# System to Integrate Fairness Transparently: An Industry Approach

Emily Dodwell, Cheryl Flynn, Balachander Krishnamurthy, Subhabrata Majumdar, Ritwik Mitra
AT&T Labs - Research, USA
{emily,cflynn,bala,subho,ritwik}@research.att.com



## ABSTRACT

Numerous Machine Learning (ML) bias-related problems have generated significant press in recent years. This has led to scrutiny of corporate failure regarding how to incorporate human oversight in thorough evaluation and prevention of bias. Companies have a responsibility to monitor ML processes for bias and mitigate any bias detected, ensure business product integrity, preserve customer loyalty, and protect brand image. In this paper, we propose SIFT (System to Integrate Fairness Transparently) as a methodological approach for integrating bias detection, mitigation, and documentation in ML projects at various stages of the ML lifecycle. To this end, SIFT involves a combination of mechanized and human-in-the-loop components. The human-in-the-loop components are inserted at points in the ML lifecycle where project managers/practitioners could benefit from guidance from experts within the company in areas such as Human Resources, Public Relations, Legal, Privacy, and Compliance. To demonstrate the value of SIFT, we present two industry use cases that may require fairness scrutiny: (1) marketing and (2) hiring. We show how SIFT can be used to identify potential biases and determine appropriate mitigation strategies.

## 1 INTRODUCTION

With the increasingly widespread use of Machine Learning (ML) in our daily lives, concomitant concerns of unintentional bias have received significant attention in the popular press and research literature. Most of the academic work defining fairness has been analytical, theoretical, and legal [6]. Technical ways to detect and remedy bias [34] and toolkits [7] have also been proposed.

In algorithmic decision making, the terms 'bias' and 'fairness' carry a multitude of implications that can vary based on the application area, and the individuals involved. Statistical definitions of bias often refer to deviation from an expected behavior which might not necessarily carry a positive or a negative connotations. Fairness also can carry a statistical and an individual-based meaning [12]. Individual fairness expresses that people with 'similar' characteristics should receive similar decisions irrespective of their demographic categorization. While this notion has immediate human appeal, a concrete mathematical formulation of the same requires stringent assumptions [22] and might be impractical as a working model. On the other hand, a statistical definition of fairness adheres more closely to the mathematical definition of bias stated above, wherein one tries to reduce the deviation of a certain measure (such as false positive rate) from parity across different demographic groups.

Thus, statistical definitions of bias or fairness can be thought of as a more quantitatively tractable expression of an ideal notion of fairness. In our present work, we use bias to refer to specific quantitative disparities, addressing which can get us closer to an idealized understanding of fairness as appropriate for the use case.

Numerous ML bias-related problems have surfaced in applications like image identification [48], hiring [15], and targeted advertising [18]. However, the problem of bias in an industrial setting with use cases far more diverse than the handful of academic literature exemplars remains poorly covered. Industry has to serve under-represented communities and not prioritize the targeting or delivering of services in a discriminating way. It has a responsibility to continually monitor ML processes for bias and mitigate any identified to ensure business product integrity, preserve customer loyalty, and protect brand image. However, there are several barriers to a systemic solution [28] to such problems, including the need for fairness-aware data gathering, identification of blind spots, use case specific guidance, and human oversight.

We propose SIFT (System to Integrate Fairness Transparently) as an operational framework to identify and mitigate bias at different stages of an industry ML project workflow. SIFT enables an industrial ML team to define, document, and maintain their project's bias history. SIFT guides a team via mechanized and human components to monitor fairness issues in all parts of the project. Longitudinally, SIFT lowers the cost for dealing with fairness through reuse of techniques and lessons learned from handling past fairness concerns.

SIFT draws valuable lessons from attempts to improve security and privacy on the Internet. In the security arena, many attempts were made to codify intrusion detection to prevent attacks. For example, the widely-used open-source intrusion detection and prevention system Snort [42] uses both signatures and rules while conducting real-time analysis of Internet traffic to match against attacks. Signatures that are stored, updated, and shared include methods to detect an attack, while rules detect specific vulnerabilities. ClamAV[13] is an anti-virus toolkit that uses complex routines to detect attacks via user-contributed signatures and allows automatic remote database updates to constantly enhance its functionality. SIFT mirrors notions of signatures and rules in its collection of an ML project's *bias history*. Over the course of an ML project this is updated with information on methods to detect and mitigate concerns like sparse groups, marginalized groups, and proxy variables. SIFT uses a collection of ML projects in the enterprise to record information about any bias detected in their bias histories, enabling future reuse.

Similar to Privacy by Design [9], which pushed for key privacy notions to be embedded in social networks, we want bias detection and mitigation to be considered early on by ML project managers. Just as we find new vectors of attacks on security and instances of privacy leakages, there will be new vectors of potential ML bias arising from data and model reuse, or repurposing of ML approaches for alternate use cases. Given the costs of gathering data and the time pressure associated with deploying new projects in large enterprises, this may occur with higher frequency than expected.

SIFT introduces several human-in-the-loop steps in the ML workflow to enforce risk assessment. use case specific factors to guide these steps include domain knowledge, bias history of past projects in the enterprise, legal guidelines, and cost considerations of brand impact or regulatory penalties. Similar to privacy, regulation may be proposed in the space of ML. Defensive steps to handle regulatory concerns and transparency mechanisms to demonstrate fairness are thus proactively built into the ML development lifecycle. We may not identify *all* potential sources of bias and mitigate them, but by including human-level checks we *reduce* the likelihood of fairness concerns impacting an enterprise.

ML Projects in large enterprises that are non-customer facing, or those with no demographic or geographic proxies may have *no* potential bias concerns. Managers of such projects would want to quickly move on with their work. If several risk factors are not involved in any stage of the ML lifecycle, SIFT allows the project to move ahead with more confidence that potential bias is not a concern. However, if fairness concerns are too high without a clear path to bias mitigation, we recommend an early exit from the SIFT process for the manager to reconsider the project design, for example, by collecting additional data.

## 2 MOTIVATION AND CONTRIBUTION

In this section, we provide an overview of prior art and enumerate SIFT's contributions. Section 2.1 reviews fairness-related ML research, and highlights challenges faced by industry ML practitioners. Section 2.2 summarizes the salient features of SIFT and how SIFT differs from prior art in addressing existing challenges.

### 2.1 Related work and industry challenges

There are two broad categories of research in bias and fairness.

*Methodology.* Depending on the use case and modeling objectives, several sources of bias and discrimination may exist in the data. A recent enumeration [34] lists 23 types of bias and 6 types of discrimination associated with ML models, exposing three types of fairness concerns: individual, group, and subgroup fairness. Research on bias detection and mitigation methodology include methods aimed at the pre-, in- or post-processing parts of an ML project using classification or regression modeling (see Section 3.3). Fair versions of other techniques such as clustering [4], community detection [35], and causal models [52] have also been proposed.

*Tools.* AI Fairness 360 (AIF360) [7] is a well-known tool that packages bias detection and mitigation methods in the literature for reuse. Among other similar packages, which are mostly open-source, Aequitas [44], Fairness Measures [49], FairML [1], FairTest [45], and Themis [24] offer bias detection, while Fairlearn [21] and Themis-ml [5] offer detection and mitigation through expandable platforms. In parallel to our work, LinkedIn Fairness Toolkit (LiFT) [46] was recently released as an open-source framework that is flexible enough to integrate bias metrics and mitigation strategies at any stage of the ML lifecycle and can handle web-scale ML problems. LiFT was designed for integration with LinkedIn's ML system, and to our knowledge, has not been applied to ML applications outside of LinkedIn's search and recommendation systems. The goal of our work is to develop a framework that is applicable across a wide variety of ML applications.

The above tools and methods focus on the technical requirements for integrating bias metrics and mitigation algorithms into ML projects. However, there are several other challenges for integrating fairness considerations into industry applications [28, 47].

Data Collection: Most fairness-aware methods focus on detection and mitigation algorithms applied to fixed datasets [28, 34]. Little guidance is available in the data gathering stage, e.g. identifying and documenting sensitive features, locating similar internal projects with their sensitive features and bias histories, and bias risk assessment given a project's data sources and sensitive features.

Blind spots: Teams are limited in their ability to detect bias in a new project due to lack of guidance for recognizing areas of bias concern. Tools to learn from past similar projects would enable, for example, quick identification of sensitive subpopulations relevant to a specific use case, along with previous discovery that collection of more data is required to mitigate associated sparse group issues.

Use case diversity: Fairness research has received unequal attention across ML domains and stages of the ML workflow [34]. The above challenges and lack of structured tools leaves little guidance on issues like domain-specific metrics and methods, or appropriate proxies when individual-level demographic data is missing.

Need for human oversight: When and how to rely on human decision making is a key concern. Examples include choosing a mitigation strategy, the amount of additional data collected for mitigation, risk and cost assessment for mitigation, consequences and trade-offs of optimizing for a fairness metric, and changes to project design to minimize the need for any fairness auditing.

Motivated by such needs, recently a number of data and model documentation artifacts have been proposed to enable transparency in industry ML processes, such as FactSheets [3], Datasheets [25], and Model Cards [36]. Indeed, these can be adapted to detect bias concerns in the data or at different stages of the ML workflow. Implementation of internal algorithmic audits [40] and co-designed fairness checklists [33] have also been proposed to ensure that deployed ML models conform to company values and principles. To ensure best practices in industry applications, these types of documentation artifacts should be directly integrated into the ML project lifecycle. Futhermore, resources and guidance on how to account for various fairness considerations and how to complete the necessary steps of the documentation process needs to be provided to ML practitioners and project managers.

