

Bridging the Gap between Buyers and Sellers in Data Marketplaces with Personalized Datasets

Donatella Firmani^{1,*}, Jerin George Mathew^{1,*}, Donatello Santoro^{2,*},
Giovanni Simonini^{3,*} and Luca Zecchini^{3,*}

¹*Sapienza University of Rome, Italy*

²*University of Basilicata, Italy*

³*University of Modena and Reggio Emilia, Italy*

Abstract

Sharing, discovering, and integrating data is a crucial task and poses many challenging spots and open research direction. Data owners need to know what data consumers want and data consumers need to find datasets that are satisfactory for their tasks. Several data market platforms, or *data marketplaces* (DMs), have been used so far to facilitate data transactions between data owners and customers. However, current DMs are mostly shop windows, where customers have to rely on metadata that owners manually curate to discover useful datasets and there is no automated mechanism for owners to determine if their data could be merged with other datasets to satisfy customers' desiderata. The availability of novel artificial intelligence techniques for data management has sparked a renewed interest in proposing new DMs that stray from this conventional paradigm and overcome its limitations. This paper envisions a conceptual framework called DataStreet where DMs can create *personalized datasets* by combining available datasets and presenting summarized statistics to help users make informed decisions. In our framework, owners share some of their data with a trusted DM, and customers provide a *dataset template* to fuel content-based (rather than metadata-based) search queries. Upon each query, the DM creates a preview of the personalized dataset through a flexible use of dataset discovery, integration, and value measurement, while ensuring owners' fair treatment and preserving privacy. The previewed datasets might not be pre-defined in the DM and are finally materialized upon successful transaction.

Keywords

Data Market, Data Integration, Fairness, Privacy

1. Introduction

Data marketplaces (DMs) [1] are digital platforms where data buyers and sellers can interact to exchange data. These platforms have become increasingly popular in recent years, which saw an increasing demand for huge amounts of data by practitioners—e.g., to train their machine learning (ML) models. Conventional DMs act as storefronts for data vendors, with buyers relying only on the provided metadata, documentation, and off-the-shelf sample to determine

SEBD 2023: 31st Symposium on Advanced Database Systems, July 02-05, 2023, Galzignano Terme, Padua, Italy


*Corresponding author.

✉ donatella.firmani@uniroma1.it (D. Firmani); jeringeorge.mathew@uniroma1.it (J. G. Mathew);
donatello.santoro@unibas.it (D. Santoro); giovanni.simonini@unimore.it (G. Simonini); luca.zecchini@unimore.it
(L. Zecchini)

ORCID 0000-0003-0358-3208 (D. Firmani); 0000-0002-4626-826X (J. G. Mathew); 0000-0002-5651-8584 (D. Santoro);
0000-0002-3466-509X (G. Simonini); 0000-0002-4856-0838 (L. Zecchini)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

whether the data is useful for them. This model, which puts solely on the buyer the burden of determining the usefulness of the data and defining how to combine multiple datasets, presents significant shortcomings.

First, the discovery of valuable data is based only on metadata, documentation, and occasionally small data samples. As the content of data sources is undisclosed and not indexed, buyers cannot search by content and thus cannot determine if the data on sale aligns with their proprietary data. Secondly, the potential of the data is limited, as a single data source might not meet the buyer's need—while multiple data sources from different sellers might succeed instead. Nonetheless, there is no mechanism for sellers to securely share a portion of the data to determine beforehand if it suits the buyer's needs. Finally, potential buyers cannot test the data before buying it, making it impossible for them to explore or manipulate the data on sale to select the best data source for their needs (e.g., by computing a feature correlation matrix to determine if it would be an appropriate training set for their model).

