# Certifying and Removing Disparate Impact[*]

# Auditing Black-box Models for Indirect Influence[†]

Michael Feldman[*], Sorelle A. Friedler[*†], John Moeller[*], Carlos Scheidegger[*†], Suresh Venkatasubramanian[*†], Philip Adler[†], Casey Falk[†], Gabriel Rybeck[†], Brandon Smith[†]

Presenter: Nima Shahbazi

# Outline

- Introduction
- Contributions
- Disparate Impact and Error Rates
- Predictability and Disparate Impact
- Certifying lack of Disparate Impact
- Removing Disparate Impact
- Auditing Black-box Models for Indirect Influence
- Experiments
- Discussion

# Legal doctrine of **Disparate Impact** is born...

- The Curious Case of Griggs v. Duke Power Co.
  - **Claim:** Discrimination against African-American employees in violation of Title VII of Act
    - Title VII of the Act: Prohibits discrimination based on race, color, religion, sex, or national origin
  - Transfer Policy:
    - Mechanical Test
    - IQ Test
    - Highschool Diploma
  - Setting: North Carolina, 1960s
    - Highschool Diploma: 34% White v. 18% Black
    - Aptitude Tests: 58% White v. 6% Black
  - **District Courts Verdict:** Claim dismissed!
  - **Supreme Court's Verdict:** Hiring decision illegal if it resulted in disparate impact by a sensitive attribute even if not explicitly determined based on it.
    - Duke Power Co. Guilty!
    - **Findings:** Requirements were to safeguard Duke's long-standing policy of giving job preferences to its white employees.

# Disparate Impact Effect

- Disparate Impact v. Disparate Treatment
  - Unintended Discrimination
- Not always illegal!
- Disparate Impact Effect Today
- 80% Rule: Advocated by US Equal Employment Opportunity Commission
- **Main Goal:** A mathematical definition of Disparate Impact based on 80% Rule

# Contributions

- Introducing the problem to Computer Science world
  - Introducing 80% rule of EEOC as a loss function
  - Showing any decision exhibiting Disparate Impact (DI) can be converted into one where the sensitive attribute leaks
- Certifying lack of DI on a dataset
  - Suggesting a regression algorithm which minimizes the error metric defined in the problem
- Transforming input dataset in a way that:
  - Predictability of the protected attribute is impossible
  - Preserving closeness to the original data distribution
- Detailed empirical study showing effectiveness of the approach

# Disparate Impact Mathematical Definition

- Dataset D=(X, Y, C):
  - X: Protected attribute for example race, sex, religion, etc.
  - Y: Remaining attributes
  - C: Binary class to be predicted for example “will hire”
- D has Disparate Impact if:

$$\frac{\Pr(C = YES|X = 0)}{\Pr(C = YES|X = 1)} \leq \tau = 0.8$$

For positive outcome class “C=YES” and majority protected group “X=1”

# Disparate Impact and Error Rates

- Let's reinterpret 80% rule in terms of more standard statistical measures of quality of a classifier: $\frac{c/(a+c)}{d/(b+d)} \geq 0.8$
- Can't use **Accuracy**!
- Class-Conditioned Error Metrics
  - Sensitivity a.k.a True Positive Rate: $\frac{d}{b+d}$
  - Specificity a.k.a True Negative Rate: $\frac{a}{a+c}$
- Likelihood Ratio Positive:

$$LR_+(C,X) = \frac{sensitivity}{1-specificity} = \frac{d/(b+d)}{c/(a+c)}$$

| Outcome | $X=0$ | $X=1$ |
|---|---|---|
| $C=\text{NO}$ | $a$ | $b$ |
| $C=\text{YES}$ | $c$ | $d$ |

- A data set has Disparate Impact if: $LR_+(C,X) > \frac{1}{\tau} = 1.25$
- Disparate Impact: $\frac{1}{\text{LR}_+(C,X)}$

# Computational Fairness

- Alice, an employer, uses algorithm A to decide who to hire. A takes data set D with protected attribute X and unprotected attributes Y and makes a binary decision C. By law, Alice is not allowed to use X in making decisions, and claims to use only Y. It is Bob's job to verify that on the data D, Alice's algorithm A is not liable for a claim of disparate impact.
- Assumptions:
  - Bob has no access to algorithm A.
  - Alice has good intentions.
- **Idea:** If Bob cannot predict X given Y, A is fair on D.

