# FairSight: Visual Analytics for Fairness in Decision Making

Yongsu Ahn, Yu-Ru Lin

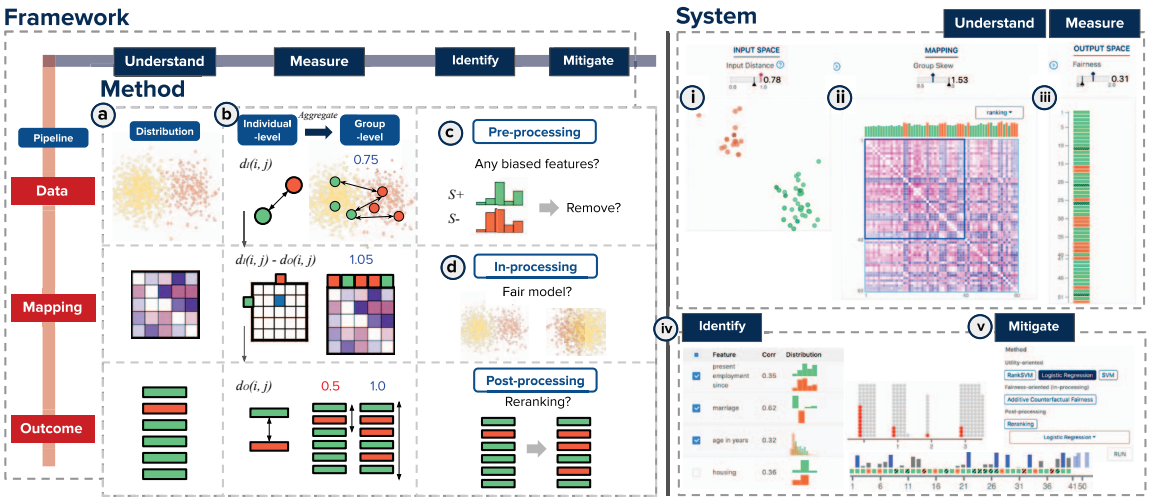


Fig. 1. We propose a design framework to protect individuals and groups from discrimination in algorithm-assisted decision making. A visual analytic system, *FairSight*, is implemented based on our proposed framework, to help data scientists and practitioners make fair decisions. The decision is made through ranking individuals who are either members of a protected group (orange bars) or a non-protected group (green bars). (a) The system provides a pipeline to help users understand the possible bias in a machine learning task as a *mapping* from the *input* space to the *output* space. (b) Different notions of fairness – *individual* fairness and *group* fairness – are measured and summarized numerically and visually. For example, the individual fairness is quantified by how pairwise distances between individuals are preserved through the mapping. The group fairness is quantified by the extent to which it leads to fair outcome distribution across groups, with (i) a 2D plot, (ii) a color-coded matrix , and (iii) a ranked-list plot capturing the pattern of potential biases. The system provides diagnostic modules to help (iv) identify and (v) mitigate biases through (c) investigating features before running a model, and (d) leveraging fairness-aware algorithms during and after the training step.

**Abstract**— Data-driven decision making related to individuals has become increasingly pervasive, but the issue concerning the potential discrimination has been raised by recent studies. In response, researchers have made efforts to propose and implement fairness measures and algorithms, but those efforts have not been translated to the real-world practice of data-driven decision making. As such, there is still an urgent need to create a viable tool to facilitate fair decision making. We propose *FairSight*, a visual analytic system to address this need; it is designed to achieve different notions of fairness in ranking decisions through identifying the required actions – understanding, measuring, diagnosing and mitigating biases – that together lead to fairer decision making. Through a case study and user study, we demonstrate that the proposed visual analytic and diagnostic modules in the system are effective in understanding the fairness-aware decision pipeline and obtaining more fair outcomes.

**Index Terms**—Fairness in Machine Learning, Visual Analytic

## 1 Introduction

Data-driven decision making about individuals has become ubiquitous nowadays. With the pervasive use of big data techniques, companies and governments increasingly rely on algorithms to assist in selecting individuals who meet certain criteria. In many cases, this process is conducted by first "ranking" the individuals based on their qualifications and then picking the top $k$ candidates based on the available resources or budgets. These ranking-based decision processes that concern rank-ordering individuals by their likelihood of success or failure have been widely adopted in many domains ranging from policing, recidivism, to job recruiting, and credit rating, which has a great impact on individuals' lives [8].

A critical issue of data-driven decision making is the possibility of intentionally or unintentionally discriminating against certain groups or individuals. While decision makers try to best utilize available information, including personal profiles such as race, sex, and age, the increasing cases of discrimination in the use of such personal profiles has been reported in real-world decision making. For example, a recent news [34] reported that Amazon's recruiting tool, trained from a 10-year historically male-dominant resume dataset, has been found biased in favor of men. Propublica [33] also reported that the recidivism scores learned from the algorithms tended to assign a higher score to an African-American defendant than to a White defendant who has been convicted of the same degree of crime. As shown in these real-world incidents, data-driven decisions are not free from existing bias. It has been pointed out that algorithmic decision making from data with inherent biases are just nothing but a systematic way of disseminating such biases to large number of people at once [30].

Data-driven decisions are criticized not only for being biased but also for lack of explanations. Even not limited to fairness problem, recent studies are increasingly aware that a black-boxed machine learning model lacks the explanation on predicted outcome [18]. The machine


- *Yongsu Ahn is with University of Pittsburgh. E-mail: yongsu.ahn@pitt.edu.*
- *Yu-Ru Lin is with University of Pittsburgh. E-mail: yurulin@pitt.edu.*


*Manuscript received 31 Mar. 2019; accepted 1 Aug. 2019.*
*Date of publication 16 Aug. 2019; date of current version 20 Oct. 2019.*
*For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.*
*Digital Object Identifier no. 10.1109/TVCG.2019.2934262*

learning models in societal decision making have assisted in judging whether individuals are qualified or not, where any results with greater performance metrics tend to be accepted without carefully examining *why*.

Fair and transparent machine learning in the real-world practice of decision making is in an urgent need; however, there is a lack of viable tools available to assist data science practitioners and decision makers in tackling the fairness problem. A variety of disciplines have made progress in developing fair algorithms and measures, but those are developed separately from decision-making contexts and not available in practice. While new tools became available recently [15, 22], none of these provide a comprehensive view and workflow to better cope with various fairness issues in the decision-making pipeline. With advancing research on measures, algorithms, and diverse perspectives on fairness, we now move one step further: to propose a viable decision-making tool to assist in fair decision making throughout the machine learning pipeline.

We argue that it is time to bring the research into real-world practice to create an impact on societal decision-making. In this paper, we present a fair decision-making framework, called *FairDM*, that identifies a set of guidelines in the algorithmic-aided decision making workflow. In this work, we focus on the problem of various high-stakes decision-making process, such as credit rating and recidivism risk prediction, which involve rank-ordering individuals. Moreover, a variety of prediction problems, such as binary or multiclass classification/prediction, can be cast as a ranking problem. The proposed *FairDM* is a model-agnostic framework that does not depend on a particular (ranking) algorithm, and it aims to provide a fairness pipeline to guide the examination of fairness at each step (from input to output) in the workflow. We develop *FairSight*, a visual analytic system that integrates the *FairDM* framework and analytic methods into a viable tool for fair decision making. Our main contributions include:

- **Fair decision making framework (Fig. 1).** We propose *FairDM* framework that facilitates the contemplative decision-making process [31] with a set of tasks to achieve fairer decision making. Our framework incorporates the different notions of fairness (including group and individual fairness) to support understanding, measuring, identifying and mitigating bias against certain individuals and groups.
- **Fairness measures and methods for explainability.** We introduce a set of measures and methods to summarize the degree of bias, evaluate the impact of features leading to bias, and mitigate possible sources of bias. Our approach supports both global- and instance-level explanation for the reasoning behind the fairness of ranking decision.
- **Fair decision making tool (Fig. 1).** We develop a viable fair decision making system, *FairSight*, to assist decision makers in achieving fair decision making through the machine learning workflow. We introduce a novel representation that visualizes the phases in a machine learning workflow as different spaces where individuals and groups are mapped from one to another.
- **Evaluation.** We present a case study to showcase the effectiveness of our system in real-world decision-making practice. Moreover, we conduct a user study to demonstrates the usefulness and understandability of our system in both an objective and subjective manner. Our study suggests that *FairSight* has a superior advantage over an existing tool [15].

