

Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence

Bettina Berendt · Sören Preibusch

Published online: 10 January 2014
© Springer Science+Business Media Dordrecht 2014

Abstract Decision makers in banking, insurance or employment mitigate many of their risks by telling “good” individuals and “bad” individuals apart. Laws codify societal understandings of which factors are legitimate grounds for differential treatment (and when and in which contexts)—or are considered unfair discrimination, including gender, ethnicity or age. Discrimination-aware data mining (DADM) implements the hope that information technology supporting the decision process can also keep it free from unjust grounds. However, constraining data mining to exclude a fixed enumeration of potentially discriminatory features is insufficient. We argue for complementing it with exploratory DADM, where discriminatory patterns are discovered and flagged rather than suppressed. This article discusses the relative merits of constraint-oriented and exploratory DADM from a conceptual viewpoint. In addition, we consider the case of loan applications to empirically assess the fitness of both discrimination-aware data mining approaches for two of their typical usage scenarios: prevention and detection. Using Mechanical Turk, 215 US-based participants were randomly placed in the roles of a bank clerk (discrimination prevention) or a citizen / policy advisor (detection). They were tasked to recommend or predict the approval or denial of a loan, across three experimental conditions: discrimination-unaware data mining, exploratory, and constraint-oriented DADM (eDADM resp. cDADM). The discrimination-aware tool support in the eDADM and cDADM treatments led to significantly higher proportions of correct decisions, which were also motivated more accurately. There is significant evidence that the relative advantage of discrimination-aware techniques depends on their intended usage. For users focussed on making and motivating their

B. Berendt (✉)
Department of Computer Science, KU Leuven, Leuven, Belgium
e-mail: bettina.berendt@cs.kuleuven.be

S. Preibusch
Microsoft Research, Cambridge, UK
e-mail: spr@microsoft.com

decisions in non-discriminatory ways, cDADM resulted in more accurate and less discriminatory results than eDADM. For users focussed on monitoring for preventing discriminatory decisions and motivating these conclusions, eDADM yielded more accurate results than cDADM.

Keywords Discrimination discovery and prevention · Data mining for decision support · Discrimination-aware data mining · Responsible data mining · Evaluation · User studies · Online experiment · Mechanical Turk

1 Introduction

In our computer-mediated lives, data supports decisions and carries value that promises unprecedented levels of convenience. The insights that can be inferred from large datasets are however not immediately accessible. They require processes of “knowledge discovery” (Shearer 2000). Knowledge discovery comprises the statistical analysis of data with the help of data mining methods. It also encompasses pre-processing and deployment, as well as the human expertise driving these sub-processes, as integral parts. Many Web users have already profited from data mining in recommender systems, which support their consumption choices or search queries. But data mining is also used when designing HIV vaccines (Heckerman 2013) or with the aim of keeping cities safe (Microsoft 2012). In e-Commerce, banking, insurance, or employment, data mining is often used to segregate “good” from “bad” individuals (Boston Consulting 2012; Duhigg 2009). Besides promising economic advantages, this raises questions of discrimination, not only within the organisations deploying data mining tools, but also among supervisory authorities and social activists.

Differentiation—making a distinction based on some features or attributes—is a fundamental characteristic of human cognition and behaviour. People apply differential treatment to other people, allowing some but not all to vote, applying certain laws to them, giving them jobs, and granting them loans—or denying them the privileges associated with these rights and decisions. Part of the social contract of any society is that certain attributes are accepted for differentiation, while others are not. Non-accepted attributes are those that violate the legal principle of equality, which has found its expression in fundamental and wide-reaching legal codifications such as Article 7 of the Universal Declaration of Human Rights. This article states that “All are equal before the law and are entitled without any discrimination to equal protection of the law.” The term ‘discrimination’ denotes a differentiation on non-accepted grounds. To avoid it, one must treat equal things equally and unequal things unequally. In many countries, individuals are protected by a range of laws against discrimination by the state and also by private actors such as employers. Along with societal notions of what constitutes wanted and unwanted differentiations, the legal demarcations between accepted and non-accepted grounds develop over time, as do the legal groupings of what is equal and what is not.

Many instruments have been proposed for fighting discrimination once it has been outlawed, but discrimination proves to be tenacious. Currently, much hope rests with information technology on which decisions increasingly rely. An appropriately modified algorithm should help to avoid discrimination. In the insurance industry, for instance, data analysis may generate gender-blind tariffs to comply with the new European Union's requirement of unisex policies.

The general *research question* we address in the present article is how to best support the monitoring, understanding, and avoidance of discrimination with the help of information technology. Specifically, we investigate how data mining can act as an instrument against discrimination. We investigate when it is better to hide discriminatory features, and when it is better to reveal and draw attention to them. We also derive recommendations for algorithm and interface design, and discuss the potentials and limitations with regard to further goals such as transparency.

Whether deliberately or unwittingly, discrimination originates in human decisions, which may be tool-supported. Our investigation therefore targets the interface between technology and its human users. We use an *empirical methodology* to quantitatively assess the ability of data mining and the tools displaying its results, to prevent discrimination in decision making. Indeed, deployment and result communication are integral parts of a data mining and knowledge discovery system. We conducted a user study where participants were equipped with data mining solutions to help them make or monitor decisions which could be discriminatory.

Our *contribution* is twofold. First, we critically discuss the emerging area of discrimination-aware data mining (DADM). We argue why the standard approach to DADM is useful and necessary, but also why it falls short of the full technical potential of data mining and also performs sub-par in fighting discrimination. We propose and evaluate a complementary form of DADM, which we call exploratory. Exploratory DADM focusses on revealing and drawing attention to discrimination in data, as opposed to traditional DADM that aims at "hiding" it. We argue that an exploratory approach is needed to find new and unexpected features and patterns of discrimination and is therefore a required complement for effectively avoiding discrimination. As our second contribution, we present empirical evidence to answer the research questions. Using a large-scale experimental user study, we uncover the relative advantages of both forms of DADM in the settings of a bank and an anti-discrimination agency. These correspond to the archetypical applications of data mining in decision support: making and monitoring decisions. To the best of our knowledge, this study represents the first user-centric evaluation of DADM described in the scientific literature; it extends on our previous small-scale exploratory study, which we briefly summarise in this paper.

The remainder of this article is structured as follows: In Sect. 2, we give an overview of related work. In particular, we propose the new classification of DADM approaches and give a brief survey of the literature structured by this framework. In Sect. 3, we discuss appropriate use cases and derive recommendations for DADM evaluation foci. We summarise an exploratory user study ($n = 20$) in which we demonstrated the effectiveness of exploratory DADM in detecting actionable patterns of differentiation and discrimination. Section 4 reports on a new, large-

scale multi-treatment user study ($n = 215$) in which we focussed on the relative advantages of the two forms of DADM in different settings. We conclude with an outlook on future work in Sect. 5.

2 Constraint-orientation versus exploration: a new framework for related work in DADM

To understand the range of DADM, we need to take a step back and ask about the fundamental relations between data mining (discrimination-aware or not) and discrimination (Sect. 2.1). From this, we derive our notion of *constraint-oriented DADM* as a description of most of the current work in the field (Sect. 2.2). While this is a very important approach, it needs to be complemented by *exploratory DADM* (Sect. 2.3).¹

2.1 Data mining and discrimination

We understand *data mining* in the more general sense of “knowledge discovery” (Shearer 2000) and therefore consider pre-processing 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” (Sykes 1982). Such “unjust” grounds are legally codified in many countries and may include further characteristics. In the following, we will call them *discrimination-indexed attributes/features*.² A comprehensive multi-disciplinary overview of discrimination research is provided in Romei and Ruggieri (2014).

Discrimination in the narrow sense may be understood as occurring if and only if one differentiates by such grounds. While *discrimination in the legal sense* often consists of a differentiation in this sense, this is not always the case. It is impossible, within the scope of this article, to describe this notion (in fact, class of notions) exhaustively. Instead, we will highlight important divergences between discrimination in the narrow sense and discrimination in the legal sense, using as an example European (EU) law on discrimination by gender. Where applicable, we will focus on the European “Gender Directive” 2004/113/EC (EU 2004) because its

¹ Sections 2 and 3.1–3.3 extend on a previous workshop paper (Berendt and Preibusch 2012), and Sect. 3.4 summarises the user study presented in detail in that paper.

² Otherwise called, e.g., “potentially discriminatory (PD) items” (Pedreschi et al. 2008) or “sensitive attributes” (Hajian and Domingo-Ferrer 2013; Kamiran et al. 2010). A *feature* or *item* is an *attribute* with a value or value range; thus for example “gender” is an attribute and “female” a feature. All three terms refer to the formal representation of *legal grounds* of discrimination (the reasons specified by the law that will serve as a basis for demanding relief) and other grounds in the databases used for data mining. While Pedreschi et al. (2008) point out that PD items may comprise more than just legally-defined sensitive attributes, they still assume a priori knowledge about these items.

application area is closest to the example setting chosen in the experiment described in Sect. 4 below.

- Whether a given differentiation in treatment amounts to discrimination may depend on the agent performing it. States are mainly bound by Article 14 of the European Convention on Human Rights and Articles 18 and 19 of the Treaty on the Functioning of the European Union, private parties in their role as suppliers of goods and services by the national implementations of the “Gender Directive” 2004/113/EC (EU 2004), and private parties in their role as employers by the national implementations of the Equal Treatment Directive 2006/54/EC (EU 2006).
- A differentiation in treatment may amount to discrimination when it is based directly on the discrimination-indexed feature (so-called “direct discrimination”), but discrimination can also result from decisions based on other, seemingly neutral features highly correlated with the discrimination-indexed features (so-called “indirect discrimination”), e.g. EU (2004, Article 2(b)).
- A differentiation in treatment is not discrimination when the situations are not comparable (EU 2004, Recital (12)). In fact, in such a situation non-differentiation may be discrimination. An example are maternity protection measures that must discriminate between women and men because only women can give birth or breastfeed. Examples include EU (2004, Recital (24)) and EU (2006, Article 15).
- A differentiation in treatment is not discrimination when it is justified by a legitimate aim and the means of achieving that aim are appropriate and necessary (“proportional”) (EU 2004, Article 4 (5)). Examples are single-sex sports clubs or shelters for abused women. In specific employment situations, a discrimination-indexed feature may actually be a “genuine occupational requirement”. For example, it is legitimate to consider only male applicants when searching for models for men’s fashion.