# Predictability and Disparate Impact

- **Basis:** Procedure that predicts X from Y
- We want to measure the quality of this predictor with two constraints:
  - Optimizable using standard predictors in ML: LR+ fails!
  - Relatable to LR+: Accuracy fails!
- A new error measure: **Balanced Error Rate (BER)**
  - **Definition:** Let f : Y → X be a predictor of X from Y. BER of f on D over the pair (X, Y) is defined as the (unweighted) average class-conditioned error of f.

$$\mathrm{BER}(f(Y), X) = \frac{\Pr[f(Y) = 0 | X = 1] + \Pr[f(Y) = 1 | X = 0]}{2}$$

- **Predictability**
  - **Definition:** X is ε-predictable from Y if there exists a function f : Y → X such that:

$$\mathrm{BER}(f(Y), X) \leq \epsilon$$

# Predictability and Disparate Impact

- **ε-Fairness**
  - **Definition:** D = (X, Y, C) is ε-fair $\text{BER}(f(Y), X) > \epsilon$ if for any classification algorithm f : Y → X
- **Theorem:** A data set is (1/2−β/8)-predictable iff it admits disparate impact, where β is the fraction of elements in the minority class (X = 0) that are selected (C = 1).
- **Proof of Theorem**
  - **Disparate Impact→Predictability**
  - **Predictability→Disparate Impact**

# Proof: Disparate Impact→Predictability

Suppose there exists some function **g:Y→C** such that **LR+(g(y), c)≥1/τ**.
We will create a function **ψ:C→X** such that **BER(ψ(g(y)), x)<ε** for **(x,y)∈D**. Thus the combined predictor **ψ◦g** satisfies the definition of predictability.
Consider the confusion matrix associated with **g**. Set **α= b/(b+d)** and **β= c/(a+c)**. Then we have:

- **LR+(g(y), X)= (1-α)/β**
- **DI(g)= β/(1-α)**

| Prediction | $X=0$ | $X=1$ |
|---|---|---|
| $g(y)=$ NO | a | b |
| $g(y)=$ YES | c | d |

We define the purely biased mapping **ψ:C→X** as **ψ(YES)=1** and **ψ(NO)=0**. Finally, let **φ:Y→X=ψ◦g**.
Confusion matrix for **φ** is identical to the matrix for g and **BER(φ)=(α+β)/2** in terms of this matrix.

| Prediction | $X=0$ | $X=1$ |
|---|---|---|
| $\phi(Y)=0$ | a | b |
| $\phi(Y)=1$ | c | d |

# Proof: Disparate Impact→Predictability

We can now express contours of the **DI** and **BER** functions as curves in the unit square $[0,1]^2$.
Reparameterizing $\pi_1=1-\alpha$ and $\pi_0=\beta$ we can express the error measures:

- **DI(g)** = $\pi_0/\pi_1$ → Any classifier **g** with **DI(g)=δ** can be represented in the $[0,1]^2$ as the line $\pi_1=\pi_0/\delta$.
- **BER(φ)**= $(1+\pi_0-\pi_1)/2$ → Any classifier **φ** with **BER(φ)=ε** can be written as the $\pi_1=\pi_0+1-2\varepsilon$.

Let us now fix the desired **DI** threshold **τ**, corresponding to the line $\pi_1=\pi_0/\tau$. Notice that the region $\{(\pi_0, \pi_1) \mid \pi_1 \geq \pi_0/\tau\}$ is the region where one would make a finding of disparate impact (for **τ=0.8**).

