MACHINE LEARNING MODELS WORKSHOP I:

METHODS FOR DETECTING & CORRECTING BIAS

AGENDA

- Introduction
- Importance of Fairness
- Definition of Algorithmic Fairness
- Detect/Measure Biasness – Theory and Python Hands on
- Remove/reduce Biasness – Theory and Python Hands on

LEGALLY RECOGNIZED PROTECTED CLASSES (USA)

RACE Civil Risk Act 1964	**Color** Civil Rights Act 1964	**Sex** Equal Pay Act 1963, Civil Rights Act 1964
Religion Civil Risk Act 1964	**National Origin** Civil Rights Act 1964	**Citizenship** Immigration Reform and Control Act
Age Age Discrimination in Employment Act 1967	**Pregnancy** Pregnancy Discrimination Act	**Family Status** Civil Rights Act 1968
Disability Status Rehabilitation Act 1973, Americans with Disabilities Act 1990	**Veterans Status** Vietnam Era Veterans' Readjustment Assistance Act 1974, Uniformed Services Employment and Reemployment Rights Act	**Genetic Information** Genetic Information Nondiscrimination Act

REGULATED DOMAINS - USA

Credit
Equal Credit Opportunity Act

Employment
Civil Rights Act 1964

Education
Civil Rights Act 1964, Education Amendments 1972

Housing
Fair Housing Act

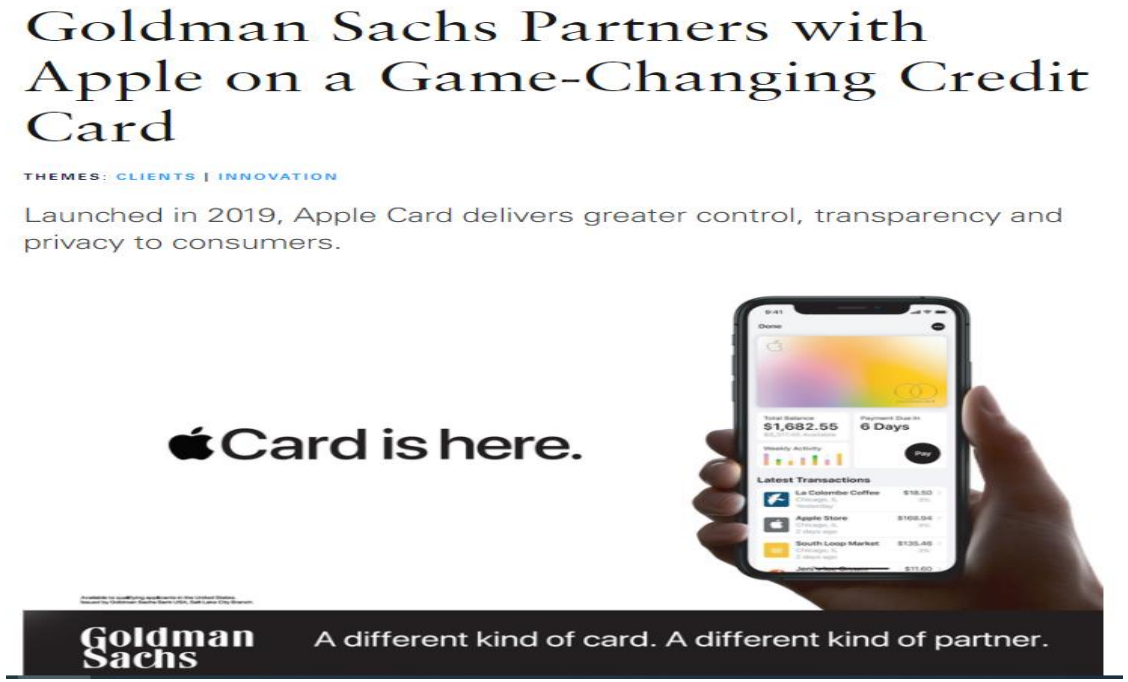
Public Accommodation
Civil Rights Act 1964

Extends to Marketing and Advertising
Not Limited to Final Decision

The above List sets aside complex web of Laws that regulates the government

EXAMPLES OF BIASNESS

Apple, Goldman Face Criticism Over Alleged Sexism in Credit Card Algorithm



Amazon scraps secret AI recruiting tool that showed bias against women

INTRODUCTION

- Bias in this context is unfairness (more or less)
- True definition is a wild good chase. Lack of concrete definition. How to develop model correctly?
- Researchers and practioners have come up with many definition of fairness
- Predictive Accuracy - Model developers/decision makers goal is to maximize accuracy subject to fairness constraint
- Standard statistical bias in machine learning (the bias in the bias vs. variance tradeoff)
- Bias = Expected value of model - true value
- Is statistical bias an adequate fairness criteria?
- Why statistical bias is not enough?
  - Error or distribution of errors
  - Data bias
- Real challenge is how to make algorithms systems support human values? We need to align with this objective

WHY FAIRNESS IS IMPORTANT

Fairness is becoming one of the most popular topics in machine learning in recent years. The main motivation is that it is highly related to our own benefits

