

# BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2024

## Lecture 5: Bias and fairness in biomedical AI



**HARVARD**  
MEDICAL SCHOOL

Marinka Zitnik  
marinka@hms.harvard.edu

# Responses to L4 Quick Check

Describe a scenario in which a predictive model is created using a biomedical dataset and the LIME explainability method is used to analyze its behavior. What can be expected from the LIME explanations?

We could develop a predictive model on a dataset of echocardiogram results to predict risk of future heart failure. The dataset would include numeric features such as myocardial wall thickness, ejection fraction, chamber size, etc. After training the prediction model, we could then use LIME by sampling points around a given example of interest (perturbing the sample data point to create new artificial samples) and seeing how predictions change; then generating a linear model that is more easily explainable on a local basis. This resulting model would give us weights (and signs of weights) of each feature to understand how they contribute to predictions at a local level for a given sample.

Scenario: predictive model for progression, regression or stability of lung disease using a biomedical dataset including multiple sociodemographic and clinical data. LIME is used post-hoc to explain the model's predictions of the 3 categories and the expected output is as such: a patient is predicted to have progression, LIME explainability model would produce a linear model explaining which features about the patient are most relevant to predict progression. For example  $f(x) = \text{age} + \text{duration of underlying disease} + \text{medication response currently}$ .

# Responses to L4 Quick Check

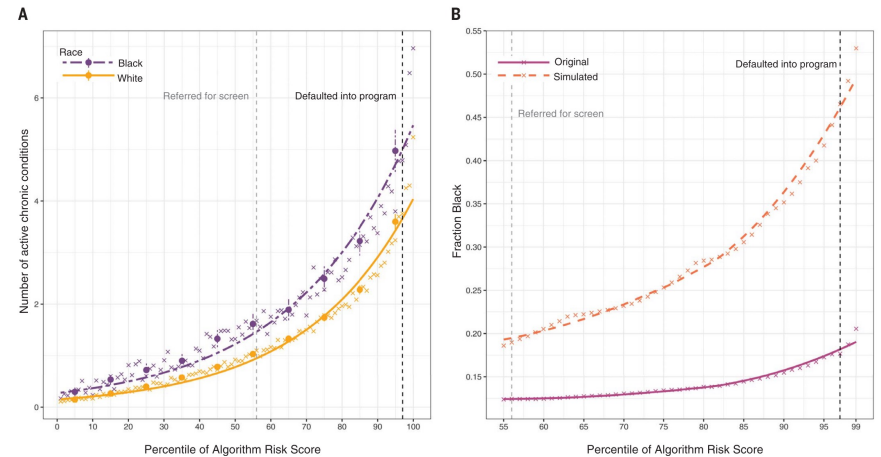
Describe a scenario in which a predictive model is created using a biomedical dataset and the Integrated Gradients explainability method is used to analyze its behavior. What can be expected from the IG explanations?

The Integrated Gradients explainability method is suitable for the images deep learning networks. And a scenario that we can diagnose the disease from chest X-rays is that the research team at the medical imaging center developing the deep learning model to diagnose various conditions from chest X-rays and use the pre-trained inception neural network, fine-tuned on the large dataset of annotated chest X-rays. We can expect from the integrated gradients is that the IG will provide the visual explanation by highlighting the regions in the chest X-ray images that are important for the model's prediction, for example, IG may highlight the areas that show lung consolidation

Scenario: predictive model for pulmonary fibrosis presence on a future CT (in one year) based on baseline CT using only the baseline CT image and no other clinical data [I think this is going to be very hard but, imagining something that could be very helpful]. IG is used post-hoc to explain which areas/pixels of the image the model is prioritizing to predict PF in one year. The IG explainability model would produce the aggregated overlay image showing highlighted parts of the CT scan the model is prioritizing on the baseline image to predict the 1 year image as fibrotic.

# Adopting AI in high-stakes areas

- Healthcare
  - Genomic medicine
  - Public health policy
  - Child welfare
- 
- Criminal risk assessment
  - Surveillance
  - Financial lending
  - Hiring



**Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race.** (A) Mean number of chronic conditions by race, plotted against algorithm risk score. (B) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated"): at each threshold of risk, defined at a given percentile on the x axis, healthier Whites above the threshold are

replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

Obermeyer et al. *Science* 2019



# Three problematic examples

## 1. High-risk Healthcare Management

- Commercial prediction models are used by large health systems to identify and help patients with complex health needs.
- These models can exhibit significant bias: At a given risk score, black patients are considerably sicker than white patients
- The bias arises because the algorithm predicts health care costs rather than illness

## 2. Criminal Risk Assessment Tools

- Defendants are assigned scores that predict the risk of re-committing crimes
- These scores inform decisions about bail, sentencing, and parole.
- Some tools have been biased against black defendants

## 3. Face Recognition Systems

- Surveillance and self-driving cars
- Systems can perform poorly for populations that are not well represented in training dataset



## The COMPAS debate

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

*Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)*

# Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

# COMPAS

- Correctional Offender Management Profiling for Alternative Sanctions
- Used in prisons across country: AZ, CO, DL, KY, LA, OK, VA, WA, WI
- “Evaluation of a defendant’s rehabilitation needs”
- Recidivism = likelihood of criminal to reoffend

# COMPAS (continued)

“Our analysis of Northpointe’s tool, called COMPAS, found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.”



