Describe bias and fairness issues in machine learning

Module Introduction

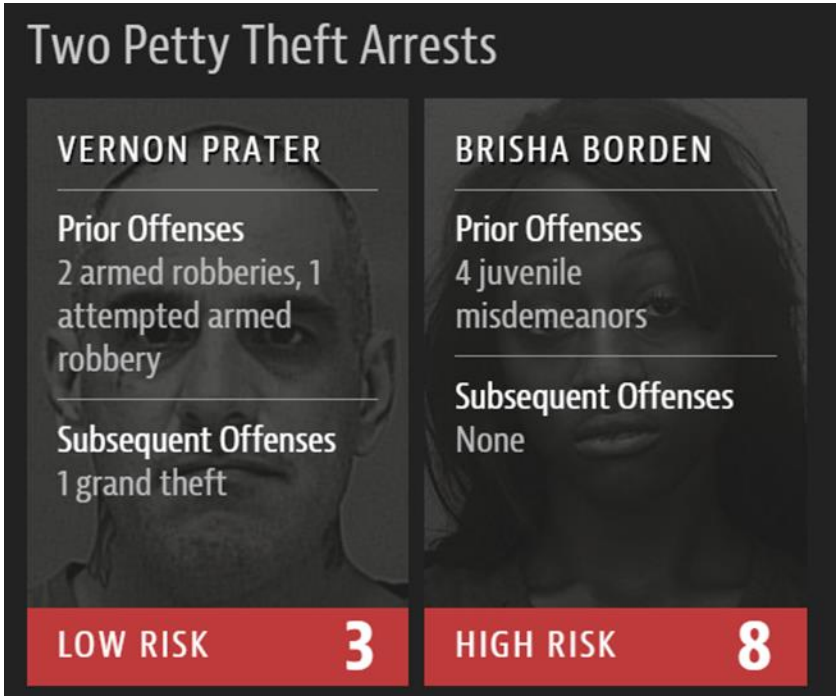
Objectives

By the end of this module, you will be prepared to:

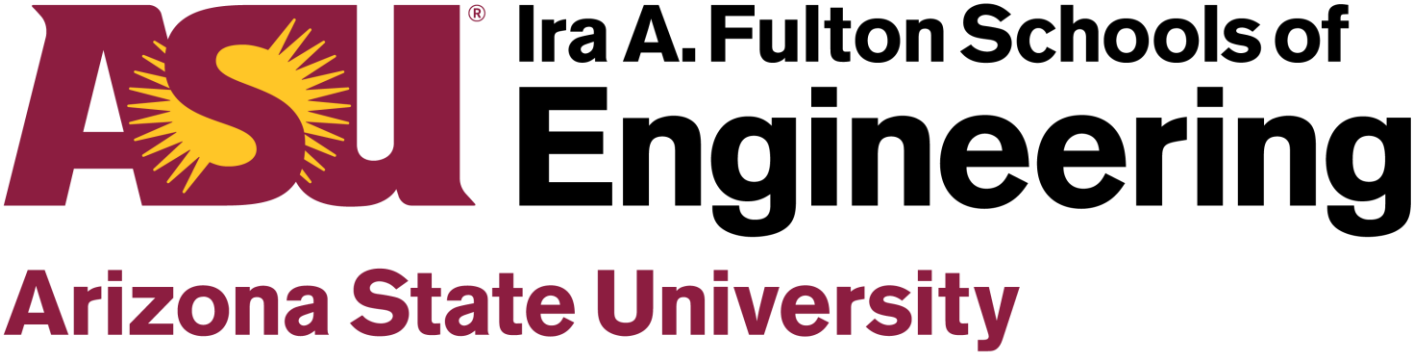
- Define “data to algorithm” bias and list several examples
- Define “algorithm to user” bias and list several examples
- Define “user to data” bias and list several examples
- Describe explainable and unexplainable discrimination in machine learning
- Define individual, group, and subgroup fairness
- List methods for fair machine learning

Unfairness in the COMPAS Risk Assessment Model

- Risk assessment model from a firm called Northpointe was designed to predict likelihood of subsequent offenses producing a “risk score”
- Software widely used by judges as a tool to aide in sentencing
- System was shown to disproportionately provide greater risk scores to African Americans
- System did not use race as a feature, but rather a combination of questions and arrest records



Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing


Ira A. Fulton Schools of
Engineering
Arizona State University

Basic Concepts in Bias and Fariness

Objectives

By the end of this module, you will be prepared to:

- Define “data to algorithm” bias and list several examples
- Define “algorithm to user” bias and list several examples
- Define “user to data” bias and list several examples
- Describe explainable and unexplainable discrimination in machine learning
- Define individual, group, and subgroup fairness
- List methods for fair machine learning

Algorithmic Fairness

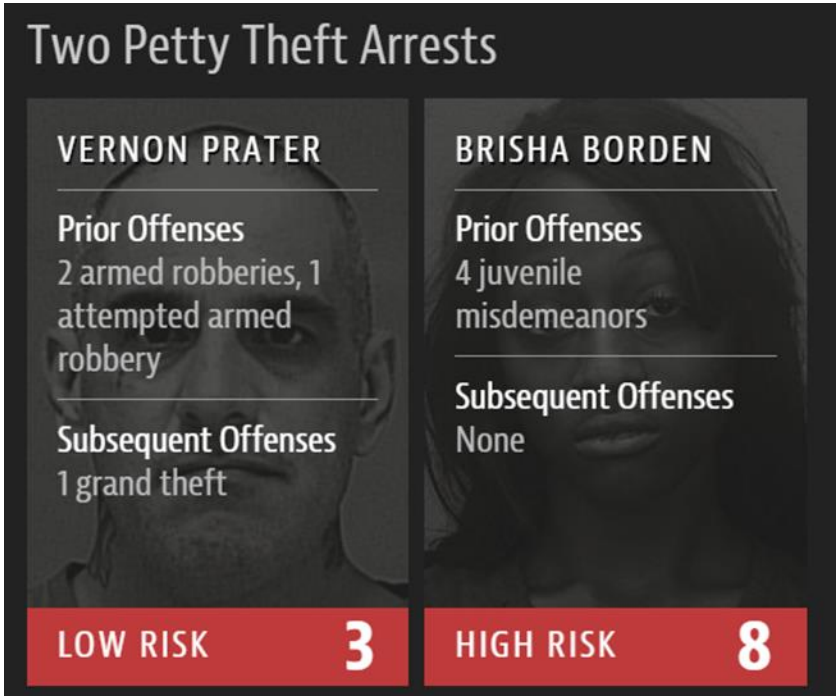
- ***Fairness*** is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics

- An ***unfair algorithm*** is one whose decisions are skewed toward a particular group of people

Definitions from Mehrabi et al., 2022 https://arxiv.org/pdf/1908.09635.pdf

Unfairness in the COMPAS Risk Assessment Model

- Risk assessment model from a firm called Northpointe was designed to predict likelihood of subsequent offenses producing a “risk score”
- Software widely used by judges as a tool to aide in sentencing
- System was shown to disproportionately provide greater risk scores to African Americans
- System did not use race as a feature, but rather a combination of questions and arrest records



Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS

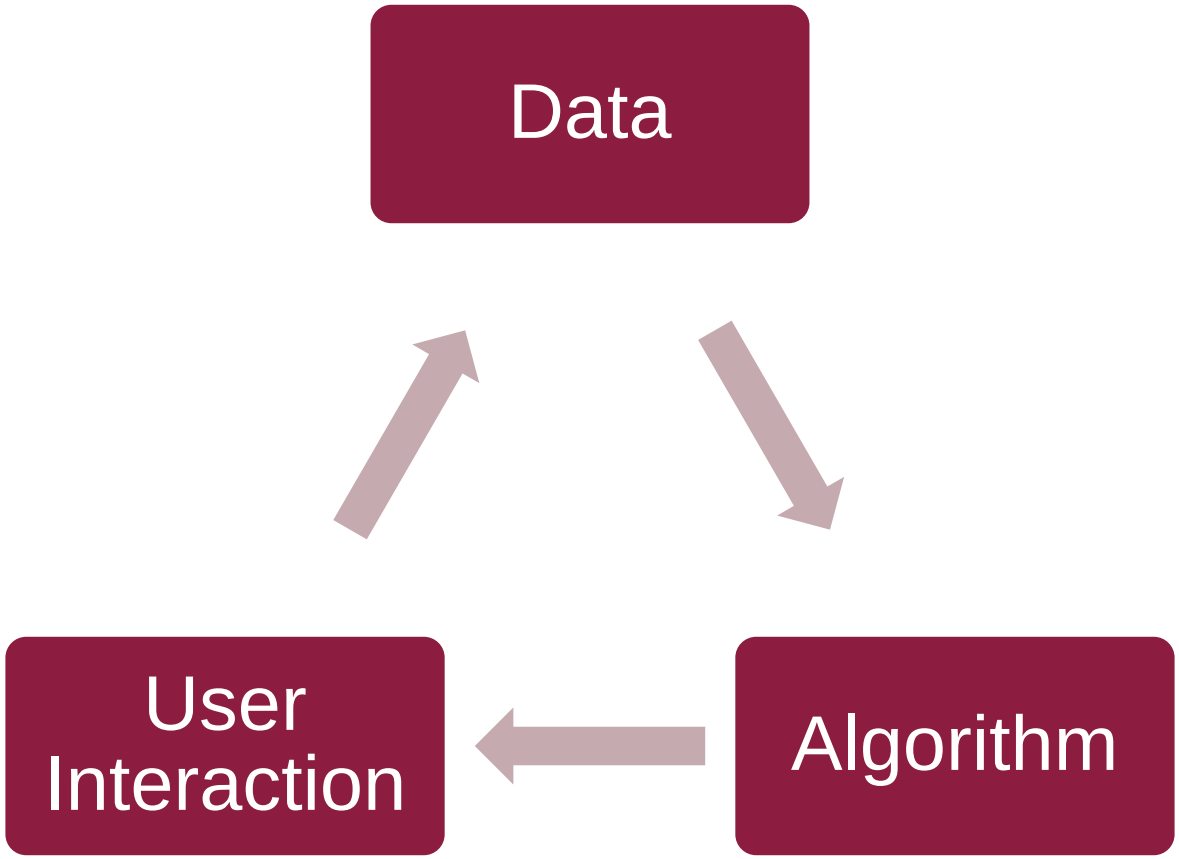
- Analysis of the model performance to identify bias also highlighted several other shortcomings
- The system used 137 features, but only 7 were shown to the user
- Subsequent studies have found the product to do no better than logistic regression and provided no improvement over human judgement

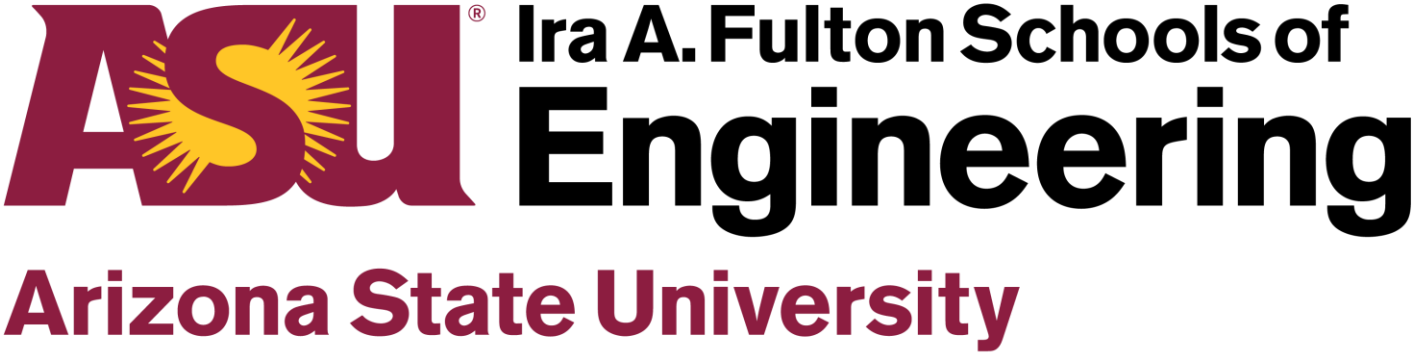
Blink Detection

- Nikon camera utilized a blink detection feature when taking portrait photographs
- Blink detection reportedly gave false positives for pictures of persons of Asian heritage
- Many suspect that this was due to the model being only trained on Caucasian people

