**Explainable-AI**

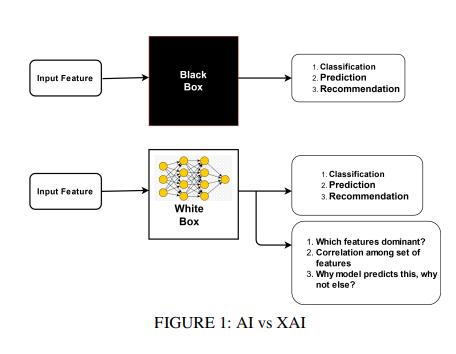
**Introduction:**

This current day AI is mainly limited to a sub-branch known as machine learning (ML). Machine learning provides a computer with a set of examples (aka training data set), and let the computer learn from the example set. Once well trained, the computer can then answer questions related to what it was taught previously. Typically, this traditional AI is a blackbox that can answer “yes” and “no” type questions without elaborating how that answer is obtained.

**In many applications, an explanation of how an answer was obtained is crucial for ensuring trust and transparency**

**Example:**An example of one such application is a medical application, where the doctors should be damn sure about a conclusion. They, for example, would like to know how AI decided whether someone is suffering from a disease by analyzing a CT scan image.

An insight of how a result was obtained will therefore not only can induce trustfulness but also can avoid life-threatening errors. In some other applications (e.g., law and order), answers to other "wh" questions (such as "why", "when", "where", etc.) could be required. The traditional AI is unable to answer these "wh" questions. This explainability requirement lead a new area of AI research, know as Explainable AI (XAI)

****

**Simple Explanation of Blackbox vs Glassbox Models:**

**Scenario:**

**A bank wants to automate its loan approval process using machine learning models. However, it's crucial for the bank to not only make accurate predictions but also to ensure transparency and fairness in its decision-making process, complying with regulations and earning the trust of its customers.**

**Blackbox Model Approach:**

**The bank could deploy a complex blackbox model like a deep neural network to predict whether a loan application should be approved or denied. This model might achieve high accuracy, but it's challenging to explain its decisions to customers or regulators. If the model denies a loan application, it's difficult to provide a clear explanation as to why, which could lead to distrust or legal issues.**

**Glassbox Model Approach:**

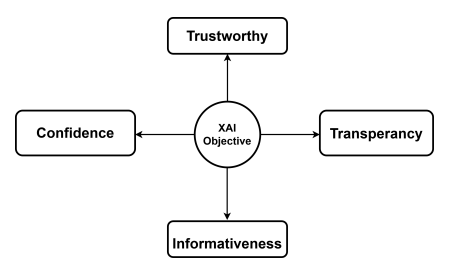
**Instead, the bank chooses to use a glassbox model such as a decision tree or a rule-based system for loan approval. While these models might not achieve the same level of accuracy as blackbox models, they offer transparency and explainability. For example, if the decision tree denies a loan application, it can easily show the specific criteria (e.g., low credit score, high debt-to-income ratio) that led to that decision. This transparency helps both customers and regulators understand why a particular decision was made, increasing trust and compliance.**

**----------x----------**

**Blackbox models of machine learning are rapidly being used with a tag of AI-enabled technology** for various critical domains of human life. The list of domains varies from socioeconomic justice, cyber forensics, criminal justice, etc. But these AI-powered models are lagging to win the trust of naive people because these models are less transparent and less accountable.

**For example,** there are cases in criminal justice where the AI-enabled justice model release criminals on parole and grants bail. This leads to serious consequences among people and the government.

**Objective of X-AI:**

****

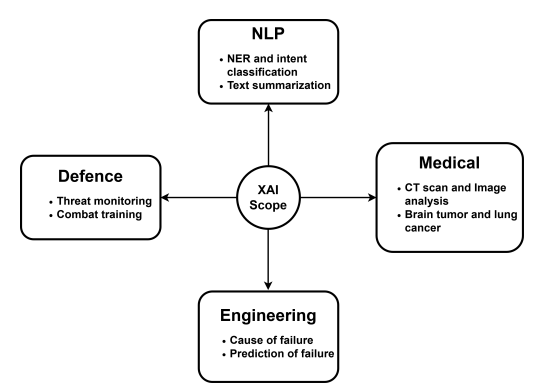
**The main objective of XAI is to answer the "wh" questions related to an obtained answer**.