# What are protected classes?

- **Protected classes in the US:**
  - Race
  - Sex
  - Religion
  - National origin
  - Citizenship
  - Pregnancy
  - Disability status
  - Genetic information
- **Regulated domains in the US:**
  - Credit (Equal Credit Opportunity Act)
  - Education (Civil Rights Act of 1964; Education Amend. of 1972)
  - Employment (Civil Rights Act of 1964)
  - Housing (Fair Housing Act)

# Fairness in ML

- **It does not necessarily mean being malicious:** Bias can occur even when everyone, from data generators to engineers to clinical staff, has the best intentions
- **It is not one and done:** Just because an algorithm has no bias now does not mean it has no potential bias later
- **It is not new:** Researchers have raised concerns about it over the last 50 years
  
- It is defined in many ways, for example, **disparate treatment** or **impact of algorithm**
- It can be a **culmination of a flawed system**
  - Biases in data collection processes
  - Biases in algorithmic design
  - Bias in model implementation/deployment
- It is the **vigilance** of how technology can amplify/create bias

# Outline for today's class

1. Quantitative definitions of fairness in AI
2. Framework for fair AI
3. Algorithmic fairness criteria
  - Individual fairness
  - Group fairness
4. Auditing AI systems
  - Auditing input data
  - Auditing ML model



# Part I

## Quantitative definitions of fairness in AI

# How to define fairness in ML?

- Fairness through unawareness
- Group fairness
- Calibration
- Error rate balance
- Representational fairness
- Counterfactual fairness
- Individual fairness

# Fairness through unawareness

- **Idea:** Don't record protected attributes, and don't use them in your algorithm
  - Predict risk  $Y$  from features  $X$  and group  $S$  using  $P(\hat{Y} = Y|X)$  instead of  $P(\hat{Y} = Y|X, S)$
- **Pros:** Guaranteed to not be making a judgement on protected attribute
- **Cons:** Other proxies may still be included in a “race-blind” setting, e.g. zip code or conditions



# Group fairness

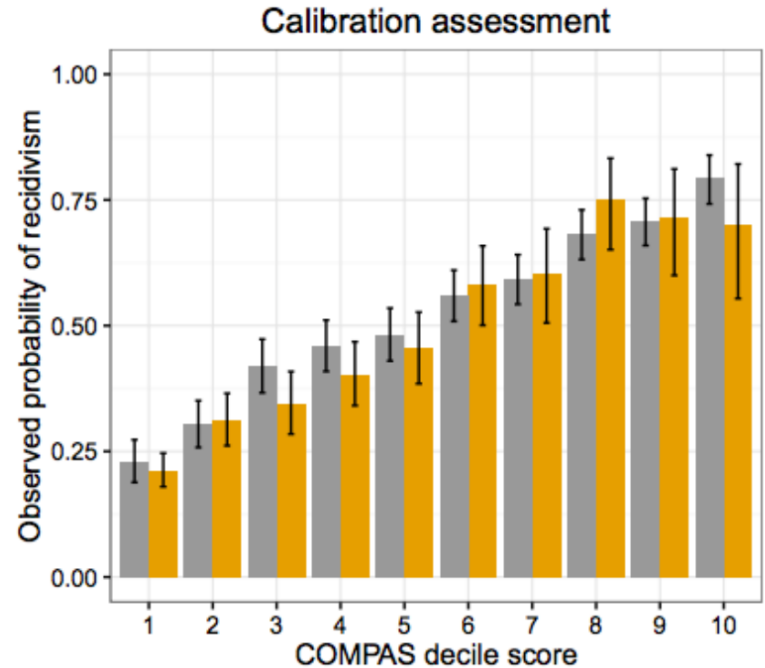
- **Idea:** Require prediction rate be the same across protected groups
  - E.g. “20% of the resources should go to the group that has 20% of population”
- Predict risk  $Y$  from features  $X$  and group  $S$  such that  $P(\hat{Y} = 1|S = 1) = P(\hat{Y} = 1|S = 0)$
- **Pros:** Literally treats each race equally
- **Cons:**
  - Too strong: Groups might have different base rates. Then, even a perfect classifier wouldn't qualify as “fair”
  - Too weak: Doesn't control error rate. Could be perfectly biased (correct for  $S = 0$  and wrong for  $S = 1$ ) and still satisfy

# Calibration

- **Idea:** Same positive predictive value across groups
- Predict  $Y$  from features  $X$  and group  $S$  with score  $R$ :

$$\frac{P(Y = 1 | R = r, A = 1)}{P(Y = 1 | R = r, A = 0)} =$$

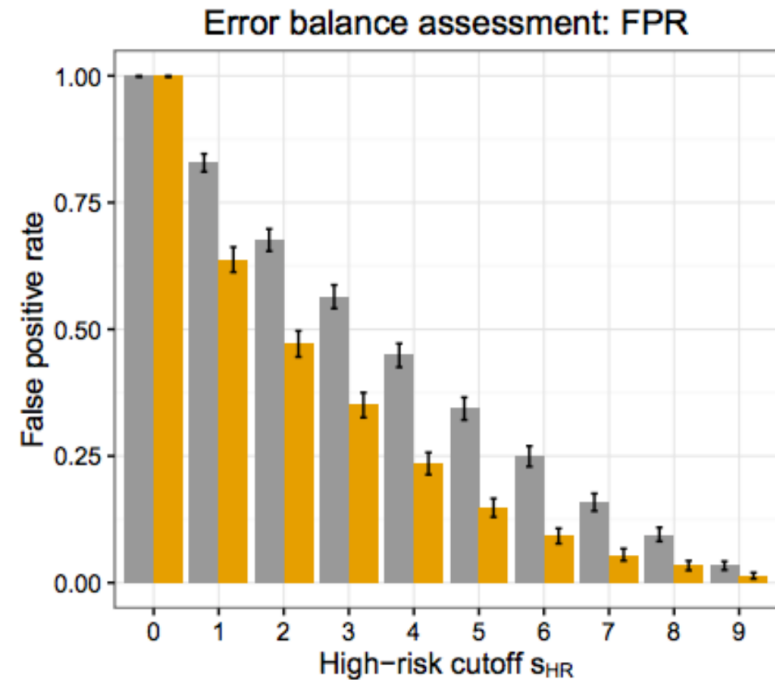
- **Pros:** “Equally right across groups”
- **Cons:** Not compatible with error rate balance (next slide)





