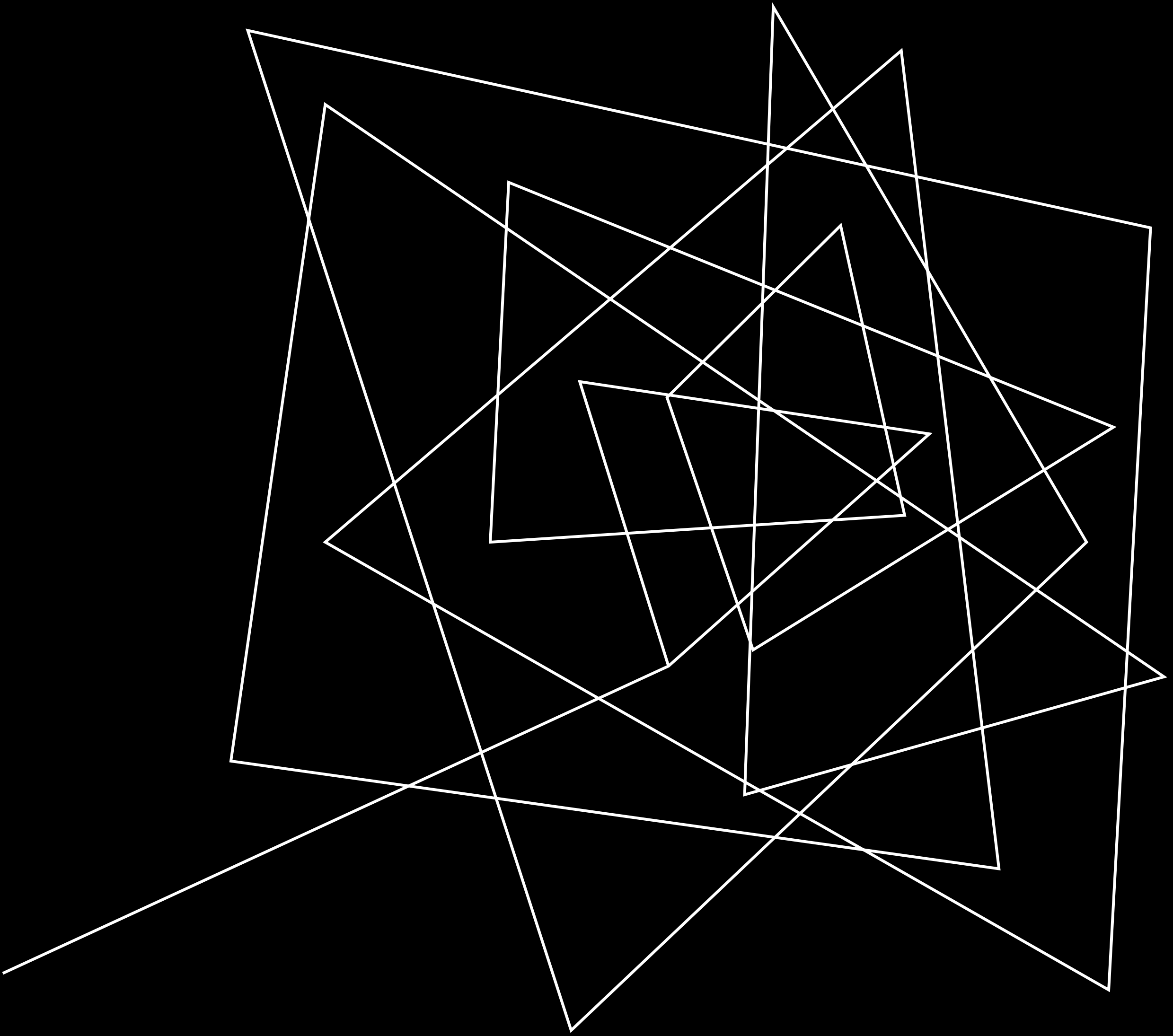
RESPONSIBLE AI

Week 3: Replication Project Part 0 - Introduction

Nandita Rahman & Emily Ramond

TODAY'S OBJECTIVES

- Reading Presentation #1
- Reading Presentation #2
- Introduce the Q1 (replication) project objectives, the AI Fairness 360 model overview, and the medical expenditure tutorial
- **Workshop**: Initiate logistics for the replication project
- **Survey**: What type of role do you see yourself in a data science team, and skills?



