AI & SOCIETY (2023) 38:721–731
https://doi.org/10.1007/s00146-022-01472-5

OPEN FORUM

Fairness & friends in the data science era

Barbara Catania[1] · **Giovanna Guerrini**[1] · **Chiara Accinelli**[1]

Received: 29 July 2021 / Accepted: 13 April 2022 / Published online: 9 June 2022

Abstract
The data science era is characterized by *data-driven automated decision systems* (ADS) enabling, through data analytics and machine learning, automated decisions in many contexts, deeply impacting our lives. As such, their downsides and potential risks are becoming more and more evident: technical solutions, alone, are not sufficient and an interdisciplinary approach is needed. Consequently, ADS should evolve into *data-informed ADS*, which take *humans in the loop* in all the data processing steps. Data-informed ADS should deal with data *responsibly*, guaranteeing *nondiscrimination* with respect to protected groups of individuals. Nondiscrimination can be characterized in terms of different types of properties, like fairness and diversity. While fairness, i.e., absence of bias against minorities, has been widely investigated in machine learning, only more recently this issue has been tackled by considering all the steps of data processing pipelines at the basis of ADS, from data acquisition to analysis. Additionally, fairness is just one point of view of nondiscrimination to be considered for guaranteeing equity: other issues, like diversity, are raising interest from the scientific community due to their relevance in society. This paper aims at critically surveying how nondiscrimination has been investigated in the context of complex data science pipelines at the basis of data-informed ADS, by focusing on the specific data processing tasks for which nondiscrimination solutions have been proposed.

Keywords Data-informed automated decision system · Processing pipeline · Nondiscrimination · Fairness · Diversity

1 Introduction

The impact of data on our society is getting higher and higher, with data about people being more and more often exploited as the basis to make decisions that might impact people's lives: we are inside the data science era. Decisions based on data are taken, e.g., to determine whether to admit a student to a school, to hire an employee, to offer a loan to an applicant, and event to grant socially useful work to an arrested person. Consequently, the downsides and potential risks of a deep use of people-related data in decision processes are becoming more and more evident: the probability of inequities is likely to increase and be amplified by *data-driven automated decision systems* (ADS), relying on data to take guided decisions. Such systems should therefore evolve into *data-informed ADS*, that take *humans in the loop* in all the data processing steps, from acquisition to analysis (Stoyanovich et al. 2020). Indeed, to obtain insights, data from disparate sources first need to be organized in a clean unified dataset, then undergoing a *data processing pipeline*, i.e., a sequence of complex tasks usually involving, besides learning tasks, many data preparation steps like data cleaning, transformation, matching, and merging (Doan 2018).

Data-informed ADS should deal with data in a *responsible* way: besides ensuring *transparency* and *interpretability,* allowing people to understand the process and the decisions, they should guarantee *nondiscrimination* with respect to all the considered groups of individuals (Stoyanovich et al. 2020). More concretely, consider a population upon which a data processing (either operational or analytical) task is to be applied. Suppose that a subset of our population shares some characteristics that should not be employed for discrimination (e.g., race, gender, disability status). It is important to guarantee that the result of the processing task is not discriminating with respect to the considered *sensitive* attributes. This may include ensuring a fair probability

✉ Barbara Catania
barbara.catania@unige.it

Giovanna Guerrini
giovanna.guerrini@unige.it

Chiara Accinelli
chiara.accinelli@dibris.unige.it

[1] University of Genoa, Genoa, Italy

of selection, not giving undue relevance to specific groups of individuals sharing these properties, or other related constraints.

Nondiscrimination can be characterized in terms of different properties. In social sciences, one key concept to guarantee nondiscrimination is *equity,* i.e., absence of bias against minorities. As stated by Jagadish et al. (2021), equity promotes *fairness* by treating people differently depending on their endowments and needs (equality of outcome), whereas equality aims to achieve fairness through equal treatment regardless of need (equality of opportunity). Besides fairness, other issues like *diversity*, i.e., the degree to which different kinds of objects are represented in a dataset, are raising interest from the scientific community, due to their relevance in society.

Fairness and diversity are not new concepts. The relevance of fairness is well recognized by the machine learning and data mining communities (Mehrabi et al. 2021). On the other hand, diversity is one of the main relevant concepts in recommender systems (Kaminskas and Bridge 2017). This *last mile* of data analysis, i.e., the decision-making components, is indeed the most visible part of data science. More recently, the importance of a lifecycle view of data science lead to realize that the achieved results are not enough (Asudeh 2021). As first pointed out in Abiteboul et al. (2016), algorithmic fairness has to be tackled by developing a holistic treatment of nondiscrimination, tailored to incrementally enforcing non-discriminating constraints along the pipeline at the basis of ADS, through individual independent choices, rather than as a constraint on the set of final results. Any automated task can indeed introduce *technical bias* by exacerbating pre-existing bias that may lead to inequity in society. This type of requirements is not only made desirable by the ethical need to take responsibility, but also mandatory by the recent General Data Protection Regulation (GDPR) of the European Union (Bonatti and Kirrane 2019). The GDPR imposes that this type of guarantees is provided *by design*, i.e., intrinsically embedded in the mechanisms of the data processing workflow.

Starting from those considerations, the aim of this paper is to critically survey how nondiscrimination can be modeled and how it can be guaranteed in the context of complex data science pipelines at the basis of data-informed ADS, thus complementing the already existing reviews on fairness in machine learning and AI, not further discussed in this paper, with a broader focus on all data processing tasks. In particular, this paper focuses on the following research questions:

- RQ1: For which data processing tasks have nondiscrimination solutions been proposed?
- RQ2: Which communities (defined in terms of author geographical location) have been most active in this field?

In the remainder of the paper, we first present the typical structure of data processing pipelines (Sect. 2) and we classify the main properties proposed for modeling nondiscrimination (Sect. 3). Then, we present the methodology used in our literature review and we briefly survey the main achievements for each data processing task (Sect. 4). A discussion on the review results concludes the paper (Sect. 5).

