Next-generation Challenges of Responsible Data Integration

Fatemeh Nargesian
University of Rochester
fnargesian@rochester.edu

Abolfazl Asudeh
University of Illinois 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**.

KEYWORDS

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

ACM Reference Format:
Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish. 2023. Next-generation Challenges of Responsible Data Integration. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3539597.3572727

1 INTRODUCTION

AI is as good as the data it is built on [3, 8, 29]. 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 information needs, no model can achieve a high-enough performance [29]. As a result, responsible AI requires the collection of *responsible data*. Since relevant data is often scattered across multiple sources, *responsible data integration* is required for collecting a responsible dataset. 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.

Related tutorials: This proposal is a modified version of our previously presented tutorial in SIGMOD 2022, with a focus on web and information retrieval. The Web, by providing cyber-infrastructure to remove physical barriers between people, has affected every corner of human life and society. It is rich with lakes of data that are being used for AI and need to be responsibly integrated into AI-ready data with minimal bias. 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 dxone during model training [45]. Similarly, a tutorial presented in KDD 2022, WWW 2022, and WSDM 2022 focuses on algorithmic bias [10]. A SIGMOD 2021 tutorial with a different perspective was looking at the systems' challenges of deep learning, including data storage, data movement, and cost of computation [44]. The scope of our tutorial, however, is particularly the data integration challenges under AI-responsibility constraints.

Audience and Schedule This tutorial is designed for attendees with a wide range of interests and backgrounds, including ML and web data researchers interested in knowing about data management efforts for addressing fairness concerns, as well as practitioners interested in implementing fairness-aware data algorithms for ML applications. The tutorial does not assume any background in fairness to ensure that the material is accessible to all WSDM attendees. The tutorial slides and video with linked citations and additional material will be publicly available. The tutorial consists of four main parts: 1) responsible-AI requirements (40 minutes), 2) revisiting relevant data integration topics (40 minutes), and 3) fairness-aware integration (40 minutes), and 4) open challenges and opportunities (20 minutes). In addition to a 5-minute Q&A separating different parts of the tutorial, we allocate 20 minutes to a roundtable discussion with the audience.

2 REQUIREMENTS OF RESPONSIBLE-AI

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.

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,

[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.

WSDM '23, February 27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3572727

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. In cases where the underlying distribution is known, one can use off-the-shelf techniques [25, 34] 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.

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 [38]. 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 [9, 18]. Similarly, when we are interested in characterizing rare events, we may need training data that has rare events over-represented. The Group Representation requirement may require (almost) *equal representation* of different groups (demographic parity) [24, 39]. A more liberal metric is *data coverage* [4, 6, 23, 28]. 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 [4]. 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 [6]. The uncovered region then is the universe of query points that are not covered.

Unbiased and Informative Features. A dataset is a collection of tuples, over a set of (observation) attributes, $\mathbf{x} = \{x_1, \cdots, x_m\}$ that are used for decision making. The data set may also include a set of target (a.k.a label) attributes $\mathbf{y}$. The performance of ML models and data-driven algorithms depends on the set of attributes a data set contains. 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 that they generate fair outcomes. Despite its importance, it is often challenging to collect such information. 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 [7]) try to minimize their impact by de-biasing those attributes or minimizing their impact on the model outcomes. These resolutions, however, are in trade-off with algorithm performance and model accuracy. Therefore, it is important to find attributes that are unbiased (minimally correlated with sensitive attributes) and at the same time informative (highly correlated with target attributes).

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 since incomplete and incorrect data can further increase data bias. 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.

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 satisfies the Underlying Distribution Representation may not 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 [41]. 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 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.

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 [13]. 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 [11, 14, 20, 21, 32, 47, 48]. A complementary alternative to the point-query style of search is navigation in a hierarchical structure or a linkage graph [20, 31, 33]. 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 [36]. In addition to the research on how to efficiently search, Tae et al. study identifying problematic slices and selectively acquiring the right amount of data for those slices [43]. 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.