READING PRESENTATION #1

Datasheets for Datasets (Gebru et al.)

(OPTIONAL) DISCUSSION

- What are some ways that data can introduce bias into an AI/ML model? List as many as you can think of.

REPLICATION PROJECT: INTRO & OBJECTIVES

Overview

- Quarter One Project (70%), The Quarter 1 Project is due at the end of Week 10.
- Introduce students to the area in which they will do their project by replicating a known result.
- Students will complete coding tasks related to the replication project and are also responsible for creating a final writeup
- Create written material and code that serves as a foundation for work in quarter-2's projects.
- Full details of the requirements for the Q1 project can be found in the **Capstone Program Syllabus**

CHECKPOINT is **due on Monday, November 6th at 11:59PM (at the start of Week 6).** It Includes:

- The title, abstract, and introduction sections of your report.

WHOLE PROJECT is **The whole project is due on Monday, December 11th at 11:59PM (at the start of Week 11)**

Focus Area: AI and Fairness

Data Quality

- **Is data representative of the population?**
- Are certain groups under-represented or over-represented?
- **Does the data contain protected attributes?**
- Does it contain the attributes themselves? Does it contain attributes that can be used as proxies for protected attributes (e.g., ZIP codes)? Does it contain variables that vary with protected attributes (e.g., credit scores)?

Algorithmic Fairness

- **Is the algorithm disparately accurate for certain groups?**
- Can it predict outcomes for one group better than it can for another group?
- **Does the algorithm make different types of errors across certain groups?**
- Does it have differing false positive or false negative rates for certain groups?

BIAS DETECTION WITHIN AI-SYSTEMS.

Model bias is commonly thought to occur **primarily** in the data collection step.[1]