For example, XAI should be able to answer

* "why a particular answer was obtained?"
* "how a particular answer was obtained"
* "when a particular AI-based system can fail?"

XAI can enhance transparency as well as fairness by providing a justification that can be understood by a layman. The minimum criteria for a transparent AI model are it should be **expressive enough to be human-understandable.**

**Scope of X-AI:**

****

1. **Data Protection:** European Union and its regulatory body have a ’right to explanation’ clause. That makes to enables explanation from XAI algorithms.

2. **Medical:** XAI can diagnose a patient by observing his/her past medical records. Using AI/ML algorithms in **the medical image processing domain** it is easier for medical experts to diagnose patients with malignant cancer tumors and other lung diseases.

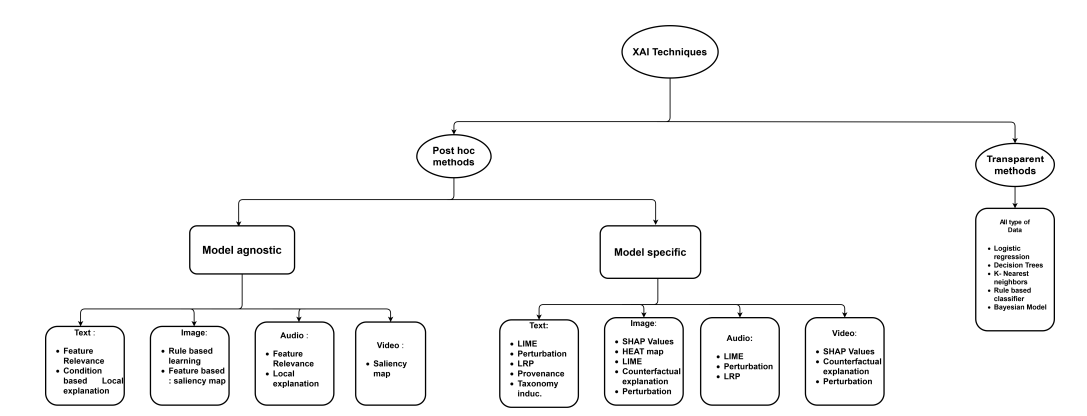
3. **Defense:** XAI in defense practices becomes crucial because of automated weapon and surveillance systems. XAI also provides **good second-hand support** during combat mode training and real-time combat tactics.

4**. Banking**: The banking system is one of the biggest financial sectors which affects human life the most. In dayto-day life, there are many fraud transactions and cones by cheaters. Well-trained XAI models can help to investigate fraudulent transactions and **help to reduce false positives cases.**

**CLASSIFICATION TREE**

XAI techniques are classified in two categories of **transparent and post-hoc methods.** Transparent methods are such methods where the inner working and decision-making process of the model is simple to interpret and represent. Bayesian model, decision trees, linear regression, and fuzzy inference systems are examples of transparent models. Transparent methods are useful where internal feature correlations are not that much complex or linear in nature

A **post-hoc XAI method** receives a trained and/or tested AI model as input, then generates useful approximations of the model’s inner working and decision logic by producing understandable representations in the form of feature importance scores, rule sets, heat maps, or natural language

****

**Enhancing Transparency and Fairness in AI:**

In our exploration of Explainable AI (XAI), we've emphasized the critical need for transparency and fairness in AI systems. Transparency ensures that the inner workings of AI models are understandable to humans, fostering trust and accountability. Fairness, on the other hand, guarantees that AI systems treat all individuals equally, without perpetuating biases or discrimination.

**The Impact of Bias in AI:**

One significant challenge in achieving fairness is addressing bias within AI models. Bias can manifest in various forms, leading to unjust outcomes and reinforcing societal inequalities. Let's delve into how bias can infiltrate AI systems and its ramifications:

1. **Sampling Bias**

* With sampling bias, one population is overrepresented or underrepresented in a training data set. An example of this would be a digital credit app in a market in which **men are more likely than women to have smartphones.** If all of the customer data is used to train the algorithm, the algorithm will **rely more on men’s data than on women’s.**