Challenges. An enduring challenge in the DM field is the definition of a horizontal framework for implementing the operations described above to support the matching between the supply of sellers and the demand of buyers. Recently, few works addressed this issue by proposing new models that deviate from the conventional DM-as-a-shop window paradigm to reduce the time and effort required by searching and preparing datasets for both buyers and sellers, streamlining the process of price negotiation, promoting trust in data management, and facilitating data trading across sectors and countries.

Our approach. Following the aforementioned line of research that strays from the usual DM-as-a-shop window paradigm, we propose a novel conceptual framework called DataStreet, which aims at facilitating matching buyers' needs with sellers' offers. DataStreet envisions a new DM model where sellers share their data with a trusted DM, with the assurance that it will only be disclosed after an agreement with the buyer. Buyers provide the DM with a template of the dataset they wish to purchase and can specify quality metrics the dataset should meet. The DM then creates *personalized datasets* by combining sellers' datasets and presents summarized statistics to help buyers to select the dataset that best suits their needs. The concept of personalized datasets is crucial to our system and entails an integrated view over various datasets in the DM that fit with the buyer's query. These personalized datasets are only materialized upon a completed transaction, resulting in no additional storage overhead.

2. Related Work

We can identify three related lines of work that our contribution builds upon and extends:

Conventional approaches to DMs. The concept of data vendors and public data has been present since the beginning of the web; however, in recent years there has been a remarkable surge in the volume of generated and processed data. The practice of trading data has long caught the interest of both economists and the database community, with discussions on this topic going on for decades [2]. As a result, several proposals for data market platforms gained significant popularity in enabling interactions between data buyers and sellers. Some of these proposals can be regarded as *barter data markets*, in which online providers offer services (such

as email, web searches, and social networks) to users in exchange for their data. On the other hand, *broker data markets* gather anonymized customer data and sell it to marketers, allowing for more relevant ads to be delivered to consumers with better measurement (e.g., Acxiom¹, Experian²). In both cases, the market controller has a dominant position and typically retains most of the exchange benefits. Lately, online DMs have emerged [3] with the aim of enhancing the value of data for a wider range of participants. These platforms are primarily designed to connect data sellers and buyers, serving as intermediaries to facilitate data exchanges and transactions. This two-sided nature of the platforms allows organizations with a high demand for data to extract value from the available datasets that are offered for purchase.

Limitations of traditional approaches to DMs. One of the main shortcomings of current DMs (e.g., AWS Marketplace³, Dawex⁴, Snowflake⁵, etc.) is that they act as a storefront, where sellers can display their data and buyers can browse a centralized collection to find appealing datasets. Under this conventional model, the responsibility of integrating and matching relevant data purchased from multiple providers to meet the desired target is on the buyer. In other words, suppliers offer consumers “raw data”, leaving the task of effectively converting this raw data into valuable insights to the consumers, who might not dispose of the required expertise or resources. We debate that this imbalance between providers and customers is one of the primary factors restricting DM adoption. It is also important to note that several data markets are focused on specific industries (such as martech, automotive, and energy) and data types (like spatio-temporal data or data sourced from IoT sensors). These niche DMs are more difficult to locate and access compared to general-purpose DMs. In most situations, buyers might have to interact with multiple platforms to fulfill their requirements [1]. Furthermore, every transaction requires a one-on-one negotiation, resulting in increased final costs and, more significantly, making it challenging for the buyer to determine the quality of the final integrated dataset.

Recent approaches to improve DMs. Recent vision papers such as [4, 5] recognize that moving away from the DM-as-a-shopwindow paradigm is a crucial step towards the next generation of DMs. These papers emphasize the importance of data sharing, discovery, and integration, and propose an intermediary framework between sellers and buyers known as *arbiter* or *broker*. However, we note that these works have limited support in terms of (i) real-time performance, (ii) revenue sharing among sellers, (iii) reducing the effort on the consumer side, and (iv) enforcing responsibility by design. The DataStreet framework envisions the use of recent advancements in data management based on the application of artificial intelligence tools, such as [6, 7, 8], to enable the aforementioned advancements while also addressing the current limitations of these vision papers.