## 2 Related Work

### 2.1 Fair Ranking

Today's decision making has increasingly relied on machine learning algorithms such as classification and ranking methods. We mainly discuss ranking in decision making due to its broad applications. In fair ranking, early studies mainly focused on quantifying the discrimination with proposed ranking measures in the top-$k$ list [32], or indirect discrimination [43]. Recently, fair ranking methods have been proposed [2, 41, 23]. Asudeh et al. [2] scored items based on a set of desired attribute weights to achieve fairness. On the other hand, Karako and Manggala [23] presented a fairness-aware Maximal Marginal Relevance method to re-rank the representation of demographic groups based on their dissimilarity as a post-hoc approach. Zehlike et al. [41] also proposed a re-ranking method of picking candidates from the pools of multiple groups with the desired probability. Online ranking systems, such as search engines or recommender systems, use ranking algorithms to generate an ordered list of items such as documents, goods, or individuals. The fairness problem here is to pursue the degree of exposure and attention fairly for groups [37, 39] or individuals [4].

In this work, we go beyond the issue of fair exposure/attention in ranking systems and broadly consider more broadly how a system can and should best help decision makers to rank items fairly when considering the trade-offs among different notions of fairness and utility.

### 2.2 Explainable Machine Learning

Machine learning and AI approaches have been recently criticized for the lack of capability in reasoning and diagnosing the logic behind the produced decisions [18]. With the increasing awareness associated with this problem, explainable machine learning techniques have been proposed. A number of studies have focused on interpreting the interaction between inputs and predictions from the original model, by training a secondary interpretable model to capture instance-level [35] or global-level pattern [1, 14]. For example, feature-level auditing methods seek to analyze the feature importance by permutation [11] or quantifying feature interaction [6, 13, 25]. Instance-level explanations that identify instances such as counterfactual examples or prototypes [24] seek to generate an explanation with a single instance. Recent research in visual analytics integrated machine learning tools with intuitive and manipulable representation and interface. Examples include RuleMatrix [29]'s rule-based visual interface for explaining decision rules based on the secondary decision tree model, the distribution-based visual representation for global-level explanation, and Rivelo's instance- and feature-level representation [38]. Mainfold [42] suggested a model-agnostic framework to interpret the outcome, inspect a subset of instances, and refine the feature set or model, to facilitate the comparison of models.

None of the aforementioned approaches have addressed the explainability with respect to fairness. In this work, we leverage state-of-the-art techniques, including feature auditing, to capture the feature-induced bias at both global and instance levels. We further propose new metrics via neighborhood comparison to capture both the global- and instance-level fairness with evidence of potential unfair outcomes.

### 2.3 Frameworks for Promoting Fair Machine Learning

Fair decision-making aid is an emerging topic. With the increasing awareness of the importance of fair machine learning, a number of new tools, including API [3] and interface [40], as well as integrated systems such as the What-if tool [15] or AI Fairness 360 [22], have been developed. While these tools offered a combination of explainable machine learning techniques and fair algorithms, none of them provides a comprehensive guideline to help users take proper actions to address various fairness issues throughout the machine learning decision pipeline. In contrast, *FairSight* is developed based on the *FairDM* framework, with a goal to empower users with a better understanding of various potential biases and a suite of tools to identify and mitigate the biases. In our evaluation study, we compare *FairSight* with the What-if tool and demonstrate several strengths of our design.

## 3 Fair Decision Making Framework

In this section, we present *FairDM*, a decision-making framework that aims to support a better understanding of fair decision-making processes. We consider such a process as a series of required actions that decision makers need to take in order to ensure the fairness of the decision-making process and outcome is in check as shown in Fig. 1. We start by formulating the top-$k$ ranking problem, followed by elaborating on the stages and the rationale behind them.

### 3.1 Top-$k$ Ranking Problem

In the framework, we assume that a decision maker requires to select the best top-$k$ individuals in the pool of $n$ candidates $C = \{1, 2, 3, ...n\}$.

The goal is to rank the $n$ individuals by the probability of being classified as qualified (positive) through learning a predicted value $\hat{y}_i$, where $\hat{y}_i$ represents the predicted level of qualification for an individual $i$. The learning is based on a subset of $p$ features $\mathbf{X}' = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$ selected from full feature set $\mathbf{X}$. The final outcome is an ordered list of ranked individuals $R = \langle r_i \rangle_{i \in C}$.

The following concepts will be involved when discussing the fairness in ranking decisions. A *sensitive attribute* is a characteristic of a group of individuals (e.g., gender or race) where, within existing social or cultural settings, the group is likely to be disadvantaged/discriminated and hence needs to be protected (as often regulated by a non-discrimination policy). We refer to a *protected* group, denoted as $S^+$, as a group that is likely to be discriminated in the decision-making process, and we refer to the remaining as *non-protected* group, denoted as $S^-$. In addition, a *proxy variable* is a variable correlated with a sensitive attribute whose use in a decision procedure can result in indirect discrimination [9].

### 3.2 Machine Learning Pipeline

We consider the decision making process as a simple machine learning pipeline consisting of three phases: Data, Model, and Outcome. The primary machine learning task in this context is to select the top-$k$ candidates based on available features.

**Data.** Given a feature set $\mathbf{X}$, a decision maker selects a set of features $\mathbf{X}' \subset \mathbf{X}$ to represent the qualification of candidates and seeks to learn the candidates' true property (e.g., qualified or not), denoted as a target value $y$. Each individual $i$ is represented by a set of qualification information, $\mathbf{X}'_i = \{\mathbf{x}_{i1}, ..., \mathbf{x}_{ip}\}$, and a target value $y_i$.

**Model.** A machine learning model is a function $f(\mathbf{X}')$ that maps the individuals' features to the ranking outcome.

**Outcome.** The outcome of the machine learning task is a ranking $R = \langle r_1, r_2, ..., r_n \rangle$, where $n$ individuals are ordered by their predicted qualification.

### 3.3 Fairness Pipeline

Given the machine learning pipeline, we propose a comprehensive workflow that consists of required actions to be supported by a fair decision making tool. For all three phases in the machine learning pipeline, a fair decision making tool should support the four fairness-aware actions: (1) Understand how every step in the machine learning process could potentially lead to biased/unfair decision making, (2) Measure the existing or potential bias, (3) Identify the possible sources of bias, and (4) Mitigate the bias by taking diagnostic actions. We provide the rationale for each action in the following.

#### 3.3.1 Understand

The first action of the fairness pipeline is to clearly understand the machine learning process and its consequences to fairness in decision making. The challenge is how to facilitate such an understanding as many practitioners do not fully recognize how every step in the process could potentially lead to biased decision making [21]. To address this, we propose that a fair decision-making tool should take proactive action to help decision makers understand the possible unfairness at each machine learning stage, by providing an overview with a step-by-step workflow to guide users to examine different notions of fairness.

#### 3.3.2 Measure

With an overall understanding of various fairness issues, the next step is to quantify the degree of fairness and utility and evaluate how each machine learning phase impacts fairness. While many studies working on proposing the measures primarily focus on measuring the outcome bias, we argue that quantifying bias throughout all phases of the pipeline should be made available to users to detect not only *consequential* bias in Outcome, but also *procedural* bias in Data and Model phase as well.

#### 3.3.3 Identify

Upon understanding and measuring bias, decision makers need to remove the potential bias. As a crucial step to achieve this, we emphasize the importance of identifying bias from features in the dataset. In data-driven decision making, feature selection is an important step that captures information based on which individuals' qualification should be evaluated. Feature inspection tools should help identify potential bias with respect to not only the sensitive attribute but also the likely proxy variables.

In our framework, we incorporate per-feature auditing modules to investigate bias being involved in all phases of the machine learning process: pre-processing, in-processing, and post-processing bias within features.

#### 3.3.4 Mitigate

Informed by the aforementioned diagnostic actions, Mitigate is where decision makers take actions to remove bias within the machine learning pipeline. We consider mitigating the bias in each of the following phases: (1) Data/pre-processing: How does one incorporate fairness-aware feature selection? (2) Model/in-processing: How does one select machine learning model with less bias? (3) Outcome/post-processing: How does one adjust ranking outcome to make it fairer?

