Exploring discrimination: A user-centric evaluation of discrimination-aware data mining

Bettina Berendt[1] and Sören Preibusch[2]
[1] *Dept. of Computer Science; KU Leuven; Leuven, Belgium;* firstname.lastname@cs.kuleuven.be *
[2] *Computer Laboratory; University of Cambridge; Cambridge, UK;* soeren.preibusch@cl.cam.ac.uk

IN: 2012 IEEE 12TH INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS. IEEE COMPUTER SCIENCE PRESS, 2012.

***Abstract*—Discrimination-aware data mining (DADM) aims at deriving patterns that do *not* discriminate on "unjust grounds" such as gender, ethnicity or nationality. DADM safeguards can be very helpful for decision-support applications in fields such as banking or employment. However, *constraining* data mining to exclude a fixed enumeration of potentially discriminatory features is too restrictive. It should be complemented by *exploratory* DADM. We discuss these two forms of DADM and their requirements for evaluation, and we discuss and refine our DCUBE-GUI tool as a system for exploratory DADM. In a user study administered via Mechanical Turk, we show that tools such as DCUBE-GUI can successfully assist novice users in exploring discrimination in data mining.**

***Keywords*-Discrimination Discovery; Evaluation; User studies; Responsible data mining; Mechanical Turk.**

I. INTRODUCTION

"Are your methods for discrimination-aware data mining being used, and if so, how do you know this works?" A good answer to this question is the litmus test for the success of discrimination-aware data mining methods. Still, it remains surprisingly hard to get an answer. One reason may be that answering this question requires a comprehensive understanding of the use cases for DADM, evaluation measures and methods, and empirical studies.

The contributions of this paper are twofold: (1) a critical re-evaluation of key assumptions about deployment that underlie DADM today; and (2) a user study of DADM that, to the best of our knowledge, is the first described in the scientific literature. Section II will consider the first issue and subsume the relevant related work under our new classification into *constraint-oriented* vs. *exploratory* DADM. We will also contrast these methods' assumptions about users, use cases, and evaluation methodology. The second contribution is the topic of Section III, in which we derive a deployment and evaluation setting for exploratory DADM and then test it in a user study. Section IV concludes with an outlook on future work.

II. CONSTRAINT-ORIENTATION VS. EXPLORATION: A NEW FRAMEWORK FOR RELATED WORK IN DADM

To understand the range of use cases of DADM, we need to take a step back and ask about the fundamental relations between data mining (discrimination-aware or not) and discrimination (Section II-A). From this, we derive our notion of *constraint-oriented DADM* as a description of most of the current work in the field (Section II-B). While this is a very important approach, it needs to be complemented by *exploratory DADM* (Section II-C). We analyse the different assumptions that these two approaches to DADM need to make about users, use cases, and evaluation criteria (Sections II-B and II-D). We instantiate and operationalise these for our exploratory system DCUBE-GUI in Section III.

A. Data mining and discrimination

We understand *data mining* in the more general sense of "knowledge discovery" [1] and therefore consider preprocessing and deployment as integral parts. Data mining includes descriptive aspects (when it is used as exploratory data analysis) as well as prescriptive aspects (when it is used for decision support, in recommender systems, etc.).

In a wide sense, *discrimination* is to "make a distinction [...] on grounds of [some feature]"; in a narrow sense one "make[s] a distinction, esp. unjustly on grounds of race or colour or sex" [2]; a multi-disciplinary overview is provided in [3]. Such "unjust" grounds are legally codified in many countries and may include further characteristics. In the following, we will call them *discrimination-indexed features*[1]. Thus, discrimination is not the existence of some statistical imbalance (e.g., more men than women have jobs in higher management). It is a property of a decision that may lead to such an imbalance in the population or disadvantage a specific individual (such as a woman not getting a job just because of her gender).

In its descriptive role, data mining may *detect* discrimination in a data set, when statistical imbalances originate in earlier decisions. If imbalances result from something else (such as a law of nature), the detected patterns are not discrimination. DADM leverages background knowledge about discrimination-indexed features in order to detect discrimination in the narrow sense, thus going beyond standard mining.

* We thank the IWT SBO project SPION (www.spion.me) for support.

[1] otherwise called, e.g., "potentially discriminatory (PD) items" [4] or "sensitive attributes" [5], [6]. While Pedreschi et al. [4] point out that PD items may comprise more than just legally-defined sensitive attributes, they still assume the a priori existence of knowledge about these items.

In its prescriptive role, the very point of data mining is to *create* discrimination – in the wider sense: a decision rule by definition makes distinctions based on some features. The basic idea of DADM was then to turn this around and use an analysis of its patterns to *prevent creating* discrimination in the narrow sense: If discrimination per se is allowed and desired, but discrimination based on a well-circumscribed set of grounds is forbidden, then data-mining methods must prevent the generation of "bad patterns" or identify them and filter them out.[2] The remaining patterns are by definition "good" ones. Prevention is realised by a number of pre-processing and in-processing methods for DADM, and identification/filtering by a number of post-processing methods, e.g., [5-10].