Now given a classification that admits a finding of disparate impact, we can compute **β**.
Consider the point **(β, β/τ)** at which the line $\pi_0=\beta$ intersects the **DI** curve $\pi_1=\pi_0/\tau$. This point lies on the **BER** contour $(1+\beta-\beta/\tau)/2=\varepsilon$, yielding $\varepsilon=1/2-\beta(1/\tau-1)/2$ in particular for the **DI** threshold of **τ=0.8**, the desired **BER** threshold is: $\varepsilon=1/2-\beta/8$

# Proof: Predictability→Disparate Impact

Suppose there is a function **f:Y→X** such that **BER(f(y),x) ≤ ε**.
Let **$\psi^{-1}$:X→ C** be the inverse purely biased mapping i.e. **$\psi^{-1}$(1)=YES** and **$\psi^{-1}$(0)=NO**.
Let **g:Y→C=$\psi^{-1}$∘f**. This gives us **$\pi_1 \geq 1+\pi_0-2\varepsilon$** and therefore:

$$\frac{\pi_0}{\pi_1} \leq \frac{\pi_0}{1+\pi_0-2\epsilon} = 1 - \frac{1-2\epsilon}{\pi_0+1-2\epsilon}$$

Recall that **DI(g)=$\pi_0/\pi_1$** and **$\pi_0=\beta$** yields:

$$\mathrm{DI}(g) \leq 1 - \frac{1-2\epsilon}{\beta+1-2\epsilon} = \tau$$

For **τ=0.8** this gives us **BER** threshold of: **ε=1/2-β/8**

# A Few Observations

- As ε approaches ½ (β tends to 0) the bound tends towards the trivial which introduces a line of attack to evade DI finding:
    - When a company is under investigation for discriminatory hiring practices, it can defeat such a finding by interviewing (but not hiring) a large proportion of applicants from the protected class. This effectively drives β down, and the observation above says that in this setting their discriminatory practices will be harder to detect, because our result can not guarantee that a classifier will have error significantly less than 0.5.
- **Uncertainty due to β:** In practice we only know that the true value of β lies in a range $[\beta_l, \beta_u]$. Since the BER threshold varies monotonically with β, we can merely use $\beta_l$ to obtain a conservative estimate.
- **Uncertainty due to BER estimate:** Suppose that our classifier yields an error that lies in a range [γ, γ′ ]. Again, because of monotonicity, we will obtain an interval of values [τ, τ′] for DI.

# Certifying Lack of DI

- **Goal:** Check whether there is insufficient information to detect a protected attribute from data.
- **Algorithm:**
  - We run a classifier that optimizes **BER** on the given data set, attempting to predict the **X** from **Y**. Suppose the error in this prediction is **ε**.
  - Using the estimate of **β** from the data, we can substitute this into the equation **ε=1/2-β/8** and obtain a threshold **ε′**.
  - If **ε′<ε** then data set is free from disparate impact.

# Some Math Review

- Cumulative Distribution Function (CDF)
  - Probabilities of **X** being smaller than or equal to some value **x**: $\mathbf{F_X(x)=Pr(X \leq x)=p}$
  - This function takes as input **x** and returns values from the **[0,1]** denoted as **p**.
- Quantile Function
  - The inverse of the cumulative distribution function tells you what **x** would make $\mathbf{F_X(x)}$ return some value p: $\mathbf{F^{-1}(p)=x}$

