Responsible Data Integration: Next-generation Challenges

Fatemeh Nargesian
University of Rochester
fnargesian@rochester.edu

Abolfazl Asudeh
University of Illinois at Chicago
asudeh@uic.edu

H. V. Jagadish
University of Michigan
jag@umich.edu

ABSTRACT

Data integration has been extensively studied by the data management community and is a core task in the data pre-processing step of ML pipelines. When the integrated data is used for analysis and model training, responsible data science requires addressing concerns about data quality and bias. We present a tutorial on data integration and responsibility, highlighting the existing efforts in responsible data integration along with research opportunities and challenges. In this tutorial, we encourage the community to audit data integration tasks with responsibility measures and develop integration techniques that optimize the requirements of responsible data science. We focus on three critical aspects: (1) the requirements to be considered for evaluating and auditing data integration tasks for quality and bias; (2) the data integration tasks that elicit attention to data responsibility measures and methods to satisfy these requirements; and, (3) techniques, tasks, and open problems in data integration that help achieve data responsibility.

CCS CONCEPTS

• **Information systems** → **Information integration**; • **Social and professional topics** → *User characteristics.*

KEYWORDS

data integration, responsible AI, data equity, data collection, distribution tailoring, fair ML

ACM Reference Format:
Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish. 2022. Responsible Data Integration: Next-generation Challenges. In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22), June 12–17, 2022, Philadelphia, PA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3514221.3522567

1 INTRODUCTION

AI technologies provide user-friendly solutions at a scale and efficiency that was not imaginable before. In decision making, AI can help to eliminate human bias, and to make wise decisions that benefit human beings and societies. Its many benefits have caused the AI revolution to have a huge impact on all aspects of modern human life. As AI is fusing into our lives, its potential harms have become more evident. We all have perhaps faced or heard many of these concerns. The concept of *Responsible AI* has been introduced to minimize the drawbacks of AI.

It is known that AI is as good as the data it is built on [6, 10, 36]. This has become the main focus of data-centric AI, where the goal is to collect good data rather than big data. When data does not contain enough signals to address business needs, no model can achieve a high-enough performance to address those needs [36], hence, responsible AI requires that *responsible data* be collected. Since responsible data is often scattered across multiple sources, *responsible data integration* is required for collecting the responsible data. Consider the following example in the healthcare domain.

Example 1: *Consider an AI company that would like to use Chicago health record data and build an ML model for early detection of breast cancer. The company considers building a model on an in-house data set for training the model. However, it turns out the collected data is highly skewed: due to the historical discriminatory policies, such as redlining, in the city of Chicago, racial/ethnic minorities have disproportionate (lack of) access to high-quality breast cancer care [20]. This, in turn, has resulted in the under-representation of non-white patients in the data, which needs to get resolved in responsible data collection. On the other hand, a partnership between Chicago healthcare and research institutions has been established by companies such as CAPriCORN [1] to integrate health data from multiple sources, each of which has its own skew, for reasons such as those described above. A question here is how to responsibly integrate data to find a data set that not only is complete and correct, but also contains enough informative and unbiased features for building the model, and adequately represents the minority patients.* □

Responsible AI introduces new challenges and requirements for data integration, that require revisiting different tasks in the data integration pipeline to make sure these needs are satisfied. In this tutorial, we outline (some of) these next-generation challenges, review the work to date to address these requirements, and discuss some of the open problems and opportunities to enable responsible data integration. This proposal has four parts:

- In part one, we identify a set of next-generation data requirements needed by responsible AI. These requirements provide a roadmap for this tutorial, to address them in the context of data integration.
- The second part of the tutorial describes the data integration tasks and related work. Revisiting these tasks is needed to achieve responsible data integration.
- Part three is devoted to existing work on responsible data integration. In particular, we zoom into the body of work on distribution-aware and fairness-aware measures when data is integrated from multiple sources.
- Finally, we will discuss the open problems, opportunities, and promising directions for the data community to address the challenges of responsible data integration.

SIGMOD '22, June 12–17, 2022, Philadelphia, PA, USA

ACM ISBN 978-1-4503-9249-5/22/06…$15.00
https://doi.org/10.1145/3514221.3522567

Related tutorials: A tutorial on data collection for deep learning was presented in VLDB 2020, which covers topics of fair and robust training with the assumption that model fairness improvement is usually done during model training [58]. The focus of our tutorial is on addressing responsibility issues in the data pre-processing step of an ML pipeline. Another tutorial with a different perspective was presented in SIGMOD 2021 looking at the systems' challenges of deep learning, including data storage, data movement, and cost of computation. This tutorial provides an overview of AI-responsibility concerns from a data management perspective with a broad scope [57]. The scope of our tutorial is particularly the data integration challenges under AI-responsibility constraints.

This tutorial is designed for an audience with basic data management and data science background and does not assume any background in fairness.