Source: http://content.time.com/time/business/article/0,8599,1954643,00.html

The interrelationship between data, algorithms, and user interaction




Ira A. Fulton Schools of
Engineering
Arizona State University

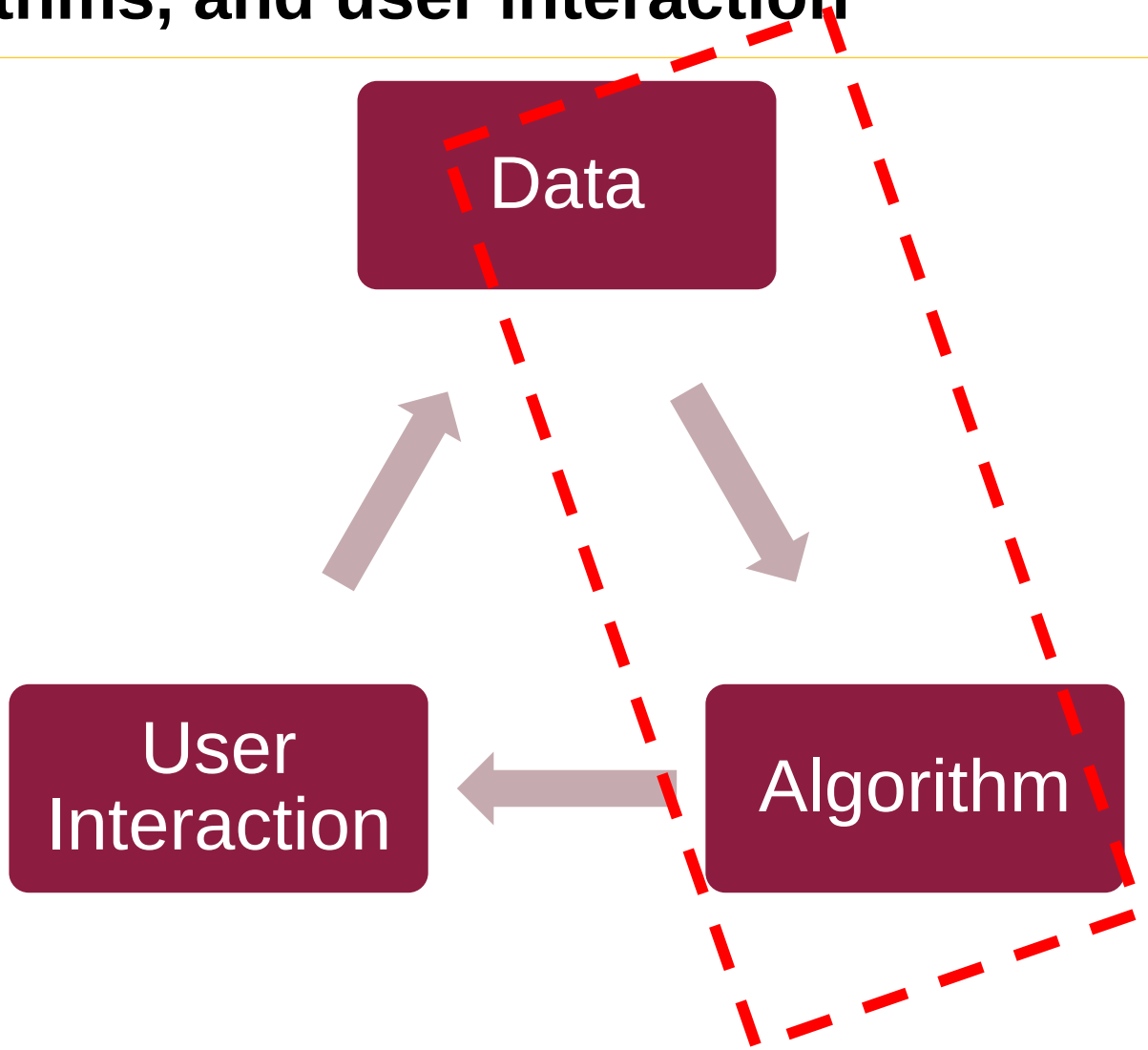
Data-to-Algorithm Bias

Objectives

By the end of this module, you will be prepared to:

- Define “data to algorithm” bias and list several examples

The interrelationship between data, algorithms, and user interaction



Data to Algorithm Bias

Data to algorithm bias is when data used to train a machine learning system leads to a biased model.

Measurement Bias

- Arises from the selection, use, and measurement of features
- COMPAS examined arrests of friends and family members, which was disproportionately high for minority communities that are policed more heavily

Omitted Variable Bias

- Omitted variable bias occurs when an important variable is omitted from a model

- A model with high precision but limited recall could potentially have omitted variable bias – as it cannot classify certain outcomes

Representation Bias

- Bias that originated from how a population is sampled
- Results from a lack of diversity in the sampling strategy – underrepresenting or omitting certain groups
- A classic example is in the ImageNet dataset (shown on the slide)

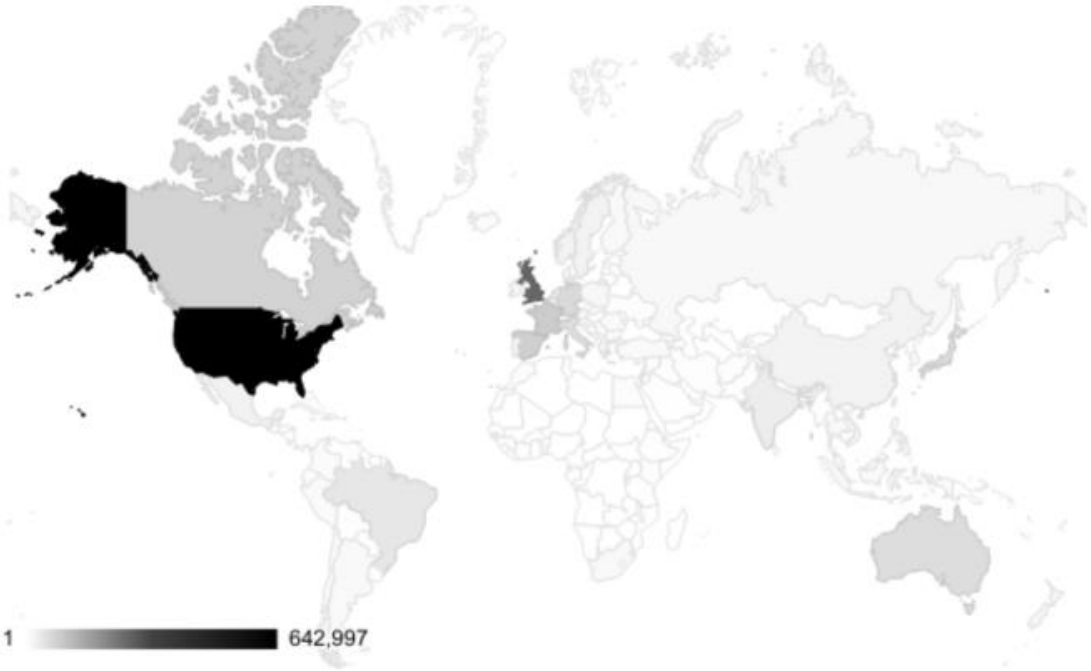


Image Source: Shankar et al., © 2017

Aggregation Bias

- False conclusions of an individual derived from a population-level observation
- Two examples:
  - ***Modifiable Areal Unit Problem*** occurs when modeling at different spatial scales yields different results
  - ***Simpson's Paradox*** (next slide)
- Example: Diabetes diagnosis using HbA1c levels has a wide variance across many genders and ethnic groups – will not be an accurate predictor if group characteristics are not considered by a model