These rules, and therefore also the definitions of which situations are comparable and which are not, and which aims are legitimate and which are not, may change over time. For example, men and women may be argued to be in non-comparable situations when it comes to statistical life expectancy or risk of illness and accidents. Until 2012, Article 5(2) of EU (2004) allowed Member States to “permit proportionate differences in individuals’ premiums and benefits [from insurance and related financial services] where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data”. On 1st March 2011, the European Court of Justice ruled that Article 5(2) was in breach of the Charter of Fundamental Rights and therefore void, after a transition period lasting until 21st December 2012 (European Court of Justice 2011).

In the employment sector, the legally admissible exclusions of women from certain professions, especially in the police and armed forces, are gradually eroding along with the assumptions that women are “by nature” not suited to them (Pitt 2009). Moreover, the legal provisions of what constitutes illegal discrimination may be quite heterogeneous even across jurisdictions governed by the same principles [concerning insurance, see Schanze (2013) for an overview of pre-2012 European

implementations and Avraham et al. (2013) for an overview of US states' legislations].

Three further aspects are needed to distinguish between the notions of discrimination and related concepts. First, discrimination in a wide sense can involve a merely cognitive making of a distinction, or a making of a distinction in treating people, or a making of a distinction in treating other creatures or things. Discrimination in the narrow and in the legal sense focus on differentiations in treating people. Second, a statistical imbalance in itself is not discrimination—discrimination is a property of a decision or decisions, which may result in statistical imbalances as well as the situation of individuals. As an example, more men than women having jobs in higher management is a statistical imbalance, although it may well be the result of discriminatory decisions. On the other hand, a woman not getting a job just because of her gender is discrimination. Third, discrimination can happen intentionally or unintentionally.

2.2 Classical discrimination-aware data mining (DADM)

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. Establishing the causal reasons of these imbalances of course requires going beyond the mere statistics of data mining. DADM methods are extensions of standard data mining that leverage background knowledge about discrimination-indexed features and their correlation with other features in order to detect discrimination in the narrow sense.

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 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.³ 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. Examples include Hajian and Domingo-Ferrer (2013), Mancuhan and Clifton (2012) (pre-processing), Calders and Verwer (2010), Kamiran et al. (2010, 2012), Kamishima et al. (2012) (in-processing), and Calders and Verwer (2010), Pedreschi et al. (2009), Ruggieri et al. (2010) (post-processing).

As an example, we consider a typical use of data mining: the analysis of 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

³ “Bad patterns” correspond to, e.g., “ α -discriminatory rules” in Pedreschi et al. (2008).

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. We have used this example of loan decisions as the basis for the user studies described in this paper (see Sects. 3.4, 4).

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.⁴ We therefore call this classical approach to DADM *constraint-oriented*.

Further constraints are imposed on this form of DADM in order to also prevent indirect discrimination such as red-lining. DADM approaches such as those of Calders and Verwer (2010), Hajian and Domingo-Ferrer (2013), and Ruggieri et al. (2010) formalise and take measures against such indirect discrimination.

2.3 The need for exploratory DADM

The constraint-oriented approach to DADM, 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. Overt discrimination against women appears to have abated relative to the past, thanks in no small measure to past efforts to detect gender discrimination, raise awareness about it, and implement equal-opportunities policies. However, it increasingly appears that *mothers* now suffer from discrimination in the workplace (Fine 2010). 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 “parenthood” is hardly predictive of gender. Such forms of discrimination can only become successful targets for classical DADM if the risks implied by “parenthood” *within* the group with feature “female” have been discovered and a new feature “mother” has been constructed. Note that such feature construction often requires background knowledge and negotiation among stakeholders. For instance, the risks implied by “lack of job experience” (another not discrimination-indexed feature) may be statistically equal to those of parenthood, but are unlikely to be accepted as unjust job-market discrimination. We call such an approach, which focusses on *discovering* features and discrimination, *exploratory DADM*.

An exploratory approach to DADM is also advantageous when it is not clear-cut whether a distinction by some attribute amounts to discrimination in the legal sense or not. Making a feature visible may allow for more open-ended interpretations and evaluations and, importantly, for an awareness of the complexity of the notion of discrimination as such. Constraint-oriented DADM requires a model in which the

⁴ See for example Hajian et al. (2011), Kamiran et al. (2010) for measures of utility.

Table 1 Data mining (*DM*), discrimination, and foci of constraint-oriented DADM (*cDADM*) and exploratory DADM (*eDADM*)

	Discrimination (wide sense)	Discrimination (narrow sense, legal sense)
<i>Descriptive DM</i>	Detection	
cDADM		Assumption-based detection
eDADM		Discovery-based detection
Not DADM-supported DM		Detection is possible
<i>Prescriptive DM</i>	Creation	
cDADM		Prevention of creation
eDADM		Feature evaluation/construction
Not DADM-supported DM		Creation is possible

distinction between discrimination and non-discrimination relies on explicit and binary distinctions between legitimate and non-legitimate attributes. However, this may not always be straightforward. First, the temporal and spatial heterogeneity of anti-discrimination legislation needs to be taken into account when, for example, a DADM software is rolled out in a large multinational company. In addition, the modelling of non-comparable situations may require measures that relate to populations⁵ or aims⁶ as well as their restrictions by legal principles⁷. The visibility of the features may remind the analyst that additional judgment must be applied before a rule is simply discarded as “illegitimate”.

The resulting relationships between data mining and discrimination, as described in Sects. 2.1–2.3, are summarised in Table 1. At this high level of abstraction, data mining has similar relationships to discrimination in the narrow and in the legal senses, even if there will be important differences in practice. We will return to this in the Conclusions.

DCUBE-GUI (Gao and Berendt 2011) is a DADM system that encompasses several of these roles of data mining for discrimination detection and prevention. DCUBE-GUI employs methods from constraint-oriented DADM (more specifically, it builds on rules mined by DCUBE (Ruggieri et al. 2010)) and complements them by risk scores defined on items or item pairs. The analysis of items addresses a descriptive question (people with what features were possibly discriminated against, or simply appear to be at more risk of bad outcomes) as

⁵ E.g. the “actuarial factors related to sex” discussed in Sect. 2.1.

⁶ E.g. “Differences in treatment may be accepted only if they are justified by a legitimate aim. A legitimate aim may, for example, be the protection of victims of sex-related violence (in cases such as the establishment of single-sex shelters), reasons of privacy and decency (in cases such as the provision of accommodation by a person in a part of that person’s home), the promotion of gender equality or of the interests of men or women (for example single-sex voluntary bodies), the freedom of association (in cases of membership of single-sex private clubs), and the organisation of sporting activities (for example single-sex sports events).” (EU 2004, Recital (16)).

⁷ E.g. “Any limitation should nevertheless be appropriate and necessary in accordance with the criteria derived from case law of the Court of Justice of the European Communities.” (EU 2004, Recital (16))

well as a prescriptive question (which of these features will be applied in decision rules to the detriment of people). The methods for classifier learning from paired instances and for the use of ontologies proposed by Luong (2011) and Luong et al. (2011) open opportunities for such exploration. DCUBE-GUI displays these results in interactive visualisations, thereby inviting users to engage in exploration and sense-making.

3 Use cases and evaluation criteria for DADM

In this section, we investigate how DADM is evaluated today (with a focus on automated evaluations, see Sect. 3.1) and how the requirements for evaluation change when DADM is seen in the larger context of knowledge discovery and in particular as part of decision support. After a general discussion of key issues (Sect. 3.2), we derive conclusions for evaluations of cDADM and eDADM (Sect. 3.3). In Sect. 3.4, we then summarise a first exploratory user study of eDADM and its limitations as a motivation for the experiment to be presented in the subsequent section.

3.1 Automated evaluations and evaluation criteria of DADM

The evaluation of DADM has so far concentrated on the automated analysis of the patterns obtained by the modified algorithms. These evaluations have a simple success criterion: Ideally, all “bad patterns” disappear. 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 employs effective methods and disables possibly found bad patterns.

The resulting success measures of non-existence include counts of successfully sanitised bad patterns, as well as numbers of missed rules and of newly emerging “ghost rules” found in the transformed dataset but not in the original one (Hajian and Domingo-Ferrer 2013). Success can also be measured by reduced discrimination scores (Kamiran et al. 2010). An overview of metrics is given in Hajian and Domingo-Ferrer (2013). Note that agreed-upon definitions of “bad patterns” are still being developed, cf. Pedreschi et al. (2012), Ruggieri et al. (2010). DCUBE (Ruggieri et al. 2010) and LP2DD (Pedreschi et al. 2009) 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 the one proposed by Berendt et al. (2008).⁸

⁸ We claim this analogy due to the focus on hiding and sanitising patterns that privacy-preserving and discrimination-aware data mining share. However, using one does not imply the other, and their relation is in general non-trivial (Hajian 2013; Hajian et al. 2012).

These success measures abstract from the concrete use cases of DADM for decision support, but the literature does suggest measures of success in such deployment scenarios, to which we turn next.

3.2 Considerations for the evaluation of data mining for decision support

Viewed simply, a decision-support system is “good” to the extent that it supports “the right” decisions. However, this concept is too vague and maybe not even definable in general. We therefore consider a number of general considerations for evaluating decision-support systems and interactive data mining and then derive specific lessons for DADM from them.

Pertinent methodology comes from design studies and visual data mining (Sedlmair et al. 2012). We follow earlier work that proposes visualisation, interaction, and information as levels of analysis (Marghescu et al. 2004), but focus more strongly on actionability of the information. Actionability is a key concept in the traditional definition of data mining: “Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”, where “useful” or “actionable” means that patterns should potentially lead to some useful action (Fayyad et al. 1996). We will therefore consider measures of the visibility and saliency (through visualisation) of discrimination-related information and measures of the actionability of patterns for application-related decisions.

It is important that evaluations take real decision-making situations into account as well as possible (Perer and Shneiderman 2009; Plaisant 2004), although the difficulties of acquiring actual decision makers and following them in their actual, often long-term professional routines are well-known. The evaluation practice in specific domains such as medical decision support therefore suggests that laboratory studies are useful and necessary as a first step on the way to evaluation in more naturalistic settings (Kaplan 2001). For these reasons, we will investigate in which real decision-making situations various forms of DADM might be useful, for whom and how. We have conducted controlled user studies with non-expert users and placed them in situations requiring decisions.

Finally, when humans decide with decision support from a machine, they often do this under conditions of uncertainty. Even with the help of data and statistics, complete information and full “rationality” cannot be achieved, and they may also not be desired. Rather, humans typically employ a number of *heuristics*, which have been found to lead to typical decision *biases* (Arnott 2006). The design of interactive decision-support systems can address well-known heuristics and biases (Chen and Lee 2003).