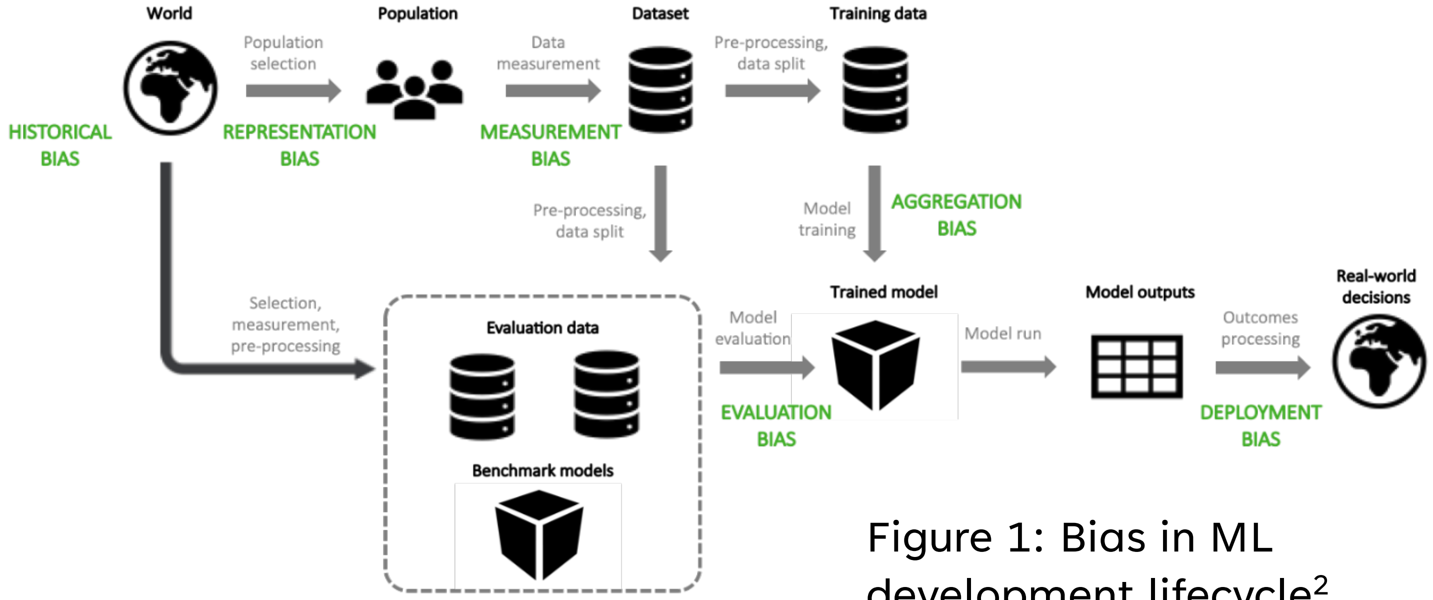


Figure 1: Bias in ML development lifecycle[2]

Recent research suggests that bias can be introduced at **any stage** in the machine learning model life cycle (see Figure above).[2]

Bias-Fairness detection and mitigation is a complex study that lies at the intersection of ethical and technical disciplines.[3]

A comprehensive bias and fairness model evaluation cannot solely attribute bias to the dataset and recommend improving data collection.[1]

Model outputs must be used to provide thoughtful mitigation steps via policy and industry knowledge, in addition to applying techniques that reshape the input data, model decision making, and model outputs.

AI fairness and transparency are at the heart of this, with the goal of empowering data-science practitioners and affected groups to detect bias in AI-systems using model outputs.

Clinical Decision Support | Examples of Bias in Healthcare

A growing number of healthcare experts are concerned about potential harmful racial bias in clinical decision support (CDS) tools and how it affects health equity.

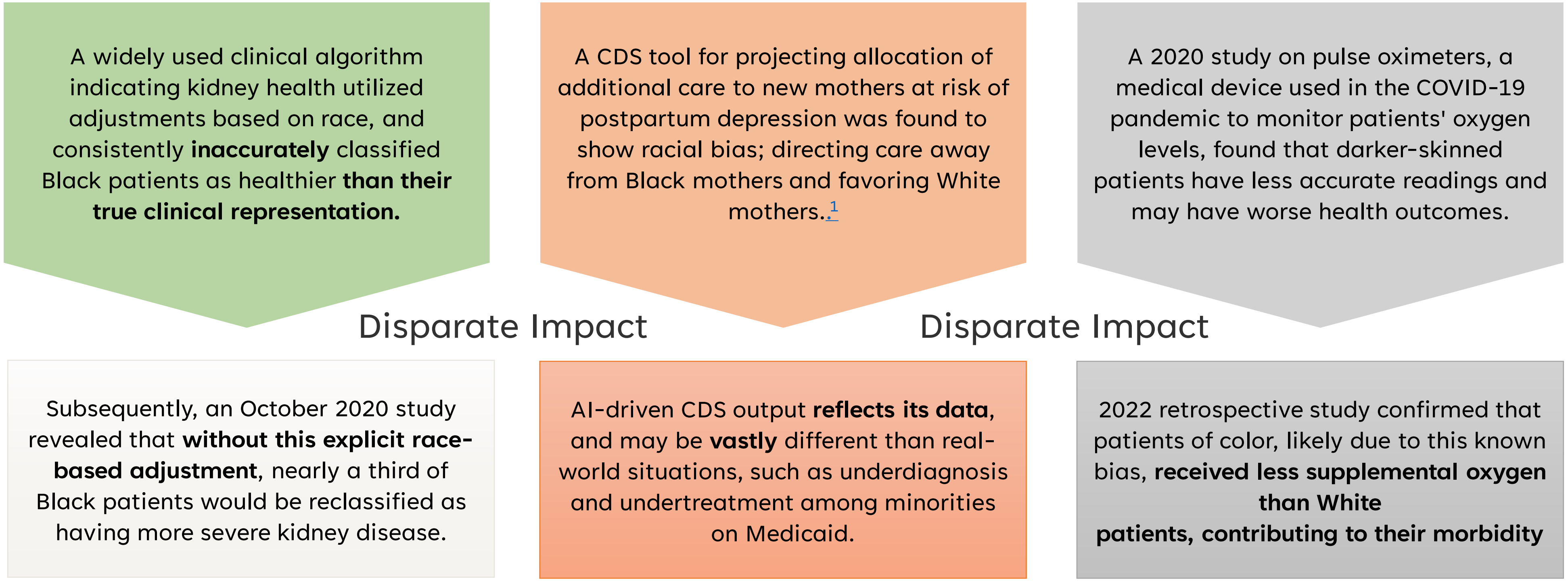
An ACLU legal report outlines how widespread use of **"racial correction" modifiers** in clinical guidelines and algorithms to account for **biological differences** between races and ethnicities.