### 2.2 Our contributions

The tools and artifacts proposed in the area until now do not *actively* enjoin data scientists and program managers to consider bias at *every* stage of the ML workflow. In contrast, SIFT starts by defining a bias history object, then points out *specific* steps in the project for bias considerations, allowing practitioners to use relevant existing methods to detect and mitigate bias. Guided by the use case and team requirements, these can be implemented either directly from the literature, or from toolkits like AIF360. The bias history object sequentially records all such instances of bias-related checks, methods, and decisions. SIFT further brings a diverse set of key players (e.g. compliance, legal, PR) into the equation and reaps the benefit of domain expertise, along with reusing practical lessons from past ML projects.

Human oversight and blind spot checks are an integral part of SIFT because of its use of mechanized and human-in-the-loop components in tandem. It further allows the modeling process to complement important non-technical considerations. This flexibility can be valuable for the development of fairness-aware ML systems in industry areas that are less explored in the fairness literature until now, e.g. spatial models, and speech recognition, and for new areas that will arise in the future.

SIFT is intended for enterprises with multiple ML application areas, where cost considerations are central to ML project planning. In those settings, many applications may not have human oversight requirements, but the ones who need it are the ones that are likely to cause serious problems to the enterprise if not handled properly. Each new project using SIFT looks for similar projects internally and maintains its own model history and bias history. As more projects use SIFT, finding similar projects (along with their model/bias histories) becomes more likely. Bias detection and mitigation becomes increasingly proactive, thereby reducing cost of fairness-aware project planning over time. The diversity of use cases and the holistic attempt to translate the skills of multiple non-technical enterprise experts makes SIFT broadly applicable. This is a key advantage of our work over existing tools, including the most recently released toolkit, LiFT.

To summarize the above, the key contributions in this paper are:

- Propose the SIFT framework consisting of four pipelines, each of which directly incorporates both mechanized and human-in-the-loop functions, documents each stage of bias detection and mitigation process, and takes advantage of information from prior projects where possible. (Section 3)
- Demonstrate the functionality of SIFT and the use of the bias history object on two example use cases inspired by real enterprise problems. (Section 4)
- Discuss challenges and propose solutions to operationalize SIFT, including the concept of human oversight guideline documents. These documents offer a way to incorporate guidance from a diverse set of industry experts in a scalable manner. Example seed questions for creating these documents are provided (Section 5).

## 3 THE SIFT FRAMEWORK

We now describe the four pipelines of SIFT and their key components. We assume the existence of a database of existing ML projects in the company with the schema discussed below. We refer to the team working on the ML project as the SIFT user.

The SIFT framework has four classes: project, data, bias history and model history. The **`sift_project`** class is the top-level class for the system with all project-related information. We assume that a SIFT user starts a new project with (1) a `name`, (2) a `description` summarizing the background and objective, and (3) `data_location` of the database that stores data for the project, such as a remote directory, network drive, or URL. Based on these three input arguments, the user initiates a `sift_project` (Section 3.1). We assume that these inputs can be used at the outset to extract any available information regarding sensitive features, project personnel, and older versions of the same project. The full list of components of a `sift_project` are:

- `name`, `description`, and `data_location`,
- `project_id` – a unique identifier for the ML project,
- `sift_data`, `bias_history`, and `model_history` are objects of different classes described below and in Appendix A,
- `metadata` – dictionary with project personnel/status information
- `model_flow` – 'Standard' or 'Custom' bias-aware model building strategy (see Section 3.3),
- `similar_projects` – pointers to similar projects in projects database,
- `older_versions` – pointers to previous versions of the project in projects database,
- `timeout` – set to a company-specified time frame for terminated projects to be removed from the project database; defaults to None otherwise.

The **`sift_data`** class stores information about data used in the project. This includes raw data, data definitions, a list of feature variables and the target variable, and the predicted outcomes from any existing pre-built model. This class also contains the list of sensitive features relevant to the project, and a summary variable that is a *dictionary of dictionaries* with any additional information relevant to bias investigation. While we consider three such categories of information (sparse groups, proxy features, and marginalized groups) in this paper, specific project teams may decide to include other categories. See Appendix A.1 for a full list of `sift_data` class components.

We do not assume that every project has all of the components in the `sift_data` class available at the outset. For example, a new project may not initially know the relevant sensitive features. SIFT's identification of similar projects helps with this part of the project discovery, thereby reducing the overall cost of addressing bias.

The **`bias_history`** class is a novel contribution of the SIFT framework and is used to track each stage in the bias and mitigation process. The components of the `bias_history` class are:

- `step` – a counter capturing the place in the sequence of bias and mitigation tasks performed,
- `sift_pipeline` – name of pipeline this step belongs to,
- `bias_features` – the sensitive features under consideration in the current step,
- `bias_detection_function`,
- `bias_mitigation_function`,
- `mitigation_success_status` – indicates whether bias is not detected after implementing the mitigation algorithm,
- `details` – additional information such as the results of the bias investigation or the actions taken by the SIFT user.

Steps in the bias history are added to document each stage of the bias detection and mitigation process. Methods associated with the class allow SIFT to access the current step of the bias history, add components to the current step, and add a next step in the history. We describe these methods in Appendix A.2.

We record longitudinal bias-related information across a large number of projects through `bias_history`, adding to the transparency on exactly *how* fairness is weaved into the ML lifecycle in an enterprise. This has numerous advantages. First, it enables information reuse for future projects and lowers cost across the enterprise in handling bias. Second, the precise nature of bias detected, the specific pipeline where it was first detected, the algorithm that helped locate it and its mitigation success status suggest the first

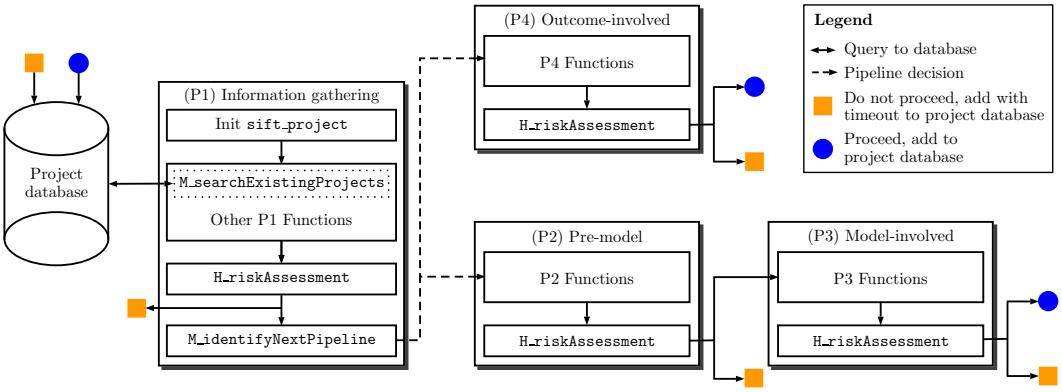


**Figure 1: The four pipelines of SIFT.**

steps for a new project. The success status is key in deciding if a recommended mitigation algorithm should be reused. Note that this value is not dispositive; a different mitigation algorithm might work better for a different use case or project. However it is still useful information that is traditionally not tracked in the ML lifecyle. The `bias_history` is returned as a result for queries for similar projects, indicating that the manner of resolution is visible to anyone in the future. Such transparency helps organizations defend the actions they have taken to address bias.

The **`model_history`** class tracks the history of the ML model through development and training. This class includes information on the training and test sets, the fitted model object, the performance metric(s) used to evaluate the model, and its deployment status at each stage of the modeling process. The class components are described in Appendix A.3.

SIFT assists the user through four pipelines (Figure 1, see Figure B.3 in the Appendix for expanded version). Progress through the pipelines is *not* sequential: After initiating a new project in Information gathering (P1), users move to Outcome-involved (P4) if the model is already deployed, otherwise to Pre-model (P2) followed by Model-involved (P4) pipelines. We present below more details on each pipeline with pseudo-code (written with respect to a single sensitive feature, so note that functions are run for each when there are multiple). Function names with an `M_` prefix indicate mechanizability while `H_` implies need for a human in the loop; auxiliary functions are described in Appendix B.

## 3.1 Information gathering pipeline

This pipeline checks in three steps if a new project has fairness concerns based on its description and intended audience.

**(1) Project initialization.** The pipeline starts by initializing the `sift_project` object which adds the prior model history (if any) obtained from the `data_location`, and any details on the data and metadata. It initializes the `bias_history` object, which is updated alongside bias-related steps taken throughout the project.

**(2) Similar project identification.** We now search the ML projects database for similar projects, its location specifed by `db_location`. The database can have a simple Web query interface whence `db_location` would be a URL. The search uses normal information retrieval steps: remove non-alphanumeric characters, normalize case, lemmatize words, remove stop words before vectorizing the data and calculating a cosine similarity score.

```
def M_searchExistingProjects(current_project, db_location,
  ngram, threshold):
  matched_projects_list = []; results_indices = []
  removeNonAlphaNumericChars (current_project)
  lowerCase (current_project)
  removeStopWords (current_project)
  lemmatize (current_project)
  ComputeTFIDFVectors_with_Bigrams (db_location, current_project)
  computeCosineSimilarity (db_location, current_project)
  if cosine_similarity > threshold: results_indices.append()
  matched_projects_list = database[results_indices]
  return matched_projects_list
```

The SIFT user human-verifies the list of matched projects via `H_verifySimilarity` and adds relevant ones to `similar_projects`. The user finds a relevant set of sensitive features by considering those in `similar_projects` and other factors (e.g., legal constraints and domain knowledge). `H_identifySensitiveCategories` captures this action and stores them as `sens_features`. In absence of similar projects, external considerations can identify sensitive features. `H_verifySimilarity` and `H_identifySensitiveCategories` are the *Other P1 Functions* in Figure 1.