A particularly pervasive heuristic is that of *availability*: an outcome will be considered more likely to happen the easier it is to think of it or its examples. Design guidelines for decision-support systems have emphasized the need to address this, usually by making *more* information available through presentation in the digital system. Translated into our setting, we expect an availability heuristic of the following kind: a factor (e.g., a piece of discriminatory information) will be considered more important in a decision situation the easier it is to think of it.

DADM (and related fields such as privacy-preserving data mining) have, interestingly, led to a situation in which two completely different approaches to availability are being proposed: cDADM focusses on making bad patterns *less* available or completely unavailable, whereas eDADM focusses on making them *more* available (or available at all) through various forms of highlighting. In the following, we will explore these two approaches to availability as design choices and in their role of co-determining evaluation choices. We will also ask to what extent the cDADM approach of making discrimination less visible by “hiding” it will indeed make it less cognitively available.

3.3 Use cases and evaluations of DADM decision support

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

Typical use cases of such systems will involve decision makers as users. An example are 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.

The automated-evaluation criteria of non-existence can be directly translated into measures of *invisibility* of bad patterns in decision-making situations. However, one also needs to ask whether this system-given invisibility still creates *actionable* patterns and leads to the correct or desired human decisions. Thus, *decision quality* should be measured as part of actionability. Of course, evaluation also has to integrate appropriate measures of usability.

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.

Typical use cases of such systems will involve actors and users who focus on monitoring other decision makers. Examples are societal organisations such as anti-discrimination centres and commissions, or enforcement authorities. Others could be individuals potentially affected by discrimination or their representatives such as lawyers or social workers, judges having to rule on discrimination-related complaints, and last but not least researchers and activists interested in discovering and investigating patterns of discrimination.

An effective data-mining method for preventing discrimination in eDADM applies an agreed-upon definition of bad patterns and guarantees that it finds (or highlights) them. An effective system architecture for preventing discrimination employs effective methods and makes “bad patterns” visible, interactive, and actionable. Evaluation methods must therefore be based on *visibility*, *interactivity*, and *actionability*. Again, *decision quality* should be measured as part of actionability. As in constraint-oriented DADM, system evaluation also has to integrate appropriate measures of usability.

3.4 Can eDADM support non-expert users in exploring items associated with discrimination? A first, exploratory user study

We conducted an exploratory user study to test whether the DCUBE-GUI (Gao and Berendt 2011) 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 setting. 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.

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 (e.g. being a young foreign worker) become and can lead to action (giving advice to a member of the social worker's community). 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.

In a series of nine scenarios describing the features of a loan applicant and his or her loan request, participants chose a “best recommendation” for the client. The results showed that the highlighting of the relative risk factors by the eDADM tool DCUBE-GUI enabled participants to readily identify negative and positive risk factors and from them to correctly identify recommendations—a sign of high decision quality.

In addition, the answers and the comments indicated that most participants took the task very seriously and thought about the scenarios. The answers and comments also indicated 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 took the life context of scenario personnel's age, family, or business into account, and they commented on the ethics of actors' behaviours in the scenario.

The results show that DCUBE-GUI is effective in making the results of DADM visible and actionable. 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.

This exploratory study also presented evidence that eDADM is suitable for detecting discrimination, including new forms of it. Still, there are four aspects of DADM usage that were not addressed. (1) This first study only asked people for interpretations of result configurations that were by design quite clear-cut. Also, users were offered decision options, but not asked to motivate their decisions. (2) The study only used one tool. This restricts the interpretation of its results to an evaluation of the effectiveness of eDADM. As a first extension, eDADM and

cDADM should be compared using decision-support interfaces that are as similar and information-equivalent as possible. (3) The study considered only one user role and use case: a social worker whose task is to detect and advise potentially concerned individuals in the face of given discrimination. This spectrum needs to be extended by the users and use cases we have described as characteristic of DADM's role in preventing and monitoring discrimination. (4) The first study was deliberately exploratory and employed only a small sample.

4 How do eDADM and cDADM support decision-making and reasoning in different settings? A large-scale experimental user study

To address the open questions after the first exploratory user study, we conducted a larger study. In this section, we first specify our hypotheses (Sect. 4.1), then give a non-technical overview of the study's method (Sect. 4.2), followed by a detailed description (Sects. 4.3, 4.4). We describe and interpret the results in Sect. 4.5. A discussion of its limitations will be the subject of the general conclusions of this paper.

4.1 Hypotheses

The purpose of the study was to further investigate the role of DADM for the detection and prevention of discrimination. In particular, we were interested in the relative value of eDADM and cDADM for decision quality in different typical settings. These settings are characterised by different foci on discrimination detection and (non-)creation as outlined above. We also wanted to investigate not only the decisions being made, but also the reasoning towards them. This led to the following hypotheses.

The first two hypotheses concern the role of DADM, exploratory and constraint-based, in supporting and motivating decisions.

H1: DADM supports users in making non-discriminatory decisions based on data-mining results, with more accurate results than not DADM-supported data mining.

H2: DADM supports users in motivating their conclusions in non-discriminatory ways with more accurate results than not DADM-supported data mining.

The third and fourth hypotheses concern the differential advantages of cDADM and eDADM for different settings.

H3: For users focussed on making and motivating their decisions in non-discriminatory ways, cDADM supports more accurate and less discriminatory results than eDADM.

H4: For users focussed on monitoring for preventing discriminatory decisions and motivating these conclusions, eDADM supports more accurate results than cDADM.

4.2 Study overview

We created experimental conditions that differed along the dimensions “mining form” and “setting”. As mining forms, we chose cDADM, eDADM and, as control

Bank: You work in a bank, and your responsibility is to prepare loan decisions for your manager: Based on an applicant's data, you propose to either grant or deny a specific loan request. The bank's policy is to draw on data analyses of past loan data.

ADA: You work in a citizen-advice / company-watch center, and your responsibility is to prepare decisions for your manager: Based on an applicant's data, you predict likely outcomes. Your manager will use your predictions to derive "alerts" as to which cases to follow up. Various citizens have turned to the center for help: they want a loan from the bank, but are not sure whether they will get it. The center has access to data analyses of past loan data, and it uses these analyses both to advise individuals and to monitor patterns of lending.

Bank, ADA: Thus, you will receive (a) data describing the requested loan and the applicant, and (b) statistical rules that argue for or against granting the loan, given specific data from (a). Based on the information from (a) and the decision support from (b), you will 1. propose a decision for your manager and 2. motivate that decision.

Fig. 1 The overall task descriptions for the bank conditions (*top* and *bottom*) and for the ADA conditions (*middle* and *bottom*)

conditions, non-DADM data mining (DM for short). As settings, we chose a bank and an anti-discrimination agency (ADA), both focussing on the granting of loans. These correspond to the archetypical applications of data mining in decision support: making and monitoring decisions. This results in 3 (mining forms) \times 2 (settings), i.e. 6 experimental conditions. The settings were introduced to participants via instructions about how to use data-mining results for reaching decisions (see Fig. 1) and instructions to avoid discrimination in the process (see Fig. 2).

The 215 participants of our user study, randomly and approximately equally distributed over the 6 conditions, were then asked to consider a series of loan requests. They were given features of the request and the applicant, and provided with decision-supporting rules of a data-mining tool that was fictitious but based on the principles of the mining form. Bank participants were asked to decide whether to grant the loan or not, and to motivate their decision. ADA participants were asked to conclude whether they considered it likely that the loan would be granted or not, and to motivate their conclusion. Examples of the tool and answer choices are shown in Figs. 3 and 4.

We then analysed the decisions as well as the motivations. The results were analysed with a view to testing the hypotheses H1–H4. In addition, these answers and free-form comments were analysed in an exploratory fashion for further insights into how cDADM and eDADM could help against discrimination, and where potential pitfalls lie.

4.3 Method: notes on operationalisation and terminology

We applied some simplifications when operationalizing the constructs in order to (a) test the formalisations of discrimination employed in today's DADM, (b) maximise experimental control, (c) make the tasks feasible for participants, and (d) obtain a first baseline of results.

Note that banks are not allowed to discriminate against applicants based on gender, marital status, nationality, or age. It is important for { **Bank:** the bank | **ADA:** the center to monitor } that decisions that discriminate based on these features not be taken – even if data from the past suggest it.

If any of your answers (the { **Bank:** recommendation to grant or deny the loan | **ADA:** conclusion as to the likelihood of the loan being granted or denied }, or your motivations) need explanation with regard to possible discrimination, please note this in the free-form answer box.

Fig. 2 Instructions against discrimination in the bank resp. ADA conditions

Dabiku is a Kenyan national. She is single and has no children. She has been employed as a manager for the past 10 years. She now asks for a loan of \$10,000 for 24 months to set up her own business. She has \$100 in her checking account and no other debts. There have been some delays in paying back past loans.

Fig. 3 Example vignette describing the loan request, used in all conditions. Another example is shown in Fig. 4

First, we applied a simplified definition of the “discrimination” we asked participants to avoid: we restricted the specified attributes to four (gender, marital status, nationality, and age), and we declared any discrimination by these attributes as illegitimate, without exceptions. This was done in order to give our non-expert participants a task of manageable difficulty and a clear-cut instruction (“do not discriminate based on these attributes”). The four specific attributes were chosen (a) as typical discrimination-indexed attributes in many jurisdictions (including the European provisions described above and the US-American Equal Credit Opportunity Act ECOA, which applies to most of our participants) and (b) as compatible with a dataset commonly used in DADM (see Sect. 4.4.4). Like its European counterparts, the ECOA previews exceptions to an absolute prohibition to discriminate based on the listed grounds, and a valid identification of whether some decision is legally discriminatory will generally need to involve a legal expert. To avoid this, we gave the simplified instruction.

Second, we wanted to avoid obtaining results confounded by the choice of any specific data mining algorithm. We therefore decided to implement only the key difference between cDADM and eDADM: whether to hide/remove or to highlight discrimination-indexed features in rules.

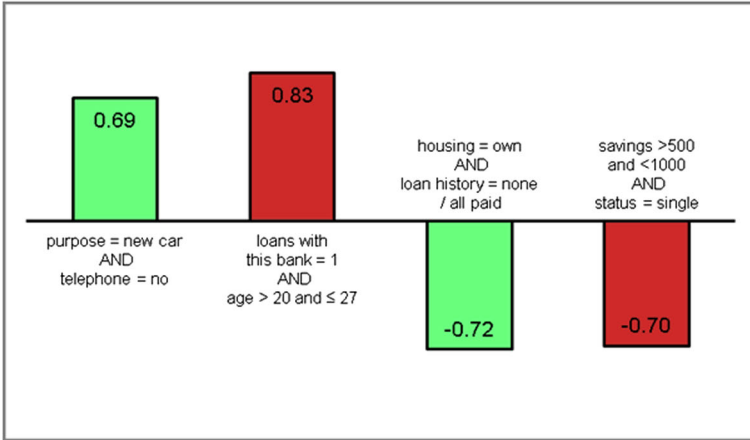
Future work will be able to build on our results and introduce higher legal as well as computational and interface complexity into our tasks and materials, in particular through exceptions/legitimate grounds for making distinctions based on discrimination-indexed features.