2 PART 1: REQUIREMENTS OF RESPONSIBLE-AI

Data is the central component of data-driven systems, as "an algorithm is only as good as the data it works with" [10]. As a result, the first step to achieve responsible AI is "*responsible data*". Consider fairness in machine learning as an example. It is true that subsequent steps of building ML models have the potential to add bias and unfairness to the outcome, but almost all the reported unfairness issues have roots in *biased data*. Furthermore, it has been shown that in the presence of biased data, it is impossible to achieve complete fairness on its different definitions [24].

Responsible AI poses new requirements for the data preparation pipeline, including data integration. In the following, we discuss a set of "next-generation requirements" to enable responsible AI[1]. Addressing these requirements in the context of data integration will be our focus in the next sections.

2.1 Underlying Distribution Representation

The standard assumption in AI and machine learning is that the data used for building models and algorithms is a representative sample of the data that will be seen in production. Formally, the assumption is that the training data is a set of *i.i.d random samples* drawn from the distribution that query points follow. This fundamental assumption, however, is not always easy to satisfy and is often violated specifically when it comes to social data. That is due to the fact that local distributions of social data often differ from the global underlying distribution; hence, data collection and integration processes can generate data that does not satisfy this assumption. For example, surveys may be sent out to a carefully chosen random sample, but only a fraction of surveys are returned, with the return rate not being completely random. Survey statistics has developed sophisticated techniques to handle such a lack of randomness [26]. Similar issues arise when analyzing online comments or tweets to gauge popular opinion. We wish that the opinions expressed be representative of the target population of interest (e.g. all voters or all customers), but we know that we only have a skewed sample with the most vocal individuals, potentially skewing young and more tech-savvy individuals. One recent example of violating the Underlying Distribution Representation assumption with significant consequences is the *Pulse Oximeters*, being less accurate for dark skin [44]. The low accuracy of such tools for dark skin is due to poor data collection from white-dominant populations, with different distribution than the entire society. This case is often called sampling bias. There are numerous examples of sampling bias, including the HP webcams that were not able to detect black faces [51] since they collected data from their "mostly-white male" employees [56], violating the Underlying Distribution Representation assumption.

In cases where the underlying distribution is known, one can use off-the-shelf techniques [29, 43] to ensure that collected data follows the distribution. However, sometimes the underlying distribution may not be known, making it challenging to verify the Underlying Distribution Representation assumption.

2.2 Group Representation

Beyond the need for Underlying Distribution Representation, it may sometimes be important to show adequate consideration of minority groups, to ensure reliable outcomes for such groups [49]. Otherwise, when the "behavior" of under-represented groups is different from the others, trained models will poorly perform for such groups. Note that this requirement is different from (and sometimes in trade-off) the Underlying Distribution Representation assumption. That is because to ensure that minority entities are adequately considered, we may need to train with data in which small minorities are intentionally over-represented [11, 19]. Similarly, when we are interested in characterizing rare events, we may need training data that has rare events over-represented. For example, to learn how to handle emergencies, we need car-driving data with accidents and near-accidents over-represented: representative driving data may involve few challenging scenarios [45].

The Group Representation requirement may require (almost) *equal representation* of different groups (demographic parity) [28, 50]. A more liberal metric is *data coverage* [7, 8, 27, 33]. Generally speaking, a group is covered by a given data set, if there are "enough" samples from that group in the data set. The uncovered region of the data set is the set of groups that are not covered by it. For non-ordinal categorical attributes, the set of uncovered patterns specifies the uncovered region [7]. For example, the uncovered pattern {`gender:female, race:black`} indicates that the data set does not contain enough black female samples. For cases where attributes of interest are ordinal, given a distance measure and a neighborhood radius, a query point q is covered if there are enough samples in the data set in the neighborhood of q [8]. The uncovered region then is the universe of query points that are not covered.

2.3 Unbiased and Informative Features

A dataset is a collection of tuples, each defined over a set of (observation) attributes, a.k.a features, $\mathbf{x} = \{x_1, \cdots, x_m\}$ that are used for decision making. The data set may also include a set of target (a.k.a label) attributes $\mathbf{y}$. In addition, responsible AI requires data to include information about the sensitive attributes to identify the demographic groups. These attributes are required to make sure the data-driven algorithms and ML models are built responsibly and

[1] In this tutorial, we focus on requirements specific to responsible AI. Other requirements such as environmental impact, while critical in general are out of the scope of this tutorial.

that they generate fair outcomes. Despite its importance, it is often challenging to collect such information. For example, consider a data set of users registered in a shopping website. It is usually the case that these websites do not require information about the race or age group of their users. In such cases, one may consider other attributes, such as name, as a "proxy" to learn the demographic information, which by itself can be problematic and add bias to data.

The performance of ML models and data-driven algorithms depends on the set of attributes a data set contains. Take classification as an example, where the objective is to predict the value of a target variable y, using attributes $\mathbf{x}$. The high accuracy of the model directly depends on how much "information" $\mathbf{x}$ contains about y and if y is learnable from it. As a result, one has to make sure that the data set contains attributes *highly correlated with* y.