¹<https://www.acxiom.com>

²<https://www.experian.in>

³<https://aws.amazon.com/data-exchange>

⁴<https://www.dawex.com>

⁵<https://www.snowflake.com/en/data-cloud/marketplace>

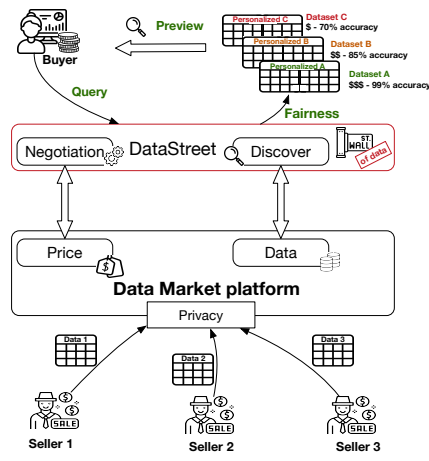


Figure 1: An overview of the DataStreet approach

3. The DataStreet Framework

For the sake of clarity, the discussion mainly focuses on structured tabular data, where the terms *attributes* and *features* are used interchangeably to refer to table columns, while the terms *records* and *instances* are used to refer to table rows. However, the proposed framework can also be adapted to semi-structured data with minor modifications. Our DM paradigm, illustrated in Figure 1, comprises four primary steps.

The first one involves sellers sharing with the broker details of their dataset features along with a selection or projection of instances. If sellers lack confidence in sharing certain parts of the dataset, they are handled separately using privacy-preserving mechanisms. The second step involves buyers submitting to the broker a query that includes a view of the desired dataset along with a combination of attribute names and preferences, which can be expressed in both structured data and natural language—e.g., a query could be the following:

TARGET SCHEMA: City, Country, Salary, Age and MortgageApproval.
 PREFERENCES: recent data, with at least half from USA and $\geq 20\%$ from New York

Buyers can also search for instances, e.g., a given city, country, salary/age range, and mortgage status, by including them in the query. The third step involves the broker returning a preview of different datasets that match the query. Each preview contains a sample of the dataset’s instances, depending on the disclosure preferences of the sellers. The datasets are ranked based on a utility function that ensures fair treatment of sellers and provides adequate economic opportunities. The fourth and final step involves each dataset having an estimated value used to determine the price. If one of the datasets satisfies the buyer, the broker proposes a fair share to all sellers: a transaction is completed if they are satisfied with the share.

It should be noted that the datasets pre-defined in the DM might not immediately match the buyer’s preferences, while multiple datasets might need to be integrated to accommodate them. Thus, in the DataStreet framework, the broker identifies all relevant datasets owned

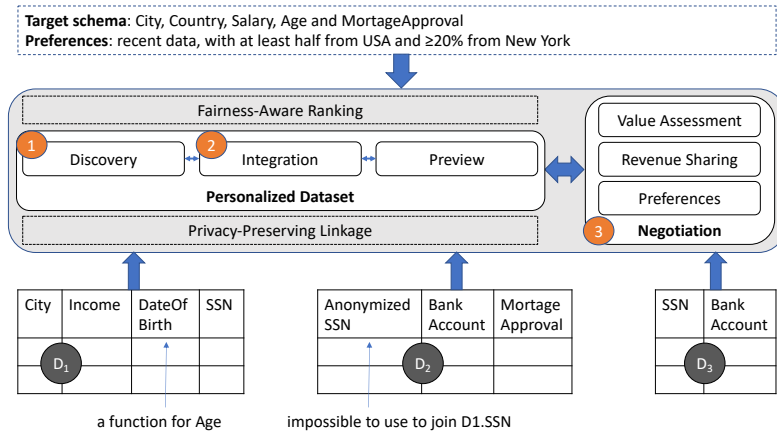


Figure 2: A conceptual architecture of the DataStreet framework

by various sellers and combines them into a preview that aligns with the buyer’s preferences. These merged datasets are referred to as *personalized datasets* and are only created upon the successful completion of a transaction, requiring no additional storage overhead. To illustrate the framework’s architecture, we present a diagram in Figure 2, along with a group of sample datasets that will be used as a reference throughout the discussion.