In the materials, the loan applicant and request were described in terms of *features*. The data-mining rules given to participants as a decision basis, as well as the motivations they could select for their decisions, were based on *risk factors* that subsumed features. For example, “age = 37” is a feature, and “age > 30” is a risk factor. We call features, risk factors, motivations and choices *discriminatory* versus *legitimate* (or *non-discriminatory*) depending on whether or not they involve age,

Assessed task 3 of 6

Information on the loan applicant and the loan application: Frank is a single 26-year old. He lives in his own house. He is a skilled employee with savings of \$800. He has neither telephone nor checking account. Currently, he has one existing loan at the bank, and he has paid back all previous loans. He asks for a loan of \$2000 for a new car.

Decision support: The data-analysis tool found four applicable rules.



Your conclusion: Should the loan be granted?

- yes
- no

Your motivation: The loan should be granted / denied because:

	<u>favorable</u>	<u>unfavorable</u>	<u>irrelevant</u>
Frank lives in his own house.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Frank is between 20 and 27 years old.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Frank has had no previous loans, or they have all been paid back.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The second column must not be checked here.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Frank is single.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Frank is male	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Fig. 4 Example screenshot with vignette, rules for data-mining decision support, decision, and motivations choice (partial view)

nationality, gender, or marital status. For example, “age > 30” is a discriminatory motivation, and “loan duration > 30” is a legitimate motivation. We call decisions based on legitimate motivations *non-discriminatory* decisions. Note that “discriminatory motivation” is used as a technical term and implies no statements about the psychological motives of the participant.

4.4 Method: details

4.4.1 Participants

In total, 215 US-based participants were recruited over Amazon Mechanical Turk. They received USD 6.00 for full participation and up to USD 1.50 as an additional performance-dependent payoff (bonus). Basic demographics were self-reported in an exit questionnaire (see Sect. 4.5.1).

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” (Goodman et al. 2012). To reduce cultural confounds, we recruited only US participants. We also heeded factors for quality control that have been observed to drastically reduce the occurrence of cheating on mTurk (Eickhoff and de Vries 2013). We included attention-check questions whose cross-evaluation can help identify users who checked answer options randomly. All participants obtained a check score of at least 50 % of the possible maximum. Further analyses of our results gave no indication of cheaters either. Based on these findings, we considered recruitment through mTurk an adequate choice for our study.

4.4.2 Design

The factors *setting* (Bank, ADA as short for anti-discrimination agency) and *mining form* (eDADM, cDADM, DM) were manipulated between subjects.

4.4.3 Procedure

Participants were given a series of scenarios with multiple answer options each. In each scenario, participants ticked exactly one answer corresponding to what they considered the best response for the decision and the relevance of each possible motivation. Three training tasks were presented first after an introductory page with the instructions. The correct answers for the training tasks were shown on the following page, so that participants could check theirs. Six assessed tasks, without information on the correct answers, followed this stage.

An exit questionnaire completed the study. First, we asked for impressions about the task and the tool. Twelve statements were rated on a 7-point Likert scale anchored in “strongly agree” and “strongly disagree”. As a simple reliability check, all items came in pairs, with one reverse-coded. The statements build on standard usability questionnaires (Lewis 1995). Subsequently, participants were asked for some basic demographics and personality traits (reciprocity).

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 decisions and motivations, the opinions, and the demographics) had to be filled in; all free-form answers were optional.

4.4.4 Tasks and materials

All tasks had the same basic scenario and overall task, which varied by setting, see Fig. 1. This was given at the beginning. Within this top-level instruction, each participant had to solve three exercise tasks and six assessed tasks.

Each task consisted of four parts. The first was a vignette in which a loan applicant was described briefly, for example by the text shown in Fig. 3. This was identical across all conditions. The second part was the output of a fictitious data-mining tool. In the third part of each task, participants were asked to decide whether to grant the loan request or not (Bank) resp. whether they considered it likely that the request would be granted or not (ADA). Fourth, they judged 12 possible motivations for their decision/conclusion by checking whether these were “favourable”, “unfavourable”, or “irrelevant” for the decision/conclusion. An example screenshot is shown in Fig. 4.

The tool output consisted of visualisations of decision rules in an intentionally minimalistic way that (a) follows the basic logic of the rule miners that inspired DADM and (b) implements the spirit of the DADM forms and standard data mining. In particular, the tool suggests a “voting” by rules of different strengths for the final decision as in CPAR (Yin and Han 2003), which is also used in DADM (Pedreschi et al. 2008); however it does not perform the last step of calculating the scores that makes the miner decide between two classes (“yes”/“no”). This calculation was left as a task for the user. The tool in its three versions also implements the basic spirit of cDADM (eliminate discriminatory rules), eDADM (highlight discriminatory features in rules), and data mining without DADM support (show all rules, whether they contain discriminatory features or not). Figure 5 shows an example of the three versions.

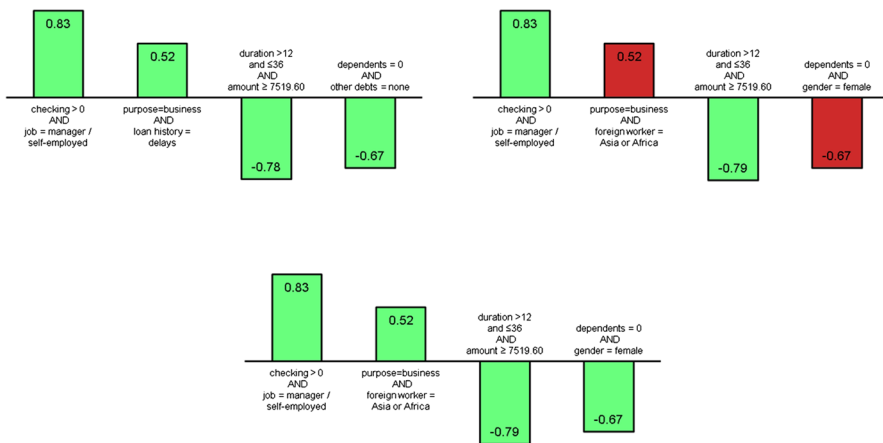


Fig. 5 The tool interfaces for (top left) cDADM, (top right) eDADM, and (bottom) DM. The visualization is identical between cDADM and DM, and the risk factors are identical between DM and eDADM. eDADM highlights rules with discriminatory features in red (second and fourth bar in the example). Identical visualisations were used for the Bank and ADA settings. (Color figure online)

Exercise task (ET) 1 explained the basic logic of rule certainties: Each bar is a rule with one or two risk factors in its premises. All of these must hold in order for the rule to be applied. If the positive risk factors (always above the line) outweigh the negative risk factors, the correct decision is yes, otherwise it is no. ET2 introduced more complex decision settings with several positive and negative rules (two of each). The task explained the basic logic of voting that consists of averaging the certainties of the positive and negative rules, respectively. The materials in ETs 1 and 2 were identical over all conditions. ET2 also gave participants the instruction: “For the following tasks, please remember to answer in line with the policy of your employer of relying on statistically validated results. However, you need not follow the statistical analyses blindly: please exercise judgment where needed.” ET3 introduced the topic of discrimination and alerted participants to the need to avoid it, see Fig. 2. As before, feedback was only given on the correctness of the decision.

Assessed tasks (AT) 1 to 6 were like ET3, but without feedback. All assessed tasks were designed equally and with no intentional differences in difficulty.

Risk factors and rule certainties were designed as follows: We created a pool of 17 legitimate attributes and 4 attributes that were explicitly described as discriminatory: nationality, age, gender and marital status. The legitimate attributes comprised further characteristics of the loan applicant (e.g. job status or duration of residence) and of the loan (e.g. loan purpose or duration). These attributes were given a total of 82 values to create features to describe the risk factors in the tasks.⁹

For each task from ET3 to AT6, we randomly chose 3 discriminatory plus 9 legitimate features to describe the applicant and the loan request. The descriptions in ET1 and ET2 had 4 resp. 8 legitimate features. Each feature in any given scenario referred to a different attribute.

From all features describing an applicant, 8 (ET3–AT6) resp. 6 (ET2) or 2 (ET1) were chosen as risk factors for the rules. The risk factors were distributed over the rules to produce 4 rules with 2 risk factors each (ET3–AT6) resp. 2 rules with 1 risk factor each (ET1) or 4 rules with 1, 1, 2, and 2 risk factors (ET2). Distribution was random, except that in both eDADM and DM, 1 positive rule contained 1 discriminatory feature and 1 negative rule contained another discriminatory feature. In cDADM, no rule contained any discriminatory feature. This is shown in Fig. 6.

In ET1 and ET2, the rule certainties implied one correct decision (yes resp. no). In ET3–AT6, the rule certainties were designed such that taking all risk factors and rules into account produced one decision, whereas taking only the legitimate ones into account produced the reverse decision. Thus, the first of these decisions was correct for the cDADM mining form (which had no discriminatory features and thus required that all risk factors and rules be considered), and the reverse was correct for

⁹ Our focus was not on analysing any specific true lending data, but on how people deal with data mining results that in reality often are or seem to be non-causal, with correlations often going against common sense and referring to features that act as a positive risk factor in one rule and as a negative risk factor in another one. However, we wanted to create a *possible* loan-related model. We therefore used the attributes of the German Credit Dataset (Newman et al. 1998) as well as their values, and added further values to create a sufficient number of features (for example, we converted the binary “foreign worker” attribute into a multi-valued attribute specifying the country of origin of the loan applicant).