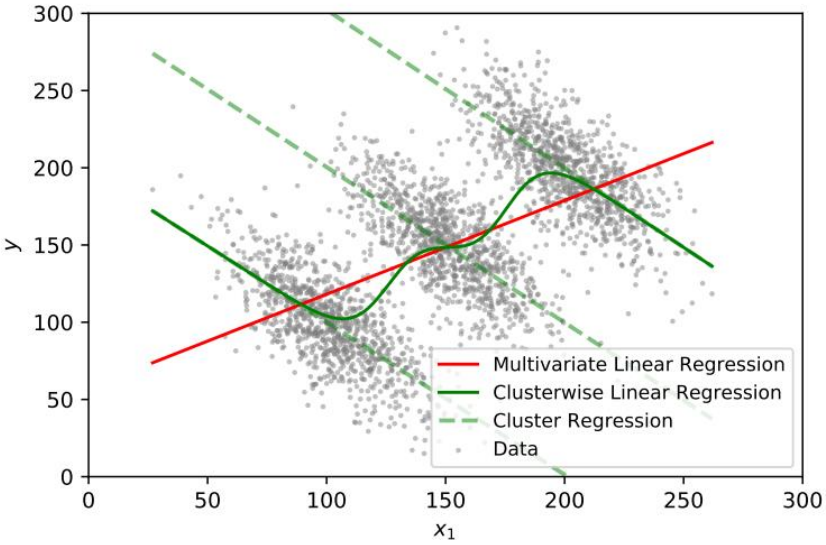
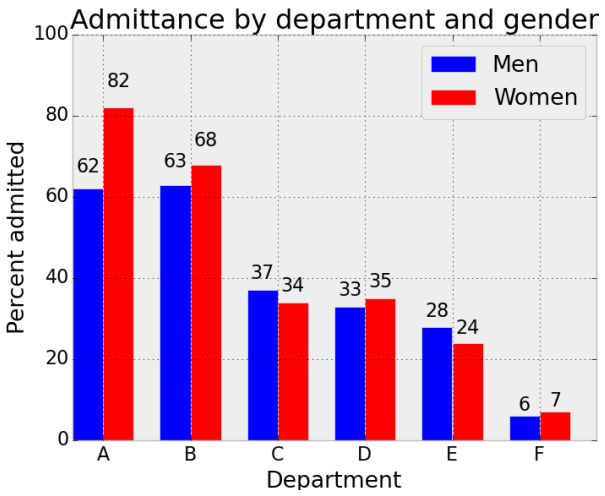
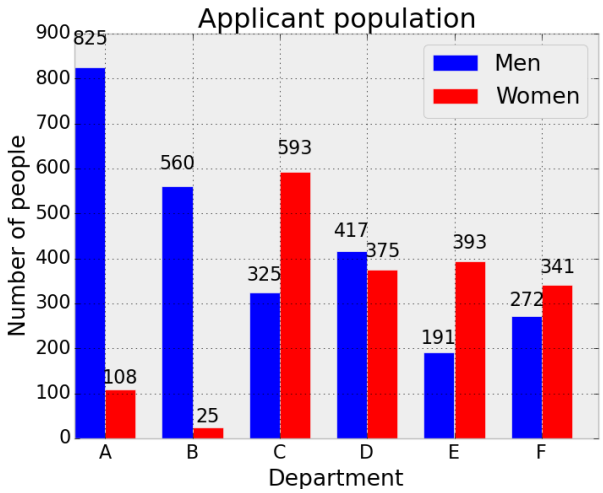


Image Source: Alipourfard et al., *WSDM,* 2018

Aggregation Bias: Simpson's Paradox

- An observed trend in the aggregate disappears or reverses for subgroups
- Arises form heterogenous groups represented in the data
- Famous example: 1973 UC Berkeley was sued for gender discrimination
  - Male applications were more likely to be accepted than female (44% vs. 35%)
  - However, when viewed by department, women were accepted at a higher rate
  - Women tended to apply to more selective departments

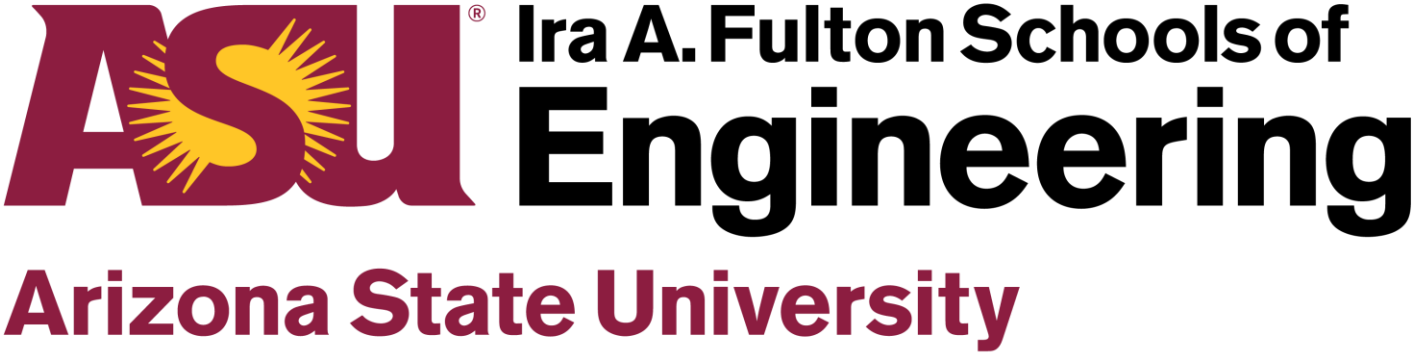




Source: http://corysimon.github.io/articles/simpsons-paradox/

Other Important forms of Data-to-Algorithm Bias

- ***Sampling bias.*** Arises from non-random sampling methods (sometimes related to representation bias)
- ***Longitudinal Data Fallacy.*** Changes over time for an aggregate group differ significantly when such changes are studied in cohorts
- ***Linking Bias.*** Network attributes from users in a social network mis-represent actual attributes in the real world


