RESPONSIBLE AI

Week 1: Introduction

TODAYS OBJECTIVES

- Quick lab (hey otter.ai!)
- Reading Presentation 1 & 2
- Lecture content
- Introduce Q1 and team forming
- AI Fairness 360 model overview
- Workshop (if we have time)
    - Implement logistics for replication project. What are your strengths and weaknesses?
        - Reminder to check out your dsc80 projects!!! Brush up on coding!

RESPECTFUL OF PRIVACY

FAIR AND IMPARTIAL, SAFE AND SECURE

TRANSPARENT AND EXPLAINABLE

RESPONSIBLE AND ACCOUNTABLE

ROBOUST AND RELIABLE

REVISTING LAST WEEK

Deloitte’s AI Framework

REVISTING LAST WEEK

Deloitte’s AI Framework

RESPECTFUL OF PRIVACY

FAIR AND IMPARTIAL, SAFE AND SECURE

TRANSPARENT AND EXPLAINABLE

RESPONSIBLE AND ACCOUNTABLE

ROBOUST AND RELIABLE

Otter.ai

REVISTING LAST WEEK

Deloitte's AI Framework – Otter.ai

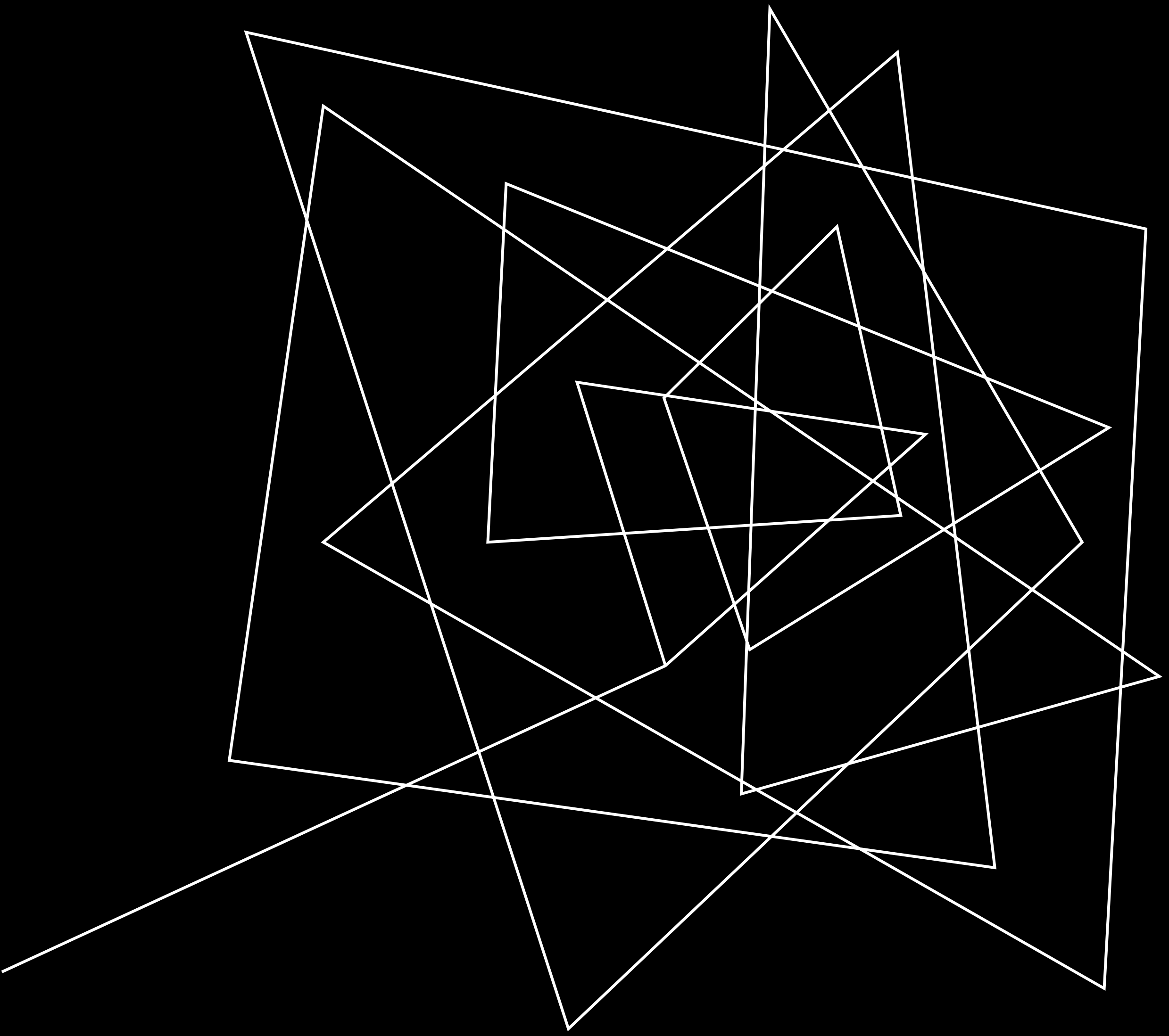
RESPECTFUL OF PRIVACY — Transcribing audio is private. There must be clear consent mechanisms for all parties. Users (who are the users…?) should have control over how their data is stored, shared, or deleted.