**(3) Preliminary risk assessment.** The function `H_riskAssessment` now considers the information collected in the current `sift_project` object, with the list of similar projects, factoring in business, contractual, and legal constraints, as well as customer impact (see Section 2.2). The `bias_history` object documents the outcome of the risk assessment. If the decision is not to proceed due to absence of fairness concerns or high risk of bias, the project object is correspondingly updated. A `timeout` component is set based on company guidelines and the project is added to the project database before exiting SIFT. Else, the next pipeline is determined by the function `M_identifyNextPipeline`, which sets it to Outcome-involved if the model is deployed or to Pre-model otherwise. A deployed model is one that is beyond model training and development (for example if the model is field-trial ready or already in production). Based on the `sift_pipeline` determination, the project is moved to either the

Pre-model pipeline (Section 3.2) or the Outcome-involved pipeline (Section 3.4).

## 3.2 Pre-model pipeline

The Pre-model pipeline (P2) checks the raw data for issues that could lead to a biased ML model. To simplify the flow through the pipelines, we implement all algorithmic mitigation strategies in the next (Model-involved) pipeline, but allow the user to make changes to the raw data in this pipeline. A typical example of P2 would involve the following steps. At the end of each step, the bias history is updated with the task performed and the results, and a new step is added to the bias history. Detected sparse groups, proxy features, and marginalized groups are added to the `sens_features_summary`. Below we present details for each of these steps. Steps 1-4 among them are the *P2 Functions* of Figure 1.

**(1) Data preparation.** Standard practice in ML workflows, this step involves data cleaning and feature engineering and is typically completed prior to performing any bias checks. While some aspects of this could be mechanizable, data preparation often involves some input and inspection of the data by the ML practitioner, handled by `H_prepareData`. The final data for the project should be updated and stored in the `sift_data` object.

**(2) Sparse group detection.** To train a fair ML model we need a sufficient number of training samples for each of the subgroups defined by the sensitive attributes. Otherwise, the ML model can have poor performance when predicting results for samples of the under-represented subgroup in practice. For example, Amazon abandoned an ML system intended to automate the hiring process by identifying resumes of top technical talent; its training on past resumes penalized female applicants due to historical gender imbalance within the tech industry [15]. We check for sparse group representation in the data using the function defined below.

```
def M_detectSparseGroups(sens_feature):
  sparsityFunc, threshold = M_selectSparsityFunction()
  subgroup_list = []; sparse_groups_dict = {}
  subgroups = sens_feature.unique()
  for g in subgroup.iteritems():
    sparsity = sparsityFunc(sens_feature, g)
    if sparsity > threshold: subgroup_list.append(g)
  if subgroup_list: sparse_groups_dict[sens_feature] =
       subgroup_list
  return sparse_group_dict, sparsityFunc
```

If sparse groups are detected, the user can collect additional data or terminate the project with `H_verifyGetMoreData`. Else, the Model-involved pipeline will attempt to address this issue using an algorithmic pre-processing strategy such as reweighing or resampling [29].

**(3) Proxy feature detection.** Removing sensitive attributes from the set of features will not guarantee an unbiased ML model. One way bias may remain in the data is through the existence of proxy variables. For example, in the context of targeted advertising on Facebook, [43] found that many features provided on the Facebook ad platform were strongly correlated with sensitive attributes like gender and race. SIFT checks for strong pairwise correlations between sensitive and non-sensitive attributes.

```
def M_detectProxyFeatures(sens_feature, nonsens_feature):
  depFunc, threshold =
    M_selectDependenceFunction(sens_feature, nonsens_feature)
  feature_list = [];  proxy_dict = {}
  for x in nonsens_features.itercols():
    dep = depFunc(sens_feature, nonsens_features[x])
    if dep > threshold: feature_list.append(x)
  if proxy_list: proxy_dict[proxy_features] = feature_list
  return proxy_dict, depFunc
```

Pairwise correlation checks do not guarantee the removal of all proxy variables: further bias checks are needed in the Model-involved pipeline. In particular, when the number of sensitive attributes and non-sensitive attributes is large, combinations of non-sensitive attributes could create a proxy for a sensitive attribute even when individual variables don't. Such multivariate proxies would not be detected at this step. When univariate proxy variables are detected, the SIFT user can drop the proxy variable from consideration in later modeling steps via `H_verifyDropProxy`.

**(4) Marginalized group detection.** Bias present in the target variable will be learned by the ML model. This step checks the target variable for marginalized groups to alert the SIFT user to this potential issue using the function `M_detectBias` (see Section 3.3). If marginalized groups are detected, then an algorithmic mitigation strategy can be implemented later in the Model-involved pipeline.

**(5) Pre-model risk assessment.** The last step of this pipeline asks the SIFT user to perform a risk assessment given the information learned in this pipeline– information that could fundamentally change the project plan. For example, if a key input feature is found to be a proxy for a sensitive attribute, the user may not wish to proceed. The `bias_history` object captures the steps and results of each bias-related action taken in the pipeline. The user reviews this in `H_riskAssessment` and `bias_history` object documents the outcome of the risk assessment. If the user decides not to proceed, then project status is set to 'Terminated', `timeout` is set pursuant to company guidelines, and the project is added to the project database. Else, SIFT begins the Model-involved pipeline.

## 3.3 Model-involved pipeline

The Model-involved pipeline (P3) checks for bias introduced when training the ML model and implements mitigation strategies if bias is detected. This pipeline attempts to mitigate bias detected in Section 3.2 (P2) that could not be addressed by dropping variables or collecting additional data. It further tries to mitigate bias detected in the model outcome, which could arise even if the raw data passes previous checks due to complex data patterns. For example, a deep learning model could learn a non-linear function of non-sensitive features that creates a proxy variable for a sensitive feature.

A typical example of this pipeline can be broken down into six steps: pre-processing mitigation, model training, model outcome bias detection, in-processing mitigation, post-processing mitigation and model-involved risk assessment. We describe each step in detail shortly. We refer to functions in steps 1-5 as *P3 Functions* in Figure 1. Unlike the previous pipelines, the user may not proceed sequentially through all the six steps. Constraints of time or computational resources may influence the choice and ordering of mitigation strategies. For example, if the ML model is computationally expensive to retrain and time-to-market is a concern, then the user may limit focus to only post-processing strategies. There is also no guarantee that any one mitigation strategy will resolve detected bias issues; multiple mitigation strategies may be required. Further, a mitigation strategy may not exist that will address the source of bias, requiring designing of a novel mitigation strategy.

We thus allow for a choice of flow processes in this pipeline: standard or custom. The enterprise would determine the appropriate *standard* flow that a majority of projects would follow. For example,

it could be set to closely follow the framework of D'Alessandro et al. [16], which works sequentially through the steps listed above. The *custom* flow allows the user control over the sequencing and implementation of the steps of the pipeline. The user can restrict attention to a specific set of bias detection metrics and mitigation algorithms, copy a routine from a similar project, or run a novel bias detection and mitigation strategy designed for the application. The ability to copy bias detection metrics and mitigation strategies used in similar ML projects is a key feature of SIFT that helps reduce the cost of reducing bias as more projects are added to the database. The `model_flow` input to the `sift_project` object indicates the selected flow for the project.

Below we present details for each step of the pipeline. We assume that default settings for the functions would be standardized across the enterprise but could depend on the project application. These functions would make use of open-source bias mitigation algorithms or designed for company-specific projects. For reference, we provide examples of pre-, in-, and post-processing mitigation algorithms, that have open-source code available through [7]. An example standard flow process that works sequentially through the six steps is provided in Appendix C.

**(1) Pre-processing mitigation.** Pre-processing algorithms [8, 23, 29, 50] transform the raw data to reduce if not remove bias. These algorithms address bias in the raw data, for example, due to an under-representation of samples from a protected group, and do not require access to the training model or the model output.

As an example of a standard flow process, the system first summarizes the information collected during Pre-model pipeline using `M_getPreModelBiasStatus`. In case bias is detected, the system runs a pre-processing mitigation strategy using `M_preProcessingMitigation`. This function selects and implements the pre-processing strategy using the information collected in the Pre-model pipeline. It then returns the transformed dataset, and the pre-processing function, which is recorded in the `bias_history` object. The pre-processing function involves a transformation of the raw data, so the user will need to train the ML model regardless of prior model availability.

**(2) Model training.** This step of the pipeline uses the function `M_trainModel` to train the ML model.

```
def M_trainModel(current_project):
  seed, train_index, test_index =
    M_dataSplit(current_project.data.rawdata, split_ratio)
  modelFunc, modelMetricFunc = M_selectModel(current_project)
  fitted_model = modelFunc(current_project.data, train_index)
  perf_metric =  modelMetricFunc(current_project.data,
    fitted_model_object, test_index)
  return seed, train_index, test_index, fitted_model, perf_metric
```

All iterations in this model development process will be documented in the `model_history` object.

**(3) Model outcome bias detection.** The existence of bias in the model outcome should be quantified using one or more bias detection metrics [7]. Often such metrics check that the model outputs are equivalent across different values of a sensitive feature or demographic subpopulations. The function `M_detectBias` computes these metrics with the model output as the `feature`.

```
def M_detectBias(current_project, feature, sens_feature):
  biasFunc, fairness_range = M_selectBiasFunction(current_project)
  biased_groups_list = []
  subgroups = sens_feature.unique()
  for g in subgroups.iteritems():
    bias = biasFunc(feature, sens_feature, g)
    if bias not in fairness_range: biased_groups_list.append(g)
  return biased_groups_list, biasFunc
```

In the above, `fairness_range` is the interval for which no bias is detected. Examples of bias detection metrics include disparate impact, equalized odds, demographic parity, and statistical parity [7, 19, 27].

**(4) In-processing mitigation.** In-processing algorithms incorporate one or more bias metrics directly into the prediction model and ensure that prediction accuracy is attained only under predefined fairness constraints stipulated by those bias metrics [11, 31, 51]. In SIFT, the function `M_inProcessingMitigation` selects and implements the in-processing strategy. It then returns information about the newly fitted model, which is stored as a new step in the `model_history` object, and the in-processing function, which is recorded in `bias_history`.