2 Data processing pipelines in data science

As pointed out in Jagadish et al. (2014), creating value from Big Data is a multi-step process: data acquisition, data cleaning, data integration, and analysis.[1] The steps of a typical data processing pipeline in ADS are graphically depicted in Fig. 1 and discussed in what follows (with reference example a college admission system). The figure also contains the term *back-end*, commonly used in the data warehousing context to refer to all the extraction and transformation processes data undergoes before feeding the centralized repository on which the *front-end* components, with which the decision-maker interacts, perform analyses.

Data acquisition. Data at the basis of data science is a record of some underlying activity of interest. It can be gathered as any effect of any interaction with or observation of the world around us, ranging from any application relying on an operational database, to logs of user-activity on a website or event-logs in a software, to physical sensors in Internet of Things systems. Much of these data can be filtered and aggregated without compromising our ability to reason about the underlying activity of interest. One challenge is to define these "on-line" filters in such a way they do not discard useful information. Effective data-driven decisions can be enabled by acquiring data from multiple heterogeneous data sources.

Data cleaning. Collected data can be structured, semi-structured, or unstructured, and rarely are in a format ready for analysis. Even limiting to structured data, most data sources are notoriously unreliable: data entry can be partial, sensors can be faulty, humans may provide biased opinions, remote websites might be stale, and so on. As a result, data may suffer of many data quality issues. In the context of a college admission system, for example, some grades or personal information can be missing, grades can refer to different scales or be out-of-scale; the same student can be modeled by distinct records. We cannot leave the data in this form and still effectively analyze it. Rather, data cleaning techniques, facing the possible sources of errors, are applied

[1] They devise five stages in the Big Data pipeline, out of which we focus on the first four, since the last one (interpretation) is not automatable rather is by the decision-maker.

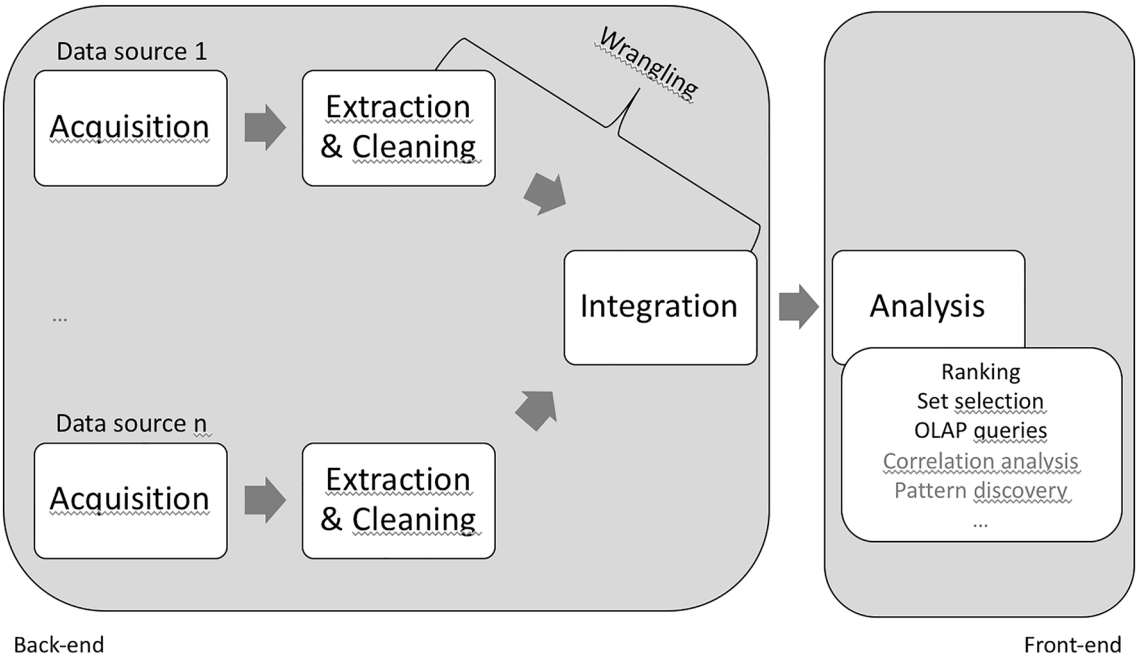

Fig. 1 Steps in a data processing pipeline

to the acquired data, removing or alleviating the data quality issues.

Data integration. Effective large-scale analysis often requires the collection of heterogeneous data from multiple sources. For example, in the context of a college admission system, to obtain the best-informed decisions, it would be useful to have a 360 view of applicants (including scores and assessments, but also, e.g., the essay from which we can derive information about interests) and information about the academic careers of previously enrolled students. The issue here is to resolve heterogeneities in data structure and semantics, obtaining a reconciled integrated dataset that is uniformly interpretable within a community, fitting its analytical needs. This is achieved through a number of *transformations*, which can be supported by integration tools.

Since the cost of full integration is often formidable and the analysis needs quickly shift, "pay-as-you-go" integration techniques (Madhavan et al. 2007) perform much of this work on-the-fly in support of ad hoc exploration. In this case, especially when the data sources are not databases with well-known schemas, the term *data wrangling* (Rattenbury et al. 2017) is used, referring to the activities for cleaning, integrating, and organizing data before it can be analyzed. The result of data wrangling can provide important metadata for further insights about the data.

Analysis. All the methods for querying and mining data, to extract valuable information for decision-making and strategical actionable knowledge, fall in front-end components. Here, we focus on analytical queries rather than on mining and learning algorithms. Specifically, we consider ranking and set selection techniques, as well as OLAP queries.

The aim of *ranking* is, given a set of items, to produce a rank for each item in the set. In score-based ranking, a given set of candidates is sorted on the score attribute, which may itself be computed on the fly, and returned in sorted order.[2] We typically return the best-ranked k candidates, the top-k. As an example, taken from Zehlike et al. (2021), consider an admissions officer at a university who selects candidates from a large applicant pool. Each applicant submits several quantitative scores, all of them can transformed to a discrete scale of 1 (worst) through 5 (best): the high school GPA (grade point average), the verbal portion of the SAT (Scholastic Assessment Test) score, and the mathematics portion of the SAT score. The score of each candidate can be obtained based on a formula that the admissions officer gives, and then return some number of highest-scoring applicants in ranked order. This scoring formula may, for example, specify the score as a linear combination of the applicant's high-school GPA and the two components of the SAT score, each carrying an equal weight. The admissions officer will take the order in which the candidates appear in

[2] As an alternative, in supervised learning, a preference-enriched training set of candidates is given, with preferences among them stated in the form of scores, preference pairs, or lists. This training set is used to train a model that predicts the ranking of unseen candidates. In our analysis, we focus on score-based methods and disregard supervised ones.

the ranking when deciding whom to consider more closely, interview, and admit.

