels per instance	123456
Algorithmic Labels	123456789
Human Labels	123456789
Other	123456
<write here>	

DESCRIPTION OF CONTENT

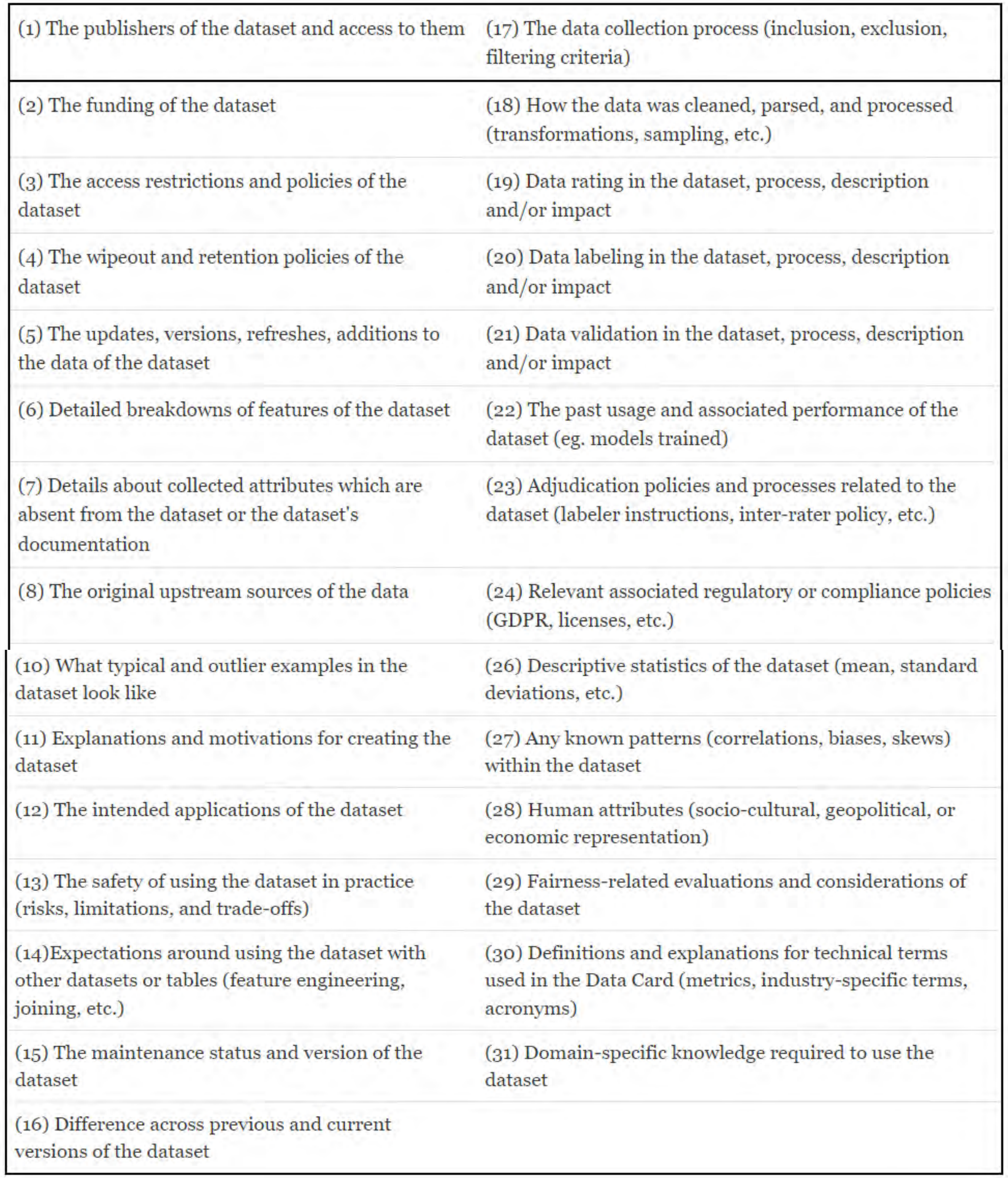
Provide a short summary *of the dataset content. Include links where applicable.*

E.g. bounding-box annotations and labels in images of coarse-and fine-grained objects.

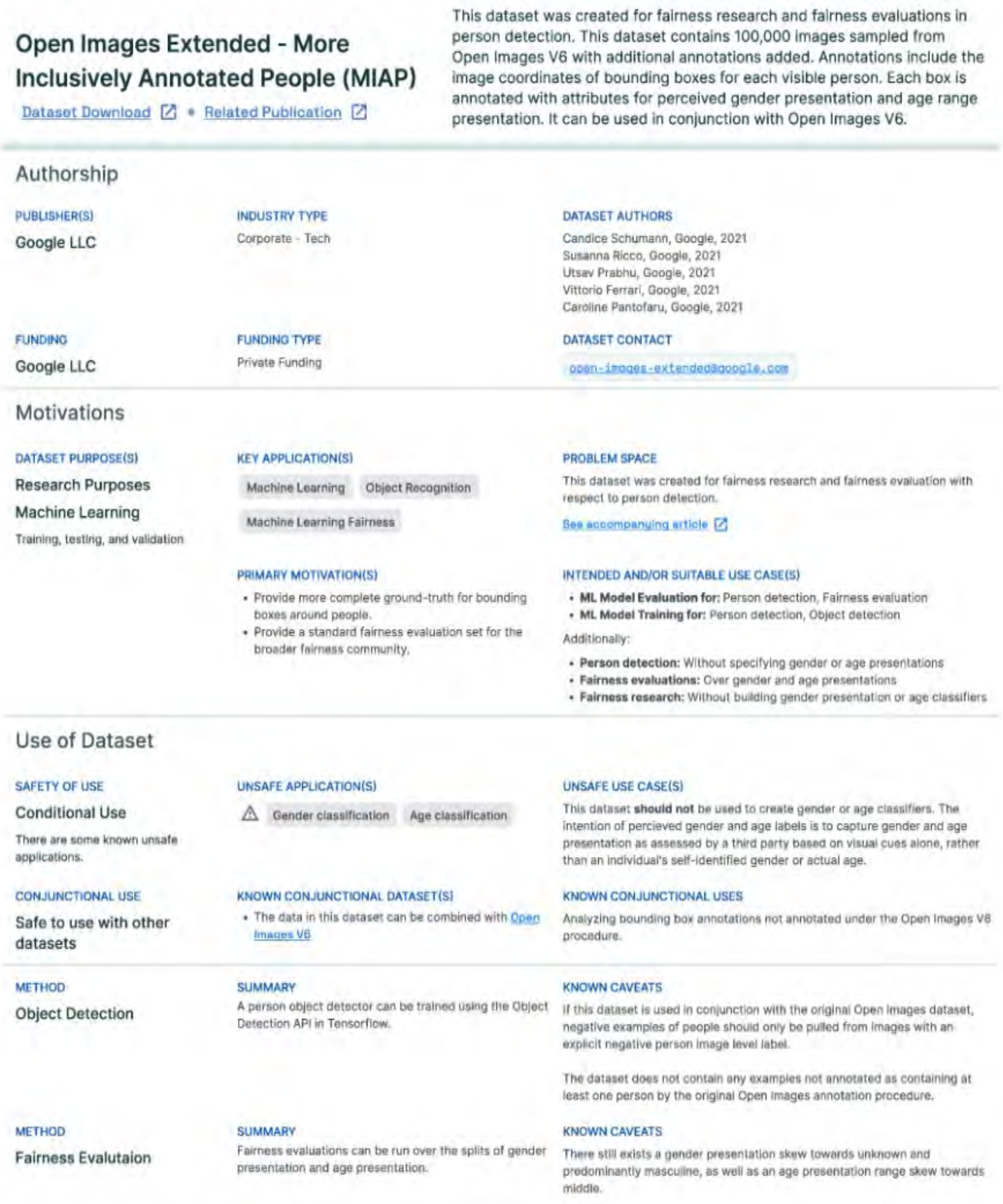
DESCRIPTIVE STATISTICS

Figure 2: A Data Card Template Section: This section is titled ”Dataset Overview”, and contains two rows. The first row has three blocks, whereas the second row spans the entire width of the section. Blocks contain (A) A Title, (B) A prompting question, and (C) an answer input space populated with predetermined choices or suggested answer structures.

Blocks are arranged thematically and hierarchically on a grid to enable an “*overview first, zoom-and-filter, details-on-demand*” [27] presentation of the dataset, to accomplish principle (**P4**). In our template, blocks with related questions are organized into *rows*, and rows are stacked to create *sections* using meaningful and descriptive titles (Figure 2). Each row is thematically self-contained so readers can effectively navigate multiple facets of a dataset in a Data Card. Answers increase in both detail and specificity across columns in the direction of the language in which the Data Card is written, allowing readers to find information at the appropriate fidelity for



(1) The publishers of the dataset and access to them	(17) The data collection process (inclusion, exclusion, filtering criteria)
(2) The funding of the dataset	(18) How the data was cleaned, parsed, and processed (transformations, sampling, etc.)
(3) The access restrictions and policies of the dataset	(19) Data rating in the dataset, process, description and/or impact
(4) The wipeout and retention policies of the dataset	(20) Data labeling in the dataset, process, description and/or impact
(5) The updates, versions, refreshes, additions to the data of the dataset	(21) Data validation in the dataset, process, description and/or impact
(6) Detailed breakdowns of features of the dataset	(22) The past usage and associated performance of the dataset (eg. models trained)
(7) Details about collected attributes which are absent from the dataset or the dataset's documentation	(23) Adjudication policies and processes related to the dataset (labeler instructions, inter-rater policy, etc.)
(8) The original upstream sources of the data	(24) Relevant associated regulatory or compliance policies (GDPR, licenses, etc.)
(10) What typical and outlier examples in the dataset look like	(26) Descriptive statistics of the dataset (mean, standard deviations, etc.)
(11) Explanations and motivations for creating the dataset	(27) Any known patterns (correlations, biases, skews) within the dataset
(12) The intended applications of the dataset	(28) Human attributes (socio-cultural, geopolitical, or economic representation)
(13) The safety of using the dataset in practice (risks, limitations, and trade-offs)	(29) Fairness-related evaluations and considerations of the dataset
(14)Expectations around using the dataset with other datasets or tables (feature engineering, joining, etc.)	(30) Definitions and explanations for technical terms used in the Data Card (metrics, industry-specific terms, acronyms)
(15) The maintenance status and version of the dataset	(31) Domain-specific knowledge required to use the dataset
(16) Difference across previous and current versions of the dataset	



Open Images Extended - More Inclusively Annotated People (MIAP)

Dataset Download • Related Publication

This dataset was created for fairness research and fairness evaluations in person detection. This dataset contains 100,000 images sampled from Open Images V6 with additional annotations added. Annotations include the image coordinates of bounding boxes for each visible person. Each box is annotated with attributes for perceived gender presentation and age range presentation. It can be used in conjunction with Open Images V6.

Authorship

PUBLISHER(S)
Google LLC

INDUSTRY TYPE
Corporate - Tech

DATASET AUTHORS
Candice Schumann, Google, 2021
Susanna Ricco, Google, 2021
Utsav Prabhu, Google, 2021
Vittorio Ferrari, Google, 2021
Caroline Pantofaru, Google, 2021

FUNDING
Google LLC

FUNDING TYPE
Private Funding

DATASET CONTACT
open-images-extended@google.com

Motivations

DATASET PURPOSE(S)
Research Purposes
Machine Learning
Training, testing, and validation

KEY APPLICATION(S)
Machine Learning
Object Recognition
Machine Learning Fairness

PROBLEM SPACE
This dataset was created for fairness research and fairness evaluation with respect to person detection.
See accompanying article

PRIMARY MOTIVATION(S)
- Provide more complete ground-truth for bounding boxes around people.
- Provide a standard fairness evaluation set for the broader fairness community.

INTENDED AND/OR SUITABLE USE CASE(S)
- **ML Model Evaluation for:** Person detection, Fairness evaluation
- **ML Model Training for:** Person detection, Object detection

Additionally:

- **Person detection:** Without specifying gender or age presentations
- **Fairness evaluations:** Over gender and age presentations
- **Fairness research:** Without building gender presentation or age classifiers

Use of Dataset

SAFETY OF USE
Conditional Use
There are some known unsafe applications.

UNSAFE APPLICATION(S)
⚠ Gender classification
Age classification

UNSAFE USE CASE(S)
This dataset **should not** be used to create gender or age classifiers. The intention of percieved gender and age labels is to capture gender and age presentation as assessed by a third party based on visual cues alone, rather than an individual's self-identified gender or actual age.

CONJUNCTIONAL USE
Safe to use with other datasets

KNOWN CONJUNCTIONAL DATASET(S)
- The data in this dataset can be combined with Open Images V6

KNOWN CONJUNCTIONAL USES
Analyzing bounding box annotations not annotated under the Open Images V6 procedure.

METHOD
Object Detection

SUMMARY
A person object detector can be trained using the Object Detection API in Tensorflow.

KNOWN CAVEATS
If this dataset is used in conjunction with the original Open Images dataset, negative examples of people should only be pulled from images with an explicit negative person image level label.

The dataset does not contain any examples not annotated as containing at least one person by the original Open Images annotation procedure.

METHOD
Fairness Evalutaion

SUMMARY
Fairness evaluations can be run over the splits of gender presentation and age presentation.

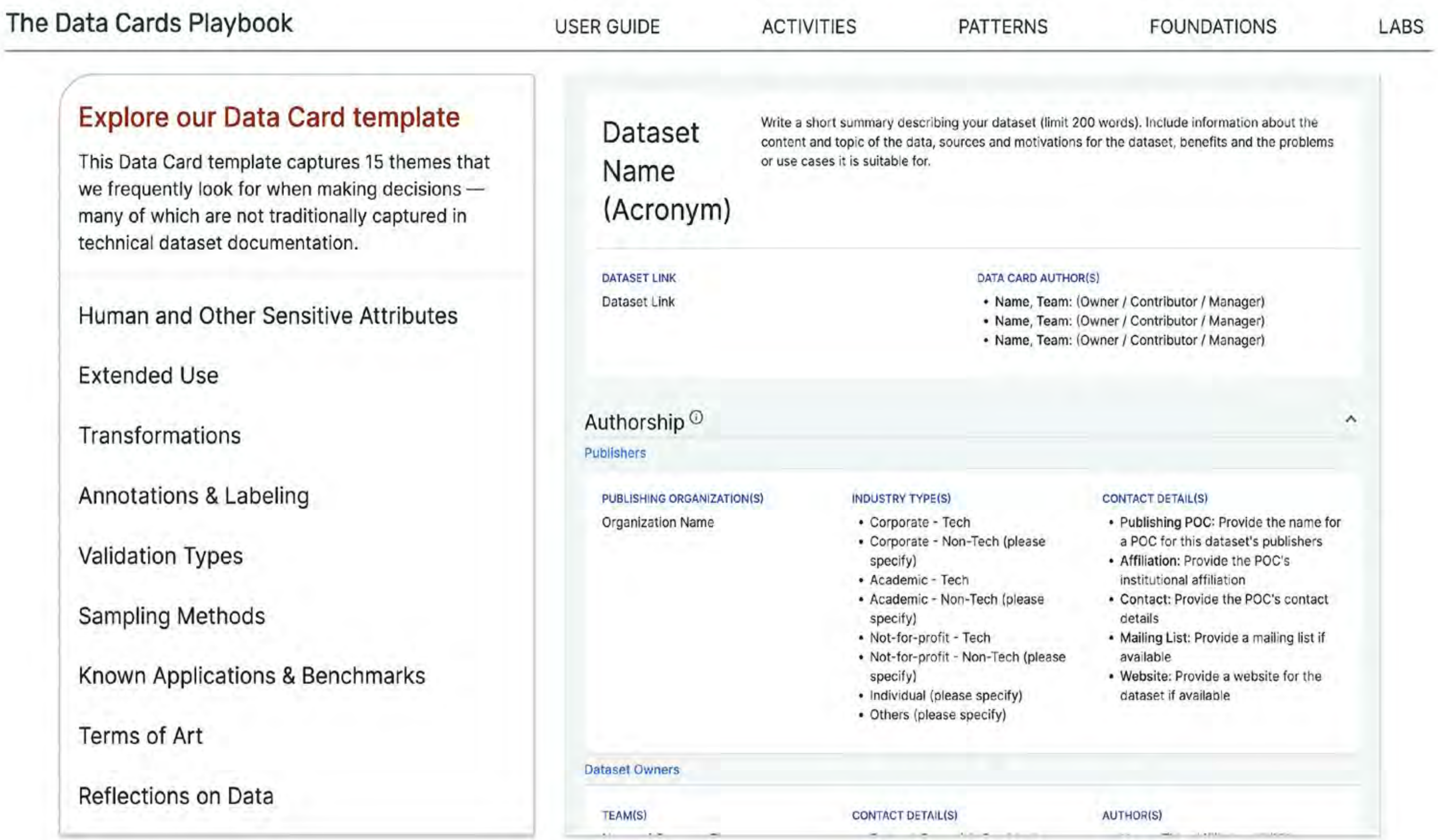
KNOWN CAVEATS
There still exists a gender presentation skew towards unknown and predominantly masculine, as well as an age presentation range skew towards middle.

https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533231

Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI

Mahima Pushkarna, Google Research, Canada, mahimap@google.com
Andrew Zaldivar, Google Research, USA, andrewzaldivar@google.com
Oddur Kjartansson, Google Research, United Kingdom, oddur@google.com

DOI: https://doi.org/10.1145/3531146.3533231
FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 2022



Explore our Data Card template

This Data Card template captures 15 themes that we frequently look for when making decisions — many of which are not traditionally captured in technical dataset documentation.

- Human and Other Sensitive Attributes
- Extended Use
- Transformations
- Annotations & Labeling
- Validation Types
- Sampling Methods
- Known Applications & Benchmarks
- Terms of Art
- Reflections on Data

Dataset Name (Acronym)

Write a short summary describing your dataset (limit 200 words). Include information about the content and topic of the data, sources and motivations for the dataset, benefits and the problems or use cases it is suitable for.

DATASET LINK
Dataset Link

DATA CARD AUTHOR(S)
- **Name, Team:** (Owner / Contributor / Manager)
- **Name, Team:** (Owner / Contributor / Manager)
- **Name, Team:** (Owner / Contributor / Manager)

Authorship

Publishers

PUBLISHING ORGANIZATION(S)
Organization Name

INDUSTRY TYPE(S)
- Corporate - Tech
- Corporate - Non-Tech (please specify)
- Academic - Tech
- Academic - Non-Tech (please specify)
- Not-for-profit - Tech
- Not-for-profit - Non-Tech (please specify)
- Individual (please specify)
- Others (please specify)

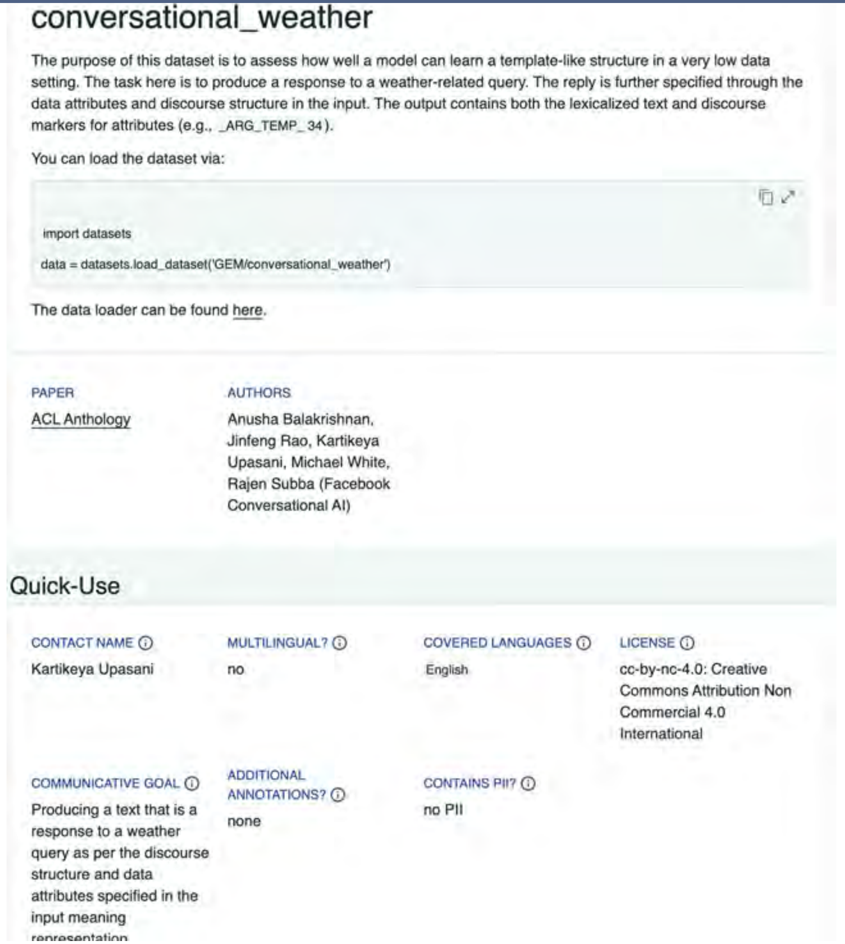
CONTACT DETAIL(S)
- **Publishing POC:** Provide the name for a POC for this dataset's publishers
- **Affiliation:** Provide the POC's institutional affiliation
- **Contact:** Provide the POC's contact details
- **Mailing List:** Provide a mailing list if available
- **Website:** Provide a website for the dataset if available

Dataset Owners

TEAM(S)

CONTACT DETAIL(S)

AUTHOR(S)



conversational_weather

The purpose of this dataset is to assess how well a model can learn a template-like structure in a very low data setting. The task here is to produce a response to a weather-related query. The reply is further specified through the data attributes and discourse structure in the input. The output contains both the lexicalized text and discourse markers for attributes (e.g., _ARG_TEMP_ 34).

You can load the dataset via:

```
import datasets
data = datasets.load_dataset('GEM/conversational_weather')
```

The data loader can be found here.

PAPER
ACL Anthology

AUTHORS
Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, Rajen Subba (Facebook Conversational AI)

Quick-Use

CONTACT NAME ⓘ
Kartikeya Upasani

MULTILINGUAL? ⓘ
no

COVERED LANGUAGES ⓘ
English

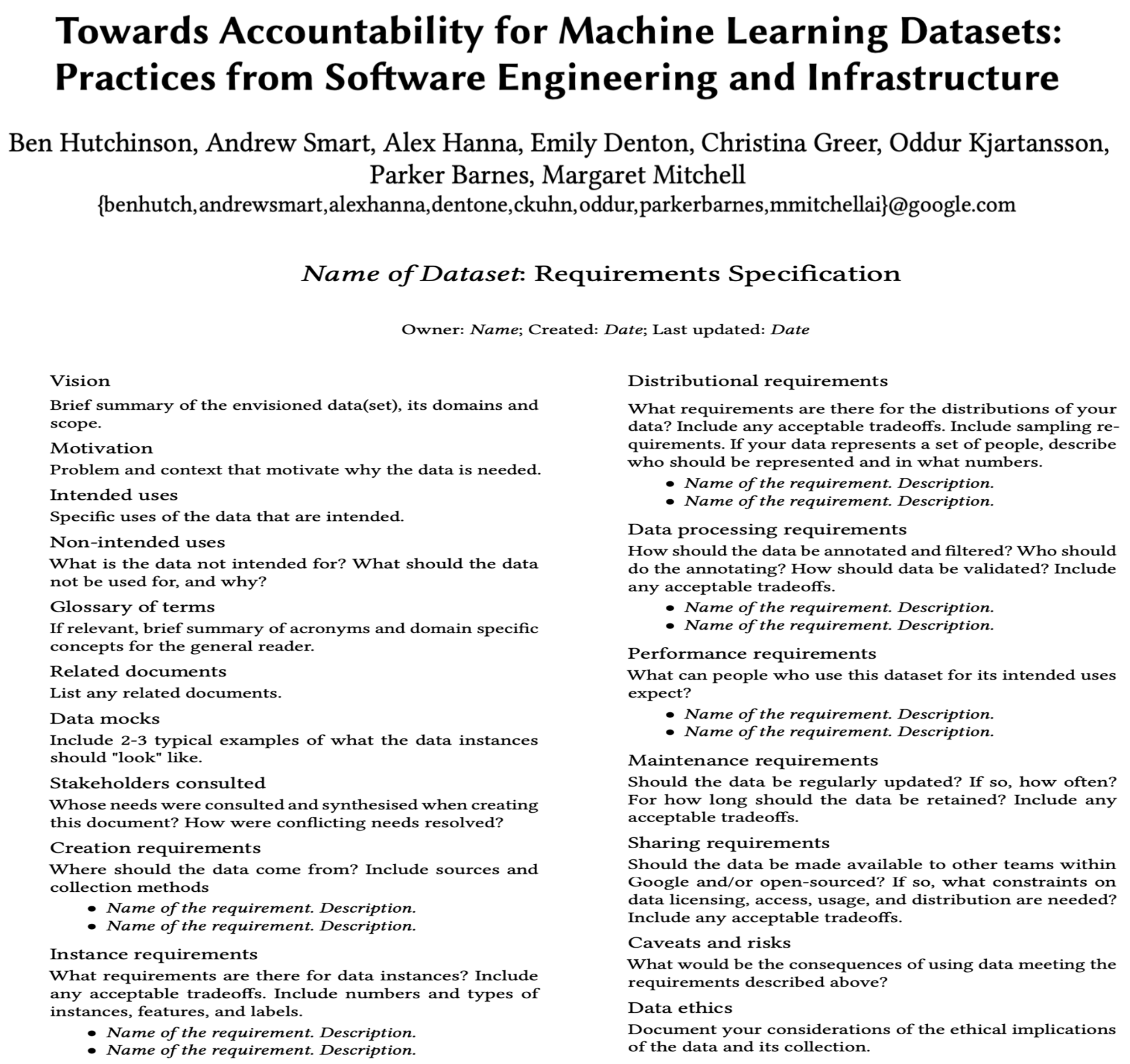
LICENSE ⓘ
cc-by-nc-4.0: Creative Commons Attribution Non Commercial 4.0 International

COMMUNICATIVE GOAL ⓘ
Producing a text that is a response to a weather query as per the discourse structure and data attributes specified in the input meaning representation

ADDITIONAL ANNOTATIONS? ⓘ
none

CONTAINS PII? ⓘ
no PII

https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533231



Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, Margaret Mitchell
{benhutch,andrewsmart,alexhanna,dentone,ckuhn,oddur,parkerbarnes,mmitchellai}@google.com

Name of Dataset: Requirements Specification

Owner: *Name*; Created: *Date*; Last updated: *Date*

Vision

Brief summary of the envisioned data(set), its domains and scope.