**(5) Post-processing mitigation.** Post-processing algorithms [27, 30, 37] transform the outputs from a specific trained model and are model-agnostic in the sense that they do not require access to the training data or the trained model. These methods are particularly useful when there is a high cost for re-training the underlying model. The function `M_postProcessingMitigation` selects and implements the post-processing mitigation strategy and subsequently returns the new predicted outcome and the selected post-processing mitigation function, which is recorded in the `bias_history` object.

**(6) Model-involved risk assessment.** The last step of this pipeline asks the SIFT user to perform a risk assessment given the results of this pipeline using the function `H_riskAssessment`. Since no mitigation strategy is guaranteed to remove all forms of bias, this is a key step in the process and will depend on factors such as the extent to which the ML model might impact customers and the potential risk to the enterprise's brand image. This assessment should also take into account any degradation in utility due to bias mitigation steps. If the user decides to proceed, then SIFT will mark the project as scheduled for deployment and record the project in the project database. Otherwise, SIFT marks the project as terminated and records the project in the project database with the appropriate timeout specified. In both cases, the `bias_history` object documents the risk assessment decision.

## 3.4 Outcome-involved pipeline

If the project focuses on a deployed model, then the user comes directly here from Information gathering. Examples of deployed models include ones in production that require periodic checks for bias, models that have passed training and development and are to be tried in the field, or third-party models used by the company.

This pipeline checks for bias in the model outcome and may mitigate detected bias using the post-processing algorithms described in Section 3.3. In this pipeline, we do not require access to either the training data or the trained model. Unlike the workflow in the Pre-model and Model-involved pipelines, the data fed through the model may be independent of the model's original training and test datasets. If an earlier version of the project exists, we assume that the previous iteration has been documented as a `sift_project` object, the project passed the appropriate bias risk assessments, and that the project has at least one sensitive feature identified in its

data input. When an earlier version exists, this pipeline also checks for a distributional shift in the data between the two time points.

As a typical example, this pipeline would involve: (1) Data change detection, (2) Model outcome bias detection, (3) Post-processing mitigation, and (4) Outcome-involved risk assessment. Steps 1-3 are the *P4 Functions* in Figure 1. Steps 2 and 3 are same as ones in Section 3.3; details for steps 1 and 4 are provided below.

**(1) Data change detection.** Relevant variables may change over time, triggering a check against recorded statistics about the model's original training data for similarity [36]. For example, features could be deleted, missing values or new categories could be introduced, or there could be a distributional shift in one or more variables. If an earlier model iteration that was trained on a separate set of samples exists, then a typical example of this step would check for a covariate shift in the feature variables via the function:

```
def M_detectCovariateShift(current_data, prior_data, features):
  covariateShiftFunc, threshold =
    M_selectCovariateShiftFunction(current_data, prior_data)
  covariate_shift_list = []
  for x in features.itercols():
    shift = covariateShiftFunc(current_data[x],
      prior_data[x])
    if shift > threshold:
      covariate_shift_list.append(x)
  return covariate_shift_list, covariateShiftFunc
```

The prior data can be accessed through the list of pointers in the `older_versions` component of the `sift_project` object. If necessary, summary statistics about the current data and prior data can be compared if the full prior data set is not available. It's important to note that checking for a covariate shift in the data may not detect all changes to the underlying data, and additional safeguards should be built-in that are specific to the company's applications. If a change to the underlying data is detected, then the user may decide to exit SIFT and retrain the model, or to continue through the pipeline to see if the change to the data results in a biased outcome. This decision is captured by the function `H_verifyRetrainModel`.

**(4) Outcome-involved risk assessment.** Like previous pipelines, the last step of the Outcome-involved pipeline asks the user to make a final risk assessment via `H_riskAssessment` and `bias_history` object is updated to reflect the final decision. Any fairness concerns that remain in the ML project must be weighed against the cost and feasibility of developing, training, and deploying a new ML model or working with an alternative third-party source. If the user decides to proceed, then the project remains in deployment. Else, the project is terminated and added to the project database with an appropriate timeout specified.

# 4 INDUSTRY USE CASES

Several recent examples of industries deploying ML-based products and realizing biases as they manifest later have occurred. A reactive approach of companies redeploying their product with corrective measures leads to cost inflation and negative PR. Our proactive approach in handling bias throughout the ML lifecycle can reduce these concerns. We now show how SIFT addresses potential biases in three examples covering two representative use cases. As part of each use case we present their `bias_history` objects. We give all `bias_history` fields in the first example (Listing 1), and only non-empty fields for the other two (Listings 2 and 3) for brevity. Note that the bias detection and mitigation functions in these bias history objects are auxiliary functions, and are defined in Appendix B.

## 4.1 Marketing

As targeted advertising has become standard in the digital landscape, industrial concerns of unethical or illegal advertising have also arisen. Historically marginalized groups have lost visibility into information in ads related to high paying jobs [17, 18]. Such discrimination may be unintentional on the part of the advertiser or the ad platform, but nevertheless does occur when targeting systems and ad delivery algorithms are applied without careful evaluation of unexpectedly introduced bias [32] along the way [10]. SIFT provides a thorough framework through which such an evaluation can proceed.

Suppose a company wants to identify customers likely to be early adopters of a new service being rolled out among its existing customer base to receive exclusive discounts as part of a promotional marketing campaign. To build an early adopter model, a project team surveys a small sample of customers on the likelihood of service adoption. The team constructs the binary target variable, **y**, indicating whether each customer is likely to be an early adopter based on their response. As features for the model, the team merges the survey data with marketing data purchased externally. This data is available for the entire customer base, and includes demographic information and consumer segmentation data constructed from social media, online browsing, and purchase data.

Below we describe two sample projects under this setup and their steps through the SIFT pipelines. We summarize these steps in Figure 2. Note that both projects skip the Outcome-involved pipeline because neither involve a deployed model.

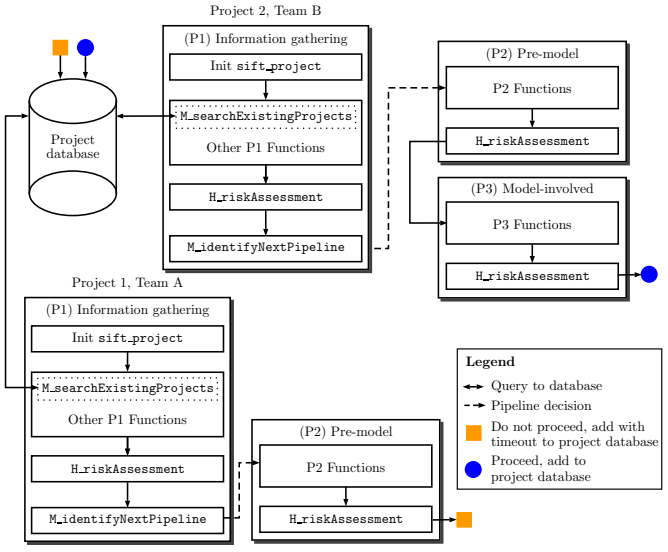


**Figure 2: SIFT flow for the two marketing-related projects**

As data for these projects, we use the demographic data from the UCI Adult dataset [20] and simulate 50 binary features designed to represent consumer segmentation data. We simulate **y** as a binomial random variable with probability depending on a subset of the features. The simulation details are provided in Appendix D. A small subsample is selected that contains only 5% non-white samples in Project 1. The full dataset is used in Project 2.

#### 4.1.1 SIFT implementation - Project 1.

**Information gathering.** Team A initializes a `sift_project` object with project id Svc2020. The corresponding `sift_data` object is populated with the target variable, demographic data, and consumer segmentation data. Then `M_searchExistingProjects` and `H_verifySimilarity` obtain a list of similar projects in the project database. The team connects with other business units working on the identified similar projects and learns from their subject matter experts (SME). After considering the collected information, marital status, race, and sex are determined to be sensitive features.

The company would prefer to avoid the bias of offering discounts for the service disproportionally to any of the demographic subgroups identified by the sensitive features. After consulting legal and compliance, `H_riskAssessment` determines that the project should proceed through the SIFT system due to a risk of potential bias. Prior projects used in the company's standard model flow, and there are no additional cost or computational constraints for this project. Thus, the standard model flow is selected. Disparate impact is the bias detection metric specified in the standard flow for marketing applications. The specified fairness range is $(0.8, 1.2)$, outside of which bias is detected. No model is deployed, thus `M_identifyNextPipeline` sets the next stage as Pre-model, and the `bias_history` and `model_history` objects are initialized.

**Pre-model.** After data preparation, `M_detectSparseGroups` checks that each subgroup defined by each sensitive feature makes up at least 10% of the data samples. The output of this function indicates that there is an under-representation of non-white customers in the dataset. The SIFT user decides to collect additional data for the project through the function `H_verifyGetMoreData`. The project is terminated and the `bias_history` is updated accordingly. The project is added to the Project database with a timeout set to 1-year pursuant to the company's settings.

**Listing 1: Bias history of marketing project 1**

```
{"bias_history": [
  { 'step' : 0,
    'sift_pipeline' : 'Information gathering',
    'bias_features' : '',
    'bias_detection_function' : '',
    'bias_mitigation_function : '',
    'mitigation_success_status' : '',
    'details' : 'Risk assessment indicates project should proceed
         through SIFT.'},
  { 'step' : 1,
    'sift_pipeline' : 'Pre-model',
    'bias_features' : '{'sex','race','marital_status'}',
    'bias_detection_function' : 'computeSampProportion',
    'bias_mitigation_function : '',
    'mitigation_success_status' : '',
    'details' : 'Get additional data.'},
  { 'step' : 2,
    'sift_pipeline' : 'Exit SIFT',
    'bias_features' : '',
    'bias_detection_function' : '',
    'bias_mitigation_function : '',
    'mitigation_success_status' : '',
    'details' : 'Team will collect additional data.  Project
         terminated and added to project database.'}]}
```

#### 4.1.2 SIFT implementation - Project 2.

**Information gathering.** Six-months later, after collecting additional data from a new survey, Team B initializes a `sift_project` object with project id NewSvc2020, populating `sift_data` with the new data. `M_searchExistingProjects` and `H_verifySimilarity` obtain a list of similar projects in the project database, and a pointer to Svc2020 is added to the list in `older_versions`. Based on the information available from Svc2020, `H_riskAssessment` quickly determines that NewSvc2020 should proceed through SIFT. Project Svc2020's `sens_features` and model flow selection are copied into the new project. No model is deployed, thus `M_identifyNextPipeline` sets the next stage as Pre-model, and new `bias_history` and `model_history` objects are initialized.