The main goal of *set selection* is to select a specific number of items from a given set of items, according to some specific conditions. More precisely, assume that we have a set of items, each with associated attributes. From this set, we wish to select k items to maximize a utility score, defined in terms of the attributes. In the university admission example, we could select the k candidates with the highest score computed as a function of the candidates' attribute (GPA and SAT scores). The items in the set may be presented to us together or one at a time. Set selection is thus a special case of ranking that ignores the relative order among the top-k, returning them as a set.

Online analytical processing (OLAP) is an essential task of decision-support systems. *OLAP queries* are queries defined against a data cube that introduce a multidimensional range (via specifying an interval for each dimension of the data cube) and a SQL aggregate operator and return as output the aggregate value computed over cells of the data cube contained in that range. With reference to the university admission domain, examples of OLAP queries are those returning the average GPA of candidates that completed a public high school in Italy in 2020, or the average GPA of candidates admitted to a certain university in 2021, per gender and major.

3 Fairness & friends

To ensure a responsible data processing, any automated task in a data processing pipeline should guarantee nondiscrimination with respect to minorities and protected groups of individuals, defined in terms of specific sensitive attributes. Nondiscrimination can be characterized in terms of different properties, briefly surveyed in what follows.

3.1 Fairness

Fairness is generally associated with the lack of discrimination; it can be broadly defined as the impartial treatment of individuals and of demographic groups. A multitude of interpretations and technical definitions have been provided, initially tailored to supervised learning tasks (Verma and Rubin 2018) and nowadays used for any processing activity. A first distinction is between *individual fairness* and *group fairness*. Individual fairness, first defined in Dwork et al. (2012), refers to the requirement that similar individuals, defined in terms of a given metric, are treated similarly; group fairness, also known as *statistical parity* or *demographic parity*, refers to the requirement that demographics of those receiving a particular positive outcome are identical to the demographics of the population as a whole (Dwork et al. 2012). Demographics are defined in terms of a given sensitive attribute, like gender or race, and statistical parity requires the distribution of values of the considered attribute in the result of the task be the same as its distribution in the input population. As an example, consider a set with 10 students (5 paying reduced fees and 5 paying standard fees) and suppose you should select 4 of them. The result of the selection will satisfy group fairness only if the distribution of the two groups of students in the result and in the population coincide, thus both groups will receive the same treatment. This means that 2 students paying reduced fees and 2 paying standard fees should be selected. Now suppose that, in the first group, there are 2 Italian and 3 non-Italian students paying reduced fees and, in the second group, there are 2 non-Italian and 3 Italian students paying standard fees. The result of the selection satisfies individual fairness if all individuals are treated similarly, thus, in this example, if the 2 selected students paying reduced fees are Italian and the 2 selected students paying standard fees are non-Italian, respectively.

Many variations of group fairness have been proposed, all together referred to as *associational fairness* (see, e.g., Salimi et al. 2019a) since they all focus on correlating the outcome of a classification task with the values for the considered sensitive attributes. Among them, we recall (Chouldechova 2017): *conditional statistical parity*, a variation of group fairness that checks for a set of admissible factors, e.g., we want our individuals to be from Italy (Corbett-Davies et al. 2017); *equalized odds*, which requires that both protected and privileged groups have the same false positive rate and the same false negative rate; *predictive parity*, requiring that both protected and unprotected groups have the same predicted positive value; *intersectional fairness*, introduced as a way for extending group fairness to subgroups, defined by considering the intersection of several demographical variables (e.g., gender, race, age).

Unfortunately, associational fairness constraints are context-specific and might fail to distinguish the right correlations between a sensitive attribute and the outcome of a given algorithm (Dwork et al. 2012). Additionally, they can be mutually exclusive, thus they lack universality; as an example, it has been shown that equalized odd and predictive parity are incompatible (Chouldechova 2017). This observation has motivated the definition or alternative *causal fairness* constraints, under which fairness holds when the sensitive attribute has no causal influence on the outcome of a given task, thus avoiding anomalous correlations. Even in this case, many variations have been proposed. For example, under *counterfactual fairness*, the sensitive attribute should not be the cause of the outcome at the individual level; it coincides with individual fairness only under some strong assumptions (Kusner et al. 2017) and cannot be estimated from data. To avoid this limitation, *proxy fairness* considers

counterfactuals at the group level, but it does not capture group fairness as defined above (Gupta et al. 2018; Kilbertus et al. 2017). Other causal fairness notions rely on various properties of the causal graphs with the aim of avoiding specific paths from the sensitive attribute to the outcome (see, e.g., Nabi and Shpitser 2018) but often rely on very restrictive assumptions and are of limited use. In general, causal fairness constraints rely on background information regarding the underlying causal model, which might not be available in practice. An alternative causal fairness notion that does not need such knowledge, capture at the same time group-level fairness and can be easily tested on data, is *interventional fairness* (Salimi et al. 2019b): besides the sensitive attribute and the outcome variable, it relies on a set of admissible variables through which it is admitted for the protected attribute to influence the outcome.

There is currently no consensus as to which classes of fairness constraint, and which specific formulations, are appropriate for the various data processing tasks. Rather, a lot of ongoing work is devoted to understanding the relationships between the various definitions and their usage inside specific data processing tasks (see, e.g., Salimi et al. 2019a; Asudeh 2019; Zehlike et al. 2021).

3.2 Diversity

Diversity is a general term used to capture the quality of a collection of items with respect to the variety of its constituent elements. Like fairness, diversity is inherently a sociotechnical concept that gives rise to a multitude of interpretations and has an important impact from an ethic-point of view; indeed, a lack of diversity can lead to exclusion. Diversity has been extensively considered in the contexts of information retrieval and content recommendation, with the aim of designing models and algorithms enforcing diversity in the output of an algorithmic task (see Kaminskas and Bridge 2017 for a survey). Only recently such property has been considered in the context of other processing tasks.