Finally, to ensure fairness in downstream data science tasks, it is important to find attributes that are *not biased*, i.e., those are (almost) independent from the sensitive attributes, or at least those *minimally correlated with the sensitive attributes*. In cases where $\mathbf{x}$ is biased, the later steps of responsible AI (including in-process, or post-process techniques [9]) try to minimize their impact by debiasing those attributes or minimizing their impact in the algorithm or model outcomes. These resolutions, however, are in trade-off with algorithm performance and model accuracy. Therefore, it is important to find attributes that are not biased (minimally correlated with sensitive attributes) and at the same time informative (highly correlated with the target attributes).

2.4 Completeness and Correctness

Collecting complete and correct data has always been a critical requirement in the data processing pipeline. This requirement becomes even more critical for responsible AI. That is because incomplete and incorrect data typically hurt minorities, further increasing the data bias in such cases. To see why, consider a data set with two groups, where most tuples belong to the majority group while a small portion is from the minority group(s). Now consider, for example, an AVG operation over a specific attribute of the data set. An incorrect value in one majority tuple does not significantly impact the value of the average, but it may significantly change the outcome for a minority group with fewer members. Similar observations can be made for other data science tasks and ML model training. Besides correctness, completeness also becomes more critical for responsible AI. That is because the way missing value issues are resolved in downstream tasks can further increase bias in data. For example, consider two resolutions where (i) rows with missing values are removed and (ii) missing values are replaced with the column average. In (i), while removing the tuples from a majority row may not have a significant impact on the majority population in data, removing a minority row further decreases the data coverage for that minority group. In (ii) also, the data bias may get increased, since the average value is mostly affected by the tuples of the large group.

2.5 Scope-of-use Augmentation

Collecting data that fully satisfies all requirements is not often possible in practice. Additionally, some of the requirements may conflict with others. For example, a data set that fully satisfies the Underlying Distribution Representation may not fully satisfy Group Representation. In the end, every data set has a limited scope of use, and no data set is good for all tasks. As a result, to ensure *transparency*, it is important to embed data with the meta-data and information that describe its collection process, its limitations, and its fitness for use [53]. Such meta-data, for example, should include the information about the underlying distribution the data has been collected from, existing biases both on the groups it fails to represent and its features that are biased, as well as the information related to correctness and completeness of data.

3 PART 2: REVISITING DATA INTEGRATION

Satisfying the requirements of § 2 in data obtained from integration introduces new challenges, which require revisiting different data integration tasks. In this part, we describe the integration tasks together with some of the related works, old and new, that ought to be revisited to develop the piece-part technologies needed to meet the responsibility requirements.

3.1 Data Set Discovery

In data-centric AI, the focus is on collecting and improving the data to improve model accuracy. For data collection, data discovery techniques can be used to discover and augment data sets. With the popularity of data lakes, data set discovery has gained interest in the data management community. Data set discovery is normally formulated as a search problem. In one version of the problem, the query is a set of keywords and the goal is to find tables relevant to the keywords in an IR-style of search [14]. Alternatively, the query can be a table and the problem is to find other tables that can be integrated with the query table with union and join operations [12, 15, 22, 23, 39, 61, 62]. A complementary alternative to the point-query style of search is navigation in a hierarchical structure or a linkage graph [22, 38, 42].

The new generation of data set discovery techniques focuses on feature discovery to improve ML models by using distribution-aware measures such as join-correlation [47]. Given a target column (containing class labels) and a join column from a query table, the goal is to retrieve candidates from a repository such that a candidate table is joinable with the query on the join column and contains a column (feature) that is correlated with the target column. Correlation measures, such as mutual information, are evaluated on the sketches built from the random samples of data sets.

In addition to the research on how to efficiently search, recent work studies what queries to ask. Tae et al. propose a way of identifying problematic slices and selectively acquiring the right amount of data for slices of data that cause bias [55].

Data set discovery is the first step towards finding informative tuples and features (Unbiased and Informative Features). Since there often does not exist one particular source that satisfies the required distribution, data discovery enables collecting sources for integration to tailor data sets that satisfy the **Underlying Distribution Representation** and **Group Representation** requirements.

3.2 Data Profiling

The body of work on data profiling [2] and more specifically nutritional labels [53] that ensure transparency by including fairness-aware fields and widgets in meta-data are steps taken for satisfying the **Scope-of-use Augmentation** requirement. MithraLabel augments the traditional profiling information with information about the fitness of a data set for responsible data science [53]. Such information includes correlation between attributes, functional dependencies between sensitive attributes and target variables, association rules to capture bias, maximal uncovered patterns (MUPs) [7] to identify under-represented subgroups, attributes with maximum/minimum demographic parity with respect to sensitive attributes, and the most diverse attributes on demographic groups. In the same line of work, Datasheets proposes that every ML data set be accompanied with a datasheet that documents its collection process and recommended uses, to increase transparency and accountability and facilitate reproducibility [25]. More concretely, datasheets consist of a set of questions on the collection process and motivation of creation that aim to cover the needs of data set consumers and producers.