As an example, we consider a typical use of data mining: one analyses old loan data to derive rules for future loan decisions. The descriptive and prescriptive roles of data mining are linked by a set of assumptions: (a) the descriptive analysis revealed imbalances that identify certain features to be predictive of undesirable outcomes (e.g., loan applicants with these properties often default on their loan), (b) existing customers and potential future customers are drawn from the same population, and thus (c) decision rules that discriminate against customers with features that have been found to be predictive of undesirable outcomes in step (a) will reduce the occurrence of these undesirable outcomes.

In this view, *DADM* is therefore but a *constraint on step (c)*, and the reduced utility of forgoing some rules must be outweighed by the (legal or otherwise) need to prevent discrimination in the narrow sense.[3] We therefore call this (classical) approach to DADM *constraint-oriented.*

B. Use cases and evaluation of constraint-oriented DADM

To the extent that discrimination is static and well-defined in terms of a fixed set of discrimination-indexed features that decisions must not be based on, and DADM's role is to act as a constraint, its best use case is a black-box approach. Ideally, the decision-maker should not even get to see the bad patterns (because they might unduly influence her).

It is important to also prevent indirect discrimination such as red-lining. Thus, discriminating against someone because he has an innocuously-seeming feature (such as living in a certain neighbourhood) is also not allowed if this feature is highly correlated with or predicts a discrimination-indexed one (such as being black). Most DADM approaches formalise and take measures against such indirect discrimination, see for example [8], [9], [5].

In this view of DADM, an effective data-mining *method* for preventing discrimination applies an agreed-upon definition of bad patterns and guarantees that it either does not find any such patterns or finds all of them and filters them out. An effective *system* architecture for preventing discrimination (a) employs effective methods and (b) disables possibly found bad patterns. Evaluation methods must therefore be based on measures of *non-existence* and *invisibility* of bad patterns. Typical use cases of such systems will involve decision makers as actors/users (for example, employees of a bank who decide on whether to give a loan or not). These may be the original data owners or third parties receiving the data. Of course, evaluation also has to integrate appropriate measures of usability.

Classical DADM approaches focus on creating effective methods in this sense and evaluating them by measures of non-existence and thereby invisibility, such as counts of successfully sanitised bad patterns, missed rules, and newly emerging "ghost rules" found in the transformed dataset [10], [5] or reduced discrimination scores [6]. Note that agreed-upon definitions of "bad patterns" are still being developed, cf. [8], [11]. DCUBE [12] and LP2DD [13] are systems that focus on detecting all assumption-based bad patterns. Systems focussing on making them invisible/ineffective could be modelled on analogous architectures proposed for privacy-protection such as [14].

C. The need for exploratory DADM

This approach, however, forgoes the advantages inherent in descriptive data mining: the exploration of data that may lead to new insights and new hypotheses to be tested. This is of utmost importance in the field of discrimination too. An exploration of data may lead to insights about new or changing forms of or grounds for discrimination, and it may lead to a pinpointing of (sub-)groups at risk within groups more obviously in danger of discrimination.

One example that is currently being discussed in sociology are the changing challenges that women face in the workplace. Much overt discrimination against women appears to have abated, thanks in no small measure to past efforts to detect discrimination against this discrimination-indexed feature, raise awareness about it, and implement measures against it. However, it increasingly appears that *mothers* now suffer from discrimination in the workplace [15]. This is not only socially relevant, but also a prime example of an emerging pattern that even a typical indirect-discrimination analysis may not notice, since the (not discrimination-indexed) feature "having children" is hardly predictive of gender. Such forms of discrimination can only become successful targets for classical DADM if the risks implied by "having children" *within* the group with feature "female" have been discovered and a new feature "mother" has been constructed. Note that such feature constructions often require background knowledge and negotiation among stakeholders. For example, the risks implied by "having no job experience" (another not discrimination-indexed feature) may be statistically equal to those of having children, but are unlikely to be accepted as unjust job-market discrimination.

[2] "Bad patterns" correspond to, e.g., "α-discriminatory rules" in [4].

[3] see for example [6], [10] for measures of utility

	discrimination (wide sense)	*discrimination (narrow sense)*
descriptive DM	detection	[no systematic relationship]
cDADM		assumption-based detection
eDADM		discovery-based detection
prescriptive DM	creation	creation is possible
cDADM		prevention of creation
eDADM		feature evaluation/construction

Table I
DATA MINING (*DM*), DISCRIMINATION, AND FOCI OF CONSTRAINT-ORIENTED (*c*) AND EXPLORATORY (*e*) DADM

(This may also be the case for a feature found by searching for indirectly discriminating rules: In an applied setting, "no known savings" [16, p.59] are likely to be regarded as a legitimate ground for rejecting a loan application, rather than as discrimination.)