Motivation

Problem and context that motivate why the data is needed.

Intended uses

Specific uses of the data that are intended.

Non-intended uses

What is the data not intended for? What should the data not be used for, and why?

Glossary of terms

If relevant, brief summary of acronyms and domain specific concepts for the general reader.

Related documents

List any related documents.

Data mocks

Include 2-3 typical examples of what the data instances should "look" like.

Stakeholders consulted

Whose needs were consulted and synthesised when creating this document? How were conflicting needs resolved?

Creation requirements

Where should the data come from? Include sources and collection methods

- *Name of the requirement. Description.*
- *Name of the requirement. Description.*

Instance requirements

What requirements are there for data instances? Include any acceptable tradeoffs. Include numbers and types of instances, features, and labels.

- *Name of the requirement. Description.*
- *Name of the requirement. Description.*

Distributional requirements

What requirements are there for the distributions of your data? Include any acceptable tradeoffs. Include sampling requirements. If your data represents a set of people, describe who should be represented and in what numbers.

- *Name of the requirement. Description.*
- *Name of the requirement. Description.*

Data processing requirements

How should the data be annotated and filtered? Who should do the annotating? How should data be validated? Include any acceptable tradeoffs.

- *Name of the requirement. Description.*
- *Name of the requirement. Description.*

Performance requirements

What can people who use this dataset for its intended uses expect?

- *Name of the requirement. Description.*
- *Name of the requirement. Description.*

Maintenance requirements

Should the data be regularly updated? If so, how often? For how long should the data be retained? Include any acceptable tradeoffs.

Sharing requirements

Should the data be made available to other teams within Google and/or open-sourced? If so, what constraints on data licensing, access, usage, and distribution are needed? Include any acceptable tradeoffs.

Caveats and risks

What would be the consequences of using data meeting the requirements described above?

Data ethics

Document your considerations of the ethical implications of the data and its collection.

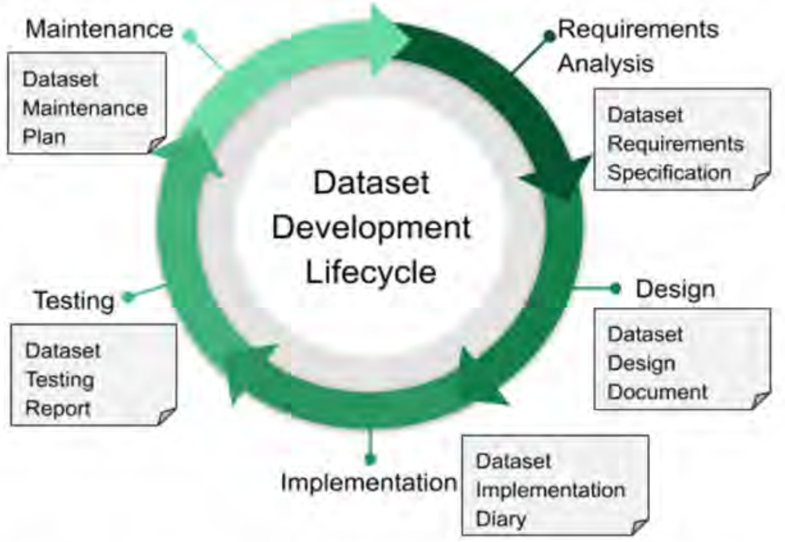


Figure 1: The Dataset Development Lifecycle requires documentation for each stage. See Table 3 for descriptions of each stage, and Table 1 for document types.

Appendix A

The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards

Sarah Holland[1]*, Ahmed Hosny[2]*, Sarah Newman[3], Joshua Joseph[4], and Kasia Chmielinski[1*†]

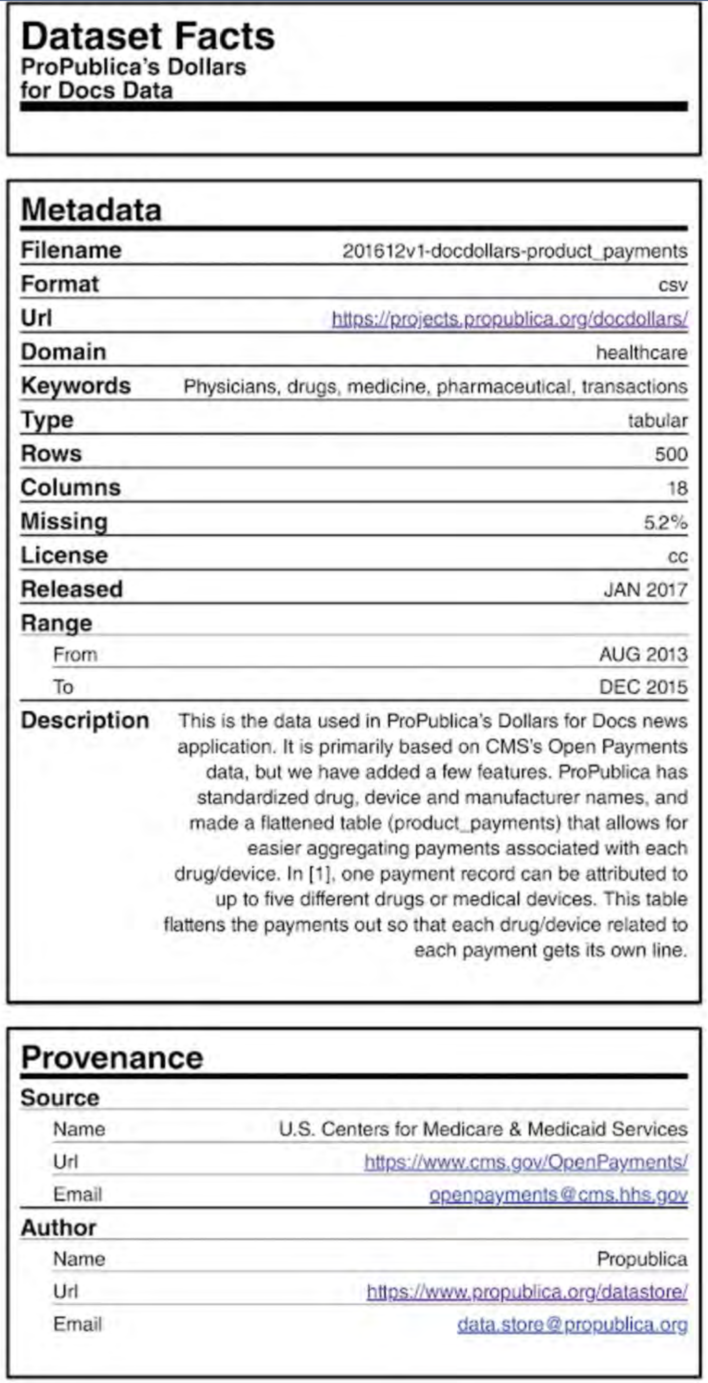
[1]Assembly, MIT Media Lab and Berkman Klein Center at Harvard University, [2]Dana-Farber Cancer Institute, Harvard Medical School, [3]metaLAB (at) Harvard, Berkman Klein Center for Internet & Society, Harvard University, [4]33x.ai

**authors contributed equally*

†nutrition@media.mit.edu

Module Name	Description	Contents
Metadata	Meta information. This module is the only required module. It represents the absolute minimum information to be presented	Filename, file format, URL, domain, keywords, type, dataset size, % of missing cells, license, release date, collection range, description
Provenance	Information regarding the origin and lineage of the dataset	Source and author contact information with version history
Variables	Descriptions of each variable (column) in the dataset	Textual descriptions
Statistics	Simple statistics for all variables, in addition to stratifications into ordinal, nominal, continuous, and discrete	Least/most frequent entries, min/max, median, mean,.etc
Pair Plots	Distributions and linear correlations between 2 chosen variables	Histograms and heatmaps
Probabilistic Model	Synthetic data generated using distribution hypotheses from which the data was drawn - leverages a probabilistic programming backend	Histograms and other statistical plots
Ground Truth Correlations	Linear correlations between a chosen variable in the dataset and variables from other datasets considered to be "ground truth", such as Census Data	Heatmaps

Table 1. Table illustrating 7 modules of the Dataset Nutrition Label, together with their description, role, and contents.



Dataset Facts

ProPublica's Dollars for Docs Data

Metadata

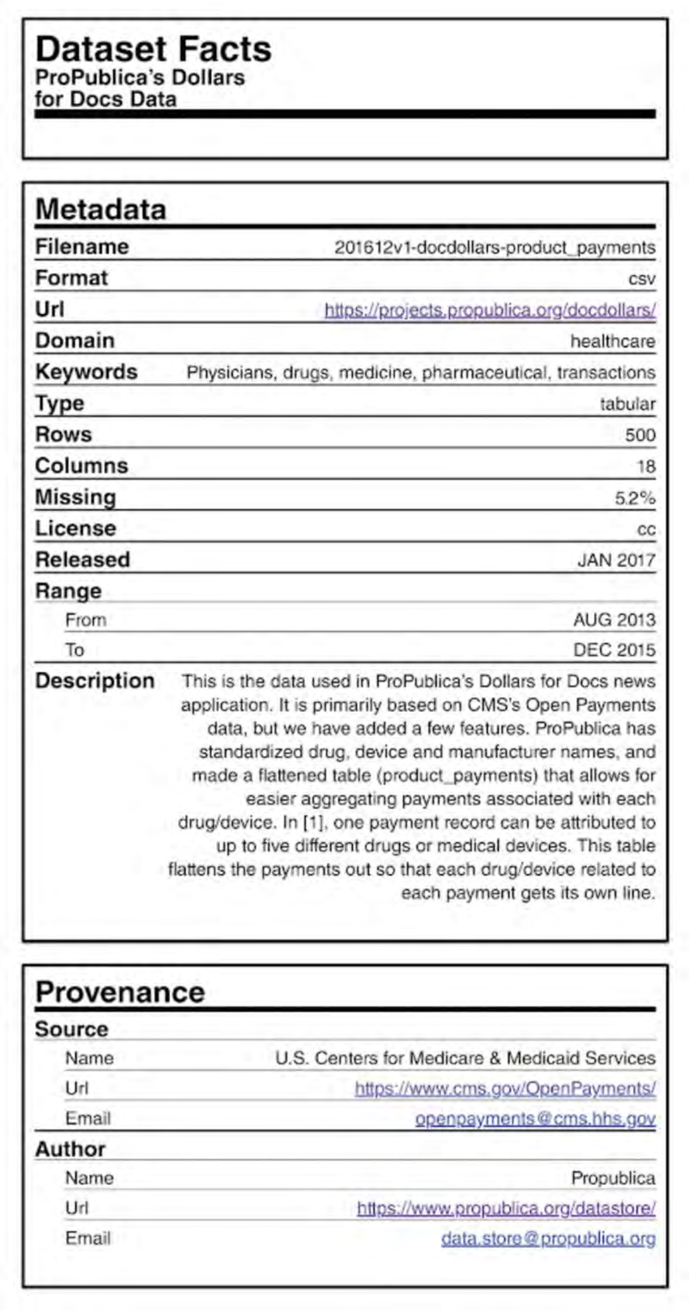
Filename	201612v1-docdollars-product_payments
Format	csv
Url	https://projects.propublica.org/docdollars/
Domain	healthcare
Keywords	Physicians, drugs, medicine, pharmaceutical, transactions
Type	tabular
Rows	500
Columns	18
Missing	5.2%
License	cc
Released	JAN 2017
Range	
From	AUG 2013
To	DEC 2015
Description	This is the data used in ProPublica's Dollars for Docs news application. It is primarily based on CMS's Open Payments data, but we have added a few features. ProPublica has standardized drug, device and manufacturer names, and made a flattened table (product_payments) that allows for easier aggregating payments associated with each drug/device. In [1], one payment record can be attributed to up to five different drugs or medical devices. This table flattens the payments out so that each drug/device related to each payment gets its own line.

Provenance

Source	
Name	U.S. Centers for Medicare & Medicaid Services
Url	https://www.cms.gov/OpenPayments/
Email	openpayments@cms.hhs.gov
Author	
Name	Propublica
Url	https://www.propublica.org/datastore/
Email	data.store@propublica.org

Slido

What are the benefits of using a food label approach to a data set? What is missing in this example?



Dataset Facts

ProPublica's Dollars for Docs Data

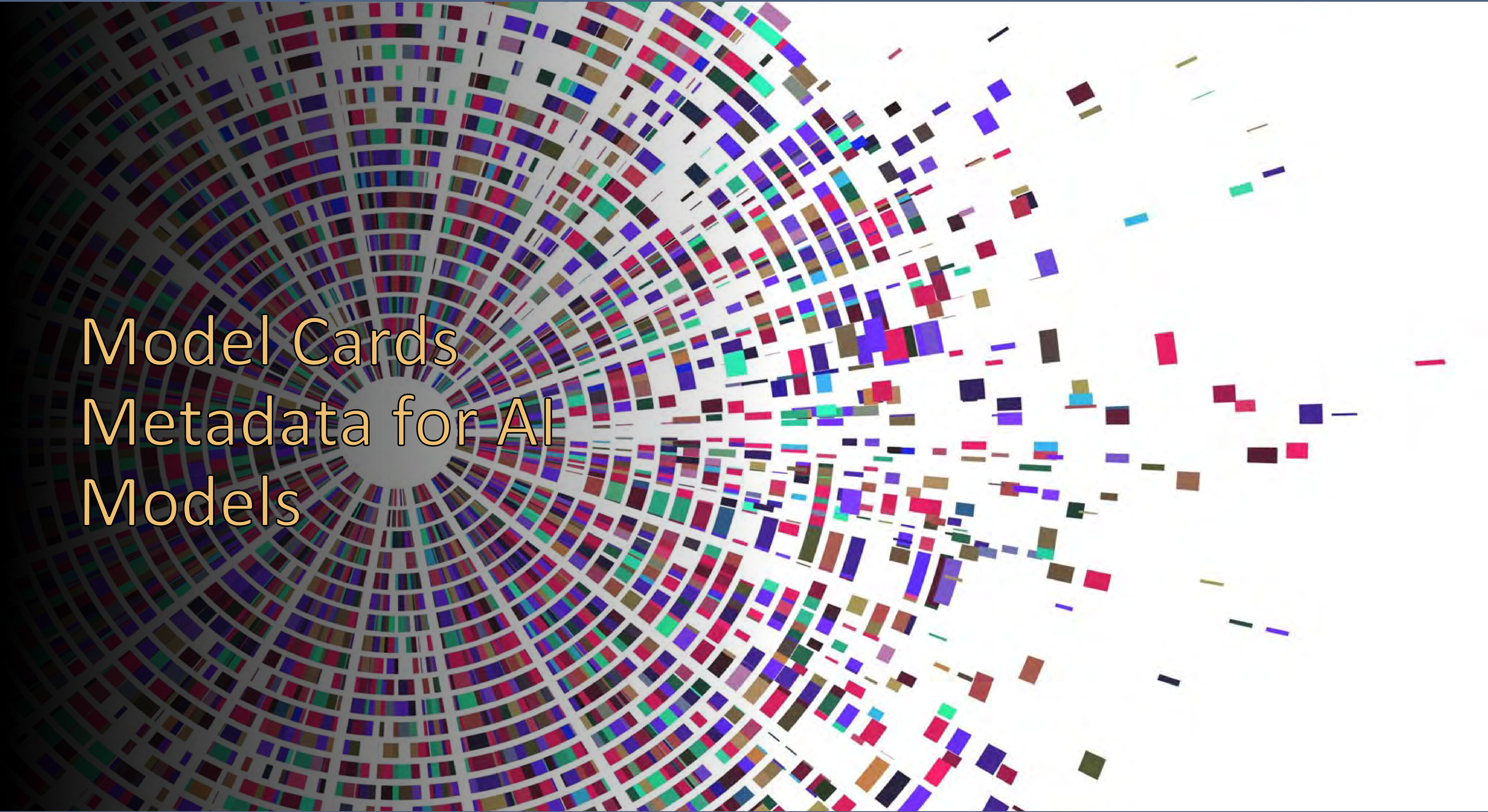
Metadata

Filename	201612v1-docdollars-product_payments
Format	csv
Url	https://projects.propublica.org/docdollars/
Domain	healthcare
Keywords	Physicians, drugs, medicine, pharmaceutical, transactions
Type	tabular
Rows	500
Columns	18
Missing	5.2%
License	cc
Released	JAN 2017
Range	
From	AUG 2013
To	DEC 2015
Description	This is the data used in ProPublica's Dollars for Docs news application. It is primarily based on CMS's Open Payments data, but we have added a few features. ProPublica has standardized drug, device and manufacturer names, and made a flattened table (product_payments) that allows for easier aggregating payments associated with each drug/device. In [1], one payment record can be attributed to up to five different drugs or medical devices. This table flattens the payments out so that each drug/device related to each payment gets its own line.

Provenance

Source	
Name	U.S. Centers for Medicare & Medicaid Services
Url	https://www.cms.gov/OpenPayments/
Email	openpayments@cms.hhs.gov
Author	
Name	Propublica
Url	https://www.propublica.org/datastore/
Email	data.store@propublica.org

Model Cards Metadata for AI Models

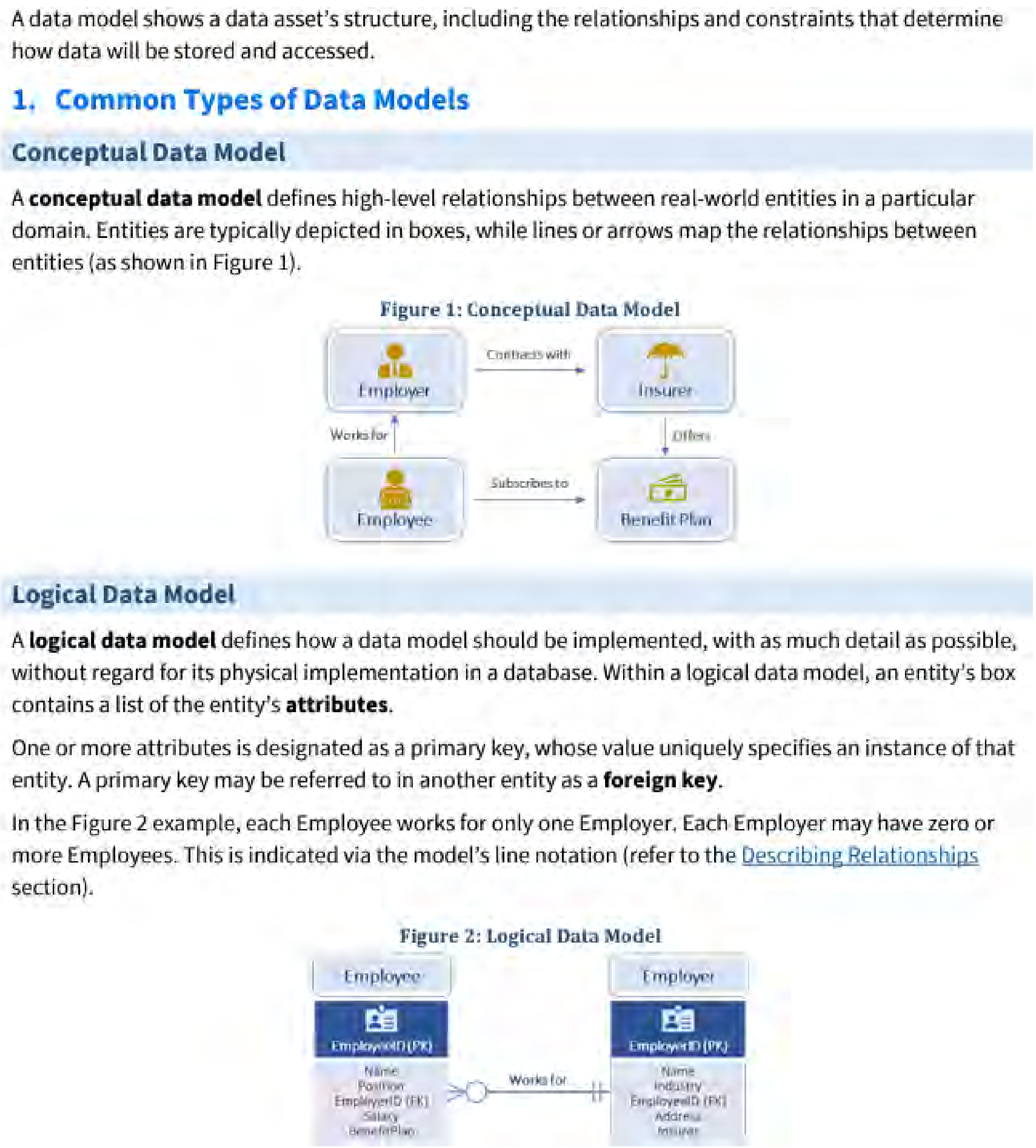


Model Cards Conveys Key AI/ML Information

A **model card** is a short document that provides key information about a machine learning model. Model cards increase transparency by communicating information about trained models to broad audiences.

Model cards - introduced in a 2019 paper - are one way for teams to communicate key information about their AI system to a broad audience. This information generally includes intended uses for the model, how the model works, and how the model performs in different situations.

For the Audience, a model card should strike a balance between being easy-to-understand and communicating important technical information. When writing a model card, you should consider your audience: the groups of people who are most likely to read your model card-varies according to the AI system's purpose. Most users are data scientist and AI researchers.



A data model shows a data asset's structure, including the relationships and constraints that determine how data will be stored and accessed.

1. Common Types of Data Models

Conceptual Data Model

A **conceptual data model** defines high-level relationships between real-world entities in a particular domain. Entities are typically depicted in boxes, while lines or arrows map the relationships between entities (as shown in Figure 1).