- We are at an age where many things have become or are becoming automated by ML systems. e.g.
  - Driverless cars are around the corner and are estimated to be widely used within 5–10 years;
  - Employers use ML system to select job applicants;
  - Courts in United States use COMPAS algorithm for recidivism prediction;
  - Linked-in uses ML to rank job candidates queried;
  - Amazon uses recommender system to recommend items and decide the order of items appearing on a page.
  - Netflix uses recommender system to present customized page for every user
- Machine learning systems have been an inseparable part of our daily lives. They will be even more widely used in the near future as more and more fields begin to integrate AI into their existing practice/products
- AI is good, but it can be used incorrectly. ML, the most widely used AI technique, relies heavily on data
- It is a common misconception that AI is absolutely objective. AI is objective only in the sense of learning what human teaches. The data provided by human can be highly-biased

FAIRNESS TERMS ONE SHOULD KNOW

Protected Attribute

An attribute that partitions a population into groups whose outcomes should have parity (e.g. race, caste, gender, religion)

Group Fairness

Groups defined by protected attributes receiving similar treatments or outcomes

Fairness Metric

Measure of unwanted bias in training data or models

Privileged Protected Attribute

A protected attribute value indicating a group that has historically been at systematics advantage

Individual Fairness

Similar Individuals receiving similar treatments or outcomes

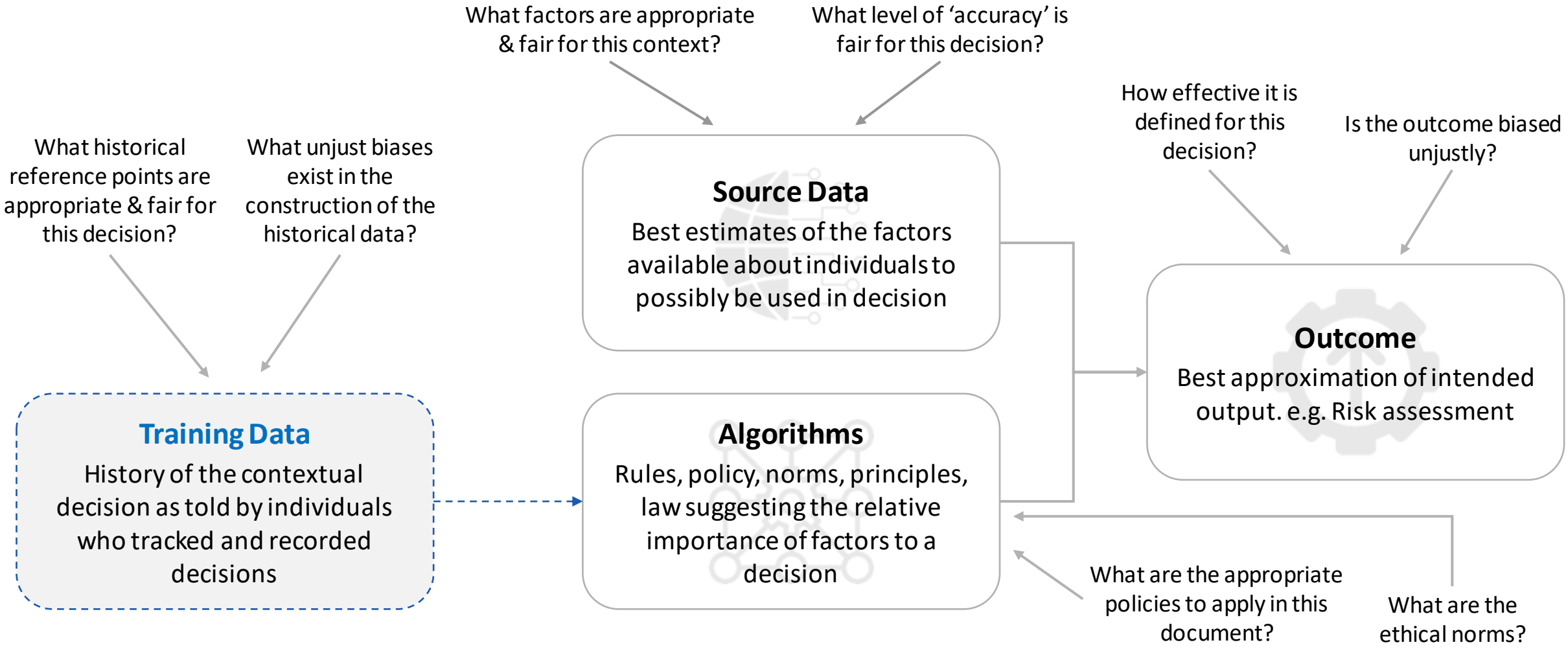
Favourable Label

A label whose value corresponds to an outcome that provides an advantage to recipient