Figure 2 depicts a scenario where a buyer B submits a query to build an ML classifier that predicts whether a mortgage request is accepted or rejected. The label or class for each instance is represented by the MortgageApproval attribute. Additionally, the figure displays three datasets, D1, D2, and D3, belonging to three sellers, S1, S2, and S3, respectively. Each dataset has distinct attributes, with D1 containing information about City, Income, DateOfBirth, and SSN, D2 presenting the AnonymizedSSN, BankAccount, and MortgageApproval attributes, and D3 consisting of the SSN and BankAccount features.

Our conceptual framework consists of three main blocks, dubbed ① *Personalized Data Discovery*, ② *Personalized Data Integration*, and ③ *Negotiation*. The first module will discover the datasets D1 and D2 as relevant for the target schema. Features do not need to appear exactly as specified. For instance, Salary might appear under the name Income and in place of Age we might have DateOfBirth, from which we can infer the Age column. The second module will merge and integrate the datasets with the help of other datasets in the DM. For example, since the SSN attribute of D1 cannot be directly linked to the AnonymizedSSN attribute of D2, we first need to link D2 to D3 via a BankAccount attribute in order to get a SSN-MortgageApproval association. Both the BankAccount and the SSN attributes might contain errors. This module will also identify a sample of instances that can be previewed and rank the resulting personalized datasets with respect to potential other datasets. Finally, the third module will be responsible for assessing the value of the personalized dataset and sharing the revenue among the sellers. Value assessment will be based on the quality of the personalized dataset – including matching accuracy and seller reputation – and possibly on the results of testing a model trained on the dataset, if the buyer specifies so. Revenue sharing will include all the sellers contributing to the dataset, both with actual rows and columns (i.e., S1 and S2) and with auxiliary information

(e.g. S3 – in fact, none of the data owned by S3 appears in the personalized dataset). While in current DMs a buyer has to search in the catalogs of existing datasets and the effort of discovery is split between the buyer (that has to assess the relevance of each dataset) and the sellers (that have to make their datasets discoverable), DMs equipped with our framework will have a major advantage to solve the problem of matching offer and demand with low user effort, without compromising on privacy and fairness. We will now discuss the modules in more detail.

3.1. Personalized Data Discovery

In our DM model, each buyer's query has a target schema, consisting of a set of attributes that may not readily align with the datasets owned by the sellers. Attributes might appear transformed or under a different name and some of them might be owned by different sellers, hence requiring a subsequent linkage of the instances [9]. Thus, the first step in DataStreet is to automatically discover the relevant sources and to find mappings that match the target schema.

Research challenge 1. Equally valid solutions to the mapping problem can have different correspondences with the buyer's subjective preferences.

In our running example, preferences are related to the freshness of data and some attribute values. Other preferences might include a constraint on the model accuracy or the presence of specific rows in the data. Continuing the example, there might be in the DM multiple datasets with the same features as D2, dubbed D2.1, D2.2, ..., D2.n, each with a different distribution of cities, ages and labels. This module is responsible for identifying D1 and the datasets among D2.1, D2.2, ..., D2.n that when combined with D1 can result in a dataset D^* that better matches with the preferences specified in the buyer's query. In principle, this can be achieved even using simple preference criteria, such as instance diversity [10, 11]. However, enabling the framework to incorporate subjective preferences (i.e., those expressed by buyers in their own words, such as "recent data") is known to be challenging tasks. In fact, despite the existence of several research directions in data discovery [12, 6, 13, 14], which are capable of dealing with natural language ambiguity and the diverse ways in which humans represent information, none of these techniques can handle subjective data. While these methods can assist in identifying multiple relevant datasets, it is up to the end users to select a subset that meets their needs. In our proposed framework we aim to reduce human effort, which is currently beyond the scope of the existing algorithms.