Given a set of N elements, obtained as output of a given processing task, a number k of elements to be selected, and a function quantifying diversity between elements, diversity aims at selecting the k elements out of the N that maximize such function. Diversity can be defined in terms of different types of measures to be computed over a set, usually obtained as result of a data processing task. According to Drosou et al. (2017), we can distinguish distance-based diversity, novelty-based diversity, and coverage-based diversity. *Distance-based diversity* relies on a pairwise distance or similarity measure between elements and the diversity of a set is the aggregate (usually the average or the minimum) distance value among its pairs (Agrawal et al. 2009). The problem here is how to identify the right distance function, which also has an impact on the performance of diversity computation. When the diversity function is a metric, efficient approximation algorithms exist for the diversification problem. *Coverage-based diversity* relies on the existence of a predefined number of aspects and measures the extent to which the elements of the set cover the considered aspects (Agrawal et al. 2009; Clarke et al. 2008). *Novelty-based diversity* aims at reducing redundancy and is defined with respect to the elements seen in the past. The elements of the result set are selected one at the time, maximizing the distance-based diversity between the new element and those selected in the past (Lathia et al. 2010). Popularity and serendipity are notions related to novelty: novel elements are defined as the most unpopular (and thus, probably not seen in the past) (Ziegler et al. 2005) and as the most unusual or surprising elements (Herlocker et al. 2004), respectively.

3.3 Fairness and diversity: two definitions, one objective

While fairness is generally linked to the notion of lack of bias, diversity refers to the degree to which different kinds of objects are represented in a dataset. Even if they allow the representation of different kinds of nondiscrimination constraints, sometimes fairness and diversity can lead to the generation of similar results. For example, statistical parity is a fairness objective, but it can also be interpreted according to diversity since, like diversity, it is stated as a property of the value distribution inside a collection of items. Fairness and diversity are however slightly different. Consider for example a population including 10% Italian and 90% non-Italian individuals. While group fairness will preserve the same rate in the output obtained by a processing task, diversity might require the same result to contain 50% Italian individuals and 50% non-Italian.

Coverage is another example of nondiscrimination constraint that recently received attention and that can be associated with both fairness and diversity (Asudeh et al. 2019b). It has been initially introduced in the context of diversity with the aim of re-balancing the distribution of categories in a task outcome. However, it is also related to the concept of intersectional fairness (Chouldechova 2017): coverage constraints can be used to avoid an under-representation of protected categories of interest in a dataset, defined in terms of one or many sensitive attributes, possibly introducing bias in following analyses, by specifying how many items of a given protected category should be available inside the result of a data processing step. Lack of coverage in a dataset opens the door to adversarial attacks (Biggio et al. 2013): poorly covered regions in the training dataset provide an adversary with opportunities to create examples that are misclassified by a trained model.

Table 1 The Scopus search query

Field	Keyword search
TITLE	(((ethic* OR *discrimination OR responsibl* OR equity OR fair* OR *coverage OR divers* OR bias) AND ("data quality" OR "big data" OR "data science" OR dataset OR database OR "data management" OR "data engineering" OR "data preparation" OR "*prep" OR "data pre-processing" OR "data preprocessing" OR "data processing" OR "data wrangling" OR "data transformation" OR rewriting OR (data integration) OR "data cleaning" OR "database repair" OR *rank* OR olap OR analytics OR "data analysis" OR pipeline OR "data-driven application" OR rewriting OR "data acquisition" OR "decision support")) OR ("data equity" OR (data responsibly) OR "data coverage" OR "diversity constraint" OR (divers* *fair*) OR (bias AND (fair* OR *coverage OR divers*)))
REF	(*fair*)
PUBYEAR	(PUBYEAR > 2015)
SUBJAREA	(SUBJAREA(COMP) OR SRCTITLE(conference on information knowledge management)

4 Contributions of data processing tasks to nondiscrimination

To investigate nondiscrimination issues in the context of the main data processing tasks, we conducted a literature review over Scopus, a widely used search engine for literature review. To make the search effective, we combined conditions on the paper titles, abstracts, references, publication years, and subject areas as follows (see Table 1 for the precise query specification):

- Paper titles include one general keyword related to nondiscrimination and one more specific keyword related to ADS pipelines. Additionally, papers containing specific combinations of those words, relevant for the considered field, are included. This made the search more effective since most papers dealing with nondiscrimination in machine learning and AI do not satisfy such conditions.
- Paper references include a fairness-related keyword: this helps in excluding papers using the specified keywords under different semantics, unrelated to nondiscrimination.
- Papers have been published from 2016, year in which the issue of nondiscrimination in data management has been first identified (Abiteboul et al. 2016).
- Only papers of the computer science area are considered since this is the reference field of our research. A single exception is related to papers appearing in the proceedings of the Int. Conf. on Information and Knowledge Management (CIKM), since it is a relevant computer science conference, classified in a different way in Scopus.

This initial search returned 335 papers.[3] The papers have then been carefully inspected, looking at their title, abstract, and context, with the aim of excluding those that either: (1) include the search keywords under a meaning unrelated to nondiscrimination; (2) do not refer to the tasks introduced in Sect. 2; or (3) propose solutions based on supervised approaches, more tailored to the machine learning context. After this step, we obtained 61 papers, further classified into: (1) seminal papers, general architectures, and surveys (G); (2) papers related to data acquisition (A); (3) papers related to data cleaning, integration, and wrangling (W); (4) papers related to analytical queries (Q); (5) papers related to techniques, demos, and systems for the analysis of (portion of) analytical pipelines (P).

The number of papers for each considered group and publication year is presented in Table 2. Each group will be discussed in the following.

4.1 Seminal papers, general architectures, and surveys

Among the retrieved general papers, 6 out of 17 deal with ethic problems in data science at a very high level. The other 11 address this issue from a more technical point of view and are briefly discussed in what follows.