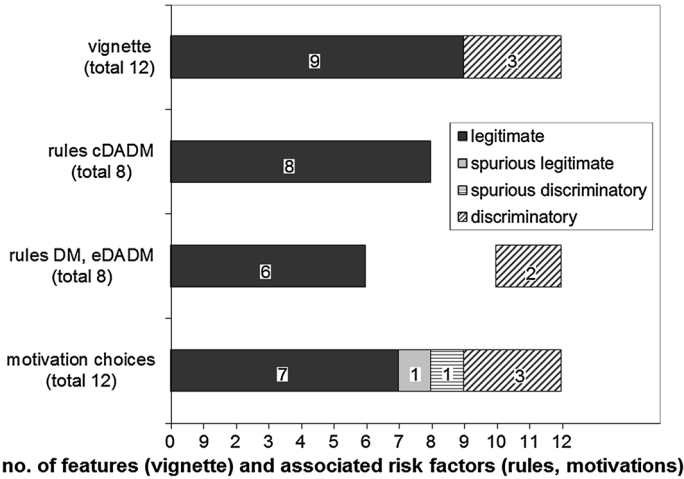


Fig. 6 The construction of features for vignette, rules and motivation choices. The figure gives the numbers of features of the different types. Thus, for example, in eDADM and DM the rules contained 6 legitimate features and 2 discriminatory features taken from the vignette. The motivation choices included all these features, plus 1 extra legitimate and 1 extra discriminatory from the vignette, and 1 extra spurious. (These numbers refer to ET3–AT6; ET1 and ET2 were slightly smaller and simplified)

the eDADM mining form (in which 1 positive and 1 negative rule had to be disregarded to reach a non-discriminatory decision). For 2 of the assessed tasks, the correct cDADM answer was “yes”, and thus for 4 of the assessed tasks, the correct eDADM answer was “yes”.

For each task, the features mentioned in the vignette, in the rules, and the motivation choices were chosen to ensure that all rules were applicable because they referred to features of the applicant or request. The possible motivations included correct choices (in the vignette, in the rules, and legitimate), irrelevant choices (not in the rules, or in a rule that was irrelevant because its premise also involved a discriminatory feature), discriminatory choices (involving discriminatory features), and spurious choices (not in the vignette). All vignette/rules/motivations designs followed the same schema, illustrated in Fig. 6. All vignette, rule, and spurious choices were random.

Remarks on the unavoidably larger complexity of the ADA task In a sense, the bank setting is more straightforward than the ADA setting. A bank clerk has data and rules (or other data-mining patterns) given by a tool and should make a decision based on this, but not on discriminatory features. An ADA clerk, on the other hand, is faced with an inherently epistemic task in the sense that she has data and patterns and has to make assumptions about somebody else’s reasoning and behaviour. These include assumptions about tool access and use, about motivations and decisions, and about one’s own role.

Assumptions about tool access and use assumptions could be “I have access to this tool, the bank has and uses the same tool” or “I have access to this tool, the bank has and uses a different tool”. Assumptions about motivations and decisions

could be “The bank tries to act ethically” or “The bank does not try to act ethically”. One’s own role could be perceived more as regulating (“I have to help the bank make ethical decisions”) or as monitoring (“I have to detect when unethical decisions were made”).

These inherently more complex task aspects are difficult to disentangle and more difficult still to manipulate experimentally. In addition, trying to do so would result in a large increase in the number of experimental conditions, in a situation in which we have no prior empirical knowledge about the workings of DADM in an ADA setting. We therefore decided (a) to use a simple baseline in this first experiment that was as similar as possible to the bank task and designed to draw participants’ attention to non-discriminatory decisions, (b) to allow for a certain openness in participants’ own interpretation of the setting, and (c) to reflect this in our analysis and interpretation of results.

4.5 Results and discussion

In this section, we describe the results of analysing the decisions and motivations given for the assessed tasks by the 215 participants, divided over the six conditions as shown in Table 2. Additional analyses (Sect. 4.5.6) also investigated exercise-task results. No decision or motivation restricted any other. Also, no indication of dependencies between decisions or between motivations were found in the results.

4.5.1 Participant demographics

Basic demographics were self-reported in an exit questionnaire: 43 % (56 %) of participants reported being female (male). Age ranged from 18 to 69, with a median of 31 years. Among all participants, 12 % reported high school graduate (or equivalent) as their highest grade of schooling, 40 % reported some college (1–4 years, no degree), 38 % a Bachelor’s degree, 6 % a Master’s degree or a Professional degree, and 2 % “Other”.¹⁰ 7 % reported that they “speak a language other than English at home”.

A quarter (24 %) reported that they are “dealing with data mining or statistics in [their] job or have done so in the past”. 25 % reported that they are “dealing with financial information in [their] job (e.g., banking, insurance, finance industry) or have done so in the past”. 13 % reported both. Together, these constituted 36 % of the sample.

Three quarters of participants stated that they had “applied for a loan at least once in [their] life” (73 %, validated by a reverse-coded question), with 50 % of these having at least once been denied a loan. Also, 50 % of all participants reported that they had “experienced discrimination in [their] own life”. These proportions mirror those found in our earlier study (Berendt and Preibusch 2012).

¹⁰ The US Census 2012 reports: 85 % (compared to our 98 %) “high school or more”, 28 % (compared to our 44 %) “Bachelor’s degree or more”, 10 % (compared to our 6 %) “advanced degree or more”. (<http://www.census.gov/compendia/statab/2012/tables/12s0233>).

Table 2 Numbers of participants, decisions, and motivations, over all tasks resp. assessed tasks (ATs)

	ADA- cDADM	ADA- DM	ADA- eDADM	Bank- cDADM	Bank- DM	Bank- eDADM
Participants	40	32	32	37	33	41
Decisions (all)	360	288	288	333	297	369
Motivations (all)	3,840	3,072	3,072	3,552	3,144	3,936
Decisions (ATs)	240	192	192	222	198	246
Motivations (ATs)	2,880	2,304	2,304	2,664	2,352	2,952

4.5.2 Decisions [H1]

To analyse decision quality, we investigated the impact of setting and mining form on the number of correct decisions.

We encoded the proportion of “correct decisions” in the assessed tasks as a $2 \times 3 \times 2$ contingency table (2 settings, 3 mining forms, correct/incorrect decisions) and analysed this with log-linear modelling including pairwise comparisons with Bonferroni corrections (Bresnahan and Shapiro 1966). The data are given in Fig. 7. Thus, for example in ADA-cDADM, 240 decisions were made, out of which 184 were correct (as defined in Sect. 4.4.4), which amounts to 76.7 %. Mining form was found to have a clear effect on decision correctness (significant at $\alpha = .01$).¹¹ Both cDADM and eDADM led to significantly higher proportions of correct decisions than DM, in both settings. No other main effects or interactions were significant.

Participants came to better decisions without taking longer: An investigation of times needed to come to the decisions and give the motivations showed a high variability between participants. On average, the bank setting led to longer response times, with a close-to-significant result in an ANOVA analysis of time-per-task ($p = .06$), and no other significant relationships. However, the DM average was higher for ADA. We were not able to find any other results that correlate with the higher times in the bank conditions.

Taken together, these results support H1: DADM supports users in making non-discriminatory decisions based on data-mining results, with more accurate results than not DADM-supported data mining.

4.5.3 Motivations: overview

But *why* did participants decide or conclude in the ways they did? We analysed the motivations and how they were judged. We partitioned all motivations into x different types and encoded the proportions of these different types in the assessed tasks as a $2 \times 3 \times x$ contingency table, with 2 the number of settings and 3 the number of mining forms. Two different partitionings were designed to take into account the different starting points of the two settings. The first, with $x = 3$ types, is described in Sect. 4.5.4 and the second, with $x = 2$ types, in Sect. 4.5.5. We

¹¹ All results reported as significant in the following were significant at $\alpha = .01$.

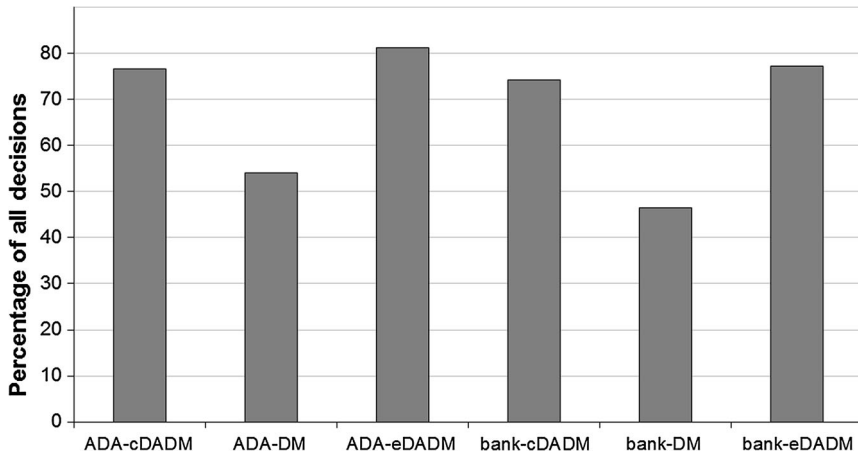


Fig. 7 Percentage of correct decisions by condition

analysed the partitions, including pairwise comparisons, with log-linear modelling, employing Bonferroni corrections.

In addition, we found that discriminatory features were mentioned by participants as relevant for their decisions or conclusions across all conditions. We present and discuss the results of this exploratory analysis in Sect. 4.5.6.

4.5.4 Motivations: the correct specification of legitimate motivations [H2, H3]

The first analysis focusses on the role of DADM for discrimination avoidance. Ideally, DADM would comprehensively ban discriminatory features from the decision discourse and allow decision makers to focus on other reasons for granting or withholding desired treatments. Such avoidance is in line with the major reason for banks to use DADM.

We partitioned the participants' motivations into three groups. (a) *Discriminatory* motivations, as defined in Sect. 4.3, involve nationality, gender, age or marital status. A motivation is discriminatory if the feature was deemed “favourable” or “unfavourable”, regardless of whether the applicant has this feature, of whether it is mentioned in a rule, and of whether it is a negative or a positive risk factor. (b) *Avoidance-correct* motivations are features that are legitimate, that the applicant possesses, that are mentioned in one of the task's admissible rules as a positive or negative risk factor, and that the participant correctly identifies as favourable resp. unfavourable. (c) *Avoidance-incorrect* motivations are all others.

The data are shown in Fig. 8. The three-way and all two-way interactions in the contingency table were significant. All pairwise differences except one were significant. Using $>$ to denote a significantly better performance and \sim an insignificant difference, we can summarize:

Bank: cDADM $>$ eDADM $>$ DM

ADA: cDADM \sim eDADM $>$ DM

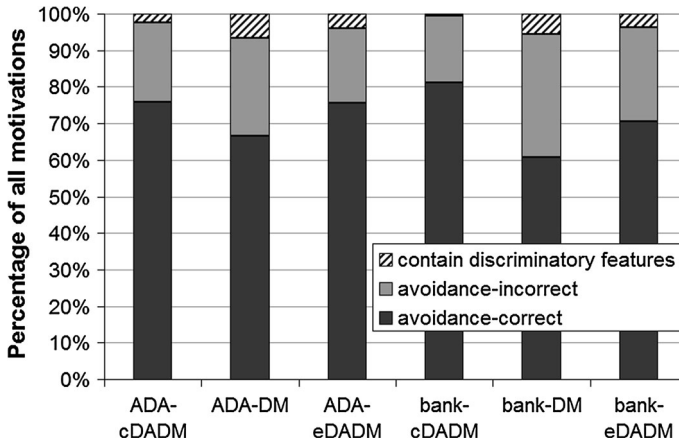


Fig. 8 Detection of correct and non-discriminatory motivations

The bank motivations profited from DADM more and suffered from DM more than the ADA motivations.