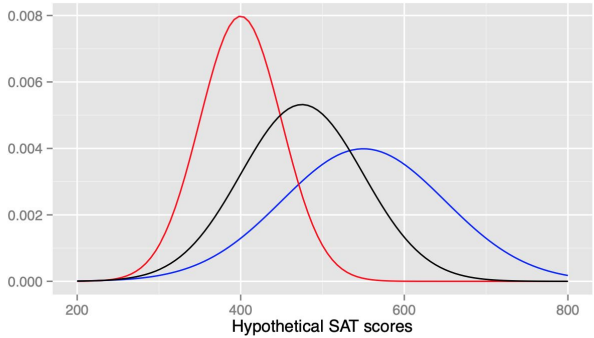
# Removing Disparate Impact

- Alice and Bob taking it to the next stage!
- **Goal**: Construct such a set D′ = (X, Y′, C) such that D′ does not have disparate impact in terms of protected attribute X.
- **Precondition:** It is very important to change the data in such a way that predicting the class is still possible but **how?**
  - We need to preserve the relative per-attribute ordering as follows:
    - Given protected attribute **X** and a single numerical attribute **Y**, let $\mathbf{Y_x = Pr(Y|X = x)}$.
    - Let $\mathbf{F_x : Y_x \rightarrow [0,1]}$ be the cumulative distribution function for values $\mathbf{y \in Y_x}$.
    - Let $\mathbf{F^{-1}_x : [0,1] \rightarrow Y_x}$ be the associated quantile function (i.e. $\mathbf{F^{-1}(1/2)}$ is the value of **y** such that **Pr(Y≥y|X = x) = 1/2)**.
    - We will say that $\mathbf{F_x}$ ranks the values of $\mathbf{Y_x}$.
    - Let **Y′** be the repaired version of **Y** in **D′**. We will say that **D′** strongly preserves rank if for any $\mathbf{y \in Y_x}$ and **x∈X**, its “repaired” counterpart $\mathbf{y' \in Y_x}$ has $\mathbf{F_x(y) = F_x(y')}$.
    - Strongly preserving rank in this way, despite changing the true values of Y, appears to allow Alice’s algorithm to continue choosing stronger (higher ranked) applicants over weaker ones.

# Full Repair

- We define a “median” distribution **A** in terms of its quantile function $\mathbf{F_A^{-1}}$:$\mathbf{F_A^{-1}(u)=median_{x\in X}\ F_X^{-1}(u)}$.
- **Lemma:** Let **A** be a distribution such that $\mathbf{F_A^{-1}(u)=median_{x\in X}\ F_x^{-1}(u)}$. Then **A** is also the distribution minimizing $\sum_{x\in X}\mathbf{d(Y_x,C)}$ over all distributions **C**, where **d(·,·)** is the earth-mover distance on **R**.
- **Algorithm:** Repair algorithm creates **Y′**, such that for all $\mathbf{y\in Y_x}$, the corresponding $\mathbf{y'= F_A^{-1}(F_x(y))}$. The resulting **D′** changes only **Y** while the protected attribute and class remain same as in the original data, thus preserving the ability to predict the class.

- Blue curve: Distribution of SAT scores for X = female, with μ = 550,σ = 100
- Red curve: Distribution of SAT scores for X = male, with μ = 400, σ = 50
- Black curve: Fully repaired data is the distribution in black, with μ = 475, σ =75
- Male score in 95th percentile: 500→625
- Female score in 95th percentile: 750→625




**Repair’s Effect:** Consider **Y** values at some rank **z**. Probability of the occurrence of a data item with attribute $\mathbf{x\in X}$ is the same as the probability of the occurrence of **x** in the full population. This observation gives the intuitive backing for lack of predictability of **X** from **Y′** which means lack of **DI** in **D′**.

# Partial Repair

- Full repair is likely to degrade the ability to classify accurately.
- Partial repair creates a tradeoff between accuracy and fairness of the resulting data.
- Tradeoff can be achieved by simply moving each inverse quantile distribution only part way towards the median distribution.
- We create a different distribution $\mathbf{A_x}$ for each protected value $\mathbf{x \in X}$ and setting $\mathbf{y' = F_{Ax}^{-1}(F_x(y))}$. Consider the ordered set of all **y** at rank **u** in their respective conditional distributions i.e the set $\mathbf{U(u) = \{F_x^{-1}(u) | x \in X\}}$. We can associate with **U** the cumulant function $\mathbf{UF(u,y) = |\{y' \geq y | y \in U(u)\}| / |U(u)|}$ and define the associated quantile function $\mathbf{UF^{-1}(u,\alpha) = y}$ where $\mathbf{UF(u,y) = \alpha}$. We can restate the full repair algorithm in this formulation as follows: for any $\mathbf{(x,y)}$, $\mathbf{y' = UF^{-1}(F_x(y), 1/2)}$.