3.2. Personalized Data Integration

This module is responsible of producing a ranking of datasets according to the specifications of the buyers. It is important to recall that the datasets returned by the broker may not match any data that currently exists in the DM catalog. Additionally, the datasets presented in the ranking may contain only a sample of the instances, in accordance with the data disclosure preferences of sellers. Thus, we must integrate discovered datasets from previous steps into previews.

Research challenge 2. Executing data integration on the entire raw data and subsequently applying sampling methods to generate the preview may not be computationally feasible.

This module must incorporate appropriate data integration methods that can minimize the computation time by only executing the necessary operations for generating the preview of the

personalized datasets. These methods must deliver real-time results to a buyer’s query as if the personalized datasets were already available and materialized in the DM. To achieve this, we will build on existing data integration [15, 16, 17, 18] as well as other recent techniques, such as join paths [19] to devise a custom indexing systems to ensure rapid access to a collection of potentially relevant views in the DM, while still protecting data privacy. This group of potential data views will be represented as a knowledge graph (KG) in DataStreet and will be maintained using KG completion techniques based on representation learning, such as those in [20].

Research challenge 3. Another crucial aspect in this context is preserving the privacy of the data during the linkage (i.e., integration) step.

In fact, attributes such as SSN and BankAccount in Figure 2, which serve as identifiers or quasi-identifiers in our running example, play a crucial role in locating records in different datasets. However, due to their sensitive nature, some sellers may be reluctant to share these attributes values with the DM broker. To address this issue, the *Personalized Data Integration* module will incorporate appropriate privacy-preserving record linkage (PPRL) techniques to meet the diverse privacy requirements of various sellers. PPRL techniques enable data linkage while safeguarding sensitive information against being exposed during the linkage process. Notably, there exists a vast array of PPRL methods in the literature, which leverage techniques such as cryptography [21], ML [22], and probabilistic data structures [23, 24]; however, these methods can be challenging to use and require substantial effort on the part of the data owner [25].

Finally, there can be multiple ways to generate a personalized dataset D^* that satisfies a buyer’s query, and therefore they need to be ordered based on utility criteria such as value, accuracy, and cost. Placing a dataset in a top position increases the chances of a data transaction.

Research challenge 4. Regardless of how utility is defined, ranking systems have a responsibility not only to buyers but also to sellers being ranked.

Even small differences in utility can result in significant differences in economic opportunities across seller groups. These groups can be based on demographic categories or any other category at risk of systematic discrimination. To achieve our vision, it is essential to ensure fair exposure of sellers in the DM without compromising the utility of the ranking for the buyer.

3.3. Value Assessment and Revenue Sharing

Although personalized datasets can reduce the cost of finding useful data in a DM, negotiating price and overcoming information asymmetry between buyers and sellers remains a major challenge. The value of a dataset may differ between a seller and a buyer. For instance, sellers may consider the effort they put in acquiring and preparing the data, while buyers might be interested in how much the data will enhance a process and its overall value. Negotiations can significantly affect real-world markets, and the notion of dataset value is fundamental for data discovery, integration, and ranking. Pricing data is a complex issue that has been studied in various fields such as economics, law, and business [26]; however, to the best of our knowledge, there has been a limited emphasis in a data-driven approaches to facilitate appropriate pricing.

Research challenge 5. Develop data-driven measures that enable transactions between buyers and sellers without explicitly determining the individual components of the transaction price.

In this context, we identify three main actions.