3.3 Data Cleaning

The rich body of work in data cleaning has a lot to offer to the process of obtaining complete and correct data sets and satisfying **Completeness and Correctness**. To build robust, fair, and clean models, recent works focus on the data pre-processing step in the ML pipeline [48, 54]. To enable the best practices of ML experimentation, FairPrep proposes a design and evaluation framework for fairness-enhancing interventions [48]. In particular, FairPrep is concerned with extending the data processing pipeline with fairness-specific evaluation metrics as well as quantifying and validating the effects of fairness-enhancing interventions.

3.4 Uniform and Independent Sampling

Random sampling is widely applied in statistical analysis over large data sets. A random sample of size k is selected from one data set such that each element in the underlying population is picked with equal probability and the draw procedure is done k times. Random sampling is non-trivial when the underlying distribution is not known explicitly, which is often the case when data from multiple sources in the wild need to be integrated.

In the data management community, sampling is mostly studied for the result of join to ensure the **Underlying Distribution Representation** requirement needed for approximate query answering. Join operations are inherently expensive. Luckily, in many applications, a random sample from the full result of a join suffices to do analyses such as approximate query processing (estimating COUNT, SUM, AVG, medians and quantiles aggregates), statistical inference, clustering, etc.

Initially, two seminal works [5, 18] proposed the problem of random sampling from join. The main observation at the time was that sampling cannot be pushed down in join, $sample(R) \bowtie sample(S) \neq sample(R \bowtie S)$. This implies that if independent samples are taken from data sets, then they are joined, the result is indeed uniformly sampled from the full join, but the tuples that end up in the result of join are in fact highly correlated.

To obtain uniform and independent samples, Chaudhuri et al. proposed an accept-reject strategy that first samples a tuple from one data set then uniformly samples a tuple from all tuples in the second data set that can be joined with the first sampled tuples. Finally, the join of samples is returned with a probability estimated based on value frequencies in the second data set, otherwise rejected [18, 40]. This problem was originally studied for joins on two data sets [18]. Later, two works followed up on the problem. Ripple join studies on how to obtain a random but non-independent sample and use it for online aggregation (estimating aggregates such as COUNT, SUM, and AVG) [35]. Wander join obtains independent but non-uniform samples from the chain, cyclic, and acyclic join paths. The samples can again be used for estimating aggregates [30].

Recently, Zhao et al. proposed a general sampling framework for uniform and independent sampling that handles a broad family of joins (multi-way, acyclic, and cyclic) [60]. This framework can be instantiated with various distributions for value frequencies depending on the available information about the underlying data. For example, a special instantiation of this framework is the work by Chaudhuri et al. that assumes prior knowledge about frequencies and joinability. The main consideration of this framework is the trade-off between latency and throughput.

4 PART 3: DISTRIBUTION/FAIRNESS-AWARE DATA INTEGRATION

In this part, we zoom into the works explicitly designed for *data collection* with distribution-aware and fairness-aware measures in mind, particularly those related to satisfying the **Group Representation** requirement when sampling data from one data source or integrating data from multiple sources. In § 5, we discuss the research opportunities in a relevant topic, i.e. unbiased query answering over the data that does not satisfy the **Underlying Distribution Representation** requirement.

The rapid increase of the number and variety of data sources available on the web and data portals has made secondary data analysis attractive to data scientists. As such, data scientists often rely on secondary data that have been collected previously for some other purpose to fulfill the data responsibility requirements. However, since each data source is collected in some manner over some population, it will have its own distribution, which may differ from the desired distribution. The question to ask then is how data from multiple sources can be integrated to achieve the desired distribution. This has been the central problem studied in the data tailoring line of work [34, 37] as well as crowd-sourced entity collection [16, 17, 21].

4.1 Entity Collection

In crowd-sourced entity collection, the crowd is asked to complete missing data (specific attributes or entities) in a data set or a knowledge base. The challenge of crowd-sourced data collection is the open-world nature of crowd-sourcing. As such, users define distribution requirements on the entities collected from the crowd. For example, in a crowd-sourced point-of-interest (POI) collection, the desired distribution is that POIs are evenly distributed in an area [16, 21].

In distribution-aware crowd-sourced entity collection, given distribution on an attribute, the goal is to collect a set of entities via crowd-sourcing such that the difference of the distribution of collected entities from the expected distribution is minimized. Since the distribution of entities submitted by crowd workers is not known apriori, Fan et al. propose an approach that iterates between worker selection and entity distribution estimation [21]. An adaptive worker selection approach is proposed to estimate the underlying entity distribution of workers on the fly based on the collected entities then adaptively selecting the optimal set of workers that minimizes the difference (e.g. Kullback-Leibler (KL) divergence) between the expected distribution and current distribution. Distribution adjustment is done once workers submit their answers. For each worker, a statistical method is developed to estimate its particular underlying entity distribution based on the worker's history of collected entities so far. Unlike the literature of cost-effective crowd-sourced entity resolution [17], distribution-aware crowd-sourcing is agnostic to the cost of using the crowd.