One of the first attempt to draw the attention of the data management community to the various facets of responsibility was a tutorial proposed at EDBT 2016 (Stoyanovich et al. 2016). Right after, in Stoyanovich et al. (2017), fairness (but also accountability and transparency) properties are advocated to be considered as database system issues, since bias may be introduced at any processing steps. The *Fides*

[3] The list of papers and their classification are available at https://bit.ly/fairness_scopus_search.

Table 2 Paper distribution with respect to the data processing task and the publication year

	2016	2017	2018	2019	2020	2021	
G	1	3	1	7	2	3	17
A	0	0	0	2	3	3	8
W	0	0	0	2	2	4	8
Q	0	1	3	4	2	6	16
P	0	1	1	4	4	2	12
	1	5	5	19	13	18	61

platform was proposed with features to encourage (and, in some cases, enforce) best practices at all stages of the data science lifecycle. In the same year, the role of diversity on Big Data management ethics was discussed in Drosou et al. (2017).

In the next years, many special events of the main data management conferences were devoted to this issue (Stoyanovich et al. 2018a) and many further papers have been published (Stoyanovich 2019; Abiteboul and Stoyanovich 2019; Firmani et al. 2019a; Jagadish et al. 2021). The considered ethic-related properties can be interpreted as special social-minded dimensions for the more general data quality issue: this is the focus of Firmani et al. (2019b), Pitoura (2020).

A very recent survey on machine learning and data management approaches for measuring and mitigating bias in data-driven decision support systems is presented in Balayn et al. (2021).

4.2 Data acquisition

Fairness can be considered during data acquisition to guarantee to start the processing with a dataset that does not lead to bias. A specific data management approach is *repairing*, i.e., modifying, the input dataset so that the new dataset satisfies the considered fairness constraints and the distance between the two datasets is minimized.

Causal fairness, and specifically interventional fairness, has been considered for repairing datasets to be used by classifiers in Salimi et al. (2019a, 2020), Getoor (2020). The repaired training dataset can be seen as a sample from a *hypothetical fair world* in which the effect of any discriminatory causal relationship between the sensitive attribute and the classifier outcome is removed.

Data repair solutions based on coverage constraints have been first introduced in Asudeh et al. (2019b). Specifically, efficient techniques for determining the least amount of additional data to be collected for guaranteeing coverage with respect to multiple sensitive attributes are proposed. An efficient approach for coverage analysis, given a set of attributes across multiple tables, is presented in Lin et al. (2020). The previous proposals are limited to categorical attributes with low-cardinality. In Asudeh et al. (2021), the coverage-based data repair problem is addressed by considering ordinal and continuous-valued attributes.

An alternative approach to detect and correct biases and discrimination in datasets exploits the notion of functional dependency, a particular type of constraint on the data, to recognize cases where the value of a certain attribute (e.g., gender, ethnicity or religion) frequently determines the value of another one (such as the range of the proposed salary or the social state) (Azzalini et al. 2021a, b).

4.3 Data cleaning, integration, and wrangling

Fairness-enhancing data cleaning interventions have been considered in Tae et al. (2019), that mitigate unfairness during data sanitization, considering demographic parity as the reference nondiscrimination constraint.

Fairness has been considered in data wrangling, in the context of an approach for the automatic identification of ways for integrating the data, in Mazilu et al. (2020, 2021). They consider two potential sources of dataset bias: those arising from unequal representation of sensitive groups and those arising from hidden biases through proxies for sensitive attributes. Both proposals analyze problems that may arise during data wrangling and lead to bias in downstream analyses and propose an approach to respond to them in a system automating the generation of data wrangling pipelines. Discriminatory bias has been considered in Yan and Howe (2021), where a learning approach to generate integrated representations (EquiTensors) of heterogeneous datasets is proposed and adversarial learning is used to remove correlations with a sensitive attribute. The impact of widely adopted data preparation procedures and of the sensitive attribute usage on the fairness of machine learning approaches is further considered in Valentim et al. (2019).

Coverage-based data transformations are considered in Accinelli et al. (2020, 2021b). In this case, the focus is on back-end transformations defined in terms of a Select-Project-Join query, whose result violates coverage constraints. In this case, the transformation is rewritten into the "closest" one satisfying those constraints. Coverage is also considered in Nargesian et al. (2021), investigating how to acquire, in the most cost-effective manner, new data for integration

when the desired distribution requirements are not satisfied by the dataset at hand.

4.4 Analytical queries

Rankings are at the basis of many important decision processes and have a potentially enormous impact on the livelihood and well-being of individuals. Thus, most of the proposed analytical querying approaches taking nondiscrimination into account are ranking approaches. They have been recently surveyed in Pitoura et al. (2021a) and tutorials have been proposed in Pitoura et al. (2020, 2021b), demonstrating the research activeness in the area. Non-discriminatory ranking approaches address many issues: (1) the design of ranking schemes (Yang and Stoyanovich 2017; Asudeh et al. 2019a; Yang et al. 2020; Kuhlman et al. 2019, 2021; García-Soriano and Bonchi, 2021); (2) the design of ranking schemes for specific domains (e.g., online job marketplaces) (Elbassuoni et al. 2019); (3) the design of approaches for intervening on the ranked outcome (Celis et al. 2018; Yang et al. 2019). Most non-discriminating ranking approaches consider group and associational fairness. Recently, there was an interest on causally fair ranking schemes (Yang et al. 2020) and coverage-based diversity (Celis et al. 2018; Yang et al. 2019).

Ethic-based set selection guarantees that the selected set satisfies specific nondiscrimination constraints. Specifically, coverage-based diversity and group fairness constraints have been considered in Stoyanovich et al. (2018b) whereas (Moumoulidou et al. 2021) focuses on maximizing diversity in set selection, while offering fairness guarantees.

In the context of OLAP queries, causal fairness has been considered for detecting bias in OLAP queries and limiting it through rewriting (Salimi et al. 2018b). Vázquez-Ingelmo et al. (2020) focus on the role of visual tools in assisting decision-making processes and raising awareness regarding potential data issues.

4.5 Pipelines and systems

One important issue in ethic-based data processing concerns the effective and efficient use of existing ethic-based approaches inside complex data processing pipelines, e.g., those provided by data processing environments like Pandas, scikit-learn, and Tableau.