ACLU

NEWS & COMMENTARY

Algorithms Are Making Decisions About Health Care, Which May Only Worsen Medical Racism

Unclear regulation and a lack of transparency increase the risk that AI and algorithmic tools that exacerbate racial biases will be used in medical settings.



A widely used clinical algorithm indicating kidney health utilized adjustments based on race, and consistently **inaccurately** classified Black patients as healthier **than their true clinical representation.**

A CDS tool for projecting allocation of additional care to new mothers at risk of postpartum depression was found to show racial bias; directing care away from Black mothers and favoring White mothers.[1]

A 2020 study on pulse oximeters, a medical device used in the COVID-19 pandemic to monitor patients' oxygen levels, found that darker-skinned patients have less accurate readings and may have worse health outcomes.

Disparate Impact

Disparate Impact

Subsequently, an October 2020 study revealed that **without this explicit race-based adjustment**, nearly a third of Black patients would be reclassified as having more severe kidney disease.

AI-driven CDS output **reflects its data**, and may be **vastly** different than real-world situations, such as underdiagnosis and undertreatment among minorities on Medicaid.

2022 retrospective study confirmed that patients of color, likely due to this known bias, **received less supplemental oxygen than White patients, contributing to their morbidity**

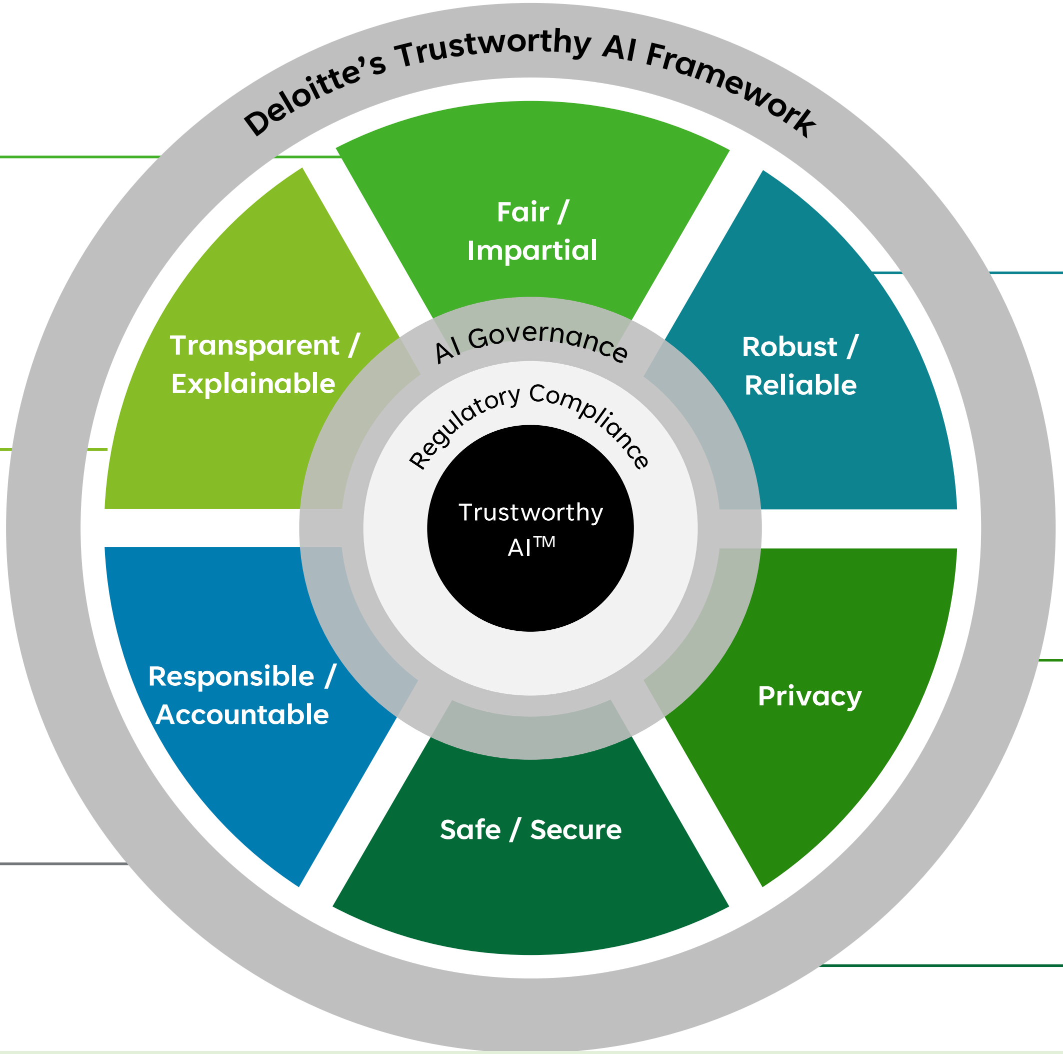
RESPONSIBLE AI PILLARS

