A Technical Perspective on Data Bias and AI Fairness

Mahmoud Rahat

School of Information Technology

Halmstad University

mahmoud.rahat@hh.se

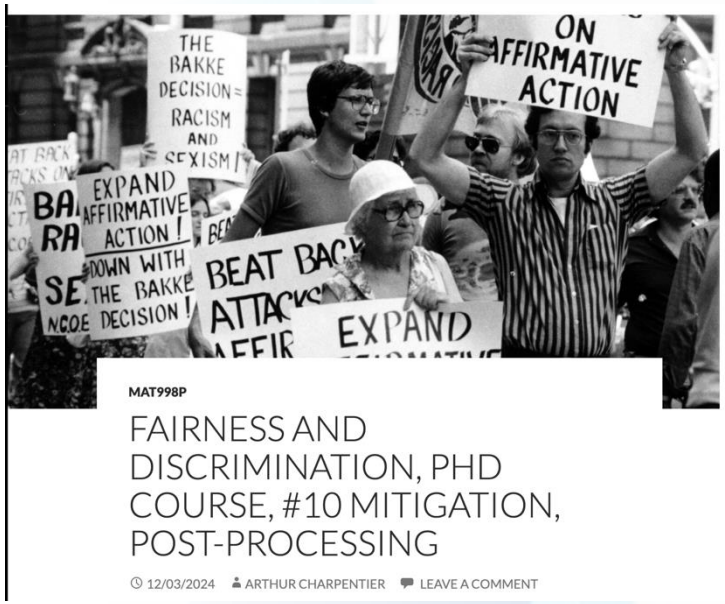
Outline

- Modern AI is not Fair
- Sensitive Attributes
- Algorithmic Fairness

Reference Courses

Some of the slides are borrowed from or inspired by these two excellent courses:

- Data Feminism @KTH
- Fairness and Discrimination @UQAM



https://freakonometrics.hypotheses.org/69650

FID3216 Data Feminism 7.5 credits

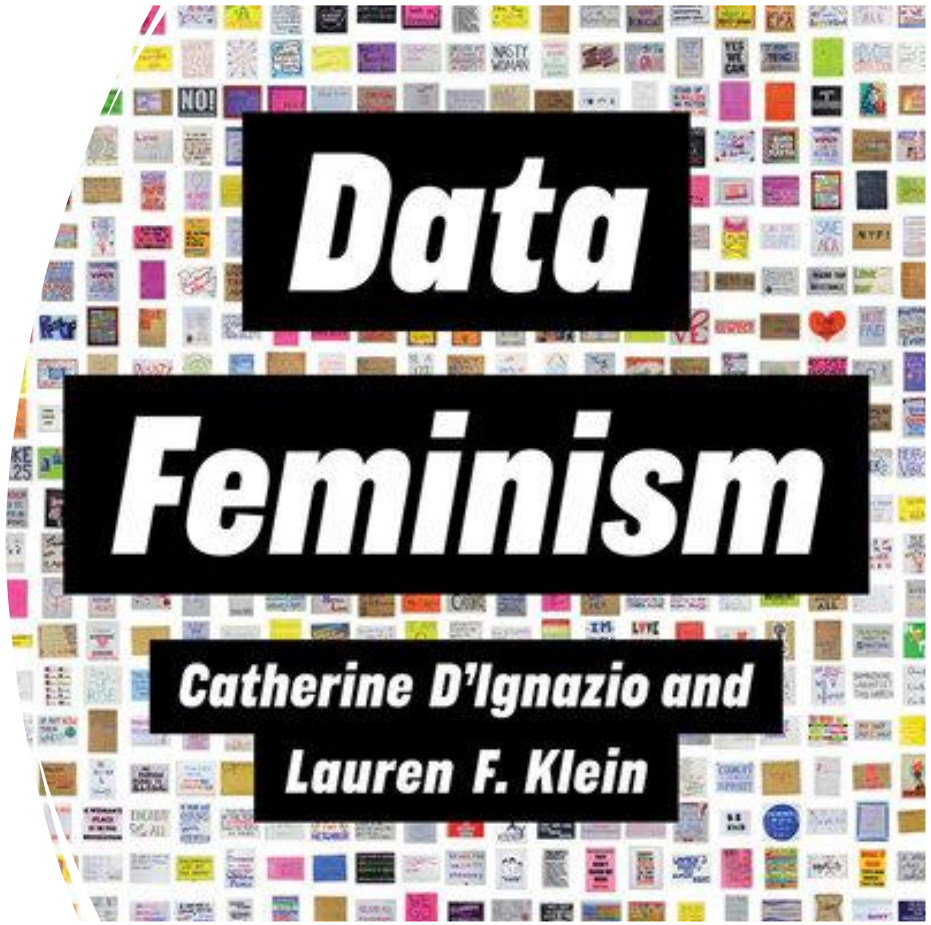
Before course selection

The "Data Feminism" course bridges the gap between data science and the crucial aspects of "equality, diversity, and equitable conditions (JML)". With a comprehensive exploration of these themes, the course delves deeply into both theoretical concepts and technical considerations surrounding data ethics, data justice, and data sustainability. The course is mainly inspired by the book "Data Feminism", which presents a paradigm that re-imagines the concept of data and its applications while acknowledging the inherent power imbalances within data science. Upon completing the course, students will be able to use data and data science to challenge and mitigate injustices amplified by data-driven practices. Moreover, they will gain the analytical skills to identify and address biases inherent in various data science practices.

https://www.kth.se/student/kurser/kurs/FID3216?l=en

Literature

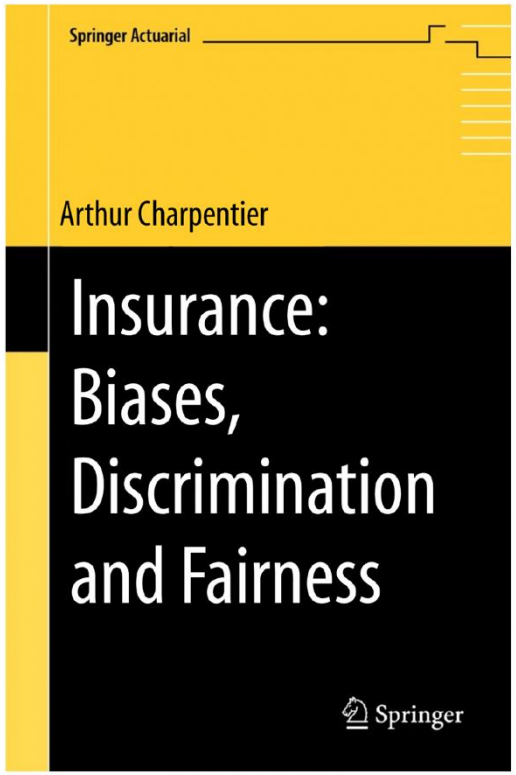
- **Data Feminism**
  - By Catherine D'Ignazio, Lauren F. Klein
  - Open Access: https://direct.mit.edu/books/book/4660/Data-Feminism
  - MIT Press
  - **DOI:** https://doi.org/10.7551/mitpress/11805.001.0001
  - **ISBN electronic:**9780262358521
  - **Publication date:** 2020



Literature

Insurance, Biases, Discrimination and Fairness
ISBN : 978-3-031-49782-7

Pitch: Discrimination and fairness of predictive models, in insurance, in the context of data enrichment ("big data") and opaque models ("machine learning", not to say "artificial intelligence").



AI is Making Decisions

- AI systems are algorithms that make decisions
- Modern AI (ML, deep learning) is used to:
  - Diagnose diseases
  - Recommend medicine
  - Approve loans
  - Set insurance prices
  - Rank job candidates
  - Recommend content
- But what has changed with Modern AI to make it so good?

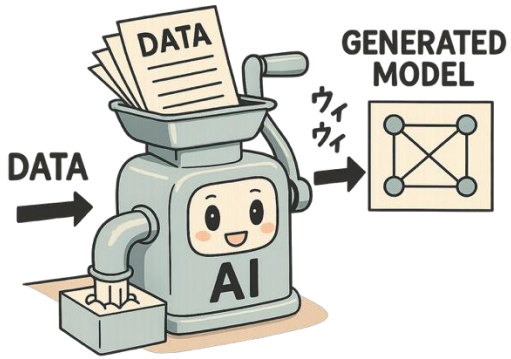
Machine learning algorithms have become particularly good at spotting patterns from medical images.

https://www.medicaldevice-network.com/features/ai-diagnosis/?cf-view

Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates. Photograph: Brian Snyder/Reuters

https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

What has changes with Modern AI



- AI has been around for quite some time, so **why is AI only now taking off** and performing so well?
- One of the main driving factors is the **availability of big data**.
- Modern AI is largely **data-driven**, meaning that machine learning models are trained **at scale**, primarily based on data rather than **hand-crafted rules**.
- While this has led to impressive performance, it also raises important **concerns around fairness and bias**.

Four Driving Factors…

Big Data Availability

- Larger Datasets
- Easier Collection & Annotation & Storage

IMAGENET 15 millions of labeled images

facebook 350 millions images uploaded per day

YouTube 100 hours of video uploaded every minute

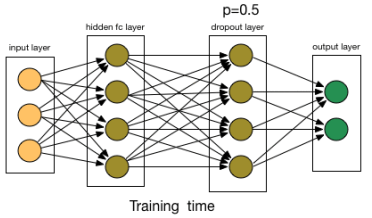
Computational Power (Hardware)

- Graphics Processing Units (GPUs)
- Massively Parallelizable
- Price decreases
- In every year, computational power is increasing exponentially.