In this respect, a framework for evaluating different types of fairness guarantees for pipelines is proposed in Dwork et al. (2020) while in Biswas and Rajan (2021), the impact of fairness on pre-processing stages in ML pipelines and, through composition, on the global fairness of the pipelines is investigated.

Many systems have also been developed for detecting nondiscrimination along the data processing pipeline. Among them, the open-source Python toolkit for algorithmic fairness, AI Fairness 360 (Bellamy et al. 2019), and FairTest (Tramer et al. 2017) support the user in checking algorithmic fairness and associations between application outcomes (such as prices or premiums) and sensitive user attributes (such as race or gender) with a debugging focus. Other systems refer to specific data processing tasks, prototyping many techniques discussed in the previous sections:

- *Data acquisition*: MithraLabel, providing a user with information, in the form of "nutritional labels", helping in determining the fitness of the dataset for the task at hand (Sun et al. 2019); MithraCoverage, investigating population bias in terms of coverage over the intersection of multiple attributes (Jin et al. 2020).
- *Data cleaning, integration, and wrangling*: FairPrep (Schelter et al. 2020), an environment for investigating the impact of fairness-enhancing interventions inside data processing pipelines, with a special reference to data cleaning; covRew (Accinelli et al. 2021a), a Python toolkit for pre-processing pipeline rewriting ensuring coverage constraint satisfaction.
- *Analytical* queries: HypDB, detecting, explaining, and resolving bias in decision-support queries (Salimi et al. 2018a); FairSight, a visual analytic system designed to achieve different notions of fairness in ranking decisions (Ahn and Lin 2019); FairRank, an interactive system to explore fairness of ranking in online job marketplaces (Ghizzawi et al. 2019); MithraRanking, a system for interactive ranking design, analysis, and repair (Guan et al. 2019).

5 Discussion

To answer the research questions RQ1 and RQ2, pointed out in Sect. 1, we grouped the 61 considered papers with respect to the publication year (Table 2) and the geographical locations of the authors (Table 3).

Table 2 shows that in the first years of the considered period (up to 2019), most contributions refer to either general papers (seminal papers, papers describing general architectures, and surveys), due to the need of positioning the research area inside the data management community, or analytical queries, due to their relationship with tasks already investigated in other areas (e.g., recommender systems). More recently, the number of papers proposing specific discrimination-aware technical solutions, related to single data processing tasks or the whole pipeline, has increased.

The community analysis (RQ2) results in two main findings. First, only few research communities are currently active in the considered research area. Indeed, as shown in

Table 3 Paper distribution with respect to the author geographical location and data processing task

COUNTRY	# PAPERS
United States	37
Italy	8
Greece	5
France	4
United Kingdom	4
Finland	3
Germany	2
Lebanon	2
Portugal	2
Switzerland	2
Austria	1
Netherlands	1
South Korea	1
Spain	1
Undefined	1

	G	A	W	Q	P	
North America	10	6	2	9	10	**37**
Europe	13	1	5	11	3	**33**
Others	0	1	1	1	1	**4**

Table 3, only 6 countries, in the considered period, have contributed to publishing at least 3 papers. Second, most of such communities are in the United States. As a consequence, the developed solutions often rely on specific US-tailored laws and case studies (e.g., employment, specific steps of the US judicial system); less approaches have been designed starting from European case studies even if the EU GDPR now calls for such kind of proposals (Bonatti and Kirrane 2019).

In this respect, our group at the University of Genoa has recently started a project aiming at proposing responsible data processing pipelines, with a special reference to data wrangling, in the higher education domain, also relying on data generated from online learning activities. The education context is just one possible example but, to design effective responsible ADS approaches, additional reference domains and real-world scenarios are needed: we hope that further data management communities, with a special reference to Europe, will invest resources in this relevant field soon.

Funding Open access funding provided by Università degli Studi di Genova within the CRUI-CARE Agreement.

Availability of data and material Result of the Scopus search discussed in Sect. 4 and paper classification: https://bit.ly/fairness_scopus_search.

Code availability Not applicable.

Declarations

Conflicts of interest Not applicable.

References

Abiteboul S, Stoyanovich J (2019) Transparency, fairness, data protection, neutrality: data management challenges in the face of new regulation. J Data Inf Qual 11(3):1–9

Abiteboul S, Arenas M, Barceló P, Bienvenu M, Calvanese D, David C, Schwentick M et al (2016) Research directions for principles of data management (abridged). SIGMOD Rec 45(4):5–17

Accinelli C, Minisi S, Catania B (2020) Coverage-based rewriting for data preparation. In: Proceedings of the EDBT/ICDT workshops, p 2578. CEUR-WS.org

Accinelli C, Catania B, Guerrini G, Minisi S (2021a) covRew: a Python toolkit for pre-processing pipeline rewriting ensuring coverage constraint satisfaction. In: Proceedings of the international conference on extending database technology (pp 698–701). OpenProceedings.org

Accinelli C, Catania B, Guerrini G, Minisi S (2021b) The impact of rewriting on coverage constraint satisfaction. In: Proceedings of the EDBT/ICDT workshops, p 2841

Agrawal R, Gollapudi S, Halverson A, Ieong S (2009) Diversifying search results. In: Proceedings of the international conference on web search and web data mining (pp 5–14), ACM

Ahn Y, Lin Y-R (2019) Fairsight: visual analytics for fairness in decision making. IEEE Trans Visual Comput Graph 26(1):1086–1095