Deloitte's **Trustworthy AI™** framework[4] is an effective first step in having an approach to categorizing AI risks and integrated into modeling approaches.

Fair / Impartial
AI applications include automated and manual checks to help enable equitable application across all participants

Transparent / Explainable
Participants can understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations are open to inspection

Responsible / Accountable
Policies are in place to determine who is held responsible for the output of AI system decisions



Robust / Reliable
AI systems should build to learn from humans and other systems and produce consistent and reliable outputs

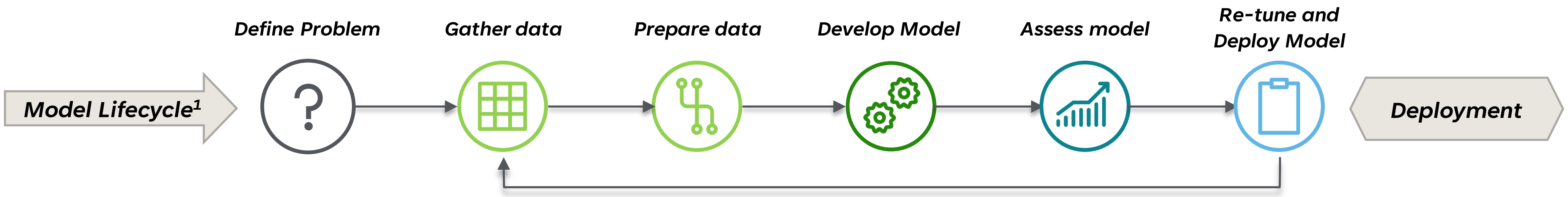
Privacy
Consumer privacy is respected, and customer data is not used beyond its intended and stated use; consumers are able to opt in/ out of sharing their data

Safe / Secure
AI systems can be protected from risks (including cyber risks) that may cause physical and/or digital harm

TAI principles are not mutually exclusive, and tradeoffs often exist when applying them.[5]

REPLICATION PROJECT: RESPONSIBLE AI IN ACTION

OPERATIONALIZING RESPONSIBLE AI ISN'T JUST USING CODE.
IT ALSO REQUIRES HUMANS IN THE LOOP FOR EACH STAGE OF THE MODEL LIFECYCLE



Model Lifecycle[1]

Define Problem → *Gather data* → *Prepare data* → *Develop Model* → *Assess model* → *Re-tune and Deploy Model* → *Gather data*

Deployment

Start: Week 03 (10/20/23)

Part 0: Overview of Data, Scope, and Preparation

- Dataset
- Scope Definition

Week 04 (10/27/23)