Figure 1: Conceptual Data Model

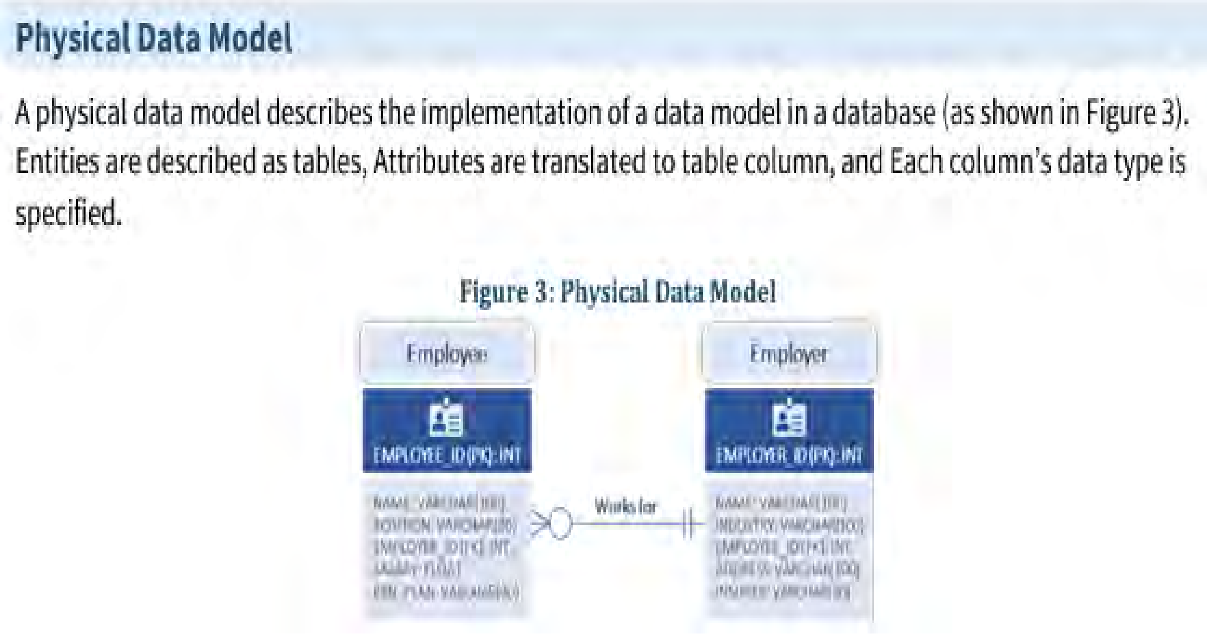
Logical Data Model

A **logical data model** defines how a data model should be implemented, with as much detail as possible, without regard for its physical implementation in a database. Within a logical data model, an entity's box contains a list of the entity's **attributes**.

One or more attributes is designated as a primary key, whose value uniquely specifies an instance of that entity. A primary key may be referred to in another entity as a **foreign key**.

In the Figure 2 example, each Employee works for only one Employer. Each Employer may have zero or more Employees. This is indicated via the model's line notation (refer to the Describing Relationships section).

Figure 2: Logical Data Model

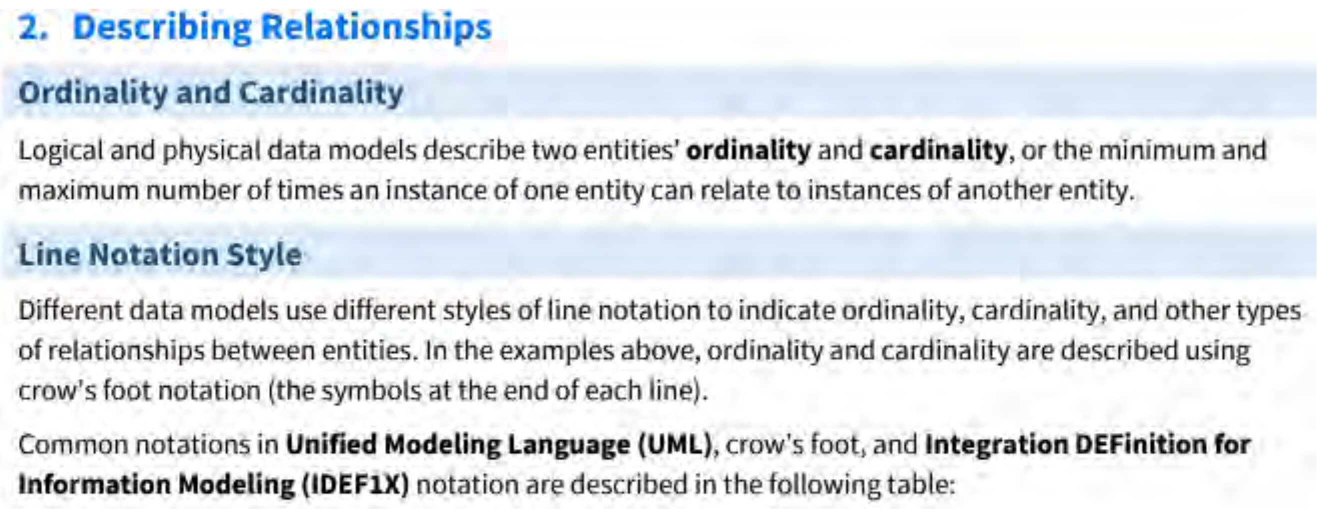


Physical Data Model

A physical data model describes the implementation of a data model in a database (as shown in Figure 3). Entities are described as tables, Attributes are translated to table column, and Each column's data type is specified.

Figure 3: Physical Data Model

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.cms.gov/files/document/laymans-guide-reading-data-models.pdf



2. Describing Relationships

Ordinality and Cardinality

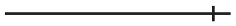
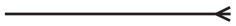
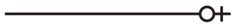
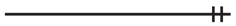
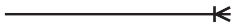
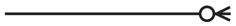
Logical and physical data models describe two entities' **ordinality** and **cardinality**, or the minimum and maximum number of times an instance of one entity can relate to instances of another entity.

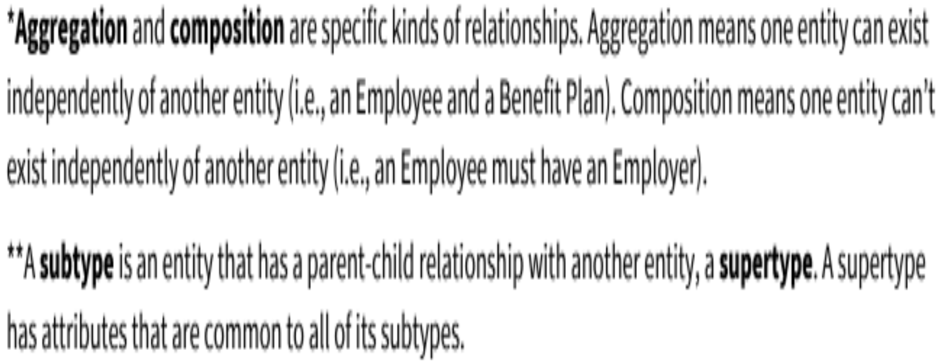
Line Notation Style

Different data models use different styles of line notation to indicate ordinality, cardinality, and other types of relationships between entities. In the examples above, ordinality and cardinality are described using crow's foot notation (the symbols at the end of each line).

Common notations in **Unified Modeling Language (UML)**, crow's foot, and **Integration DEFinition for Information Modeling (IDEF1X)** notation are described in the following table:

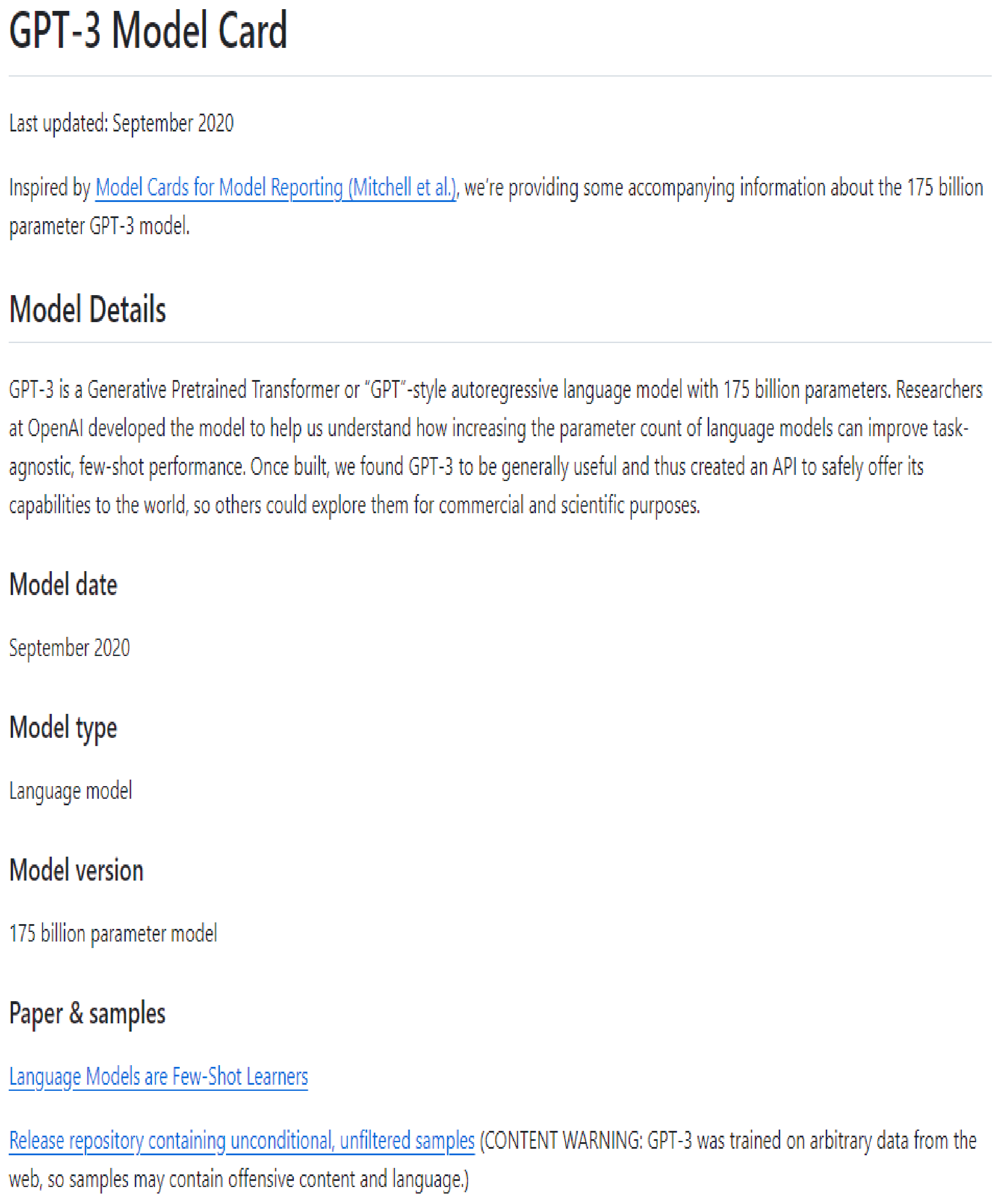
Table 1: Syntax in Common Data Modeling Notation Styles

Notation	Crow's Foot	UML	IDEF1X
One		N/A	N/A
Many		N/A	N/A
Zero or one		0..1	● Z
One only		1	● 1
One or more		1..*	● P
Zero or more		0..*	●
(Specific range)	N/A	3..7	N/A
Composition*	N/A	Part —◆ Whole	"Is part of"
Aggregation*	N/A	Part —◇ Whole	"Is part of"
Subtype**	N/A	Subtype —▷ Supertype	Subtype —○— Supertype



*__Aggregation__ and **composition** are specific kinds of relationships. Aggregation means one entity can exist independently of another entity (i.e., an Employee and a Benefit Plan). Composition means one entity can't exist independently of another entity (i.e., an Employee must have an Employer).

A **subtype is an entity that has a parent-child relationship with another entity, a **supertype**. A supertype has attributes that are common to all of its subtypes.



GPT-3 Model Card

Last updated: September 2020

Inspired by Model Cards for Model Reporting (Mitchell et al.), we're providing some accompanying information about the 175 billion parameter GPT-3 model.

Model Details

GPT-3 is a Generative Pretrained Transformer or "GPT"-style autoregressive language model with 175 billion parameters. Researchers at OpenAI developed the model to help us understand how increasing the parameter count of language models can improve task-agnostic, few-shot performance. Once built, we found GPT-3 to be generally useful and thus created an API to safely offer its capabilities to the world, so others could explore them for commercial and scientific purposes.

Model date

September 2020

Model type

Language model

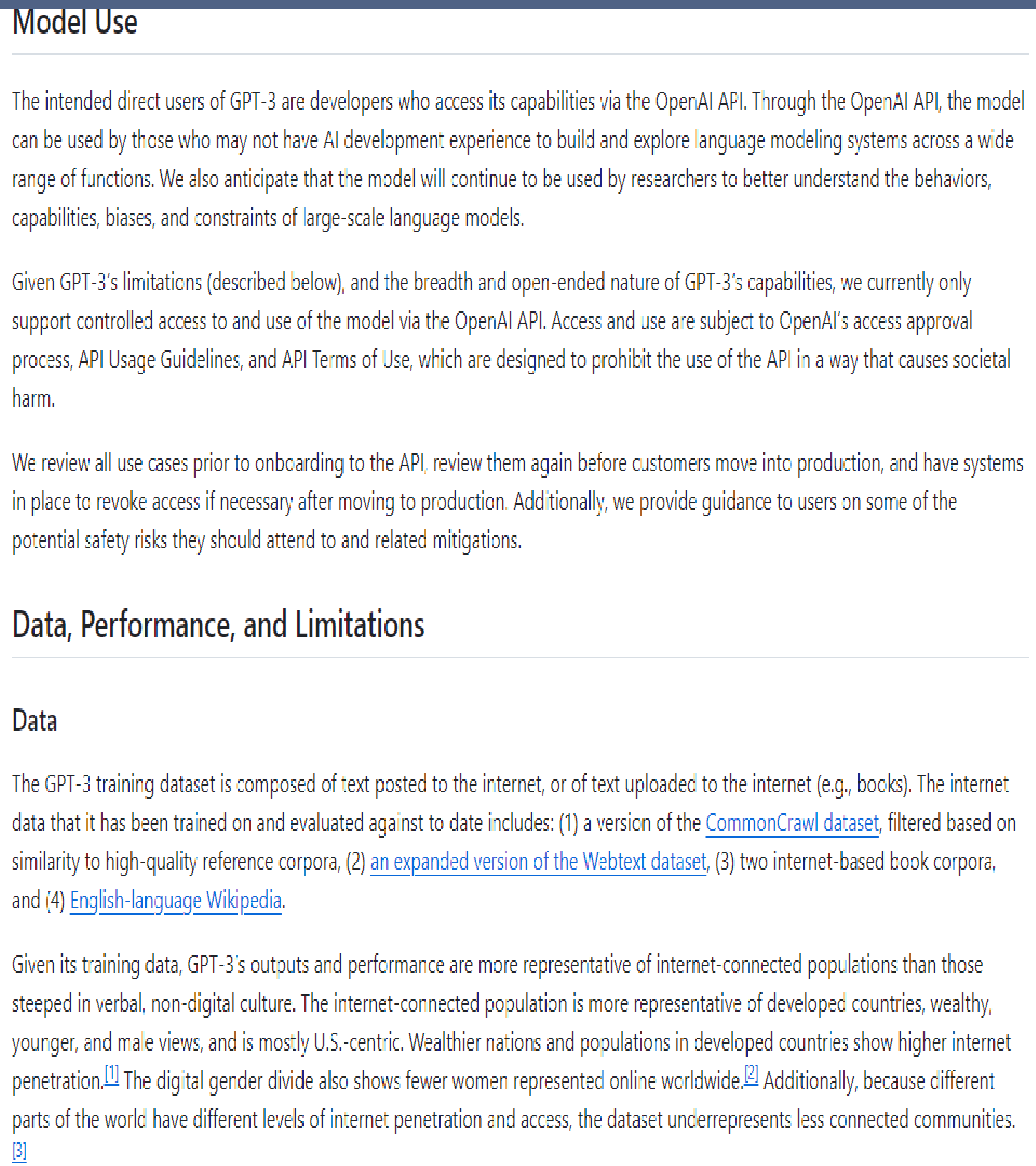
Model version

175 billion parameter model

Paper & samples

Language Models are Few-Shot Learners

Release repository containing unconditional, unfiltered samples (CONTENT WARNING: GPT-3 was trained on arbitrary data from the web, so samples may contain offensive content and language.)



Model Use

The intended direct users of GPT-3 are developers who access its capabilities via the OpenAI API. Through the OpenAI API, the model can be used by those who may not have AI development experience to build and explore language modeling systems across a wide range of functions. We also anticipate that the model will continue to be used by researchers to better understand the behaviors, capabilities, biases, and constraints of large-scale language models.

Given GPT-3's limitations (described below), and the breadth and open-ended nature of GPT-3's capabilities, we currently only support controlled access to and use of the model via the OpenAI API. Access and use are subject to OpenAI's access approval process, API Usage Guidelines, and API Terms of Use, which are designed to prohibit the use of the API in a way that causes societal harm.

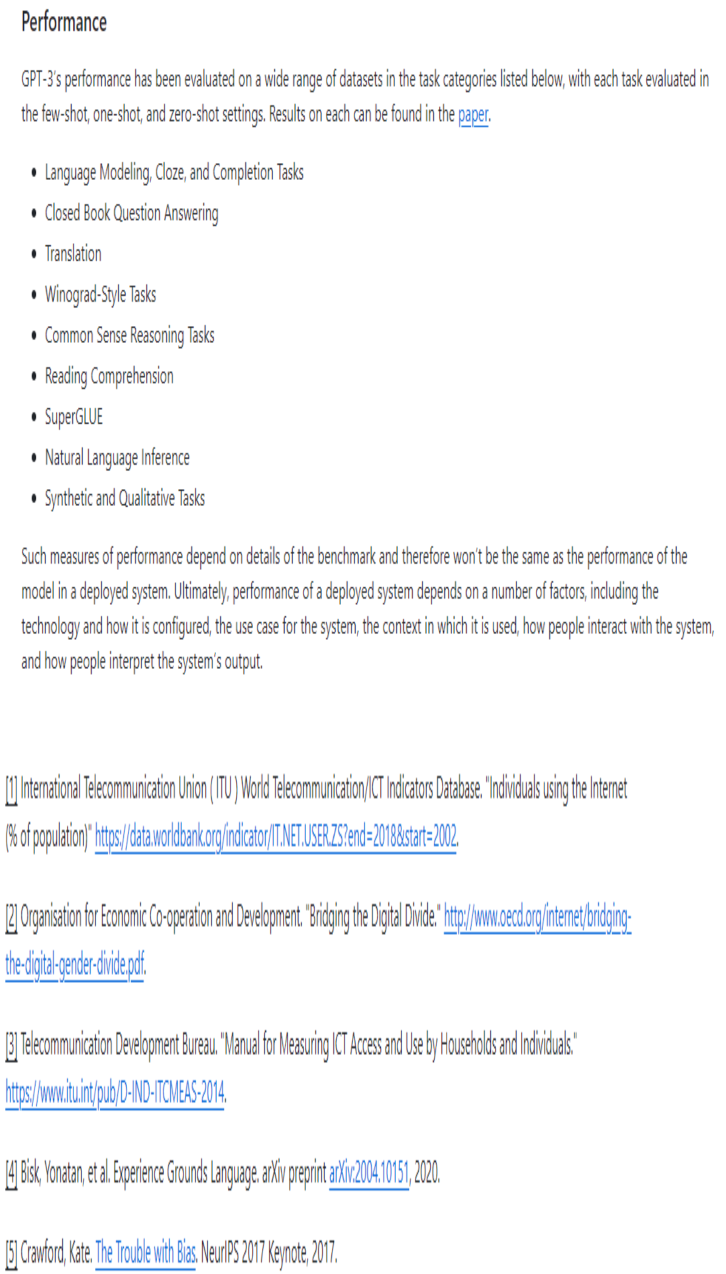
We review all use cases prior to onboarding to the API, review them again before customers move into production, and have systems in place to revoke access if necessary after moving to production. Additionally, we provide guidance to users on some of the potential safety risks they should attend to and related mitigations.

Data, Performance, and Limitations

Data

The GPT-3 training dataset is composed of text posted to the internet, or of text uploaded to the internet (e.g., books). The internet data that it has been trained on and evaluated against to date includes: (1) a version of the CommonCrawl dataset, filtered based on similarity to high-quality reference corpora, (2) an expanded version of the Webtext dataset, (3) two internet-based book corpora, and (4) English-language Wikipedia.

Given its training data, GPT-3's outputs and performance are more representative of internet-connected populations than those steeped in verbal, non-digital culture. The internet-connected population is more representative of developed countries, wealthy, younger, and male views, and is mostly U.S.-centric. Wealthier nations and populations in developed countries show higher internet penetration.[1] The digital gender divide also shows fewer women represented online worldwide.[2] Additionally, because different parts of the world have different levels of internet penetration and access, the dataset underrepresents less connected communities. [3]



Performance

GPT-3's performance has been evaluated on a wide range of datasets in the task categories listed below, with each task evaluated in the few-shot, one-shot, and zero-shot settings. Results on each can be found in the paper.

- Language Modeling, Cloze, and Completion Tasks
- Closed Book Question Answering
- Translation
- Winograd-Style Tasks
- Common Sense Reasoning Tasks
- Reading Comprehension
- SuperGLUE
- Natural Language Inference
- Synthetic and Qualitative Tasks

Such measures of performance depend on details of the benchmark and therefore won't be the same as the performance of the model in a deployed system. Ultimately, performance of a deployed system depends on a number of factors, including the technology and how it is configured, the use case for the system, the context in which it is used, how people interact with the system, and how people interpret the system's output.

[1] International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database. "Individuals using the Internet (% of population)" https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2018&start=2002.

[2] Organisation for Economic Co-operation and Development. "Bridging the Digital Divide." http://www.oecd.org/internet/bridging-the-digital-gender-divide.pdf.

[3] Telecommunication Development Bureau. "Manual for Measuring ICT Access and Use by Households and Individuals." https://www.itu.int/pub/D-IND-ITCMEAS-2014.

[4] Bisk, Yonatan, et al. Experience Grounds Language. arXiv preprint arXiv:2004.10151, 2020.

[5] Crawford, Kate. The Trouble with Bias. NeurIPS 2017 Keynote, 2017.

Limitations

GPT-3 and our analysis of it have a number of limitations. Some of these limitations are inherent to any model with machine learning (ML) components that can have high-bandwidth, open-ended interactions with people (e.g. via natural language): ML components have limited robustness; ML components are biased; open-ended systems have large surface areas for risk; and safety is a moving target for ML systems. GPT-3 has the propensity to generate text that contains falsehoods and expresses them confidently, and, like any model with ML components, it can only be expected to provide reasonable outputs when given inputs similar to the ones present in its training data. In addition to these fundamental limitations, we outline some of the technical limitations evaluated in the paper below.

Repetition: GPT-3 samples sometimes repeat themselves semantically at the document level, and can lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs. Our release repository contains 500 unconditional, unfiltered 2048 token samples (CONTENT WARNING: GPT-3 was trained on arbitrary data from the web, so samples may contain offensive content and language).