Algorithms (Software)

- Improved regularization (e.g., dropout) and optimization techniques
- Novel Methods
- Toolboxes

TensorFlow



Marketing

- Fancy naming
- New breakthroughs in robotics and AI

AI learns from data — and data reflects society

- AI models learn patterns from historical data. That data may contain:
  - Historical discrimination
  - Social inequalities
  - Biased measurements
  - Missing or underrepresented groups
- As a result, **AI can learn and reproduce unfair patterns**, even without explicit intent.

https://davenussbaum.com/blog/forget-about-fairness

Example:

1. Past hiring favored men → AI learns to favor male candidates

2. Medical data overrepresents light skin → AI performs worse on darker skin

Classical AI	Modern AI
Explicit Rules	Learned Patterns
Easier to Inspect	Often Opaque (“black box”)
Smaller Scale	Deployed at massive Scale
Limited Data	Massive uncontrolled Data

https://www.linkedin.com/pulse/why-fairness-ai-more-complex-than-youve-heard-dena-neek-l3zhc/

This means:
- Unfairness can be **hidden**
- Decisions are **hard to explain**
- Bias can affect **millions of people**

Two Scenarios

FID3216 - DATA FEMINISM COURSE (2024)

The Course Examiner

Amir H. Payberah

Scenario 1: Medical Research on Heart Attacks

- A researcher builds a model to detect heart attacks.
- Model trained on medical records and labeled data of prior patients.
- Observes higher false negative rate for women

What could have gone wrong here?

https://tinyurl.com/22nv9suy

Scenario 1: Addressing The Issue

- Hypothesis: Model lacks sufficient examples of heart attacks in women.
- Solution: Augment dataset with more data on women who experienced heart attacks.
- Result: Improved performance for female patients.

https://tinyurl.com/4a39xpek

Scenario 2: Hiring Lab Technicians

- A researcher builds a model to evaluate the candidates from their CVs.
- Model trained on CVs and human-assigned ratings.
- Notices women are less likely to be predicted as suitable candidates.

What could have gone wrong here?

https://tinyurl.com/b3c8m4sk

Scenario 2: Addressing The Issue

- Hypothesis: Model lacks sufficient women's resumes.
- Solution: Attempts to collect more samples of women to add to the dataset.
- Result: Disappointment as model behavior does not change.

What's the Difference?

- The sources of issues were different in each scenario.
- Medical research issue: Lack of data on women, resolved by adding more data.
- Hiring issue: Human assessment bias, additional data did not help.

https://tinyurl.com/4k88kuvh

Sensitive Attributes

Sensitive Attributes

- Almost everywhere, we can find a list of variables that are considered, by law, as sensitive, since they could lead to discrimination.
- Sensitive variable might change with time, and across regions...

Context

- There exists a list of variables considered (by law) as sensitive (e.g., in Québec)
  - race,
  - color,
  - sex,
  - gender identity or expression,
  - pregnancy,
  - sexual orientation,
  - civil status,
  - age,
  - religion,
  - political convictions,
  - language,
  - ethnic or national origin,
  - social condition,
  - disability

https://freakonometrics.hypotheses.org/71560

Arthur Charpentier (February 16, 2024). Fairness and discrimination, PhD Course, #7 Sensitive attributes and proxies. Freakonometrics. Retrieved December 9, 2025 from https://doi.org/10.58079/vuwh

Racism

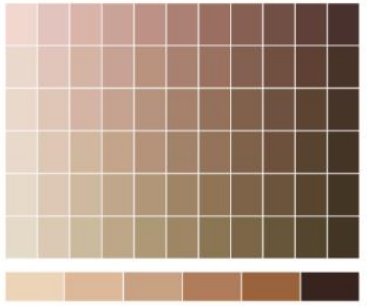
- The first sensitive attribute is probably the race.
- One should keep in mind that race is a social information, and most of the time, it is based on self-identification
- Racism is usually related to "colourism" (discrimination based on skin tone)

Racial Discrimination

Definition 5.2: Colourism, Merriam-Webster (2022)

Prejudice or discrimination especially within a racial or ethnic group favoring people with lighter skin over those with darker skin.

› Fitzpatrick Skin Scale (six levels), Telles (2014).



Racism

- It has been observed that African Americans, in the U.S. were usually asked a higher insurance premium.

Racial Discrimination

- In auto insurance, Heller (2015) observed that African American neighbourhood pay 70% more, on average, for auto insurance premiums than other neighbourhoods.

Figure 6-1. Average Premium by Company and Percentage of African American Residents

Company	<25% African American	25-49% African American	50-75% African American	≥75% African American	National Average	Percent Increase from <25% to ≥75% African American
Allstate	$658	$800	$848	$1,024	$674	56%
Farmers	662	757	795	1,271	676	92%
GEICO	575	713	793	876	591	53%
Progressive	694	852	911	1,332	717	93%
State Farm	543	697	771	882	561	63%
Top Five Companies	**$622**	**$769**	**$834**	**$1,060**	**$640**	**70%**

Source: CFA analysis of data provided by Quadrant Information Services, US Census

via https://www.michiganautolaw.com/wp-content/uploads/2017/08/Consumer-Federation-of-America-High-Price-of-Mandatory-Auto-Insurance-in-Predominantly...

Color

- In the widely used HAM10000 medical image dataset: **Less than 5%** of images depict darker skin tones
- Common conditions seen in **black patients** were missing entirely from the dataset
- This imbalance creates racial **bias in AI dermatology** tools

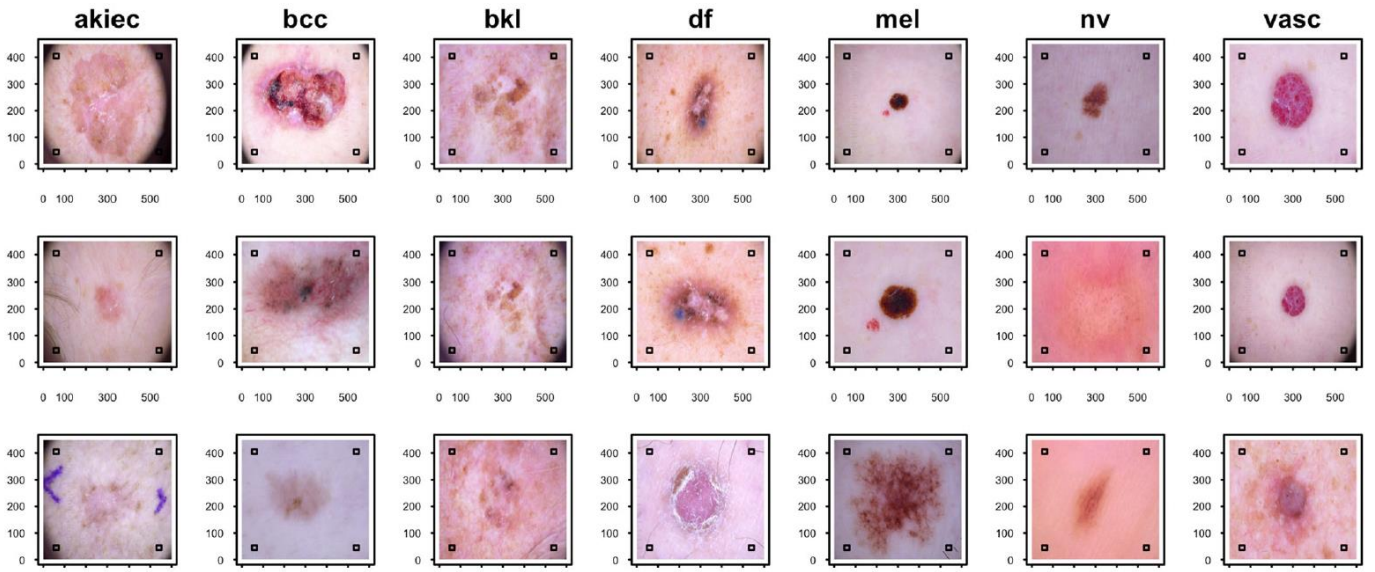


FIGURE 1 Visual overview of corner pixel sampling (the four squares at each corner) in the HAM10000 by lesion type (columns).

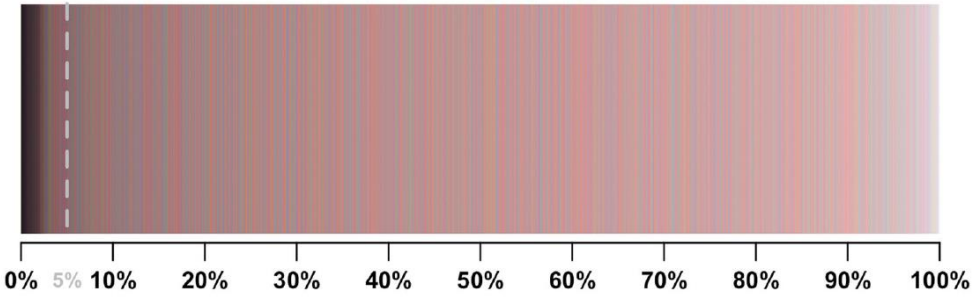


FIGURE 2 Proportional comparison of dark versus light skin images using the 10,015 pictures in the HAM10000 database.

Morales-Forero, Andres, et al. "An insight into racial bias in dermoscopy repositories: A HAM10000 data set analysis." *JEADV Clinical Practice* 3.3 (2024): 836-843.

Sexism

- Sexism is another popular exemple of discrimination, related to sex, or gender.

Sex and Gender Discrimination

› See slides with life tables per gender (exist since 1720, see Struyck (1912))

Definition 5.3: Sexism, Merriam-Webster (2022)

Prejudice or discrimination based on sex especially, discrimination against women; *also* behavior, conditions, or attitudes that foster stereotypes of social roles based on sex.