4.2 Distribution Tailoring

In data distribution tailoring (DT), the goal is to enable integrating data from multiple sources to construct a target data set that follows the desired distribution [37]. The DT problem originally considers group distribution requirements on data sets defined in terms of (minimum) counts of samples from different groups [7]. In DT, a user query consists of a target schema, consisting of a collection of attributes, and group distribution requirements over groups. For example, groups can be identified as the intersection of domain values in some sensitive attributes in the target schema. Sources can be external, accessible through limited interfaces or APIs, or data views that are the outcome of the discovery and integration over data lakes. Each source is associated with a cost for sampling either monetary or in the form of computation, memory access, or network access cost. During distribution tailoring, different sources are queried in a sequential manner, in order to collect samples that fulfill the input count description, while the expected total query cost is minimized [37]. Depending on the availability of knowledge about the data source distributions, two versions of DT have been considered. Assuming the availability of group distributions, the task is to select an optimal data source to query each time based upon the set of tuples that have been already acquired. When group count aggregates are unknown and distribution estimation is expensive, solving the DT problem requires source selection without explicitly learning distributions. This is done by balancing the trade-off of exploration and exploitation using customized reward functions.

Li et al. consider the distribution tailoring problem in a data market setting where a consumer queries one data provider for data to enhance the accuracy of an ML model [32]. The assumption here is that the provider maintains a collection of data that follows the same distribution as the target distribution (the conditional probability distribution of the feature space and class labels), but the distribution is invisible to the consumer. The consumer has an initial ML model trained on some data that is not representative. The consumer is also restricted to a budget in terms of the number of records that can be queried from the provider. Upon receiving a query the provider selects a random sample without replacement from the result of the query. The goal of the consumer is to issue an optimal series of queries by adjusting filtering predicates, such that the collected data incurs accuracy improvement. Li et al. define the utility of a predicate as the anticipated accuracy improvement that the result of the query brings to the model. To quantify predicate utility, the notion of novelty is used which measures the difference between the result of the query and the data that the consumer currently possesses. The higher the difference, the more information the predicate incurs. Li et al. propose strategies that consider the exploration-exploitation trade-off. During exploration, data is obtained to learn the distribution of the provider's data and during exploitation, the query predicates are optimized based on the existing knowledge. Moreover, in the data marketplace setting, Li et al. aim at finding tuples from a join graph of a collection of data sets such that the join result of tuples incurs high correlation between the desired attributes while optimizing data quality and budget [31].

5 OPPORTUNITIES

Data responsibility has become an important topic in data integration and data collection within the data management community. While there is some excellent work, as described above, there remain many challenges yet to address. In the following, we highlight some of the many contributions the database community can make in the general area of responsible data integration.

Data Cleaning: Completeness and correctness of data is an important requirement for data-driven decision making and model training. Removing bias from data can be viewed as a special case of data cleaning where the goal is to repair problematic tuples or values that cause bias [46]. Moreover, the community has a lot to offer in auditing the existing cleaning techniques and coming up with task-specific fairness measures. For example, the new trend of data cleaning (including entity resolution) leverage pre-trained models and injects expert knowledge into models to improve the cleaning task. Since the bias in these external sources can potentially introduce bias in the linked data, fairness-aware measures can potentially pinpoint the root cause of bias in the cleaning process. The cleaning techniques are not themselves safe from data bias. The existence of missing values in a data set can lead to biased findings and deteriorate the performance of data analytics. Zhang and Long propose a novel notion (imputation accuracy parity difference) that measures the fairness of imputation results across sensitive groups [59]. This work offers insights on the connection between missing data and sample imbalance with the unfairness of imputation and prediction.

Interpretability and Transparency: There have been some efforts in annotating and reusing data-processing pipelines [13, 52]. From the system building perspective, incorporating these functionalities within data profiles in data science platforms is an important step towards improving the transparency of data integration pipelines.

Distribution Tailoring on Data Lakes: The DT problem proposes a way of collecting data from homogeneous sources that have almost similar schemas. The problem becomes more interesting

when the source of data is a data lake containing heterogeneous data sets. The ultimate goal of DT is an end-to-end system for discovering and integrating data from data lakes, in a cost-effective manner, into a data set that meets user-provided schema and distribution requirements. The pipeline includes a specialized and distribution-aware index structure for discovering sensitive attributes, efficient sampling of sources, and cost-effective and scalable integration.

Extensions of Distribution Tailoring: The distribution tailoring problem can be relaxed to consider sophisticated distribution requirements on groups. For example, a count requirement may be a range, i.e. as soon as the count of a group becomes equal to or greater than the lower bound of a range interval, the requirement is satisfied and the algorithm must start discarding samples of this group once the count becomes equal to the upper bound. Moreover, the count requirements may be on multiple groups individually instead of intersectional groups. For example, we may need 100 of `gender=F` and 100 of `gender=M` as well as 100 of `race=W` and 100 of `race=NW`. In the real world, data sources may or may not have overlap and it is necessary to design algorithms that optimize the integration cost, using the information about source overlaps.