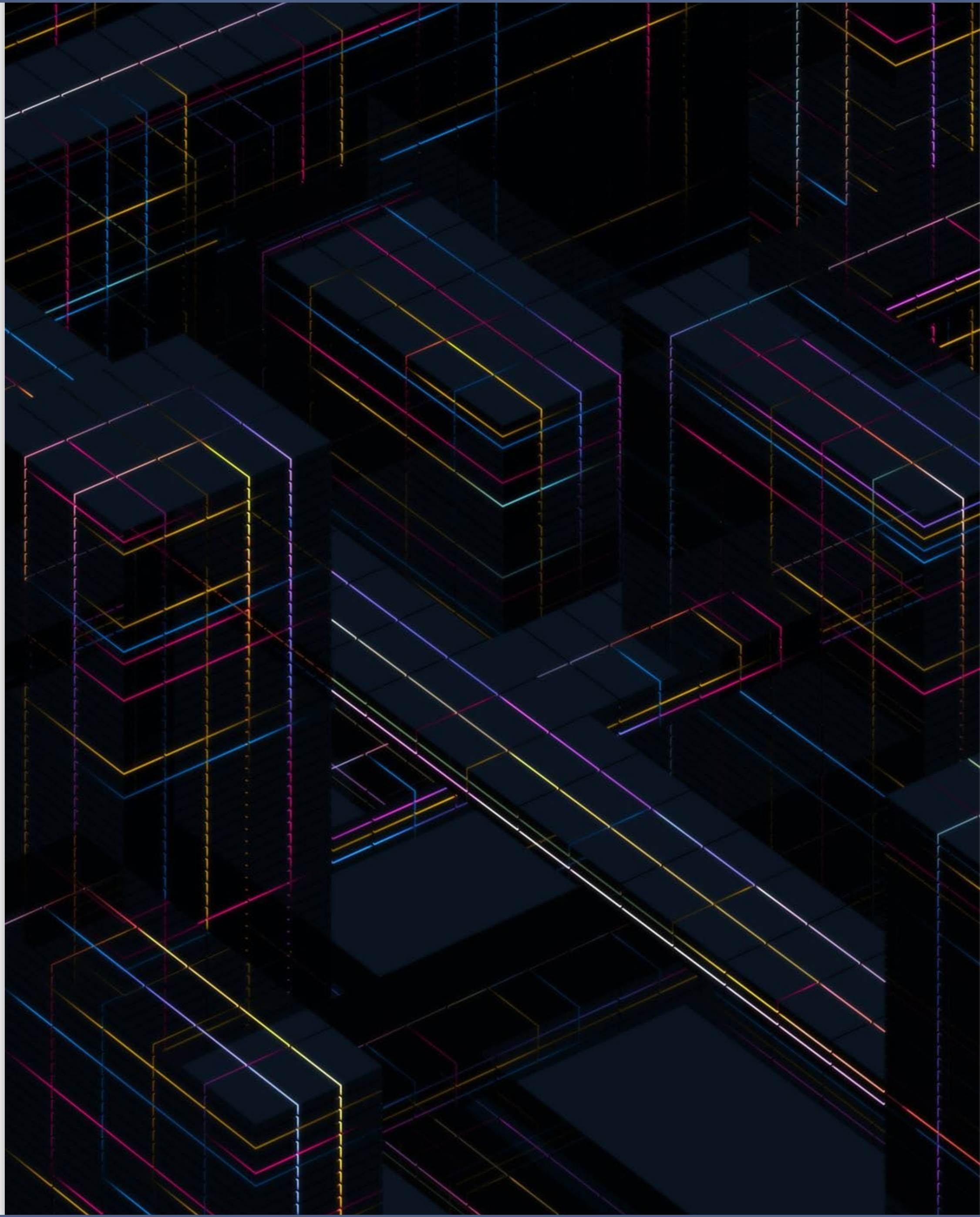
Lack of world grounding: GPT-3, like other large pretrained language models, is not grounded in other modalities of experience, such as video, real-world physical interaction, or human feedback, and thus lacks a large amount of context about the world.[4]
Predominantly English: GPT-3 is trained largely on text in the English language, and is best suited for classifying, searching, summarizing, or generating such text. GPT-3 will by default perform worse on inputs that are different from the data distribution it is trained on, including non-English languages as well as specific dialects of English that are not as well-represented in training data.

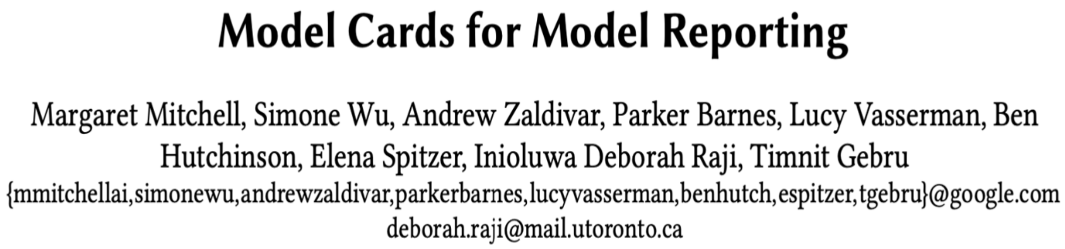
Interpretability & predictability: the capacity to interpret or predict how GPT-3 will behave is very limited, a limitation common to most deep learning systems, especially in models of this scale. High variance on novel inputs: GPT-3 is not necessarily well-calibrated in its predictions on novel inputs. This can be observed in the much higher variance in its performance as compared to that of humans on standard benchmarks. Creation date of training corpora: The May 2020 version of GPT-3 was trained on a dataset created in November 2019, so has not been trained on any data more recent than that. The September 2020 version of the model was retrained to reflect data up to August 2020.

Biases: GPT-3, like all large language models trained on internet corpora, will generate stereotyped or prejudiced content. The model has the propensity to retain and magnify biases it inherited from any part of its training, from the datasets we selected to the training techniques we chose. This is concerning, since model bias could harm people in the relevant groups in different ways by entrenching existing stereotypes and producing demeaning portrayals amongst other potential harms.[5] This issue is of special concern from a societal perspective, and is discussed along with other issues in the paper section on Broader Impacts.

Proprietary Tension

- Complexity of ML models give them the ability to learn deeper patterns in data.
- This complexity makes models hard to interpret but most ML is a function of data + ML architecture.
- If the models cannot be transparent, then we need to be transparent about things around the ML models as much as possible.





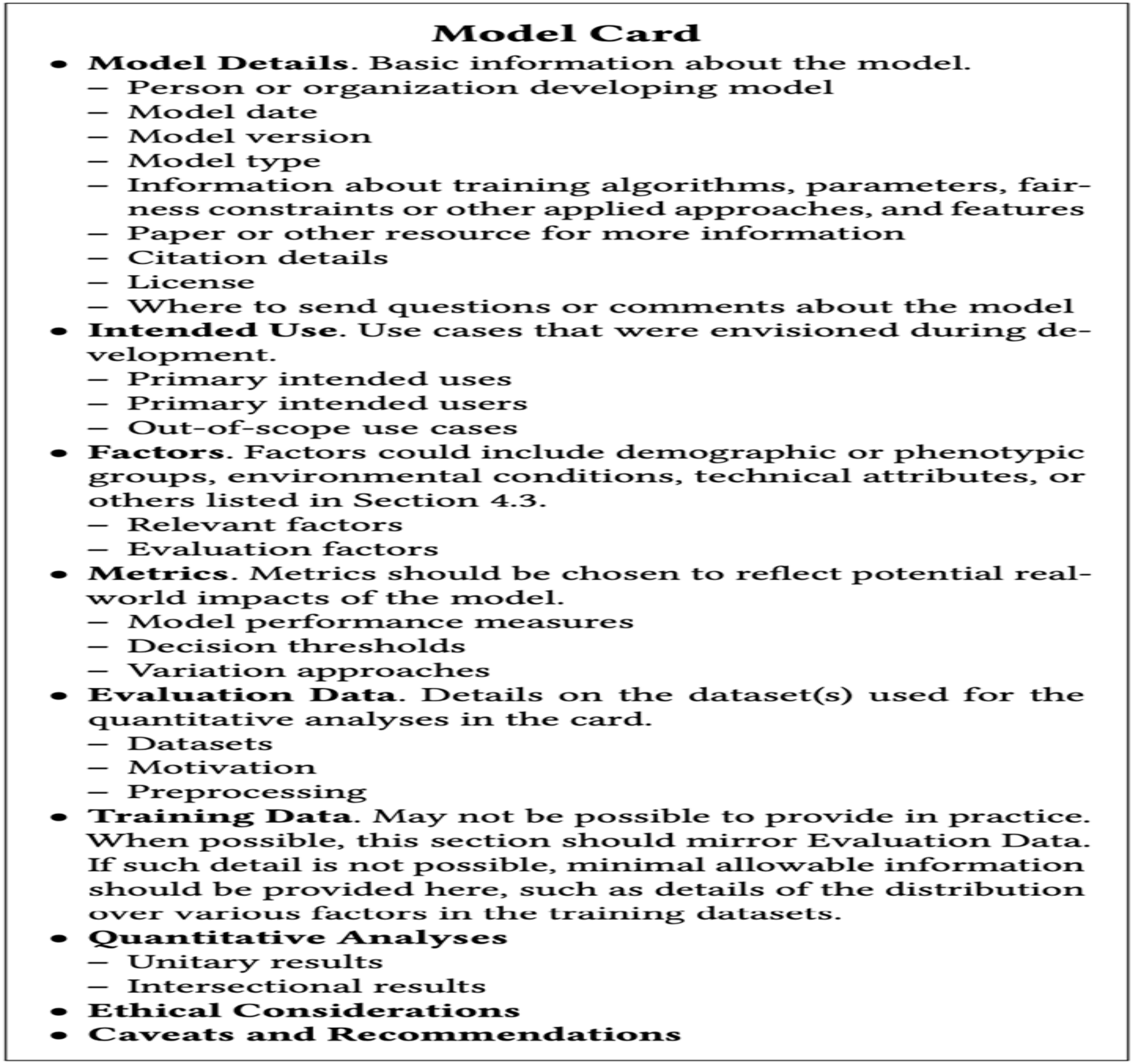
Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

Stakeholders targeted:

- ML/AI Practitioners + Developers
- Policymakers
- ML-Knowledgeable individuals
- Impacted individuals

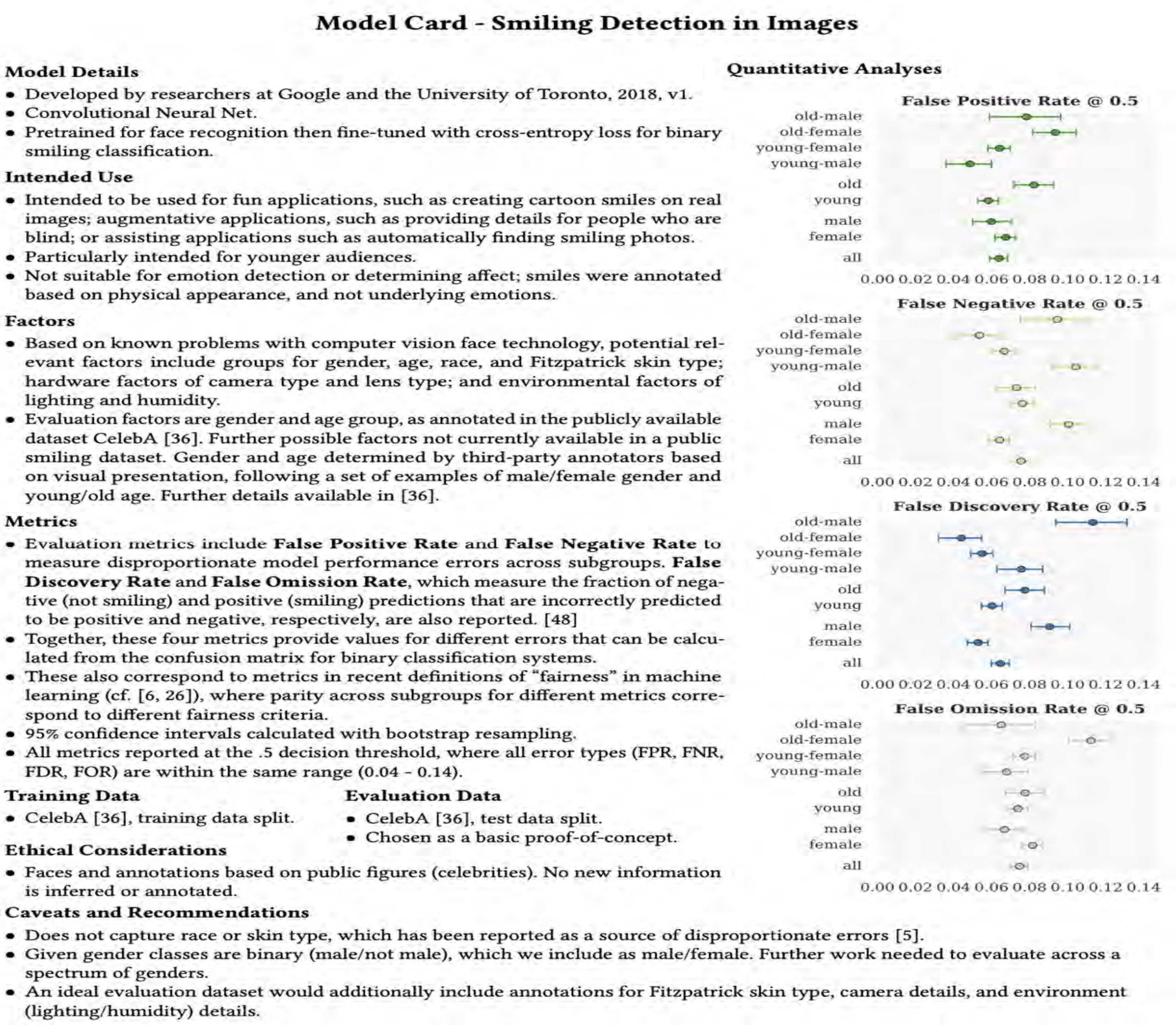
https://www.kaggle.com/code/var0101/model-cards



Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 1: Summary of model card sections and suggested prompts for each.



Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

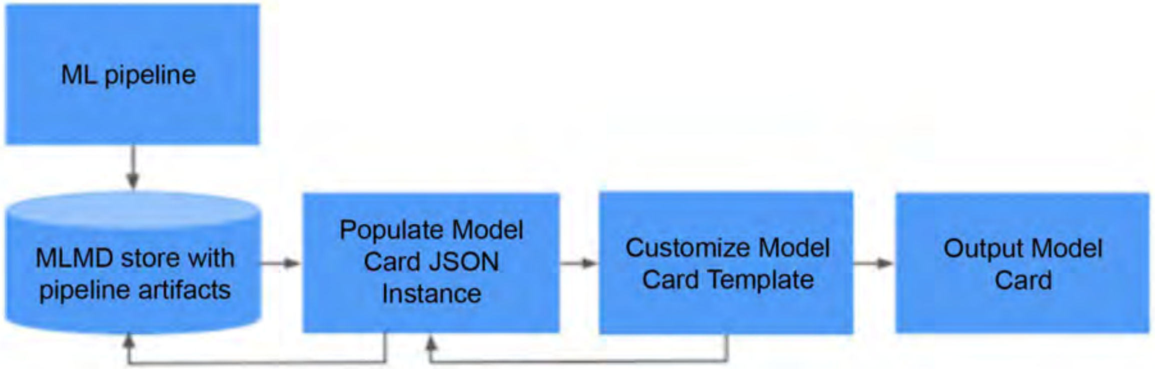
Google Research   Philosophy   Research Areas   Publications   People   Resources

BLOG ›

Introducing the Model Card Toolkit for Easier Model Transparency Reporting

ÇARŞAMBA, TEMMUZ 29, 2020

Posted by Huanming Fang and Hui Miao, Software Engineers, Google Research



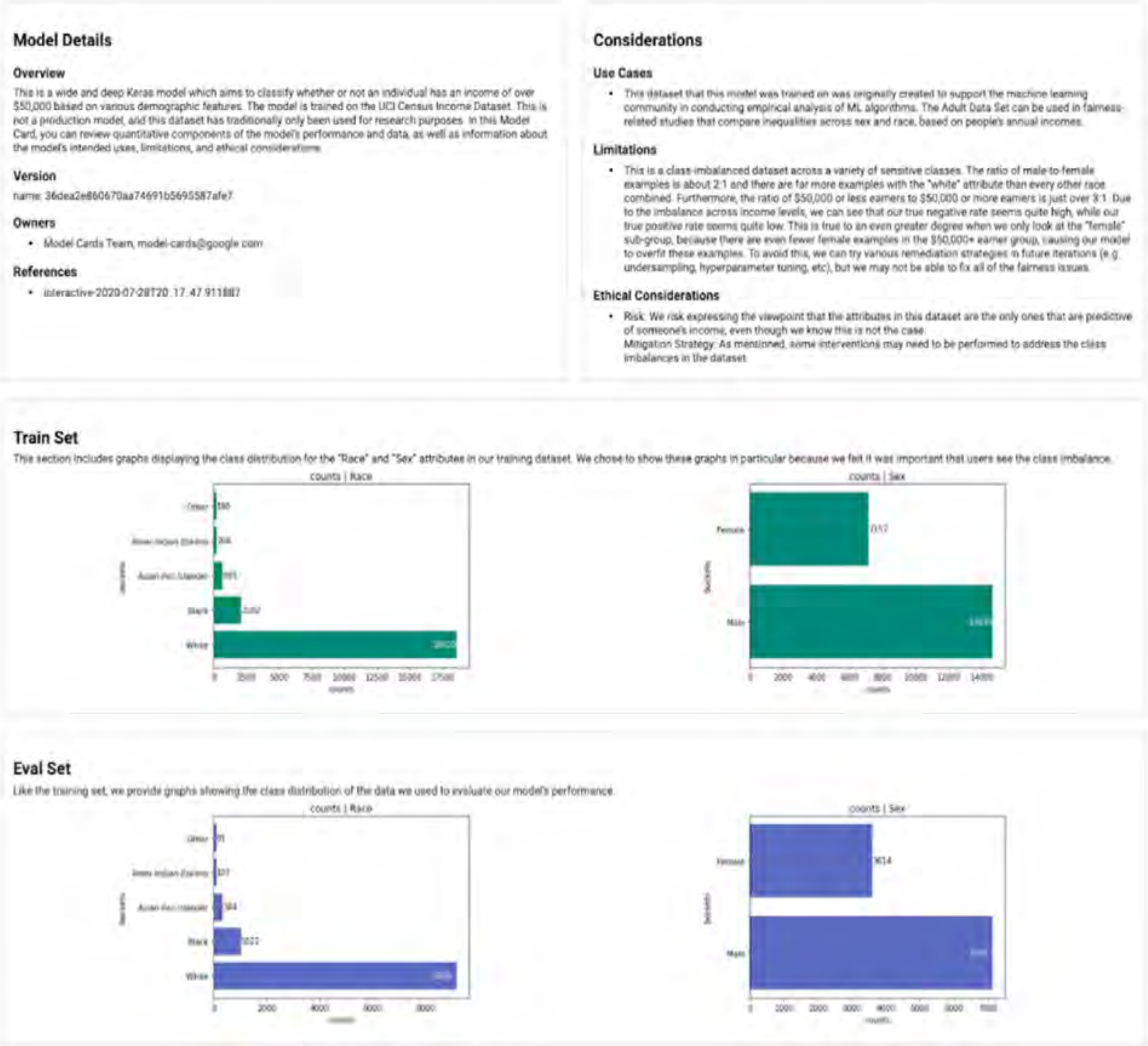
ML pipeline

MLMD store with pipeline artifacts

Populate Model Card JSON Instance

Customize Model Card Template

Output Model Card



Model Details

Overview

This is a wide and deep Keras model which aims to classify whether or not an individual has an income of over $50,000 based on various demographic features. The model is trained on the UCI Census Income Dataset. This is not a production model, and this dataset has traditionally only been used for research purposes. In this Model Card, you can review quantitative components of the model's performance and data, as well as information about the model's intended uses, limitations, and ethical considerations.

Version

name: 36dea2e860670aa74691b5695587afe7

Owners

- Model Cards Team, model-cards@google.com

References

- interactive-2020-07-28T20_17_47.911887

Considerations

Use Cases

- This dataset that this model was trained on was originally created to support the machine learning community in conducting empirical analysis of ML algorithms. The Adult Data Set can be used in fairness-related studies that compare inequalities across sex and race, based on people's annual incomes.

Limitations

- This is a class-imbalanced dataset across a variety of sensitive classes. The ratio of male-to-female examples is about 2:1 and there are far more examples with the "white" attribute than every other race combined. Furthermore, the ratio of $50,000 or less earners to $50,000 or more earners is just over 3:1. Due to the imbalance across income levels, we can see that our true negative rate seems quite high, while our true positive rate seems quite low. This is true to an even greater degree when we only look at the "female" sub-group, because there are even fewer female examples in the $50,000+ earner group, causing our model to overfit these examples. To avoid this, we can try various remediation strategies in future iterations (e.g. undersampling, hyperparameter tuning, etc), but we may not be able to fix all of the fairness issues.

Ethical Considerations

- Risk: We risk expressing the viewpoint that the attributes in this dataset are the only ones that are predictive of someone's income, even though we know this is not the case.
  Mitigation Strategy: As mentioned, some interventions may need to be performed to address the class imbalances in the dataset.

Train Set

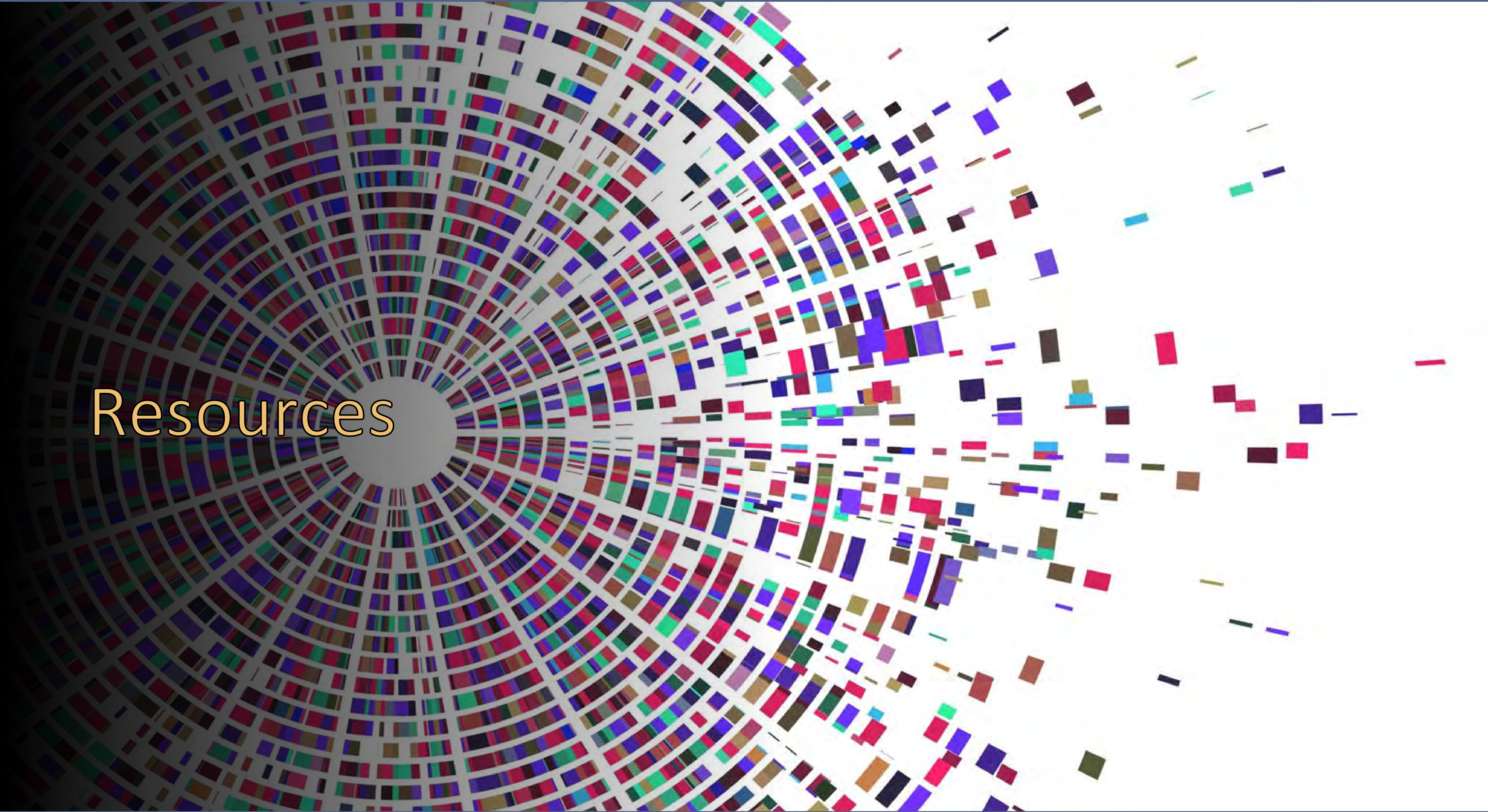
This section includes graphs displaying the class distribution for the "Race" and "Sex" attributes in our training dataset. We chose to show these graphs in particular because we felt it was important that users see the class imbalance.

Eval Set

Like the training set, we provide graphs showing the class distribution of the data we used to evaluate our model's performance.