Taken together, these results support H2: DADM supports users in motivating their conclusions in non-discriminatory ways with more accurate results than not DADM-supported data mining.

They also support H3: For users focussed on making and motivating their decisions in non-discriminatory ways, cDADM supports more accurate and less discriminatory results than eDADM.

4.5.5 Motivations: the correct detection of discriminatory motivations [H2, H4]

Attention to a discriminatory motivation may mean different things depending on context. For example, some ADA participants indicated, in the free-form comments, that they saw their role as a kind of consultant for the described bank. In such a role, it would be important for them to spot a discriminatory feature/rule *in order to be able to advise, prospectively*, the bank to use other information. An ADA participant may also consider her role to be that of a watchdog who assumes that banks do not necessarily act ethically and therefore needs to spot a discriminatory feature/rule *in order to be able to demonstrate, retrospectively*, that a bank used it. In all such roles, it is key to pay close attention to all rules and risk factors in them.

The second analysis of all motivations therefore focusses on the role of DADM for discrimination detection. Ideally, DADM would comprehensively “spot” discriminatory features in the decision discourse and allow decision makers to focus on the workings of these reasons for granting or withholding desired treatments. Such detection is in line with a major reason for ADAs to use DADM.

We therefore partitioned the motivations slightly differently: (1) *Detection-correct* motivations are all risk factors suggested by the rules, if they are specified with the polarity as indicated in the rule. These comprise all avoidance-correct motivations in the sense of (b) above, and subsets of sets (a) and (c) above.

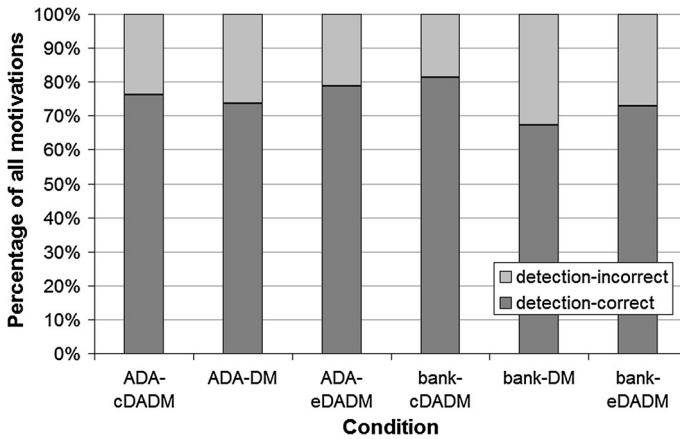


Fig. 9 Detection of given motivations (including discriminatory ones)

(2) *Detection-incorrect* motivations are all others. For cDADM, detection-correct coincides with avoidance-correct, and detection-incorrect covers discriminatory and avoidance-incorrect.

The data are shown in Fig. 9. The three-way and all two-way interactions in the contingency table were significant. All pairwise differences except two were significant. Using the same operators as above and $>\sim$ to denote a near-significantly better performance ($p = .02$), we can summarize:

$$\text{Bank: cDADM} > \text{eDADM} > \text{DM}$$

$$\text{ADA: eDADM} >\sim \text{cDADM} \text{ and } \text{eDADM} > \text{DM}$$

Taken together, these results support H2 and also H4: For users focussed on monitoring for preventing discriminatory decisions and motivating these conclusions, eDADM supports more accurate results than cDADM.

4.5.6 Motivations: signs of persisting discrimination?

Although H3 was supported, “less discriminatory” does not mean “not discriminatory”. On the contrary, discriminatory motivations were named as relevant (i.e. “favourable” or “unfavourable”) across all conditions, including all cDADM conditions in which deciding based on the data mining rules would have involved no discriminatory features, and all bank conditions in which using a discriminatory features clearly violated the bank’s obligations. In this section, we report the results of an exploratory analysis of these observations.

Figure 10 shows a further breakdown of the discriminatory motivations. It distinguishes between discriminatory features mentioned in the vignette and in the rules of a task, discriminatory features mentioned only in the vignette, and spurious features, present neither in the vignette nor in the rules. By the construction of the materials (see Fig. 6), vignette-and-rules features constituted half of the possible discriminatory choices in the motivation checklist in DM and eDADM (two of four)

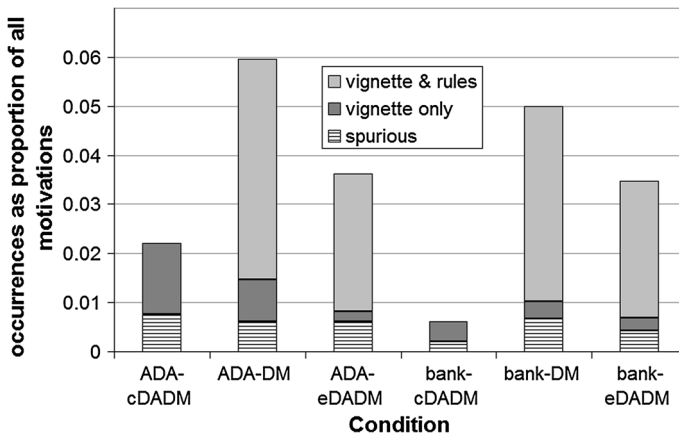


Fig. 10 Discriminatory motivation types

and 0 % in cDADM; vignette-only constituted 25 % resp. 75 %; and spurious features constituted 25 %.

The over-representation of vignette-and-rules features relative to these “prior probabilities” may indicate that motivation specifications were subject to an availability bias. Expressed differently, that the eDADM choice of highlighting rather than hiding a problematic feature may provoke discriminatory thoughts. The presence of spurious features in all conditions may indicate that pre-existing cognitive associations can be activated when judging other people, the typical working of prejudice. The semantics of some spurious discriminatory features suggested this. For example, participants appear to have inferred being married from having children. Alternatively, it may indicate a vulnerability to another cognitive bias, the so-called “Moses illusion” (Erickson and Mattson 1981; Park and Reder 2004): when words and with them thoughts are “put into people’s mouth”, they are prone to operating with them.¹²

The proportion of discriminatory motivations chosen within the set of all motivations is, fortunately, small. However, the data also indicate that it is persistent: Fig. 11 shows how many participants used at least one discriminatory motivation. Even in bank-cDADM, between 3 and 14 % of participants did this. The figure also suggests that the feedback after ET3 reduced the incidence of such mentions.

Given that we formulated the issue of persisting discrimination as a question rather than as a hypothesis, and that the numbers are relatively small, we do not investigate this subdivision in further statistical detail.

An analysis of the free-form comments revealed possible reasons for checking discriminatory motivations. First, discrimination may be seen only when it is explicitly negative—thus, a rule in which a discriminatory feature is named as a positive risk factor is not considered problematic. In other words, the fact that this

¹² The original observation was that when asked “How many animals of each kind did Moses take on the Ark,” most people respond “two,” even though they know that it was Noah, not Moses, who took the animals on the Ark.

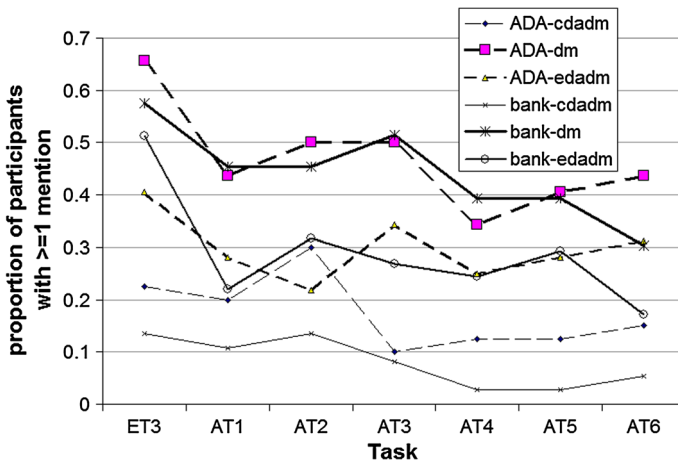


Fig. 11 Participants who mentioned at least one discriminatory motivation

very rule discriminates against people with a different value of the same attribute is not perceived. The data show some evidence of this: 80 % of the discriminatory motivations were rated as “irrelevant” when these features had been mentioned as a negative risk factor, compared to 75 % when they had been mentioned as a positive risk factor. Second, comments indicated a focus on nationality and gender as discriminatory, such that age and marital status were sometimes not identified as problematic. Third, some participants indicated their willingness to “reduce discrimination”. One participant remarked: “I dropped the -0.67 number a little bit because it included her being a female as a reason”. Fourth, background assumptions about loan collateral, job status, and prospects of repayment sometimes obscured the view on discrimination.

Of course, these observations should not be over-interpreted as indicating that any of our participants thought or acted in a sexist, ageist, or in any other way discriminatory fashion. Rather, we want to point out the effects that different data mining tools and the cues given by them may have on the cognitive saliency of discriminatory motivations. Even if a tool (such as our cDADM visualization) does not by itself give cues, the environment in which it is used may. For example, a company may internally and/or externally announce that they “are now using a discrimination-safe data-mining tool”. Such an announcement, mimicked by the instructions in our experiment, is in itself a possible cue-giver. What follows from cognitive saliency of discriminatory motivations is of course a question for further research.

In sum, even if cDADM’s hiding of discriminatory features from data mining *improves* decision making with respect to discrimination, it may not *eliminate* discrimination. Future work should investigate how to reduce the cognitive availability of discriminatory reasoning for decision-making situations like those of our fictitious bank clerk further, and how to reduce the generation of spurious, discriminatory reasoning across all settings.

4.5.7 Opinions on the tool, the task and the participant's own performance

In addition to measuring participants' performance with the tool, we also asked them for usability feedback. Building on a standard instrument (Lewis 1995), participants had to rate twelve statements on a 7-point Likert scale anchored in "strongly agree" and "strongly disagree". They covered the ease of understanding the vignettes, questions and the interface; enjoyment of the task and self-assessed performance at it; as well as intent to reuse the tool for future applications. The items were presented in a randomised order and consisted of six pairs, with a positively and a negatively worded version each. Cronbach's alpha of the overall instrument was $\alpha = 0.90$. The pairwise Pearson correlations between the items and their reverse-coded equivalents were between 0.53 and 0.77, suggesting an overall good reliability.