Unbiased Feature Discovery: During feature discovery through join, it is important to design index structures that enable the efficient discovery of attributes that are not biased or at least are minimally correlated with the sensitive attributes, while ensuring a high correlation with target attributes.

Uniform Sampling over Data Lakes: Distribution tailoring focuses on satisfying count requirements by sampling from a collection of homogeneous sources. Besides data collection for representativeness, obtaining iid samples from data scattered in multiple heterogeneous sources enables unbiased analysis over data lakes.

Fairness-aware Query Answering: Bias in sources can be propagated into data-driven applications through the result of query answering. In the open-world query answering, the database is considered as a sample. A data scientist provides a sample; then aggregates and approximate results are calculated as if the queries were issued on the true population [41].

Recent work studies fairness-aware range query answering [50]. The fair queries are considered as *an alternative* to the initial query provided by the user, which is useful when the user specifies filtering conditions intuitively and hence is flexible to accept similar predicates that generate fair outcomes. This work proposes a declarative system that allows specifying some fairness requirements, along with range predicates, and a similarity requirement in SQL selection queries. The fairness requirements are defined over count differences between the tuples from different groups in the query outcome. It then aims to minimally change range query conditions such that the query output is fair. If the discovered range by the system is not satisfactory for users, they can change the fairness and similarity requirements, and *explore different choices* until the final result is found responsibly. A similar problem is coverage-aware query reformulation, where the goal is to minimally relax a query in order to provide data coverage for different groups [3, 4]. Despite some work in fairness-aware query answering, there is a need to fully integrate fairness-aware query answering operations into database systems, to build proper indices, and to revisit core operations such as relational join to enable efficient processing of such operations.

6 PRESENTERS

Fatemeh Nargesian is an assistant professor in the Department of Computer Science, at the University of Rochester. She got her PhD at the University of Toronto and was a research intern at IBM Watson in 2014 and 2016. Before the University of Toronto, she worked at Clinical Health and Informatics Group at McGill University. Her primary research interests are in data intelligence focused on data discovery and (fairness-aware) data integration.

Abolfazl Asudeh is an assistant professor at the Computer Science department of the University of Illinois at Chicago and the director of Innovative Data Exploration Laboratory (InDeX Lab). His research spans different aspects of Big Data Exploration and Data Science, including data management, information retrieval, and data mining, for which he finds efficient, accurate, and scalable algorithmic solutions. Responsible Data Science, Data Equity, and Algorithmic Fairness is his current research focus.

H. V. Jagadish is Bernard A Galler Collegiate Professor of Electrical Engineering and Computer Science, ACM Fellow (since 2003), AAAS Fellow (since 2018), the director of Michigan Institute for Data Science (MIDAS), and the Steering Committee Chair for MS in Data Science at the University of Michigan. Professor Jagadish served on the board of the Computing Research Association (CRA) 2009-2018 and on the board of the Very Large Database Endowment (2004-2010). He was the founding Editor-in-Chief of the Proceedings of the VLDB Endowment (2008-2014), and he serves on the CS Advisory Committee for the University of the People. In 2013, he was recognized with a Contributions Award by the ACM SIG on Management of Data, and in 2019, he was recognized by the University of Michigan with a Distinguished Faculty Award. He has developed a MOOC on "Data Science Ethics", carried by EdX, Coursera, and Futurelearn. Professor Jagadish has been studying issues of representation, diversity, fairness, transparency, and validity in the general area of Data Equity Systems.

ACKNOWLEDGEMENTS

Abolfazl Asudeh was supported in part by the National Science Foundation under grant 2107290 and by the Google research scholar award. H. V. Jagadish was supported in part by the National Science Foundation under grants 1741022 and 1934565.

REFERENCES

[1] [n.d.]. The Chicago Area Patient Centered Outcomes Research Network (CAPriCORN). https://www.capricorncdrn.org/.
[2] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. 2015. Profiling relational data: a survey. *VLDBJ* 24, 4 (2015), 557–581.
[3] Chiara Accinelli, Barbara Catania, Giovanna Guerrini, and Simone Minisi. 2021. The impact of rewriting on coverage constraint satisfaction.. In *EDBT/ICDT Workshops*.
[4] Chiara Accinelli, Simone Minisi, and Barbara Catania. 2020. Coverage-based Rewriting for Data Preparation. In *EDBT/ICDT Workshops*.
[5] Swarup Acharya, Phillip B. Gibbons, Viswanath Poosala, and Sridhar Ramaswamy. 1999. Join Synopses for Approximate Query Answering. In *SIGMOD*, Alex Delis, Christos Faloutsos, and Shahram Ghandeharizadeh (Eds.). 275–286.

[6] Abolfazl Asudeh and HV Jagadish. 2020. Fairly evaluating and scoring items in a data set. *PVLDB* 13, 12 (2020), 3445–3448.