Ira A. Fulton Schools of
Engineering
Arizona State University

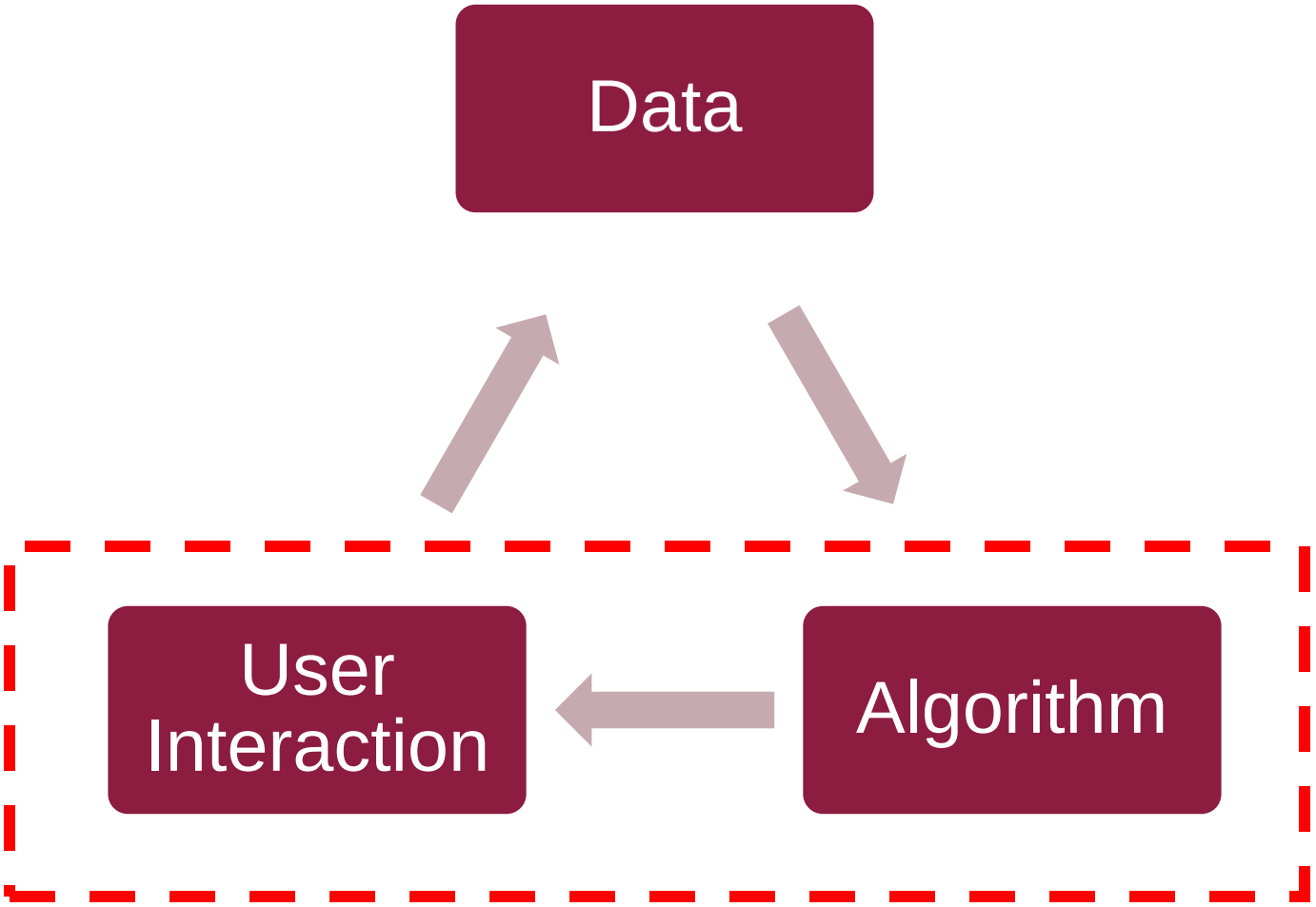
Algorithm-to-User Bias

Objectives

By the end of this module, you will be prepared to:

- Define “algorithm to user” bias and list several examples

The interrelationship between data, algorithms, and user interaction



Algorithm to User Bias

An ***algorithm to user bias*** is a bias in an algorithmic result that may influence the user's behavior.

Algorithmic Bias

- ***Algorithmic bias*** occurs when a bias not present in the data emerges after the algorithm has processed it.
- In supervised and reinforcement learning, such a bias will be present in the model
- In unsupervised learning, the bias will be present in the resulting analysis (e.g., clusters or rules)
- Some contributors:
  - Optimization functions
  - Regularization strategies
  - Selection of target class

User Interaction Bias

- ***User Interaction Bias*** results from the user interface and how the user manipulates the resulting output

- ***Presentation Bias*** is bias resulting from how the information is presented in the interface

- ***Ranking Bias*** results from ranking algorithms (e.g. showing the “most popular” results) – such results can become self-referential

Popularity Bias

- ***Popularity Bias*** results from more popular items being exposed more

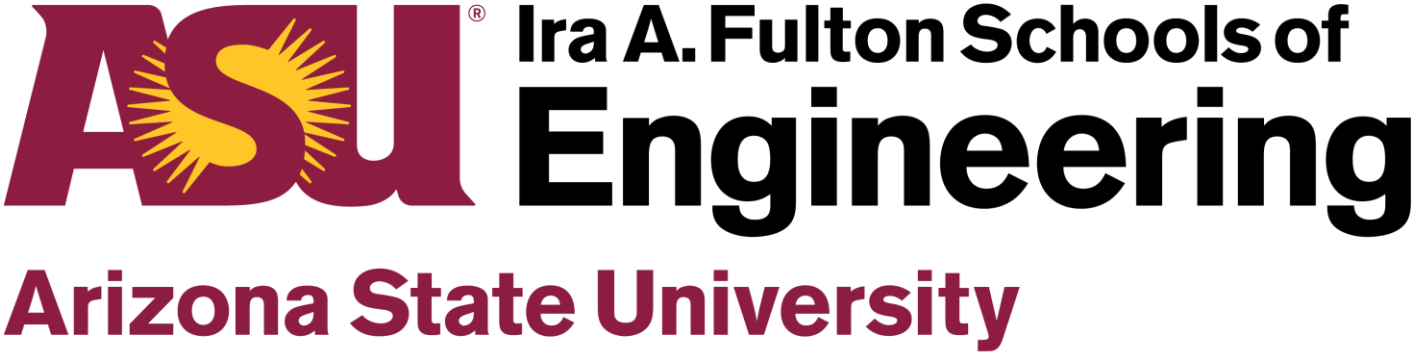
- However, such a bias is not necessarily a result of human interaction, but also can be due to social bots, fake reviews, and other forms of manipulation

Emergent Bias

- ***Emergent Bias*** occurs as a result of a system interacting with a changing population
- Often user-interface centric
- Normally occurs after a system has been deployed for a period of time and results from cultural changes in the population

Evaluation Bias

- ***Evaluation bias*** occurs when the evaluation criteria used to measure algorithm/model performance is biased
- Often occurs during validation and testing of a model
- Can also result by the selection of target class for which is the focus of measurement
- Note that a selection of an objective function – which a model is optimized for during training – is algorithmic bias while the selection of an evaluation metric during validation or testing is evaluation bias