› Martin (1977), Hedges (1977) and Myers (1977) in the U.S. In Los Angeles, Department of Water and Power vs. Manhart, the Supreme Court considered a pension system in which female employees made higher contributions than males for the same monthly benefit because of longer life expectancy.

› See slides about the "Gender Directive" in Europe (and Thiery and Van Schoubroeck (2006)).

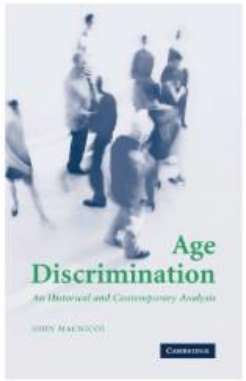
Ageism

- Age is another possible sensitive attribute, but it is more complicated. First, it is not a “club” and second, it is (somehow) clearly related to risk.
- Mental Health in elderly receives far less attention compared to other age groups [1].
  - Suicide prediction has been relatively **well studied among younger populations yet** remains understudied in older adults.
  - This **gap** exists despite **clear evidence** that suicide risk is also significant among the elderly.

Age-based Discrimination

› Age is not a club in which one enters at birth, and it will change with time, Macnicol (2006)

› “*If you are not already part of a group disadvantaged by prejudice, just wait a couple of decades—you will be*,” Robbins (2015).



Definition 5.4: Ageism, Merriam-Webster (2022)

Prejudice or discrimination against a particular age-group and especially the elderly.

› COVID-19 Decision Support Tool used in England, in March 2020, provided by the NHS (National Health System).
https://www.nhsdghandbook.co.uk/wp-content/uploads/2020/04/COVID-Decision-Support-Tool.pdf

[1] Karlsson, Oliver, et al. "Suicide Prediction among Older Adults in Sweden using Survival Analysis and Sequential Machine Learning Models." *Artificial Intelligence in Medicine and Healthcare* (2025): 110.

Table showing the number of suicides and the number of suicides per 100,000 inhabitants divided by gender and age groups in 2023 in Sweden.

Age group	Number of men (number of suicides among men per 100,000)	Number of women (number of suicides among women per 100,000)
15–29	138 (14)	69 (8)
30–44	223 (20)	81 (8)
45–64	300 (23)	128 (10)
65–84	222 (24)	84 (9)
85+	**55 (53)**	**18 (10)**

10 were children under the age of 15.

Source: Cause of Death Register, National Board of Health and Welfare.

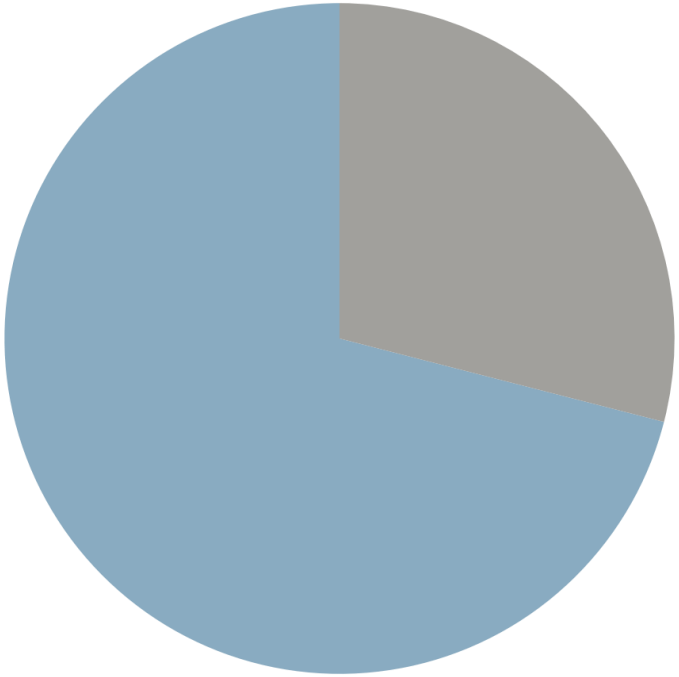
Suicide Deaths in Sweden: 1,328 Lives Lost in 2023

According to statistics from Folkhälsomyndigheten (Swedish Public Health Agency)

71%

Men

943 Cases



29%

Women

385 Cases

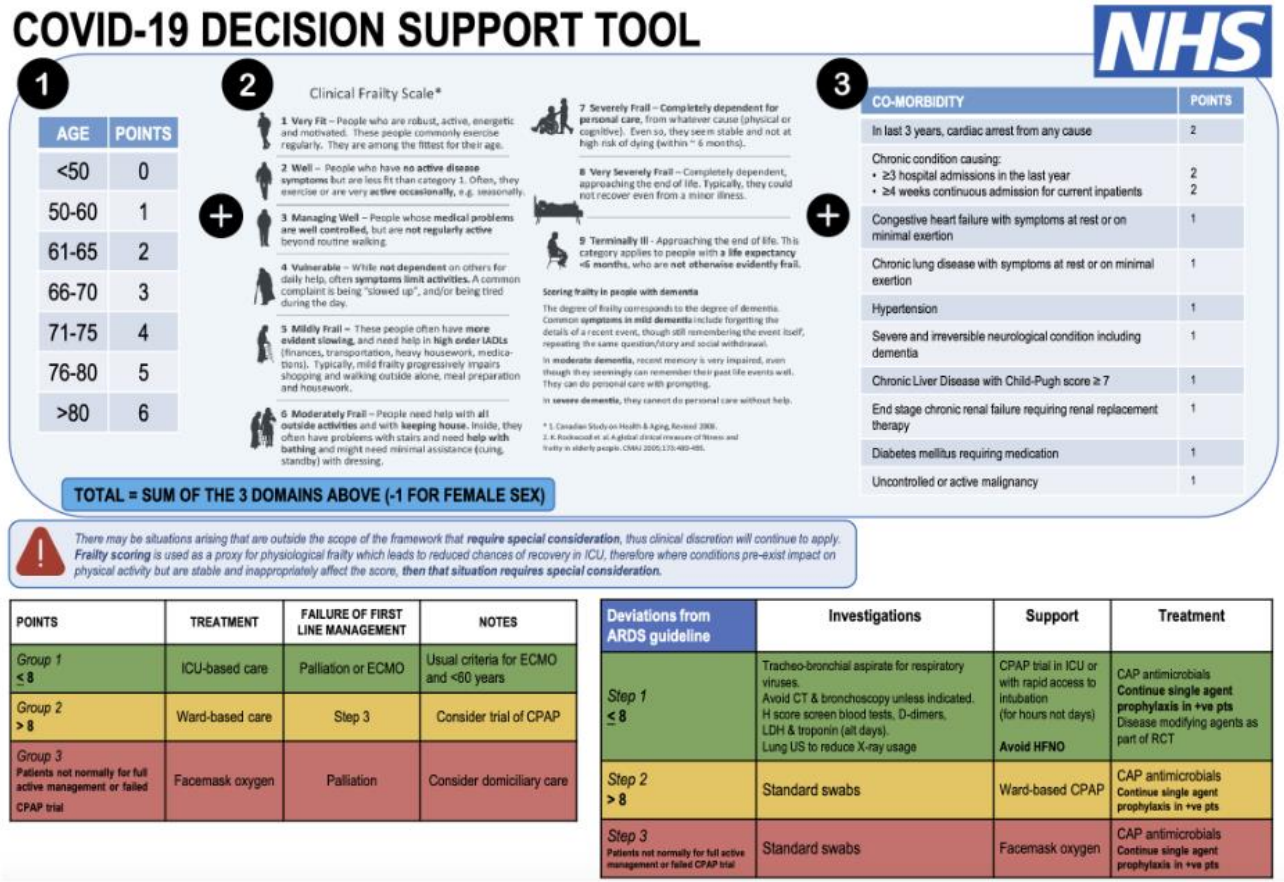
A further 289 cases were registered where there was suspicion of suicide, but where the intention could not be proven.

Source: Cause of Death Register, National Board of Health and Welfare.

Ageism in Healthcare

Age-based Discrimination

- In dataset, there can also be selection bias, related to the age. For instance, during the **COVID pandemic**, triage was based on the age of patients. Treatments and tests can be related to the age of patients. So, this bias will probably have an impact on observed risks.



COVID-19 DECISION SUPPORT TOOL

NHS

1

AGE	POINTS
<50	0
50-60	1
61-65	2
66-70	3
71-75	4
76-80	5
>80	6

2 Clinical Frailty Scale*

1 Very Fit – People who are robust, active, energetic and motivated. These people commonly exercise regularly. They are among the fittest for their age.

2 Well – People who have no active disease symptoms but are less fit than category 1. Often, they exercise or are very active occasionally, e.g. seasonally.

3 Managing Well – People whose medical problems are well controlled, but are not regularly active beyond routine walking.

4 Vulnerable – While not dependent on others for daily help, often symptoms limit activities. A common complaint is being "slowed up", and/or being tired during the day.

5 Mildly Frail – These people often have more evident slowing, and need help in high order IADLs (finances, transportation, heavy housework, medications). Typically, mild frailty progressively impairs shopping and walking outside alone, meal preparation and housework.

6 Moderately Frail – People need help with all outside activities and with keeping house. Inside, they often have problems with stairs and need help with bathing and might need minimal assistance (cuing, standby) with dressing.

7 Severely Frail – Completely dependent for personal care, from whatever cause (physical or cognitive). Even so, they seem stable and not at high risk of dying (within ~ 6 months).