Asudeh A (2019) Towards responsible data-driven decision making in score-based systems. IEEE Bull 42(3):76–87
Asudeh A (2021) Enabling responsible data science in practice. In: ACM SIGMOD blog
Asudeh A, Jagadish HV, Stoyanovich J, Das G (2019a) Designing fair ranking schemes. In: Proceedings of the international conference on management of data (pp 1259–1276), ACM
Asudeh A, Jin Z, Jagadish HV (2019b) Assessing and remedying coverage for a given dataset. In: Proceedings of the international conference on data engineering (pp 554–565), IEEE
Asudeh A, Shahbazi N, Jin Z, Jagadish HV (2021) Identifying insufficient data coverage for ordinal continuous-valued attributes. In: Proceedings of the international conference on management of data (pp 129–141), ACM
Azzalini F, Criscuolo C, Tanca L (2021a) A short account of FAIR-DB: a system to discover data bias (discussion paper). In: Proceedings of the Italian symposium on advanced database systems, vol 2994, pp 192–199. CEUR-WS.org
Azzalini F, Criscuolo C, Tanca L (2021b) FAIR-DB: FunctionAl dependencIes to discoveR Data Bias. In: Proceedings of the EDBT/ICDT workshops, p 2841, CEUR-WS.org
Balayn A, Lofi C, Houben G-J (2021) Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. VLDB J 30(5):738–768
Bellamy RK et al (2019) AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 63(4/5):4:1-4:15
Biggio B, Corona I, Maiorca D, Nelson B, Srndic N, Laskov P, Roli F et al (2013) Evasion attacks against machine learning at test time. In: Proceedings of the European conference on machine learning and knowledge discovery in databases, vol 8190, pp 387–402, Springer
Biswas S, Rajan H (2021) Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In: Proceedings of the joint European software engineering conference and symposium on the foundations of software engineering (pp 981–993), ACM
Bonatti PA, Kirrane S (2019) Big Data and analytics in the age of the GDPR. In: Proceedings of the international congress on big data (pp 7–16), IEEE
Celis LE, Straszak D, Vishnoi NK (2018) Ranking with fairness constraints. In: Proceedings of the international colloquium on automata, languages, and programming, vol 107, pp 28:1–28:15. Schloss Dagstuhl—Leibniz-Zentrum für Informatik
Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163
Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: Proceedings of the international conference on research and development in information retrieval (pp 659–666), ACM
Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the international conference on knowledge discovery and data mining (pp 797–806), ACM
Doan A (2018) Human-in-the-loop data analysis: a personal perspective. In: Proceedings of the workshop on human-in-the-loop data analytics (pp 1:1–1:6), ACM
Drosou M, Jagadish HV, Pitoura E, Stoyanovich J (2017) Diversity in big data: a review. Big Data 5(2):73–84
Dwork C, Hardt M, Pitassi T, Reingold O, Zemel RS (2012) Fairness through awareness. In: Proceedings of the international conference on innovations in theoretical computer science (pp 214–226), ACM
Dwork C, Ilvento C, Jagadeesan M (2020) Individual fairness in pipelines. In: Proceedings of the international symposium on foundations of responsible computing, vol 156, pp 7:1–7:22. Schloss Dagstuhl—Leibniz-Zentrum für Informatik
Elbassuoni S, Amer-Yahia S, Atie CE, Ghizzawi A, Oualha B (2019) Exploring fairness of ranking in online job marketplaces. In: Proceedings of the international conference on extending database technology (pp 646–649). OpenProceedings.org
Firmani D, Tanca L, Torlone R (2019a) Data processing: reflections on ethics. In: Proceedings of the international workshop on processing information ethically, co-located with CAISE, p 2417. CEUR-WS.org
Firmani D, Tanca L, Torlone R (2019b) Ethical dimensions for data quality. J Data Inf Qual 12(1):21–25
García-Soriano D, Bonchi F (2021) Maxmin-fair ranking: individual fairness under group-fairness constraints. In: Proceedings of the international conference on knowledge discovery and data mining (pp 436–446), ACM
Getoor L (2020) Technical perspective: database repair meets algorithmic fairness. SIGMOD Rec 49(1):33
Ghizzawi A, Marinescu J, Elbassuoni S, Amer-Yahia S, Bisson G (2019) FaiRank: An interactive system to explore fairness of ranking in online job marketplaces. In: Proceedings of the international conference on extending database technology (pp 582–585). OpenProceedings.org
Guan Y, Asudeh A, Mayuram P, Jagadish HV, Stoyanovich J, Miklau G, Das G (2019) MithraRanking: a system for responsible ranking design. In: Proceedings of the international conference on management of data (pp 1913–1916), ACM
Gupta M, Cotter A, Fard MM, Wang S (2018) Proxy fairness. CoRR abs/1806.11212
Herlocker JL, Konstan JA, Terveen LG, Riedl J (2004) Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst 22(1):5–53
Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, Shahabi C (2014) Big data and its technical challenges. Commun ACM 57(7):86–94
Jagadish HV, Stoyanovich J, Howe B (2021) The many facets of data equity. In: Proceedings of the EDBT/ICDT workshops, p 2841. CEUR-WS.org
Jin Z, Xu M, Sun C, Asudeh A, Jagadish HV (2020) MithraCoverage: a system for investigating population bias for intersectional fairness. In: Proceedings of the international conference on management of data (pp 2721–2724), ACM
Kaminskas M, Bridge D (2017) Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Trans Interact Intell Syst 7(1):2:1-2:42
Kilbertus N, Rojas-Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B (2017) Avoiding discrimination through causal reasoning. CoRR, abs/1706.02744
Kuhlman C, Valkenburg MV, Rundensteiner EA (2019) FARE: diagnostics for fair ranking using pairwise error metrics. In: Proceedings of the world wide web conference (pp 2936–2942), ACM
Kuhlman C, Gerych W, Rundensteiner EA (2021) Measuring group advantage: A comparative study of fair ranking metrics. In: Proceedings of the international conference on AI, Ethics, and Society (pp 674–682), ACM
Kusner MJ, Loftus JR, Russell C, Silva R (2017) Counterfactual fairness. CoRR abs/1703.06856
Lathia N, Hailes S, Capra L, Amatriain X (2010) Temporal diversity in recommender systems. In: Proceeding of the international conference on research and development in information retrieval (pp 210–217), ACM

Lin Y, Guan Y, Asudeh A, Jagadish HV (2020) Identifying insufficient data coverage in databases with multiple relations. Proc VLDB Endow 13(11):2229–2242

Madhavan J, Jeffery SR, Cohen S, Dong XL, Ko D, Yu C, Halevy A (2007) Web-scale data integration: you can afford to pay as you go. In: Proceedings of the biennial conference on innovative data systems research (pp 342–350)

Mazilu L, Paton NW, Konstantinou N, Fernandes AA (2020) Fairness in data wrangling. In: Proceedings of the international conference on information reuse and integration for data science (pp 341–348), IEEE