## 4 Method

Based on the *FairDM* framework, we propose *FairSight* to enable fair decision making in machine learning workflow. This section presents our analytical methods to support each of four required actions (Understand, Measure, Identify, and Mitigate), and the system design will be introduced in the next section. Table. 1 provides a summary of the requirements, tasks, and corresponding methods.

### 4.1 Understanding Bias

*FairSight* seeks to facilitate an effective understanding of machine learning process by introducing a novel "*space-mapping*" representation. Inspired by Friedler et al. [12], we consider the Data and Outcome phases as two *metric spaces* (Input and Output spaces), and the machine learning model as "*the interactions between different spaces that make up the decision pipeline for a task*" (Fig. 1a). Then, biases can be introduced in each space or through the mapping between two spaces.

**Input space.** We denote the Input space as $\mathcal{I} = (\mathbf{X}, d_{\mathbf{X}})$, where $\mathbf{X}$ is the feature space and $d_{\mathbf{X}}$ is a distance metric defined on $\mathbf{X}$.

**Output space.** The Output space is noted as $\mathcal{O} = (O, d_O)$, where $O$ is an ordered list where individuals are ranked by a decision process, and $d_O$ is a distance metric defined on $O$. Both ranking and classification algorithms may be involved in such a process. When a classification model is used, the probability of each instance being classified as positive can be used to generate the ranking.

**Mapping.** A machine learning model is a map $f : \mathcal{I} \rightarrow \mathcal{O}$ from the Input space $\mathcal{I}$ to the Output space $\mathcal{O}$.

### 4.2 Measuring Bias

*FairSight* provides a comprehensive set of measures to support the Measure requirement (Fig. 1b) covering the following aspects: (1) These measures are defined and organized consistently with the aforementioned the space-mapping representation (Section 4.1). (2) It covers different (complimentary) notions of fairness, including *individual* and *group* fairness (details below). (3) The fairness individuals and groups are measured at both instance-level and global-level to enable examining the detailed and summative evidence. (4) In addition to fairness measures, we also introduce utility measures in the context of the ranking decision that allows for trade-off comparison. We first describe the metrics defined in spaces and then present the various measures.

#### 4.2.1 Distance and Distortion

Distance and distortion are two fundamental notions in *FairSight*. Given two individuals $i$ and $j$, a pairwise distance $d(i, j)$ indicates how the two individuals are dissimilar to each other in a space. When two individuals are mapped from one space to another, the pairwise distance may not be preserved. The distortion is defined as the discrepancy between two pairwise distances, indicating the degree of which the mapping is not preserved. We incorporate two different distance metrics for the Input and Output spaces. For Input space, we calculate the Gower distance $d_{\mathcal{I}}(\cdot)$ [16] between $\mathbf{X}'_i$ and $\mathbf{X}'_j$, which measures the pairwise distance of the two individuals' feature representations. For Output space, the pairwise distance is computed using the absolute ranking difference $d_{\mathcal{O}}(\cdot)$ of two individuals $r_i$ and $r_j$. Then, the

| Pipeline | Task | Data (D) | Model (M) | Outcome (O) |
|---|---|---|---|---|
| **Understand (U)** (sec 4.1) | Can we understand how individuals and groups are distributed? | Input space | Mapping space | Output space |
| **Measure (M)** | Can we measure and summarize biases? | | | |
| Instance-level (sec 4.2.1) | Is an individual not treated similarly?<br>Is the distortion advantaged or disadvantaged? | - | rNN (eq. 2)<br>rNN-gain (eq. 3) | - |
| Global-level | | | rNN-S+,- (eq. 5) | |
| Individual fairness (sec 4.2.2) | Are similar individuals treated similarly? | - | rNN-mean (eq. 4) | - |
| Group fairness (sec 4.2.3) | Are groups treated equally? | Group Separation (eq. 5) | Group Skew (eq. 6) | Between: GFDCG (eq. 8)<br>Within: Statistical parity |
| Utility (sec 4.2.4) | How is the model accurate? | - | - | Between: utility@k (eq. 9)<br>Within: Precision@k |
| **Identify (I)** (sec 4.3) | Can we identify features as a potential source of bias? | Feature correlation | Feature distortion | Feature perturbation |
| **Mitigate (MT)**(sec 4.4) | Can we mitigate biases? | Feature selection | Fair algorithm (ACF) | Reranking algorithm (FA*IR) |

Table 1. *FairDM* tasks & *FairSight* measures

distortion of two spaces is computed as the absolute difference of two pairwise distances between the Input space and Output space, as:

$$distortion_{(i,j)} = |d_{\mathcal{I}}(\mathbf{X}'_i, \mathbf{X}'_j) - d_{\mathcal{O}}(r_i, r_j)|, \tag{1}$$

### 4.2.2 Individual Fairness

By the definition of individual fairness, "*similar individuals should be treated similarly*" [10], we measure individual fairness based on the degree to which the pairwise distances in Input space is preserved in Output space through the mapping, i.e., based on the notion of distortion (Section 4.2.1). *FairSight* provides both instance- and global-level examinations for individual fairness as follows.

**Instance-level bias.** Instance-level bias is measured as the amount of distortion with respect to an individual compared with other similar individuals. We capture the "similar individuals" of an individual $i$ based on $i$'s $h$ nearest neighbors (i.e., the closest neighbors in the Input space), denoted as $NN_h$ ($h = 4$ in this work). Then, fairness with respect to an individual $i$ is measured based on how $i$ and the nearest neighbors in Input space are close to each other in the Output space (based on the ranking outcome), as follows:

$$rNN(i) = 1 - \frac{1}{h} \sum_{j \in NN_h(i)} \frac{|r_i - r_j|}{|C|}, \tag{2}$$

where the absolute ranking difference is normalized both by the number of individuals $|C|$ and the number of nearest neighbors $h$.

While $rNN$ quantifies the degree of bias/fairness with respect to an individual, the individual may be disadvantaged (i.e., ranked much lower than the neighbors), or advantaged (i.e., ranked much higher than the neighbors). To understand the differences in bias, a signed ranking distortion for an individual $i$ is defined as:

$$rNN_{gain}(i) = 1 - \frac{1}{h} \sum_{j \in NN_h(i)} \frac{(r_i - r_j)}{|C|}. \tag{3}$$

**Global-level bias.** The global-level measure of individual fairness can be obtained by aggregating (averaging) over all instance-level measures, as:

$$rNN_{mean} = \frac{1}{|C|} \sum_{i \in C} rNN(i). \tag{4}$$

### 4.2.3 Group Fairness

Group fairness relates to equalizing outcomes across the protected $S^+$ and non-protected $S^-$ groups in Data and Outcome phase, and the mapping between two spaces. The analysis of group fairness can be richer than that of individual fairness because, by definition, individual fairness concerns "*similar treatment for similar individuals*", which indicates the consistency in the mapping between the Input and Output spaces, while group fairness concerns the distribution across groups, in terms of data representation (data bias), mapping data to outcome (mapping bias), and prediction (outcome bias), as detailed below.

**Data bias.** Data bias regarding group fairness seeks to uncover any bias already inherent in the input dataset. It is captured by the degree of separation between the two groups in the Input space – as ideally, the group membership should not be uncovered until revealing the sensitive attribute. Here, we adopt the symmetric Hausdorff distance [19], referred to as *Group Separation* score, to measure the separation between the individuals from the two groups A and B:

$$h(A, B) = max(\tilde{h}(A, B), \tilde{h}(B, A)), \tag{5}$$

where $\tilde{h}(A, B) = max_{a \in A}\{min_{b \in B}\ d(a, b)\}$ is the one-sided Hausdorff distance from $A$ to $B$.

**Mapping bias.** Mapping bias regarding group fairness seeks to uncover any unfair distortion between the Input and Output spaces at the group level. It is defined based on comparing the pairwise distortions, $distortion(i, j)$, between the two groups (for $i \in S^+$ and $j \in S^-$) against the distortion within the groups (for $i, j \in S^{(\cdot)}$). For example, when the pairwise distances between Men and Women are distorted (i.e., when greater between-group distortion is observed), the mapping has a systemic, or *structural* bias. We adopt the *Group Skew* concept [12] to measure such bias as:

$$Group\ Skew = \frac{Distortion_{BTN}}{Distortion_{WTN}}, \tag{6}$$

where $Distortion_{BTN} = \sum_{i \in S^+, j \in S^-, i \neq j} distortion(i, j)$ and $Distortion_{WTN} = \sum_{i,j \in S^{(\cdot)}, i \neq j} distortion(i, j)$.