FAIR AND IMPARTIAL, SAFE AND SECURE — Are there biases in speech recognition? How is that measured? How does this affect outcomes for individuals who are not being recognized?

TRANSPARENT AND EXPLAINABLE — How does the algorithm work? What at the accuracy rates across different groups? What CAN'T this product do?

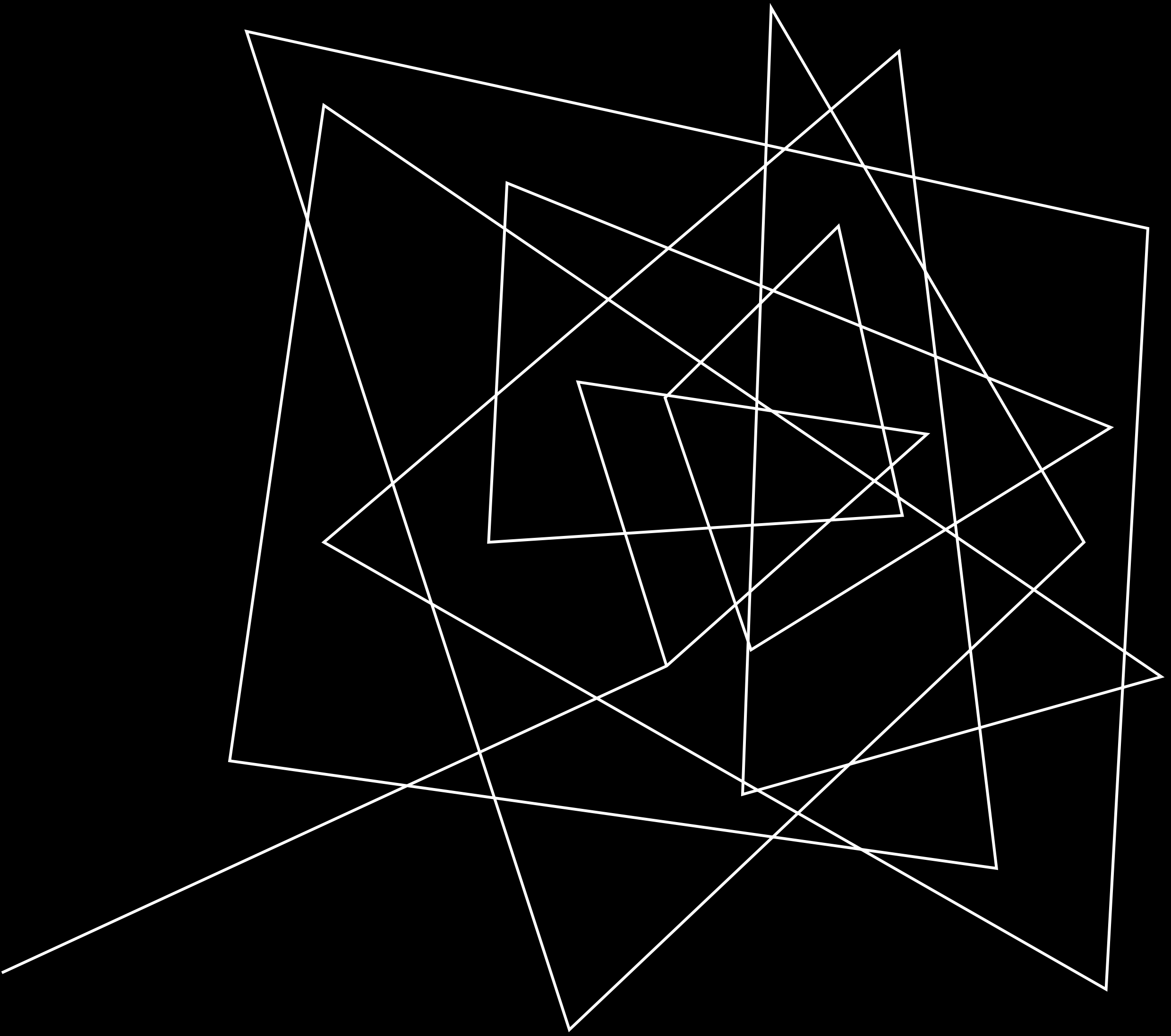
RESPONSIBLE AND ACCOUNTABLE — Take accountability for inaccuracies and the ability to report bias. How are they managing ethical concerns?