# Error rate balance

- **Idea:** Equal false positive rates (FPR) across groups  
$$P(\hat{Y} = 1 | Y = 0, S = 1)$$
$$= P(\hat{Y} = 1 | Y = 0, S = 0)$$
- **Pros:** “Equally wrong across groups”
- **Cons:** Incompatible with calibration and false negative rates (FNR), could dilute with easy cases



# Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg \*

Sendhil Mullainathan †

Manish Raghavan ‡

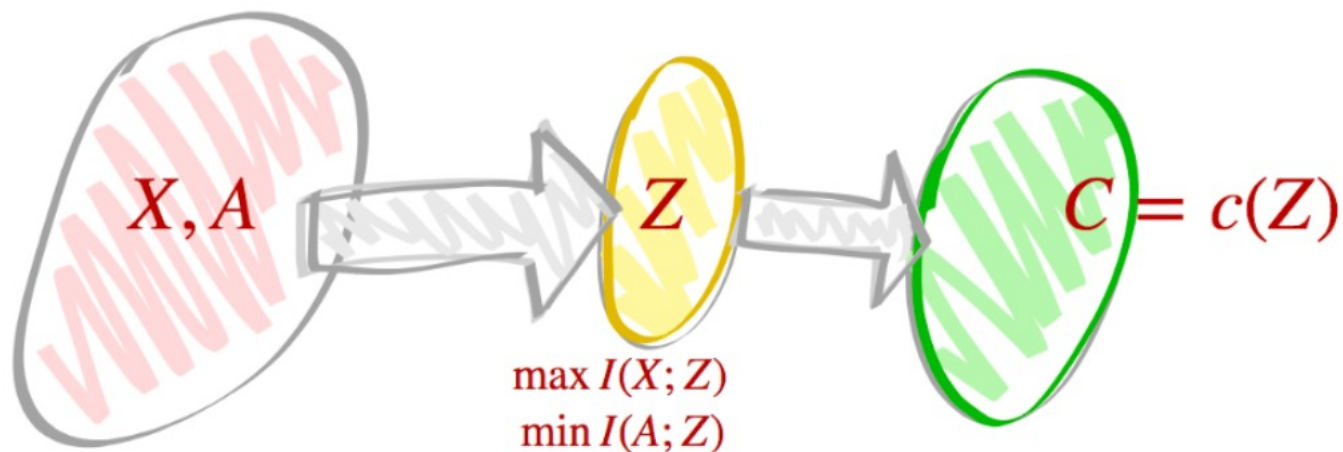
## Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

framework for thinking about the trade-offs between them.

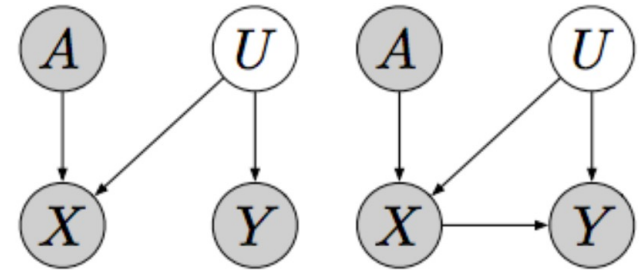
# Representational fairness

- **Idea:** Transform input feature vectors in “fair representations  $Z$  to minimize group information
- **Pros:** Reduce information given to model while still keeping important information
- **Cons:** Trade-off between accuracy and fairness



# Counterfactual fairness

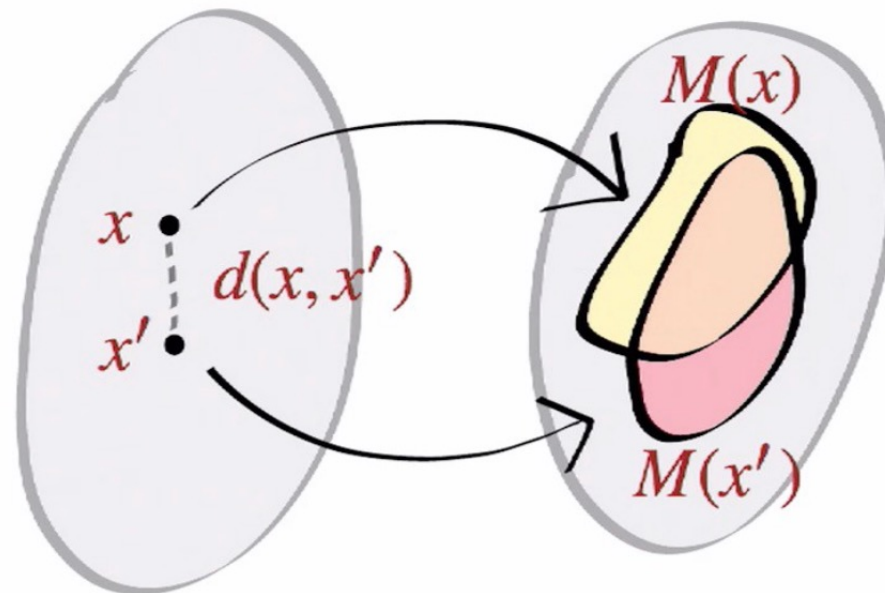
- **Idea:** Group  $A$  should not cause prediction  $\hat{Y}$
- **Pros:** Can model explicit dependencies between features
- **Cons:**
  - Dependency graphs may not represent real world
  - Inference assumes observed confounders



$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) \\ = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$$

# Individual fairness

- **Idea:** Similar individuals should be treated similarly
- **Pros:** Can model heterogeneity within each group
- **Cons:** Notion of “similar” is hard to define mathematically, especially in high dimensions



# How to define “fairness” in ML?

- ~~■ Fairness through unawareness~~ Not useful
  - Group fairness
  - Calibration
  - Error rate balance
  - Representational fairness
  - Counterfactual fairness
  - Individual fairness
- Established strategies
- Ongoing and cutting-edge research

# One fairness definition or one framework

## 21 Fairness Definitions and Their Politics. Arvind Narayanan.

ACM Conference on Fairness, Accountability, and Transparency Tutorial (2018)

S. Mitchell, E. Potash, and S. Barocas (2018)  
P. Gajane and M. Pechenizkiy (2018)  
S. Verma and J. Rubin (2018)

Differences/connections between fairness definitions are difficult to grasp.