**Mapping bias per group.** The group-specific mapping bias can be quantified based on how individuals in the group receive the mapping bias (ref. Equation 2). It is thus defined by averaging the mapping bias of individuals in either the protected group $S^+$ or non-protected group $S^-$:

$$\begin{aligned} rNN_{S^+} &= \frac{1}{|S^+|} \sum_{i \in S^+} rNN(i), \\ rNN_{S^-} &= \frac{1}{|S^-|} \sum_{i \in S^-} rNN(i). \end{aligned} \tag{7}$$

**Outcome bias.** Outcome bias regarding group fairness should capture how decision outcomes are fairly distributed across groups. In the context of ranking decisions, fairness should be evaluated between different rankings (different ordered lists). Furthermore, fairness should also be evaluated based on the choice about the top-$k$ threshold within a given ranking list (i.e., to choose the most qualified $k$ individuals from the list). A popular method for comparing rankings is $nDCG$ [7], which involves logarithmic discount to favor items at the top ranking positions. While this method has been widely adopted in online ranking systems to reflect users' scarce attention toward only a limited few top items, in *FairSight*, our primary concern is about whether an individual being ranked on the top-$k$ is fair or not, rather than how much attention the individual received from the ranking position. Therefore, we consider a linear rather than a logarithmic discount in order to differentiate rankings with different orders without heavily favoring a very few top items. Our measure, called *GFDCG* (Group-Fairness DCG) is defined based on comparing the quality of the top-$k$ ordering in one group against another, as:

$$GFDCG = \frac{linear\ DCG@k(S^+)}{linear\ DCG@k(S^-)}, \tag{8}$$

where $linear\ DCG@k(S^{(\cdot)}) = \sum_{i \in S^{(\cdot)} \cap R^{(k)}} y_i \times \frac{n-r_i}{n}$, $R^{(k)}$ is the top-$k$ list truncated from the whole ranking list $R$, $r_i$ is the ranking position of an individual $i$, $y_i$ is the true qualification of $i$, and $n$ is the total number of individuals in the dataset.

While $GFDCG$ is useful for comparing a different threshold $k$ within the same ranking (referred as *within-ranking* comparison), it is less effective in comparing different rankings when $k$ is large due to the discounting effect. Therefore, to compare the fairness between different rankings, we adopt the statistical parity, which calculates the ratio of two groups without ranking discount.

#### 4.2.4 Utility

In this work, we consider the utility of a ranking as the quality of the ranking. Similar to the GFDCG, the quality is evaluated based on the extent to which the top-$k$ ordering captures truly qualified individuals, and thus a linear discount is also adopted. The utility is defined by comparing the *linear DCG* from the predicted top-$k$ ordering against the *ideal* top-$k$ ordering (*IDCG*):

$$utility@k = \frac{linear\ DCG@k}{linear\ IDCG@k}, \tag{9}$$

where $linear\ DCG@k = \sum_{i \in R^{(k)}} y_i \times \frac{n-r_i}{n}$, $linear\ IDCG@k = \sum_{i \in R_I^{(k)}} y_i \times \frac{n-r_i}{n}$, $R^{(k)}$ is the top-$k$ list from the predicted ranking and $R_I^{(k)}$ is the top-$k$ list from the true ranking (based on individuals' true qualifications). In the same manner, we define the within-ranking utility measure by adopting the widely used information retrieval measure $Precision@k$.

### 4.3 Identifying Bias

*FairSight* provides three different strategies in each of the machine learning phases to detect possible biases due to feature selection: (1) Data: feature correlation between sensitive attribute and other features, (2) Model: per-feature impact on outliers that received most distortions from the model, and (3) Outcome: feature importance by ranking perturbation.

**Feature correlation.** At the Data phase, feature correlation analysis offers a way to detect bias regarding group fairness, by removing the information that can reveal the group membership of individuals. In *FairSight*, a highly correlated feature to a sensitive attribute is detected by comparing the distributions of the feature values by groups. If two distributions from the protected and non-protected groups are very distinct, the features are subject to be used as a proxy for the sensitive attribute. In this work, we compute the difference between the two distributions using Wasserstein distance [36], which measures the transportation cost from one distribution to another. The greater the distance, the more likely the feature can be used to distinguish the two groups and can lead to indirect discrimination.

**Feature impact on outlier distortions.** At the Model phase, outlier distortion analysis measures how a feature is correlated with the overall distortion from the mapping. To compute this, a distortion distribution is first generated from the distortions of all instances. Outliers are those having greater distortions than other instances (the right tail of the distribution). For a given feature, the distance between the distortions received by the outliers and by the rest of individuals are computed to reveal the impact of the feature on the outlier distortions. Here, the Wasserstein distance is used to compute the distance between the two distortion distributions.

**Feature perturbation.** At the Outcome phase, the feature importance of each feature to the fairness and utility of ranking outcome is analyzed. We adopt the feature perturbation method [5], a widely-used feature auditing technique to evaluate the feature impact on a model. To compute this, we permute the values of a feature $\mathbf{x}_q$ into $\tilde{\mathbf{x}}_q$ from the selected feature set $\mathbf{X}'$ and re-train the model with $\mathbf{X}'_{\tilde{q}} = \left[\mathbf{x}_1, \ldots, \tilde{\mathbf{x}}_q, \ldots\right]$. The permutation is performed by swapping the feature values $\{\mathbf{x}_{q1}, \mathbf{x}_{q2}, ..., \mathbf{x}_{q(2/n)}\}$ with $\{\mathbf{x}_{q(2/n+1)}, ..., \mathbf{x}_{qn}\}$ as suggested by [11]. Then, the impact of the feature is measured by how much the fairness and utility measures (ref. Equation 8 and 9) drop compared with the model with the original non-permuted features.

### 4.4 Mitigating Bias

*FairSight* provides three types of methods for reducing biases in each of the three phases in the machine learning pipeline: pre-processing, in-processing, and post-processing methods.

**Pre-processing.** Pre-processing method can be considered as a pre-emptive action to make the feature selection step as free from bias as possible. To achieve group fairness, decision makers should avoid using the sensitive attribute as part of the selected features. In addition, any other features that are highly correlated with any of the sensitive attributes, if used intentionally or unintentionally, can lead to indirect discrimination and should be avoided as much as possible. Such features can be detected during the Identify stage as described in Section 4.3.

**In-processing.** This method seeks to mitigate bias by selecting fair machine learning algorithms. *FairSight* incorporates Additive Counterfactually Fair (ACF) model [26]. ACF assumes that the counterfactual probabilities of two individuals choosing from either group should be the same with respect to a given (non-sensitive) feature. We also provide a plug-in architecture in *FairSight* to allow a variety of fairness-aware algorithms to be added in the system, which can be utilized to compare multiple fair algorithms and to choose one that better mitigates the bias while having a high utility value.

**Post-processing.** Here, we aim to achieve a fair ranking outcome independently of the Data and Model phases by adjusting the ranking. This approach is especially useful in situations where decision makers do not have full control of phases before the outcome. With access to the ranking outcome, a post-processing method provides a safeguard against biases in the outcome. *FairSight* incorporates a fair ranking algorithm proposed by [41], which re-ranks a ranking list by randomly choosing an individual from either of two group rankings with a pre-defined probability. Other re-ranking algorithms can also be included through the plug-in architecture in the system.

## 5 Design Goals

We identify a set of requirements based on *FairDM* that integrates the aforementioned methods with relevant tasks.

**R1. Enable examining different notions of fairness in the data-driven decision with consistent visual components and interactive tools.**

- **T1.** A fair decision making tool should enable users to select a sensitive attribute and a protected group to pursue different notions of fairness, including individual fairness and group fairness. The system also should provide an integrated interface and intuitive representation consistent with various fairness notions.

**R2. Facilitate the understanding and measuring of bias.**

- **T2.** The system should help users to understand the distribution of individuals and groups to figure out whether or to what extent each phase of machine learning process is biased.
- **T3.** The system should provide the degree of fairness and utility in a summary to help users to obtain fair decisions among the trade-offs.
- **T4.** The system should enable the instance-level exploration to help understand the reason behind how individuals and groups are processed fairly/unfairly.