Part 1: Exploratory Data Analysis

- Review of protected classes
- Analysis insights and explanation
- Data imbalances

Week 06* (11/7/23)

Part 2: Model Development and Fairness Evaluation (Bias Detection)

- Model Development
- Bias detection metrics
- Fairness metrics, assessment and visualizations

***RESCHEDULED TO HOLD DURING OFFICE HOURS (VETERANS DAY)**

Week 07 (11/17)

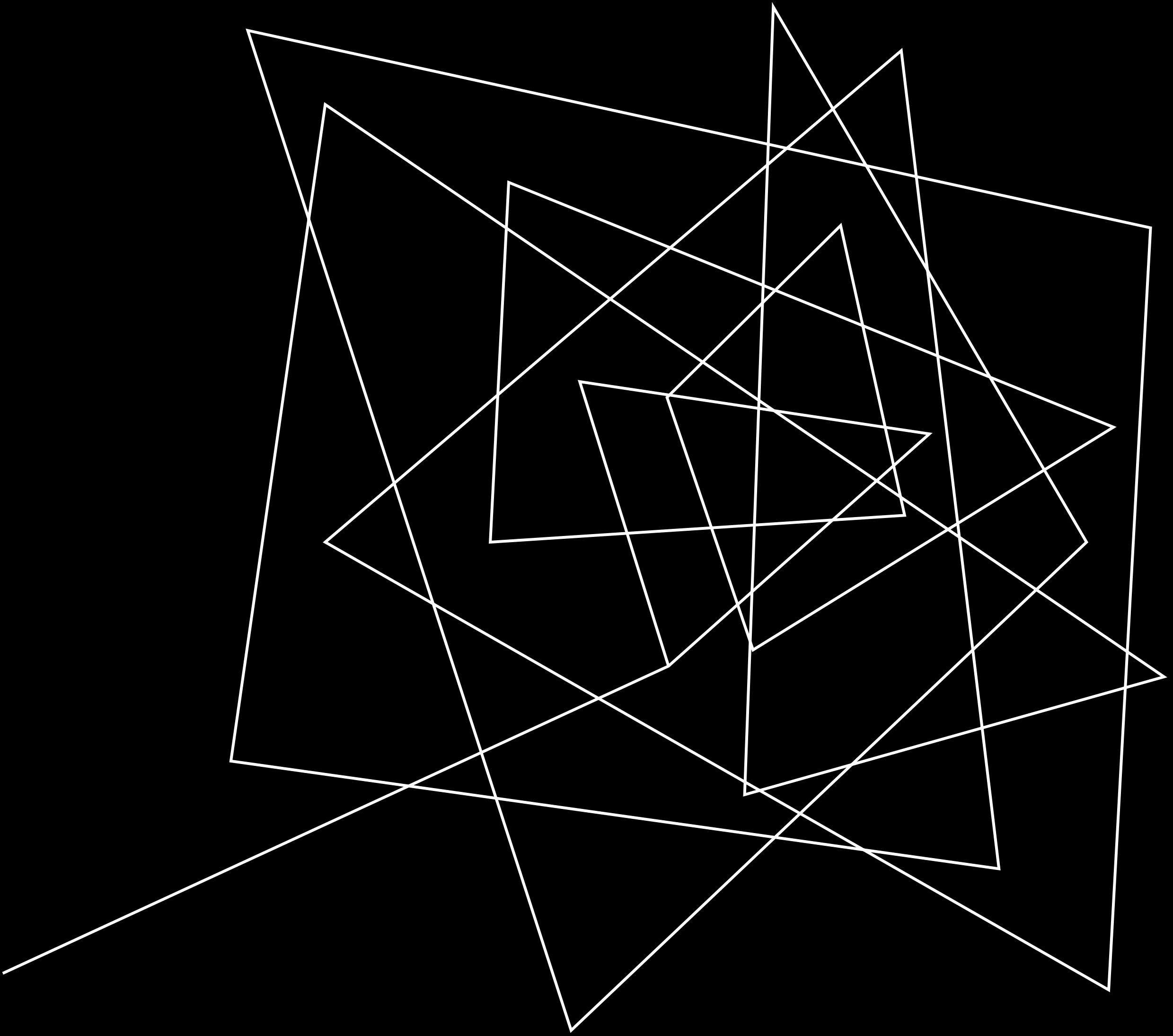
Part 3: Bias Mitigation Techniques

- Pre, in, post-processing bias mitigation methods
- Discussion

End: Week 09 (Slides Only)