Slido

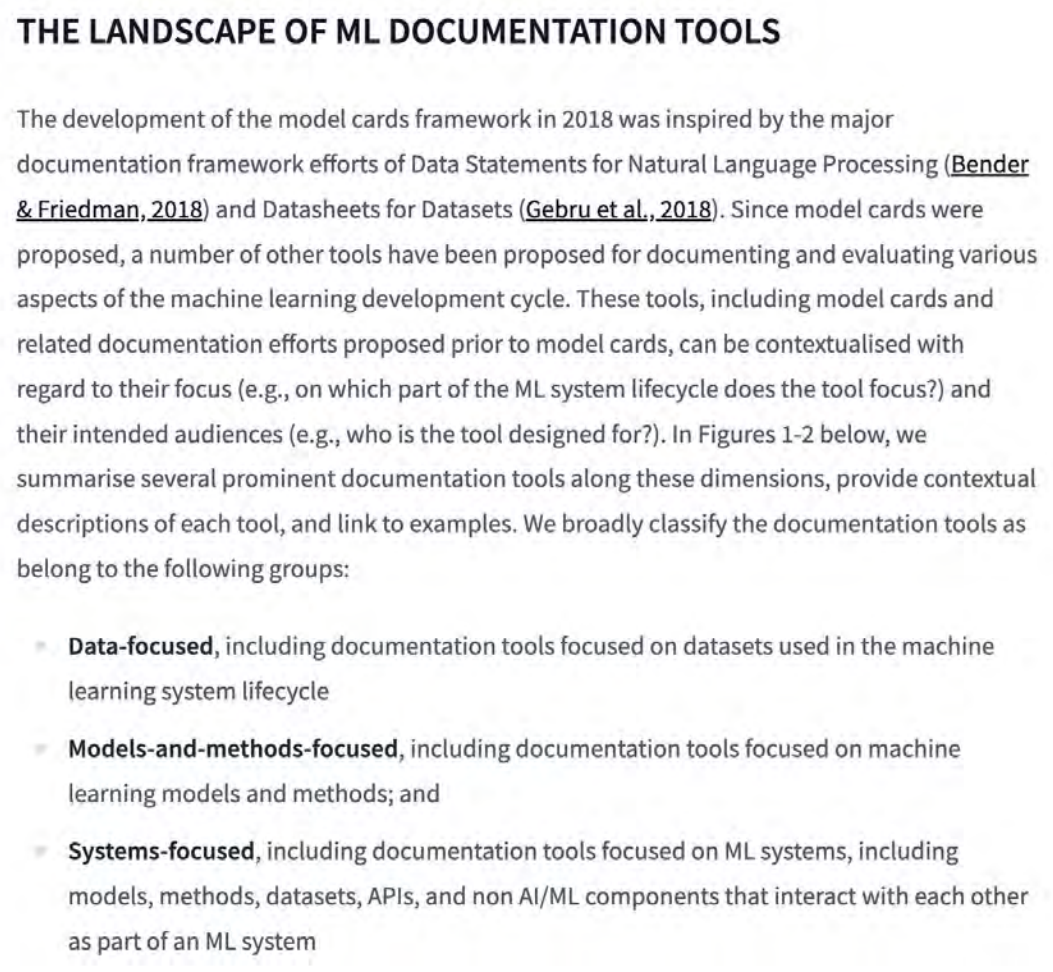
How would you design a model card for community-engaged users of AI? What would you include to make the AI application understandable and adoptable?



Resources

🤗 Hugging Face

- An AI startup originally focused on making a chatbot for teens.
- Pivoted towards trying to build a community and ecosystem of tools for accelerating AI research
- Started mainly in the NLP space.
- Provided easy to use interfaces to Text-based DL models (transformers models that worked with TF/PyTorch)
- Evolved to include LLMs



THE LANDSCAPE OF ML DOCUMENTATION TOOLS

The development of the model cards framework in 2018 was inspired by the major documentation framework efforts of Data Statements for Natural Language Processing (Bender & Friedman, 2018) and Datasheets for Datasets (Gebru et al., 2018). Since model cards were proposed, a number of other tools have been proposed for documenting and evaluating various aspects of the machine learning development cycle. These tools, including model cards and related documentation efforts proposed prior to model cards, can be contextualised with regard to their focus (e.g., on which part of the ML system lifecycle does the tool focus?) and their intended audiences (e.g., who is the tool designed for?). In Figures 1-2 below, we summarise several prominent documentation tools along these dimensions, provide contextual descriptions of each tool, and link to examples. We broadly classify the documentation tools as belong to the following groups:

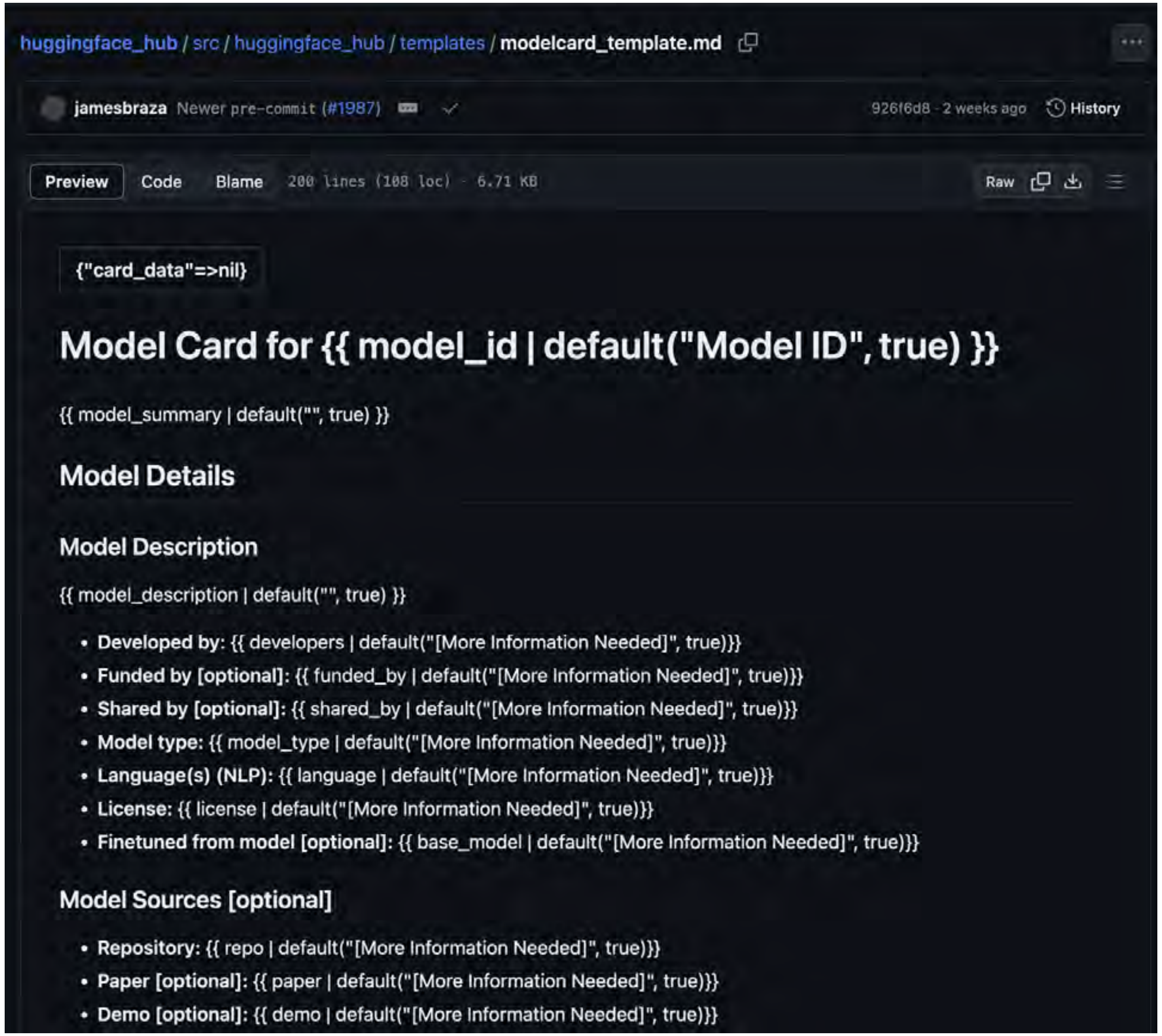
- **Data-focused**, including documentation tools focused on datasets used in the machine learning system lifecycle
- **Models-and-methods-focused**, including documentation tools focused on machine learning models and methods; and
- **Systems-focused**, including documentation tools focused on ML systems, including models, methods, datasets, APIs, and non AI/ML components that interact with each other as part of an ML system

User Study Details

We selected people from a variety of different backgrounds relevant to machine learning and model documentation. Below, we detail their demographics, the questions they were asked, and the corresponding insights from their responses. Full details on responses are available in Appendix A.

Respondent Demographics

- Tech & Regulatory Affairs Counsel
- ML Engineer (x2)
- Developer Advocate
- Executive Assistant
- Monetization Lead
- Policy Manager/AI Researcher
- Research Intern



huggingface_hub / src / huggingface_hub / templates / modelcard_template.md
jamesbraza Newer pre-commit (#1987)
926f6d8 · 2 weeks ago
History
Preview
Code
Blame
200 lines (108 loc) · 6.71 KB
Raw

{"card_data"=>nil}

Model Card for {{ model_id | default("Model ID", true) }}

{{ model_summary | default("", true) }}

Model Details

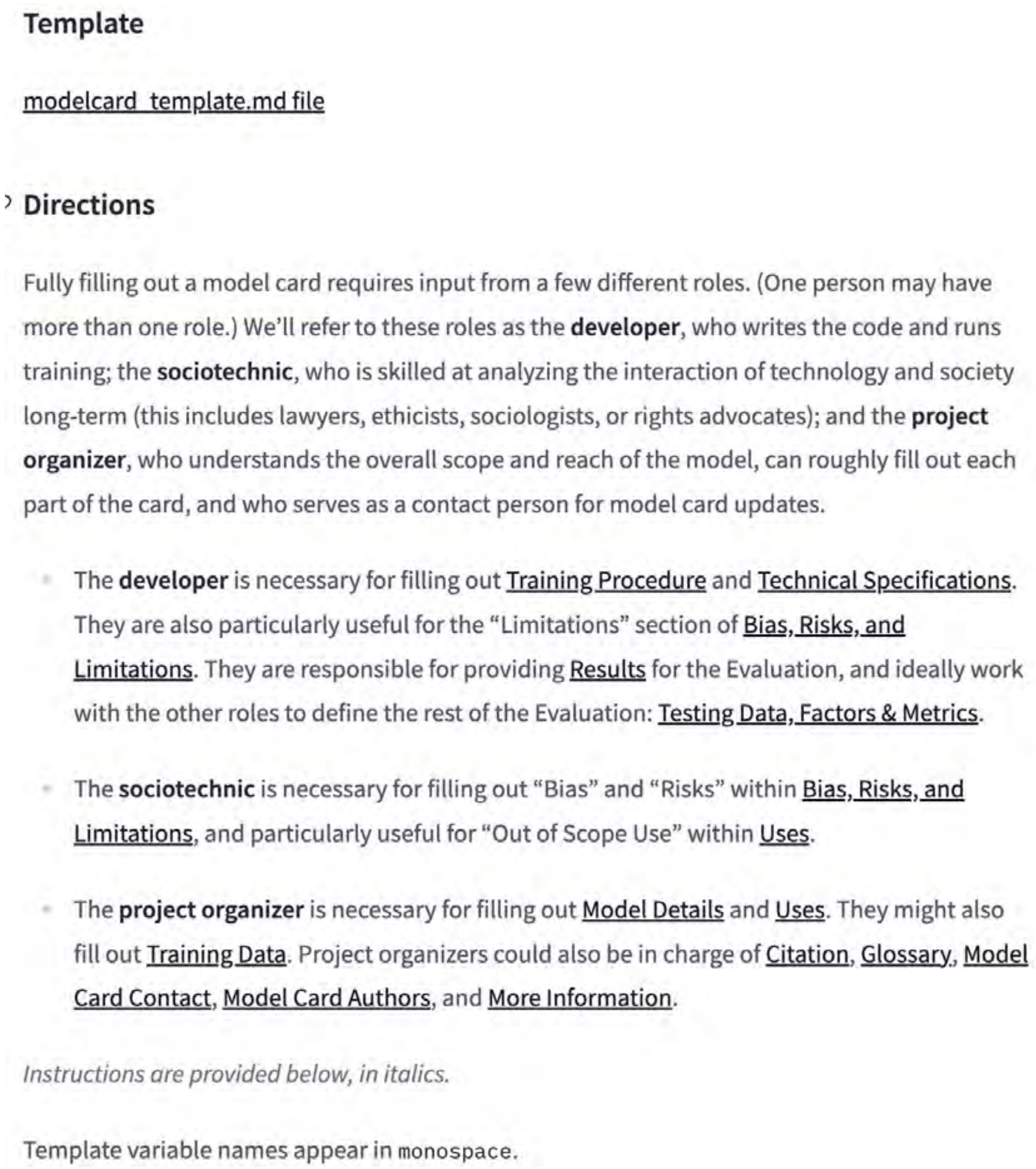
Model Description

{{ model_description | default("", true) }}

- **Developed by:** {{ developers | default("[More Information Needed]", true)}}
- **Funded by [optional]:** {{ funded_by | default("[More Information Needed]", true)}}
- **Shared by [optional]:** {{ shared_by | default("[More Information Needed]", true)}}
- **Model type:** {{ model_type | default("[More Information Needed]", true)}}
- **Language(s) (NLP):** {{ language | default("[More Information Needed]", true)}}
- **License:** {{ license | default("[More Information Needed]", true)}}
- **Finetuned from model [optional]:** {{ base_model | default("[More Information Needed]", true)}}

Model Sources [optional]

- **Repository:** {{ repo | default("[More Information Needed]", true)}}
- **Paper [optional]:** {{ paper | default("[More Information Needed]", true)}}
- **Demo [optional]:** {{ demo | default("[More Information Needed]", true)}}



Template

modelcard_template.md file

Directions

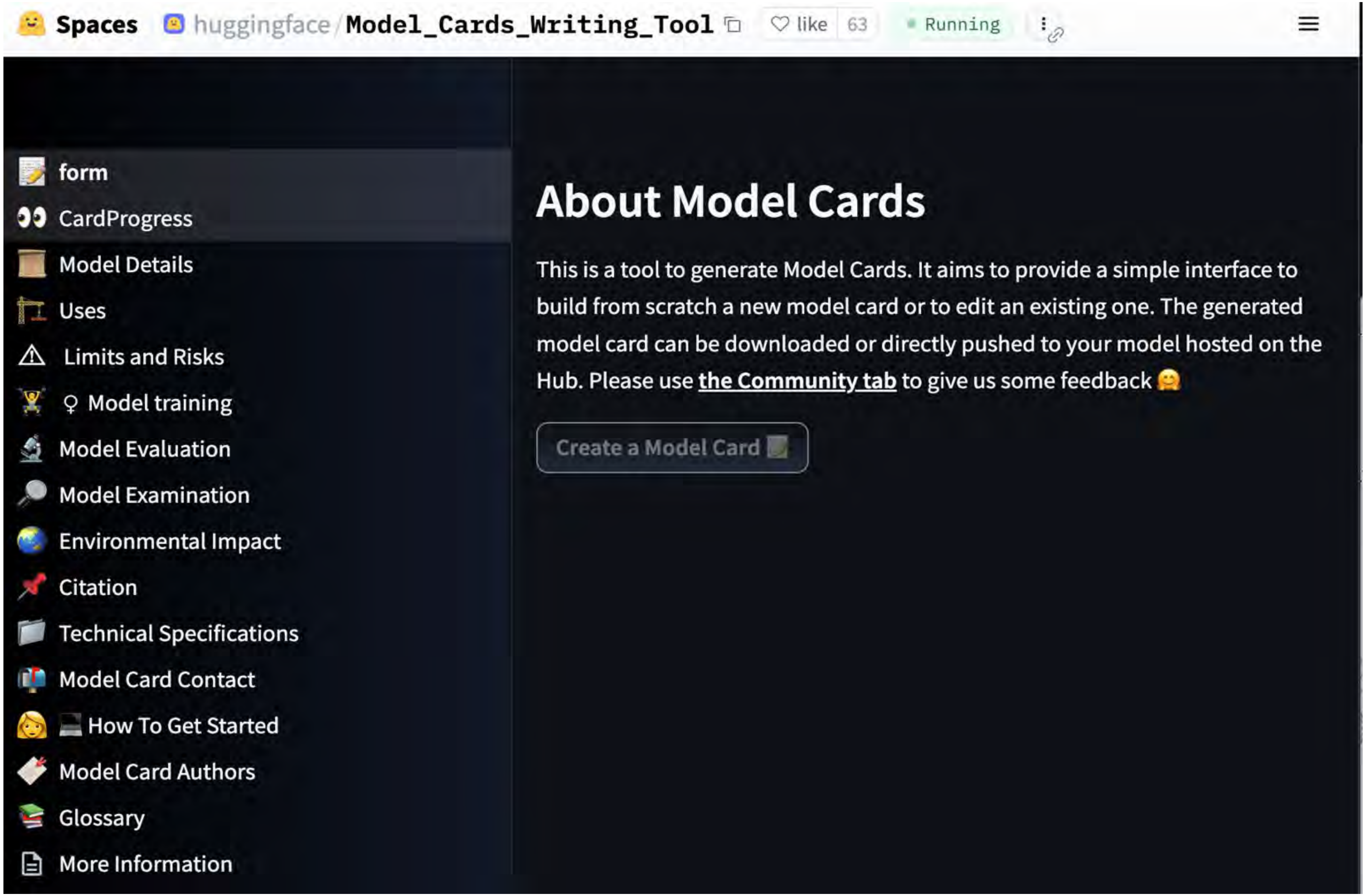
Fully filling out a model card requires input from a few different roles. (One person may have more than one role.) We'll refer to these roles as the **developer**, who writes the code and runs training; the **sociotechnic**, who is skilled at analyzing the interaction of technology and society long-term (this includes lawyers, ethicists, sociologists, or rights advocates); and the **project organizer**, who understands the overall scope and reach of the model, can roughly fill out each part of the card, and who serves as a contact person for model card updates.

- The **developer** is necessary for filling out Training Procedure and Technical Specifications. They are also particularly useful for the "Limitations" section of Bias, Risks, and Limitations. They are responsible for providing Results for the Evaluation, and ideally work with the other roles to define the rest of the Evaluation: Testing Data, Factors & Metrics.
- The **sociotechnic** is necessary for filling out "Bias" and "Risks" within Bias, Risks, and Limitations, and particularly useful for "Out of Scope Use" within Uses.
- The **project organizer** is necessary for filling out Model Details and Uses. They might also fill out Training Data. Project organizers could also be in charge of Citation, Glossary, Model Card Contact, Model Card Authors, and More Information.

Instructions are provided below, in italics.

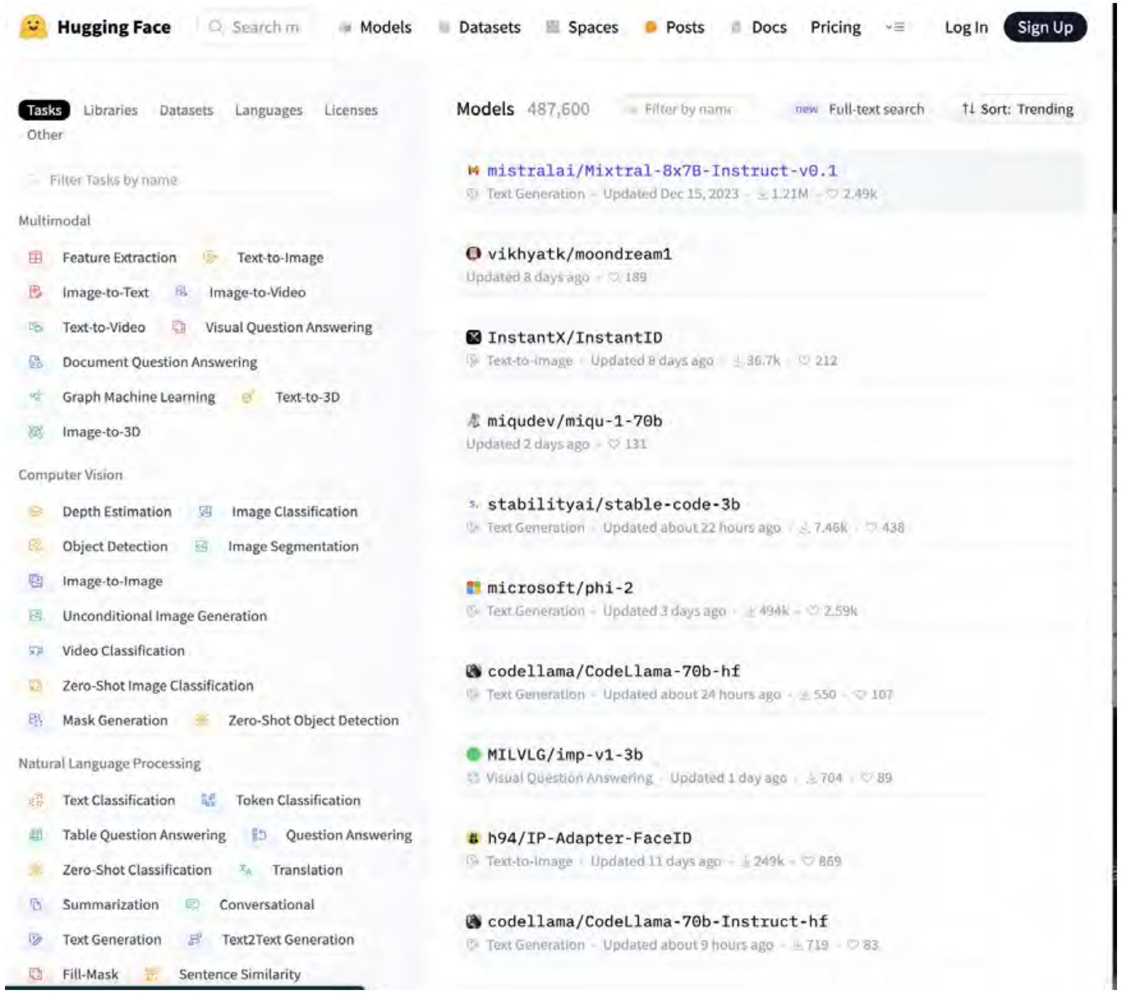
Template variable names appear in `monospace`.

https://arxiv.org/abs/1810.03993



About Model Cards

This is a tool to generate Model Cards. It aims to provide a simple interface to build from scratch a new model card or to edit an existing one. The generated model card can be downloaded or directly pushed to your model hosted on the Hub. Please use **the Community tab** to give us some feedback 🤗



Multimodal

Feature Extraction · Text-to-Image · Image-to-Text · Image-to-Video · Text-to-Video · Visual Question Answering · Document Question Answering · Graph Machine Learning · Text-to-3D · Image-to-3D

Computer Vision

Depth Estimation · Image Classification · Object Detection · Image Segmentation · Image-to-Image · Unconditional Image Generation · Video Classification · Zero-Shot Image Classification · Mask Generation · Zero-Shot Object Detection

Natural Language Processing

Text Classification · Token Classification · Table Question Answering · Question Answering · Zero-Shot Classification · Translation · Summarization · Conversational · Text Generation · Text2Text Generation · Fill-Mask · Sentence Similarity

Models 487,600 Filter by name new Full-text search Sort: Trending

mistralai/Mixtral-8x7B-Instruct-v0.1
Text Generation · Updated Dec 15, 2023 · 1.21M · 2.49k

vikhyatk/moondream1
Updated 8 days ago · 189

InstantX/InstantID
Text-to-Image · Updated 8 days ago · 36.7k · 212

miqudev/miqu-1-70b
Updated 2 days ago · 131

stabilityai/stable-code-3b
Text Generation · Updated about 22 hours ago · 7.46k · 438

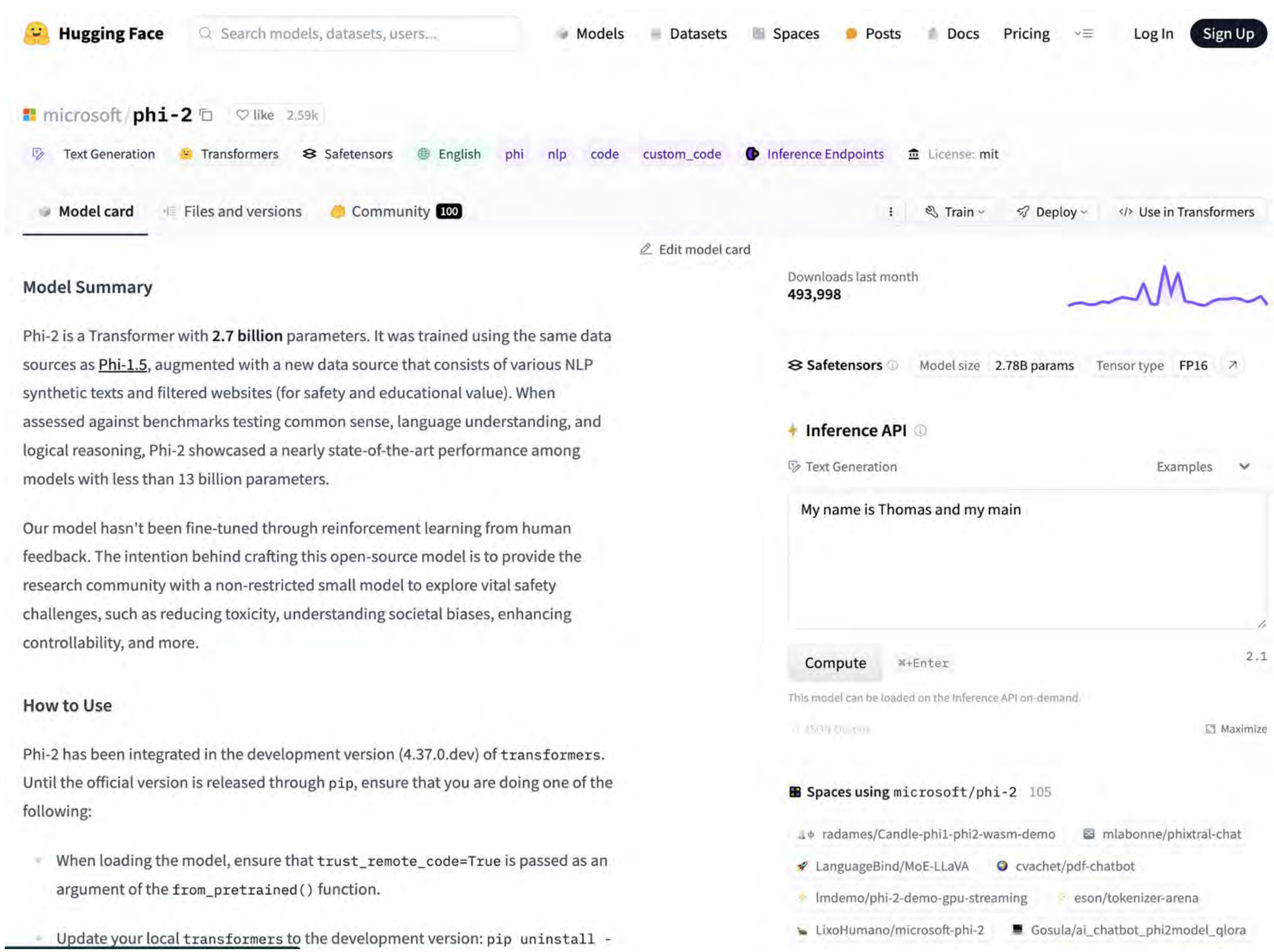
microsoft/phi-2
Text Generation · Updated 3 days ago · 494k · 2.59k

codellama/CodeLlama-70b-hf
Text Generation · Updated about 24 hours ago · 550 · 107

MILVLG/imp-v1-3b
Visual Question Answering · Updated 1 day ago · 704 · 89

h94/IP-Adapter-FaceID
Text-to-Image · Updated 11 days ago · 249k · 869

codellama/CodeLlama-70b-Instruct-hf
Text Generation · Updated about 9 hours ago · 719 · 83



microsoft/phi-2

Text Generation · Transformers · Safetensors · English · phi · nlp · code · custom_code · Inference Endpoints · License: mit

Model Summary

Phi-2 is a Transformer with **2.7 billion** parameters. It was trained using the same data sources as Phi-1.5, augmented with a new data source that consists of various NLP synthetic texts and filtered websites (for safety and educational value). When assessed against benchmarks testing common sense, language understanding, and logical reasoning, Phi-2 showcased a nearly state-of-the-art performance among models with less than 13 billion parameters.