In general, participants appeared to like the tool, although their feedback was not overly enthusiastic. Of all participants, 62 % agreed or strongly agreed they found the interface easy to understand. 65 % found the questions understandable. More than half of the participants believed they had answered the questions correctly. This self-assessment correlated at $\rho = 0.26$ with their actual performance as the number of correct decisions ($R^2 = 0.07$). There were only weak correlations with the other per-item or overall usability ratings.

No clear picture emerged when we compared the usability ratings across the different experimental conditions. In particular, there is no setting or data mining form that scored systematically better.

4.5.8 Free-form comments

Participants made good use of their chance to comment. Every task had a field for free-form comments, and in addition there was the chance to give general feedback at the end. This led to a maximum of 10 comments per person (based on the data, we aggregated the two general-feedback data items into one). On average, each participant gave 3.3 comments. No clear differences emerged between the settings, but the fewest comments were given in the cDADM conditions, more in the DM conditions, and most in the eDADM conditions. The increase towards eDADM was clearer for ADA than for bank. Averages per condition were: 2.6 (ADA-cDADM), 2.4 (ADA-DM), 4.0 (ADA-eDADM), 2.7 (bank-cDADM), 2.9 (bank-DM), and 4.1 (bank-eDADM).¹³

The comments could be grouped into a number of main content categories, which all occurred in all conditions. (Additional specific content points are described in Sect. 4.5.6.) (a) Some comments just described how arithmetic was applied, such as "The negative risk factors outweigh the positive certainty", some of them enhanced: "Sum of balances is positive after removing discriminatory factors". (b) Many comments indicated that people had been thinking about the scenarios in depth, commenting on the features of the applicant and application and giving (sensible) real-world appraisals of them. They also commented on information that

¹³ Due to the exploratory nature of this analysis, we did not test these values for statistical significance.

was *not* mentioned in the rules. Examples of commenting, appraisals, and non-supplied information include “The length of the loan and its small size make it seem acceptable”, “Owns a car, so there’s collateral”, “If it’s a business loan, as a lender I’d want to see a business plan before approval”. (c) Some comments explicitly described the avoidance of discrimination, such as “Age and nationality must be disregarded, thus the middle two rules are ignored in the analysis” or “If we took into account some of his unfavorable factors we would be discriminating and we don’t want that.”

Several comments indicated that some participants perceived the study as a test of a new banking tool (and some then commented or complained about the unrealistic rules). Only one explicitly wondered whether this might instead be a “study on how people would react when given the choices presented”. Some comments showed visual thinking, i.e. the effectiveness of our interface choices: “Anything that contributed in the RED I marked irrelevant because legally you have to ignore discriminatory attributes.” There was a small number of comments on the tool itself, with proposals for interface improvements such as avoiding the need to scroll up and down. 32 participants stated that they had found the attention checks confusing, some indicating worries that they might have given the wrong answers to them, and five more commented on their content otherwise.

Many participants expressed their appreciation of the tasks, for example through “This was unique, interesting, and difficult” or “This was one of the most interesting and enjoyable studies I have done.”

5 Conclusions and future work

In this paper, we have investigated how computational methods can help enforce fairness in the knowledge society. Our focus has been on reducing discrimination as a key element of greater societal fairness, and on data mining as one of today’s most influential computational methods. In particular, we have presented a conceptual and an empirical analysis of the emerging area of DADM, with a special focus on data mining for decision support.

We have argued for the need to supplement classical, constraint-oriented discrimination-aware data mining by more exploratory forms. We have analysed how constraint-oriented and exploratory forms of DADM are likely to be deployed in practice and what this implies for evaluation. We have summarised the results of a first, exploratory user study, which 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.

In the subsequently described large-scale experimental user study, we have investigated how different forms of DADM can support data mining. We addressed the accuracy and actionability of the conclusions and the reasoning process. The results suggest that both constraint-oriented and exploratory DADM support correct conclusions and reasoning. The results also underline the differential merits of (a) the approach proposed by constraint-oriented DADM to *hide* discriminatory information and thus reduce its cognitive availability and (b) the approach proposed

by exploratory DADM to *highlight* discriminatory information and thus increase users' cognitive awareness. We used decision-making scenarios of a bank and of an anti-discrimination agency as typical examples of two relevant perspectives on whether people are granted loans or not. The results indicate that (a) constraint-oriented DADM can better support users focussed on directly preventing discriminatory decisions, whereas (b) exploratory DADM better supports users focussed on monitoring for preventing that discriminatory decisions are made. We therefore conclude that both forms of DADM complement each other and that appropriate combinations of them will be needed in future real-world tools.

There are of course many aspects of DADM usage that we have not addressed in this study. To conclude, we sketch four aspects as topics of future work.

1. *Tools and study design*: Our studies 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 interface choices may support or hinder correct interpretations in such cases. It will be particularly interesting to see how the “recall rather than recognition” requirements of open answers will affect cognitive availability and other heuristics and biases.

Also, participants studied tool output visualisations, but did not interact with the tools. The first reason for this was to make conditions as similar as possible, to reduce cognitive load, and to maximise experimental control. In addition, we believe that this accords well with the current state of the art in DADM, where far more algorithms exist than integrated, interactive deployments of these algorithms in tools. We expect a shift towards more full-fledged tools in the future. It will then be interesting to see how a sequence of exploratory activities and the need to integrate their results in such complex environments will influence visibility and actionability. Extending our methodology of crowdsourcing user-study participants along these lines will be a research challenge that can build on recent work on the evaluation of interactive tools with crowdsourcing (Zucon et al. 2013).

2. *Notion of discrimination*: As explained and motivated in Sects. 2.1 and 4.3, our study defined the discrimination to be avoided in an intentionally simplified way. The discrimination to be avoided in practice—the one in a legal or even in a sociological sense—is more complex and can often not be reduced to the mandate to avoid differentiating by one or several given features. Future DADM decision-support systems will have to go beyond data mining to be able to deal with decision context, exceptions, and other legally relevant circumscriptions of discrimination, and future DADM research should become a dedicated interdisciplinary area.
3. *Transparency*: eDADM in particular, by its focus on making decision grounds and valuations attached to them visible, can serve as a transparency tool (Gutwirth and De Hert 2006)—an instrument that can make the decision-making of institutions (private or governmental) more understandable. First, it could help make the decisions of monitored institutions (as in the ADA setting)

or of one's own institution (as in the Bank setting) more transparent. Second, it could not only increase understandability for people directly involved in decision-making or in monitoring decision-making, but also for citizens in general. These are the intended beneficiaries of the transparency called for today throughout the world, including the EU and the US, e.g. (European Commission 2012; Federal Trade Commission 2012). The purpose of such transparency tools is to “compel government and private actors to ‘good practices’ by focusing on the transparency of governmental or private decision-making and action” (Gutwirth and De Hert 2006, p. 9). This can also help achieve more accountability (Alhadeff et al. 2011). To realise this potential, future work on eDADM will need to develop methods that can present data and decision-making to citizens in a usable way and at the same time respect the data-privacy and intellectual-property constraints under which decision-making institutions operate.

eDADM also has the potential to enhance transparency in another sense. Recently, cDADM authors have observed that some patterns of differentiations may be explainable by correlations of discrimination-indexed features with legitimate grounds for differentiation—for example, “no known savings” (Luong 2011, p. 59) as a legitimate ground for rejecting a loan application, or women on average missing specific requirements for a job (Kamiran et al. 2013; Kamishima et al. 2012). The authors have proposed modifications to their algorithms that essentially split an observed pattern of differentiation that appears to be discriminatory into the variance explained by these legitimate grounds and the residual variance that expresses the “real” discrimination by a discrimination-indexed attribute. However, such real-life patterns can also be interpreted in terms of the “intersectionality” of real-life discrimination: the observation that multiple factors of societal disadvantages tend to intersect (Knudsen 2006) (such as specific ethnicities, genders, and ages, low educational level, and poverty). The cDADM approach to “explain away” differentiation may often guard against inappropriate assumptions about decision makers' intentions, but it also effectively hides patterns of intersectionality. In contrast, the eDADM approach can serve to make these very patterns of intersectionality more transparent.

4. *The role of data mining:* We have concentrated on how data mining can contribute to, or help prevent, discrimination by virtue of how patterns are processed and/or presented. However, data mining may also contribute to discrimination in the narrow sense by virtue of its features rather than its patterns.

First, using an attribute at all draws attention to a differentiation that may as well not be made, whereas not storing and/or using an attribute such as nationality would avoid this. This claim is supported by findings from domains as different as peer-reviewing in science and job applications without gender, where the evidence suggests that a decision maker who does not know an attribute's value (the name of the paper's author, the gender of the job applicant) may make choices that are less biased and ultimately lead to better-quality publications or applicant short-lists. On

the other hand, if these features are also unknown to monitoring stakeholders, these may not be able any more to find patterns of indirect discrimination. This might be addressed by sophisticated solutions of differentially disclosed information.

Second, data mining not only uses, but also often produces features. An example are the “profiles” found as patterns in uses such as user/customer modelling. Such profiles are then ascribed as features to new individuals, and this may perpetuate or introduce new discrimination (Berendt 2012). eDADM, by its exploratory nature, can also increase transparency by making such by-products of data mining and new forms of discrimination visible—and thus contribute to more reflection, societal discussion and ultimately better decision making. However, “fairness-aware” computational methods (Kamishima et al. 2012) by themselves cannot ensure social fairness, and they may have side-effects. For example, when insurance tariffs may no longer discriminate by sex, but new sensors and data (of eating habits, sports performance, driving style, etc.) are readily available and their use permitted, the data mining of such data becomes very attractive. Especially when the notion of distributional justice underlying the use of data mining remain stable (for example, premiums based on individual risk factors rather than ability to pay), “eradicating” one form of discrimination may merely shift problems. To the extent that the newly identified desired behaviours indeed are under the control of the individual, new social norms (of eating, movement, and other behaviours) get created and enforced, which can severely limit individual freedoms. To the extent that behaviours are not or only partially under the control of the individual and/or that multiple factors of societal disadvantages intersect, patterns of exclusion will be maintained or shift only marginally. Bringing transparency into *these* patterns is an interesting challenge for fairness-aware data mining—but changing the reality of these patterns also requires legal reasoning and concrete decisions beyond the choice of computational approaches.

Acknowledgements We thank Brendan Van Alsenoy and Albrecht Zimmermann for many inspiring discussions and valuable comments on an earlier version of the paper, and the Flemish Agency for Innovation through Science and Technology (IWT) and the Fonds Wetenschappelijk Onderzoek—Vlaanderen (FWO) for support through the projects SPION (Grant Number 100048) resp. Data Mining for Privacy in Social Networks (Grant Number 65269).