FAIRNESS IN MACHINE LEARNING – SOME STANDARD TERMINOLOGY

- A **favorable label** is a label whose value corresponds to an outcome that provides an advantage to the recipient. Examples are receiving a loan, being hired for a job, and not being arrested.
- A **protected attribute** is an attribute that partitions a population into groups that have parity in terms of benefit received. Examples include race, gender, caste, and religion. Protected attributes are not universal, but are application specific.
- A **privileged value** of a protected attribute indicates a group that has historically been at a systematic advantage
- **Group fairness** is the goal of groups defined by protected attributes receiving similar treatments or outcomes
- **Individual fairness** is the goal of similar individuals receiving similar treatments or outcomes
- **Bias** is a systematic error. In the context of fairness, we are concerned with unwanted bias that places privileged groups at a systematic advantage and unprivileged groups at a systematic disadvantage.
- A **fairness metric** is a quantification of unwanted bias in training data or models
- A bias mitigation algorithm is a procedure for reducing unwanted bias in training data or models

ML MODEL DEVELOPMENT



ML/AI MODELS: FTC, FRCA AND ECOA EXPECTATIONS

Be Transparent

- Don't deceive consumers about how you use your ML models
- Be transparent when collecting sensitive data
- If you make automated decisions based on information from a third-party vendor, you may be required to provide the consumer with an "adverse action" notice.

Explainability

- Explain if you deny consumers something based on algorithmic decision-making
- For algorithm-based risk scores to consumers, disclose the key factors that affected the score, rank ordered for importance
- Tell consumers if terms of a deal changes based on automated tools

Sound Data and Models

- Written policies and procedures to ensure that the data they furnish is accurate and has integrity
- Ensuring that the data is accurate and up to date
- Make sure that ML/AI models are validated and revalidated to ensure that they work as intended, and do not illegally discriminate

Ensure Fairness

- Don't discriminate based on protected classes
- Focus on inputs, but also on outcomes
- Give consumers access & an opportunity to correct information used to make decisions about them

Accountability

- Ask questions:

a) How representative is your data set?
b) Does your data model account for biases?
c) How accurate are your predictions based on big data?
d) Does your reliance on big data raise ethical or fairness concerns?

- Consider your accountability mechanism
- Protect your algorithm from unauthorized use.

MACHINE LEARNING: BIAS IN – BIAS OUT

Bias in Data

Data is often heterogeneous, generated by subgroups with their own characteristics and behaviors.

- Historical bias
- Representation bias
- Measurement bias
- Population bias
- Collection bias
- External source bias

Data Cleansing and Variable Selection

Algorithmic bias is added purely by the algorithm

- Sampling bias
- Missing values & outliers bias
- Omitted variable bias
- Cause and effect bias
- Inappropriate qualitative analysis bias
- Transformation, cleansing and normalization bias

Algorithmic Bias

Algorithmic bias is added purely by the algorithm

- Choice of methodology
- Estimation methodology bias
- Aggregation bias
- Non-interpretable and non-explainable models
- Model automation bias

Evaluation Bias

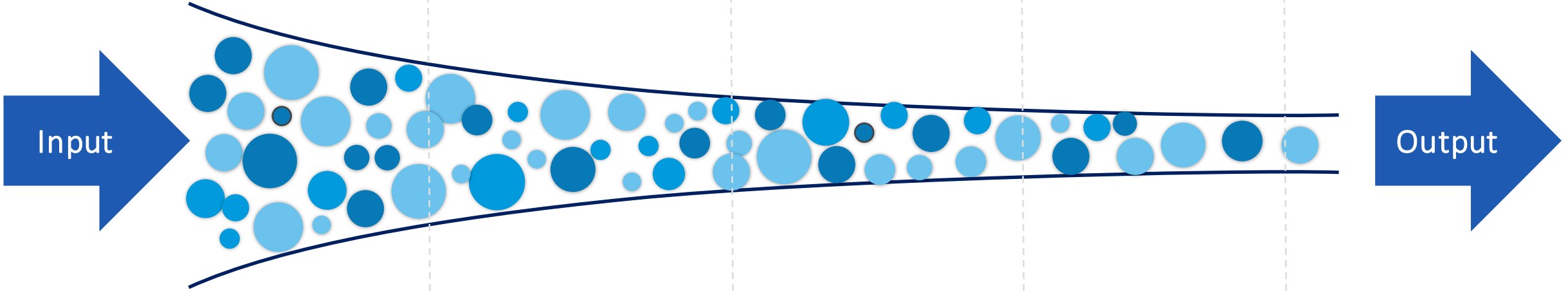
Evaluation bias happens during model evaluation

- Ranking bias
- Equal Metrics bias
- Inappropriate benchmark
- Performance metrics bias
- Replicability and reproducibility bias

Interaction Bias

Triggered from two sources - the user interface and through the user

- Presentation bias
- Linking bias
- Content Production bias
- Emergent bias



IMPACT OF BIAS IN FINANCIAL SERVICES ML MODELS

Consumer Financial Protection Problem

- Data and ML bias can lead to problems: product steering, discriminatory pricing, unfair credit rationing, exclusionary filtering, accountability and digital redlining
- The stakes become higher when developers does not care of fairness and ethical issues from the start
- Innovative tools, oversight policy and frameworks would be useful for consumer financial protection regulatory regimes such as FTC, FCRA, ECOA, BSA and AML

Discrimination

- If the data and ML models are biased, they could disparate impact the way millions of consumers borrow, save and manage their money.
- Differentiate customers w.r.t to age, ethnicity, color, race, religion, gender, locality, profession, legal etc.,
- Discrimination on the basis of any protected class is illegal under the US Equal Credit Opportunity Act of 1974.