Our model hasn't been fine-tuned through reinforcement learning from human feedback. The intention behind crafting this open-source model is to provide the research community with a non-restricted small model to explore vital safety challenges, such as reducing toxicity, understanding societal biases, enhancing controllability, and more.

How to Use

Phi-2 has been integrated in the development version (4.37.0.dev) of `transformers`. Until the official version is released through `pip`, ensure that you are doing one of the following:

- When loading the model, ensure that `trust_remote_code=True` is passed as an argument of the `from_pretrained()` function.
- Update your local `transformers` to the development version: `pip uninstall -`

Downloads last month
493,998

Safetensors · Model size 2.78B params · Tensor type FP16

Inference API

Text Generation · Examples

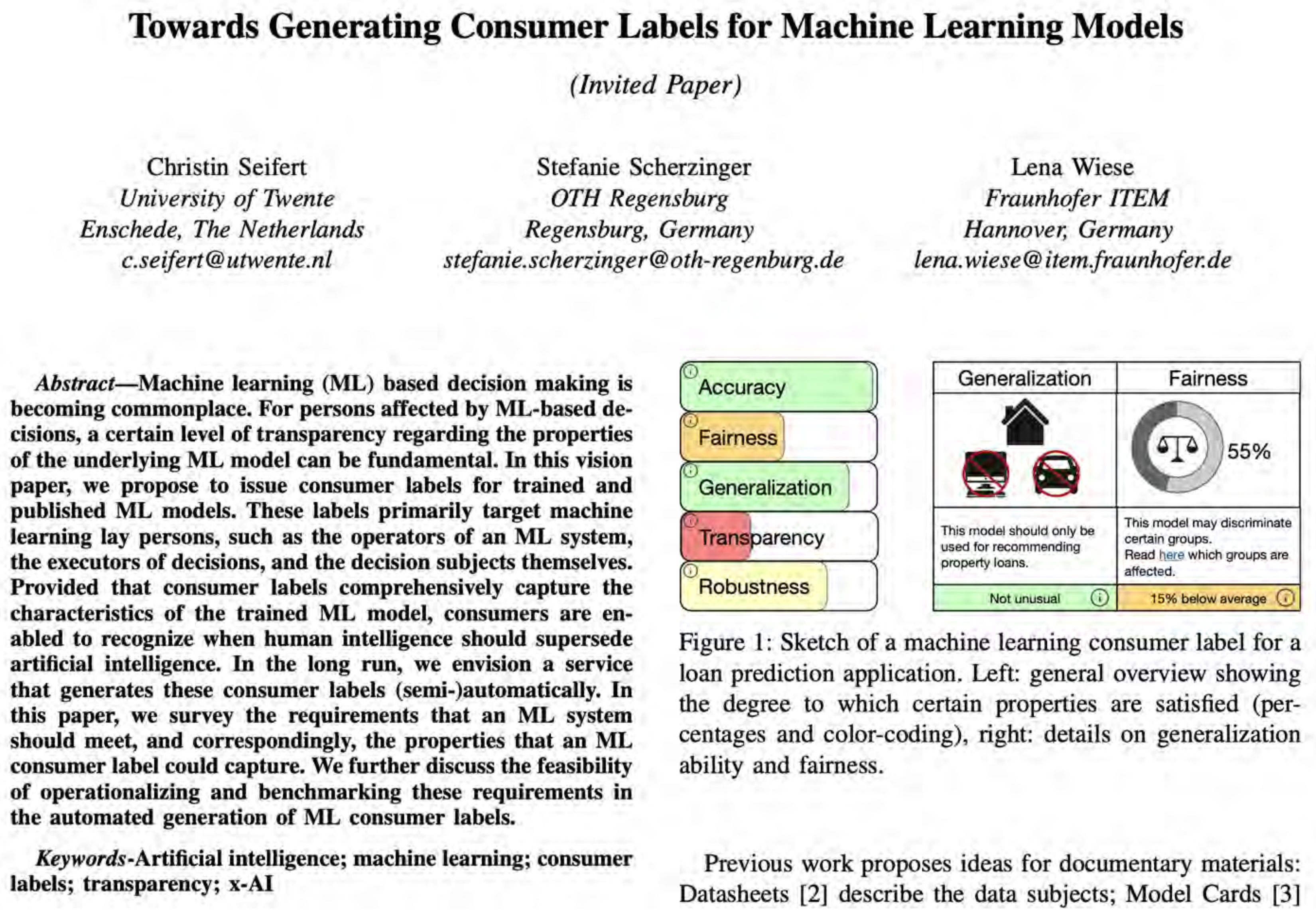
My name is Thomas and my main

2.1

This model can be loaded on the Inference API on-demand.

Spaces using `microsoft/phi-2` 105

radames/Candle-phi1-phi2-wasm-demo · mlabonne/phixtral-chat · LanguageBind/MoE-LLaVA · cvachet/pdf-chatbot · lmdemo/phi-2-demo-gpu-streaming · eson/tokenizer-arena · LixoHumano/microsoft-phi-2 · Gosula/ai_chatbot_phi2model_qlora



Towards Generating Consumer Labels for Machine Learning Models

(Invited Paper)

Christin Seifert
University of Twente
Enschede, The Netherlands
c.seifert@utwente.nl

Stefanie Scherzinger
OTH Regensburg
Regensburg, Germany
stefanie.scherzinger@oth-regenburg.de

Lena Wiese
Fraunhofer ITEM
Hannover, Germany
lena.wiese@item.fraunhofer.de

***Abstract*—Machine learning (ML) based decision making is becoming commonplace. For persons affected by ML-based decisions, a certain level of transparency regarding the properties of the underlying ML model can be fundamental. In this vision paper, we propose to issue consumer labels for trained and published ML models. These labels primarily target machine learning lay persons, such as the operators of an ML system, the executors of decisions, and the decision subjects themselves. Provided that consumer labels comprehensively capture the characteristics of the trained ML model, consumers are enabled to recognize when human intelligence should supersede artificial intelligence. In the long run, we envision a service that generates these consumer labels (semi-)automatically. In this paper, we survey the requirements that an ML system should meet, and correspondingly, the properties that an ML consumer label could capture. We further discuss the feasibility of operationalizing and benchmarking these requirements in the automated generation of ML consumer labels.**

***Keywords*-Artificial intelligence; machine learning; consumer labels; transparency; x-AI**

Accuracy
Fairness
Generalization
Transparency
Robustness
Generalization
Fairness
55%
This model should only be used for recommending property loans.
This model may discriminate certain groups.
Read here which groups are affected.
Not unusual
15% below average

Figure 1: Sketch of a machine learning consumer label for a loan prediction application. Left: general overview showing the degree to which certain properties are satisfied (percentages and color-coding), right: details on generalization ability and fairness.

Previous work proposes ideas for documentary materials: Datasheets [2] describe the data subjects; Model Cards [3]

ScHARe

Research gaps

Identifying research gaps and opportunities: a multifaceted approach

Identifying **research gaps and opportunities** is crucial for selecting a impactful project topic. Here's a multifaceted approach to achieve this:

1. **Comprehensive Literature Review:** Conduct a thorough review of existing research in your chosen field, focusing on:
   - **Recent publications (past 5-10 years):** Stay updated with current trends and emerging areas of inquiry.
   - **Recurring themes, controversies, and unanswered questions:** Identify potential areas for further investigation or knowledge gaps requiring deeper exploration.
   - **Limitations of existing studies:** Analyze methodological limitations, unexplored variables, or inconclusive findings that warrant further research.

Leveraging Google Scholar for targeted research inspiration

- Utilize **Google Scholar** and other academic search engines to explore current research trends and identify potential topics.

- Employ **effective search strategies** to refine your results:

  - Relevant **keywords and phrases**: Use quotation marks around specific phrases for accurate matching.
  - **Boolean operators** (AND, OR, NOT): Combine keywords to narrow down your search and focus on specific aspects of your research interest.
  - **Advanced search features**: Utilize **filters** offered by search engines to refine results by publication date, author, or specific fields within the research article.

Identifying research gaps and opportunities: a multifaceted approach

- Pat**ent databases** offer insights into technological advancements and can spark research ideas.

- Explore patent databases to identify novel technologies, potential research gaps, and unmet needs.

Tips:

- Utilize keywords relevant to your field of interest.
- Search by inventor, assignee, or publication date.
- Analyze the claims section of patents to understand the innovative aspects of the invention.

ScHARe

Datasets

Selecting the right dataset for your project

Choosing the **appropriate dataset** is critical for the success of your research project.

Consider the following **factors** when selecting a dataset:

1. **Relevance:** Ensure the dataset aligns with your research question and objectives.
    - **Tips:**
        - Clearly define your research question and identify the variables needed to address it.
        - Evaluate the dataset description and variable list to ensure they match your research needs.
2. **Quality:** Assess the data for accuracy, completeness, and consistency.
    - **Tips:**
        - Look for information about data collection methods, quality control procedures, and potential limitations.
        - Check for missing values, outliers, and inconsistencies in the data.

Selecting the right dataset for your project

3. **Accessibility:** Determine if the dataset is publicly available or requires permission for access.

- **Tips:**
  - Explore **data repositories and platforms** relevant to your field of research.
  - Contact data owners or custodians if permission is required for access.

Importance of in-depth dataset analysis before selection

Before diving into analysis or model building, conducting an in-depth analysis of the dataset is crucial.

Here’s why:

1. **Understanding Data Characteristics:**

- **Data quality:** Assess completeness, consistency, accuracy, and representativeness of the data. Identify potential issues like missing values, outliers, and biases.
- **Data suitability:** Evaluate whether the data aligns with your research question or model objective. Consider factors like variable relevance, data granularity, and temporal coverage.
- **Data exploration:** Gain insights into the data distribution, relationships between variables, and potential patterns through techniques like visualization and summary statistics.

Importance of in-depth dataset analysis before selection

2. Allow Informed Decision-Making:

Data selection: Based on the analysis, you might decide to:

- Refine your research question or model objective based on the data's limitations.
- Seek alternative datasets that better suit your needs.
- Implement data cleaning and preprocessing techniques to address identified issues.

Feature engineering: The analysis informs decisions about creating new features from existing ones, potentially enhancing model performance.

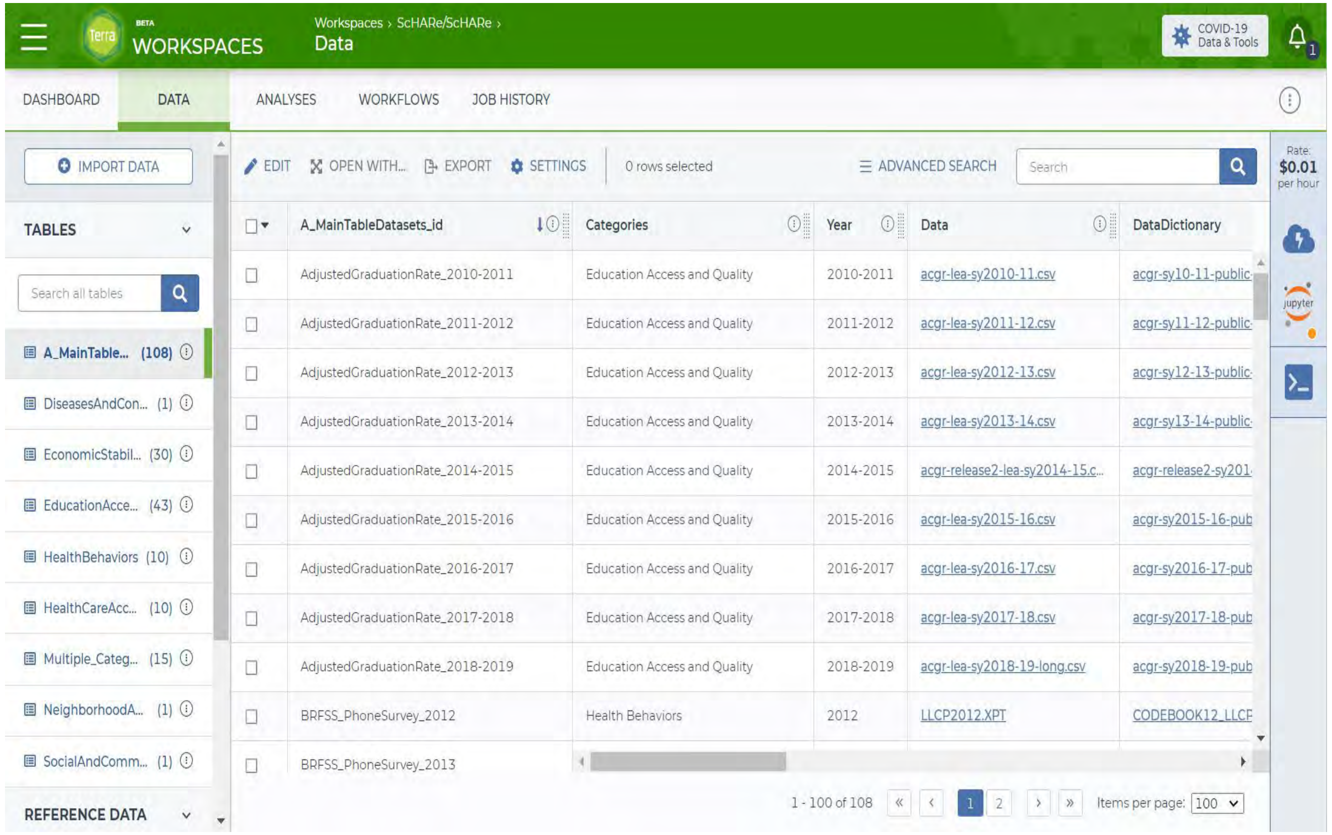
Model selection: Understanding the data characteristics helps choose appropriate machine learning algorithms or statistical methods for your analysis.

SDoH-related datasets available on ScHARe: a valuable resource

ScHARe provides a valuable platform for researchers seeking **SDoH-related data.**

Explore the available datasets to identify potential resources that align with your research interests in social determinants of health and their impact on various health outcomes.

Remember to consider the relevance of each dataset to your specific research question and aims.



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

ScHARe Ecosystem

The ScHARe Data Ecosystem is comprised of:

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

   Example: *American Community Survey (ACS)*

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

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

3. **ScHARe Hosted Project Datasets:** 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*

ScHARe Ecosystem: Google hosted datasets

Examples of interesting datasets include:

- **American Community Survey** (U.S. Census Bureau)
- **US Census Data** (U.S. Census Bureau)
- **Area Deprivation Index** (BroadStreet)
- **GDP and Income by County** (Bureau of Economic Analysis)
- **US Inflation and Unemployment** (U.S. Bureau of Labor Statistics)
- **Quarterly Census of Employment and Wages** (U.S. Bureau of Labor Statistics)
- **Point-in-Time Homelessness Count** (U.S. Dept. of Housing and Urban Development)
- **Low Income Housing Tax Credit Program** (U.S. Dept. of Housing and Urban Development)
- **US Residential Real Estate Data** (House Canary)
- **Center for Medicare and Medicaid Services - Dual Enrollment** (U.S. Dept. of Health & Human Services)
- **Medicare** (U.S. Dept. of Health & Human Services)
- **Health Professional Shortage Areas** (U.S. Dept. of Health & Human Services)
- **CDC Births Data Summary** (Centers for Disease Control)
- **COVID-19 Data Repository by CSSE at JHU** (Johns Hopkins University)
- **COVID-19 Mobility Impact** (Geotab)
- **COVID-19 Open Data** (Google BigQuery Public Datasets Program)
- **COVID-19 Vaccination Access** (Google BigQuery Public Datasets Program)

ScHARe Ecosystem: ScHARe hosted datasets

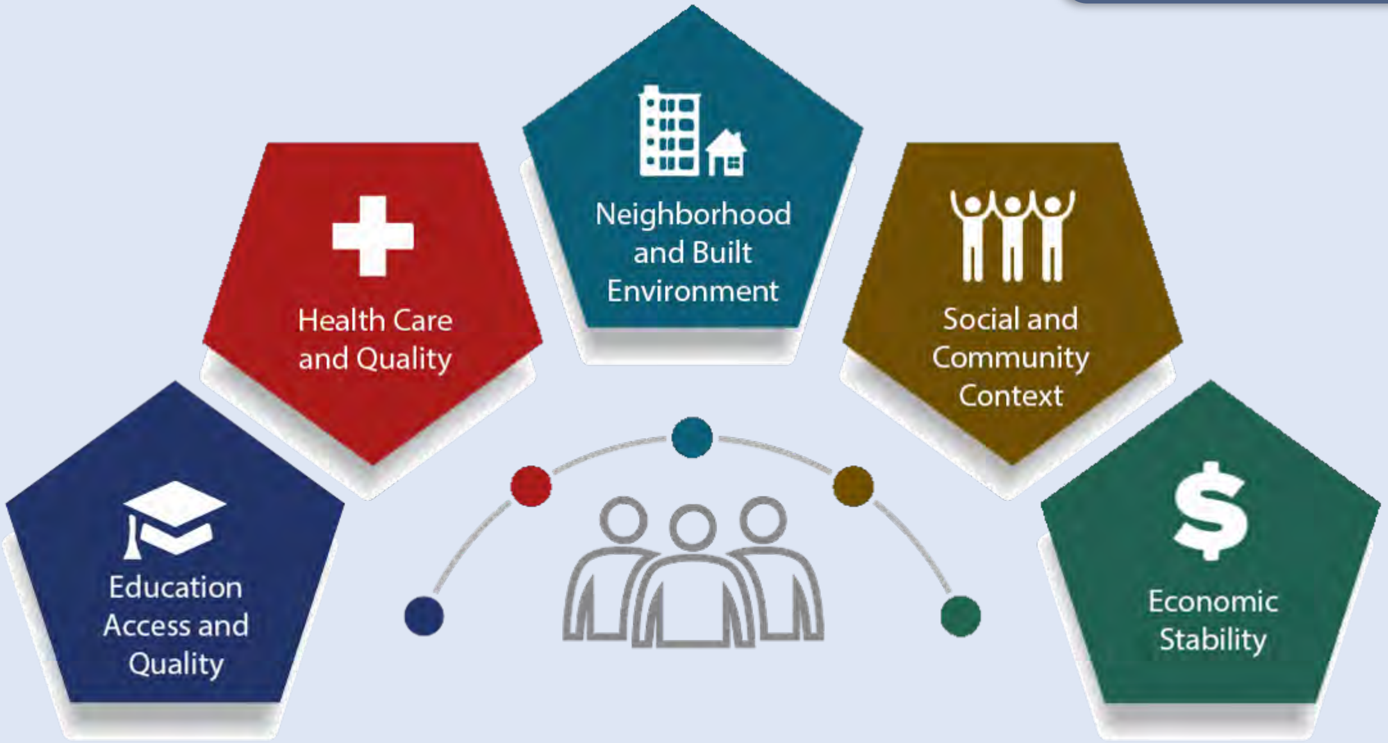
Organized based on the **CDC SDoH categories**, with the addition of *Health Behaviors* and *Diseases and Conditions*:

200+ datasets

- **What are the Social Determinants of Health?**

Social determinants of health (SDoH) are the **nonmedical factors that influence health outcomes.**

They are the **conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life.**


Health Care and Quality
Neighborhood and Built Environment
Social and Community Context
Education Access and Quality
Economic Stability

www.cdc.gov/about/sdoh/index.html

ScHARe Ecosystem: ScHARe hosted datasets

Examples of datasets for each category include:

Education access and quality

Data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.

Examples:

- **EDFacts Data Files** (U.S. Dept. of Education) - Graduation rates and participation/proficiency assessment
- **NHES - National Household Education Surveys Program** (U.S. Dept. of Education) – Educational activities

ScHARe Ecosystem: ScHARe hosted datasets

Health care access and quality

Data on health literacy, use of health IT, emergency room waiting times, preventive healthcare, health screenings, treatment of substance use disorders, family planning services, access to a primary care provider and high quality care, access to telehealth and electronic exchange of health information, access to health insurance, adequate oral care, adequate prenatal care, STD prevention measures, etc.

Example:

- **MEPS - Medical Expenditure Panel Survey** (AHRQ) - Cost and use of healthcare and health insurance coverage
- **Dartmouth Atlas Data -** Selected Primary Care Access and Quality Measures - Measures of primary care utilization, quality of care for diabetes, mammography, leg amputation and preventable hospitalizations

ScHARe Ecosystem: ScHARe hosted datasets

Neighborhood and built environment

Data on access to broadband internet, access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, deaths from motor vehicle crashes, asthma and COPD cases and hospitalizations, noise exposure, smoking, mass transit use, etc.

Examples:

- **National Environmental Public Health Tracking Network** (CDC) - Environmental indicators and health, exposure, and hazard data
- **LATCH - Local Area Transportation Characteristics for Households** (U.S. Dept. of Transportation) – Local transportation characteristics for households

ScHARe Ecosystem: ScHARe hosted datasets

Social and community context

Data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc.