**Pre-model.** After data preparation, `M_detectSparseGroups` does not identify any sparse groups in the new dataset. `M_detectProxyFeatures` performs a Chi-Square test for independence on each (sensitive feature, non-sensitive feature) pair and compares the p-value against a Bonferroni corrected threshold of $0.01/m$, where $m$ is the number of non-sensitive features. No proxy features are identified by the function. Lastly, `M_detectBias` computes Disparate Impact between $y$ and each sensitive feature (Table 1, Column 2). All results are within the fairness range of (0.8, 1.2), so no marginalized groups are detected. The pipeline updates the `bias_history` after each of these steps with the corresponding bias detection algorithm and result. The SIFT user decides to proceed to the Model-involved pipeline through `H_riskAssessment`.

**Table 1: Bias detection metric results for Project 2**

| | Disparate Impact | | |
|---|---|---|---|
| Sensitive Feature | $y$ | Original Model | Debiased Model |
| marital_status | 0.85 | 0.82 | 0.83 |
| race | 0.96 | 0.97 | 1.00 |
| sex | 0.84 | 0.79 | 0.88 |

**Listing 2: Bias history of marketing project 2**

```
{"bias_history": [
  { 'step' : 0,
    'sift_pipeline' : 'Information gathering',
    'details' : 'Risk assessment indicates project should proceed
         through SIFT.'},
  { 'step' : 1,
    'sift_pipeline' : 'Pre-model',
    'bias_features' : '{'sex','race','marital_status'}',
    'bias_detection_function' : 'computeSampProportion',
    'details' : 'No sparse groups detected.'},
  { 'step' : 2,
    'sift_pipeline' : 'Pre-model',
    'bias_features' : '{'sex','race','marital_status'}',
    'bias_detection_function' : 'computeChiSqTest',
    'details' : 'No proxy features detected.'},
  { 'step' : 3,
    'sift_pipeline' : 'Pre-model',
    'bias_features' : '{'sex','race','marital_status'}',
    'bias_detection_function' : 'computeDispImpact',
    'details' : 'No marginalized groups detected.'},
  { 'step' : 4,
    'sift_pipeline' : 'Model-involved',
    'bias_features' : '{'sex','race','marital_status'}',
    'bias_detection_function' : 'computeDispImpact',
    'bias_mitigation_function' : 'adversarialDebiasing',
    'mitigation_success_status' : 'TRUE',
    'details' : 'Bias detected in model outcome. In-processing
         strategy implemented.'},
  { 'step' : 5,
    'sift_pipeline' : 'Exit SIFT',
    'details' : 'Project scheduled for deployment and added to
         project database.'}]}
```

**Model-involved.** `M_getPreModelBiasStatus` does not indicate any bias was detected in the Pre-model pipeline. The SIFT system proceeds to `M_trainModel`, which splits the data evenly into a training and test set, and trains a logistic regression model. The model has a test-set accuracy of 77.6%. The `model_history`

object is updated accordingly. The predicted outcomes for the test-set are passed to `M_detectBias`, which computes Disparate Impact as set by the company's standard flow settings (Table 1, Column 3). The results show that bias is detected on the basis of the sensitive feature `'sex'` since 0.79 is outside of the fairness range.

Next, as part of standard flow, `M_inProcessingMitigation` applies Adversarial Debiasing [51] (`adversarialDebiasing`) to correct the detected bias. The `model_history` object is updated accordingly. A final check using `M_detectBias` confirms the absence of any bias in the predicted outcomes of the debiased model (Table 1, Column 4). The test-set accuracy of the debiased model is 76.2%. These steps are recorded in `bias_history`.

`H_riskAssesment` confirms only a minimal drop in accuracy between the original and debiased models and shares `bias_history` with legal and compliance, who confirm that the bias has been addressed. The `sift_project` object is updated now and returned to the ML projects database as scheduled for deployment.

Given the importance of time-to-market in campaign deployment, copying the information collected by Team A and other similar projects was an important time saving measure for Team B.

## 4.2 Hiring

Many studies in the past few decades have found evidence of discrimination in hiring practices, especially against women and minorities [38, 39]. A study [39] of 18 algorithmic hiring vendors found that empirical analysis of their algorithms was challenging due to model opaqueness and limited public information.

We apply SIFT in a hiring scenario adapting the post-processing mitigation approach of [26] and applying it on the probability scores from a logistic regression model built on the UCI Adult dataset. Suppose a company performs its initial candidate screening using ML models that score and rank all applicants. Based on an existing model used to select candidates for a past position, the company selects a list of top $k = 500$ candidates, say $C_k$, for further interviews. However, the company first wants to ensure that this list proportionately represents candidates from all relevant sensitive categories, as compared to the full set of candidates $C$.

*4.2.1 SIFT implementation.* As the deployed model is available, the project skips the Pre-model and Model-involved pipelines.

**Information gathering.** The ML team initializes a new `sift_project` and `sift_data`. `M_searchExistingProjects` and `H_verifySimilarity` return an older project that built an ML model to screen applicants for a past position as similar, which had race, sex, and country of origin as sensitive features. Consulting past team and HR, `H_identifySensitiveCategories` lets team keep same features. `M_identifyNextPipeline` returns "Outcome-involved" as the next stage as model is deployed.

**Outcome-involved.** The team uses `M_detectCovariateShift` to identify shifted sensitive features. Suppose this detects only race as a shifted feature. It is assumed that the previous project performed its own bias detection and mitigation, so the new project need only consider the shifted feature. A demographic parity check (`M_detectBias`) compares non-White candidate proportions:

$$P(\text{race}(c) \neq \text{White} | c \in C_k) = 0.068,$$
$$P(\text{race}(c) \neq \text{White} | c \in C) = 0.1399.$$

Then `M_postProcessingMitigation` performs bias mitigation using the re-ranking approach given in Algorithm 1 of [26] (`reRankingPostProc`) to come up with $C'_k$, a new list of $k$ candidates. Reapplying `M_detectBias` gives the new proportion:

$$P(\text{race}(c) \neq \text{White} | c \in C'_k) = 0.14.$$

A final risk assessment considers the full bias history, and moves the candidates in $C'_k$ to the next phase of the recruiting process.

Possible risk factors considered in `H_riskAssessment` at the end of the pipeline include any compliance guidelines (such as [14]), the legal threshold for demographic parity comparisons, and the difference between predicted accuracy in the lists $C_k$ and $C'_k$. Model reuse from the older, deployed project shows SIFT's cost saving advantage enabled by information reuse.

**Listing 3: Bias history of hiring project**

```
{"bias_history": [
  { 'step' : 0,
    'sift_pipeline' : 'Information gathering',
    'details' : 'Risk assessment indicates project should proceed
         through SIFT.'},
  { 'step' : 1,
    'sift_pipeline' : 'Outcome-involved',
    'bias_features' : '{'race','sex','country_of_origin'}',
    'bias_detection_function' : 'computeKS2SampTest',
    'details' : 'Covariate shift detected.'},
  { 'step' : 2,
    'sift_pipeline' : 'Outcome-involved',
    'bias_features' : '{'race'}',
    'bias_detection_function' : 'computeDemographicParity',
    'bias_mitigation_function : 'reRankingPostProc',
    'mitigation_success_status' : 'TRUE',
    'details' : 'Bias detected in model outcome.  Post-processing
         strategy implemented.'},
  { 'step' : 3,
    'sift_pipeline' : 'Exit SIFT',
    'details' : 'Project added to project database.'}]}
```

# 5 CHALLENGES IN OPERATIONALIZATION

Adoption and operationalization of SIFT relies on three key components of the `sift_project` object; first, a database of ML projects with the documented bias history in building and deploying them. Second, codification within SIFT of the human oversight guidelines, best practices and restrictions that are specific to the industry and use cases that ML practitioners must adhere to while building their models. Last, providing the data scientists options in choosing the right tools to address potential biases, that can be tailored to the needs of each new use case and letting them integrate those tools into SIFT pipelines seamlessly. In the following we discuss the operationalization of these three components, and some limitations.

## 5.1 Operationalizing ML database

Identifying the list of ML projects and resources allocated to them within large enterprises from outside is impossible due to the proprietary nature of such information. But valuable proxies demonstrate the growing investment in and importance of ML to the everyday operations and business decisions of various companies. Algorithmia's blind survey of 303 employees across companies representing a range of ML maturity found that 28% of companies increased their ML budgets by more than 25% in 2019 [2]. Internal ML-focused conferences can draw hundreds if not thousands of employees to present their work and learn about projects across the company. For example, UberML, Uber's annual internal ML conference, drew 500 employees from 50 groups, and an internal Amazon AI/ML

event drew thousands [41]. Large tech companies with dedicated research organizations including Google, Facebook, Uber, and Amazon regularly publish original research and present at leading ML conferences, and the list of papers published by researchers across these companies number in hundreds.

Two of the three top ML use cases, generating customer insights and intelligence, and improving customer experience [2], are inherently customer-facing. Because decisions based on ML projects have the potential to impact billions of external customers, ensuring fairness in product offerings, customer service treatment, and customer analytics is key. Therefore, significant savings are associated with a streamlined process that can detect bias across new and existing projects. Among Algorithmia's surveyed companies employing ML solutions, 45% had models in production for at least 1 year [2]. So it is reasonable to assume that documentation of data and methodology already exist and would be available for addition to a bias and model history database.