Interpretability

- Data bias and bias in ML model development stages can lead to ubiquity, unfair, opaque and non-interpretable models
- Transparency around ML models are regulators' greatest obstacle
- Need of robust oversight to ensure ML applications remain accountable to society, the people and government and circumvent discriminatory bias

Vendor Management

- The black box and complex ML algorithms offered by third-parties and opens sources may not perform as intended and may lack data protection, interpretability and quality documentation.
- Such gaps calls for conducting third-party risk management (SR 13-19), data protection standards, SR 11-7 compliant documentation, and technology management guidelines geared toward AI applications.

Supervision

- As data set, ML algorithms and model complexity increases, so the models will become even more inscrutable
- The shift toward relatively more sophisticated, non-linear (often metaphorically described as "black box") models necessitates new governance processes
- Call for continuously monitoring performance and timely maintenance of developed models.

Coffee Break

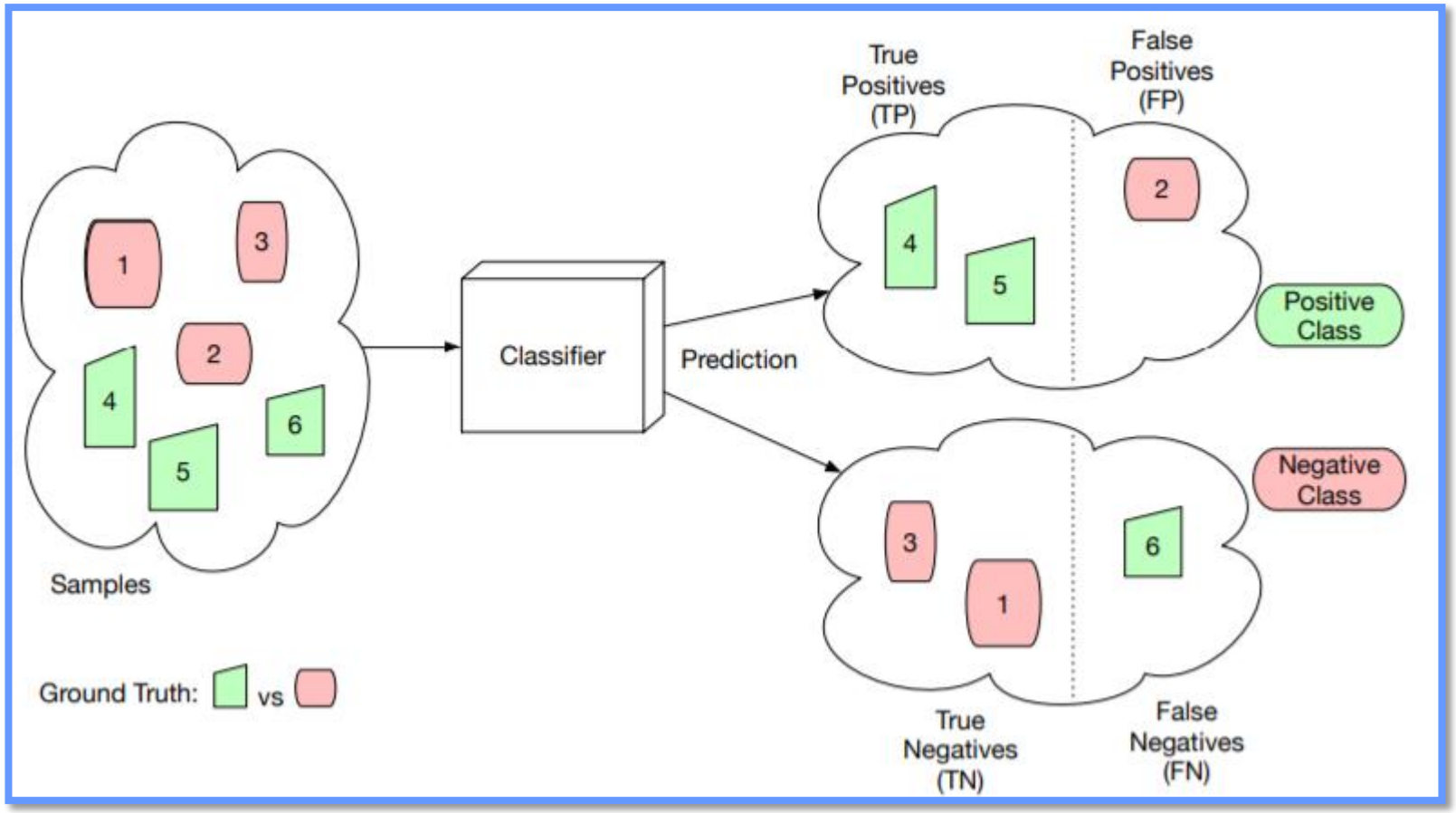
Lets Meet in 15 Minutes

FAIRNESS METRICS EVALUATION SETTING

- $X \in \mathbb{R}^d$ denote a non-sensitive feature vector describing an individual.
- $g \in \{0,1\}$ denote sensitive binary feature describing the gender of the individual. g is assumed to be a binary variable representing the sex (male or female) of the credit seeker.
- $Y \in \{0,1\}$ represent the actual or true labels. Y is either an approved or rejected decision (or good vs. bad credit). As such, it must be denoted as a binary variable in the dataset.
- $S \in [0,1]$ is the predicted probability. For a logistic regression model, it is defined as:

- $S = Pr(Y = 0, 1 \mid X, g) = \exp[\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \beta g g] / (1 + \exp[\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \beta g g])$
  where α, β_i, and βg are the estimated regression coefficients.
- $\hat{y} \in \{0,1\}$ is the prediction decision of the algorithm.
- It depends on the predicted probability (S) of an applicant having good or bad credit. For a threshold value of S^*, $\hat{y} = 1$ when $S > S^*$, and $\hat{y} = 0$ when $S < S^*$.

OVERVIEW OF CLASSIFICATION



DISCRIMINATION IN MACHINE LEARNING

Disparate Treatment	▪ Situation where a policy or decision is made that explicitly discriminates against a demographic group of people based on race, age, gender etc. ▪ Due to the fairness blind optimization functions often used in practice, a machine learning algorithm may end up making decisions based on a protected attribute like gender
Disparate Impact	▪ Disparate Impact involves policies or decisions that implicitly discriminate against a certain group of people. ▪ A decision making system suffers from disparate impact if it provides outputs that benefits (hurt) a group of people sharing a value of a sensitive feature more frequently than other groups of people . ▪ With the massive amounts of data we feed our machine learning algorithms, they are likely to find features that correlate with sensitive attributes such as race. ▪ Therefore even if an algorithm is not explicitly trained to make unequal decisions for different demographic groups, it may do just so.

GROUP LEVEL FAIRNESS

Group-level notions of fairness require just (equal) distributions of benefit or harm across groups. Metrics belonging to this category are based on observational criteria as they depend on one of the following joint distributions

Subcategory 1

- Distribution between predicted outcome $\hat{y}$ and sensitive attribute g (marginalizing features X and the actual outcome Y)
- There are two metrics that belong to this category commonly known as **demographic parity** and **conditional demographic parity**

Subcategory 2

- Distribution between predicted outcome $\hat{y}$, sensitive attribute g, and the actual outcome Y (marginalizing features X)
- Metrics belonging to this subcategory look beyond the absolute distribution of predictions $\hat{y}$ by conditioning fairness on accuracy or imperfections of the model

Subcategory 3

- The fairness notions in this subcategory are based on some distribution between predicted probability S, sensitive attribute g, and the actual outcome Y

Source: Verma and Rubin (2018), Fairness Definition Explained, 2018 ACM/IEEE International Workshop on Software Fairness

SUBCATEGORY1 FAIRNESS METRICS

Fairness Metrics

- **Demographic parity** requires that the positive prediction between the two groups of males and females be equal
- **Conditional demographic parity** extends the definition of demographic parity by conditioning the outcome to be based on certain legitimate factors (L)
- While there is no unified agreement on what legitimate factors are, it is useful to mention that conventionally, legitimacy o f loan approval decisions is based on evaluating key factors, commonly known as credit history, character, collateral, and capacity

	Metric	Mathematical notation	Requires
Independence	Demographic parity	$Pr(\hat{y} = 1 \mid g = 0) = Pr(\hat{y} = 1 \mid g = 1)$	Equal probability of positive predictions for both males and females.
	Conditional demographic parity	$Pr(\hat{y} = 1 \mid g = 0, L) = Pr(\hat{y} = 1 \mid g = 1, L)$	Equal probability of positive predictions for both males and females, but condition on a subset of legitimate input features.

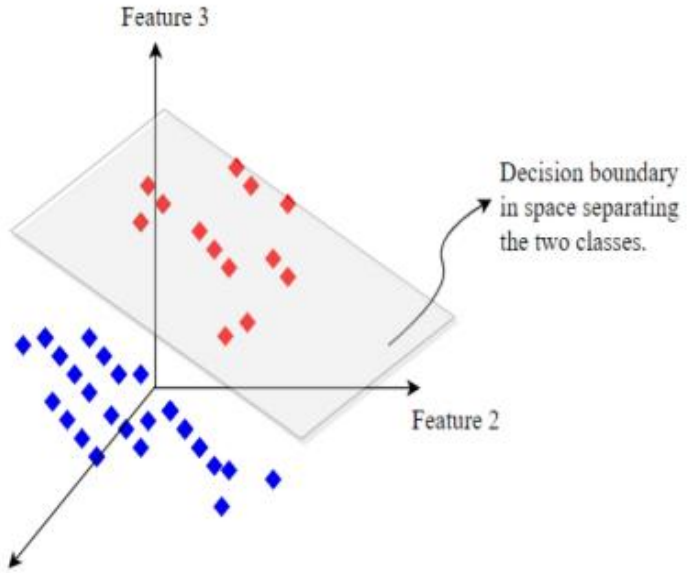
- The two metrics require that the decision outcomes $\hat{y}$ be uncorrelated with sensitive features g: $\hat{y} \perp g$.
- Parity measures can be problematic?