# Partial Repair: Combinatorial Repair

- Two approaches to partial repair:
  - Combinatorial space: A combinatorial repair
  - Geometric space: A geometric repair
- Combinatorial repair
  - Each item, rather than being moved to the median of its associated distribution, is only moved part of the way there, with the amount moved being proportional (in rank) to its distance from the median.
  - Fix an **x** and consider any pair **(x, y)**. Let **$r = F_x(y)$** be the rank of **y** conditioned on **X=x**. Suppose that in the set **U(r)** the rank of **y** is **ρ**.Then we replace **y** by **$y' \in U(r)$** whose rank in **U(r)** is **$\rho' = \lfloor (1-\lambda)\rho + \lambda/2 \rfloor$**. Formally, **$y' = UF^{-1}(r, \rho')$**. We call the resulting data set **$D'_\lambda$** where **$\lambda \in [0,1]$** is the amount of repair desired.
  - **Pros:** Easy to implement
  - **Cons:** Not satisfying the property of strong rank preservation (affecting quality of resulting data but not fairness properties of the repair).

# Partial Repair: Geometric Repair

- Geometric repair
  - Combinatorial repair does not admit functional interpretation as an optimization of a certain distance function i.e. for λ=½ the modified distribution Y′ is not equidistant between the unrepaired distributions and the full repair.
  - Geometric repair does have this property! The intuition is that rather than doing a linear interpolation in **rank space** between the original item and the fully repaired value, it does a linear interpolation in the original **data space**.
  - Let $\mathbf{F_A}$ be the cumulative distribution associated with **A**. Given a conditional distribution $\mathbf{F_x(y)}$, its **λ-partial** repair is given by: $\mathbf{F'_x{}^{-1}(\alpha)=(1-\lambda)F_x^{-1}(\alpha)+(\lambda)F_A^{-1}(\alpha)}$
  - Linear interpolation allows us to connect this repair to the underlying earthmover distance between repaired and unrepaired distributions. In particular for any **x**, $\mathbf{d(Y_x,Y'_x) = \lambda d(Y_x,Y_A)}$ where $\mathbf{Y_A}$ is the distribution on **Y** in the full repair, and $\mathbf{Y_x}$ is the **λ-partial** repair. Moreover, the repair strongly preserves rank (by observing that the repair is a linear interpolation between the original data and the full repair).

# Fairness/Utility Tradeoff

- Partial repair is desired because increasing fairness may result in loss of utility.
- We make this intuition precise by the definition of **Utility**:
  - **Utility:** The utility of a classifier $\mathbf{g'_\lambda : Y' \to C}$ with respect to some partially repaired data set $\mathbf{D'_\lambda}$ is: $\mathbf{\gamma(g'_\lambda, D_\lambda') = 1 - BER(g'_\lambda(y'), c)}$

# Auditing Black-box Models

- Now let's use these algorithmic fairness ideas to study how features influence the the outcome of model without knowing how the models work.
- Direct v. Indirect Influence
  - Direct: replace the feature by random noise and test how model accuracy deteriorates.
  - Indirect: Classic case of Redlining! Race has indirect influence via Zip code.
    - Remove Race? Zip code still generates signal!
    - Remove Race and Zip code? Eliminates other task-specific value of Zip code!
    - Perturbing feature?
      - Randomly
        - Can also remove useful task-related information in proxy features that would degrade the quality of classification
        - Prevents us from cleanly quantifying the relative effect of the feature being perturbed on related proxy variables
      - Obscuring
        - In a directed and deterministic manner, with minimal change organized around the question: "Can we predict the value of feature j from the remaining features?"

# Computing Influence

- Let **f:X** $\rightarrow$**Y** be a classifier, and let **(X,Y)={($X_i$,$y_i$)}** be a set of examples. Let $\mathbf{X^{(i)} = (x_{1i}, x_{2i},..., x_{ni})}$ denote the column corresponding to the $i^{th}$ feature**.**
  **ε-obscure:** We define $\mathbf{X\backslash_{\varepsilon}X_i}$ as the ε-obscure version of **X** with respect to feature $\mathbf{X_i}$ if $\mathbf{X^{(i)}}$ cannot be predicted from $\mathbf{X\backslash_{\varepsilon}X_i}$.
- **Indirect influence:** The indirect influence **II(i)** of a feature **i** on a classifier **f** applied to data **(X,Y)** is the difference in accuracy when **f** is run on **X** versus when it is run on $\mathbf{X\backslash_{\varepsilon}X_i}$:

$$\mathbf{II(i)=acc(X,Y,f)-acc(X\backslash_{\varepsilon}X_i,Y,f)}$$

- **Gradient Feature Audit (GFA): An algorithm to estimate indirect influence**
  For each feature:
  1) Remove indirect influence of feature on other features in data **(How?)**
  2) Run model on modified test data
  3) Calculate influence using II(i)

- GFA works one feature at a time cannot guarantee that all influence can be removed.