Ira A. Fulton Schools of
Engineering
Arizona State University

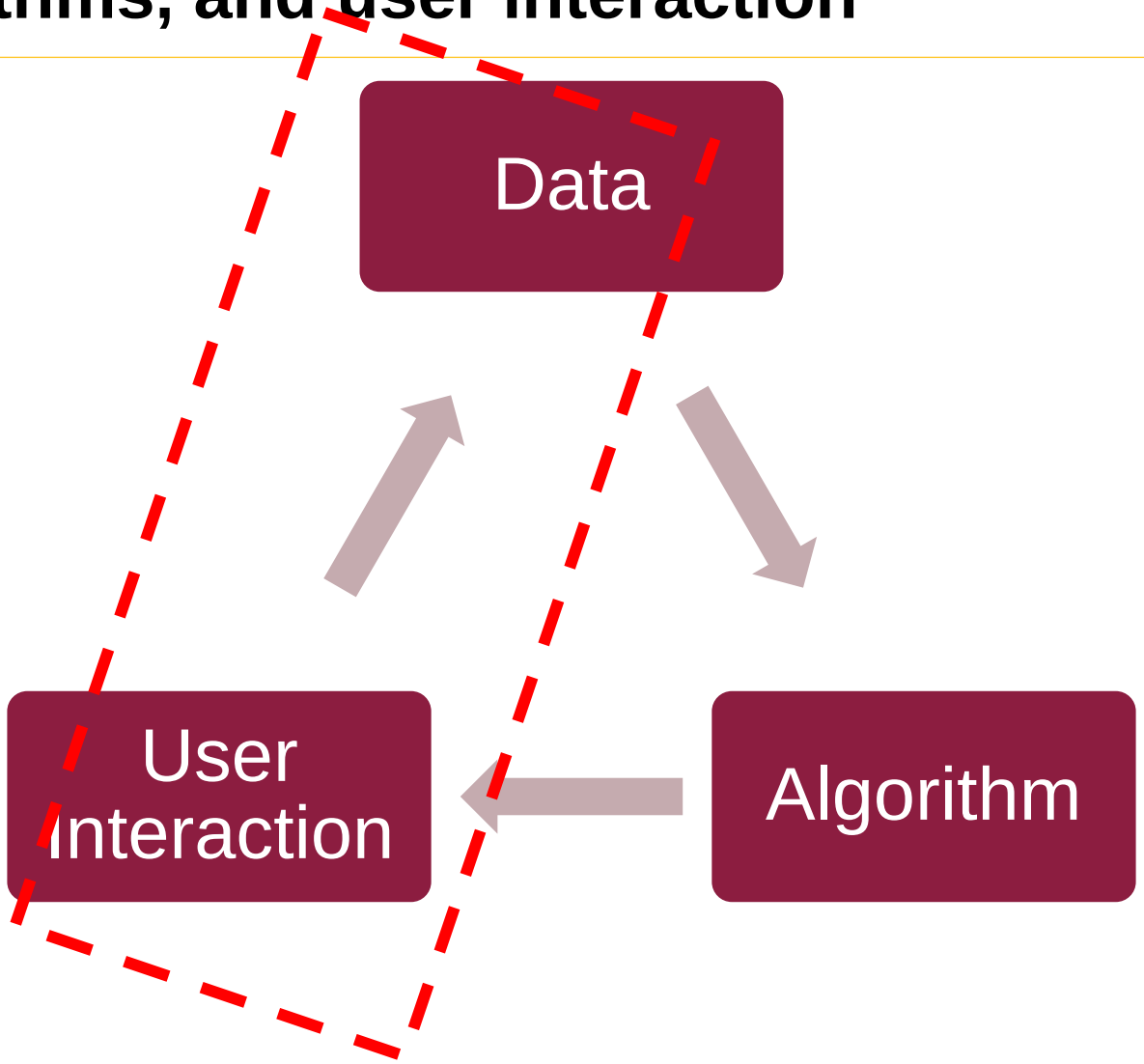
User-to-Data Bias

Objectives

By the end of this module, you will be prepared to:

- Define “user to data” bias and list several examples

The interrelationship between data, algorithms, and user interaction



User to Data Bias

- Certain applications such as social media, web search, and advertising leverage user-created data to train models

- Biases the user have, whether pre-existing or induced by an algorithm can be reflected in the data they generate

- Such biases are in the category of ***user to data bias***

Historical Bias

- ***Historical bias*** occurs when pre-existing bias in a population leads to user-generated data possessing that bias – which in turn affects downstream algorithms

Population Bias

- ***Population bias*** occurs when the population from which data is derived does not reflect the population as a whole.

- Example: women are over-represented in Facebook and Pinterest while men are over-represented in Reddit

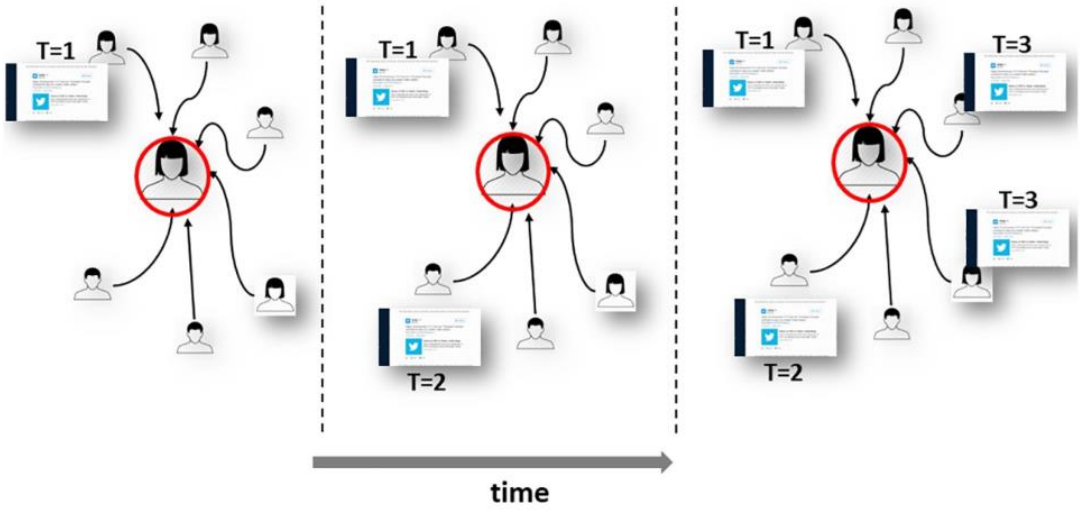
Self-Selection Bias

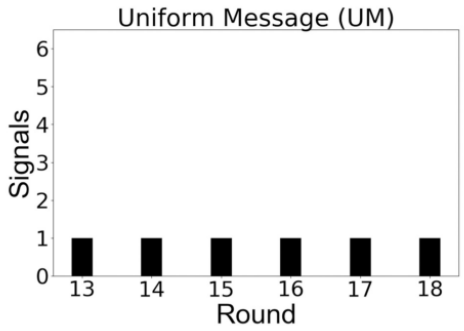
- When a subject with certain characteristic contributing to data has an incentive to self-select, such characteristics will be over-represented in the data. This is called ***self-selection bias.***

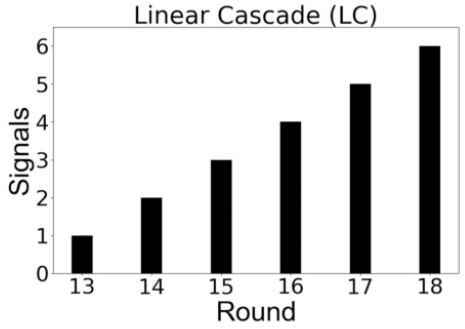
- The canonical example are individuals who are excited to vote for a certain political candidate, and are therefore more likely to respond to a polling inquiry

Social Bias

- ***Social Bias*** occurs when a user's contribution to a dataset is influenced by other users
- Classic example: social media influence
- We have observed this in experimental studies







Behavioral Bias

- ***Behavior bias*** occurs when user behavior changes across platforms, contexts, or datasets

- Example: emoji difference across different devices/platform led to different interpretations by users