8 Very Severely Frail – Completely dependent, approaching the end of life. Typically, they could not recover even from a minor illness.

9 Terminally Ill - Approaching the end of life. This category applies to people with a life expectancy <6 months, who are not otherwise evidently frail.

Scoring frailty in people with dementia

The degree of frailty corresponds to the degree of dementia. Common symptoms in mild dementia include forgetting the details of a recent event, though still remembering the event itself, repeating the same question/story and social withdrawal.

In moderate dementia, recent memory is very impaired, even though they seemingly can remember their past life events well. They can do personal care with prompting.

In severe dementia, they cannot do personal care without help.

* 1. Canadian Study on Health & Aging, Revised 2008.
2. K. Rockwood et al. A global clinical measure of fitness and frailty in elderly people. CMAJ 2005;173:489-495.

3

CO-MORBIDITY	POINTS
In last 3 years, cardiac arrest from any cause	2
Chronic condition causing: • ≥3 hospital admissions in the last year • ≥4 weeks continuous admission for current inpatients	2 2
Congestive heart failure with symptoms at rest or on minimal exertion	1
Chronic lung disease with symptoms at rest or on minimal exertion	1
Hypertension	1
Severe and irreversible neurological condition including dementia	1
Chronic Liver Disease with Child-Pugh score ≥ 7	1
End stage chronic renal failure requiring renal replacement therapy	1
Diabetes mellitus requiring medication	1
Uncontrolled or active malignancy	1

TOTAL = SUM OF THE 3 DOMAINS ABOVE (-1 FOR FEMALE SEX)

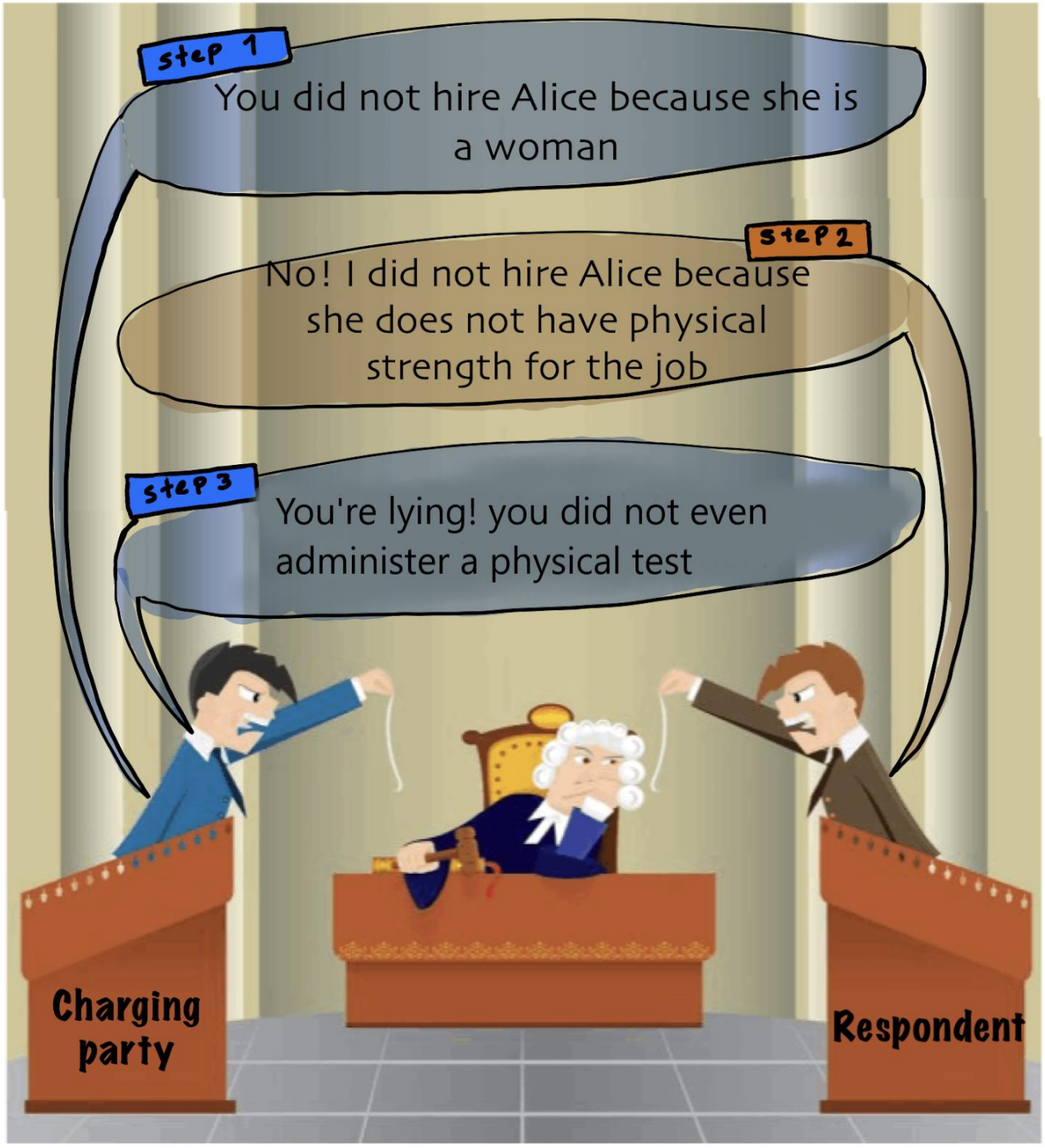
*There may be situations arising that are outside the scope of the framework that **require special consideration**, thus clinical discretion will continue to apply. **Frailty scoring** is used as a proxy for physiological frailty which leads to reduced chances of recovery in ICU, therefore where conditions pre-exist impact on physical activity but are stable and inappropriately affect the score, **then that situation requires special consideration.***

POINTS	TREATMENT	FAILURE OF FIRST LINE MANAGEMENT	NOTES
Group 1 ≤ 8	ICU-based care	Palliation or ECMO	Usual criteria for ECMO and <60 years
Group 2 > 8	Ward-based care	Step 3	Consider trial of CPAP
Group 3 Patients not normally for full active management or failed CPAP trial	Facemask oxygen	Palliation	Consider domiciliary care

Deviations from ARDS guideline	Investigations	Support	Treatment
Step 1 ≤ 8	Tracheo-bronchial aspirate for respiratory viruses. Avoid CT & bronchoscopy unless indicated. H score screen blood tests, D-dimers, LDH & troponin (alt days). Lung US to reduce X-ray usage	CPAP trial in ICU or with rapid access to intubation (for hours not days) **Avoid HFNO**	CAP antimicrobials **Continue single agent prophylaxis in +ve pts** Disease modifying agents as part of RCT
Step 2 > 8	Standard swabs	Ward-based CPAP	CAP antimicrobials **Continue single agent prophylaxis in +ve pts**
Step 3 Patients not normally for full active management or failed CPAP trial	Standard swabs	Facemask oxygen	CAP antimicrobials **Continue single agent prophylaxis in +ve pts**

Proxy Discrimination

- Even if you remove a sensitive attribute (like gender or race), the model can still **reconstruct it indirectly** using other variables that are correlated with it. As a result, discrimination can persist—or even look "legitimate" because the sensitive attribute is no longer explicit.
- **Sensitive attributes leave "footprints" in the data**
- Attributes such as gender, race, age, or ethnicity are often correlated with many other variables:
  - Occupation
  - Income
  - Location
  - Driving behavior
  - Education
  - Medical history
  - Weight
  - Height
- These variables act as **proxies**.



Models exploit Correlations, Not Intentions

- Machine-learning models are designed to:

  maximize predictive accuracy using **any available signal**

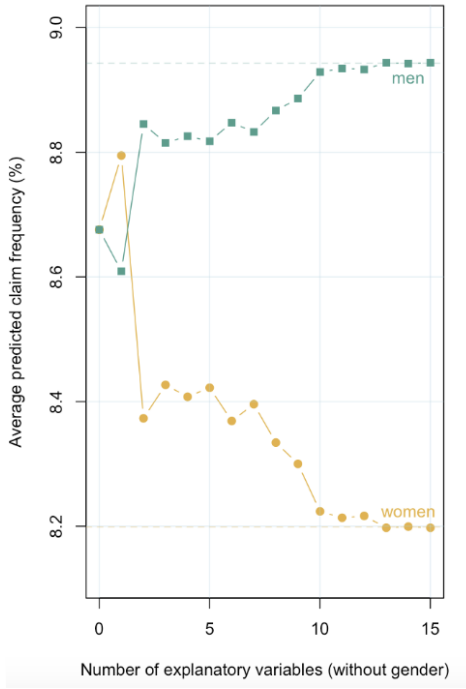
- They don't know what is "fair" or "ethical".

On Proxy Discrimination

› On `frenchmotor` dataset, average claim frequencies are 8.94% (men) 8.20% (women).

› Consider some logistic regression to estimate annual claim frequency, on k explanatory variables excluding gender.