# That **How?** Question

- This process differs according to the feature types:
  - O Feature to be **removed** is **Categorical** and W feature to be **obscured** is **Numerical**
    - Modify the distribution of **W** by “moving” values so as to mimic the median distribution **A**. By Doing so **O** is maximally obscured and **W** minimally changed. The reason is **A** also minimizes the function $\Sigma_{x \in O} d(W_x, A)$ where $d(\cdot,\cdot)$ is the earthmover distance between the distributions using $\ell_2$ as the base metric.
    - Just works if if features to be obscured and removed are numerical and categorical respectively
  - Removing numerical features
    - Remove higher order bits of a number.
    - Bin the numerical feature and use the bins as categorical labels and use previous approach.
    - Bins are chosen using the Freedman-Diaconis rule for choosing histogram bin sizes.
  - Obscuring categorical features
    - Since the procedure relies on being able to compute cumulative density functions for the feature **W** being obscured, if it is categorical, we no longer have an ordered domain on which to define the cumulative distributions $F_w$.
    - However, we do have a base metric: the exact metric **1** where $1(x,w) = 1 \leftrightarrow x = w$.
    - We can therefore define **A** as before, as the distribution minimizing the function $\Sigma_{x \in O} d(W_x, A)$ with respect to metric **1** and **A** can be found by taking a **component-wise median** for each value **w**.

# Experiments

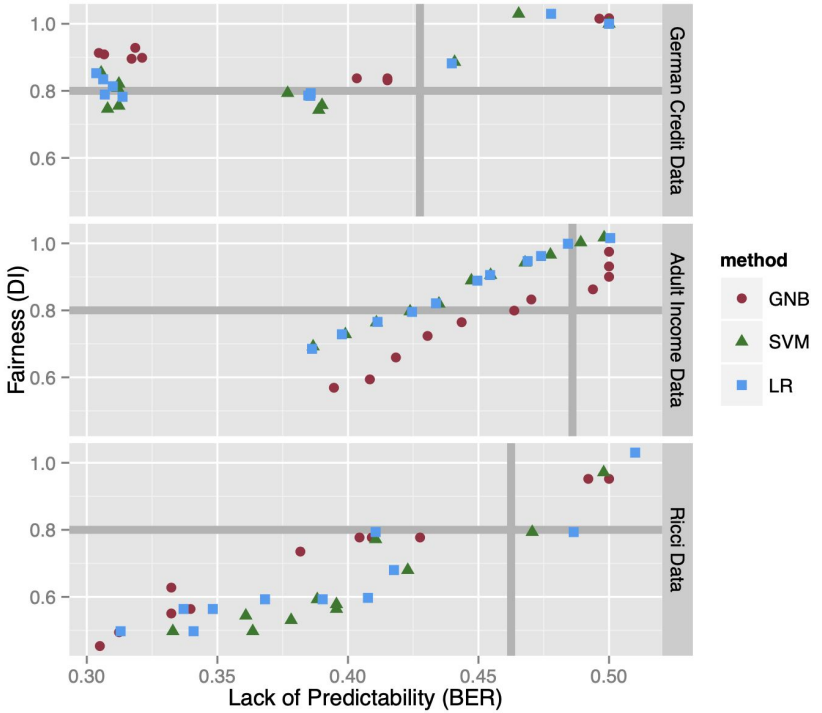
- Evaluate certification and repair algorithms' fairness/utility tradeoff
- On three data sets:
  - Ricci data set:
    - 118 instances
    - Features: Firefighter exam promotion taken, Oral section score, Written section score, Combined score, Race (group Black and Hispanic into a single non-white category)
    - Test takers promoted have a score of at least 70%
  - German credit cards
    - 1000 instances
    - 20 attributes, categorized GOOD/BAD, Protected attribute Age (discretized into two categories YOUNG/OLD at age 25)
  - Adult income
    - 48,842 instances
    - 14 attributes, categorized more or less than $50K annually,  Protected attribute Gender
- 21 versions of the data, the original data set plus ($\lambda \in [0,1]$ at increments of 0.1) 10 partially or fully repaired attribute sets for each of the combinatorial and geometric partial repairs.
- Preprocessing:
  - Remove all protected attributes from Y
  - Remove all unordered categorical features and ordered categories converted to integers
  - Scale to [0,1]
- 3 Classifiers used for measureing discrimination: LR, SVM, GNB