We lack common language/framework.

*“Nobody has found a definition which is widely agreed as a good definition of fairness in the same way we have for, say, the security of a random number generator.”*

*“There are a number of definitions and research groups are not on the same page when it comes to the definition of fairness.”*

*“The search for one true definition is not a fruitful direction, as technical considerations cannot adjudicate moral debates.”*

# Outline for today's class

- ✓ 1. Quantitative definitions of fairness in AI
2. Framework for fair AI
3. Algorithmic fairness criteria
  - Individual fairness
  - Group fairness
4. Auditing AI systems
  - Auditing input data
  - Auditing ML model





# Part II

# Framework for fair AI

## Data Regulator

Determines fairness criteria, determines data source(s), audits results



AUTHORITY

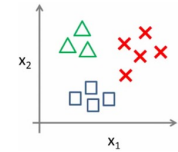
01

03

02

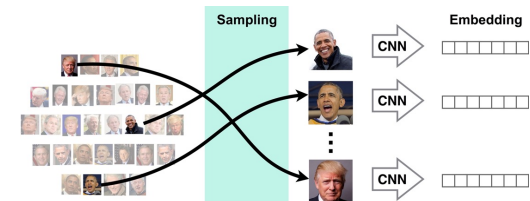
## Data User

Computes ML model given sanitized data



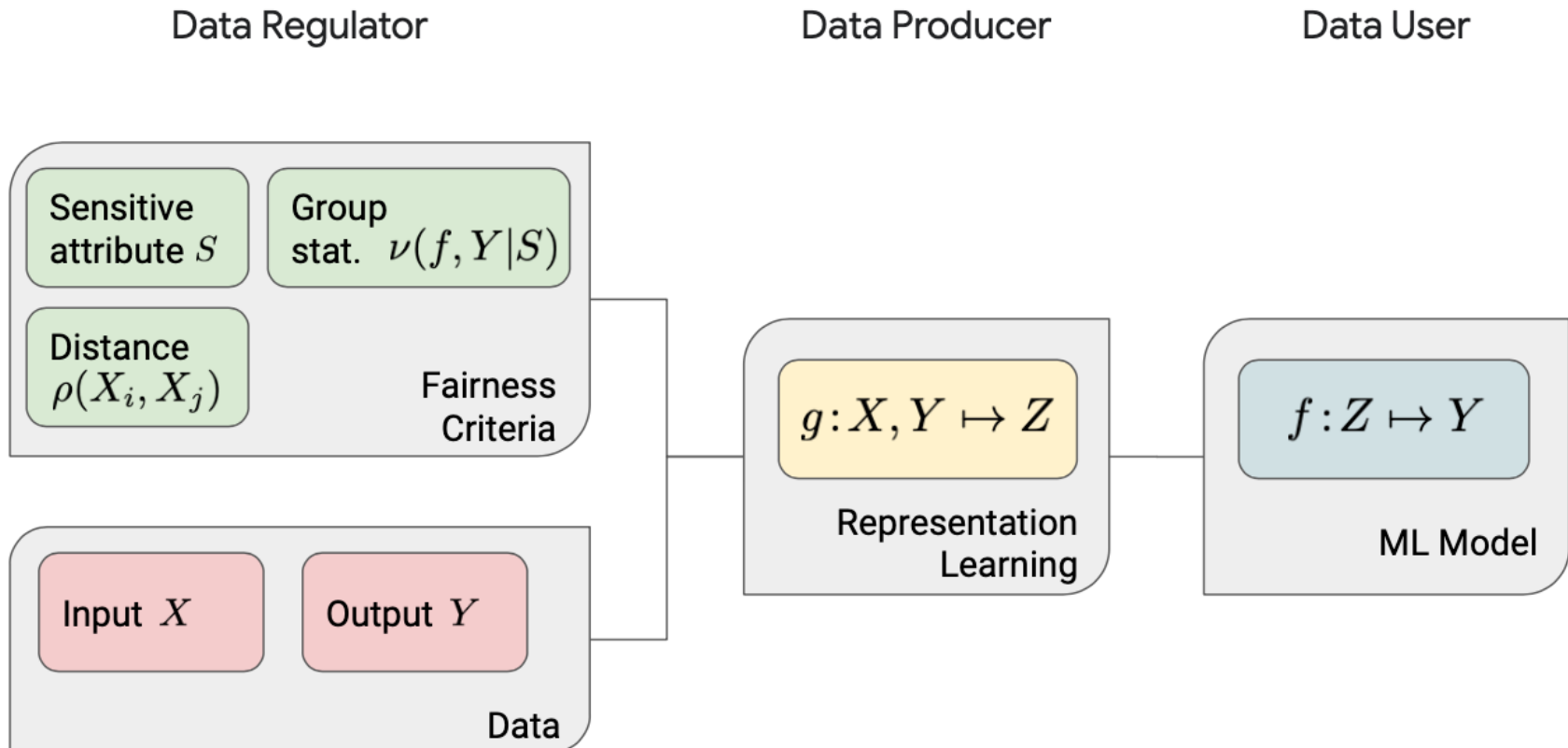
## Data Producer

Computes the fair representation given data regulator criteria



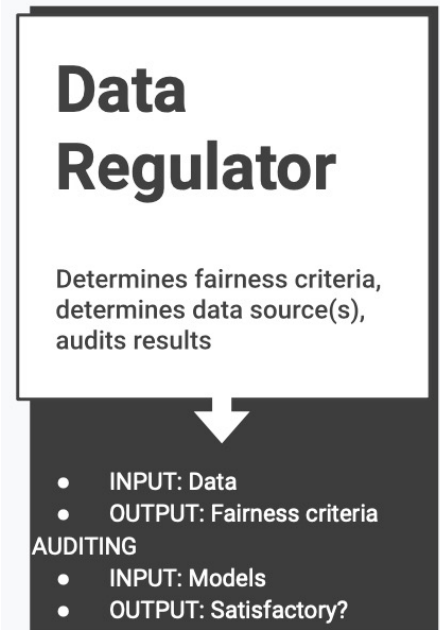
# Framework for fair AI/ML

- **Data regulator:** determines fairness measures, audits results
- **Data producer:** creates “fair” feature vectors (i.e., “fair” representations)
- **Data user:** agnostically trains an ML model using “fair” feature vectors