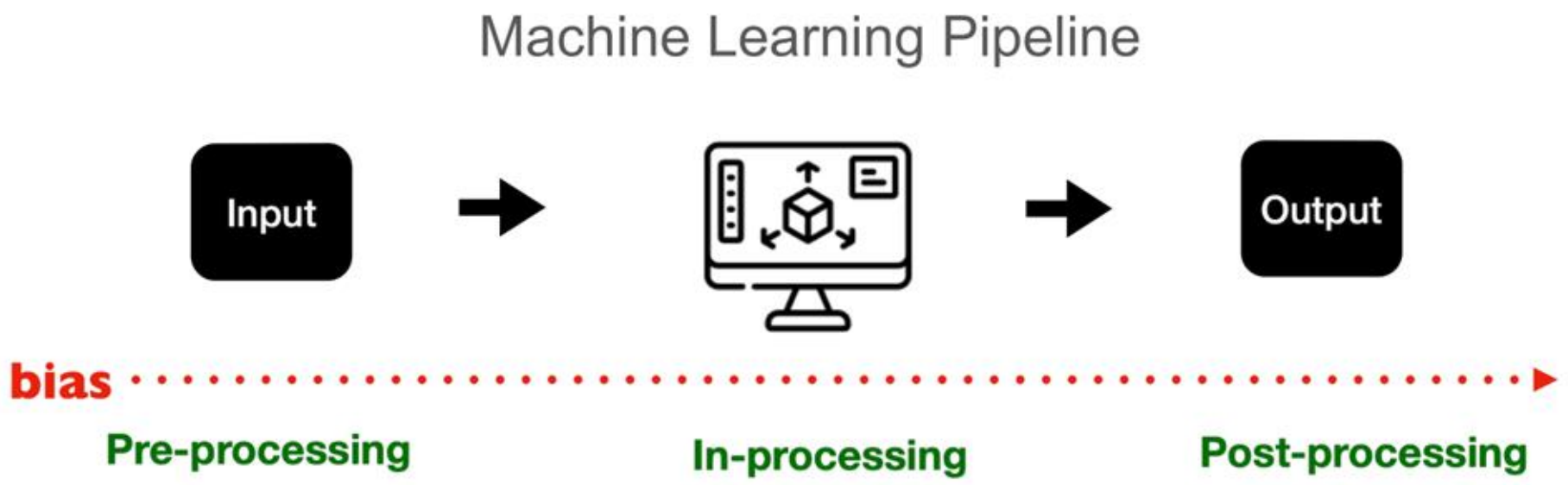
	men	women
$k = 0$	8.68%	8.68%
$k = 2$	8.85%	8.37%
$k = 8$	8.87%	8.33%
$k = 15$	8.94%	8.20%
empirical	8.94%	8.20%



Algorithmic Fairness

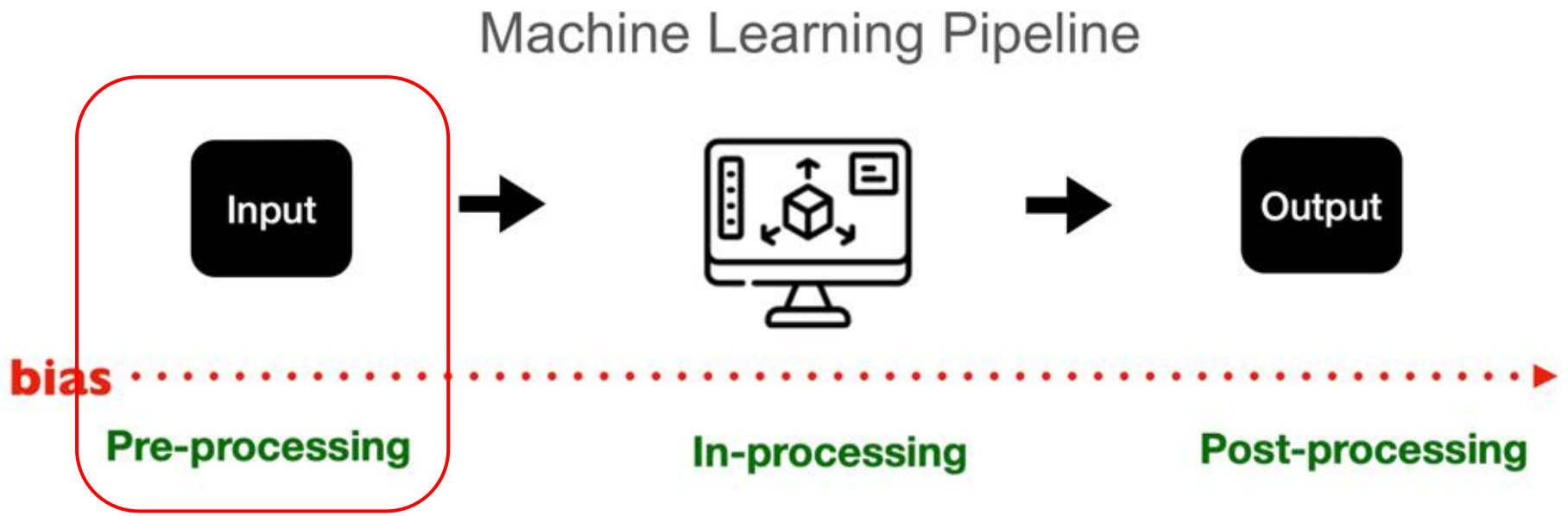
How To Mitigate Bias in ML Pipeline?

- Pre-processing
- In-processing
- Post-processing



How To Mitigate Bias in ML Pipeline?

- Pre-processing
- In-processing
- Post-processing



Pre-Processing Bias Mitigation

- Data collection
- Relabeling (reweighting) and perturbation
- Sampling

Data Collection

- Data collection is inherently costly.

- Minimizing data collection costs.
  - Acquire the least amount of additional data necessary.
  - Identify and add only essential data points.

Relabeling and Perturbation

- Adjusts or modifies the labels or features of data points to reduce bias.
- Relabeling (reweighting): change the labels for protected groups to balance outcomes.
- Perturbation: introduces small changes to features or labels to break biased patterns

Relabeling and Perturbation

- Relabeling: some females with higher education and long working hours are labeled as <=50K, while their male counterparts are labeled as >50K.

Gender	Education Level	Work Hours/Week	Income Category
Female	Bachelors	40	<=50K
Male	Bachelors	40	>50K
Female	Masters	45	<=50K
Male	Masters	45	>50K
Female	Bachelors	50	<=50K

Gender	Education Level	Work Hours/Week	Original Income	New Income Label
Female	Bachelors	40	<=50K	<=50K
Male	Bachelors	40	>50K	>50K
Female	Masters	45	<=50K	**>50K**
Male	Masters	45	>50K	>50K
Female	Bachelors	50	<=50K	<=50K

Relabeling and Perturbation

- Perturbation: add some noise into the gender attribute by randomly flipping the gender for a small percentage of individuals (e.g., 5%).

Gender	Education Level	Work Hours/Week	Income Category
Female	Bachelors	40	<=50K
Male	Bachelors	40	>50K
Female	Masters	45	<=50K
Male	Masters	45	>50K
Female	Bachelors	50	<=50K

Gender	Education Level	Work Hours/Week	Original Income	Perturbed Features
Female	Bachelors	40	<=50K	Male
Male	Bachelors	40	>50K	Male
Female	Masters	45	<=50K	Female
Male	Masters	45	>50K	Male
Female	Bachelors	50	<=50K	Female

Sampling

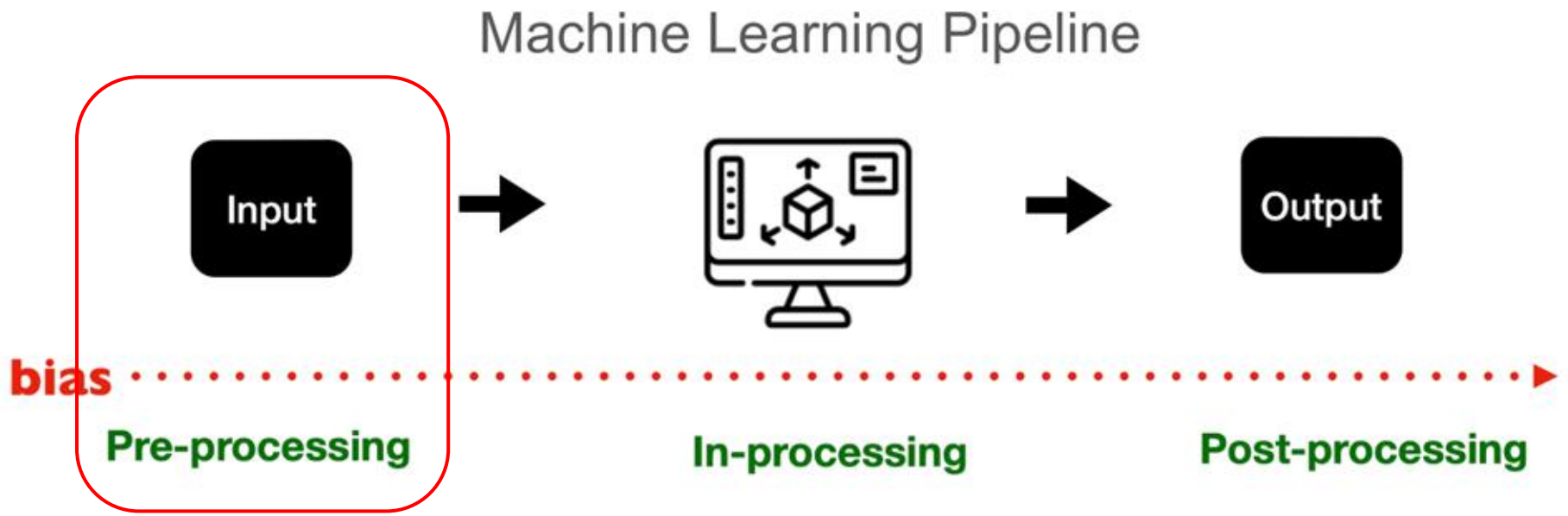
- Modifies the distribution of training data, ensuring the model is trained on more balanced data
- Oversampling: duplicate instances from underrepresented groups.
- Undersampling: removes instances from overrepresented groups.