ROBOUST AND RELIABLE — Noisy backgrounds, different languages, etc?



READING PRESENTATION #1

AJ FALAK



READING PRESENTATION #1

ETHAN LIN

QUESTIONS / DISCUSSION?

WHAT ARE ALL THE WAYS THAT DATA CAN INTRODUCE BIAS INTO AN AI/ML MODEL? LIST AS MANY AS YOU CAN THINK OF.

REPLICATION PROJECT (Q1 PROJECT) OVERVIEW

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**

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?

REPLICATION PROJECT: INTRO AND OBJECTIVES

Overview

- Q1 Project (65%), due end of week 10.
- Introduce students to the area in which they will do their project by replicating ga 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 Q2 projects.
- Full details of the requirements for Q1 project can be found in the Capstone Program Syllabus

CHECKPOINT is **due Monday, November 4th at 11:59pm**. It includes:

- Title, abstract, and introduction sections of your report.

WHOLE PROJECT is **due Monday, December 2nd at 11:59pm.**

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? (zip codes)? Does it contain variables that vary with protected attributes? (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]

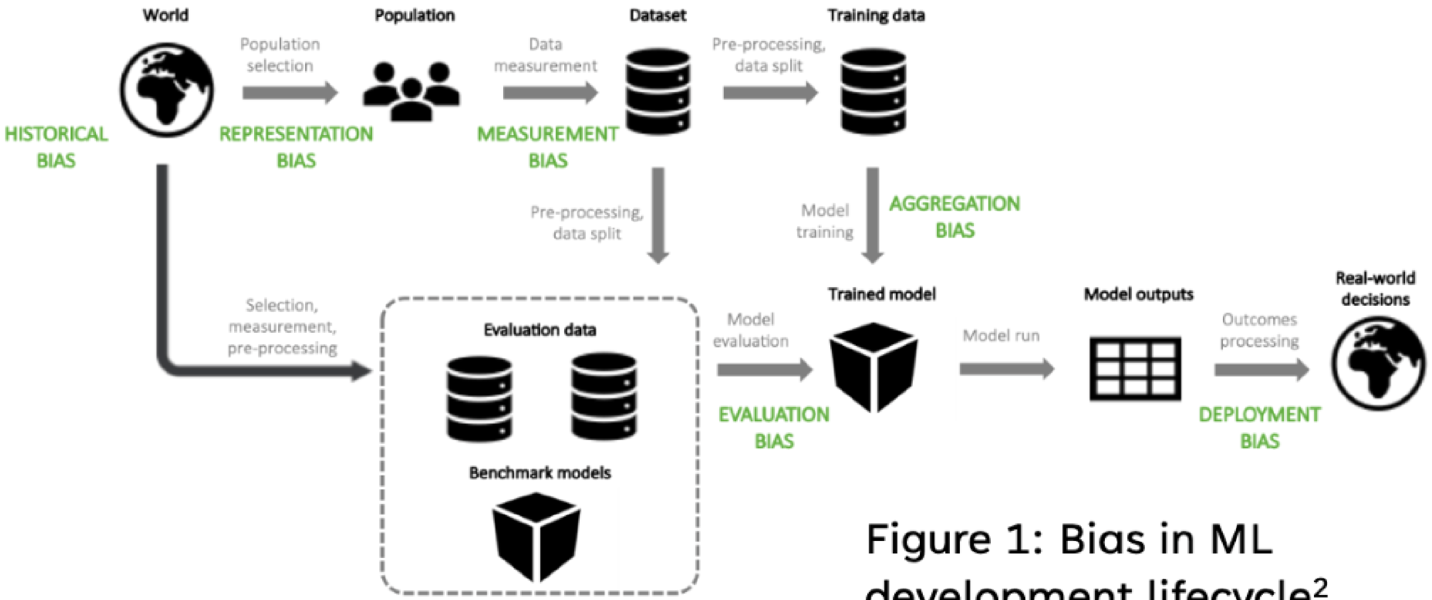


Figure 1: Bias in ML development lifecycle[2]

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.