# Roles of different parties

- **Data regulator** determines which fairness criteria to use, and (optionally) audits the results
- When training:
  - Input: interaction with users/experts/judges/policy to determine fairness criteria
  - Output: fairness criteria
- When auditing the ML model:
  - Input (for auditing the **data producer**):
    - “Fair” representations
  - Input (for auditing the **data user**):
    - Data and model predictions
  - Output:
    - Are fairness criteria satisfied?



# How to achieve fairness?

- **Post-processing:** Post-process the model outputs  
Doherty et al. (2012), Feldman (2015), Hardt et al. (2016), Kusner et al. (2018), Jiang et al. (2019)
- **Pre-processing:** Pre-process the data to remove bias, or extract representations that do not contain sensitive information during training  
Kamiran and Calder (2012), Zemel et al. (2013), Feldman et al. (2015), Fish et al. (2015), Louizos et al. (2016), Lum and Johndrow (2016), Adler et al. (2016), Edwards and Storkey (2016)
- **In-processing:** Enforce fairness notions by imposing constraints into the optimization, or by using an adversary  
Goh et al. (2016), Corbett-Davies et al. (2017), Agarwal et al. (2018), Cotter et al. (2018), Komiyama et al. (2018), Narasimhan (2018), Wu et al. (2018), Zhang et al. (2018), Jiang et al. (2019)

# Outline for today's class

- ✓ 1. Quantitative definitions of fairness in AI
- ✓ 2. Framework for fair AI
3. Algorithmic fairness criteria
  - Individual fairness
  - Group fairness
4. Auditing AI systems
  - Auditing input data
  - Auditing ML model



# Part III

# Algorithmic fairness criteria

# Algorithmic fairness criteria

1) Individual Fairness



2) Group Fairness



# Individual fairness: Similar individuals should be treated similarly



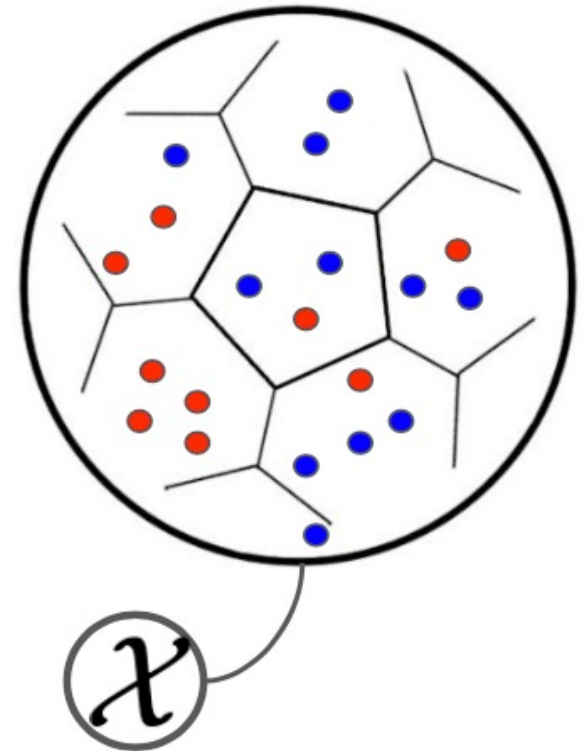
Problem: Pairs of **similar individuals playing the same sport** classified differently. The model is biased against individuals with certain characteristics

Shown are pairs of pictures (columns) sampled over the Internet along with their prediction by a ResNet-10.

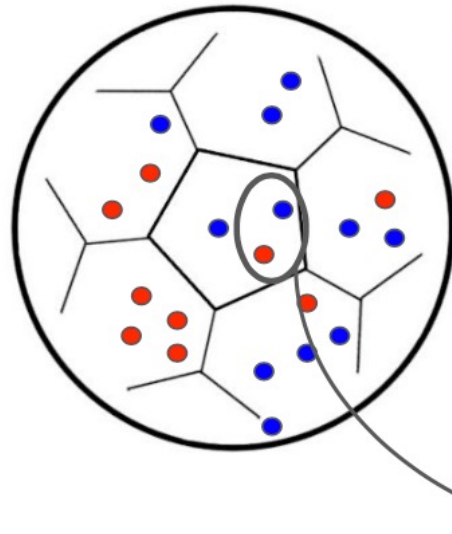
Explore biases of a neural net by analyzing the distance of a sample to the decision boundary using adversarial samples. The distance to the decision boundary is closely related to the magnitude of the perturbation necessary to make a sample cross it.

# Individual fairness: Similar individuals should be treated similarly

- **Data Regulator:** Which individuals are similar? equiv., which individuals should be treated similarly?
- One approach:
  - Define a **partition** of the space into disjoint cells such that similar individuals are in the same cell
  - Individuals in the **same cell** should be **treated similarly** even if they are apparently different (e.g., dots with different colored attributes)



# Individual fairness: Similar individuals should be treated similarly



**Data Regulator:** Which individuals are similar?  
quiv., which individuals should be treated similarly?

An algorithm  $\mathcal{A}_{\mathcal{D}}$  is  $(B, \epsilon(\mathcal{D}))$ -individually fair if  $\mathcal{X}$  can be partitioned into  $B$  disjoint subsets denoted  $\{C_i\}_{i=1}^B$  such that  $\forall x_1 \in \mathcal{X}$ :

$$x_1, x_2 \in C_i \Rightarrow |l(\mathcal{A}_{\mathcal{D}}, x_1) - l(\mathcal{A}_{\mathcal{D}}, x_2)| \leq \epsilon(\mathcal{D})$$