**R3. Provide diagnostic modules to identify and mitigate bias.**

- **T5.** The system should support feature evaluation with respect to fairness in three machine learning phases. This task seeks the evidence of features selection in pre-processing the data.
- **T6.** The system should allow users to mitigate the bias to obtain better rankings. This module also should provide the mitigating methods in all three phases to achieve the procedural fairness.

**R4. Facilitate the comparison of multiple rankings to evaluate ranking decisions.**

- **T7.** The system should allow users to repeat the process to generate multiple rankings and evaluate the trade-off between fairness and utility across them.

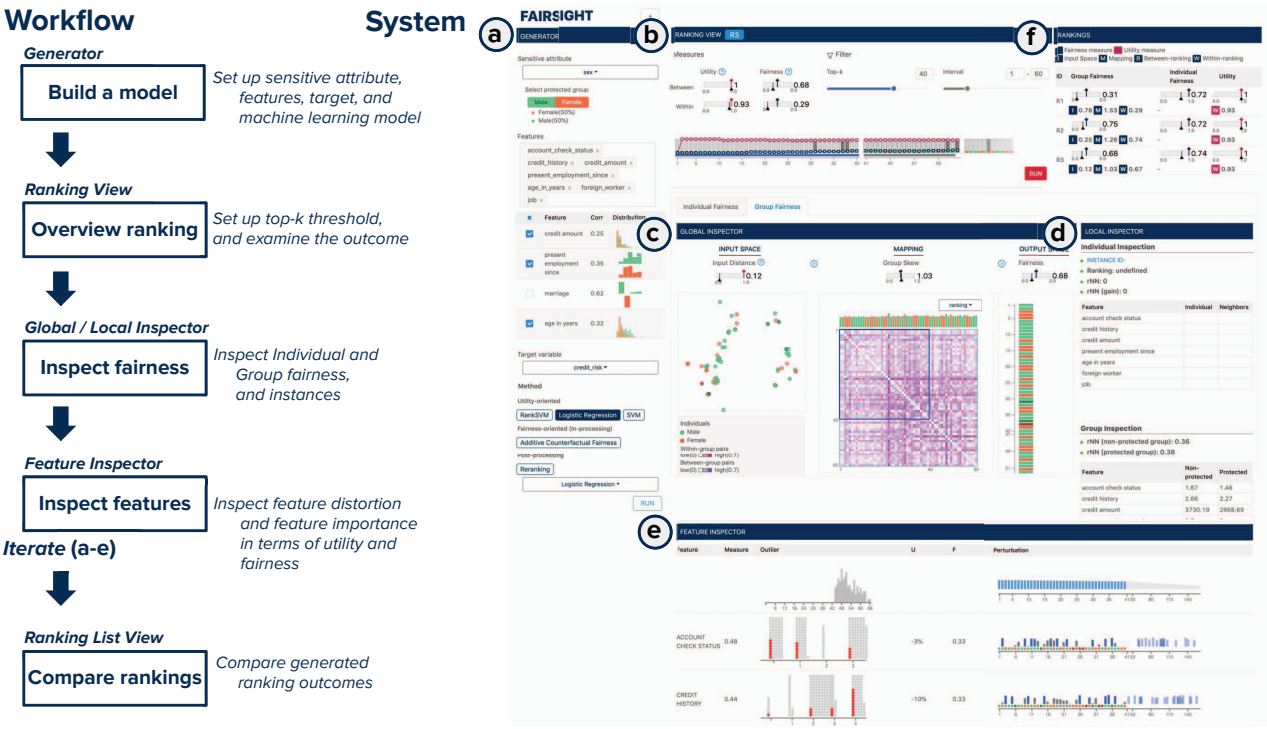


Fig. 2. The workflow of fair decision making in *FairSight*. (a) It starts with setting up inputs including the sensitive attribute and protected group. (b) After running a model, the ranking outcome and measures are represented in Ranking View. (c) Global Inspection View visualizes the two spaces and the mapping process of Individual and Group fairness provided in the separate tap. (d) When an individual is hovered, Local Inspection View provides the instance- and group-level exploration. (e) In Feature Inspection View, users can investigate the feature distortion and feature perturbation to identify features as the possible source of bias. (f) All generated ranking outcomes are listed and plotted in the Ranking List View.

## 6 FairSight - System Overview

In this section, we discuss the system architecture of *FairSight* and present the major six visual analytic components.

### 6.1 Generator

Generator allows users to set up all required inputs for fair decision making (Fig. 2a) (**R1**). Users can start the setting with the selection of the sensitive attribute and protected group (**T1**). In the feature table, we provide the feature correlation measures (ref. Section 4) to aid the feature selection. Each feature has two bar charts which indicate the by-group feature value distribution (e.g., orange bar chart for Male, and green bar chart for Female), and the correlation measure (i.e., how the two distributions are dissimilar to each other). Users can scrutinize how each feature has a potential to be used as proxy variable of the sensitive attribute (e.g., 'marriage' of the feature table in Fig. 2a – where the distributions across groups are quite different and hence can be used as a proxy for the sensitive attribute) (**T5**).

Generator also provides a list of available machine learning algorithms, including popular classification and ranking models such as Logistic Regression, Support Vector Machine, and Ranking SVM. We also include the in-processing model (Additive Counterfactual Fairness (ACF)) [26]) and the post-processing method (FA*IR [41]) as a way of mitigating bias (**R3**). The system also has a plug-in architecture that allows users to flexibly include more machine learning algorithms.

### 6.2 Ranking View

Ranking View (Fig. 2b) provides an overview of the current ranking outcome. First, we report all outcome measures of fairness and utility, which includes between-ranking and within-ranking measures. While the between-ranking measures help determine whether the current ranking is better than other generated rankings, the within-ranking measures are useful to find the best top-$k$ threshold. Along with the top-$k$ slider is the interval slider to determine how many individuals are to be represented in Global Inspection View and Feature Inspection View. Ranking View also visually represents the current ranking outcome as shown in Fig. 3. Each individual is encoded as a bar, with the group membership as a small rectangle at the bottom. A bar is filled with diagonal patch if the individual has false target label (e.g., "not likely to pay back a loan" in credit risk prediction) from the target variable. On top of the bars are the trend line of within-ranking utility (dark red) and fairness (dark blue).

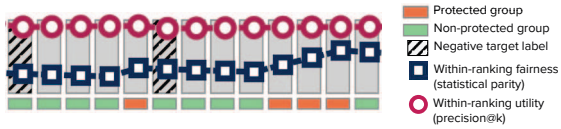


Fig. 3. The visualization of ranking outcome supports the exploration of cumulative ranking quality in terms of fairness and utility. The individuals are encoded as vertical bars with the group indicator at the bottom, with a blue rectangle (fairness) and a red circle (utility). It facilitates the selection of different $k$ values, by helping users recognize the score change as $k$ increases, as the trade-off between the fairness and utility.

### 6.3 Global Inspection View

Global Inspection View (Fig. 2c) provides an overview of the fairness in three phases of the machine learning process to help Understand and Measure bias (**R2**). We present two notions of fairness, Individual and Group fairness, in the separate tabs to represent individuals and groups with corresponding measures independently.

This view consists of the visual components of three spaces. Each space visualizes the distribution of individuals in each phase (**T2**). The Input Space View visualizes the feature representation of individuals as circles in a 2D plot using t-SNE [28]. For Mapping space, Matrix View represents all pairs of individuals in the mapping process with the amount of pairwise distortion between two spaces. As mentioned in Section 4.2.3, two kinds of pairs (between-group and within-group pairs) are colored as purple and pink. Darker colors indicate greater distortion, as opposed to no distortion with white color). Along with the two spaces and the mapping, fairness measures of each space are presented to provide the summary of bias in each phase (**T3**).

### 6.4 Local Inspection View

Local Inspection View (Fig. 2d) supports the exploration of information on a specific individual (**T4**). As shown in Fig. 4, when users hover an individual in any three spaces views of Global Inspection View, the individual is highlighted with black stroke, and its nearest neighbors with blue stroke. Local Inspection View displays the detailed information: Instance-level bias ($rNN$) and gain ($rNN_{gain}$), and feature value of the individual and its nearest neighbors. The feature table enables comparing the feature value of the individual, and the average feature value of nearest neighbors, so that users can do reasoning about what feature contributed to bias and gains. We also support the comparison of two groups. We provide group-level bias ($rNN_{S+}$ and $rNN_{S-}$) for users to compare which group has more bias, but also the average of feature values for each group to show the difference (Fig. 2d).