Given that ML projects routinely reuse data (often with modifications or sampling) and tweaked models, it is essential to examine the provenance and reliability of datasets, as well as any past applications and concerns. Incentivizing project managers to enforce adequate data and model documentation helps maintain the accuracy of components in a `sift_project` object. The cost for doing so is traded off against financial risk metrics that approximate the cost of negative PR and brand impact. Long-term adoption of SIFT in the enterprise will reduce projected overall cost. Such explicit analysis of risk vs. return will drive better institutional decision-making.

## 5.2 Operationalizing human oversight

Apart from the mechanizable functions (which can be implemented as code), SIFT identifies seven `H_` (human in the loop) functions: `H_prepareData`, `H_identifySensitiveCategories`, `H_verifyGetMoreData`, `H_riskAssessment`, `H_verifyDropProxy`, `H_verifySimilarity`, and `H_verifyRetrainModel`. These `H_` functions are not mechanizable by design and require human oversight. Some of them relate to critical bias detection and mitigation tasks and could benefit from the knowledge of a subject matter expert (SME) covering specific areas in a large enterprise, such as Human Resources, Communications (PR), Legal, Privacy, and Compliance. Most large enterprises have experts in these areas who have dealt with problems arising out of inadvertent actions in the past, and thus are in a position to better anticipate inadvertent ML bias. For example, a compliance expert will ensure that the guidelines provided around customer privacy are properly addressed in compliance with local, state, federal, or even international regulations (such as GDPR). The oversight guides are written with data scientists as target audience (as opposed to high level enterprise-wide principles) and assists them and project managers (PM) to address potential concerns in a timely manner.

The large number of ML projects in medium and large enterprises coupled with limited number of SMEs in any policy area, could raise questions on the efficacy of the `H_` functions and their potential to decelerate the pace of development and deployment of an ML project. Such interventions, while inconvenient, safeguard against future unintended consequences of rapid deployment of a ML product both on society and to the enterprise's brand. Moreover, the curation of SME-guided questions beforehand helps scale the simultaneous use of SIFT on a large number of ML projects as it obviates the need to consult SMEs at every stage. If project-specific issues pop up repeatedly that are not covered by a particular oversight guide, the relevant SME could be asked to revise it. See Appendix E for a sample list of seed questions.

SIFT automatically points to the relevant SME guides at the right pipeline and allows the PM to decide if they need to communicate with the SME for additional clarifications. Before deploying SIFT, each SME is presented with a relevant and tailored set of seed questions to assist them in converging on their oversight guideline.

## 5.3 Integration of existing tools and artifacts

The holistic human-in-the-loop framework of SIFT allows data scientists and project managers to be flexible in choosing bias monitoring tools and artifacts optimal for their own team and enterprise. Data scientists can get started on bias detection and mitigation with methods implemented in the open-source libraries like AIF360 or LiFT, then code up other methods on an on-demand basis. Existing implementations of DataSheets [25], Factsheets [3] or Model cards [36] in the enterprise can inform or enrich components of the `sift_project` object. The `H_riskAssessment` steps can be amply facilitated by algorithmic audits [40] and fairness checklists [33].

## 5.4 Limitations

Our current proposal of SIFT has a few limitations. Firstly, during the early stages of implementing SIFT there will be few 'similar' projects leading to higher per-project cost for bias detection and mitigation. However, as time progresses SIFT's novel reuse of bias history will be progressively more effective, thus significantly diminishing the *overall* cost of bias monitoring to the enterprise. Secondly, it is non-trivial to quickly determine how changes to data or model may impact bias at different stages of the lifecycle. Research remains to be done to compartmentalize the potential impact of such modifications. Effectively, we would like to limit the parts of the lifecycle that would be impacted by any modifications to the input data, changes to the model, application to new use cases etc. Next, for specific use cases, possible ways to mechanize one or more human components in the SIFT pipelines are worth exploring. Lastly, SIFT in its current form will not handle ML models such as adaptive algorithms that are constantly updating and learning from streaming data, or unsupervised learning problems. These types of problems will require different class structures and pipelines, but would still benefit from the knowledge-sharing aspects of SIFT.

# 6 CONCLUSION

In this paper, we have shown how bias detection and mitigation in ML may be done in an enterprise setting in a holistic and transparent manner. Industry's technical challenges include diversity of use cases, datasets, and audiences. Further, companies have to ensure that no demographic bias arises even well after deployment while adhering to policy recommendations and meeting compliance requirements. While SIFT does not handle all these problems, it shows how a large class of ML projects can follow a framework to reduce chances of bias going undetected until it is too late.

# REFERENCES


[1] J. A. Adebayo. 2016. *FairML : Toolbox for diagnosing bias in predictive modeling.* Master's thesis. MIT. https://github.com/adebayoj/fairml.

[2] Algorithmia. 2020. *2020 State of Enterprise Machine Learning.* https://info.algorithmia.com/hubfs/2019/Whitepapers/The-State-of-Enterprise-ML-2020/Algorithmia_2020_State_of_Enterprise_ML.pdf

[3] M. Arnold et al. 2018. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *arXiv:1808.07261* (2018).

[4] A. Backurs et al. 2019. Scalable Fair Clustering. In *ICML-2019*.

[5] N. Bantilan. 2018. Themis-ml: A Fairness-Aware ML Interface for End-To-End Discrimination Discovery and Mitigation. *J. Tech. Hum. Serv.* 36, 1 (2018), 15–30.

[6] S. Barocas and A. D. Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[7] R. K. E. Bellamy et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv:1810.01943* (2018).

[8] F. Calmon et al. 2017. Optimized Pre-Processing for Discrimination Prevention. In *NIPS-2017*.

[9] A. Cavoukian, S. Taylor, and M. E. Abrams. 2010. Privacy by Design: essential for organizational accountability and strong business practices. *Identity Inform. Soc.* 3, 2 (2010), 405–413.

[10] E. Celis, A. Mehrotra, and N. Vishnoi. 2019. Toward Controlling Discrimination in Online Ad Auctions. In *ICML-2019*.

[11] L. E. Celis et al. 2018. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. *arXiv:1806.06055* (2018).

[12] A. Chouldechova and Roth A. 2020. A Snapshot of the Frontiers of Fairness in Machine Learning. *Commun. ACM* 63 (2020), 82–89.

[13] ClamAV. 2020. Available at: https://www.clamav.net/about.

[14] Code of Federal Regulations. 2017. *Uniform Guidelines on Employee Selection Procedures.* https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml

[15] J. Cook. 2018. *Amazon scraps 'sexist AI' recruiting tool that showed bias against women.* https://www.telegraph.co.uk/technology/2018/10/10/amazon-scraps-sexist-ai-recruiting-tool-showed-bias-against/

[16] B. D'Alessandro, C. O'Neil, and T. LaGatta. 2017. Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. *Big Data* 5, 2 (2017), 120–134.

[17] A. Datta et al. 2018. Discrimination in Online Advertising: A Multidisciplinary Inquiry. In *FAT-2018*.

[18] A. Datta, M. C. Tschantz, and A. Datta. 2015. Automated Experiments on Ad Privacy Settings. *Priv. Enh. Technologies* 2015, 1 (2015), 102–112.

[19] S.-C. Davies, E. Pierson, A. Feller, et al. 2017. Algorithmic decision making and the cost of fairness. In *KDD-2017*.

[20] D. Dua and C. Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[21] M. Dudik et al. 2020. *Fairlearn.* https://github.com/fairlearn/fairlearn

[22] C. Dwork et al. 2012. Fairness through awareness. In *ITCS-2012*. 214–226.

[23] M. Feldman et al. 2015. Certifying and Removing Disparate Impact. In *KDD-2015*.

[24] S. Galhotra, Y. Brun, and A. Meliou. 2017. Fairness Testing: Testing software for discrimination. In *ESEC/FSE-2017*.

[25] T. Gebru et al. 2018. Datasheets for datasets. *arXiv:1803.09010* (2018).

[26] S. C. Geyik, S. Ambler, and K. Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *KDD-2019*.

[27] M. Hardt, E. Price, and N. Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS-2016*.

[28] K. Holstein et al. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *CHI-2019*.

[29] F. Kamiran and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[30] F. Kamiran, A. Karim, and X. Zhang. 2016. Decision Theory for Discrimination-Aware Classification. In *NIPS-2016*.

[31] T. Kamishima et al. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*.

[32] A. Lambrecht and C. E. Tucker. 2019. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Manage. Sci.* 65, 7 (2019), 2966–2981.

[33] M. A. Madaio et al. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *CHI-2020*.

[34] N. Mehrabi et al. 2019. A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635* (2019).

[35] N. Mehrabi et al. 2019. Debiasing Community Detection: The Importance of Lowly Connected Nodes. In *ASONAM-2019*.

[36] M. Mitchell et al. 2019. Model Cards for Model Reporting. In *FAT*-2019*.

[37] G. Pleiss et al. 2017. On Fairness and Calibration. In *NIPS-2017*.

[38] L. Quillian et al. 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proc. Natl. Acad. Sci.* 114 (2017), 10870–10875.

[39] M. Raghavan et al. 2020. Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices. In *FAT*-2020*.

[40] I. D. Raji et al. 2019. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *FAT*-2019*.

[41] J. Robinson. 2019. *How Uber Organizes Around Machine Learning.* https://medium.com/@jamal.robinson/how-uber-organizes-around-artificial-intelligence-machine-learning-665cdeb946bc

[42] Snort. 2015. Available at: https://www.snort.org/faq/what-is-snort.

[43] T. Speicher et al. 2018. Potential for Discrimination in Online Targeted Advertising. In *FAT-2018*.

[44] A. Stevens et al. 2018. *Aequitas: Bias and fairness audit.* Technical Report. Center for Data Science and Public Policy, The University of Chicago. https://github.com/dssg/aequitas.

[45] A. Tramer et al. 2017. FairTest: Discovering unwarranted associations in data-driven applications. In *EuroS&P-2017*.