**Remark:** Individual fairness implies **algorithmic robustness** (c.f. Xu & Mannor '11)

# Individual fairness: Pros and Cons


## ■ Advantages:

- Intuitive and **easy to explain** to data producers (and non-experts)
- Individual fairness **implies generalization** (c.f. Xu & Mannor, '12)
- Individual fairness **implies statistical parity** given regularity conditions (Dwork et al., '12)

## ■ Challenges:

- Regulator **must provide a metric** or a set of examples to be treated similarly
- Constructing a metric requires **significant domain expertise and human insight**
- Fairness of the representation **heavily depends on the quality of the metric** chosen by the regulator
- Optimizing and measuring individual fairness is **generally more computationally expensive** than other measures

# Algorithmic fairness criteria

 1) Individual Fairness

 2) Group Fairness

# Group fairness: Similar classifier statistics across groups

- **Regulator:** Which statistic  $v(f, Y|S)$  should be equalized across groups  $S$ ?

- Typical **fairness measure** is a of the ML model performance:
  - Eq. of opportunity (Hardt et al., '16)

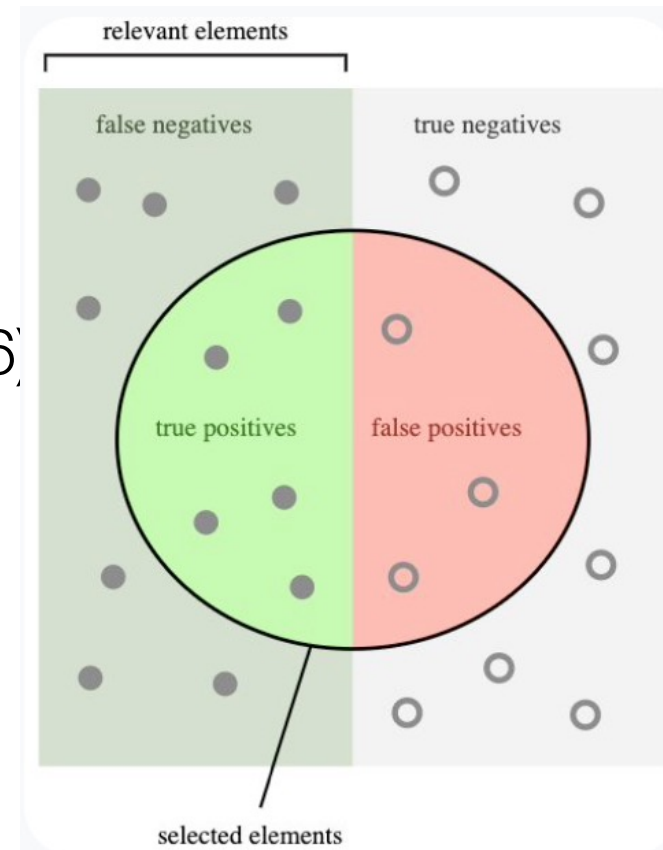
$$TP_S = P(Y = 1, f = 1|S)$$

- Equalized odds (Hardt et al., '16)

$$\{TP_S; FP_S\}$$

- Statistical parity (Dwork et al., '12)

$$TP_S + FP_S = P(f(Z) = 1|S)$$



# Details #1: Statistical Parity

- Statistical parity is a popular measure of group fairness
- **Setup:**
  - Population is a set  $X$
  - Subset  $S \subset X$  that is a “protected” subset of the population
- **Example:**
  - $X$  is people
  - $S$  is people who dye their hair blue
  - We are afraid that banks give fewer loans to the blues because of hair-colorism, despite blue-haired people being just as creditworthy as the general population on average

# Details #2: Statistical parity

- **Assumption:** There is some distribution  $D$  over  $X$  which represents the probability that any individual will be drawn for evaluation
- Example:
  - Some people will have no reason to apply for a loan (maybe they're filthy rich, or don't like homes, cars, or expensive colleges)
  - $D$  takes that into account
  - Generally, we impose no restrictions on  $D$ , and the definition of fairness will work no matter what  $D$  is



# Details #3: Statistical parity

- Classifier  $f: X \rightarrow \{0,1\}$  gives labels to  $X$ 
  - When given a person  $x$  as input  $f(x) = 1$  if  $x$  gets a loan and 0 otherwise
- **Statistical imparity** of  $f$  on  $S$  with respect to  $X, D$ :

$$\text{imparity}_f(X, S, D) = \underbrace{P(f(x) = 1 | x \in S^c)}_{\substack{\text{Probability that a random} \\ \text{individual from the complement} \\ S^c \text{ is labeled 1}}} - \underbrace{P(f(x) = 1 | x \in S)}_{\substack{\text{Probability that a random} \\ \text{individual drawn from } S \\ \text{is labeled 1}}}$$

- This is the statistical equivalent of **adverse impact**
  - It measures the difference that the majority and protected classes get a particular outcome

# Details #4: Statistical parity

- Statistical imparity measures the difference that the majority and protected classes get a certain outcome
- When the difference is small, the classifier has **statistical parity**, it conforms to this notion of fairness
- **Definition:** ML model  $f: X \rightarrow \{0,1\}$  achieves statistical parity on  $D$  with respect to  $S$  up to bias  $\epsilon$  if  $|\text{imparity}_f(X, S, D)| < \epsilon$
- If  $f$  achieves statistical parity, it treats the general population statistically similarly as the protected class
  - If 30% of normal-hair-colored people get loans, statistical parity requires roughly 30% of blue also get loans

# Group fairness: Pros and Cons

## ■ Advantages:

- Efficient to compute, measure and enforce for data producer and regulator
- Often easier to explain to policy-makers (as in terms of population behavior)