Data Profiling. Data profiling [1] and more specifically nutritional labels [41] 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 [41]. Such information includes correlation between attributes, functional dependencies between sensitive attributes and target variables, association rules to capture bias, representation biases [4], etc. Similarly, Datasheets proposes that every ML data set be accompanied by a datasheet that documents its collection process and recommended uses, to increase transparency and accountability and facilitate reproducibility [22].

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 [37, 42]. To enable the best practices of ML experimentation, FairPrep proposes a design and evaluation framework for fairness-enhancing interventions [37]. 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.

4 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.

Entity Collection. In crowd-sourced entity collection, the crowd is asked to complete missing data 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 [15, 19]. In distribution-aware crowd-sourced entity collection, given distribution on an attribute, the goal is to collect a set of entities 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 adaptive worker selection approach to estimate the underlying entity distribution of workers on the fly. Distribution adjustment is done once workers submit their answers. Unlike cost-effective crowd-sourced entity resolution [16], distribution-aware crowd-sourcing is agnostic to the cost of using the crowd.

Distribution Tailoring. In data distribution tailoring (DT), the goal is to enable the integration of data from multiple sources to construct a target data set that follows the desired distribution [5, 30]. The DT problem originally considers group distribution requirements in terms of (minimum) counts of samples from different groups [4]. In DT, a user query consists of a target schema, consisting of a collection of attributes, and group distribution requirements. 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 [30]. Similarly, 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 [27].

5 OPPORTUNITIES

While there is some excellent related work to data responsibility, as described above, there remain many challenges yet to address. Here, we highlight some of the many contributions the WSDM community can make in the area of responsible data.

Data Cleaning: Removing bias from data can be viewed as a special case of data cleaning where the goal is to repair problematic tuples that cause bias [35]. 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. The community has a lot to offer in auditing the existing cleaning techniques and coming up with task-specific fairness measures.

Interpretability and Transparency: Existing work on annotating and reusing data-processing pipelines includes [12, 40]. From the system-building perspective, incorporating these functionalities within data profiles in data science platforms is an important step toward improving the transparency of data integration pipelines.

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.

Distribution Tailoring on Data Lakes: DT 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.

Uniform Sampling over Data Lakes: In the DB community, random sampling is mostly studied for the result of join to ensure the *Underlying Distribution Representation* requirement needed for approximate query answering [2, 17, 26, 46]. Join operations are inherently expensive. Besides data collection for group representativeness, obtaining iid samples from data scattered in multiple heterogeneous sources enables unbiased analysis over data lakes.

6 PRESENTERS

Fatemeh Nargesian is an assistant professor in the Department of Computer Science, at the University of Rochester. 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 Chicago. His research focus is on Responsible Data Science, Data Equity Systems, and

Algorithmic Fairness.

H. V. Jagadish is Bernard A Galler Collegiate Professor of Electrical Engineering and Computer Science, ACM Fellow, and AAAS Fellow. 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 2106176 and 1934565. Fatemeh Nargesian was supported in part by the National Science Foundation under grant 2107050.

REFERENCES

[1] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. 2015. Profiling relational data: a survey. *VLDBJ* 24, 4 (2015), 557–581.

[2] 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.

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

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

[5] Abolfazl Asudeh and Fatemeh Nargesian. 2022. Towards Distribution-aware Query Answering in Data Markets. *PVLDB* 15, 11 (2022).

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

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

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

[9] 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.

[10] Sarah Bird, Krishnaram Kenthapadi, Emre Kiciman, and Margaret Mitchell. 2019. Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned. In *WSDM*, J. Shane Culpepper, Alistair Moffat, Paul N. Bennett, and Kristina Lerman (Eds.). ACM, 834–835.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

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

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

[28] 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.

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

[30] 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.

[31] 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.

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

[33] 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.

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

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

[36] 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.

[37] 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.

[38] 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

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

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

[41] 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.

[42] 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.

[43] 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.

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

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

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

[47] 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.

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