1. **Labeling Bias**

* Labeling is how data scientists annotate and classify certain properties and characteristics of a data point in order to make it searchable by an algorithm. An example of this would be labeling loan applicant occupations as **“doctor” versus “nurse” rather than as “healthcare worker.”** Doctor and nurse would quickly become proxies for gender among loan applicants, whereas **“healthcare worker” would continue to mask gender.**

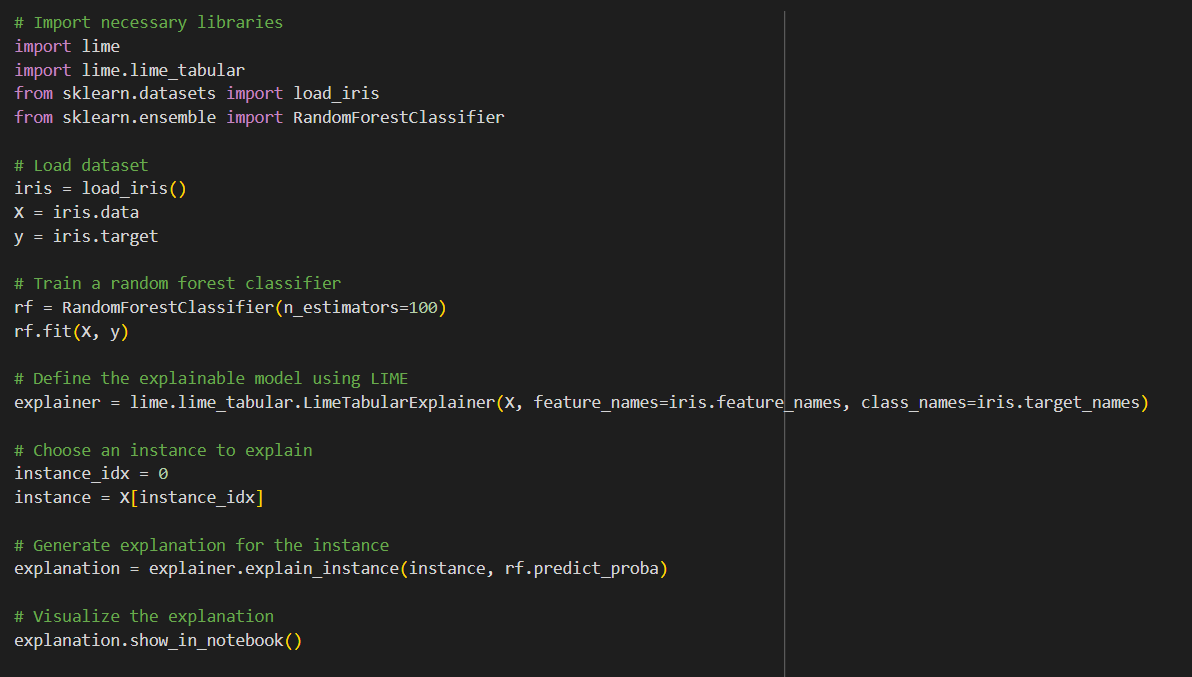
1. **Outcome Proxy Bias**

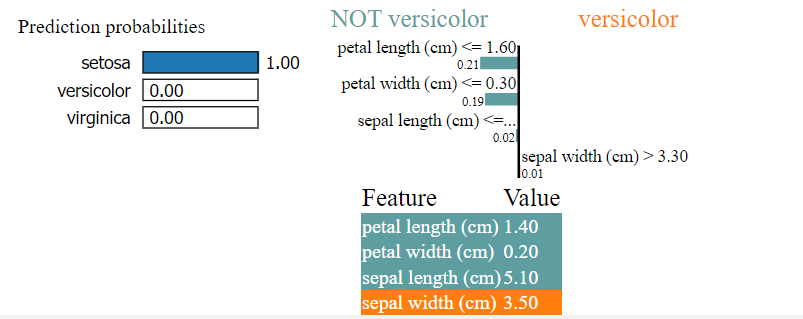
* Outcome proxy bias occurs when the machine learning assignment is not well-defined. For example, if an algorithm uses address of residence as a proxy to predict the likelihood of credit default, the decision-making suffers from outcome proxy bias. The data is biased because default **might be higher in neighborhoods with lower incomes**, but this correlation does not guarantee the individual will default.