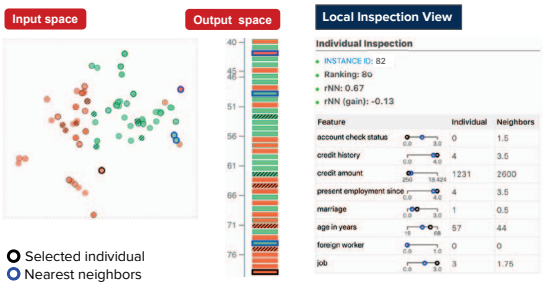


Fig. 4. Local inspection View. Once users hover over an instance (selected individual as black), nearest neighbors are also highlighted as blue. The feature table shows the difference of feature values (the individual's vs. the average of neighbors').

### 6.5 Feature Inspection View

Feature Inspection View (Fig. 2e) lists all selected features to support the identification of the feature bias in Model and Outcome phase (**T5**). It is composed of two components: Feature distortion and feature perturbation (ref. Section 4.3).

For the feature distortion, we plot the overall distribution of instances with respect to their distortions. We then identify outliers that have greater distortion within 5% of the right tail. For each feature, we represent the whole individuals (gray circle) with outliers (red circle) in a histogram along with feature correlation score. The more distinct two distributions are, with respect to certain feature, such feature is likely to be a source of bias.

Also, the result of feature perturbation for each feature is represented as the visual component of perturbed ranking in the feature table as shown in Fig. 5. Each individual is represented as a bar, which is ordered by after-perturbation ranking in the $x$-axis. To represent the ranking difference by perturbation, we color the individual bars based on whether they were previously in the top-$k$ (blue) or not (gray), and set the height as before-perturbation ranking. We also encode the information of group membership as a small rectangle (orange or green) and target label as a black striped patch (negative label). Any individuals who were in the top-$k$ in before-perturbation ranking are represented with a semi-transparent blue bar to indicate how they are ranked.

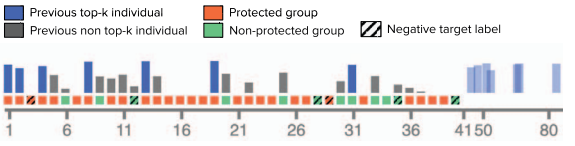


Fig. 5. Visual ranking component of feature perturbation. This view represents how the ranking changes after perturbation (x-axis) compared to one before perturbation (y-axis: the height of bars as previous ranking).

### 6.6 Ranking List View

The Ranking List View provides the summary of all generated ranking outcomes as shown in Fig. 2f. This view serves to compare the fairness and utility measures of rankings so that users can consider the trade-off between fairness and utility in their decision making. In the table, we list rankings which consist of Group fairness, Individual fairness, and Utility measures as columns. In particular, Fig. 2f presents a number of representative measures, one for each type of fairness and utility, via multiple linear gauge plots. Each of the plots consists of an ideal score for the corresponding measure (encoded by a diamond shape) and the current score (as a triangle marker). Additional fairness or utility measures are numerically presented (Fig. 2f).

## 7 Case Study

We present a case study of the loan approval decision to showcase the effectiveness of *FairSight* in achieving the fair data-driven decision (Fig. 6). In this scenario, a credit analyst, Julia, working for a credit rating company aims to pick the best qualified $k$ customers to grant a loan. She has a fairness goal to protect female customers as the persistent discrimination against women in financing has been reported.

**Settings.** We utilize the German Credit Dataset published in UCI dataset [20]. For this case study, we randomly select 250 instances with 10 features. We sample the same number of individuals from two groups (Men:Women = 5:5) and keep the ratio of target label (Credit risk: Yes or No) of each group the same as the original dataset. For the initial run, we select 9 features (Fig. 6g) with $k = 45$ as the top-$k$ threshold. We illustrate this case by showing how Julia iteratively went over the machine learning pipeline for six iterations. In the following, we use the abbreviated notation $S_F \times S_M$ for $S_F \in \{\mathsf{U}, \mathsf{M}, \mathsf{I}, \mathsf{MT}\}$ and $S_M \in \{\mathsf{D}, \mathsf{M}, \mathsf{O}\}$ to highlight how each action correspond to the machine learning and *FairDM* stages, e.g., U×D denotes the action of using tools at the Data phase to meet the Understand requirement.

**Iteration 1.** Julia started the initial run with 9 features (Fig. 6g) including "Sex" feature selected with Logistic Regression model. After the model running, she realized that the ranking outcome was significantly discriminated against women ($GFDCG$ = 0.3) in Ranking View (Fig. 6d-i1). When she took a closer look at the distribution of individuals, the ranking outcome was severely favorable towards men, especially within the top-15. In Global Inspection View, all Data and Model phases were biased. Specifically, two groups were clearly separated due to inherent group bias in Input space (U×D) (Fig. 6a-i1). In Mapping space, Group skew was over 1 indicating there is a structural bias (Fig. 6e-i1). There was also a gap between the amount of bias per group as well ($rNN_{S+}$: 0.4, $rNN_{S-}$: 0.32). By excluding the sensitive attribute from the decision making, she deleted the "Sex" feature and generated the second ranking (I×D).

**Iteration 2.** Without "Sex" feature involved, she checked that two groups are more scattered throughout Input space though she could detect two groups that formed the clusters ($Group\ separation$ = 0.27) (U×D) (Fig. 6a-i2). There was still inherent bias, so she decided to continue examining the other potential features that brings in bias in Data phase. While investigating Feature Correlation table (I×D) (Fig. 6b), she found that "Marriage" feature is highly correlated to the sensitive attribute with the score of 0.62 (M×D). The distribution plot showed that almost all men are in single status, whereas most women are in the status of married or divorced. She judged that "Marriage" feature could be a potential proxy variable to discriminate against certain group.

She also noticed in Input Space that there is an individual (Woman, ID: 82) plotted in a distance from other female individuals (Fig. 4). When she hovered over the individual, she found that the individual had the distortion ($rNN = 0.67$), but it turned out to be the disadvantage against the individual ($rNN_{gain}$ = -0.13). She investigated the feature value table to see how the individual is different from neighbors. She noticed that the individual was significantly different in "Account check status", "Marriage", and "Job". Especially, the individual had a significantly different account check status (0: No account), and was in "Married" status (Marriage = 1) compared to the average of neighbors' marriage status 0.5, which is closer to "Single" status (0: Single). This instance-level exploration enabled her to explore how each individual was biased or disadvantaged, with the difference of feature values explained. Taking all pieces of evidence from this iteration, she decided to remove "Marriage" feature (MT×D).

**Iteration 3.** She instantly noticed that removing the "Marriage" feature improved the fairness score. For Input space, Group separation score dropped to 0.12 (M×D) (R3 in Fig. 6h). In Input space, she