Determining the value of the data. Measures of intrinsic value, based on traditional data quality literature such as freshness and completeness [27], as well as more recent social-minded measures [28, 29], become important for D^* only if buyers specify their preferences in their query. On the other hand, extrinsic measures can accommodate different sellers' needs and expectations. Examples include the buyers' demand for a dataset and the sellers' reputation.

Defining revenue sharing schemes. If all sellers contribute to the personalized dataset D^* with instances, we could use game-theory measures such as the Shapley value [30] to allocate revenue to each row. However, this is not always the case, as shown in our example where datasets D_1 and D_2 have different identifier attributes, `SSN` and `AnonymizedSSN`, which cannot be directly linked. Therefore, we need to use a dataset from another seller, S_3 , which owns the attribute `BankAccount` and serves as an auxiliary identifier to disambiguate `SSN` and `AnonymizedSSN`. Even though S_3 does not contribute to D^* with any row or column, it plays a crucial role. To address this, our vision can be accomplished by developing sharing algorithms based on the traditional notion of provenance [31]. We will also draw upon the concepts of necessity and sufficiency from the field of explainable ML [32, 33]. In the context of DM, instances can be considered necessary if their absence results in the model's inability to produce the desired prediction, and attributes sufficiently important if they contribute to the model's ability to make accurate predictions.

4. Conclusions and Future Works

In this paper, we proposed a conceptual framework called DataStreet to address the limitations of conventional data marketplaces (DMs) by enabling personalized dataset creation through a flexible use of dataset discovery, integration, and data value measurement. Our framework allows DMs to combine sellers' data into *personalized datasets* and present summarized statistics to help buyers in making informed decisions, while also ensuring fair treatment of sellers and privacy. Personalized datasets are created upon successful transactions and may correspond to combination of multiple datasets in the DM.

The proposed framework is still in the conceptual stage, and there are several technical challenges, arising from such a novel user-centric perspective. As illustrated in Section 3, state-of-the-art techniques are designed to be applied in scenarios where the value of data artifacts is somewhat objective, as opposed to DataStreet, where the value of the data artifact is also subjective and depends on the buyers' and sellers' perspectives. We believe that such extension might enable critical human-in-the-loop interactions to ensure that the final product is tailored to the needs and requirements of all the involved parties.

Acknowledgments

This work was supported in part by the Sapienza Research grant B83C22007180001 and the SEED PNR 2021 grant FLOWER. Jerin George Mathew is financed by the Italian National PhD Program in AI.