**Tools to explore X-AI:**

Let's delve deeper into the technical aspects, particularly focusing on how XAI techniques are implemented, including **LIME** (Local Interpretable Model-agnostic Explanations) and **SHAP** (SHapley Additive exPlanations), and their significance in enhancing transparency and accountability in AI systems:

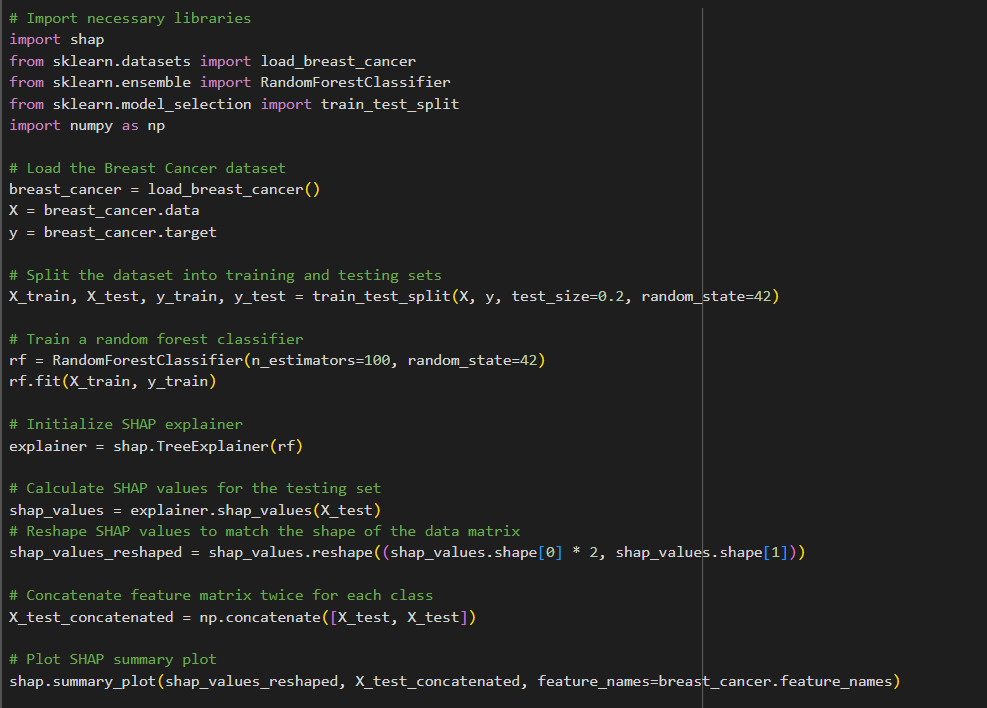
1. **Local Interpretable Model-agnostic Explanations (LIME):**
   * LIME is a technique used to explain the predictions of complex machine learning models on individual instances.
   * It works by generating locally faithful interpretations of the model's predictions by approximating the model's behavior in the vicinity of a specific data point.
   * LIME provides human-understandable explanations for individual predictions, helping users comprehend why a model made a particular decision.
   * For instance, in the context of loan approval, LIME could highlight the specific features of an applicant's profile that contributed most to the model's decision to approve or deny the loan.
2. **SHapley Additive exPlanations (SHAP):**
   * SHAP is another popular method for explaining the output of machine learning models.
   * It's based on cooperative game theory and assigns each feature an importance score indicating its contribution to the model's prediction.
   * SHAP values offer a global perspective on feature importance, showing how each feature impacts predictions across the entire dataset.
   * In the banking scenario, SHAP could reveal which features, such as **income level or credit history, have the most significant influence on loan approval decisions.**
3. **Implementation in Code:**
   * Both LIME and SHAP have implementations in various programming languages, including Python.
   * For example, the lime package in Python allows users to generate explanations for individual predictions using LIME.
   * Similarly, the shap library provides tools to compute SHAP values for different machine learning models, enabling users to interpret model predictions effectively.