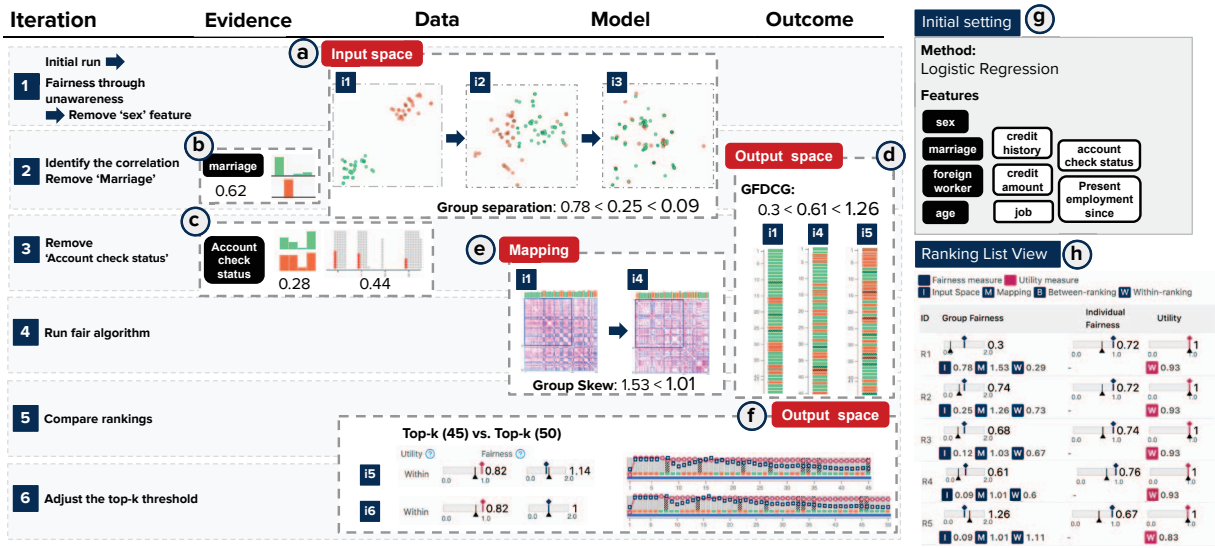


Fig. 6. Summary of case study in loan approval (g) with the initial features and method. Users can understand (a) the distribution of groups and individuals in Data phase. With the feature auditing modules to detect (b) feature correlation and (c) feature distortion, the visual components of (d) Matrix View in Mapping and (d) Ranking View in Outcome phase represent that the fairness in each phase improves. (f) Within-ranking module helps adjusting better top-$k$ threshold. (h) Ranking List View displays the trade-off between fairness and utility to help users comparing the rankings.

found that two groups were not clearly separated by their cluster anymore (Fig. 6a-i3). She also found that Group Skew score in Model phase improved by 0.05. But GFDCG score in Outcome phase was still severely biased towards men, which still left much room to improve (score = 0.53). At this time, she observed in Feature Inspection View that "Account check status" feature had the highest feature distortion (score = 0.44) and high feature correlation bias (score = 0.28) (Fig. 6c). She finally decided to remove this qualification feature.

**Iteration 4.** She immediately found that Group Separation and Group Skew score improved by 0.03 and 0.02 (From R3 to R4 in Fig. 6h), and Individual fairness slightly increased by 0.02, but GFDCG score was still dragged around 0.6. After all feature exploration, she decided to finalize the feature set with 7 features and run the fair method to make an improvement (MT×M).

**Iteration 5.** When she ran the in-processing model, it improved the fairness of ranking outcome ($GFDCG$ = 1.26) (Fig. 6d-i5) with a fair number of women within the top-20 shown in Ranking View. Finally, she found that the overall fairness score highly improved in two spaces and also Mapping ($Group\ Separation$ = 0.09, $Group\ skew$ = 1.01, $GFDCG$ = 1.26). The amount of biases per group was also closer to each other ($rNN_{S^+}$: 0.34, $rNN_{S^-}$: 0.33). In Ranking List View, she was able to compare all generated rankings with Group fairness, Individual fairness, and Utiliy measures. While there was a trade-off between rankings (R5 in Fig. 6h), the last ranking outcome achieved both higher fairness and utility scores.

**Iteration 6.** While she settled down with 5th ranking, she had to decide how many individuals she should pick. As she moved on to Ranking View, the ranking visualization conveyed the information of within-ranking fairness and utility trend (Fig. 6f). Observing the nearby positions, she found that within-ranking fairness improved by 0.14 when she slightly increased the threshold to $k = 50$ while within-ranking utility remains the same as 0.82. She decided to finalize the ranking decision with the last ranking by selecting 50 candidates based on the top-$k$ threshold.

## 8 User Study

We evaluate *FairSight*'s design in terms of its understandability and usefulness in decision-making by conducting a user study. We compared *FairSight* with 'What-if' tool [15] developed by Google, which is one of existing tools for probing the machine learning models on their performance and fairness.

We recruited 20 participants (age: 23–30; gender: 8 female and 12 male participants), a majority of which were graduate students who study Information and Computer Science and have the knowledge of machine learning to ensure they are familiar with the typical machine learning workflow and terminologies. We conducted a within-subject study where each participant was asked to use both tools in a random order. We gave participants the tutorial (25 mins) and let them explore the two systems (15 mins).

**Questions and tasks.** Participants were given tasks in a decision-making situation similar to our case study (Section 7) based on the German credit rating dataset [20]. The task is to predict which candidates are most likely to pay back a loan, with a fairness goal of protecting female customers against discrimination. Due to the differences in the two systems' output decisions (i.e., FairSight: ranking; What-if: classification), the participants were asked to conduct the task differently where (1) with FairSight: to select $k$ qualified customers among $n$ candidates ($n = 250$), while (2) with What-if: to classify qualified customers. Participants were asked to start with seven out of the ten features (Fig. 6g) with Logistic Regression as initial setting.

Participants were asked to conduct 12 sub-tasks (4 fairness stage x 3 machine learning phases). These tasks correspond to the tasks listed in Fig.2 but with more specific question that has a correct answer, e.g., MT×O: "Can you quantify the degree of fairness in the ranking outcome?". The users were expected to correctly identify the directly relevant information offered by the system (e.g., the answer to this question could be "a fairness score of 0.85"). We also asked users 3 additional questions for Decision (e.g., ask users to compare the differences in two iterations) and Explain (e.g., ask users to find the explanation on instances or features). The accuracy was measured based on whether a user can correctly point out the directly relevant information. We also asked users to rate the understandability "*How well does the system intuitively capture and represent the bias/fairness in the decision process*?" and usefulness "*How is the information provided by system useful for fair decision-making*?" in a Likert scale from 1 to 5 for each task. Since the two tools have different functionality (What-if tool is able to support 9 out of 15 tasks while FairSight provides all the functionality), we measured and compared the accuracy on 9 questions which can be answered by both systems. We also collected the subjective feedback after completing the tasks.

**Result.** The result is presented in Fig. 7. The overall accuracy of two tools was 95% for FairSight, and 80% for What-if. Per-criteria accuracy with their average accuracy is shown in Fig. 7. Fig. 7a summarizes the evaluation result of each stage of fairness pipeline. The result, based on the t-test, indicated that *FairSight* significantly outperformed the What-if tool in terms of understandability and usefulness (Fig. 7a).

We also measured the result when ratings are aggregated by three criteria: Fair, Decision, Explain, as shown in Fig. 7b. The statistical test

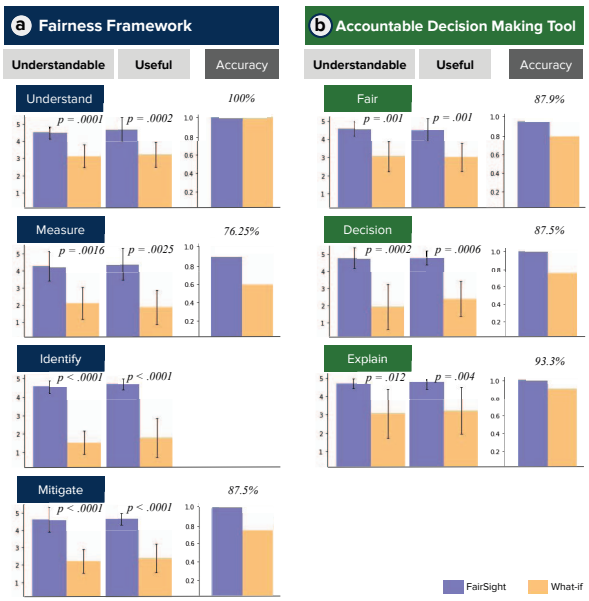


Fig. 7. Subjective ratings (understandability and usefulness) and accuracy (a) in the four stages of the fairness framework and (b) in three criteria of the decision making tool: Fairness, Decision, and Explain.

proved that FairSight was more effective in terms of understandability and usefulness (Fig. 7b). What-if tool was relatively good at providing reasoning behind instance-level inspection using counterfactual example, and feature importance with partial dependence plot, with the score of 3.5, but lacked in comparing multiple outcomes.