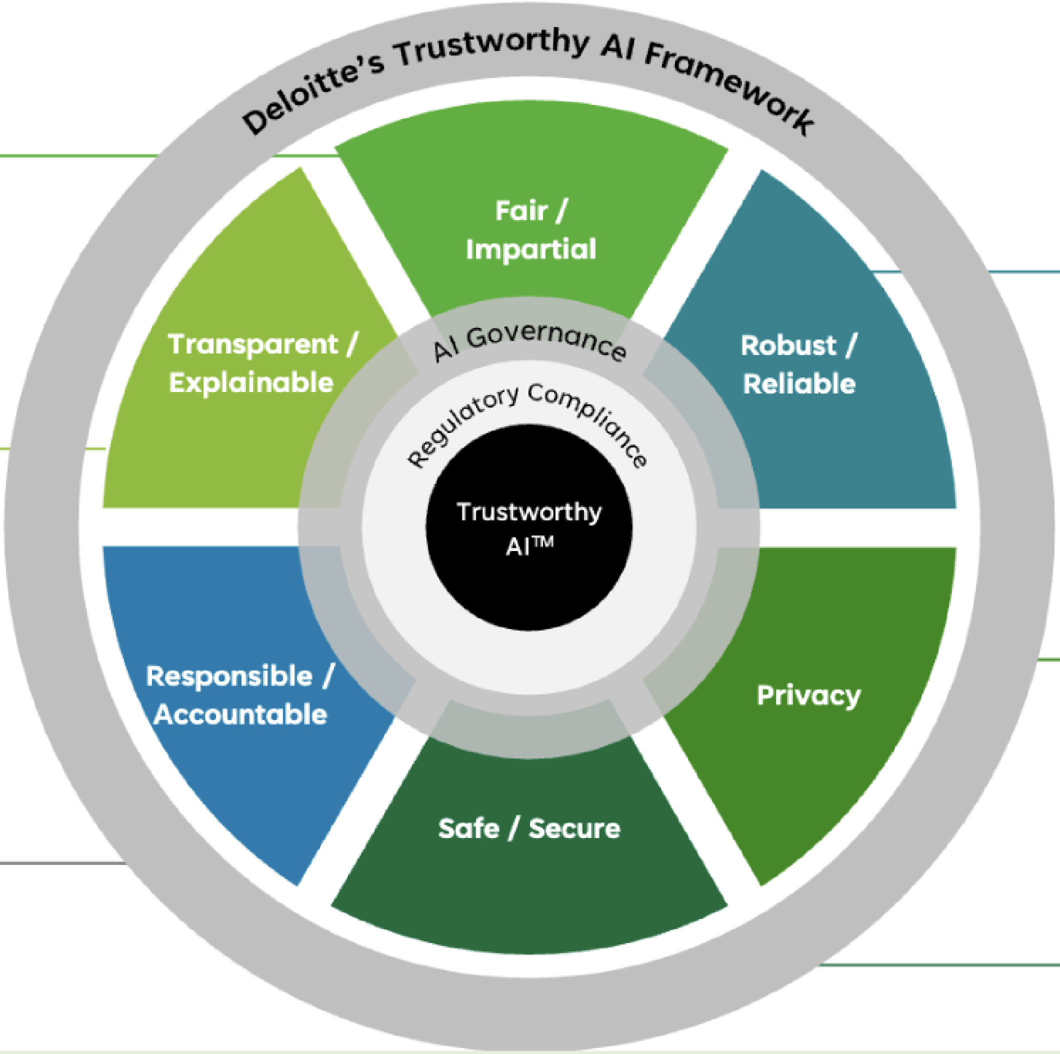
RESPONSIBLE AI PILLARS

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

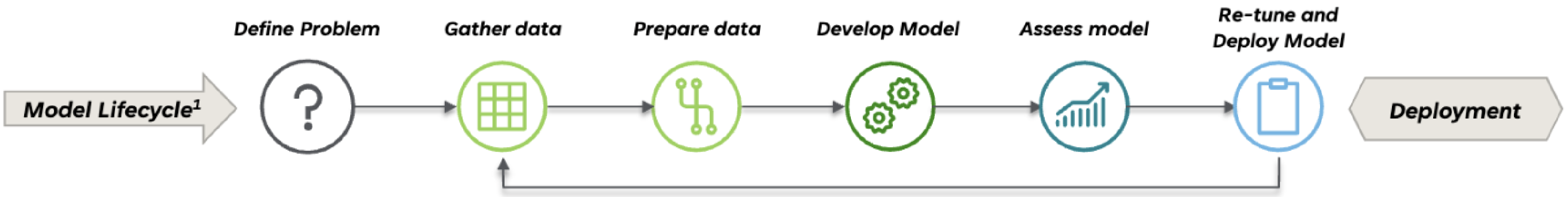
Privacy
Consumer privacy is respected, and customer data is not used beyond its intended and stated use; consumers are able to opt in/ out of sharing their data

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



Start: Week 03 (10/20/23)	Week 04 (10/27/23)	Week 06* (11/7/23)	Week 07 (11/17)	**End**: Week 09 (Slides Only)
Part 0: Overview of Data, Scope, and Preparation	**Part 1: Exploratory Data Analysis**	**Part 2: Model Development and Fairness Evaluation (Bias Detection)**	**Part 3: Bias Mitigation Techniques**	**Part 4: Presentation, Conclusions and Explainability**
Dataset	Review of protected classes	Model Development	Pre, in, post-processing bias mitigation methods	Local, Global Explainability Results
Scope Definition	Analysis insights and explanation	Bias detection metrics	Discussion	Explainability integration into Model Lifecycles
	Data imbalances	Fairness metrics, assessment and visualizations		Future perspectives on explainability

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

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.

TEAM CREATION

FOR NEXT WEEK

- Complete next weeks **reading**
  - If you sign up to present next week's reading, come prepared to present by 2:00PM PT on Tuesday, October 15th
  - Presentations are 5 minutes – short and simple. You are not graded on the slides and do not need to send them to me.
- Submit your answers to next week’s default **participation questions** to Gradescope by 2PM PT on Monday, October 14th
- Review class website and email any questions to Emily Ramond (eramond@ucsd.edu)
- **Office hours Friday 12:30-1pm and Mondays 4-4:30pm.**
- Explore replication project and review syllabus. Pull data and do a preliminary EDA. Find two times a week you and your team can meet.
  - See writeup 3 for Quickstart guide. Send a photo of it completed by next class!!!

Please see the UCSD Capstone Syllabus and the Capstone Program Website for a detailed description of the assignment weights and rubric.