We call such an approach, which focusses on *discovering* features and discrimination, *exploratory DADM*. The resulting relationships between data mining and discrimination are summarised in Table I.

D. Use cases and evaluation criteria of exploratory DADM

In the exploratory view of DADM, the *visibility* of patterns and interactive use cases are key – users must be supported in exploring, making sense of, and inspecting bad patterns further, as well as given the possibility of constructing new features for future analysis.

An effective data-mining *method* for preventing discrimination applies an agreed-upon definition of bad patterns and guarantees that it finds (or transforms) them. However, patterns are often complex and hard to interpret, and transformed patterns may be more indicative of future decisions than of past ills. Thus, not only should "bad patterns" be found, but also "dangerous items". The latter address a descriptive question (people with what features were possibly discriminated against, or simply appear to be at more risk of bad outcomes) as well as a prescriptive question (which of these features will be applied in decision rules to the detriment of people).[4] The methods for classifier learning from paired instances and for the use of ontologies proposed in [16], [17] open opportunities for such exploration.

An effective *system* architecture for preventing discrimination (a) employs effective methods and (b) makes possibly found "bad patterns" and "dangerous items" as visible, interactive, and actionable as possible. Evaluation methods must therefore be based on *visibility*, *interactivity*, and *actionability*. As in constraint-oriented DADM, system evaluation also has to integrate appropriate measures of usability.

[4]We use both *item* and *feature* to denote "an attribute with a certain value range"; where "item" is used in its technical DADM sense and "feature" for less technical remarks on discrimination.

DCUBE-GUI [18] is a DADM system that focusses on these issues. DCUBE-GUI employs methods from constraint-oriented DADM (more specifically, it builds on rules mined by DCUBE [12]) and complements them by risk scores defined on items or item pairs. It displays these results in interactive visualisations, thereby inviting users to engage in exploration and sense-making. One of DCUBE-GUI's item-oriented displays is shown in Fig. 1.

We extended DCUBE-GUI by a new score, established in exploratory statistics: the Pearson χ^2 value of the contingency table of (a) the scored item with (b) the target variable, measured within the population defined by the group with the discrimination-indexed feature. For example, a high value for the scored item "having children" within the discrimination-indexed feature "female" and target variable "chance of a job" means that mothers have a substantially higher / lower chance of not getting a job than other women. When the chance is higher (which is usually the case as a consequence of item selection), we call the scored item a *negative risk factor*, when it is lower, we call it a *positive risk factor*. The χ^2 values thus obtained [19] need to be adjusted for multiple comparisons, for example by Bonferoni corrections, in order to determine whether they are statistically significant. This does not affect the score order on risk factors; it may only filter out some of them. Since DCUBE-GUI focuses on displaying the largest risk factors and encourages ordinal rather than metrical comparisons between them, this produces diagrams that lend themselves to correct interpretations.[5]

However, the visibility and actionability of these patterns and items have not yet been measured or evaluated empirically. In data mining, such domain-oriented criteria are harder to assess than merely technical interactivity. Pertinent methodology comes from design studies and visual data mining [20]. We follow earlier work that proposes visualisation, interaction, and information as levels of analysis [21], but focus more strongly on actionability of the information.

Another challenge is that the classes of interested users and therefore the possible use cases are likely to multiply relative to constraint-oriented DADM. Not only decision makers, but also other stakeholders are likely to be interested in such analyses of discrimination and data mining. These may include sociologists, "applied sociologists" such as social workers, lawyers, and also the possibly concerned individuals themselves. Each of them may use DADM in different ways, and each of these uses may be associated with different knowledge and action goals.

[5]The χ^2 score is related to the odds ratio, cf. [7]. It is also related to [17]: it is possible that the latter learns a classifier "if female and with-children, then job=no" when the former learns that "with-children" is a significant negative risk factor within the "female" group. However, we differ from these approaches by focussing on items rather than rules.

III. User study

We conducted an exploratory user study to test whether the DCUBE-GUI interface can support non-expert users in exploring items associated with discrimination. To make the study more engaging and relevant, we embedded the interpretation of DADM results into a fictitious but realistic scenario, based on the following choices.

Users – social workers: In contrast to concerned individuals or lawyers, social workers are likely to be interested in a whole population rather than in the effects of the specific features that they themselves or their client have. Compared to (many) scientific sociologists, they are likely to favour actionability of the gained insights over the insights themselves.