Part 4: Presentation, Conclusions and Explainability

- Local, Global Explainability Results
- Explainability integration into Model Lifecycles
- Future perspectives on explainability



READING PRESENTATION #1

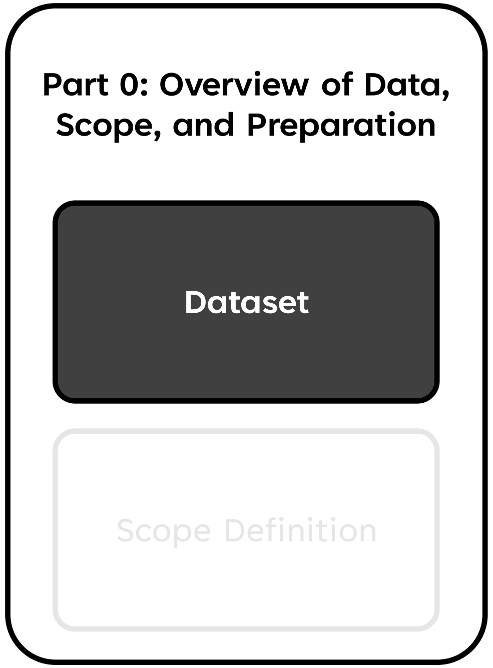
AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias (Bellamy et al.)[4]

(OPTIONAL) DISCUSSION

- Walk through the AIF360 demo(s) and explore the functionality. When comparing to your list of how data can introduce bias in an AI/ML models, did the tool help understand how data can introduce bias?
- After going through the AIF360 demo, what are some missing features that you think are important? What is still unclear?

REPLICATION PROJECT: DATASET

DATASET: MEDICAL EXPENDITURE PANEL SURVEY (MEPS)[4]



Part 0: Overview of Data, Scope, and Preparation

Dataset

Scope Definition

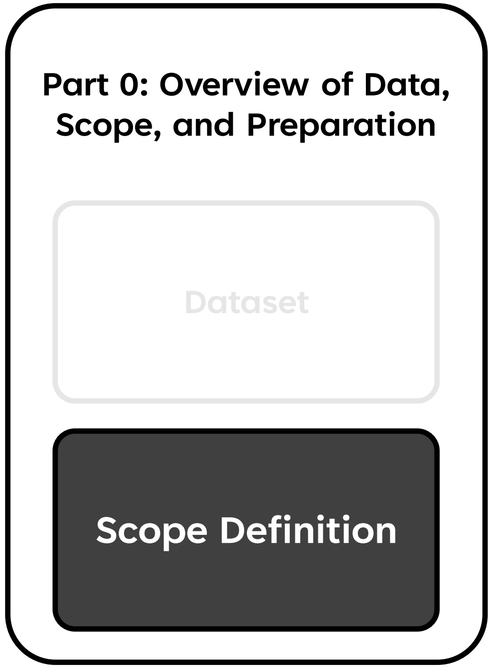
- The **Medical Expenditure Panel Survey** (MEPS) provides nationally representative estimates of health expenditure, utilization, payment sources, health status, and health insurance coverage among the noninstitutionalized U.S. population.
- These government-produced data sets examine how people use the US healthcare system.
- MEPS is administered by the Agency for Healthcare Research and Quality (AHRQ) and is divided into three **components:**
  - (1) Household, (2) Insurance/Employer, and (3) Medical Provider.
  - Only the **Household Component (HC)** is available for download on the Internet.