References

- [1] S. A. Azcoitia, N. Laoutaris, A survey of data marketplaces and their business models, *ACM SIGMOD Record* (2022).
- [2] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, D. Suciu, Query-based data pricing, *JACM* (2015).
- [3] F. Schomm, F. Stahl, G. Vossen, Marketplaces for data: An initial survey, *ACM SIGMOD Record* (2013).
- [4] R. Castro Fernandez, P. Subramaniam, M. J. Franklin, Data market platforms: Trading data assets to solve data problems, *PVLDB* (2020).
- [5] J. Kennedy, P. Subramaniam, S. Galhotra, R. C. Fernandez, Revisiting online data markets in 2022, *ACM SIGMOD Record* (2022).
- [6] R. Castro Fernandez, E. Mansour, A. A. Qahtan, A. Elmagarmid, I. Ilyas, S. Madden, M. Ouzzani, M. Stonebraker, N. Tang, Seeping semantics: Linking datasets using word embeddings for data discovery, in: *ICDE*, 2018.
- [7] Y. Li, J. Li, Y. Suhara, A. Doan, W. Tan, Deep entity matching with pre-trained language models, *PVLDB* (2020).
- [8] E. K. Rezig, A. Bhandari, A. Fariha, B. Price, A. Vanterpool, V. Gadepally, M. Stonebraker, DICE: Data discovery by example, *PVLDB* (2021).
- [9] G. Mecca, G. Rull, D. Santoro, E. Teniente, Ontology-based mappings, *DKE* (2015).
- [10] A. Asudeh, F. Nargesian, Towards distribution-aware query answering in data markets, *PVLDB* (2022).
- [11] D. Firmani, M. Mecella, M. Scannapieco, C. Batini, On the meaningfulness of “big data quality”, *DSE* (2016).
- [12] A. Bogatu, A. A. Fernandes, N. W. Paton, N. Konstantinou, Dataset discovery in data lakes, in: *ICDE*, 2020.
- [13] R. Castro Fernandez, Z. Abedjan, F. Koko, G. Yuan, S. Madden, M. Stonebraker, Aurum: A data discovery system, in: *ICDE*, 2018.
- [14] A. Giuzio, G. Mecca, E. Quintarelli, M. Roveri, D. Santoro, L. Tanca, Indiana: An interactive system for assisting database exploration, *Information Systems* (2019).
- [15] J. F. Sequeda, D. P. Miranker, A pay-as-you-go methodology for ontology-based data access, *IEEE Internet Computing* (2017).
- [16] D. Firmani, B. Saha, D. Srivastava, Online entity resolution using an oracle, *PVLDB* (2016).
- [17] G. Simonini, L. Gagliardelli, S. Bergamaschi, H. V. Jagadish, Scaling entity resolution: A loosely schema-aware approach, *Information Systems* (2019).
- [18] G. Simonini, L. Zecchini, S. Bergamaschi, F. Naumann, Entity Resolution On-Demand, *Proceedings of the VLDB Endowment (PVLDB)* 15 (2022) 1506–1518. doi:10.14778/3523210.3523226.
- [19] P. Ouellette, A. Sciortino, F. Nargesian, B. G. Bashardoost, E. Zhu, K. Q. Pu, R. J. Miller, Ronin: Data lake exploration, *PVLDB* (2021).
- [20] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, P. Merialdo, Knowledge graph embedding for link prediction: A comparative analysis, *ACM TKDD* (2021).
- [21] A. Inan, M. Kantarcioglu, G. Ghinita, E. Bertino, Private record matching using differential privacy, in: *EDBT*, 2010.

- [22] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: IEEE SP, IEEE, 2017.
- [23] G. Cormode, M. Muthukrishnan, Approximating data with the count-min sketch, IEEE software (2011).
- [24] G. Cormode, D. Firmani, A unifying framework for L0-sampling algorithms, Distributed Parallel Databases (2014).
- [25] M. Franke, Z. Sehili, E. Rahm, Primat: A toolbox for fast privacy-preserving matching, PVLDB (2019).
- [26] M. Y. Carriere-Swallow, M. V. Haksar, The economics and implications of data: an integrated perspective, International Monetary Fund, 2019.
- [27] F. Geerts, G. Mecca, P. Papotti, D. Santoro, Cleaning data with llunatic, The VLDB Journal (2019).
- [28] E. Pitoura, Social-minded measures of data quality: fairness, diversity, and lack of bias, ACM JDIQ (2020).
- [29] D. Firmani, L. Tanca, R. Torlone, Ethical dimensions for data quality, ACM JDIQ (2020). URL: <https://doi.org/10.1145/3362121>. doi:10.1145/3362121.
- [30] L. S. Shapley, Quota solutions of n-person games, Technical Report, 1952.
- [31] S. B. Davidson, J. Freire, Provenance and scientific workflows: challenges and opportunities, in: SIGMOD, 2008.
- [32] D. S. Watson, L. Gultchin, A. Taly, L. Floridi, Local explanations via necessity and sufficiency: Unifying theory and practice, in: Uncertainty in Artificial Intelligence, 2021.
- [33] T. Teofili, D. Firmani, N. Koudas, V. Martello, P. Merialdo, D. Srivastava, Effective explanations for entity resolution models, in: ICDE, 2022. URL: <https://doi.org/10.1109/ICDE53745.2022.00248>. doi:10.1109/ICDE53745.2022.00248.