**Sample Code Snippet: LIME IMPLEMENTATION**

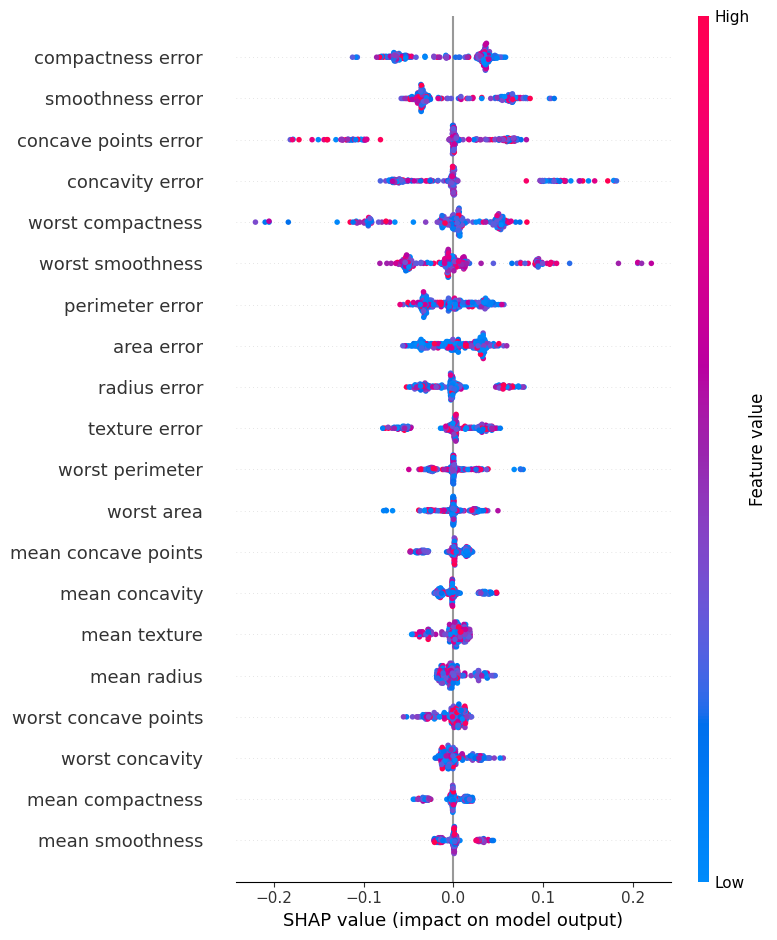
**Results:  
  
**

1. **Prediction Probabilities:**
   * The model predicts with high confidence (probability of 1.00) that the instance belongs to the class "setosa".
   * It assigns negligible probabilities (close to 0.00) to the classes "versicolor" and "virginica".
2. **Explanation:**
   * The LIME explanation highlights the most influential features that contributed to the model's prediction:
     + Petal Width (cm): The feature value of 0.20 indicates that the petal width is relatively small.
     + Petal Length (cm): The feature value of 1.40 suggests that the petal length is shorter.
     + Sepal Length (cm): The feature value of 5.10 indicates an average sepal length.
     + Sepal Width (cm): The feature value of 3.50 suggests a moderate sepal width**.**

**Sample Code Snippet: SHAP IMPLEMENTATION**

****

**RESULTS:**



**The key insights from this plot are:**

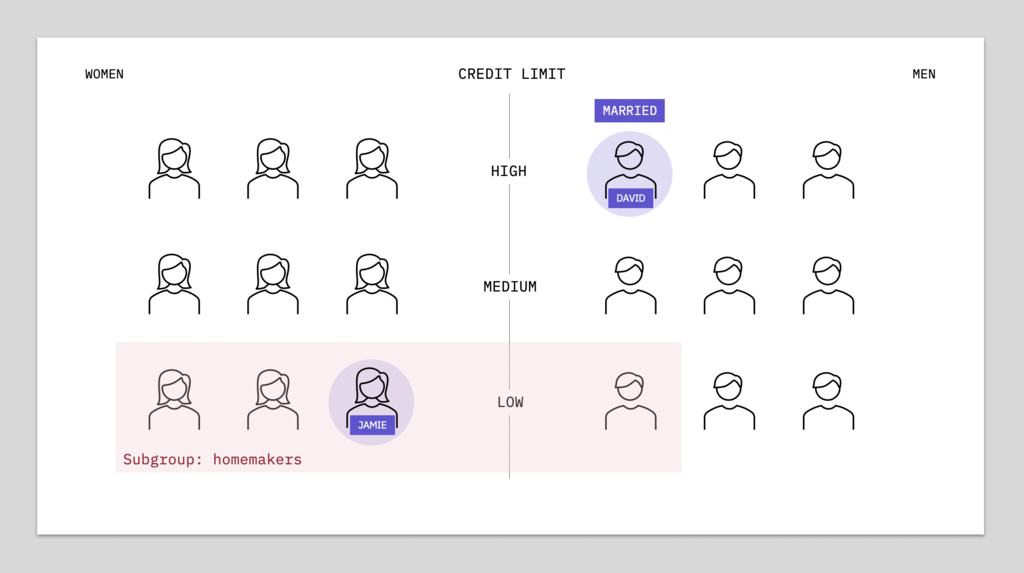
1. The feature with the highest positive impact on the model's output is "**high compactness error**", indicating that higher compactness error is associated with a higher predicted output (e.g., a higher likelihood of cancer).
2. Other features with a substantial positive impact include "smoothness error", "concave points error", and "concavity error".
3. Features with a negative impact include "mean concave points", "mean concavity", "mean texture", and "mean radius", suggesting that higher values of these features are associated with a lower predicted output (e.g., lower likelihood of cancer).
4. The features with the overall strongest positive and negative impacts are "high compactness error" and "mean concave points", respectively.