## ■ Challenges:

- Data regulator must determine which classifier statistic(s) to equalize
- Fairness of the representation depends on the quality of the fairness metric chosen by the regulator
- Group fairness can lead to (more) violated individual fairness, e.g., intersectionality
- It can lead to fairness gerrymandering (Kearns et. al., '18), and other issues (McNamara et. al., '19)

# Algorithmic fairness criteria

✓ 1) Individual Fairness

✓ 2) Group Fairness

# Data regulator: Measures (un-)fairness

- Regulator must choose how to measure (un-)fairness:
  - For individual fairness: must choose the distance metric
  - For group fairness: must choose the classifier statistic to equalize
- However, remember that there are **no magic metrics**:
  - **Measurement 101**: all measures have **blind spots**
  - *“When a measure becomes a target, it ceases to be a good measure”*
- For ML, we generally specify all measures apriori and optimize them
  - However, **all** metrics will have **failure cases**, i.e., unusual situations with non-ideal behavior
- One productive approach is to select measures that best capture tradeoffs relevant to the context

# Outline for today's class

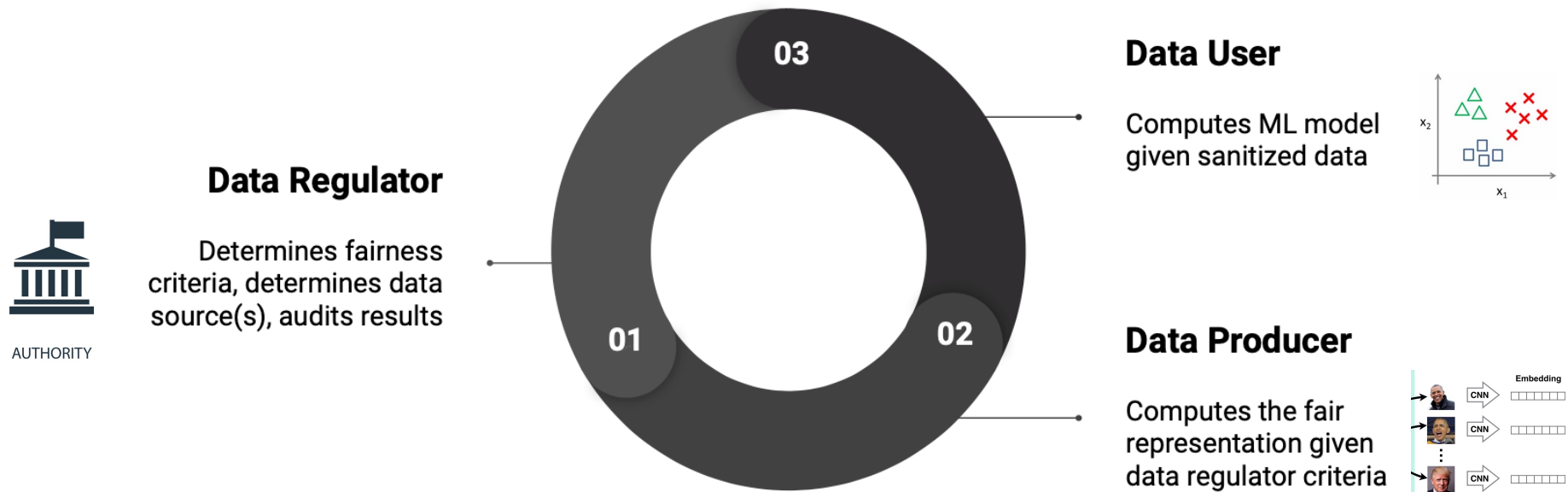
- ✓ 1. Quantitative definitions of fairness in AI
- ✓ 2. Framework for fair AI
- ✓ 3. Algorithmic fairness criteria
  - Individual fairness
  - Group fairness
4. Auditing AI systems
  - Auditing input data
  - Auditing ML model



# Part IV

# Auditing AI systems

# Recall: Framework for fair AI



- How to ensure that our implemented ML model is fair?
- Data regulator (e.g., health department, office for sentencing and incarceration) needs to audit the ML model!

**Who should be audited to ensure that ML predictions are fair and unbiased?**

- a) **Data producer:** The regulator audits input data representations for fairness
- b) **Data user:** The regulator audits the final ML model for fairness

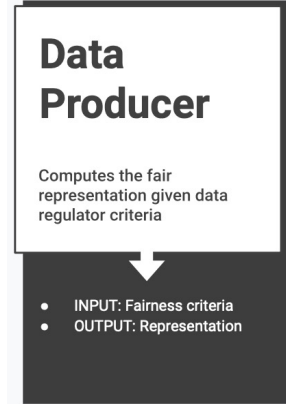


# Who should be audited by the data regulator to ensure fairness?

- Key task of the **data regulator** is to **audit** the learning system (e.g., Madras et al., '18)
- For complex label-dependent settings, or for an adversarial **data user**, the **data regulator** must audit the final model, i.e., the **data user**
- The most efficient approach is to audit input data representations, i.e., the **data producer**

Next: How to produce fair input data representations?

# How to compute fair representations?



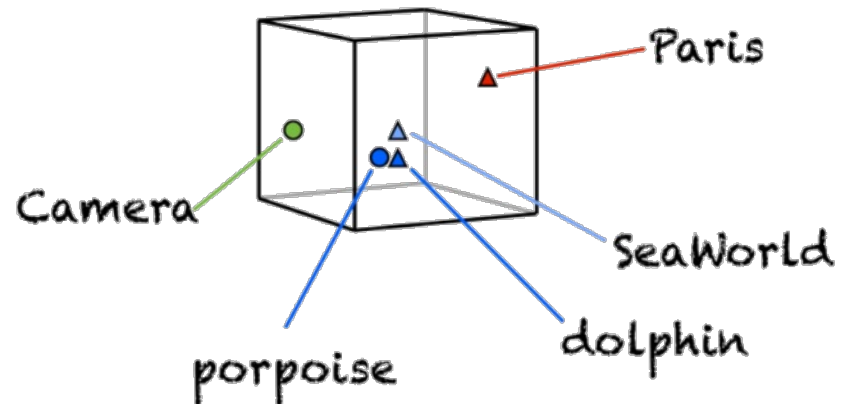
- Data producer computes representations  $Z$  given the fairness criteria and raw input data
- Inputs:
  - Data  $X, Y$
  - Fairness criteria:
    - For individual fairness: similarity metric  $\rho(X_i, X_j)$
    - For group fairness: classifier statistic  $\nu(f, Y|S)$  to equalize across groups  $S$ , e.g., statistical parity
- Output:
  - Fair representations,  $g: X, Y \rightarrow Z$
- There are many **feature/representation learning methods** with fairness constraints that can serve as  $g$

How to design function  $g$  that can produce fair representations from raw input data?