[46] S. Vasudevan and K. Kenthapadi. 2020. LiFT: A Scalable Framework for Measuring Fairness in ML Applications. In *CIKM-2020*.

[47] M. Veale, M. Van Cleek, and R. Binns. 2018. Fairness and accountability design needs for algorithmic support in highstakes public sector decision-making. In *CHI-2018*.

[48] J. Vincent. 2018. *Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech.* https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

[49] M. Zehlike et al. 2017. *Fairness Measures: Datasets and software for detecting algorithmic discrimination.* http://fairness-measures.org

[50] R. Zemel et al. 2013. Learning Fair Representations. In *ICML-2013*.

[51] B. Zhang, B. Lemoine, and M. Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *AAAI-2018*.

[52] L. Zhang, Y. Wu, and X. Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *IJCAI-2017*.

# APPENDIX

## A CLASS COMPONENTS AND METHODS

We provide additional information about the `sift_data` class components, `bias_history` methods, and the `model_history` class components here.

### A.1 SIFT data class components

The list of components of the `sift_data` class are:

- `raw_data` – a connection to the data for the project, such as a dataframe or a connection to a distributed file system that provides access to the data,
- `data_definitions` – a dictionary with definitions for each variable in the dataset,
- `y` – a variable with the name of the response variable,
- `X` – a list with the names of the predictors,
- `outcome` – a connection to the predicted outcomes from the ML model; This will be pre-populated if the model is deployed else populated after model training,
- `sens_features` – a list with the names of the sensitive features that contain categories that might suffer from biased treatment,
- `sens_features_summary` – a dictionary of dictionaries where each key denotes a specific characteristic relevant to bias-investigation, such as sparse groups, proxy features, or marginalized groups, and each corresponding value defines a dictionary with keys denoting sensitive features as identified in `sens_features` and values being lists describing the corresponding characteristics.

As a concrete example of the `sens_features_summary`, assume income is a sensitive feature and we are trying to identify its proxy features, which are education and race. This will be represented in the `sens_features_summary` as

```
sens_features_summary = {
  proxy_features: {income: [education, race]}
}
```

If multiple features were to jointly act as proxy for a sensitive feature they could be represented as a list within the list of proxy features.

The components of the `sift_data` class may be populated manually or via mechanizable extraction functions that take the `data_location` and project description as arguments.

### A.2 Bias history class methods

During the course of the project, the `bias_history` object is updated through methods associated with the class to reflect the bias detection and mitigation steps performed. To get the current step we use the method:

```
def getLatestStep(sift_project):
  return sift_project.bias_history[-1]['step']
```

Components are added to the current step to track bias investigation at each step using the class method:

```
def insertBiasHistoryAt(sift_project, insert_at, **kwargs):
  if (insert_at > sift_project.getLatestStep()):
    return 'cannot insert outside current history range'
  #      fill in components and corresponding values from **kwargs
  for key, value in kwargs.item():
    if key in list(sift_project.bias_history[0].keys()):
      sift_project.bias_history[insert_at][key] = value
    else:
      print(f'{key} is not an attribute of bias history')
```

Lastly, the next step in the bias history is added using the class method:

```
def addBiasHistoryStep(sift_project, **kwargs):
  # append new step
  sift_project.bias_history.append(
    {'step'                    :sift_project.getLatestStep() + 1,
     'sift_pipeline'           :None,
     'bias_features'           :None,
     'bias_detection_function' :None,
     'bias_mitigation_function':None,
     'mitigation_success_status':None,
     'details'                 :None})

  #      fill in components and corresponding values from **kwargs
  for key, value in kwargs.item():
    if (key != 'step') and
      (key in  list(sift_project.bias_history[0].keys())):
      insert_at = sift_project.getLatestStep()
      sift_project.bias_history[insert_at][key] = values
```

### A.3 Model history class components

The SIFT model history class keeps track of the modeling efforts. The list of components of the `model_history` class are:

- `step` – a counter capturing the place in the sequence of modeling tasks performed,
- `seed` – the random seed used in the modeling process to ensure reproducibility,
- `train_index` – the set of indices in the raw data used for training the ML model,
- `test_index` – the set of indices in the raw data used for testing the ML model,
- `fitted_model` – includes the loss function to be optimized and its value for the fitted model, the tuning parameters, and the estimated model,
- `perf_metric` – a dictionary that includes the name of the performance metric used to evaluate the model and the performance metric value for the test-set; For example, the performance metric for a classification problems could be accuracy, precision, recall, etc.,
- `is_deployed` – a flag indicating if the model is deployed. If `True` at step 0, then this indicates that the project was initiated with an earlier model already deployed.

## B AUXILIARY FUNCTIONS

We provide descriptions of the auxiliary functions that appear in Section 3, and illustrate them in B.3. We assume these functions would be standardized by the company, but could depend on the specific data or project application and could be overridden by the user when necessary. We provide examples of these functions based on their implementation in the use cases in Section 4.

**`M_selectSparsityFunction`:** This selects the sparsity function and its threshold. The sparsity function calculates the sparsity of a particular subgroup of a sensitive feature. For example, the sparsity function could compute the percent of samples that belong to the subgroup (`computeSampProportion`).

**`M_selectDependenceFunction`:** This selects the dependence function and its threshold. The dependence function computes the statistical dependence between a sensitive attribute and non-sensitive attribute in order to identify non-sensitive features that might act

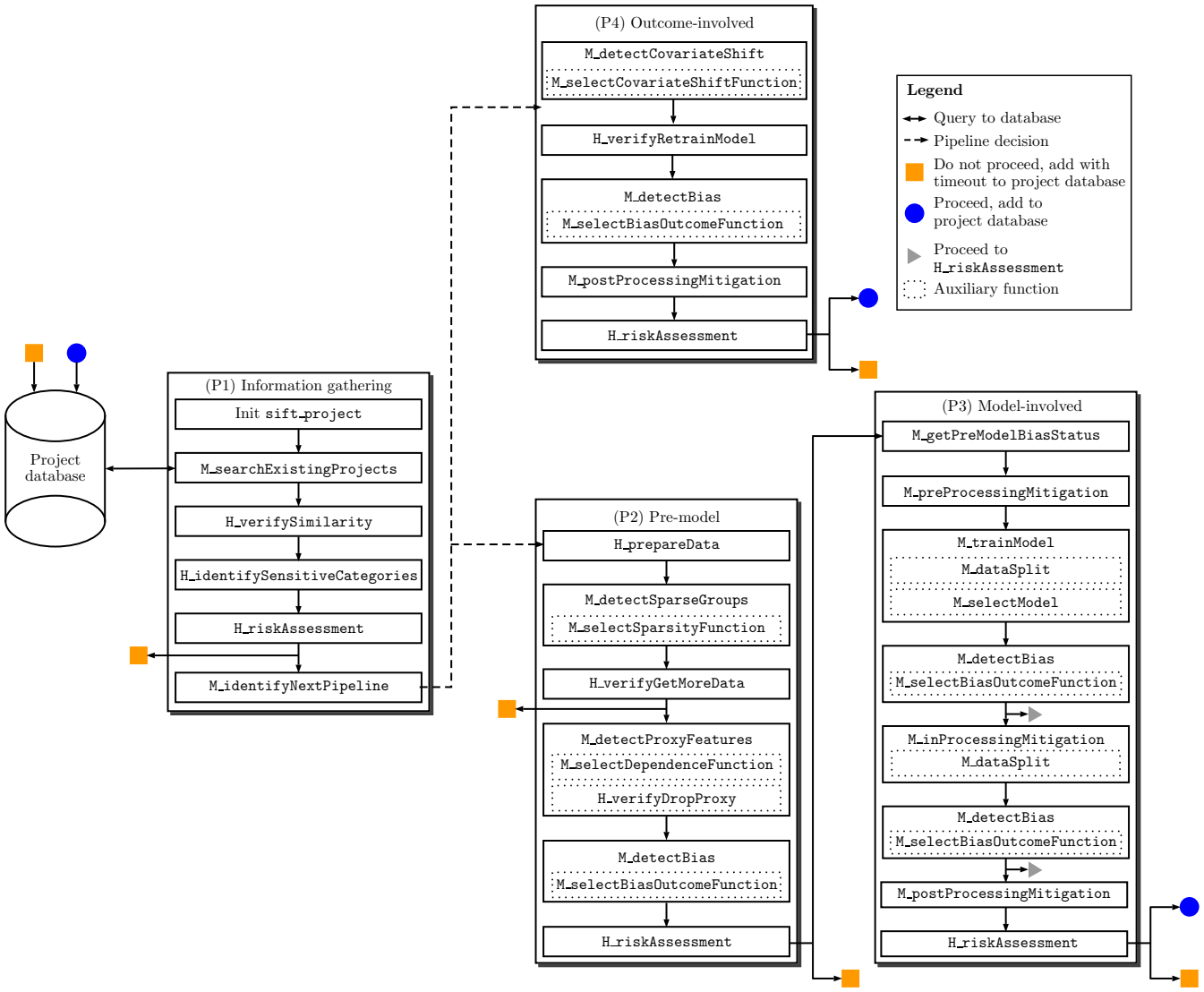


**Figure B.3: A full diagram of the SIFT pipelines and functions.**

as a proxy for other sensitive categories. For example, the dependence function could return the p-value from a Chi-Square test of independence between the sensitive feature and a non-sensitive feature (`computeChiSqTest`).

**`M_selectBiasOutcomeFunction`:** This selects the function to compute the bias detection metric for the model outcome and its corresonding fairness ranges. As examples, `M_selectBiasOutcomeFunction` selects a function that computes Disparate Impact in the marketing use case (`computeDispImpact`) and selects a function that computes Demographic Parity in the hiring use case (`computeDemographicParity`).

**`M_dataSplit`:** This function sets the random seed and splits the data into a training and test set using the specified split ratio.