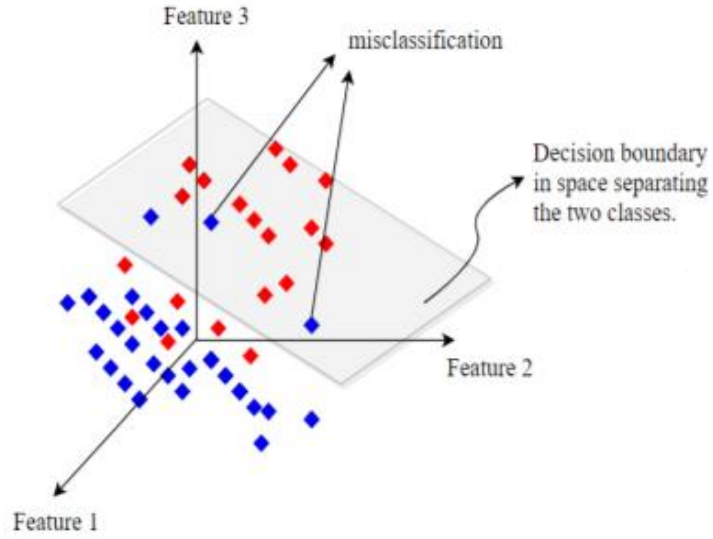
Source: Verma and Rubin (2018), Amirali Khaleghi, M. S Thesis, Delft University of Technology, March 2020

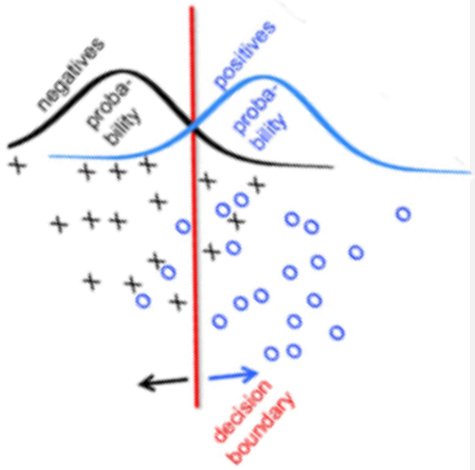
SUBCATEGORY2 FAIRNESS METRICS (1/6)

Fairness Metrics

- Metrics belonging to this **subcategory 2** look beyond the absolute distribution of predictions ŷ by conditioning fairness on accuracy or imperfections of the model.
- In an ideal case, a logistic regression model or other ML model may find a decision boundary such that the two binary classes are linearly separable and that output decisions are 100% accurate.







- If the items in the training data are not perfectly, linearly separable, the hyper plane/linear boundary estimated will misclassify some of the data points.

Source: Verma and Rubin (2018), Amirali Khaleghi, M. S Thesis, Delft University of Technology, March 2020

SUBCATEGORY2 FAIRNESS METRICS (2/6)

Do outcomes systematically different between demographic group or other population group?

Predicted	Actual			
	Positive Class		Negative Class	
Positive Class	**True Positive**	Hit	**False Positive**	Type I Error (False Alarm)
	Reality	Money laundering transaction	**Reality**	Not a money laundering transaction
	Prediction	**Bad** transaction predicted	**Prediction**	**Bad** Transaction predicted
	Results	Good	**Results**	Unnecessary work and cost
Negative Class	**False Negative**	Type II Error (Miss)	**True Negative**	Correct reject
	Reality	Money laundering transaction	**Reality**	Not a money laundering transaction
	Prediction	**Good** transaction predicted	**Prediction**	**Good** transaction predicted
	Results	Not a good sign	**Results**	Good

- Applicable to many context – e.g. Loan (default versus non default), Hiring (Succeed at Job vs. did not)
- Yields useful insights
- What do different stakeholders want of the binary classifier (Logistic regression, ANN, GBM, SVM etc)?
- E.g. Decision maker/model developers and Defendants (wrongly classified as future criminals)

SUBCATEGORY2 FAIRNESS METRICS (3/6)

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

If the model satisfies predictive parity but the prevalence differs between groups, the model can achieve equal false positive rates and equal false negative rates across the groups.

SUBCATEGORY2 FAIRNESS METRICS (4/6)

✓	**True positive (TP)**	A case when the predicted and actual outcomes are both in the positive class
✓	**False positive (FP)**	A case predicted to be in the positive class when the actual outcome belongs to the negative class
✓	**False negative (FN)**	A case predicted to be in the negative class when the actual outcome belongs to the positive class
✓	**True negative (TN)**	A case when the predicted and actual outcomes are both in the negative class
✓	**Positive predictive value (PPV)**	The fraction of positive cases correctly predicted to be in the positive class out of all predicted positive cases, TP/(TP+FP). PPV is often referred to as precision, and represents the probability of a subject with a positive predictive value to truly belong to the positive class, P(Y = 1|d = 1). In our example, it is the probability of an applicant with a good predicted credit score to actually have a good credit score
✓	**False discovery rate (FDR)**	The fraction of negative cases incorrectly predicted to be in the positive class out of all predicted positive cases, FP/(TP+FP) . FDR represents the probability of false acceptance, P(Y = 0|d = 1), e.g., the probability of an applicant with a good predicted credit score to actually have a bad credit score
✓	**False omission rate (FOR)**	The fraction of positive cases incorrectly predicted to be in the negative class out of all predicted negative cases, FN/(TN +FN) . FOR represents the probability of a positive case to be incorrectly rejected, (P(Y = 1|d = 0)), e.g, the probability of an applicant with a bad predicted credit score to actually have a good score