Use case – Analysing visualisations that display relative discrimination scores of items: This was chosen as the simplest use case (for example, because it concentrates on items rather than on item pairs). To make the best use of the exploratory affordances of the tool, we conceived of a use case that lets the user explore risk factors.

Scenario: We asked people to imagine they were social workers giving advice to a client regarding risk factors for a loan. The idea was to have participants recognise the relative risk of different factors and to transform this into a recommendation to the client – to ask for a loan in a way that avoids the most important negative risk factors and, if applicable, take advantage of positive risk factors. Thus, our hypothesis was that the interface supports these steps (comparison of risk factors, identification of important ones, and translation into a correct and useful recommendation), i.e. that it makes the DADM results visible and actionable.

Questions: By postulating a scenario, we take a previous definition of top-level item as given (e.g., being female, being a foreign worker) and then investigate how visible problematic second-level items (such as being a young foreign worker) become and can lead to action (giving advice to a member of the social worker's community).

Measures: To limit the complexity of the study and confounding of factors, we restricted the interaction with the tool severely by giving participants screenshots rather than asking them to interact with the tool. This enabled us to focus on measures of *visibility* and *actionability*. In addition, we measured basic *usability* indicators.

A. Method

1) Participants: 20 US-based participants were recruited over Amazon Mechanical Turk (mTurk, mturk.com). They received USD 5.00 for full participation and up to USD 2.00 as an additional performance-dependent payoff (bonus). Sampling through mTurk has attracted some scrutiny with respect to self-selection recently, but it does appear to produce "reliable results consistent with standard decision-making biases" [22]. Through the design of our tasks and the restriction of participants' location (which in our case was done in order to reduce cultural confounds), we also heeded factors for quality control that have been observed to drastically reduce the occurrence of cheating on mTurk [23]. Later analyses of our results gave no indication of cheaters. Based on these findings, we considered recruitment through mTurk an adequate choice for our study.

Basic demographics were self-reported in an exit questionnaire: 14 (6) participants reported being female (male). 10 reported being between 18 and 25 years of age; the others were nearly evenly split between higher age groups: 25–30 (2), 30–40 (3), 40–50 (2), and 50–60 (3). Half reported a Bachelor's degree as their highest grade of schooling, 7 reported some college without a degree (5 of these were in the 18–25 age bracket), and 1 each reported being a High School graduate, having a Master's degree, or having a Professional degree. Only 1 participant, 18–25 and in the college/no degree (yet) group, reported that he "speak[s] a language other than English at home".

20% reported that they are "dealing with data mining or statistics in [their] job or have done so in the past", but 3 of these 4 were in the 18–25 age bracket with "some college without a degree". On the other hand, 40% reported that they are "dealing with financial information in [their] job (e.g., banking, insurance, finance industry) or have done so in the past", and 7 of these 8 were in higher age brackets. (Only one person, male, 18–25 and in the college/no degree (yet) group, answered "yes" to both questions.) This may be interpreted as a large number of participants having job experience with financial information, but only a small number having experience with data mining, and these mostly in education.

Three quarters of participants (15) stated that they had "applied for a loan at least once in [their] life", and 9 that they had "experienced discrimination in [their] own life".

2) Materials and procedure: As part of their task on mTurk, participants were given a series of scenarios with multiple answer options each. Participants ticked exactly one answer corresponding to what they considered the best response for each scenario. Three training tasks were presented first after an introductory page with the instructions. The correct answers were shown and explained on the following page, so that participants could check theirs. Five assessed tasks, without information on the correct answers, followed this page. All tasks featured a basic scenario, which was expanded upon as required: *"You are a social worker. In your neighborhood, it has been quite difficult recently for foreign workers to get a loan. The bank makes its decisions based on data-mining their existing credit records, and they publish their decision rules."* In addition, the screenshots were explained at the start: *"The tool FIGHT-DISCR visualizes the risk factors within population groups: The circles show the effects of properties of a loan applicant."* This was followed by the explanation that both circle size and colour encoded the relative *"statistical risk for someone in*

Figure 1. Assessed task 1 (screenshot)

the population group to not get a loan"[6] and that circle position was irrelevant.

The assessed tasks were more complex than the training task, featuring twice as many answers options as the training tasks (6 vs. 3) and up to twice as many potentially discriminating items (14 labelled items maximum). Figure 1 shows Task 1 as a prototypical task and at the same time the one that used the original data and their measures.

A two-phase exit questionnaire completed the study. First, we asked for impressions about the task, the tool, and their use. Ten statements were rated on a 7-point Likert scale anchored in "strongly agree" to "strongly disagree". As a simple reliability check, 8 of the 10 items came in pairs, with one reverse-coded. The statements build on standard usability questionnaires [25]. In the second questionnaire, participants were asked for some basic demographics (see "Participants" in Section III-A1 above).