Example:

- **Hate crime statistics** (FBI) - Data on crimes motivated by bias against race, gender identity, religion, disability, sexual orientation, or ethnicity
- **General Social Survey** (GSS) - Data on a wide range of characteristics, attitudes, and behaviors of Americans.

ScHARe Ecosystem: ScHARe hosted datasets

Economic stability

Data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.

Examples:

- **Current Population Survey (CPS) Annual Social and Economic Supplement** (U.S. Bureau of Labor Statistics) - Labor force statistics: annual work activity, income, health insurance, and health
- **Food Access Research Atlas** (U.S. Dept. of Agriculture) – Food access indicators for low-income and other census tracts

ScHARe Ecosystem: ScHARe hosted datasets

Health behaviors

Data on health-related practices that can directly affect health outcomes.

Examples:

- **BRFSS - Behavioral Risk Factor Surveillance System** (CDC) - State-level data on health-related risk behaviors, chronic health conditions, and use of preventive services
- **YRBSS - Youth Risk Behavior Surveillance System** (CDC) – Health behaviors that contribute to the leading causes of death, disability, and social problems among youth and adults

ScHARe Ecosystem: ScHARe hosted datasets

Diseases and conditions

Data on incidence and prevalence of specific diseases and health conditions.

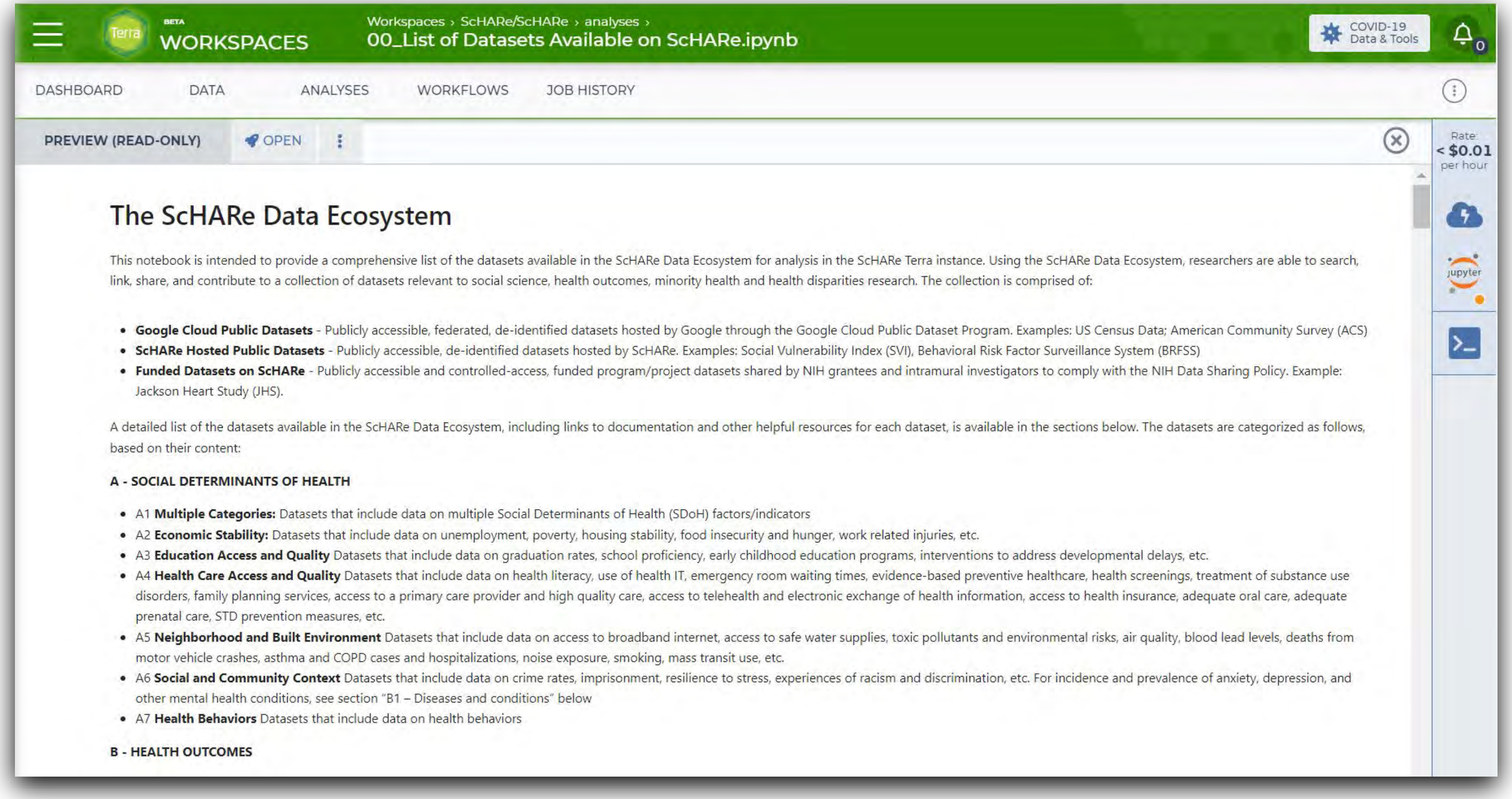
Examples:

- **U.S. CDI - Chronic Disease Indicators** (CDC) - 124 chronic disease indicators important to public health practice
- **UNOS - United Network of Organ Sharing** (Health Resources and Services Administration) – Organ transplantation: cadaveric and living donor characteristics, survival rates, waiting lists and organ disposition

How to check what data is available on ScHARe

Analyses tab

In the **Analyses** tab in the ScHARe workspace, the notebook **00_List of Datasets Available on ScHARe** lists all of the datasets available in the ScHARe Datasets collection



The ScHARe Data Ecosystem

This notebook is intended to provide a comprehensive list of the datasets available in the ScHARe Data Ecosystem for analysis in the ScHARe Terra instance. Using the ScHARe Data Ecosystem, researchers are able to search, link, share, and contribute to a collection of datasets relevant to social science, health outcomes, minority health and health disparities research. The collection is comprised of:

- **Google Cloud Public Datasets** - Publicly accessible, federated, de-identified datasets hosted by Google through the Google Cloud Public Dataset Program. Examples: US Census Data; American Community Survey (ACS)
- **ScHARe Hosted Public Datasets** - Publicly accessible, de-identified datasets hosted by ScHARe. Examples: Social Vulnerability Index (SVI), Behavioral Risk Factor Surveillance System (BRFSS)
- **Funded Datasets on ScHARe** - Publicly accessible and controlled-access, funded program/project datasets shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy. Example: Jackson Heart Study (JHS).

A detailed list of the datasets available in the ScHARe Data Ecosystem, including links to documentation and other helpful resources for each dataset, is available in the sections below. The datasets are categorized as follows, based on their content:

A - SOCIAL DETERMINANTS OF HEALTH

- A1 **Multiple Categories:** Datasets that include data on multiple Social Determinants of Health (SDoH) factors/indicators
- A2 **Economic Stability:** Datasets that include data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.
- A3 **Education Access and Quality** Datasets that include data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.
- A4 **Health Care Access and Quality** Datasets that include data on health literacy, use of health IT, emergency room waiting times, evidence-based preventive healthcare, health screenings, treatment of substance use disorders, family planning services, access to a primary care provider and high quality care, access to telehealth and electronic exchange of health information, access to health insurance, adequate oral care, adequate prenatal care, STD prevention measures, etc.
- A5 **Neighborhood and Built Environment** Datasets that include data on access to broadband internet, access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, deaths from motor vehicle crashes, asthma and COPD cases and hospitalizations, noise exposure, smoking, mass transit use, etc.
- A6 **Social and Community Context** Datasets that include data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc. For incidence and prevalence of anxiety, depression, and other mental health conditions, see section "B1 – Diseases and conditions" below
- A7 **Health Behaviors** Datasets that include data on health behaviors

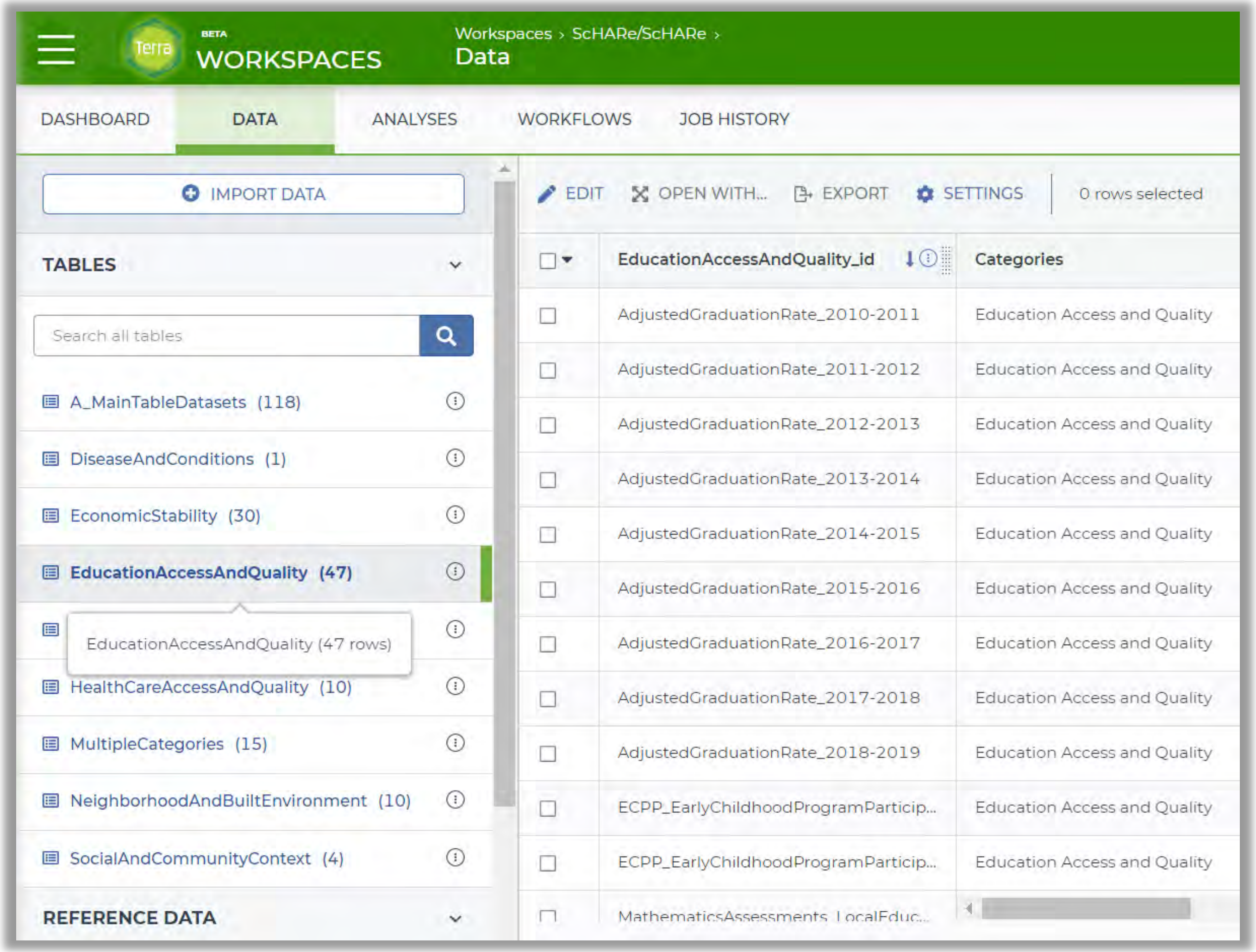
B - HEALTH OUTCOMES

How to access available data on ScHARe

Data tab

In the **Data** tab in the ScHARe workspace, **data tables help access ScHARe data and keep track of your project data:**

- In the ScHARe workspace, click on the Data tab
- Under Tables, you will see a list of dataset categories
- If you click on a category, you will see a list of relevant datasets
- Scroll to the right to learn more about each dataset



A_MainTableDatasets (118)
DiseaseAndConditions (1)
EconomicStability (30)
EducationAccessAndQuality (47)
EducationAccessAndQuality (47 rows)
HealthCareAccessAndQuality (10)
MultipleCategories (15)
NeighborhoodAndBuiltEnvironment (10)
SocialAndCommunityContext (4)

□	EducationAccessAndQuality_id	Categories
□	AdjustedGraduationRate_2010-2011	Education Access and Quality
□	AdjustedGraduationRate_2011-2012	Education Access and Quality
□	AdjustedGraduationRate_2012-2013	Education Access and Quality
□	AdjustedGraduationRate_2013-2014	Education Access and Quality
□	AdjustedGraduationRate_2014-2015	Education Access and Quality
□	AdjustedGraduationRate_2015-2016	Education Access and Quality
□	AdjustedGraduationRate_2016-2017	Education Access and Quality
□	AdjustedGraduationRate_2017-2018	Education Access and Quality
□	AdjustedGraduationRate_2018-2019	Education Access and Quality
□	ECPP_EarlyChildhoodProgramParticip...	Education Access and Quality
□	ECPP_EarlyChildhoodProgramParticip...	Education Access and Quality
□	MathematicsAssessments_LocalEduc...	

ScHARe

Common data elements (CDEs)

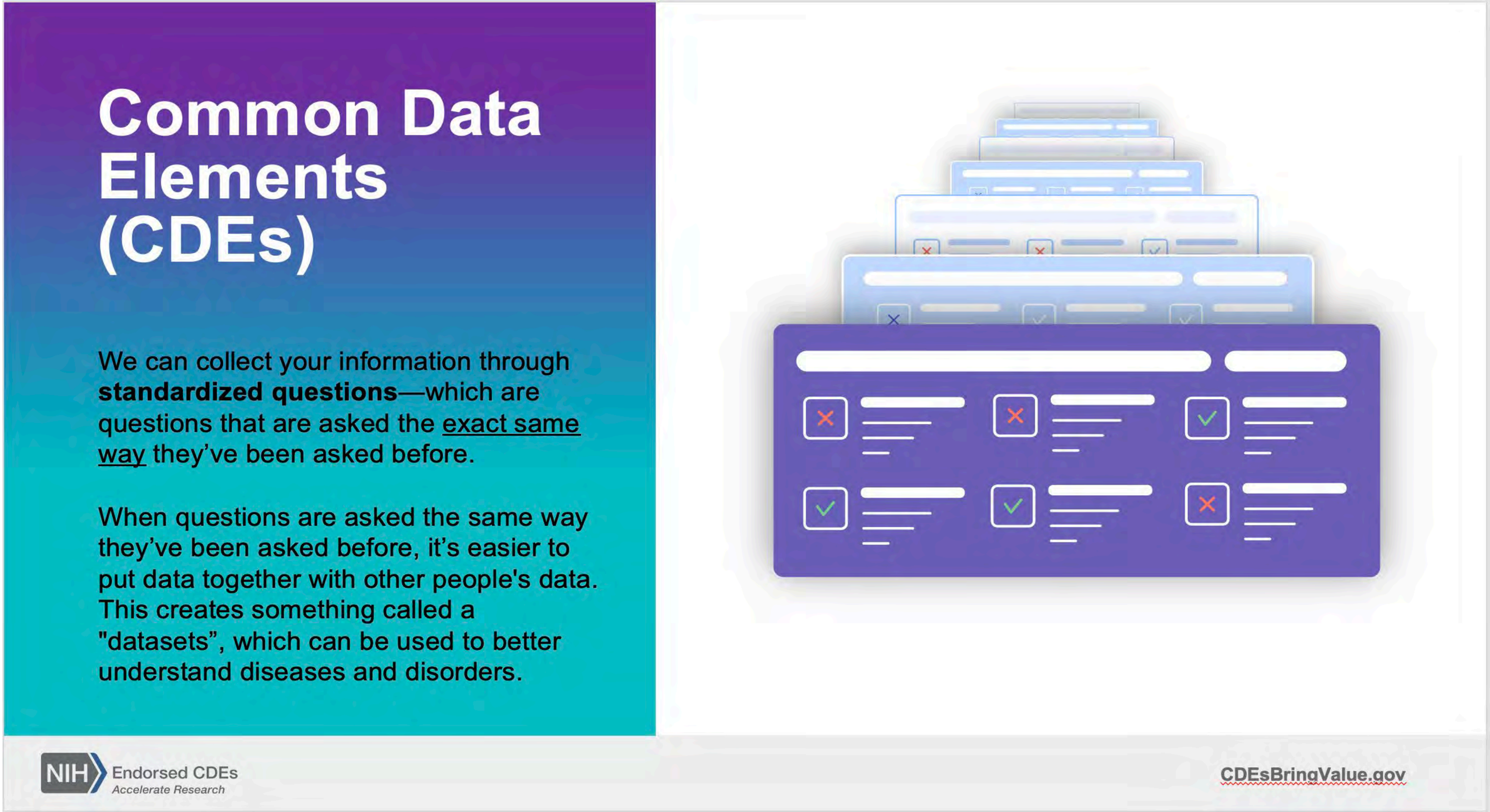
Understanding Common Data Elements (CDEs): ensuring standardization

- **Common Data Elements (CDEs)** are standardized data elements used in research studies.

- **Definition:** Common Data Elements (CDEs) are standardized, precisely defined data points used consistently across different research studies or clinical trials. They act as building blocks for collecting and sharing data in a comparable and interoperable manner.

NIH CDE
Repository

Understanding Common Data Elements (CDEs)



Common Data Elements (CDEs)

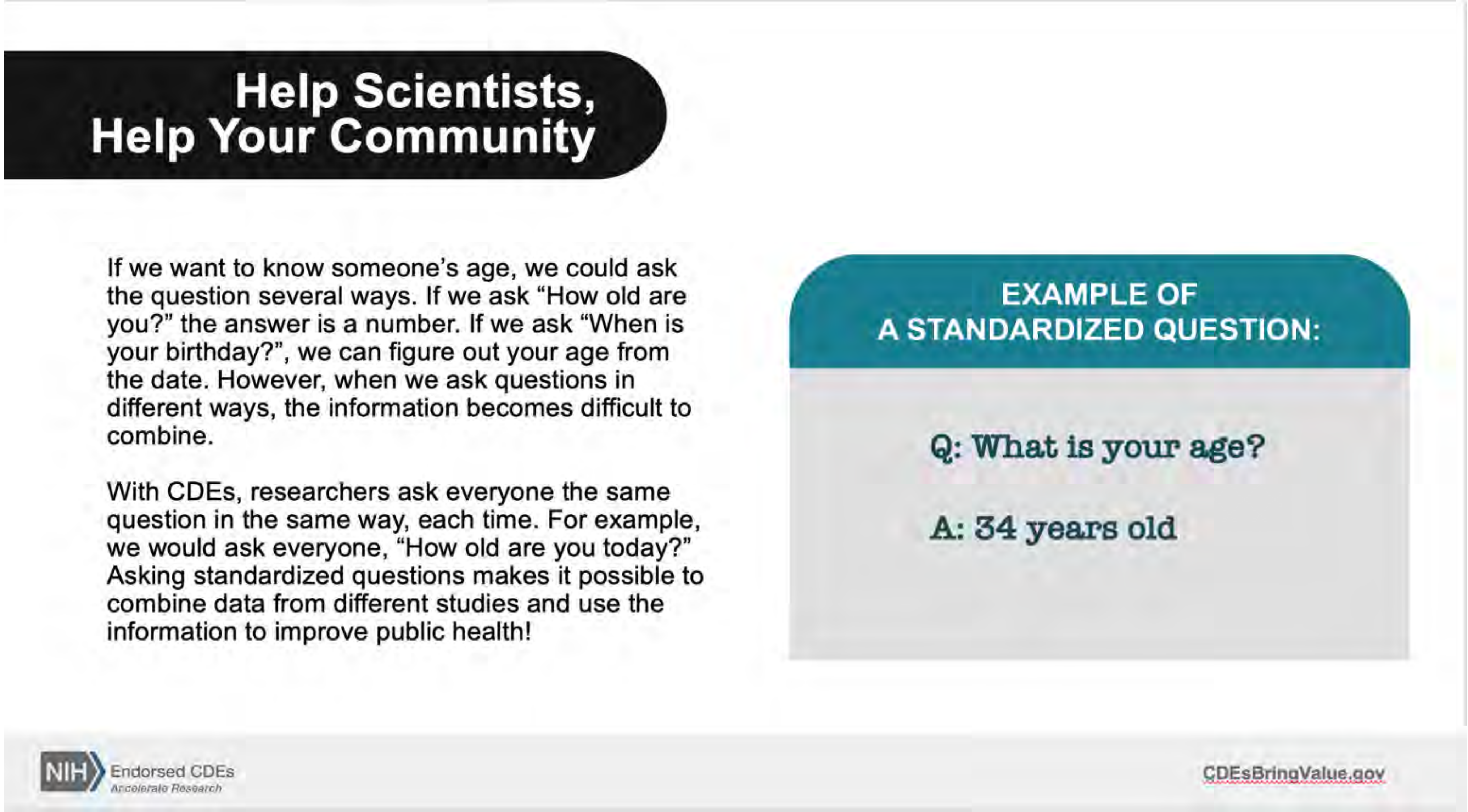
We can collect your information through **standardized questions**—which are questions that are asked the exact same way they've been asked before.

When questions are asked the same way they've been asked before, it's easier to put data together with other people's data. This creates something called a "datasets", which can be used to better understand diseases and disorders.

NIH Endorsed CDEs *Accelerate Research*

CDEsBringValue.gov

Understanding Common Data Elements (CDEs)



Help Scientists, Help Your Community

If we want to know someone's age, we could ask the question several ways. If we ask "How old are you?" the answer is a number. If we ask "When is your birthday?", we can figure out your age from the date. However, when we ask questions in different ways, the information becomes difficult to combine.

With CDEs, researchers ask everyone the same question in the same way, each time. For example, we would ask everyone, "How old are you today?" Asking standardized questions makes it possible to combine data from different studies and use the information to improve public health!

EXAMPLE OF
A STANDARDIZED QUESTION:

Q: What is your age?

A: 34 years old

NIH Endorsed CDEs
Accelerate Research

CDEsBringValue.gov

Importance of CDEs

- **Enhanced comparability of findings:** CDEs enable researchers to compare findings from different studies more easily, leading to a broader understanding of the research topic.
- **Reduced redundancy:** By using standardized elements, researchers avoid collecting the same data repeatedly, saving time and resources.
- **Improved data quality:** CDEs come with clear definitions and standardized formats, minimizing errors and inconsistencies in data collection.
- **Facilitated data sharing:** CDEs promote open science by enabling researchers to share and reuse data efficiently across institutions and disciplines.

Labels of ScHARe Core Common Data Elements

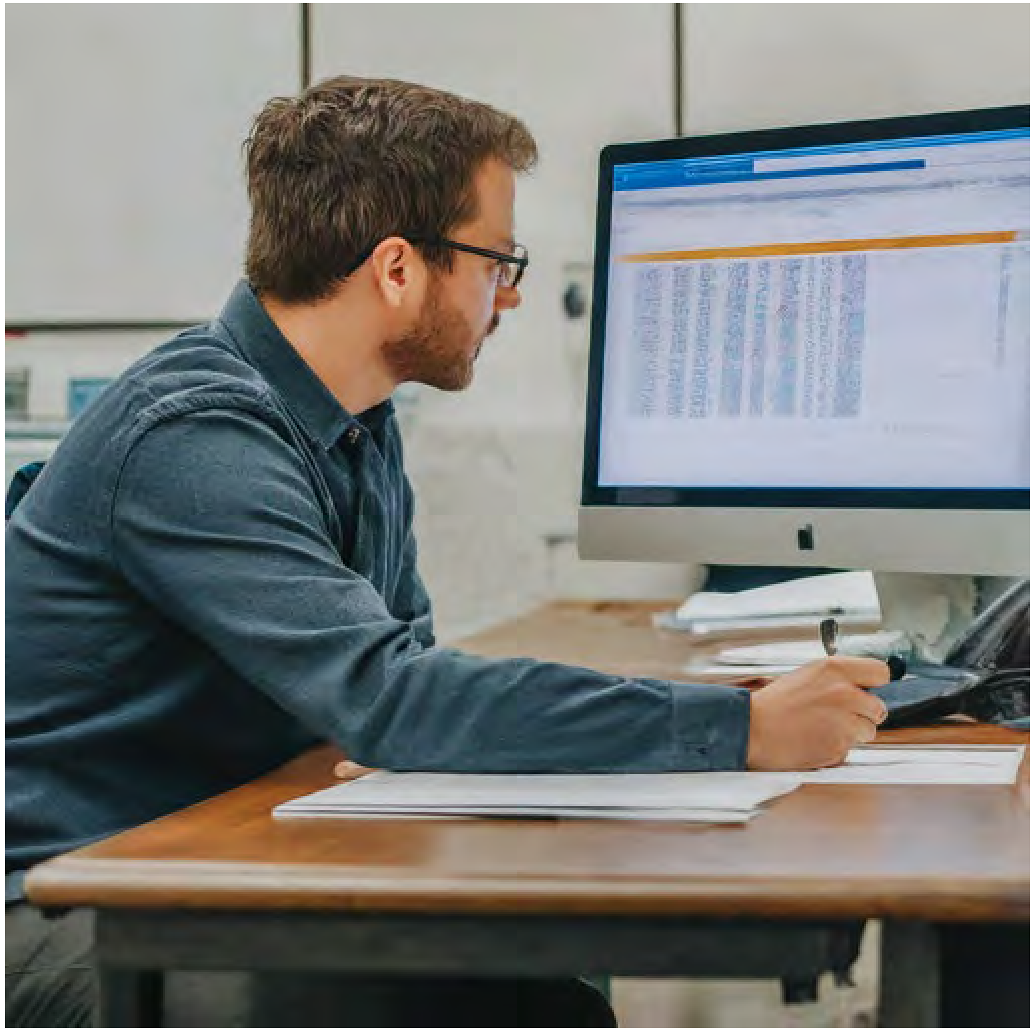
NIH CDE Repository: https://cde.nlm.nih.gov/home

NIH Endorsed

- 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 (and Associated Medications)

- **NIMHD Framework**
- **Health Disparity Outcomes**

Checking datasets for CDEs: promoting compatibility and collaboration



When selecting a dataset, check if it utilizes CDEs relevant to your research field.