SUBCATEGORY2 FAIRNESS METRICS (5/6)

✓	**Negative predictive value (NPV)**	The fraction of negative cases correctly predicted to be in the negative class out of all predicted negative cases, TN/(TN +FN) . NPV represents the probability of a subject with a negative prediction to truly belong to the negative class, P(Y = 0|d = 0), e.g., the probability of an applicant with a bad predicted credit score to actually have such score
✓	**True positive rate (TPR)**	The fraction of positive cases correctly predicted to be in the positive class out of all actual positive cases, TP/(TP+FN) . TPR is often referred to as sensitivity or recall; it represents the probability of the truly positive subject to be identified as such, P(d = 1|Y = 1). In our example, it is the probability of an applicant with a good credit score to be correctly assigned with such score
✓	**False positive rate (FPR)**	The fraction of negative cases incorrectly predicted to be in the positive class out of all actual negative cases, FP/(FP+TN) . FPR represents the probability of false alarms – falsely accepting a negative case, P(d = 1|Y = 0), e.g., the probability of an applicant with a actual bad credit score to be incorrectly assigned with a good credit score
✓	**False negative rate (FNR)**	The fraction of positive cases incorrectly predicted to be in the negative class out of all actual positive cases, FN/(TP+FN). FNR represents the probability of a negative result given an actually positive subject, P(d = 0|Y = 1), e.g., the probability of an applicant with a good credit score to be incorrectly assigned with a bad credit score
✓	**True negative rate (TNR)**	The fraction of negative cases correctly predicted to be in the negative class out of all actual negative cases, TN/(FP+TN) . TNR represents the probability of a subject from the negative class to be assigned to the negative class, P(d = 0|Y = 0), e.g., the probability of an applicant with a bad credit score to be correctly assigned with such score

SUBCATEGORY2 FAIRNESS METRICS (6/6)

	Metric	Mathematical notation/Alias	Requires
Accuracy/Sufficiency	Predictive parity	$Pr(Y=1 \mid \hat{y}=1, g=0) = Pr(Y=1 \mid \hat{y}=1, g=1)$ *False discovery rate:* $Pr(Y=0 \mid \hat{y}=1, g=0) = Pr(Y=0 \mid \hat{y}=1, g=1)$	Equal fraction of correct positive predictions for both male and females.
	Overall accuracy equality	$Pr(\hat{y}=Y, g=0) = Pr(\hat{y}=Y, g=1)$ *Overall misclasification rate:* $Pr(Y \neq \hat{y} \mid g=0) = Pr(Y \neq \hat{y} \mid g=1)$	Equality of positive and negative predictive values for both males and females.
	Conditional accuracy equality	$Pr(Y=1 \mid \hat{y}=1, g=0) = Pr(Y=1 \mid \hat{y}=1, g=1) \cap$ $Pr(Y=0 \mid \hat{y}=0, g=0) = Pr(Y=0 \mid \hat{y}=0, g=1)$	Equal correct positive predictions and equal correct negative predictions for both males and females.
Misrepresentation/Separation	False positive error rate	$Pr(\hat{y}=1 \mid Y=0, g=0) = Pr(\hat{y}=1 \mid Y=0, g=1)$ *True negative rate:* $Pr(\hat{y}=0 \mid Y=0, g=0) = Pr(\hat{y}=0 \mid Y=0, g=1)$	Equal incorrect positive predictions for both male and females.
	False negative error rate	$Pr(\hat{y}=0 \mid Y=1, g=0) = Pr(\hat{y}=0 \mid Y=1, g=1)$ *True negative rate:* $Pr(\hat{y}=0 \mid Y=0, g=0) = Pr(\hat{y}=0 \mid Y=0, g=1)$	Equal incorrect negative predictions for both male and females.
	Equalized odds	$Pr(\hat{y}=1 \mid Y=0, g=0) = Pr(\hat{y}=1 \mid Y=0, g=1) \cap$ $Pr(\hat{y}=0 \mid Y=1, g=0) = Pr(\hat{y}=0 \mid Y=1, g=1)$	Equality of false positive and false negative error rates for both males and females.
	Treatment equality	(FNR/FPR) for males = (FNR/FPR) for females	Equal ratio of false positive and false negative error rates for both males and females.

Source: Verma and Rubin (2018), Amirali Khaleghi, M. S Thesis, Delft University of Technology, March 2020

Two statistical fairness criteria, separation and sufficiency are introduced.

Sufficiency requires the true label Y to be statistically independent of group-membership conditional on prediction outcome $\hat{y}$ (or decision):
$Y \perp g \mid \hat{y}$.

Individuals about whom the same decision is made have the same statistical prospects of being either true label, regardless of their irrelevant feature.