[6]The study did not focus on any objective truth about such statistical risks. Yet, to make it as realistic and comparable to other studies in DADM as possible, we built on the German Credit Dataset [24], adjusted by oversampling the non-foreign workers, who are heavily under-represented in this dataset, 5 times. We determined critical items by mining for rules with DCUBE [12], fixing the item "foreign worker" as the PD item, and used default settings of rule interestingness measures. As explained in Section II-D, risk scores were the χ^2 values of items with the target outcome (a successful loan) within this group. The negative risk factors thus obtained were used to generate the display for Task 1; the other figures used in the study were permutations of these data and/or included also the largest positive risk factors.

Participants were also given the option to comment on the materials, explain their answers, or give any other kind of feedback, by the chance to fill in free-form text fields at the end of each Web page.

All multiple-choice questions (the advice given, the opinions, and the demographics) had to be filled in; all free-form answers were optional.

3) Design: The tasks were experimentally varied in a within-subjects design. The tasks roughly corresponded to different levels or characteristics of difficulty in the scenario's analysis-plus-recommendation combination. *Task 1* focused on the straightforward identification of the biggest negative risk factor and its negation to make it a recommendation. (For example, the risk factor "checking $\leq$ 0" had to be transformed into the recommendation to "put some money into one's checking account".) *Task 2* focused on the impossibility of changing some risk factors and thus the need to recommend changing the second-biggest. The scenario and the semantics of the biggest risk factor had to be understood to detect this impossibility. *Task 3* had a similar constraint on the biggest risk factor plus the possibility of recommending a decision that eliminated the second-biggest *and* the third-biggest risk factors together. Scenario and answer options had to be read carefully to make the right choice between different conjunctive options.

Task 4 introduced positive risk factors. This requires the decision maker to not negate the item, but use it as is, such as translating "loan duration $\leq$ 17" into a preference of a 10-months loan over a 20-months loan. Only one positive risk factor was shown, and its score was sufficiently big to make it stand out. The second-biggest addressable (negative) risk factor was a constraint on the same variable (loan duration). It was chosen such that (a) the correct answer option satisfied the positive risk and the eliminated the negative risk; (b) another answer option violated the positive risk, but eliminated the negative risk; and (c) a third answer option violated both. Only non-conjunctive answer options were given. *Task 5* contained two positive risk factors and distracting items with inequalities, some of which had to be negated. The correct solution was conjunctive and involved two variables.

B. Results and discussion

The analysed results were the participants' answers and times taken to solve the tasks. Correctness and speed were regarded as indicators of visibility and actionability.

No systematic relationships between demographics and any other results were observed.

1) Training tasks: The three training tasks were solved correctly by 13, 8 and 18 participants, respectively. This may be taken to indicate that even without prior training, the interface led a majority of participants to correct interpretations (training tasks 1 and 3), and that the integration of real-world knowledge was obvious to most (in task 2, the

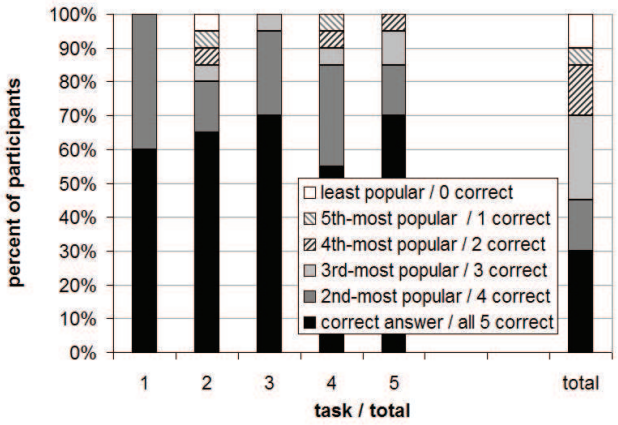

Figure 2. Proportion of participants by answer correctness/popularity

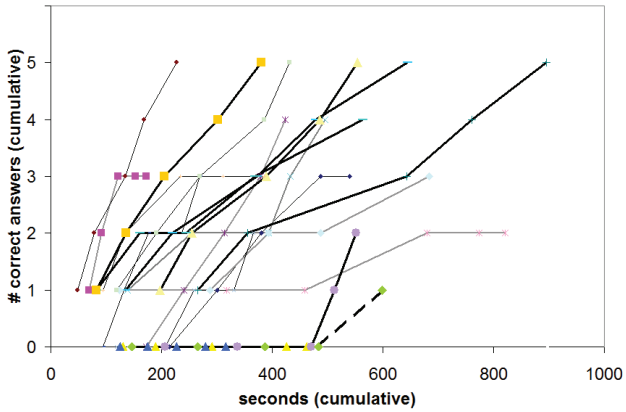

Figure 3. Answer times (x axis), correctness of answers (y axis), and commenting activity (line thickness and style).

wilful change of one's age in order to eliminate a risk factor was chosen only by 3). The preference for a wrong answer in training task 2 (9 participants) may indicate some problems with the semantics of the answer options. 4 participants solved all training tasks correctly, 11 solved 2, and 5 solved 1. Performance on the training tasks was not predictive of performance on the tasks (Spearman correlation $\rho = 0.24$).