Question for Reflection

In cases where collecting unbiased data is impractical or impossible, what are some ways to adjust model training or outputs to prevent harm?

How To Mitigate Bias in ML Pipeline?

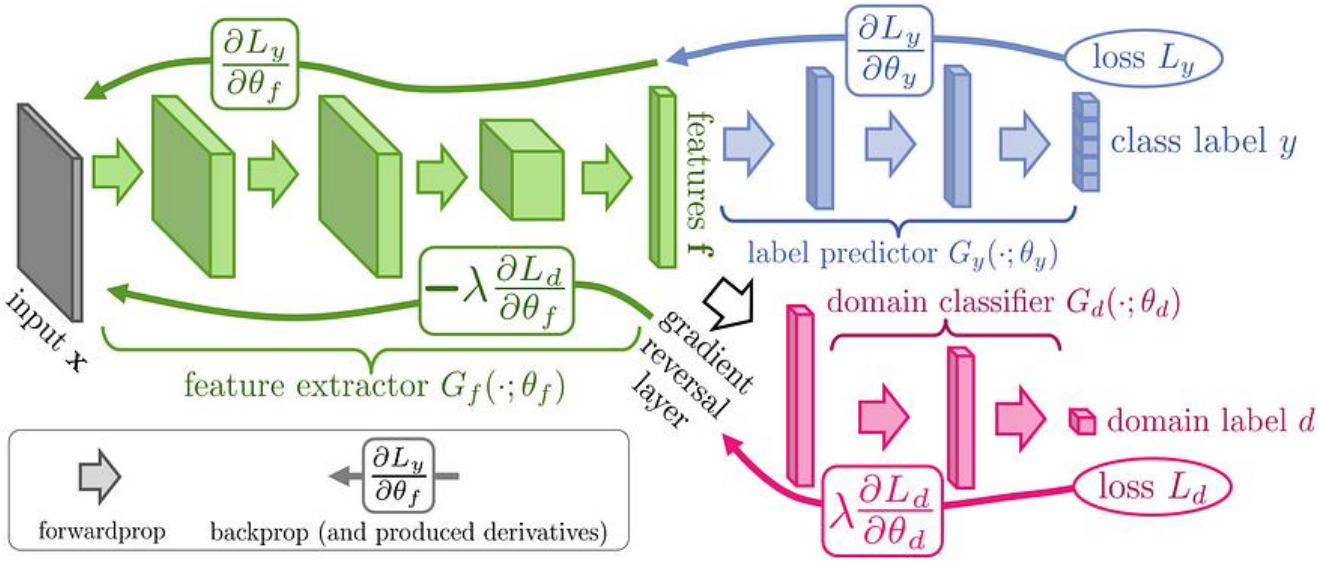
- Pre-processing
- In-processing
- Post-processing



Adversarial Training for Fairness

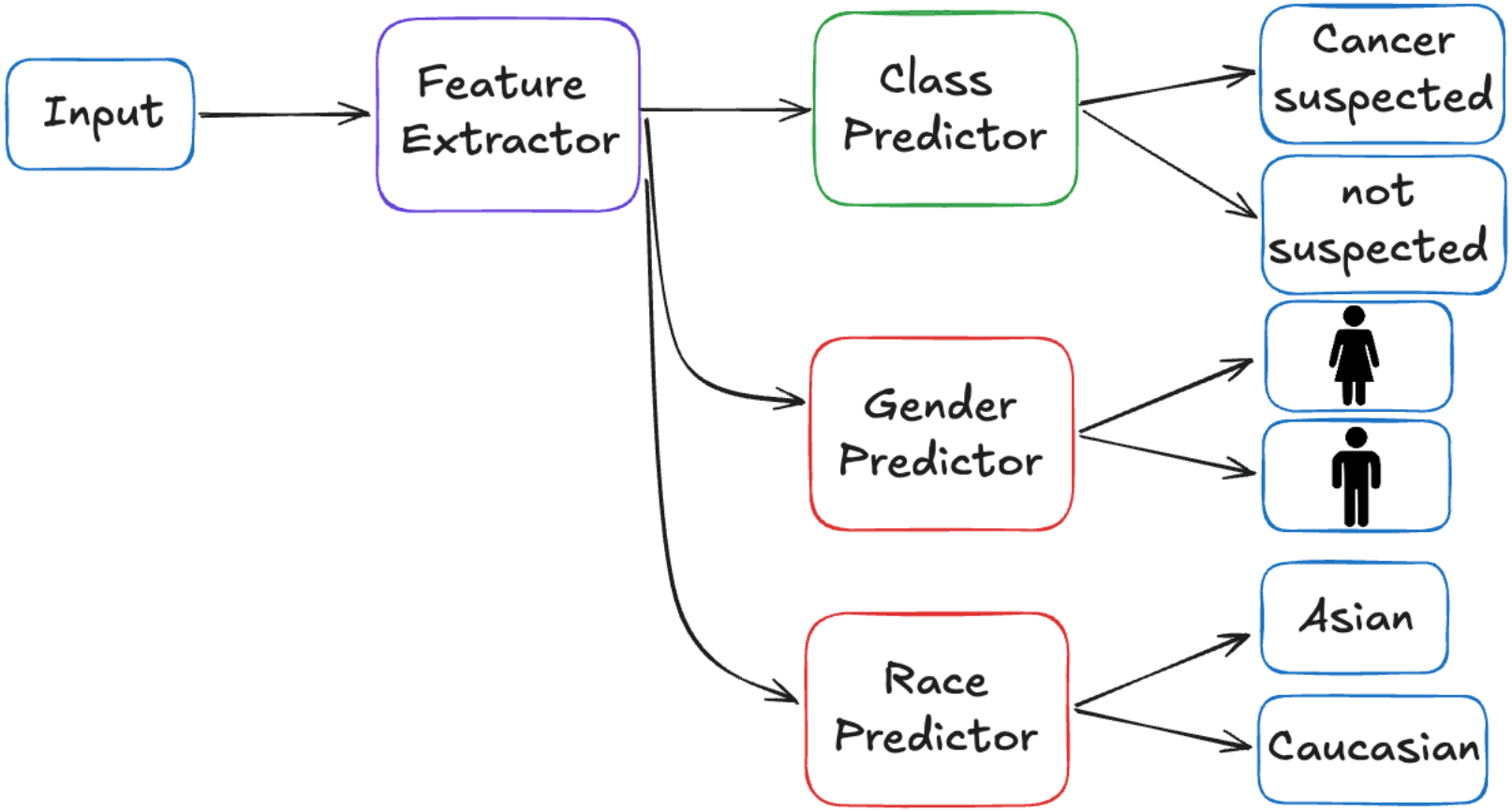
Adversarial training for fairness aims to **learn representations** that support accurate predictions while **preventing** the model from **encoding sensitive attributes**.

Domain-Adversarial Neural Networks (DANN)

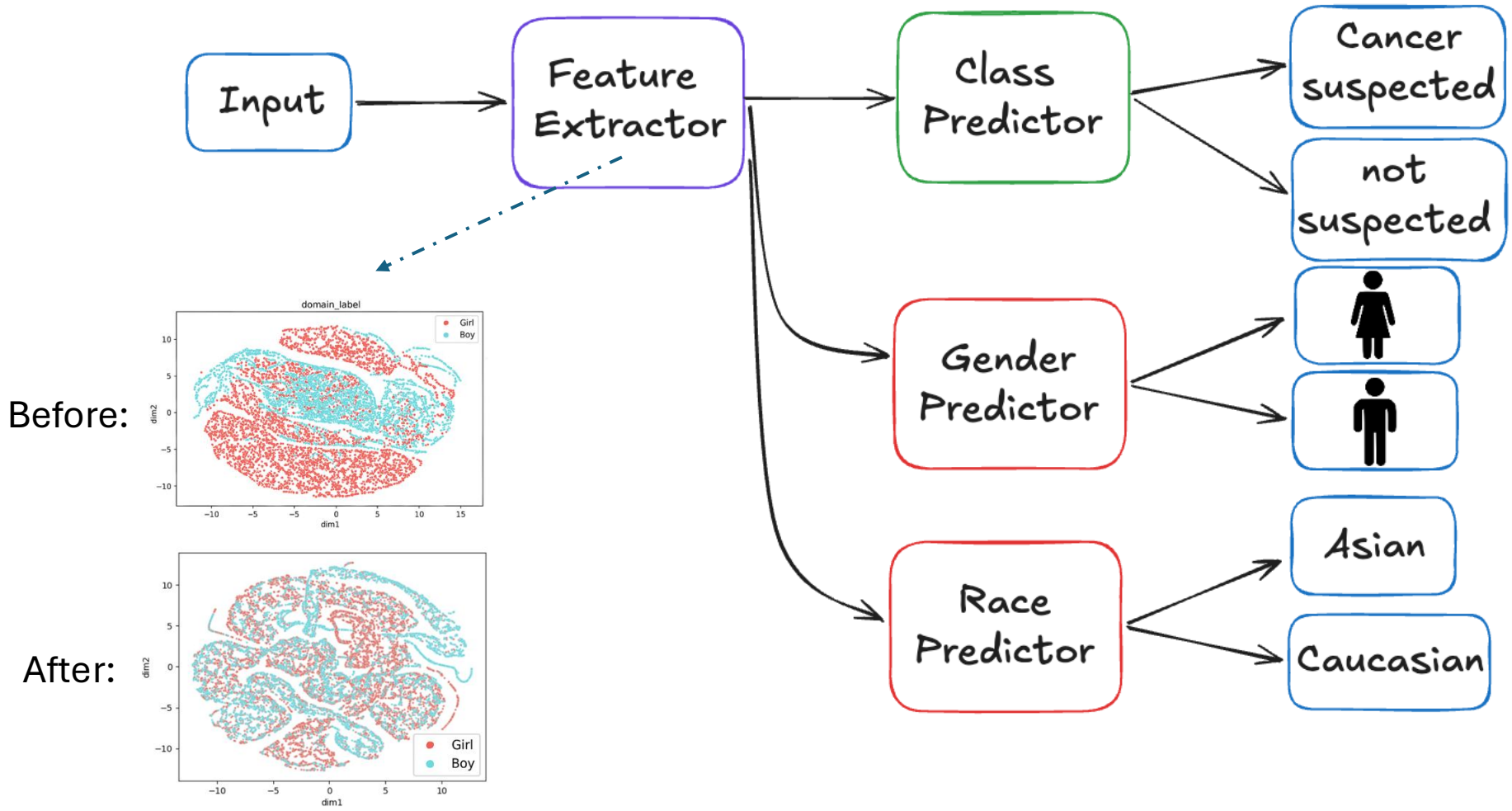


Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." *Journal of machine learning research* 17.59 (2016): 1-35.