Look for information about the specific CDEs used in the dataset and their corresponding variable names.

- When selecting a dataset, check if it utilizes CDEs relevant to your research field.
- Look for information about the specific CDEs used in the dataset and their corresponding variable names.

- If CDEs are not used, consider the potential challenges in data comparability with other studies.

Identifying opportunities for merging datasets

Merging datasets using CDEs can unlock valuable insights and enhance research power. Here are some strategies for identifying opportunities:

- **Content overlap:** Assess whether datasets share similar content areas or address related research questions.
- **Variable comparability:** Evaluate if datasets contain variables with comparable definitions and formats that can be meaningfully merged.
- **Temporal alignment:** Consider the timeframes covered by each dataset and ensure they align for meaningful integration.
- **Complementary information:** Identify if datasets offer complementary information that can enrich the combined analysis when merged.

Challenges and considerations

Merging datasets using CDEs also presents challenges:

- **Data quality discrepancies:** Ensure data quality is addressed before merging to minimize biases and inconsistencies.
- **Harmonization efforts:** Careful harmonization of data elements and formats might be necessary for successful integration.
- **Ethical considerations:** Adhere to ethical guidelines regarding data privacy, informed consent, and responsible data sharing practices.

CDE poll

In your opinion, which factor presents the biggest obstacle to the widespread adoption of CDEs in research?

ScHARe

Making data AI-ready

Making datasets AI-ready: a multifaceted approach

Making datasets AI-ready involves ensuring they are suitable for use in machine learning and artificial intelligence applications.

Key aspects of AI-ready datasets:

- **Data quality:** Ensure data accuracy, completeness, and consistency. Address missing values, outliers, and inconsistencies that could impact model performance.
- **Data cleaning and pre-processing:** Apply techniques like normalization, scaling, and encoding to prepare the data for machine learning algorithms.
- **Feature engineering:** Create new features from existing data or transform existing features to improve model performance.
- **Documentation:** Provide clear and detailed documentation about the dataset, including variable definitions, data collection methods, and any transformations applied.

Why quality checks are essential for AI-ready data

Datasets are the lifeblood of AI models. Their quality directly impacts the performance, fairness, and reliability of the resulting models.

Poor quality data can lead to:

- **Biased models:** Unrepresentative or skewed data can lead to models that perpetuate existing biases and produce discriminatory outcomes.
- **Inaccurate results:** Inconsistent or erroneous data can cause models to learn incorrect patterns and generate unreliable predictions.
- **Wasted resources:** Training models on low-quality data is a waste of time, computational power, and financial resources.

Overview of quality checks

Quality checks for AI-ready datasets encompass various aspects, categorized into these key areas:

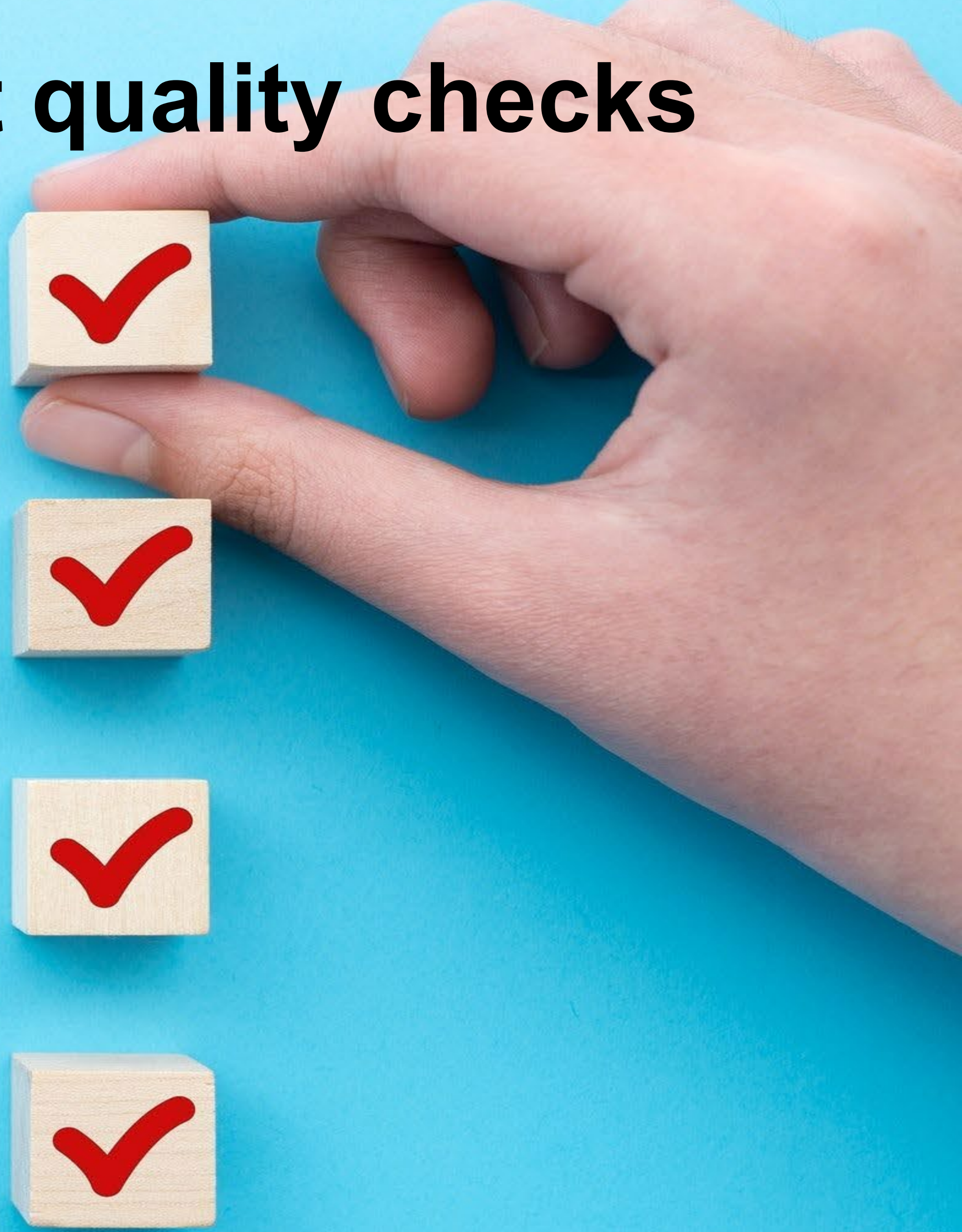
1. **Data completeness:**
   1. **Missing values:** Identifying and handling missing data points through imputation or removal.
   2. **Outliers:** Detecting and addressing unusual data points that might skew model training.
2. **Data consistency:**
   1. **Formatting:** Ensuring consistent data formats across the entire dataset.
   2. **Units and labels:** Maintaining consistency in units of measurement and data labeling.
3. **Data accuracy:**
   1. **Verification:** Cross-checking data with reliable sources to identify and correct errors.
   2. **Validation:** Comparing data against expected values or domain knowledge to ensure accuracy.

Overview of quality checks

4. **Data representativeness:**
    1. **Bias:** Analyzing the data for potential biases in sampling, labeling, or other aspects.
    2. **Generalizability:** Assessing whether the data adequately represents the target population for the intended AI application.

5. **Data documentation:**
    1. **Metadata:** Providing comprehensive information about the data, including its origin, collection methods, and usage guidelines.
    2. **Version control:** Maintaining clear versioning of the data to track changes and ensure consistency.

Checklist for AI-ready dataset quality checks



Data completeness:

Check for missing values and implement appropriate handling strategies. Identify and address outliers.

Data consistency:

Ensure consistent formatting throughout the dataset.

Verify consistency in units and labels.

Data accuracy:

Perform data verification against reliable sources.

Validate data against expected values or domain knowledge.

Data representativeness:

Analyze the data for potential biases.

Assess the generalizability of the data to the target population.

Data documentation:

Create comprehensive metadata describing the data.

Implement version control mechanisms.

Importance of completeness and data dictionaries for AI-ready datasets

Quality checks for AI-ready datasets encompass various aspects, categorized into these key areas:

1. **Data completeness:**
    1. **Missing values:** Identifying and handling missing data points through imputation or removal.
    2. **Outliers:** Detecting and addressing unusual data points that might skew model training.

2. **Data consistency:**
    1. **Formatting:** Ensuring consistent data formats across the entire dataset.
    2. **Units and labels:** Maintaining consistency in units of measurement and data labeling.

3. **Data accuracy:**
    1. **Verification:** Cross-checking data with reliable sources to identify and correct errors.
    2. **Validation:** Comparing data against expected values or domain knowledge to ensure accuracy.

Importance of completeness and data dictionaries for AI-ready datasets

Two critical aspects of ensuring datasets are AI-ready are completeness and data dictionaries. Let's explore why each is crucial:

1. Completeness:

A complete dataset refers to one with minimal missing values or outliers that could significantly impact the training and performance of AI models. Missing data can lead to:

- **Biased models:** if specific data points are consistently missing, the model might learn skewed patterns and produce unfair results.
- **Inaccurate predictions:** missing data can hinder the model's ability to capture the full picture and lead to unreliable outputs.
- **Inefficient training:** training models on incomplete data can be computationally expensive and inefficient, yielding suboptimal results.

Importance of completeness and data dictionaries for AI-ready datasets

2. Data dictionaries:

Data dictionaries act as the instruction manuals for your dataset, providing crucial information about each variable. They define:

- **Variable names:** clear and consistent names that facilitate understanding and avoid confusion.
- **Data types:** specifying the format of data (e.g., Numerical, categorical, text) ensures proper interpretation by the model.
- **Descriptions:** explanations of the meaning and potential values of each variable, promoting clarity and reducing ambiguity.
- **Units of measurement:** standardizing units (e.g., Meters, kilometers) ensures consistent interpretation and analysis.

Addressing missing data: strategies for imputation

- Missing data is a common challenge in datasets, and how you handle it can significantly impact your research findings.
- Strategies for handling missing data:

  - **Deletion:** Remove rows or columns with a high percentage of missing values, but this can lead to information loss.
  - **Mean/median imputation:** Replace missing values with the mean or median of the respective variable.
  - **Model-based imputation:** Use statistical models to predict missing values based on other variables in the dataset.

Understanding and addressing missing data

Data Missingness Strategies: Understanding and Addressing Missing Data

Missing data, where values are absent from a dataset, is a prevalent challenge in various fields. It can significantly impact the results of data analysis and machine learning models. Fortunately, various strategies exist to address missing data

Understanding Missing Data:

Before delving into strategies, it's crucial to understand the **types of missing data**:

- **Missing Completely at Random (MCAR):** Missingness occurs randomly and is unrelated to any other variables in the dataset.
- **Missing at Random (MAR):** Missingness depends on observable variables in the dataset but not on the missing values themselves.
- **Missing Not at Random (MNAR):** Missingness is related to the missing values themselves, often due to unobserved factors.

Understanding and addressing missing data

Addressing Missing Data:

Several strategies can be employed to handle missing data, depending on the nature and extent of missingness:

1. **Deletion:**

❑ **Listwise deletion:** Removes entire rows with missing values, potentially reducing sample size and introducing bias if MCAR doesn't hold.

❑ **Pairwise deletion:** Removes only the data points with missing values for the variable being analyzed, potentially wasting information.

Understanding and addressing missing data

2. Imputation:

❑ **Mean/Median/Mode imputation:** Replaces missing values with the average, median, or most frequent value of the variable, respectively. Simple but may introduce bias, especially for skewed distributions.

❑ **Hot Deck imputation:** Replaces missing values with values from existing observations with similar characteristics, reducing bias but potentially introducing noise.

❑ **Model-based imputation:** Uses statistical models like regression or machine learning to predict missing values based on other variables, potentially more accurate but computationally expensive.

Dealing with proxies and small sample sizes: alternative approaches

- **Not all research questions may have readily available data for every variable**. In such cases, researchers might need to employ **proxy variables** or navigate situations with small sample sizes.

- Strategies for addressing proxies and small sample sizes:

  - **Proxy variables:** Carefully select proxy variables that are demonstrably linked to the desired variable, but be mindful of potential limitations and biases.

  - **Small sample size analysis:** Utilize appropriate statistical methods designed for small datasets, such as non-parametric tests or bootstrapping techniques.

Exploring the ethical considerations of using synthetic data

- Synthetic data generation involves creating artificial data that resembles real-world data, but protects privacy and confidentiality.

- While synthetic data offers certain advantages, its use raises **ethical considerations** that researchers must address responsibly:

    - **Transparency and disclosure:** Clearly communicate the use of synthetic data, including the number of actual people used to generate it, and its limitations to avoid misinterpretations.

    - **Responsible use:** Ensure the synthetic data is used ethically and does not perpetuate harmful stereotypes or discriminatory practices.

    - **Potential biases:** Be mindful of generalizability limitations and potential biases that might be introduced during the synthetic data generation process, and take steps to mitigate them.

Avoiding perpetuating bad AI: mitigating bias in datasets

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

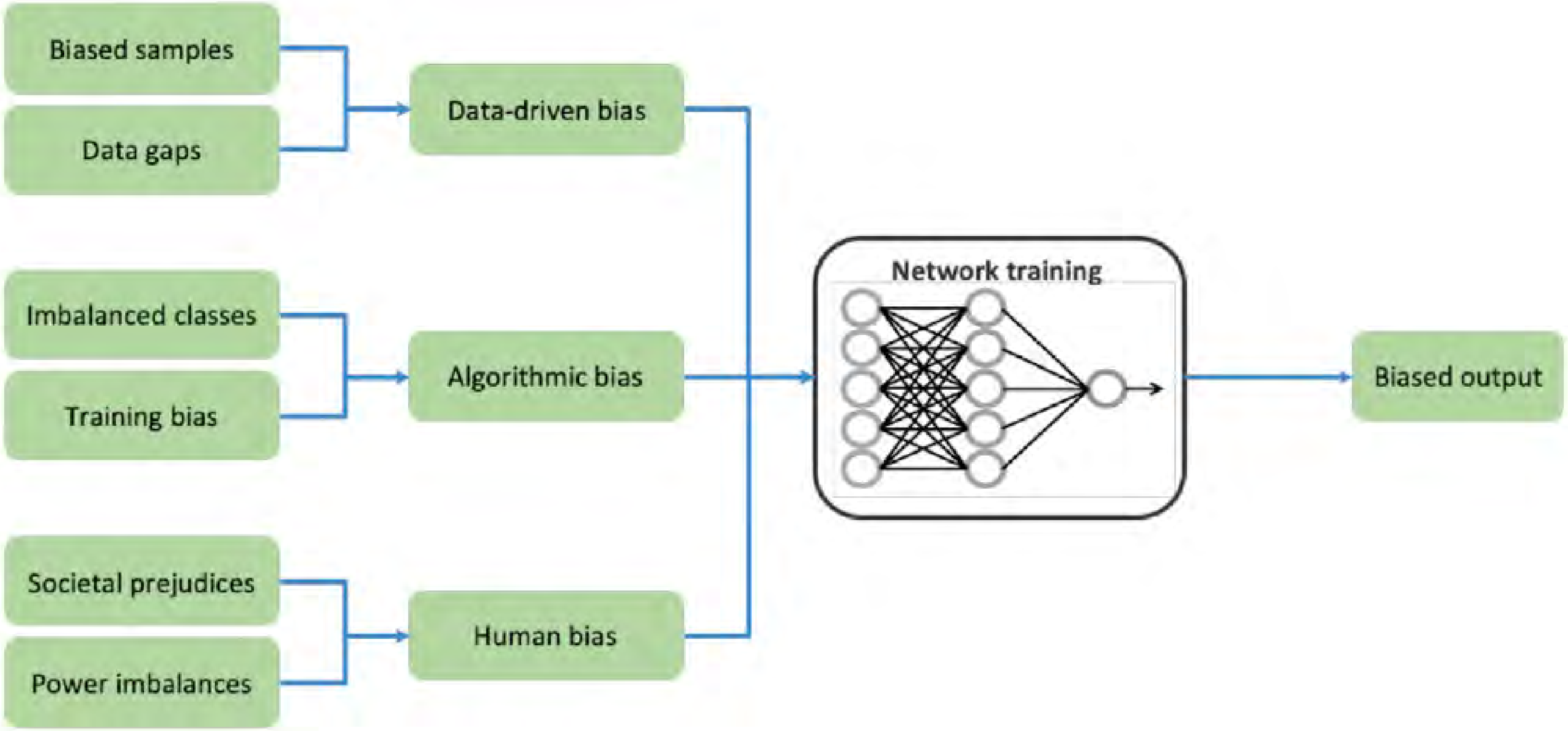
- Training data that specifies what the correct outputs are for some people/objects is used to learn a model which is then applied to other people/objects to make predictions about the correct outputs for them

- **Datasets can perpetuate bias if they contain inherent biases reflecting societal inequalities or discriminatory practices.**

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

Algorithmic bias mechanisms

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



Example: 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**.
- However, the algorithms used are basing most of their knowledge on a repository of **skin images from primarily fair-skinned populations.**
- **Bias emanates from unrepresentative training data that reflects historical inequalities:** decades of clinical research have focused primarily on people with light skin.
- The solution: **expand the archive to include as many skin types as possible**

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.

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

Avoiding perpetuating bad AI: mitigating bias in datasets

Strategies to mitigate bias in datasets:

1. **Identify potential sources of bias:** Analyze data collection methods, sampling procedures, and variable selection for potential biases. Testing for biases in datasets and algorithmic models is **crucial for ensuring fairness and reliability** in data science.

2. **Utilize bias mitigation techniques:** Apply techniques like data balancing, weighting, or fairness-aware algorithms to mitigate bias in the data.

3. **Promote transparency and responsible AI practices:** Document the limitations of the data and potential biases to ensure responsible use of AI models trained on the dataset.

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

AI-ready data poll

In your opinion, what are the biggest challenges researchers face in ensuring their datasets are truly 'AI-ready' beyond the technical aspects?

ScHARe

Research project

Formulating your research question

- **Clearly define your research question.** This forms the foundation of your entire project.

- Characteristics of a **good research question**:

  - **Specific:** Focuses on a particular aspect of a broader topic.
  - **Feasible:** Achievable within the constraints of available data and resources.
  - **Measurable:** Allows for data collection and analysis to answer the question.
  - **Relevant:** Addresses a significant gap in knowledge or has practical implications.

Launching your successful secondary data analysis project

- By following these steps and considerations, you can effectively launch your secondary data analysis research project.
- Remember:
  - Define your **research question and aims**.
  - Conduct a thorough **literature review** and explore diverse data sources.
  - **Select a dataset** that aligns with your research needs and ensures data quality.
  - Apply appropriate **data cleaning** and pre-processing techniques.
  - Address **potential biases and limitations** in the data.
  - Utilize **ethical practices throughout** your research process.

Next time

- Selecting computational strategies
- Algorithm testing and implementation
- Publishing research
- Research Think-a-Thons: brainstorming projects

ScHARe

Resources

ScHARe resources

Support made available to users:

ScHARe-specific

- ScHARe documentation
- Email support

Platform-specific

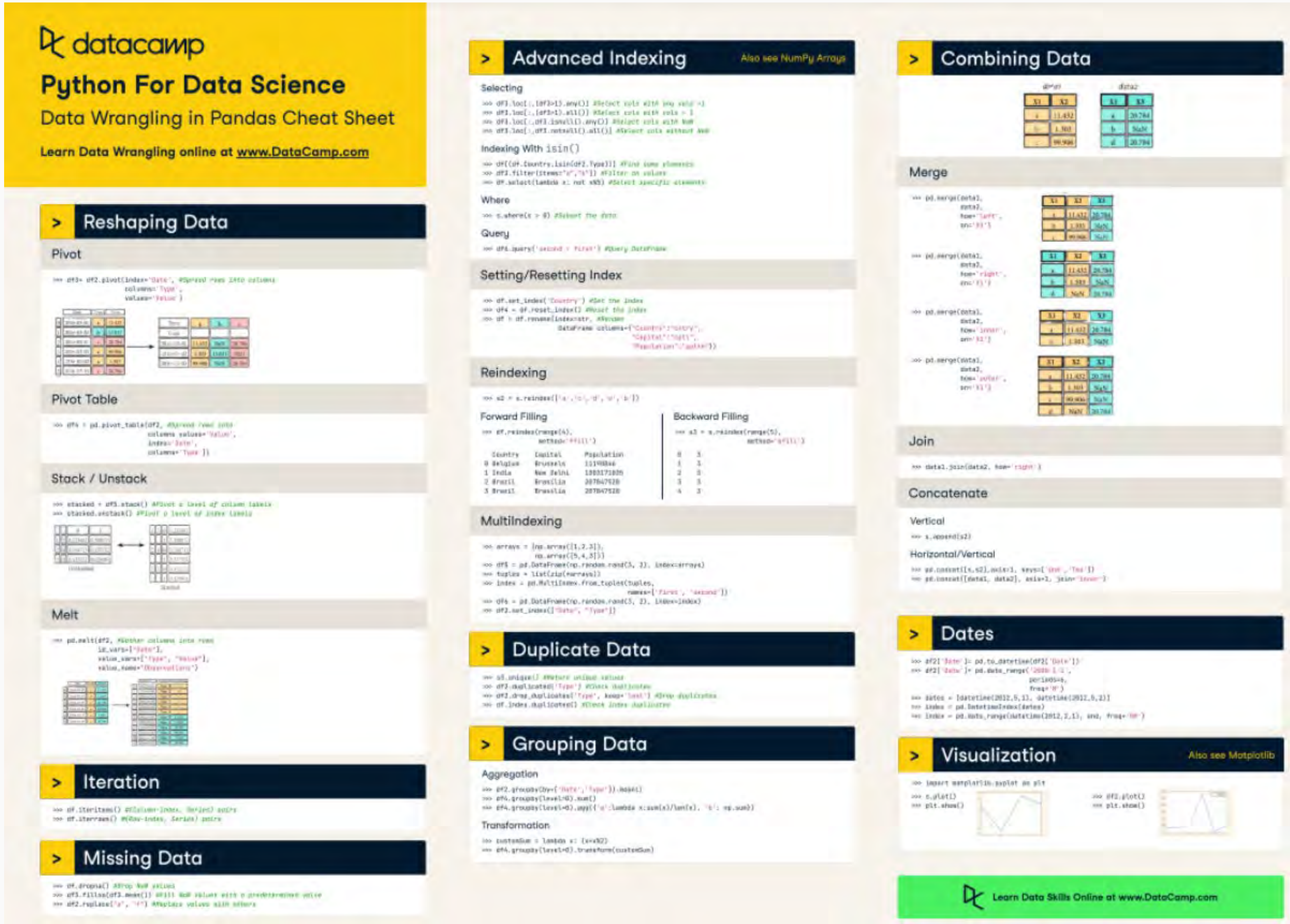
- Terra-specific support
- Terra-specific documentation

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     Brasília   207847528
 3 Brazil     Brasília   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