References

- Alhadef J, Van Alsenoy B, Dumortier J (2011) The accountability principle in data protection regulation: origin, development and future directions. Presented at the privacy and accountability 2011 conference, Berlin, 5–6 Apr 2011. <http://ssrn.com/abstract=1933731>. 11 Oct 2013
- Arnott D (2006) Cognitive biases and decision support systems development: a design science approach. *Inf Syst J* 16(1):55–78
- Avraham R, Logue KD, Schwarcz D (2013) Understanding insurance anti-discrimination laws. Technical report. U of Michigan law & econ research paper no. 12-017; U of Michigan public law research paper no. 289; U of Texas Law. Law and econ research paper no. 234; Minnesota legal studies research paper no. 12-45. <http://dx.doi.org/10.2139/ssrn.2135800>. 20 Aug 2013
- Berendt B (2012) More than modelling and hiding: towards a comprehensive view of web mining and privacy. *Data Min Knowl Discov* 24(3):697–737
- Berendt B, Preibusch S (2012) Exploring discrimination: a user-centric evaluation of discrimination-aware data mining. In: Vreeken et al. (2012), pp 344–351

- Berendt B, Preibusch S, Teltzrow M (2008) A privacy-protecting business-analytics service for online transactions. *Int J Electron Commer* 12:115–150
- Boston Consulting Group (2012) The value of our digital identity. Liberty global policy series. <http://www.lgi.com/PDF/public-policy/The-Value-of-Our-Digital-Identity.pdf>. 20 Aug 2013
- Bresnahan J, Shapiro M (1966) A general equation and technique for the exact partitioning of chi-square contingency tables. *Psychol Bull* 66:252–262
- Calders T, Verwer S (2010) Three naive Bayes approaches for discrimination-free classification. *Data Min Knowl Discov* 21(2):277–292
- Chen JQ, Lee SM (2003) An exploratory cognitive DSS for strategic decision making. *Decis Support Syst* 36(2):147–160
- Duhigg C (2009) What does your credit-card company know about you? *New York Times*, 12 May 2009. http://www.nytimes.com/2009/05/17/magazine/17credit-t.html?pagewanted=all&_r=0. 20 Aug 2013
- Eickhoff C, de Vries AP (2013) Increasing cheat robustness of crowdsourcing tasks. *Inf Retr* 16(2): 121–137
- Ericksen TA, Mattson ME (1981) From words to meaning: a semantic illusion. *J Verbal Learn Verbal Behav* 20:540–552
- EU (2004/2012) Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:373:0037:0043:EN:PDF>. 20 Aug 2013
- EU (2006) Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:204:0023:0036:EN:PDF>. 20 Aug 2013
- European Commission (2012) How does the data protection reform strengthen citizens' rights? http://ec.europa.eu/justice/data-protection/document/review2012/factsheets/2_en.pdf. 20 Aug 2013
- European Court of Justice (2011) Case C-236/09, Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres. <http://curia.europa.eu/juris/liste.jsf?language=en&num=C-236/09>. 20 Aug 2013
- Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) *Advances in knowledge discovery and data mining*. MIT Press, Cambridge, MA, pp 1–34
- Federal Trade Commission (2012) Protecting consumer privacy in an era of rapid change: recommendations for businesses and policymakers. FTC report. <http://www.ftc.gov/os/2012/03/120326privacyreport.pdf>. 20 Aug 2013
- Fine C (2010) *Delusions of gender. The real science behind sex differences*. Icon Books, London
- Gao B, Berendt B (2011) Visual data mining for higher-level patterns: discrimination-aware data mining and beyond. In: *Proceedings of the 20th machine learning conference of Belgium and The Netherlands*. <http://www.benelearn2011.org/>. 20 Aug 2013
- Goodman J, Cryder C, Cheema A (2012) Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. *J Behav Decis Mak* 26:213–224
- Gutwirth S, De Hert P (2006) Privacy, data protection and law enforcement. Opacity of the individual and transparency of power. In: Claes E, Duff A, Gutwirth S (eds) *Privacy and the criminal law*. Intersentia, Antwerp, pp 61–104
- Hajjan S (2013) Simultaneous discrimination prevention and privacy protection in data publishing and mining. PhD thesis, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Tarragona, Catalonia
- Hajjan S, Domingo-Ferrer J (2013) Direct and indirect discrimination prevention methods. In: Custers B, Caldere T, Schermer B, Zarsky TZ (eds) *Discrimination and privacy in the information society, studies in applied philosophy, epistemology and rational ethics*, vol 3. Springer, Berlin, pp 241–254
- Hajjan S, Domingo-Ferrer J (2013) A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans Knowl Data Eng* 25(7):1445–1459
- Hajjan S, Domingo-Ferrer J, Martínez-Ballesté A (2011) Discrimination prevention in data mining for intrusion and crime detection. In: *IEEE SSCI 2011*
- Hajjan S, Monreale A, Pedreschi D, Domingo-Ferrer J, Giannotti F (2012) Injecting discrimination and privacy awareness into pattern discovery. In: *Vreeken et al. (2012)*, pp 360–369
- Heckerman D (2013) From wet to dry: how machine learning and big data are changing the face of biological sciences. <http://research.microsoft.com/apps/video/default.aspx?id=189426>

- Kamiran F, Calders T, Pechenizkiy M (2010) Discrimination aware decision tree learning. In: Proceedings of ICDM'10, pp 869–874
- Kamiran F, Karim A, Verwer S, Goudriaan H (2012) Classifying socially sensitive data without discrimination: an analysis of a crime suspect dataset. In: Vreeken et al. (2012), pp 370–377
- Kamiran F, Zliobaite I, Calders T (2013) Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl Inf Syst* 35(3):613–644
- Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Considerations on fairness-aware data mining. In: Vreeken et al. (2012), pp 378–385
- Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Fairness-aware classifier with prejudice remover regularizer. In: ECML/PKDD (2), LNCS, vol 7524, pp 35–50. Springer
- Kaplan B (2001) Evaluating informatics applications—clinical decision support systems literature review. *Int J Med Inform* 64(1):15–37
- Knudsen S (2006) Intersectionality—a theoretical inspiration in the analysis of minority cultures and identities in textbooks. In: Caught in the web or lost in the textbook, pp 61–76. http://iartem.no/documents/caught_in_the_web.pdf. 20 Aug 2013
- Lewis JR (1995) IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int J Hum-Comput Interact* 7(1):57–78. <http://hcibib.org/perlman/question.cgi>. 31 July 2012
- Luong BT (2011) Generalized discrimination discovery on semi-structured data supported by ontology. PhD thesis, IMT Institute for Advanced Studies, Lucca, Italy
- Luong BT, Ruggieri S, Turini F (2011) k-nn as an implementation of situation testing for discrimination discovery and prevention. In: KDD, pp 502–510. ACM
- Mancuhan K, Clifton C (2012) Discriminatory decision policy aware classification. In: Vreeken et al. (2012), pp 386–393
- Marghescu D, Rajanen M, Back B (2004) Evaluating the quality of use of visual data-mining tools. In: Proceedings of 11th European conference on IT evaluation, 11–12 Nov 2004, Amsterdam, pp 239–250. Academic Conferences Limited
- Microsoft (2012) New York City Police Department and Microsoft partner to bring real-time crime prevention and counterterrorism technology solution to global law enforcement agencies. <http://www.microsoft.com/en-us/news/Press/2012/Aug12/08-08NYPDPR.aspx>. 20 Aug 2013
- Newman DJ, Hettich S, Blake CL, Merz CJ (1998) UCI repository of machine learning databases. GCD at <http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29>. 20 Aug 2013
- Park H, Reder ML (2004) Moses illusion. In: Pohl FR (ed) *Cognitive illusions*, pp 275–291. Psychology Press, London
- Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of KDD'08, pp 560–568. ACM
- Pedreschi D, Ruggieri S, Turini F (2009) Integrating induction and deduction for finding evidence of discrimination. In: ICAIL, pp 157–166. ACM
- Pedreschi D, Ruggieri S, Turini F (2009) Measuring discrimination in socially-sensitive decision records. In: SDM, pp 581–592
- Pedreschi D, Ruggieri S, Turini F (2012) A study of top-k measures for discrimination discovery. In: SAC '12, pp 126–131. ACM, New York, NY, USA
- Perer A, Shneiderman B (2009) Integrating statistics and visualization for exploratory power: from long-term case studies to design guidelines. *IEEE Comput Graphics Appl* 29(3):39–51
- Pitt G (2009) Genuine occupational requirements. EC anti-discrimination legislation for legal practitioners, 27–28 Apr 2009, Trier, Germany. http://www.era-comm.eu/oldoku/Adiskri/05_Occupational_requirements/2009_Pitt_EN.pdf. 20 Aug 2013
- Plaisant C (2004) The challenge of information visualization evaluation. In: Costabile MF (ed) *AVI*, pp 109–116. ACM Press, New York
- Romei A, Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. *Knowl Eng Rev* (to appear). doi:10.1017/S0269888913000039
- Ruggieri S, Pedreschi D, Turini F (2010) Data mining for discrimination discovery. *TKDD ACM Trans Knowl Discov* 4(2):1–40
- Ruggieri S, Pedreschi D, Turini F (2010) DCUBE: discrimination discovery in databases. In: Proceedings of SIGMOD'10, pp 1127–1130
- Schanze E (2013) Injustice by generalization. Notes on the Test-Achats decision of the European Court of Justice. *Ger Law J* 14(2):423–433

- Sedlmair M, Meyer M, Munzner T (2012) Design study methodology: reflections from the trenches and the stacks. *IEEE Trans Vis Comput Graphics* 18(12):2431–2440
- Shearer C (2000) The CRISP-DM model: the new blueprint for data mining. *J Data Warehous* 5(4):13–22
- Sykes JB (ed) (1982) *The concise Oxford dictionary*, 7th edn. Oxford University Press, Oxford
- Vreeken J, Ling C, Zaki MJ, Siebes A, Yu JX, Goethals B, Webb GI, Wu X (eds) (2012) 12th IEEE ICDM workshops, Brussels, Belgium, 10 Dec 2012. IEEE Computer Society
- Yin X, Han J (2003) Cpar: classification based on predictive association rules. In: Barabási D, Kamath C (eds) *SDM*. SIAM, Philadelphia, PA
- Zuccon G, Leelanupab T, Whiting S, Yilmaz E, Jose JM, Azzopardi L (2013) Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Inf Retr* 16(2):267–305