[7] Abolfazl Asudeh, Zhongjun Jin, and H. V. Jagadish. 2019. Assessing and Remedying Coverage for a Given Dataset. In *ICDE*. 554–565.

[8] Abolfazl Asudeh, Nima Shahbazi, Zhongjun Jin, and HV Jagadish. 2021. Identifying Insufficient Data Coverage for Ordinal Continuous-Valued Attributes. In *SIGMOD*. 129–141.

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and machine learning: Limitations and opportunities. fairmlbook.org.

[10] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[11] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter* 6, 1 (2004), 20–29.

[12] Alex Bogatu, Alvaro A. A. Fernandes, Norman W. Paton, and Nikolaos Konstantinou. 2020. Dataset Discovery in Data Lakes. In *ICDE*. 709–720.

[13] Mike Brachmann, Carlos Bautista, Sonia Castelo, Su Feng, Juliana Freire, Boris Glavic, Oliver Kennedy, Heiko Mueller, Rémi Rampin, William Spoth, and Ying Yang. 2019. Data Debugging and Exploration with Vizier. In *SIGMOD*. 1877–1880.

[14] Dan Brickley, Matthew Burgess, and Natasha F. Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *WWW*. 1365–1375.

[15] Sonia Castelo, Rémi Rampin, Aécio S. R. Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. 2021. Auctus: A Dataset Search Engine for Data Discovery and Augmentation. *PVLDB* 14, 12 (2021), 2791–2794.

[16] Chengliang Chai, Ju Fan, and Guoliang Li. 2018. Incentive-Based Entity Collection Using Crowdsourcing. In *ICDE*. 341–352.

[17] Chengliang Chai, Guoliang Li, Jian Li, Dong Deng, and Jianhua Feng. 2016. Cost-Effective Crowdsourced Entity Resolution: A Partial-Order Approach. In *SIGMOD*. 969–984.

[18] Surajit Chaudhuri, Rajeev Motwani, and Vivek R. Narasayya. 1999. On Random Sampling over Joins. In *SIGMOD*, Alex Delis, Christos Faloutsos, and Shahram Ghandeharizadeh (Eds.). 263–274.

[19] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[20] Beti Thompson et. al. 2018. Breast cancer disparities among women in underserved communities in the USA. *Current breast cancer reports* 10, 3 (2018), 131–141.

[21] Ju Fan, Zhewei Wei, Dongxiang Zhang, Jingru Yang, and Xiaoyong Du. 2019. Distribution-Aware Crowdsourced Entity Collection. *IEEE Trans. Knowl. Data Eng.* 31, 7 (2019), 1312–1326.

[22] Raul Castro Fernandez, Essam Mansour, Abdulhakim Ali Qahtan, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and Nan Tang. 2018. Seeping Semantics: Linking Datasets Using Word Embeddings for Data Discovery. In *ICDE*. 989–1000.

[23] Raul Castro Fernandez, Jisoo Min, Demitri Nava, and Samuel Madden. 2019. Lazo: A Cardinality-Based Method for Coupled Estimation of Jaccard Similarity and Containment. In *ICDE*. 1190–1201.

[24] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[25] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[26] David Holt and David Elliot. 1991. Methods of weighting for unit non-response. *Journal of the Royal Statistical Society: Series D (The Statistician)* 40, 3 (1991), 333–342.

[27] Zhongjun Jin, Mengjing Xu, Chenkai Sun, Abolfazl Asudeh, and HV Jagadish. 2020. Mithracoverage: a system for investigating population bias for intersectional fairness. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2721–2724.

[28] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS (LIPIcs, Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 43:1–43:23.

[29] Solomon Kullback. 1987. Letter to the editor: The Kullback-Leibler distance. (1987).

[30] Feifei Li, Bin Wu, Ke Yi, and Zhuoyue Zhao. 2016. Wander Join: Online Aggregation via Random Walks. In *SIGMOD*. 615–629.

[31] Yanying Li, Haipei Sun, Boxiang Dong, and Wendy Hui Wang. 2018. Cost-efficient Data Acquisition on Online Data Marketplaces for Correlation Analysis. *PVLDB* 12, 4 (2018), 362–375.

[32] Yifan Li, Xiaohui Yu, and Nick Koudas. 2021. Data Acquisition for Improving Machine Learning Models. *PVLDB* 14, 10 (2021), 1832–1844.

[33] Yin Lin, Yifan Guan, Abolfazl Asudeh, and HV Jagadish. 2020. Identifying insufficient data coverage in databases with multiple relations. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2229–2242.

[34] Jiabin Liu, Fu Zhu, Chengliang Chai, Yuyu Luo, and Nan Tang. 2021. Automatic Data Acquisition for Deep Learning. *Proc. VLDB Endow.* 14, 12 (2021), 2739–2742.

[35] Gang Luo, Curt J. Ellmann, Peter J. Haas, and Jeffrey F. Naughton. 2002. A scalable hash ripple join algorithm. In *SIGMOD*. 252–262.