# Experiments

- **Repair details**
  - The repair procedure requires a ranking of each attribute
  - The numeric and categorical attributes were ordered and then quantiles were used as the ranks.
  - Since the repair assumes that there is a point at each quantile value in each protected class, the quantiles were determined in the following way:
    - For each attribute, the protected class with the smallest number of members was determined.
    - This size determined how many quantile buckets to create.
    - The other protected classes were then appropriately divided into the same number of quantile buckets, with the median value in each bucket chosen as a representative value for that quantile.
      - Each quantile value in the fully repaired version is the median of the representative values for that quantile.
      - The combinatorial partial repair determines all valid values for an attribute and moves the original data part way to the fully repaired data within this space.
      - The geometric repair assumes all numeric values are allowed for the partial repair.

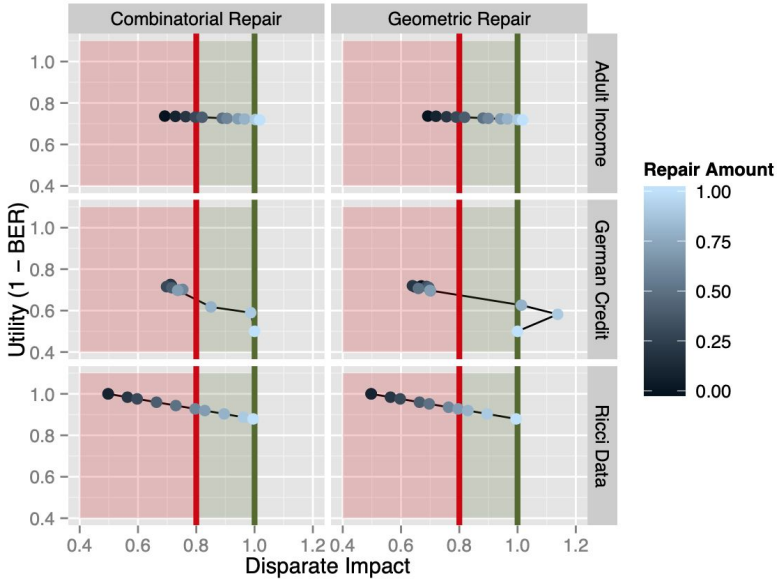
# Certification

- We predict the protected attribute from the remaining attributes. BER is compared to DI(g) where g:Y→C
- BER threshold ε=1/2−β/8
- **(Bottom-right quadrant) No False Positives!** few due to error in β measure
- **(Upper-left quadrant) Some False Negatives!?** However, certification algorithm guarantees lack of disparate impact over any classifier, so these are not false negatives in the traditional sense. In fact, when a single data set is considered over all classifiers, we see that all such data sets below the BER threshold have some classifier that has DI close to or below τ=0.8.



# Fairness/Utility Tradeoff

- Each unrepaired data set begins with DI < 0.8 meaning it fails the 80% rule, and we are able to repair it to a legal value.
- Drop in utility when fully repaired is different in each data set and this difference in decay is inherent to the class decisions in the data set.
- **DI>1:** Since DI is calculated with respect to fixed majority and minority classes, this happens when the classifier has given a good outcome to proportionally more minority than majority class members and should be considered unfair to the majority class.



# Discussion

- Mathematical definition of disparate impact
- Fairness/Utility tradeoff
- Just considering binary class attributed, how about ethnicity?
    - A more general treatment of joint discrimination among multiple classes
- Multiple proxy attributes
    - Repair each attribute individually

# Questions?