DANN Architecture with Sensitive Attribute Predictors



DANN Architecture with Sensitive Attribute Predictors

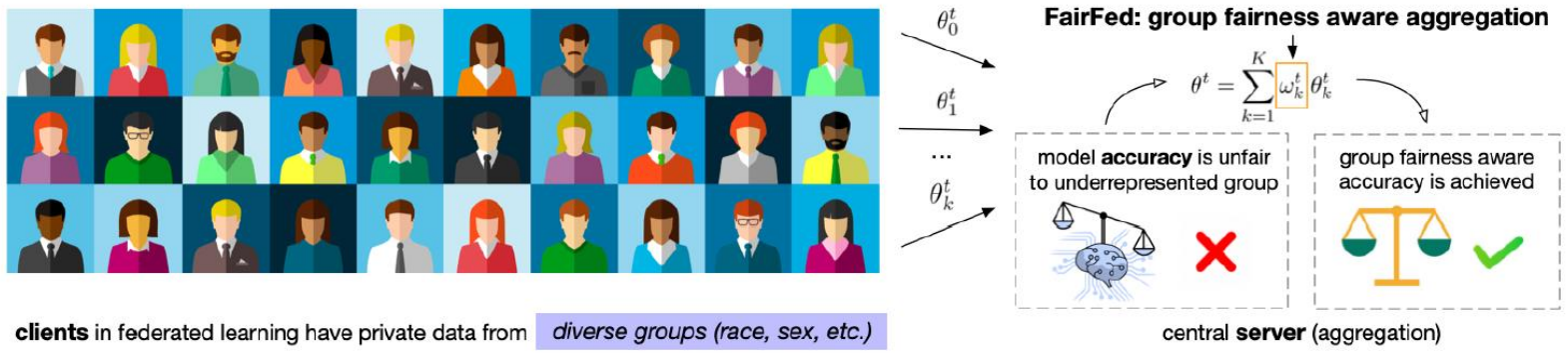


Federated Learning for Fairness

Federated learning for fairness trains models **collaboratively** on **decentralized data** to reduce bias while **preserving privacy**.

NVIDIA Research: First Privacy-Preserving Federated Learning System for Medical Imaging

FairFed: Group fairness-aware FL framework



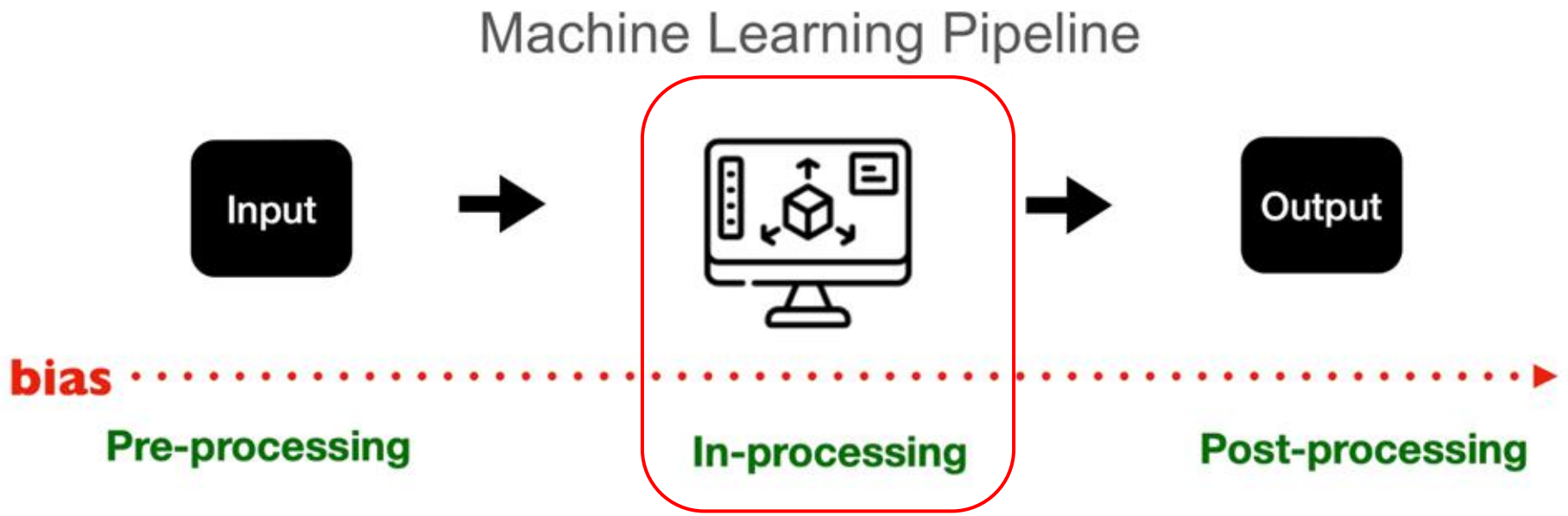
Federated learning (FL) can help improve **AI fairness** and reduce **data bias** in several keyways:

- Inclusive learning from **diverse populations**
- Privacy-preserving access to **sensitive data**
- Mitigation of **geographic and institutional bias**
- Fairness-aware local updates
- Continuous bias monitoring in dynamic settings

Ezzeldin, Yahya H., et al. "Fairfed: Enabling group fairness in federated learning." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. No. 6. 2023.

How To Mitigate Bias in ML Pipeline?

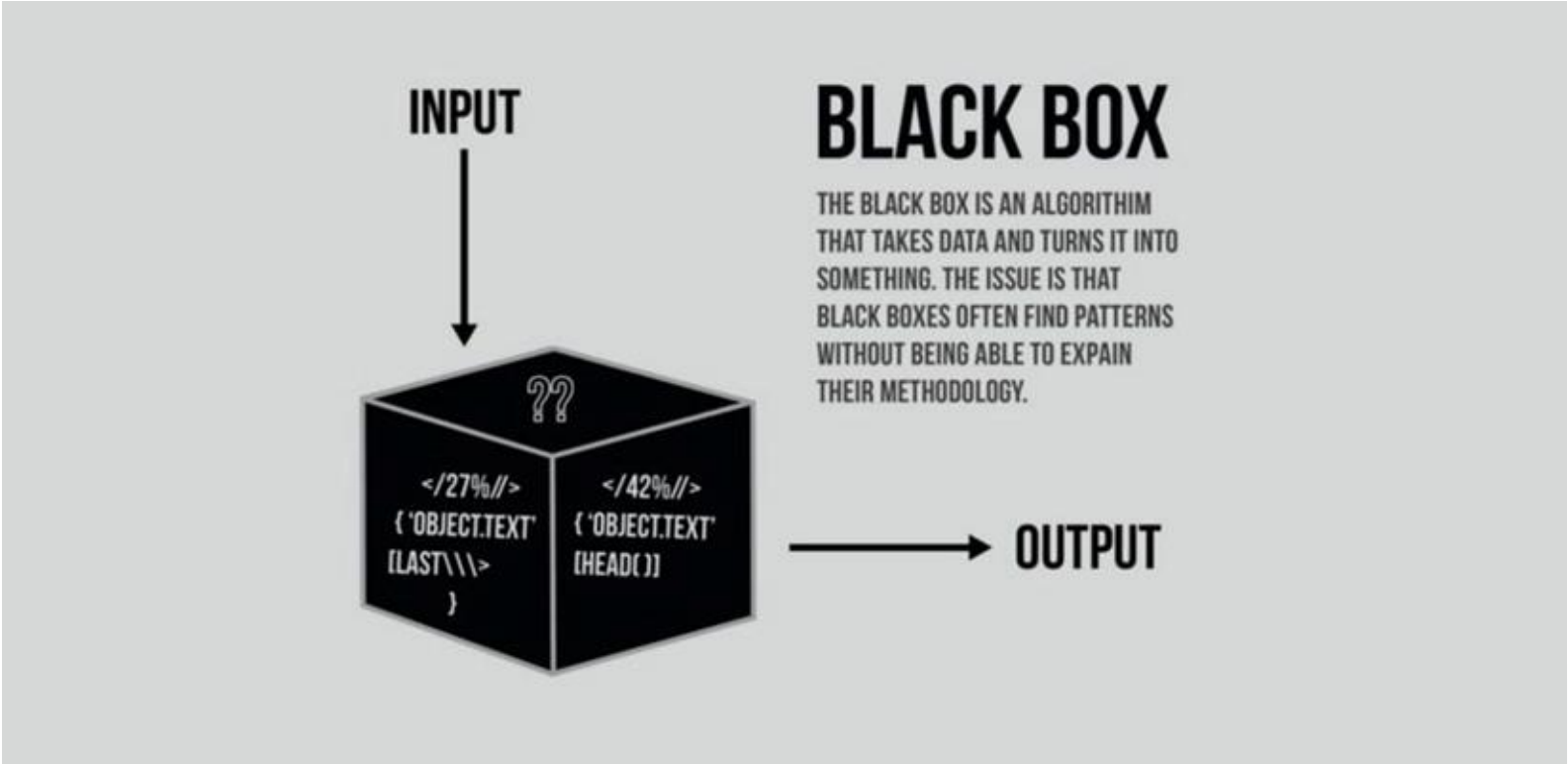
- Pre-processing
- In-processing
- Post-processing



XAI for Fairness

- XAI promotes transparency.
- Transparency enables bias detection.
- Bias mitigation can improve fairness.