**`M_selectModel`:** This function selects the ML model and performance metric. For example, in the marketing use case, this selects logistic regression as the ML model and classification accuracy as the performance metric.

**`M_selectCovariateShiftFunction`:** This selects the covariate shift function, which calculates the shift in distribution of a feature from one period to the next. For example, the covariate shift function could compute the p-value for a two sample Kolmogorov-Smirnov test based on the two features (`computeKS2SampTest`).

## C EXAMPLE STANDARD FLOW

Here we provide a sample standard flow process that works sequentially through the six steps described in Section 3.3. For ease of presentation, this example assumes there is only one sensitive feature.

```
premodel_status = M_getPreModelStatus(
  current_project.data.sens_features_summary)
if premodel_status:
```

```
  # Step: Pre-processing mitigation
  current_project.data, preProcFunc =
    M_preProcessingMitigation(current_project)
  ## Update bias history and add a step
  current_bias_step = getLatestStep(current_project)
  current_project.insertBiasHistoryAt(current_bias_step,
    **{'bias_features' = current_project.data.sens_features,
    'bias_mitigation_function' = preProcFunc.__name__})
  current_project.addBiasHistoryStep(
    **{'sift_pipeline' = "Model-involved"})
# Step: Model training
fitted_model = M_trainModel(current_project)
## Update model history and add a step
...
# Step: Model outcome bias detection
biased_groups_list, biasFunc = M_detectBiasModelOutput(
  current_project.data.outcome,
  current_project.data.raw_data[current_project.data.sens_features
      ])
## Update bias history
current_bias_step = getLatestStep(current_project)
current_project.insertBiasHistoryAt(current_bias_step,
  **{'bias_features' = current_project.data.sens_features,
  'bias_detection_function' = biasFunc.__name__})
if not biased_groups_list:
  ## Set mitigation_success_status in previous step to TRUE
  ...
else:
  ## Set mitigation_success_status in previous step to FALSE
  current_project.insertBiasHistoryAt(current_bias_step - 1,
    **{'mitigation_success_status' = FALSE})
  # Step: In-processing mitigation
  new_fitted_model, inProcFunc =
    M_inProcessingMitigation(current_project)
  ## Update bias history and add a step
  current_project.insertBiasHistoryAt(current_bias_step,
    **{'bias_mitigation_function' = inProcFunc.__name__}
  current_project.addBiasHistoryStep(
    **{'sift_pipeline' = "Model-involved"})
  ## Update model history
  ...
  # Step: Model outcome bias detection
  biased_groups_list, biasFunc = M_detectBiasModelOutput(
    current_project.data.outcome,
    current_project.data.raw_data[current_project.data.
        sens_features])
  ## Update bias history
  current_bias_step = getLatestStep(current_project)
  current_project.insertBiasHistoryAt(current_bias_step,
    **{bias_features' = current_project.data.sens_features,
    'bias_detection_function' = biasFunc.__name__})
  if not biased_groups_list:
    ## Set mitigation_success_status in previous step to TRUE
    ...
  else:
    ## Set mitigation_success_status in previous step to FALSE
    current_project.insertBiasHistoryAt(current_bias_step - 1,
      **{'mitigation_success_status' = FALSE})
    # Step: Post-processing mitigation
    current_project.data.outcome, postProcFunc =
      M_postProcessingMitigation(current_project)
    ## Update bias history
    current_project.insertBiasHistoryAt(current_bias_step,
      **{'bias_mitigation_function' = postProcFunc.__name__})
    # Step: Model outcome bias detection
    bias_status, biasFunc = M_detectBiasModelOutput(
      current_project.data.outcome,
      current_project.data.raw_data[current_project.data.
          sens_features])
    ## Update bias history
    current_bias_step = getLatestStep(current_project)
    current_project.insertBiasHistoryAt(current_bias_step,
      **{'bias_features' = current_project.data.sens_features,
      'bias_detection_function' = biasFunc.__name__,
      'mitigation_success_status' = (len(biased_groups_list) == 0)
          })
# Step: Model-involved risk assessment
go_ahead_status = H_riskAssessment(current_project)
if go_ahead_status is "Do not proceed":
  current_project.metadata['project_status'] = "Terminated"
  current_project.timeout = company_timeout
else:
  current_project.metadata.project_status =
    "Scheduled for deployment"
```

# D SIMULATION DETAILS FOR MARKETING USE CASE

We use demographic data from the UCI Adult dataset and remove all examples with missing information, resulting in $n = 45,222$ examples. In our experiments, we use income, sex,

- age - binned as $[17,25]$, $[26,35]$, $[36,45]$, $[46,55]$, $[56,65]$, $[66,75]$, or 75+, and converted to its one hot encoding,
- marital_status - converted to "married" or "single",
- race - converted to "white" or "non-white".

We treat {marital_status, race, sex} as the set of sensitive features.

To generate consumer segments that are correlated with the sensitive features, we simulate $C^c_{i,j} \sim Bin(\tilde{p_i})$ for $j = 1,\ldots,5$ and $i = 1,\ldots,n$, where

$$\tilde{p_i} = \frac{\exp\left(-1 \ + \ \mathbb{1}_{\{marital\_status_i=Married\}} \ + \ \mathbb{1}_{\{sex_i=Male\}}\right)}{1+\exp\left(-1 \ + \ \mathbb{1}_{\{marital\_status_i=Married\}} \ + \ \mathbb{1}_{\{sex_i=Male\}}\right)}.$$

In addition, we simulate consumer segments $C^u_{i,j} \sim Bin(p_j)$ for $j = 1,\ldots,45$ and $i = 1,\ldots,n$, where $p_j \sim U(0.2,0.8)$, to be consumer segments that are uncorrelated with the sensitive features.

We define $\mathbf{X} = [\text{age}, \text{income}, \mathbf{C^c}, \mathbf{C^u}]$ to be the set of features for model training. To simulate the target variable, $\mathbf{y}$, we define $\boldsymbol{\beta}$ to be a vector of coefficients corresponding to the features in $\mathbf{X}$. We simulate coefficients for income, $\mathbf{C^c}$, and the first 10 features in $\mathbf{C^u}$ from $U(-2,2.5)$. We set all other coefficients to zero. Then $y_i \sim Bin(p^y_i)$ for $i = 1,\ldots,n$, where

$$p^y_i = \frac{exp(-0.5 \ + \ \mathbf{X}_i\boldsymbol{\beta} \ + \ z_i)}{1+exp(-0.5 \ + \ \mathbf{X}_i\boldsymbol{\beta} \ + \ z_i)}.$$

Here $z_i \sim N(0,1)$, for $i = 1,\ldots,n$, prevents a perfect model fit.

# E SAMPLE SEED QUESTIONS FOR THE SME

Each large organization has SMEs in fields relevant to bias detection and mitigation, including e.g. Legal, Privacy, Compliance, Public Relations, and Human Resources. The below seed questions may be shared with a SME in each field. Each SME provides guidance to the data scientist according to their expertise in the form of a human oversight guide – a question and answer document organized by project pipeline, the appropriate section of which SIFT prompts the data scientist to consult when and where necessary. Where warranted, we include in parenthesis the specific H_ human function to which the question refers.

**P1 Information gathering pipeline**

- What types of data fall under the purview of the SME?
- What laws/regulations are in place for this data and its use? (H_riskAssessment)
- When do machine learning projects typically raise concern within the SME's field? Are there external examples related to ML bias a data scientist should be aware of? (H_riskAssessment)

- What qualities would enable the data scientist to assess whether a project is low/medium/high risk? Are there ways to mitigate related risks? (`H_riskAssessment`)
- Are there data elements that we are not allowed to look at or need specific approval to use in bias mitigation strategies? (`H_riskAssessment`)
- What metrics are typically used to evaluate fairness? Is there a standard accepted threshold for each metric?
- What vetting is done for 3rd party data and what liabilities do we have in using the data? (`H_riskAssessment`)

**P2 Pre-model pipeline**

- How would use of data (i.e. descriptive vs. predictive) impact whether a project is considered low/medium/high risk? (`H_riskAssessment`)
- If a project requires additional data, what are the necessary approval steps? (`H_verifyGetAdditionalData`)
- Are proxy features a concern? Are there cases where proxy features are acceptable and/or appropriate from a business perspective? (`H_verifyDropProxy/H_riskAssessment`)

**P4 Outcome-involved pipeline**

- What vetting is done for internal models? 3rd party models?
- Are there outcomes that always carry risk from the SME's perspective relative to their field? (`H_riskAssessment`)
- Who should a data scientist contact for additional information?

Questions may require iteration in consultation with a specific SME to further elaborate upon the question and capture information most relevant from that field for the SIFT user. For example, a Pipeline 1 question to `H_identifySensitiveCategories` for Human Resources may be *"What are the protected classes that may inform identification of sensitive categories?"*. A related question for Privacy may be, *"Are there privacy concerns in identifying protected classes? Do these concerns vary depending on the data subjects?"*

Similarly, additional focused questions may be created for field. A Pipeline 1 question specific to HR may be *"What laws/regulations are in place for hiring or employee related data?"*, while one posted to Privacy may be, *"Are there privacy concerns around the reuse of data/models/bias history across business units?"*

The guidance provided by a SME in the form of an answer to each question may be similar to the following:

**Human Resources**

**Q:** *What types of data fall under the purview of the SME?*

**A:** There are some key issues surrounding the use of people data, including the importance of having a deep understanding of data elements used. During modeling there might be a correlation between a school and some outcome, but discrimination in education exists. Performance-related discrimination may be good but race-related is bad. Working with HR will provide a clear understanding of the people data elements.

**Public Relations**

**Q:** *Are there external examples related to ML bias a data scientist should be aware of?*

**A:** Media often identifies cases where individuals don't seem to be treated fairly and/or seem to have the same opportunities. The output is judged more than the input for, e.g. job applicant screening, better services in some neighborhoods, and best offers and targeted ads going to certain demographics.