2) Assessed tasks: Correctness on the tasks was high: Each task was answered correctly by a majority of participants (12, 13, 14, 11, and 14, respectively). Interestingly, for each task but #2 and #5, there was a clear second answer, chosen by 8, 3, 5, 6, and 3 participants, respectively. Figure 2 shows that for every question viewed individually, at least 55% of participants chose the correct options, and that this percentage was increasing over tasks. In each task, at least 80% chose the correct or a second-most popular option. On the right of the figure, we show the percentages of participants who answered all tasks correctly (30%), 4 out of 5 (another 15%), etc. 70% of all participants answered the majority of questions (at least 3) correctly.

In general, the performance of a given participant remained equal or increased over tasks. The largest drops in learning curves were observed for task 4, in which 25% of all participants gave an incorrect answer after a correct one in the previous task. In tasks 2, 3, and 5, similar performance drops were observed for 2, 2 and 1 users.

All questions were answered in average times of 2.5 minutes or less. The ranges and average values were, in seconds: 49–318 / 144.25 (1), 22–164 / 79.3 (2), 31–288 / 105.8 (3), 31–134 / 82.95 (4), 19–196 / 76.75 (5), 172–895 / 489.05 (total). The generally decreasing average may be interpreted as a learning effect with a specific challenge in task 3, probably caused by the need to first process conjunctive answer options. In terms of processing time, task 4 appears to have been perceived as simpler than the actual performance warrants.

Figure 3 shows the number of correct answers, answer times, and the number of comments. Each participant is plotted on a line starting at (x, y), where x is the time needed to answer question 1 and $y = 1$ if that question was answered correctly (0 else). The next point on the line shows the cumulative time and correctness for task 2, etc. It shows that the 10 people who gave no comments on any of the questions (thin black lines) needed less time and mostly had good performance. Two participants had 0 correct answers and did not comment, but their times appeared within the normal range (463/316 seconds in total). The figure also shows that the 7 participants who commented on every or nearly every question (thick black lines, dashed: 4 or solid: 5 comments) tended to need more time and mostly had good performance. However, there were also two participants who appeared to learn how to answer the tasks more slowly and only had correct answers on tasks 4 and/or 5; their comments indicated that they had been thinking in depth about the questions, but then relied on world knowledge rather than focussing on the statistical information alone.[7] The 3 participants who commented on 1 or 2 tasks (grey lines of intermediate thickness) scored in mid-range with respect to correctness and time needed. This integrated view of performance, times and comments did not show evidence of any participant cheating.

An analysis of the second-most popular answer options revealed that these were also plausible choices (task 1) or resulted from a failure to take into account real-life semantics (task 2), a failure to see that a conjunctive advice would eliminate two risk factors at once (tasks 3 and 5), and the misinterpretation of a positive risk factor as a negative risk factor (tasks 4 and 5). Thus, the difficulties introduced by the task design increased difficulty / decreased

[7]One example is the comment on why the applicant should get a telephone. The diagram and scenario showed that not having a phone was the biggest risk, and that it was remediable. Rather than relying on this information, this participant explained: "Loans require communication with the applicant. At least one phone contact is required."

	EASY+	EASY-	PERFORM+	PERFORM-	Q-EASY-	INFO+	INFO-	REUSE+	REUSE-	ENJOY+
#correct	-.26	.14	-.05	.18	-.15	-.26	.11	-.51	.46	-.13
EASY+		**-.85**	.61	-.65	-.56	**.80**	**-.80**	**.76**	-.63	**.74**
EASY-			-.62	**.70**	.66	-.89	**.77**	**-.70**	**.74**	**-.86**
PERFORM+				**-.75**	**-.78**	**.81**	**-.77**	.50	-.42	**.78**
PERFORM-					.68	**-.79**	**.81**	**-.71**	.49	**-.70**
Q-EASY-						-.64	.65	-.47	.26	**-.83**
INFO+							**-.82**	**.72**	**-.76**	**.80**
INFO-								-.64	.58	**-.74**
REUSE+									**-.71**	.58
REUSE-										-.49

Table II
PEARSON CORRELATIONS ρ BETWEEN OPINIONS. FOR EACH OPINION, + IDENTIFIES THE POSITIVE WORDING AND - THE NEGATIVE.

correctness as expected, although only for a minority of participants.

Taken together, all answers – including the wrong ones – indicate that people read the task instructions and inspected the figures, even if they misinterpreted them.