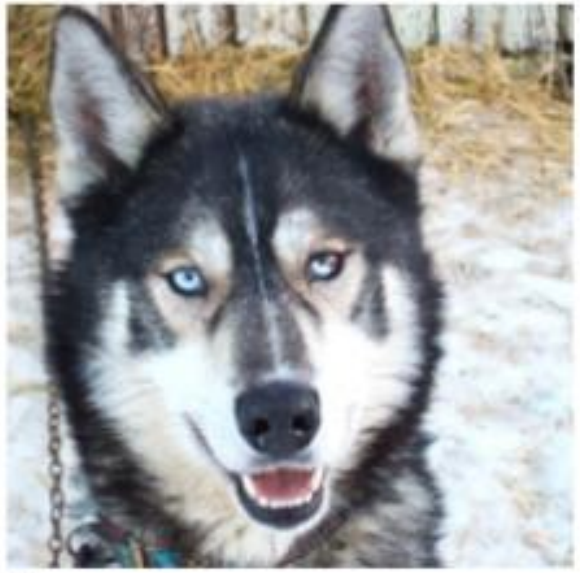
XAI is not inherently a fairness technique — it is a tool that can support fairness auditing or bias detection. They don't guarantee fair results, but they enable fairness evaluation.



HALMSTAD UNIVERSITY

https://www.topbots.com/interpretable-machine-learning/

When the Machine Learns Something Unexpected



(a) Husky classified as wolf

(b) Explanation

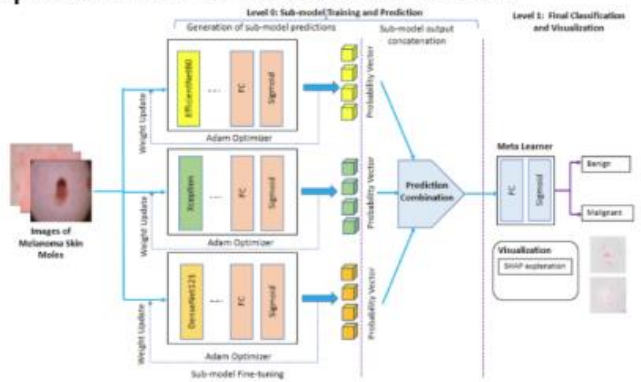
HALMSTAD UNIVERSITY

Explain The Prediction

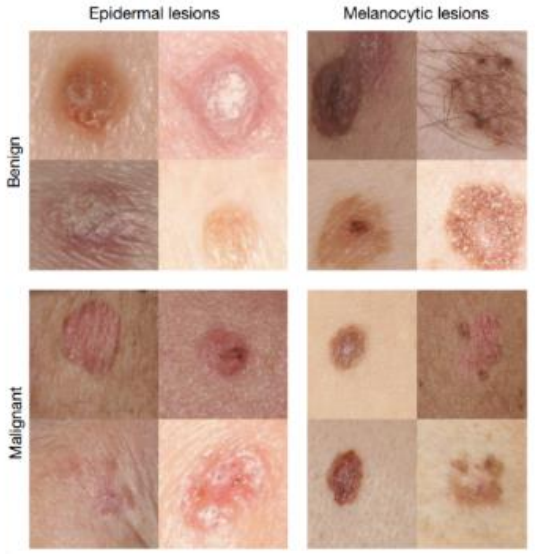
- Another issue with black boxes is that it might be hard to assess if they are related to **sensitive attribute.**
- In order to extract information in pictures, algorithms might use information that could be considered as **sensitive**.
- This can also be the case for health issues, where classifiers can be influenced by the color of the skin (or possibly some unexpected information!)

Explainability

Esteva et al. (2017) and Winkler et al. (2019) use deep-classifiers to detect skin cancer



> "*So in the set of biopsy images, if an image had a ruler in it, the algorithm was more likely to call a tumor malignant, because the presence of a ruler correlated with an increased likelihood a lesion was cancerous,*" Patel (2017)



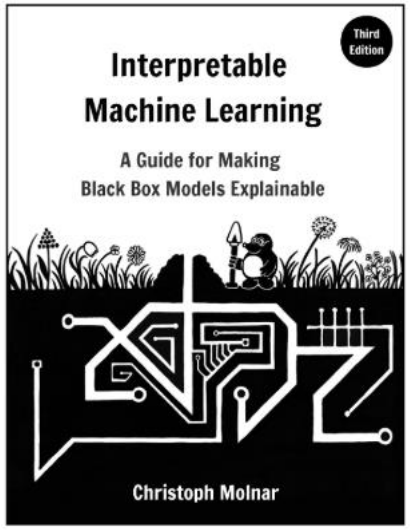
https://freakonometrics.hypotheses.org/71560

Counterfactual for Fairness

Counterfactual fairness is a notion of fairness where a model's prediction remains the same in a hypothetical world in which an individual's **sensitive attributes** (such as gender or race) were different, while all other factors are held constant.

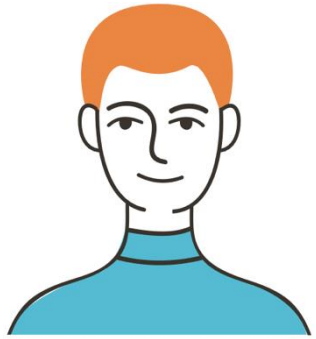
Counterfactual Explanations: Definition

"A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output."



Christoph Molnar (Interpretable Machine Learning)

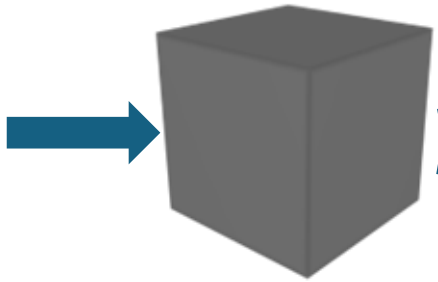
Counterfactual Explanations: Motivation



Applicant

Bank account data:

- Age
- Credit_years
- # of credit cards
- Salary
- people_liable
- …

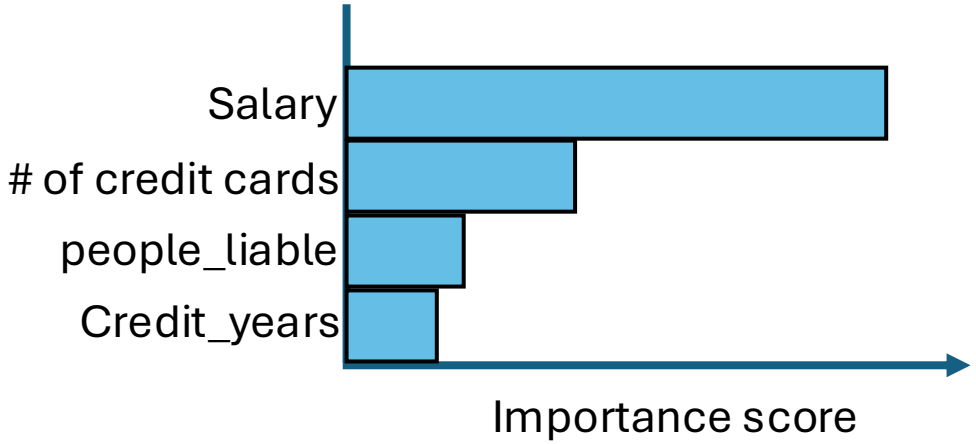


Predictive Model

f(x)

Loan granted

Loan rejected

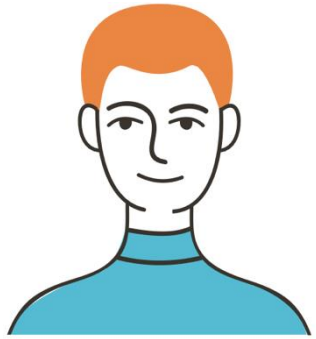


Decision makers: why this person was not eligible?

Person: What should I do next time to get the loan?

what to do to flip the class??

Counterfactual Explanations: Motivation



Applicant

Bank account data:

- Age
- Credit_years
- # of credit cards
- Salary
- people_liable
- …

Predictive Model

f(x)

Loan granted

Loan rejected

Counterfactual Generation Algorithm

Increase your salary by 10K
And pay your credit cards on time for next 2 month

Decision makers: why this person was not eligible
Person: What should I do next time to get the loan
what to do to flip the class??

1. But which features are allowed to be changed?
2. What if the counterfactual analysis suggests that changing a sensitive attribute like gender would flip the decision?

Conclusion

- Modern AI systems are powerful, but they inherit biases present in data and society.
- Understanding sensitive attributes is essential to detecting and mitigating unfair outcomes.
- Building fair AI requires careful data practices, transparent models, and ongoing evaluation.

The End

- Thanks for Listening
- Any Questions?