**The 2019 Goldman Sachs Incident :**

On November 7th, 2019, the creator of Ruby on Rails [tweeted that the Goldman Sachs Apple Card was sexist](https://twitter.com/dhh/status/1192540900393705474?s=21) because his wife received 1/20th of the credit limit he received. The tweet went viral, with many others saying the same thing, including Apple [co-founder Steve Wozniak](https://twitter.com/stevewoz/status/1193424787248279552). Just three days later, the New York State Department of Financial Services announced a legal investigation. One year later, if you search Goldman Sachs women credit, or Apple women and credit, the searches are full of allegations of discrimination. Thus the incident brought a significant reputational, legal, and financial cost for both companies.

Here’s a fictional example to illustrate one possible explanation of what may have happened.

Suppose we have an algorithm based largely on personal income. We have high, medium, and low credit limit groups. We constrain the algorithm to successfully ensure perfect parity between men and women. Now we consider fictional individuals David and Jamie. **David has a high personal income. Jamie has none.** Therefore our algorithm gives them very different credit limits.



**In the case where David and Jamie, who are married, share all household income and assets, including David's paid job and Jamie's role as a homemaker, their creditworthiness becomes very similar.**

**Considering the household's joint finances and shared assets, individuals in the homemaker subgroup, which is predominantly women, could reasonably expect fair treatment relative to their partners. However, if they face unequal treatment, such as in credit evaluations, they may perceive it as gender discrimination, potentially leading to public complaints, particularly on platforms like Twitter.**

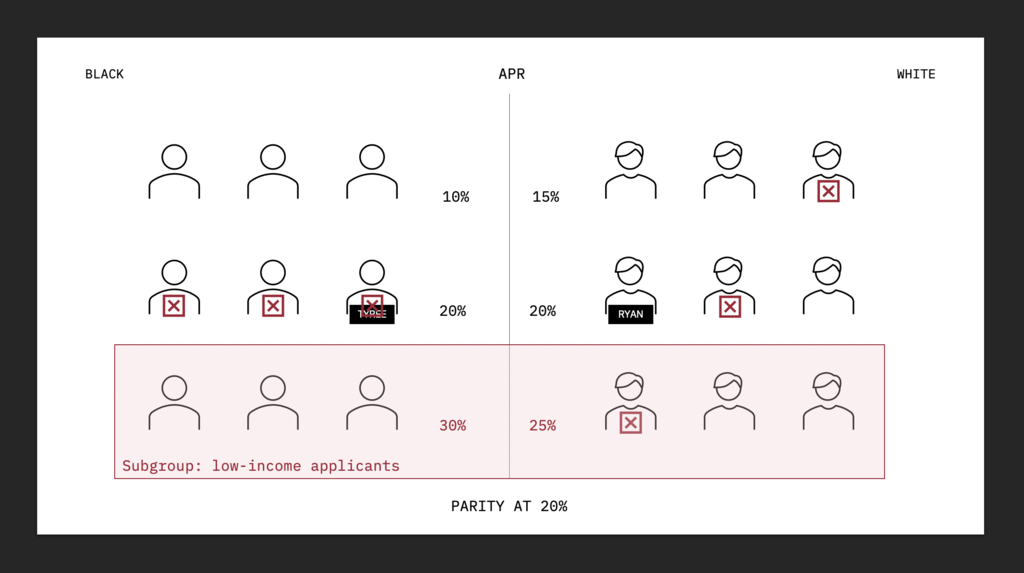
**---- x -----**

**Now, on the topic of racial discrimination:**

Let’s consider loan applications for Tyree and Ryan, which are names from a dataset of common African American and Caucasian names.