These components provide comprehensive national estimates of health care use and payment by individuals, families, and any other demographic group of interest.[1]

The specific data used is the 2015 Full Year Consolidated Data File[6] as well as the 2016 Full Year Consolidated Data File[7]

REPLICATION PROJECT: SCOPE

SCOPE: MEDICAL EXPENDITURE PANEL SURVEY (MEPS) AND UTILIZATION, CARE MANAGEMENT[4]



Part 0: Overview of Data, Scope, and Preparation

Dataset

Scope Definition

- We will be *adapting* IBM AI Fairness 360's Medical expenditure tutorial, which is a comprehensive tutorial demonstrating the interactive exploratory nature of a data scientist detecting and mitigating racial bias in a **medical care management** scenario.

- **Specifically, it walks through the scenario of a data scientist** who is tasked with developing a 'fair' healthcare **utilization** scoring model with respect to defined protected classes.

> In this context, the model classification task is to predict whether a person would have **'high'** utilization (defined as UTILIZATION >= 10, roughly the average utilization for the considered population). A high utilizer of medical care could signal a need for additional care, any risk factors or comorbidity trends.

- As shown in previous lectures, evaluating fairness for labels such as utilization, may be driven by legal or government regulations. An example could be new a requirement that additional care decisions **are not** predicated on factors such as race of the patient.

- It also demonstrates how **explanations can be generated** for predictions made by models learned with the toolkit using LIME.

REPLICATION PROJECT: MODEL USE CASE

MEDICAL EXPENDITURE PANEL SURVEY (MEPS)[4]

Initial deployment is simulated, demonstrating how classification scores (utilization) would be used to identify potential candidates for additional care management. We assume that the model is initially built and tuned using the 2015 Panel data.

- For each dataset, the **sensitive attribute** is 'RACE' constructed as follows: 'Whites' (privileged class) defined by the features RACEV2X = 1 (White) and HISPANX = 2 (non-Hispanic); 'Non-Whites' that included everyone else.

- Along with race as the sensitive feature, other features used for modeling include **demographics** (such as **age**, **gender**, **active-duty status**), **physical/mental** health assessments, **diagnosis codes** (such as history of diagnosis of cancer, or diabetes), and **limitations** (such as cognitive or hearing or vision limitation).

- The model classification task is to predict whether a person would have **'high**' utilization (defined as UTILIZATION >= 10, roughly the average utilization for the considered population). We will also investigate how to evaluate **which type of machine learning model** (e.g., Linear Regression versus Random Forest, etc.) would be best used, and who the relevant stakeholders are in this scenario.

REPLICATION PROJECT LOGISTICS

- **Data**: Medical Expenditure Panel Survey
- **Setup**: Refer to QuickStart Guide
- **Code**: AI Fairness 360 (AIF360) toolkit, EDA Python Notebook (*Available Next Week*)
- **Teams**: Fill out survey; Instructors will use results to build two equally sized teams.

Replication Project Contact: Emily Ramond (eramond@deloitte.com) and cc Nandita Rahman (nanrahman@deloitte.com)

Replication Project Office hours: Use regular office hours

FOR NEXT WEEK

- Complete next week's **readings**
  - If you signed up to present **Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? (Holstein et al.)** come prepared to present next week.
  - Submit your presentation to Gradescope by 10 AM PT, Friday, October 27th

- **Participation Questions/Points: See Write up #3**
  - **Replication Pre-Work (Part #0):** Using the QuickStart Guide to set up your workspace and launch the example AIF360 notebook listed in the QuickStart Guide.
    - Submit screenshot of completed notebook
  - Answer listed questions (at least 500 words total)
  - Fill out the Data Science Team Persona survey by 10 AM PT, Thursday, October 26th

References

1. Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ACM, Honolulu HI USA, 1–14. DOI:https://doi.org/10.1145/3313831.3376445

2. Michelle Seng Ah Lee and Jatinder Singh. 2021. Risk Identification Questionnaire for Unintended Bias in Machine Learning Development Lifecycle. *SSRN Journal* (2021). DOI:https://doi.org/10.2139/ssrn.3777093

3. Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]* (November 2016). Retrieved November 4, 2021 from http://arxiv.org/abs/1609.05807

4. https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html

5. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv*. https://doi.org/10.48550/arXiv.1810.01943

6. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181

7. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192