3) Opinions on the tool, the task and the participant's own performance: The participants uttered fairly good opinions on the tool and tasks, and with a small number of exceptions answered the question pairs consistently. All Pearson correlations between question pairs (positive/negative wording) lay between -0.7 and -0.85, and their standardised Cronbach's α [26] values were good ($\geq$ 0.8: PERFORM, REUSE) or excellent ($\geq$ 0.9: EASY, INFO). This supports our earlier interpretation that participants did not cheat.

15 people strongly agreed or agreed with the statement that they "enjoyed doing this task" (ENJOY). All the following statistics report the numbers of people who agreed or strongly agreed. 12 resp. 14 (positive/negative wording) found the interface easy to understand (EASY). 15 indicated that they believed to have understood the questions (Q-EASY). 10 resp. 11 (positive/negative wording) found the tool helpful for making sense of the information (INFO). 12 resp. 9 (positive/negative wording) would use the tool again for solving similar problems (REUSE). 11 resp. 9 (positive/negative wording) believed they answered the questions correctly (PERFORM).

Table II indicates that the judgement of these usability criteria did not stand in a systematic relationship to actual performance. However, other relationships emerged. The table shows correlations that were significant in bold font ($df = 18$; $\alpha = 0.05$ two-tailed and Bonferoni-corrected). All of them had the expected sign and a plausible interpretation, such as the positive association between perceived ease of understanding the interface and its perceived helpfulness in interpreting the information, or the associations between the intention to reuse the tool and its perceived good features as well as perceived performance.

4) Free-form comments: 14 of the 20 participants gave free-form comments; all of which showed that they had been thinking about the scenarios in depth – which gave us further indications regarding visibility and actionability. This was seen in particular in remarks about the second training task, in which the purpose to buy a car with the loan was targeted. Several participants commented on the answer option to *ask for a loan for a motorcycle instead of a car*, either by referring to the scenario, asking whether a delivery service can really be operated with a motorcycle, or by referring to ethical questions, writing that no advice should be given that would involve lying about the purpose of the loan application.

Many free-form comments indicated "visual thinking" (e.g., "best option to knock out two of the bubbles at once") or learning ("I just realized I got the previous question wrong, I believe. I think I mixed up the direction of the less-than-or-equal-to sign with respect to duration."). Five participants expressed their appreciation of the visualisation/tool and of the tasks, one asked for information on his task performance, and two gave implicit recommendations for improving the interface or study design (a simultaneous display of tasks and instructions, a comparison of different tools).

IV. CONCLUSIONS AND OUTLOOK

In this paper, we have argued for the need to supplement classical, constraint-oriented discrimination-aware data mining by more exploratory forms. As an example, we discussed our development of the DCUBE-GUI tool and presented an exploratory user study.

The results of our study suggest that DADM can be presented in ways that make it relevant and interesting to people, help them understand facets of discrimination and draw correct and actionable conclusions from DADM results. The results indicate that the DCUBE-GUI interface is useful for the task, engaging and well-liked. For most users, this form of displaying the results of discrimination-aware data mining for further exploration supported correct interpretations of the data and useful behaviour in a realistic scenario. This is evidence that DCUBE-GUI is effective in making the results of DADM visible and actionable.

The answers and the comments indicate that most participants took the task very seriously and thought about the scenarios. The results however also show that all but the simplest of result patterns (and derived action possibilities) do need more interface support than the mere display of the relative sizes of risk scores. In particular, positive vs. negative risk factors as well as conjunctions proved to be challenging. Tool instructions as well as scenario descriptions need to be as specific and unambiguous as possible. The results of our relatively small and deliberately formative study will be used for improving the tool and developing further user studies.

The answers and comments also indicate that many people prefer to think about an application scenario of data mining in a more holistic way than only in terms of numbers and risk scores. They take the life context of scenario personnel's age, family, or business into account, and they comment on the ethics of actors' behaviour in the scenario.

There are many aspects of DADM usage that we have not addressed in this study. These will be subjects of future work: (1) This first study only asked people for interpretations of result configurations that were by design quite clear-cut. Also, users were offered answer options rather than asked to produce answers. In many datasets, less clear-cut relations are likely to hold, and it remains to be seen how the interface choices may support or hinder correct interpretations in such cases. (2) Participants of our study studied tool output visualisations, but did not interact with the tool. It remains to be seen how a sequence of exploratory activities and the need to integrate their results will influence visibility and actionability. (3) The effects of different scenarios including different user roles and use cases will be explored. (4) Future studies should investigate the differences and complementarities between different tools for DADM and their effects on performance and satisfaction. These should compare information-equivalent presentations (e.g., scores in numerical form or as bar charts or circles) and complementary information. Such comparison studies may also help reduce possible social-desirability biases in the opinions given about the tool. As we have argued above, constraint-oriented DADM and exploratory DADM follow different goals and therefore need to target different evaluation criteria, but ultimately, only the effective cooperation between systems with these different roles and quality criteria will help DADM reach a wide audience and create a sizeable impact.