Mazilu L, Konstantinou N, Paton NW, Fernandes AA (2021) Data wrangling for fair classification. In: Proceedings of the EDBT/ICDT workshops, vol 2841. CEUR-WS.org

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):115:1-115:35

Moumoulidou Z, McGregor A, Meliou A (2021) Diverse data selection under fairness constraints. In: Proceedings of the international conference on database theory, vol 186, pp 13:1–13:25. Schloss Dagstuhl—Leibniz-Zentrum für Informatik

Nabi R, Shpitser I (2018) Fair inference on outcomes. Proc AAAI Conf Artif Intell 32:1931–1940

Nargesian F, Asudeh A, Jagadish HV (2021) Tailoring data source distributions for fairness-aware data integration. Proc VLDB Endow 14(11):2519–2532

Pitoura E (2020) Social-minded measures of data quality: fairness, diversity, and lack of bias. ACM J Data Inf Qual 12(3):121–128

Pitoura E, Koutrika G, Stefanidis K (2020) Fairness in rankings and recommenders. In: Proceedings of the international conference on extending database technology (pp 651–654). OpenProceedings.org

Pitoura E, Stefanidis K, Koutrika G (2021a) Fairness in rankings and recommendations: an overview. VLDB J 2021:5

Pitoura E, Stefanidis K, Koutrika G (2021b) Fairness in rankings and recommenders: models, methods and research directions. In: Proceedings of the international conference on data engineering (pp 2358–2361), IEEE

Rattenbury T, Hellerstein JM, Heer J, Kandel S, Carreras C (2017) Principles of data wrangling: practical techniques for data preparation. O'Reilly Media, Inc

Salimi B, Cole C, Li P, Gehrke J, Suciu D (2018a) HypDB: a demonstration of detecting, explaining and resolving bias in OLAP queries. Proc VLDB Endow 11(12):2062–2065

Salimi B, Gehrke J, Suciu D (2018b) Bias in OLAP queries: detection, explanation, and removal. In: Proceedings of the international conference on management of data (pp 1021–1035), ACM

Salimi B, Howe B, Suciu D (2019a) Data management for causal algorithmic fairness. IEEE Data Eng Bull 42(3):24–35

Salimi B, Rodriguez L, Howe B, Suciu D (2019b) Interventional fairness: causal database repair for algorithmic fairness. In: Proceedings of the international conference on management of data (pp 793–810), ACM

Salimi B, Howe B, Suciu D (2020) Database repair meets algorithmic fairness. SIGMOD Rec 49(1):34–41

Schelter S, He Y, Khilnani J, Stoyanovich J (2020) FairPrep: promoting data to a first-class citizen in studies on fairness-enhancing interventions. In: Proc. of the international conference on extending database technology (pp 395–398)

Stoyanovich J, Abiteboul S, Miklau G (2016) Data responsibly: fairness, neutrality and transparency in data analysis. In: Proceedings of the international conference on extending database technology (pp 718–719). OpenProceedings.org

Stoyanovich J, Howe B, Abiteboul S, Miklau G, Sahuguet A, Weikum G (2017) Fides: towards a platform for responsible data science. In: Proceedings of the international conference on scientific and statistical database management (pp 26:1–26:6)

Stoyanovich J, Howe B, Jagadish HV (2018a) Special session: a technical research agenda in data ethics and responsible data management. In: Proceedings of the international conference on management of data (pp 1635–1636), ACM

Stoyanovich J, Yang K, Jagadish HV (2018b) Online set selection with fairness and diversity constraints. In: Proc. of the international conference on extending database technology (pp 241–252). OpenProceedings.org

Stoyanovich J (2019) TransFAT: translating fairness, accountability and transparency into data science practice. In: Proceedings of the international workshop on processing information ethically co-located with 31st International conference on advanced information systems engineering, p 2417. CEUR Workshop Proceedings

Stoyanovich J, Howe B, Jagadish HV (2020) Responsible data management. PVLDB 13(12):3474–3488

Sun C, Asudeh A, Jagadish HV, Howe B, Stoyanovich J (2019) MithraLabel: flexible dataset nutritional labels for responsible data science. In: Proceedings of the ACM international conference on information and knowledge management (pp 2893–2896), ACM

Tae KH, Roh Y, Oh YH, Kim H, Whang SE (2019) Data cleaning for accurate, fair, and robust models: a big data-AI integration approach. In: Proceedings of the international workshop on data management for end-to-end machine learning (pp 1–4)

Tramer F, Atlidakis V, Geambasu R, Hsu D, Hubaux J-P, Humbert M, Lin H et al (2017) Fairtest: discovering unwarranted associations in data-driven applications. In: Proceedings of the European symposium on security and privacy (pp 401–416), IEEE

Valentim I, Lourenço N, Antunes N (2019) The impact of data preparation on the fairness of software systems. In: Proceedings of the international symposium on software reliability engineering (pp 391–401), IEEE

Vázquez-Ingelmo A, García-Peñalvo FJ, Therón R (2020) Aggregation bias: a proposal to raise awareness regarding inclusion in visual analytics.In: Trends and innovations in information systems and technologies—volume 3.1161, pp 409–417, Springer

Verma S, Rubin J (2018) Fairness definitions explained. In: Proceedings of the international workshop on software fairness (pp 1–7), ACM

Yan A, Howe B (2021) EquiTensors: learning fair integrations of heterogeneous urban data. In: Proceedings of the international conference on management of data (pp 2338–2347), ACM

Yang K, Stoyanovich J (2017) Measuring fairness in ranked outputs. In: Proceedings of the international conference on scientific and statistical database management (pp 22:1–22:6), ACM

Yang K, Gkatzelis V, Stoyanovich J (2019) Balanced ranking with diversity constraints. In: Proceedings of the international joint conference on artificial intelligence (pp 6035–6042). ijcai.org

Yang K, Loftus JR, Stoyanovich J (2020) Causal intersectionality for fair ranking, CoRR, abs/2006.08688

Zehlike M, Yang K, Stoyanovich J (2021) Fairness in ranking: a survey. CoRR abs/2103.14000

Ziegler C-N, McNee SM, Konstan JA, Lausen G (2005) Improving recommendation lists through topic diversification.In: Proceedings of the international conference on World Wide Web (pp 22–32), ACM