Let’s say Tyree and Ryan are very similar in every way except for race. Tyree is African American. Ryan is Caucasian. We run a fictional profit-maximizing algorithm that satisfies group fairness. Three applicants from each group are rejected, the others are approved. But look at Tyree and Ryan.

We’ve got a problem of counterfactual robustness here in which the model is not invariant to race, given that Tyree and Ryan are similar but experience different outcomes. Tyree was denied while Ryan was approved. That doesn’t seem fair at all.

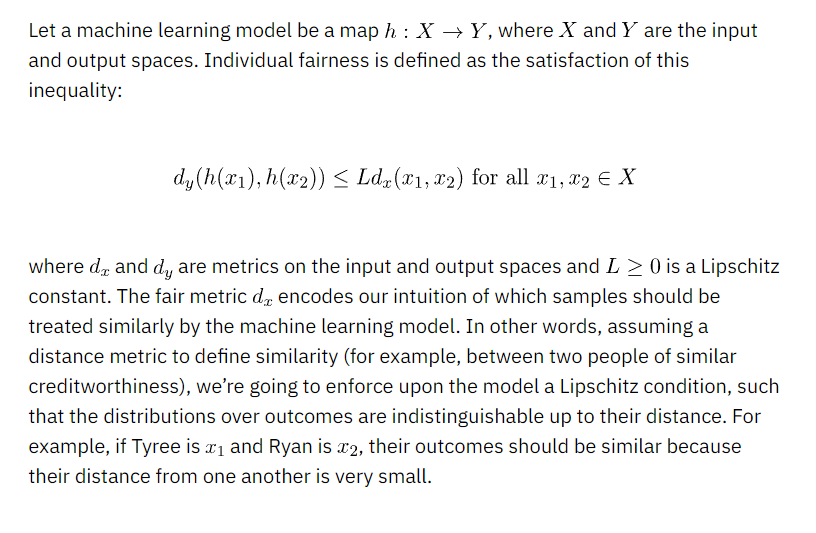


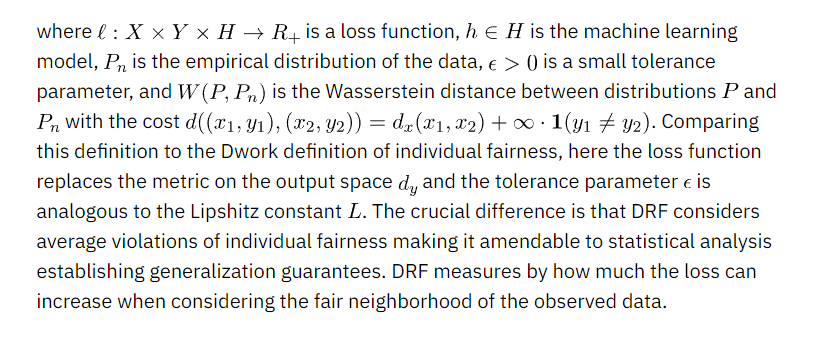
**Maybe this subgroup represents middle-income applicants and the algorithm has decided that for some reason black middle-income applicants aren’t worth the risk while white middle-income applicants are.**

To understand why this might be, let’s go deeper and look at interest rates. Here we’re assigning annual percentage rates to the upper, middle, and lower income groups. We’ve achieved group fairness between black and white borrowers with a shared average of 20%, but notice the discrepancy for **low-income applicants at the bottom.** Black borrowers are paying significantly more than white borrowers.

Introducing individual fairness

Because we can’t anticipate all possible subgroups, an approach called individual fairness emerges as the way we can deal with subgroup complexity to ensure that similar individuals like Tyree and Ryan are in fact treated similarly.



****