REFERENCES

[1] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13–22, 2000.

[2] J. B. Sykes, Ed., *The Concise Oxford Dictionary*, 7th ed. Oxford: Oxford University Press, 1982.

[3] A. Romei and S. Ruggieri, "Discrimination data analysis: A multi-disciplinary bibliography," in *Discrimination and Privacy in the Information Society*. Springer, 2013, pp. 109–135.

[4] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proc. KDD'08*, Y. Li, B. Liu, and S. Sarawagi, Eds. ACM, 2008, pp. 560–568.

[5] S. Hajian, J. Domingo-Ferrer, and A. Martínez-Ballesté, "Rule protection for indirect discrimination prevention in data mining," in *Proc. MDAI*, LNCS 6820. Springer, 2011, pp. 211–222.

[6] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *Proc. ICDM'10*, 2010, pp. 869–874.

[7] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring discrimination in socially-sensitive decision records," in *Proc. SDM*, 2009, pp. 581–592.

[8] S. Ruggieri, D. Pedreschi, and F. Turini, "Data mining for discrimination discovery," *TKDD: ACM Transactions on Knowledge Discovery*, vol. 4, no. 2, 2010.

[9] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification," *Data Min. Knowl. Discov.*, vol. 21, no. 2, pp. 277–292, 2010.

[10] S. Hajian, J. Domingo-Ferrer, and A. Martínez-Ballesté, "Discrimination prevention in data mining for intrusion and crime detection," in *Proc. IEEE SSCI 2011*, 2011.

[11] D. Pedreschi, S. Ruggieri, and F. Turini, "A study of top-k measures for discrimination discovery," in *Proc. 27th SAC*. New York, NY, USA: ACM, 2012, pp. 126–131.

[12] S. Ruggieri, D. Pedreschi, and F. Turini, "DCUBE: discrimination discovery in databases," in *Proc. SIGMOD'10*, 2010, pp. 1127–1130.

[13] D. Pedreschi, S. Ruggieri, and F. Turini, "Integrating induction and deduction for finding evidence of discrimination," in *Proc. ICAIL*. ACM, 2009, pp. 157–166.

[14] B. Berendt, S. Preibusch, and M. Teltzrow, "A privacy-protecting business-analytics service for online transactions," *International Journal of Electronic Commerce*, vol. 12, pp. 115–150, 2008.

[15] C. Fine, *Delusions of Gender. The Real Science Behind Sex Differences*. London: Icon Books, 2010.

[16] B. L. Thanh, "Generalized discrimination discovery on semi-structured data supported by ontology," Ph.D. dissertation, IMT Institute for Advanced Studies, Lucca, Italy, 2011.

[17] B. L. Thanh, S. Ruggieri, and F. Turini, "k-NN as an implementation of situation testing for discrimination discovery and prevention," in *Proc. KDD*, ACM, 2011, pp. 502–510.

[18] B. Gao and B. Berendt, "Visual data mining for higher-level patterns: Discrimination-aware data mining and beyond," in *Proc. 20^{th} Benelearn*, 2011, http://www.benelearn2011.org/.

[19] J. Bresnahan and M. Shapiro, "A general equation and technique for the exact partitioning of chi-square contingency tables," *Psychological Bulletin*, vol. 66, pp. 252–262, 1966.

[20] M. Sedlmair, M. Meyer, and T. Munzner, "Design Study Methodology: Reflections from the Trenches and the Stacks," *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis)*, 2012.

[21] D. Marghescu, M. Rajanen, and B. Back, "Evaluating the quality of use of visual data-mining tools," in *Proc. 11th Europ. Conf. IT Evaluation*, Academic Conferences Limited, 2004, pp. 239–250.

[22] J. Goodman, C. Cryder, and A. Cheema, "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples," *Journal of Behavioral Decision Making*, 2012, DOI 10.1002/bdm.1753.

[23] C. Eickhoff and A. P. de Vries, "Increasing cheat robustness of crowdsourcing tasks," *Information Retrieval*, 2012, DOI 10.1007/s10791-011-9181-9.

[24] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, "UCI repository of Machine Learning databases," 1998, GCD at http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29.

[25] J. R. Lewis, "IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use," *International Journal of Human-Computer Interaction*, vol. 7, no. 1, pp. 57–78, 1995, see http://hcibib.org/perlman/question.cgi [2012-07-31].

[26] C. Hulin, "Measurement. III.A. Cronbach's alpha on two-item scales," *Journal of Consumer Psychology*, vol. 10, no. 1&2, pp. 55–69, 2001.