Separation is satisfied when the prediction/decision $\hat{y}$ is statistically independent of group-membership g conditional on the true label
$Y: \hat{y} \perp g \mid Y$.

Individuals with the same true label have the same statistical prospects of either decision, regardless of their irrelevant feature

Coffee Break

Lets Meet in 15 Minutes

EVALUATE FOR FAIRNESS

Female Applicants Results	
True Positive (TP) = 10	False Positive (FP)=1
False Negative (FN)=1	True Negative (TN)=488

Male Applicants Results	
True Positive (TP) = 6	False Positive (FP)=3
False Negative (FN)=5	True Negative (TN)=48

Precision =	TP/(TP+FP) =	10/(10+1) =	0.9090

Predictive parity

Recall=	TP/(TP+FN) =	10/(10+1) =	0.9090

Equality of Opportunity

Precision =	TP/(TP+FP) =	6/(6+3) =	0.6666

Recall=	TP/(TP+FN) =	6/(6+5) =	0.5454

SUBCATEGORY3 FAIRNESS METRICS

Fairness Metrics

- The fairness notions in this subcategory are based on some distribution between predicted probability S, sensitive attribute g, and the actual outcome Y.
- The fairness criteria imposed on these metrics are in correspondence to their similar metric in Subcategory 2

Metric	Mathematical Notation	Requires	Similar to
Calibration	$Pr(Y=1 \mid S=s, g=0) = Pr(Y=1 \mid S=s, g=1)$	Equal probability to belong to true positive label.	Predictive parity (except it considers fraction of positive predictions)
Well-calibration	$Pr(Y=1 \mid S=s, g=0) = Pr(Y=1 \mid S=s, g=1) = s$	Equal probability to belong to true positive label and the probability to equal *S*.	Predictive parity (except it considers fraction of positive predictions)
Balance for positive class	$E(S \mid Y=1, g=0) = E(S \mid Y=1, g=1)$	Equal expected value of probability S for individuals with positive true label.	• False negative error rate • True positive rate.
Balance for negative class	$E(S \mid Y=0, g=0) = E(S \mid Y=0, g=1)$	Equal expected value of probability S for individuals with negative true label.	• False positive error rate balance • true negative rate

Source: Verma and Rubin (2018), Amirali Khaleghi, M. S Thesis, Delft University of Technology, March 2020

Ideally, one would desire to achieve an algorithmic decision-making model that satisfies all the group level notions

But this is not achievable unless for some improbable circumstances.

Consequently, careful choices have to be made between one subcategory and another

EVALUATE FOR FAIRNESS

Female Applicants Results	
True Positive (TP) = 10	False Positive (FP)=1
False Negative (FN)=1	True Negative (TN)=488

Male Applicants Results	
True Positive (TP) = 6	False Positive (FP)=3
False Negative (FN)=5	True Negative (TN)=48

Precision =	TP/(TP+FP) =	10/(10+1) =	0.9090

Predictive parity

Recall=	TP/(TP+FN) =	10/(10+1) =	0.9090

Equality of Opportunity

Precision =	TP/(TP+FP) =	6/(6+3) =	0.6666

Recall=	TP/(TP+FN) =	6/(6+5) =	0.5454

INDIVIDUAL FAIRNESS

"Treat similar examples similarly" i.e. similar individuals should be treated similarly

Flaws

- It is hard to determine what is an appropriate metric function to measure the similarity of two inputs
- In our case, it is hard to quantify the difference between two job candidates. Imagine three job applicants, A, B and C
  - A has a bachelor degree and 1-year related work experience
  - B has a master degree and 1-year related work experience
  - C has a master degree but no related work experience
- Is A closer to B than C? If so, by how much? Such question is hard to answer
- It becomes even worse when the sensitive attribute(s) comes into the play

SIMILARITY –BASED MEASURES (INDIVIDUAL LEVEL)

Metric	Mathematical notation	Requires
Causal discrimination	$(X_i = X_j) \cap (g_i \mathrel{!}= g_j) \rightarrow \hat{y}_i = \hat{y}_j$	Two identical individuals that otherwise have different gender, receive the same classification.
Fairness through awareness	$(X_i = X_j) \rightarrow \hat{y}_i = \hat{y}_j$	Two identical individuals that otherwise have different gender, receive the same classification, and no sensitive feature is explicitly used in the model.
Fairness through unawareness	$(X_i \sim X_j) \cap (g_i \mathrel{!}= g_j) \rightarrow \hat{y}_i = \hat{y}_j$	Similar individuals receive similar classification

Source: Verma and Rubin (2018), Amirali Khaleghi, M. S Thesis, Delft University of Technology, March 2020

- Causal discrimination requires the model to be trained on the sensitive feature g
- Fairness through unawareness excludes sensitive attribute g from the feature set
- Ensuring that identical or similar individuals that only differ by gender receive similar classification outcomes and failing to do so is an indication of direct discrimination or indirect discrimination

AGENDA – SESSION II

- Estimating Fairness Metrics Using Python
- Pre-Processing
- In-Processing
- Post processing
- Exercise of Some Popular Techniques - Python

Session 2

Python Demo – Fairness Metrics

Pre- Processing

In - processing

Pro-procesisng