Temporal Bias

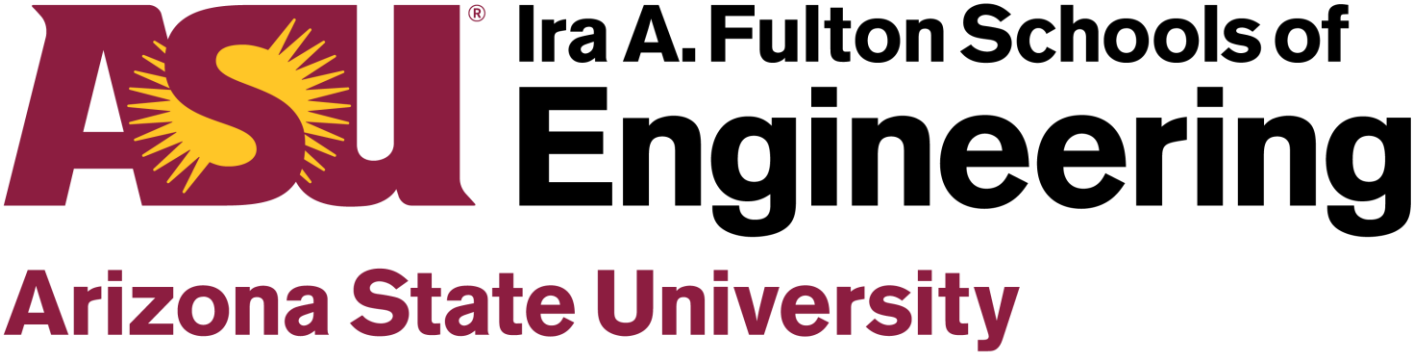
- ***Temporal bias*** arises from a change in population behavior over time

- Example: Twitter hashtag use becomes pronounced only in certain portions of a lifecycle of a topic

Content Production Bias

- Structural, lexical, semantic, and syntactic differences in user-generated content lead to ***content production bias.***

- Language, gender, and age all lead to such differences


Ira A. Fulton Schools of
Engineering
Arizona State University

Machine Learning and Discrimination

Objectives

By the end of this module, you will be prepared to:

- Describe explainable and unexplainable discrimination in machine learning

Review: Algorithmic Fairness

- ***Fairness*** is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics

- An ***unfair algorithm*** is one whose decisions are skewed toward a particular group of people

Definitions from Mehrabi et al., 2022 https://arxiv.org/pdf/1908.09635.pdf

Discrimination and Bias

- ***Discrimination*** is a source for unfairness that is due to human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally

- ***Bias*** can be considered as a source for unfairness that is due to the data collection, sampling, and measurement

- Note: it is possible for bias in data collection to be driven by discrimination

Definitions from Mehrabi et al., 2022 https://arxiv.org/pdf/1908.09635.pdf

Explainable Discrimination

- ***Explainable discrimination*** is when differences between groups in a dataset can be shown to come from attributes within the data
- Example: Differences between salaries among groups that are attributed to working hours
- Explainable discrimination is not considered illegal
- The removal of explainable discrimination can result in reverse discrimination

Unexplainable Discrimination

- When differences between groups are unjustified due to attributes of the data, such discrimination is ***unexplainable discrimination***
- Unexplainable discrimination is considered illegal
- Unexplainable discrimination consists of direct and indirect discrimination

Unexplainable Discrimination: Direct Discrimination

- ***Direct discrimination*** occurs when protected attributes are explicitly used to provide a non-favorable outcome to those possessing such attributes
- Often, the use of such attributes is prohibited by law (protected attributes under US FHA and ECOA shown)

Attribute	FHA	ECOA
Race	✓	✓
Color	✓	✓
National origin	✓	✓
Religion	✓	✓
Sex	✓	✓
Familial status	✓	
Disability	✓	
Exercised rights under CCPA		✓
Marital status		✓
Recipient of public assistance		✓
Age		✓

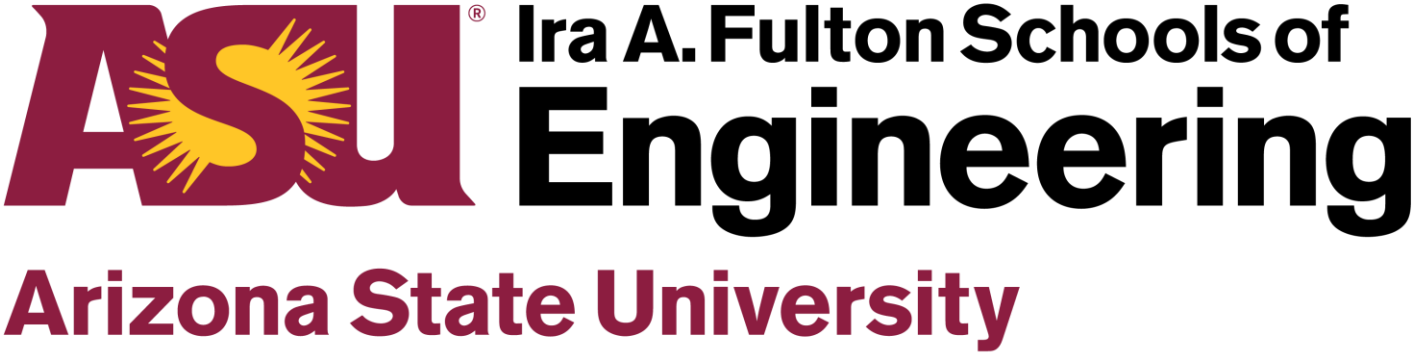
Protected attributes specified in the U.S. Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA)

Definitions from Mehrabi et al., 2022 https://arxiv.org/pdf/1908.09635.pdf
Table source: Chen et al., *Porc. ACM Conf. Fairness, Accountability, and Transparency, 2019.*

Unexplainable Discrimination: Indirect Discrimination

- ***Indirect discrimination*** occurs when individuals appear to be treated based on seemingly neutral and non-protected attributes; however, protected groups, or individuals still get to be treated unjustly as a result of implicit effects from their protected attributes
- Example: geographic location may correspond with race in certain areas

Definitions from Mehrabi et al., 2022 https://arxiv.org/pdf/1908.09635.pdf


Ira A. Fulton Schools of
Engineering
Arizona State University

Fairness Criteria

Objectives

By the end of this module, you will be prepared to:

- Define individual, group, and subgroup fairness

Types of Fairness

1. Individual Fairness

2. Group Fairness

3. Subgroup Fairness

Individual Fairness

- A system has ***individual fairness*** when it provides similar results for similar individuals

- Some common interpretations
  - ***Fairness through awareness.*** Individuals with similar attributes are near each other in the feature space
  - ***Fairness through unawareness.*** Protected attributes are not explicitly used in the decision-making process
  - ***Counterfactual fairness.*** An individual is treated in the same fashion if that individual were to belong to a different group

Group Fairness

- When a system treats different groups equally, it is said to possess ***group fairness***