# Feature/Representation learning

- Representation learning methods produce condensed data summaries (i.e., feature vectors, embeddings), usually implemented as low-dimensional data transformations

$$g: X, Y \rightarrow Z$$

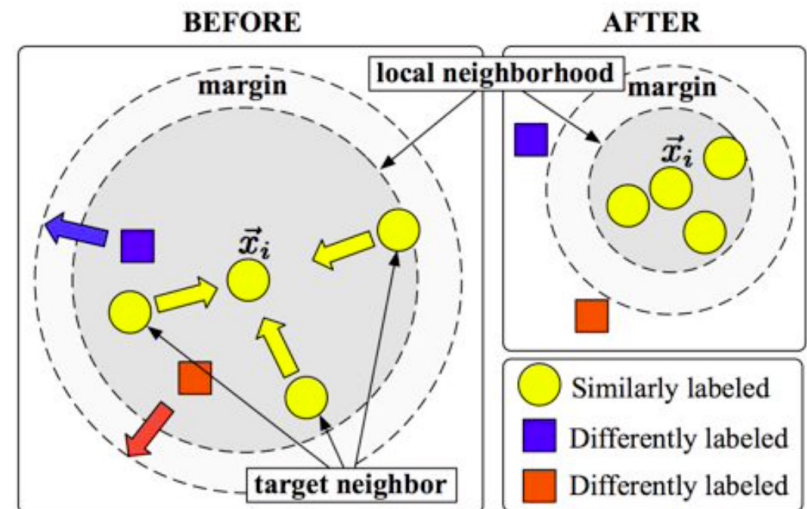


- Approaches in common use include PCA, non-linear autoencoders, deep embeddings (more on this in the next lecture)

# Individual fairness: Metric learning approach

- **Regulator** (to the data producer):
  - Provides sets of examples which should **be treated similarly** (e.g., similarly labeled points)
- Producer: **Learns distance metric  $\rho$**  such that individuals which should be treated similarly are closer to each other

Find a metric  $\rho$  such that  $\forall(x_1, x_2, x_3)$ :  
 $x_1, x_2 \in C_i$  and  $x_3 \in C_j (j \neq i)$   
 $\Rightarrow \rho(x_1, x_2) \leq \rho(x_1, x_3)$

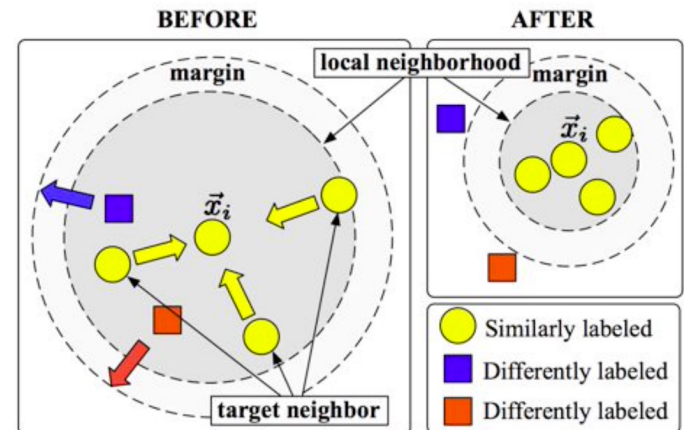


# Individual fairness: Metric learning approach

- **Regulator** (to the data producer):
  - Provides sets of examples which should **be treated similarly** (e.g., similarly labeled points)
- Producer: **Equivalently, learns a representation** such that individuals which should be treated similarly are embedded close together in the embedding space

Find a metric  $\rho$  such that  $\forall(x_1, x_2, x_3)$ :  
 $x_1, x_2 \in C_i$  and  $x_3 \in C_j (j \neq i)$   
 $\Rightarrow \|z_1 - z_2\|_2 \leq \|z_1 - z_3\|_2$

where  $z_i = Lx_i$  and  $\rho(x_i, x_j) = x_i^T L^T Lx_j$ .



# Outline for today's class

- ✓ 1. Quantitative definitions of fairness in AI
- ✓ 2. Framework for fair AI
- ✓ 3. Algorithmic fairness criteria
  - Individual fairness
  - Group fairness
- ✓ 4. Auditing AI systems
  - Auditing input data
  - Auditing ML model

# Quick Check

<https://forms.gle/Nv6E3E5hda2FzSs57>

## BMI 702: Biomedical Artificial Intelligence

*Foundations of Biomedical Informatics II, Spring 2024*

Quick check quiz for lecture 5: Bias and fairness in biomedical AI

Course website and slides: <https://zitniklab.hms.harvard.edu/BMI702>

\* Indicates required question

First and last name \*

Your answer \_\_\_\_\_

Harvard email address \*

Your answer \_\_\_\_\_

Using the framework for fair AI, describe a biomedical AI application and explain the roles of data regulators, data users, and data producers. Which individuals in a clinic, research lab, biomedical institution or health system would take on these roles? \*

Your answer \_\_\_\_\_

Give a biomedical example where you think that ensuring **individual fairness** is necessary. \*

Your answer \_\_\_\_\_

Give a biomedical example where you think that ensuring **group fairness** is necessary. \*

Your answer \_\_\_\_\_