**Subjective feedback.** We gathered the subjective feedbacks from users. Those are summarized in three aspects: (1) *FairDM* as a guideline of increasing the awareness of fairness, (2) Visualization as a fairness-aware overview of machine learning task, and (3) Comprehensive diagnosis of discriminatory features. First, most of participants appreciated how the framework and system enhanced the understanding and awareness of fairness. Several participants provided feedback on how the system improves their awareness, e.g., "*It was the first time I recognized/learned how machine learning can result in the fairness problem.*" Second, *FairSight* with visual components not only served as a diagnostic tool for fairness, but also helped users understand how the distribution of individuals changes with the fairness improved in three machine learning phases, as mentioned by a user, "*Three space views are intuitively show how the process is going with the degree of fairness*". Lastly, most of the users were surprised by how the system supports evaluating features as possible sources of direct or indirect discrimination in each phase. As a user mentioned, "*Feature selection is sometimes arbitrary, but it provides the feature-level measures as evidence of fairness-aware decision.*" – this demonstrated how the system can help decision makers to achieve fair decision making through better explainability.

## 9 Discussion

In this section, we discuss our observations on the use of *FairSight*, and extend it to the general context of fair decision making. We also summarize the limitations of our system based on findings from the user study.

**Importance of pre-processing stage.** Although all stages of *FairDM* have an important role in achieving fair decision making, the most critical part was found to be the pre-processing stage. As shown in the case study, the first 4 iterations were primarily concerned with the pre-processing stage, where the fairness scores can be significantly improved. We also observed that participants in the user study spent 80% of their exploring sessions in detecting and removing bias from features. It is also the case of the real-world practice that, according to [17], the data collecting and processing is the most preferable and time-consuming task. Based on our study, a fair decision tool that simply offers a package of fairness-aware algorithms and outcome measures will be not sufficient to meet the needs of data scientists and practitioners to combat the various bias and fairness issues in real-world practice, and our proposed design helps address this challenge with comprehensive support at the pre-processing stage.

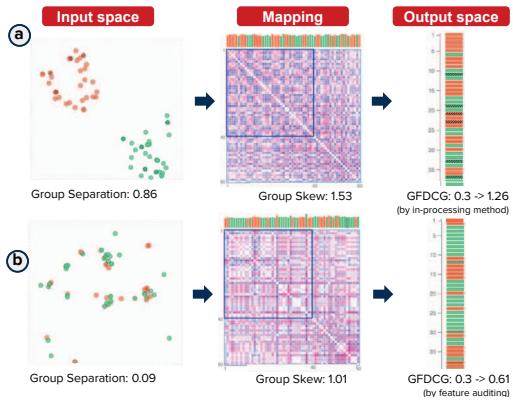


Fig. 8. Visualization of three stages from (a) 9 features (Iteration 1 in Case study) with the in-processing method (b) 6 features (Iteration 5 in Case study) with Logistic regression.

**Interaction between spaces.** *FairSight* represented the machine learning process as the mapping between Data and Outcome. With the perspective of space, our observation is that the entire pipeline coordinates in such a way bias is reinforced from data to outcome through the mapping. As illustrated in Fig. 8, features without the pre-processing step create more bias in the mapping (Fig. 8a), whereas features from fairer data representation were found to be less biased in the mapping and ultimately resulted in greater fairness in the ranking outcome (Fig. 8b).

**Subjectiveness of feature selection.** While it is possible to identify feature-level bias through the well-defined metrics provided by our system, human scrutiny through the interactive visualization is still required. Feature selection often requires domain knowledge; determining how a feature is important and fair may differ across contexts and domains [27], and is also subjective to people's perception on fairness [16]. There is no generally acceptable criteria for evaluating the trade-off between fairness and utility over decision outcomes. Therefore, it is desirable to have a decision-making tool that helps incorporate the domain knowledge and human judgment to achieve fair decision making.

**Limitation.** Despite the comprehensive framework and system implementation in our study to go towards fair decision making, we observe that a few drawbacks still exist. First, our visualization creates visual clutters as a number of instances increase while it enables the instance-level exploration. Second, the visualization can be misleading depending on group population. For example, in the case when the sensitive attribute has a skewed ratio of two groups (e.g., Men:Women = 8:2), the visualization of linearly ordered ranking outcome may look unfair even for a fair ranking. Our system also treats the sensitive attribute as dichotomy between protected group and non-protected group, which may not fit into some cases.

## 10 Conclusion

In this work, we presented *FairDM*, a decision making framework to aid fair data-driven decision making. We also developed *FairSight*, a visual analytic system with viable methods and visual components to facilitate a real-world practice of comprehensive fair decision making. The evaluation of our system demonstrated the usefulness of the system in fairness work over the existing tool. In the case study, we illustrated how the system was effective to measure and mitigate bias using a well-known dataset. For future work, we plan to extend the current binary representation of sensitive attribute in *FairSight* to handle multiple groups and sub-groups, as well as user-defined groups. Furthermore, to tackle industry-scale dataset, we will develop a scalable visual representation of rankings (e.g., how to make the matrix representation or reordering to efficiently present the fairness).

## Acknowledgement

The authors would like to acknowledge the support from NSF #1637067 and #1739413.

## REFERENCES


[1] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18(1):8753–8830, 2017.

[2] A. Asudehy, H. Jagadishy, J. Stoyanovichz, and G. Das. Designing fair ranking schemes. *arXiv preprint arXiv:1712.09752*, 2017.

[3] N. Bantilan. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services*, 36(1):15–30, 2018.

[4] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. *arXiv preprint arXiv:1805.01788*, 2018.

[5] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] M. Brooks, S. Amershi, B. Lee, S. M. Drucker, A. Kapoor, and P. Simard. Featureinsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 105–112. IEEE, 2015.

[7] R. Busa-Fekete, G. Szarvas, T. Elteto, and B. Kégl. An apple-to-apple comparison of learning-to-rank algorithms in terms of normalized discounted cumulative gain. In *ECAI 2012-20th European Conference on Artificial Intelligence: Preference Learning: Problems and Applications in AI Workshop*, vol. 242. Ios Press, 2012.

[8] D. K. Citron and F. Pasquale. The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89:1, 2014.

[9] A. Datta, M. Fredrikson, G. Ko, P. Mardziel, and S. Sen. Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*, 2017.

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

[11] A. Fisher, C. Rudin, and F. Dominici. Model class reliance: Variable importance measures for any machine learning model class, from the" rashomon" perspective. *arXiv preprint arXiv:1801.01489*, 2018.

[12] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[13] J. H. Friedman, B. E. Popescu, et al. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

[14] N. Frosst and G. Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

[15] Google. What-if tool. `https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html`, 2018. Accessed: 2018-12-30.

[16] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pp. 857–871, 1971.

[17] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, vol. 1, p. 2, 2016.

[18] D. Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[19] F. Hausdorff. *Set theory*, vol. 119. American Mathematical Soc., 2005.

[20] H. Hofmann. UCI machine learning repository, 2017.

[21] K. Holstein, J. W. Vaughan, H. DaumÃ© III, M. DudÃk, and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? doi: 10.1145/3290605.3300830

[22] IBM. Fairness 360. `https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/`, 2018. Accessed: 2018-12-30.

[23] C. Karako and P. Manggala. Using image fairness representations in diversity-based re-ranking for recommendations. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, pp. 23–28. ACM, 2018.

[24] B. Kim, R. Khanna, and O. O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pp. 2280–2288, 2016.

[25] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5686–5697. ACM, 2016.

[26] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

[27] Z. C. Lipton. The doctor just won't accept that! *arXiv preprint arXiv:1711.08037*, 2017.

[28] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[29] Y. Ming, H. Qu, and E. Bertini. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics*, 25(1):342–352, 2019.

[30] C. O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

[31] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive research: principles and implications*, 3(1):29, 2018.

[32] D. Pedreschi, S. Ruggieri, and F. Turini. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 126–131. ACM, 2012.

[33] Propublica. Machine bias, 2016.

[34] Reuters. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

[36] L. Rüschendorf. Wasserstein metric. *Hazewinkel, Michiel, Encyclopaedia of Mathematics, Springer*, 2001.

[37] A. Singh and T. Joachims. Fairness of exposure in rankings. *arXiv preprint arXiv:1802.07281*, 2018.

[38] P. Tamagnini, J. Krause, A. Dasgupta, and E. Bertini. Interpreting black-box classifiers using instance-level visual explanations. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, p. 6. ACM, 2017.

[39] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, p. 22. ACM, 2017.

[40] K. Yang, J. Stoyanovich, A. Asudeh, B. Howe, H. Jagadish, and G. Miklau. A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data*, pp. 1773–1776. ACM, 2018.

[41] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1569–1578. ACM, 2017.

[42] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert. Manifold: a model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, 25(1):364–373, 2018.

[43] I. Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.