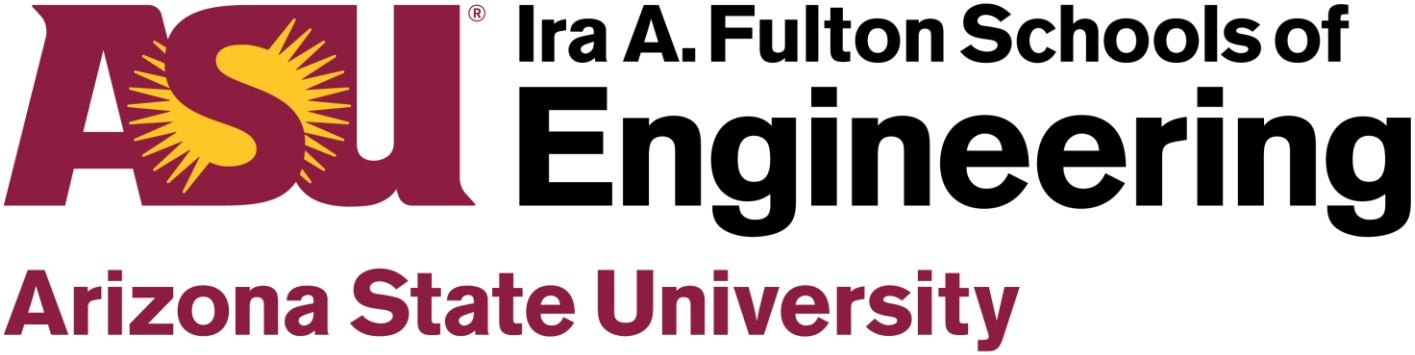
- Some common interpretations
  - ***Demographic parity.*** When the likelihood of a positive outcome should be the same regardless of if an individual is in a protected group
  - ***Conditional Statistical Parity.*** Individuals in both protected and non-protected groups should have the same probability of being assigned a positive outcome when conditioned on the set of legitimate attributes
  - ***Equalized Odds.*** Members of protected and non-protected groups should have equal rates for true positives and false positives
  - ***Equal Opportunity.*** Protected and unprotected groups have equal rates for true positives
  - ***Treatment Equality.*** Is when the ratio of false negatives to false positives is the same between groups
  - ***Test Fairness.*** For a given confidence value returned by the system, the probability of correctly belonging to the positive class is the same across all groups

Subgroup Fairness

- ***Subgroup fairness*** intends to obtain the best properties of the group and individual notions of fairness.
- Uses group and individual and group fairness definitions
- Intuition: Select a fairness constraint and examine whether it holds over a large collection of subgroups

Notes on Fairness

- Not all criteria can be simultaneously satisfied in real world problems
- Understanding impact on fairness measurements over time
- Measurement errors can skew fairness results
- Regardless of approach to ensuring fairness, important to understand the source of bias


Ira A. Fulton Schools of
Engineering
Arizona State University

Fair Machine Learning Practices

Objectives

By the end of this module, you will be prepared to:

- List methods for fair machine learning

Approaches to Fairness in Machine Learning Pre-Processing

- ***Pre-processing.*** Transform underlying data to remove discrimination.
- Can be leveraged in cases where the algorithm can modify underlying data

Approaches to Fairness in Machine Learning In-Processing

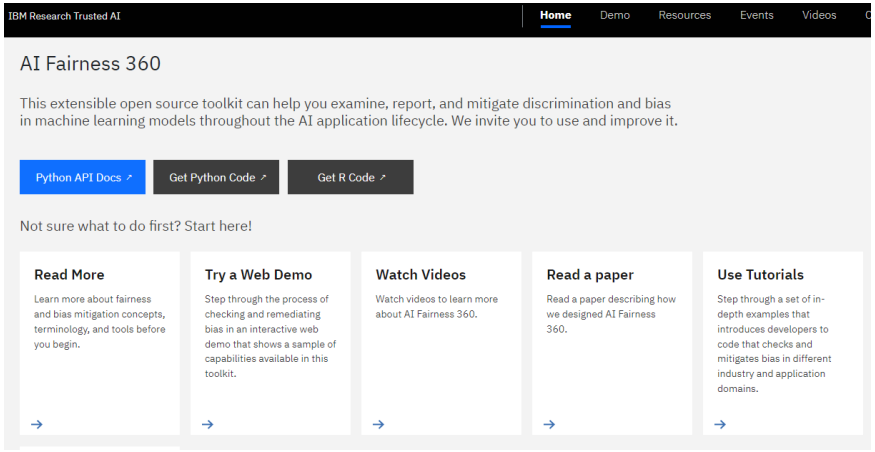
- ***In-processing.*** Involves changes to the algorithm to remove discrimination during model training
- Includes modifications of the objective function, adding constraints, or forms of regularization

Approaches to Fairness in Machine Learning Post-Processing

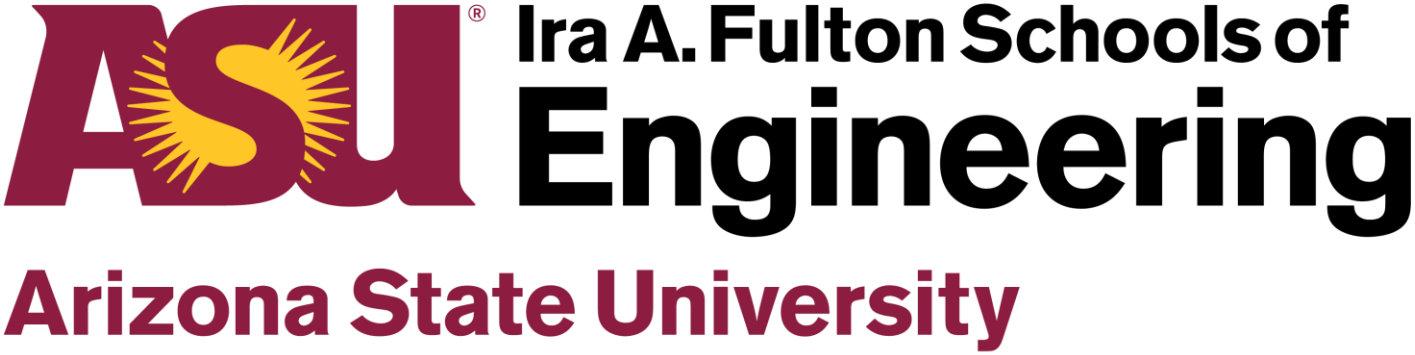
- ***Post-processing.*** Performed after training by accessing a holdout set which was not involved during the training of the model
- Concept: labels assigned by model are re-assigned based on a function during post-processing
- Useful in cases where the data or algorithm cannot be modified

Fairness Testing

- Key is evaluation of your machine learning system and the underlying data to ensure fairness
- Fairness definitions based on metrics such as TP/FP/TN/FN relatively easy to implement
- An open source project “AI Fairness 360” (curated by IBM) has many fairness methods implemented in Python and R



Screenshot Source: https://aif360.mybluemix.net/


Ira A. Fulton Schools of
Engineering
Arizona State University