[36] Piero Molino and Christopher Ré. 2021. Declarative Machine Learning Systems. *arXiv preprint arXiv:2107.08148* (2021).

[37] Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish. 2021. Tailoring Data Source Distributions for Fairness-aware Data Integration. *Proc. VLDB Endow.* 14, 11 (2021), 2519–2532.

[38] Fatemeh Nargesian, Ken Q. Pu, Erkang Zhu, Bahar Ghadiri Bashardoost, and Renée J. Miller. 2020. Organizing Data Lakes for Navigation. In *SIGMOD*, David Maier, Rachel Pottinger, AnHai Doan, Wang-Chiew Tan, Abdussalam Alawini, and Hung Q. Ngo (Eds.). 1939–1950.

[39] Fatemeh Nargesian, Erkang Zhu, Ken Q. Pu, and Renée J. Miller. 2018. Table Union Search on Open Data. *PVLDB* 11, 7 (2018), 813–825.

[40] Frank Olken. 1993. *Random Sampling from Databases*. Ph.D. Dissertation. University of California at Berkeley.

[41] Laurel J. Orr, Magdalena Balazinska, and Dan Suciu. 2020. Sample Debiasing in the Themis Open World Database System. In *SIGMOD*. 257–268.

[42] Paul Ouellette, Aidan Sciortino, Fatemeh Nargesian, Bahar Ghadiri Bashardoost, Erkang Zhu, Ken Pu, and Renée J. Miller. 2021. RONIN: Data Lake Exploration. *PVLDB* 14, 12 (2021), 2863–2866.

[43] Leandro Pardo. 2018. *Statistical inference based on divergence measures*. CRC press.

[44] Tara Parker-Pope. Dec. 23 2020. Pulse Oximeters May Be Less Accurate for Black People. Should You Use One? The NewYork Times.

[45] Amir Bahador Parsa, Homa Taghipour, Sybil Derrible, and Abolfazl Kouros Mohammadian. 2019. Real-time accident detection: coping with imbalanced data. *Accident Analysis & Prevention* 129 (2019), 202–210.

[46] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *SIGMOD*. ACM, 793–810.

[47] Aécio S. R. Santos, Aline Bessa, Fernando Chirigati, Christopher Musco, and Juliana Freire. 2021. Correlation Sketches for Approximate Join-Correlation Queries. In *SIGMOD*, Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava (Eds.). 1531–1544.

[48] Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. 2020. FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions. In *EDBT*. 395–398.

[49] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2022. A Survey on Techniques for Identifying and Resolving Representation Bias in Data. *CoRR* abs/2203.11852 (2022). https://doi.org/10.48550/arxiv.2203.11852

[50] Suraj Shetiya, Ian Swift, Abolfazl Asudeh, and Gautam Das. 2022. Fairness-Aware Range Queries for Selecting Unbiased Data. *ICDE* (2022).

[51] Mallory Simon. 2009. HP looking into claim webcams can't see black people. CNN.

[52] William Spoth, Poonam Kumari, Oliver Kennedy, and Fatemeh Nargesian. [n.d.]. Loki: Streamlining Integration and Enrichment. In *HILDA@SIGMOD*.

[53] Chenkai Sun, Abolfazl Asudeh, H. V. Jagadish, Bill Howe, and Julia Stoyanovich. 2019. MithraLabel: Flexible Dataset Nutritional Labels for Responsible Data Science. In *CIKM*. 2893–2896.

[54] Ki Hyun Tae, Yuji Roh, Young Hun Oh, Hyunsu Kim, and Steven Euijong Whang. 2019. Data Cleaning for Accurate, Fair, and Robust Models: A Big Data - AI Integration Approach. In *DEEM@SIGMOD*. 5:1–5:4.

[55] Ki Hyun Tae and Steven Euijong Whang. 2021. Slice Tuner: A Selective Data Acquisition Framework for Accurate and Fair Machine Learning Models. In *SIGMOD*. 1771–1783.

[56] Tess Townsend. 2017. Most engineers are white and so are the faces they use to train software. Recode.

[57] Abdul Wasay, Subarna Chatterjee, and Stratos Idreos. 2021. Deep Learning: Systems and Responsibility. In *SIGMOD*. 2867–2875.

[58] Steven Whang and Jae-Gil Lee. 2020. Data Collection and Quality Challenges for Deep Learning. *PVLDB* 13, 12 (2020), 3429–3432.

[59] Yiliang Zhang and Qi Long. 2021. Assessing Fairness in the Presence of Missing Data. In *NeurIPS*.

[60] Zhuoyue Zhao, Robert Christensen, Feifei Li, Xiao Hu, and Ke Yi. 2018. Random Sampling over Joins Revisited. In *SIGMOD*. 1525–1539.

[61] Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J. Miller. 2019. JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. In *SIGMOD*. 847–864.

[62] Erkang Zhu, Fatemeh Nargesian, Ken Q. Pu, and Renée J. Miller. 2016. LSH Ensemble: Internet-Scale Domain Search. *PVLDB* 9, 12 (2016), 1185–1196.